Kwantitatieve analyse

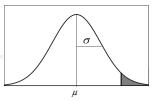
Inferentiële statistiek

Prof. dr. Jannick Demanet
Academiejaar: 2019-2020

Inferentiële statistiek

- Drie verdelingen
  - Populatieverdeling
    - Verdeling van waarden van een variabele over de eenheden van een **populatie**
  - Steekproefverdeling
    - Verdeling van waarden van een variabele over de eenheden van een **steekproef**
  - SteekproeVENverdeling
    - Verdeling van een **steekproefgrootheid**, over **alle mogelijke steekproeven** met dezelfde *n*, getrokken uit dezelfde populatie
      - (gemiddelde, fractie, chi², …)

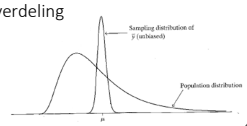
Inferentiële statistiek



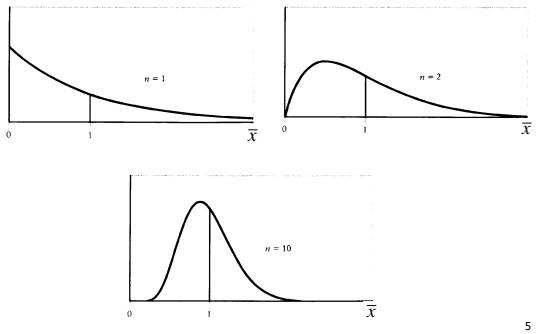
- Schat populatiewaarden obv steekproefwaarden
  - $\bar{x}$, *s*, *b*, *r*, … → μ, σ, β, ρ, …
- Stochastische variabele
  - Variabele waarbij waarde bepaald wordt door toeval
    - Steekproefgrootheden
  - Steekproevenverdeling gekend bij
    - Eenvoudig aselecte steekproef (EAS)
    - Grote *n*
  - Veel steekproefgrootheden volgen normaalverdeling

Centrale limietstelling

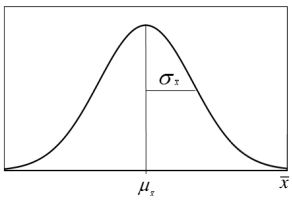
- Hoe groter *n*, hoe meer de *steekproevenverdeling* ***van het steekproefgemiddelde*** de normaalverdeling benadert
- Zelfs als de populatieverdeling niet normaal is
  - Hoe groter afwijking normaalverdeling, hoe groter *n* moet zijn
- In praktijk
  - Gebruik de perfecte normaalverdeling en je gezond verstand



CLS in actie



Inferentie voor het gemiddelde



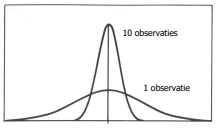
$$\mu_{\bar{x}} = \mu_x$$

$$\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}}$$

Inferentiële statistiek

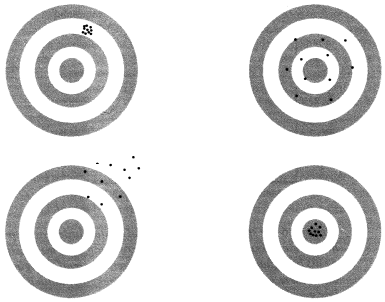
- Steekproevenverdeling
  - Centraliteit
    - Onvertekende schatter
  - Spreiding
    - Variantie van de steekproevenverdeling
    - Hoe groter *n*, hoe lager spreiding
      - Wet van de grote getallen

$$\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}}$$



7

Centraliteit en spreiding in actie



8

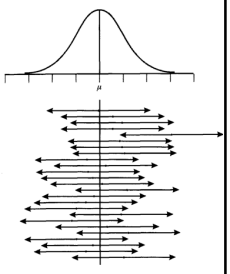
Inferentiële statistiek

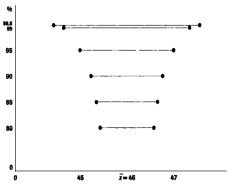
- Twee benaderingen:
  1. Betrouwbaarheidsintervallen
     - Schat de populatieparameter aan de hand van de grootheden van de steekproef
  2. Significantietoetsen
     - Test een assumptie over de populatieparameter aan de hand van de grootheden van de steekproef

9

1. Betrouwbaarheidsintervallen

- Puntschatting <-> intervalschatting
  - **BI = puntschatting + foutenmarge**
  - Betrouwbaarheidsniveau
    - 80%, 90%, 95%, 99%





10

1. Betrouwbaarheidsintervallen

- Componenten van de foutenmarge
  - Standaardafwijking steekproevenverdeling: $\sigma_{\bar{x}}$

$$\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}}$$

  - Betrouwbaarheidsniveau: C = 1 - α
    - (α= kans op fout)
      - Bepaald door *z** (van de standaardnormaalverdeling)

$$m = z^* \frac{\sigma_x}{\sqrt{n}}$$

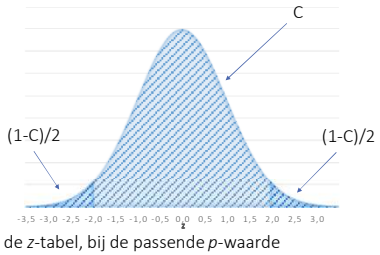
  - Volledige formule:

$$\bar{x} - z^* \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z^* \frac{\sigma}{\sqrt{n}}$$

11

1. Betrouwbaarheidsinterval

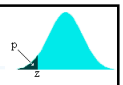
- Hoe bepalen we *z**?



*z** uit de *z*-tabel, bij de passende *p*-waarde

12

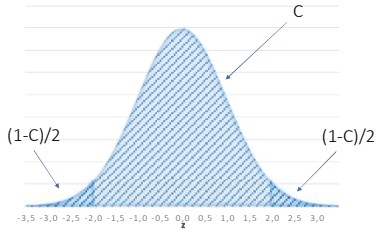
z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
-3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
-3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
-3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
-3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
-2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
-2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
-2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
-2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
-2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
-2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
-2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
-2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
-2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
-2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
-1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
-1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
-1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
-1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
-1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
-1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
-1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
-1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
-1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
-1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
-0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
-0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
-0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
-0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
-0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
-0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
-0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
-0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
-0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641



13

1. Betrouwbaarheidsintervallen

- Hoe bepalen we z^*?



z^* uit de z-tabel, bij de passende p-waarde

C = 90% → z* = 1,645

C = 95% → z* = 1,96

C = 99% → z* = 2,576

14

1. Betrouwbaarheidsintervallen

- Keuze van het betrouwbaarheidsinterval
  - Accuraatheid
  - Informatie
- Spreiding bepaald door
  - Betrouwbaarheidsniveau
  - Standaardafwijking van de steekproevenverdeling
    - Steekproefgrootte n

$$\bar{x} - z^* \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z^* \frac{\sigma}{\sqrt{n}}$$

15

1. Betrouwbaarheidsintervallen

- Voorwaarden
  - Normaliteit in de steekproevenverdeling
    - EAS met onafhankelijke waarnemingen
    - Centrale limietstelling
  - σ is gekend
    - Nogal vreemd…
    - Lossen we later op!

$$\bar{x} - z^* \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z^* \frac{\sigma}{\sqrt{n}}$$

16

1. Betrouwbaarheidsintervallen

- Voorbeeld 1
  - Huur voor een garagebox in Gent in een EAS van 500
    - $\bar{x}$ = 210
    - σ = 100 (verondersteld gekend)
  - Vraag: Schat de gemiddelde huur in de populatie met een 95% BI.

$$\bar{x} - z^* \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z^* \frac{\sigma}{\sqrt{n}}$$

  - Schatting?
    - $\bar{x}$ = 210
  - Foutenmarge?
    - z^* = 1,96
    - $\sigma_{\bar{x}} = \frac{100}{\sqrt{500}} = 4{,}5$
    - m = 1,96*4,5 = 8,82
  - Betrouwbaarheidsinterval:
    - 210 – 8,82 < μ < 210 + 8,82
    - 201,18 < μ < 218,82

17

1. Betrouwbaarheidsintervallen

- Voorbeeld 2

Greenpeace Vlaanderen voert een onderzoek uit. Ze willen onderzoeken hoeveel bomen er gemiddeld staan in de bossen van Vlaanderen. Uit vorig onderzoek weet men dat de standaardafwijking van het aantal bomen per bos 2483 bedraagt. Uit een EAS van 83 bossen berekent men het gemiddelde van 8496.

1) Schat het gemiddeld aantal bomen met een 95%-BI.
2) Schat het gemiddeld aantal bomen met een 80%-BI.
3) Zonder het te berekenen, zal het 85%-BI breder of smaller zijn dan het 80%-BI?
4) Zonder het te berekenen, zal een BI berekend op een steekproef van 100 bossen breder of smaller zijn dan die van Greenpeace?

18

1. Betrouwbaarheidsintervallen

- Voorbeeld 2

  Greenpeace Vlaanderen

  $\bar{x} = 8496$

  $\sigma = 2483$

  $n = 83$

$$\bar{x} - z^* \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z^* \frac{\sigma}{\sqrt{n}}$$

1) 95%-BI

   8496 +/- 1,96*2483/ √83 → [7961,91; 9030,19]

2) 80%-BI

   8496 +/- 1,285*2483/√83 → [8145.78; 8846,22]

3) 85%?

   Breder

4) n = 100?

   Smaller

19

2. Significantietoetsen

- Test of een assumptie over een populatieparameter waar is, gebaseerd op steekproefgrootheden

  Hier: over μ

  Voorbeelden

  De gemiddelde huurprijs van een garagebox in Gent is hoger dan het nationaal gemiddelde van € 254

  Proximus weet dat mensen gemiddeld 6 personen per dag bellen. Flair denkt dat vrouwen daarvan verschillen.

20

2. Significantietoetsen

- 5 stappen bij elke significantietoets

1. Hypotheses

   Nulhypothese

   Wat je wil testen **is niet waar**

   H_0: $\mu = \mu_0$

   Alternatieve hypothese

   Wat je wil testen **is waar**

   Drie vormen:

   H_a: $\mu \neq \mu_0$

   H_a: $\mu > \mu_0$

   H_a: $\mu < \mu_0$

   Kies op basis van theorie, niet steekproefresultaat

   Voor de rest van de significantietoets ga je ervan uit dat de nulhypothese waar is

21

2. Significantietoetsen

- 5 stappen bij elke significantietoets

1. Hypotheses

   Voorbeelden:

   Een melkproducent garandeert dat er minstens 1000 ml melk in elke fles zit. Wij denken minder.

   H_0: $\mu = 1000$

   H_a: $\mu < 1000$

   We weten dat Europeanen gemiddeld 3 uur per dag naar schermen kijken. We willen testen of Europese studenten daarvan verschillen.

   H_0: $\mu = 3$

   H_a: $\mu \neq 3$

22

2. Significantietoetsen

cf. $z = \frac{x - \mu}{\sigma}$

2. Steekproevenverdeling

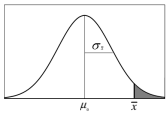
   Doen alsof nulhypothese klopt

   Hier: $N(\mu_0, \sigma/\sqrt{n})$

3. Toetsingsgrootheid

   Hoe ver is de steekproefgrootheid verwijderd van de populatiegrootheid, verondersteld in H_0

   Gestandaardiseerde afstand



$$z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$$

23

2. Significantietoetsen

4. Staartkans

   Kans om deze waarde op de steekproefparameter of extremer te vinden onder de voorwaarde dat H_0 waar is

   Zoek p-waarde in de tabel (hier: z-tabel)

5. Conclusie

   In praktijk:

   Als $p < 0{,}05$ → verwerp H_0

   Hoe kleiner p hoe meer bewijs tegen H_0

   Significantieniveau (α)

   $p <$,05: 0,05-niveau of 5%-niveau (*)

   $p <$,01: 0,01-niveau of 1%-niveau (**)

   $p <$,001: 0,001-niveau or 0,1%-niveau (***)

24

2. Significantietoetsen

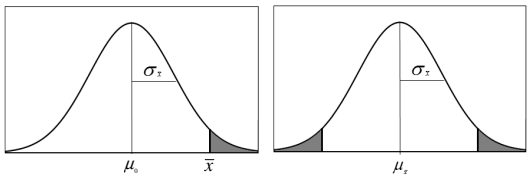
- Eénzijdige versus tweezijdige toetsen

  Hangt af van alternatieve hypothese

  H_a: $\mu > \mu_0$: staartkans: $P(Z \geq z)$

  H_a: $\mu < \mu_0$: staartkans: $P(Z \leq z)$

  H_a: $\mu \neq \mu_0$: staartkans: $P(Z \leq -z \text{ or } Z \geq z) = 2P(Z \geq |z|)$



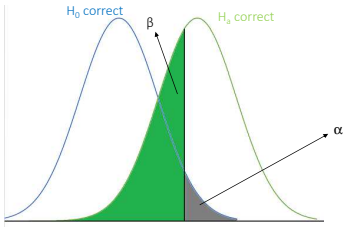
25

2. Significantietoetsen

- Twee types fouten

  Type I: H_0 verwerpen terwijl ze correct is (kans: α)

  Type II: H_0 aanvaarden terwijl ze fout is (kans: β)



26

2. Significantietoetsen

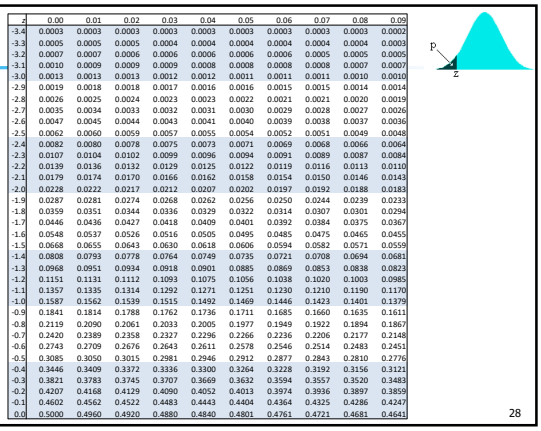
- Voorbeeld:

  We veronderstellen dat in Brussel de huur voor een garagebox hoger is dan in de rest van België, waar de huur gemiddeld € 215 bedraagt, met een standaardafwijking van 40. We nemen een EAS van 16 boxen in Brussel, en we berekenen daarin het gemiddelde van € 245.

  1. Hypothesen
     H_0: $\mu = 215$
     H_a: $\mu > 215$
  2. Steekproevenverdeling
     $N(215, 40/\sqrt{16})$
  3. Toetsingsgrootheid
     $z = (245-215)/(40/\sqrt{16}) = 3$
  4. Staartkans

$$z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$$

27



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
-3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
-3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
-3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
-3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
-2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
-2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
-2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
-2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
-2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
-2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
-2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
-2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
-2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
-2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
-1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
-1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
-1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
-1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
-1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
-1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
-1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
-1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
-1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
-1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
-0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
-0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
-0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
-0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
-0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
-0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
-0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
-0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
-0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641

28

2. Significantietoetsen

- Voorbeeld:

  We veronderstellen dat in Brussel de huur voor een garagebox hoger is dan in de rest van België, waar de huur gemiddeld € 215 bedraagt, met een standaardafwijking van 40. We nemen een EAS van 16 boxen in Brussel, en we berekenen daarin het gemiddelde van € 245.

  1. Hypothesen
     H_0: $\mu = 215$
     H_a: $\mu > 215$
  2. Steekproevenverdeling
     $N(215, 40/\sqrt{16})$
  3. Toetsingsgrootheid
     $z = (245-215)/(40/\sqrt{16}) = 3$
  4. Staartkans
     $p(z>3) = 0{,}0013$
  5. Conclusie?

$$z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$$

29

2. Significantietoetsen

- Voorbeeld:

  De toeristische dienst van de Belgische kust publiceert haar jaarlijkse rapport na de zomer. Ze beweren dat een gemiddelde van 267 ijsjes per dag werden verkocht, met een standaardafwijking van 15. De burgemeester van Oostende verwacht echter dat dat getal niet klopt voor zijn stad. Hij beveelt een onderzoek. In de EAS van 46 dagen bleek het gemiddelde 285 te zijn.

  1. Hypothesen
     H_0: $\mu = 267$
     H_a: $\mu \neq 267$
  2. Steekproevenverdeling
     $N(267, 15/\sqrt{46})$
  3. Toetsingsgrootheid
     $z = (285-267)/(15/\sqrt{46}) = 8{,}14$
  4. Staartkans
     $p(z>8{,}14) < 0{,}0002$
     → Tweezijdig → $p < 2*0{,}0002$ → $p < 0{,}0004$
  5. Conclusie?

30

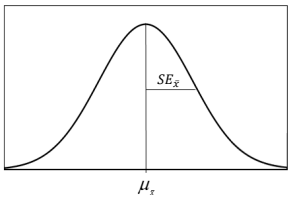
2. Significantietoetsen

- Gebruik met verstand!
  - EAS, *n* groot genoeg
- Significantieniveau
  - Sterkte van bewijs
  - Maar afhankelijk van steekproefgrootte
- Belang van betrouwbaarheidsintervallen
  - Significantietoetsen
    - Is er een verschil
  - Betrouwbaarheidsintervallen
    - Schatting van populatiewaarde
- Let op voor herhaaldelijk testen!
  - Type I en II fouten worden over testen opgeteld

31

Het σ-probleem

- Tot nu toe, kenden we σ
  - Redelijk onrealistisch
- Hoe lossen we dit op?

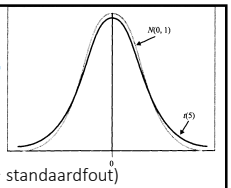


$$\mu_{\bar{x}} = \mu_x$$

$$\sigma_{\bar{x}} = \frac{S_x}{\sqrt{n}}$$

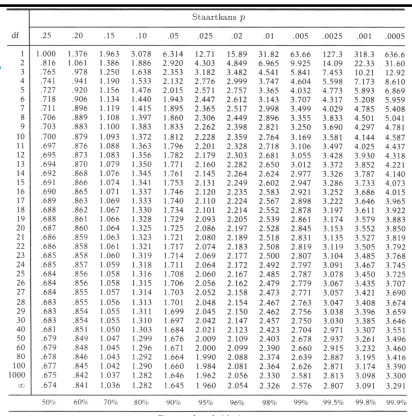
32

t-verdeling



- Implicatie schatten *σ* door *s*
  - Meer onzekerheid
    - σ_x → SE (standard error → standaardfout)
    - *z*-verdeling → *t*-verdeling
- Kenmerken
  - Hogere spreiding dan *z*-verdeling
  - Bepaald door μ, σ, en ***df***
    - df = *n* – 1
      - Hoe hoger, hoe meer *t* lijkt op *z*

33



	Staartkans *p*											
df	.25	.20	.15	.10	.05	.025	.02	.01	.005	.0025	.001	.0005
1	1.000	1.376	1.963	3.078	6.314	12.71	15.89	31.82	63.66	127.3	318.3	636.6
2	.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925	14.09	22.33	31.60
3	.765	.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.21	12.92
4	.741	.941	1.190	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173	8.610
5	.727	.920	1.156	1.476	2.015	2.571	2.757	3.365	4.032	4.773	5.893	6.869
6	.718	.906	1.134	1.440	1.943	2.447	2.612	3.143	3.707	4.317	5.208	5.959
7	.711	.896	1.119	1.415	1.895	2.365	2.517	2.998	3.499	4.029	4.785	5.408
8	.706	.889	1.108	1.397	1.860	2.306	2.449	2.896	3.355	3.833	4.501	5.041
9	.703	.883	1.100	1.383	1.833	2.262	2.398	2.821	3.250	3.690	4.297	4.781
10	.700	.879	1.093	1.372	1.812	2.228	2.359	2.764	3.169	3.581	4.144	4.587
11	.697	.876	1.088	1.363	1.796	2.201	2.328	2.718	3.106	3.497	4.025	4.437
12	.695	.873	1.083	1.356	1.782	2.179	2.303	2.681	3.055	3.428	3.930	4.318
13	.694	.870	1.079	1.350	1.771	2.160	2.282	2.650	3.012	3.372	3.852	4.221
14	.692	.868	1.076	1.345	1.761	2.145	2.264	2.624	2.977	3.326	3.787	4.140
15	.691	.866	1.074	1.341	1.753	2.131	2.249	2.602	2.947	3.286	3.733	4.073
16	.690	.865	1.071	1.337	1.746	2.120	2.235	2.583	2.921	3.252	3.686	4.015
17	.689	.863	1.069	1.333	1.740	2.110	2.224	2.567	2.898	3.222	3.646	3.965
18	.688	.862	1.067	1.330	1.734	2.101	2.214	2.552	2.878	3.197	3.611	3.922
19	.688	.861	1.066	1.328	1.729	2.093	2.205	2.539	2.861	3.174	3.579	3.883
20	.687	.860	1.064	1.325	1.725	2.086	2.197	2.528	2.845	3.153	3.552	3.850
21	.686	.859	1.063	1.323	1.721	2.080	2.189	2.518	2.831	3.135	3.527	3.819
22	.686	.858	1.061	1.321	1.717	2.074	2.183	2.508	2.819	3.119	3.505	3.792
23	.685	.858	1.060	1.319	1.714	2.069	2.177	2.500	2.807	3.104	3.485	3.768
24	.685	.857	1.059	1.318	1.711	2.064	2.172	2.492	2.797	3.091	3.467	3.745
25	.684	.856	1.058	1.316	1.708	2.060	2.167	2.485	2.787	3.078	3.450	3.725
26	.684	.856	1.058	1.315	1.706	2.056	2.162	2.479	2.779	3.067	3.435	3.707
27	.684	.855	1.057	1.314	1.703	2.052	2.158	2.473	2.771	3.057	3.421	3.690
28	.683	.855	1.056	1.313	1.701	2.048	2.154	2.467	2.763	3.047	3.408	3.674
29	.683	.854	1.055	1.311	1.699	2.045	2.150	2.462	2.756	3.038	3.396	3.659
30	.683	.854	1.055	1.310	1.697	2.042	2.147	2.457	2.750	3.030	3.385	3.646
40	.681	.851	1.050	1.303	1.684	2.021	2.123	2.423	2.704	2.971	3.307	3.551
50	.679	.849	1.047	1.299	1.676	2.009	2.109	2.403	2.678	2.937	3.261	3.496
60	.679	.848	1.045	1.296	1.671	2.000	2.099	2.390	2.660	2.915	3.232	3.460
80	.678	.846	1.043	1.292	1.664	1.990	2.088	2.374	2.639	2.887	3.195	3.416
100	.677	.845	1.042	1.290	1.660	1.984	2.081	2.364	2.626	2.871	3.174	3.390
1000	.675	.842	1.037	1.282	1.646	1.962	2.056	2.330	2.581	2.813	3.098	3.300
∞	.674	.841	1.036	1.282	1.645	1.960	2.054	2.326	2.576	2.807	3.091	3.291
	50%	60%	70%	80%	90%	95%	96%	98%	99%	99.5%	99.8%	99.9%

Betrouwbaarheidsniveau

34

t-verdeling

- Voorbeeld:
  De toeristische dienst van de Belgische kust publiceert haar jaarlijkse rapport na de zomer. Ze beweren dat een gemiddelde van 267 ijsjes per dag werden verkocht. De burgemeester van Oostende verwacht echter dat dat getal niet klopt voor zijn stad. Hij beveelt een onderzoek. In de EAS van dagen bleek het gemiddelde 285 te zijn, met een standaardafwijking van 45.
  1. Hypothesen
     H_0: μ = 267
     H_a: μ ≠ 267
  2. Steekproevenverdeling
     t(267, 45/√46) met df=45
  3. Toetsingsgrootheid
     t = (285-267)/(45/√46) = 2,71
  4. Staartkans
     p(*t*>2,71) < 0,005
     → Tweezijdig → *p* < 2*0,005 → *p* < 0,01
  5. Conclusie?

35

t-verdeling

- Voorbeeld:
  We weten nu dat het gemiddelde in Oostende verschillend is van het nationaal gemiddelde. Schat het populatiegemiddelde voor Oostende met 95% betrouwbaarheid. (*n*=46, weet je nog)

Schatting:

$\bar{x}$ = 285

Foutenmarge:

*t** = ?

36

	Staartkans p											
df	.25	.20	.15	.10	.05	.025	.02	.01	.005	.0025	.001	.0005
1	1.000	1.376	1.963	3.078	6.314	12.71	15.89	31.82	63.66	127.3	318.3	636.6
2	.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925	14.09	22.33	31.60
3	.765	.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.21	12.92
4	.741	.941	1.190	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173	8.610
5	.727	.920	1.156	1.476	2.015	2.571	2.757	3.365	4.032	4.773	5.893	6.869
6	.718	.906	1.134	1.440	1.943	2.447	2.612	3.143	3.707	4.317	5.208	5.959
7	.711	.896	1.119	1.415	1.895	2.365	2.517	2.998	3.499	4.029	4.785	5.408
8	.706	.889	1.108	1.397	1.860	2.306	2.449	2.896	3.355	3.833	4.501	5.041
9	.703	.883	1.100	1.383	1.833	2.262	2.398	2.821	3.250	3.690	4.297	4.781
10	.700	.879	1.093	1.372	1.812	2.228	2.359	2.764	3.169	3.581	4.144	4.587
11	.697	.876	1.088	1.363	1.796	2.201	2.328	2.718	3.106	3.497	4.025	4.437
12	.695	.873	1.083	1.356	1.782	2.179	2.303	2.681	3.055	3.428	3.930	4.318
13	.694	.870	1.079	1.350	1.771	2.160	2.282	2.650	3.012	3.372	3.852	4.221
14	.692	.868	1.076	1.345	1.761	2.145	2.264	2.624	2.977	3.326	3.787	4.140
15	.691	.866	1.074	1.341	1.753	2.131	2.249	2.602	2.947	3.286	3.733	4.073
16	.690	.865	1.071	1.337	1.746	2.120	2.235	2.583	2.921	3.252	3.686	4.015
17	.689	.863	1.069	1.333	1.740	2.110	2.224	2.567	2.898	3.222	3.646	3.965
18	.688	.862	1.067	1.330	1.734	2.101	2.214	2.552	2.878	3.197	3.611	3.922
19	.688	.861	1.066	1.328	1.729	2.093	2.205	2.539	2.861	3.174	3.579	3.883
20	.687	.860	1.064	1.325	1.725	2.086	2.197	2.528	2.845	3.153	3.552	3.850
21	.686	.859	1.063	1.323	1.721	2.080	2.189	2.518	2.831	3.135	3.527	3.819
22	.686	.858	1.061	1.321	1.717	2.074	2.183	2.508	2.819	3.119	3.505	3.792
23	.685	.858	1.060	1.319	1.714	2.069	2.177	2.500	2.807	3.104	3.485	3.768
24	.685	.857	1.059	1.318	1.711	2.064	2.172	2.492	2.797	3.091	3.467	3.745
25	.684	.856	1.058	1.316	1.708	2.060	2.167	2.485	2.787	3.078	3.450	3.725
26	.684	.856	1.058	1.315	1.706	2.056	2.162	2.479	2.779	3.067	3.435	3.707
27	.684	.855	1.057	1.314	1.703	2.052	2.158	2.473	2.771	3.057	3.421	3.690
28	.683	.855	1.056	1.313	1.701	2.048	2.154	2.467	2.763	3.047	3.408	3.674
29	.683	.854	1.055	1.311	1.699	2.045	2.150	2.462	2.756	3.038	3.396	3.659
30	.683	.854	1.055	1.310	1.697	2.042	2.147	2.457	2.750	3.030	3.385	3.646
40	.681	.851	1.050	1.303	1.684	2.021	2.123	2.423	2.704	2.971	3.307	3.551
50	.679	.849	1.047	1.299	1.676	2.009	2.109	2.403	2.678	2.937	3.261	3.496
60	.679	.848	1.045	1.296	1.671	2.000	2.099	2.390	2.660	2.915	3.232	3.460
80	.678	.846	1.043	1.292	1.664	1.990	2.088	2.374	2.639	2.887	3.195	3.416
100	.677	.845	1.042	1.290	1.660	1.984	2.081	2.364	2.626	2.871	3.174	3.390
1000	.675	.842	1.037	1.282	1.646	1.962	2.056	2.330	2.581	2.813	3.098	3.300
∞	.674	.841	1.036	1.282	1.645	1 960	2.054	2.326	2.576	2.807	3.091	3.291
	50%	60%	70%	80%	90%	95%	96%	98%	99%	99.5%	99.8%	99.9%

Betrouwbaarheidsniveau

37

t-distribution

- Voorbeeld:

  We weten nu dat het gemiddelde in Oostende verschillend is van het nationaal gemiddelde. Schat het populatiegemiddelde voor Oostende met 95% betrouwbaarheid. (*n*=46, weet je nog)

  Schatting:

  $\bar{x} = 285$

  Foutenmarge:

  $t^* = 2{,}021$

  $SE = 45/\sqrt{46} = 6{,}63$

  $m = 2{,}021 * 6{,}63 = 13{,}41$

  BI:

  $285 - 13{,}41 < \mu < 285 + 13{,}41$

  $271{,}59 < \mu < 298{,}41$

38

t-verdeling

We weten dat, voor de hele populatie studenten aan de Universiteit Gent, het gemiddeld aantal pintjes per week 6,5 bedraagt. We vragen ons nu af of de studenten sociologie hiervan verschillen. We namen een EAS van 132 sociologiestudenten. Het gemiddelde in deze steekproef was 8.19, met een standaardafwijking van 11. Kunnen we op het 0.05-niveau besluiten dat de sociologiestudenten van alle UGent-studenten verschillen?

1. Hypothesen

   $H_0: \mu = 6{,}5$

   $H_a: \mu \neq 6{,}5$
2. Steekproevenverdeling

   $t(6{,}5,\ 11/\sqrt{132})$ met df=131
3. Toetsingsgrootheid

   $t = (8{,}19-6{,}5)/(11/\sqrt{132}) = 1{,}77$
4. Staartkans

   $p(t>1{,}77) < 0{,}05$

   → Tweezijdig → $p < 2*0{,}05$ → $p < 0{,}1$
5. Conclusie?

39

Moore, McCabe, & Craig (2009). Introduction to the practice of statistics. New York: Freeman (selected chapters)

5.33 Multiple-choice tests. Here is a simple probability model for multiple-choice tests. Suppose that each student has probability p of correctly answering a question chosen at random from a universe of possible questions. (A strong student has a higher p than a weak student.) The correctness of an answer to a question is independent of the correctness of answers to other questions. Jodi is a good student for whom $p = 0.85$.

(a) Use the Normal approximation to find the probability that Jodi scores 80% or lower on a 100-question test.

(b) If the test contains 250 questions, what is the probability that Jodi will score 80% or lower?

(c) How many questions must the test contain in order to reduce the standard deviation of Jodi's proportion of correct answers to half its value for a 100-item test?

(d) Laura is a weaker student for whom $p = 0.75$. Does the answer you gave in (c) for the standard deviation of Jodi's score apply to Laura's standard deviation also?

5.34 Tossing a die. You are tossing a balanced die that has probability 1/6 of coming up 1 on each toss. Tosses are independent. We are interested in how long we must wait to get the first 1.

(a) The probability of a 1 on the first toss is 1/6. What is the probability that the first toss is not a 1 and the second toss is a 1?

(b) What is the probability that the first two tosses are not 1s and the third toss is a 1? This is the probability that the first 1 occurs on the third toss.

(c) Now you see the pattern. What is the probability that the first 1 occurs on the fourth toss? On the fifth toss?

5.35 The geometric distribution. Generalize your work in Exercise 5.34. You have independent trials, each resulting in a success or a failure. The probability of a success is p on each trial. The binomial distribution describes the count of successes in a fixed number of trials. Now the number of trials is not fixed; instead, continue until you get a success. The random variable Y is the number of the trial on which the first success occurs. What are the possible values of Y? What is the probability $P(Y = k)$ for any of these values? (*Comment:* The distribution of the number of trials to the first success is called a **geometric distribution**.)

5.2 The Sampling Distribution of a Sample Mean

Counts and proportions are discrete random variables that describe categorical data. The statistics most often used to describe quantitative data, on the other hand, are continuous random variables. The sample mean, percentiles, and standard deviation are examples of statistics based on quantitative data. Statistical theory describes the sampling distributions of these statistics. In this section we will concentrate on the sample mean. Because sample means are just averages of observations, they are among the most common statistics.

EXAMPLE

5.17 Sample means are approximately Normal. Figure 5.8 illustrates two striking facts about the sampling distribution of a sample mean. Figure 5.8(a) displays the distribution of customer service call lengths for a bank service center for a month. There are more than 30,000 calls in this population.[6] (We omitted a few extreme outliers, calls that lasted more than 20 minutes.) The distribution is extremely skewed to the right. The population mean is $\mu = 173.95$ seconds.

Table 1.1 (page 8) contains the lengths of a sample of 80 calls from this population. The mean of these 80 calls is $\bar{x} = 196.6$ seconds. If we take more samples of size 80, we will get different values of $\bar{x}$. To find the sampling distribution of $\bar{x}$, take many random samples of size 80 and calculate $\bar{x}$

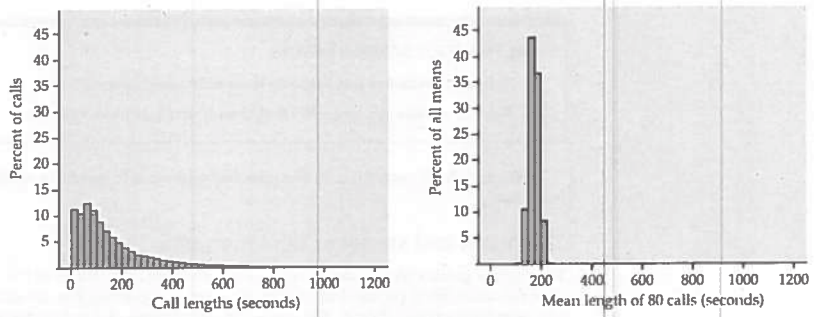


FIGURE 5.8 (a) The distribution of lengths of all customer service calls received by a bank in a month. (b) The distribution of the sample means $\bar{x}$ for 500 random samples of size 80 from this population. The scales and histogram classes are exactly as in Figure 5.8(a).

for each sample. Figure 5.8(b) is the distribution of the values of $\bar{x}$ for 500 samples. The scales and choice of classes are exactly the same as in Figure 5.8(a), so that we can make a direct comparison.

The sample means are much less spread out than the individual call lengths. What is more, the distribution in Figure 5.8(b) is roughly symmetric rather than skewed. The Normal quantile plot in Figure 5.9 confirms that the distribution is close to Normal.

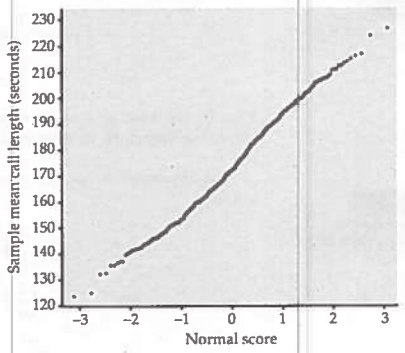


FIGURE 5.9 Normal quantile plot of the 500 sample means in Figure 5.8(b). The distribution is close to Normal.

This example illustrates two important facts about sample means that we will discuss in this section.

FACTS ABOUT SAMPLE MEANS

1. Sample means are less variable than individual observations.
2. Sample means are more Normal than individual observations.

These two facts contribute to the popularity of sample means in statistical inference.

The mean and standard deviation of $\bar{x}$

The sample mean $\bar{x}$ from a sample or an experiment is an estimate of the mean μ of the underlying population, just as a sample proportion $\hat{p}$ is an estimate of a population proportion p. The sampling distribution of $\bar{x}$ is determined by the design used to produce the data, the sample size n, and the population distribution.

Select an SRS of size n from a population, and measure a variable X on each individual in the sample. The n measurements are values of n random variables $X_1, X_2, \ldots, X_n$. A single X_i is a measurement on one individual selected at random from the population and therefore has the distribution of the population. If the population is large relative to the sample, we can consider $X_1, X_2, \ldots, X_n$ to be independent random variables each having the same distribution. This is our probability model for measurements on each individual in an SRS.

The sample mean of an SRS of size n is

$$\bar{x} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$$

LOOK BACK
addition rule for means, page 278

If the population has mean μ, then μ is the mean of the distribution of each observation X_i. The addition rule for means of random variables,

$$\mu_{\bar{x}} = \frac{1}{n}(\mu_{X_1} + \mu_{X_2} + \cdots + \mu_{X_n})$$
$$= \frac{1}{n}(\mu + \mu + \cdots + \mu) = \mu$$

That is, *the mean of $\bar{x}$ is the same as the mean of the population.* The sample mean $\bar{x}$ is therefore an unbiased estimator of the unknown population mean μ.

LOOK BACK
addition rule for variances, page 282

The observations are independent, so the addition rule for variances also applies:

$$\sigma^2_{\bar{x}} = \left(\frac{1}{n}\right)^2 (\sigma^2_{X_1} + \sigma^2_{X_2} + \cdots + \sigma^2_{X_n})$$
$$= \left(\frac{1}{n}\right)^2 (\sigma^2 + \sigma^2 + \cdots + \sigma^2)$$
$$= \frac{\sigma^2}{n}$$

Just as in the case of a sample proportion $\hat{p}$, the variability of the sampling distribution of a sample mean decreases as the sample size grows. Because the

standard deviation of $\bar{x}$ is $\sigma/\sqrt{n}$, it is again true that the standard deviation of the statistic decreases in proportion to the square root of the sample size. Here is a summary of these facts.

MEAN AND STANDARD DEVIATION OF A SAMPLE MEAN

Let $\bar{x}$ be the mean of an SRS of size n from a population having mean μ and standard deviation σ. The mean and standard deviation of $\bar{x}$ are

$$\mu_{\bar{x}} = \mu$$
$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

How accurately does a sample mean $\bar{x}$ estimate a population mean μ? Because the values of $\bar{x}$ vary from sample to sample, we must give an answer in terms of the sampling distribution. We know that $\bar{x}$ is an unbiased estimator of μ, so its values in repeated samples are not systematically too high or too low. Most samples will give an $\bar{x}$-value close to μ if the sampling distribution is concentrated close to its mean μ. So the accuracy of estimation depends on the spread of the sampling distribution.

EXAMPLE

5.18 Standard deviations for sample means of service call lengths. The standard deviation of the population of service call lengths in Figure 5.8(a) is $\sigma = 184.81$ seconds. The length of a single call will often be far from the population mean. If we choose an SRS of 20 calls, the standard deviation of their mean length is

$$\sigma_{\bar{x}} = \frac{184.81}{\sqrt{20}} = 41.32 \text{ seconds}$$

Averaging over more calls reduces the variability and makes it more likely that $\bar{x}$ is close to μ. Our sample size of 80 calls is 4 times 20, so the standard deviation will be half as large:

$$\sigma_{\bar{x}} = \frac{184.81}{\sqrt{80}} = 20.66 \text{ seconds}$$

USE YOUR KNOWLEDGE

5.36 Find the mean and the standard deviation of the sampling distribution. You take an SRS of size 25 from a population with mean 200 and standard deviation 10. Find the mean and standard deviation of the sampling distribution of your sample mean.

5.37 The effect of increasing the sample size. In the setting of the previous exercise, repeat the calculations for a sample size of 100. Explain the effect of the increase on the sample mean and standard deviation.

The central limit theorem

We have described the center and spread of the probability distribution of a sample mean $\bar{x}$, but not its shape. The shape of the distribution of $\bar{x}$ depends on the shape of the population distribution. Here is one important case: if the population distribution is Normal, then so is the distribution of the sample mean.

SAMPLING DISTRIBUTION OF A SAMPLE MEAN

If a population has the $N(\mu, \sigma)$ distribution, then the sample mean $\bar{x}$ of n independent observations has the $N(\mu, \sigma/\sqrt{n})$ distribution.

This is a somewhat special result. Many population distributions are not Normal. The service call lengths in Figure 5.8(a), for example, are strongly skewed. Yet Figures 5.8(b) and 5.9 show that means of samples of size 80 are close to Normal. One of the most famous facts of probability theory says that, for large sample sizes, the distribution of $\bar{x}$ is close to a Normal distribution. This is true no matter what shape the population distribution has, as long as the population has a finite standard deviation σ. This is the **central limit theorem.** It is much more useful than the fact that the distribution of $\bar{x}$ is exactly Normal if the population is exactly Normal.

central limit theorem

CENTRAL LIMIT THEOREM

Draw an SRS of size n from any population with mean μ and finite standard deviation σ. When n is large, the sampling distribution of the sample mean $\bar{x}$ is approximately Normal:

$$\bar{x} \text{ is approximately } N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

EXAMPLE

5.19 How close will the sample mean be to the population mean? With the Normal distribution to work with, we can better describe how accurately a random sample of 80 calls estimates the mean length of all the calls in the population. The population standard deviation for the more than 30,000 calls in the population of Figure 5.8(a) is $\sigma = 184.81$ seconds. From Example 5.18 we know $\sigma_{\bar{x}} = 20.66$ seconds. By the 95 part of the 68–95–99.7 rule, 95% of all samples will have mean $\bar{x}$ within two standard deviations of μ, that is, within ± 41.32 seconds of μ.

LOOK BACK
68–95–99.7 rule, page 59

USE YOUR KNOWLEDGE

5.38 Use the 68–95–99.7 rule. You take an SRS of size 100 from a population with mean 200 and standard deviation 10. According to the central limit theorem, what is the approximate sampling distribution of the sample mean? Use the 95 part of the 68–95–99.7 rule to describe the variability of this sample mean.

For the sample size of $n = 80$ in Example 5.19, the sample mean is not very accurate. The population is very spread out, so the sampling distribution of $\bar{x}$ is still quite variable.

EXAMPLE

5.20 How can we reduce the standard deviation? In the setting of Example 5.19, if we want to reduce the standard deviation of $\bar{x}$ by a factor of 4, we must take a sample 16 times as large, $n = 16 \times 80$, or 1280. Then

$$\sigma_{\bar{x}} = \frac{184.81}{\sqrt{1280}} = 5.165 \text{ seconds}$$

For samples of size 1280, 95% of the sample means will be within twice 5.165, or 10.33 seconds, of the population mean μ.

USE YOUR KNOWLEDGE

5.39 The effect of increasing the sample size. In the setting of Exercise 5.38, suppose we increase to the sample size to 400. Use the 95 part of the 68–95–99.7 rule to describe the variability of this sample mean. Compare your results with those you found in Exercise 5.38.

Example 5.20 reminds us that if the population is very spread out, the $\sqrt{n}$ in the standard deviation of $\bar{x}$ implies that even large samples will not estimate the population mean accurately. But the big point of the example is that the central limit theorem allows us to use Normal probability calculations to answer questions about sample means even when the population distribution is not Normal. How large a sample size n is needed for $\bar{x}$ to be close to Normal depends on the population distribution. More observations are required if the shape of the population distribution is far from Normal. Even for the very skewed call length population, however, samples of size 80 are large enough. Here is a more detailed example.

EXAMPLE

5.21 The central limit theorem in action. Figure 5.10 shows the central limit theorem in action for another very non-Normal population. Figure 5.10(a) displays the density curve of a single observation, that is, of the population. The distribution is strongly right-skewed, and the most probable outcomes are near 0. The mean μ of this distribution is 1, and its standard deviation σ is also 1. This particular continuous distribution is called an **exponential distribution.** Exponential distributions are used as models for how long an electronic component will last and for the time required to serve a customer or repair a machine.

exponential distribution

Figures 5.10(b), (c), and (d) are the density curves of the sample means of 2, 10, and 25 observations from this population. As n increases, the shape be-

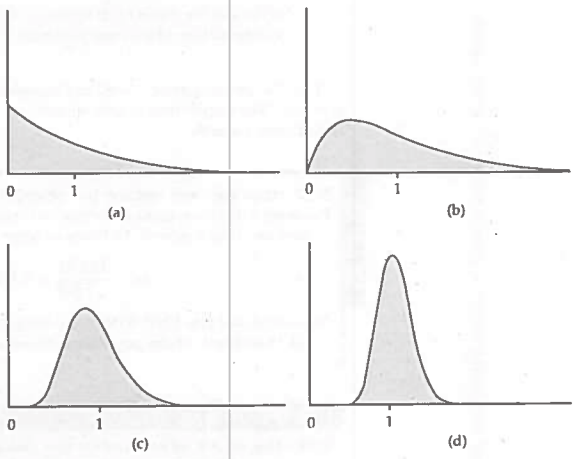


FIGURE 5.10 The central limit theorem in action: the distribution of sample means from a strongly non-Normal population becomes more Normal as the sample size increases. (a) The distribution of 1 observation. (b) The distribution of $\bar{x}$ for 2 observations. (c) The distribution of $\bar{x}$ for 10 observations. (d) The distribution of $\bar{x}$ for 25 observations.

comes more Normal. The mean remains at $\mu = 1$, but the standard deviation decreases, taking the value $1/\sqrt{n}$. The density curve for 10 observations is still somewhat skewed to the right but already resembles a Normal curve having $\mu = 1$ and $\sigma = 1/\sqrt{10} = 0.32$. The density curve for $n = 25$ is yet more Normal. The contrast between the shapes of the population distribution and of the distribution of the mean of 10 or 25 observations is striking.

APPLET

The *Central Limit Theorem* applet animates Figure 5.10. You can slide the sample size n from 1 to 100 and watch both the exact density curve of $\bar{x}$ and the Normal approximation. As you increase n, the two curves move closer together.

EXAMPLE

5.22 Preventive maintenance on an air-conditioning unit. The time X that a technician requires to perform preventive maintenance on an air-conditioning unit is governed by the exponential distribution whose density curve appears in Figure 5.10(a). The mean time is $\mu = 1$ hour and the standard deviation is $\sigma = 1$ hour. Your company operates 70 of these units. What is the probability that their average maintenance time exceeds 50 minutes?

The central limit theorem says that the sample mean time $\bar{x}$ (in hours) spent working on 70 units has approximately the Normal distribution with mean equal to the population mean $\mu = 1$ hour and standard deviation

$$\frac{\sigma}{\sqrt{70}} = \frac{1}{\sqrt{70}} = 0.12 \text{ hour}$$

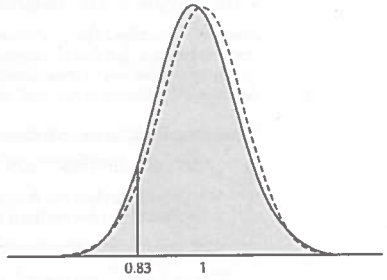


FIGURE 5.11 The exact distribution (dashed) and the Normal approximation from the central limit theorem (solid) for the average time needed to maintain an air conditioner, for Example 5.22.

The distribution of $\bar{x}$ is therefore approximately $N(1, 0.12)$. Figure 5.11 shows this Normal curve (solid) and also the actual density curve of $\bar{x}$ (dashed).

Because 50 minutes is 50/60 of an hour, or 0.83 hour, the probability we want is $P(\bar{x} > 0.83)$. A Normal distribution calculation gives this probability as 0.9222. This is the area to the right of 0.83 under the solid Normal curve in Figure 5.11. The exactly correct probability is the area under the dashed density curve in the figure. It is 0.9294. The central limit theorem Normal approximation is off by only about 0.007.

USE YOUR KNOWLEDGE

5.40 Find a probability. Refer to the example above. Find the probability that the mean time spent working on 70 units is less than 1.1 hours.

EXAMPLE

5.23 Convert the results to the total maintenance time. In Example 5.22 what can we say about the total maintenance time for 70 units? According to the central limit theorem

$$P(\bar{x} > 0.83) = 0.9222$$

We know that the sample mean is the total maintenance time divided by 70, so the event $\{\bar{x} > 0.83\}$ is the same as the event $\{70\bar{x} > 70(0.83)\}$. We can say that the probability is 0.9222 that the total maintenance time is $70(0.83) = 58.1$ hours or greater.

Figure 5.12 summarizes the facts about the sampling distribution of $\bar{x}$ in a way that emphasizes the big idea of a sampling distribution.

- Keep taking random samples of size n from a population with mean μ.
- Find the sample mean $\bar{x}$ for each sample.
- Collect all the $\bar{x}$'s and display their distribution.

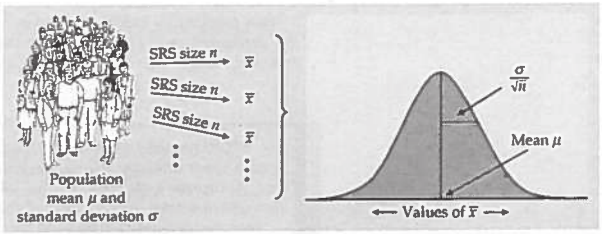


FIGURE 5.12 The sampling distribution of a sample mean $\bar{x}$ has mean μ and standard deviation $\sigma/\sqrt{n}$. The distribution is Normal if the population distribution is Normal; it is approximately Normal for large samples in any case.

That's the sampling distribution of $\bar{x}$. Sampling distributions are the key to understanding statistical inference. Keep this figure in mind as you go forward.

A few more facts

The central limit theorem is the big fact of this section. Here are three useful smaller facts related to our topic.

The Normal approximation for sample proportions and counts is an example of the central limit theorem. This is true because a sample proportion can be thought of as a sample mean. Recall the idea that we used to find the mean and variance of a binomial random variable X. We wrote the count X as a sum

$$X = S_1 + S_2 + \cdots + S_n$$

of random variables S_i that take the value 1 if a success occurs on the ith trial and the value 0 otherwise. The variables S_i take only the values 0 and 1 and are far from Normal. The proportion $\hat{p} = X/n$ is the sample mean of the S_i and, like all sample means, is approximately Normal when n is large.

The fact that the sample mean of an SRS from a Normal population has a Normal distribution is a special case of a more general fact: **any linear combination of independent Normal random variables is also Normally distributed.** That is, if X and Y are independent Normal random variables and a and b are any fixed numbers, $aX + bY$ is also Normally distributed, and so it is for any number of Normal variables. In particular, the sum or difference of independent Normal random variables has a Normal distribution. The mean and standard deviation of $aX + bY$ are found as usual from the addition rules for means and variances. These facts are often used in statistical calculations.

LOOK BACK
rules for means, page 278
rules for variances, page 282

EXAMPLE

5.24 Who will win? Tom and George are playing in the club golf tournament. Their scores vary as they play the course repeatedly. Tom's score X has the $N(110, 10)$ distribution, and George's score Y varies from round to round according to the $N(100, 8)$ distribution. If they play independently, what is the probability that Tom will score lower than George and thus do better in the tournament? The difference $X - Y$ between their scores is Normally

CHAPTER 6

Introduction to Inference

Undergraduate student loan debt has been increasing steadily during the past decade. Is the debt becoming too much of a burden upon graduation? Example 6.4 discusses the average debt of undergraduate borrowers.

Introduction

Statistical inference draws conclusions about a population or process based on sample data. It also provides a statement, expressed in terms of probability, of how much confidence we can place in our conclusions. Although there are many specific recipes for inference, there are only a few general types of statistical inference. This chapter introduces the two most common types: *confidence intervals* and *tests of significance*.

Because the underlying reasoning for these types of inference remains the same across different settings, this chapter considers a single simple setting: inference about the mean of a Normal population whose standard deviation is known. Later chapters will present the recipes for inference in other situations.

In this setting, we can address questions like:

- What is the average loan debt among undergraduate borrowers?
- What is the average miles per gallon (mpg) for a hybrid car?
- Is the average cholesterol level of undergraduate women at your university below the national average?

Overview of Inference

The purpose of statistical inference is to draw conclusions from data. We have already examined data and arrived at conclusions many times. Formal inference emphasizes substantiating our conclusions by probability calculations. Probability allows us to take chance variation into account. Here is an example.

EXAMPLE

6.1 Clustering of trees in a forest. The Wade Tract in Thomas County, Georgia, is an old-growth forest of longleaf pine trees (*Pinus palustris*) that has survived in a relatively undisturbed state since before the settlement of the area by Europeans. Foresters who study these trees are interested in how the trees are distributed in the forest. Is there some sort of clustering, resulting in regions of the forest with more trees than others? Or are the tree locations random, resulting in no particular patterns?

Figure 6.1 gives a plot of the locations of all 584 longleaf pine trees in a 200-meter by 200-meter region in the Wade Tract.[1] Do the locations appear to be random, or do there appear to be clusters of trees? One approach to the analysis of these data indicates that a pattern as clustered as, or more clustered than, the one in Figure 6.1 would occur only 4% of the time if, in fact, the locations of longleaf pine trees in the Wade Tract are random. Because this chance is fairly small, we conclude that there is some clustering of these trees.

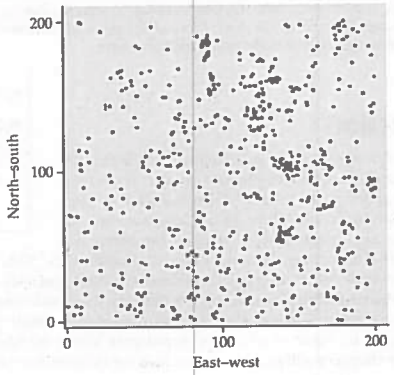


FIGURE 6.1 The distribution of longleaf pine trees, for Example 6.1.

Our probability calculation helps us to distinguish between patterns that are consistent or inconsistent with the random location scenario. Here is an example comparing two drug treatments with a different conclusion.

EXAMPLE

6.2 Effectiveness of a new drug. Researchers want to know if a new drug is more effective than a placebo. Twenty patients receive the new drug, and 20 receive a placebo. Twelve (60%) of those taking the drug show improvement versus only 8 (40%) of the placebo patients.

Our unaided judgment would suggest that the new drug is better. However, probability calculations tell us that a difference this large or larger between the results in the two groups would occur about one time in five simply because of chance variation. Since this probability is not very small, it is better to conclude that the observed difference is due to chance rather than a real difference between the two treatments.

LOOK BACK
sampling distributions, page 215

In this chapter we introduce the two most prominent types of formal statistical inference. Section 6.1 concerns *confidence intervals* for estimating the value of a population parameter. Section 6.2 presents *tests of significance*, which assess the evidence for a claim. Both types of inference are based on the sampling distributions of statistics. That is, both report probabilities that state *what would happen if we used the inference method many times*. This kind of probability statement is characteristic of standard statistical inference. Users of statistics must understand the nature of this reasoning and the meaning of the probability statements that appear, for example, in newspaper and journal articles as well as statistical software output.

CAUTION

Because the methods of formal inference are based on sampling distributions, they require a probability model for the data. Trustworthy probability models can arise in many ways, but the model is most secure and inference is most reliable when the data are produced by a properly randomized design. *When you use statistical inference, you are acting as if the data come from a random sample or a randomized experiment.* If this is not true, your conclusions may be open to challenge. Do not be overly impressed by the complex details of formal inference. This elaborate machinery cannot remedy basic flaws in producing the data such as voluntary response samples and confounded experiments. Use the common sense developed in your study of the first three chapters of this book, and proceed to detailed formal inference only when you are satisfied that the data deserve such analysis.

The primary purpose of this chapter is to describe the reasoning used in statistical inference. We will discuss only a few specific inference techniques, and these require an unrealistic assumption: that we know the standard deviation σ. Later chapters will present inference methods for use in most of the settings we met in learning to explore data. There are libraries—both of books and of computer software—full of more elaborate statistical techniques. Informed use of any of these methods requires an understanding of the underlying reasoning. A computer will do the arithmetic, but you must still exercise judgment based on understanding.

6.1 Estimating with Confidence

The SAT test is a widely used measure of readiness for college study. Originally, there were two sections, one for verbal reasoning ability (SATV) and one for mathematical reasoning ability (SATM). In April 1995, section scores were *recentered* so that the mean is approximately 500 in a large "standardized group." This scale has been maintained since then so that scores have a constant interpretation.

LOOK BACK
linear transformations, page 45

In 2005, the College Board changed the test, renaming the verbal section "Critical Reading" and adding a third section on writing ability. These changes increased the total possible score to 2400, extended the exam an additional 35 minutes, and increased the cost to register for the exam by $12. The changes also raised concerns about the constant-interpretation assumption.

EXAMPLE

6.3 Estimating the mean SATM score for seniors in California. Suppose you want to estimate the mean SATM score for the more than 420,000 high school seniors in California. You know better than to trust data from the students who choose to take the SAT. Only about 45% of California students take the SAT. These self-selected students are planning to attend college and are not representative of all California seniors. At considerable effort and expense, you give the test to a simple random sample (SRS) of 500 California high school seniors. The mean score for your sample is $\bar{x} = 461$. What can you say about the mean score μ in the population of all 420,000 seniors?

The sample mean $\bar{x}$ is the natural estimator of the unknown population mean μ. We know that $\bar{x}$ is an unbiased estimator of μ. More important, the law of large numbers says that the sample mean must approach the population mean as the size of the sample grows. The value $\bar{x} = 461$ therefore appears to be a reasonable estimate of the mean score μ that all 420,000 students would achieve if they took the test. But how reliable is this estimate? A second sample would surely not give 461 again. Unbiasedness says only that there is no systematic tendency to underestimate or overestimate the truth. Could we plausibly get a sample mean of 410 or 510 on repeated samples? An estimate without an indication of its variability is of little value.

Statistical confidence

Just as unbiasedness of an estimator concerns the center of its sampling distribution, questions about variation are answered by looking at the spread. We know that if the entire population of SAT scores has mean μ and standard deviation σ, then in repeated samples of size 500 the sample mean $\bar{x}$ follows the $N(\mu, \sigma/\sqrt{500})$ distribution. Let us suppose that we know that the standard deviation σ of SATM scores in our California population is $\sigma = 100$. (This is not realistic. We will see in the next chapter how to proceed when σ is not known. For now, we are more interested in statistical reasoning than in details of realistic methods.) In repeated sampling the sample mean $\bar{x}$ has a Normal distribution centered at the unknown population mean μ and having standard deviation

LOOK BACK
distribution of the sample mean, page 339

$$\sigma_{\bar{x}} = \frac{100}{\sqrt{500}} = 4.5$$

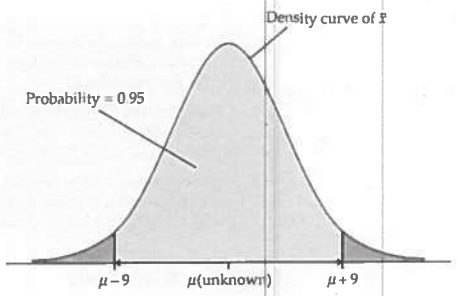


FIGURE 6.2 $\bar{x}$ lies within ± 9 of μ in 95% of all samples, so μ also lies within ± 9 of $\bar{x}$ in those samples.

Now we are in business. Consider this line of thought, which is illustrated by Figure 6.2:

- The 68–95–99.7 rule says that the probability is about 0.95 that $\bar{x}$ will be within 9 points (two standard deviations of $\bar{x}$) of the population mean score μ.
- To say that $\bar{x}$ lies within 9 points of μ is the same as saying that μ is within 9 points of $\bar{x}$.
- So 95% of all samples will capture the true μ in the interval from $\bar{x} - 9$ to $\bar{x} + 9$.

We have simply restated a fact about the sampling distribution of $\bar{x}$. *The language of statistical inference uses this fact about what would happen in the long run to express our confidence in the results of any one sample.* Our sample gave $\bar{x} = 461$. We say that we are *95% confident* that the unknown mean score for all California seniors lies between

$$\bar{x} - 9 = 461 - 9 = 452$$

and

$$\bar{x} + 9 = 461 + 9 = 470$$

Be sure you understand the grounds for our confidence. There are only two possibilities for our SRS:

1. The interval between 452 and 470 contains the true μ.
2. The interval between 452 and 470 does not contain the true μ.

We cannot know whether our sample is one of the 95% for which the interval $\bar{x} \pm 9$ catches μ or one of the unlucky 5% that does not catch μ. The statement that we are 95% confident is shorthand for saying, "We arrived at these numbers by a method that gives correct results 95% of the time."

USE YOUR KNOWLEDGE

6.1 How much do you spend on lunch? The average amount you spend on a lunch during the week is not known. Based on past experience,

you are willing to assume that the standard deviation is about $2. If you take a random sample of 36 lunches, what is the value of the standard deviation for $\bar{x}$?

6.2 **Applying the 68–95–99.7 rule.** In the setting of the previous exercise, the 68–95–99.7 rule says that the probability is about 0.95 that $\bar{x}$ is within $\$\_\_\_\_\_$ of the population mean μ. Fill in the blank.

6.3 **Constructing a 95% confidence interval.** In the setting of the previous two exercises, about 95% of all samples will capture the true mean in the interval $\bar{x}$ plus or minus $\$\_\_\_\_\_$. Fill in the blank.

Confidence intervals

The interval of numbers between the values $\bar{x} \pm 9$ is called a *95% confidence interval* for μ. Like most confidence intervals we will discuss, this one has the form

$$\text{estimate} \pm \text{margin of error}$$

margin of error

The estimate ($\bar{x}$ in this case) is our guess for the value of the unknown parameter. The **margin of error** (9 here) reflects how accurate we believe our guess is, based on the variability of the estimate, and how confident we are that the procedure will catch the true population mean μ.

Figure 6.3 illustrates the behavior of 95% confidence intervals in repeated sampling. The center of each interval is at $\bar{x}$ and therefore varies from sample to sample. The sampling distribution of $\bar{x}$ appears at the top of the figure to show

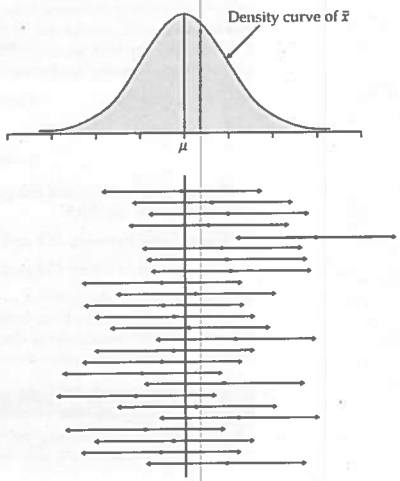


FIGURE 6.3 Twenty-five samples from the same population gave these 95% confidence intervals. In the long run, 95% of all samples give an interval that covers μ.

the long-term pattern of this variation. The 95% confidence intervals, $\bar{x} \pm 9$, from 25 SRSs appear below. The center $\bar{x}$ of each interval is marked by a dot. The arrows on either side of the dot span the confidence interval. All except one of the 25 intervals cover the true value of μ. In a very large number of samples, 95% of the confidence intervals would contain μ. With the *Confidence Interval* applet, you can construct many diagrams similar to the one displayed in Figure 6.3.

APPLET

Statisticians have constructed confidence intervals for many different parameters based on a variety of designs for data collection. We will meet a number of these in later chapters. You need to know two important things about a confidence interval:

1. It is an interval of the form (a, b), where a and b are numbers computed from the data.
2. It has a property called a confidence level that gives the probability of producing an interval that contains the unknown parameter.

Users can choose the confidence level, but 95% is the standard for most situations. Occasionally, 90% or 99% is used. We will use C to stand for the confidence level in decimal form. For example, a 95% confidence level corresponds to $C = 0.95$.

CONFIDENCE INTERVAL

A level C **confidence interval** for a parameter is an interval computed from sample data by a method that has probability C of producing an interval containing the true value of the parameter.

USE YOUR KNOWLEDGE

6.4 **80% confidence intervals.** The idea of an 80% confidence interval is that the interval captures the true parameter value in 80% of all samples. That's not high enough confidence for practical use, but 80% hits and 20% misses make it easy to see how a confidence interval behaves in repeated samples from the same population.

APPLET

(a) Set the confidence level in the *Confidence Interval* applet to 80%. Click "Sample" to choose an SRS and calculate the confidence interval. Do this 10 times. How many of the 10 intervals captured the true mean μ? How many missed?

(b) You see that we can't predict whether the next sample will capture μ or miss. The confidence level, however, tells us what percent will capture μ in the long run. Reset the applet and click "Sample 50" to get the confidence intervals from 50 SRSs. How many hit? Keep clicking "Sample 50" and record the percent of hits among 100, 200, 300, 400, 500, 600, 700, 800, and 1000 SRSs. As the number of samples increases, we expect the percent of captures to get closer to the confidence level, 80%.

Confidence interval for a population mean

We will now construct a level C confidence interval for the mean μ of a population when the data are an SRS of size n. The construction is based on the sampling distribution of the sample mean $\bar{x}$. This distribution is exactly $N(\mu, \sigma/\sqrt{n})$ when the population has the $N(\mu, \sigma)$ distribution. The central limit theorem says that this same sampling distribution is approximately correct for large samples whenever the population mean and standard deviation are μ and σ.

LOOK BACK
central limit theorem, page 339

Our construction of a 95% confidence interval for the mean SATM score began by noting that any Normal distribution has probability about 0.95 within ±2 standard deviations of its mean. To construct a level C confidence interval we first catch the central C area under a Normal curve. That is, we must find the number z^* such that any Normal distribution has probability C within $\pm z^*$ standard deviations of its mean. Because all Normal distributions have the same standardized form, we can obtain everything we need from the standard Normal curve. Figure 6.4 shows how C and z^* are related. Values of z^* for many choices of C appear in the row labeled z^* at the bottom of Table D at the back of the book. Here are the most important entries from that row:

z^*	1.645	1.960	2.576
C	90%	95%	99%

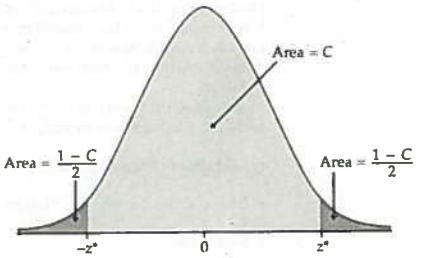


FIGURE 6.4 The area between $-z^*$ and z^* under the standard normal curve is C.

As Figure 6.4 reminds us, any Normal curve has probability C between the point z^* standard deviations below the mean and the point z^* standard deviations above the mean. The sample mean $\bar{x}$ has the Normal distribution with mean μ and standard deviation $\sigma/\sqrt{n}$. So there is probability C that $\bar{x}$ lies between

$$\mu - z^*\frac{\sigma}{\sqrt{n}} \quad \text{and} \quad \mu + z^*\frac{\sigma}{\sqrt{n}}$$

This is exactly the same as saying that the unknown population mean μ lies between

$$\bar{x} - z^*\frac{\sigma}{\sqrt{n}} \quad \text{and} \quad \bar{x} + z^*\frac{\sigma}{\sqrt{n}}$$

That is, there is probability C that the interval $\bar{x} \pm z^*\sigma/\sqrt{n}$ contains μ. That is our confidence interval. The estimate of the unknown μ is $\bar{x}$, and the margin of error is $z^*\sigma/\sqrt{n}$.

CONFIDENCE INTERVAL FOR A POPULATION MEAN

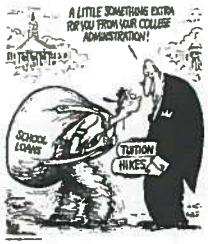
Choose an SRS of size n from a population having unknown mean μ and known standard deviation σ. The **margin of error** for a level C confidence interval for μ is

$$m = z^*\frac{\sigma}{\sqrt{n}}$$

Here z^* is the value on the standard Normal curve with area C between the critical points $-z^*$ and z^*. The level C **confidence interval** for μ is

$$\bar{x} \pm m$$

This interval is exact when the population distribution is Normal and is approximately correct when n is large in other cases.



EXAMPLE

6.4 Average debt of undergraduate borrowers. The National Student Loan Survey collects data to examine questions related to the amount of money that borrowers owe. The survey selected a sample of 1280 borrowers who began repayment of their loans between four and six months prior to the study.[2] The mean of the debt for undergraduate study was $18,900 and the standard deviation was about $49,000. This distribution is clearly skewed but because our sample size is quite large, we can rely on the central limit theorem to assure us that the confidence interval based on the Normal distribution will be a good approximation. Let's compute a 95% confidence interval for the true mean debt for all borrowers. Although the standard deviation is estimated from the data collected, we will treat it as a known quantity for our calculations here.

For 95% confidence, we see from Table D that $z^* = 1.960$. The margin of error for the 95% confidence interval for μ is therefore

$$\begin{aligned} m &= z^*\frac{\sigma}{\sqrt{n}} \\ &= 1.960\frac{49{,}000}{\sqrt{1280}} \\ &= 2684 \end{aligned}$$

We have computed the margin of error with more digits than we really need. Our mean is rounded to the nearest $100, so we will do the same for the margin of error. Keeping additional digits would provide no additional useful information. Therefore, we will use $m = 2700$. The 95% confidence interval is

$$\begin{aligned} \bar{x} \pm m &= 18{,}900 \pm 2700 \\ &= (16{,}200,\ 21{,}600) \end{aligned}$$

We are 95% confident that the mean debt for all borrowers is between \$16,200 and \$21,600.

Suppose the researchers who designed the National Student Loan Survey had used a different sample size. How would this affect the confidence interval? We can answer this question by changing the sample size in our calculations and assuming that the mean and standard deviation are the same.

EXAMPLE

6.5 How sample size affects the confidence interval. Let's assume that the sample mean of the debt for undergraduate study is \$18,900 and the standard deviation is about \$49,000, as in Example 6.4. But suppose that the sample size is only 320. The margin of error for 95% confidence is

$$m = z^* \frac{\sigma}{\sqrt{n}}$$
$$= 1.960 \frac{49{,}000}{\sqrt{320}}$$
$$= 5400$$

and the 95% confidence interval is

$$\bar{x} \pm m = 18{,}900 \pm 5400$$
$$= (13{,}500,\ 24{,}300)$$

Notice that the margin of error for this example is twice as large as the margin of error that we computed in Example 6.4. The only change that we made was to assume that the sample size is 320 rather than 1280. This sample size is exactly one-fourth of the original 1280. Thus, we double the margin of error when we reduce the sample size to one-fourth of the original value. Figure 6.5 illustrates the effect in terms of the intervals.

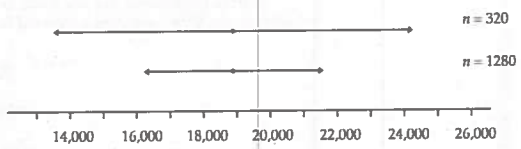


FIGURE 6.5 Confidence intervals for $n = 1280$ and $n = 320$, for Examples 6.4 and 6.5.

The argument leading to the form of confidence intervals for the population mean μ rested on the fact that the statistic $\bar{x}$ used to estimate μ has a Normal distribution. Because many sample estimates have Normal distributions (at least approximately), it is useful to notice that the confidence interval has the form

$$\text{estimate} \pm z^* \sigma_{\text{estimate}}$$

The estimate based on the sample is the center of the confidence interval. The margin of error is $z^* \sigma_{\text{estimate}}$. The desired confidence level determines z^* from Table D. The standard deviation of the estimate is found from a knowledge of the sampling distribution in a particular case. When the estimate is $\bar{x}$ from an SRS, the standard deviation of the estimate is $\sigma_{\text{estimate}} = \sigma/\sqrt{n}$.

USE YOUR KNOWLEDGE

6.5 College freshmen anxiety level. An SRS of 100 incoming freshmen was taken to look at their college anxiety level. The mean score of the sample was 83.5 (out of 100). Assuming a standard deviation of 4, give the 95% confidence interval for μ, the average anxiety level among all freshmen.

6.6 Changing the confidence level. In the setting of the previous exercise, would the margin of error for 99% confidence be larger or smaller? Verify your answer by performing the calculations.

How confidence intervals behave

The margin of error $z^*\sigma/\sqrt{n}$ for the mean of a Normal population illustrates several important properties that are shared by all confidence intervals in common use. The user chooses the confidence level, and the margin of error follows from this choice. High confidence is desirable and so is a small margin of error. High confidence says that our method almost always gives correct answers. A small margin of error says that we have pinned down the parameter quite precisely.

Suppose that you calculate a margin of error and decide that it is too large. Here are your choices to reduce it:

- Use a lower level of confidence (smaller C).
- Increase the sample size (larger n).
- Reduce σ.

For most problems you would choose a confidence level of 90%, 95%, or 99%. So z^* will be 1.645, 1.960, or 2.576. Figure 6.4 shows that z^* will be smaller for lower confidence (smaller C). The bottom row of Table D also shows this. If n and σ are unchanged, a smaller z^* leads to a smaller margin of error. Similarly, increasing the sample size n reduces the margin of error for any fixed confidence level. The square root in the formula implies that we must multiply the number of observations by 4 in order to cut the margin of error in half. The standard deviation σ measures the variation in the population. You can think of the variation among individuals in the population as noise that obscures the average value μ. It is harder to pin down the mean μ of a highly variable population; that is why the margin of error of a confidence interval increases with σ. In practice, we can sometimes reduce σ by carefully controlling the measurement process or by restricting our attention to only part of a large population.

EXAMPLE

6.6 How confidence level affects the confidence interval. Suppose that for the student loan data in Example 6.4, we wanted 99% confidence. Table D tells us that for 99% confidence, $z^* = 2.576$. The margin of error for 99% confidence based on 1280 observations is

$$m = z^* \frac{\sigma}{\sqrt{n}}$$
$$= 2.576 \frac{49{,}000}{\sqrt{1280}}$$
$$= 3500$$

and the 99% confidence interval is

$$\bar{x} \pm m = 18{,}900 \pm 3500$$
$$= (15{,}400,\ 22{,}400)$$

Requiring 99%, rather than 95%, confidence has increased the margin of error from 2700 to 3500. Figure 6.6 compares the two intervals.

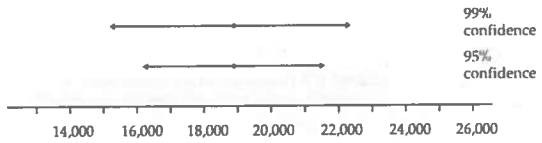


FIGURE 6.6 Confidence intervals for Examples 6.4 and 6.6.

Choosing the sample size

A wise user of statistics never plans data collection without at the same time planning the inference. You can arrange to have both high confidence and a small margin of error. The margin of error of the confidence interval for a population mean is

$$m = z^* \frac{\sigma}{\sqrt{n}}$$

To obtain a desired margin of error m, plug in the value of σ and the value of z^* for your desired confidence level, and solve for the sample size n. Here is the result.

SAMPLE SIZE FOR DESIRED MARGIN OF ERROR

The confidence interval for a population mean will have a specified margin of error m when the sample size is

$$n = \left(\frac{z^* \sigma}{m}\right)^2$$

a 95% confidence interval, we know that the probability that the interval we compute will cover the parameter is 0.95. That's the meaning of 95% confidence. If we use several such intervals, however, our confidence that *all* of them give correct results is less than 95%. Suppose we take independent samples each month for five months and report a 95% confidence interval for each set of data.

(a) What is the probability that all five intervals cover the true means? This probability (expressed as a percent) is our overall confidence level for the five simultaneous statements.

(b) What is the probability that at least four of the five intervals cover the true means?

6.34 **Telemarketing wages.** An advertisement in the student newspaper asks you to consider working for a telemarketing company. The ad states, "Earn between $500 and $1000 per week." Do you think that the ad is describing a confidence interval? Explain your answer.

6.35 **Like your job?** A Gallup Poll asked working adults about their job satisfaction. One question was "All in all, which best describes how you feel about your job?" The possible answers were "love job," "like job," "dislike job," and "hate job." Fifty-nine percent of the sample responded that they liked their job. Material provided with the results of the poll noted:

Results are based on telephone interviews with 1,001 national adults, aged 18 and older, conducted Aug. 8–11, 2005. For results based on the total sample of national adults, one can say with 95% confidence that the maximum margin of sampling error is ±3 percentage points.

The Gallup Poll uses a complex multistage sample design, but the sample percent has approximately a Normal sampling distribution.

(a) The announced poll result was 59% ± 3%. Can we be certain that the true population percent falls in this interval?

(b) Explain to someone who knows no statistics what the announced result 59% ± 3% means.

(c) This confidence interval has the same form we have met earlier:

$$\text{estimate} \pm z^* \sigma_{\text{estimate}}$$

What is the standard deviation σ_{estimate} of the estimated percent?

(d) Does the announced margin of error include errors due to practical problems such as undercoverage and nonresponse?

6.2 Tests of Significance

The confidence interval is appropriate when our goal is to estimate population parameters. The second common type of inference is directed at a quite different goal: to assess the evidence provided by the data in favor of some claim about the population parameters.

The reasoning of significance tests

A significance test is a formal procedure for comparing observed data with a hypothesis whose truth we want to assess. The hypothesis is a statement about the population parameters. The results of a test are expressed in terms of a probability that measures how well the data and the hypothesis agree. We use the following examples to illustrate these concepts.

EXAMPLE

6.8 Debt levels of private and public college borrowers. One purpose of the National Student Loan Survey described in Example 6.4 (page 361) is to compare the debt of different subgroups of students. For example, the 525 borrowers who last attended a private four-year college had a mean debt of $21,200, while those who last attended a public four-year college had a mean debt of $17,100. The difference of $4100 is fairly large, but we know that these

numbers are estimates of the true means. If we took different samples, we would get different estimates. Can we conclude from these data that the average debt of borrowers who attended a private college is different than the average debt of borrowers who attended a public college?

One way to answer this question is to compute the probability of obtaining a difference as large or larger than the observed $4100 assuming that, in fact, there is no difference in the true means. This probability is 0.17. Because this probability is not particularly small, we conclude that observing a difference of $4100 is not very surprising when the true means are equal. The data do not provide evidence for us to conclude that the mean debts for private four-year borrowers and public four-year borrowers are different.

Here is an example with a different conclusion.

EXAMPLE

6.9 Change in average debt levels between 1997 and 2002. Another purpose of the National Student Loan Survey is to look for changes over time. For example, in 1997, the survey found that the mean debt for undergraduate study was $11,400. How does this compare with the value of $18,900 in the 2002 study? The difference is $7500. As we learned in the previous example, an observed difference in means is not necessarily sufficient for us to conclude that the true means are different. Do the data provide evidence that there is an increase in borrowing? Again, we answer this question with a probability calculated under the assumption that there is *no difference in the true means*. The probability is 0.00004 of observing an increase in mean debt that is $7500 or more when there really is no difference. Because this probability is so small, we have sufficient evidence in the data to conclude that there has been a change in borrowing between 1997 and 2002.

What are the key steps in these examples?

- We started each with a question about the difference between two mean debts. In Example 6.8, we compare private four-year borrowers with public four-year borrowers. In Example 6.9, we compare borrowers in 2002 with borrowers in 1997. In both cases, we ask whether or not the data are compatible with no difference, that is, a difference of $0.
- Next we compared the data, $4100 in the first case and $7500 in the second, with the value that comes from the question, $0.
- The results of the comparisons are probabilities, 0.17 in the first case and 0.00004 in the second.

The 0.17 probability is not particularly small, so we have no evidence to question the possibility that the true difference is zero. In the second case, however, the probability is quite small. Something that happens with probability 0.00004 occurs only about 4 times out of 100,000. In this case we have two possible explanations:

1. we have observed something that is very unusual, or
2. the assumption that underlies the calculation, no difference in mean debt, is not true.

Because this probability is so small, we prefer the second conclusion: there has been a change in the mean debt between 1997 and 2002.

The probabilities in Examples 6.8 and 6.9 are measures of the compatibility of the data (a difference in means of $4100 and $7500) with the *null hypothesis* that there is no difference in the true means. Figures 6.7 and 6.8 compare the two results graphically. For each a Normal curve centered at 0 is the sampling distribution. You can see that we are not particularly surprised to observe the difference $4100 in Figure 6.7, but the difference $7500 in Figure 6.8 is clearly an unusual observation. We will now consider some of the formal aspects of significance testing.

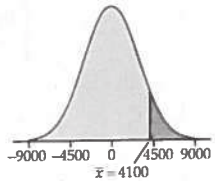


FIGURE 6.7 Comparison of the sample mean in Example 6.8 relative to the null hypothesized value 0.

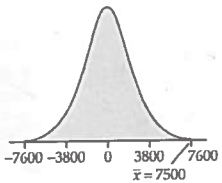


FIGURE 6.8 Comparison of the sample mean in Example 6.9 relative to the null hypothesized value 0.

Stating hypotheses

In Examples 6.8 and 6.9, we asked whether the difference in the observed means is reasonable if, in fact, there is no difference in the true means. To answer this, we begin by supposing that the statement following the "if" in the previous sentence is true. In other words, we suppose that the true difference is $0. We then ask whether the data provide evidence against the supposition we have made. If so, we have evidence in favor of an effect (the means are different) we are seeking. The first step in a test of significance is to state a claim that we will try to find evidence *against*.

NULL HYPOTHESIS

The statement being tested in a test of significance is called the null hypothesis. The test of significance is designed to assess the strength of the evidence against the null hypothesis. Usually the null hypothesis is a statement of "no effect" or "no difference."

We abbreviate "null hypothesis" as H_0. A null hypothesis is a statement about the population parameters. For example, our null hypothesis for Example 6.8 is

H_0: there is no difference in the true means

Note that the null hypothesis refers to the *true* means for all borrowers from either a four-year private or public college, including those for whom we do not have data.

alternative hypothesis

It is convenient also to give a name to the statement we hope or suspect is true instead of H_0. This is called the **alternative hypothesis** and is abbreviated as H_a. In Example 6.8, the alternative hypothesis states that the means are different. We write this as

H_a: the true means are not the same

Hypotheses always refer to some populations or a model, not to a particular outcome. For this reason, we must state H_0 and H_a in terms of population parameters.

Because H_a expresses the effect that we hope to find evidence *for*, we often begin with H_a and then set up H_0 as the statement that the hoped-for effect is not present. Stating H_a is often the more difficult task. It is not always clear, in particular, whether H_a should be **one-sided** or **two-sided**, which refers to whether a parameter differs from its null hypothesis value in a specific direction or in either direction.

one-sided or two-sided alternatives

The alternative hypothesis should express the hopes or suspicions we bring to the data. *It is cheating to first look at the data and then frame H_a to fit what the data show.* If you do not have a specific direction firmly in mind in advance, you must use a two-sided alternative. Moreover, some users of statistics argue that we should always use a two-sided alternative.

USE YOUR KNOWLEDGE

6.36 Food court survey. The food court at your dormitory has been redesigned. A survey is planned to determine whether or not students think that the new design is an improvement. Sampled students will respond on a seven-point scale with scores less than 4 favoring the old design and scores greater than 4 favoring the new design (to varying degrees). State the null and alternative hypotheses that provide a framework for examining whether or not the new design is an improvement.

6.37 DXA scanners. A dual-energy X-ray absorptiometry (DXA) scanner is used to measure bone mineral density for people who may be at risk for osteoporosis. To ensure its accuracy, the company uses an object called a "phantom" that has known mineral density $\mu = 1.4$ grams per square centimeter. Once installed, the company scans the phantom 10 times and compares the sample mean reading $\bar{x}$ with the theoretical mean μ using a significance test. State the null and alternative hypotheses for this test.

Test statistics

We will learn the form of significance tests in a number of common situations. Here are some principles that apply to most tests and that help in understanding these tests:

- The test is based on a statistic that estimates the parameter that appears in the hypotheses. Usually this is the same estimate we would use in a confidence interval for the parameter. When H_0 is true, we expect the estimate to take a value near the parameter value specified by H_0.
- Values of the estimate far from the parameter value specified by H_0 give evidence against H_0. The alternative hypothesis determines which directions count against H_0.
- To assess how far the estimate is from the parameter, standardize the estimate. In many common situations the test statistic has the form

$$z = \frac{\text{estimate} - \text{hypothesized value}}{\text{standard deviation of the estimate}}$$

Let's return to our student loan example.

EXAMPLE

6.10 Debt levels of private and public college borrowers: the hypotheses. In Example 6.8, the hypotheses are stated in terms of the difference in debt between borrowers who attended a private college and those who attended a public college:

H_0: there is no difference in the true means

H_a: there is a difference in the true means

Because H_a is two-sided, large values of both positive and negative differences count as evidence against the null hypothesis.

test statistic

A **test statistic** measures compatibility between the null hypothesis and the data. We use it for the probability calculation that we need for our test of significance. It is a random variable with a distribution that we know.

EXAMPLE

6.11 Debt levels of private and public college borrowers: the test statistic. In Example 6.8, we can state the null hypothesis as H_0: the true mean difference is 0. The estimate of the difference is \$4100. Using methods that we will discuss in detail later, we can determine that the standard deviation of the estimate is \$3000. For this problem the test statistic is

$$z = \frac{\text{estimate} - \text{hypothesized value}}{\text{standard deviation of the estimate}}$$

For our data,

$$z = \frac{4100 - 0}{3000} = 1.37$$

We have observed a sample estimate that is about one and a third standard deviations away from the hypothesized value of the parameter. Because the sample sizes are sufficiently large for us to conclude that the distribution of the sample estimate is approximately Normal, the standardized test statistic z will have approximately the $N(0, 1)$ distribution.

LOOK BACK
Normal distribution, page 58

We will use facts about the Normal distribution in what follows.

P-values

If all test statistics were Normal, we could base our conclusions on the value of the z test statistic. In fact, the Supreme Court of the United States has said that "two or three standard deviations" ($z = 2$ or 3) is its criterion for rejecting H_0 (see Exercise 6.42 on page 381), and this is the criterion used in most applications involving the law. Because not all test statistics are Normal, we translate the value of test statistics into a common language, the language of probability.

A test of significance finds the probability of getting an outcome *as extreme or more extreme than the actually observed outcome*. "Extreme" means "far from what we would expect if H_0 were true." The direction or directions that count as "far from what we would expect" are determined by H_a and H_0.

In Example 6.8 we want to know if the debt of private college borrowers is different from the debt of public college borrowers. The difference we calculated based on our sample is \$4100, which corresponds to 1.37 standard deviations away from zero—that is, $z = 1.37$. Because we are using a two-sided alternative for this problem, the evidence against H_0 is measured by the probability that we observe a value of Z as extreme or more extreme than 1.37. More formally, this probability is

$$P(Z \le -1.37 \text{ or } Z \ge 1.37)$$

where Z has the standard Normal distribution $N(0, 1)$.

P-VALUE

The probability, assuming H_0 is true, that the test statistic would take a value as extreme or more extreme than that actually observed is called the ***P*-value** of the test. The smaller the *P*-value, the stronger the evidence against H_0 provided by the data.

The key to calculating the *P*-value is the sampling distribution of the test statistic. For the problems we consider in this chapter, we need only the standard Normal distribution for the test statistic z.

EXAMPLE

6.12 Debt levels of private and public college borrowers: the *P*-value. In Example 6.11 we found that the test statistic for testing

H_0: the true mean difference is 0

versus

H_a: there is a difference in the true means

is

$$z = \frac{4100 - 0}{3000} = 1.37$$

If H_0 is true, then z is a single observation from the standard Normal, $N(0, 1)$, distribution. Figure 6.9 illustrates this calculation. The *P*-value is the probability of observing a value of Z at least as extreme as the one that we observed, $z = 1.37$. From Table A, our table of standard Normal probabilities, we find

$$P(Z \ge 1.37) = 1 - 0.9147 = 0.0853$$

The probability for being extreme in the negative direction is the same:

$$P(Z \le -1.37) = 0.0853$$

So the *P*-value is

$$P = 2P(Z \ge 1.37) = 2(0.0853) = 0.1706$$

This is the value that was reported on page 373. There is a 17% chance of observing a difference as extreme as the \$4100 in our sample if the true population difference is zero. The *P*-value tells us that our outcome is not particularly extreme, so we conclude that the data do not provide evidence that would cause us to doubt the validity of the null hypothesis.

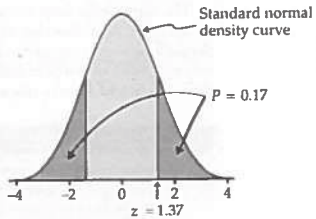


FIGURE 6.9 The *P*-value for Example 6.12. The *P*-value here is the probability (when H_0 is true) that $\bar{x}$ takes a value as extreme or more extreme than the actual observed value.

USE YOUR KNOWLEDGE

6.38 Normal curve and the *P*-value. A test statistic for a two-sided significance test for a population mean is $z = 2.7$. Sketch a standard Normal curve and mark this value of z on it. Find the *P*-value and shade the appropriate areas under the curve to illustrate your calculations.

6.39 More on the Normal curve and the *P*-value. A test statistic for a two-sided significance test for a population mean is $z = -1.2$. Sketch a standard Normal curve and mark this value of z on it. Find the *P*-value and shade the appropriate areas under the curve to illustrate your calculations.

Statistical significance

We started our discussion of the reasoning of significance tests with the statement of null and alternative hypotheses. We then learned that a test statistic is the tool used to examine the compatibility of the observed data with the null hypothesis. Finally, we translated the test statistic into a *P*-value to quantify the evidence against H_0. One important final step is needed: to state our conclusion.

We can compare the *P*-value we calculated with a fixed value that we regard as decisive. This amounts to announcing in advance how much evidence against H_0 we will require to reject H_0. The decisive value of P is called the **significance level.** It is commonly denoted by α. If we choose $\alpha = 0.05$, we are requiring that the data give evidence against H_0 so strong that it would happen no more than 5% of the time (1 time in 20) when H_0 is true. If we choose $\alpha = 0.01$, we are insisting on stronger evidence against H_0, evidence so strong that it would appear only 1% of the time (1 time in 100) if H_0 is in fact true.

significance level

STATISTICAL SIGNIFICANCE

If the *P*-value is as small or smaller than α, we say that the data are **statistically significant at level α.**

"Significant" in the statistical sense does not mean "important." The original meaning of the word is "signifying something." In statistics the term is used to indicate only that the evidence against the null hypothesis reached the standard set by α. Significance at level 0.01 is often expressed by the statement "The results were significant ($P < 0.01$)." Here P stands for the *P*-value. The *P*-value is more informative than a statement of significance because we can then assess significance at any level we choose. For example, a result with $P = 0.03$ is significant at the $\alpha = 0.05$ level but is not significant at the $\alpha = 0.01$ level.

A test of significance is a process for assessing the significance of the evidence provided by data against a null hypothesis. The four steps common to all tests of significance are as follows:

1. State the *null hypothesis* H_0 and the *alternative hypothesis* H_a. The test is designed to assess the strength of the evidence against H_0; H_a is the statement that we will accept if the evidence enables us to reject H_0.
2. Calculate the value of the *test statistic* on which the test will be based. This statistic usually measures how far the data are from H_0.
3. Find the *P-value* for the observed data. This is the probability, calculated assuming that H_0 is true, that the test statistic will weigh against H_0 at least as strongly as it does for these data.

4. State a conclusion. One way to do this is to choose a *significance level* α, how much evidence against H_0 you regard as decisive. If the *P*-value is less than or equal to α, you conclude that the alternative hypothesis is true; if it is greater than α, you conclude that the data do not provide sufficient evidence to reject the null hypothesis. Your conclusion is a sentence that summarizes what you have found by using a test of significance.

We will learn the details of many tests of significance in the following chapters. The proper test statistic is determined by the hypotheses and the data collection design. We use computer software or a calculator to find its numerical value and the *P*-value. The computer will not formulate your hypotheses for you, however. Nor will it decide if significance testing is appropriate or help you to interpret the *P*-value that it presents to you. The most difficult and important step is the last one: stating a conclusion.

EXAMPLE

6.13 Debt levels of private and public college borrowers: significance. In Example 6.12 we found that the *P*-value is 0.1706. There is a 17% chance of observing a difference as extreme as the $4100 in our sample if the true population difference is zero. The *P*-value tells us that our outcome is not particularly extreme. We could report the result as "the data do not provide evidence that would cause us to conclude that there is a difference in student loan debt between private college borrowers and public college borrowers ($z = 1.37$, $P = 0.17$)."

If the *P*-value is small, we reject the null hypothesis. Here is an example.

EXAMPLE

6.14 Change in mean debt levels: significance. In Example 6.9 we found that the average debt has risen by $7500 from 1997 to 2002. Since we would have a prior expectation that the debt would increase over this period because of rising costs of a college education, it is appropriate to use a one-sided alternative in this situation. So, our hypotheses are

H_0: the true mean difference is 0

versus

H_a: the mean debt has increased between 1997 and 2002

The standard deviation is $1900 (again we defer details regarding this calculation), and the test statistic is

$$z = \frac{\text{estimate} - \text{hypothesized value}}{\text{standard deviation of the estimate}}$$

$$z = \frac{7500 - 0}{1900}$$

$$= 3.95$$

Because only increases in debt count against the null hypothesis, the one-sided alternative leads to the calculation of the *P*-value using the upper tail

of the Normal distribution. The *P*-value is

$$P = P(Z \geq 3.95)$$
$$= 0.00004$$

The calculation is illustrated in Figure 6.10. There is about a 4 in 100,000 chance of observing a difference as large or larger than the $7500 in our sample if the true population difference is zero. This *P*-value tells us that our outcome is extremely rare. We conclude that the null hypothesis must be false. Here is one way to report the result: "The data clearly show that the mean debt for college loans has increased between 1997 and 2002 ($z = 3.95$, $P < 0.001$)."

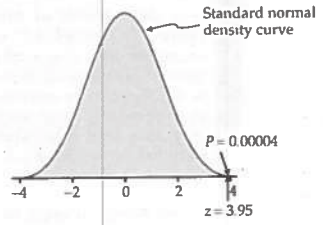


FIGURE 6.10 The *P*-value for Example 6.14. The *P*-value here is the probability (when H_0 is true) that $\bar{x}$ takes a value as large or larger than the actual observed value.

Note that the calculated *P*-value for this example is 0.00004 but we reported the result as $P < 0.001$. The value 0.001, 1 in 1000, is sufficiently small to force a clear rejection of H_0. Standard practice is to report very small *P*-values as simply less than 0.001.

USE YOUR KNOWLEDGE

6.40 Finding significant z-scores. Consider a significance test of the true mean based on an SRS of 30 observations from a Normal population. The alternative hypothesis is that the true mean is different from 1000. What values of the z statistic are statistically significant at the $\alpha = 0.05$ level?

6.41 More on finding significant z-scores. Consider a significance test of the true mean based on an SRS of 30 observations from a Normal population. The alternative hypothesis is that the true mean is larger than 1000. What values of the z statistic are statistically significant at the $\alpha = 0.05$ level?

6.42 The Supreme Court speaks. The Supreme Court has said that z-scores beyond $z^* = 2$ or 3 are generally convincing statistical evidence. For a two-sided test, what significance level corresponds to $z^* = 2$? To $z^* = 3$?

Tests for a population mean

Our discussion has focused on the reasoning of statistical tests, and we have outlined the key ideas for one type of procedure. Here is a summary. We want to test the hypothesis that a parameter has a specified value. This is the null hypothesis. For a test of a population mean μ, the null hypothesis is

$$H_0: \text{the true population mean is equal to } \mu_0$$

which often is expressed as

$$H_0: \mu = \mu_0$$

where μ_0 is the specified value of μ that we would like to examine.

The test is based on data summarized as an estimate of the parameter. For a population mean this is the sample mean $\bar{x}$. Our test statistic measures the difference between the sample estimate and the hypothesized parameter in terms of standard deviations of the test statistic:

$$z = \frac{\text{estimate} - \text{hypothesized value}}{\text{standard deviation of the estimate}}$$

LOOK BACK
distribution of sample mean, page 339
central limit theorem, page 339

Recall from Chapter 5 that the standard deviation of $\bar{x}$ is $\sigma/\sqrt{n}$. Therefore, the test statistic is

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

Again recall from Chapter 5 that, if the population is Normal, then $\bar{x}$ will be Normal and z will have the standard Normal distribution when H_0 is true. By the central limit theorem both distributions will be approximately Normal when the sample size is large even if the population is not Normal.

Suppose we have calculated a test statistic $z = 1.7$. If the alternative is one-sided on the high side, then the *P*-value is the probability that a standard Normal random variable Z takes a value as large or larger than the observed 1.7. That is,

$$P = P(Z \geq 1.7)$$
$$= 1 - P(Z < 1.7)$$
$$= 1 - 0.9554$$
$$= 0.0446$$

Similar reasoning applies when the alternative hypothesis states that the true μ lies below the hypothesized μ_0 (one-sided). When H_a states that μ is simply unequal to μ_0 (two-sided), values of z away from zero in either direction count against the null hypothesis. The *P*-value is the probability that a standard Normal Z is at least as far from zero as the observed z. Again, if the test statistic is $z = 1.7$, the two-sided *P*-value is the probability that $Z \leq -1.7$ or $Z \geq 1.7$. Because the standard Normal distribution is symmetric, we calculate this probability by finding $P(Z \geq 1.7)$ and *doubling* it:

$$P(Z \leq -1.7 \text{ or } Z \geq 1.7) = 2P(Z \geq 1.7)$$
$$= 2(1 - 0.9554) = 0.0892$$

We would make exactly the same calculation if we observed $z = -1.7$. It is the absolute value $|z|$ that matters, not whether z is positive or negative. Here is a statement of the test in general terms.

z TEST FOR A POPULATION MEAN

To test the hypothesis $H_0: \mu = \mu_0$ based on an SRS of size n from a population with unknown mean μ and known standard deviation σ, compute the test statistic

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

In terms of a standard Normal random variable Z, the P-value for a test of H_0 against

$H_a: \mu > \mu_0$ is $P(Z \geq z)$

$H_a: \mu < \mu_0$ is $P(Z \leq z)$

$H_a: \mu \neq \mu_0$ is $2P(Z \geq |z|)$

These P-values are exact if the population distribution is Normal and are approximately correct for large n in other cases.

EXAMPLE

6.15 Cholesterol level of sedentary female undergraduates. In 1999, it was reported that the mean serum cholesterol level for female undergraduates was 168 mg/dl with a standard deviation of 27 mg/dl. A recent study at Baylor University investigated the lipid levels in a cohort of sedentary university students.[12] The mean total cholesterol level among $n = 71$ females was $\bar{x} = 173.7$. Is this evidence that cholesterol levels of sedentary students differ from the previously reported average?

The null hypothesis is "no difference" from the published mean $\mu_0 = 168$. The alternative is two-sided because the researcher did not have a particular direction in mind before examining the data. So the hypotheses about the unknown mean μ of the sedentary population are

$$H_0: \mu = 168$$
$$H_a: \mu \neq 168$$

As usual in this chapter, we make the unrealistic assumption that the population standard deviation is known, in this case that sedentary female students have the same $\sigma = 27$ as the general population of female undergraduates. The z test requires that the 71 students in the sample are an SRS from the population of all sedentary female students. We check this assumption by asking how the data were produced. In this case, all participants were enrolled in a health class at Baylor, so there may be some concerns about whether the sample is an SRS. We will press on for now.

We compute the test statistic:

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{173.7 - 168}{27/\sqrt{71}}$$
$$= 1.78$$

Figure 6.11 illustrates the P-value, which is the probability that a standard Normal variable Z takes a value at least 1.78 away from zero. From Table A we find that this probability is

$$P = 2P(Z \geq 1.78) = 2(1 - 0.9625) = 0.075$$

That is, more than 7% of the time an SRS of size 71 from the general undergraduate female population would have a mean cholesterol level at least as far from 168 as that of the sedentary sample. The observed $\bar{x} = 173.7$ is therefore not strong evidence that the sedentary female undergraduate population differs from the general female undergraduate population.

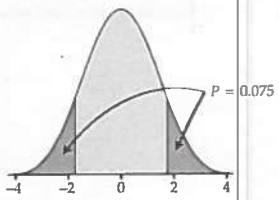


FIGURE 6.11 The P-value for the two-sided test in Example 6.15.

The data in Example 6.15 do *not* establish that the mean cholesterol level μ for the sedentary population is 168. We sought evidence that μ differed from 168 and failed to find convincing evidence. That is all we can say. No doubt the mean cholesterol level of the entire sedentary population is not exactly equal to 168. A large enough sample would give evidence of the difference, even if it is very small. Tests of significance assess the evidence *against* H_0. If the evidence is strong, we can confidently reject H_0 in favor of the alternative. Failing to find evidence against H_0 means only that the data are consistent with H_0, not that we have clear evidence that H_0 is true.

EXAMPLE

6.16 Significance test of the mean SATM score. In a discussion of SAT Mathematics (SATM) scores, someone comments: "Because only a minority of California high school students take the test, the scores overestimate the ability of typical high school seniors. I think that if all seniors took the test, the mean score would be no more than 450." You decided to test this claim (H_0) and gave the SAT to an SRS of 500 seniors from California (Example 6.3). These students had a mean SATM score of $\bar{x} = 461$. Is this good evidence against this claim? Because the claim states the mean is "no more than 450," the alternative hypothesis is one-sided. The hypotheses are

$$H_0: \mu = 450$$
$$H_a: \mu > 450$$

As we did in the discussion following Example 6.3, we assume that $\sigma = 100$. The z statistic is

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{461 - 450}{100/\sqrt{500}}$$
$$= 2.46$$

Because H_a is one-sided on the high side, large values of z count against H_0. From Table A, we find that the P-value is

$$P = P(Z \geq 2.46) = 1 - 0.9931 = 0.0069$$

Figure 6.12 illustrates this P-value. A mean score as large as that observed would occur fewer than seven times in 1000 samples if the population mean were 450. This is convincing evidence that the mean SATM score for all California high school seniors is higher than 450.

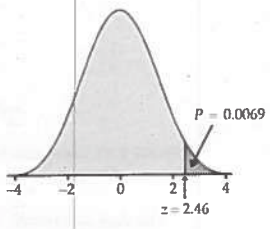


FIGURE 6.12 The P-value for the one-sided test in Example 6.16.

USE YOUR KNOWLEDGE

6.43 Computing the test statistic and P-value. You will perform a significance test of $H_0: \mu = 25$ based on an SRS of $n = 25$. Assume $\sigma = 5$.

(a) If $\bar{x} = 27$, what is the test statistic z?

(b) What is the P-value if $H_a: \mu > 25$?

(c) What is the P-value if $H_a: \mu \neq 25$?

6.44 Testing a random number generator. Statistical software has a "random number generator" that is supposed to produce numbers uniformly distributed between 0 to 1. If this is true, the numbers generated come from a population with $\mu = 0.5$. A command to generate 100 random numbers gives outcomes with mean $\bar{x} = 0.522$ and $s = 0.316$. Because the sample is reasonably large, take the population standard deviation also to be $\sigma = 0.316$. Do we have evidence that the mean of all numbers produced by this software is not 0.5?

Two-sided significance tests and confidence intervals

Recall the basic idea of a confidence interval, discussed in the first section of this chapter. We constructed an interval that would include the true value of μ with a specified probability C. Suppose we use a 95% confidence interval ($C = 0.95$). Then the values of μ that are not in our interval would seem to be incompatible with the data. This sounds like a significance test with $\alpha = 0.05$ (or 5%) as our standard for drawing a conclusion. The following examples demonstrate that this is correct.

EXAMPLE

6.17 Testing a pharmaceutical product. The Deely Laboratory analyzes specimens of a pharmaceutical product to determine the concentration of the active ingredient. Such chemical analyses are not perfectly precise. Repeated measurements on the same specimen will give slightly different results. The results of repeated measurements follow a Normal distribution quite closely. The analysis procedure has no bias, so that the mean μ of the population of all measurements is the true concentration in the specimen. The standard deviation of this distribution is a property of the analytical procedure and is known to be $\sigma = 0.0068$ grams per liter. The laboratory analyzes each specimen three times and reports the mean result.

The Deely Laboratory has been asked to evaluate the claim that the concentration of the active ingredient in a specimen is 0.86 grams per liter. The true concentration is the mean μ of the population of repeated analyses. The hypotheses are

$$H_0: \mu = 0.86$$
$$H_a: \mu \neq 0.86$$

The lab chooses the 1% level of significance, $\alpha = 0.01$.

Three analyses of one specimen give concentrations

0.8403 0.8363 0.8447

The sample mean of these readings is

$$\bar{x} = \frac{0.8403 + 0.8363 + 0.8447}{3} = 0.8404$$

CHAPTER

7

Inference for Distributions

Some people feel that a full moon causes strange and aggressive behavior in people. Is there any scientific evidence to support this? Example 7.7 describes one such study.

Introduction

We began our study of data analysis in Chapter 1 by learning graphical and numerical tools for describing the distribution of a single variable and for comparing several distributions. Our study of the practice of statistical inference begins in the same way, with inference about a single distribution and comparison of two distributions. Comparing more than two distributions requires more elaborate methods, which are presented in Chapters 12 and 13.

Two important aspects of any distribution are its center and spread. If the distribution is Normal, we describe its center by the mean μ and its spread by the standard deviation σ. In this chapter, we will meet confidence intervals and significance tests for inference about a population mean μ and for comparing the means or spreads of two populations. The previous chapter emphasized the reasoning of tests and confidence intervals; now we emphasize statistical practice, so we no longer assume that population standard deviations are known. The t procedures for inference about means are among the most common statistical methods. Inference about the spreads, as we will see, poses some difficult practical problems.

The methods in this chapter will allow us to address questions like:

- Does cellular phone use, specifically the number of hours listening to music tracks, differ between cell phone users in the United States and the United Kingdom?
- Do male and female college students differ in terms of "social insight," the ability to appraise other people?
- Does the daily number of disruptive behaviors in dementia patients change when there is a full moon?

7.1 Inference for the Mean of a Population

LOOK BACK
sampling distribution of $\bar{x}$, page 339

Both confidence intervals and tests of significance for the mean μ of a Normal population are based on the sample mean $\bar{x}$, which estimates the unknown μ. The sampling distribution of $\bar{x}$ depends on σ. This fact causes no difficulty when σ is known. When σ is unknown, however, we must estimate σ even though we are primarily interested in μ. The sample standard deviation s is used to estimate the population standard deviation σ.

The t distributions

Suppose that we have a simple random sample (SRS) of size n from a Normally distributed population with mean μ and standard deviation σ. The sample mean $\bar{x}$ is then Normally distributed with mean μ and standard deviation $\sigma/\sqrt{n}$. When σ is not known, we estimate it with the sample standard deviation s, and then we estimate the standard deviation of $\bar{x}$ by $s/\sqrt{n}$. This quantity is called the *standard error* of the sample mean $\bar{x}$ and we denote it by $\mathrm{SE}_{\bar{x}}$.

STANDARD ERROR

When the standard deviation of a statistic is estimated from the data, the result is called the **standard error** of the statistic. The standard error of the sample mean is

$$\mathrm{SE}_{\bar{x}} = \frac{s}{\sqrt{n}}$$

The term "standard error" is sometimes used for the actual standard deviation of a statistic. The estimated value is then called the "estimated standard error." In this book we will use the term "standard error" only when the standard deviation of a statistic is estimated from the data. The term has this meaning in the output of many statistical computer packages and in research reports that apply statistical methods.

The standardized sample mean, or one-sample z statistic,

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

is the basis of the z procedures for inference about μ when σ is known. This statistic has the standard Normal distribution $N(0, 1)$. When we substitute the standard error $s/\sqrt{n}$ for the standard deviation $\sigma/\sqrt{n}$ of $\bar{x}$, the statistic does *not* have a Normal distribution. It has a distribution that is new to us, called a *t distribution*.

THE t DISTRIBUTIONS

Suppose that an SRS of size n is drawn from an $N(\mu, \sigma)$ population. Then the **one-sample t statistic**

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

has the **t distribution with $n - 1$ degrees of freedom.**

LOOK BACK
degrees of freedom, page 42

A particular t distribution is specified by giving the *degrees of freedom*. We use $t(k)$ to stand for the t distribution with k degrees of freedom. The degrees of freedom for this t statistic come from the sample standard deviation s in the denominator of t. We showed earlier that s has $n - 1$ degrees of freedom. Thus, there is a different t distribution for each sample size. There are also other t statistics with different degrees of freedom, some of which we will meet later in this chapter.

The t distributions were discovered in 1908 by William S. Gosset. Gosset was a statistician employed by the Guinness brewing company, which prohibited its employees from publishing their discoveries that were brewing related. In this case, the company let him publish under the pen name "Student" using an example that did not involve brewing. The t distribution is often called "Student's t" in his honor.

The density curves of the $t(k)$ distributions are similar in shape to the standard Normal curve. That is, they are symmetric about 0 and are bell-shaped. Figure 7.1 compares the density curves of the standard Normal distribution and the t distributions with 5 and 10 degrees of freedom. The similarity in shape is apparent, as is the fact that the t distributions have more probability in the tails and less in the center. This greater spread is due to the extra variability caused by substituting the random variable s for the fixed parameter σ. Figure 7.1 also shows that as the degrees of freedom k increase, the $t(k)$ density curve gets closer to the $N(0, 1)$ curve. This reflects the fact that s will likely be closer to σ as the sample size increases.

Table D in the back of the book gives critical values t^* for the t distributions. For convenience, we have labeled the table entries both by the value of p needed for significance tests and by the confidence level C (in percent) required for confidence intervals. The standard Normal critical values are in the bottom row of entries and labeled z^*. As in the case of the Normal table (Table A), computer software often makes Table D unnecessary.

USE YOUR KNOWLEDGE

7.1 Apartment rents. You randomly choose 15 unfurnished one-bedroom apartments from a large number of advertisements in your local newspaper. You calculate that their mean monthly rent is \$570 and their standard deviation is \$105.

(a) What is the standard error of the mean?

(b) What are the degrees of freedom for a one-sample t statistic?

7.2 Finding critical t^* values. What critical value t^* from Table D should be used to construct

(a) a 95% confidence interval when $n = 12$?

(b) a 99% confidence interval when $n = 24$?

(c) a 90% confidence interval when $n = 200$?

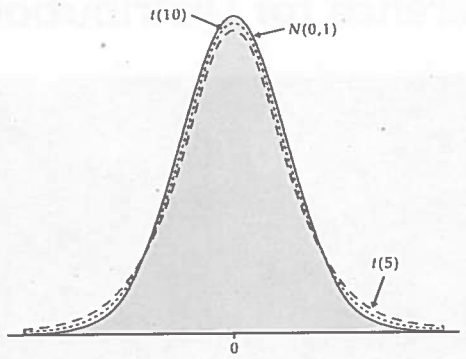


FIGURE 7.1 Density curves for the standard Normal, $t(10)$, and $t(5)$ distributions. All are symmetric with center 0. The t distributions have more probability in the tails than the standard Normal distribution.

The one-sample t confidence interval

LOOK BACK
z confidence interval, page 361

With the t distributions to help us, we can now analyze a sample from a Normal population with unknown σ. The one-sample t confidence interval is similar in both reasoning and computational detail to the z confidence interval of Chapter 6. There, the margin of error for the population mean was $z^*\sigma/\sqrt{n}$. Here, we replace σ by its estimate s and z^* by t^*. This means that the margin of error for the population mean when we use the data to estimate σ is $t^*s/\sqrt{n}$.

THE ONE-SAMPLE t CONFIDENCE INTERVAL

Suppose that an SRS of size n is drawn from a population having unknown mean μ. A level C **confidence interval** for μ is

$$\bar{x} \pm t^*\frac{s}{\sqrt{n}}$$

where t^* is the value for the $t(n-1)$ density curve with area C between $-t^*$ and t^*. The quantity

$$t^* \frac{s}{\sqrt{n}}$$

is the **margin of error.** This interval is exact when the population distribution is Normal and is approximately correct for large n in other cases.

EXAMPLE

7.1 Listening to music on cell phones. Founded in 1998, Telephia provides a wide variety of information on cellular phone use. In 2006, Telephia reported that, on average, United Kingdom (U.K.) subscribers with third-generation technology (3G) phones spent an average of 8.3 hours per month listening to full-track music on their cell phones.[1] Suppose we want to determine a 95% confidence interval for the U.S. average and draw the following random sample of size 8 from the U.S. population of 3G subscribers:

5 6 0 4 11 9 2 3

The sample mean is $\bar{x} = 5$ and the standard deviation is $s = 3.63$ with degrees of freedom $n - 1 = 7$. The standard error is

$$\text{SE}_{\bar{x}} = s/\sqrt{n} = 3.63/\sqrt{8} = 1.28$$

From Table D we find $t^* = 2.365$. The 95% confidence interval is

$$\begin{aligned} \bar{x} \pm t^* \frac{s}{\sqrt{n}} &= 5.0 \pm 2.365 \frac{3.63}{\sqrt{8}} \\ &= 5.0 \pm (2.365)(1.28) \\ &= 5.0 \pm 3.0 \\ &= (2.0, 8.0) \end{aligned}$$

We are 95% confident that the U.S. population's average time spent listening to full-track music on a cell phone is between 2.0 and 8.0 hours per month. Since this interval does not contain 8.3 hours, these data suggest that, on average, a U.S. subscriber listens to less full-track music.

In this example we have given the actual interval (2.0, 8.0) as our answer. Sometimes we prefer to report the mean and margin of error: the mean time is 5.0 hours per month with a margin of error of 3.0 hours.

The use of the t confidence interval in Example 7.1 rests on assumptions that appear reasonable here. First, we assume our random sample is an SRS from the U.S. population of cell phone users. Second, we assume the distribution of listening times is Normal. With only 8 observations, this assumption cannot be effectively checked. In fact, because the listening time cannot be negative, we might expect this distribution to be skewed to the right. With these data, however, there are no extreme outliers to suggest a severe departure from Normality.

USE YOUR KNOWLEDGE

7.3 More on apartment rents. Recall Exercise 7.1 (page 419). Construct a 95% confidence interval for the mean monthly rent of all advertised one-bedroom apartments.

7.4 90% versus 95% confidence interval. If you were to use 90% confidence, rather than 95% confidence, would the margin of error be larger or smaller? Explain your answer.

The one-sample t test

LOOK BACK
z significance test, page 383

Significance tests using the standard error are also very similar to the z test that we studied in the last chapter.

THE ONE-SAMPLE t TEST

Suppose that an SRS of size n is drawn from a population having unknown mean μ. To test the hypothesis $H_0: \mu = \mu_0$ based on an SRS of size n, compute the one-sample t statistic

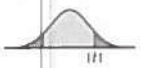
$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

In terms of a random variable T having the $t(n-1)$ distribution, the P-value for a test of H_0 against

$H_a: \mu > \mu_0$ is $P(T \geq t)$

$H_a: \mu < \mu_0$ is $P(T \leq t)$

$H_a: \mu \neq \mu_0$ is $2P(T \geq |t|)$



These P-values are exact if the population distribution is Normal and are approximately correct for large n in other cases.

EXAMPLE

7.2 Significance test for cell phone use. Suppose that, for the U.S. data in Example 7.1, we want to test whether the U.S. average is different from the reported U.K. average. Specifically, we want to test

$$H_0: \mu = 8.3$$
$$H_a: \mu \neq 8.3$$

df = 7

p	0.02	0.01
t^*	2.517	2.998

Recall that $n = 8$, $\bar{x} = 5.0$, and $s = 3.63$. The t test statistic is

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{5.0 - 8.3}{3.63/\sqrt{8}}$$
$$= -2.57$$

This means that the sample mean $\bar{x} = 5.0$ is slightly over 2.5 standard deviations away from the null hypothesized value $\mu = 8.3$. Because the degrees of freedom are $n - 1 = 7$, this t statistic has the $t(7)$ distribution. Figure 7.2 shows that the P-value is $2P(T \geq 2.57)$, where T has the $t(7)$ distribution. From Table D we see that $P(T \geq 2.517) = 0.02$ and $P(T \geq 2.998) = 0.01$. Therefore, we conclude that the P-value is between $2 \times 0.01 = 0.02$ and $2 \times 0.02 = 0.04$. Software gives the exact value as $P = 0.037$. These data are incompatible with a mean of 8.3 hours per month at the $\alpha = 0.05$ level.

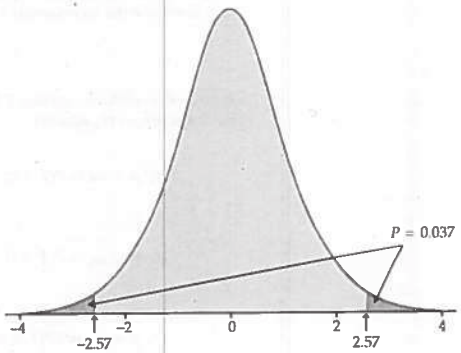


FIGURE 7.2 The P-value for Example 7.2.

In this example we tested the null hypothesis $\mu = 8.3$ hours per month against the two-sided alternative $\mu \neq 8.3$ hours per month because we had no prior suspicion that the average in the United States would be larger or smaller. If we had suspected that the U.S. average would be smaller, we would have used a one-sided test. *It is wrong, however, to examine the data first and then decide to do a one-sided test in the direction indicated by the data.* If in doubt, use a two-sided test. In the present circumstance, however, we could use our results from Example 7.2 to justify a one-sided test for *another* sample from the same population.

EXAMPLE

7.3 One-sided test for cell phone use. For the cell phone problem described in the previous example, we want to test whether the U.S. average is smaller than the U.K. average. Here we test

$$H_0: \mu = 8.3$$

versus

$$H_a: \mu < 8.3$$

The t test statistic does not change: $t = -2.57$. As Figure 7.3 illustrates, however, the P-value is now $P(T \leq -2.57)$, half of the value in the previous example. From Table D we can determine that $0.01 < P < 0.02$; software gives the exact value as $P = 0.0185$. At the $\alpha = 0.05$ level, we conclude that the U.S. average is smaller than the U.K. average.

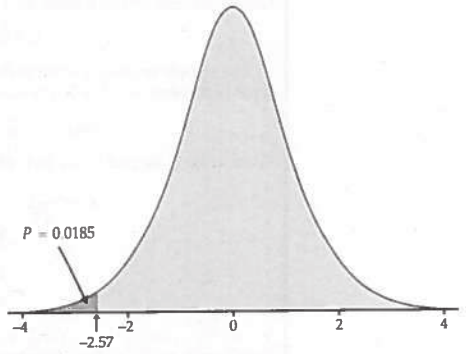


FIGURE 7.3 The P-value for Example 7.3.

For small data sets, such as the one in Example 7.1, it is easy to perform the computations for confidence intervals and significance tests with an ordinary calculator. For larger data sets, however, we prefer to use software or a statistical calculator.

EXAMPLE

7.4 Stock portfolio diversification? An investor with a stock portfolio worth several hundred thousand dollars sued his broker and brokerage firm because lack of diversification in his portfolio led to poor performance. Table 7.1 gives the rates of return for the 39 months that the account was managed by the broker.[2] Figure 7.4 gives a histogram for these data and Figure 7.5 gives the Normal quantile plot. There are no outliers and the distribution shows no strong skewness. We are reasonably confident that the

TABLE 7.1

Monthly rates of return on a portfolio (percent)

−8.36	1.63	−2.27	−2.93	−2.70	−2.93	−9.14	−2.64
6.82	−2.35	−3.58	6.13	7.00	−15.25	−8.66	−1.03
−9.16	−1.25	−1.22	−10.27	−5.11	−0.80	−1.44	1.28
−0.65	4.34	12.22	−7.21	−0.09	7.34	5.04	−7.24
−2.14	−1.01	−1.41	12.03	−2.56	4.33	2.35	

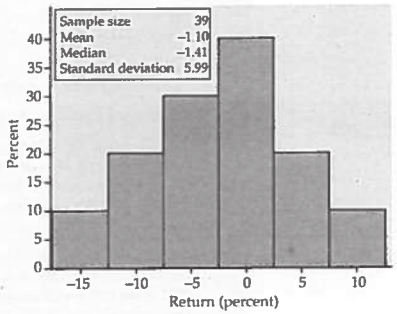


FIGURE 7.4 Histogram for Example 7.4.

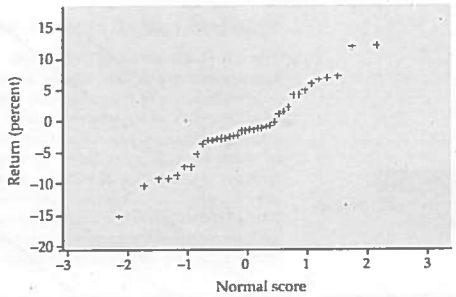


FIGURE 7.5 Normal quantile plot for Example 7.4.

distribution of $\bar{x}$ is approximately Normal, and we proceed with our inference based on Normal theory.

The arbitration panel compared these returns with the average of the Standard and Poor's 500 stock index for the same period. Consider the 39 monthly returns as a random sample from the population of monthly returns the

brokerage would generate if it managed the account forever. Are these returns compatible with a population mean of $\mu = 0.95\%$, the S&P 500 average? Our hypotheses are

$$H_0: \mu = 0.95$$
$$H_a: \mu \neq 0.95$$

Minitab and SPSS outputs appear in Figure 7.6. Output from other software will look similar.

Here is one way to report the conclusion: the mean monthly return on investment for this client's account was $\bar{x} = -1.1\%$. This differs significantly from the performance of the S&P 500 stock index for the same period ($t = -2.14$, df = 38, $P < 0.039$).

Minitab

```
Test of mu = 0.950 vs mu not = 0.950

Variable   N     Mean     StDev    SE Mean        T        P
RETURN    39   -1.100     5.991      0.959    -2.14    0.039
```

SPSS

One-Sample Test

	Test Value = 0.95					
					95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
RETURN	−2.137	38	.039	−2.0497	−3.9918	−.1077

FIGURE 7.6 Minitab and SPSS output for Example 7.4.

The hypothesis test in Example 7.4 leads us to conclude that the mean return on the client's account differs from that of the S&P 500 stock index. Now let's assess the return on the client's account with a confidence interval.

EXAMPLE

7.5 Estimating the mean monthly return. The mean monthly return on the client's portfolio was $\bar{x} = -1.1\%$ and the standard deviation was $s = 5.99\%$. Figure 7.7 gives the Minitab, SPSS, and Excel outputs for a 95% confidence interval for the population mean μ. Note that Excel gives the margin of error next to the label "Confidence Level (95.0%)" rather than the actual confidence interval. We see that the 95% confidence interval is (−3.04, 0.84), or (from Excel) -1.0997 ± 1.9420.

Because the S&P 500 return, 0.95%, falls outside this interval, we know that μ differs significantly from 0.95% at the $\alpha = 0.05$ level. Example 7.4 gave the actual P-value as $P = 0.039$.

Minitab

Variable	N	Mean	StDev	SE Mean	95.0% C.I.
RETURN	39	-1.100	5.991	0.959	(-3.042, 0.842)

SPSS

Descriptives

			Statistic	Std. Error
RETURN	MEAN		-1.0997	.9593
	95% Confidence Interval for Mean	Lower Bound	-3.0418	
		Upper Bound	.8423	

Excel

	A	B
1	Mean	-1.09974359
2	Standard Error	0.95930991
3	Standard Deviation	5.990888471
4	Count	39
5	Confidence Level(95.0%)	1.942021452

FIGURE 7.7 Minitab, SPSS, and Excel output for Example 7.5.

The confidence interval suggests that the broker's management of this account had a long-term mean somewhere between a loss of 3.04% and a gain of 0.84% per month. We are interested not in the actual mean but in the difference between the performance of the client's portfolio and that of the diversified S&P 500 stock index.

EXAMPLE

7.6 Estimating the difference from a standard. Following the analysis accepted by the arbitration panel, we are considering the S&P 500 monthly average return as a constant standard. (It is easy to envision scenarios where we would want to treat this type of quantity as random.) The difference between the mean of the investor's account and the S&P 500 is $\bar{x} - \mu = -1.10 - 0.95 = -2.05\%$. In Example 7.5 we found that the 95% confidence interval for the investor's account was (−3.04, 0.84). To obtain the corresponding interval for the difference, subtract 0.95 from each of the endpoints. The resulting interval is (−3.04 − 0.95, 0.84 − 0.95), or (−3.99, −0.11). We conclude with 95% confidence that the underperformance was between −3.99% and −0.11%. This interval is presented in the SPSS output of Figure 7.6. This estimate helps to set the compensation owed the investor.

The assumption that these 39 monthly returns represent an SRS from the population of monthly returns is certainly questionable. If the monthly S&P 500 returns were available, an alternative analysis would be to compare the average difference between the monthly returns for this account and for the S&P 500. This method of analysis is discussed next.

USE YOUR KNOWLEDGE

7.5 Significance test using the *t* distribution. A test of a null hypothesis versus a two-sided alternative gives $t = 2.35$.

(a) The sample size is 15. Is the test result significant at the 5% level? Explain how you obtained your answer.

(b) The sample size is 6. Is the test result significant at the 5% level? Explain how you obtained your answer.

(c) Sketch the two t distributions to illustrate your answers.

7.6 Significance test for apartment rents. Recall Exercise 7.1 (page 419). Does this SRS give good reason to believe that the mean rent of all advertised one-bedroom apartments is greater than $550? State the hypotheses, find the t statistic and its P-value, and state your conclusion.

7.7 Using software. In Example 7.1 (page 421) we calculated the 95% confidence interval for the U.S. average of hours per month spent listening to full-track music on a cell phone. Use software to compute this interval and verify that you obtain the same interval.

Matched pairs *t* procedures

The cell phone problem of Example 7.1 concerns only a single population. We know that comparative studies are usually preferred to single-sample investigations because of the protection they offer against confounding. For that reason, inference about a parameter of a single distribution is less common than comparative inference. One common comparative design, however, makes use of single-sample procedures. In a matched pairs study, subjects are matched in pairs and the outcomes are compared within each matched pair. The experimenter can toss a coin to assign two treatments to the two subjects in each pair. Matched pairs are also common when randomization is not possible. One situation calling for matched pairs is when observations are taken on the same subjects, under different conditions.

LOOK BACK
matched pairs design, page 189

EXAMPLE

7.7 Does a full moon affect behavior? Many people believe that the moon influences the actions of some individuals. A study of dementia patients in nursing homes recorded various types of disruptive behaviors every day for 12 weeks. Days were classified as moon days if they were in a three-day period centered at the day of the full moon. For each patient the average number of disruptive behaviors was computed for moon days and for all other days. The data for the 15 subjects whose behaviors were classified as aggressive are presented in Table 7.2.[3] The patients in this study are not a

that you studied in the previous exercise are given in Exercise 7.40. Answer the questions given in the previous exercise for TBBMD.

7.48 Sign test for assessment of foreign-language institute. Use the sign test to assess whether the summer institute of Exercise 7.41 improves French listening skills. State the hypotheses, give the *P*-value using the binomial table (Table C), and report your conclusion.

7.49 Sign test for fuel efficiency comparison. Use the sign test to assess whether the computer calculates a higher mpg than the driver in Exercise 7.35. State the hypotheses, give the *P*-value using the binomial table (Table C), and report your conclusion.

7.50 Insulation study. A manufacturer of electric motors tests insulation at a high temperature (250°C) and records the number of hours until the insulation fails.[16] The data for 5 specimens are

446 326 372 377 310

The small sample size makes judgment from the data difficult, but engineering experience suggests that the logarithm of the failure time will have a Normal distribution. Take the logarithms of the 5 observations, and use t procedures to give a 90% confidence interval for the mean of the log failure time for insulation of this type.

7.51 Power of the comparison of DXA machine operators. Suppose that the bone researchers in Exercise 7.39 wanted to be able to detect an alternative mean difference of 0.002. Find the power for this alternative for a sample size of 15. Use the standard deviation that you found in Exercise 7.39 for these calculations.

7.52 Sample size calculations. You are designing a study to test the null hypothesis that $\mu = 0$ versus the alternative that μ is positive. Assume that σ is 10. Suppose that it would be important to be able to detect the alternative $\mu = 2$. Perform power calculations for a variety of sample sizes and determine how large a sample you would need to detect this alternative with power of at least 0.80.

7.53 Determining the sample size. Consider Example 7.9 (page 434). What is the minimum sample size needed for the power to be greater than 80% when $\mu = 1.0$?

7.2 Comparing Two Means

A nutritionist is interested in the effect of increased calcium on blood pressure. A psychologist wants to compare male and female college students' impressions of personality based on selected photographs. A bank wants to know which of two incentive plans will most increase the use of its credit cards. Two-sample problems such as these are among the most common situations encountered in statistical practice.

TWO-SAMPLE PROBLEMS

- The goal of inference is to compare the responses in two groups.
- Each group is considered to be a sample from a distinct population.
- The responses in each group are independent of those in the other group.

LOOK BACK
randomized comparative experiment, page 183

A two-sample problem can arise from a randomized comparative experiment that randomly divides the subjects into two groups and exposes each group to a different treatment. Comparing random samples separately selected from two populations is also a two-sample problem. Unlike the matched pairs designs studied earlier, there is no matching of the units in the two samples, and the two samples may be of different sizes. Inference procedures for two-sample data differ from those for matched pairs.

We can present two-sample data graphically by a back-to-back stemplot (for small samples) or by side-by-side boxplots (for larger samples). Now we will apply the ideas of formal inference in this setting. When both population distributions are symmetric, and especially when they are at least approximately Normal, a comparison of the mean responses in the two populations is most often the goal of inference.

We have two independent samples, from two distinct populations (such as subjects given a treatment and those given a placebo). The same variable is measured for both samples. We will call the variable x_1 in the first population and x_2 in the second because the variable may have different distributions in the two populations. Here is the notation that we will use to describe the two populations:

Population	Variable	Mean	Standard deviation
1	x_1	μ_1	σ_1
2	x_2	μ_2	σ_2

We want to compare the two population means, either by giving a confidence interval for $\mu_1 - \mu_2$ or by testing the hypothesis of no difference, H_0: $\mu_1 = \mu_2$.

Inference is based on two independent SRSs, one from each population. Here is the notation that describes the samples:

Population	Sample size	Sample mean	Sample standard deviation
1	n_1	$\bar{x}_1$	s_1
2	n_2	$\bar{x}_2$	s_2

Throughout this section, the subscripts 1 and 2 show the population to which a parameter or a sample statistic refers.

The two-sample *z* statistic

LOOK BACK
addition rule for means, page 278
addition rule for variances, page 282

The natural estimator of the difference $\mu_1 - \mu_2$ is the difference between the sample means, $\bar{x}_1 - \bar{x}_2$. If we are to base inference on this statistic, we must know its sampling distribution. First, the mean of the difference $\bar{x}_1 - \bar{x}_2$ is the difference of the means $\mu_1 - \mu_2$. This follows from the addition rule for means and the fact that the mean of any $\bar{x}$ is the same as the mean of the population. Because the samples are independent, their sample means $\bar{x}_1$ and $\bar{x}_2$ are independent random variables. The addition rule for variances says that the variance of the difference $\bar{x}_1 - \bar{x}_2$ is the sum of their variances, which is

$$\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

We now know the mean and variance of the distribution of $\bar{x}_1 - \bar{x}_2$ in terms of the parameters of the two populations. If the two population distributions are both Normal, then the distribution of $\bar{x}_1 - \bar{x}_2$ is also Normal. This is true

because each sample mean alone is Normally distributed and because a difference of independent Normal random variables is also Normal.

EXAMPLE

7.13 Heights of 10-year-old girls and boys. A fourth-grade class has 12 girls and 8 boys. The children's heights are recorded on their 10th birthdays. What is the chance that the girls are taller than the boys? Of course, it is very unlikely that all of the girls are taller than all of the boys. We translate the question into the following: what is the probability that the mean height of the girls is greater than the mean height of the boys?

Based on information from the National Health and Nutrition Examination Survey,[17] we assume that the heights (in inches) of 10-year-old girls are $N(56.4, 2.7)$ and the heights of 10-year-old boys are $N(55.7, 3.8)$. The heights of the students in our class are assumed to be random samples from these populations. The two distributions are shown in Figure 7.12(a).

The difference $\bar{x}_1 - \bar{x}_2$ between the female and male mean heights varies in different random samples. The sampling distribution has mean

$$\mu_1 - \mu_2 = 56.4 - 55.7 = 0.7 \text{ inch}$$

and variance

$$\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} = \frac{2.7^2}{12} + \frac{3.8^2}{8} = 2.41$$

The standard deviation of the difference in sample means is therefore $\sqrt{2.41} = 1.55$ inches.

If the heights vary Normally, the difference in sample means is also Normally distributed. The distribution of the difference in heights is shown in Figure 7.12(b). We standardize $\bar{x}_1 - \bar{x}_2$ by subtracting its mean (0.7) and dividing by its standard deviation (1.55). Therefore, the probability that the girls are taller than the boys is

$$P(\bar{x}_1 - \bar{x}_2 > 0) = P\left(\frac{(\bar{x}_1 - \bar{x}_2) - 0.7}{1.55} > \frac{0 - 0.7}{1.55}\right)$$
$$= P(Z > -0.45) = 0.67$$

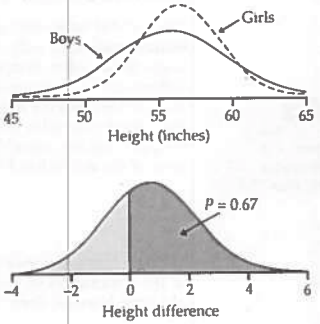


FIGURE 7.12 Distributions for Example 7.13. (a) Distributions of heights of 10-year-old boys and girls. (b) Distribution of the difference between mean heights of 12 girls and 8 boys.

Even though the population mean height of 10-year-old girls is greater than the population mean height of 10-year-old boys, the probability that the sample mean of the girls is greater than the sample mean of the boys in our class is only 67%. *Large samples are needed to see the effects of small differences.*

As Example 7.13 reminds us, any Normal random variable has the $N(0, 1)$ distribution when standardized. We have arrived at a new z statistic.

TWO-SAMPLE z STATISTIC

Suppose that $\bar{x}_1$ is the mean of an SRS of size n_1 drawn from an $N(\mu_1, \sigma_1)$ population and that $\bar{x}_2$ is the mean of an independent SRS of size n_2 drawn from an $N(\mu_2, \sigma_2)$ population. Then the **two-sample z statistic**

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

has the standard Normal $N(0, 1)$ sampling distribution.

In the unlikely event that both population standard deviations are known, the two-sample z statistic is the basis for inference about $\mu_1 - \mu_2$. Exact z procedures are seldom used, however, because σ_1 and σ_2 are rarely known. In Chapter 6, we discussed the one-sample z procedures in order to introduce the ideas of inference. Here we move directly to the more useful t procedures.

The two-sample t procedures

Suppose now that the population standard deviations σ_1 and σ_2 are not known. We estimate them by the sample standard deviations s_1 and s_2 from our two samples. Following the pattern of the one-sample case, we substitute the standard errors for the standard deviations used in the two-sample z statistic. The result is the *two-sample t statistic:*

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Unfortunately, this statistic does *not* have a t distribution. A t distribution replaces the $N(0, 1)$ distribution only when a single standard deviation (σ) in a z statistic is replaced by its sample standard deviation (s). In this case, we replace two standard deviations (σ_1 and σ_2) by their estimates (s_1 and s_2), which does not produce a statistic having a t distribution.

approximations for the degrees of freedom

Nonetheless, we can approximate the distribution of the two-sample t statistic by using the $t(k)$ distribution with an **approximation for the degrees of freedom k.** We use these approximations to find approximate values of t^* for confidence intervals and to find approximate P-values for significance tests. Here are two approximations:

1. Use a value of k that is calculated from the data. In general, it will not be a whole number.
2. Use k equal to the smaller of $n_1 - 1$ and $n_2 - 1$.

Most statistical software uses the first option to approximate the $t(k)$ distribution for two-sample problems unless the user requests another method. Use of this approximation without software is a bit complicated; we will give the details later in this section. If you are not using software, the second approximation is preferred. This approximation is appealing because it is conservative.[18] Margins of error for confidence intervals are a bit larger than they need to be, so the true confidence level is larger than C. For significance testing, the true P-values are a bit smaller than those we obtain from the approximation; for tests at a fixed significance level, we are a little less likely to reject H_0 when it is true. In practice, the choice of approximation rarely makes a difference in our conclusion.

The two-sample t significance test

THE TWO-SAMPLE t SIGNIFICANCE TEST

Suppose that an SRS of size n_1 is drawn from a Normal population with unknown mean μ_1 and that an independent SRS of size n_2 is drawn from another Normal population with unknown mean μ_2. To test the hypothesis $H_0: \mu_1 = \mu_2$, compute the **two-sample t statistic**

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

and use P-values or critical values for the $t(k)$ distribution, where the degrees of freedom k are either approximated by software or are the smaller of $n_1 - 1$ and $n_2 - 1$.

EXAMPLE

7.14 Directed reading activities assessment. An educator believes that new directed reading activities in the classroom will help elementary school pupils improve some aspects of their reading ability. She arranges for a third-grade class of 21 students to take part in these activities for an eight-week period. A control classroom of 23 third-graders follows the same curriculum without the activities. At the end of the eight weeks, all students are given a Degree of Reading Power (DRP) test, which measures the aspects of reading ability that the treatment is designed to improve. The data appear in Table 7.4.[19]

TABLE 7.4

DRP scores for third-graders

Treatment Group				Control Group			
24	61	59	46	42	33	46	37
43	44	52	43	43	41	10	42
58	67	62	57	55	19	17	55
71	49	54		26	54	60	28
43	53	57		62	20	53	48
49	56	33		37	85	42	

First examine the data:

Control		Treatment
970	1	
860	2	4
773	3	3
8632221	4	3334689
5543	5	23467789
20	6	127
	7	1
5	8	

A back-to-back stemplot suggests that there is a mild outlier in the control group but no deviation from Normality serious enough to forbid use of t procedures. Separate Normal quantile plots for both groups (Figure 7.13) confirm that both are approximately Normal. The scores of the treatment group appear to be somewhat higher than those of the control group. The summary statistics are

Group	n	$\bar{x}$	s
Treatment	21	51.48	11.01
Control	23	41.52	17.15

Because we hope to show that the treatment (Group 1) is better than the control (Group 2), the hypotheses are

$$H_0: \mu_1 = \mu_2$$
$$H_a: \mu_1 > \mu_2$$

The two-sample t test statistic is

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{51.48 - 41.52}{\sqrt{\frac{11.01^2}{21} + \frac{17.15^2}{23}}}$$
$$= 2.31$$

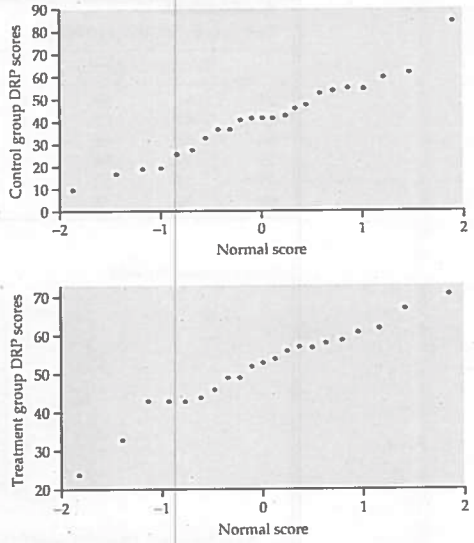


FIGURE 7.13 Normal quantile plots of the DRP scores in Table 7.4.

The *P*-value for the one-sided test is $P(T \geq 2.31)$. Software gives the approximate *P*-value as 0.0132 and uses 37.9 as the degrees of freedom. For the second approximation, the degrees of freedom k are equal to the smaller of

$$n_1 - 1 = 21 - 1 = 20 \quad \text{and} \quad n_2 - 1 = 23 - 1 = 22$$

Comparing 2.31 with the entries in Table D for 20 degrees of freedom, we see that *P* lies between 0.02 and 0.01. The data strongly suggest that directed reading activity improves the DRP score ($t = 2.31$, df = 20, $P < 0.02$).

df = 20

p	0.02	0.01
t^*	2.197	2.528

Note that when we report a result such as this with $P < 0.02$, we imply that the result is *not* significant at the 0.01 level.

If your software gives *P*-values for only the two-sided alternative, $2P(T \geq |t|)$, you need to divide the reported value by 2 after checking that the means differ in the direction specified by the alternative hypothesis.

USE YOUR KNOWLEDGE

7.54 Comparison of two Web designs. You want to compare the daily number of hits for two different Web designs that advertise your Internet business. You assign the next 50 days to either Design A or Design B, 25 days to each.

(a) Would you use a one-sided or two-sided significance test for this problem? Explain your choice.

(b) If you use Table D to find the critical value, what are the degrees of freedom using the second approximation?

(c) If you perform the significance test using $\alpha = 0.05$, how large (positive or negative) must the t statistic be to reject the null hypothesis that the two designs result in the same average hits?

7.55 More on the comparison of two Web designs. Consider the previous problem. If the t statistic for comparing the mean hits were 2.75, what *P*-value would you report? What would you conclude using $\alpha = 0.05$?

The two-sample *t* confidence interval

The same ideas that we used for the two-sample t significance tests also apply to give us *two-sample t confidence intervals*. We can use either software or the conservative approach with Table D to approximate the value of t^*.

THE TWO-SAMPLE *t* CONFIDENCE INTERVAL

Suppose that an SRS of size n_1 is drawn from a Normal population with unknown mean μ_1 and that an independent SRS of size n_2 is drawn from another Normal population with unknown mean μ_2. The **confidence interval for $\mu_1 - \mu_2$** given by

$$(\bar{x}_1 - \bar{x}_2) \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

has confidence level at least C no matter what the population standard deviations may be. Here, t^* is the value for the $t(k)$ density curve with area C between $-t^*$ and t^*. The value of the degrees of freedom k is approximated by software or we use the smaller of $n_1 - 1$ and $n_2 - 1$.

To complete the analysis of the DRP scores we examined in Example 7.14, we need to describe the size of the treatment effect. We do this with a confidence interval for the difference between the treatment group and the control group means.

EXAMPLE

7.15 How much improvement? We will find a 95% confidence interval for the mean improvement in the entire population of third-graders. The interval is

$$(\bar{x}_1 - \bar{x}_2) \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = (51.48 - 41.52) \pm t^* \sqrt{\frac{11.01^2}{21} + \frac{17.15^2}{23}}$$
$$= 9.96 \pm 4.31 t^*$$

Using software, the degrees of freedom are 37.9 and $t^* = 2.025$. This approximation gives

$$9.96 \pm (4.31 \times 2.025) = 9.96 \pm 8.72 = (1.2, 18.7)$$

The conservative approach uses the $t(20)$ distribution. Table D gives $t^* = 2.086$. With this approximation we have

$$9.96 \pm (4.31 \times 2.086) = 9.96 \pm 8.99 = (1.0, 18.9)$$

We can see that the conservative approach does, in fact, give a larger interval than the more accurate approximation used by software. However, the difference is pretty small.

We estimate the mean improvement to be about 10 points, but with a margin of error of almost 9 points with either method. Although we have good evidence of some improvement, the data do not allow a very precise estimate of the size of the average improvement.

The design of the study in Example 7.14 is not ideal. Random assignment of students was not possible in a school environment, so existing third-grade classes were used. The effect of the reading programs is therefore confounded with any other differences between the two classes. The classes were chosen to be as similar as possible—for example, in terms of the social and economic status of the students. Extensive pretesting showed that the two classes were on the average quite similar in reading ability at the beginning of the experiment. To avoid the effect of two different teachers, the researcher herself taught reading in both classes during the eight-week period of the experiment. We can therefore be somewhat confident that the two-sample test is detecting the effect of the treatment and not some other difference between the classes. This example is typical of many situations in which an experiment is carried out but randomization is not possible.

USE YOUR KNOWLEDGE

7.56 Two-sample t confidence interval. Assume $\bar{x}_1 = 100$, $\bar{x}_2 = 120$, $s_1 = 10$, $s_2 = 12$, $n_1 = 50$, and $n_2 = 50$. Find a 95% confidence interval for the difference in the corresponding values of μ using the second approximation for degrees of freedom. Does this interval include more or fewer values than a 99% confidence interval? Explain your answer.

7.57 Another two-sample t confidence interval. Assume $\bar{x}_1 = 100$, $\bar{x}_2 = 120$, $s_1 = 10$, $s_2 = 12$, $n_1 = 10$, and $n_2 = 10$. Find a 95% confidence interval for the difference in the corresponding values of μ using the second approximation for degrees of freedom. Would you reject the null hypothesis that the population means are equal in favor of the two-sided alternative at significance level 0.05? Explain.

Robustness of the two-sample procedures

The two-sample t procedures are more robust than the one-sample t methods. When the sizes of the two samples are equal and the distributions of the two populations being compared have similar shapes, probability values from the t table are quite accurate for a broad range of distributions when the sample sizes are as small as $n_1 = n_2 = 5$.[20] When the two population distributions have different shapes, larger samples are needed. The guidelines for the use of one-sample t procedures can be adapted to two-sample procedures by replacing "sample size" with the "sum of the sample sizes" $n_1 + n_2$. These guidelines are rather conservative, especially when the two samples are of equal size. *In planning a two-sample study, choose equal sample sizes if you can.* The two-sample t procedures are most robust against non-Normality in this case, and the conservative probability values are most accurate.

Here is an example with moderately large sample sizes that are not equal. Even if the distributions are not Normal, we are confident that the sample means will be approximately Normal. The two-sample t test is very robust in this case.

EXAMPLE

7.16 Wheat prices. The U.S. Department of Agriculture (USDA) uses sample surveys to produce important economic estimates. One pilot study estimated wheat prices in July and in September using independent samples of wheat producers in the two months. Here are the summary statistics, in dollars per bushel:[21]

Month	n	$\bar{x}$	s
September	45	\$3.61	\$0.19
July	90	\$2.95	\$0.22

The September prices are higher on the average. But we have data from only a sample of producers each month. Can we conclude that national average prices in July and September are not the same? Or are these differences merely what we would expect to see due to random variation?

Because we did not specify a direction for the difference before looking at the data, we choose a two-sided alternative. The hypotheses are

$$H_0: \mu_1 = \mu_2$$
$$H_a: \mu_1 \neq \mu_2$$

Because the samples are moderately large, we can confidently use the t procedures even though we lack the detailed data and so cannot verify the Normality condition.

The two-sample t statistic is

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{3.61 - 2.95}{\sqrt{\frac{0.19^2}{45} + \frac{0.22^2}{90}}}$$
$$= 18.03$$

df = 40

p	0.0005
t^*	3.551

The conservative approach finds the P-value by comparing 18.03 to critical values for the $t(44)$ distribution because the smaller sample has 45 observations. We must double the table tail area p because the alternative is two-sided. Table D does not have entries for 44 degrees of freedom. When this happens, we use the next smaller degrees of freedom. Our calculated value of t is larger than the $p = 0.0005$ entry in the table. Doubling 0.0005, we conclude that the P-value is less than 0.001. The data give conclusive evidence that the mean wheat prices were higher in September than they were in July ($t = 18.03$, df $= 44$, $P < 0.001$).

In this example the exact P-value is very small because $t = 18$ says that the observed difference in means is 18 standard errors above the hypothesized difference of zero ($\mu_1 = \mu_2$). This is so unlikely that the probability is zero for all practical purposes. The difference in mean prices is not only highly significant but large enough (66 cents per bushel) to be important to producers.

In this and other examples, we can choose which population to label 1 and which to label 2. After inspecting the data, we chose September as Population 1 because this choice makes the t statistic a positive number. This avoids any possible confusion from reporting a negative value for t. *Choosing the population labels is not the same as choosing a one-sided alternative after looking at the data.* Choosing hypotheses after seeing a result in the data is a violation of sound statistical practice.

CAUTION

Inference for small samples

Small samples require special care. We do not have enough observations to examine the distribution shapes, and only extreme outliers stand out. The power of significance tests tends to be low, and the margins of error of confidence intervals tend to be large. Despite these difficulties, we can often draw important conclusions from studies with small sample sizes. If the size of an effect is as large as it was in the wheat price example, it should still be evident even if the n's are small.

EXAMPLE

7.17 More about wheat prices. In the setting of Example 7.16, a quick survey collects prices from only 5 producers each month. The data are

Month	Price of wheat ($/bushel)				
September	$3.5900	$3.6150	$3.5950	$3.5725	$3.5825
July	$2.9200	$2.9675	$2.9175	$2.9250	$2.9325

The prices are reported to the nearest quarter of a cent. First, examine the distributions with a back-to-back stemplot after rounding each price to the nearest cent.

sum $n_1 + n_2 - 2$ of the two individual degrees of freedom. The number of degrees of freedom is generally not a whole number. There is a t distribution with any positive degrees of freedom, even though Table D contains entries only for whole-number degrees of freedom. When df is small and is not a whole number, interpolation between entries in Table D may be needed to obtain an accurate critical value or P-value. Because of this and the need to calculate df, we do not recommend regular use of this approximation if a computer is not doing the arithmetic. With a computer, however, the more accurate procedures are painless.

USE YOUR KNOWLEDGE

7.58 Calculating the degrees of freedom. Assume $s_1 = 10$, $s_2 = 12$, $n_1 = 20$, and $n_2 = 18$. Find the software approximate degrees of freedom.

The pooled two-sample t procedures*

There is one situation in which a t statistic for comparing two means has exactly a t distribution. Suppose that the two Normal population distributions have the *same* standard deviation. In this case we need substitute only a single standard error in a z statistic, and the resulting t statistic has a t distribution. We will develop the z statistic first, as usual, and from it the t statistic.

Call the common—and still unknown—standard deviation of both populations σ. Both sample variances s_1^2 and s_2^2 estimate σ^2. The best way to combine these two estimates is to average them with weights equal to their degrees of freedom. This gives more weight to the information from the larger sample, which is reasonable. The resulting estimator of σ^2 is

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

pooled estimator of σ^2

This is called the **pooled estimator of σ^2** because it combines the information in both samples.

When both populations have variance σ^2, the addition rule for variances says that $\bar{x}_1 - \bar{x}_2$ has variance equal to the *sum* of the individual variances, which is

$$\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2} = \sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$$

The standardized difference of means in this equal-variance case is therefore

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sigma\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$$

This is a special two-sample z statistic for the case in which the populations have the same σ. Replacing the unknown σ by the estimate s_p gives a t statistic.

*This section can be omitted if desired, but it should be read if you plan to read Chapters 12 and 13.

The degrees of freedom are $n_1 + n_2 - 2$, the sum of the degrees of freedom of the two sample variances. This statistic is the basis of the pooled two-sample t inference procedures.

THE POOLED TWO-SAMPLE t PROCEDURES

Suppose that an SRS of size n_1 is drawn from a Normal population with unknown mean μ_1 and that an independent SRS of size n_2 is drawn from another Normal population with unknown mean μ_2. Suppose also that the two populations have the same standard deviation. A level C confidence interval for $\mu_1 - \mu_2$ is

$$(\bar{x}_1 - \bar{x}_2) \pm t^* s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Here t^* is the value for the $t(n_1 + n_2 - 2)$ density curve with area C between $-t^*$ and t^*.

To test the hypothesis $H_0: \mu_1 = \mu_2$, compute the pooled two-sample t statistic

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

In terms of a random variable T having the $t(n_1 + n_2 - 2)$ distribution, the P-value for a test of H_0 against

$$H_a: \mu_1 > \mu_2 \text{ is } P(T \geq t)$$
$$H_a: \mu_1 < \mu_2 \text{ is } P(T \leq t)$$
$$H_a: \mu_1 \neq \mu_2 \text{ is } 2P(T \geq |t|)$$

EXAMPLE

7.19 Calcium and blood pressure. Does increasing the amount of calcium in our diet reduce blood pressure? Examination of a large sample of people revealed a relationship between calcium intake and blood pressure, but such observational studies do not establish causation. Animal experiments, however, showed that calcium supplements do reduce blood pressure in rats, justifying an experiment with human subjects. A randomized comparative experiment gave one group of 10 black men a calcium supplement for 12 weeks. The control group of 11 black men received a placebo that appeared identical. (In fact, a block design with black and white men as the blocks was used. We will look only at the results for blacks, because the earlier survey suggested that calcium is more effective for blacks.) The experiment was double-blind. Table 7.5 gives the seated systolic (heart contracted) blood pressure for all subjects at the beginning and end of the 12-week period, in millimeters (mm) of mercury. Because the researchers were interested in decreasing blood pressure, Table 7.5 also shows the decrease for each subject. An increase appears as a negative entry.[22]

TABLE 7.5

Seated systolic blood pressure

Calcium Group			Placebo Group		
Begin	End	Decrease	Begin	End	Decrease
107	100	7	123	124	−1
110	114	−4	109	97	12
123	105	18	112	113	−1
129	112	17	102	105	−3
112	115	−3	98	95	3
111	116	−5	114	119	−5
107	106	1	119	114	5
112	102	10	114	112	2
136	125	11	110	121	−11
102	104	−2	117	118	−1
			130	133	−3

As usual, we first examine the data. To compare the effects of the two treatments, take the response variable to be the amount of the decrease in blood pressure. Inspection of the data reveals that there are no outliers. Normal quantile plots (Figure 7.15) give a more detailed picture. The calcium group has a somewhat short left tail, but there are no departures from Normality that will prevent use of t procedures. To examine the question of the researchers who collected these data, we perform a significance test.

EXAMPLE

7.20 Does increased calcium reduce blood pressure? Take Group 1 to be the calcium group and Group 2 to be the placebo group. The evidence that calcium lowers blood pressure more than a placebo is assessed by testing

$$H_0: \mu_1 = \mu_2$$
$$H_a: \mu_1 > \mu_2$$

Here are the summary statistics for the decrease in blood pressure:

Group	Treatment	n	$\bar{x}$	s
1	Calcium	10	5.000	8.743
2	Placebo	11	−0.273	5.901

The calcium group shows a drop in blood pressure, and the placebo group has a small increase. The sample standard deviations do not rule out equal population standard deviations. A difference this large will often arise by chance in samples this small. We are willing to assume equal population standard

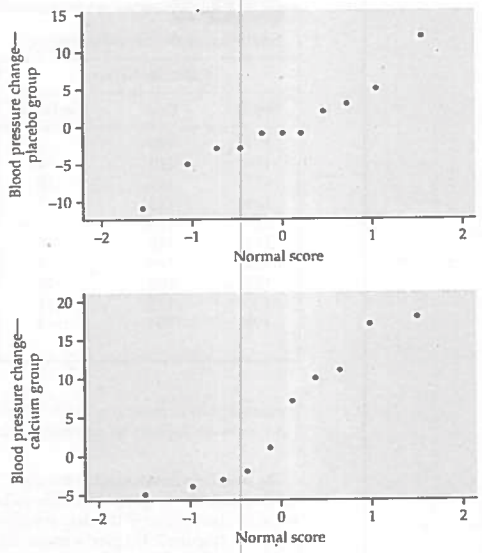


FIGURE 7.15 Normal quantile plots of the change in blood pressure from Table 7.5.

deviations. The pooled sample variance is

$$s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$$

$$= \frac{(10-1)8.743^2 + (11-1)5.901^2}{10+11-2} = 54.536$$

so that

$$s_p = \sqrt{54.536} = 7.385$$

The pooled two-sample t statistic is

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$$

$$= \frac{5.000 - (-0.273)}{7.385\sqrt{\frac{1}{10}+\frac{1}{11}}}$$

$$= \frac{5.273}{3.227} = 1.634$$

df = 19

p	0.10	0.05
t^*	1.328	1.729

The P-value is $P(T \geq 1.634)$, where T has the $t(19)$ distribution. From Table D we can see that P falls between the $\alpha = 0.10$ and $\alpha = 0.05$ levels. Statistical software gives the exact value $P = 0.059$. The experiment found evidence that calcium reduces blood pressure, but the evidence falls a bit short of the traditional 5% and 1% levels.

Sample size strongly influences the P-value of a test. An effect that fails to be significant at a specified level α in a small sample can be significant in a larger sample. In the light of the rather small samples in Example 7.20, the evidence for some effect of calcium on blood pressure is rather good. The published account of the study combined these results for blacks with the results for whites and adjusted for pretest differences among the subjects. Using this more detailed analysis, the researchers were able to report the P-value $P = 0.008$.

Of course, a P-value is almost never the last part of a statistical analysis. To make a judgment regarding the size of the effect of calcium on blood pressure, we need a confidence interval.

EXAMPLE

7.21 How different are the calcium and placebo groups? We estimate that the effect of calcium supplementation is the difference between the sample means of the calcium and the placebo groups, $\bar{x}_1 - \bar{x}_2 = 5.273$ mm. A 90% confidence interval for $\mu_1 - \mu_2$ uses the critical value $t^* = 1.729$ from the $t(19)$ distribution. The interval is

$$(\bar{x}_1 - \bar{x}_2) \pm t^* s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} = [5.000 - (-0.273)] \pm (1.729)(7.385)\sqrt{\frac{1}{10}+\frac{1}{11}}$$

$$= 5.273 \pm 5.579$$

We are 90% confident that the difference in means is in the interval

$$(-0.306, 10.852)$$

The calcium treatment reduced blood pressure by about 5.3 mm more than a placebo on the average, but the margin of error for this estimate is 5.6 mm.

CAUTION

The pooled two-sample t procedures are anchored in statistical theory and so have long been the standard version of the two-sample t in textbooks. *But they require the assumption that the two unknown population standard deviations are equal.* As we shall see in Section 7.3, this assumption is hard to verify. The pooled t procedures are therefore a bit risky. They are reasonably robust against both non-Normality and unequal standard deviations when the sample sizes are nearly the same. When the samples are quite different in size, the pooled t procedures become sensitive to unequal standard deviations and should be used with caution unless the samples are large. Unequal standard deviations are quite common. In particular, it is not unusual for the spread of data to increase when the center gets larger. Statistical software often calculates both the pooled and the unpooled t statistics, as in Figure 7.14.

USE YOUR KNOWLEDGE

7.59 Wheat prices revisited. Figure 7.14 (page 458) gives the outputs from four software packages for comparing prices received by wheat producers in July and September for small samples of 5 producers in each month. Some of the software reports both pooled and unpooled analyses. Which outputs give the pooled results? What are the pooled t and its P-value?

7.60 More on wheat prices. The software outputs in Figure 7.14 give the *same value* for the pooled and unpooled t statistics. Do some simple algebra to show that this is always true when the two sample sizes n_1 and n_2 are the same. In other cases, the two t statistics usually differ.

SECTION 7.2 Summary

Significance tests and confidence intervals for the difference of the means μ_1 and μ_2 of two Normal populations are based on the difference $\bar{x}_1 - \bar{x}_2$ of the sample means from two independent SRSs. Because of the central limit theorem, the resulting procedures are approximately correct for other population distributions when the sample sizes are large.

When independent SRSs of sizes n_1 and n_2 are drawn from two Normal populations with parameters μ_1, σ_1 and μ_2, σ_2 the **two-sample z statistic**

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

has the $N(0, 1)$ distribution.

The **two-sample t statistic**

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

does *not* have a t distribution. However, good approximations are available.

Conservative inference procedures for comparing μ_1 and μ_2 are obtained from the two-sample t statistic by using the $t(k)$ distribution with degrees of freedom k equal to the smaller of $n_1 - 1$ and $n_2 - 1$.

More accurate probability values can be obtained by estimating the degrees of freedom from the data. This is the usual procedure for statistical software.

An approximate level C **confidence interval** for $\mu_1 - \mu_2$ is given by

$$(\bar{x}_1 - \bar{x}_2) \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Here, t^* is the value for the $t(k)$ density curve with area C between $-t^*$ and t^*, where k is computed from the data by software or is the smaller of $n_1 - 1$ and $n_2 - 1$. The quantity

$$t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

is the **margin of error.**

Significance tests for $H_0: \mu_1 = \mu_2$ use the **two-sample t statistic**

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

The P-value is approximated using the $t(k)$ distribution where k is estimated from the data using software or is the smaller of $n_1 - 1$ and $n_2 - 1$.

The guidelines for practical use of two-sample t procedures are similar to those for one-sample t procedures. Equal sample sizes are recommended.

If we can assume that the two populations have equal variances, **pooled two-sample t procedures** can be used. These are based on the **pooled estimator**

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

of the unknown common variance and the $t(n_1 + n_2 - 2)$ distribution.

SECTION 7.2 Exercises

For Exercises 7.54 and 7.55, see pages 453 and 454; for Exercises 7.56 and 7.57, see page 455; for Exercise 7.58, see page 461; and for Exercises 7.59 and 7.60, see page 466.

In exercises that call for two-sample t procedures, you may use either of the two approximations for the degrees of freedom that we have discussed: the value given by your software or the smaller of $n_1 - 1$ and $n_2 - 1$. Be sure to state clearly which approximation you have used.

7.61 Comparison of blood lipid levels in males and females. A recent study at Baylor University investigated the lipid levels in a cohort of sedentary university students.[23] A total of 108 students volunteered for the study and met the eligibility criteria. The following table summarizes the blood lipid levels, in milligrams per deciliter (mg/dl), of the participants broken down by gender:

	Females ($n = 71$)		Males ($n = 37$)	
	$\bar{x}$	s	$\bar{x}$	s
Total cholesterol	173.70	34.79	171.81	33.24
LDL	96.38	29.78	109.44	31.05
HDL	61.62	13.75	46.47	7.94

(a) Is it appropriate to use the two-sample t procedures that we studied in this section to analyze these data for gender differences? Give reasons for your answer.

(b) Describe appropriate null and alternative hypotheses for comparing male and female total cholesterol levels.

(c) Carry out the significance test. Report the test statistic with the degrees of freedom and the P-value. Write a short summary of your conclusion.

(d) Find a 95% confidence interval for the difference between the two means. Compare the information given by the interval with the information given by the significance test.

(e) The participants in this study were all taking an introductory health class. To what extent do you think the results can be generalized to other populations?

7.62 More on blood lipid levels. Refer to the previous exercise. LDL is also known as "bad" cholesterol. Suppose the researchers wanted to test the hypothesis that LDL levels are higher in sedentary males than in sedentary females. Describe appropriate null and alternative hypotheses and carry out the significance test using $\alpha = 0.05$. Report

CHAPTER

12

One-Way Analysis of Variance

Which brand of tires lasts the longest under city driving conditions? The methods described in this chapter allow us to compare the average wear of each brand.

Introduction

Many of the most effective statistical studies are comparative. For example, we may wish to compare customer satisfaction of men and women using an online fantasy football site or compare the responses to various treatments in a clinical trial. We display these comparisons with back-to-back stemplots or side-by-side boxplots, and we measure them with five-number summaries or with means and standard deviations.

When only two groups are compared, Chapter 7 provides the tools we need to answer the question "Is the difference between groups statistically significant?" Two-sample t procedures compare the means of two Normal populations, and we saw that these procedures, unlike comparisons of spread, are sufficiently robust to be widely useful.

In this chapter, we will compare any number of means by techniques that generalize the two-sample t and share its robustness and usefulness. These methods will allow us to address comparisons such as

- Which of 4 advertising offers mailed to sample households produces the highest dollar sales?
- Which of 10 brands of automobile tires wears longest?
- How long do cancer patients live under each of 3 therapies for their lung cancer?

12.1 Inference for One-Way Analysis of Variance

When comparing different populations or treatments, the data are subject to sampling variability. For example, we would not expect the same sales data if we mailed various advertising offers to a different sample of households. We therefore pose the question for inference in terms of the *mean* response. In Chapter 7 we met procedures for comparing the means of two populations. We are now ready to extend those methods to problems involving more than two populations. The statistical methodology for comparing several means is called **analysis of variance,** or simply **ANOVA.** In the sections that follow, we will examine the basic ideas and assumptions that are needed for ANOVA. Although the details differ, many of the concepts are similar to those discussed in the two-sample case.

LOOK BACK
comparing two means, page 447

ANOVA

We will consider two ANOVA techniques. When there is only one way to classify the populations of interest, we use **one-way ANOVA** to analyze the data. For example, to compare the survival times for three different lung cancer therapies we use one-way ANOVA. This chapter presents the details for one-way ANOVA.

one-way ANOVA

In many other comparison studies, there is more than one way to classify the populations. For the advertising study, the company may also consider mailing the offers using two different envelope styles. Will each offer draw more sales on the average when sent in an attention-grabbing envelope? Analyzing the effect of advertising offer and envelope layout together requires **two-way ANOVA.** This technique will be discussed in Chapter 13. While adding yet more factors necessitates even higher-way ANOVA techniques, most of the new ideas in ANOVA with more than one factor already appear in two-way ANOVA.

two-way ANOVA

Data for one-way ANOVA

One-way analysis of variance is a statistical method for comparing several population means. We draw a simple random sample (SRS) from each population and use the data to test the null hypothesis that the population means are all equal. Consider the following two examples:

EXAMPLE

12.1 Choosing the best magazine layout. A magazine publisher wants to compare three different layouts for a magazine that will be offered for sale at supermarket checkout lines. She is interested in whether there is a layout that better catches shoppers' attention and results in more sales. To investigate, she randomly assigns each of 60 stores to one of the three layouts and records the number of magazines that are sold in a one-week period.

EXAMPLE

12.2 Average age of bookstore customers. How do five bookstores in the same city differ in the demographics of their customers? Are certain bookstores more popular among teenagers? Do upper-income shoppers tend to go to one store? A market researcher asks 50 customers of each store to respond to a questionnaire. Two variables of interest are the customer's age and income level.

These two examples are similar in that

- There is a single quantitative response variable measured on many units; the units are stores in the first example and customers in the second.
- The goal is to compare several populations: stores displaying three magazine layouts in the first example and customers of five bookstores in the second.

LOOK BACK
observation versus experiment, page 175

There is, however, an important difference. Example 12.1 describes an experiment in which stores are randomly assigned to layouts. Example 12.2 is an observational study in which customers are selected during a particular time period and not all agree to provide data. We will treat our samples of customers as random samples even though this is only approximately true.

In both examples, we will use ANOVA to compare the mean responses. The same ANOVA methods apply to data from random samples and to data from randomized experiments. *It is important to keep the data-production method in mind when interpreting the results. A strong case for causation is best made by a randomized experiment.*

Comparing means

The question we ask in ANOVA is "Do all groups have the same population mean?" We will often use the term *groups* for the populations to be compared in a one-way ANOVA. To answer this question we compare the sample means. Figure 12.1 displays the sample means for Example 12.1. It appears that Layout 2 has the highest average sales. But is the observed difference in sample means just the result of chance variation? We should not expect sample means to be equal, even if the population means are all identical.

LOOK BACK
standard deviation of $\bar{x}$, page 338

The purpose of ANOVA is to assess whether the observed differences among sample means are *statistically significant*. Could a variation among the three sample means this large be plausibly due to chance, or is it good evidence for a difference among the population means? This question can't be answered from the sample means alone. Because the standard deviation of a sample mean $\bar{x}$ is the population standard deviation σ divided by $\sqrt{n}$, the answer also depends upon both the variation within the groups of observations and the sizes of the samples.

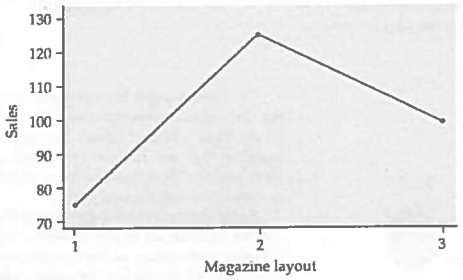


FIGURE 12.1 Mean sales of magazines for three different magazine layouts.

Side-by-side boxplots help us see the within-group variation. Compare Figures 12.2(a) and 12.2(b). The sample medians are the same in both figures, but the large variation within the groups in Figure 12.2(a) suggests that the differences among the sample medians could be due simply to chance variation. The data in Figure 12.2(b) are much more convincing evidence that the populations differ. Even the boxplots omit essential information, however. To assess the observed differences, we must also know how large the samples are. Nonetheless, boxplots are a good preliminary display of the data. While ANOVA compares means and boxplots display medians, we expect the data to be approximately Normal and will consider a transformation if they are not. For distributions that are nearly symmetric, these two measures of center will be close together.

LOOK BACK
transforming data, page 435

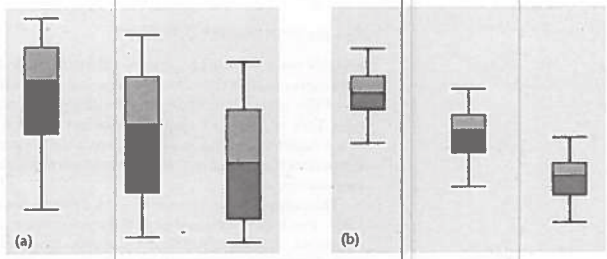


FIGURE 12.2 (a) Side-by-side boxplots for three groups with large within-group variation. The differences among centers may be just chance variation. (b) Side-by-side boxplots for three groups with the same centers as in Figure 12.2(a) but with small within-group variation. The differences among centers are more likely to be significant.

The two-sample *t* statistic

Two-sample t statistics compare the means of two populations. If the two populations are assumed to have equal but unknown standard deviations and the sample sizes are both equal to n, the t statistic is

LOOK BACK
pooled two-sample t statistic, page 462

$$t = \frac{\bar{x} - \bar{y}}{s_p\sqrt{\frac{1}{n} + \frac{1}{n}}} = \frac{\sqrt{\frac{n}{2}}(\bar{x} - \bar{y})}{s_p}$$

The square of this t statistic is

$$t^2 = \frac{\frac{n}{2}(\bar{x} - \bar{y})^2}{s_p^2}$$

If we use ANOVA to compare two populations, the ANOVA F statistic is exactly equal to this t^2. We can therefore learn something about how ANOVA works by looking carefully at the statistic in this form.

between-group variation

The numerator in the t^2 statistic measures the variation **between** the groups in terms of the difference between their sample means $\bar{x}$ and $\bar{y}$. It includes a

factor for the common sample size n. The numerator can be large because of a large difference between the sample means or because the sample sizes are large. The denominator measures the variation **within** groups by s_p^2, the pooled estimator of the common variance. If the within-group variation is small, the same variation between the groups produces a larger statistic and thus a more significant result.

within-group variation

Although the general form of the F statistic is more complicated, the idea is the same. To assess whether several populations all have the same mean, we compare the variation *among* the means of several groups with the variation *within* groups. Because we are comparing variation, the method is called *analysis of variance.*

An overview of ANOVA

ANOVA tests the null hypothesis that the population means are *all* equal. The alternative is that they are not all equal. This alternative could be true because all of the means are different or simply because one of them differs from the rest. This is a more complex situation than comparing just two populations. If we reject the null hypothesis, we need to perform some further analysis to draw conclusions about which population means differ from which others and by how much.

The computations needed for an ANOVA are more lengthy than those for the t test. For this reason we generally use computer programs to perform the calculations. Automating the calculations frees us from the burden of arithmetic and allows us to concentrate on interpretation. *Complicated computations do not guarantee a valid statistical analysis. We should always start our ANOVA with a careful examination of the data using graphical and numerical summaries.*

CAUTION

EXAMPLE

12.3 Workplace safety. In a study of workplace safety, workers were asked to rate various elements of safety, and a composite score called the Safety Climate Index (SCI) was calculated.[1] The index is the sum of the responses to 10 different questions about safety. The response for each of these questions is an integer ranging from 0 to 10, so the SCI has values from 0 to 100. The workers were classified according to their job category as unskilled, skilled, and supervisor. Here is a summary of the data:

Job category	n	$\bar{x}$	s
Unskilled workers	448	70.42	18.27
Skilled workers	91	71.21	18.83
Supervisors	51	80.51	14.58

Histograms and descriptive statistics for the three groups of workers are given in Figure 12.3. Note that the heights of the bars in the histograms are percents rather than counts. If we had used counts with the same scale on the y-axis, then the bars for the skilled workers and the supervisors would be very small because of the smaller sample sizes in these groups. Figure 12.4 gives side-by-side boxplots for these data. We see that the largest and the smallest possible values are present in the data. The distributions are somewhat

skewed toward lower values. Our sample sizes, however, are sufficiently large that we are confident that the sample means are approximately Normal.

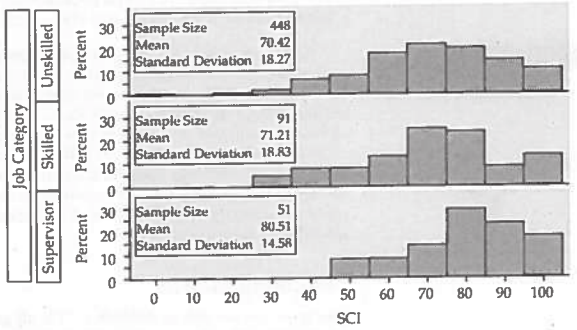


FIGURE 12.3 Histograms and descriptive statistics for the worker safety example.

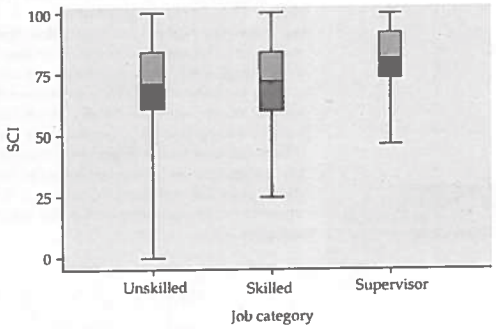


FIGURE 12.4 Side-by-side boxplots for the worker safety example.

The three sample means are plotted in Figure 12.5. It appears that the means for the unskilled workers and the skilled workers are similar, while the supervisors have a higher mean. To apply ANOVA in this setting, we view the three samples that we have as three independent random samples from three distinct populations. Each of these populations has a mean and our inference asks questions about these means.

CAUTION

Formulating a clear definition of the populations being compared with ANOVA can be difficult, as in our example. Often some expert judgment is required, and different consumers of the results may have differing opinions. The workers in this study all worked in the same industry in a particular region. They certainly

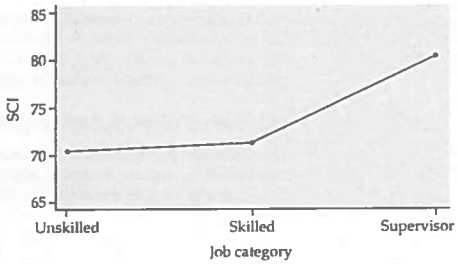


FIGURE 12.5 SCI means for the worker safety example.

do represent some larger population of similar workers. We are more confident in generalizing our conclusions to similar populations when the results are clearly significant than when the level of significance just barely passes the standard of $P = 0.05$.

We first ask whether or not there is sufficient evidence in the data to conclude that the corresponding population means are not all equal. Our null hypothesis here states that the population mean SCI is the same for all three groups of workers. The alternative is that they are not all the same.

Our inspection of the data for our example suggests that the means for the skilled workers and the unskilled workers may be the same while the mean for the supervisors is higher. *Rejecting the null hypothesis that the means are all the same using ANOVA is not the same as concluding that all of the means are different from one another.* The ANOVA null hypothesis can be false in many different ways. Additional analysis is required to distinguish among these possibilities.

contrasts

When there are particular versions of the alternative hypothesis that are of interest, we use **contrasts** to examine them. In our example, we might want to compare the supervisors with all of the other workers. *Note that, to use contrasts, it is necessary that the questions of interest be formulated before examining the data. It is cheating to make up these questions after analyzing the data.*

multiple comparisons

If we have no specific relations among the means in mind before looking at the data, we instead use a **multiple-comparisons** procedure to determine which pairs of population means differ significantly. In later sections we will explore both contrasts and multiple comparisons in detail.

USE YOUR KNOWLEDGE

12.1 What's wrong? For each of the following, explain what is wrong and why.

(a) ANOVA tests the null hypothesis that the sample means are all equal.

(b) A strong case for causation is best made in an observational study.

(c) You use one-way ANOVA when the response variable has only two possible values.

12.2 What's wrong? For each of the following, explain what is wrong and why.

(a) In rejecting the null hypothesis, one can conclude that all the means are different from each other.

(b) A one-way ANOVA can be used only when there are fewer than five means to be compared.

(c) A two-way ANOVA is used when comparing two populations.

The ANOVA model

When analyzing data, the following equation reminds us that we look for an overall pattern and deviations from it:

$$\text{DATA} = \text{FIT} + \text{RESIDUAL}$$

LOOK BACK
DATA = FIT + RESIDUAL, page 564

In the regression model of Chapter 10, the FIT was the population regression line, and the RESIDUAL represented the deviations of the data from this line. We now apply this framework to describe the statistical models used in ANOVA. These models provide a convenient way to summarize the assumptions that are the foundation for our analysis. They also give us the necessary notation to describe the calculations needed.

LOOK BACK
Normal distributions, page 58

First, recall the statistical model for a random sample of observations from a single Normal population with mean μ and standard deviation σ. If the observations are

$$x_1, x_2, \ldots, x_n$$

we can describe this model by saying that the x_i are an SRS from the $N(\mu, \sigma)$ distribution. Another way to describe the same model is to think of the x's varying about their population mean. To do this, write each observation x_i as

$$x_i = \mu + \epsilon_i$$

The ϵ_i are then an SRS from the $N(0, \sigma)$ distribution. Because μ is unknown, the ϵ's cannot actually be observed. This form more closely corresponds to our

$$\text{DATA} = \text{FIT} + \text{RESIDUAL}$$

way of thinking. The FIT part of the model is represented by μ. It is the systematic part of the model, like the line in a regression. The RESIDUAL part is represented by ϵ_i. It represents the deviations of the data from the fit and is due to random, or chance, variation.

There are two unknown parameters in this statistical model: μ and σ. We estimate μ by $\bar{x}$, the sample mean, and σ by s, the sample standard deviation. The differences $e_i = x_i - \bar{x}$ are the sample residuals and correspond to the ϵ_i in the statistical model.

The model for one-way ANOVA is very similar. We take random samples from each of I different populations. The sample size is n_i for the ith population. Let x_{ij} represent the jth observation from the ith population. The I population means are the FIT part of the model and are represented by μ_i. The random

variation, or RESIDUAL, part of the model is represented by the deviations ϵ_{ij} of the observations from the means.

THE ONE-WAY ANOVA MODEL

The **one-way ANOVA model** is

$$x_{ij} = \mu_i + \epsilon_{ij}$$

for $i = 1, \ldots, I$ and $j = 1, \ldots, n_i$. The ϵ_{ij} are assumed to be from an $N(0, \sigma)$ distribution. The **parameters of the model** are the population means $\mu_1, \mu_2, \ldots, \mu_I$ and the common standard deviation σ.

Note that the sample sizes n_i may differ, but the standard deviation σ is assumed to be the same in all of the populations. Figure 12.6 pictures this model for $I = 3$. The three population means μ_i are different, but the shapes of the three Normal distributions are the same, reflecting the assumption that all three populations have the same standard deviation.

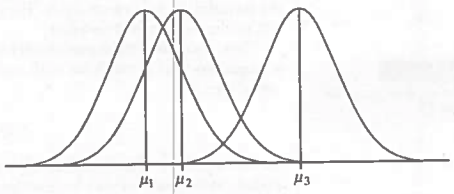


FIGURE 12.6 Model for one-way ANOVA with three groups. The three populations have Normal distributions with the same standard deviation.

EXAMPLE

12.4 ANOVA model for worker safety study. In our worker safety example there are three groups of workers that we want to compare, so $I = 3$. The population means μ_1, μ_2, and μ_3 are the mean SCI values for unskilled workers, for skilled workers, and for supervisors, respectively. The sample sizes n_i are 448, 91, and 51.

The observation $x_{1,1}$ is the SCI score for the first unskilled worker. The data for the other unskilled workers are denoted by $x_{1,2}, x_{1,3}, \ldots, x_{1,448}$. Similarly, the data for the other two groups have a first subscript indicating the group and a second subscript indicating the worker in that group.

According to our model, the SCI for the first worker is $x_{1,1} = \mu_1 + \epsilon_{1,1}$, where μ_1 is the average for *all* unskilled workers and $\epsilon_{1,1}$ is the chance variation due to this particular worker. The ANOVA model assumes that the ϵ_{ij} are independent and Normally distributed with mean 0 and standard deviation σ. We have clear evidence that the data are not Normal in our example. The values are numbers ranging from 0 to 100, and we saw some skewness for all three groups in Figures 12.3 and 12.4. However, because our inference is based on the sample means, which will be approximately Normal, we are not overly concerned about this violation of our assumptions.

It is common to use numerical subscripts to distinguish the different means, and some software requires that levels of factors in ANOVA be specified as numerical values. An alternative is to use subscripts that suggest the actual groups. In our example, we could replace μ_1, μ_2, and μ_3 by μ_{UN}, μ_{SK}, and μ_{SU}.

Estimates of population parameters

The unknown parameters in the statistical model for ANOVA are the I population means μ_i and the common population standard deviation σ. To estimate μ_i we use the sample mean for the ith group:

$$\bar{x}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij}$$

The residuals $e_{ij} = x_{ij} - \bar{x}_i$ reflect the variation about the sample means that we see in the data.

The ANOVA model assumes that the population standard deviations are all equal. If we have unequal standard deviations, we generally try to transform the data so that they are approximately equal. We might, for example, work with $\sqrt{x_{ij}}$ or $\log x_{ij}$. Fortunately, we can often find a transformation that *both* makes the group standard deviations more nearly equal and also makes the distributions of observations in each group more nearly Normal. If the standard deviations are markedly different and cannot be made similar by a transformation, inference requires different methods that are beyond the scope of this book.

Unfortunately, formal tests for the equality of standard deviations in several groups share the lack of robustness against non-Normality that we noted in Chapter 7 for the case of two groups. Because ANOVA procedures are not extremely sensitive to unequal standard deviations, we do *not* recommend a formal test of equality of standard deviations as a preliminary to the ANOVA. Instead, we will use the following rule as a guideline.

LOOK BACK
F test for equality of spread, page 474

RULE FOR EXAMINING STANDARD DEVIATIONS IN ANOVA

If the largest standard deviation is less than twice the smallest standard deviation, we can use methods based on the assumption of equal standard deviations, and our results will still be approximately correct.[2]

When we assume that the population standard deviations are equal, each sample standard deviation is an estimate of σ. To combine these into a single estimate, we use a generalization of the pooling method introduced in Chapter 7.

POOLED ESTIMATOR OF σ

Suppose we have sample variances $s_1^2, s_2^2, \ldots, s_I^2$ from I independent SRSs of sizes $n_1, n_2, \ldots, n_I$ from populations with common variance σ^2. The **pooled sample variance**

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \cdots + (n_I - 1)s_I^2}{(n_1 - 1) + (n_2 - 1) + \cdots + (n_I - 1)}$$

is an unbiased estimator of σ^2. The **pooled standard deviation**

$$s_p = \sqrt{s_p^2}$$

is the estimate of σ.

Pooling gives more weight to groups with larger sample sizes. If the sample sizes are equal, s_p^2 is just the average of the I sample variances. *Note that s_p is not the average of the I sample standard deviations.*

EXAMPLE

12.5 Population estimates for worker safety study. In the worker safety study there are $I = 3$ groups and the sample sizes are $n_1 = 448$, $n_2 = 91$, and $n_3 = 51$. The sample standard deviations are $s_1 = 18.27$, $s_2 = 18.83$, and $s_3 = 14.58$.

Because the largest standard deviation (18.83) is less than twice the smallest ($2 \times 14.58 = 29.16$), our rule indicates that we can use the assumption of equal population standard deviations.

The pooled variance estimate is

$$s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2 + (n_3-1)s_3^2}{(n_1-1)+(n_2-1)+(n_3-1)}$$

$$= \frac{(447)(18.27)^2 + (90)(18.83)^2 + (50)(14.58)^2}{447+90+50}$$

$$= \frac{191{,}745}{587} = 326.7$$

The pooled standard deviation is

$$s_p = \sqrt{326.7} = 18.07$$

This is our estimate of the common standard deviation σ of the SCI scores in the three populations of workers.

USE YOUR KNOWLEDGE

12.3 Computing the pooled standard deviation. An experiment was run to compare three groups. The sample sizes were 25, 22, and 19, and the corresponding estimated standard deviations were 22, 20, and 18.

(a) Is it reasonable to use the assumption of equal standard deviations when we analyze these data? Give a reason for your answer.

(b) Give the values of the variances for the three groups.

(c) Find the pooled variance.

(d) What is the value of the pooled standard deviation?

12.4 Visualizing the ANOVA model. For each of the following situations, draw a picture of the ANOVA model similar to Figure 12.6 (page 645). Use numerical values for the μ_i. To sketch the Normal curves, you may want to review the 68–95–99.7 rule on page 59.

(a) $\mu_1 = 15$, $\mu_2 = 16$, $\mu_3 = 21$, and $\sigma = 6$.

(b) $\mu_1 = 10$, $\mu_2 = 15$, $\mu_3 = 20$, $\mu_4 = 20.1$, and $\sigma = 2.5$.

(c) $\mu_1 = 15$, $\mu_2 = 16$, $\mu_3 = 21$, and $\sigma = 2$.

Testing hypotheses in one-way ANOVA

Comparison of several means is accomplished by using an F statistic to compare the variation among groups with the variation within groups. We now show how the F statistic expresses this comparison. Calculations are organized in an ANOVA table, which contains numerical measures of the variation among groups and within groups.

LOOK BACK
ANOVA table, page 582

First we must specify our hypotheses for one-way ANOVA. As usual, I represents the number of populations to be compared.

HYPOTHESES FOR ONE-WAY ANOVA

The **null and alternative hypotheses** for one-way ANOVA are

$$H_0: \mu_1 = \mu_2 = \cdots = \mu_I$$

H_a: not all of the μ_i are equal

We will now use our worker safety example to illustrate how to do a one-way ANOVA. Because the calculations are generally performed using statistical software, we focus on interpretation of the output.

EXAMPLE

12.6 Reading software output. Figure 12.7 gives descriptive statistics generated by SPSS for the ANOVA of the worker safety example. Summaries for each group are given on the first three lines. In addition to the sample size, the mean, and the standard deviation, this output also gives the minimum and maximum observed value, standard error of the mean and the 95% confidence interval for the mean of each group. The three sample means $\bar{x}_i$ given in the output are estimates of the three unknown population means μ_i.

The output gives the estimates of the standard deviations for each group, the s_i, but does not provide s_p, the pooled estimate of the model standard deviation, σ. We could perform the calculation using a calculator, as we did in Example 12.5. We will see an easier way to obtain this quantity from the ANOVA table in Figure 12.8. Some software packages report s_p as part of the standard ANOVA output. *Sometimes you are not sure whether or not a quantity given by*

SPSS

Descriptives

SCI

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
1	448	70.42	18.269	.863	68.73	72.12	0	100
2	91	71.21	18.830	1.974	67.29	75.13	25	100
3	51	80.51	14.577	2.041	76.41	84.61	46	100
Total	590	71.42	18.260	.752	69.94	72.89	0	100

FIGURE 12.7 Software output with descriptive statistics for the worker safety example.

SPSS

ANOVA

SCI

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	4662.233	2	2331.116	7.137	.001
Within Groups	191729.2	587	326.626		
Total	196391.4	589			

FIGURE 12.8 Software output giving the ANOVA table for the worker safety example.

software is what you think it is. A good way to resolve this dilemma is to do a sample calculation with a simple example to check the numerical results.

Note that s_p is not the standard deviation given in the Total row of Figure 12.7. This quantity is the standard deviation that we would obtain if we viewed the data as a single sample of 590 workers and ignored the possibility that the group means could be different. As we have mentioned many times before, it is important to use care when reading and interpreting software output.

EXAMPLE

12.7 Reading software output, continued. Additional output generated by SPSS for the ANOVA of the worker safety example is given in Figure 12.8. We will discuss some details in the next section. For now, we observe that the results of our significance test are given in the last two columns of the output. The null hypothesis that the three population means are the same is tested by the statistic $F = 7.137$, and the associated P-value is reported as $P = 0.001$. The data provide clear evidence to support the claim that these three groups of workers have different mean SCI values.

The ANOVA table

The information in an analysis of variance is organized in an ANOVA table. To understand the table, it is helpful to think in terms of our

$$\text{DATA} = \text{FIT} + \text{RESIDUAL}$$

view of statistical models. For one-way ANOVA, this corresponds to

$$x_{ij} = \mu_i + \epsilon_{ij}$$

We can think of these three terms as sources of variation. The ANOVA table separates the variation in the data into two parts: the part due to the fit and the remainder, which we call residual.

EXAMPLE

12.8 ANOVA table for worker safety study. The SPSS output in Figure 12.8 gives the sources of variation in the first column. Here, FIT is called Between Groups, RESIDUAL is called Within Groups, and DATA is the last entry, Total. Different software packages use different terms for these sources of variation but the basic concept is common to all. In place of FIT, some software packages use Between Groups, Model, or the name of the factor. Similarly, terms like Within Groups or Error are frequently used in place of RESIDUAL.

variation among groups

The Between Groups row in the table gives information related to the variation **among** group means. In writing ANOVA tables we will use the generic label "groups" or some other term that describes the factor being studied for this row.

variation within groups

The Within Groups row in the table gives information related to the variation **within** groups. We noted that the term "error" is frequently used for this source of variation, particularly for more general statistical models. This label is most appropriate for experiments in the physical sciences where the observations within a group differ because of measurement error. In business and the biological and social sciences, on the other hand, the within-group variation is often due to the fact that not all firms or plants or people are the same. This sort of variation is not due to errors and is better described as "residual" or "within-group" variation. Nevertheless, we will use the generic label "error" for this source of variation in writing ANOVA tables.

Finally, the Total row in the ANOVA table corresponds to the DATA term in our DATA = FIT + RESIDUAL framework. So, for analysis of variance,

$$\text{DATA} = \text{FIT} + \text{RESIDUAL}$$

translates into

$$\text{Total} = \text{Between Groups} + \text{Within Groups}$$

LOOK BACK
sum of squares, page 580

The second column in the software output given in Figure 12.8 is labeled Sum of Squares. As you might expect, each sum of squares is a sum of squared deviations. We use SSG, SSE, and SST for the entries in this column, corresponding to groups, error, and total. Each sum of squares measures a different type of variation. SST measures variation of the data around the overall mean, $x_{ij} - \bar{x}$. Variation of the group means around the overall mean $\bar{x}_i - \bar{x}$ is measured

by SSG. Finally, SSE measures variation of each observation around its group mean, $x_{ij} - \bar{x}_i$.

EXAMPLE

12.9 ANOVA table for worker safety study, continued. The Sum of Squares column in Figure 12.8 gives the values for the three sums of squares.

$$\text{SST} = 196391.4$$
$$\text{SSG} = 4662.2$$
$$\text{SSE} = 191729.2$$

Verify that SST = SSG + SSE.

This fact is true in general. The total variation is always equal to the among-group variation plus the within-group variation. Note that software output frequently gives many more digits than we need, as in this case. In this example it appears that most of the variation is coming from within groups.

LOOK BACK
degrees of freedom, page 42

Associated with each sum of squares is a quantity called the degrees of freedom. Because SST measures the variation of all N observations around the overall mean, its degrees of freedom are DFT $= N - 1$. This is the same as the degrees of freedom for the ordinary sample variance with sample size N. Similarly, because SSG measures the variation of the I sample means around the overall mean, its degrees of freedom are DFG $= I - 1$. Finally, SSE is the sum of squares of the deviations $x_{ij} - \bar{x}_i$. Here we have N observations being compared with I sample means, and DFE $= N - I$.

EXAMPLE

12.10 Degrees of freedom for worker safety study. In our worker safety example, we have $I = 3$ and $N = 590$. Therefore,

$$\text{DFT} = N - 1 = 590 - 1 = 589$$
$$\text{DFG} = I - 1 = 3 - 1 = 2$$
$$\text{DFE} = N - I = 590 - 3 = 587$$

These are the entries in the df column of Figure 12.8.

Note that the degrees of freedom add in the same way that the sums of squares add. That is, DFT = DFG + DFE.

LOOK BACK
mean square, page 581

For each source of variation, the mean square is the sum of squares divided by the degrees of freedom. You can verify this by doing the divisions for the values given on the output in Figure 12.8.

SUMS OF SQUARES, DEGREES OF FREEDOM, AND MEAN SQUARES

Sums of squares represent variation present in the data. They are calculated by summing squared deviations. In one-way ANOVA there are three **sources of variation:** groups, error, and total. The sums of squares are related by the formula

$$\text{SST} = \text{SSG} + \text{SSE}$$

Thus, the total variation is composed of two parts, one due to groups and one due to error.

Degrees of freedom are related to the deviations that are used in the sums of squares. The degrees of freedom are related in the same way as the sums of squares are:

$$\text{DFT} = \text{DFG} + \text{DFE}$$

To calculate each **mean square**, divide the corresponding sum of squares by its degrees of freedom.

We can use the error mean square to find s_p, the pooled estimate of the parameter σ of our model. It is true in general that

$$s_p^2 = \text{MSE} = \frac{\text{SSE}}{\text{DFE}}$$

In other words, the error mean square is an estimate of the within-group variance, σ^2. The estimate of σ is therefore the square root of this quantity. So,

$$s_p = \sqrt{\text{MSE}}$$

EXAMPLE

12.11 MSE for worker safety study. From the SPSS output in Figure 12.8 we see that the MSE is reported as 326.626. The pooled estimate of σ is therefore

$$s_p = \sqrt{\text{MSE}}$$
$$= \sqrt{326.626} = 18.07$$

The *F* test

If H_0 is true, there are no differences among the group means. The ratio MSG/MSE is a statistic that is approximately 1 if H_0 is true and tends to be larger if H_a is true. This is the ANOVA F statistic. In our example, MSG = 2331.116 and MSE = 326.626, so the ANOVA F statistic is

$$F = \frac{\text{MSG}}{\text{MSE}} = \frac{2331.116}{326.626} = 7.137$$

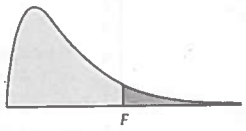
When H_0 is true, the F statistic has an F distribution that depends upon two numbers: the *degrees of freedom for the numerator* and the *degrees of freedom for the denominator.* These degrees of freedom are those associated with the mean squares in the numerator and denominator of the F statistic. For one-way ANOVA, the degrees of freedom for the numerator are DFG $= I - 1$, and the degrees of freedom for the denominator are DFE $= N - I$. We use the notation $F(I - 1, N - I)$ for this distribution.

The *One-Way ANOVA* applet available on the Web site www.whfreeman.com/ips is an excellent way to see how the value of the F statistic and the P-value depend upon the variability of the data within the groups and the differences between the means. See Exercises 12.18 and 12.19 for use of this applet.

THE ANOVA F TEST

To test the null hypothesis in a one-way ANOVA, calculate the **F statistic**

$$F = \frac{\text{MSG}}{\text{MSE}}$$



When H_0 is true, the F statistic has the $F(I-1, N-I)$ distribution. When H_a is true, the F statistic tends to be large. We reject H_0 in favor of H_a if the F statistic is sufficiently large.

The **P-value** of the F test is the probability that a random variable having the $F(I-1, N-I)$ distribution is greater than or equal to the calculated value of the F statistic.

Tables of F critical values are available for use when software does not give the P-value. Table E in the back of the book contains the F critical values for probabilities $p = 0.100, 0.050, 0.025, 0.010$, and 0.001. For one-way ANOVA we use critical values from the table corresponding to $I-1$ degrees of freedom in the numerator and $N-I$ degrees of freedom in the denominator.

EXAMPLE

12.12 The ANOVA F test for the worker safety study. In the study of worker safety, we found $F = 7.14$. (Note that it is standard practice to round F statistics to two places after the decimal point.) There were three populations, so the degrees of freedom in the numerator are $\text{DFG} = I - 1 = 2$. For this example the degrees of freedom in the denominator are $\text{DFE} = N - I = 590 - 3 = 587$. In Table E we first find the column corresponding to 2 degrees of freedom in the numerator. For the degrees of freedom in the denominator, we see that there are entries for 200 and 1000. These entries are very close. To be conservative we use critical values corresponding to 200 degrees of freedom in the denominator since these are slightly larger.

p	Critical value
0.100	2.33
0.050	3.04
0.025	3.76
0.010	4.71
0.001	7.15

We have $F = 7.14$. This is very close to the critical value for $P = 0.001$. Using the table, however, we can conclude only that $P < 0.010$ because our calculated F does not exceed 7.15. (Note that the more accurate calculations performed by software indicated that, in fact, $P < 0.001$.) For this example, we reject H_0 and conclude that the population means are not all the same.

When determining the P-value, remember that the F test is always one-sided because any differences among the group means tend to make F large. The ANOVA F test shares the robustness of the two-sample t test. It is relatively insensitive to moderate non-Normality and unequal variances, especially when the sample sizes are similar.

The following display shows the general form of a one-way ANOVA table with the F statistic. The formulas in the sum of squares column can be used for calculations in small problems. There are other formulas that are more efficient for hand or calculator use, but ANOVA calculations are usually done by computer software.

Source	Degrees of freedom	Sum of squares	Mean square	F
Groups	$I-1$	$\sum_{\text{groups}} n_i(\bar{x}_i - \bar{x})^2$	SSG/DFG	MSG/MSE
Error	$N-I$	$\sum_{\text{groups}} (n_i - 1)s_i^2$	SSE/DFE	
Total	$N-1$	$\sum_{\text{obs}} (x_{ij} - \bar{x})^2$		

coefficient of determination

One other item given by some software for ANOVA is worth noting. For an analysis of variance, we define the **coefficient of determination** as

$$R^2 = \frac{\text{SSG}}{\text{SST}}$$

LOOK BACK
multiple correlation squared, page 614

The coefficient of determination plays the same role as the squared multiple correlation R^2 in a multiple regression. We can easily calculate the value from the ANOVA table entries.

EXAMPLE

12.13 Coefficient of determination for the worker safety study. The software-generated ANOVA table for the worker safety study is given in Figure 12.8. From that display, we see that SSG = 4662.233 and SST = 196,391.4. The coefficient of determination is

$$R^2 = \frac{\text{SSG}}{\text{SST}} = \frac{4662.233}{196,391.4} = 0.02$$

About 2% of the variation in SCI scores is explained by membership in the groups of workers: unskilled workers, skilled workers, and supervisors. The other 98% of the variation is due to worker-to-worker variation within each of the three groups. We can see this in the histograms of Figure 12.3. Each of the groups has a large amount of variation, and there is a substantial amount of overlap in the distributions. *The fact that we have strong evidence ($P < 0.001$)*

against the null hypothesis that the three population means are not all the same does not tell us that the distributions of values are far apart.

USE YOUR KNOWLEDGE

12.5 What's wrong? For each of the following, explain what is wrong and why.

(a) Within-group variation is the variation in the data due to the differences in the sample means.

(b) The mean squares in an ANOVA table will add, that is, MST = MSG + MSE.

(c) The pooled estimate s_p is a parameter of the ANOVA model.

12.6 Determining the critical value of *F*. For each of the following situations, state how large the F statistic needs to be for rejection of the null hypothesis at the 0.05 level.

(a) Compare 5 groups with 3 observations per group.

(b) Compare 5 groups with 6 observations per group.

(c) Compare 5 groups with 9 observations per group.

(d) Summarize what you have learned about F distributions from this exercise.

12.2 Comparing the Means

Contrasts

The ANOVA F test gives a general answer to a general question: are the differences among observed group means significant? Unfortunately, a small P-value simply tells us that the group means are not all the same. It does not tell us specifically which means differ from each other. Plotting and inspecting the means give us some indication of where the differences lie, but we would like to supplement inspection with formal inference.

In the ideal situation, specific questions regarding comparisons among the means are posed before the data are collected. We can answer specific questions of this kind and attach a level of confidence to the answers we give. We now explore these ideas through our worker safety example.

EXAMPLE

12.14 Reporting the results. In the worker safety study we compared the SCI scores for three groups of workers: unskilled workers, skilled workers, and supervisors. Let's use $\bar{x}_{UN}$, $\bar{x}_{SK}$, and $\bar{x}_{SU}$ to represent the three sample means and a similar notation for the population means. From Figure 12.7 we see that the three sample means are

$$\bar{x}_{UN} = 70.42,\ \bar{x}_{SK} = 71.21,\ \text{and}\ \bar{x}_{SU} = 80.51$$

The null hypothesis we tested was

$$H_0: \mu_{UN} = \mu_{SK} = \mu_{SU}$$

Multiple comparisons

multiple-comparisons methods

In many studies, specific questions cannot be formulated in advance of the analysis. If H_0 is not rejected, we conclude that the population means are indistinguishable on the basis of the data given. On the other hand, if H_0 is rejected, we would like to know which pairs of means differ. **Multiple-comparisons methods** address this issue. It is important to keep in mind that multiple-comparisons methods are used only *after rejecting* the ANOVA H_0.

EXAMPLE

12.22 Comparing each pair of groups. Return once more to the worker safety data with three groups of workers. We can make three comparisons between pairs of means: unskilled workers versus skilled workers, unskilled workers versus supervisors, and skilled workers versus supervisors. We can write a t statistic for each of these pairs. For example, the statistic

$$t_{12} = \frac{\bar{x}_1 - \bar{x}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{70.42 - 71.21}{18.07\sqrt{\frac{1}{448} + \frac{1}{91}}} = -0.38$$

compares populations 1 and 2. The subscripts on t specify which groups are compared.

The t statistics for the other two pairs are

$$t_{13} = \frac{\bar{x}_1 - \bar{x}_3}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_3}}} = \frac{70.42 - 80.51}{18.07\sqrt{\frac{1}{448} + \frac{1}{51}}} = -3.78$$

and

$$t_{23} = \frac{\bar{x}_2 - \bar{x}_3}{s_p\sqrt{\frac{1}{n_2} + \frac{1}{n_3}}} = \frac{71.21 - 80.51}{18.07\sqrt{\frac{1}{91} + \frac{1}{51}}} = -2.94$$

LOOK BACK
pooled two-sample t procedures, page 462

We performed the first calculation when we analyzed the contrast $\psi_2 = \mu_1 - \mu_2$ in the previous section. These t statistics are very similar to the pooled two-sample t statistic for comparing two population means. The difference is that we now have more than two populations, so each statistic uses the pooled estimator s_p from all groups rather than the pooled estimator from just the two groups being compared. This additional information about the common σ increases the power of the tests. The degrees of freedom for all of these statistics are DFE = 587, those associated with s_p.

Because we do not have any specific ordering of the means in mind as an alternative to equality, we must use a two-sided approach to the problem of deciding which pairs of means are significantly different.

MULTIPLE COMPARISONS

To perform a **multiple-comparisons procedure,** compute **t statistics** for all pairs of means using the formula

$$t_{ij} = \frac{\bar{x}_i - \bar{x}_j}{s_p\sqrt{\frac{1}{n_i} + \frac{1}{n_j}}}$$

If

$$|t_{ij}| \geq t^{**}$$

we declare that the population means μ_i and μ_j are different. Otherwise, we conclude that the data do not distinguish between them. The value of t^{**} depends upon which multiple-comparisons procedure we choose.

One obvious choice for t^{**} is the upper $\alpha/2$ critical value for the t(DFE) distribution. This choice simply carries out as many separate significance tests of fixed level α as there are pairs of means to be compared. The procedure based on this choice is called the **least-significant differences method,** or simply LSD.

LSD method

CAUTION

LSD has some undesirable properties, particularly if the number of means being compared is large. Suppose, for example, that there are $I = 20$ groups and we use LSD with $\alpha = 0.05$. There are 190 different pairs of means. If we perform 190 t tests, each with an error rate of 5%, our overall error rate will be unacceptably large. We expect about 5% of the 190 to be significant even if the corresponding population means are the same. Since 5% of 190 is 9.5, we expect 9 or 10 false rejections.

The LSD procedure fixes the probability of a false rejection for each single pair of means being compared. It does not control the overall probability of *some* false rejection among all pairs. Other choices of t^{**} control possible errors in other ways. The choice of t^{**} is therefore a complex problem, and a detailed discussion of it is beyond the scope of this text. Many choices for t^{**} are used in practice. One major statistical package allows selection from a list of over a dozen choices.

Bonferroni method

We will discuss only one of these, called the **Bonferroni method.** Use of this procedure with $\alpha = 0.05$, for example, guarantees that the probability of *any* false rejection among all comparisons made is no greater than 0.05. This is much stronger protection than controlling the probability of a false rejection at 0.05 for *each separate* comparison.

EXAMPLE

12.23 Applying the Bonferroni method. We apply the Bonferroni multiple-comparisons procedure with $\alpha = 0.05$ to the data from the worker safety study. The value of t^{**} for this procedure (from software or special tables) is 2.13. Of the statistics $t_{12} = -0.38$, $t_{13} = -3.78$, and $t_{23} = -2.94$ calculated in the beginning of this section, only t_{13} and t_{23} are significant. These two statistics compare supervisors with each of the other two groups.

Of course, we prefer to use software for the calculations.

EXAMPLE

12.24 Interpreting software output. The output generated by SPSS for Bonferroni comparisons appears in Figure 12.10. The software uses an asterisk to indicate that the difference in a pair of means is statistically significant. These results agree with the calculations that we performed in Examples 12.22 and 12.23. Note that each comparison is given twice in the output.

SPSS

Multiple Comparisons

Dependent Variable: SCI
Bonferroni

(I) JOBC	(J) JOBC	Mean Difference (I − J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
1	2	−.78	2.078	1.000	−5.77	4.20
	3	−10.09*	2.671	.001	−16.50	−3.67
2	1	.78	2.078	1.000	−4.20	5.77
	3	−9.30*	3.161	.010	−16.89	−1.71
3	1	10.09*	2.671	.001	3.67	16.50
	2	9.30*	3.161	.010	1.71	16.89

*. The mean difference is significant at the .05 level.

FIGURE 12.10 Software output giving the multiple-comparisons analysis for the worker safety example.

The data in the worker safety study provided a clear result: the supervisors have the highest mean SCI score, and we are unable to see a difference between the unskilled workers and the skilled workers. Unfortunately, this type of clarity does not always emerge from a multiple-comparisons analysis. For example, with three groups, we can (*a*) fail to detect a difference between Groups 1 and

2, (*b*) fail to detect a difference between Groups 2 and 3, and (*c*) conclude that Groups 1 and 3 are not the same. *This kind of apparent contradiction points out dramatically the nature of the conclusions of statistical tests of significance.* The conclusion appears to be illogical. If μ_1 is the same as μ_2 and μ_2 is the same as μ_3, doesn't it follow that μ_1 is the same as μ_3? Logically, the answer must be Yes.

Some of the difficulty can be resolved by noting the choice of words used. In describing the inferences, we talk about failing to detect a difference or concluding that two groups are different. In making logical statements, we say things like "is the same as." There is a big difference between the two modes of thought. Statistical tests ask, "Do we have adequate evidence to distinguish two means?" It is not illogical to conclude that we have sufficient evidence to distinguish μ_1 from μ_3, but not μ_1 from μ_2 or μ_2 from μ_3.

simultaneous confidence intervals

One way to deal with these difficulties of interpretation is to give confidence intervals for the differences. The intervals remind us that the differences are not known exactly. We want to give **simultaneous confidence intervals,** that is, intervals for *all* differences among the population means at once. Again, we must face the problem that there are many competing procedures—in this case, many methods of obtaining simultaneous intervals.

SIMULTANEOUS CONFIDENCE INTERVALS FOR DIFFERENCES BETWEEN MEANS

Simultaneous confidence intervals for all differences $\mu_i - \mu_j$ between population means have the form

$$(\bar{x}_i - \bar{x}_j) \pm t^{**} s_p \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}$$

The critical values t^{**} are the same as those used for the multiple-comparisons procedure chosen.

The confidence intervals generated by a particular choice of t^{**} are closely related to the multiple-comparisons results for that same method. If one of the confidence intervals includes the value 0, then that pair of means will not be declared significantly different, and vice versa.

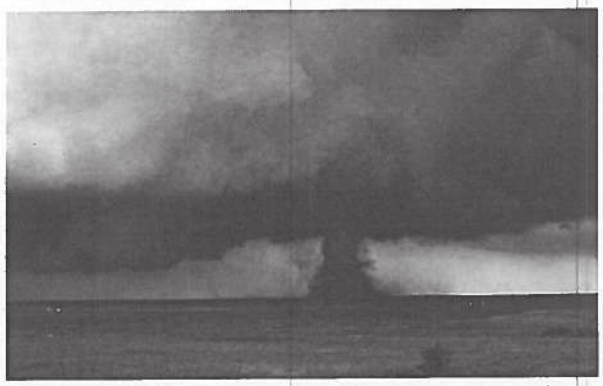
EXAMPLE

12.25 Interpreting software output, continued. The SPSS output for the Bonferroni multiple-comparisons procedure given in Figure 12.10 includes the simultaneous 95% confidence intervals. We can see, for example, that the interval for $\mu_1 - \mu_2$ is −5.77 to 4.20. The fact that the interval includes 0 is consistent with the fact that we failed to detect a difference between these two means using this procedure. Note that the interval for $\mu_2 - \mu_1$ is also provided. This is not really a new piece of information, because it can be obtained from the other interval by reversing the signs and reversing the order, that is, −4.20 to 5.77. So, in fact, we really have only three intervals. Use of the Bonferroni procedure provides us with 95% confidence that *all three* intervals simultaneously contain the true values of the population mean differences.

CHAPTER

10

Inference for Regression



Previously we looked at the average property damage per year due to tornadoes. What about the frequency of tornadoes? Has the annual number of reported tornadoes increased over time? See Exercises 10.23 and 10.24 for more details.

Introduction

In this chapter we describe methods for inference when there is a single quantitative response variable and a single quantitative explanatory variable. The descriptive tools we learned in Chapter 2—scatterplots, least-squares regression, and correlation—are essential preliminaries to inference and also provide a foundation for confidence intervals and significance tests.

We first met the sample mean $\bar{x}$ in Chapter 1 as a measure of the center of a collection of observations. Later we learned that when the data are a random sample from a population, the sample mean is an estimate of the population mean μ. In Chapters 6 and 7, we used $\bar{x}$ as the basis for confidence intervals and significance tests for inference about μ.

Now we will follow the same approach for the problem of fitting straight lines to data. In Chapter 2 we met the least-squares regression line $\hat{y} = b_0 + b_1 x$ as a description of a straight-line relationship between a response variable y and an explanatory variable x. At that point we did not distinguish between sample and population. Now we will think of the least-squares line computed from a sample as an estimate of a *true* regression line for the population.

Following the common practice of using Greek letters for population parameters, we will write the population line as $\beta_0 + \beta_1 x$. This notation reminds us that the intercept b_0 of the fitted line estimates the intercept β_0 of the population line, and the slope b_1 estimates the slope β_1.

The methods detailed in this chapter will help us answers questions such as:

- Is the trend in the annual number of tornadoes reported in the United States linear? If so, what is the average yearly increase in the number of tornadoes? How many are predicted for next year?
- What is the relationship between the selling price of a home and the number of bathrooms that it contains?
- Among North American universities, is there a strong correlation between the binge-drinking rate and the average price for a bottle of beer at establishments within a 2-mile radius of campus?

10.1 Simple Linear Regression

Statistical model for linear regression

Simple linear regression studies the relationship between a response variable y and a single explanatory variable x. We expect that different values of x will produce different mean responses. We encountered a similar but simpler situation in Chapter 7 when we discussed methods for comparing two population means. Figure 10.1 illustrates the statistical model for a comparison of blood pressure change in two groups of experimental subjects, one group taking a calcium supplement and the other a placebo. We can think of the treatment (placebo or calcium) as the explanatory variable in this example. This model has two important parts:

- The mean change may be different in the two populations. These means are labeled μ_1 and μ_2 in Figure 10.1.
- Individual changes in blood pressure vary within each population according to a Normal distribution. The two Normal curves in Figure 10.1 describe the individual responses. These Normal distributions have the same spread, indicating that the population standard deviations are assumed to be equal.

In linear regression the explanatory variable x is quantitative and can have many different values. Imagine, for example, giving different amounts x of calcium to different groups of subjects. We can think of the values of x as defining

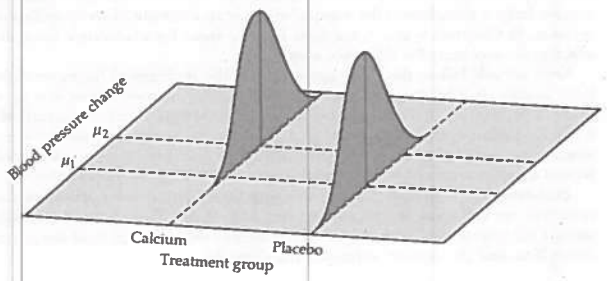


FIGURE 10.1 The statistical model for comparing responses to two treatments; the mean response varies with the treatment.

subpopulations

different **subpopulations,** one for each possible value of x. Each subpopulation consists of all individuals in the population having the same value of x. If we conducted an experiment with five different amounts of calcium, we could view these values as defining five different subpopulations.

The statistical model for simple linear regression also assumes that for each value of x the observed values of the response variable y are Normally distributed with a mean that depends on x. We use μ_y to represent these means. In general, the means μ_y can change as x changes according to any sort of pattern. In **simple linear regression** we assume the means all lie on a line when plotted against x. To summarize, this model also has two important parts:

simple linear regression

- The mean of the response variable y changes as x changes. The means all lie on a straight line. That is, $\mu_y = \beta_0 + \beta_1 x$.
- Individual responses of y with the same x vary according to a Normal distribution. These Normal distributions all have the same standard deviation.

This statistical model is pictured in Figure 10.2. Rather than just two means μ_1 and μ_2, we are interested in how the many means μ_y change as x changes. The simple linear regression model assumes that they all lie on a line when plotted against x. The equation of the line is

$$\mu_y = \beta_0 + \beta_1 x$$

population regression line

with intercept β_0 and slope β_1. This is the **population regression line;** it describes how the mean response changes with x. The line in Figure 10.2 is the population regression line. Observed y's will vary about these means. The three Normal curves show how the response y will vary for three different values of the explanatory variable x. The model assumes that this variation, measured by the standard deviation σ, is the same for all values of x.

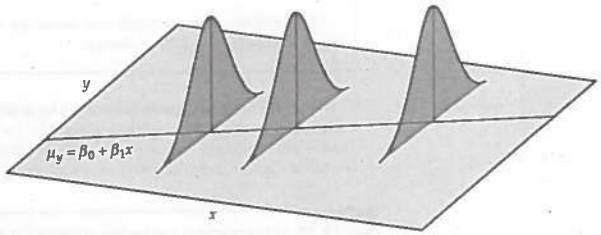


FIGURE 10.2 The statistical model for linear regression; the mean response is a straight-line function of the explanatory variable.

LOOK BACK
random variable, page 259

Data for simple linear regression

The data for a linear regression are observed values of y and x. The model takes each x to be a fixed known quantity. In practice, x may not be exactly known. *If the error in measuring x is large, more advanced inference methods are needed.* The response y to a given x is a random variable. The linear regression model describes the mean and standard deviation of this random variable y. These unknown parameters must be estimated from the data.

We will use the following example to explain the fundamentals of simple linear regression. Because regression calculations in practice are always done by statistical software, we will rely on computer output for the arithmetic. In the next section, we give an example that illustrates how to do the work with a calculator if software is unavailable.

EXAMPLE

10.1 Relationship between speed driven and fuel efficiency. Computers in some vehicles calculate various quantities related to the vehicle's performance. One of these is the fuel efficiency, or gas mileage, expressed as miles per gallon (mpg). Another is the average speed in miles per hour (mph). For one vehicle equipped in this way, mpg and mph were recorded each time the gas tank was filled, and the computer was then reset.[1] How does the speed at which the vehicle is driven affect the fuel efficiency? There are 234 observations available. We will work with a simple random sample of size 60.

Before starting our analysis, it is appropriate to consider the extent to which our results can reasonably be generalized. Because we have a simple random sample from a population of size 234, we are on firm ground in making inferences about this particular vehicle. However, as a practical matter, no one really cares about this particular vehicle. Our results are interesting only if they can be applied to other similar vehicles that are driven under similar conditions. Our statistical modeling for this data set is concerned about the process by which speed affects the fuel efficiency. Although we would not expect the parameters that describe the relationship between speed and fuel efficiency to be *exactly* the same for similar vehicles, we would expect to find qualitatively similar results.

In the statistical model for predicting fuel efficiency from speed, subpopulations are defined by the explanatory variable, speed. For a particular value of speed, say 30 mph, we can think about operating this vehicle repeatedly at this average speed. Variation in driving conditions and the behavior of the driver would be sources of variation that would give different values of mpg for this subpopulation.

EXAMPLE

10.2 Graphical display of the fuel efficiency relationship. We start our analysis with a graphical display of the data. Figure 10.3 is a plot of fuel efficiency versus speed for our sample of 60 observations. We use the variable names MPG and MPH. The least-squares regression line and a smooth function are also shown in the plot. Although there is a positive association between MPG and MPH, the fit is not linear. The smooth function shows us that the relationship levels off somewhat with increasing speed.

CAUTION

Always start with a graphical display of the data. There is no point in trying to do statistical inference if our data do not, at least approximately, meet the assumptions that are the foundation for our inference. At this point we need to make a choice. One possibility would be to confine our interest to speeds that are 30 mph or less, a region where it appears that a line would be a good fit to the data. Another possibility is to make some sort of transformation that will

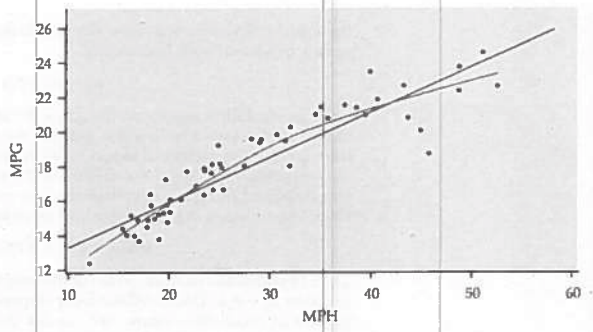


FIGURE 10.3 Scatterplot of MPG versus MPH with a smooth function and the least-squares line, for Example 10.2.

make the relationship approximately linear for the entire set of data. We will choose the second option.

EXAMPLE

10.3 Is this relationship linear? One type of function that looks similar to the smooth-function fit in Figure 10.3 is a logarithm. Therefore, we will examine the effect of transforming speed by taking the natural logarithm. The result is shown in Figure 10.4. In this plot the smooth function and the line are quite close. We are satisfied that the relationship between the log of speed and fuel efficiency is approximately linear for this set of data.

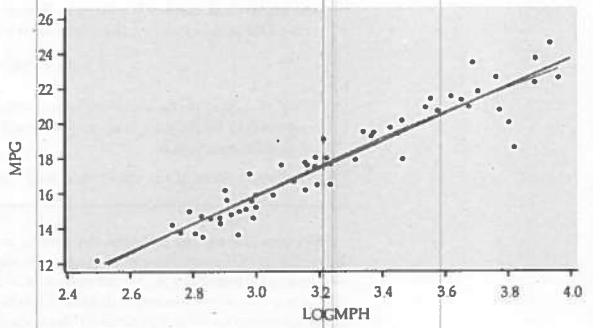


FIGURE 10.4 Scatterplot of MPG versus logarithm of MPH with a smooth function and the least-squares line, for Example 10.3.

Now that we have an approximate linear relationship, we return to predicting fuel efficiency for different subpopulations, defined by the explanatory variable speed. Consider a particular value of speed, for example 30 mph, which in Figure 10.4 would be $x = \log(30) = 3.4$. Our statistical model assumes that

these fuel efficiencies are Normally distributed with a mean μ_y that depends upon x in a linear way. Specifically,

$$\mu_y = \beta_0 + \beta_1 x$$

This population regression line gives the mean fuel efficiency for all values of x. We cannot observe this line, because the observed responses y vary about their means. The statistical model for linear regression consists of the population regression line and a description of the variation of y about the line. This was displayed in Figure 10.2 with the line and the three Normal curves. The following equation expresses this idea in an equation:

$$\text{DATA} = \text{FIT} + \text{RESIDUAL}$$

The FIT part of the model consists of the subpopulation means, given by the expression $\beta_0 + \beta_1 x$. The RESIDUAL part represents deviations of the data from the line of population means. We assume that these deviations are Normally distributed with standard deviation σ. We use ϵ (the Greek letter epsilon) to stand for the RESIDUAL part of the statistical model. A response y is the sum of its mean and a chance deviation ϵ from the mean. The deviations ϵ represent "noise," that is, variation in y due to other causes that prevent the observed (x, y)-values from forming a perfectly straight line on the scatterplot.

SIMPLE LINEAR REGRESSION MODEL

Given n observations of the explanatory variable x and the response variable y,

$$(x_1, y_1),\ (x_2, y_2),\ \ldots,\ (x_n, y_n)$$

the **statistical model for simple linear regression** states that the observed response y_i when the explanatory variable takes the value x_i is

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

Here $\beta_0 + \beta_1 x_i$ is the mean response when $x = x_i$. The deviations ϵ_i are assumed to be independent and Normally distributed with mean 0 and standard deviation σ.

The parameters of the model are β_0, β_1, and σ.

Because the means μ_y lie on the line $\mu_y = \beta_0 + \beta_1 x$, they are all determined by β_0 and β_1. Once we have estimates of β_0 and β_1, the linear relationship determines the estimates of μ_y for all values of x. Linear regression allows us to do inference not only for subpopulations for which we have data but also for those corresponding to x's not present in the data. We will learn how to do inference about

- the slope β_1 and the intercept β_0 of the population regression line,
- the mean response μ_y for a given value of x, and
- an individual future response y for a given value of x.

Estimating the regression parameters

LOOK BACK
least-squares regression, page 112

The method of least squares presented in Chapter 2 fits a line to summarize a relationship between the observed values of an explanatory variable and a response variable. Now we want to use the least-squares line as a basis for inference about a population from which our observations are a sample. We can do this only when the statistical model just presented holds. In that setting, the slope b_1 and intercept b_0 of the least-squares line

$$\hat{y} = b_0 + b_1 x$$

estimate the slope β_1 and the intercept β_0 of the population regression line.

LOOK BACK
least-squares equations, page 114
correlation, page 102
unbiased estimator, page 217

Using the formulas from Chapter 2, the slope of the least-squares line is

$$b_1 = r\frac{s_y}{s_x}$$

and the intercept is

$$b_0 = \bar{y} - b_1 \bar{x}$$

Here, r is the correlation between y and x, s_y is the standard deviation of y, and s_x is the standard deviation of x. Some algebra based on the rules for means of random variables (Section 4.4) shows that b_0 and b_1 are unbiased estimators of β_0 and β_1. Furthermore, b_0 and b_1 are Normally distributed with means β_0 and β_1 and standard deviations that can be estimated from the data. Normality of these sampling distributions is a consequence of the assumption that the ϵ_i are distributed Normally. A general form of the central limit theorem tells us that the distributions of b_0 and b_1 will still be approximately Normal even if the ϵ_i are not. On the other hand, outliers and influential observations can invalidate the results of inference for regression.

LOOK BACK
central limit theorem, page 339

The predicted value of y for a given value x^* of x is the point on the least-squares line $\hat{y} = b_0 + b_1 x^*$. This is an unbiased estimator of the mean response μ_y when $x = x^*$. The **residual** is

residual

$$\begin{aligned} e_i &= \text{observed response} - \text{predicted response} \\ &= y_i - \hat{y}_i \\ &= y_i - b_0 - b_1 x_i \end{aligned}$$

The residuals e_i correspond to the model deviations ϵ_i. The e_i sum to 0, and the ϵ_i come from a population with mean 0.

The remaining parameter to be estimated is σ, which measures the variation of y about the population regression line. Because this parameter is the standard deviation of the model deviations, it should come as no surprise that we use the residuals to estimate it. As usual, we work first with the variance and take the square root to obtain the standard deviation. For simple linear regression, the estimate of σ^2 is the average squared residual

$$\begin{aligned} s^2 &= \frac{\sum e_i^2}{n-2} \\ &= \frac{\sum (y_i - \hat{y}_i)^2}{n-2} \end{aligned}$$

LOOK BACK
sample variance, page 40

We average by dividing the sum by $n - 2$ in order to make s^2 an unbiased estimate of σ^2. The sample variance of n observations uses the divisor $n - 1$ for this same reason. The quantity $n - 2$ is called the **degrees of freedom** for s^2. The estimate of σ is given by

degrees of freedom

$$s = \sqrt{s^2}$$

We will use statistical software to calculate the regression for predicting fuel efficiency with the log of speed for Example 10.3. In entering the data, we chose the names LOGMPH for the log of speed and MPG for fuel efficiency. *It is good practice to use names, rather than just x and y, to remind yourself which data the output describes.*

EXAMPLE

10.4 Statistical software output for fuel efficiency. Figure 10.5 gives the outputs for four commonly used statistical software packages and Excel. Other software will give similar information. The SPSS output reports estimates of our three parameters as $b_0 = -7.796$, $b_1 = 7.874$, and $s = 0.9995$. Be sure that you can find these entries in this output and the corresponding values in the other outputs.

The least-squares regression line is the straight line that is plotted in Figure 10.4. We would report it as

$$\widehat{\text{MPG}} = -7.80 + 7.87\text{LOGMPH}$$

with a model standard deviation of $s = 1.00$. Note that the number of digits provided varies with the software used and we have rounded off the values to three significant digits. *It is important to avoid cluttering up your report of the results of a statistical analysis with many digits that are not relevant.* Software often reports many more digits than are meaningful or useful.

The outputs contain other information that we will ignore for now. Computer outputs often give more information than we want or need. *The experienced user of statistical software learns to ignore the parts of the output that are not needed for the current problem.* This is done to reduce user frustration when a software package does not print out the particular statistics wanted for an analysis.

Now that we have fitted a line, we should examine the residuals for Normality and any remaining patterns in the data. We usually plot the residuals both against the case number (especially if this reflects the order in which the observations were collected) and against the explanatory variable. For this example, in place of case number, we prefer another variable that is similar but is recorded in a more useful scale. It is the total number of miles that the vehicle has been driven.

Figure 10.6 gives a plot of the residuals versus miles driven with a smooth-function fit. The smooth function suggests that the residuals increase slightly up to about 50,000 miles and then tend to decrease somewhat. With the data that we have for this example, it is difficult to decide if this effect is real or due to chance variation. It is not unreasonable to think that the vehicle performance decreases with age. Since the effect does not appear to be particularly large, we

(b) Explain clearly what this slope says about the change in the mean of y for a change in x.

(c) What is the subpopulation mean when $x = 10$?

(d) Between what 2 values would approximately 95% of the observed responses, y, fall when $x = 10$?

10.2 More on speed's effect on fuel efficiency. Refer to Example 10.4.

(a) What is the predicted mpg for the car when it averages 35 mph?

(b) If the observed mpg when $x = 35$ mph were 21.0, what is the residual?

(c) Suppose you wanted to use the estimated population regression line to examine the average mpg at 45, 55, 65, and 75 mph. Discuss the appropriateness of using the equation to predict mpg for each of these speeds.

Confidence intervals and significance tests

Chapter 7 presented confidence intervals and significance tests for means and differences in means. In each case, inference rested on the standard errors of estimates and on t distributions. Inference for the intercept and slope in a linear regression is similar in principle. For example, the confidence intervals have the form

$$\text{estimate} \pm t^* \text{SE}_{\text{estimate}}$$

where t^* is a critical point of a t distribution. It is the formulas for the estimate and standard error that are more complicated.

Confidence intervals and tests for the slope and intercept are based on the Normal sampling distributions of the estimates b_1 and b_0. Standardizing these estimates gives standard Normal z statistics. The standard deviations of these estimates are multiples of σ, the model parameter that describes the variability about the true regression line. Because we do not know σ, we estimate it by s, the variability of the data about the least-squares line. When we do this, we get t distributions with degrees of freedom $n - 2$, the degrees of freedom of s. We give formulas for the standard errors SE_{b_1} and SE_{b_0} in Section 10.2. For now we will concentrate on the basic ideas and let the computer do the computations.

CONFIDENCE INTERVALS AND SIGNIFICANCE TESTS FOR REGRESSION SLOPE AND INTERCEPT

A level C confidence interval for the intercept β_0 is

$$b_0 \pm t^* \text{SE}_{b_0}$$

A level C confidence interval for the slope β_1 is

$$b_1 \pm t^* \text{SE}_{b_1}$$

In these expressions t^* is the value for the $t(n-2)$ density curve with area C between $-t^*$ and t^*.

To test the hypothesis $H_0: \beta_1 = 0$, compute the **test statistic**

$$t = \frac{b_1}{SE_{b_1}}$$

The **degrees of freedom** are $n - 2$. In terms of a random variable T having the $t(n-2)$ distribution, the P-value for a test of H_0 against

$H_a: \beta_1 > 0$ is $P(T \geq t)$

$H_a: \beta_1 < 0$ is $P(T \leq t)$

$H_a: \beta_1 \neq 0$ is $2P(T \geq |t|)$

There is a similar significance test about the intercept β_0 that uses SE_{b_0} and the $t(n-2)$ distribution. Although computer outputs often include a test of $H_0: \beta_0 = 0$, this information usually has little practical value. From the equation for the population regression line, $\mu_y = \beta_0 + \beta_1 x$, we see that β_0 is the mean response corresponding to $x = 0$. In many practical situations, this subpopulation does not exist or is not interesting.

On the other hand, the test of $H_0: \beta_1 = 0$ is quite useful. When we substitute $\beta_1 = 0$ in the model, the x term drops out and we are left with

$$\mu_y = \beta_0$$

This model says that the mean of y does not vary with x. All of the y's come from a single population with mean β_0, which we would estimate by $\bar{y}$. The hypothesis $H_0: \beta_1 = 0$ therefore says that there is no straight-line relationship between y and x and that linear regression of y on x is of no value for predicting y.

EXAMPLE

10.5 Statistical software output, continued. The computer outputs in Figure 10.5 for the fuel efficiency problem contain the information needed for inference about the regression slope and intercept. Let's look at the SPSS output. The column labeled Std. Error gives the standard errors of the estimates. The value of SE_{b_1} appears on the line labeled with the variable name for the explanatory variable, LOGMPH. It is given as 0.354. In a summary we would report that the regression coefficient for the log of speed is 7.87 with a standard error of 0.35.

The t statistic and P-value for the test of $H_0: \beta_1 = 0$ against the two-sided alternative $H_a: \beta_1 \neq 0$ appear in the columns labeled t and Sig. We can verify the t calculation from the formula for the standardized estimate:

$$t = \frac{b_1}{SE_{b_1}} = \frac{7.874}{0.354} = 22.24$$

The P-value is given as 0.000. This is a rounded number and from that information we can conclude that $P < 0.0005$. The other outputs in Figure 10.5 also indicate that the P-value is very small. We will report the result as $P < 0.001$ because 1 chance in 1000 is sufficiently small for us to decisively reject the null hypothesis.

We have found a statistically significant linear relationship between fuel efficiency and log speed. The estimated slope is more than 22 standard deviations away from zero. Because this is extremely unlikely to happen if the true slope is zero, we have strong evidence for our claim. Note, however, that this is not the same as concluding that we have found a strong relationship between the response and explanatory variables in this example. *A very small P-value for the significance test for a zero slope does not necessarily imply that we have found a strong relationship.* A confidence interval will provide additional information about the relationship.

CAUTION

EXAMPLE

10.6 Confidence interval for the slope. A confidence interval for β_1 requires a critical value t^* from the $t(n-2) = t(58)$ distribution. In Table D there are entries for 50 and 60 degrees of freedom. The values for these rows are very similar. To be conservative, we will use the larger critical value, for 50 degrees of freedom. Find the confidence level values at the bottom of the table. In the 95% confidence column the entry for 50 degrees of freedom is $t^* = 2.009$.

To compute the 95% confidence interval for β_1 we combine the estimate of the slope with the margin of error:

$$b_1 \pm t^* SE_{b_1} = 7.874 \pm (2.009)(0.354)$$
$$= 7.874 \pm 0.711$$

The interval is (7.16, 8.58). This agrees with the value given by the software outputs that provide this information in Figure 10.5. We estimate that an increase of 1 in the logarithm of speed is associated with an increase of between 7.16 and 8.58 mpg.

To interpret the interval, it is useful to translate the statement back to the original mph scale. From Figure 10.4 we can see that the values for LOGMPH range from about 2.5 to 3.9. Let's translate the increase of 1 unit in LOGMPH to the mph scale by considering a change from 2.8 to 3.8. Since $\log(16.4) = 2.8$ and $\log(44.7) = 3.8$, the change corresponds to an increase in speed from 16.4 to 44.7 mph. An increase in average speed from 16.4 to 44.7 mph is associated with an increase of 7.8 ± 0.7 in mpg.

Note that the intercept in this example is not of practical interest. It estimates mpg when the logarithm of mph (that's x) is 0, a value that cannot occur. For this reason, we do not compute a confidence interval for β_0.

Confidence intervals for mean response

For any specific value of x, say x^*, the mean of the response y in this subpopulation is given by