Poglavlje 1

Uvod

Uporaba riječi **statistika** u svakodnevnom životu najčešće je povezana s brojčanim vrijednostima kojima pokušavamo opisati bitne karakteristike nekog skupa podataka. Na službenim mrežnim stranicama Državnog zavoda za statistiku Republike Hrvatske možemo pročitati (http://www.dzs.hr, 5. rujna 2012.):

Prosječna mjesečna isplaćena neto plaća po zaposlenome u pravnim osobama Republike Hrvatske za lipanj 2012. iznosila je 5492 *kune.*

Minimalna plaća za razdoblje od 1. lipnja 2012. do 31. svibnja 2012. u Republici Hrvatskoj iznosila je 2814 *kuna.*

Stopa registrirane nezaposlenosti za srpanj 2012. iznosila je 17.5%.

Udio aktivnog stanovništva u radno sposobnom stanovništvu (stopa aktivnosti) za siječanj, veljaču i ožujak 2012. iznosila je 51.7%, *istovremeno* 42.9% *radno sposobnih osoba je zaposleno (stopa zaposlenosti), a* 17% *radne snage je nezaposleno (stopa nezaposlenosti).*

Temelj statistike kao znanstvene discipline, kao i svih istraživanja koja se koriste statističkim metodama, čine skupovi podataka.

Statistika kao znanstvena disciplina bavi se razvojem metoda prikupljanja, opisivanja i analiziranja podataka te primjenom tih metoda u procesu donošenja zaključaka na temelju prikupljenih podataka.

Statističko istraživanje fokusirano je na skup **objekata**, tj. **jedinki** (ljudi, životinja, biljaka, stvari, država, gradova, poduzeća, itd.) i skup odabranih veličina koje

se na njima promatraju. Veličine koje se promatraju zovemo **varijablama**. Sve jedinke koje se žele obuhvatiti istraživanjem, tj. o kojima se želi zaključivati, čine **populaciju**.

Primjer 1.1. *Bavimo se istraživanjem uspjeha studenata jedne generacije na ispitu iz kolegija Statistika na nekom sveučilištu (tablica 1.1).*

Jedinke	osobe, imenom i prezimenom ili nekom šifrom
Varijabla	ocjena iz Statistike

Tablica 1.1: Primjer jedinki i varijabli obuhvaćenih opisanim istraživanjem.

U tom primjeru navedena je samo jedna varijabla koja se analizira na jedinkama populacije, tj. uspjeh iz kolegija Statistika. Međutim, često nas zanima nekoliko varijabli i/ili veze među njima. Primjerice, želimo li ispitati ovisi li uspjeh iz kolegija u prethodnom primjeru o spolu studenta, potrebno je u istraživanju populacije za svaku jedinku zabilježiti i vrijednost varijable spol (M ili Ž), a želimo li ispitati ovisi li uspjeh o pripadnosti pojedinoj grupi vježbi, potrebno je za svaku jedinku zabilježiti koju je grupu vježbi pohađala. Zbog preglednosti prikupljene podatke prikazujemo tablično tako da jedan redak odgovara točno jednoj jedinki, a stupac točno jednoj varijabli.

Primjer 1.2. *Bavimo se istraživanjem uspjeha studenata jedne generacije na ispitu iz kolegija Statistika na nekom sveučilištu u ovisnosti o spolu ispitanika i grupi vježbi koju je student pohađao. U ovom slučaju istraživanje se temelji na jedinkama i varijablama prikazanima u tablici 1.2.*

Jedinke	studenti, identificirani svojim matičnim brojem
Varijable	ocjena iz Statistike, spol, grupa vježbi

Tablica 1.2: Istraživanje uspjeha studenata - jedinke i varijable.

Tablicu za bilježenje prikupljenih podataka treba organizirati na način prikazan tablicom 1.3.

Matični broj studenta	Ocjena iz Statistike	Spol	Grupa vježbi
1206	5	Ž	A
1326	2	Ž	B
942	4	Ž	C
⋮	⋮	⋮	⋮

Tablica 1.3: Istraživanje uspjeha studenata - tablica prikupljenih podataka.

U prethodnim primjerima možemo lako istražiti cijelu populaciju s obzirom da generacija koju proučavamo broji konačno mnogo studenata (npr. 83 studenta). Međutim, istražujemo li prije izbora za predsjednika neke države preferencije građana prema nekom od kandidata, ne možemo ispitati sve osobe populacije (tj. sve državljane koji imaju pravo glasa) jer bi to bilo provođenje izbora. Kada nije moguće istražiti veličine koje nas zanimaju na svim jedinkama populacije, potrebno je iz populacije izdvojiti **uzorak** na kojemu će biti prikupljeni podaci. S obzirom da se o cijeloj populaciji želi zaključivati na temelju podataka prikupljenih na uzorku, za istraživanje je vrlo važno znati kako kreirati kvalitetan uzorak.

Primjena statistike u istraživanju podrazumijeva da se u pripremi istraživanja izabranog problema poštuju sljedeća pravila:

Populaciju koja je predmet istraživanja i ciljeve potrebno je jasno odrediti (detaljno proučiti populaciju, zabilježiti njene osnovne karakteristike i ciljeve istraživanja).

Kreirati kvalitetan uzorak i odabrati metodu za prikupljanje podataka.

Izabrati prikladne metode za opis skupa prikupljenih podataka (deskriptivna statistika).

Izabrati prikladne statističke metode za zaključivanje o populaciji na temelju prikupljenih podataka na uzorku.

U skladu s tim u ovom ćemo se kolegiju baviti nekim **metodama prikupljanja podataka i kreiranja uzorka**, **metodama deskriptivne statistike** i **metodama statističkog zaključivanja**. S obzirom da se metode kojima se kreira uzorak i metode statističkog zaključivanja temelje na poznavanju osnovnih pojmova teorije vjerojatnosti, u kolegiju ćemo također navesti temeljne pojmove i zakone teorije vjerojatnosti potrebne za razumijevanje osnovnog statističkog aparata.

Poglavlje 2

Prikupljanje i organizacija podataka

2.1 Populacija i uzorak

Statističko istraživanje usmjereno je na skup jedinki koje zadovoljavaju neka svojstva bitna za obilježje koje se istražuje, tj. **populaciju**. Dakle, **populaciju čine sve jedinke koje su predmet istraživanja.**

Primjer 2.1. *Istražujemo razlike u prehrambenim navikama između stanovnika Slavonije i Baranje i stanovnika Dalmacije. Populaciju čine svi stanovnici Slavonije, Baranje i Dalmacije. Međutim, ako nas zanimaju samo prehrambene navike studenata iz tih područja, onda populaciju čine samo studenti iz Slavonije, Baranje i Dalmacije.*

Populacija može sadržavati vrlo velik broj jedinki i stoga je često teško, ili čak nemoguće, istraživanje provesti na svim jedinkama populacije. Rješenje tog problema sastoji se u odabiru jednog podskupa populacije, koji nazivamo **uzorak**, na kojemu je osigurano kvalitetno provođenje istraživanja.

Da bi zaključci prilikom istraživanja o populaciji na temelju podataka iz uzorka bili ispravni, nužno je da uzorak bude **reprezentativan**, tj. u njemu moraju biti zastupljene tipične karakteristike populacije bitne za istraživanje.

Primjer 2.2. *U prethodnom primjeru, ako populaciju čine svi stanovnici Slavonije, Baranje i Dalmacije, istraživanje ne možemo provesti samo na uzorku djece koja pohađaju srednju školu. To bi možda bilo praktično, ali takav uzorak nije reprezentativan za zaključivanje o cijeloj populaciji.*

Jedan od načina izbora jedinki iz populacije u uzorak jest formiranje takozvanog **slučajnog uzorka** , uz poštivanje zahtjeva da svaka jedinka populacije ima jednaku vjerojatnost (šansu) ući u uzorak.

S obzirom da se u gornjoj definiciji pojavljuje pojam **vjerojatnost**, metodu formiranja slučajnog uzorka ostavljamo za sljedeća poglavlja, nakon što pojasnimo pojam vjerojatnosti.

2.2 Izvori podataka

Način prikupljanja podataka ovisi o karakteristikama obilježja koje je predmet proučavanja. Najčešće korišteni načini prikupljanja podataka jesu sljedeći:

Podaci iz javnih izvora (knjige, časopisi, novine, Internet).

Podaci iz dizajniranog eksperimenta (istraživač raspoređuje eksperimentalne jedinke u skupine s kojima provodi eksperimente te bilježi podatke za varijable koje ga zanimaju).

Podaci iz ankete (istraživač sastavlja anketni upitnik, izabire skupinu ljudi koju anketira i na osnovi njihovih odgovora prikuplja podatke).

Podaci prikupljeni promatranjem (istraživač promatra eksperimentalne jedinke u njihovu prirodnom okruženju i bilježi podatke za varijable od interesa).

Primjer 2.3. *Jedno medicinsko istraživanje proučava snagu nekog lijeka u prevenciji moždanog udara. Ljude s kojima će se provesti istraživanje istraživač dijeli na dvije skupine: tretiranu i kontrolnu. Ljudima u tretiranoj skupini daje se lijek, dok se ljudima u kontrolnoj skupini daje placebo (nadomjestak koji izgleda isto kao lijek, ali zapravo nije ništa što može imati bilo kakav utjecaj na organizam). To istraživanje primjer je dizajniranog eksperimenta kojim se mogu prikupiti određeni podaci o ispitanicima.*

2.3 Tipovi varijabli

U statističkim istraživanjima razlikujemo nekoliko osnovnih tipova varijabli koje se međusobno razlikuju po svojstvima vrijednosti koje mogu poprimiti.

2.3.1 Kvalitativne varijable

Karakteristika je kvalitativnih varijabli da njihove vrijednosti nisu, po svojim svojstvima korištenim u istraživanju, realni brojevi. Tipičan je primjer takve varijable

spol osobe. Vrijednosti kvalitativne varijable uobičajeno svrstavamo u kategorije. Kategorije kvalitativnih varijabli mogu biti definirane u skladu s potrebama statističkog istraživanja.

Primjer 2.4. *Sljedeće su varijable kvalitativnog tipa:*

- *radna mjesta u školi (spremačica, domar, tajnik, nastavnik, pedagog, ravnatelj)*
- *opisne ocjene (ništa, malo, srednje, puno)*
- *boja očiju (plava, smeđa, zelena)*
- *krvne grupe (A, B, AB, 0)*
- *spol (m ili ž).*

2.3.2 Numeričke varijable

Numeričke varijable prirodno primaju vrijednosti iz skupa realnih brojeva. Tipičan primjeri numeričkih varijabli jesu tjelesna masa i visina osobe. Međutim, treba naglasiti da se i kategorije kvalitativnih varijabli mogu izražavati brojevima, što ih ne čini numeričkim varijablama. Primjerice, spol osobe je jedna kvalitativna varijabla. Kategoriju "ženski spol" možemo označiti npr. oznakom "1", a kategoriju "muški spol" npr. oznakom "2", što može biti korisno prilikom unošenja podataka u bazu. Time smo kategorijama kvalitativne varijable pridružili numeričke vrijednosti, ali samu varijablu nismo učinili numeričkom po njenim svojstvima.

Primjer 2.5. *Sljedeće su varijable numeričkog tipa:*

- *postotak prolaznosti na pojedinim ispitima tijekom jedne akademske godine*
- *broj bodova na državnoj maturi iz matematike*
- *broj ulovljenih komaraca u klopku*
- *temperatura mora*
- *koncentracija soli u morskoj vodi.*

Među numeričkim varijablama razlikujemo **diskretne** i **neprekidne** varijable.

Diskretne numeričke varijable mogu poprimiti samo konačno ili prebrojivo mnogo vrijednosti, dok je skup mogućih vrijednosti neprekidnih numeričkih varijabli cijeli skup realnih brojeva ili neki interval.

Primjer 2.6. *Sljedeće su numeričke varijable diskretne:*

- *broj bodova na državnoj maturi iz matematike*
- *broj ulovljenih komaraca u klopku*
- *broj dana u godini s temperaturom zraka većim od $35^o\,C$.*

Primjer 2.7. *Sljedeće su numeričke varijable neprekidne:*

- *postotak prolaznosti na pojedinim ispitima tijekom jedne akademske godine*
- *temperatura mora*
- *vodostaj neke rijeke.*

Radi prikaza podataka i nekih statističkih analiza vrijednosti numeričke varijable također se mogu svrstati u kategorije. Za razliku od kategorija kvalitativne varijable, među kategorijama numeričke varijable uvijek se može prepoznati prirodan poredak.

Primjer 2.8. (auto-centar.sta)
Svrha ovog primjera je prikazati mogućnost kategorizacije numeričke varijable. Taj se postupak najčešće rješava stvaranjem nove kvalitativne varijable čije su vrijednosti svrstane u kategorije kojih je (znatno) manje nego svih mogućih vrijednosti odgovarajuće diskretne numeričke varijable. Baza podataka auto-centar.sta *sastoji se od sljedećih varijabli:*

automobili - *diskretna numerička varijabla koja sadrži podatke o broju prodanih automobila u jednom danu za sto promatranih dana. Budući da broj prodanih automobila u jednom danu može biti vrlo mali (npr. samo nekoliko osobnih automobila), ali i vrlo velik (npr. narudžbe automobila za vozni park nekog poduzeća), zaključujemo da varijabla* automobili *može poprimiti velik broj različitih vrijednosti iz skupa prirodnih brojeva. Zato je u nekim situacijama korisno kategorizirati vrijednosti ove varijable prema točno određenom kriteriju. Na primjer, kategrizacija prema broju prodanih automobila u jednom danu može se realizirati stvaranjem nove varijable* kategorija.

kategorija - *kvalitativna varijabla koja podatke iz varijable* automobili *svrstava u pet kategorija prema kriteriju prikazanom u tablici 2.8.*

broj prodanih automobila	kategorija
0 - 9	E
10 i 11	D
12 i 13	C
14 i 15	B
16 i više	A

Tablica 2.1: Primjer kategorizacije diskretne numeričke varijable automobili.

2.3.3 Ordinalne varijable

Karakteristika je ordinalnih varijabli da su one po svom karakteru kvalitativne, ali među kategorijama se može uspostaviti prirodan poredak. Tipičan je primjer takve varijable stručna sprema osobe.

Primjer 2.9. (matematika.sta)
Baza podataka matematika.sta *sadrži podatke prikupljene anketiranjem studenata nakon održanih predavanja, vježbi, kolokvija te usmenog ispita iz jednog matematičkog kolegija. Prikupljeni podaci organizirani su na sljedeći način:*

prosjek - *varijabla koja sadrži podatke o prosječnoj ocjeni studiranja za* 49 *anketiranih studenata,*

polozeno - *varijabla koja studente svrstava u dvije kategorije s obzirom na to jesu li položili ispit iz promatranog kolegija prema kriteriju prikazanom u tablici 2.2.*

položen/nepoložen ispit	kategorija
položen ispit	1
nepoložen ispit	0

Tablica 2.2: Kategorizacija studenata prema položenosti ispita.

predavanja, vjezbe - *dvije varijable koje prisutnost studenata na predavanjima/vježbama (p/v) svrstavaju u tri kategorije na način prikazan u tablici 2.3.*

prisutnost studenta na p/v	kategorija
student s p/v nije nikada izostao	1
student je s p/v izostao samo jednom	2
student je s p/v izostao barem dva puta	3

Tablica 2.3: Kategorizacija studenata prema broju izostanaka s predavanja/vježbi.

tezina kolegija, materijali - *dvije varijable koje sadrže subjektivne ocjene (u standardnoj skali od 1 do 5) studenata o težini kolegija i dostatnosti dostupnih materijala za pripremanje ispita iz promatranog kolegija.*

Uočimo da se varijabla prosjek *može promatrati kao neprekidna numerička varijabla, varijabla* položeno *je kvalitativna, dok se varijable* predavanja, vjezbe, tezina kolegija i materijali *mogu svrstati u ordinalne varijable.*

2.4 Organizacija baze podataka

Podaci u bazi podataka mogu biti organizirani na različite načine ovisno o informacijama koje želimo dobiti istraživanjem. Za ilustraciju navodimo jedan primjer niza podataka koji su organizirani na dva različita načina.

Primjer 2.10. (student.sta, student-grupe.sta)
Svrha je ovog primjera pokazati kako isti podaci u bazi podataka mogu biti organizirani na različite načine. Način organizacije ovisi o informacijama koje iz podataka želimo dobiti statističkom analizom. Baza podataka student.sta *sastoji se od sljedećih varijabli:*

klasicno studiranje - *neprekidna numerička varijabla koja sadrži podatke o godinama starosti studenata koji studiraju na klasičan način (stanuju u gradu u kojem studiraju ili putuju na predavanja)*

e-learning - *neprekidna numerička varijabla koja sadrži podatke o godinama starosti studenata koji studiraju putem interneta (tzv. e-learning).*

Baza podataka student-grupe.sta *sastoji se od sljedećih varijabli:*

dob studenta - *neprekidna numerička varijabla koja sadrži podatke o godinama starosti za sto studenata koji studiraju ili na klasičan način ili putem interneta*

nacin studiranja - *kvalitativna varijabla koja studente, bez obzira na podatke sadržane u varijabli* dob studenta*, svrstava u dvije kategorije prema kriteriju prikazanom u tablici 2.4.*

način studiranja	kategorija
student studira na klasičan način	1
student studira putem interneta	0

Tablica 2.4: Primjer kategorizacije studenata prema načinu studiranja.

Dakle, baze podataka student.sta *i* student-grupe.sta *sadrže iste podatke (godine starosti sto promatranih studenata) i daju informaciju o načinu studiranja za svakog studenta:*

- *u bazi podataka* student.sta *podaci o dobi studenata organizirani su u dvije varijable, ovisno o tome studira li student na klasičan način (*klasicno studiranje*) ili putem interneta (*e-learning*)*
- *u bazi podataka* student-grupe.sta *varijabla* dob studenta *sadrži podatke o dobi studenata, dok binarna varijabla* nacin studiranja *za svakog studenta sadrži informaciju o načinu studiranja (tablica 2.4).*

2.5 Zadaci

Zadatak 2.1. (stanovnistvo.sta)
Pretpostavimo da želite saznati starosnu strukturu (prema godinama starosti) stanovništva u svom gradu te da ste u tu svrhu prikupili podatke koji su dani u bazi stanovnistvo.sta. Navedena baza sadrži četiri varijable:

osnovna škola - varijabla koja sadrži podatke o godinama starosti za pedeset slučajno odabranih učenika jedne osnovne škole u vašem gradu

kafić - varijabla koja sadrži podatke o godinama starosti za pedeset slučajno odabranih gostiju popularnog kafića u vašem gradu

gradska knjižnica - varijabla koja sadrži podatke o godinama starosti za pedeset slučajno odabranih posjetitelja gradske knjižnice u vašem gradu

telefonska anketa - varijabla koja sadrži podatke o godinama starosti za pedeset osoba iz vašeg grada čije ste telefonske brojeve na slučajan način izabrali iz telefonskog imenika.

Nakon kratke analize baze podataka stanovnistvo.sta komentirajte reprezentativnost uzorka. Razmislite o mogućim načinima prikupljanja podataka kojima biste kreirali reprezentativan uzorak za proučavanje starosne strukture populacije.

Zadatak 2.2. (glukoza.sta)
Baza podataka glukoza.sta sastoji se od sljedećih varijabli:

dob - neprekidna numerička varijabla koja sadrži podatke o godinama starosti 102 promatrane osobe.

koncentracija - neprekidna numerička varijabla koja sadrži podatke o koncentraciji glukoze u krvi za svaku od 102 promatrane osobe.

kategorija - kvalitativna varijabla koja podatke iz varijable koncentracija glukoze svrstava u dvije kategorije (svaka je kategorija jedan interval pozitivnih realnih brojeva) na način prikazan u tablici 2.5.

interval koncentracije glukoze	kategorija
koncentracija < 6 mMol/L	N - normalna koncentracija
koncentracija ≥ 6 mMol/L	P - povišena koncenracija

Tablica 2.5: Primjer kategorizacije neprekidne numeričke varijable koncentracija.

Predložite neku drugu kategorizaciju varijable koncentracija i usporedite je s varijablom kategorija koju je u istu svrhu formirao istraživač u pokusu.

Zadatak 2.3. (kolegij.sta)
Baza podataka sastoji se od sljedećih varijabli:

godina upisa - *kvalitativna varijabla koja sadrži podatke o akademskoj godini upisa na studij za sto promatranih studenata*

kategorija - *kvalitativna varijabla koja podatke iz varijable* godina upisa *svrstava u tri kategorije (svaka je kategorija jedan konačan skup) na način prikazan u tablici 2.6.*

godina upisa	kategorija
student upisan prije 1990. godine	1
student upisan 1990., 1991. ili 1992. godine	2
student upisan 1993. ili 1994. godine	3

Tablica 2.6: Primjer kategorizacije kvalitativne varijable godina upisa.

opća kemija, organska kemija, anorganska kemija, mikrobiologija - *četiri ordinalne varijable koje sadrže podatke o postignutim ocjenama na ispitima iz spomenutih kolegija za svakog od sto promatranih studenata*

prosjek - *neprekidna numerička varijabla koja sadrži prosječne ocjene iz četiriju spomenuta kolegija za svakog od sto promatranih studenata*

uspjeh - *kvalitativna varijabla koja vrijednosti varijable* prosjek *svrstava u četiri kategorije prema kriteriju prikazanom u tablici 2.7.*

prosjek	uspjeh	prosjek	uspjeh	prosjek	uspjeh	prosjek	uspjeh
$[2, 2.5>$	dovoljan	$[2.5, 3.5>$	dobar	$[3.5, 4.5>$	vrlo dobar	$[4.5, 5]$	izvrstan

Tablica 2.7: Primjer kategorizacije neprekidne numeričke varijable prosjek.

Predložite drugačije kategorizacije varijabli godina upisa *i* uspjeh *i obrazložite svoj prijedlog kategorizacije.*

Zadatak 2.4. *Na sličan način proanalizirajte i odredite tipove varijabli u sljedećim bazama podataka:*

a) baza podataka komarci.sta *sadrži dio rezultata proučavanja komaraca u jednom močvarnom području (dostupni su podaci za* 210 *mjerenja na istoj lokaciji):*

varijable brojM *i* brojZ *redom sadrže broj muških i ženskih jedinki komaraca*

varijabla mjesec *sadrži mjesečevu mijenu (M - mlađak, U - uštap) za svako mjerenje*

varijabla doba dana *sadrži doba dana u kojem je mjerenje obavljeno (P - predvečerje, N - noć, S - svitanje)*

varijabla svjetlost *sadrži tip osvjetljenja pri mjerenju*

varijabla temperatura *sadrži temperaturu pri kojoj je mjerenje izvršeno*

varijabla rel vlaznost *sadrži relativnu vlažnost zraka za vrijeme mjerenja*

b) u bazi podataka navike.sta *nalaze se rezultati praćenja nekih životnih navika u jednom danu za svakog od* 300 *ispitanika iz uzorka:*

varijabla dnevne novine *sadrži broj prelistanih različitih dnevnih novina*

varijabla tv vijesti *sadrži broj pogledanih televizijskih vijesti na dostupnim televizijskim kanalima*

varijabla kava *sadrži broj ispijenih kava*

varijabla troskovi *sadrzi informaciju o troškovima hrane za promatrani dan*

varijabla vrijeme *sadrži ispitanikov subjektivan doživljaj vremenskih prilika u njegovu mjestu stanovanja (O - oblačno, S - sunčano)*

varijabla raspolozenje *sadrži ispitanikovu subjektivnu ocjenu vlastitog raspoloženja (L - loše, D - dobro, O - odlično)*

c) u bazi podataka posao.sta *nalaze se podaci o udaljenosti mjesta stanovanja od radnog mjesta (varijabla* udaljenost*) i mjesečnim troškovima putovanja do radnog mjesta (varijabla* troskovi*) za* 100 *slučajno odabranih zaposlenih ljudi*

d) baza podataka TV-program.sta *sastoji se od sljedećih varijabli:*

varijabla spol *sadrži informaciju o spolu ispitanika*

varijable P1, P2, P3 *i* P4 *sadrže subjektivne ocjene kvalitete ljetne programske sheme televizijskih programa P1, P2, P3 i P4*

varijabla prosjek *sadrži prosječnu ocjenu kvalitete ljetne programske sheme navedenih televizijskih programa*

e) u bazi podataka zdravlje.sta *nalaze se neki zdravstveni podaci anketiranih ispitanika:*

varijable godine *i* spol *sadrže podatke o starosti u godinama i spolu ispitanika*

vrijednosti varijable zdravlje *su subjektivne ocjene vlastitog zdravstvenog stanja ispitanika*

varijabla broj pregleda *sadrži informacije o ukupnom broju zdravstvenih pregleda svakog ispitanika u tekućoj kalendarskoj godini*

varijabla dodatno zdravstveno *sadrži podatke o dodatnom zdravstvenom osiguranju svakog ispitanika (*1 - *ispitanik je dodatno osiguran;* 0 - *ispitanik nije dodatno osiguran)*

varijabla cijena *sadrži cijenu u kunama najskupljeg zdravstvenog pregleda svakog ispitanika (u tekućoj kalendarskoj godini)*

f) baza podataka djelatnici.sta *sadrži podatke o uzorcima djelatnika dviju konkurentskih tvornica - tvornice A i tvornice B. U tablici s imenom "tvornica A" zabilježene su vrijednosti sljedećih varijabli za djelatnike* **tvornice** *A:*

varijabla spol *sadrži informaciju o spolu (*M - *muški spol,* Z - *ženski spol)*

varijabla odjel *sadrži naziv odjela u kojem je djelatnik zaposlen (*TR - *transport,* P- *pakiranje,* IS - *isporuka)*

varijabla obrazovanje *sadrži stručnu spremu djelatnika (*SSS - *srednja stručna sprema,* VŠSS - *viša stručna sprema,* VSS - *visoka stručna sprema)*

varijabla dob *sadrži starost djelatnika u godinama*

varijabla visina *sadrži visinu djelatnika u centimetrima*

varijabla rukovostvo *sadrži broj godina rada koje je djelatnik proveo na nekoj od rukovodećih pozicija u toj tvornici*

varijabla placa prije *sadrži iznos godišnje plaće djelatnika prije reorganizacije poslovnog sustava*

varijabla placa poslije *sadrži iznos godišnje plaće djelatnika nakon reorganizacije poslovnog sustava.*

U tablici s imenom "tvornica B", u varijabli placa konkurencija*, zabilježeni su iznosi godišnje plaće za svakog djelatnika iz uzorka iz* **tvornice** *B.*

Chapter 1

Introduction to Statistics and Data Analysis

1.1 Overview: Statistical Inference, Samples, Populations, and the Role of Probability

Beginning in the 1980s and continuing into the 21st century, an inordinate amount of attention has been focused on *improvement of quality* in American industry. Much has been said and written about the Japanese "industrial miracle," which began in the middle of the 20th century. The Japanese were able to succeed where we and other countries had failed–namely, to create an atmosphere that allows the production of high-quality products. Much of the success of the Japanese has been attributed to the use of *statistical methods* and statistical thinking among management personnel.

Use of Scientific Data

The use of statistical methods in manufacturing, development of food products, computer software, energy sources, pharmaceuticals, and many other areas involves the gathering of information or **scientific data**. Of course, the gathering of data is nothing new. It has been done for well over a thousand years. Data have been collected, summarized, reported, and stored for perusal. However, there is a profound distinction between collection of scientific information and **inferential statistics**. It is the latter that has received rightful attention in recent decades.

The offspring of inferential statistics has been a large "toolbox" of statistical methods employed by statistical practitioners. These statistical methods are designed to contribute to the process of making scientific judgments in the face of **uncertainty** and **variation**. The product density of a particular material from a manufacturing process will not always be the same. Indeed, if the process involved is a batch process rather than continuous, there will be not only variation in material density among the batches that come off the line (batch-to-batch variation), but also within-batch variation. Statistical methods are used to analyze data from a process such as this one in order to gain more sense of where in the process changes may be made to improve the **quality** of the process. In this process, qual-

ity may well be defined in relation to closeness to a target density value in harmony with *what portion of the time* this closeness criterion is met. An engineer may be concerned with a specific instrument that is used to measure sulfur monoxide in the air during pollution studies. If the engineer has doubts about the effectiveness of the instrument, there are two **sources of variation** that must be dealt with. The first is the variation in sulfur monoxide values that are found at the same locale on the same day. The second is the variation between values observed and the **true** amount of sulfur monoxide that is in the air at the time. If either of these two sources of variation is exceedingly large (according to some standard set by the engineer), the instrument may need to be replaced. In a biomedical study of a new drug that reduces hypertension, 85% of patients experienced relief, while it is generally recognized that the current drug, or "old" drug, brings relief to 80% of patients that have chronic hypertension. However, the new drug is more expensive to make and may result in certain side effects. Should the new drug be adopted? This is a problem that is encountered (often with much more complexity) frequently by pharmaceutical firms in conjunction with the FDA (Federal Drug Administration). Again, the consideration of variation needs to be taken into account. The "85%" value is based on a certain number of patients chosen for the study. Perhaps if the study were repeated with new patients the observed number of "successes" would be 75%! It is the natural variation from study to study that must be taken into account in the decision process. Clearly this variation is important, since variation from patient to patient is endemic to the problem.

Variability in Scientific Data

In the problems discussed above the statistical methods used involve dealing with variability, and in each case the variability to be studied is that encountered in scientific data. If the observed product density in the process were always the same and were always on target, there would be no need for statistical methods. If the device for measuring sulfur monoxide always gives the same value and the value is accurate (i.e., it is correct), no statistical analysis is needed. If there were no patient-to-patient variability inherent in the response to the drug (i.e., it either always brings relief or not), life would be simple for scientists in the pharmaceutical firms and FDA and no statistician would be needed in the decision process. Statistics researchers have produced an enormous number of analytical methods that allow for analysis of data from systems like those described above. This reflects the true nature of the science that we call inferential statistics, namely, using techniques that allow us to go beyond merely reporting data to drawing conclusions (or inferences) about the scientific system. Statisticians make use of fundamental laws of probability and statistical inference to draw conclusions about scientific systems. Information is gathered in the form of **samples**, or collections of **observations**. The process of sampling is introduced in Chapter 2, and the discussion continues throughout the entire book.

Samples are collected from **populations**, which are collections of all individuals or individual items of a particular type. At times a population signifies a scientific system. For example, a manufacturer of computer boards may wish to eliminate defects. A sampling process may involve collecting information on 50 computer boards sampled randomly from the process. Here, the population is all

computer boards manufactured by the firm over a specific period of time. If an improvement is made in the computer board process and a second sample of boards is collected, any conclusions drawn regarding the effectiveness of the change in process should extend to the entire population of computer boards produced under the "improved process." In a drug experiment, a sample of patients is taken and each is given a specific drug to reduce blood pressure. The interest is focused on drawing conclusions about the population of those who suffer from hypertension.

Often, it is very important to collect scientific data in a systematic way, with planning being high on the agenda. At times the planning is, by necessity, quite limited. We often focus only on certain properties or characteristics of the items or objects in the population. Each characteristic has particular engineering or, say, biological importance to the "customer," the scientist or engineer who seeks to learn about the population. For example, in one of the illustrations above the quality of the process had to do with the product density of the output of a process. An engineer may need to study the effect of process conditions, temperature, humidity, amount of a particular ingredient, and so on. He or she can systematically move these **factors** to whatever levels are suggested according to whatever prescription or **experimental design** is desired. However, a forest scientist who is interested in a study of factors that influence wood density in a certain kind of tree cannot necessarily design an experiment. This case may require an **observational study** in which data are collected in the field but **factor levels** can not be preselected. Both of these types of studies lend themselves to methods of statistical inference. In the former, the quality of the inferences will depend on proper planning of the experiment. In the latter, the scientist is at the mercy of what can be gathered. For example, it is sad if an agronomist is interested in studying the effect of rainfall on plant yield and the data are gathered during a drought.

The importance of statistical thinking by managers and the use of statistical inference by scientific personnel is widely acknowledged. Research scientists gain much from scientific data. Data provide understanding of scientific phenomena. Product and process engineers learn a great deal in their off-line efforts to improve the process. They also gain valuable insight by gathering production data (on-line monitoring) on a regular basis. This allows them to determine necessary modifications in order to keep the process at a desired level of quality.

There are times when a scientific practitioner wishes only to gain some sort of summary of a set of data represented in the sample. In other words, inferential statistics is not required. Rather, a set of single-number statistics or **descriptive statistics** is helpful. These numbers give a sense of center of the location of the data, variability in the data, and the general nature of the distribution of observations in the sample. Though no specific statistical methods leading to **statistical inference** are incorporated, much can be learned. At times, descriptive statistics are accompanied by graphics. Modern statistical software packages allow for computation of **means**, **medians**, **standard deviations**, and other single-number statistics as well as production of graphs that show a "footprint" of the nature of the sample. Definitions and illustrations of the single-number statistics and graphs, including histograms, stem-and-leaf plots, scatter plots, dot plots, and box plots, will be given in sections that follow.

The Role of Probability

In this book, Chapters 2 to 6 deal with fundamental notions of probability. A thorough grounding in these concepts allows the reader to have a better understanding of statistical inference. Without some formalism of probability theory, the student cannot appreciate the true interpretation from data analysis through modern statistical methods. It is quite natural to study probability prior to studying statistical inference. Elements of probability allow us to quantify the strength or "confidence" in our conclusions. In this sense, concepts in probability form a major component that supplements statistical methods and helps us gauge the strength of the statistical inference. The discipline of probability, then, provides the transition between descriptive statistics and inferential methods. Elements of probability allow the conclusion to be put into the language that the science or engineering practitioners require. An example follows that will enable the reader to understand the notion of a P-value, which often provides the "bottom line" in the interpretation of results from the use of statistical methods.

Example 1.1: Suppose that an engineer encounters data from a manufacturing process in which 100 items are sampled and 10 are found to be defective. It is expected and anticipated that occasionally there will be defective items. Obviously these 100 items represent the sample. However, it has been determined that in the long run, the company can only tolerate 5% defective in the process. Now, the elements of probability allow the engineer to determine how conclusive the sample information is regarding the nature of the process. In this case, the **population** conceptually represents all possible items from the process. Suppose we learn that *if the process is acceptable*, that is, if it does produce items no more than 5% of which are defective, there is a probability of 0.0282 of obtaining 10 or more defective items in a random sample of 100 items from the process. This small probability suggests that the process does, indeed, have a long-run rate of defective items that exceeds 5%. In other words, under the condition of an acceptable process, the sample information obtained would rarely occur. However, it did occur! Clearly, though, it would occur with a much higher probability if the process defective rate exceeded 5% by a significant amount. ◢

From this example it becomes clear that the elements of probability aid in the translation of sample information into something conclusive or inconclusive about the scientific system. In fact, what was learned likely is alarming information to the engineer or manager. Statistical methods, which we will actually detail in Chapter 10, produced a P-value of 0.0282. The result suggests that the process **very likely is not acceptable**. The concept of a ***P*-value** is dealt with at length in succeeding chapters. The example that follows provides a second illustration.

Example 1.2: Often the nature of the scientific study will dictate the role that probability and deductive reasoning play in statistical inference. Exercise 9.40 on page 294 provides data associated with a study conducted at the Virginia Polytechnic Institute and State University on the development of a relationship between the roots of trees and the action of a fungus. Minerals are transferred from the fungus to the trees and sugars from the trees to the fungus. Two samples of 10 northern red oak seedlings were planted in a greenhouse, one containing seedlings treated with nitrogen and

the other containing seedlings with no nitrogen. All other environmental conditions were held constant. All seedlings contained the fungus *Pisolithus tinctorus*. More details are supplied in Chapter 9. The stem weights in grams were recorded after the end of 140 days. The data are given in Table 1.1.

Table 1.1: Data Set for Example 1.2

No Nitrogen	Nitrogen
0.32	0.26
0.53	0.43
0.28	0.47
0.37	0.49
0.47	0.52
0.43	0.75
0.36	0.79
0.42	0.86
0.38	0.62
0.43	0.46

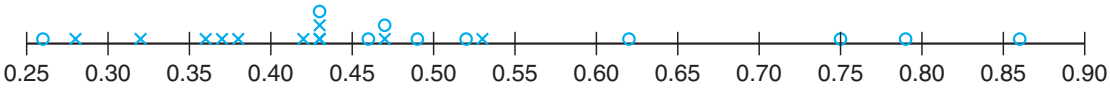


Figure 1.1: A dot plot of stem weight data.

In this example there are two samples from two **separate populations**. The purpose of the experiment is to determine if the use of nitrogen has an influence on the growth of the roots. The study is a comparative study (i.e., we seek to compare the two populations with regard to a certain important characteristic). It is instructive to plot the data as shown in the dot plot of Figure 1.1. The ∘ values represent the "nitrogen" data and the × values represent the "no-nitrogen" data.

Notice that the general appearance of the data might suggest to the reader that, on average, the use of nitrogen increases the stem weight. Four nitrogen observations are considerably larger than any of the no-nitrogen observations. Most of the no-nitrogen observations appear to be below the center of the data. The appearance of the data set would seem to indicate that nitrogen is effective. But how can this be quantified? How can all of the apparent visual evidence be summarized in some sense? As in the preceding example, the fundamentals of probability can be used. The conclusions may be summarized in a probability statement or P-value. We will not show here the statistical inference that produces the summary probability. As in Example 1.1, these methods will be discussed in Chapter 10. The issue revolves around the "probability that data like these could be observed" *given that nitrogen has no effect*, in other words, given that both samples were generated from the same population. Suppose that this probability is small, say 0.03. That would certainly be strong evidence that the use of nitrogen does indeed influence (apparently increases) average stem weight of the red oak seedlings. ∎

How Do Probability and Statistical Inference Work Together?

It is important for the reader to understand the clear distinction between the discipline of probability, a science in its own right, and the discipline of inferential statistics. As we have already indicated, the use or application of concepts in probability allows real-life interpretation of the results of statistical inference. As a result, it can be said that statistical inference makes use of concepts in probability. One can glean from the two examples above that the sample information is made available to the analyst and, with the aid of statistical methods and elements of probability, conclusions are drawn about some feature of the population (the process does not appear to be acceptable in Example 1.1, and nitrogen does appear to influence average stem weights in Example 1.2). Thus for a statistical problem, **the sample along with inferential statistics allows us to draw conclusions about the population, with inferential statistics making clear use of elements of probability**. This reasoning is *inductive* in nature. Now as we move into Chapter 2 and beyond, the reader will note that, unlike what we do in our two examples here, we will not focus on solving statistical problems. Many examples will be given in which no sample is involved. There will be a population clearly described with all features of the population known. Then questions of importance will focus on the nature of data that might hypothetically be drawn from the population. Thus, one can say that **elements in probability allow us to draw conclusions about characteristics of hypothetical data taken from the population, based on known features of the population**. This type of reasoning is *deductive* in nature. Figure 1.2 shows the fundamental relationship between probability and inferential statistics.

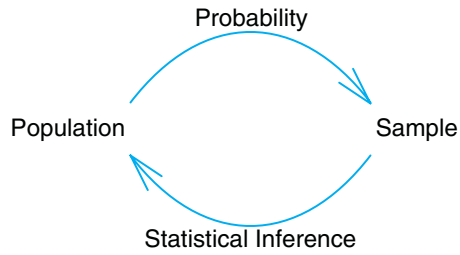


Figure 1.2: Fundamental relationship between probability and inferential statistics.

Now, in the grand scheme of things, which is more important, the field of probability or the field of statistics? They are both very important and clearly are complementary. The only certainty concerning the pedagogy of the two disciplines lies in the fact that if statistics is to be taught at more than merely a "cookbook" level, then the discipline of probability must be taught first. This rule stems from the fact that nothing can be learned about a population from a sample until the analyst learns the rudiments of uncertainty in that sample. For example, consider Example 1.1. The question centers around whether or not the population, defined by the process, is no more than 5% defective. In other words, the conjecture is that **on the average** 5 out of 100 items are defective. Now, the sample contains 100 items and 10 are defective. Does this support the conjecture or refute it? On the

surface it would appear to be a refutation of the conjecture because 10 out of 100 seem to be "a bit much." But without elements of probability, how do we know? Only through the study of material in future chapters will we learn the conditions under which the process is acceptable (5% defective). The probability of obtaining 10 or more defective items in a sample of 100 is 0.0282.

We have given two examples where the elements of probability provide a summary that the scientist or engineer can use as evidence on which to build a decision. The bridge between the data and the conclusion is, of course, based on foundations of statistical inference, distribution theory, and sampling distributions discussed in future chapters.

1.2 Sampling Procedures; Collection of Data

In Section 1.1 we discussed very briefly the notion of sampling and the sampling process. While sampling appears to be a simple concept, the complexity of the questions that must be answered about the population or populations necessitates that the sampling process be very complex at times. While the notion of sampling is discussed in a technical way in Chapter 8, we shall endeavor here to give some common-sense notions of sampling. This is a natural transition to a discussion of the concept of variability.

Simple Random Sampling

The importance of proper sampling revolves around the degree of confidence with which the analyst is able to answer the questions being asked. Let us assume that only a single population exists in the problem. Recall that in Example 1.2 two populations were involved. **Simple random sampling** implies that any particular sample of a specified *sample size* has the same chance of being selected as any other sample of the same size. The term **sample size** simply means the number of elements in the sample. Obviously, a table of random numbers can be utilized in sample selection in many instances. The virtue of simple random sampling is that it aids in the elimination of the problem of having the sample reflect a different (possibly more confined) population than the one about which inferences need to be made. For example, a sample is to be chosen to answer certain questions regarding political preferences in a certain state in the United States. The sample involves the choice of, say, 1000 families, and a survey is to be conducted. Now, suppose it turns out that random sampling is not used. Rather, all or nearly all of the 1000 families chosen live in an urban setting. It is believed that political preferences in rural areas differ from those in urban areas. In other words, the sample drawn actually confined the population and thus the inferences need to be confined to the "limited population," and in this case confining may be undesirable. If, indeed, the inferences need to be made about the state as a whole, the sample of size 1000 described here is often referred to as a **biased sample**.

As we hinted earlier, simple random sampling is not always appropriate. Which alternative approach is used depends on the complexity of the problem. Often, for example, the sampling units are not homogeneous and naturally divide themselves into nonoverlapping groups that are homogeneous. These groups are called *strata*,

and a procedure called *stratified random sampling* involves random selection of a sample *within* each stratum. The purpose is to be sure that each of the strata is neither over- nor underrepresented. For example, suppose a sample survey is conducted in order to gather preliminary opinions regarding a bond referendum that is being considered in a certain city. The city is subdivided into several ethnic groups which represent natural strata. In order not to disregard or overrepresent any group, separate random samples of families could be chosen from each group.

Experimental Design

The concept of randomness or random assignment plays a huge role in the area of **experimental design**, which was introduced very briefly in Section 1.1 and is an important staple in almost any area of engineering or experimental science. This will be discussed at length in Chapters 13 through 15. However, it is instructive to give a brief presentation here in the context of random sampling. A set of so-called **treatments** or **treatment combinations** becomes the populations to be studied or compared in some sense. An example is the nitrogen versus no-nitrogen treatments in Example 1.2. Another simple example would be "placebo" versus "active drug," or in a corrosion fatigue study we might have treatment combinations that involve specimens that are coated or uncoated as well as conditions of low or high humidity to which the specimens are exposed. In fact, there are four treatment or factor combinations (i.e., 4 populations), and many scientific questions may be asked and answered through statistical and inferential methods. Consider first the situation in Example 1.2. There are 20 diseased seedlings involved in the experiment. It is easy to see from the data themselves that the seedlings are different from each other. Within the nitrogen group (or the no-nitrogen group) there is considerable **variability** in the stem weights. This variability is due to what is generally called the **experimental unit**. This is a very important concept in inferential statistics, in fact one whose description will not end in this chapter. The nature of the variability is very important. If it is too large, stemming from a condition of excessive nonhomogeneity in experimental units, the variability will "wash out" any detectable difference between the two populations. Recall that in this case that did not occur.

The dot plot in Figure 1.1 and P-value indicated a clear distinction between these two conditions. What role do those experimental units play in the data-taking process itself? The common-sense and, indeed, quite standard approach is to assign the 20 seedlings or experimental units **randomly to the two treatments or conditions**. In the drug study, we may decide to use a total of 200 available patients, patients that clearly will be different in some sense. They are the experimental units. However, they all may have the same chronic condition for which the drug is a potential treatment. Then in a so-called **completely randomized design**, 100 patients are assigned randomly to the placebo and 100 to the active drug. Again, it is these experimental units within a group or treatment that produce the variability in data results (i.e., variability in the measured result), say blood pressure, or whatever drug efficacy value is important. In the corrosion fatigue study, the experimental units are the specimens that are the subjects of the corrosion.

Why Assign Experimental Units Randomly?

What is the possible negative impact of not randomly assigning experimental units to the treatments or treatment combinations? This is seen most clearly in the case of the drug study. Among the characteristics of the patients that produce variability in the results are age, gender, and weight. Suppose merely by chance the placebo group contains a sample of people that are predominately heavier than those in the treatment group. Perhaps heavier individuals have a tendency to have a higher blood pressure. This clearly biases the result, and indeed, any result obtained through the application of statistical inference may have little to do with the drug and more to do with differences in weights among the two samples of patients.

We should emphasize the attachment of importance to the term **variability**. Excessive variability among experimental units "camouflages" scientific findings. In future sections, we attempt to characterize and quantify measures of variability. In sections that follow, we introduce and discuss specific quantities that can be computed in samples; the quantities give a sense of the nature of the sample with respect to center of location of the data and variability in the data. A discussion of several of these single-number measures serves to provide a preview of what statistical information will be important components of the statistical methods that are used in future chapters. These measures that help characterize the nature of the data set fall into the category of **descriptive statistics**. This material is a prelude to a brief presentation of pictorial and graphical methods that go even further in characterization of the data set. The reader should understand that the statistical methods illustrated here will be used throughout the text. In order to offer the reader a clearer picture of what is involved in experimental design studies, we offer Example 1.3.

Example 1.3: A corrosion study was made in order to determine whether coating an aluminum metal with a corrosion retardation substance reduced the amount of corrosion. The coating is a protectant that is advertised to minimize fatigue damage in this type of material. Also of interest is the influence of humidity on the amount of corrosion. A corrosion measurement can be expressed in thousands of cycles to failure. Two levels of coating, no coating and chemical corrosion coating, were used. In addition, the two relative humidity levels are 20% relative humidity and 80% relative humidity.

The experiment involves four treatment combinations that are listed in the table that follows. There are eight experimental units used, and they are aluminum specimens prepared; two are assigned randomly to each of the four treatment combinations. The data are presented in Table 1.2.

The corrosion data are averages of two specimens. A plot of the averages is pictured in Figure 1.3. A relatively large value of cycles to failure represents a small amount of corrosion. As one might expect, an increase in humidity appears to make the corrosion worse. The use of the chemical corrosion coating procedure appears to reduce corrosion.

In this experimental design illustration, the engineer has systematically selected the four treatment combinations. In order to connect this situation to concepts with which the reader has been exposed to this point, it should be assumed that the

Table 1.2: Data for Example 1.3

Coating	Humidity	Average Corrosion in Thousands of Cycles to Failure
Uncoated	20%	975
	80%	350
Chemical Corrosion	20%	1750
	80%	1550

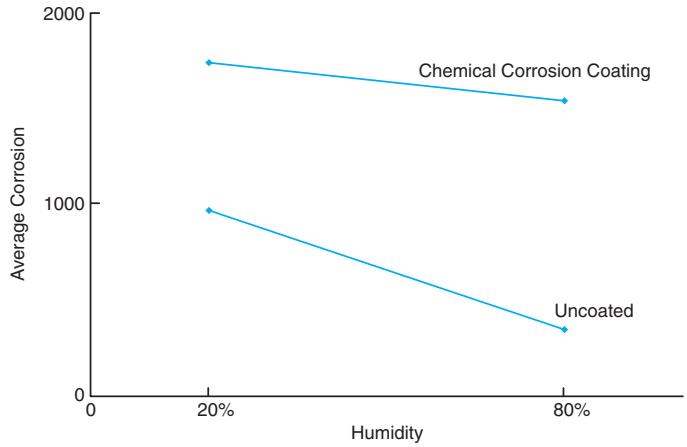


Figure 1.3: Corrosion results for Example 1.3.

conditions representing the four treatment combinations are four separate populations and that the two corrosion values observed for each population are important pieces of information. The importance of the average in capturing and summarizing certain features in the population will be highlighted in Section 1.3. While we might draw conclusions about the role of humidity and the impact of coating the specimens from the figure, we cannot truly evaluate the results from an analytical point of view without taking into account the *variability around* the average. Again, as we indicated earlier, if the two corrosion values for each treatment combination are close together, the picture in Figure 1.3 may be an accurate depiction. But if each corrosion value in the figure is an average of two values that are widely dispersed, then this variability may, indeed, truly "wash away" any information that appears to come through when one observes averages only. The foregoing example illustrates these concepts:

(1) random assignment of treatment combinations (coating, humidity) to experimental units (specimens)

(2) the use of sample averages (average corrosion values) in summarizing sample information

(3) the need for consideration of measures of variability in the analysis of any sample or sets of samples

⊙ **Guided Practice 1.1** Of the 224 patients in the treatment group, 45 had a stroke by the end of the first year. Using these two numbers, compute the proportion of patients in the treatment group who had a stroke by the end of their first year. (Please note: answers to all in-text exercises are provided using footnotes.)[2]

We can compute summary statistics from the table. A **summary statistic** is a single number summarizing a large amount of data.[3] For instance, the primary results of the study after 1 year could be described by two summary statistics: the proportion of people who had a stroke in the treatment and control groups.

Proportion who had a stroke in the treatment (stent) group: $45/224 = 0.20 = 20\%$.

Proportion who had a stroke in the control group: $28/227 = 0.12 = 12\%$.

These two summary statistics are useful in looking for differences in the groups, and we are in for a surprise: an additional 8% of patients in the treatment group had a stroke! This is important for two reasons. First, it is contrary to what doctors expected, which was that stents would *reduce* the rate of strokes. Second, it leads to a statistical question: do the data show a "real" difference between the groups?

This second question is subtle. Suppose you flip a coin 100 times. While the chance a coin lands heads in any given coin flip is 50%, we probably won't observe exactly 50 heads. This type of fluctuation is part of almost any type of data generating process. It is possible that the 8% difference in the stent study is due to this natural variation. However, the larger the difference we observe (for a particular sample size), the less believable it is that the difference is due to chance. So what we are really asking is the following: is the difference so large that we should reject the notion that it was due to chance?

While we don't yet have our statistical tools to fully address this question on our own, we can comprehend the conclusions of the published analysis: there was compelling evidence of harm by stents in this study of stroke patients.

Be careful: do not generalize the results of this study to all patients and all stents. This study looked at patients with very specific characteristics who volunteered to be a part of this study and who may not be representative of all stroke patients. In addition, there are many types of stents and this study only considered the self-expanding Wingspan stent (Boston Scientific). However, this study does leave us with an important lesson: we should keep our eyes open for surprises.

1.2 Data basics

Effective presentation and description of data is a first step in most analyses. This section introduces one structure for organizing data as well as some terminology that will be used throughout this book.

1.2.1 Observations, variables, and data matrices

Table 1.3 displays rows 1, 2, 3, and 50 of a data set concerning 50 emails received during early 2012. These observations will be referred to as the `email50` data set, and they are a random sample from a larger data set that we will see in Section 1.7.

[2] The proportion of the 224 patients who had a stroke within 365 days: $45/224 = 0.20$.

[3] Formally, a summary statistic is a value computed from the data. Some summary statistics are more useful than others.

	spam	num_char	line_breaks	format	number
1	no	21,705	551	html	small
2	no	7,011	183	html	big
3	yes	631	28	text	none
⋮	⋮	⋮	⋮	⋮	⋮
50	no	15,829	242	html	small

Table 1.3: Four rows from the `email50` data matrix.

variable	description
`spam`	Specifies whether the message was spam
`num_char`	The number of characters in the email
`line_breaks`	The number of line breaks in the email (not including text wrapping)
`format`	Indicates if the email contained special formatting, such as bolding, tables, or links, which would indicate the message is in HTML format
`number`	Indicates whether the email contained no number, a small number (under 1 million), or a large number

Table 1.4: Variables and their descriptions for the `email50` data set.

Each row in the table represents a single email or **case**.[4] The columns represent characteristics, called **variables**, for each of the emails. For example, the first row represents email 1, which is a not spam, contains 21,705 characters, 551 line breaks, is written in HTML format, and contains only small numbers.

In practice, it is especially important to ask clarifying questions to ensure important aspects of the data are understood. For instance, it is always important to be sure we know what each variable means and the units of measurement. Descriptions of all five email variables are given in Table 1.4.

The data in Table 1.3 represent a **data matrix**, which is a common way to organize data. Each row of a data matrix corresponds to a unique case, and each column corresponds to a variable. A data matrix for the stroke study introduced in Section 1.1 is shown in Table 1.1 on page 8, where the cases were patients and there were three variables recorded for each patient.

Data matrices are a convenient way to record and store data. If another individual or case is added to the data set, an additional row can be easily added. Similarly, another column can be added for a new variable.

⊙ **Guided Practice 1.2** We consider a publicly available data set that summarizes information about the 3,143 counties in the United States, and we call this the `county` data set. This data set includes information about each county: its name, the state where it resides, its population in 2000 and 2010, per capita federal spending, poverty rate, and five additional characteristics. How might these data be organized in a data matrix? Reminder: look in the footnotes for answers to in-text exercises.[5]

Seven rows of the `county` data set are shown in Table 1.5, and the variables are summarized in Table 1.6. These data were collected from the US Census website.[6]

[4] A case is also sometimes called a **unit of observation** or an **observational unit**.

[5] Each county may be viewed as a case, and there are eleven pieces of information recorded for each case. A table with 3,143 rows and 11 columns could hold these data, where each row represents a county and each column represents a particular piece of information.

[6] quickfacts.census.gov/qfd/index.html

	name	state	pop2000	pop2010	fed_spend	poverty	homeownership	multiunit	income	med_income	smoking_ban
1	Autauga	AL	43671	54571	6.068	10.6	77.5	7.2	24568	53255	none
2	Baldwin	AL	140415	182265	6.140	12.2	76.7	22.6	26469	50147	none
3	Barbour	AL	29038	27457	8.752	25.0	68.0	11.1	15875	33219	none
4	Bibb	AL	20826	22915	7.122	12.6	82.9	6.6	19918	41770	none
5	Blount	AL	51024	57322	5.131	13.4	82.0	3.7	21070	45549	none
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
3142	Washakie	WY	8289	8533	8.714	5.6	70.9	10.0	28557	48379	none
3143	Weston	WY	6644	7208	6.695	7.9	77.9	6.5	28463	53853	none

Table 1.5: Seven rows from the `county` data set.

variable	description
`name`	County name
`state`	State where the county resides (also including the District of Columbia)
`pop2000`	Population in 2000
`pop2010`	Population in 2010
`fed_spend`	Federal spending per capita
`poverty`	Percent of the population in poverty
`homeownership`	Percent of the population that lives in their own home or lives with the owner (e.g. children living with parents who own the home)
`multiunit`	Percent of living units that are in multi-unit structures (e.g. apartments)
`income`	Income per capita
`med_income`	Median household income for the county, where a household's income equals the total income of its occupants who are 15 years or older
`smoking_ban`	Type of county-wide smoking ban in place at the end of 2011, which takes one of three values: `none`, `partial`, or `comprehensive`, where a `comprehensive` ban means smoking was not permitted in restaurants, bars, or workplaces, and `partial` means smoking was banned in at least one of those three locations

Table 1.6: Variables and their descriptions for the `county` data set.

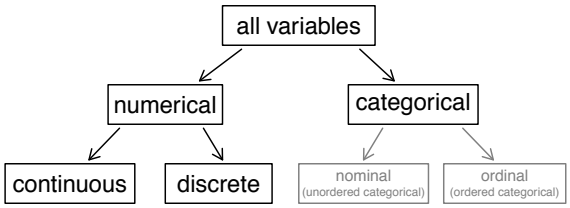


Figure 1.7: Breakdown of variables into their respective types.

1.2.2 Types of variables

Examine the `fed_spend`, `pop2010`, `state`, and `smoking_ban` variables in the `county` data set. Each of these variables is inherently different from the other three yet many of them share certain characteristics.

First consider `fed_spend`, which is said to be a **numerical** variable since it can take a wide range of numerical values, and it is sensible to add, subtract, or take averages with those values. On the other hand, we would not classify a variable reporting telephone area codes as numerical since their average, sum, and difference have no clear meaning.

The `pop2010` variable is also numerical, although it seems to be a little different than `fed_spend`. This variable of the population count can only take whole non-negative numbers (`0`, `1`, `2`, ...). For this reason, the population variable is said to be **discrete** since it can only take numerical values with jumps. On the other hand, the federal spending variable is said to be **continuous**.

The variable `state` can take up to 51 values after accounting for Washington, DC: `AL`, ..., and `WY`. Because the responses themselves are categories, `state` is called a **categorical** variable, and the possible values are called the variable's **levels**.

Finally, consider the `smoking_ban` variable, which describes the type of county-wide smoking ban and takes values `none`, `partial`, or `comprehensive` in each county. This variable seems to be a hybrid: it is a categorical variable but the levels have a natural ordering. A variable with these properties is called an **ordinal** variable, while a regular categorical variable without this type of special ordering is called a **nominal** variable. To simplify analyses, any ordinal variables in this book will be treated as categorical variables.

Example 1.3 Data were collected about students in a statistics course. Three variables were recorded for each student: number of siblings, student height, and whether the student had previously taken a statistics course. Classify each of the variables as continuous numerical, discrete numerical, or categorical.

The number of siblings and student height represent numerical variables. Because the number of siblings is a count, it is discrete. Height varies continuously, so it is a continuous numerical variable. The last variable classifies students into two categories – those who have and those who have not taken a statistics course – which makes this variable categorical.

Guided Practice 1.4 Consider the variables `group` and `outcome` (at 30 days) from the stent study in Section 1.1. Are these numerical or categorical variables?[7]

[7] There are only two possible values for each variable, and in both cases they describe categories. Thus, each is a categorical variable.

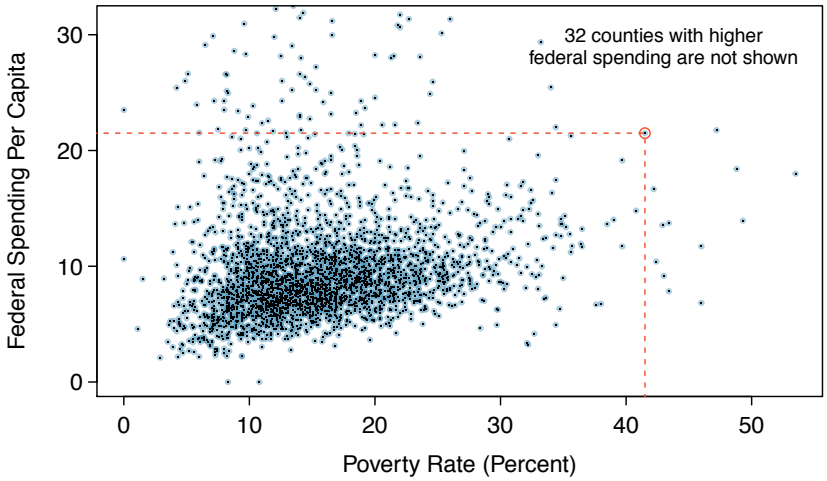


Figure 1.8: A scatterplot showing `fed_spend` against `poverty`. Owsley County of Kentucky, with a poverty rate of 41.5% and federal spending of $21.50 per capita, is highlighted.

1.2.3 Relationships between variables

Many analyses are motivated by a researcher looking for a relationship between two or more variables. A social scientist may like to answer some of the following questions:

(1) Is federal spending, on average, higher or lower in counties with high rates of poverty?

(2) If homeownership is lower than the national average in one county, will the percent of multi-unit structures in that county likely be above or below the national average?

(3) Which counties have a higher average income: those that enact one or more smoking bans or those that do not?

To answer these questions, data must be collected, such as the `county` data set shown in Table 1.5. Examining summary statistics could provide insights for each of the three questions about counties. Additionally, graphs can be used to visually summarize data and are useful for answering such questions as well.

Scatterplots are one type of graph used to study the relationship between two numerical variables. Figure 1.8 compares the variables `fed_spend` and `poverty`. Each point on the plot represents a single county. For instance, the highlighted dot corresponds to County 1088 in the `county` data set: Owsley County, Kentucky, which had a poverty rate of 41.5% and federal spending of $21.50 per capita. The scatterplot suggests a relationship between the two variables: counties with a high poverty rate also tend to have slightly more federal spending. We might brainstorm as to why this relationship exists and investigate each idea to determine which is the most reasonable explanation.

⊙ **Guided Practice 1.5** Examine the variables in the `email50` data set, which are described in Table 1.4 on page 10. Create two questions about the relationships between these variables that are of interest to you.[8]

[8] Two sample questions: (1) Intuition suggests that if there are many line breaks in an email then there also would tend to be many characters: does this hold true? (2) Is there a connection between whether an email format is plain text (versus HTML) and whether it is a spam message?

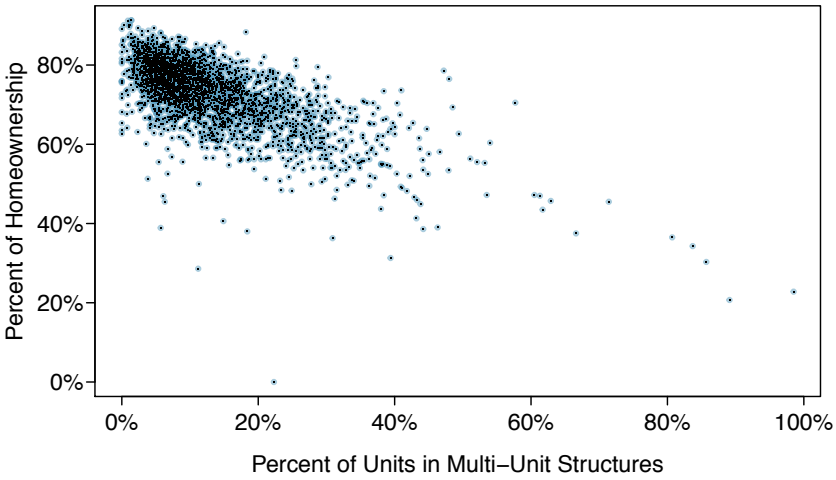


Figure 1.9: A scatterplot of homeownership versus the percent of units that are in multi-unit structures for all 3,143 counties. Interested readers may find an image of this plot with an additional third variable, county population, presented at www.openintro.org/stat/down/MHP.png.

The `fed_spend` and `poverty` variables are said to be associated because the plot shows a discernible pattern. When two variables show some connection with one another, they are called **associated** variables. Associated variables can also be called **dependent** variables and vice-versa.

Example 1.6 This example examines the relationship between homeownership and the percent of units in multi-unit structures (e.g. apartments, condos), which is visualized using a scatterplot in Figure 1.9. Are these variables associated?

It appears that the larger the fraction of units in multi-unit structures, the lower the homeownership rate. Since there is some relationship between the variables, they are associated.

Because there is a downward trend in Figure 1.9 – counties with more units in multi-unit structures are associated with lower homeownership – these variables are said to be **negatively associated**. A **positive association** is shown in the relationship between the `poverty` and `fed_spend` variables represented in Figure 1.8, where counties with higher poverty rates tend to receive more federal spending per capita.

If two variables are not associated, then they are said to be **independent**. That is, two variables are independent if there is no evident relationship between the two.

Associated or independent, not both

A pair of variables are either related in some way (associated) or not (independent). No pair of variables is both associated and independent.

1.3 Overview of data collection principles

The first step in conducting research is to identify topics or questions that are to be investigated. A clearly laid out research question is helpful in identifying what subjects or cases should be studied and what variables are important. It is also important to consider *how* data are collected so that they are reliable and help achieve the research goals.

1.3.1 Populations and samples

Consider the following three research questions:

1. What is the average mercury content in swordfish in the Atlantic Ocean?
2. Over the last 5 years, what is the average time to complete a degree for Duke undergraduate students?
3. Does a new drug reduce the number of deaths in patients with severe heart disease?

Each research question refers to a target **population**. In the first question, the target population is all swordfish in the Atlantic ocean, and each fish represents a case. Often times, it is too expensive to collect data for every case in a population. Instead, a sample is taken. A **sample** represents a subset of the cases and is often a small fraction of the population. For instance, 60 swordfish (or some other number) in the population might be selected, and this sample data may be used to provide an estimate of the population average and answer the research question.

⊙ **Guided Practice 1.7** For the second and third questions above, identify the target population and what represents an individual case.[9]

1.3.2 Anecdotal evidence

Consider the following possible responses to the three research questions:

1. A man on the news got mercury poisoning from eating swordfish, so the average mercury concentration in swordfish must be dangerously high.

2. I met two students who took more than 7 years to graduate from Duke, so it must take longer to graduate at Duke than at many other colleges.

3. My friend's dad had a heart attack and died after they gave him a new heart disease drug, so the drug must not work.

Each conclusion is based on data. However, there are two problems. First, the data only represent one or two cases. Second, and more importantly, it is unclear whether these cases are actually representative of the population. Data collected in this haphazard fashion are called **anecdotal evidence**.

[9] (2) Notice that the first question is only relevant to students who complete their degree; the average cannot be computed using a student who never finished her degree. Thus, only Duke undergraduate students who have graduated in the last five years represent cases in the population under consideration. Each such student would represent an individual case. (3) A person with severe heart disease represents a case. The population includes all people with severe heart disease.

Figure 1.10: In February 2010, some media pundits cited one large snow storm as valid evidence against global warming. As comedian Jon Stewart pointed out, "It's one storm, in one region, of one country."

Anecdotal evidence

Be careful of data collected in a haphazard fashion. Such evidence may be true and verifiable, but it may only represent extraordinary cases.

Anecdotal evidence typically is composed of unusual cases that we recall based on their striking characteristics. For instance, we are more likely to remember the two people we met who took 7 years to graduate than the six others who graduated in four years. Instead of looking at the most unusual cases, we should examine a sample of many cases that represent the population.

1.3.3 Sampling from a population

We might try to estimate the time to graduation for Duke undergraduates in the last 5 years by collecting a sample of students. All graduates in the last 5 years represent the *population*, and graduates who are selected for review are collectively called the *sample*. In general, we always seek to *randomly* select a sample from a population. The most basic type of random selection is equivalent to how raffles are conducted. For example, in selecting graduates, we could write each graduate's name on a raffle ticket and draw 100 tickets. The selected names would represent a random sample of 100 graduates.

Why pick a sample randomly? Why not just pick a sample by hand? Consider the following scenario.

● **Example 1.8** Suppose we ask a student who happens to be majoring in nutrition to select several graduates for the study. What kind of students do you think she might collect? Do you think her sample would be representative of all graduates?

Perhaps she would pick a disproportionate number of graduates from health-related fields. Or perhaps her selection would be well-representative of the population. When selecting samples by hand, we run the risk of picking a *biased* sample, even if that bias is unintentional or difficult to discern.

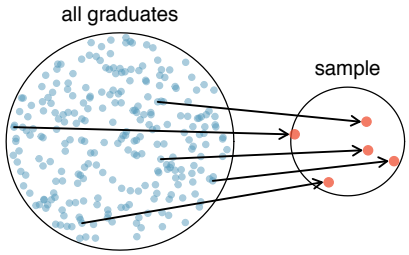


Figure 1.11: In this graphic, five graduates are randomly selected from the population to be included in the sample.

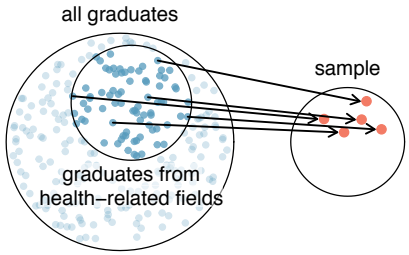


Figure 1.12: Instead of sampling from all graduates equally, a nutrition major might inadvertently pick graduates with health-related majors disproportionately often.

If someone was permitted to pick and choose exactly which graduates were included in the sample, it is entirely possible that the sample could be skewed to that person's interests, which may be entirely unintentional. This introduces **bias** into a sample. Sampling randomly helps resolve this problem. The most basic random sample is called a **simple random sample**, and which is equivalent to using a raffle to select cases. This means that each case in the population has an equal chance of being included and there is no implied connection between the cases in the sample.

The act of taking a simple random sample helps minimize bias, however, bias can crop up in other ways. Even when people are picked at random, e.g. for surveys, caution must be exercised if the **non-response** is high. For instance, if only 30% of the people randomly sampled for a survey actually respond, then it is unclear whether the results are **representative** of the entire population. This **non-response bias** can skew results.

Another common downfall is a **convenience sample**, where individuals who are easily accessible are more likely to be included in the sample. For instance, if a political survey is done by stopping people walking in the Bronx, this will not represent all of New York City. It is often difficult to discern what sub-population a convenience sample represents.

⊙ **Guided Practice 1.9** We can easily access ratings for products, sellers, and companies through websites. These ratings are based only on those people who go out of their way to provide a rating. If 50% of online reviews for a product are negative, do you think this means that 50% of buyers are dissatisfied with the product?[10]

[10] Answers will vary. From our own anecdotal experiences, we believe people tend to rant more about products that fell below expectations than rave about those that perform as expected. For this reason, we suspect there is a negative bias in product ratings on sites like Amazon. However, since our experiences may not be representative, we also keep an open mind.

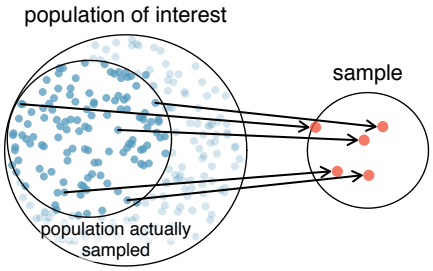


Figure 1.13: Due to the possibility of non-response, surveys studies may only reach a certain group within the population. It is difficult, and often times impossible, to completely fix this problem.

1.3.4 Explanatory and response variables

Consider the following question from page 13 for the `county` data set:

(1) Is federal spending, on average, higher or lower in counties with high rates of poverty?

If we suspect poverty might affect spending in a county, then poverty is the **explanatory** variable and federal spending is the **response** variable in the relationship.[11] If there are many variables, it may be possible to consider a number of them as explanatory variables.

TIP: Explanatory and response variables
To identify the explanatory variable in a pair of variables, identify which of the two is suspected of affecting the other and plan an appropriate analysis.

explanatory variable → might affect → response variable

Caution: association does not imply causation
Labeling variables as *explanatory* and *response* does not guarantee the relationship between the two is actually causal, even if there is an association identified between the two variables. We use these labels only to keep track of which variable we suspect affects the other.

In some cases, there is no explanatory or response variable. Consider the following question from page 13:

(2) If homeownership is lower than the national average in one county, will the percent of multi-unit structures in that county likely be above or below the national average?

It is difficult to decide which of these variables should be considered the explanatory and response variable, i.e. the direction is ambiguous, so no explanatory or response labels are suggested here.

[11] Sometimes the explanatory variable is called the **independent** variable and the response variable is called the **dependent** variable. However, this becomes confusing since a *pair* of variables might be independent or dependent, so we avoid this language.

1.3.5 Introducing observational studies and experiments

There are two primary types of data collection: observational studies and experiments.

Researchers perform an **observational study** when they collect data in a way that does not directly interfere with how the data arise. For instance, researchers may collect information via surveys, review medical or company records, or follow a **cohort** of many similar individuals to study why certain diseases might develop. In each of these situations, researchers merely observe the data that arise. In general, observational studies can provide evidence of a naturally occurring association between variables, but they cannot by themselves show a causal connection.

When researchers want to investigate the possibility of a causal connection, they conduct an **experiment**. Usually there will be both an explanatory and a response variable. For instance, we may suspect administering a drug will reduce mortality in heart attack patients over the following year. To check if there really is a causal connection between the explanatory variable and the response, researchers will collect a sample of individuals and split them into groups. The individuals in each group are *assigned* a treatment. When individuals are randomly assigned to a group, the experiment is called a **randomized experiment**. For example, each heart attack patient in the drug trial could be randomly assigned, perhaps by flipping a coin, into one of two groups: the first group receives a **placebo** (fake treatment) and the second group receives the drug. See the case study in Section 1.1 for another example of an experiment, though that study did not employ a placebo.

TIP: association $\neq$ causation
In general, association does not imply causation, and causation can only be inferred from a randomized experiment.

1.4 Observational studies and sampling strategies

1.4.1 Observational studies

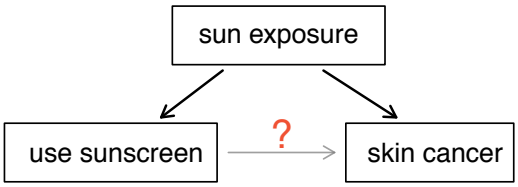
Generally, data in observational studies are collected only by monitoring what occurs, while experiments require the primary explanatory variable in a study be assigned for each subject by the researchers.

Making causal conclusions based on experiments is often reasonable. However, making the same causal conclusions based on observational data can be treacherous and is not recommended. Thus, observational studies are generally only sufficient to show associations.

⊙ **Guided Practice 1.10** Suppose an observational study tracked sunscreen use and skin cancer, and it was found that the more sunscreen someone used, the more likely the person was to have skin cancer. Does this mean sunscreen *causes* skin cancer?[12]

Some previous research tells us that using sunscreen actually reduces skin cancer risk, so maybe there is another variable that can explain this hypothetical association between sunscreen usage and skin cancer. One important piece of information that is absent is sun exposure. If someone is out in the sun all day, she is more likely to use sunscreen *and* more likely to get skin cancer. Exposure to the sun is unaccounted for in the simple investigation.

[12] No. See the paragraph following the exercise for an explanation.



Sun exposure is what is called a **confounding variable**,[13] which is a variable that is correlated with both the explanatory and response variables. While one method to justify making causal conclusions from observational studies is to exhaust the search for confounding variables, there is no guarantee that all confounding variables can be examined or measured.

In the same way, the `county` data set is an observational study with confounding variables, and its data cannot easily be used to make causal conclusions.

⊙ **Guided Practice 1.11** Figure 1.9 shows a negative association between the homeownership rate and the percentage of multi-unit structures in a county. However, it is unreasonable to conclude that there is a causal relationship between the two variables. Suggest one or more other variables that might explain the relationship visible in Figure 1.9.[14]

Observational studies come in two forms: prospective and retrospective studies. A **prospective study** identifies individuals and collects information as events unfold. For instance, medical researchers may identify and follow a group of similar individuals over many years to assess the possible influences of behavior on cancer risk. One example of such a study is The Nurses' Health Study, started in 1976 and expanded in 1989.[15] This prospective study recruits registered nurses and then collects data from them using questionnaires. **Retrospective studies** collect data after events have taken place, e.g. researchers may review past events in medical records. Some data sets, such as `county`, may contain both prospectively- and retrospectively-collected variables. Local governments prospectively collect some variables as events unfolded (e.g. retails sales) while the federal government retrospectively collected others during the 2010 census (e.g. county population counts).

1.4.2 Four sampling methods (special topic)

Almost all statistical methods are based on the notion of implied randomness. If observational data are not collected in a random framework from a population, these statistical methods – the estimates and errors associated with the estimates – are not reliable. Here we consider four random sampling techniques: simple, stratified, cluster, and multistage sampling. Figures 1.14 and 1.15 provide graphical representations of these techniques.

Simple random sampling is probably the most intuitive form of random sampling. Consider the salaries of Major League Baseball (MLB) players, where each player is a member of one of the league's 30 teams. To take a simple random sample of 120 baseball players and their salaries from the 2010 season, we could write the names of that season's

[13] Also called a **lurking variable**, **confounding factor**, or a **confounder**.

[14] Answers will vary. Population density may be important. If a county is very dense, then this may require a larger fraction of residents to live in multi-unit structures. Additionally, the high density may contribute to increases in property value, making homeownership infeasible for many residents.

[15] www.channing.harvard.edu/nhs

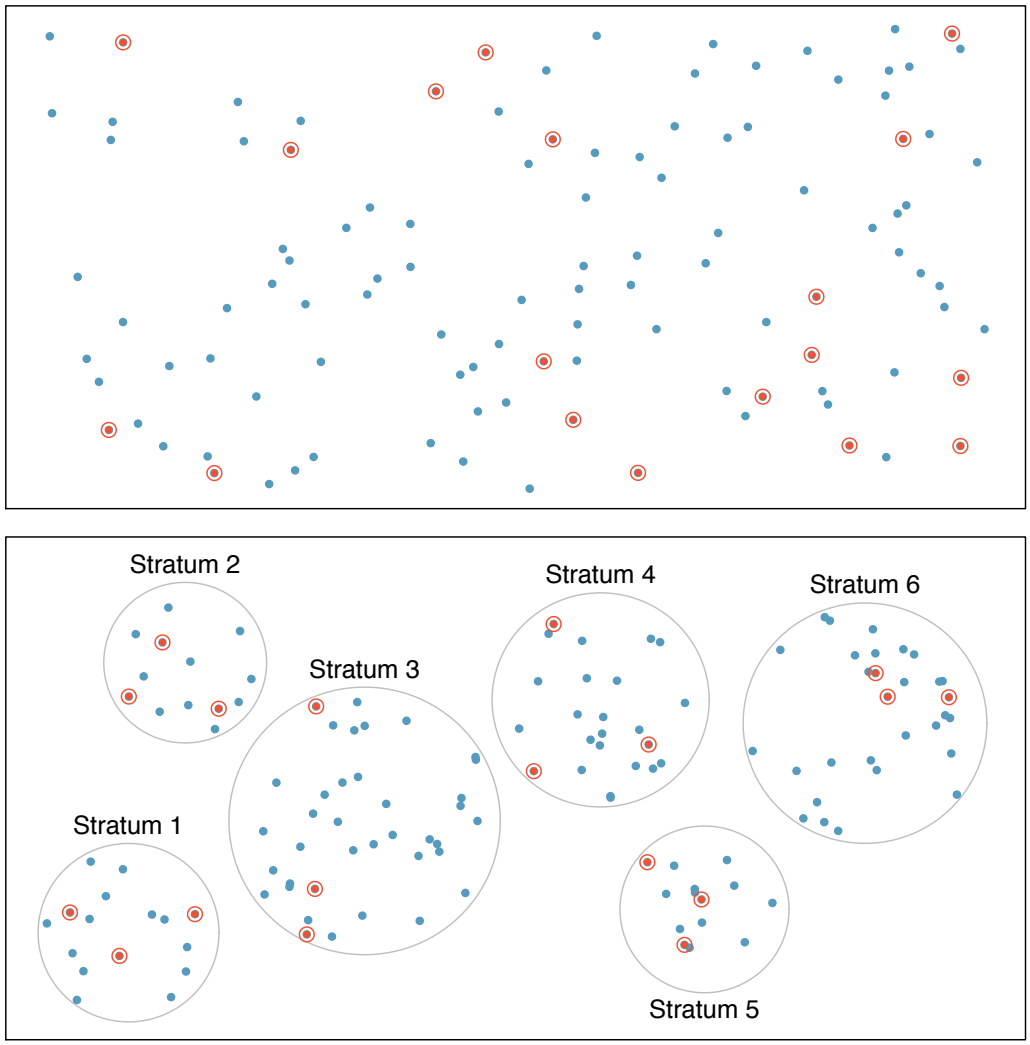


Figure 1.14: Examples of simple random and stratified sampling. In the top panel, simple random sampling was used to randomly select the 18 cases. In the bottom panel, stratified sampling was used: cases were grouped into strata, then simple random sampling was employed within each stratum.

828 players onto slips of paper, drop the slips into a bucket, shake the bucket around until we are sure the names are all mixed up, then draw out slips until we have the sample of 120 players. In general, a sample is referred to as "simple random" if each case in the population has an equal chance of being included in the final sample *and* knowing that a case is included in a sample does not provide useful information about which other cases are included.

Stratified sampling is a divide-and-conquer sampling strategy. The population is divided into groups called **strata**. The strata are chosen so that similar cases are grouped together, then a second sampling method, usually simple random sampling, is employed within each stratum. In the baseball salary example, the teams could represent the strata, since some teams have a lot more money (up to 4 times as much!). Then we might randomly sample 4 players from each team for a total of 120 players.

Stratified sampling is especially useful when the cases in each stratum are very similar with respect to the outcome of interest. The downside is that analyzing data from a stratified sample is a more complex task than analyzing data from a simple random sample. The analysis methods introduced in this book would need to be extended to analyze data collected using stratified sampling.

Example 1.12 Why would it be good for cases within each stratum to be very similar?

We might get a more stable estimate for the subpopulation in a stratum if the cases are very similar. These improved estimates for each subpopulation will help us build a reliable estimate for the full population.

In a **cluster sample**, we break up the population into many groups, called **clusters**. Then we sample a fixed number of clusters and include all observations from each of those clusters in the sample. A **multistage sample** is like a cluster sample, but rather than keeping all observations in each cluster, we collect a random sample within each selected cluster.

Sometimes cluster or multistage sampling can be more economical than the alternative sampling techniques. Also, unlike stratified sampling, these approaches are most helpful when there is a lot of case-to-case variability within a cluster but the clusters themselves don't look very different from one another. For example, if neighborhoods represented clusters, then cluster or multistage sampling work best when the neighborhoods are very diverse. A downside of these methods is that more advanced analysis techniques are typically required, though the methods in this book can be extended to handle such data.

Example 1.13 Suppose we are interested in estimating the malaria rate in a densely tropical portion of rural Indonesia. We learn that there are 30 villages in that part of the Indonesian jungle, each more or less similar to the next. Our goal is to test 150 individuals for malaria. What sampling method should be employed?

A simple random sample would likely draw individuals from all 30 villages, which could make data collection extremely expensive. Stratified sampling would be a challenge since it is unclear how we would build strata of similar individuals. However, cluster sampling or multistage sampling seem like very good ideas. If we decided to use multistage sampling, we might randomly select half of the villages, then randomly select 10 people from each. This would probably reduce our data collection costs substantially in comparison to a simple random sample, and this approach would still give us reliable information.

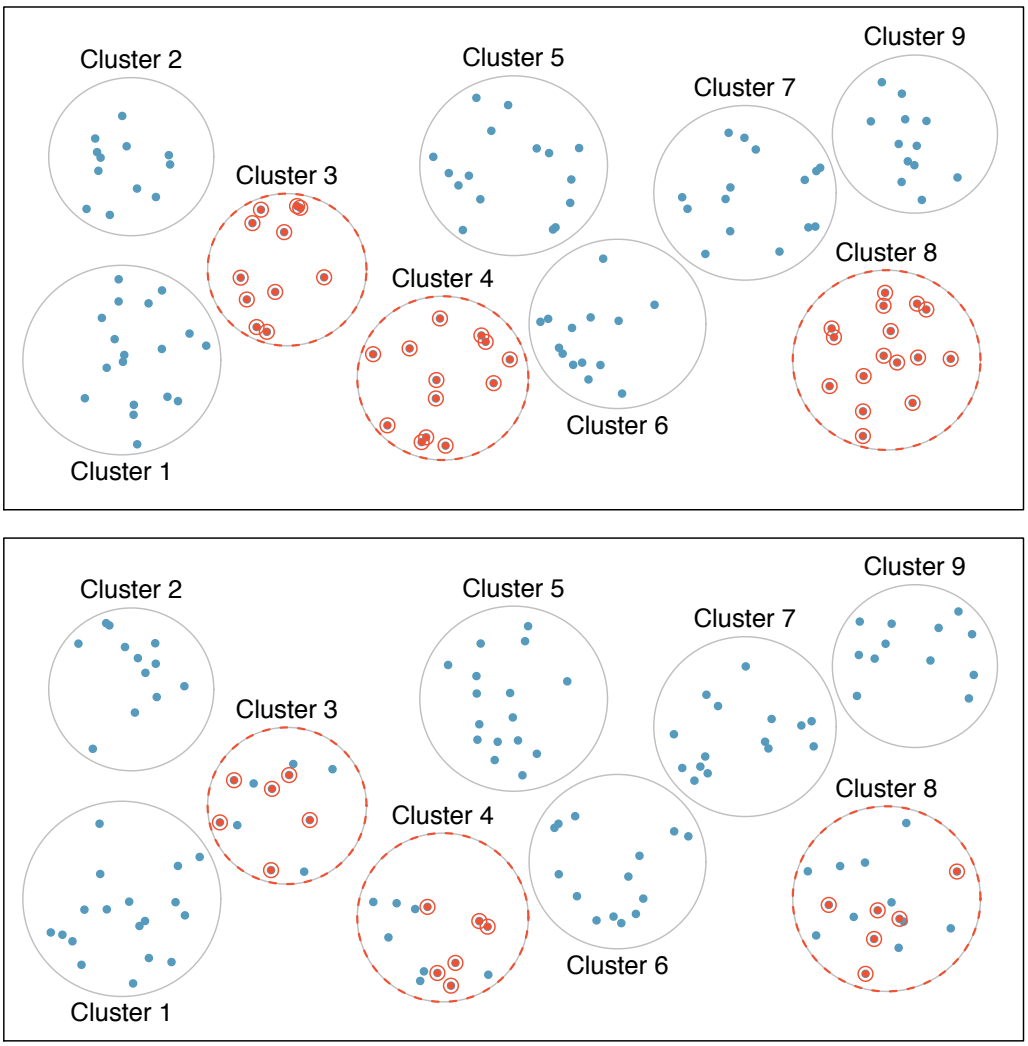


Figure 1.15: Examples of cluster and multistage sampling. In the top panel, cluster sampling was used. Here, data were binned into nine clusters, three of these clusters were sampled, and all observations within these three cluster were included in the sample. In the bottom panel, multistage sampling was used. It differs from cluster sampling in that of the clusters selected, we randomly select a subset of each cluster to be included in the sample.

Poglavlje 3

Deskriptivna statistika

3.1 Metode opisivanja kvalitativnih podataka

Kvalitativne varijable primaju vrijednosti koje su razvrstane u kategorije. Pri proučavanju takvih varijabli pažnju usmjeravamo na zastupljenost pojedine kategorije u uzorku na kojem provodimo istraživanje. Primjer 3.1 uvodi nas u problematiku opisivanja kvalitativnih varijabli.

Primjer 3.1. *Svaki čovjek prema spolu pripada jednoj od dviju kategorija (ženskom spolu (Ž) ili muškom spolu (M)), a prema tipu svoje krvne grupe jednoj od četiriju kategorija (A, B, AB ili 0). Tablica 3.1 sadrži podatke o spolu i tipu krvne grupe za deset ispitanika iz nekog medicinskog istraživanja.*

ispitanik	spol	krvna grupa
1	Ž	A
2	Ž	B
3	M	0
4	Ž	0
5	M	AB
6	M	B
7	Ž	B
8	M	A
9	Ž	AB
10	Ž	A

Tablica 3.1: Tablični prikaz podataka o spolu i krvnoj grupi.

Iz tablice 3.1 vidimo da za svakog ispitanika iz promatranog uzorka vrijednost varijable spol *pripada kategoriji M ili kategoriji Ž, a vrijednost varijable* krvna grupa *jednoj od kategorija A, B, AB ili*

0. *Prema tome, varijable* spol *i* krvna grupa *jesu* kvalitativne varijable. *Informacije koje je moguće dobiti iz prethodne tablice vezane su uz zastupljenost pojedine kategorije u promatranom uzorku. Tako je npr. moguće dobiti odgovore na sljedeća i slična pitanja:*

Koliko ispitanika ženskog spola ima u promatranom uzorku?

Koliki je udio ispitanika s krvnom grupom 0 *u promatranom uzorku?*

Koliko ispitanika ženskog spola iz promatranog uzorka ima krvnu grupu A?

Koliki udio ispitanika muškog spola iz promatranog uzorka ima krvnu grupu B ili AB?

Kako izmjeriti zastupljenost pojedine kategorije u uzorku?
Osnovna mjera kojom opisujemo zastupljenost jedne kategorije u uzorku jest **frekvencija** kategorije.
Neka varijabla, koju ćemo označiti s X, ima k kategorija (recimo $k = 4$ znači da varijabla ima 4 kategorije - npr. krvne grupe). **Označimo pojedine kategorije s $x_1, x_2, \ldots, x_k$, odnosno u drugom zapisu $\{x_i : i = 1, \ldots, k\}$. Frekvencija kategorije x_i je broj izmjerenih vrijednosti varijable koje pripadaju kategoriji x_i, $i = 1, \ldots, k$. Frekvenciju kategorije x_i označavamo s**

$$f_i.$$

Frekvencija pojedine kategorije ovisi o broju izvršenih mjeranja, tj. veličini uzorka. Da bismo lakše usporedili i tumačili rezultate raznih istraživanja, u opisu zastupljenosti jedne kategorije u uzorku često koristimo i **relativnu frekvenciju** kategorije. **Relativna frekvencija kategorije x_i je broj izmjerenih vrijednosti varijable koje pripadaju kategoriji x_i podijeljen ukupnim brojem izmjerenih vrijednosti za ispitivanu varijablu, $i = 1, \ldots, k$. Ako je n veličina uzorka, tj. broj svih izmjerenih vrijednosti ispitivane varijable, relativnu frekvenciju kategorije x_i računamo kao**

$$\frac{f_i}{n}.$$

Relativna frekvencija kategorije je mjera zastupljenosti koja daje informaciju o udjelu kategorije u uzorku poznate veličine i često se izražava kao postotak. **Frekvencije i relativne frekvencije pojedinih kategorija prikazujemo tablično i grafički.**

3.1.1 Tablični prikaz frekvencija i relativnih frekvencija

U tabličnom prikazu frekvencija i relativnih frekvencija trebaju biti zastupljene sve kategorije promatrane varijable.

Primjer 3.2. *Frekvencije i relativne frekvencije svih kategorija varijabli* spol *i* krvna grupa *iz primjera 3.1 prikazane su u tablicama 3.2 i 3.3.*

spol	frekvencija	relativna frekvencija
Ž	6	$6/10 = 0.6 = 60\%$
M	4	$4/10 = 0.4 = 40\%$

Tablica 3.2: Tablica frekvencija i relativnih frekvencija svih kategorija varijable spol.

krvna grupa	frekvencija	relativna frekvencija
A	3	$3/10 = 0.3 = 30\%$
B	3	$3/10 = 0.3 = 30\%$
AB	2	$2/10 = 0.2 = 20\%$
0	2	$2/10 = 0.2 = 20\%$

Tablica 3.3: Tablica frekvencija i relativnih frekvencija svih kategorija varijable krvna grupa.

Primjer 3.3. *Od velike su važnosti u mnogim istraživanjima i kategorizirane tablice frekvencija i relativnih frekvencija. Frekvencije i relativne frekvencije za izmjerene vrijednosti varijable* krvna grupa *iz primjera 3.1 kategorizirane prema spolu ispitanika dane su u tablicama 3.4 (za ženski spol) i 3.5 (za muški spol).*

spol = Ž		
krvna grupa	frekvencija	relativna frekvencija
A	2	2/6
B	2	2/6
AB	1	1/6
0	1	1/6

Tablica 3.4: Frekvencije i relativne frekvencije krvnih grupa za ženski spol.

spol = M		
krvna grupa	frekvencija	relativna frekvencija
A	1	$1/4 = 0.25 = 25\%$
B	1	$1/4 = 0.25 = 25\%$
AB	1	$1/4 = 0.25 = 25\%$
0	1	$1/4 = 0.25 = 25\%$

Tablica 3.5: Frekvencije i relativne frekvencije krvnih grupa za muški spol.

Na temelju prethodnih dviju tablica i tablica iz primjera 3.2 možemo redom odgovoriti na pitanja postavljena u primjeru 3.1:

U uzorku ima šest ispitanika ženskog spola (tj. frekvencija žena u uzorku je šest).

U uzorku ima 20% *ispitanika s krvnom grupom* 0 *(tj. relativna frekvenicja krvne grupe nula u uzorku je* 20%*).*

U uzorku ima dvije žene s krvnom grupom A (tj. frekvencija žena s krvnom grupom A u uzorku je dva).

Od svih ispitanika muškog spola njih 50% *ima krvnu grupu B ili AB.*

Primjer 3.4. (krvne-grupe.sta)
U ovom primjeru naučit ćemo kako bazu podataka te tablice frekvencija i relativnih frekvencija napraviti u programskom paketu Statistica. Rezultat postupka u tom programskom paketu prikazan je za varijable krvna grupa *i* spol *iz primjera 3.1, tj. iz baze podataka* krvne-grupe.sta. *Tablične prikaze frekvencija i relativnih frekvencija u programskom paketu Statistica možemo dobiti provodeći sljedeći postupak (koji provodimo slijedeći navedeni niz opcija u izborniku):*

Statistics → Basic Statistics/Tables → Freq. Tables → Variables → Summary.

Rezultat provedbe prethodnog postupka jesu tablice prikazane na slici 3.1.

	Frequency table: krvna_grupa (krvne-grupe.sta)			
Category	Count	Cumulative Count	Percent	Cumulative Percent
0	2	2	20,00	20,00
A	3	5	30,00	50,00
B	3	8	30,00	80,00
AB	2	10	20,00	100,00
Missing	0	10	0,00	100,00

(a) krvna grupa

	Frequency table: spol (krvne-grupe.sta)			
Category	Count	Cumulative Count	Percent	Cumulative Percent
Ž	6	6	60,00	60,00
M	4	10	40,00	100,00
Missing	0	10	0,00	100,00

(b) spol

Slika 3.1: Frekvencije i relativne frekvencija svih kategorija varijabli krvna grupa i spol.

Promatranje vrijednosti varijable spol *kategorizirane prema krvnoj grupi ispitanika omogućuju kategorizirane tablice frekvencija i relativnih frekvencija. Za izradu takvih tablica podatke iz varijabli od interesa moramo profiltrirati, tj. moramo zadati uvjet prema kojemu će u daljnju analizu biti uključena samo uvjetom određena kategorija podataka. Kategorizirane tablice frekvencija i relativnih frekvencija u programskom paketu Statistica možemo dobiti provodeći sljedeći postupak:*

Selection → označiti Enable Selection Conditions → pod Include Cases odabrati opciju "Specific, selected by expression" (u polje za unos teksta upisati krvna grupa="A" ako želimo u obzir uzeti samo ispitanike s krvnom grupom A; analogno se postavlja uvjet krvna grupa="B" za krvnu grupu B, krvna grupa="AB" za krvnu grupu AB, krvna grupa="0" za krvnu grupu 0) → OK.

Rezultat provedbe prethodnog postupka jesu tablice prikazane na slici 3.2.

Category	Frequency table: spol (krvne-grupe.sta) Include condition: krvna_grupa="A"			
	Count	Cumulative Count	Percent	Cumulative Percent
Ž	2	2	66,67	66,67
M	1	3	33,33	100,00
Missing	0	3	0,00	100,00

(a) kategorija: krvna grupa A

Category	Frequency table: spol (krvne-grupe.sta) Include condition: krvna_grupa="B"			
	Count	Cumulative Count	Percent	Cumulative Percent
Ž	2	2	66,67	66,67
M	1	3	33,33	100,00
Missing	0	3	0,00	100,00

(b) kategorija: krvna grupa B

Category	Frequency table: spol (krvne-grupe.sta) Include condition: krvna_grupa="AB"			
	Count	Cumulative Count	Percent	Cumulative Percent
Ž	1	1	50,00	50,00
M	1	2	50,00	100,00
Missing	0	2	0,00	100,00

(c) kategorija: krvna grupa AB

Category	Frequency table: spol (krvne-grupe.sta) Include condition: krvna_grupa=0			
	Count	Cumulative Count	Percent	Cumulative Percent
Ž	1	1	50,00	50,00
M	1	2	50,00	100,00
Missing	0	2	0,00	100,00

(d) kategorija: krvna grupa 0

Slika 3.2: Frekvencije i relativne frekvencije kategorija varijable spol za krvne grupe A, B, AB i 0.

3.1.2 Grafički prikazi frekvencija i relativnih frekvencija

Frekvencije i relativne frekvencije kategorija kvalitativnih varijabli grafički prikazujemo korištenjem **stupčastog dijagrama** (eng. Bar Chart ili Bar Plot) frekvencija i stupčastog dijagrama relativnih frekvencija. U istu svrhu može se koristiti i **kružni dijagram** (eng. Pie Chart) frekvencija i relativnih frekvencija . Popularni naziv za isti grafički prikaz je "pita").

Primjer 3.5. (hormon.sta)
Grafičke prikaze frekvencija i relativnih frekvencija kvalitativnih varijabli prikazat ćemo na primjeru varijable dijagnoza *iz baze podataka* hormon.sta *(koja je opisana u zadatku 3.1). Stupčasti dijagram frekvencija u programskom paketu Statistica možemo dobiti provodeći sljedeći postupak:*

Statistics → Basic Statistics/Tables → Frequency Tables → Choose variables → Histograms.

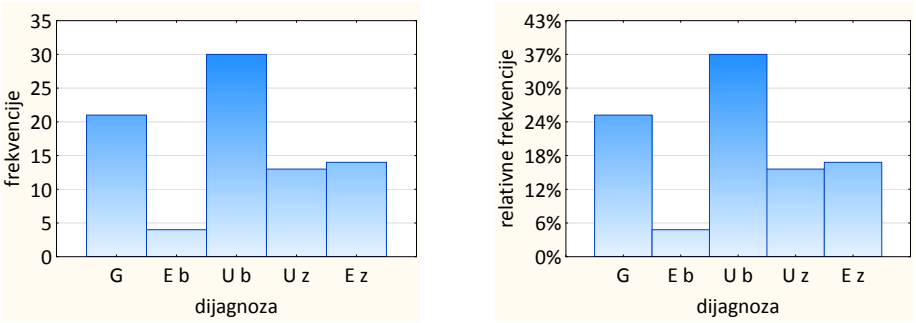
Stupčasti dijagram koji prikazuje i frekvencije i relativne frekvencije u programskom paketu Statistica možemo dobiti provodeći sljedeći postupak:

Graphs → Histograms → Choose variables → Advanced → Pod "Y axis" uključiti "% and N" → OK.

Stupčasti dijagrami frekvencija i relativnih frekvencija svih kategorija varijable dijagnoza *prikazani su na slici 3.3. Drugi način grafičkog prikazivanja mjera zastupljenosti pojedinih kategorija neke kvalitativne varijable u uzorku jesu kružni dijagrami frekvencija i relativnih frekvencija koje u programskom paketu Statistica možemo dobiti provodeći sljedeći postupak:*

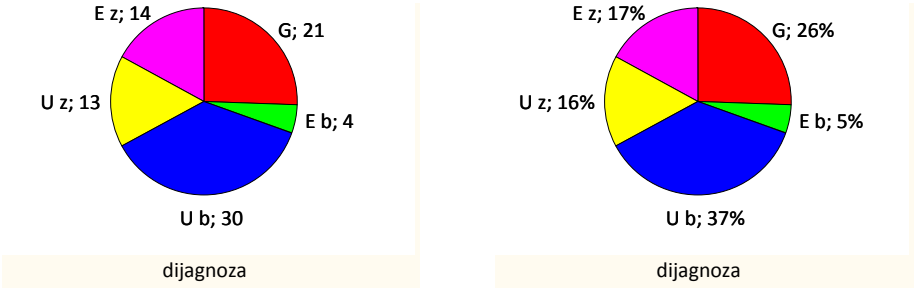
Graphs → 2D Graphs → Graph type (opcija "Pie Chart - Counts") → Choose variables → Advanced → Pie Legend - odabrati opciju "Text and Value" za kružni dijagram frekvencija, a opciju "Text and Percent" za kružni dijagram relativnih frekvencija → OK.

Kružni dijagrami frekvencija i relativnih frekvencija kategorija varijable dijagnoza *prikazani su na slici 3.4.*



(a) frekvencije (b) relativne frekvencije

Slika 3.3: Stupčasti dijagrami frekvencija i relativnih frekvencija svih kategorija varijable dijagnoza.



(a) frekvencije (b) relativne frekvencije

Slika 3.4: Kružni dijagrami frekvencija i relativnih frekvencija svih kategorija varijable dijagnoza.

Primjer 3.6. (djelatnici.sta)
Često se u praksi pokazuje korisnim poznavanje zastupljenosti kategorija jedne varijable za svaku od kategorija neke druge kvalitativne varijable proučavane na istom uzorku. U ovom ćemo primjeru tablično i grafički prikazati frekvencije i relativne frekvencije svih kategorija varijable obrazovanje *iz baze podataka* djelatnici.sta *opisane u primjeru 2.4 posebno za ispitanike ženskog spola, a posebno za ispitanike muškog spola. Tablice tako kategoriziranih frekvencija i relativnih frekvencija varijable* obrazovanje *prikazane su u tablici 3.5.*

Category	Frequency table: obrazovanje (djelatnici.sta) Include condition: spol="Z"			
	Count	Cumulative Count	Percent	Cumulative Percent
SSS	21	21	51.22	51.22
VŠSS	18	39	43.90	95.12
VSS	2	41	4.88	100.00
Missing	0	41	0.00	100.00

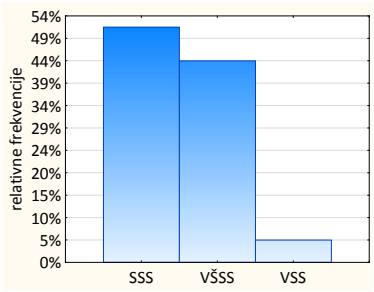
(a) spol = Z

Category	Frequency table: obrazovanje (djelatnici.sta) Include condition: spol="M"			
	Count	Cumulative Count	Percent	Cumulative Percent
SSS	30	30	50.85	50.85
VŠSS	25	55	42.37	93.22
VSS	4	59	6.78	100.00
Missing	0	59	0.00	100.00

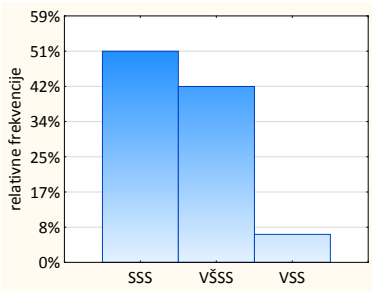
(b) spol = M

Slika 3.5: Tablica frekvencija i relativnih frekvencija svih kategorija varijable obrazovanje posebno za svaku kategoriju varijable spol.

Stupčasti dijagrami frekvencija i relativnih frekvencija svih kategorija varijable obrazovanje *za kategorije* Z *i* M *varijable* spol *prikazani su na slici 3.6, a kružni dijagramovi na slici 3.7.*

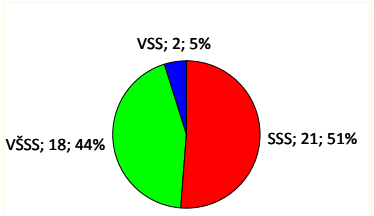


(a) spol=Z

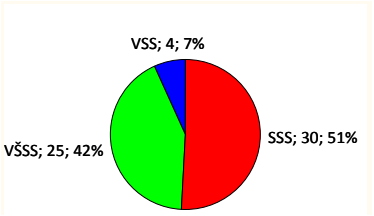


(b) spol=M

Slika 3.6: Stupčasti dijagrami relativnih frekvencija svih kategorija varijable obrazovanje posebno za svaku kategoriju varijable spol.



(a) spol=Z



(b) spol=M

Slika 3.7: Kružni dijagram frekvencija i relativnih frekvencija svih kategorija varijable obrazovanje posebno za svaku kategoriju varijable spol.

3.2 Metode opisivanja numeričkih podataka

Numerički podaci mogu biti prikupljeni promatranjem (mjerenjem) numeričke ili ordinalne varijable. Ordinalne varijable najčešće se zadaju tako da mogu primiti samo nekoliko međusobno različitih vrijednosti, dok kod numeričkih varijabli to vrlo često nije slučaj. Numeričke varijable, po svojoj prirodi, mogu biti diskretne ili neprekidne, kao što je opisano u poglavlju 2.3.2. U oba slučja, a posebno kod neprekidnih varijabli, može se dogoditi da u prikupljenim podacima postoji mnogo međusobno različitih vrijednosti. U takvim slučajevima tablični i grafički prikazi uvedeni za kvalitativne varijable mogu biti nedovoljno informativni. Ilustracija tog problema dana je sljedećim primjerom.

Primjer 3.7. (cijena.sta, hormon.sta, komarci.sta, matematika.sta)

Baza podataka cijena.sta *sadrži informacije o prodajnim mjestima (varijabla* trgovina*) i cijenama nekog proizvoda na tim prodajnim mjestima (varijabla* cijena*). Evidentirane vrijednosti obje varijable jesu brojevi, ali varijabla* trgovina *je, po svojoj prirodi, kvalitativna, a varijabla* cijena *neprekidna. Uočite da su svi prikupljeni podaci za varijablu* cijena *međusobno različiti.*

U bazi podataka komarci.sta *(opisanoj u zadatku 3.1) varijable* brojM *i* brojZ *su diskretne numeričke varijable, a varijable* temperatura *i* rel-vlaznost *neprekidne numeričke varijable. Uočite da se u podacima za sve te varijable pojavljuje mnogo međusobno različitih vrijednosti.*

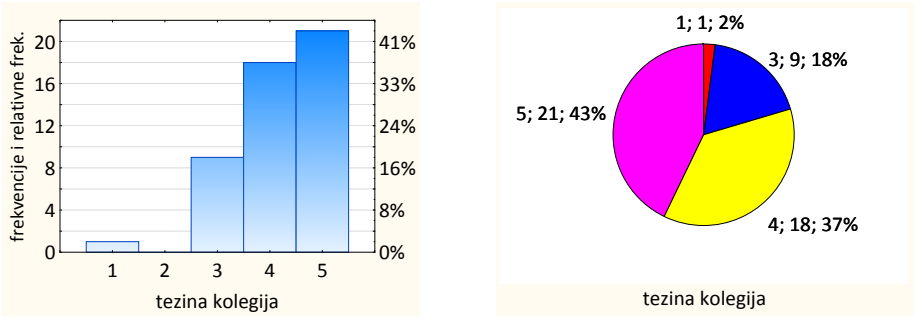
Ako su numeričke varijable diskretne s malo mogućih vrijednosti ili ako su varijable ordinalne, za opis podataka možemo koristiti iste metode kao pri opisivanju kvalitativnih podataka, tj. frekvencije i relativne frekvencije te ih grafički prikazivati stupčastim dijagramima i kružnim dijagramima.

Primjer 3.8. (matematika.sta)

Tablični i grafički prikazi (stupčasti dijagram i kružni dijagram) frekvencija i relativnih frekvencija svih vrijednosti ordinalne varijable tezina-kolegija *prikazani su na slikama 3.8 i 3.9.*

	Frequency table: tezina kolegija (matematika.sta)			
Category	Count	Cumulative Count	Percent	Cumulative Percent
1	1	1	2.04	2.04
3	9	10	18.37	20.41
4	18	28	36.73	57.14
5	21	49	42.86	100.00
Missing	0	49	0.00	100.00

Slika 3.8: Tablica frekvencija i relativnih frekvencija za varijablu tezina-kolegija.



(a) stupčasti dijagram (b) kružni dijagram

Slika 3.9: Grafički prikazi frekvencija i relativnih frekvencija za varijablu tezina-kolegija.

Iz prikazanih opisa varijable tezina-kolegija *možemo dobiti npr. sljedeće informacije:*

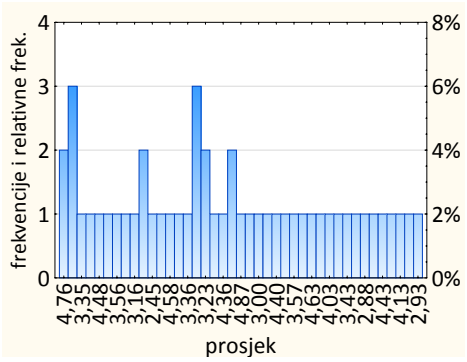
Ocjenom većom od 3 *težinu kolegija ocijenilo je čak* 39 *ispitanika, tj. čak* $39/49 \approx 79.59\%$ *od ukupnog broja ispitanika.*

Ocjenom 3 *težinu kolegija ocijenilo je* 9 *(*$9/49 \approx 18.37\%$*), a ocjenom* 4 *čak* 18 *(*$18/49 \approx 36.73\%$*) ispitanika. Dakle, dvostruko više ispitanika težinu kolegija ocijenilo je ocjenom* 4 *nego ocjenom* 3*.*

U sljedećem primjeru prikazano je šta se događa ako koristimo uobičajeni stupčasti dijagram za prikazivanje numeričkih podataka među kojima ima velik broj različitih vrijednosti.

Primjer 3.9. (matematika.sta)
Stupčasti dijagram za podatke neprekidne numeričke varijable prosjek *iz baze podataka* matematika.sta *(vidi primjer 2.9) prikazan je na slici 3.10. Pri opisivanju ove varijable pretpostavili smo da svi međusobno različiti podaci varijable* prosjek *čine zasebne kategorije. Zbog velikog broja različitih podataka broj kategorija je prevelik i rezultat analize grafičkog prikaza 3.10 ne daje željene informacije.*

Radi dobivanja korisnijih stupčastih i kružnih dijagrama za podatke iz neprekidnih numeričkih varijabli vrijednosti je potrebno **kategorizirati**, tj. razvrstati ih u odabrane kategorije. Pri tome podatke kategoriziramo u disjunktne intervale po kriteriju za koji smatramo da će nam dati željene rezultate. Za potrebe opisivanja skupa podataka obično biramo disjunktne intervale tako da dobivenim tabličnim i grafičkim prikazima možemo ilustrirati karakteristike skupa podataka koje želimo naglasiti.



Slika 3.10: Stupčasti dijagram za podatke varijable prosjek.

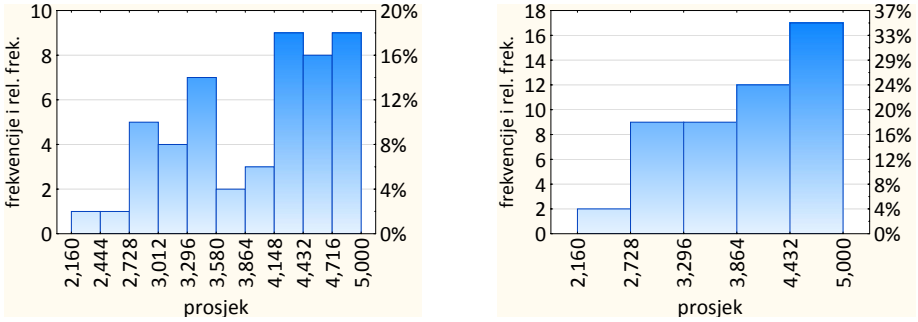
3.2.1 Postupak razvrstavanja numeričkih podataka u kategorije

Razvrstavanje vrijednosti neprekidne numeričke varijable u kategorije moguće je provesti na nekoliko načina, npr. moguće je skup svih podataka (ili nešto veći skup koji sadrži skup svih podataka, ali koji je jednostavnije podijeliti na jednake dijelove) podijeliti na disjunktne intervale jednake duljine. No, nije nužno da su intervali jednake duljine, tj. nema točno definiranog pravila po kojemu bi trebalo definirati duljine intervala niti nihov broj, ali je jasno da ih ne smije biti ni previše ni premalo da bi cijeli postupak imao smisla i služio svrsi (a to je u ovom času prikazivanje skupa podataka).

Za prikaz frekvencija ili relativnih frekvencija tako kategoriziranih podataka možemo koristiti i specifičan stupčasti dijagram koji zovemo **histogram**. Histogram mora imati stupce postavljene u koordinatni sustav nad odgovarajućim intervalima. Širina svakog stupca histograma odgovara duljini odgovarajućeg intervala, a visina frekveniciji, odnosno relativnoj frekvenciji intervala.

Primjer 3.10. (matematika.sta)
Primjerom 3.9 pokazali smo da je teško analizirati varijablu prosjek *iz baze* matematika.sta *ako za kategorije uzmemo sve različite izmjerene vrijednosti te varijable. Stoga ćemo provesti kategorizaciju izmjerenih vrijednosti.*

Dva primjera kategorizacije, tj. podjele izmjerenih vrijednosti u disjunkte intervale, rezultiraju histogramima prikazanim na slici 3.11.

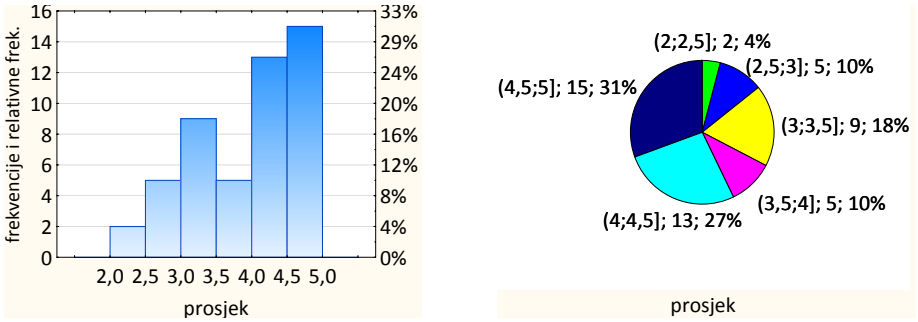


(a) kategorizacija na 10 disjunktnih intervala

(b) kategorizacija na 5 disjunktnih intervala

Slika 3.11: Stupčasti dijagrami za podatke varijable prosjek.

Kriterij kategorizacije treba biti prilagođen zahtjevima istraživanja, tj. treba omogućiti dobivanje odgovora na postavljena pitanja. Npr. ako nas zanima zastupljenost studenata s prosjekom većim od 3.5 *u promatranom uzorku, tada podatke iz varijable* prosjek *možemo kategorizirati u šest disjunktnih intervala duljine* 0.5*, počevši od* 2.0*. Iz grafičkih prikaza sa slike 3.12 očitavamo da je frekvencija takvih studenata* 33*, a relativna frekvencija* $33/49 \approx 67.35\%$.



Slika 3.12: Stupčasti i kružni dijagram za podatke varijable prosjek razvrstane u 6 disjunktnih intervala počevši od ocjene 2.0.

3.2.2 Mjere centralne tendencije i raspršenosti podataka

Karakteristika numeričkih i ordinalnih varijabli jest da među njihovim vrijednostima postoji prirodan uređaj. Na osnovi te činjenice možemo definirati numeričke karakteristike podataka iz tih varijabli koje imaju logičnu interpretaciju i mogu se iskoristiti za prikazivanje skupa podataka. U ovom poglavlju navodimo osnovne numeričke karakteristike skupa podataka te primjerima ilustriramo njihovu inter-

pretaciju u praktičnim problemima.

Aritmetička sredina

Aritmetička sredina (eng. arithmetic mean) niza podataka $x_1, x_2, \ldots, x_n$ iz varijable X definirana je izrazom

$$\bar{x}_n = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

Aritmetička sredina je numerička karakteristika koja spada u mjere centralne tendencije, tj. ona mjeri "srednju vrijednost" podataka.

Primjer 3.11. *Neka su izmjerene vrijednosti jedne varijable sljedeće:*

$$1.2, 2.1, 3.2, 4.3, 5.4, 6.5, 7.6, 8.7, 9.8.$$

S obzirom da ih ima ukupno devet, aritmetička sredina ovog skupa izmjerenih vrijednosti je

$$\frac{1.2 + 2.1 + 3.2 + 4.3 + 5.4 + 6.5 + 7.6 + 8.7 + 9.8}{9} \approx 5.42.$$

Medijan

Da bismo razumjeli i odredili medijan potrebno je prvo poredati izmjerene vrijednosti $x_1, x_2, \ldots, x_n$ varijable X po veličini (u rastućem poretku, tj. od manjeg prema većem). Medijan je također jedna mjera centralne tendencije kao i aritmetička sredina, a karakterizira ga činjenica da je barem pola podataka manje ili jednako medijanu, a istovremeno je barem pola podataka veće ili jednako od medijana. Način njegova izračuna ovisi o tome imamo li **neparan** ili **paran** broj podataka. Ako imamo **neparan broj** podataka, onda postoji vrijednost koja je na srednjoj poziciji u uređenom skupu podataka pa nju definiramo kao medijan.

Primjer 3.12. *Neka su izmjerene vrijednosti jedne varijable sljedeće:*

$$1, 2, 5, 6, 5, 1, 2, 7, 2, 2, 3.$$

Prvo ove vrijednosti poredamo po veličini:

$$1, 1, 2, 2, 2, \mathbf{2}, 3, 5, 5, 6, 7.$$

S obzirom da ih ima ukupno jedanaest, medijan je vrijednost koja je na šestoj poziciji u tako dobivenom nizu, tj. broj 2.

Ako imamo **paran broj** podataka, onda ne postoji podatak koji je na srednjoj poziciji jer srednju poziciju "zauzimaju" dva podatka. Zapravo, zahtjev na temelju kojega želimo odrediti medijan ispunjavaju svi brojevi iz intervala čije su granice dva srednja podatka. Da bismo jedinstveno odredili medijan podataka, u tom ga slučaju definiramo kao broj na polovini tog intervala, tj. kao aritmetičku sredinu tih dvaju podataka.

Primjer 3.13. *Neka su izmjerene vrijednosti jedne varijable sljedeće:*

$$1, 2, 5, 6, 5, 1, 2, 7, 2, 2, 3, 3.$$

Prvo ove vrijednosti poredamo po veličini:

$$1, 1, 2, 2, 2, \mathbf{2}, \mathbf{3}, 3, 5, 5, 6, 7.$$

S obzirom da ih ima dvanaest, "sredinu" čine šesti i sedmi podatak, tj. brojevi 2 *i* 3*. Medijan ovog skupa podataka je aritmetička sredina ta dva broja, tj. medijan je* $(2 + 3)/2 = 2.5$.

Postotna vrijednost, donji i gornji kvartil

Medijan je karakteriziran činjenicom da je barem pola (50%) podataka manje ili jednko od medijana, dok je istovremeno i barem 50% podataka veće ili jednako njemu. Analognim rezoniranjem karakterizirat ćemo postotnu vrijednost. Postotna vrijednost (eng. percentile value) za neki izabrani broj $p \in \langle 0, 100\rangle$, označimo je s x'_p, definira se poštujući zahtjev da je barem $p\%$ izmjerenih vrijednosti manje ili jednako x'_p, dok je barem $(100 - p)\%$ vrijednosti veće ili jednako x'_p. Dvadeset pet postotna vrijednost zove se donji kvartil (eng. lower quartile), a sedamdeset pet postotna vrijednost zove se gornji kvartil (eng. upper quartile). Donji i gornji kvartil su mjere koje spadaju u grupu mjera raspršenosti podataka.

Analogno kao i kod određivnja medijana, navedena karakterizacija postotne vrijednosti često ne određuje postotnu vrijednost podataka jedinstveno, tj. često postoji cijeli intarval realnih brojeva koji zadovoljava zadani kriterij. Predloženo je nekoliko metoda za određivanje postotne vrijednosti u takvim slučajevima. Programski paket *Statistica* u inačici 10 nudi šest načina računanja postotne vrijednosti čiji opis zainteresirani čitatelj može naći u elektronskom priručniku programskog paketa. Jedan od tih načina navodimo u nastavku teksta.

Postupak računanja postotne vrijednosti

Pretpostavimo da imamo n podataka i da želimo odrediti p-tu postotnu vrijednost x'_p, $p \in \langle 0, 100\rangle$. Prvo je potrebno podatke poredati u rastućem poretku i odrediti "poziciju" j koja je ključna za određivanje zadanog percentila kao $j = np/100$. Ako j nije prirodan broj, onda podatak na poziciji $j + 1$ odgovara p-toj postotnoj vrijednosti. Ako je j prirodan broj onda, se p-ta postotna vrijednost računa kao aritmetička sredina podataka na pozicijama j i $j + 1$.

Primjer 3.14. *Neka su izmjerene vrijednosti jedne varijable sljedeće:*

$$1, 2, 5, 6, 6, 1, 3, 7, 3, 3, 3, 3.$$

Prvo ove vrijednosti poredamo po veličini:

$$1, 1, 2, 3, 3, 3, 3, 3, 5, 6, 6, 7.$$

Želimo li odrediti donji kvartil, potrebno je prvo odrediti četvrtinu podataka (25%). S obzirom da imamo 12 podataka, četvrtinu (25%) čine tri podatka. Treći podatak u gornjem skupu je broj 2, a četvrti 3. Donji kvartil je 2.5. Deveti broj u gornjem skupu podataka je broj 5, a deseti 6 pa je gornji kvartil 5.5.

Najmanja i najveća vrijednost, raspon podataka

Raspon (eng. range) podataka je mjera koja pokazuje koliko su podaci raspršeni, tj. to je jedna od mjera raspršenosti podataka. Definiran je kao razlika najveće i najmanje vrijednosti u skupu mjerenih vrijednosti varijable (tj. razlika maksimalne i minimalne izmjerene vrijednosti varijable). Ako su $x_1, x_2, \ldots, x_n$ izmjerene vrijednosti varijable X, označimo najmanju od njih (minimum) s $x_{\min}$, a najveću s $x_{\max}$.

Primjer 3.15. *Neka su izmjerene vrijednosti jedne varijable sljedeće:*

$$1, 2, 5, 6, 5, 1, 2, 7, 2, 2, 3, 3.$$

Vidimo da je vrijednost 1 najmanja izmjerena vrijednost, a 7 najveća. Prema tome, raspon ovog skupa izmjerenih vrijednosti je $7 - 1 = 6$.

U mnogim primjerima zanimljivo je promatrati **maksimalno odstupanje izmjerenih vrijednosti varijable od "prosjeka", tj. aritmetičke sredine**, izmjerenih vrijednosti. Ta je numerička karakteristika definirana kao veći od brojeva $(\bar{x}_n - x_{\min})$ i $(x_{\max} - \bar{x}_n)$, tj. broj

$$\max\{(\bar{x}_n - x_{\min}), (x_{\max} - \bar{x}_n)\}.$$

Primjer 3.16. *Neka su $1, 2, 5, 6, 5, 1, 2, 7, 2, 2, 3, 3$ izmjerene vrijednosti neke varijable X. Tada je*

$$x_{\min} = 1, \quad x_{\max} = 7, \quad \bar{x}_n = \frac{1+2+5+6+5+1+2+7+2+2+3+3}{12} = 3.25.$$

Maksimalno odstupanje izmjerenih vrijednosti ove varijable od prosjeka izmjerenih vrijednosti je

$$\max\{3.25 - 1, 7 - 3.25\} = \max\{2.25, 3.75\} = 3.75.$$

Varijanca i standardna devijacija

Varijanca i standardna devijacija također spadaju u grupu mjera raspršenosti podataka. One karakteriziraju raspršenost podataka oko aritmetičke sredine. Varijanca niza izmjerenih vrijednosti $x_1, x_2, \ldots, x_n$ varijable X definirana je izrazom:

$$s_n^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2,$$

a standardna devijacija je kvadratni korijen varijance, tj.

$$s_n = \sqrt{s_n^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

Primjer 3.17. *Neka su izmjerene vrijednosti jedne varijable sljedeće:*

$$1.2, 2.1, 3.2, 4.3, 5.4, 6.5, 7.6, 8.7, 9.8.$$

Iz primjera 3.11 znamo da je aritmetička sredina ovog skupa podataka približno jednaka 5.42. *Varijanca ovog skupa podataka jest*

$$s_n^2 \approx \frac{1}{9}\sum_{i=1}^{9}(x_i - 5.42)^2 \approx 7.87,$$

a standardna devijacija

$$s_n \approx \sqrt{\frac{1}{9}\sum_{i=1}^{9}(x_i - 5.42)^2} \approx 2.81.$$

Mod

Mod je vrijednost iz niza izmjerenih vrijednosti varijable X kojoj pripada najveća frekvencija, tj. izmjerena je najviše puta. Mod ne mora biti jedinstven.

Primjer 3.18. *Neka su izmjerene vrijednosti jedne varijable sljedeće:*

$$1, 2, 5, 6, 5, 1, 2, 7, 2, 2, 3, 3.$$

Vidimo da je vrijednost 2 *izmjerena najviše puta (četiri puta) pa je* 2 *mod ovog skupa podataka.*

Primjer 3.19. *Neka su izmjerene vrijednosti jedne varijable sljedeće:*

$$1, 2, 5, 6, 5, 3, 1, 2, 7, 2, 2, 3, 3.$$

Vidimo da su najviše puta izmjerene dvije vrijednosi - 2 *i* 3 *su obje izmjerene točno četiri puta. Dakle, mod ovog skupa podataka nije jedinstven. U programskom paketu Statistica za mod ovog skupa izmjerenih vrijednosti pisalo bi* mod = multiple *te bismo u tom slučaju sve vrijednosti moda saznali analizom pripadne tablice frekvencija.*

Korištenjem numeričkih karakteristika podataka skup podataka može se prikazati grafički pomoću **kutijastog dijagrama** (eng. box plot, boxplot ili box-and-whiskers plot).

Kutijastim dijagramom prikazujemo odnos pet numeričkih karakteristika skupa izmjerenih vrijednosti: minimalnu vrijednost, donji kvartil, medijan, gornji kvartil i maksimalnu vrijednost. Na kutijastom dijagramu također se označavaju takozvane stršeće vrijednosti (eng. outliers) ako postoje.

Primjer 3.20. (trgovacki-centri.sta)
Pažljivim proučavanjem kretanja cijena prehrambenih proizvoda analitičar tržišta uočio je da isti proizvodi nemaju jednaku cijenu u različitim trgovačkim centrima. Promatrajući deset trgovačkih centara, zabilježio je cijene proizvoda kod kojega su razlike bile najizraženije (tablica 3.6).

trg. centar	1	2	3	4	5	6	7	8	9	10
cijena	45.52	44.64	39.99	48.95	51.59	46.89	52.02	56.89	50.21	49.99

Tablica 3.6: Cijene jednog proizvoda u deset različitih trgovačkih centara.

Numeričke karakteristike ovog skupa izmjerenih vrijednosti u programskom paketu Statistica možemo izračunati koristeći bazu podataka trgovacki-centri.sta *i provodeći sljedeći postupak:*

Statistics → Basic Statistics/Tables → Descriptive Statistics → Variables → Advanced → označiti mean (aritmetička sredina), mod, range (raspon), variance, standard deviation, median, minimum & maximum i lower & upper quartiles (donji i gornji kvartil) → Summary.

Rezultat ovog postupka (mjere deskriptivne statistike promatranog skupa izmjerenih vrijednosti) jesu tablice prikazane na slici 3.13.

	Descriptive Statistics (trgovacki-centri.sta)						
Variable	Valid N	Mean	Mode	Frequency of Mode	Range	Variance	Std.Dev.
cijena-proizvoda	10	48,66900	Multiple	1	16,90000	21,79821	4,668855

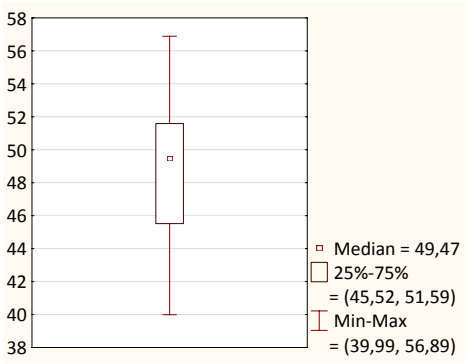
	Descriptive Statistics (trgovacki-centri.sta)						
Variable	Valid N	Median	Minimum	Maximum	Lower Quartile	Upper Quartile	Range
cijena-proizvoda	10	49,47000	39,99000	56,89000	45,52000	51,59000	16,90000

Slika 3.13: Deskriptivna statistika cijena iz tablice 3.6.

Uočimo da mod nije jedinstven - naime sve su izmjerene vrijednosti međusobno različite, tj. svaka je vrijednost izmjerena točno jedanput.

Za analiziranje raspršenosti cijena iz tablice 3.6 korisno je skicirati kutijasti dijagram na bazi medijana (slika 3.14) koji prikazuje odnos numeričkih karakteristika iz donje tablice sa slike 3.13 i koji u programskom paketu Statistica možemo napraviti provodeći sljedeći postupak:

Statistics → Basic Statistics/Tables → Descriptive Statistics → Variables → Options → pod "Options for Box-Whisker Plots" označiti opciju "Median/Quartiles/ Range" → Quick → Box and whisker Plot for all variables.



Slika 3.14: Kutijasti dijagram na bazi medijana za cijene iz tablice 3.6.

3.2.3 Detekcija stršećih vrijednosti

Podatak koji je značajno veći ili manji u odnosu na druge izmjerene vrijednosti jedne varijable nazivamo **stršeća vrijednost** (eng. outlier). Pojavljivanje stršećih vrijednosti najčešće je vezano uz jedan od sljedećih razloga:

- podatak je ili netočno izmjerena ili krivo unesen u bazu podataka
- podatak dolazi iz druge populacije (ne iz populacije koju promatramo u kontekstu problema koji proučavamo) - npr. ako u varijablu čije su izmjerene vrijednosti godišnje plaće 1000 poreznih obveznika u Hrvatskoj upišemo godišnju plaću Microsoftovog managera iz SAD-a, taj će podatak biti stršeća vrijednost
- podatak je točno izmjeren i unesen u bazu, ali predstavlja rijetku pojavu u populaciji - npr. ako se u varijabli čije su izmjerene vrijednosti koncentracije glukoze u krvi za 1000 osoba nađe točno izmjerena vrijednost 46.7, taj ćemo podatak smatrati stršećom vrijednošću jer se radi o vrlo visokoj koncentraciji glukoze koja se rijetko pojavljuje.

Vrlo korisna grafička metoda za detekciju stršećih vrijednosti jest kutijasti dijagram na bazi medijana. U programskom paketu Statistica kutijasti dijagrami osjetljivi na stršeće vrijednosti izrađuju se na sljedeći način:

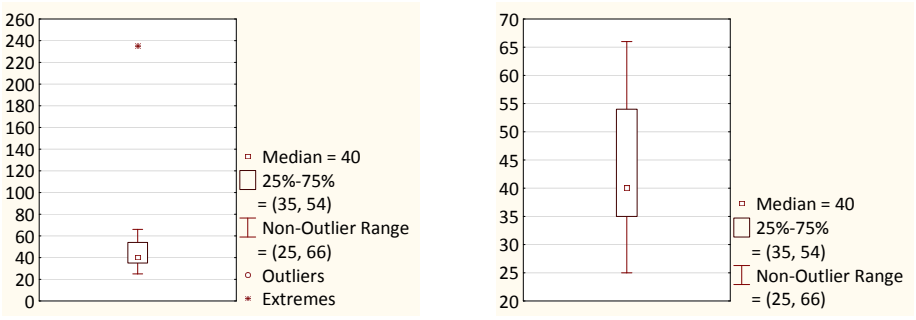
Graphs → 2D Graphs → BoxPlots → Variables → Advanced → pod Whisker odabrati "Non-outlier range" → pod Outliers odabrati "Outl. & Extremes" → OK.

Primjer 3.21. (zdravlje.sta)
Baza podataka zdravlje.sta *sadrži neke zdravstvene podatke za* 51 *ispitanika. Kratkom analizom mjera deskriptivne statistike možemo uočiti da je maksimum skupa izmjerenih vrijednosti* 235*, što u ovom primjeru znači da naš najstariji ispitanik ima* 235 *godina (slika 3.15).*

	Descriptive Statistics (zdravlje.sta)								
Variable	Valid N	Mean	Median	Mode	Frequency of Mode	Minimum	Maximum	Lower Quartile	Upper Quartile
godine	51.00	46.61	40.00	39.00000	7.00	25.00	235.00	35.00	54.00

Slika 3.15: Deskriptivna statistika izmjerenih vrijednosti varijable godine.

Taj je podatak stršeća vrijednost skupa izmjerenih vrijednosti varijable godine*. Međutim, ovaj način analize i detekcije stršećih vrijednosti nije prikladan za velike skupove podataka. Zato za detekciju stršećih vrijednosti često koristimo kutijaste dijagrame. Na slici 3.16 prikazan je kutijasti dijagram za varijablu* godine *sa stršećom vrijednošću te kutijasti dijagram koji dobivamo kad uklonimo stršeće vrijednosti.*



(a) uključena stršeća vrijednost (b) uklonjena stršeća vrijednost

Slika 3.16: Kutijasti dijagrami na bazi medijana za varijablu godine.

Uklanjanjem stršeće vrijednosti mijenjaju se i vrijednosti mjera deskriptivne statistike. Iz tablica sa slike 3.17 vidimo da su se uklanjanjem stršeće vrijednosti aritmetička sredina i gornji kvartil smanjili, dok su mod, medijan i donji kvartil ostali nepromijenjeni. Općenito, uklanjanjem stršećih vrijednosti mod će najčešće ostati nepromijenjen.

	Descriptive Statistics (zdravlje.sta) Include condition: godine < 150								
Variable	Valid N	Mean	Median	Mode	Frequency of Mode	Minimum	Maximum	Lower Quartile	Upper Quartile
godine	50.00	42.84	39.50	39.00000	7.00	25.00	66.00	35.00	53.00

Slika 3.17: Deskriptivna statistika izmjerenih vrijednosti varijable godine nakon uklanjanja stršeće vrijednosti.

3.3 Zadaci

Zadatak 3.1. (hormon.sta, nalaz.sta)
Baza podataka hormon.sta sadrži neke informacije i rezultate nekih medicinskih testova za svakog od 82 ispitanika:

varijabla spol sadrži informaciju o spolu ispitanika (m - ispitanik je muškog spola, z - ispitanik je ženskog spola)

varijable gastrS, somatS i somatZ sadrže izmjerene koncentracije određenih enzima utvrđene prilikom medicinske analize ispitanika

varijable pusenje, alkohol i kava sadrže informaciju o tome konzumira li ispitanik cigarete, alkohol i kavu (0 - ne konzumira, 1 - konzumira)

varijabla CLOtest sadrži rezultate testa na zarazu bakterijom helicobacter pilory (0 - test je negativan, 1 - test je pozitivan)

varijabla dijagnoza sadrži oznake dijagnoze ispitanika.

Baza podataka nalaz.sta sadrži neke informacije i rezultate testova o koncentraciji nekih tvari u krvi za svakog od 102 ispitanika:

varijabla skupina sadrži informaciju o pripadnosti ispitanika jednoj od devet dobnih skupina (g1 - g9)

varijable k1 - k8 sadrže izmjerene koncentracije promatranih tvari u krvi

varijabla stupanj sadrži stupnjevanje rezultata provedenih testova s obzirom na dobnu skupinu kojoj ispitanik pripada (u skali od 1 do 10).

Proučite varijable u prethodno opisanim bazama podataka te pomoću programskog paketa Statistica odredite frekvencije i relativne frekvencije svih kategorija za varijable koje smatrate kvalitativnima. Rezultate prikažite tablično.

Rješenje. *Tablice frekvencija i relativnih frekvencija za kvalitativne varijable s najvećim brojem kategorija - varijable* dijagnoza *iz baze podataka* hormon.sta *i varijable* stupanj *iz baze podataka* nalaz.sta *prikazane su na slici 3.18.*

	Frequency table: dijagnoza (hormon.sta)			
Category	Count	Cumulative Count	Percent	Cumulative Percent
G	21	21	25,61	25,61
E b	4	25	4,88	30,49
U b	30	55	36,59	67,07
U z	13	68	15,85	82,93
E z	14	82	17,07	100,00
Missing	0	82	0,00	100,00

(a) varijabla dijagnoza (hormon.sta)

	Frequency table: stupanj (nalaz.sta)			
Category	Count	Cumulative Count	Percent	Cumulative Percent
1	12	12	11,76	11,76
2	11	23	10,78	22,55
3	12	35	11,76	34,31
4	9	44	8,82	43,14
5	11	55	10,78	53,92
6	10	65	9,80	63,73
7	12	77	11,76	75,49
8	8	85	7,84	83,33
9	8	93	7,84	91,18
10	9	102	8,82	100,00
Missing	0	102	0,00	100,00

(b) varijabla stupanj (nalaz.sta)

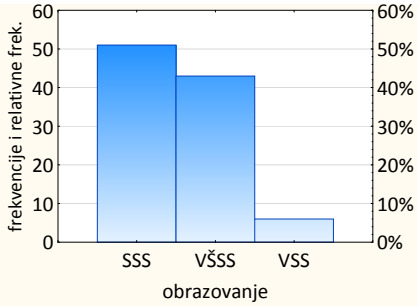
Slika 3.18: Frekvencije i relativne frekvencije svih kategorija varijabli dijagnoza i stupanj.

Zadatak 3.2. (djelatnici.sta)
Baza podataka djelatnici.sta opisana je u primjeru 2.4. Za kvalitativnu varijablu obrazovanje, čije su vrijednosti svrstane u tri kategorije: SSS - srednja stručna sprema, VŠSS - viša stručna sprema, VSS - visoka stručna sprema, odredite zastupljenost tih kategorija u promatranom uzorku od 100 djelatnika.

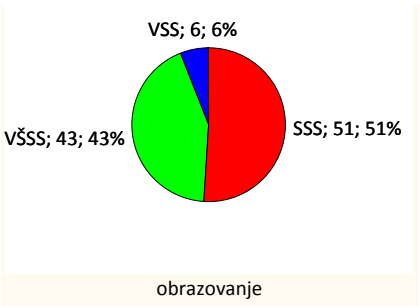
Rješenje. *Zastupljenost kategorija opisana je tablicom frekvencija i relativnih frekvencija 3.19 te stupčastim dijagramom i kružnim dijagramom frekvencija i relativnih frekvencija koji su prikazani na slici 3.20.*

	Frequency table: obrazovanje (djelatnici.sta)			
Category	Count	Cumulative Count	Percent	Cumulative Percent
SSS	51	51	51.00	51.00
VŠSS	43	94	43.00	94.00
VSS	6	100	6.00	100.00
Missing	0	100	0.00	100.00

Slika 3.19: Frekvencije i relativne frekvencije svih kategorija varijabli obrazovanje.



(a) stupčasti dijagram



(b) kružni dijagram

Slika 3.20: Grafički prikazi podataka varijable obrazovanje.

Zadatak 3.3. (nalaz.sta)
U bazi podataka nalaz.sta (opisanoj u zadatku 3.1) odredite frekvencije i relativne frekvencije svih kategorija za varijable koje smatrate kvalitativnima.

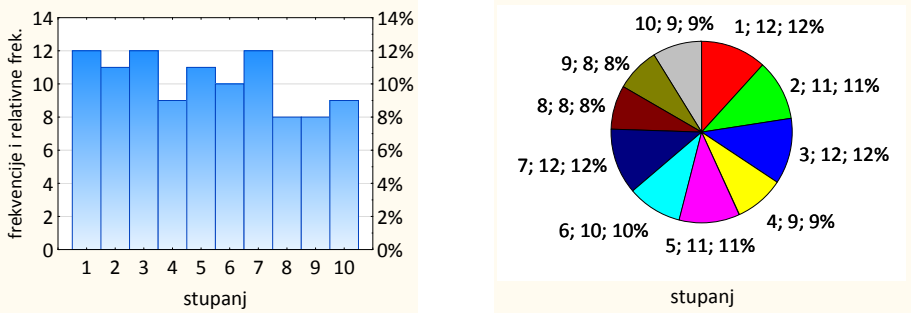
a) Rezultate prikažite grafički koristeći programski paket Statistica.

b) Za koliko je ispitanika vrijednost varijable stupanj manja od tri, za koliko je vrijednost barem četiri, ali manja od sedam, a za koliko je vrijednost barem osam?

c) Za frekvencije iz zadatka b) odredite pripadne relativne frekvencije.

Rješenje.

a) Grafički prikazi frekvencija i relativnih frekvencija kategorija kvalitativne varijable stupanj *prikazani su na slici 3.21.*



Slika 3.21: Grafički prikazi frekvencija i relativnih frekvencija svih kategorija varijable stupanj.

b) Frekvencija ispitanika za koje je vrijednost varijable stupanj *manja od tri je* 23*, frekvencija ispitanika za koje je vrijednost barem četiri, ali manja od sedam je* 30*, a frekvencija ispitanika za koje je vrijednost barem osam je* 25*.*

c) Pripadne relativne frekvencije su redom $23/102 \approx 22.55\%$*,* $30/102 \approx 29.41\%$ *i* $25/102 \approx 24.51\%$*.*

Zadatak 3.4. (djeca.sta)

U bazi podataka djeca.sta nalazi se dio podataka o nekim ocjenama novorođenčeta, načinu poroda i majci iz istraživanja koje je provedeno u jednoj bolnici:

varijabla spol sadrži spol novorođenčeta

varijabla nacin-poroda informaciju o načinu poroda

varijable RM, apgar1 i apgar5 izmjerene vrijednosti nekih obilježja novorođenčeta

varijabla majka-dob godine starosti majke

varijabla majka-bolest informaciju o bolesti majke tijekom trudnoće (N - nije bila bolesna, D - bila je bolesna)

varijabla komplikacije stupanj komplikacija za vrijeme trudnoće (u skali od 0, što označava da komplikacija nije bilo, do 7)

varijabla konvulzije informaciju o konvulzijama kod novorođenčeta (N - konvulzija nije bilo, D - konvulzije su bile prisutne)

varijabla uzv jednu ocjenu ultrazvučnog pregleda mozga novorođenčeta (u skali od 1 do 4).

Odredite frekvencije i relativne frekvencije svih kategorija za varijable koje smatrate kvalitativnima.

a) Rezultate prikažite tablično i grafički koristeći programski paket Statistica.

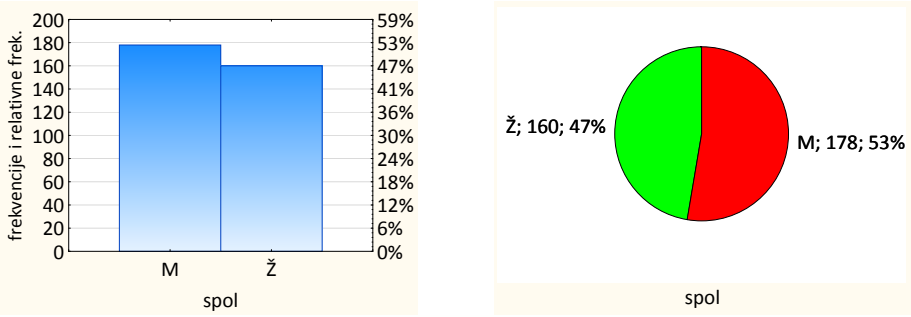
b) Broji li ovaj uzorak više djevojčica ili dječaka? Koliki je udio majki starijih od 35 godina?

Rješenje.

a) Tablični i grafički prikazi frekvencija i relativnih frekvencija svih kategorija varijable spol *prikazani su na slikama 3.22 i 3.23.*

	Frequency table: spol (djeca.sta)			
Category	Count	Cumulative Count	Percent	Cumulative Percent
M	178	178	52,66	52,66
Ž	160	338	47,34	100,00
Missing	0	338	0,00	100,00

Slika 3.22: Tablica frekvencija i relativnih frekvencija svih kategorija varijable spol.



(a) stupčasti dijagram (b) kružni dijagram

Slika 3.23: Grafički prikazi frekvencija i relativnih frekvencija svih kategorija varijable spol.

b) Uzorkom je obuhvaćeno 338 *novorođenčadi* - 160 *djevojčica i* 178 *dječaka. Dakle, u uzorku ima više dječaka. Majki starijih od* 35 *godina ima* $29/338 \approx 8.58\%$.

Zadatak 3.5. (navike.sta)
U bazi podataka navike.sta (opisanoj u zadatku 2.4) odredite frekvencije i relativne frekvencije svih kategorija za varijable koje smatrate kvalitativnima.

a) Rezultate prikažite tablično i grafički koristeći programski paket Statistica.

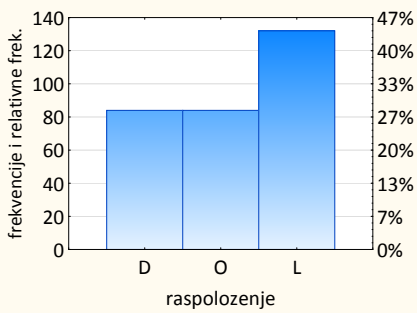
b) Koliko je ispitanika dobro raspoloženo? Je li više ispitanika raspoloženo dobro ili osrednje ili ih je najviše lošeg raspoloženja?

Rješenje.

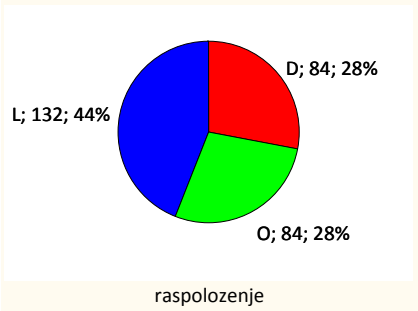
a) Tablični i grafički prikazi frekvencija i relativnih frekvencija svih kategorija varijable raspolozenje *prikazani su na slikama 3.24 i 3.25.*

Category	Frequency table: raspolozenje (navike.sta)			
	Count	Cumulative Count	Percent	Cumulative Percent
D	84	84	28,00	28,00
O	84	168	28,00	56,00
L	132	300	44,00	100,00
Missing	0	300	0,00	100,00

Slika 3.24: Tablica frekvencija i relativnih frekvencija svih kategorija varijable raspolozenje.



(a) stupčasti dijagram



(b) kružni dijagram

Slika 3.25: Grafički prikazi frekvencija i relativnih frekvencija svih kategorija varijable raspolozenje.

b) Uzorkom je obuhvaćeno 300 *ispitanika. Dobro je raspoloženo njih* 84*, što čini* $84/300 = 28\%$ *od ukupnog broja ispitanika. Osrednje je raspoloženo također* 84 *(*28%*) ispitanika, a loše njih* 132 *(*44%*). Dakle, više je ispitanika koji su raspoloženi dobro ili osrednje - u te dvije kategorije spada* 168 *(56 %) ispitanika.*

Zadatak 3.6. (zdravlje.sta)
Često ima smisla analizirati frekvencije i relativne frekvencije numeričkih ili ordinalnih varijabli za pojedine kategorije zadane kvalitativne varijable. Na primjer, korisno je analizirati određene zdravstvene karakteristike posebno za osobe ženskog, a posebno za osobe muškog spola. Analizirajte ordinalnu varijablu zdravlje po kvalitativnoj varijabli spol iz baze podataka zdravlje.sta koja je opisana u zadatku 2.4.

Rješenje. *Prvo ćemo tablično i grafički prikazati frekvencije i relativne frekvencije za podatke sadržane u varijablama* zdravlje *i* spol *(slike 3.26, 3.27 i 3.28).*

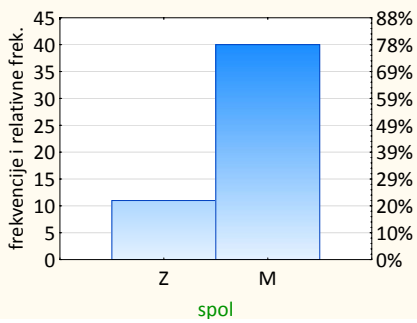
Category	Frequency table: spol (zdravlje.sta) Count	Cumulative Count	Percent	Cumulative Percent
Z: žena	11	11	21,57	21,57
M: muškarac	40	51	78,43	100,00
Missing	0	51	0,00	100,00

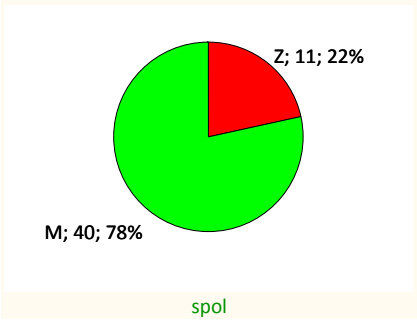
(a) varijabla spol

Category	Frequency table: zdravlje (zdravlje.sta) Count	Cumulative Count	Percent	Cumulative Percent
1	4	4	7,84	7,84
2	8	12	15,69	23,53
3	18	30	35,29	58,82
4	12	42	23,53	82,35
5	9	51	17,65	100,00
Missing	0	51	0,00	100,00

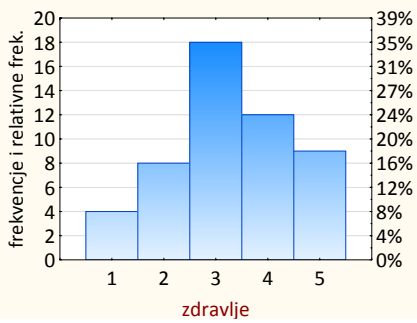
(b) varijabla zdravlje

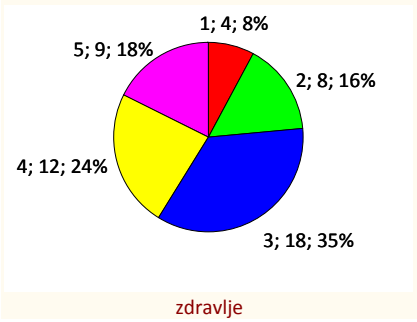
Slika 3.26: Tablice frekvencija i relativnih frekvencija svih podataka varijabli spol i zdravlje.





Slika 3.27: Grafički prikazi frekvencija i relativnih frekvencija svih podataka varijable spol.





Slika 3.28: Grafički prikazi frekvencija i relativnih frekvencija svih podataka varijable zdravlje.

Tablični i grafički prikazi podataka sadržanih u varijabli zdravlje *posebno za kategoriju ispitanika ženskog spola, a posebno za kategoriju ispitanika muškog spola prikazani su na slikama 3.29, 3.30 i 3.31. Kružne dijagrame relativnih frekvencija sa slike 3.31 u programskom paketu Statistica možemo dobiti provodeći sljedeći postupak:*

Graphs → Categorized Graphs → Pie Charts → Graph Type: Pie Chart - Counts → Variables (Vars - zdravlje, X-Category - spol) → Advanced → Pie Legend (Text and Value za kružne dijagrame

frekvencija, Text and Percent za kružne dijagrame relativnih frekvencija).

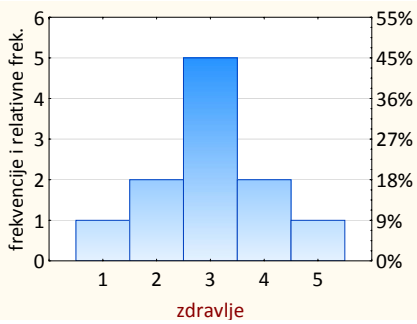
Category	Frequency table: zdravlje (zdravlje.sta) Include condition: spol="Z"			
	Count	Cumulative Count	Percent	Cumulative Percent
1	1	1	9,09	9,09
2	2	3	18,18	27,27
3	5	8	45,45	72,73
4	2	10	18,18	90,91
5	1	11	9,09	100,00
Missing	0	11	0,00	100,00

(a) žene (spol=Z)

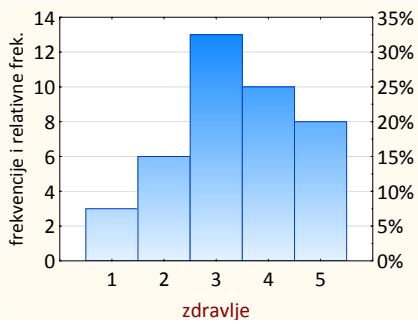
Category	Frequency table: zdravlje (zdravlje.sta) Include condition: spol="M"			
	Count	Cumulative Count	Percent	Cumulative Percent
1	3	3	7,50	7,50
2	6	9	15,00	22,50
3	13	22	32,50	55,00
4	10	32	25,00	80,00
5	8	40	20,00	100,00
Missing	0	40	0,00	100,00

(b) muškarci (spol=M)

Slika 3.29: Tablični prikaz podataka za varijablu zdravlje kategoriziranih prema spolu ispitanika.

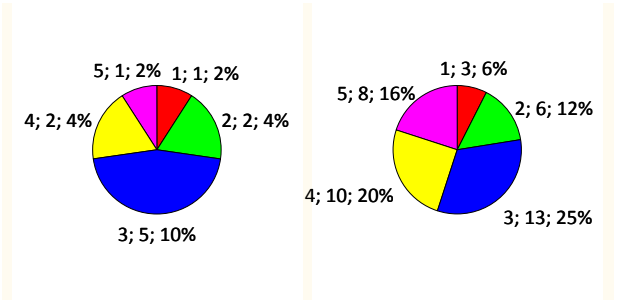


(a) žene (spol=Z)



(b) muškarci (spol=M)

Slika 3.30: Stupčasti dijagrami podataka varijable zdravlje kategoriziranih prema spolu ispitanika.



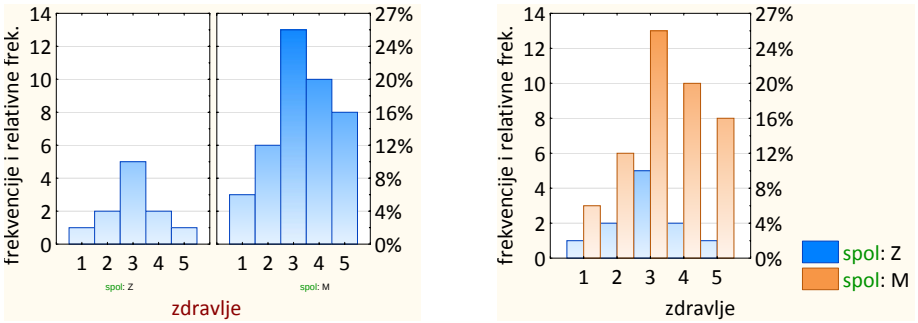
žene (spol=Z) muškarci (spol=M)

Slika 3.31: Kružni dijagrami podataka varijable zdravlje kategoriziranih prema spolu ispitanika.

Radi uspoređivanja rezultata po spolu korisno je stupčaste dijagrame frekvencija i relativnih frek-

vencija podataka sadržanih u varijabli zdravlje *kategoriziranih prema spolu ispitanika prikazati na jednoj slici, tj. grafu (slika 3.32). Objedinjene dijagramske prikaze frekvencija i relativnih frekvencija neke varijable čije su vrijednosti kategorizirane po nekom kriteriju možemo dobiti u programskom paketu Statistica provodeći sljedeći postupak:*

Graphs → Categorized Graphs → Histograms → Variables (Variable - zdravlje, X-Category - spol) → Layout (Separate - za odvojene stupčaste dijagrame kategorija varijable zdravlje kategoriziranih s obzirom na vrijednosti varijable spol; Overlaid - za prikaz frekvencija kategorija varijable zdravlje kategoriziranih s obzirom na vrijednosti varijable spol na istom stupčastom dijagramu)



Slika 3.32: Stupčasti dijagrami podataka varijable zdravlje kategoriziranih prema spolu ispitanika.

Zadatak 3.7. (TV-program.sta)
Za varijable iz baze podataka TV-program.sta napravite sljedeće tablične i grafičke prikaze:

a) napravite tablice i nacrtajte stupčaste dijagrame frekvencija i relativnih frekvencija za podatke sadržane u varijablama spol i P1,

b) napravite tablice i nacrtajte stupčaste dijagrame frekvencija i relativnih frekvencija za podatke sadržane u varijabli P1 posebno za kategoriju ispitanika ženskog spola, a posebno za kategoriju ispitanika muškog spola,

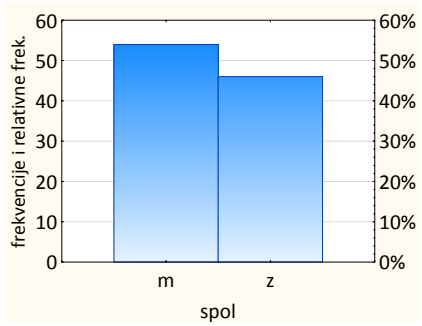
c) nacrtajte kružne dijagrame frekvencija i relativnih frekvencija za podatke sadržane u varijablama spol i P3,

d) nacrtajte kružne dijagrame frekvencija i relativnih frekvencija za podatke sadržane u varijabli P3 posebno za kategoriju ispitanika ženskog spola, a posebno za kategoriju ispitanika muškog spola.

Rješenje.

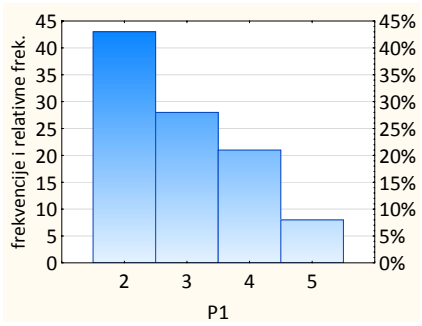
a) Tablični i grafički prikazi frekvencija i relativnih frekvencija svih kategorije varijable spol *i svih različitih vrijednosti varijable* P1 *prikazani su na slikama 3.33 i 3.34.*

Category	Frequency table: spol (TV-program.sta)			
	Count	Cumulative Count	Percent	Cumulative Percent
m	54	54	54,00	54,00
z	46	100	46,00	100,00
Missing	0	100	0,00	100,00



Slika 3.33: Tablica i stupčasti dijagram za podatke varijable spol.

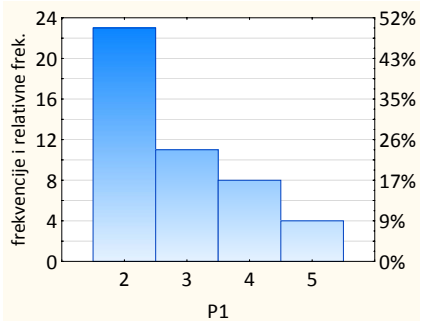
Category	Frequency table: P1 (TV-program.sta)			
	Count	Cumulative Count	Percent	Cumulative Percent
2	43	43	43,00	43,00
3	28	71	28,00	71,00
4	21	92	21,00	92,00
5	8	100	8,00	100,00
Missing	0	100	0,00	100,00



Slika 3.34: Tablica i stupčasti dijagram za podatke varijable P1.

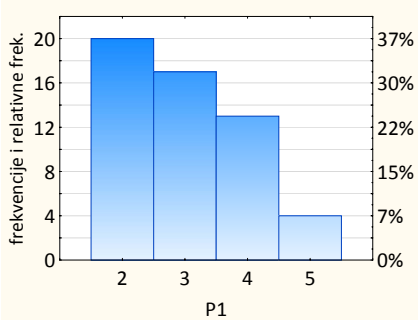
b) Tablični i grafički prikazi frekvencija i relativnih frekvencija svih kategorija varijable P1 *kategoriziranih prema spolu ispitanika prikazani su na slikama 3.35, 3.36 i 3.37.*

Category	Frequency table: P1 (TV-program.sta) Include condition: spol="z"			
	Count	Cumulative Count	Percent	Cumulative Percent
2	23	23	50,00	50,00
3	11	34	23,91	73,91
4	8	42	17,39	91,30
5	4	46	8,70	100,00
Missing	0	46	0,00	100,00

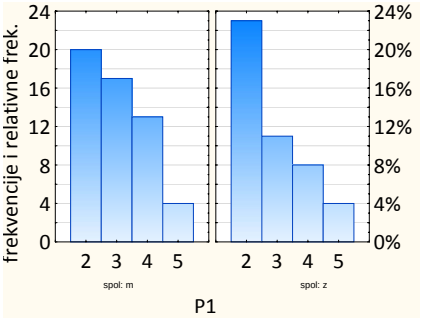


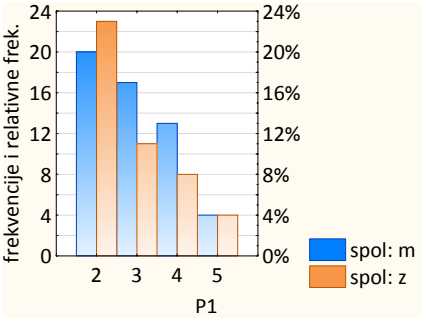
Slika 3.35: Tablica i stupčasti dijagram za podatke varijable P1 za ženski spol.

Category	Frequency table: P1 (TV-program.sta) Include condition: spol="m"			
	Count	Cumulative Count	Percent	Cumulative Percent
2	20	20	37,04	37,04
3	17	37	31,48	68,52
4	13	50	24,07	92,59
5	4	54	7,41	100,00
Missing	0	54	0,00	100,00



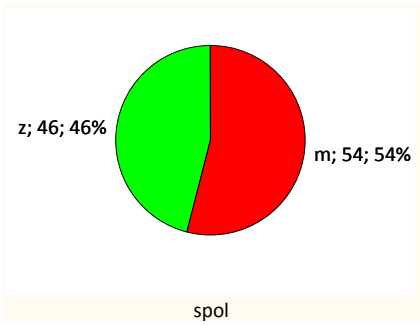
Slika 3.36: Tablica i stupčasti dijagram za podatke varijable P1 za muški spol.



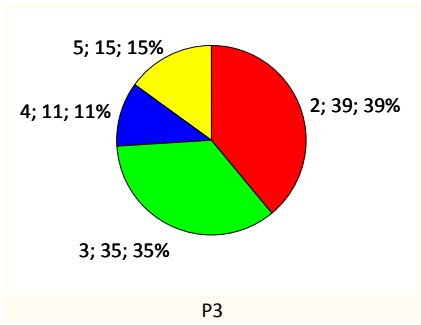


Slika 3.37: Stupčasti dijagrami za podatke varijable P1 kategorizirane prema spolu ispitanika.

d) Kružni dijagrami frekvencija i relativnih frekvencija svih kategorija varijable spol *i svih različitih vrijednosti varijable* P3 *prikazani su na slici 3.38.*



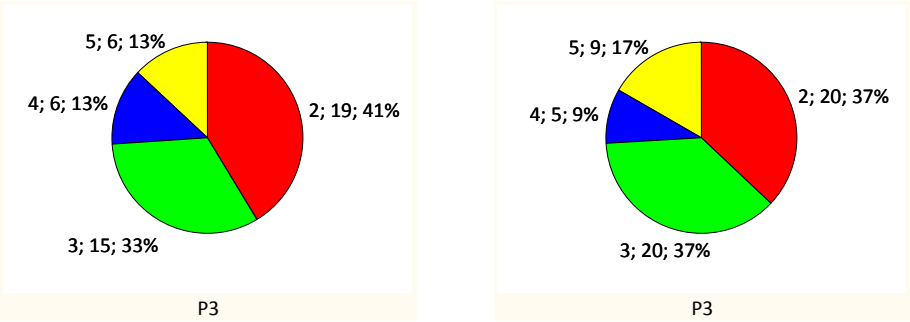
(a) varijabla spol



(b) varijabla P3

Slika 3.38: Kružni dijagrami za podatke varijabli spol i P3.

e) Kružni dijagrami relativnih frekvencija za podatke iz varijable P3 *kategorizirane prema spolu ispitanika prikazani su na slici 3.39.*



(a) spol = z (b) spol = m

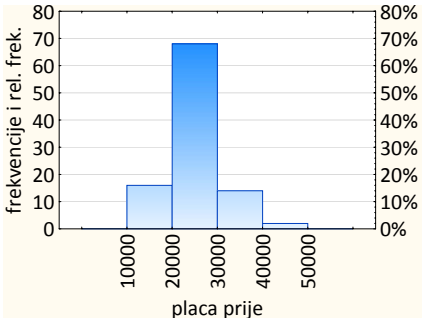
Slika 3.39: Kružni dijagrami za podatke varijable P3 kategorizirane prema spolu ispitanika.

Zadatak 3.8. (djelatnici.sta)
Promotrite varijablu placa prije iz baze podataka djelatnici.sta opisane u primjeru 2.4. Razvrstajte vrijednosti u disjunktne intervale duljine 10000 počevši od nule te prikažite podatke tablično i histogramom.

Rješenje. *Tablični prikaz frekvencija i relativnih frekvencija dan je tablicom 3.7, a pripadni histogram slikom 3.40. Ovakakv histogram jasno ilustrira činjenicu da najviše djelatnika u uzorku ima godišnju plaću od* 20000 *do* 30000 *novčanih jedinica, dok je plaća iz intervala* 40000 *do* 50000 *rijetkost. Intervale za kategorizaciju u ovakvim i sličnim slučajevima obično radimo tako da bismo zadovoljili potrebe za prezentiranjem informacija koje želimo istaknuti.*

iznos plaće	frekvencija	relativna frekvencija
$[0, 10000\rangle$	0	0
$[10000, 20000\rangle$	15	0.15
$[20000, 30000\rangle$	69	0.69
$[30000, 40000\rangle$	14	0.14
$[40000, 50000\rangle$	2	0.02

Tablica 3.7: Tablica frekvencija i relativnih frekvencija kategoriziranih podataka varijable placa prije.



Slika 3.40: Histogram frekvencija i relativnih frekvencija kategoriziranih podataka varijable placa prije.

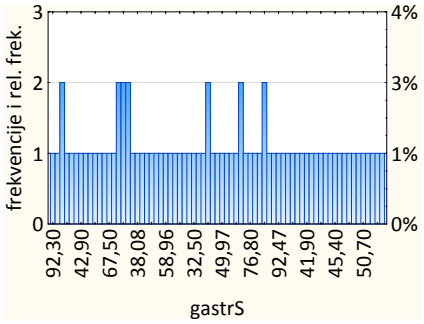
Zadatak 3.9. (hormon.sta)

a) Odredite tablicu frekvencija i stupčasti dijagram za neprekidnu numeričku varijablu gastrS iz baze podataka hormon.sta (koja je opisana u zadatku 3.1) tako da za kategorije uzmete sve međusobno različite izmjerene vrijednosti.

b) Iskoristite izmjerene vrijednosti varijable gastrS, kategorizirajte podatke i prikažite ih histogramom. Mijenjajte broj intervala na koji dijelite skup vrijednosti. Proučavajte što se događa i pribilježite svoj zaključak.

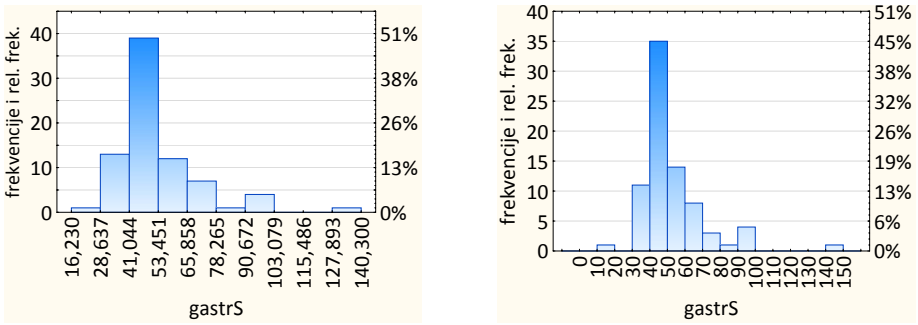
Rješenje.

a) Stupčasti dijagram frekvencija i relativnih frekvencija te kružni dijagram izmjerenih vrijednosti varijable gastrS *u kojima su kao kategorije uzete sve različite izmjerene vrijednosti prikazani su na slici 3.41.*



Slika 3.41: Stupčasti dijagram svih izmjerenih vrijednosti varijable gastr S.

b) Kategorizacija izmjerenih vrijednosti varijable gastrS *na disjunktne intervale daje preglednije grafičke prikaze iz kojih je lakše analizirati izmjerene vrijednosti i donijeti neke zaključke. Grafički prikazi frekvencija i relativnih fekvencija izmjerenih vrijednosti varijable* gastrS *razvrstanih u* 10 *i* 15 *disjunktnih intervala prikazani su na slici 3.42.*



Slika 3.42: Histogram za podatke varijable gastrS.

Zadatak 3.10. (djelatnici.sta)
Odredite numeričke karakteristike skupa izmjerenih vrijednosti varijable placa prije iz baze podataka djelatnici.sta opisane u primjeru 2.4.

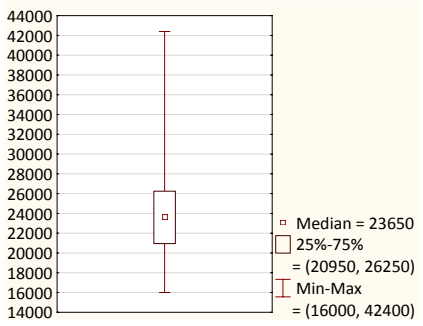
Rješenje. *Numeričke karakteristike prikazane su u tablicama na slici 3.43.*

Variable	Descriptive Statistics (djelatnici.sta)					
	Valid N	Mean	Mode	Frequency of Mode	Variance	Std.Dev.
placa prije	100.00	24522.00	24600.00	4.00	26069208.08	5105.80

Variable	Descriptive Statistics (djelatnici.sta)					
	Median	Minimum	Maximum	Lower Quartile	Upper Quartile	Range
placa prije	23650.00	16000.00	42400.00	20950.00	26250.00	26400.00

Slika 3.43: Deskriptivna statistika izmjerenih vrijednosti varijable placa prije.

Odnos minimuma, donjeg kvartila, medijana, gornjeg kvartila i maksimuma izmjerenih vrijednosti varijable placa prije *prikazani su kutijastim dijagramom 3.44.*



Slika 3.44: Kutijasti dijagram na bazi medijana za varijablu placa prije.

Iz tablice 3.43 i kutijastog dijagrama 3.44 možemo izvesti sljedeće i slične zaključke:

- *najniža godišnja plaća u uzorku iznosi* 16000, *a najviša* 42400
- *bar* 25% *ispitanika iz uzorka ima plaću manju ili jednaku* 20950
- *bar* 25% *ispitanika iz uzorka ima plaću veću ili jednaku* 26250
- *bar* 50% *ispitanika iz uzorka ima plaću manju ili jednaku medijanu, tj.* 23650
- *bar* 50% *ispitanika iz uzorka ima plaću veću ili jednaku* 23650.

Zadatak 3.11. (nastava.sta)
Baza podataka nastava.sta sadrži ocjene u skali od 0 (najniža ocjena) do 10 (najviša ocjena) različitih komponenti probnog nastavnog sata za 65 studenata (budućih nastavnika):

varijabla znanje sadrži ocjene znanja studenta o temi nastavnog sata

varijabla literatura sadrži ocjene primjerenosti korištene literature za pripremu nastavnog sata

varijabla predavac sadrži ocjene predavačeva stava i nastupa pred razredom

varijabla atmosfera sadrži ocjene radne atmosfere na nastavnom satu

varijabla govor sadrži ocjene studentova izražavanja tijekom nastavnog sata

varijabla interes sadrži ocjene pobuđenosti interesa kod učenika za temu nastavnog sata

varijabla bitan sadrzaj sadrži ocjene naglašenosti bitnih sadržaja tijekom nastavnog sata

varijabla primjeri sadrži ocjene odabira i primjerenosti primjera prezentiranih tijekom nastavnog sata

varijabla ukupno sadrži ocjene koje odražavaju ukupan ocjenjivačev dojam o održanom nastavnom satu.

Ako želimo donijeti opći zaključak o uspješnosti budućih nastavnika u stvarnoj nastavnoj situaciji, logično je pažnju usmjeriti na analizu varijable ukupno. Odredite numeričke karakteristike te varijable i kutijasti dijagram na bazi medijana. Diskutirajte o rezultatima.

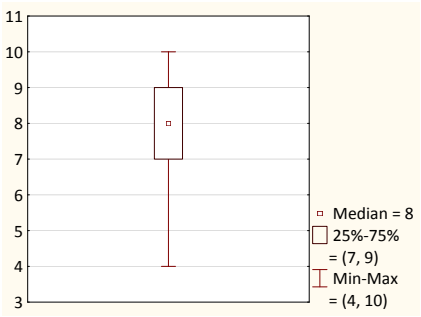
Rješenje. *Numeričke karakteristike te varijable prikazane su u tablici 3.45.*

	Descriptive Statistics (nastava.sta)					
Variable	Valid N	Mean	Mode	Frequency of Mode	Variance	Std.Dev.
ukupno	65.00	8.11	Multiple	19.00	2.16	1.47

	Descriptive Statistics (nastava.sta)					
Variable	Median	Minimum	Maximum	Lower Quartile	Upper Quartile	Range
ukupno	8.00	4.00	10.00	7.00	9.00	6.00

Slika 3.45: Deskriptivna statistika podataka za varijablu ukupno.

Iz tablice frekvencija za varijablu ukupno *lako se vidi da skup podataka te varijable ima dva moda - to su ocjene* 8 *i* 9. *Dakle, probno je predavanje za čak* 19 *studenata ocijenjeno visokom ocjenom* 8 *te za isto toliko ocjenom* 9, *dok je prosječna ocjena ukupnog dojma probnog nastavnog sata* 8.11. *Analizu raspršenosti ocjena napravit ćemo pomoću kutijastog dijagrama (slika 3.46).*



Slika 3.46: Kutijasti dijagram na bazi medijana za podatke varijable ukupno.

Analiza kutijastog dijagrama sugerira sljedeće zaključke: nitko od ispitanika predavanje nije ocijenio ocjenom nižom od četiri, barem 25% *ispitanika predavanje je ocijenilo ocjenama* 4, 5, 6 *ili* 7, *barem* 25% *ocjenama* 7 *ili* 8, *barem* 25% *ocjenama* 8 *ili* 9 *te barem* 25% *ocjenama* 9 *ili* 10. *Zanimljivo je uočiti da je barem* 75% *ispitanika predavanje ocijenilo ocjenom* 7 *i više.*

Zadatak 3.12. (matematika.sta)
Baza podataka matematika.sta (opisana u primjeru 2.9) sadrži rezultate ankete o kvaliteti izvođenja nekog matematičkog kolegija. Ukoliko nas zanima prilagođenost težine sadržaja kolegija predznanju studenata, analizirat ćemo varijablu tezina kolegija. Odredite numeričke karakteristike podataka te varijable i prikažite ih kutijastim dijagramom.

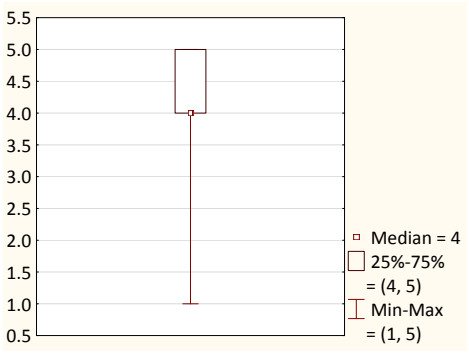
Rješenje. *Mjere deskriptivne statistike varijable* tezina kolegija *prikazane su u tablici na slici 3.47.*

	Descriptive Statistics (matematika.sta)					
Variable	Valid N	Mean	Mode	Frequency of Mode	Variance	Std.Dev.
tezina kolegija	49.00	4.18	5.00	21.00	0.78	0.88

	Descriptive Statistics (matematika.sta)					
Variable	Median	Minimum	Maximum	Lower Quartile	Upper Quartile	Range
tezina kolegija	4.00	1.00	5.00	4.00	5.00	4.00

Slika 3.47: Deskriptivna statistika podataka varijable tezina kolegija.

Uočimo da je čak 21 *ispitanik prilagođenost težine kolegija predznanju studenata ocijenio ocjenom* 5 *(ocjena* 5 *je mod ovog skupa podataka) te da je prosječna ocjena* 4.18. *Za analizu raspršenosti ocjena koristimo kutijasti dijagram prikazan na slici 3.48.*



Slika 3.48: Kutijasti dijagram na bazi medijana za varijablu tezina kolegija.

Analizom kutijastog dijagrama donosimo sljedeći zaključak: barem 25% *ispitanika težinu kolegija ocijenilo je ocjenama* 1, 2, 3 *ili* 4, *barem* 50% *ocjenom* 4 *te barem* 25% *ocjenama* 4 *ili* 5. *Zanimljivo je uočiti da je barem* 75% *ispitanika težinu kolegija ocijenilo ocjenam* 4 *ili* 5.

Zadatak 3.13. (djelatnici.sta)
Varijabla dob iz baze podataka djelatnici.sta opisane u primjeru 2.4 za svakog ispitanika iz uzorka djelatnika promatranog poduzeća sadrži informaciju o dobi u godinama. Odredite numeričke karakteristike podataka iz te varijable, analizirajte postojanje stršećih vrijednosti, prikažite podatke kutijastim dijagramom i diskutirajte o rezultatima.

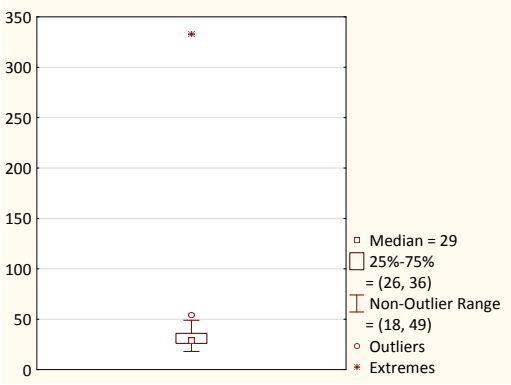
Rješenje. *Iz deskriptivne statistike varijable* dob *(tablica 3.49) vidimo da je maksimalna podatak za dob 333 godine pa je očigledno da postoji stršeći podatak koji je pogrešno upisan u bazu podataka.*

	Descriptive Statistics (djelatnici.sta)					
Variable	Valid N	Mean	Mode	Frequency of Mode	Variance	Std.Dev.
dob	100.00	33.83	28.00000	12.00	964.28	31.05

	Descriptive Statistics (djelatnici.sta)					
Variable	Median	Minimum	Maximum	Lower Quartile	Upper Quartile	Range
dob	29.00	18.00	333.00	26.00	36.00	315.00

Slika 3.49: Deskriptivna statistika podataka varijable dob.

Osim iz tablice 3.49, stršeće vrijednosti među podacima varijable dob *mogli smo detektirati i pomoću kutijastog dijagrama na bazi medijana.*



Slika 3.50: Kutijasti dijagram na bazi medijana s prikazom stršećih vrijednosti varijable dob.

Kao što vidimo iz kutijastog dijagrama 3.50, i dob od 54 *godine prepoznata je kao stršeća vrijednost. Budući da je sasvim razumljivo da promatrano poduzeće može imati djelatnika starog* 54 *godine, taj podatak smatramo točnim, no radi se o dobi koja se rijetko pojavljuje u populaciji djelatnika tog poduzeća.*

Zadatak 3.14. (glukoza.sta)
Varijabla dob baze podataka glukoza.sta sadrži godine starosti, a varijabla koncentracija izmjerene vrijednosti koncentracije glukoze u krvi za 102 ispitanika. Korištenjem programskog paketa Statistica riješite sljedeće zadatke:

a) Napravite deskriptivnu statistiku podataka sadržanih u varijabli koncentracija. Grafičkom metodom odredite stršeću vrijednost u ovom skupu podataka. Možete li se složiti s tvrdnjom da je identificirani podatak moguća izmjerena vrijednost ili ipak sumnjate u dobiveni rezultat? Obrazložite svoj odgovor.

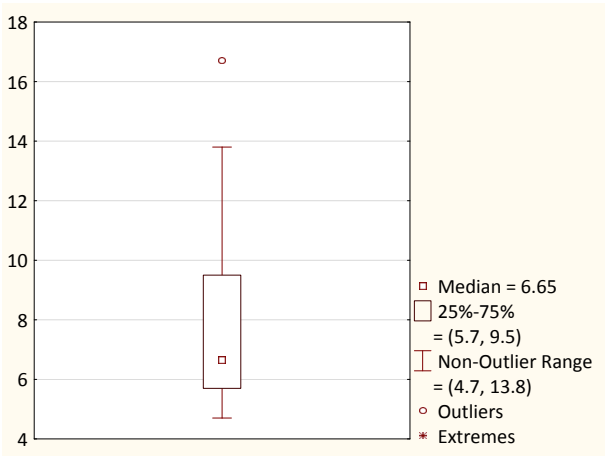
b) Grafičkom metodom identificirajte stršeće vrijednosti među podacima u varijabli dob. Što se događa s numeričkim karakteristikama podataka nakon uklanjanja stršeće vrijednosti?

Rješenje.

a) Deskriptivna statistika i kutijasti dijagram s označenim stršećim vrijednostima skupa izmjerenih vrijednosti varijable koncentracija *prikazani su na slikama 3.51 i 3.52.*

	Descriptive Statistics (glukoza.sta)							
Variable	Valid N	Mean	Median	Mode	Frequency of Mode	Minimum	Maximum	Lower Quartile
koncentracija	102.00	7.70	6.65	5.500000	14.00	4.70	16.70	5.70

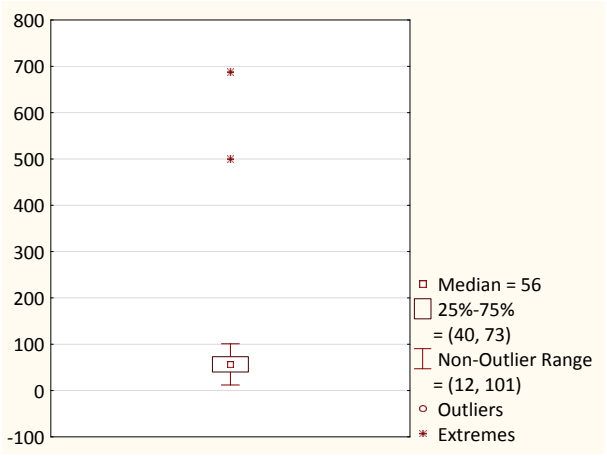
Slika 3.51: Deskriptivna statistika izmjerenih vrijednosti varijable koncentracija.



Slika 3.52: Kutijasti dijagram na bazi medijana s prikazom stršećih vrijednosti varijable dob.

Statistica je kao stršeću vrijednost detektirala podatak 16.7. *Kako se ta koncentracija glukoze u krvi može zaista pojaviti pri mjerenjima, taj podatak nećemo tretirati kao stršeću vrijednost.*

b) Kutijasti dijagram s označenim stršećim vrijednostima i deskriptivna statistika skupa izmjerenih vrijednosti varijable dob *prikazani su na slikama 3.53 i 3.54.*



Slika 3.53: Kutijasti dijagram na bazi medijana s prikazom stršećih vrijednosti varijable dob.

Variable	Descriptive Statistics (glukoza.sta)								
	Valid N	Mean	Median	Mode	Frequency of Mode	Minimum	Maximum	Lower Quartile	Upper Quartile
dob	102	66.73	56.00	Multiple	4.00	12.00	688.00	40.00	73.00

(a) uključene stršeće vrijednosti

Variable	Descriptive Statistics (glukoza.sta)								
	Valid N	Mean	Median	Mode	Frequency of Mode	Minimum	Maximum	Lower Quartile	Upper Quartile
dob	100	56.18	55.50	Multiple	4.00	12.00	101.00	40.00	71.50

(b) uklonjene stršeće vrijednosti

Slika 3.54: Deskriptivna statistika izmjerenih vrijednosti varijable dob.

Statistica je kao stršeće vrijednosti među izmjerenim vrijednostima varijable dob *detektirala podatke* 500 *i* 688. *Zaključujemo da uklanjanjem tih vrijednosti dolazi do smanjenja aritmetičke sredine i medijana izmjerenih vrijednosti.*

Zadatak 3.15. (komarci.sta)
Proučite bazu podataka komarci.sta koja je opisana u zadatku 2.4. Odredite tablicu i histogram frekvencija i relativnih frekvencija varijable brojM tako da za kategorije uzmete sve međusobno različite izmjerene vrijednosti te varijable. Zatim podijelite skup izmjerenih vrijednosti na određen broj disjunktnih intervala i ponovno odredite frekvencije i relativne frekvencije pojedinih kategorija (tj. intervala). Mijenjajte broj intervala, proučavajte što se događa i pribilježite svoj zaključak.

Zadatak 3.16. Koristeći javne izvore podataka ili podatke koje ste prikupljali u sklopu nekog istraživanja formirajte jednu bazu podataka koja će sadržavati najmanje dvije kvalitativne varijable, najmanje jednu diskretnu numeričku varijablu i jednu neprekidnu numeričku varijablu. Opišite o kakvom se istraživanju radi i zašto se mjere vrijednosti navedenih varijabli. Vodite računa da baza sadrži što više jedinki. Navedite točan izvor podataka. Iskoristite prethodno opisane postupke i pojmove te opišite svoju bazu podataka.

This example suggests the need for what follows in Sections 1.3 and 1.4, namely, descriptive statistics that indicate measures of center of location in a set of data, and those that measure variability.

1.3 Measures of Location: The Sample Mean and Median

Measures of location are designed to provide the analyst with some quantitative values of where the center, or some other location, of data is located. In Example 1.2, it appears as if the center of the nitrogen sample clearly exceeds that of the no-nitrogen sample. One obvious and very useful measure is the **sample mean**. The mean is simply a numerical average.

Definition 1.1:

Suppose that the observations in a sample are $x_1, x_2, \ldots, x_n$. The **sample mean**, denoted by $\bar{x}$, is

$$\bar{x} = \sum_{i=1}^{n} \frac{x_i}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n}.$$

There are other measures of central tendency that are discussed in detail in future chapters. One important measure is the **sample median**. The purpose of the sample median is to reflect the central tendency of the sample in such a way that it is uninfluenced by extreme values or outliers.

Definition 1.2:

Given that the observations in a sample are $x_1, x_2, \ldots, x_n$, arranged in **increasing order** of magnitude, the sample median is

$$\tilde{x} = \begin{cases} x_{(n+1)/2}, & \text{if } n \text{ is odd,} \\ \frac{1}{2}(x_{n/2} + x_{n/2+1}), & \text{if } n \text{ is even.} \end{cases}$$

As an example, suppose the data set is the following: 1.7, 2.2, 3.9, 3.11, and 14.7. The sample mean and median are, respectively,

$$\bar{x} = 5.12, \quad \tilde{x} = 3.9.$$

Clearly, the mean is influenced considerably by the presence of the extreme observation, 14.7, whereas the median places emphasis on the true "center" of the data set. In the case of the two-sample data set of Example 1.2, the two measures of central tendency for the individual samples are

$$\begin{aligned} \bar{x} \text{ (no nitrogen)} &= 0.399 \text{ gram}, \\ \tilde{x} \text{ (no nitrogen)} &= \frac{0.38 + 0.42}{2} = 0.400 \text{ gram}, \\ \bar{x} \text{ (nitrogen)} &= 0.565 \text{ gram}, \\ \tilde{x} \text{ (nitrogen)} &= \frac{0.49 + 0.52}{2} = 0.505 \text{ gram}. \end{aligned}$$

Clearly there is a difference in concept between the mean and median. It may be of interest to the reader with an engineering background that the sample mean

is the **centroid of the data** in a sample. In a sense, it is the point at which a fulcrum can be placed to balance a system of "weights" which are the locations of the individual data. This is shown in Figure 1.4 with regard to the with-nitrogen sample.

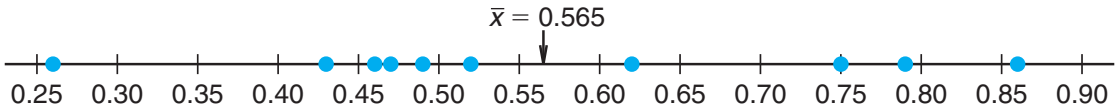


Figure 1.4: Sample mean as a centroid of the with-nitrogen stem weight.

In future chapters, the basis for the computation of $\bar{x}$ is that of an **estimate** of the **population mean**. As we indicated earlier, the purpose of statistical inference is to draw conclusions about population characteristics or **parameters** and **estimation** is a very important feature of statistical inference.

The median and mean can be quite different from each other. Note, however, that in the case of the stem weight data the sample mean value for no-nitrogen is quite similar to the median value.

Other Measures of Locations

There are several other methods of quantifying the center of location of the data in the sample. We will not deal with them at this point. For the most part, alternatives to the sample mean are designed to produce values that represent compromises between the mean and the median. Rarely do we make use of these other measures. However, it is instructive to discuss one class of estimators, namely the class of **trimmed means**. A trimmed mean is computed by "trimming away" a certain percent of both the largest and the smallest set of values. For example, the 10% trimmed mean is found by eliminating the largest 10% and smallest 10% and computing the average of the remaining values. For example, in the case of the stem weight data, we would eliminate the largest and smallest since the sample size is 10 for each sample. So for the without-nitrogen group the 10% trimmed mean is given by

$$\bar{x}_{\text{tr}(10)} = \frac{0.32 + 0.37 + 0.47 + 0.43 + 0.36 + 0.42 + 0.38 + 0.43}{8} = 0.39750,$$

and for the 10% trimmed mean for the with-nitrogen group we have

$$\bar{x}_{\text{tr}(10)} = \frac{0.43 + 0.47 + 0.49 + 0.52 + 0.75 + 0.79 + 0.62 + 0.46}{8} = 0.56625.$$

Note that in this case, as expected, the trimmed means are close to both the mean and the median for the individual samples. The trimmed mean is, of course, more insensitive to outliers than the sample mean but not as insensitive as the median. On the other hand, the trimmed mean approach makes use of more information than the sample median. Note that the sample median is, indeed, a special case of the trimmed mean in which all of the sample data are eliminated apart from the middle one or two observations.

Exercises

1.1 The following measurements were recorded for the drying time, in hours, of a certain brand of latex paint.

3.4	2.5	4.8	2.9	3.6
2.8	3.3	5.6	3.7	2.8
4.4	4.0	5.2	3.0	4.8

Assume that the measurements are a simple random sample.

(a) What is the sample size for the above sample?

(b) Calculate the sample mean for these data.

(c) Calculate the sample median.

(d) Plot the data by way of a dot plot.

(e) Compute the 20% trimmed mean for the above data set.

(f) Is the sample mean for these data more or less descriptive as a center of location than the trimmed mean?

1.2 According to the journal *Chemical Engineering*, an important property of a fiber is its water absorbency. A random sample of 20 pieces of cotton fiber was taken and the absorbency on each piece was measured. The following are the absorbency values:

18.71	21.41	20.72	21.81	19.29	22.43	20.17
23.71	19.44	20.50	18.92	20.33	23.00	22.85
19.25	21.77	22.11	19.77	18.04	21.12	

(a) Calculate the sample mean and median for the above sample values.

(b) Compute the 10% trimmed mean.

(c) Do a dot plot of the absorbency data.

(d) Using only the values of the mean, median, and trimmed mean, do you have evidence of outliers in the data?

1.3 A certain polymer is used for evacuation systems for aircraft. It is important that the polymer be resistant to the aging process. Twenty specimens of the polymer were used in an experiment. Ten were assigned randomly to be exposed to an accelerated batch aging process that involved exposure to high temperatures for 10 days. Measurements of tensile strength of the specimens were made, and the following data were recorded on tensile strength in psi:

No aging:	227	222	218	217	225
	218	216	229	228	221
Aging:	219	214	215	211	209
	218	203	204	201	205

(a) Do a dot plot of the data.

(b) From your plot, does it appear as if the aging process has had an effect on the tensile strength of this polymer? Explain.

(c) Calculate the sample mean tensile strength of the two samples.

(d) Calculate the median for both. Discuss the similarity or lack of similarity between the mean and median of each group.

1.4 In a study conducted by the Department of Mechanical Engineering at Virginia Tech, the steel rods supplied by two different companies were compared. Ten sample springs were made out of the steel rods supplied by each company, and a measure of flexibility was recorded for each. The data are as follows:

Company A:	9.3	8.8	6.8	8.7	8.5
	6.7	8.0	6.5	9.2	7.0
Company B:	11.0	9.8	9.9	10.2	10.1
	9.7	11.0	11.1	10.2	9.6

(a) Calculate the sample mean and median for the data for the two companies.

(b) Plot the data for the two companies on the same line and give your impression regarding any apparent differences between the two companies.

1.5 Twenty adult males between the ages of 30 and 40 participated in a study to evaluate the effect of a specific health regimen involving diet and exercise on the blood cholesterol. Ten were randomly selected to be a control group, and ten others were assigned to take part in the regimen as the treatment group for a period of 6 months. The following data show the reduction in cholesterol experienced for the time period for the 20 subjects:

Control group:	7	3	−4	14	2
	5	22	−7	9	5
Treatment group:	−6	5	9	4	4
	12	37	5	3	3

(a) Do a dot plot of the data for both groups on the same graph.

(b) Compute the mean, median, and 10% trimmed mean for both groups.

(c) Explain why the difference in means suggests one conclusion about the effect of the regimen, while the difference in medians or trimmed means suggests a different conclusion.

1.6 The tensile strength of silicone rubber is thought to be a function of curing temperature. A study was carried out in which samples of 12 specimens of the rubber were prepared using curing temperatures of 20°C and 45°C. The data below show the tensile strength values in megapascals.

20°C:	2.07	2.14	2.22	2.03	2.21	2.03
	2.05	2.18	2.09	2.14	2.11	2.02
45°C:	2.52	2.15	2.49	2.03	2.37	2.05
	1.99	2.42	2.08	2.42	2.29	2.01

(a) Show a dot plot of the data with both low and high temperature tensile strength values.

(b) Compute sample mean tensile strength for both samples.

(c) Does it appear as if curing temperature has an influence on tensile strength, based on the plot? Comment further.

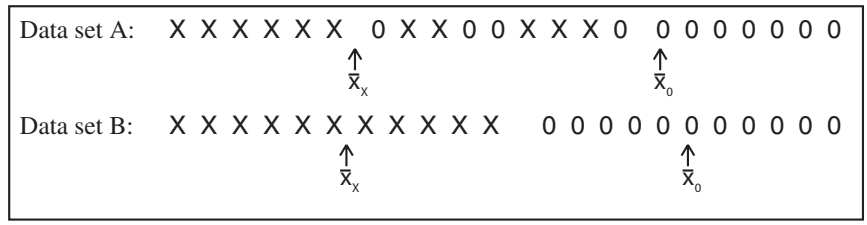
(d) Does anything else appear to be influenced by an increase in curing temperature? Explain.

1.4 Measures of Variability

Sample variability plays an important role in data analysis. Process and product variability is a fact of life in engineering and scientific systems: The control or reduction of process variability is often a source of major difficulty. More and more process engineers and managers are learning that product quality and, as a result, profits derived from manufactured products are very much a function of **process variability**. As a result, much of Chapters 9 through 15 deals with data analysis and modeling procedures in which sample variability plays a major role. Even in small data analysis problems, the success of a particular statistical method may depend on the magnitude of the variability among the observations in the sample. Measures of location in a sample do not provide a proper summary of the nature of a data set. For instance, in Example 1.2 we cannot conclude that the use of nitrogen enhances growth without taking sample variability into account.

While the details of the analysis of this type of data set are deferred to Chapter 9, it should be clear from Figure 1.1 that variability among the no-nitrogen observations and variability among the nitrogen observations are certainly of some consequence. In fact, it appears that the variability within the nitrogen sample is larger than that of the no-nitrogen sample. Perhaps there is something about the inclusion of nitrogen that not only increases the stem height ($\bar{x}$ of 0.565 gram compared to an $\bar{x}$ of 0.399 gram for the no-nitrogen sample) but also increases the variability in stem height (i.e., renders the stem height more inconsistent).

As another example, contrast the two data sets below. Each contains two samples and the difference in the means is roughly the same for the two samples, but data set B seems to provide a much sharper contrast between the two populations from which the samples were taken. If the purpose of such an experiment is to detect differences between the two populations, the task is accomplished in the case of data set B. However, in data set A the large variability *within* the two samples creates difficulty. In fact, it is not clear that there is a distinction *between* the two populations.



Sample Range and Sample Standard Deviation

Just as there are many measures of central tendency or location, there are many measures of spread or variability. Perhaps the simplest one is the **sample range** $X_{max} - X_{min}$. The range can be very useful and is discussed at length in Chapter 17 on *statistical quality control*. The sample measure of spread that is used most often is the **sample standard deviation**. We again let $x_1, x_2, \ldots, x_n$ denote sample values.

Definition 1.3:

The **sample variance**, denoted by s^2, is given by

$$s^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}.$$

The **sample standard deviation**, denoted by s, is the positive square root of s^2, that is,

$$s = \sqrt{s^2}.$$

It should be clear to the reader that the sample standard deviation is, in fact, a measure of variability. Large variability in a data set produces relatively large values of $(x - \bar{x})^2$ and thus a large sample variance. The quantity $n - 1$ is often called the **degrees of freedom associated with the variance** estimate. In this simple example, the degrees of freedom depict the number of independent pieces of information available for computing variability. For example, suppose that we wish to compute the sample variance and standard deviation of the data set (5, 17, 6, 4). The sample average is $\bar{x} = 8$. The computation of the variance involves

$$(5-8)^2 + (17-8)^2 + (6-8)^2 + (4-8)^2 = (-3)^2 + 9^2 + (-2)^2 + (-4)^2.$$

The quantities inside parentheses sum to zero. In general, $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$ (see Exercise 1.16 on page 31). Then the computation of a sample variance does not involve n **independent squared deviations** from the mean $\bar{x}$. In fact, since the last value of $x - \bar{x}$ is determined by the initial $n - 1$ of them, we say that these are $n - 1$ "pieces of information" that produce s^2. Thus, there are $n - 1$ degrees of freedom rather than n degrees of freedom for computing a sample variance.

Example 1.4: In an example discussed extensively in Chapter 10, an engineer is interested in testing the "bias" in a pH meter. Data are collected on the meter by measuring the pH of a neutral substance (pH = 7.0). A sample of size 10 is taken, with results given by

7.07 7.00 7.10 6.97 7.00 7.03 7.01 7.01 6.98 7.08.

The sample mean $\bar{x}$ is given by

$$\bar{x} = \frac{7.07 + 7.00 + 7.10 + \cdots + 7.08}{10} = 7.0250.$$

The sample variance s^2 is given by

$$s^2 = \frac{1}{9}[(7.07 - 7.025)^2 + (7.00 - 7.025)^2 + (7.10 - 7.025)^2 + \cdots + (7.08 - 7.025)^2] = 0.001939.$$

As a result, the sample standard deviation is given by

$$s = \sqrt{0.001939} = 0.044.$$

So the sample standard deviation is 0.0440 with $n - 1 = 9$ degrees of freedom.

Units for Standard Deviation and Variance

It should be apparent from Definition 1.3 that the variance is a measure of the average squared deviation from the mean $\bar{x}$. We use the term *average squared deviation* even though the definition makes use of a division by degrees of freedom $n - 1$ rather than n. Of course, if n is large, the difference in the denominator is inconsequential. As a result, the sample variance possesses units that are the square of the units in the observed data whereas the sample standard deviation is found in linear units. As an example, consider the data of Example 1.2. The stem weights are measured in grams. As a result, the sample standard deviations are in grams and the variances are measured in grams^2. In fact, the individual standard deviations are 0.0728 gram for the no-nitrogen case and 0.1867 gram for the nitrogen group. Note that the standard deviation does indicate considerably larger variability in the nitrogen sample. This condition was displayed in Figure 1.1.

Which Variability Measure Is More Important?

As we indicated earlier, the sample range has applications in the area of statistical quality control. It may appear to the reader that the use of both the sample variance and the sample standard deviation is redundant. Both measures reflect the same concept in measuring variability, but the sample standard deviation measures variability in linear units whereas the sample variance is measured in squared units. Both play huge roles in the use of statistical methods. Much of what is accomplished in the context of statistical inference involves drawing conclusions about characteristics of populations. Among these characteristics are constants which are called **population parameters**. Two important parameters are the **population mean** and the **population variance**. The sample variance plays an explicit role in the statistical methods used to draw inferences about the population variance. The sample standard deviation has an important role along with the sample mean in inferences that are made about the population mean. In general, the variance is considered more in inferential theory, while the standard deviation is used more in applications.

Exercises

1.7 Consider the drying time data for Exercise 1.1 on page 13. Compute the sample variance and sample standard deviation.

1.8 Compute the sample variance and standard deviation for the water absorbency data of Exercise 1.2 on page 13.

1.9 Exercise 1.3 on page 13 showed tensile strength data for two samples, one in which specimens were exposed to an aging process and one in which there was no aging of the specimens.

(a) Calculate the sample variance as well as standard deviation in tensile strength for both samples.

(b) Does there appear to be any evidence that aging affects the variability in tensile strength? (See also the plot for Exercise 1.3 on page 13.)

1.10 For the data of Exercise 1.4 on page 13, compute both the mean and the variance in "flexibility" for both company A and company B. Does there appear to be a difference in flexibility between company A and company B?

1.11 Consider the data in Exercise 1.5 on page 13. Compute the sample variance and the sample standard deviation for both control and treatment groups.

1.12 For Exercise 1.6 on page 13, compute the sample standard deviation in tensile strength for the samples separately for the two temperatures. Does it appear as if an increase in temperature influences the variability in tensile strength? Explain.

1.5 Discrete and Continuous Data

Statistical inference through the analysis of observational studies or designed experiments is used in many scientific areas. The data gathered may be **discrete** or **continuous**, depending on the area of application. For example, a chemical engineer may be interested in conducting an experiment that will lead to conditions where yield is maximized. Here, of course, the yield may be in percent or grams/pound, measured on a continuum. On the other hand, a toxicologist conducting a combination drug experiment may encounter data that are binary in nature (i.e., the patient either responds or does not).

Great distinctions are made between discrete and continuous data in the probability theory that allow us to draw statistical inferences. Often applications of statistical inference are found when the data are *count data*. For example, an engineer may be interested in studying the number of radioactive particles passing through a counter in, say, 1 millisecond. Personnel responsible for the efficiency of a port facility may be interested in the properties of the number of oil tankers arriving each day at a certain port city. In Chapter 5, several distinct scenarios, leading to different ways of handling data, are discussed for situations with count data.

Special attention even at this early stage of the textbook should be paid to some details associated with binary data. Applications requiring statistical analysis of binary data are voluminous. Often the measure that is used in the analysis is the *sample proportion*. Obviously the binary situation involves two categories. If there are n units involved in the data and x is defined as the number that fall into category 1, then $n - x$ fall into category 2. Thus, x/n is the sample proportion in category 1, and $1 - x/n$ is the sample proportion in category 2. In the biomedical application, 50 patients may represent the sample units, and if 20 out of 50 experienced an improvement in a stomach ailment (common to all 50) after all were given the drug, then $\frac{20}{50} = 0.4$ is the sample proportion for which

the drug was a success and $1 - 0.4 = 0.6$ is the sample proportion for which the drug was not successful. Actually the basic numerical measurement for binary data is generally denoted by either 0 or 1. For example, in our medical example, a successful result is denoted by a 1 and a nonsuccess a 0. As a result, the sample proportion is actually a sample mean of the ones and zeros. For the successful category,

$$\frac{x_1 + x_2 + \cdots + x_{50}}{50} = \frac{1 + 1 + 0 + \cdots + 0 + 1}{50} = \frac{20}{50} = 0.4.$$

What Kinds of Problems Are Solved in Binary Data Situations?

The kinds of problems facing scientists and engineers dealing in binary data are not a great deal unlike those seen where continuous measurements are of interest. However, different techniques are used since the statistical properties of sample proportions are quite different from those of the sample means that result from averages taken from continuous populations. Consider the example data in Exercise 1.6 on page 13. The statistical problem underlying this illustration focuses on whether an intervention, say, an increase in curing temperature, will alter the population mean tensile strength associated with the silicone rubber process. On the other hand, in a quality control area, suppose an automobile tire manufacturer reports that a shipment of 5000 tires selected randomly from the process results in 100 of them showing blemishes. Here the sample proportion is $\frac{100}{5000} = 0.02$. Following a change in the process designed to reduce blemishes, a second sample of 5000 is taken and 90 tires are blemished. The sample proportion has been reduced to $\frac{90}{5000} = 0.018$. The question arises, "Is the decrease in the sample proportion from 0.02 to 0.018 substantial enough to suggest a real improvement in the population proportion?" Both of these illustrations require the use of the statistical properties of sample averages—one from samples from a continuous population, and the other from samples from a discrete (binary) population. In both cases, the sample mean is an **estimate** of a population parameter, a population mean in the first illustration (i.e., mean tensile strength), and a population proportion in the second case (i.e., proportion of blemished tires in the population). So here we have sample estimates used to draw scientific conclusions regarding population parameters. As we indicated in Section 1.3, this is the general theme in many practical problems using statistical inference.

1.6 Statistical Modeling, Scientific Inspection, and Graphical Diagnostics

Often the end result of a statistical analysis is the estimation of parameters of a **postulated model**. This is natural for scientists and engineers since they often deal in modeling. A statistical model is not deterministic but, rather, must entail some probabilistic aspects. A model form is often the foundation of **assumptions** that are made by the analyst. For example, in Example 1.2 the scientist may wish to draw some level of distinction between the nitrogen and no-nitrogen populations through the sample information. The analysis may require a certain model for

the data, for example, that the two samples come from **normal** or **Gaussian distributions**. See Chapter 6 for a discussion of the normal distribution.

Obviously, the user of statistical methods cannot generate sufficient information or experimental data to characterize the population totally. But sets of data are often used to learn about certain properties of the population. Scientists and engineers are accustomed to dealing with data sets. The importance of characterizing or *summarizing* the nature of collections of data should be obvious. Often a summary of a collection of data via a graphical display can provide insight regarding the system from which the data were taken. For instance, in Sections 1.1 and 1.3, we have shown dot plots.

In this section, the role of sampling and the display of data for enhancement of **statistical inference** is explored in detail. We merely introduce some simple but often effective displays that complement the study of statistical populations.

Scatter Plot

At times the model postulated may take on a somewhat complicated form. Consider, for example, a textile manufacturer who designs an experiment where cloth specimen that contain various percentages of cotton are produced. Consider the data in Table 1.3.

Table 1.3: Tensile Strength

Cotton Percentage	Tensile Strength
15	7, 7, 9, 8, 10
20	19, 20, 21, 20, 22
25	21, 21, 17, 19, 20
30	8, 7, 8, 9, 10

Five cloth specimens are manufactured for each of the four cotton percentages. In this case, both the model for the experiment and the type of analysis used should take into account the goal of the experiment and important input from the textile scientist. Some simple graphics can shed important light on the clear distinction between the samples. See Figure 1.5; the sample means and variability are depicted nicely in the scatter plot. One possible goal of this experiment is simply to determine which cotton percentages are truly distinct from the others. In other words, as in the case of the nitrogen/no-nitrogen data, for which cotton percentages are there clear distinctions between the populations or, more specifically, between the population means? In this case, perhaps a reasonable model is that each sample comes from a normal distribution. Here the goal is very much like that of the nitrogen/no-nitrogen data except that more samples are involved. The formalism of the analysis involves notions of hypothesis testing discussed in Chapter 10. Incidentally, this formality is perhaps not necessary in light of the diagnostic plot. But does this describe the real goal of the experiment and hence the proper approach to data analysis? It is likely that the scientist anticipates the existence of a *maximum population mean tensile strength* in the range of cotton concentration in the experiment. Here the analysis of the data should revolve

around a different type of model, one that postulates a type of structure relating the population mean tensile strength to the cotton concentration. In other words, a model may be written

$$\mu_{t,c} = \beta_0 + \beta_1 C + \beta_2 C^2,$$

where $\mu_{t,c}$ is the population mean tensile strength, which varies with the amount of cotton in the product C. The implication of this model is that for a fixed cotton level, there is a population of tensile strength measurements and the population mean is $\mu_{t,c}$. This type of model, called a **regression model**, is discussed in Chapters 11 and 12. The functional form is chosen by the scientist. At times the data analysis may suggest that the model be changed. Then the data analyst "entertains" a model that may be altered after some analysis is done. The use of an empirical model is accompanied by **estimation theory**, where β_0, β_1, and β_2 are estimated by the data. Further, statistical inference can then be used to determine model adequacy.

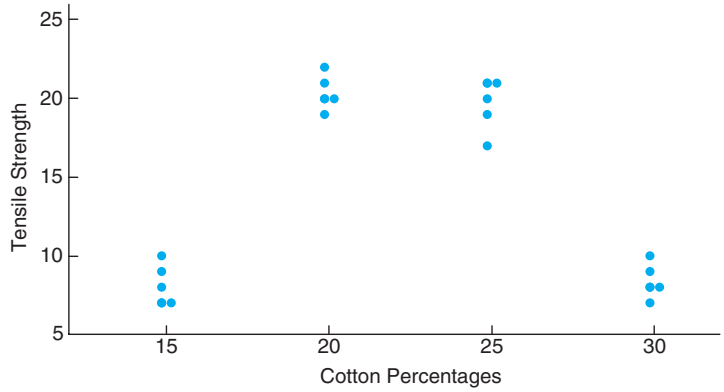


Figure 1.5: Scatter plot of tensile strength and cotton percentages.

Two points become evident from the two data illustrations here: (1) The type of model used to describe the data often depends on the goal of the experiment; and (2) the structure of the model should take advantage of nonstatistical scientific input. A selection of a model represents a **fundamental assumption** upon which the resulting statistical inference is based. It will become apparent throughout the book how important graphics can be. Often, plots can illustrate information that allows the results of the formal statistical inference to be better communicated to the scientist or engineer. At times, plots or **exploratory data analysis** can teach the analyst something not retrieved from the formal analysis. Almost any formal analysis requires assumptions that evolve from the model of the data. Graphics can nicely highlight **violation of assumptions** that would otherwise go unnoticed. Throughout the book, graphics are used extensively to supplement formal data analysis. The following sections reveal some graphical tools that are useful in exploratory or descriptive data analysis.

Stem-and-Leaf Plot

Statistical data, generated in large masses, can be very useful for studying the behavior of the distribution if presented in a combined tabular and graphic display called a **stem-and-leaf plot**.

To illustrate the construction of a stem-and-leaf plot, consider the data of Table 1.4, which specifies the "life" of 40 similar car batteries recorded to the nearest tenth of a year. The batteries are guaranteed to last 3 years. First, split each observation into two parts consisting of a stem and a leaf such that the stem represents the digit preceding the decimal and the leaf corresponds to the decimal part of the number. In other words, for the number 3.7, the digit 3 is designated the stem and the digit 7 is the leaf. The four stems 1, 2, 3, and 4 for our data are listed vertically on the left side in Table 1.5; the leaves are recorded on the right side opposite the appropriate stem value. Thus, the leaf 6 of the number 1.6 is recorded opposite the stem 1; the leaf 5 of the number 2.5 is recorded opposite the stem 2; and so forth. The number of leaves recorded opposite each stem is summarized under the frequency column.

Table 1.4: Car Battery Life

2.2	4.1	3.5	4.5	3.2	3.7	3.0	2.6
3.4	1.6	3.1	3.3	3.8	3.1	4.7	3.7
2.5	4.3	3.4	3.6	2.9	3.3	3.9	3.1
3.3	3.1	3.7	4.4	3.2	4.1	1.9	3.4
4.7	3.8	3.2	2.6	3.9	3.0	4.2	3.5

Table 1.5: Stem-and-Leaf Plot of Battery Life

Stem	**Leaf**	**Frequency**
1	69	2
2	25669	5
3	00111122233344455677889 9	25
4	11234577	8

The stem-and-leaf plot of Table 1.5 contains only four stems and consequently does not provide an adequate picture of the distribution. To remedy this problem, we need to increase the number of stems in our plot. One simple way to accomplish this is to write each stem value twice and then record the leaves 0, 1, 2, 3, and 4 opposite the appropriate stem value where it appears for the first time, and the leaves 5, 6, 7, 8, and 9 opposite this same stem value where it appears for the second time. This modified double-stem-and-leaf plot is illustrated in Table 1.6, where the stems corresponding to leaves 0 through 4 have been coded by the symbol $\star$ and the stems corresponding to leaves 5 through 9 by the symbol $\cdot$.

In any given problem, we must decide on the appropriate stem values. This decision is made somewhat arbitrarily, although we are guided by the size of our sample. Usually, we choose between 5 and 20 stems. The smaller the number of data available, the smaller is our choice for the number of stems. For example, if

the data consist of numbers from 1 to 21 representing the number of people in a cafeteria line on 40 randomly selected workdays and we choose a double-stem-and-leaf plot, the stems will be 0⋆, 0·, 1⋆, 1·, and 2⋆ so that the smallest observation 1 has stem 0⋆ and leaf 1, the number 18 has stem 1· and leaf 8, and the largest observation 21 has stem 2⋆ and leaf 1. On the other hand, if the data consist of numbers from \$18,800 to \$19,600 representing the best possible deals on 100 new automobiles from a certain dealership and we choose a single-stem-and-leaf plot, the stems will be 188, 189, 190, ..., 196 and the leaves will now each contain two digits. A car that sold for \$19,385 would have a stem value of 193 and the two-digit leaf 85. Multiple-digit leaves belonging to the same stem are usually separated by commas in the stem-and-leaf plot. Decimal points in the data are generally ignored when all the digits to the right of the decimal represent the leaf. Such was the case in Tables 1.5 and 1.6. However, if the data consist of numbers ranging from 21.8 to 74.9, we might choose the digits 2, 3, 4, 5, 6, and 7 as our stems so that a number such as 48.3 would have a stem value of 4 and a leaf of 8.3.

Table 1.6: Double-Stem-and-Leaf Plot of Battery Life

Stem	Leaf	Frequency
1·	69	2
2⋆	2	1
2·	5669	4
3⋆	001111222333444	15
3·	5567778899	10
4⋆	11234	5
4·	577	3

The stem-and-leaf plot represents an effective way to summarize data. Another way is through the use of the **frequency distribution**, where the data, grouped into different classes or intervals, can be constructed by counting the leaves belonging to each stem and noting that each stem defines a class interval. In Table 1.5, the stem 1 with 2 leaves defines the interval 1.0–1.9 containing 2 observations; the stem 2 with 5 leaves defines the interval 2.0–2.9 containing 5 observations; the stem 3 with 25 leaves defines the interval 3.0–3.9 with 25 observations; and the stem 4 with 8 leaves defines the interval 4.0–4.9 containing 8 observations. For the double-stem-and-leaf plot of Table 1.6, the stems define the seven class intervals 1.5–1.9, 2.0–2.4, 2.5–2.9, 3.0–3.4, 3.5–3.9, 4.0–4.4, and 4.5–4.9 with frequencies 2, 1, 4, 15, 10, 5, and 3, respectively.

Histogram

Dividing each class frequency by the total number of observations, we obtain the proportion of the set of observations in each of the classes. A table listing relative frequencies is called a **relative frequency distribution**. The relative frequency distribution for the data of Table 1.4, showing the midpoint of each class interval, is given in Table 1.7.

The information provided by a relative frequency distribution in tabular form is easier to grasp if presented graphically. Using the midpoint of each interval and the

Table 1.7: Relative Frequency Distribution of Battery Life

Class Interval	Class Midpoint	Frequency, f	Relative Frequency
1.5–1.9	1.7	2	0.050
2.0–2.4	2.2	1	0.025
2.5–2.9	2.7	4	0.100
3.0–3.4	3.2	15	0.375
3.5–3.9	3.7	10	0.250
4.0–4.4	4.2	5	0.125
4.5–4.9	4.7	3	0.075

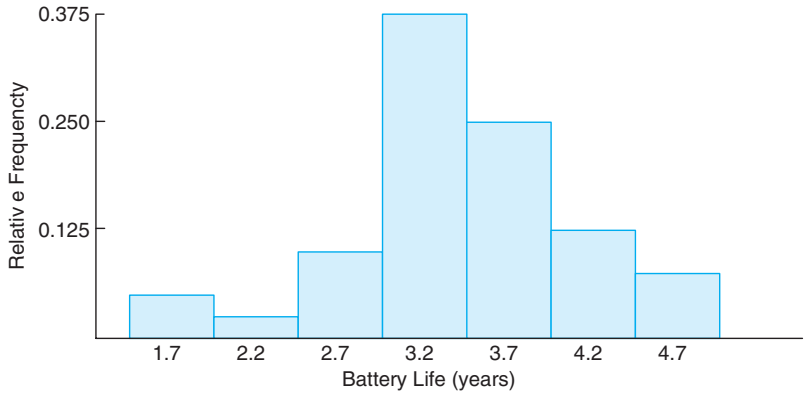


Figure 1.6: Relative frequency histogram.

corresponding relative frequency, we construct a **relative frequency histogram** (Figure 1.6).

Many continuous frequency distributions can be represented graphically by the characteristic bell-shaped curve of Figure 1.7. Graphical tools such as what we see in Figures 1.6 and 1.7 aid in the characterization of the nature of the population. In Chapters 5 and 6 we discuss a property of the population called its **distribution**. While a more rigorous definition of a distribution or **probability distribution** will be given later in the text, at this point one can view it as what would be seen in Figure 1.7 in the limit as the size of the sample becomes larger.

A distribution is said to be **symmetric** if it can be folded along a vertical axis so that the two sides coincide. A distribution that lacks symmetry with respect to a vertical axis is said to be **skewed**. The distribution illustrated in Figure 1.8(a) is said to be skewed to the right since it has a long right tail and a much shorter left tail. In Figure 1.8(b) we see that the distribution is symmetric, while in Figure 1.8(c) it is skewed to the left.

If we rotate a stem-and-leaf plot counterclockwise through an angle of 90°, we observe that the resulting columns of leaves form a picture that is similar to a histogram. Consequently, if our primary purpose in looking at the data is to determine the general shape or form of the distribution, it will seldom be necessary

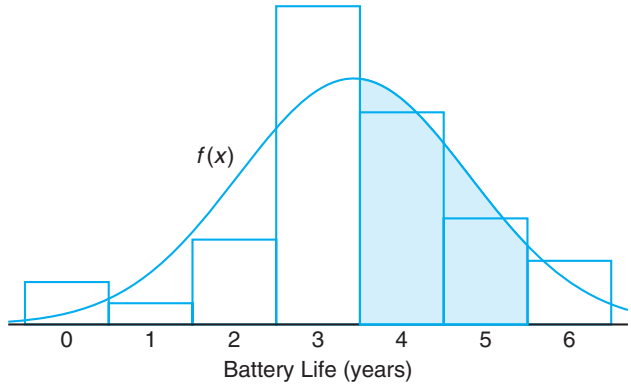


Figure 1.7: Estimating frequency distribution.

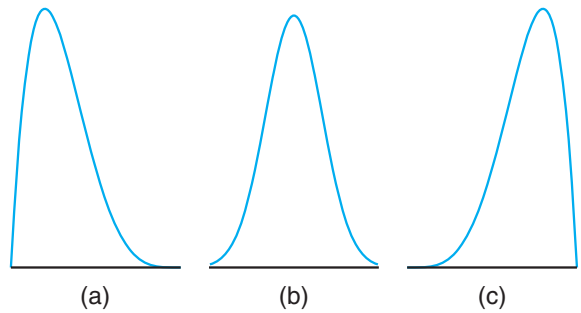


Figure 1.8: Skewness of data.

to construct a relative frequency histogram.

Box-and-Whisker Plot or Box Plot

Another display that is helpful for reflecting properties of a sample is the **box-and-whisker plot**. This plot encloses the *interquartile range* of the data in a box that has the median displayed within. The interquartile range has as its extremes the 75th percentile (upper quartile) and the 25th percentile (lower quartile). In addition to the box, "whiskers" extend, showing extreme observations in the sample. For reasonably large samples, the display shows center of location, variability, and the degree of asymmetry.

In addition, a variation called a **box plot** can provide the viewer with information regarding which observations may be **outliers**. Outliers are observations that are considered to be unusually far from the bulk of the data. There are many statistical tests that are designed to detect outliers. Technically, one may view an outlier as being an observation that represents a "rare event" (there is a small probability of obtaining a value that far from the bulk of the data). The concept of outliers resurfaces in Chapter 12 in the context of regression analysis.

The visual information in the box-and-whisker plot or box plot is not intended to be a formal test for outliers. Rather, it is viewed as a diagnostic tool. While the determination of which observations are outliers varies with the type of software that is used, one common procedure is to use a **multiple of the interquartile range**. For example, if the distance from the box exceeds 1.5 times the interquartile range (in either direction), the observation may be labeled an outlier.

Example 1.5: Nicotine content was measured in a random sample of 40 cigarettes. The data are displayed in Table 1.8.

Table 1.8: Nicotine Data for Example 1.5

1.09	1.92	2.31	1.79	2.28	1.74	1.47	1.97
0.85	1.24	1.58	2.03	1.70	2.17	2.55	2.11
1.86	1.90	1.68	1.51	1.64	0.72	1.69	1.85
1.82	1.79	2.46	1.88	2.08	1.67	1.37	1.93
1.40	1.64	2.09	1.75	1.63	2.37	1.75	1.69

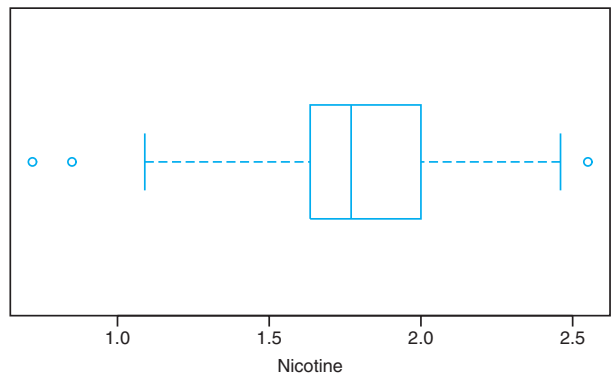


Figure 1.9: Box-and-whisker plot for Example 1.5.

Figure 1.9 shows the box-and-whisker plot of the data, depicting the observations 0.72 and 0.85 as mild outliers in the lower tail, whereas the observation 2.55 is a mild outlier in the upper tail. In this example, the interquartile range is 0.365, and 1.5 times the interquartile range is 0.5475. Figure 1.10, on the other hand, provides a stem-and-leaf plot. ◢

Example 1.6: Consider the data in Table 1.9, consisting of 30 samples measuring the thickness of paint can "ears" (see the work by Hogg and Ledolter, 1992, in the Bibliography). Figure 1.11 depicts a box-and-whisker plot for this asymmetric set of data. Notice that the left block is considerably larger than the block on the right. The median is 35. The lower quartile is 31, while the upper quartile is 36. Notice also that the extreme observation on the right is farther away from the box than the extreme observation on the left. There are no outliers in this data set. ◢

```
The decimal point is 1 digit(s) to the left of the |
 7 | 2
 8 | 5
 9 |
10 | 9
11 |
12 | 4
13 | 7
14 | 07
15 | 18
16 | 3447899
17 | 045599
18 | 2568
19 | 0237
20 | 389
21 | 17
22 | 8
23 | 17
24 | 6
25 | 5
```

Figure 1.10: Stem-and-leaf plot for the nicotine data.

Table 1.9: Data for Example 1.6

Sample	Measurements	Sample	Measurements
1	29 36 39 34 34	16	35 30 35 29 37
2	29 29 28 32 31	17	40 31 38 35 31
3	34 34 39 38 37	18	35 36 30 33 32
4	35 37 33 38 41	19	35 34 35 30 36
5	30 29 31 38 29	20	35 35 31 38 36
6	34 31 37 39 36	21	32 36 36 32 36
7	30 35 33 40 36	22	36 37 32 34 34
8	28 28 31 34 30	23	29 34 33 37 35
9	32 36 38 38 35	24	36 36 35 37 37
10	35 30 37 35 31	25	36 30 35 33 31
11	35 30 35 38 35	26	35 30 29 38 35
12	38 34 35 35 31	27	35 36 30 34 36
13	34 35 33 30 34	28	35 30 36 29 35
14	40 35 34 33 35	29	38 36 35 31 31
15	34 35 38 35 30	30	30 34 40 28 30

There are additional ways that box-and-whisker plots and other graphical displays can aid the analyst. Multiple samples can be compared graphically. Plots of data can suggest relationships between variables. Graphs can aid in the detection of anomalies or outlying observations in samples.

There are other types of graphical tools and plots that are used. These are discussed in Chapter 8 after we introduce additional theoretical details.

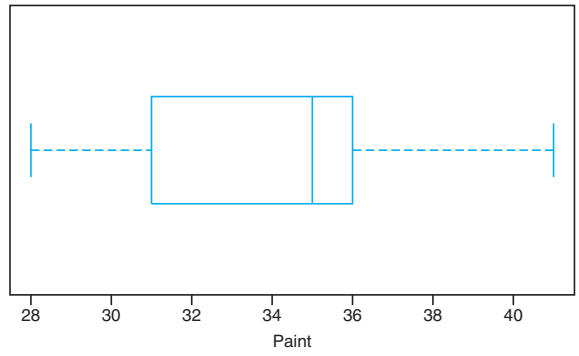


Figure 1.11: Box-and-whisker plot for thickness of paint can "ears."

Other Distinguishing Features of a Sample

There are features of the distribution or sample other than measures of center of location and variability that further define its nature. For example, while the median divides the data (or distribution) into two parts, there are other measures that divide parts or pieces of the distribution that can be very useful. Separation is made into four parts by *quartiles*, with the third quartile separating the upper quarter of the data from the rest, the second quartile being the median, and the first quartile separating the lower quarter of the data from the rest. The distribution can be even more finely divided by computing percentiles of the distribution. These quantities give the analyst a sense of the so-called *tails* of the distribution (i.e., values that are relatively extreme, either small or large). For example, the 95th percentile separates the highest 5% from the bottom 95%. Similar definitions prevail for extremes on the lower side or *lower tail* of the distribution. The 1st percentile separates the bottom 1% from the rest of the distribution. The concept of percentiles will play a major role in much that will be covered in future chapters.

1.7 General Types of Statistical Studies: Designed Experiment, Observational Study, and Retrospective Study

In the foregoing sections we have emphasized the notion of sampling from a population and the use of statistical methods to learn or perhaps affirm important information about the population. The information sought and learned through the use of these statistical methods can often be influential in decision making and problem solving in many important scientific and engineering areas. As an illustration, Example 1.3 describes a simple experiment in which the results may provide an aid in determining the kinds of conditions under which it is not advisable to use a particular aluminum alloy that may have a dangerous vulnerability to corrosion. The results may be of use not only to those who produce the alloy, but also to the customer who may consider using it. This illustration, as well as many more that appear in Chapters 13 through 15, highlights the concept of designing or controlling experimental conditions (combinations of coating conditions and humidity) of

These researchers designed a randomized experiment because they wanted to draw causal conclusions about the drug's effect. Study volunteers[17] were randomly placed into two study groups. One group, the **treatment group**, received the drug. The other group, called the **control group**, did not receive any drug treatment.

Put yourself in the place of a person in the study. If you are in the treatment group, you are given a fancy new drug that you anticipate will help you. On the other hand, a person in the other group doesn't receive the drug and sits idly, hoping her participation doesn't increase her risk of death. These perspectives suggest there are actually two effects: the one of interest is the effectiveness of the drug, and the second is an emotional effect that is difficult to quantify.

Researchers aren't usually interested in the emotional effect, which might bias the study. To circumvent this problem, researchers do not want patients to know which group they are in. When researchers keep the patients uninformed about their treatment, the study is said to be **blind**. But there is one problem: if a patient doesn't receive a treatment, she will know she is in the control group. The solution to this problem is to give fake treatments to patients in the control group. A fake treatment is called a **placebo**, and an effective placebo is the key to making a study truly blind. A classic example of a placebo is a sugar pill that is made to look like the actual treatment pill. Often times, a placebo results in a slight but real improvement in patients. This effect has been dubbed the **placebo effect**.

The patients are not the only ones who should be blinded: doctors and researchers can accidentally bias a study. When a doctor knows a patient has been given the real treatment, she might inadvertently give that patient more attention or care than a patient that she knows is on the placebo. To guard against this bias, which again has been found to have a measurable effect in some instances, most modern studies employ a **double-blind** setup where doctors or researchers who interact with patients are, just like the patients, unaware of who is or is not receiving the treatment.[18]

⊙ **Guided Practice 1.14** Look back to the study in Section 1.1 where researchers were testing whether stents were effective at reducing strokes in at-risk patients. Is this an experiment? Was the study blinded? Was it double-blinded?[19]

1.6 Examining numerical data

In this section we will be introduced to techniques for exploring and summarizing numerical variables. The `email50` and `county` data sets from Section 1.2 provide rich opportunities for examples. Recall that outcomes of numerical variables are numbers on which it is reasonable to perform basic arithmetic operations. For example, the `pop2010` variable, which represents the populations of counties in 2010, is numerical since we can sensibly discuss the difference or ratio of the populations in two counties. On the other hand, area codes and zip codes are not numerical, but rather they are categorical variables.

[17] Human subjects are often called **patients**, **volunteers**, or **study participants**.

[18] There are always some researchers involved in the study who do know which patients are receiving which treatment. However, they do not interact with the study's patients and do not tell the blinded health care professionals who is receiving which treatment.

[19] The researchers assigned the patients into their treatment groups, so this study was an experiment. However, the patients could distinguish what treatment they received, so this study was not blind. The study could not be double-blind since it was not blind.

1.6.1 Scatterplots for paired data

A **scatterplot** provides a case-by-case view of data for two numerical variables. In Figure 1.8 on page 13, a scatterplot was used to examine how federal spending and poverty were related in the `county` data set. Another scatterplot is shown in Figure 1.17, comparing the number of line breaks (`line_breaks`) and number of characters (`num_char`) in emails for the `email50` data set. In any scatterplot, each point represents a single case. Since there are 50 cases in `email50`, there are 50 points in Figure 1.17.

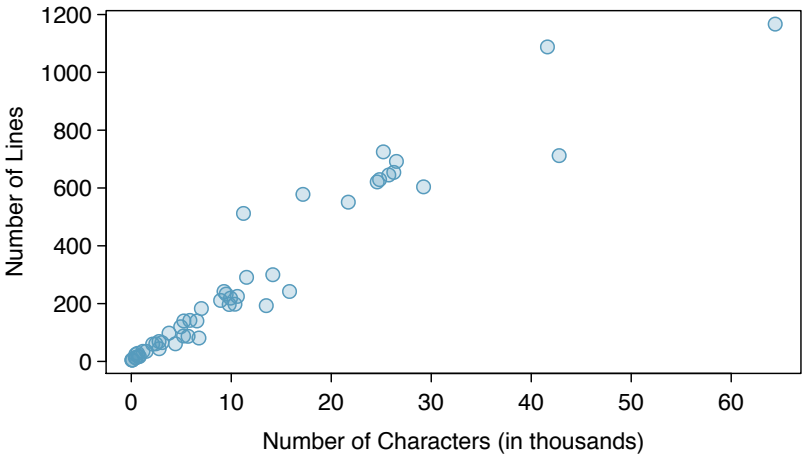


Figure 1.17: A scatterplot of `line_breaks` versus `num_char` for the `email50` data.

To put the number of characters in perspective, this paragraph has 363 characters. Looking at Figure 1.17, it seems that some emails are incredibly verbose! Upon further investigation, we would actually find that most of the long emails use the HTML format, which means most of the characters in those emails are used to format the email rather than provide text.

⊙ **Guided Practice 1.15** What do scatterplots reveal about the data, and how might they be useful?[20]

● **Example 1.16** Consider a new data set of 54 cars with two variables: vehicle price and weight.[21] A scatterplot of vehicle price versus weight is shown in Figure 1.18. What can be said about the relationship between these variables?

The relationship is evidently nonlinear, as highlighted by the dashed line. This is different from previous scatterplots we've seen, such as Figure 1.8 on page 13 and Figure 1.17, which show relationships that are very linear.

⊙ **Guided Practice 1.17** Describe two variables that would have a horseshoe shaped association in a scatterplot.[22]

[20] Answers may vary. Scatterplots are helpful in quickly spotting associations relating variables, whether those associations come in the form of simple trends or whether those relationships are more complex.

[21] Subset of data from www.amstat.org/publications/jse/v1n1/datasets.lock.html

[22] Consider the case where your vertical axis represents something "good" and your horizontal axis represents something that is only good in moderation. Health and water consumption fit this description since water becomes toxic when consumed in excessive quantities.

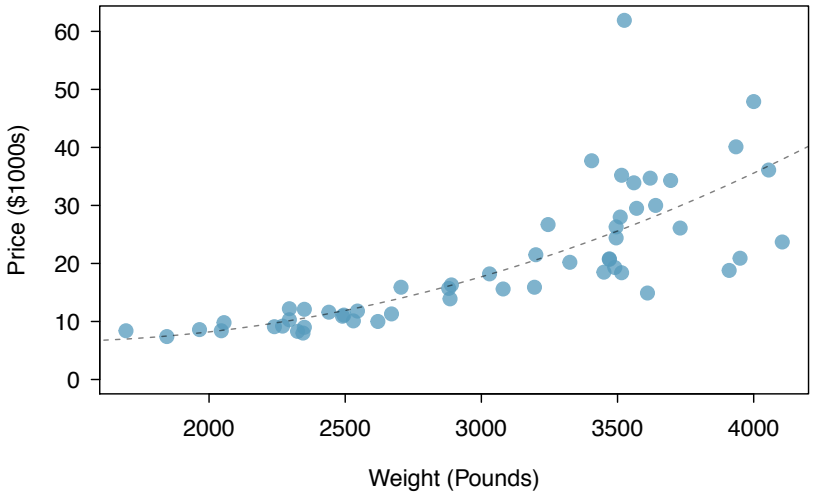


Figure 1.18: A scatterplot of `price` versus `weight` for 54 cars.

1.6.2 Dot plots and the mean

Sometimes two variables are one too many: only one variable may be of interest. In these cases, a dot plot provides the most basic of displays. A **dot plot** is a one-variable scatterplot; an example using the number of characters from 50 emails is shown in Figure 1.19. A stacked version of this dot plot is shown in Figure 1.20.

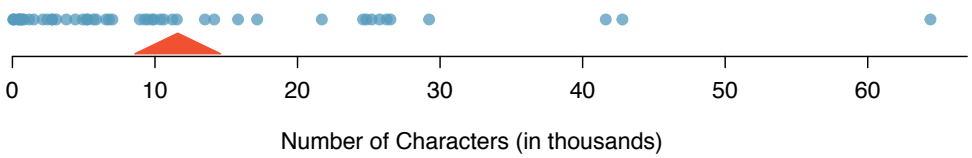


Figure 1.19: A dot plot of `num_char` for the `email50` data set.

The **mean**, sometimes called the average, is a common way to measure the center of a **distribution** of data. To find the mean number of characters in the 50 emails, we add up all the character counts and divide by the number of emails. For computational convenience, the number of characters is listed in the thousands and rounded to the first decimal.

$$\bar{x} = \frac{21.7 + 7.0 + \cdots + 15.8}{50} = 11.6 \tag{1.18}$$

$\bar{x}$
sample mean

The sample mean is often labeled $\bar{x}$. The letter x is being used as a generic placeholder for the variable of interest, `num_char`, and the bar over on the x communicates that the average number of characters in the 50 emails was 11,600. It is useful to think of the mean as the balancing point of the distribution. The sample mean is shown as a triangle in Figures 1.19 and 1.20.

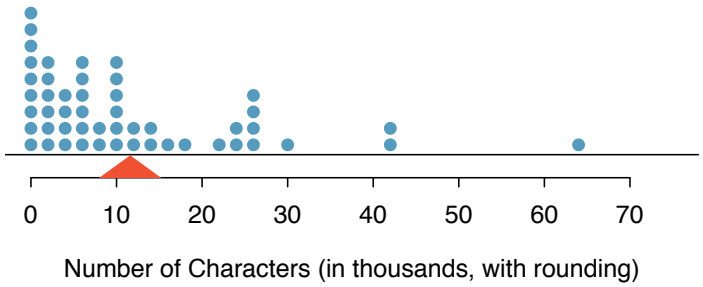


Figure 1.20: A stacked dot plot of `num_char` for the `email50` data set. The values have been rounded to the nearest 2,000 in this plot.

Mean

The sample mean of a numerical variable is computed as the sum of all of the observations divided by the number of observations:

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} \tag{1.19}$$

where $x_1, x_2, \ldots, x_n$ represent the n observed values.

n
sample size

Guided Practice 1.20 Examine Equations (1.18) and (1.19) above. What does x_1 correspond to? And x_2? Can you infer a general meaning to what x_i might represent?[23]

Guided Practice 1.21 What was n in this sample of emails?[24]

The `email50` data set represents a sample from a larger population of emails that were received in January and March. We could compute a mean for this population in the same way as the sample mean, however, the population mean has a special label: μ. The symbol μ is the Greek letter *mu* and represents the average of all observations in the population. Sometimes a subscript, such as $_x$, is used to represent which variable the population mean refers to, e.g. μ_x.

μ
population mean

Example 1.22 The average number of characters across all emails can be estimated using the sample data. Based on the sample of 50 emails, what would be a reasonable estimate of μ_x, the mean number of characters in all emails in the `email` data set? (Recall that `email50` is a sample from `email`.)

The sample mean, 11,600, may provide a reasonable estimate of μ_x. While this number will not be perfect, it provides a *point estimate* of the population mean. In Chapter 4 and beyond, we will develop tools to characterize the accuracy of point estimates, and we will find that point estimates based on larger samples tend to be more accurate than those based on smaller samples.

[23] x_1 corresponds to the number of characters in the first email in the sample (21.7, in thousands), x_2 to the number of characters in the second email (7.0, in thousands), and x_i corresponds to the number of characters in the i^{th} email in the data set.

[24] The sample size was $n = 50$.

Example 1.23 We might like to compute the average income per person in the US. To do so, we might first think to take the mean of the per capita incomes across the 3,143 counties in the `county` data set. What would be a better approach?

The `county` data set is special in that each county actually represents many individual people. If we were to simply average across the `income` variable, we would be treating counties with 5,000 and 5,000,000 residents equally in the calculations. Instead, we should compute the total income for each county, add up all the counties' totals, and then divide by the number of people in all the counties. If we completed these steps with the `county` data, we would find that the per capita income for the US is $27,348.43. Had we computed the *simple* mean of per capita income across counties, the result would have been just $22,504.70!

Example 1.23 used what is called a **weighted mean**, which will not be a key topic in this textbook. However, we have provided an online supplement on weighted means for interested readers:

www.openintro.org/stat/down/supp/wtdmean.pdf

1.6.3 Histograms and shape

Dot plots show the exact value for each observation. This is useful for small data sets, but they can become hard to read with larger samples. Rather than showing the value of each observation, we prefer to think of the value as belonging to a *bin*. For example, in the `email50` data set, we create a table of counts for the number of cases with character counts between 0 and 5,000, then the number of cases between 5,000 and 10,000, and so on. Observations that fall on the boundary of a bin (e.g. 5,000) are allocated to the lower bin. This tabulation is shown in Table 1.21. These binned counts are plotted as bars in Figure 1.22 into what is called a **histogram**, which resembles the stacked dot plot shown in Figure 1.20.

Characters (in thousands)	0-5	5-10	10-15	15-20	20-25	25-30	···	55-60	60-65
Count	19	12	6	2	3	5	···	0	1

Table 1.21: The counts for the binned `num_char` data.

Histograms provide a view of the **data density**. Higher bars represent where the data are relatively more common. For instance, there are many more emails with fewer than 20,000 characters than emails with at least 20,000 in the data set. The bars make it easy to see how the density of the data changes relative to the number of characters.

Histograms are especially convenient for describing the shape of the data distribution. Figure 1.22 shows that most emails have a relatively small number of characters, while fewer emails have a very large number of characters. When data trail off to the right in this way and have a longer right tail, the shape is said to be **right skewed**.[25]

Data sets with the reverse characteristic – a long, thin tail to the left – are said to be **left skewed**. We also say that such a distribution has a long left tail. Data sets that show roughly equal trailing off in both directions are called **symmetric**.

[25] Other ways to describe data that are skewed to the right: **skewed to the right**, **skewed to the high end**, or **skewed to the positive end**.

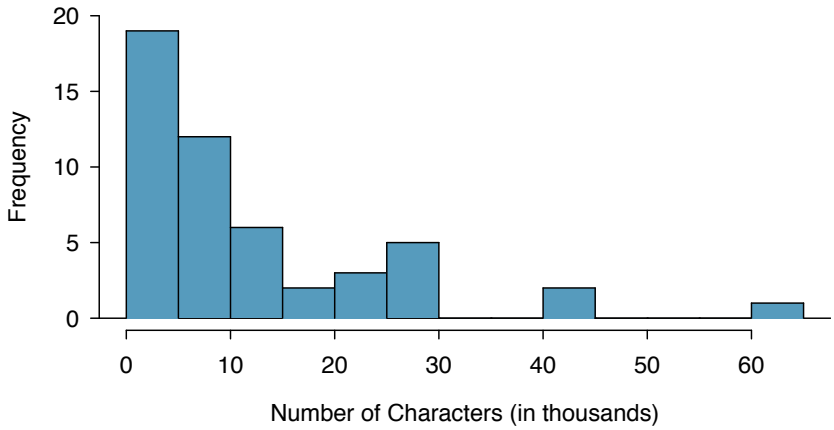


Figure 1.22: A histogram of `num_char`. This distribution is very strongly skewed to the right.

Long tails to identify skew

When data trail off in one direction, the distribution has a **long tail**. If a distribution has a long left tail, it is left skewed. If a distribution has a long right tail, it is right skewed.

⊙ **Guided Practice 1.24** Take a look at the dot plots in Figures 1.19 and 1.20. Can you see the skew in the data? Is it easier to see the skew in this histogram or the dot plots?[26]

⊙ **Guided Practice 1.25** Besides the mean (since it was labeled), what can you see in the dot plots that you cannot see in the histogram?[27]

In addition to looking at whether a distribution is skewed or symmetric, histograms can be used to identify modes. A **mode** is represented by a prominent peak in the distribution.[28] There is only one prominent peak in the histogram of `num_char`.

Figure 1.23 shows histograms that have one, two, or three prominent peaks. Such distributions are called **unimodal**, **bimodal**, and **multimodal**, respectively. Any distribution with more than 2 prominent peaks is called multimodal. Notice that there was one prominent peak in the unimodal distribution with a second less prominent peak that was not counted since it only differs from its neighboring bins by a few observations.

⊙ **Guided Practice 1.26** Figure 1.22 reveals only one prominent mode in the number of characters. Is the distribution unimodal, bimodal, or multimodal?[29]

[26] The skew is visible in all three plots, though the flat dot plot is the least useful. The stacked dot plot and histogram are helpful visualizations for identifying skew.

[27] Character counts for individual emails.

[28] Another definition of mode, which is not typically used in statistics, is the value with the most occurrences. It is common to have *no* observations with the same value in a data set, which makes this other definition useless for many real data sets.

[29] Unimodal. Remember that *uni* stands for 1 (think *uni*cycles). Similarly, *bi* stands for 2 (think *bi*cycles). (We're hoping a *multicycle* will be invented to complete this analogy.)

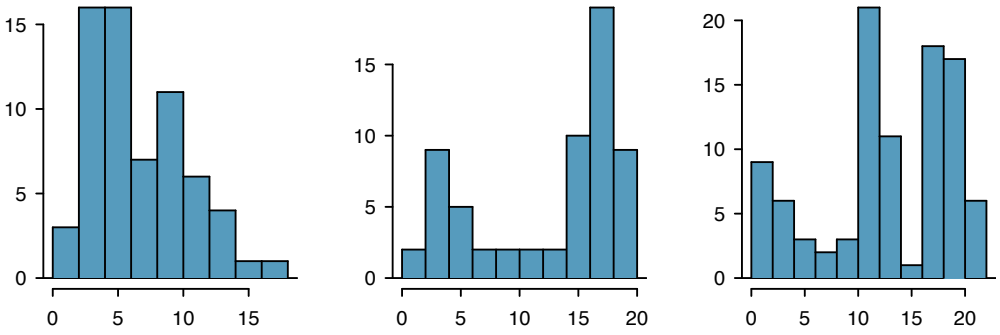


Figure 1.23: Counting only prominent peaks, the distributions are (left to right) unimodal, bimodal, and multimodal.

⊙ **Guided Practice 1.27** Height measurements of young students and adult teachers at a K-3 elementary school were taken. How many modes would you anticipate in this height data set?[30]

TIP: Looking for modes
Looking for modes isn't about finding a clear and correct answer about the number of modes in a distribution, which is why *prominent* is not rigorously defined in this book. The important part of this examination is to better understand your data and how it might be structured.

1.6.4 Variance and standard deviation

The mean was introduced as a method to describe the center of a data set, but the variability in the data is also important. Here, we introduce two measures of variability: the variance and the standard deviation. Both of these are very useful in data analysis, even though their formulas are a bit tedious to calculate by hand. The standard deviation is the easier of the two to understand, and it roughly describes how far away the typical observation is from the mean.

We call the distance of an observation from its mean its **deviation**. Below are the deviations for the 1^{st}, 2^{nd}, 3^{rd}, and 50^{th} observations in the `num_char` variable. For computational convenience, the number of characters is listed in the thousands and rounded to the first decimal.

$$\begin{aligned} x_1 - \bar{x} &= 21.7 - 11.6 = 10.1 \\ x_2 - \bar{x} &= 7.0 - 11.6 = -4.6 \\ x_3 - \bar{x} &= 0.6 - 11.6 = -11.0 \\ &\vdots \\ x_{50} - \bar{x} &= 15.8 - 11.6 = 4.2 \end{aligned}$$

[30] There might be two height groups visible in the data set: one of the students and one of the adults. That is, the data are probably bimodal.

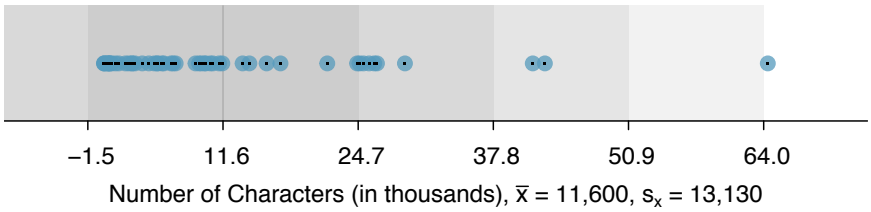


Figure 1.24: In the `num_char` data, 41 of the 50 emails (82%) are within 1 standard deviation of the mean, and 47 of the 50 emails (94%) are within 2 standard deviations. Usually about 70% of the data are within 1 standard deviation of the mean and 95% are within 2 standard deviations, though this rule of thumb is less accurate for skewed data, as shown in this example.

If we square these deviations and then take an average, the result is about equal to the sample **variance**, denoted by s^2:

s^2
sample variance

$$\begin{aligned} s^2 &= \frac{10.1^2 + (-4.6)^2 + (-11.0)^2 + \cdots + 4.2^2}{50 - 1} \\ &= \frac{102.01 + 21.16 + 121.00 + \cdots + 17.64}{49} \\ &= 172.44 \end{aligned}$$

We divide by $n - 1$, rather than dividing by n, when computing the variance; you need not worry about this mathematical nuance for the material in this textbook. Notice that squaring the deviations does two things. First, it makes large values much larger, seen by comparing 10.1^2, $(-4.6)^2$, $(-11.0)^2$, and 4.2^2. Second, it gets rid of any negative signs.

The **standard deviation** is defined as the square root of the variance:

$$s = \sqrt{172.44} = 13.13$$

s
sample standard deviation

The standard deviation of the number of characters in an email is about 13.13 thousand. A subscript of $_x$ may be added to the variance and standard deviation, i.e. s_x^2 and s_x, as a reminder that these are the variance and standard deviation of the observations represented by x_1, x_2, ..., x_n. The $_x$ subscript is usually omitted when it is clear which data the variance or standard deviation is referencing.

Variance and standard deviation

The variance is roughly the average squared distance from the mean. The standard deviation is the square root of the variance. The standard deviation is useful when considering how close the data are to the mean.

Formulas and methods used to compute the variance and standard deviation for a population are similar to those used for a sample.[31] However, like the mean, the population values have special symbols: σ^2 for the variance and σ for the standard deviation. The symbol σ is the Greek letter *sigma*.

σ^2
population variance

σ
population standard deviation

[31] The only difference is that the population variance has a division by n instead of $n - 1$.

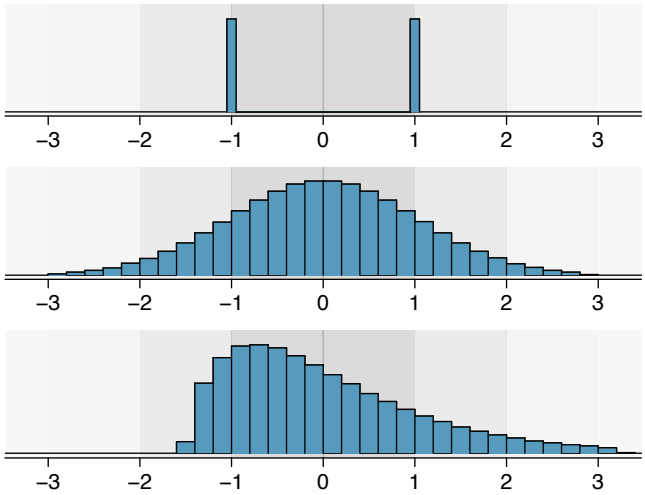


Figure 1.25: Three very different population distributions with the same mean $\mu = 0$ and standard deviation $\sigma = 1$.

TIP: standard deviation describes variability
Focus on the conceptual meaning of the standard deviation as a descriptor of variability rather than the formulas. Usually 70% of the data will be within one standard deviation of the mean and about 95% will be within two standard deviations. However, as seen in Figures 1.24 and 1.25, these percentages are not strict rules.

⊙ **Guided Practice 1.28** On page 30, the concept of shape of a distribution was introduced. A good description of the shape of a distribution should include modality and whether the distribution is symmetric or skewed to one side. Using Figure 1.25 as an example, explain why such a description is important.[32]

● **Example 1.29** Describe the distribution of the `num_char` variable using the histogram in Figure 1.22 on page 31. The description should incorporate the center, variability, and shape of the distribution, and it should also be placed in context: the number of characters in emails. Also note any especially unusual cases.

The distribution of email character counts is unimodal and very strongly skewed to the high end. Many of the counts fall near the mean at 11,600, and most fall within one standard deviation (13,130) of the mean. There is one exceptionally long email with about 65,000 characters.

In practice, the variance and standard deviation are sometimes used as a means to an end, where the "end" is being able to accurately estimate the uncertainty associated with a sample statistic. For example, in Chapter 4 we will use the variance and standard deviation to assess how close the sample mean is to the population mean.

[32] Figure 1.25 shows three distributions that look quite different, but all have the same mean, variance, and standard deviation. Using modality, we can distinguish between the first plot (bimodal) and the last two (unimodal). Using skewness, we can distinguish between the last plot (right skewed) and the first two. While a picture, like a histogram, tells a more complete story, we can use modality and shape (symmetry/skew) to characterize basic information about a distribution.

1.6.5 Box plots, quartiles, and the median

A **box plot** summarizes a data set using five statistics while also plotting unusual observations. Figure 1.26 provides a vertical dot plot alongside a box plot of the `num_char` variable from the `email50` data set.

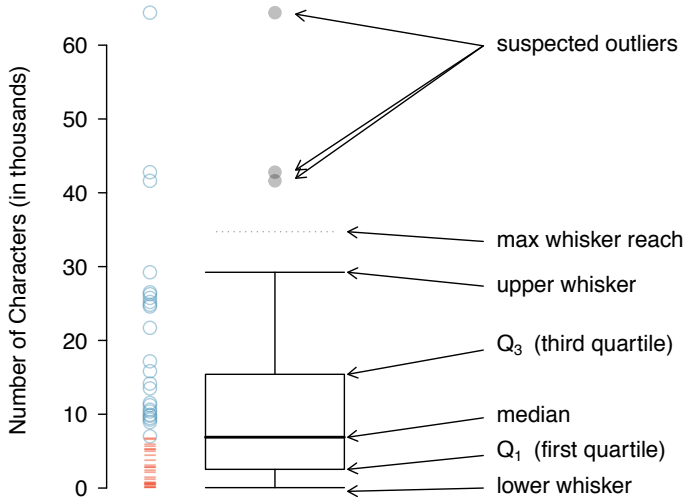


Figure 1.26: A vertical dot plot next to a labeled box plot for the number of characters in 50 emails. The median (6,890), splits the data into the bottom 50% and the top 50%, marked in the dot plot by horizontal dashes and open circles, respectively.

The first step in building a box plot is drawing a dark line denoting the **median**, which splits the data in half. Figure 1.26 shows 50% of the data falling below the median (dashes) and other 50% falling above the median (open circles). There are 50 character counts in the data set (an even number) so the data are perfectly split into two groups of 25. We take the median in this case to be the average of the two observations closest to the 50^{th} percentile: $(6{,}768 + 7{,}012)/2 = 6{,}890$. When there are an odd number of observations, there will be exactly one observation that splits the data into two halves, and in this case that observation is the median (no average needed).

Median: the number in the middle

If the data are ordered from smallest to largest, the **median** is the observation right in the middle. If there are an even number of observations, there will be two values in the middle, and the median is taken as their average.

The second step in building a box plot is drawing a rectangle to represent the middle 50% of the data. The total length of the box, shown vertically in Figure 1.26, is called the **interquartile range** (IQR, for short). It, like the standard deviation, is a measure of variability in data. The more variable the data, the larger the standard deviation and IQR. The two boundaries of the box are called the **first quartile** (the 25^{th} percentile, i.e. 25% of the data fall below this value) and the **third quartile** (the 75^{th} percentile), and these are often labeled Q_1 and Q_3, respectively.

Interquartile range (IQR)

The IQR is the length of the box in a box plot. It is computed as

$$IQR = Q_3 - Q_1$$

where Q_1 and Q_3 are the 25^{th} and 75^{th} percentiles.

⊙ **Guided Practice 1.30** What percent of the data fall between Q_1 and the median? What percent is between the median and Q_3?[33]

Extending out from the box, the **whiskers** attempt to capture the data outside of the box, however, their reach is never allowed to be more than $1.5 \times IQR$.[34] They capture everything within this reach. In Figure 1.26, the upper whisker does not extend to the last three points, which is beyond $Q_3 + 1.5 \times IQR$, and so it extends only to the last point below this limit. The lower whisker stops at the lowest value, 33, since there is no additional data to reach; the lower whisker's limit is not shown in the figure because the plot does not extend down to $Q_1 - 1.5 \times IQR$. In a sense, the box is like the body of the box plot and the whiskers are like its arms trying to reach the rest of the data.

Any observation that lies beyond the whiskers is labeled with a dot. The purpose of labeling these points – instead of just extending the whiskers to the minimum and maximum observed values – is to help identify any observations that appear to be unusually distant from the rest of the data. Unusually distant observations are called **outliers**. In this case, it would be reasonable to classify the emails with character counts of 41,623, 42,793, and 64,401 as outliers since they are numerically distant from most of the data.

Outliers are extreme

An **outlier** is an observation that appears extreme relative to the rest of the data.

TIP: Why it is important to look for outliers

Examination of data for possible outliers serves many useful purposes, including

1. Identifying strong skew in the distribution.
2. Identifying data collection or entry errors. For instance, we re-examined the email purported to have 64,401 characters to ensure this value was accurate.
3. Providing insight into interesting properties of the data.

⊙ **Guided Practice 1.31** The observation 64,401, a suspected outlier, was found to be an accurate observation. What would such an observation suggest about the nature of character counts in emails?[35]

[33] Since Q_1 and Q_3 capture the middle 50% of the data and the median splits the data in the middle, 25% of the data fall between Q_1 and the median, and another 25% falls between the median and Q_3.

[34] While the choice of exactly 1.5 is arbitrary, it is the most commonly used value for box plots.

[35] That occasionally there may be very long emails.

⊙ **Guided Practice 1.32** Using Figure 1.26, estimate the following values for `num_char` in the `email50` data set: (a) Q_1, (b) Q_3, and (c) IQR.[36]

Calculator videos
Videos covering how to create statistical summaries and box plots using TI and Casio graphing calculators are available at openintro.org/videos.

1.6.6 Robust statistics

How are the sample statistics of the `num_char` data set affected by the observation, 64,401? What would have happened if this email wasn't observed? What would happen to these summary statistics if the observation at 64,401 had been even larger, say 150,000? These scenarios are plotted alongside the original data in Figure 1.27, and sample statistics are computed under each scenario in Table 1.28.

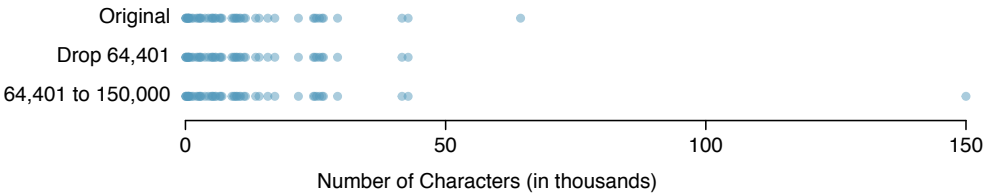


Figure 1.27: Dot plots of the original character count data and two modified data sets.

	robust		not robust	
scenario	median	IQR	$\bar{x}$	s
original `num_char` data	6,890	12,875	11,600	13,130
drop 64,401 observation	6,768	11,702	10,521	10,798
move 64,401 to 150,000	6,890	12,875	13,310	22,434

Table 1.28: A comparison of how the median, IQR, mean ($\bar{x}$), and standard deviation (s) change when extreme observations are present.

⊙ **Guided Practice 1.33** (a) Which is more affected by extreme observations, the mean or median? Table 1.28 may be helpful. (b) Is the standard deviation or IQR more affected by extreme observations?[37]

The median and IQR are called **robust estimates** because extreme observations have little effect on their values. The mean and standard deviation are much more affected by changes in extreme observations.

[36] These visual estimates will vary a little from one person to the next: $Q_1 = 3{,}000$, $Q_3 = 15{,}000$, IQR $= Q_3 - Q_1 = 12{,}000$. (The true values: $Q_1 = 2{,}536$, $Q_3 = 15{,}411$, IQR $= 12{,}875$.)

[37] (a) Mean is affected more. (b) Standard deviation is affected more. Complete explanations are provided in the material following Guided Practice 1.33.

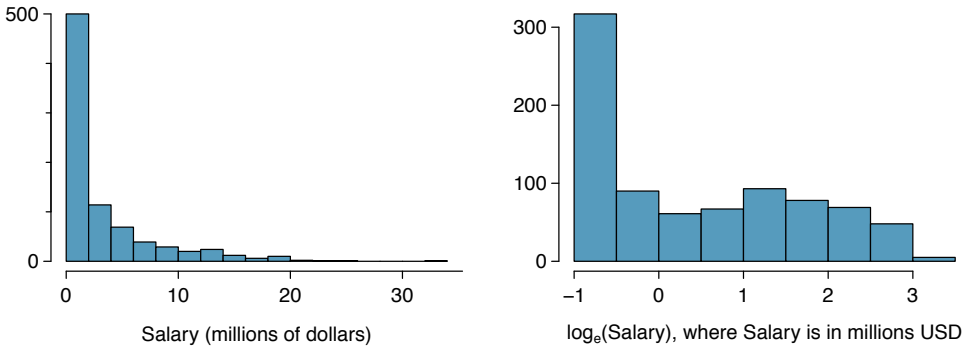
Example 1.34 The median and IQR do not change much under the three scenarios in Table 1.28. Why might this be the case?

The median and IQR are only sensitive to numbers near Q_1, the median, and Q_3. Since values in these regions are relatively stable – there aren't large jumps between observations – the median and IQR estimates are also quite stable.

Guided Practice 1.35 The distribution of vehicle prices tends to be right skewed, with a few luxury and sports cars lingering out into the right tail. If you were searching for a new car and cared about price, should you be more interested in the mean or median price of vehicles sold, assuming you are in the market for a regular car?[38]

1.6.7 Transforming data (special topic)

When data are very strongly skewed, we sometimes transform them so they are easier to model. Consider the histogram of salaries for Major League Baseball players' salaries from 2010, which is shown in Figure 1.29(a).



(a) (b)

Figure 1.29: (a) Histogram of MLB player salaries for 2010, in millions of dollars. (b) Histogram of the log-transformed MLB player salaries for 2010.

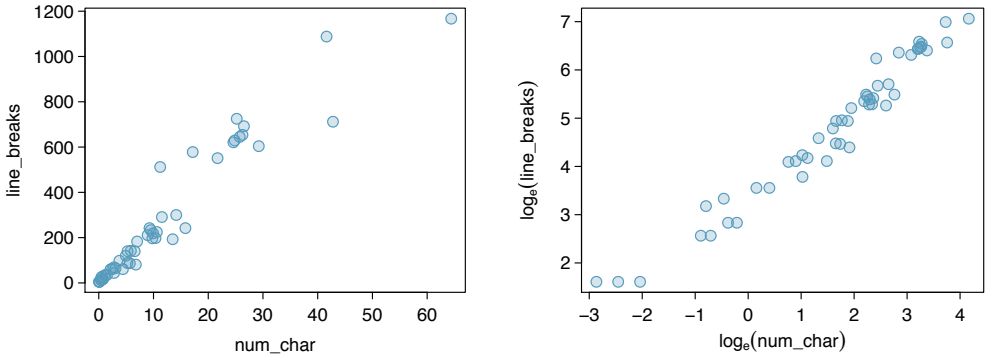
Example 1.36 The histogram of MLB player salaries is useful in that we can see the data are extremely skewed and centered (as gauged by the median) at about \$1 million. What isn't useful about this plot?

Most of the data are collected into one bin in the histogram and the data are so strongly skewed that many details in the data are obscured.

There are some standard transformations that are often applied when much of the data cluster near zero (relative to the larger values in the data set) and all observations are positive. A **transformation** is a rescaling of the data using a function. For instance, a plot of the natural logarithm[39] of player salaries results in a new histogram in Figure 1.29(b).

[38] Buyers of a "regular car" should be concerned about the median price. High-end car sales can drastically inflate the mean price while the median will be more robust to the influence of those sales.

[39] Statisticians often write the natural logarithm as log. You might be more familiar with it being written as ln.



(a) (b)

Figure 1.30: (a) Scatterplot of `line_breaks` against `num_char` for 50 emails. (b) A scatterplot of the same data but where each variable has been log-transformed.

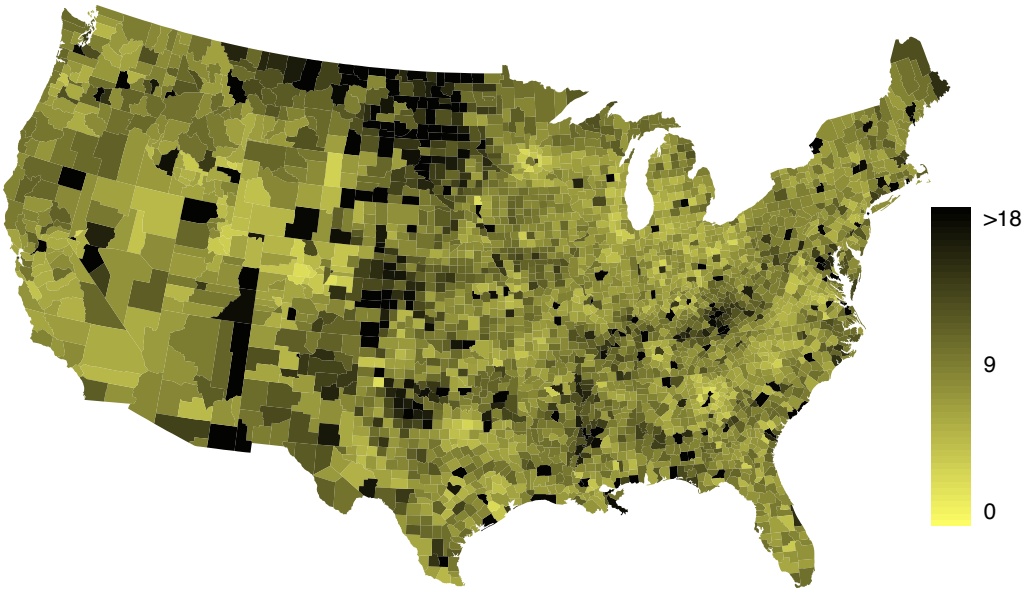
Transformed data are sometimes easier to work with when applying statistical models because the transformed data are much less skewed and outliers are usually less extreme.

Transformations can also be applied to one or both variables in a scatterplot. A scatterplot of the `line_breaks` and `num_char` variables is shown in Figure 1.30(a), which was earlier shown in Figure 1.17. We can see a positive association between the variables and that many observations are clustered near zero. In Chapter 7, we might want to use a straight line to model the data. However, we'll find that the data in their current state cannot be modeled very well. Figure 1.30(b) shows a scatterplot where both the `line_breaks` and `num_char` variables have been transformed using a log (base e) transformation. While there is a positive association in each plot, the transformed data show a steadier trend, which is easier to model than the untransformed data.

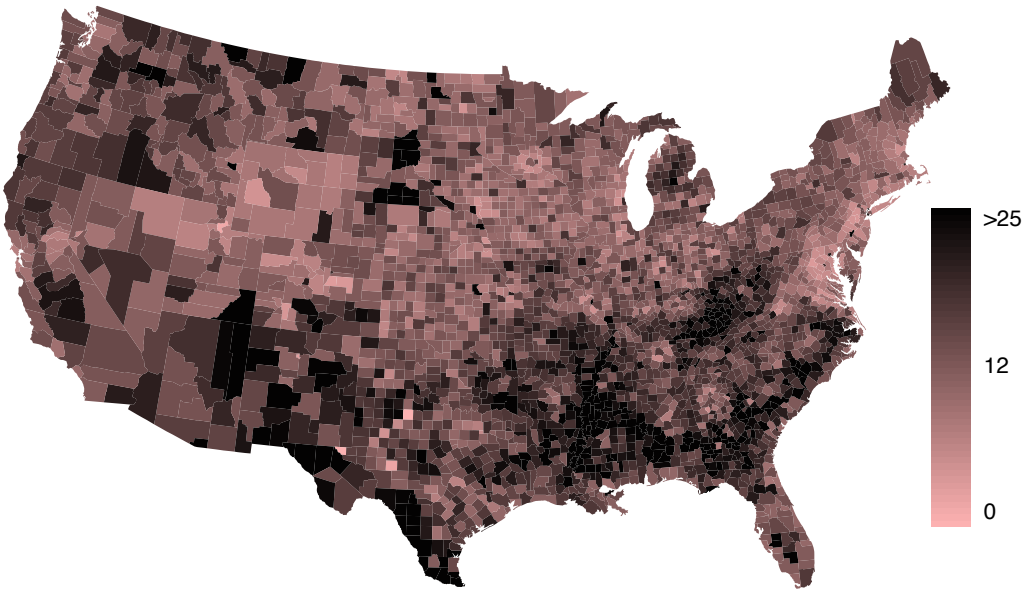
Transformations other than the logarithm can be useful, too. For instance, the square root ($\sqrt{\text{original observation}}$) and inverse ($\frac{1}{\text{original observation}}$) are used by statisticians. Common goals in transforming data are to see the data structure differently, reduce skew, assist in modeling, or straighten a nonlinear relationship in a scatterplot.

1.6.8 Mapping data (special topic)

The `county` data set offers many numerical variables that we could plot using dot plots, scatterplots, or box plots, but these miss the true nature of the data. Rather, when we encounter geographic data, we should map it using an **intensity map**, where colors are used to show higher and lower values of a variable. Figures 1.31 and 1.32 shows intensity maps for federal spending per capita (`fed_spend`), poverty rate in percent (`poverty`), homeownership rate in percent (`homeownership`), and median household income (`med_income`). The color key indicates which colors correspond to which values. Note that the intensity maps are not generally very helpful for getting precise values in any given county, but they are very helpful for seeing geographic trends and generating interesting research questions.

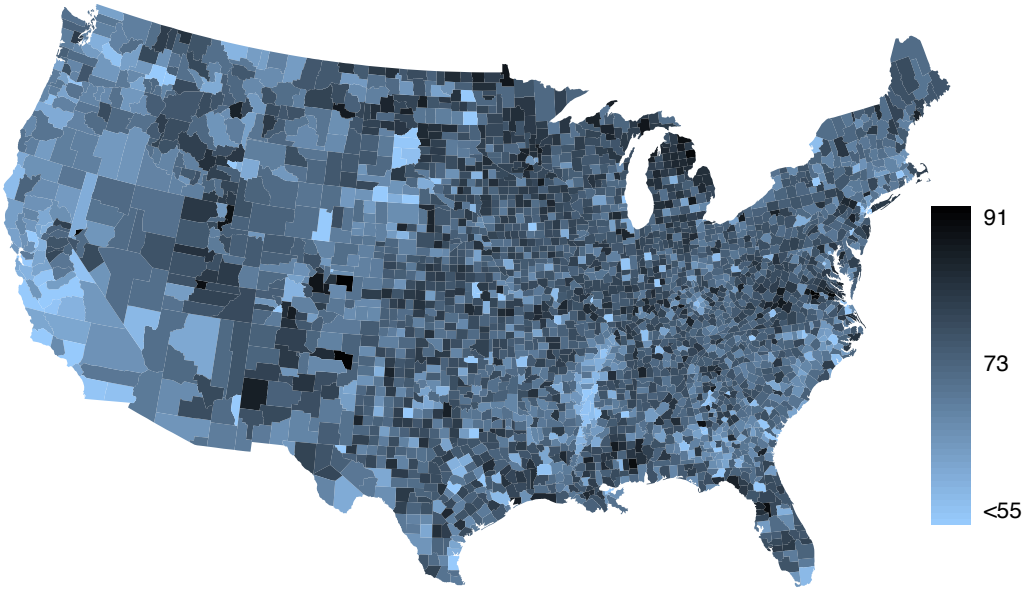


(a)

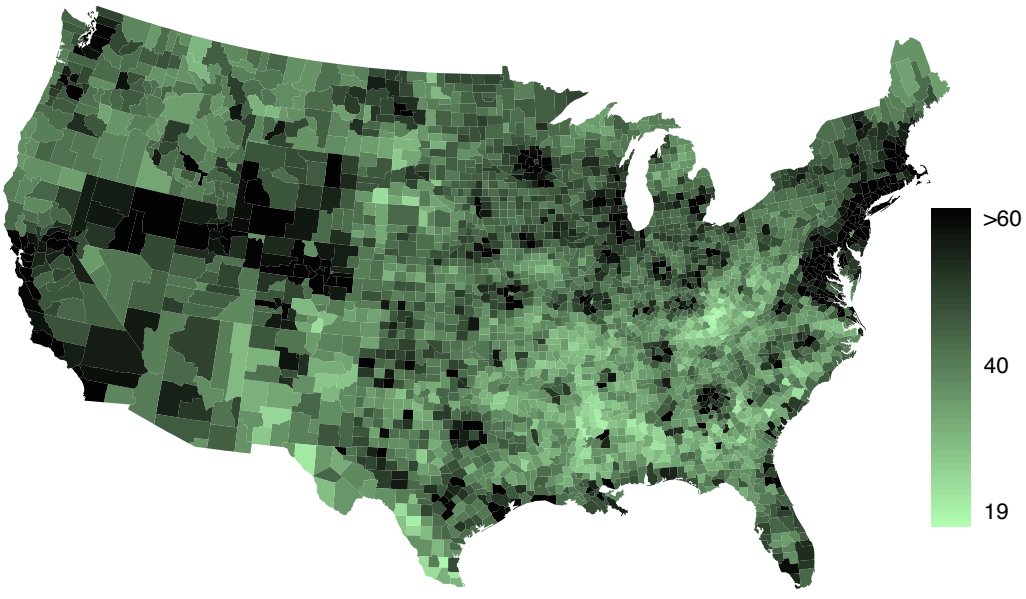


(b)

Figure 1.31: (a) Map of federal spending (dollars per capita). (b) Intensity map of poverty rate (percent).



(a)



(b)

Figure 1.32: (a) Intensity map of homeownership rate (percent). (b) Intensity map of median household income ($1000s).

Example 1.37 What interesting features are evident in the `fed_spend` and `poverty` intensity maps?

The federal spending intensity map shows substantial spending in the Dakotas and along the central-to-western part of the Canadian border, which may be related to the oil boom in this region. There are several other patches of federal spending, such as a vertical strip in eastern Utah and Arizona and the area where Colorado, Nebraska, and Kansas meet. There are also seemingly random counties with very high federal spending relative to their neighbors. If we did not cap the federal spending range at $18 per capita, we would actually find that some counties have extremely high federal spending while there is almost no federal spending in the neighboring counties. These high-spending counties might contain military bases, companies with large government contracts, or other government facilities with many employees.

Poverty rates are evidently higher in a few locations. Notably, the deep south shows higher poverty rates, as does the southwest border of Texas. The vertical strip of eastern Utah and Arizona, noted above for its higher federal spending, also appears to have higher rates of poverty (though generally little correspondence is seen between the two variables). High poverty rates are evident in the Mississippi flood plains a little north of New Orleans and also in a large section of Kentucky and West Virginia.

Guided Practice 1.38 What interesting features are evident in the `med_income` intensity map in Figure 1.32(b)?[40]

[40] Note: answers will vary. There is a very strong correspondence between high earning and metropolitan areas. You might look for large cities you are familiar with and try to spot them on the map as dark spots.

Chapter 8

Fundamental Sampling Distributions and Data Descriptions

8.1 Random Sampling

The outcome of a statistical experiment may be recorded either as a numerical value or as a descriptive representation. When a pair of dice is tossed and the total is the outcome of interest, we record a numerical value. However, if the students of a certain school are given blood tests and the type of blood is of interest, then a descriptive representation might be more useful. A person's blood can be classified in 8 ways: AB, A, B, or O, each with a plus or minus sign, depending on the presence or absence of the Rh antigen.

In this chapter, we focus on sampling from distributions or populations and study such important quantities as the *sample mean* and *sample variance*, which will be of vital importance in future chapters. In addition, we attempt to give the reader an introduction to the role that the sample mean and variance will play in statistical inference in later chapters. The use of modern high-speed computers allows the scientist or engineer to greatly enhance his or her use of formal statistical inference with graphical techniques. Much of the time, formal inference appears quite dry and perhaps even abstract to the practitioner or to the manager who wishes to let statistical analysis be a guide to decision-making.

Populations and Samples

We begin this section by discussing the notions of *populations* and *samples*. Both are mentioned in a broad fashion in Chapter 1. However, much more needs to be presented about them here, particularly in the context of the concept of random variables. The totality of observations with which we are concerned, whether their number be finite or infinite, constitutes what we call a **population**. There was a time when the word *population* referred to observations obtained from statistical studies about people. Today, statisticians use the term to refer to observations relevant to anything of interest, whether it be groups of people, animals, or all possible outcomes from some complicated biological or engineering system.

Definition 8.1: A **population** consists of the totality of the observations with which we are concerned.

The number of observations in the population is defined to be the size of the population. If there are 600 students in the school whom we classified according to blood type, we say that we have a population of size 600. The numbers on the cards in a deck, the heights of residents in a certain city, and the lengths of fish in a particular lake are examples of populations with finite size. In each case, the total number of observations is a finite number. The observations obtained by measuring the atmospheric pressure every day, from the past on into the future, or all measurements of the depth of a lake, from any conceivable position, are examples of populations whose sizes are infinite. Some finite populations are so large that in theory we assume them to be infinite. This is true in the case of the population of lifetimes of a certain type of storage battery being manufactured for mass distribution throughout the country.

Each observation in a population is a value of a random variable X having some probability distribution $f(x)$. If one is inspecting items coming off an assembly line for defects, then each observation in the population might be a value 0 or 1 of the Bernoulli random variable X with probability distribution

$$b(x; 1, p) = p^x q^{1-x}, \qquad x = 0, 1$$

where 0 indicates a nondefective item and 1 indicates a defective item. Of course, it is assumed that p, the probability of any item being defective, remains constant from trial to trial. In the blood-type experiment, the random variable X represents the type of blood and is assumed to take on values from 1 to 8. Each student is given one of the values of the discrete random variable. The lives of the storage batteries are values assumed by a continuous random variable having perhaps a normal distribution. When we refer hereafter to a "binomial population," a "normal population," or, in general, the "population $f(x)$," we shall mean a population whose observations are values of a random variable having a binomial distribution, a normal distribution, or the probability distribution $f(x)$. Hence, the mean and variance of a random variable or probability distribution are also referred to as the mean and variance of the corresponding population.

In the field of statistical inference, statisticians are interested in arriving at conclusions concerning a population when it is impossible or impractical to observe the entire set of observations that make up the population. For example, in attempting to determine the average length of life of a certain brand of light bulb, it would be impossible to test all such bulbs if we are to have any left to sell. Exorbitant costs can also be a prohibitive factor in studying an entire population. Therefore, we must depend on a subset of observations from the population to help us make inferences concerning that same population. This brings us to consider the notion of sampling.

Definition 8.2: A **sample** is a subset of a population.

If our inferences from the sample to the population are to be valid, we must obtain samples that are representative of the population. All too often we are

tempted to choose a sample by selecting the most convenient members of the population. Such a procedure may lead to erroneous inferences concerning the population. Any sampling procedure that produces inferences that consistently overestimate or consistently underestimate some characteristic of the population is said to be **biased**. To eliminate any possibility of bias in the sampling procedure, it is desirable to choose a **random sample** in the sense that the observations are made independently and at random.

In selecting a random sample of size n from a population $f(x)$, let us define the random variable X_i, $i = 1, 2, \ldots, n$, to represent the ith measurement or sample value that we observe. The random variables $X_1, X_2, \ldots, X_n$ will then constitute a random sample from the population $f(x)$ with numerical values $x_1, x_2, \ldots, x_n$ if the measurements are obtained by repeating the experiment n independent times under essentially the same conditions. Because of the identical conditions under which the elements of the sample are selected, it is reasonable to assume that the n random variables $X_1, X_2, \ldots, X_n$ are independent and that each has the same probability distribution $f(x)$. That is, the probability distributions of $X_1, X_2, \ldots, X_n$ are, respectively, $f(x_1), f(x_2), \ldots, f(x_n)$, and their joint probability distribution is $f(x_1, x_2, \ldots, x_n) = f(x_1)f(x_2)\cdots f(x_n)$. The concept of a random sample is described formally by the following definition.

Definition 8.3: Let $X_1, X_2, \ldots, X_n$ be n independent random variables, each having the same probability distribution $f(x)$. Define $X_1, X_2, \ldots, X_n$ to be a **random sample** of size n from the population $f(x)$ and write its joint probability distribution as

$$f(x_1, x_2, \ldots, x_n) = f(x_1)f(x_2)\cdots f(x_n).$$

If one makes a random selection of $n = 8$ storage batteries from a manufacturing process that has maintained the same specification throughout and records the length of life for each battery, with the first measurement x_1 being a value of X_1, the second measurement x_2 a value of X_2, and so forth, then $x_1, x_2, \ldots, x_8$ are the values of the random sample $X_1, X_2, \ldots, X_8$. If we assume the population of battery lives to be normal, the possible values of any X_i, $i = 1, 2, \ldots, 8$, will be precisely the same as those in the original population, and hence X_i has the same identical normal distribution as X.

8.2 Some Important Statistics

Our main purpose in selecting random samples is to elicit information about the unknown population parameters. Suppose, for example, that we wish to arrive at a conclusion concerning the proportion of coffee-drinkers in the United States who prefer a certain brand of coffee. It would be impossible to question every coffee-drinking American in order to compute the value of the parameter p representing the population proportion. Instead, a large random sample is selected and the proportion $\hat{p}$ of people in this sample favoring the brand of coffee in question is calculated. The value $\hat{p}$ is now used to make an inference concerning the true proportion p.

Now, $\hat{p}$ is a function of the observed values in the random sample; since many

random samples are possible from the same population, we would expect $\hat{p}$ to vary somewhat from sample to sample. That is, $\hat{p}$ is a value of a random variable that we represent by P. Such a random variable is called a **statistic**.

Definition 8.4: Any function of the random variables constituting a random sample is called a **statistic**.

Location Measures of a Sample: The Sample Mean, Median, and Mode

In Chapter 4 we introduced the two parameters μ and σ^2, which measure the center of location and the variability of a probability distribution. These are constant population parameters and are in no way affected or influenced by the observations of a random sample. We shall, however, define some important statistics that describe corresponding measures of a random sample. The most commonly used statistics for measuring the center of a set of data, arranged in order of magnitude, are the **mean**, **median**, and **mode**. Although the first two of these statistics were defined in Chapter 1, we repeat the definitions here. Let $X_1, X_2, \ldots, X_n$ represent n random variables.

(a) Sample mean:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Note that the statistic $\bar{X}$ assumes the value $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ when X_1 assumes the value x_1, X_2 assumes the value x_2, and so forth. The term *sample mean* is applied to both the statistic $\bar{X}$ and its computed value $\bar{x}$.

(b) Sample median:

$$\tilde{x} = \begin{cases} x_{(n+1)/2}, & \text{if } n \text{ is odd,} \\ \frac{1}{2}(x_{n/2} + x_{n/2+1}), & \text{if } n \text{ is even.} \end{cases}$$

The sample median is also a location measure that shows the middle value of the sample. Examples for both the sample mean and the sample median can be found in Section 1.3. The sample mode is defined as follows.

(c) The sample mode is the value of the sample that occurs most often.

Example 8.1: Suppose a data set consists of the following observations:

0.32 0.53 0.28 0.37 0.47 0.43 0.36 0.42 0.38 0.43.

The sample mode is 0.43, since this value occurs more than any other value. ◢

As we suggested in Chapter 1, a measure of location or central tendency in a sample does not by itself give a clear indication of the nature of the sample. Thus, a measure of variability in the sample must also be considered.

Variability Measures of a Sample: The Sample Variance, Standard Deviation, and Range

The variability in a sample displays how the observations spread out from the average. The reader is referred to Chapter 1 for more discussion. It is possible to have two sets of observations with the same mean or median that differ considerably in the variability of their measurements about the average.

Consider the following measurements, in liters, for two samples of orange juice bottled by companies A and B:

Sample A	0.97	1.00	0.94	1.03	1.06
Sample B	1.06	1.01	0.88	0.91	1.14

Both samples have the same mean, 1.00 liter. It is obvious that company A bottles orange juice with a more uniform content than company B. We say that the **variability**, or the **dispersion**, of the observations from the average is less for sample A than for sample B. Therefore, in buying orange juice, we would feel more confident that the bottle we select will be close to the advertised average if we buy from company A.

In Chapter 1 we introduced several measures of sample variability, including the **sample variance**, **sample standard deviation**, and **sample range**. In this chapter, we will focus mainly on the sample variance. Again, let $X_1, \ldots, X_n$ represent n random variables.

(a) Sample variance:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2. \tag{8.2.1}$$

The computed value of S^2 for a given sample is denoted by s^2. Note that S^2 is essentially defined to be the average of the squares of the deviations of the observations from their mean. The reason for using $n-1$ as a divisor rather than the more obvious choice n will become apparent in Chapter 9.

Example 8.2: A comparison of coffee prices at 4 randomly selected grocery stores in San Diego showed increases from the previous month of 12, 15, 17, and 20 cents for a 1-pound bag. Find the variance of this random sample of price increases.

***Solution*:** Calculating the sample mean, we get

$$\bar{x} = \frac{12 + 15 + 17 + 20}{4} = 16 \text{ cents}.$$

Therefore,

$$s^2 = \frac{1}{3}\sum_{i=1}^{4}(x_i - 16)^2 = \frac{(12-16)^2 + (15-16)^2 + (17-16)^2 + (20-16)^2}{3}$$
$$= \frac{(-4)^2 + (-1)^2 + (1)^2 + (4)^2}{3} = \frac{34}{3}.$$

Whereas the expression for the sample variance best illustrates that S^2 is a measure of variability, an alternative expression does have some merit and thus the reader should be aware of it. The following theorem contains this expression.

Theorem 8.1: If S^2 is the variance of a random sample of size n, we may write

$$S^2 = \frac{1}{n(n-1)}\left[n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2\right].$$

***Proof*:** By definition,

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i^2 - 2\bar{X}X_i + \bar{X}^2)$$
$$= \frac{1}{n-1}\left[\sum_{i=1}^{n} X_i^2 - 2\bar{X}\sum_{i=1}^{n} X_i + n\bar{X}^2\right].$$

As in Chapter 1, the **sample standard deviation** and the **sample range** are defined below.

(b) Sample standard deviation:

$$S = \sqrt{S^2},$$

where S^2 is the sample variance.

Let $X_{\max}$ denote the largest of the X_i values and $X_{\min}$ the smallest.

(c) Sample range:

$$R = X_{\max} - X_{\min}.$$

Example 8.3: Find the variance of the data 3, 4, 5, 6, 6, and 7, representing the number of trout caught by a random sample of 6 fishermen on June 19, 1996, at Lake Muskoka.

***Solution*:** We find that $\sum_{i=1}^{6} x_i^2 = 171$, $\sum_{i=1}^{6} x_i = 31$, and $n = 6$. Hence,

$$s^2 = \frac{1}{(6)(5)}[(6)(171) - (31)^2] = \frac{13}{6}.$$

Thus, the sample standard deviation $s = \sqrt{13/6} = 1.47$ and the sample range is $7 - 3 = 4$.

Exercises

8.1 Define suitable populations from which the following samples are selected:

(a) Persons in 200 homes in the city of Richmond are called on the phone and asked to name the candidate they favor for election to the school board.

(b) A coin is tossed 100 times and 34 tails are recorded.

(c) Two hundred pairs of a new type of tennis shoe were tested on the professional tour and, on average, lasted 4 months.

(d) On five different occasions it took a lawyer 21, 26, 24, 22, and 21 minutes to drive from her suburban home to her midtown office.

8.2 The lengths of time, in minutes, that 10 patients waited in a doctor's office before receiving treatment were recorded as follows: 5, 11, 9, 5, 10, 15, 6, 10, 5, and 10. Treating the data as a random sample, find

(a) the mean;

(b) the median;

(c) the mode.

8.3 The reaction times for a random sample of 9 subjects to a stimulant were recorded as 2.5, 3.6, 3.1, 4.3, 2.9. 2.3, 2.6, 4.1, and 3.4 seconds. Calculate

(a) the mean;

(b) the median.

8.4 The number of tickets issued for traffic violations by 8 state troopers during the Memorial Day weekend are 5, 4, 7, 7, 6, 3, 8, and 6.

(a) If these values represent the number of tickets issued by a random sample of 8 state troopers from Montgomery County in Virginia, define a suitable population.

(b) If the values represent the number of tickets issued by a random sample of 8 state troopers from South Carolina, define a suitable population.

8.5 The numbers of incorrect answers on a true-false competency test for a random sample of 15 students were recorded as follows: 2, 1, 3, 0, 1, 3, 6, 0, 3, 3, 5, 2, 1, 4, and 2. Find

(a) the mean;

(b) the median;

(c) the mode.

8.6 Find the mean, median, and mode for the sample whose observations, 15, 7, 8, 95, 19, 12, 8, 22, and 14, represent the number of sick days claimed on 9 federal income tax returns. Which value appears to be the best measure of the center of these data? State reasons for your preference.

8.7 A random sample of employees from a local manufacturing plant pledged the following donations, in dollars, to the United Fund: 100, 40, 75, 15, 20, 100, 75, 50, 30, 10, 55, 75, 25, 50, 90, 80, 15, 25, 45, and 100. Calculate

(a) the mean;

(b) the mode.

8.8 According to ecology writer Jacqueline Killeen, phosphates contained in household detergents pass right through our sewer systems, causing lakes to turn into swamps that eventually dry up into deserts. The following data show the amount of phosphates per load of laundry, in grams, for a random sample of various types of detergents used according to the prescribed directions:

Laundry Detergent	**Phosphates per Load (grams)**
A & P Blue Sail	48
Dash	47
Concentrated All	42
Cold Water All	42
Breeze	41
Oxydol	34
Ajax	31
Sears	30
Fab	29
Cold Power	29
Bold	29
Rinso	26

For the given phosphate data, find

(a) the mean;

(b) the median;

(c) the mode.

8.9 Consider the data in Exercise 8.2, find

(a) the range;

(b) the standard deviation.

8.10 For the sample of reaction times in Exercise 8.3, calculate

(a) the range;

(b) the variance, using the formula of form (8.2.1).

8.11 For the data of Exercise 8.5, calculate the variance using the formula

(a) of form (8.2.1);

(b) in Theorem 8.1.

8.12 The tar contents of 8 brands of cigarettes selected at random from the latest list released by the Federal Trade Commission are as follows: 7.3, 8.6, 10.4, 16.1, 12.2, 15.1, 14.5, and 9.3 milligrams. Calculate

(a) the mean;

(b) the variance.

8.13 The grade-point averages of 20 college seniors selected at random from a graduating class are as follows:

3.2	1.9	2.7	2.4	2.8
2.9	3.8	3.0	2.5	3.3
1.8	2.5	3.7	2.8	2.0
3.2	2.3	2.1	2.5	1.9

Calculate the standard deviation.

8.14 (a) Show that the sample variance is unchanged if a constant c is added to or subtracted from each

value in the sample.

(b) Show that the sample variance becomes c^2 times its original value if each observation in the sample is multiplied by c.

8.15 Verify that the variance of the sample 4, 9, 3, 6, 4, and 7 is 5.1, and using this fact, along with the results of Exercise 8.14, find

(a) the variance of the sample 12, 27, 9, 18, 12, and 21;

(b) the variance of the sample 9, 14, 8, 11, 9, and 12.

8.16 In the 2004-05 football season, University of Southern California had the following score differences for the 13 games it played.

11 49 32 3 6 38 38 30 8 40 31 5 36

Find

(a) the mean score difference;

(b) the median score difference.

8.3 Sampling Distributions

The field of statistical inference is basically concerned with generalizations and predictions. For example, we might claim, based on the opinions of several people interviewed on the street, that in a forthcoming election 60% of the eligible voters in the city of Detroit favor a certain candidate. In this case, we are dealing with a random sample of opinions from a very large finite population. As a second illustration we might state that the average cost to build a residence in Charleston, South Carolina, is between $330,000 and $335,000, based on the estimates of 3 contractors selected at random from the 30 now building in this city. The population being sampled here is again finite but very small. Finally, let us consider a soft-drink machine designed to dispense, on average, 240 milliliters per drink. A company official who computes the mean of 40 drinks obtains $\bar{x} = 236$ milliliters and, on the basis of this value, decides that the machine is still dispensing drinks with an average content of $\mu = 240$ milliliters. The 40 drinks represent a sample from the infinite population of possible drinks that will be dispensed by this machine.

Inference about the Population from Sample Information

In each of the examples above, we computed a statistic from a sample selected from the population, and from this statistic we made various statements concerning the values of population parameters that may or may not be true. The company official made the decision that the soft-drink machine dispenses drinks with an average content of 240 milliliters, even though the sample mean was 236 milliliters, because he knows from sampling theory that, if $\mu = 240$ milliliters, such a sample value could easily occur. In fact, if he ran similar tests, say every hour, he would expect the values of the statistic $\bar{x}$ to fluctuate above and below $\mu = 240$ milliliters. Only when the value of $\bar{x}$ is substantially different from 240 milliliters will the company official initiate action to adjust the machine.

Since a statistic is a random variable that depends only on the observed sample, it must have a probability distribution.

Definition 8.5: The probability distribution of a statistic is called a **sampling distribution**.

The sampling distribution of a statistic depends on the distribution of the population, the size of the samples, and the method of choosing the samples. In the

remainder of this chapter we study several of the important sampling distributions of frequently used statistics. Applications of these sampling distributions to problems of statistical inference are considered throughout most of the remaining chapters. The probability distribution of $\bar{X}$ is called the **sampling distribution of the mean**.

What Is the Sampling Distribution of $\bar{X}$?

We should view the sampling distributions of $\bar{X}$ and S^2 as the mechanisms from which we will be able to make inferences on the parameters μ and σ^2. The sampling distribution of $\bar{X}$ with sample size n is the distribution that results when an **experiment is conducted over and over** (always with sample size n) **and the many values of $\bar{X}$ result**. This sampling distribution, then, describes the variability of sample averages around the population mean μ. In the case of the soft-drink machine, knowledge of the sampling distribution of $\bar{X}$ arms the analyst with the knowledge of a "typical" discrepancy between an observed $\bar{x}$ value and true μ. The same principle applies in the case of the distribution of S^2. The sampling distribution produces information about the variability of s^2 values around σ^2 in repeated experiments.

8.4 Sampling Distribution of Means and the Central Limit Theorem

The first important sampling distribution to be considered is that of the mean $\bar{X}$. Suppose that a random sample of n observations is taken from a normal population with mean μ and variance σ^2. Each observation X_i, $i = 1, 2, \ldots, n$, of the random sample will then have the same normal distribution as the population being sampled. Hence, by the reproductive property of the normal distribution established in Theorem 7.11, we conclude that

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

has a normal distribution with mean

$$\mu_{\bar{X}} = \frac{1}{n}(\underbrace{\mu + \mu + \cdots + \mu}_{n \text{ terms}}) = \mu \text{ and variance } \sigma^2_{\bar{X}} = \frac{1}{n^2}(\underbrace{\sigma^2 + \sigma^2 + \cdots + \sigma^2}_{n \text{ terms}}) = \frac{\sigma^2}{n}.$$

If we are sampling from a population with unknown distribution, either finite or infinite, the sampling distribution of $\bar{X}$ will still be approximately normal with mean μ and variance σ^2/n, provided that the sample size is large. This amazing result is an immediate consequence of the following theorem, called the Central Limit Theorem.

The Central Limit Theorem

Theorem 8.2: **Central Limit Theorem:** If $\bar{X}$ is the mean of a random sample of size n taken from a population with mean μ and finite variance σ^2, then the limiting form of the distribution of

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}},$$

as $n \to \infty$, is the standard normal distribution $n(z; 0, 1)$.

The normal approximation for $\bar{X}$ will generally be good if $n \geq 30$, provided the population distribution is not terribly skewed. If $n < 30$, the approximation is good only if the population is not too different from a normal distribution and, as stated above, if the population is known to be normal, the sampling distribution of $\bar{X}$ will follow a normal distribution exactly, no matter how small the size of the samples.

The sample size $n = 30$ is a guideline to use for the Central Limit Theorem. However, as the statement of the theorem implies, the presumption of normality on the distribution of $\bar{X}$ becomes more accurate as n grows larger. In fact, Figure 8.1 illustrates how the theorem works. It shows how the distribution of $\bar{X}$ becomes closer to normal as n grows larger, beginning with the clearly nonsymmetric distribution of an individual observation ($n = 1$). It also illustrates that the mean of $\bar{X}$ remains μ for any sample size and the variance of $\bar{X}$ gets smaller as n increases.

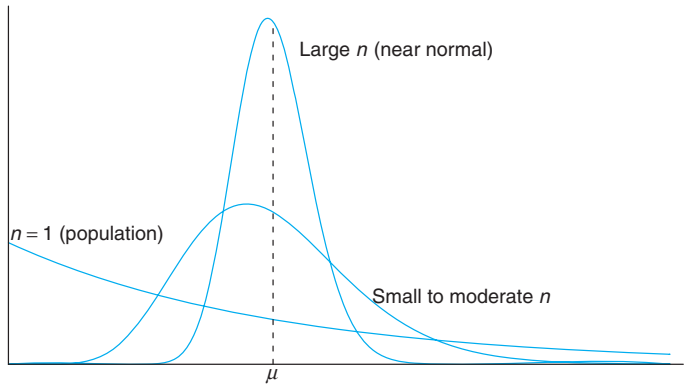


Figure 8.1: Illustration of the Central Limit Theorem (distribution of $\bar{X}$ for $n = 1$, moderate n, and large n).

Example 8.4: An electrical firm manufactures light bulbs that have a length of life that is approximately normally distributed, with mean equal to 800 hours and a standard deviation of 40 hours. Find the probability that a random sample of 16 bulbs will have an average life of less than 775 hours.

***Solution*:** The sampling distribution of $\bar{X}$ will be approximately normal, with $\mu_{\bar{X}} = 800$ and $\sigma_{\bar{X}} = 40/\sqrt{16} = 10$. The desired probability is given by the area of the shaded

region in Figure 8.2.

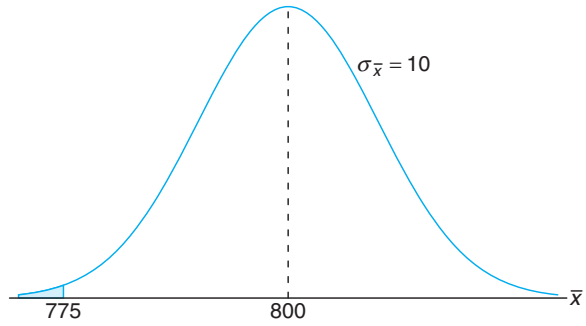


Figure 8.2: Area for Example 8.4.

Corresponding to $\bar{x} = 775$, we find that

$$z = \frac{775 - 800}{10} = -2.5,$$

and therefore

$$P(\bar{X} < 775) = P(Z < -2.5) = 0.0062.$$

Inferences on the Population Mean

One very important application of the Central Limit Theorem is the determination of reasonable values of the population mean μ. Topics such as hypothesis testing, estimation, quality control, and many others make use of the Central Limit Theorem. The following example illustrates the use of the Central Limit Theorem with regard to its relationship with μ, the mean of the population, although the formal application to the foregoing topics is relegated to future chapters.

In the following case study, an illustration is given which draws an inference that makes use of the sampling distribution of $\bar{X}$. In this simple illustration, μ and σ are both known. The Central Limit Theorem and the general notion of sampling distributions are often used to produce evidence about some important aspect of a distribution such as a parameter of the distribution. In the case of the Central Limit Theorem, the parameter of interest is the mean μ. The inference made concerning μ may take one of many forms. Often there is a desire on the part of the analyst that the data (in the form of $\bar{x}$) support (or not) some predetermined conjecture concerning the value of μ. The use of what we know about the sampling distribution can contribute to answering this type of question. In the following case study, the concept of hypothesis testing leads to a formal objective that we will highlight in future chapters.

Case Study 8.1: **Automobile Parts**:An important manufacturing process produces cylindrical component parts for the automotive industry. It is important that the process produce

parts having a mean diameter of 5.0 millimeters. The engineer involved conjectures that the population mean is 5.0 millimeters. An experiment is conducted in which 100 parts produced by the process are selected randomly and the diameter measured on each. It is known that the population standard deviation is $\sigma = 0.1$ millimeter. The experiment indicates a sample average diameter of $\bar{x} = 5.027$ millimeters. Does this sample information appear to support or refute the engineer's conjecture?

***Solution*:** This example reflects the kind of problem often posed and solved with hypothesis testing machinery introduced in future chapters. We will not use the formality associated with hypothesis testing here, but we will illustrate the principles and logic used.

Whether the data support or refute the conjecture depends on the probability that data similar to those obtained in this experiment ($\bar{x} = 5.027$) can readily occur when in fact $\mu = 5.0$ (Figure 8.3). In other words, how likely is it that one can obtain $\bar{x} \geq 5.027$ with $n = 100$ if the population mean is $\mu = 5.0$? If this probability suggests that $\bar{x} = 5.027$ is not unreasonable, the conjecture is not refuted. If the probability is quite low, one can certainly argue that the data do not support the conjecture that $\mu = 5.0$. The probability that we choose to compute is given by $P(|\bar{X} - 5| \geq 0.027)$.

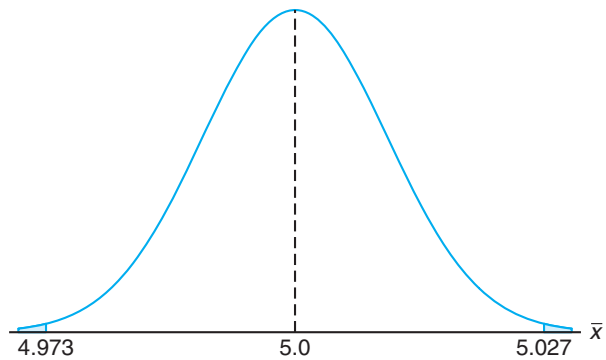


Figure 8.3: Area for Case Study 8.1.

In other words, if the mean μ is 5, what is the chance that $\bar{X}$ will deviate by as much as 0.027 millimeter?

$$P(|\bar{X} - 5| \geq 0.027) = P(\bar{X} - 5 \geq 0.027) + P(\bar{X} - 5 \leq -0.027)$$
$$= 2P\left(\frac{\bar{X} - 5}{0.1/\sqrt{100}} \geq 2.7\right).$$

Here we are simply standardizing $\bar{X}$ according to the Central Limit Theorem. If the conjecture $\mu = 5.0$ is true, $\frac{\bar{X}-5}{0.1/\sqrt{100}}$ should follow $N(0, 1)$. Thus,

$$2P\left(\frac{\bar{X} - 5}{0.1/\sqrt{100}} \geq 2.7\right) = 2P(Z \geq 2.7) = 2(0.0035) = 0.007.$$

Therefore, one would experience by chance that an $\bar{x}$ would be 0.027 millimeter from the mean in only 7 in 1000 experiments. As a result, this experiment with $\bar{x} = 5.027$ certainly does not give supporting evidence to the conjecture that $\mu = 5.0$. In fact, it strongly refutes the conjecture!

Example 8.5: Traveling between two campuses of a university in a city via shuttle bus takes, on average, 28 minutes with a standard deviation of 5 minutes. In a given week, a bus transported passengers 40 times. What is the probability that the average transport time was more than 30 minutes? Assume the mean time is measured to the nearest minute.

***Solution*:** In this case, $\mu = 28$ and $\sigma = 3$. We need to calculate the probability $P(\bar{X} > 30)$ with $n = 40$. Since the time is measured on a continuous scale to the nearest minute, an $\bar{x}$ greater than 30 is equivalent to $\bar{x} \geq 30.5$. Hence,

$$P(\bar{X} > 30) = P\left(\frac{\bar{X} - 28}{5/\sqrt{40}} \geq \frac{30.5 - 28}{5/\sqrt{40}}\right) = P(Z \geq 3.16) = 0.0008.$$

There is only a slight chance that the average time of one bus trip will exceed 30 minutes. An illustrative graph is shown in Figure 8.4.

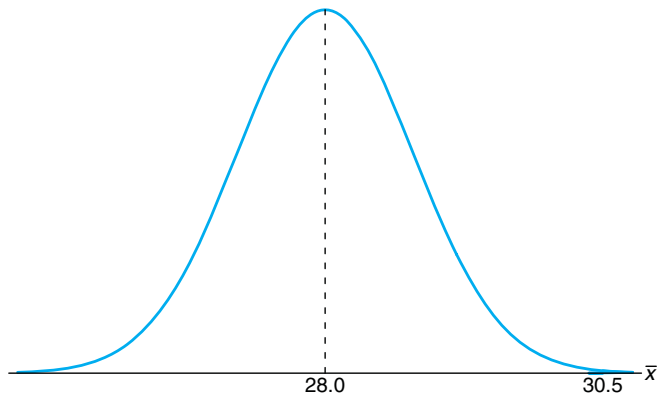


Figure 8.4: Area for Example 8.5.

Sampling Distribution of the Difference between Two Means

The illustration in Case Study 8.1 deals with notions of statistical inference on a single mean μ. The engineer was interested in supporting a conjecture regarding a single population mean. A far more important application involves two populations. A scientist or engineer may be interested in a comparative experiment in which two manufacturing methods, 1 and 2, are to be compared. The basis for that comparison is $\mu_1 - \mu_2$, the difference in the population means.

Suppose that we have two populations, the first with mean μ_1 and variance σ_1^2, and the second with mean μ_2 and variance σ_2^2. Let the statistic $\bar{X}_1$ represent the mean of a random sample of size n_1 selected from the first population, and the statistic $\bar{X}_2$ represent the mean of a random sample of size n_2 selected from

the second population, independent of the sample from the first population. What can we say about the sampling distribution of the difference $\bar{X}_1 - \bar{X}_2$ for repeated samples of size n_1 and n_2? According to Theorem 8.2, the variables $\bar{X}_1$ and $\bar{X}_2$ are both approximately normally distributed with means μ_1 and μ_2 and variances σ_1^2/n_1 and σ_2^2/n_2, respectively. This approximation improves as n_1 and n_2 increase. By choosing independent samples from the two populations we ensure that the variables $\bar{X}_1$ and $\bar{X}_2$ will be independent, and then using Theorem 7.11, with $a_1 = 1$ and $a_2 = -1$, we can conclude that $\bar{X}_1 - \bar{X}_2$ is approximately normally distributed with mean

$$\mu_{\bar{X}_1 - \bar{X}_2} = \mu_{\bar{X}_1} - \mu_{\bar{X}_2} = \mu_1 - \mu_2$$

and variance

$$\sigma^2_{\bar{X}_1 - \bar{X}_2} = \sigma^2_{\bar{X}_1} + \sigma^2_{\bar{X}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}.$$

The Central Limit Theorem can be easily extended to the two-sample, two-population case.

Theorem 8.3: If independent samples of size n_1 and n_2 are drawn at random from two populations, discrete or continuous, with means μ_1 and μ_2 and variances σ_1^2 and σ_2^2, respectively, then the sampling distribution of the differences of means, $\bar{X}_1 - \bar{X}_2$, is approximately normally distributed with mean and variance given by

$$\mu_{\bar{X}_1 - \bar{X}_2} = \mu_1 - \mu_2 \text{ and } \sigma^2_{\bar{X}_1 - \bar{X}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}.$$

Hence,

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

is approximately a standard normal variable.

If both n_1 and n_2 are greater than or equal to 30, the normal approximation for the distribution of $\bar{X}_1 - \bar{X}_2$ is very good when the underlying distributions are not too far away from normal. However, even when n_1 and n_2 are less than 30, the normal approximation is reasonably good except when the populations are decidedly nonnormal. Of course, if both populations are normal, then $\bar{X}_1 - \bar{X}_2$ has a normal distribution no matter what the sizes of n_1 and n_2 are.

The utility of the sampling distribution of the difference between two sample averages is very similar to that described in Case Study 8.1 on page 235 for the case of a single mean. Case Study 8.2 that follows focuses on the use of the difference between two sample means to support (or not) the conjecture that two population means are the same.

Case Study 8.2: **Paint Drying Time**: Two independent experiments are run in which two different types of paint are compared. Eighteen specimens are painted using type A, and the drying time, in hours, is recorded for each. The same is done with type B. The population standard deviations are both known to be 1.0.

Assuming that the mean drying time is equal for the two types of paint, find $P(\bar{X}_A - \bar{X}_B > 1.0)$, where $\bar{X}_A$ and $\bar{X}_B$ are average drying times for samples of size $n_A = n_B = 18$.

***Solution*:** From the sampling distribution of $\bar{X}_A - \bar{X}_B$, we know that the distribution is approximately normal with mean

$$\mu_{\bar{X}_A - \bar{X}_B} = \mu_A - \mu_B = 0$$

and variance

$$\sigma^2_{\bar{X}_A - \bar{X}_B} = \frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B} = \frac{1}{18} + \frac{1}{18} = \frac{1}{9}.$$

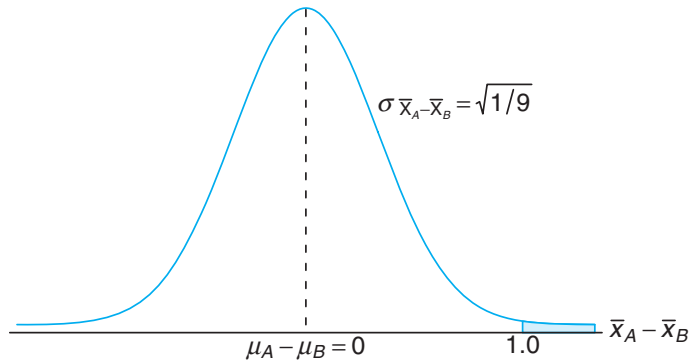


Figure 8.5: Area for Case Study 8.2.

The desired probability is given by the shaded region in Figure 8.5. Corresponding to the value $\bar{X}_A - \bar{X}_B = 1.0$, we have

$$z = \frac{1 - (\mu_A - \mu_B)}{\sqrt{1/9}} = \frac{1 - 0}{\sqrt{1/9}} = 3.0;$$

so

$$P(Z > 3.0) = 1 - P(Z < 3.0) = 1 - 0.9987 = 0.0013.$$

What Do We Learn from Case Study 8.2?

The machinery in the calculation is based on the presumption that $\mu_A = \mu_B$. Suppose, however, that the experiment is actually conducted for the purpose of drawing an inference regarding the equality of μ_A and μ_B, the two population mean drying times. If the two averages differ by as much as 1 hour (or more), this clearly is evidence that would lead one to conclude that the population mean drying time is not equal for the two types of paint. On the other hand, suppose

that the difference in the two sample averages is as small as, say, 15 minutes. If $\mu_A = \mu_B$,

$$P[(\bar{X}_A - \bar{X}_B) > 0.25 \text{ hour}] = P\left(\frac{\bar{X}_A - \bar{X}_B - 0}{\sqrt{1/9}} > \frac{3}{4}\right)$$
$$= P\left(Z > \frac{3}{4}\right) = 1 - P(Z < 0.75) = 1 - 0.7734 = 0.2266.$$

Since this probability is not low, one would conclude that a difference in sample means of 15 minutes can happen by chance (i.e., it happens frequently even though $\mu_A = \mu_B$). As a result, that type of difference in average drying times certainly *is not a clear signal* that $\mu_A \neq \mu_B$.

As we indicated earlier, a more detailed formalism regarding this and other types of statistical inference (e.g., hypothesis testing) will be supplied in future chapters. The Central Limit Theorem and sampling distributions discussed in the next three sections will also play a vital role.

Example 8.6: The television picture tubes of manufacturer A have a mean lifetime of 6.5 years and a standard deviation of 0.9 year, while those of manufacturer B have a mean lifetime of 6.0 years and a standard deviation of 0.8 year. What is the probability that a random sample of 36 tubes from manufacturer A will have a mean lifetime that is at least 1 year more than the mean lifetime of a sample of 49 tubes from manufacturer B?

***Solution*:** We are given the following information:

Population 1	Population 2
$\mu_1 = 6.5$	$\mu_2 = 6.0$
$\sigma_1 = 0.9$	$\sigma_2 = 0.8$
$n_1 = 36$	$n_2 = 49$

If we use Theorem 8.3, the sampling distribution of $\bar{X}_1 - \bar{X}_2$ will be approximately normal and will have a mean and standard deviation

$$\mu_{\bar{X}_1 - \bar{X}_2} = 6.5 - 6.0 = 0.5 \quad \text{and} \quad \sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{0.81}{36} + \frac{0.64}{49}} = 0.189.$$

The probability that the mean lifetime for 36 tubes from manufacturer A will be at least 1 year longer than the mean lifetime for 49 tubes from manufacturer B is given by the area of the shaded region in Figure 8.6. Corresponding to the value $\bar{x}_1 - \bar{x}_2 = 1.0$, we find that

$$z = \frac{1.0 - 0.5}{0.189} = 2.65,$$

and hence

$$P(\bar{X}_1 - \bar{X}_2 \geq 1.0) = P(Z > 2.65) = 1 - P(Z < 2.65)$$
$$= 1 - 0.9960 = 0.0040.$$

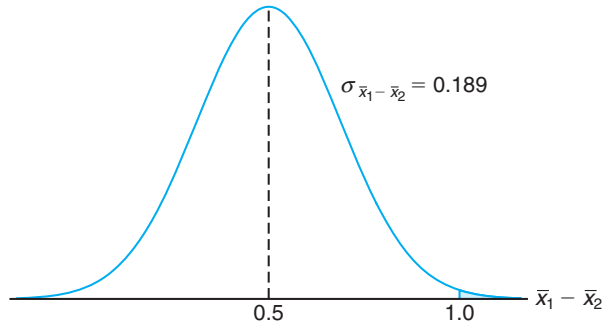


Figure 8.6: Area for Example 8.6.

More on Sampling Distribution of Means—Normal Approximation to the Binomial Distribution

Section 6.5 presented the normal approximation to the binomial distribution at length. Conditions were given on the parameters n and p for which the distribution of a binomial random variable can be approximated by the normal distribution. Examples and exercises reflected the importance of the concept of the "normal approximation." It turns out that the Central Limit Theorem sheds even more light on how and why this approximation works. We certainly know that a binomial random variable is the number X of successes in n independent trials, where the outcome of each trial is binary. We also illustrated in Chapter 1 that the proportion computed in such an experiment is an average of a set of 0s and 1s. Indeed, while the proportion X/n is an average, X is the sum of this set of 0s and 1s, and both X and X/n are approximately normal if n is sufficiently large. Of course, from what we learned in Chapter 6, we know that there are conditions on n and p that affect the quality of the approximation, namely $np \geq 5$ and $nq \geq 5$.

Exercises

8.17 If all possible samples of size 16 are drawn from a normal population with mean equal to 50 and standard deviation equal to 5, what is the probability that a sample mean $\bar{X}$ will fall in the interval from $\mu_{\bar{X}} - 1.9\sigma_{\bar{X}}$ to $\mu_{\bar{X}} - 0.4\sigma_{\bar{X}}$? Assume that the sample means can be measured to any degree of accuracy.

8.18 If the standard deviation of the mean for the sampling distribution of random samples of size 36 from a large or infinite population is 2, how large must the sample size become if the standard deviation is to be reduced to 1.2?

8.19 A certain type of thread is manufactured with a mean tensile strength of 78.3 kilograms and a standard deviation of 5.6 kilograms. How is the variance of the sample mean changed when the sample size is
(a) increased from 64 to 196?
(b) decreased from 784 to 49?

8.20 Given the discrete uniform population

$$f(x) = \begin{cases} \frac{1}{3}, & x = 2, 4, 6, \\ 0, & \text{elsewhere,} \end{cases}$$

find the probability that a random sample of size 54, selected with replacement, will yield a sample mean greater than 4.1 but less than 4.4. Assume the means are measured to the nearest tenth.

8.21 A soft-drink machine is regulated so that the amount of drink dispensed averages 240 milliliters with

a standard deviation of 15 milliliters. Periodically, the machine is checked by taking a sample of 40 drinks and computing the average content. If the mean of the 40 drinks is a value within the interval $\mu_{\bar{X}} \pm 2\sigma_{\bar{X}}$, the machine is thought to be operating satisfactorily; otherwise, adjustments are made. In Section 8.3, the company official found the mean of 40 drinks to be $\bar{x} = 236$ milliliters and concluded that the machine needed no adjustment. Was this a reasonable decision?

8.22 The heights of 1000 students are approximately normally distributed with a mean of 174.5 centimeters and a standard deviation of 6.9 centimeters. Suppose 200 random samples of size 25 are drawn from this population and the means recorded to the nearest tenth of a centimeter. Determine

(a) the mean and standard deviation of the sampling distribution of $\bar{X}$;

(b) the number of sample means that fall between 172.5 and 175.8 centimeters inclusive;

(c) the number of sample means falling below 172.0 centimeters.

8.23 The random variable X, representing the number of cherries in a cherry puff, has the following probability distribution:

x	4	5	6	7
$P(X = x)$	0.2	0.4	0.3	0.1

(a) Find the mean μ and the variance σ^2 of X.

(b) Find the mean $\mu_{\bar{X}}$ and the variance $\sigma^2_{\bar{X}}$ of the mean $\bar{X}$ for random samples of 36 cherry puffs.

(c) Find the probability that the average number of cherries in 36 cherry puffs will be less than 5.5.

8.24 If a certain machine makes electrical resistors having a mean resistance of 40 ohms and a standard deviation of 2 ohms, what is the probability that a random sample of 36 of these resistors will have a combined resistance of more than 1458 ohms?

8.25 The average life of a bread-making machine is 7 years, with a standard deviation of 1 year. Assuming that the lives of these machines follow approximately a normal distribution, find

(a) the probability that the mean life of a random sample of 9 such machines falls between 6.4 and 7.2 years;

(b) the value of x to the right of which 15% of the means computed from random samples of size 9 would fall.

8.26 The amount of time that a drive-through bank teller spends on a customer is a random variable with a mean $\mu = 3.2$ minutes and a standard deviation $\sigma = 1.6$ minutes. If a random sample of 64 customers is observed, find the probability that their mean time at the teller's window is

(a) at most 2.7 minutes;

(b) more than 3.5 minutes;

(c) at least 3.2 minutes but less than 3.4 minutes.

8.27 In a chemical process, the amount of a certain type of impurity in the output is difficult to control and is thus a random variable. Speculation is that the population mean amount of the impurity is 0.20 gram per gram of output. It is known that the standard deviation is 0.1 gram per gram. An experiment is conducted to gain more insight regarding the speculation that $\mu = 0.2$. The process is run on a lab scale 50 times and the sample average $\bar{x}$ turns out to be 0.23 gram per gram. Comment on the speculation that the mean amount of impurity is 0.20 gram per gram. Make use of the Central Limit Theorem in your work.

8.28 A random sample of size 25 is taken from a normal population having a mean of 80 and a standard deviation of 5. A second random sample of size 36 is taken from a different normal population having a mean of 75 and a standard deviation of 3. Find the probability that the sample mean computed from the 25 measurements will exceed the sample mean computed from the 36 measurements by at least 3.4 but less than 5.9. Assume the difference of the means to be measured to the nearest tenth.

8.29 The distribution of heights of a certain breed of terrier has a mean of 72 centimeters and a standard deviation of 10 centimeters, whereas the distribution of heights of a certain breed of poodle has a mean of 28 centimeters with a standard deviation of 5 centimeters. Assuming that the sample means can be measured to any degree of accuracy, find the probability that the sample mean for a random sample of heights of 64 terriers exceeds the sample mean for a random sample of heights of 100 poodles by at most 44.2 centimeters.

8.30 The mean score for freshmen on an aptitude test at a certain college is 540, with a standard deviation of 50. Assume the means to be measured to any degree of accuracy. What is the probability that two groups selected at random, consisting of 32 and 50 students, respectively, will differ in their mean scores by

(a) more than 20 points?

(b) an amount between 5 and 10 points?

8.31 Consider Case Study 8.2 on page 238. Suppose 18 specimens were used for each type of paint in an experiment and $\bar{x}_A - \bar{x}_B$, the actual difference in mean drying time, turned out to be 1.0.

(a) Does this seem to be a reasonable result if the

two population mean drying times truly are equal? Make use of the result in the solution to Case Study 8.2.

(b) If someone did the experiment 10,000 times under the condition that $\mu_A = \mu_B$, in how many of those 10,000 experiments would there be a difference $\bar{x}_A - \bar{x}_B$ that was as large as (or larger than) 1.0?

8.32 Two different box-filling machines are used to fill cereal boxes on an assembly line. The critical measurement influenced by these machines is the weight of the product in the boxes. Engineers are quite certain that the variance of the weight of product is $\sigma^2 = 1$ ounce. Experiments are conducted using both machines with sample sizes of 36 each. The sample averages for machines A and B are $\bar{x}_A = 4.5$ ounces and $\bar{x}_B = 4.7$ ounces. Engineers are surprised that the two sample averages for the filling machines are so different.

(a) Use the Central Limit Theorem to determine

$$P(\bar{X}_B - \bar{X}_A \geq 0.2)$$

under the condition that $\mu_A = \mu_B$.

(b) Do the aforementioned experiments seem to, in any way, strongly support a conjecture that the population means for the two machines are different? Explain using your answer in (a).

8.33 The chemical benzene is highly toxic to humans. However, it is used in the manufacture of many medicine dyes, leather, and coverings. Government regulations dictate that for any production process involving benzene, the water in the output of the process must not exceed 7950 parts per million (ppm) of benzene. For a particular process of concern, the water sample was collected by a manufacturer 25 times randomly and the sample average $\bar{x}$ was 7960 ppm. It is known from historical data that the standard deviation σ is 100 ppm.

(a) What is the probability that the sample average in this experiment would exceed the government limit if the population mean is equal to the limit? Use the Central Limit Theorem.

(b) Is an observed $\bar{x} = 7960$ in this experiment firm evidence that the population mean for the process exceeds the government limit? Answer your question by computing

$$P(\bar{X} \geq 7960 \mid \mu = 7950).$$

Assume that the distribution of benzene concentration is normal.

8.34 Two alloys A and B are being used to manufacture a certain steel product. An experiment needs to be designed to compare the two in terms of maximum load capacity in tons (the maximum weight that can be tolerated without breaking). It is known that the two standard deviations in load capacity are equal at 5 tons each. An experiment is conducted in which 30 specimens of each alloy (A and B) are tested and the results recorded as follows:

$$\bar{x}_A = 49.5, \quad \bar{x}_B = 45.5; \qquad \bar{x}_A - \bar{x}_B = 4.$$

The manufacturers of alloy A are convinced that this evidence shows conclusively that $\mu_A > \mu_B$ and strongly supports the claim that their alloy is superior. Manufacturers of alloy B claim that the experiment could easily have given $\bar{x}_A - \bar{x}_B = 4$ *even if* the two population means are equal. In other words, "the results are inconclusive!"

(a) Make an argument that manufacturers of alloy B are wrong. Do it by computing

$$P(\bar{X}_A - \bar{X}_B > 4 \mid \mu_A = \mu_B).$$

(b) Do you think these data strongly support alloy A?

8.35 Consider the situation described in Example 8.4 on page 234. Do these results prompt you to question the premise that $\mu = 800$ hours? Give a probabilistic result that indicates how *rare* an event $\bar{X} \leq 775$ is when $\mu = 800$. On the other hand, how rare would it be if μ truly were, say, 760 hours?

8.36 Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution that can take on only positive values. Use the Central Limit Theorem to produce an argument that if n is sufficiently large, then $Y = X_1 X_2 \cdots X_n$ has approximately a lognormal distribution.

8.5 Sampling Distribution of S^2

In the preceding section we learned about the sampling distribution of $\bar{X}$. The Central Limit Theorem allowed us to make use of the fact that

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

tends toward $N(0, 1)$ as the sample size grows large. *Sampling distributions of important statistics* allow us to learn information about parameters. Usually, the parameters are the counterpart to the statistics in question. For example, if an engineer is interested in the population mean resistance of a certain type of resistor, the sampling distribution of $\bar{X}$ will be exploited once the sample information is gathered. On the other hand, if the variability in resistance is to be studied, clearly the sampling distribution of S^2 will be used in learning about the parametric counterpart, the population variance σ^2.

If a random sample of size n is drawn from a normal population with mean μ and variance σ^2, and the sample variance is computed, we obtain a value of the statistic S^2. We shall proceed to consider the distribution of the statistic $(n-1)S^2/\sigma^2$.

By the addition and subtraction of the sample mean $\bar{X}$, it is easy to see that

$$\begin{aligned}\sum_{i=1}^{n}(X_i-\mu)^2 &= \sum_{i=1}^{n}[(X_i-\bar{X})+(\bar{X}-\mu)]^2\\ &= \sum_{i=1}^{n}(X_i-\bar{X})^2+\sum_{i=1}^{n}(\bar{X}-\mu)^2+2(\bar{X}-\mu)\sum_{i=1}^{n}(X_i-\bar{X})\\ &= \sum_{i=1}^{n}(X_i-\bar{X})^2+n(\bar{X}-\mu)^2.\end{aligned}$$

Dividing each term of the equality by σ^2 and substituting $(n-1)S^2$ for $\sum_{i=1}^{n}(X_i-\bar{X})^2$, we obtain

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2=\frac{(n-1)S^2}{\sigma^2}+\frac{(\bar{X}-\mu)^2}{\sigma^2/n}.$$

Now, according to Corollary 7.1 on page 222, we know that

$$\sum_{i=1}^{n}\frac{(X_i-\mu)^2}{\sigma^2}$$

is a chi-squared random variable with n degrees of freedom. We have a chi-squared random variable with n degrees of freedom partitioned into two components. Note that in Section 6.7 we showed that a chi-squared distribution is a special case of a gamma distribution. The second term on the right-hand side is Z^2, which is a chi-squared random variable with 1 degree of freedom, and it turns out that $(n-1)S^2/\sigma^2$ is a chi-squared random variable with $n-1$ degree of freedom. We formalize this in the following theorem.

Theorem 8.4: If S^2 is the variance of a random sample of size n taken from a normal population having the variance σ^2, then the statistic

$$\chi^2=\frac{(n-1)S^2}{\sigma^2}=\sum_{i=1}^{n}\frac{(X_i-\bar{X})^2}{\sigma^2}$$

has a chi-squared distribution with $v=n-1$ degrees of freedom.

The values of the random variable χ^2 are calculated from each sample by the

formula

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}.$$

The probability that a random sample produces a χ^2 value greater than some specified value is equal to the area under the curve to the right of this value. It is customary to let χ^2_α represent the χ^2 value above which we find an area of α. This is illustrated by the shaded region in Figure 8.7.

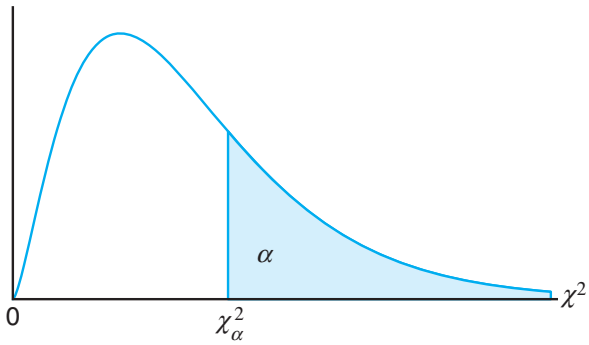


Figure 8.7: The chi-squared distribution.

Table A.5 gives values of χ^2_α for various values of α and v. The areas, α, are the column headings; the degrees of freedom, v, are given in the left column; and the table entries are the χ^2 values. Hence, the χ^2 value with 7 degrees of freedom, leaving an area of 0.05 to the right, is $\chi^2_{0.05} = 14.067$. Owing to lack of symmetry, we must also use the tables to find $\chi^2_{0.95} = 2.167$ for $v = 7$.

Exactly 95% of a chi-squared distribution lies between $\chi^2_{0.975}$ and $\chi^2_{0.025}$. A χ^2 value falling to the right of $\chi^2_{0.025}$ is not likely to occur unless our assumed value of σ^2 is too small. Similarly, a χ^2 value falling to the left of $\chi^2_{0.975}$ is unlikely unless our assumed value of σ^2 is too large. In other words, it is possible to have a χ^2 value to the left of $\chi^2_{0.975}$ or to the right of $\chi^2_{0.025}$ when σ^2 is correct, but if this should occur, it is more probable that the assumed value of σ^2 is in error.

Example 8.7: A manufacturer of car batteries guarantees that the batteries will last, on average, 3 years with a standard deviation of 1 year. If five of these batteries have lifetimes of 1.9, 2.4, 3.0, 3.5, and 4.2 years, should the manufacturer still be convinced that the batteries have a standard deviation of 1 year? Assume that the battery lifetime follows a normal distribution.

***Solution*:** We first find the sample variance using Theorem 8.1,

$$s^2 = \frac{(5)(48.26) - (15)^2}{(5)(4)} = 0.815.$$

Then

$$\chi^2 = \frac{(4)(0.815)}{1} = 3.26$$

is a value from a chi-squared distribution with 4 degrees of freedom. Since 95% of the χ^2 values with 4 degrees of freedom fall between 0.484 and 11.143, the computed value with $\sigma^2 = 1$ is reasonable, and therefore the manufacturer has no reason to suspect that the standard deviation is other than 1 year.

Degrees of Freedom as a Measure of Sample Information

Recall from Corollary 7.1 in Section 7.3 that

$$\sum_{i=1}^{n} \frac{(X_i - \mu)^2}{\sigma^2}$$

has a χ^2-distribution with n *degrees of freedom*. Note also Theorem 8.4, which indicates that the random variable

$$\frac{(n-1)S^2}{\sigma^2} = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{\sigma^2}$$

has a χ^2-distribution with $n-1$ *degrees of freedom*. The reader may also recall that the term *degrees of freedom*, used in this identical context, is discussed in Chapter 1.

As we indicated earlier, the proof of Theorem 8.4 will not be given. However, the reader can view Theorem 8.4 as indicating that when μ is not known and one considers the distribution of

$$\sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{\sigma^2},$$

there is **1 less degree of freedom**, or a degree of freedom is lost in the estimation of μ (i.e., when μ is replaced by $\bar{x}$). In other words, there are n degrees of freedom, or independent *pieces of information*, in the random sample from the normal distribution. When the data (the values in the sample) are used to compute the mean, there is 1 less degree of freedom in the information used to estimate σ^2.

8.6 t-Distribution

In Section 8.4, we discussed the utility of the Central Limit Theorem. Its applications revolve around inferences on a population mean or the difference between two population means. Use of the Central Limit Theorem and the normal distribution is certainly helpful in this context. However, it was assumed that the population standard deviation is known. This assumption may not be unreasonable in situations where the engineer is quite familiar with the system or process. However, in many experimental scenarios, knowledge of σ is certainly no more reasonable than knowledge of the population mean μ. Often, in fact, an estimate of σ must be supplied by the same sample information that produced the sample average $\bar{x}$. As a result, a natural statistic to consider to deal with inferences on μ is

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}},$$

since S is the sample analog to σ. If the sample size is small, the values of S^2 fluctuate considerably from sample to sample (see Exercise 8.43 on page 259) and the distribution of T deviates appreciably from that of a standard normal distribution.

If the sample size is large enough, say $n \geq 30$, the distribution of T does not differ considerably from the standard normal. However, for $n < 30$, it is useful to deal with the exact distribution of T. In developing the sampling distribution of T, we shall assume that our random sample was selected from a normal population. We can then write

$$T = \frac{(\bar{X} - \mu)/(\sigma/\sqrt{n})}{\sqrt{S^2/\sigma^2}} = \frac{Z}{\sqrt{V/(n-1)}},$$

where

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

has the standard normal distribution and

$$V = \frac{(n-1)S^2}{\sigma^2}$$

has a chi-squared distribution with $v = n-1$ degrees of freedom. In sampling from normal populations, we can show that $\bar{X}$ and S^2 are independent, and consequently so are Z and V. The following theorem gives the definition of a random variable T as a function of Z (standard normal) and χ^2. For completeness, the density function of the t-distribution is given.

Theorem 8.5: Let Z be a standard normal random variable and V a chi-squared random variable with v degrees of freedom. If Z and V are independent, then the distribution of the random variable T, where

$$T = \frac{Z}{\sqrt{V/v}},$$

is given by the density function

$$h(t) = \frac{\Gamma[(v+1)/2]}{\Gamma(v/2)\sqrt{\pi v}} \left(1 + \frac{t^2}{v}\right)^{-(v+1)/2}, \quad -\infty < t < \infty.$$

This is known as the ***t*-distribution** with v degrees of freedom.

From the foregoing and the theorem above we have the following corollary.

Corollary 8.1: Let $X_1, X_2, \ldots, X_n$ be independent random variables that are all normal with mean μ and standard deviation σ. Let

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad \text{and} \qquad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

Then the random variable $T = \frac{\bar{X}-\mu}{S/\sqrt{n}}$ has a t-distribution with $v = n - 1$ degrees of freedom.

The probability distribution of T was first published in 1908 in a paper written by W. S. Gosset. At the time, Gosset was employed by an Irish brewery that prohibited publication of research by members of its staff. To circumvent this restriction, he published his work secretly under the name "Student." Consequently, the distribution of T is usually called the Student t-distribution or simply the t-distribution. In deriving the equation of this distribution, Gosset assumed that the samples were selected from a normal population. Although this would seem to be a very restrictive assumption, it can be shown that nonnormal populations possessing nearly bell-shaped distributions will still provide values of T that approximate the t-distribution very closely.

What Does the *t*-Distribution Look Like?

The distribution of T is similar to the distribution of Z in that they both are symmetric about a mean of zero. Both distributions are bell shaped, but the t-distribution is more variable, owing to the fact that the T-values depend on the fluctuations of two quantities, $\bar{X}$ and S^2, whereas the Z-values depend only on the changes in $\bar{X}$ from sample to sample. The distribution of T differs from that of Z in that the variance of T depends on the sample size n and is always greater than 1. Only when the sample size $n \to \infty$ will the two distributions become the same. In Figure 8.8, we show the relationship between a standard normal distribution ($v = \infty$) and t-distributions with 2 and 5 degrees of freedom. The percentage points of the t-distribution are given in Table A.4.

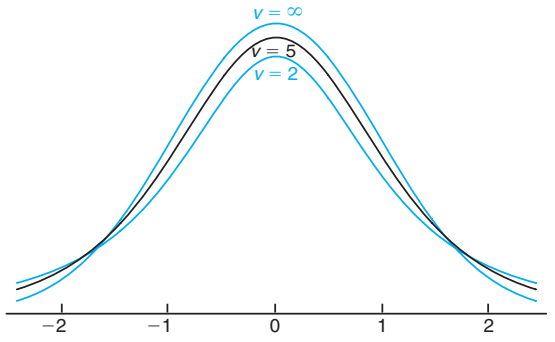


Figure 8.8: The t-distribution curves for $v = 2, 5,$ and ∞.

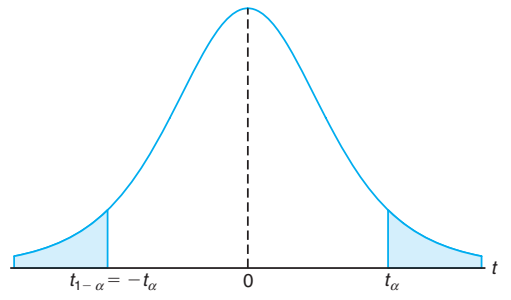


Figure 8.9: Symmetry property (about 0) of the t-distribution.

It is customary to let t_α represent the t-value above which we find an area equal to α. Hence, the t-value with 10 degrees of freedom leaving an area of 0.025 to the right is $t = 2.228$. Since the t-distribution is symmetric about a mean of zero, we have $t_{1-\alpha} = -t_\alpha$; that is, the t-value leaving an area of $1 - \alpha$ to the right and therefore an area of α to the left is equal to the negative t-value that leaves an area of α in the right tail of the distribution (see Figure 8.9). That is, $t_{0.95} = -t_{0.05}$, $t_{0.99} = -t_{0.01}$, and so forth.

Example 8.8: The t-value with $v = 14$ degrees of freedom that leaves an area of 0.025 to the left, and therefore an area of 0.975 to the right, is

$$t_{0.975} = -t_{0.025} = -2.145.$$

Example 8.9: Find $P(-t_{0.025} < T < t_{0.05})$.

***Solution*:** Since $t_{0.05}$ leaves an area of 0.05 to the right, and $-t_{0.025}$ leaves an area of 0.025 to the left, we find a total area of

$$1 - 0.05 - 0.025 = 0.925$$

between $-t_{0.025}$ and $t_{0.05}$. Hence

$$P(-t_{0.025} < T < t_{0.05}) = 0.925.$$

Example 8.10: Find k such that $P(k < T < -1.761) = 0.045$ for a random sample of size 15 selected from a normal distribution and $\frac{\overline{X}-\mu}{s/\sqrt{n}}$.

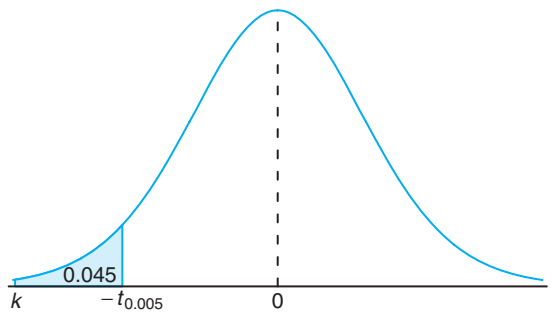


Figure 8.10: The t-values for Example 8.10.

***Solution*:** From Table A.4 we note that 1.761 corresponds to $t_{0.05}$ when $v = 14$. Therefore, $-t_{0.05} = -1.761$. Since k in the original probability statement is to the left of $-t_{0.05} = -1.761$, let $k = -t_\alpha$. Then, from Figure 8.10, we have

$$0.045 = 0.05 - \alpha, \text{ or } \alpha = 0.005.$$

Hence, from Table A.4 with $v = 14$,

$$k = -t_{0.005} = -2.977 \text{ and } P(-2.977 < T < -1.761) = 0.045.$$

Exactly 95% of the values of a t-distribution with $v = n - 1$ degrees of freedom lie between $-t_{0.025}$ and $t_{0.025}$. Of course, there are other t-values that contain 95% of the distribution, such as $-t_{0.02}$ and $t_{0.03}$, but these values do not appear in Table A.4, and furthermore, the shortest possible interval is obtained by choosing t-values that leave exactly the same area in the two tails of our distribution. A t-value that falls below $-t_{0.025}$ or above $t_{0.025}$ would tend to make us believe either that a very rare event has taken place or that our assumption about μ is in error. Should this happen, we shall make the the decision that our assumed value of μ is in error. In fact, a t-value falling below $-t_{0.01}$ or above $t_{0.01}$ would provide even stronger evidence that our assumed value of μ is quite unlikely. General procedures for testing claims concerning the value of the parameter μ will be treated in Chapter 10. A preliminary look into the foundation of these procedure is illustrated by the following example.

Example 8.11: A chemical engineer claims that the population mean yield of a certain batch process is 500 grams per milliliter of raw material. To check this claim he samples 25 batches each month. If the computed t-value falls between $-t_{0.05}$ and $t_{0.05}$, he is satisfied with this claim. What conclusion should he draw from a sample that has a mean $\bar{x} = 518$ grams per milliliter and a sample standard deviation $s = 40$ grams? Assume the distribution of yields to be approximately normal.

***Solution*:** From Table A.4 we find that $t_{0.05} = 1.711$ for 24 degrees of freedom. Therefore, the engineer can be satisfied with his claim if a sample of 25 batches yields a t-value between -1.711 and 1.711. If μ= 500, then

$$t = \frac{518 - 500}{40/\sqrt{25}} = 2.25,$$

a value well above 1.711. The probability of obtaining a t-value, with $v = 24$, equal to or greater than 2.25 is approximately 0.02. If $\mu > 500$, the value of t computed from the sample is more reasonable. Hence, the engineer is likely to conclude that the process produces a better product than he thought. ◢

What Is the *t*-Distribution Used For?

The t-distribution is used extensively in problems that deal with inference about the population mean (as illustrated in Example 8.11) or in problems that involve comparative samples (i.e., in cases where one is trying to determine if means from two samples are significantly different). The use of the distribution will be extended in Chapters 9, 10, 11, and 12. The reader should note that use of the t-distribution for the statistic

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

requires that $X_1, X_2, \ldots, X_n$ be normal. The use of the t-distribution and the sample size consideration do not relate to the Central Limit Theorem. The use of the standard normal distribution rather than T for $n \geq 30$ merely implies that S is a sufficiently good estimator of σ in this case. In chapters that follow the t-distribution finds extensive usage.

close" to the population mean (wherever it is!), so we wish

$$P(|\bar{X} - \mu| > 0.05) = 0.99.$$

What sample size is required?

8.74 Suppose a filling machine is used to fill cartons with a liquid product. The specification that is strictly enforced for the filling machine is 9 ± 1.5 oz. If any carton is produced with weight outside these bounds, it is considered by the supplier to be defective. It is hoped that at least 99% of cartons will meet these specifications. With the conditions $\mu = 9$ and $\sigma = 1$, what proportion of cartons from the process are defective? If changes are made to reduce variability, what must σ be reduced to in order to meet specifications with probability 0.99? Assume a normal distribution for the weight.

8.75 Consider the situation in Review Exercise 8.74. Suppose a considerable effort is conducted to "tighten" the variability in the system. Following the effort, a random sample of size 40 is taken from the new assembly line and the sample variance is $s^2 = 0.188$ ounces^2. Do we have strong numerical evidence that σ^2 has been reduced below 1.0? Consider the probability

$$P(S^2 \leq 0.188 \mid \sigma^2 = 1.0),$$

and give your conclusion.

8.76 Group Project: The class should be divided into groups of four people. The four students in each group should go to the college gym or a local fitness center. The students should ask each person who comes through the door his or her height in inches. Each group will then divide the height data by gender and work together to answer the following questions.

(a) Construct a normal quantile-quantile plot of the data. Based on the plot, do the data appear to follow a normal distribution?

(b) Use the estimated sample variance as the true variance for each gender. Assume that the population mean height for male students is actually three inches larger than that of female students. What is the probability that the average height of the male students will be 4 inches larger than that of the female students in your sample?

(c) What factors could render these results misleading?

8.9 Potential Misconceptions and Hazards; Relationship to Material in Other Chapters

The Central Limit Theorem is one of the most powerful tools in all of statistics, and even though this chapter is relatively short, it contains a wealth of fundamental information about tools that will be used throughout the balance of the text.

The notion of a sampling distribution is one of the most important fundamental concepts in all of statistics, and the student at this point in his or her training should gain a clear understanding of it before proceeding beyond this chapter. All chapters that follow will make considerable use of sampling distributions. Suppose one wants to use the statistic $\bar{X}$ to draw inferences about the population mean μ. This will be done by using the observed value $\bar{x}$ from a single sample of size n. Then any inference made must be accomplished by taking into account not just the single value but rather the theoretical structure, or **distribution of all $\bar{x}$ values that could be observed from samples of size n.** Thus, the concept of a *sampling distribution* comes to the surface. This distribution is the basis for the Central Limit Theorem. The t, χ^2, and F-distributions are also used in the context of sampling distributions. For example, the t-distribution, pictured in Figure 8.8, represents the structure that occurs if all of the values of $\frac{\bar{x}-\mu}{s/\sqrt{n}}$ are formed, where $\bar{x}$ and s are taken from samples of size n from a $n(x; \mu, \sigma)$ distribution. Similar remarks can be made about χ^2 and F, and the reader should not forget that the sample information forming the statistics for all of these distributions is the normal. So it can be said that **where there is a t, F, or χ^2, the source was a sample from a normal distribution**.

The three distributions described above may appear to have been introduced in a rather self-contained fashion with no indication of what they are about. However, they will appear in practical problem-solving throughout the balance of the text.

Now, there are three things that one must bear in mind, lest confusion set in regarding these fundamental sampling distributions:

(i) One cannot use the Central Limit Theorem unless σ is known. When σ is not known, it should be replaced by s, the sample standard deviation, in order to use the Central Limit Theorem.

(ii) The T statistic is **not** a result of the Central Limit Theorem and $x_1, x_2, \ldots, x_n$ must come from a $n(x; \mu, \sigma)$ distribution in order for $\frac{\bar{x}-\mu}{s/\sqrt{n}}$ to be a t-distribution; s is, of course, merely an estimate of σ.

(iii) While the notion of **degrees of freedom** is new at this point, the concept should be very intuitive, since it is reasonable that the nature of the distribution of S and also t should depend on the amount of information in the sample $x_1, x_2, \ldots, x_n$.

Chapter 9

One- and Two-Sample Estimation Problems

9.1 Introduction

In previous chapters, we emphasized sampling properties of the sample mean and variance. We also emphasized displays of data in various forms. The purpose of these presentations is to build a foundation that allows us to draw conclusions about the population parameters from experimental data. For example, the Central Limit Theorem provides information about the distribution of the sample mean $\bar{X}$. The distribution involves the population mean μ. Thus, any conclusions concerning μ drawn from an observed sample average must depend on knowledge of this sampling distribution. Similar comments apply to S^2 and σ^2. Clearly, any conclusions we draw about the variance of a normal distribution will likely involve the sampling distribution of S^2.

In this chapter, we begin by formally outlining the purpose of statistical inference. We follow this by discussing the problem of **estimation of population parameters**. We confine our formal developments of specific estimation procedures to problems involving one and two samples.

9.2 Statistical Inference

In Chapter 1, we discussed the general philosophy of formal statistical inference. **Statistical inference** consists of those methods by which one makes inferences or generalizations about a population. The trend today is to distinguish between the **classical method** of estimating a population parameter, whereby inferences are based strictly on information obtained from a random sample selected from the population, and the **Bayesian method**, which utilizes prior subjective knowledge about the probability distribution of the unknown parameters in conjunction with the information provided by the sample data. Throughout most of this chapter, we shall use classical methods to estimate unknown population parameters such as the mean, the proportion, and the variance by computing statistics from random

samples and applying the theory of sampling distributions, much of which was covered in Chapter 8. Bayesian estimation will be discussed in Chapter 18.

Statistical inference may be divided into two major areas: **estimation** and **tests of hypotheses**. We treat these two areas separately, dealing with theory and applications of estimation in this chapter and hypothesis testing in Chapter 10. To distinguish clearly between the two areas, consider the following examples. A candidate for public office may wish to estimate the true proportion of voters favoring him by obtaining opinions from a random sample of 100 eligible voters. The fraction of voters in the sample favoring the candidate could be used as an estimate of the true proportion in the population of voters. A knowledge of the sampling distribution of a proportion enables one to establish the degree of accuracy of such an estimate. This problem falls in the area of estimation.

Now consider the case in which one is interested in finding out whether brand A floor wax is more scuff-resistant than brand B floor wax. He or she might hypothesize that brand A is better than brand B and, after proper testing, accept or reject this hypothesis. In this example, we do not attempt to estimate a parameter, but instead we try to arrive at a correct decision about a prestated hypothesis. Once again we are dependent on sampling theory and the use of data to provide us with some measure of accuracy for our decision.

9.3 Classical Methods of Estimation

A **point estimate** of some population parameter θ is a single value $\hat{\theta}$ of a statistic $\hat{\Theta}$. For example, the value $\bar{x}$ of the statistic $\bar{X}$, computed from a sample of size n, is a point estimate of the population parameter μ. Similarly, $\hat{p} = x/n$ is a point estimate of the true proportion p for a binomial experiment.

An estimator is not expected to estimate the population parameter without error. We do not expect $\bar{X}$ to estimate μ exactly, but we certainly hope that it is not far off. For a particular sample, it is possible to obtain a closer estimate of μ by using the sample median $\tilde{X}$ as an estimator. Consider, for instance, a sample consisting of the values 2, 5, and 11 from a population whose mean is 4 but is supposedly unknown. We would estimate μ to be $\bar{x} = 6$, using the sample mean as our estimate, or $\tilde{x} = 5$, using the sample median as our estimate. In this case, the estimator $\tilde{X}$ produces an estimate closer to the true parameter than does the estimator $\bar{X}$. On the other hand, if our random sample contains the values 2, 6, and 7, then $\bar{x} = 5$ and $\tilde{x} = 6$, so $\bar{X}$ is the better estimator. Not knowing the true value of μ, we must decide in advance whether to use $\bar{X}$ or $\tilde{X}$ as our estimator.

Unbiased Estimator

What are the desirable properties of a "good" decision function that would influence us to choose one estimator rather than another? Let $\hat{\Theta}$ be an estimator whose value $\hat{\theta}$ is a point estimate of some unknown population parameter θ. Certainly, we would like the sampling distribution of $\hat{\Theta}$ to have a mean equal to the parameter estimated. An estimator possessing this property is said to be **unbiased**.

Definition 9.1: A statistic $\hat{\Theta}$ is said to be an **unbiased estimator** of the parameter θ if

$$\mu_{\hat{\Theta}} = E(\hat{\Theta}) = \theta.$$

Example 9.1: Show that S^2 is an unbiased estimator of the parameter σ^2.

***Solution*:** In Section 8.5 on page 244, we showed that

$$\sum_{i=1}^{n}(X_i - \bar{X})^2 = \sum_{i=1}^{n}(X_i - \mu)^2 - n(\bar{X} - \mu)^2.$$

Now

$$E(S^2) = E\left[\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2\right]$$
$$= \frac{1}{n-1}\left[\sum_{i=1}^{n}E(X_i - \mu)^2 - nE(\bar{X} - \mu)^2\right] = \frac{1}{n-1}\left(\sum_{i=1}^{n}\sigma_{X_i}^2 - n\sigma_{\bar{X}}^2\right).$$

However,

$$\sigma_{X_i}^2 = \sigma^2, \text{ for } i = 1, 2, \ldots, n, \text{ and } \sigma_{\bar{X}}^2 = \frac{\sigma^2}{n}.$$

Therefore,

$$E(S^2) = \frac{1}{n-1}\left(n\sigma^2 - n\frac{\sigma^2}{n}\right) = \sigma^2.$$

Although S^2 is an unbiased estimator of σ^2, S, on the other hand, is usually a biased estimator of σ, with the bias becoming insignificant for large samples. This example illustrates **why we divide by $n - 1$** rather than n when the variance is estimated.

Variance of a Point Estimator

If $\hat{\Theta}_1$ and $\hat{\Theta}_2$ are two unbiased estimators of the same population parameter θ, we want to choose the estimator whose sampling distribution has the smaller variance. Hence, if $\sigma_{\hat{\theta}_1}^2 < \sigma_{\hat{\theta}_2}^2$, we say that $\hat{\Theta}_1$ is a **more efficient estimator** of θ than $\hat{\Theta}_2$.

Definition 9.2: If we consider all possible unbiased estimators of some parameter θ, the one with the smallest variance is called the **most efficient estimator** of θ.

Figure 9.1 illustrates the sampling distributions of three different estimators, $\hat{\Theta}_1$, $\hat{\Theta}_2$, and $\hat{\Theta}_3$, all estimating θ. It is clear that only $\hat{\Theta}_1$ and $\hat{\Theta}_2$ are unbiased, since their distributions are centered at θ. The estimator $\hat{\Theta}_1$ has a smaller variance than $\hat{\Theta}_2$ and is therefore more efficient. Hence, our choice for an estimator of θ, among the three considered, would be $\hat{\Theta}_1$.

For normal populations, one can show that both $\bar{X}$ and $\tilde{X}$ are unbiased estimators of the population mean μ, but the variance of $\bar{X}$ is smaller than the variance

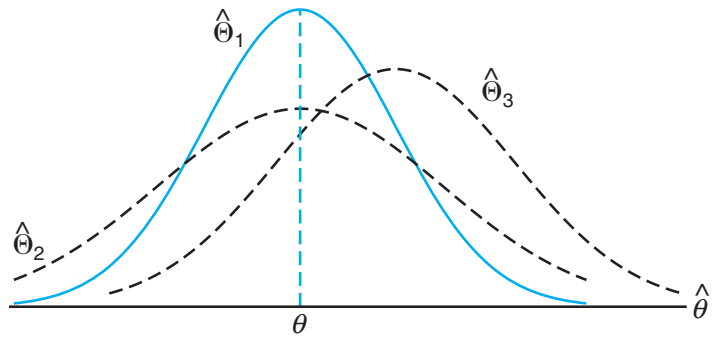


Figure 9.1: Sampling distributions of different estimators of θ.

of $\tilde{X}$. Thus, both estimates $\bar{x}$ and $\tilde{x}$ will, on average, equal the population mean μ, but $\bar{x}$ is likely to be closer to μ for a given sample, and thus $\bar{X}$ is more efficient than $\tilde{X}$.

Interval Estimation

Even the most efficient unbiased estimator is unlikely to estimate the population parameter exactly. It is true that estimation accuracy increases with large samples, but there is still no reason we should expect a **point estimate** from a given sample to be exactly equal to the population parameter it is supposed to estimate. There are many situations in which it is preferable to determine an interval within which we would expect to find the value of the parameter. Such an interval is called an **interval estimate**.

An interval estimate of a population parameter θ is an interval of the form $\hat{\theta}_L < \theta < \hat{\theta}_U$, where $\hat{\theta}_L$ and $\hat{\theta}_U$ depend on the value of the statistic $\hat{\Theta}$ for a particular sample and also on the sampling distribution of $\hat{\Theta}$. For example, a random sample of SAT verbal scores for students in the entering freshman class might produce an interval from 530 to 550, within which we expect to find the true average of all SAT verbal scores for the freshman class. The values of the endpoints, 530 and 550, will depend on the computed sample mean $\bar{x}$ and the sampling distribution of $\bar{X}$. As the sample size increases, we know that $\sigma^2_{\bar{X}} = \sigma^2/n$ decreases, and consequently our estimate is likely to be closer to the parameter μ, resulting in a shorter interval. Thus, the interval estimate indicates, by its length, the accuracy of the point estimate. An engineer will gain some insight into the population proportion defective by taking a sample and computing the *sample proportion defective*. But an interval estimate might be more informative.

Interpretation of Interval Estimates

Since different samples will generally yield different values of $\hat{\Theta}$ and, therefore, different values for $\hat{\theta}_L$ and $\hat{\theta}_U$, these endpoints of the interval are values of corresponding random variables $\hat{\Theta}_L$ and $\hat{\Theta}_U$. From the sampling distribution of $\hat{\Theta}$ we shall be able to determine $\hat{\Theta}_L$ and $\hat{\Theta}_U$ such that $P(\hat{\Theta}_L < \theta < \hat{\Theta}_U)$ is equal to any

positive fractional value we care to specify. If, for instance, we find $\hat{\Theta}_L$ and $\hat{\Theta}_U$ such that

$$P(\hat{\Theta}_L < \theta < \hat{\Theta}_U) = 1 - \alpha,$$

for $0 < \alpha < 1$, then we have a probability of $1-\alpha$ of selecting a random sample that will produce an interval containing θ. The interval $\hat{\theta}_L < \theta < \hat{\theta}_U$, computed from the selected sample, is called a $100(1-\alpha)\%$ **confidence interval**, the fraction $1-\alpha$ is called the **confidence coefficient** or the **degree of confidence**, and the endpoints, $\hat{\theta}_L$ and $\hat{\theta}_U$, are called the lower and upper **confidence limits**. Thus, when $\alpha = 0.05$, we have a 95% confidence interval, and when $\alpha = 0.01$, we obtain a wider 99% confidence interval. The wider the confidence interval is, the more confident we can be that the interval contains the unknown parameter. Of course, it is better to be 95% confident that the average life of a certain television transistor is between 6 and 7 years than to be 99% confident that it is between 3 and 10 years. Ideally, we prefer a short interval with a high degree of confidence. Sometimes, restrictions on the size of our sample prevent us from achieving short intervals without sacrificing some degree of confidence.

In the sections that follow, we pursue the notions of point and interval estimation, with each section presenting a different special case. The reader should notice that while point and interval estimation represent different approaches to gaining information regarding a parameter, they are related in the sense that confidence interval estimators are based on point estimators. In the following section, for example, we will see that $\bar{X}$ is a very reasonable point estimator of μ. As a result, the important confidence interval estimator of μ depends on knowledge of the sampling distribution of $\bar{X}$.

We begin the following section with the simplest case of a confidence interval. The scenario is simple and yet unrealistic. We are interested in estimating a population mean μ and yet σ is known. Clearly, if μ is unknown, it is quite unlikely that σ is known. Any historical results that produced enough information to allow the assumption that σ is known would likely have produced similar information about μ. Despite this argument, we begin with this case because the concepts and indeed the resulting mechanics associated with confidence interval estimation remain the same for the more realistic situations presented later in Section 9.4 and beyond.

9.4 Single Sample: Estimating the Mean

The sampling distribution of $\bar{X}$ is centered at μ, and in most applications the variance is smaller than that of any other estimators of μ. Thus, the sample mean $\bar{x}$ will be used as a point estimate for the population mean μ. Recall that $\sigma^2_{\bar{X}} = \sigma^2/n$, so a large sample will yield a value of $\bar{X}$ that comes from a sampling distribution with a small variance. Hence, $\bar{x}$ is likely to be a very accurate estimate of μ when n is large.

Let us now consider the interval estimate of μ. If our sample is selected from a normal population or, failing this, if n is sufficiently large, we can establish a confidence interval for μ by considering the sampling distribution of $\bar{X}$.

According to the Central Limit Theorem, we can expect the sampling distribution of $\bar{X}$ to be approximately normally distributed with mean $\mu_{\bar{X}} = \mu$ and

standard deviation $\sigma_{\bar{X}} = \sigma/\sqrt{n}$. Writing $z_{\alpha/2}$ for the z-value above which we find an area of $\alpha/2$ under the normal curve, we can see from Figure 9.2 that

$$P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha,$$

where

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}.$$

Hence,

$$P\left(-z_{\alpha/2} < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < z_{\alpha/2}\right) = 1 - \alpha.$$

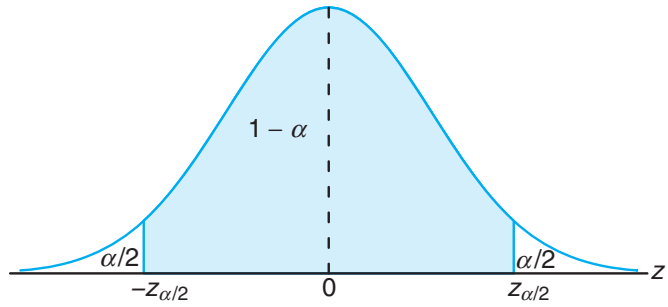


Figure 9.2: $P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha$.

Multiplying each term in the inequality by $\sigma/\sqrt{n}$ and then subtracting $\bar{X}$ from each term and multiplying by -1 (reversing the sense of the inequalities), we obtain

$$P\left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha.$$

A random sample of size n is selected from a population whose variance σ^2 is known, and the mean $\bar{x}$ is computed to give the $100(1-\alpha)\%$ confidence interval below. It is important to emphasize that we have invoked the Central Limit Theorem above. As a result, it is important to note the conditions for applications that follow.

Confidence Interval on μ, σ^2 Known

If $\bar{x}$ is the mean of a random sample of size n from a population with known variance σ^2, a $100(1-\alpha)\%$ confidence interval for μ is given by

$$\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}},$$

where $z_{\alpha/2}$ is the z-value leaving an area of $\alpha/2$ to the right.

For small samples selected from nonnormal populations, we cannot expect our degree of confidence to be accurate. However, for samples of size $n \geq 30$, with

the shape of the distributions not too skewed, sampling theory guarantees good results.

Clearly, the values of the random variables $\hat{\Theta}_L$ and $\hat{\Theta}_U$, defined in Section 9.3, are the confidence limits

$$\hat{\theta}_L = \bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \quad \text{and} \quad \hat{\theta}_U = \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

Different samples will yield different values of $\bar{x}$ and therefore produce different interval estimates of the parameter μ, as shown in Figure 9.3. The dot at the center of each interval indicates the position of the point estimate $\bar{x}$ for that random sample. Note that all of these intervals are of the same width, since their widths depend only on the choice of $z_{\alpha/2}$ once $\bar{x}$ is determined. The larger the value we choose for $z_{\alpha/2}$, the wider we make all the intervals and the more confident we can be that the particular sample selected will produce an interval that contains the unknown parameter μ. In general, for a selection of $z_{\alpha/2}$, $100(1-\alpha)\%$ of the intervals will cover μ.

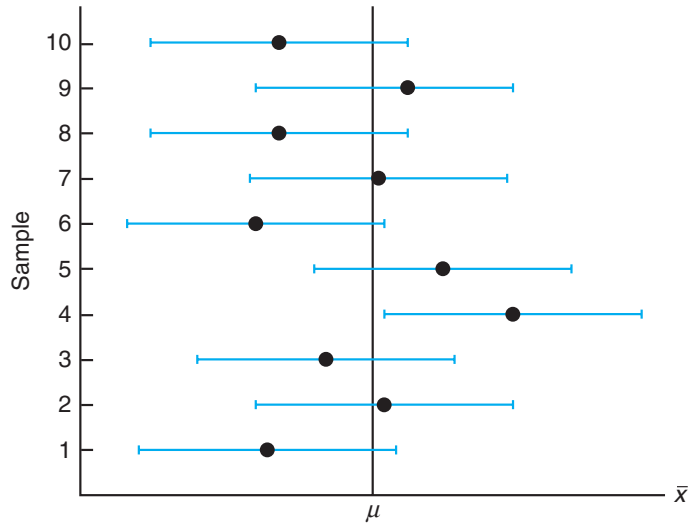


Figure 9.3: Interval estimates of μ for different samples.

Example 9.2: The average zinc concentration recovered from a sample of measurements taken in 36 different locations in a river is found to be 2.6 grams per milliliter. Find the 95% and 99% confidence intervals for the mean zinc concentration in the river. Assume that the population standard deviation is 0.3 gram per milliliter.

Solution: The point estimate of μ is $\bar{x} = 2.6$. The z-value leaving an area of 0.025 to the right, and therefore an area of 0.975 to the left, is $z_{0.025} = 1.96$ (Table A.3). Hence, the 95% confidence interval is

$$2.6 - (1.96)\left(\frac{0.3}{\sqrt{36}}\right) < \mu < 2.6 + (1.96)\left(\frac{0.3}{\sqrt{36}}\right),$$

which reduces to $2.50 < \mu < 2.70$. To find a 99% confidence interval, we find the z-value leaving an area of 0.005 to the right and 0.995 to the left. From Table A.3 again, $z_{0.005} = 2.575$, and the 99% confidence interval is

$$2.6 - (2.575)\left(\frac{0.3}{\sqrt{36}}\right) < \mu < 2.6 + (2.575)\left(\frac{0.3}{\sqrt{36}}\right),$$

or simply

$$2.47 < \mu < 2.73.$$

We now see that a longer interval is required to estimate μ with a higher degree of confidence.

The $100(1-\alpha)\%$ confidence interval provides an estimate of the accuracy of our point estimate. If μ is actually the center value of the interval, then $\bar{x}$ estimates μ without error. Most of the time, however, $\bar{x}$ will not be exactly equal to μ and the point estimate will be in error. The size of this error will be the absolute value of the difference between μ and $\bar{x}$, and we can be $100(1-\alpha)\%$ confident that this difference will not exceed $z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$. We can readily see this if we draw a diagram of a hypothetical confidence interval, as in Figure 9.4.

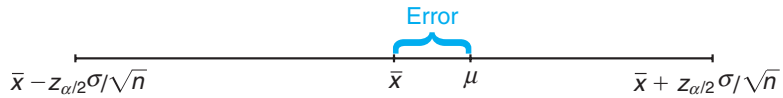


Figure 9.4: Error in estimating μ by $\bar{x}$.

Theorem 9.1: If $\bar{x}$ is used as an estimate of μ, we can be $100(1-\alpha)\%$ confident that the error will not exceed $z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$.

In Example 9.2, we are 95% confident that the sample mean $\bar{x} = 2.6$ differs from the true mean μ by an amount less than $(1.96)(0.3)/\sqrt{36} = 0.1$ and 99% confident that the difference is less than $(2.575)(0.3)/\sqrt{36} = 0.13$.

Frequently, we wish to know how large a sample is necessary to ensure that the error in estimating μ will be less than a specified amount e. By Theorem 9.1, we must choose n such that $z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = e$. Solving this equation gives the following formula for n.

Theorem 9.2: If $\bar{x}$ is used as an estimate of μ, we can be $100(1-\alpha)\%$ confident that the error will not exceed a specified amount e when the sample size is

$$n = \left(\frac{z_{\alpha/2}\sigma}{e}\right)^2.$$

When solving for the sample size, n, we round all fractional values up to the next whole number. By adhering to this principle, we can be sure that our degree of confidence never falls below $100(1-\alpha)\%$.

Strictly speaking, the formula in Theorem 9.2 is applicable only if we know the variance of the population from which we select our sample. Lacking this information, we could take a preliminary sample of size $n \geq 30$ to provide an estimate of σ. Then, using s as an approximation for σ in Theorem 9.2, we could determine approximately how many observations are needed to provide the desired degree of accuracy.

Example 9.3: How large a sample is required if we want to be 95% confident that our estimate of μ in Example 9.2 is off by less than 0.05?

***Solution*:** The population standard deviation is $\sigma = 0.3$. Then, by Theorem 9.2,

$$n = \left[\frac{(1.96)(0.3)}{0.05}\right]^2 = 138.3.$$

Therefore, we can be 95% confident that a random sample of size 139 will provide an estimate $\bar{x}$ differing from μ by an amount less than 0.05.

One-Sided Confidence Bounds

The confidence intervals and resulting confidence bounds discussed thus far are *two-sided* (i.e., both upper and lower bounds are given). However, there are many applications in which only one bound is sought. For example, if the measurement of interest is tensile strength, the engineer receives better information from a lower bound only. This bound communicates the worst-case scenario. On the other hand, if the measurement is something for which a relatively large value of μ is not profitable or desirable, then an upper confidence bound is of interest. An example would be a case in which inferences need to be made concerning the mean mercury composition in a river. An upper bound is very informative in this case.

One-sided confidence bounds are developed in the same fashion as two-sided intervals. However, the source is a one-sided probability statement that makes use of the Central Limit Theorem:

$$P\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < z_\alpha\right) = 1 - \alpha.$$

One can then manipulate the probability statement much as before and obtain

$$P(\mu > \bar{X} - z_\alpha \sigma/\sqrt{n}) = 1 - \alpha.$$

Similar manipulation of $P\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} > -z_\alpha\right) = 1 - \alpha$ gives

$$P(\mu < \bar{X} + z_\alpha \sigma/\sqrt{n}) = 1 - \alpha.$$

As a result, the upper and lower one-sided bounds follow.

One-Sided Confidence Bounds on μ, σ^2 Known

If $\bar{X}$ is the mean of a random sample of size n from a population with variance σ^2, the one-sided $100(1-\alpha)\%$ confidence bounds for μ are given by

upper one-sided bound: $\bar{x} + z_\alpha \sigma/\sqrt{n}$;

lower one-sided bound: $\bar{x} - z_\alpha \sigma/\sqrt{n}$.

Example 9.4: In a psychological testing experiment, 25 subjects are selected randomly and their reaction time, in seconds, to a particular stimulus is measured. Past experience suggests that the variance in reaction times to these types of stimuli is 4 sec^2 and that the distribution of reaction times is approximately normal. The average time for the subjects is 6.2 seconds. Give an upper 95% bound for the mean reaction time.

***Solution*:** The upper 95% bound is given by

$$\begin{aligned}\bar{x} + z_\alpha \sigma/\sqrt{n} &= 6.2 + (1.645)\sqrt{4/25} = 6.2 + 0.658 \\ &= 6.858 \text{ seconds.}\end{aligned}$$

Hence, we are 95% confident that the mean reaction time is less than 6.858 seconds.

The Case of σ Unknown

Frequently, we must attempt to estimate the mean of a population when the variance is unknown. The reader should recall learning in Chapter 8 that if we have a random sample from a *normal distribution*, then the random variable

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

has a Student t-distribution with $n - 1$ degrees of freedom. Here S is the sample standard deviation. In this situation, with σ unknown, T can be used to construct a confidence interval on μ. The procedure is the same as that with σ known except that σ is replaced by S and the standard normal distribution is replaced by the t-distribution. Referring to Figure 9.5, we can assert that

$$P(-t_{\alpha/2} < T < t_{\alpha/2}) = 1 - \alpha,$$

where $t_{\alpha/2}$ is the t-value with $n-1$ degrees of freedom, above which we find an area of $\alpha/2$. Because of symmetry, an equal area of $\alpha/2$ will fall to the left of $-t_{\alpha/2}$. Substituting for T, we write

$$P\left(-t_{\alpha/2} < \frac{\bar{X} - \mu}{S/\sqrt{n}} < t_{\alpha/2}\right) = 1 - \alpha.$$

Multiplying each term in the inequality by $S/\sqrt{n}$, and then subtracting $\bar{X}$ from each term and multiplying by -1, we obtain

$$P\left(\bar{X} - t_{\alpha/2}\frac{S}{\sqrt{n}} < \mu < \bar{X} + t_{\alpha/2}\frac{S}{\sqrt{n}}\right) = 1 - \alpha.$$

For a particular random sample of size n, the mean $\bar{x}$ and standard deviation s are computed and the following $100(1 - \alpha)\%$ confidence interval for μ is obtained.

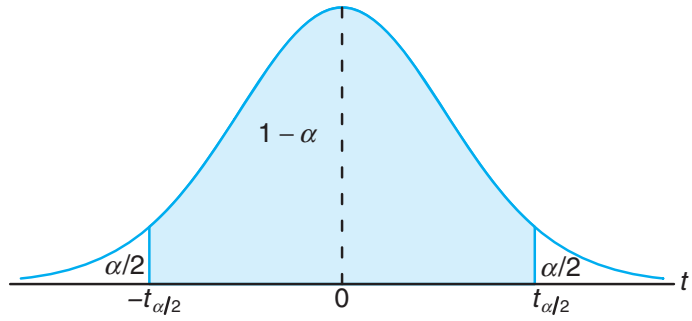


Figure 9.5: $P(-t_{\alpha/2} < T < t_{\alpha/2}) = 1 - \alpha$.

Confidence Interval on μ, σ^2 Unknown

If $\bar{x}$ and s are the mean and standard deviation of a random sample from a normal population with unknown variance σ^2, a $100(1-\alpha)\%$ confidence interval for μ is

$$\bar{x} - t_{\alpha/2}\frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{\alpha/2}\frac{s}{\sqrt{n}},$$

where $t_{\alpha/2}$ is the t-value with $v = n - 1$ degrees of freedom, leaving an area of $\alpha/2$ to the right.

We have made a distinction between the cases of σ known and σ unknown in computing confidence interval estimates. We should emphasize that for σ known we exploited the Central Limit Theorem, whereas for σ unknown we made use of the sampling distribution of the random variable T. However, the use of the t-distribution is based on the premise that the sampling is from a normal distribution. As long as the distribution is approximately bell shaped, confidence intervals can be computed when σ^2 is unknown by using the t-distribution and we may expect very good results.

Computed one-sided confidence bounds for μ with σ unknown are as the reader would expect, namely

$$\bar{x} + t_{\alpha}\frac{s}{\sqrt{n}} \qquad \text{and} \qquad \bar{x} - t_{\alpha}\frac{s}{\sqrt{n}}.$$

They are the upper and lower $100(1-\alpha)\%$ bounds, respectively. Here t_α is the t-value having an area of α to the right.

Example 9.5: The contents of seven similar containers of sulfuric acid are 9.8, 10.2, 10.4, 9.8, 10.0, 10.2, and 9.6 liters. Find a 95% confidence interval for the mean contents of all such containers, assuming an approximately normal distribution.

***Solution*:** The sample mean and standard deviation for the given data are

$$\bar{x} = 10.0 \qquad \text{and} \qquad s = 0.283.$$

Using Table A.4, we find $t_{0.025} = 2.447$ for $v = 6$ degrees of freedom. Hence, the

95% confidence interval for μ is

$$10.0 - (2.447)\left(\frac{0.283}{\sqrt{7}}\right) < \mu < 10.0 + (2.447)\left(\frac{0.283}{\sqrt{7}}\right),$$

which reduces to $9.74 < \mu < 10.26$.

Concept of a Large-Sample Confidence Interval

Often statisticians recommend that even when normality cannot be assumed, σ is unknown, and $n \geq 30$, s can replace σ and the confidence interval

$$\bar{x} \pm z_{\alpha/2}\frac{s}{\sqrt{n}}$$

may be used. This is often referred to as a *large-sample confidence interval*. The justification lies only in the presumption that with a sample as large as 30 and the population distribution not too skewed, s will be very close to the true σ and thus the Central Limit Theorem prevails. It should be emphasized that this is only an approximation and the quality of the result becomes better as the sample size grows larger.

Example 9.6: Scholastic Aptitude Test (SAT) mathematics scores of a random sample of 500 high school seniors in the state of Texas are collected, and the sample mean and standard deviation are found to be 501 and 112, respectively. Find a 99% confidence interval on the mean SAT mathematics score for seniors in the state of Texas.

***Solution*:** Since the sample size is large, it is reasonable to use the normal approximation. Using Table A.3, we find $z_{0.005} = 2.575$. Hence, a 99% confidence interval for μ is

$$501 \pm (2.575)\left(\frac{112}{\sqrt{500}}\right) = 501 \pm 12.9,$$

which yields $488.1 < \mu < 513.9$.

9.5 Standard Error of a Point Estimate

We have made a rather sharp distinction between the goal of a point estimate and that of a confidence interval estimate. The former supplies a single number extracted from a set of experimental data, and the latter provides an interval that is reasonable for the parameter, *given the experimental data*; that is, $100(1-\alpha)\%$ of such computed intervals "cover" the parameter.

These two approaches to estimation are related to each other. The common thread is the sampling distribution of the point estimator. Consider, for example, the estimator $\bar{X}$ of μ with σ known. We indicated earlier that a measure of the quality of an unbiased estimator is its variance. The variance of $\bar{X}$ is

$$\sigma^2_{\bar{X}} = \frac{\sigma^2}{n}.$$

Thus, the standard deviation of $\bar{X}$, or *standard error* of $\bar{X}$, is $\sigma/\sqrt{n}$. Simply put, the standard error of an estimator is its standard deviation. For $\bar{X}$, the computed confidence limit

$$\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \text{ is written as } \bar{x} \pm z_{\alpha/2} \text{ s.e.}(\bar{x}),$$

where "s.e." is the "standard error." The important point is that the width of the confidence interval on μ is dependent on the quality of the point estimator through its standard error. In the case where σ is unknown and sampling is from a normal distribution, s replaces σ and the *estimated standard error* $s/\sqrt{n}$ is involved. Thus, the confidence limits on μ are

Confidence Limits on μ, σ^2 Unknown

$$\bar{x} \pm t_{\alpha/2}\frac{s}{\sqrt{n}} = \bar{x} \pm t_{\alpha/2} \text{ s.e.}(\bar{x})$$

Again, the confidence interval is *no better* (in terms of width) *than the quality of the point estimate*, in this case through its estimated standard error. Computer packages often refer to estimated standard errors simply as "standard errors."

As we move to more complex confidence intervals, there is a prevailing notion that widths of confidence intervals become shorter as the quality of the corresponding point estimate becomes better, although it is not always quite as simple as we have illustrated here. It can be argued that a confidence interval is merely an augmentation of the point estimate to take into account the precision of the point estimate.

9.6 Prediction Intervals

The point and interval estimations of the mean in Sections 9.4 and 9.5 provide good information about the unknown parameter μ of a normal distribution or a nonnormal distribution from which a large sample is drawn. Sometimes, other than the population mean, the experimenter may also be interested in predicting the possible **value of a future observation**. For instance, in quality control, the experimenter may need to use the observed data to predict a new observation. A process that produces a metal part may be evaluated on the basis of whether the part meets specifications on tensile strength. On certain occasions, a customer may be interested in purchasing a **single part**. In this case, a confidence interval on the mean tensile strength does not capture the required information. The customer requires a statement regarding the uncertainty of a **single observation**. This type of requirement is nicely fulfilled by the construction of a **prediction interval**.

It is quite simple to obtain a prediction interval for the situations we have considered so far. Assume that the random sample comes from a normal population with unknown mean μ and known variance σ^2. A natural point estimator of a new observation is $\bar{X}$. It is known, from Section 8.4, that the variance of $\bar{X}$ is σ^2/n. However, to predict a new observation, not only do we need to account for the variation due to estimating the mean, but also we should account for the **variation of a future observation**. From the assumption, we know that the variance of the random error in a new observation is σ^2. The development of a

prediction interval is best illustrated by beginning with a normal random variable $x_0 - \bar{x}$, where x_0 is the new observation and $\bar{x}$ comes from the sample. Since x_0 and $\bar{x}$ are independent, we know that

$$z = \frac{x_0 - \bar{x}}{\sqrt{\sigma^2 + \sigma^2/n}} = \frac{x_0 - \bar{x}}{\sigma\sqrt{1 + 1/n}}$$

is $n(z; 0, 1)$. As a result, if we use the probability statement

$$P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha$$

with the z-statistic above and place x_0 in the center of the probability statement, we have the following event occurring with probability $1 - \alpha$:

$$\bar{x} - z_{\alpha/2}\sigma\sqrt{1 + 1/n} < x_0 < \bar{x} + z_{\alpha/2}\sigma\sqrt{1 + 1/n}.$$

As a result, computation of the prediction interval is formalized as follows.

Prediction Interval of a Future Observation, σ^2 Known

For a normal distribution of measurements with unknown mean μ and known variance σ^2, a $100(1 - \alpha)\%$ **prediction interval** of a future observation x_0 is

$$\bar{x} - z_{\alpha/2}\sigma\sqrt{1 + 1/n} < x_0 < \bar{x} + z_{\alpha/2}\sigma\sqrt{1 + 1/n},$$

where $z_{\alpha/2}$ is the z-value leaving an area of $\alpha/2$ to the right.

Example 9.7: Due to the decrease in interest rates, the First Citizens Bank received a lot of mortgage applications. A recent sample of 50 mortgage loans resulted in an average loan amount of $257,300. Assume a population standard deviation of $25,000. For the next customer who fills out a mortgage application, find a 95% prediction interval for the loan amount.

Solution: The point prediction of the next customer's loan amount is $\bar{x} = \$257,300$. The z-value here is $z_{0.025} = 1.96$. Hence, a 95% prediction interval for the future loan amount is

$$257,300 - (1.96)(25,000)\sqrt{1 + 1/50} < x_0 < 257,300 + (1.96)(25,000)\sqrt{1 + 1/50},$$

which gives the interval ($207,812.43, $306,787.57). ∎

The prediction interval provides a good estimate of the location of a future observation, which is quite different from the estimate of the sample mean value. It should be noted that the variation of this prediction is the sum of the variation due to an estimation of the mean and the variation of a single observation. However, as in the past, we first consider the case with known variance. It is also important to deal with the prediction interval of a future observation in the situation where the variance is unknown. Indeed a Student t-distribution may be used in this case, as described in the following result. The normal distribution is merely replaced by the t-distribution.

Prediction Interval of a Future Observation, σ^2 Unknown

For a normal distribution of measurements with unknown mean μ and unknown variance σ^2, a $100(1-\alpha)\%$ **prediction interval** of a future observation x_0 is

$$\bar{x} - t_{\alpha/2}s\sqrt{1+1/n} < x_0 < \bar{x} + t_{\alpha/2}s\sqrt{1+1/n},$$

where $t_{\alpha/2}$ is the t-value with $v = n-1$ degrees of freedom, leaving an area of $\alpha/2$ to the right.

One-sided prediction intervals can also be constructed. Upper prediction bounds apply in cases where focus must be placed on future large observations. Concern over future small observations calls for the use of lower prediction bounds. The upper bound is given by

$$\bar{x} + t_{\alpha}s\sqrt{1+1/n}$$

and the lower bound by

$$\bar{x} - t_{\alpha}s\sqrt{1+1/n}.$$

Example 9.8: A meat inspector has randomly selected 30 packs of 95% lean beef. The sample resulted in a mean of 96.2% with a sample standard deviation of 0.8%. Find a 99% prediction interval for the leanness of a new pack. Assume normality.

***Solution*:** For $v = 29$ degrees of freedom, $t_{0.005} = 2.756$. Hence, a 99% prediction interval for a new observation x_0 is

$$96.2 - (2.756)(0.8)\sqrt{1+\frac{1}{30}} < x_0 < 96.2 + (2.756)(0.8)\sqrt{1+\frac{1}{30}},$$

which reduces to (93.96, 98.44).

Use of Prediction Limits for Outlier Detection

To this point in the text very little attention has been paid to the concept of **outliers**, or aberrant observations. The majority of scientific investigators are keenly sensitive to the existence of outlying observations or so-called faulty or "bad data." We deal with the concept of outlier detection extensively in Chapter 12. However, it is certainly of interest here since there is an important relationship between outlier detection and prediction intervals.

It is convenient for our purposes to view an outlying observation as one that comes from a population with a mean that is different from the mean that governs the rest of the sample of size n being studied. The prediction interval produces a bound that "covers" a future single observation with probability $1-\alpha$ if it comes from the population from which the sample was drawn. As a result, a methodology for outlier detection involves the rule that **an observation is an outlier if it falls outside the prediction interval computed without including the questionable observation in the sample**. As a result, for the prediction interval of Example 9.8, if a new pack of beef is measured and its leanness is outside the interval $(93.96, 98.44)$, that observation can be viewed as an outlier.

9.7 Tolerance Limits

As discussed in Section 9.6, the scientist or engineer may be less interested in estimating parameters than in gaining a notion about where an individual *observation* or measurement might fall. Such situations call for the use of prediction intervals. However, there is yet a third type of interval that is of interest in many applications. Once again, suppose that interest centers around the manufacturing of a component part and specifications exist on a dimension of that part. In addition, there is little concern about the mean of the dimension. But unlike in the scenario in Section 9.6, one may be less interested in a single observation and more interested in where the majority of the population falls. If process specifications are important, the manager of the process is concerned about long-range performance, **not the next observation**. One must attempt to determine bounds that, in some probabilistic sense, "cover" values in the population (i.e., the measured values of the dimension).

One method of establishing the desired bounds is to determine a confidence interval on a *fixed proportion* of the measurements. This is best motivated by visualizing a situation in which we are doing random sampling from a normal distribution with known mean μ and variance σ^2. Clearly, a bound that covers the middle 95% of the population of observations is

$$\mu \pm 1.96\sigma.$$

This is called a **tolerance interval**, and indeed its coverage of 95% of measured observations is exact. However, in practice, μ and σ are seldom known; thus, the user must apply

$$\bar{x} \pm ks.$$

Now, of course, the interval is a random variable, and hence the *coverage* of a proportion of the population by the interval is not exact. As a result, a $100(1-\gamma)\%$ confidence interval must be used since $\bar{x} \pm ks$ cannot be expected to cover any specified proportion all the time. As a result, we have the following definition.

Tolerance Limits

For a normal distribution of measurements with unknown mean μ and unknown standard deviation σ, **tolerance limits** are given by $\bar{x} \pm ks$, where k is determined such that one can assert with $100(1-\gamma)\%$ confidence that the given limits contain at least the proportion $1-\alpha$ of the measurements.

Table A.7 gives values of k for $1-\alpha = 0.90, 0.95, 0.99$; $\gamma = 0.05, 0.01$; and selected values of n from 2 to 300.

Example 9.9: Consider Example 9.8. With the information given, find a tolerance interval that gives two-sided 95% bounds on 90% of the distribution of packages of 95% lean beef. Assume the data came from an approximately normal distribution.

***Solution*:** Recall from Example 9.8 that $n = 30$, the sample mean is 96.2%, and the sample standard deviation is 0.8%. From Table A.7, $k = 2.14$. Using

$$\bar{x} \pm ks = 96.2 \pm (2.14)(0.8),$$

we find that the lower and upper bounds are 94.5 and 97.9.

We are 95% confident that the above range covers the central 90% of the distribution of 95% lean beef packages. ▟

Distinction among Confidence Intervals, Prediction Intervals, and Tolerance Intervals

It is important to reemphasize the difference among the three types of intervals discussed and illustrated in the preceding sections. The computations are straightforward, but interpretation can be confusing. In real-life applications, these intervals are not interchangeable because their interpretations are quite distinct.

In the case of confidence intervals, one is attentive only to the **population mean**. For example, Exercise 9.13 on page 283 deals with an engineering process that produces shearing pins. A specification will be set on Rockwell hardness, below which a customer will not accept any pins. Here, a population parameter must take a backseat. It is important that the engineer know where the *majority of the values of Rockwell hardness are going to be*. Thus, tolerance limits should be used. Surely, when tolerance limits on any process output are tighter than process specifications, that is good news for the process manager.

It is true that the tolerance limit interpretation is somewhat related to the confidence interval. The $100(1-\alpha)\%$ tolerance interval on, say, the proportion 0.95 can be viewed as a confidence interval **on the middle 95%** of the corresponding normal distribution. One-sided tolerance limits are also relevant. In the case of the Rockwell hardness problem, it is desirable to have a lower bound of the form $\bar{x} - ks$ such that there is 99% confidence that at least 99% of Rockwell hardness values will exceed the computed value.

Prediction intervals are applicable when it is important to determine a bound on a **single value**. The mean is not the issue here, nor is the location of the majority of the population. Rather, the location of a single new observation is required.

Case Study 9.1: **Machine Quality**: A machine produces metal pieces that are cylindrical in shape. A sample of these pieces is taken and the diameters are found to be 1.01, 0.97, 1.03, 1.04, 0.99, 0.98, 0.99, 1.01, and 1.03 centimeters. Use these data to calculate three interval types and draw interpretations that illustrate the distinction between them in the context of the system. For all computations, assume an approximately normal distribution. The sample mean and standard deviation for the given data are $\bar{x} = 1.0056$ and $s = 0.0246$.

(a) Find a 99% confidence interval on the mean diameter.

(b) Compute a 99% prediction interval on a measured diameter of a single metal piece taken from the machine.

(c) Find the 99% tolerance limits that will contain 95% of the metal pieces produced by this machine.

***Solution*:** (a) The 99% confidence interval for the mean diameter is given by

$$\bar{x} \pm t_{0.005} s/\sqrt{n} = 1.0056 \pm (3.355)(0.0246/3) = 1.0056 \pm 0.0275.$$

Thus, the 99% confidence bounds are 0.9781 and 1.0331.

(b) The 99% prediction interval for a future observation is given by

$$\bar{x} \pm t_{0.005} s\sqrt{1+1/n} = 1.0056 \pm (3.355)(0.0246)\sqrt{1+1/9},$$

with the bounds being 0.9186 and 1.0926.

(c) From Table A.7, for $n = 9$, $1-\gamma = 0.99$, and $1-\alpha = 0.95$, we find $k = 4.550$ for two-sided limits. Hence, the 99% tolerance limits are given by

$$\bar{x} + ks = 1.0056 \pm (4.550)(0.0246),$$

with the bounds being 0.8937 and 1.1175. We are 99% confident that the tolerance interval from 0.8937 to 1.1175 will contain the central 95% of the distribution of diameters produced.

This case study illustrates that the three types of limits can give appreciably different results even though they are all 99% bounds. In the case of the confidence interval on the mean, 99% of such intervals cover the population mean diameter. Thus, we say that we are 99% confident that the mean diameter produced by the process is between 0.9781 and 1.0331 centimeters. Emphasis is placed on the mean, with less concern about a single reading or the general nature of the distribution of diameters in the population. In the case of the prediction limits, the bounds 0.9186 and 1.0926 are based on the distribution of a single "new" metal piece taken from the process, and again 99% of such limits will cover the diameter of a new measured piece. On the other hand, the tolerance limits, as suggested in the previous section, give the engineer a sense of where the "majority," say the central 95%, of the diameters of measured pieces in the population reside. The 99% tolerance limits, 0.8937 and 1.1175, are numerically quite different from the other two bounds. If these bounds appear alarmingly wide to the engineer, it reflects negatively on process quality. On the other hand, if the bounds represent a desirable result, the engineer may conclude that a majority (95% in here) of the diameters are in a desirable range. Again, a confidence interval interpretation may be used: namely, 99% of such calculated bounds will cover the middle 95% of the population of diameters.

Exercises

9.1 A UCLA researcher claims that the life span of mice can be extended by as much as 25% when the calories in their diet are reduced by approximately 40% from the time they are weaned. The restricted diet is enriched to normal levels by vitamins and protein. Assuming that it is known from previous studies that $\sigma = 5.8$ months, how many mice should be included in our sample if we wish to be 99% confident that the mean life span of the sample will be within 2 months of the population mean for all mice subjected to this reduced diet?

9.2 An electrical firm manufactures light bulbs that have a length of life that is approximately normally distributed with a standard deviation of 40 hours. If a sample of 30 bulbs has an average life of 780 hours, find a 96% confidence interval for the population mean of all bulbs produced by this firm.

9.3 Many cardiac patients wear an implanted pacemaker to control their heartbeat. A plastic connector module mounts on the top of the pacemaker. Assuming a standard deviation of 0.0015 inch and an approximately normal distribution, find a 95% confidence

interval for the mean of the depths of all connector modules made by a certain manufacturing company. A random sample of 75 modules has an average depth of 0.310 inch.

9.4 The heights of a random sample of 50 college students showed a mean of 174.5 centimeters and a standard deviation of 6.9 centimeters.

(a) Construct a 98% confidence interval for the mean height of all college students.

(b) What can we assert with 98% confidence about the possible size of our error if we estimate the mean height of all college students to be 174.5 centimeters?

9.5 A random sample of 100 automobile owners in the state of Virginia shows that an automobile is driven on average 23,500 kilometers per year with a standard deviation of 3900 kilometers. Assume the distribution of measurements to be approximately normal.

(a) Construct a 99% confidence interval for the average number of kilometers an automobile is driven annually in Virginia.

(b) What can we assert with 99% confidence about the possible size of our error if we estimate the average number of kilometers driven by car owners in Virginia to be 23,500 kilometers per year?

9.6 How large a sample is needed in Exercise 9.2 if we wish to be 96% confident that our sample mean will be within 10 hours of the true mean?

9.7 How large a sample is needed in Exercise 9.3 if we wish to be 95% confident that our sample mean will be within 0.0005 inch of the true mean?

9.8 An efficiency expert wishes to determine the average time that it takes to drill three holes in a certain metal clamp. How large a sample will she need to be 95% confident that her sample mean will be within 15 seconds of the true mean? Assume that it is known from previous studies that $\sigma = 40$ seconds.

9.9 Regular consumption of presweetened cereals contributes to tooth decay, heart disease, and other degenerative diseases, according to studies conducted by Dr. W. H. Bowen of the National Institute of Health and Dr. J. Yudben, Professor of Nutrition and Dietetics at the University of London. In a random sample consisting of 20 similar single servings of Alpha-Bits, the average sugar content was 11.3 grams with a standard deviation of 2.45 grams. Assuming that the sugar contents are normally distributed, construct a 95% confidence interval for the mean sugar content for single servings of Alpha-Bits.

9.10 A random sample of 12 graduates of a certain secretarial school typed an average of 79.3 words per minute with a standard deviation of 7.8 words per minute. Assuming a normal distribution for the number of words typed per minute, find a 95% confidence interval for the average number of words typed by all graduates of this school.

9.11 A machine produces metal pieces that are cylindrical in shape. A sample of pieces is taken, and the diameters are found to be 1.01, 0.97, 1.03, 1.04, 0.99, 0.98, 0.99, 1.01, and 1.03 centimeters. Find a 99% confidence interval for the mean diameter of pieces from this machine, assuming an approximately normal distribution.

9.12 A random sample of 10 chocolate energy bars of a certain brand has, on average, 230 calories per bar, with a standard deviation of 15 calories. Construct a 99% confidence interval for the true mean calorie content of this brand of energy bar. Assume that the distribution of the calorie content is approximately normal.

9.13 A random sample of 12 shearing pins is taken in a study of the Rockwell hardness of the pin head. Measurements on the Rockwell hardness are made for each of the 12, yielding an average value of 48.50 with a sample standard deviation of 1.5. Assuming the measurements to be normally distributed, construct a 90% confidence interval for the mean Rockwell hardness.

9.14 The following measurements were recorded for the drying time, in hours, of a certain brand of latex paint:

3.4	2.5	4.8	2.9	3.6
2.8	3.3	5.6	3.7	2.8
4.4	4.0	5.2	3.0	4.8

Assuming that the measurements represent a random sample from a normal population, find a 95% prediction interval for the drying time for the next trial of the paint.

9.15 Referring to Exercise 9.5, construct a 99% prediction interval for the kilometers traveled annually by an automobile owner in Virginia.

9.16 Consider Exercise 9.10. Compute the 95% prediction interval for the next observed number of words per minute typed by a graduate of the secretarial school.

9.17 Consider Exercise 9.9. Compute a 95% prediction interval for the sugar content of the next single serving of Alpha-Bits.

9.18 Referring to Exercise 9.13, construct a 95% tolerance interval containing 90% of the measurements.

9.19 A random sample of 25 tablets of buffered aspirin contains, on average, 325.05 mg of aspirin per tablet, with a standard deviation of 0.5 mg. Find the 95% tolerance limits that will contain 90% of the tablet contents for this brand of buffered aspirin. Assume that the aspirin content is normally distributed.

9.20 Consider the situation of Exercise 9.11. Estimation of the mean diameter, while important, is not nearly as important as trying to pin down the location of the majority of the distribution of diameters. Find the 95% tolerance limits that contain 95% of the diameters.

9.21 In a study conducted by the Department of Zoology at Virginia Tech, fifteen samples of water were collected from a certain station in the James River in order to gain some insight regarding the amount of orthophosphorus in the river. The concentration of the chemical is measured in milligrams per liter. Let us suppose that the mean at the station is not as important as the upper extreme of the distribution of the concentration of the chemical at the station. Concern centers around whether the concentration at the extreme is too large. Readings for the fifteen water samples gave a sample mean of 3.84 milligrams per liter and a sample standard deviation of 3.07 milligrams per liter. Assume that the readings are a random sample from a normal distribution. Calculate a prediction interval (upper 95% prediction limit) and a tolerance limit (95% upper tolerance limit that exceeds 95% of the population of values). Interpret both; that is, tell what each communicates about the upper extreme of the distribution of orthophosphorus at the sampling station.

9.22 A type of thread is being studied for its tensile strength properties. Fifty pieces were tested under similar conditions, and the results showed an average tensile strength of 78.3 kilograms and a standard deviation of 5.6 kilograms. Assuming a normal distribution of tensile strengths, give a lower 95% prediction limit on a single observed tensile strength value. In addition, give a lower 95% tolerance limit that is exceeded by 99% of the tensile strength values.

9.23 Refer to Exercise 9.22. Why are the quantities requested in the exercise likely to be more important to the manufacturer of the thread than, say, a confidence interval on the mean tensile strength?

9.24 Refer to Exercise 9.22 again. Suppose that specifications by a buyer of the thread are that the tensile strength of the material must be at least 62 kilograms. The manufacturer is satisfied if at most 5% of the manufactured pieces have tensile strength less than 62 kilograms. Is there cause for concern? Use a one-sided 99% tolerance limit that is exceeded by 95% of the tensile strength values.

9.25 Consider the drying time measurements in Exercise 9.14. Suppose the 15 observations in the data set are supplemented by a 16th value of 6.9 hours. In the context of the original 15 observations, is the 16th value an outlier? Show work.

9.26 Consider the data in Exercise 9.13. Suppose the manufacturer of the shearing pins insists that the Rockwell hardness of the product be less than or equal to 44.0 only 5% of the time. What is your reaction? Use a tolerance limit calculation as the basis for your judgment.

9.27 Consider the situation of Case Study 9.1 on page 281 with a larger sample of metal pieces. The diameters are as follows: 1.01, 0.97, 1.03, 1.04, 0.99, 0.98, 1.01, 1.03, 0.99, 1.00, 1.00, 0.99, 0.98, 1.01, 1.02, 0.99 centimeters. Once again the normality assumption may be made. Do the following and compare your results to those of the case study. Discuss how they are different and why.

(a) Compute a 99% confidence interval on the mean diameter.

(b) Compute a 99% prediction interval on the next diameter to be measured.

(c) Compute a 99% tolerance interval for coverage of the central 95% of the distribution of diameters.

9.28 In Section 9.3, we emphasized the notion of "most efficient estimator" by comparing the variance of two unbiased estimators $\hat{\Theta}_1$ and $\hat{\Theta}_2$. However, this does not take into account bias in case one or both estimators are not unbiased. Consider the quantity

$$MSE = E(\hat{\Theta} - \theta),$$

where MSE denotes **mean squared error**. The MSE is often used to compare two estimators $\hat{\Theta}_1$ and $\hat{\Theta}_2$ of θ when either or both is unbiased because (i) it is intuitively reasonable and (ii) it accounts for bias. Show that MSE can be written

$$\begin{aligned} MSE &= E[\hat{\Theta} - E(\hat{\Theta})]^2 + [E(\hat{\Theta} - \theta)]^2 \\ &= \text{Var}(\hat{\Theta}) + [\text{Bias}(\hat{\Theta})]^2. \end{aligned}$$

9.29 Let us define $S'^2 = \sum_{i=1}^{n} (X_i - \bar{X})^2/n$. Show that

$$E(S'^2) = [(n-1)/n]\sigma^2,$$

and hence S'^2 is a biased estimator for σ^2.

9.30 Consider S'^2, the estimator of σ^2, from Exercise 9.29. Analysts often use S'^2 rather than dividing $\sum_{i=1}^{n} (X_i - \bar{X})^2$ by $n-1$, the degrees of freedom in the sample.

(a) What is the bias of S'^2?

(b) Show that the bias of S'^2 approaches zero as $n \to \infty$.

9.31 If X is a binomial random variable, show that

(a) $\widehat{P} = X/n$ is an unbiased estimator of p;

(b) $P' = \frac{X+\sqrt{n}/2}{n+\sqrt{n}}$ is a biased estimator of p.

9.32 Show that the estimator P' of Exercise 9.31(b) becomes unbiased as $n \to \infty$.

9.33 Compare S^2 and S'^2 (see Exercise 9.29), the two estimators of σ^2, to determine which is more efficient. Assume these estimators are found using $X_1, X_2, \ldots, X_n$, independent random variables from $n(x; \mu, \sigma)$. Which estimator is more efficient considering only the variance of the estimators? [*Hint*: Make use of Theorem 8.4 and the fact that the variance of χ^2_v is $2v$, from Section 6.7.]

9.34 Consider Exercise 9.33. Use the MSE discussed in Exercise 9.28 to determine which estimator is more efficient. Write out

$$\frac{MSE(S^2)}{MSE(S'^2)}.$$

9.8 Two Samples: Estimating the Difference between Two Means

If we have two populations with means μ_1 and μ_2 and variances σ_1^2 and σ_2^2, respectively, a point estimator of the difference between μ_1 and μ_2 is given by the statistic $\bar{X}_1 - \bar{X}_2$. Therefore, to obtain a point estimate of $\mu_1 - \mu_2$, we shall select two independent random samples, one from each population, of sizes n_1 and n_2, and compute $\bar{x}_1 - \bar{x}_2$, the difference of the sample means. Clearly, we must consider the sampling distribution of $\bar{X}_1 - \bar{X}_2$.

According to Theorem 8.3, we can expect the sampling distribution of $\bar{X}_1 - \bar{X}_2$ to be approximately normally distributed with mean $\mu_{\bar{X}_1-\bar{X}_2} = \mu_1 - \mu_2$ and standard deviation $\sigma_{\bar{X}_1-\bar{X}_2} = \sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}$. Therefore, we can assert with a probability of $1 - \alpha$ that the standard normal variable

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

will fall between $-z_{\alpha/2}$ and $z_{\alpha/2}$. Referring once again to Figure 9.2, we write

$$P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha.$$

Substituting for Z, we state equivalently that

$$P\left(-z_{\alpha/2} < \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} < z_{\alpha/2}\right) = 1 - \alpha,$$

which leads to the following $100(1-\alpha)\%$ confidence interval for $\mu_1 - \mu_2$.

Confidence Interval for $\mu_1 - \mu_2$, σ_1^2 and σ_2^2 Known

If $\bar{x}_1$ and $\bar{x}_2$ are means of independent random samples of sizes n_1 and n_2 from populations with known variances σ_1^2 and σ_2^2, respectively, a $100(l-\alpha)\%$ confidence interval for $\mu_1 - \mu_2$ is given by

$$(\bar{x}_1 - \bar{x}_2) - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} < \mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}},$$

where $z_{\alpha/2}$ is the z-value leaving an area of $\alpha/2$ to the right.

paint were selected, and the drying times, in hours, were as follows:

Paint *A*					Paint *B*				
3.5	2.7	3.9	4.2	3.6	4.7	3.9	4.5	5.5	4.0
2.7	3.3	5.2	4.2	2.9	5.3	4.3	6.0	5.2	3.7
4.4	5.2	4.0	4.1	3.4	5.5	6.2	5.1	5.4	4.8

Assume the drying time is normally distributed with $\sigma_A = \sigma_B$. Find a 95% confidence interval on $\mu_B - \mu_A$, where μ_A and μ_B are the mean drying times.

9.50 Two levels (low and high) of insulin doses are given to two groups of diabetic rats to check the insulin-binding capacity, yielding the following data:

Low dose:	$n_1 = 8$	$\bar{x}_1 = 1.98$	$s_1 = 0.51$
High dose:	$n_2 = 13$	$\bar{x}_2 = 1.30$	$s_2 = 0.35$

Assume that the variances are equal. Give a 95% confidence interval for the difference in the true average insulin-binding capacity between the two samples.

9.10 Single Sample: Estimating a Proportion

A point estimator of the proportion p in a binomial experiment is given by the statistic $\widehat{P} = X/n$, where X represents the number of successes in n trials. Therefore, the sample proportion $\hat{p} = x/n$ will be used as the point estimate of the parameter p.

If the unknown proportion p is not expected to be too close to 0 or 1, we can establish a confidence interval for p by considering the sampling distribution of $\widehat{P}$. Designating a failure in each binomial trial by the value 0 and a success by the value 1, the number of successes, x, can be interpreted as the sum of n values consisting only of 0 and 1s, and $\hat{p}$ is just the sample mean of these n values. Hence, by the Central Limit Theorem, for n sufficiently large, $\widehat{P}$ is approximately normally distributed with mean

$$\mu_{\widehat{P}} = E(\widehat{P}) = E\left(\frac{X}{n}\right) = \frac{np}{n} = p$$

and variance

$$\sigma^2_{\widehat{P}} = \sigma^2_{X/n} = \frac{\sigma^2_X}{n^2} = \frac{npq}{n^2} = \frac{pq}{n}.$$

Therefore, we can assert that

$$P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha, \text{ with } Z = \frac{\widehat{P} - p}{\sqrt{pq/n}},$$

and $z_{\alpha/2}$ is the value above which we find an area of $\alpha/2$ under the standard normal curve. Substituting for Z, we write

$$P\left(-z_{\alpha/2} < \frac{\widehat{P} - p}{\sqrt{pq/n}} < z_{\alpha/2}\right) = 1 - \alpha.$$

When n is large, very little error is introduced by substituting the point estimate $\hat{p} = x/n$ for the p under the radical sign. Then we can write

$$P\left(\widehat{P} - z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}} < p < \widehat{P} + z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}\right) \approx 1 - \alpha.$$

On the other hand, by solving for p in the quadratic inequality above,

$$-z_{\alpha/2} < \frac{\widehat{P} - p}{\sqrt{pq/n}} < z_{\alpha/2},$$

we obtain another form of the confidence interval for p with limits

$$\frac{\hat{p} + \frac{z^2_{\alpha/2}}{2n}}{1 + \frac{z^2_{\alpha/2}}{n}} \pm \frac{z_{\alpha/2}}{1 + \frac{z^2_{\alpha/2}}{n}} \sqrt{\frac{\hat{p}\hat{q}}{n} + \frac{z^2_{\alpha/2}}{4n^2}}.$$

For a random sample of size n, the sample proportion $\hat{p} = x/n$ is computed, and the following approximate $100(1-\alpha)\%$ confidence intervals for p can be obtained.

Large-Sample Confidence Intervals for p

If $\hat{p}$ is the proportion of successes in a random sample of size n and $\hat{q} = 1 - \hat{p}$, an approximate $100(1-\alpha)\%$ confidence interval, for the binomial parameter p is given by (method 1)

$$\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}} < p < \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}$$

or by (method 2)

$$\frac{\hat{p} + \frac{z^2_{\alpha/2}}{2n}}{1 + \frac{z^2_{\alpha/2}}{n}} - \frac{z_{\alpha/2}}{1 + \frac{z^2_{\alpha/2}}{n}} \sqrt{\frac{\hat{p}\hat{q}}{n} + \frac{z^2_{\alpha/2}}{4n^2}} < p < \frac{\hat{p} + \frac{z^2_{\alpha/2}}{2n}}{1 + \frac{z^2_{\alpha/2}}{n}} + \frac{z_{\alpha/2}}{1 + \frac{z^2_{\alpha/2}}{n}} \sqrt{\frac{\hat{p}\hat{q}}{n} + \frac{z^2_{\alpha/2}}{4n^2}},$$

where $z_{\alpha/2}$ is the z-value leaving an area of $\alpha/2$ to the right.

When n is small and the unknown proportion p is believed to be close to 0 or to 1, the confidence-interval procedure established here is unreliable and, therefore, should not be used. To be on the safe side, one should require both $n\hat{p}$ and $n\hat{q}$ to be greater than or equal to 5. The methods for finding a confidence interval for the binomial parameter p are also applicable when the binomial distribution is being used to approximate the hypergeometric distribution, that is, when n is small relative to N, as illustrated by Example 9.14.

Note that although method 2 yields more accurate results, it is more complicated to calculate, and the gain in accuracy that it provides diminishes when the sample size is large enough. Hence, method 1 is commonly used in practice.

Example 9.14: In a random sample of $n = 500$ families owning television sets in the city of Hamilton, Canada, it is found that $x = 340$ subscribe to HBO. Find a 95% confidence interval for the actual proportion of families with television sets in this city that subscribe to HBO.

***Solution*:** The point estimate of p is $\hat{p} = 340/500 = 0.68$. Using Table A.3, we find that $z_{0.025} = 1.96$. Therefore, using method 1, the 95% confidence interval for p is

$$0.68 - 1.96\sqrt{\frac{(0.68)(0.32)}{500}} < p < 0.68 + 1.96\sqrt{\frac{(0.68)(0.32)}{500}},$$

which simplifies to $0.6391 < p < 0.7209$.

If we use method 2, we can obtain

$$\frac{0.68 + \frac{1.96^2}{(2)(500)}}{1 + \frac{1.96^2}{500}} \pm \frac{1.96}{1 + \frac{1.96^2}{500}} \sqrt{\frac{(0.68)(0.32)}{500} + \frac{1.96^2}{(4)(500^2)}} = 0.6786 \pm 0.0408,$$

which simplifies to $0.6378 < p < 0.7194$. Apparently, when n is large (500 here), both methods yield very similar results.

If p is the center value of a $100(1-\alpha)\%$ confidence interval, then $\hat{p}$ estimates p without error. Most of the time, however, $\hat{p}$ will not be exactly equal to p and the point estimate will be in error. The size of this error will be the positive difference that separates p and $\hat{p}$, and we can be $100(1-\alpha)\%$ confident that this difference will not exceed $z_{\alpha/2}\sqrt{\hat{p}\hat{q}/n}$. We can readily see this if we draw a diagram of a typical confidence interval, as in Figure 9.6. Here we use method 1 to estimate the error.

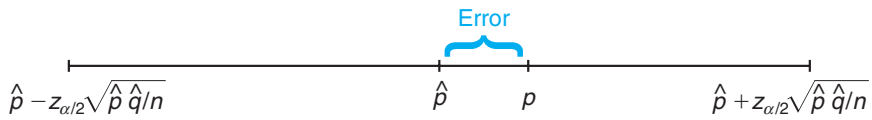


Figure 9.6: Error in estimating p by $\hat{p}$.

Theorem 9.3: If $\hat{p}$ is used as an estimate of p, we can be $100(1-\alpha)\%$ confident that the error will not exceed $z_{\alpha/2}\sqrt{\hat{p}\hat{q}/n}$.

In Example 9.14, we are 95% confident that the sample proportion $\hat{p} = 0.68$ differs from the true proportion p by an amount not exceeding 0.04.

Choice of Sample Size

Let us now determine how large a sample is necessary to ensure that the error in estimating p will be less than a specified amount e. By Theorem 9.3, we must choose n such that $z_{\alpha/2}\sqrt{\hat{p}\hat{q}/n} = e$.

Theorem 9.4: If $\hat{p}$ is used as an estimate of p, we can be $100(1-\alpha)\%$ confident that the error will be less than a specified amount e when the sample size is approximately

$$n = \frac{z_{\alpha/2}^2 \hat{p}\hat{q}}{e^2}.$$

Theorem 9.4 is somewhat misleading in that we must use $\hat{p}$ to determine the sample size n, but $\hat{p}$ is computed from the sample. If a crude estimate of p can be made without taking a sample, this value can be used to determine n. Lacking such an estimate, we could take a preliminary sample of size $n \geq 30$ to provide an estimate of p. Using Theorem 9.4, we could determine approximately how many observations are needed to provide the desired degree of accuracy. Note that fractional values of n are rounded up to the next whole number.

Example 9.15: How large a sample is required if we want to be 95% confident that our estimate of p in Example 9.14 is within 0.02 of the true value?

***Solution*:** Let us treat the 500 families as a preliminary sample, providing an estimate $\hat{p} = 0.68$. Then, by Theorem 9.4,

$$n = \frac{(1.96)^2(0.68)(0.32)}{(0.02)^2} = 2089.8 \approx 2090.$$

Therefore, if we base our estimate of p on a random sample of size 2090, we can be 95% confident that our sample proportion will not differ from the true proportion by more than 0.02. ∎

Occasionally, it will be impractical to obtain an estimate of p to be used for determining the sample size for a specified degree of confidence. If this happens, an upper bound for n is established by noting that $\hat{p}\hat{q} = \hat{p}(1 - \hat{p})$, which must be at most 1/4, since $\hat{p}$ must lie between 0 and 1. This fact may be verified by completing the square. Hence

$$\hat{p}(1 - \hat{p}) = -(\hat{p}^2 - \hat{p}) = \frac{1}{4} - \left(\hat{p}^2 - \hat{p} + \frac{1}{4}\right) = \frac{1}{4} - \left(\hat{p} - \frac{1}{2}\right)^2,$$

which is always less than 1/4 except when $\hat{p} = 1/2$, and then $\hat{p}\hat{q} = 1/4$. Therefore, if we substitute $\hat{p} = 1/2$ into the formula for n in Theorem 9.4 when, in fact, p actually differs from l/2, n will turn out to be larger than necessary for the specified degree of confidence; as a result, our degree of confidence will increase.

Theorem 9.5: If $\hat{p}$ is used as an estimate of p, we can be **at least** $100(1 - \alpha)\%$ confident that the error will not exceed a specified amount e when the sample size is

$$n = \frac{z_{\alpha/2}^2}{4e^2}.$$

Example 9.16: How large a sample is required if we want to be at least 95% confident that our estimate of p in Example 9.14 is within 0.02 of the true value?

***Solution*:** Unlike in Example 9.15, we shall now assume that no preliminary sample has been taken to provide an estimate of p. Consequently, we can be at least 95% confident that our sample proportion will not differ from the true proportion by more than 0.02 if we choose a sample of size

$$n = \frac{(1.96)^2}{(4)(0.02)^2} = 2401.$$

Comparing the results of Examples 9.15 and 9.16, we see that information concerning p, provided by a preliminary sample or from experience, enables us to choose a smaller sample while maintaining our required degree of accuracy. ∎

9.65 A certain geneticist is interested in the proportion of males and females in the population who have a minor blood disorder. In a random sample of 1000 males, 250 are found to be afflicted, whereas 275 of 1000 females tested appear to have the disorder. Compute a 95% confidence interval for the difference between the proportions of males and females who have the blood disorder.

9.66 Ten engineering schools in the United States were surveyed. The sample contained 250 electrical engineers, 80 being women; 175 chemical engineers, 40 being women. Compute a 90% confidence interval for the difference between the proportions of women in these two fields of engineering. Is there a significant difference between the two proportions?

9.67 A clinical trial was conducted to determine if a certain type of inoculation has an effect on the incidence of a certain disease. A sample of 1000 rats was kept in a controlled environment for a period of 1 year, and 500 of the rats were given the inoculation. In the group not inoculated, there were 120 incidences of the disease, while 98 of the rats in the inoculated group contracted it. If p_1 is the probability of incidence of the disease in uninoculated rats and p_2 the probability of incidence in inoculated rats, compute a 90% confidence interval for $p_1 - p_2$.

9.68 In the study *Germination and Emergence of Broccoli*, conducted by the Department of Horticulture at Virginia Tech, a researcher found that at 5°C, 10 broccoli seeds out of 20 germinated, while at 15°C, 15 out of 20 germinated. Compute a 95% confidence interval for the difference between the proportions of germination at the two different temperatures and decide if there is a significant difference.

9.69 A survey of 1000 students found that 274 chose professional baseball team A as their favorite team. In a similar survey involving 760 students, 240 of them chose team A as their favorite. Compute a 95% confidence interval for the difference between the proportions of students favoring team A in the two surveys. Is there a significant difference?

9.70 According to *USA Today* (March 17, 1997), women made up 33.7% of the editorial staff at local TV stations in the United States in 1990 and 36.2% in 1994. Assume 20 new employees were hired as editorial staff.

(a) Estimate the number that would have been women in 1990 and 1994, respectively.

(b) Compute a 95% confidence interval to see if there is evidence that the proportion of women hired as editorial staff was higher in 1994 than in 1990.

9.12 Single Sample: Estimating the Variance

If a sample of size n is drawn from a normal population with variance σ^2 and the sample variance s^2 is computed, we obtain a value of the statistic S^2. This computed sample variance is used as a point estimate of σ^2. Hence, the statistic S^2 is called an estimator of σ^2.

An interval estimate of σ^2 can be established by using the statistic

$$X^2 = \frac{(n-1)S^2}{\sigma^2}.$$

According to Theorem 8.4, the statistic X^2 has a chi-squared distribution with $n-1$ degrees of freedom when samples are chosen from a normal population. We may write (see Figure 9.7)

$$P(\chi^2_{1-\alpha/2} < X^2 < \chi^2_{\alpha/2}) = 1 - \alpha,$$

where $\chi^2_{1-\alpha/2}$ and $\chi^2_{\alpha/2}$ are values of the chi-squared distribution with $n-1$ degrees of freedom, leaving areas of $1-\alpha/2$ and $\alpha/2$, respectively, to the right. Substituting for X^2, we write

$$P\left[\chi^2_{1-\alpha/2} < \frac{(n-1)S^2}{\sigma^2} < \chi^2_{\alpha/2}\right] = 1 - \alpha.$$

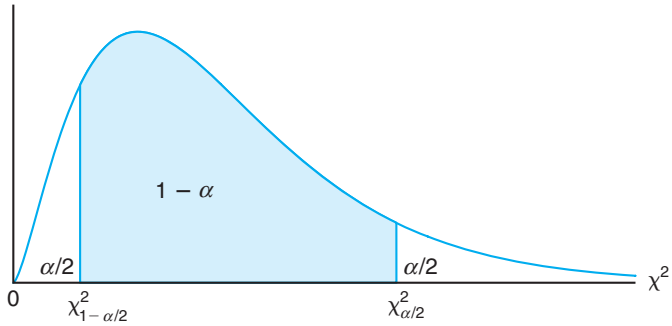


Figure 9.7: $P(\chi^2_{1-\alpha/2} < X^2 < \chi^2_{\alpha/2}) = 1 - \alpha$.

Dividing each term in the inequality by $(n-1)S^2$ and then inverting each term (thereby changing the sense of the inequalities), we obtain

$$P\left[\frac{(n-1)S^2}{\chi^2_{\alpha/2}} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{1-\alpha/2}}\right] = 1 - \alpha.$$

For a random sample of size n from a normal population, the sample variance s^2 is computed, and the following $100(1-\alpha)\%$ confidence interval for σ^2 is obtained.

Confidence Interval for σ^2

If s^2 is the variance of a random sample of size n from a normal population, a $100(1-\alpha)\%$ confidence interval for σ^2 is

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}},$$

where $\chi^2_{\alpha/2}$ and $\chi^2_{1-\alpha/2}$ are χ^2-values with $v = n-1$ degrees of freedom, leaving areas of $\alpha/2$ and $1-\alpha/2$, respectively, to the right.

An approximate $100(1-\alpha)\%$ confidence interval for σ is obtained by taking the square root of each endpoint of the interval for σ^2.

Example 9.18: The following are the weights, in decagrams, of 10 packages of grass seed distributed by a certain company: 46.4, 46.1, 45.8, 47.0, 46.1, 45.9, 45.8, 46.9, 45.2, and 46.0. Find a 95% confidence interval for the variance of the weights of all such packages of grass seed distributed by this company, assuming a normal population.

***Solution*:** First we find

$$s^2 = \frac{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}{n(n-1)} = \frac{(10)(21,273.12) - (461.2)^2}{(10)(9)} = 0.286.$$

To obtain a 95% confidence interval, we choose $\alpha = 0.05$. Then, using Table A.5 with $v = 9$ degrees of freedom, we find $\chi^2_{0.025} = 19.023$ and $\chi^2_{0.975} = 2.700$. Therefore, the 95% confidence interval for σ^2 is

$$\frac{(9)(0.286)}{19.023} < \sigma^2 < \frac{(9)(0.286)}{2.700},$$

or simply $0.135 < \sigma^2 < 0.953$.

9.13 Two Samples: Estimating the Ratio of Two Variances

A point estimate of the ratio of two population variances σ_1^2/σ_2^2 is given by the ratio s_1^2/s_2^2 of the sample variances. Hence, the statistic S_1^2/S_2^2 is called an estimator of σ_1^2/σ_2^2.

If σ_1^2 and σ_2^2 are the variances of normal populations, we can establish an interval estimate of σ_1^2/σ_2^2 by using the statistic

$$F = \frac{\sigma_2^2 S_1^2}{\sigma_1^2 S_2^2}.$$

According to Theorem 8.8, the random variable F has an F-distribution with $v_1 = n_1 - 1$ and $v_2 = n_2 - 1$ degrees of freedom. Therefore, we may write (see Figure 9.8)

$$P[f_{1-\alpha/2}(v_1, v_2) < F < f_{\alpha/2}(v_1, v_2)] = 1 - \alpha,$$

where $f_{1-\alpha/2}(v_1, v_2)$ and $f_{\alpha/2}(v_1, v_2)$ are the values of the F-distribution with v_1 and v_2 degrees of freedom, leaving areas of $1 - \alpha/2$ and $\alpha/2$, respectively, to the right.

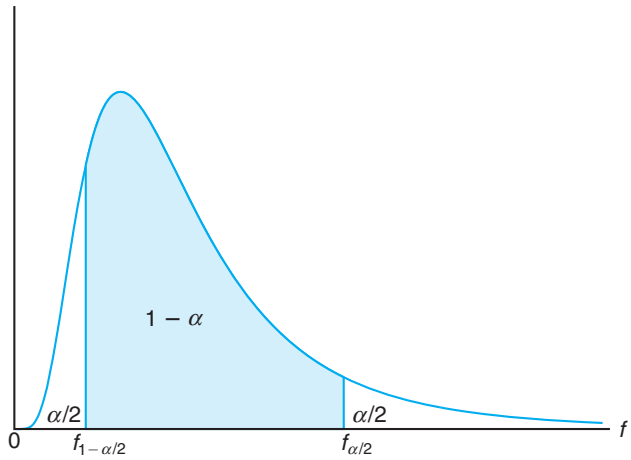


Figure 9.8: $P[f_{1-\alpha/2}(v_1, v_2) < F < f_{\alpha/2}(v_1, v_2)] = 1 - \alpha$.

1.5 Experiments

Studies where the researchers assign treatments to cases are called **experiments**. When this assignment includes randomization, e.g. using a coin flip to decide which treatment a patient receives, it is called a **randomized experiment**. Randomized experiments are fundamentally important when trying to show a causal connection between two variables.

1.5.1 Principles of experimental design

Randomized experiments are generally built on four principles.

Controlling. Researchers assign treatments to cases, and they do their best to **control** any other differences in the groups. For example, when patients take a drug in pill form, some patients take the pill with only a sip of water while others may have it with an entire glass of water. To control for the effect of water consumption, a doctor may ask all patients to drink a 12 ounce glass of water with the pill.

Randomization. Researchers randomize patients into treatment groups to account for variables that cannot be controlled. For example, some patients may be more susceptible to a disease than others due to their dietary habits. Randomizing patients into the treatment or control group helps even out such differences, and it also prevents accidental bias from entering the study.

Replication. The more cases researchers observe, the more accurately they can estimate the effect of the explanatory variable on the response. In a single study, we **replicate** by collecting a sufficiently large sample. Additionally, a group of scientists may replicate an entire study to verify an earlier finding.

Blocking. Researchers sometimes know or suspect that variables, other than the treatment, influence the response. Under these circumstances, they may first group individuals based on this variable into **blocks** and then randomize cases within each block to the treatment groups. This strategy is often referred to as **blocking**. For instance, if we are looking at the effect of a drug on heart attacks, we might first split patients in the study into low-risk and high-risk blocks, then randomly assign half the patients from each block to the control group and the other half to the treatment group, as shown in Figure 1.16. This strategy ensures each treatment group has an equal number of low-risk and high-risk patients.

It is important to incorporate the first three experimental design principles into any study, and this book describes applicable methods for analyzing data from such experiments. Blocking is a slightly more advanced technique, and statistical methods in this book may be extended to analyze data collected using blocking.

1.5.2 Reducing bias in human experiments

Randomized experiments are the gold standard for data collection, but they do not ensure an unbiased perspective into the cause and effect relationships in all cases. Human studies are perfect examples where bias can unintentionally arise. Here we reconsider a study where a new drug was used to treat heart attack patients.[16] In particular, researchers wanted to know if the drug reduced deaths in patients.

[16] Anturane Reinfarction Trial Research Group. 1980. Sulfinpyrazone in the prevention of sudden death after myocardial infarction. New England Journal of Medicine 302(5):250-256.

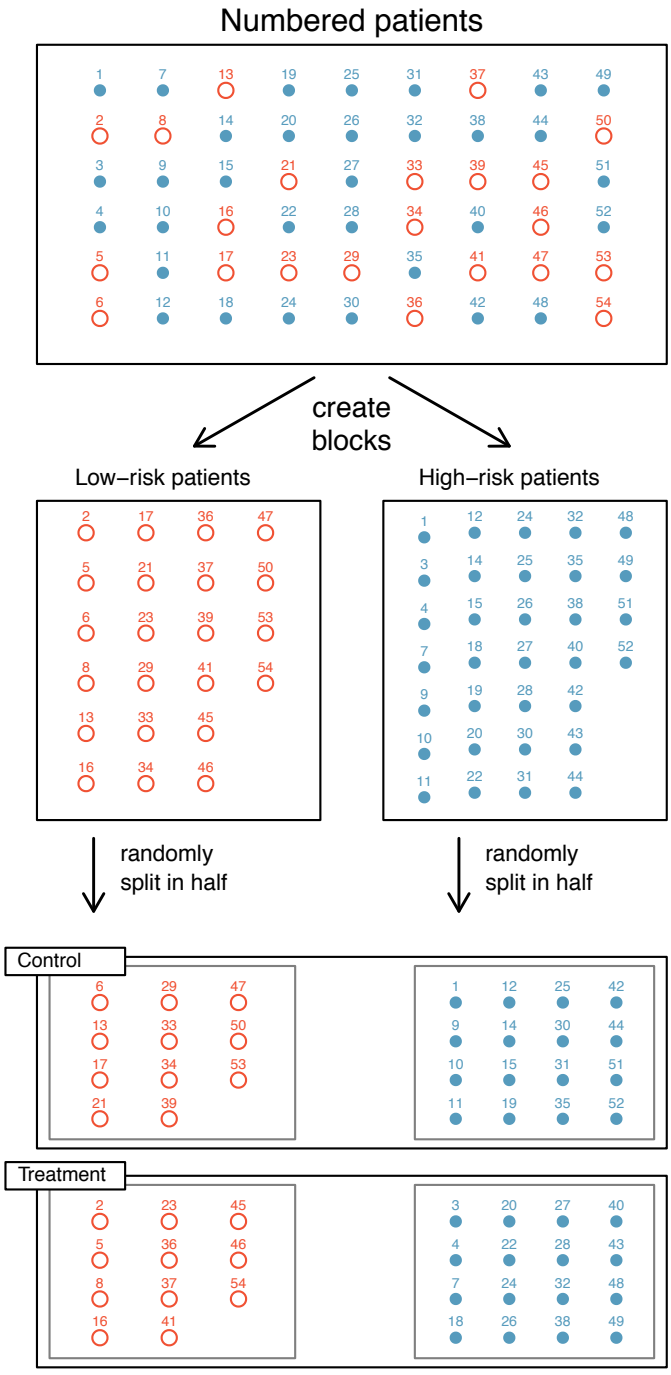


Figure 1.16: Blocking using a variable depicting patient risk. Patients are first divided into low-risk and high-risk blocks, then each block is evenly separated into the treatment groups using randomization. This strategy ensures an equal representation of patients in each treatment group from both the low-risk and high-risk categories.

These researchers designed a randomized experiment because they wanted to draw causal conclusions about the drug's effect. Study volunteers[17] were randomly placed into two study groups. One group, the **treatment group**, received the drug. The other group, called the **control group**, did not receive any drug treatment.

Put yourself in the place of a person in the study. If you are in the treatment group, you are given a fancy new drug that you anticipate will help you. On the other hand, a person in the other group doesn't receive the drug and sits idly, hoping her participation doesn't increase her risk of death. These perspectives suggest there are actually two effects: the one of interest is the effectiveness of the drug, and the second is an emotional effect that is difficult to quantify.

Researchers aren't usually interested in the emotional effect, which might bias the study. To circumvent this problem, researchers do not want patients to know which group they are in. When researchers keep the patients uninformed about their treatment, the study is said to be **blind**. But there is one problem: if a patient doesn't receive a treatment, she will know she is in the control group. The solution to this problem is to give fake treatments to patients in the control group. A fake treatment is called a **placebo**, and an effective placebo is the key to making a study truly blind. A classic example of a placebo is a sugar pill that is made to look like the actual treatment pill. Often times, a placebo results in a slight but real improvement in patients. This effect has been dubbed the **placebo effect**.

The patients are not the only ones who should be blinded: doctors and researchers can accidentally bias a study. When a doctor knows a patient has been given the real treatment, she might inadvertently give that patient more attention or care than a patient that she knows is on the placebo. To guard against this bias, which again has been found to have a measurable effect in some instances, most modern studies employ a **double-blind** setup where doctors or researchers who interact with patients are, just like the patients, unaware of who is or is not receiving the treatment.[18]

⊙ **Guided Practice 1.14** Look back to the study in Section 1.1 where researchers were testing whether stents were effective at reducing strokes in at-risk patients. Is this an experiment? Was the study blinded? Was it double-blinded?[19]

1.6 Examining numerical data

In this section we will be introduced to techniques for exploring and summarizing numerical variables. The `email50` and `county` data sets from Section 1.2 provide rich opportunities for examples. Recall that outcomes of numerical variables are numbers on which it is reasonable to perform basic arithmetic operations. For example, the `pop2010` variable, which represents the populations of counties in 2010, is numerical since we can sensibly discuss the difference or ratio of the populations in two counties. On the other hand, area codes and zip codes are not numerical, but rather they are categorical variables.

[17] Human subjects are often called **patients**, **volunteers**, or **study participants**.

[18] There are always some researchers involved in the study who do know which patients are receiving which treatment. However, they do not interact with the study's patients and do not tell the blinded health care professionals who is receiving which treatment.

[19] The researchers assigned the patients into their treatment groups, so this study was an experiment. However, the patients could distinguish what treatment they received, so this study was not blind. The study could not be double-blind since it was not blind.

približno $100\gamma\,\%$ slučajeva interval izračunat po formuli

$$\left[\hat{p} - z_\gamma \sqrt{\frac{\hat{p}\hat{q}}{n}},\ \hat{p} + z_\gamma \sqrt{\frac{\hat{p}\hat{q}}{n}}\right],$$

$\hat{p}$ — relativna frekvencija jedinice (uspjeha) u uzorku
$\hat{q}$ — relativna frekvencija nula (neuspjeha) u uzorku, $\hat{q} = 1 - \hat{p}$
z_γ — broj za koji vrijedi $P\{|Z| \leq z_\gamma\} = \gamma$
Z — standardna normalna slučajna varijabla

sadržavati pravu (nepoznatu) vrijednost vjerojatnosti p. Također se može pokazati da je broj elemenata u uzorku (n) dovoljno velik za primjenu ovakvog zaključivanja ako interval

$$\left[\hat{p} - 3\sqrt{\frac{\hat{p}\hat{q}}{n}}, \hat{p} + 3\sqrt{\frac{\hat{p}\hat{q}}{n}}\right]$$

ne sadrži ni 0 ni 1. Uočimo da iz ovog razmatranja možemo odrediti veličinu uzorka koja će osigurati zadanu preciznost procjene pouzdanim intervalom, tj. zadanu duljinu intervala.

Primjer 5.7. *Jedna tvornica hrane želi provesti istraživanje tržišta intervjuirajući* 1000 *potrošača kako bi odredila koju marku pahuljica za doručak preferiraju. Prikupljeni podaci pokazali su da* 313 *potrošača odabire upravo marku tvornice koja je provela istraživanje. Na temelju rezultata tog istraživanja možemo odrediti jednu realizaciju intervala pouzdanosti* 95 % *kojim procjenjujemo vjerojatnost da slučajno odabrani potrošač preferira pahuljice tvornice koja je provela istraživanje:*

$$\hat{p} - z_\gamma \sqrt{\frac{\hat{p}\hat{q}}{n}} = 0.313 - 1.96\sqrt{\frac{0.313 \cdot 0.687}{1000}} = 0.284,$$

$$\hat{p} + z_\gamma \sqrt{\frac{\hat{p}\hat{q}}{n}} = 0.313 + 1.96\sqrt{\frac{0.313 \cdot 0.687}{1000}} = 0.342.$$

Dakle, realizacija intervala pouzdanosti 95 % *temeljena na rezultatima istraživanja je interval realnih brojeva* $[0.284, 0.342]$. *Uočimo da taj pouzdani interval možemo interpretirati i kao pouzdani interval proporcije potrošača koji preferiraju danu marku pahuljica za doručak.*

5.4 Testiranje hipoteza

Pretpostavimo da želimo provjeriti je li očekivana vrijednost vremena čekanja u redu studentske menze u vrijeme ručka veća od pet minuta. Naime, ako je veća, onda ćemo u vrijeme ručka pokrenuti još jednu traku u menzi. U tu svrhu od sto slučajno izabranih studenata koji odlaze na ručak u studentsku menzu prikupljamo podatke

o vremenu čekanja za vrijeme ručka. Tako dolazimo do podataka $(x_1, \ldots, x_{100})$ koji su jedna realizacija slučajnog uzorka $(X_1, \ldots, X_{100})$ iz neke, nama nepoznate, distribucije. Da bismo donijeli odluku o pokretanju još jedne trake u menzi, potrebno je testirati hipotezu o iznosu očekivanog vremena čekanja u redu na temelju prikupljenih podataka $(x_1, \ldots, x_{100})$. Takvim i sličnim problemima bavi se teorija testiranja statističkih hipoteza.

Za testiranje hipoteze vezane uz varijablu koja nas zanima koristimo modeliranje varijable kao što je opisano u prethodnim poglavljima, tj. varijable u ispitavanju su slučajne varijable. Slučajna varijabla određena je svojom distribucijom. Kao što je već rečeno, distribucije nam nisu u potpunosti poznate, ali smo naučili kako možemo pribaviti neke informacije o distribuciji na osnovi teorije procjene. **Hipotezu koju želimo testirati korištenjem statističkog testa moramo izraziti u terminima hipoteze koja se odnosi na distribuciju slučajne varijable**. Tako u postupku donošenja odluke o otvaranju nove trake u studentskoj menzi treba testirati jednu hipotezu o vrijednosti očekivanja slučajne varijable koja opisuje vrijeme čekanja u redu studentske menze za vrijeme ručka. Hipotezu koja je formulirana u terminima distribucije slučajne varijable zovemo **statistička hipoteza**.

Postupak testiranja hipoteza uvijek počinje postupkom prevođenja problema koji nas zanima u statističku hipotezu. Primjerice, u uvodnom primjeru u kojem govorimo o mogućnosti otvaranja još jedne trake u studentskoj menzi, u donošenju odluke može nam pomoći testiranje statističke hipoteze da je očekivanje čekanja u redu veće od pet minuta. Statističku hipotezu standardno označavamo s $\mathcal{H}$. **Testirati hipotezu znači donijeti odluku o tome hoćemo li $\mathcal{H}$ odbaciti ili prihvatiti**. Zbog toga često govorimo o testiranju dviju hipoteza u statističkom testu. Jednu od njih zovemo **nul-hipoteza** i označavamo s $\mathcal{H}_0$, a drugu **alternativna hipoteza** i označavamo s $\mathcal{H}_1$. **Alternativna hipoteza je ona koju prihvaćamo u slučaju odbacivanja nul-hipoteze**.
Statistički test koji ćemo koristiti za testiranje statističke hipoteze dizajniran je tako da korištenjem informacija iz prikupljenih podataka o realizacijama slučajne varijable donosimo **odluku o odbacivanju nul-hipoteze** u korist alternativne hipoteze ili **neodbacivanju nul-hipoteze**. Uočimo da nul-hipoteza i alternativna hipoteza u ovoj formulaciji nisu ravnopravne, npr. nigdje nije napisano da prihvaćamo nul-hipotezu. Razlog za ovakvo neobično izražavanje leži u činjenici da se odlučivanje u statističkom testu provodi uz toleranciju malih vjerojatnosti pogrešne odluke. Da bismo bolje razumjeli ovaj koncept, opisat ćemo vrste pogrešaka statističkog testa i mogućnosti koje daje test u odnosu na njihovu kontrolu.

5.4.1 Pogreške statističkog testa

Odluka koja je donesena statističkim testom može biti ili pogrešna ili ispravna. Pri tome se mogu dogoditi dva tipa pogrešne odluke:

pogreška I. tipa: odbaciti $\mathcal{H}_0$ ako je ona istinita
pogreška II. tipa: ne odbaciti $\mathcal{H}_0$ ako je $\mathcal{H}_1$ istinita.

Vjerojatnost pogreške prvog tipa i pogreške drugog tipa ovisi o stvarnoj distribuciji slučajne varijable o kojoj testiramo hipotezu. Htjeli bismo da su te vjerojatnosti pogreške što je moguće manje. Postupak kreiranja statističkog testa, tj. definiranje pravila na osnovi kojih ćemo odlučivati, vodi računa upravo o tom zahtjevu. Statistički test dizajniran je tako da dopušta istraživaču izbor maksimalne vjerojatnosti pogreške prvog tipa koju istraživač želi prihvatiti. Te vrijednosti uglavnom se biraju između brojeva 0.01, 0.05 ili 0.1. Odabrana maksimalna vjerojatnost pogreške prvog tipa zove se **razina značajnosti testa** ili **nivo signifikantnosti testa** i standardno označava s α. Vjerojatnost pogreške drugog tipa određena je dizajnom testa uz izabrani nivo signifikantnosti. Testovi se dizajniraju uz nastojanje da se maksimalna vjerojatnost pogreške drugog tipa učini što manjom i ona se, u pravilu, ne iskazuje u primjeni statističkih testova.

Uzimajući u obzir da ćemo biti u mogućnosti birati maksimalnu vjerojatnost pogreške prilikom odbacivanja nul-hipoteze, to je informacija koju u primjeni testa referiramo. Npr. reći ćemo da **odbacujemo nul-hipotezu na nivou značajnosti α i prihvaćamo hipotezu $\mathcal{H}_1$**, što će značiti da prihvaćamo alternativnu hipotezu uz vjerojatnost najviše α da smo pri tome pogriješili. U suprotnom ćemo reći kako podaci ne podupiru tvrdnju da $\mathcal{H}_0$ treba odbaciti.

Ovakav neravnopravan odnos između nul-hipoteze i alternativne hipoteze prilikom kreiranja statističkog testa upućuje na činjenicu da nije svejedno kako smo izabrali hipoteze i pripadni test. **Ako je moguće, uputno je u primjeni birati statistički test tako da alternativna hipoteza odgovara tvrdnji koju želimo dokazati.**

5.5 Testiranje hipoteza o očekivanju

U ovom poglavlju pokazat ćemo nekoliko statističkih testova koje možemo koristiti prilikom rješavanja problema koji se mogu modelirati analogno kao problem u primjeru o otvaranju nove trake u studentskoj menzi iz prethodnog poglavlja. Način razmišljanja koji treba slijediti u problemima tog tipa objašnjen je u primjeru 5.8.

Chapter 10

One- and Two-Sample Tests of Hypotheses

10.1 Statistical Hypotheses: General Concepts

Often, the problem confronting the scientist or engineer is not so much the estimation of a population parameter, as discussed in Chapter 9, but rather the formation of a data-based decision procedure that can produce a conclusion about some scientific system. For example, a medical researcher may decide on the basis of experimental evidence whether coffee drinking increases the risk of cancer in humans; an engineer might have to decide on the basis of sample data whether there is a difference between the accuracy of two kinds of gauges; or a sociologist might wish to collect appropriate data to enable him or her to decide whether a person's blood type and eye color are independent variables. In each of these cases, the scientist or engineer *postulates* or *conjectures* something about a system. In addition, each must make use of experimental data and make a decision based on the data. In each case, the conjecture can be put in the form of a statistical hypothesis. Procedures that lead to the acceptance or rejection of statistical hypotheses such as these comprise a major area of statistical inference. First, let us define precisely what we mean by a **statistical hypothesis**.

Definition 10.1: A **statistical hypothesis** is an assertion or conjecture concerning one or more populations.

The truth or falsity of a statistical hypothesis is never known with absolute certainty unless we examine the entire population. This, of course, would be impractical in most situations. Instead, we take a random sample from the population of interest and use the data contained in this sample to provide evidence that either supports or does not support the hypothesis. Evidence from the sample that is inconsistent with the stated hypothesis leads to a rejection of the hypothesis.

The Role of Probability in Hypothesis Testing

It should be made clear to the reader that the decision procedure must include an awareness of the *probability of a wrong conclusion*. For example, suppose that the hypothesis postulated by the engineer is that the fraction defective p in a certain process is 0.10. The experiment is to observe a random sample of the product in question. Suppose that 100 items are tested and 12 items are found defective. It is reasonable to conclude that this evidence does not refute the condition that the binomial parameter $p = 0.10$, and thus it may lead one not to reject the hypothesis. However, it also does not refute $p = 0.12$ or perhaps even $p = 0.15$. As a result, the reader must be accustomed to understanding that **rejection of a hypothesis implies that the sample evidence refutes it**. Put another way, **rejection means that there is a small probability of obtaining the sample information observed when, in fact, the hypothesis is true**. For example, for our proportion-defective hypothesis, a sample of 100 revealing 20 defective items is certainly evidence for rejection. Why? If, indeed, $p = 0.10$, the probability of obtaining 20 or more defectives is approximately 0.002. With the resulting small risk of a wrong conclusion, it would seem safe to **reject the hypothesis** that $p = 0.10$. In other words, rejection of a hypothesis tends to all but "rule out" the hypothesis. On the other hand, it is very important to emphasize that acceptance or, rather, failure to reject does not rule out other possibilities. As a result, the *firm conclusion is established by the data analyst when a hypothesis is rejected.*

The formal statement of a hypothesis is often influenced by the structure of the probability of a wrong conclusion. If the scientist is interested in *strongly supporting* a contention, he or she hopes to arrive at the contention in the form of rejection of a hypothesis. If the medical researcher wishes to show strong evidence in favor of the contention that coffee drinking increases the risk of cancer, the hypothesis tested should be of the form "there is no increase in cancer risk produced by drinking coffee." As a result, the contention is reached via a rejection. Similarly, to support the claim that one kind of gauge is more accurate than another, the engineer tests the hypothesis that there is no difference in the accuracy of the two kinds of gauges.

The foregoing implies that when the data analyst formalizes experimental evidence on the basis of hypothesis testing, the formal **statement of the hypothesis** is very important.

The Null and Alternative Hypotheses

The structure of hypothesis testing will be formulated with the use of the term **null hypothesis**, which refers to any hypothesis we wish to test and is denoted by H_0. The rejection of H_0 leads to the acceptance of an **alternative hypothesis**, denoted by H_1. An understanding of the different roles played by the null hypothesis (H_0) and the alternative hypothesis (H_1) is crucial to one's understanding of the rudiments of hypothesis testing. The alternative hypothesis H_1 usually represents the *question to be answered or the theory to be tested,* and thus its specification is crucial. The null hypothesis H_0 *nullifies or opposes* H_1 and is often the logical complement to H_1. As the reader gains more understanding of hypothesis testing, he or she should note that the analyst arrives at one of the two following

conclusions:

***reject* H_0** in favor of H_1 because of sufficient evidence in the data or

***fail to reject* H_0** because of insufficient evidence in the data.

Note that the *conclusions do not involve a formal and literal "accept* H_0*."* The statement of H_0 often represents the "status quo" in opposition to the new idea, conjecture, and so on, stated in H_1, while failure to reject H_0 represents the proper conclusion. In our binomial example, the practical issue may be a concern that the historical defective probability of 0.10 no longer is true. Indeed, the conjecture may be that p exceeds 0.10. We may then state

$$H_0\text{: } p = 0.10,$$
$$H_1\text{: } p > 0.10.$$

Now 12 defective items out of 100 does not refute $p = 0.10$, so the conclusion is "fail to reject H_0." However, if the data produce 20 out of 100 defective items, then the conclusion is "reject H_0" in favor of H_1: $p > 0.10$.

Though the applications of hypothesis testing are quite abundant in scientific and engineering work, perhaps the best illustration for a novice lies in the predicament encountered in a jury trial. The null and alternative hypotheses are

$$H_0\text{: defendant is innocent,}$$
$$H_1\text{: defendant is guilty.}$$

The indictment comes because of suspicion of guilt. The hypothesis H_0 (the status quo) stands in opposition to H_1 and is maintained unless H_1 is supported by evidence "beyond a reasonable doubt." However, "failure to reject H_0" in this case does not imply innocence, but merely that the evidence was insufficient to convict. So the jury does not necessarily *accept* H_0 but *fails to reject* H_0.

10.2 Testing a Statistical Hypothesis

To illustrate the concepts used in testing a statistical hypothesis about a population, we present the following example. A certain type of cold vaccine is known to be only 25% effective after a period of 2 years. To determine if a new and somewhat more expensive vaccine is superior in providing protection against the same virus for a longer period of time, suppose that 20 people are chosen at random and inoculated. (In an actual study of this type, the participants receiving the new vaccine might number several thousand. The number 20 is being used here only to demonstrate the basic steps in carrying out a statistical test.) If more than 8 of those receiving the new vaccine surpass the 2-year period without contracting the virus, the new vaccine will be considered superior to the one presently in use. The requirement that the number exceed 8 is somewhat arbitrary but appears reasonable in that it represents a modest gain over the 5 people who could be expected to receive protection if the 20 people had been inoculated with the vaccine already in use. We are essentially testing the null hypothesis that the new vaccine is equally effective after a period of 2 years as the one now commonly used. The alternative

hypothesis is that the new vaccine is in fact superior. This is equivalent to testing the hypothesis that the binomial parameter for the probability of a success on a given trial is $p = 1/4$ against the alternative that $p > 1/4$. This is usually written as follows:

$$H_0: \ p = 0.25,$$
$$H_1: \ p > 0.25.$$

The Test Statistic

The **test statistic** on which we base our decision is X, the number of individuals in our test group who receive protection from the new vaccine for a period of at least 2 years. The possible values of X, from 0 to 20, are divided into two groups: those numbers less than or equal to 8 and those greater than 8. All possible scores greater than 8 constitute the **critical region**. The last number that we observe in passing into the critical region is called the **critical value**. In our illustration, the critical value is the number 8. Therefore, if $x > 8$, we reject H_0 in favor of the alternative hypothesis H_1. If $x \leq 8$, we fail to reject H_0. This decision criterion is illustrated in Figure 10.1.

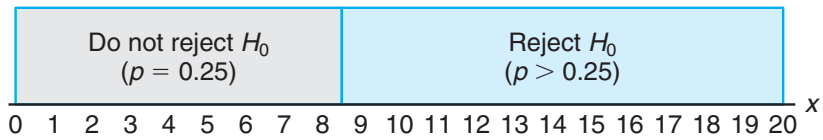


Figure 10.1: Decision criterion for testing $p = 0.25$ versus $p > 0.25$.

The Probability of a Type I Error

The decision procedure just described could lead to either of two wrong conclusions. For instance, the new vaccine may be no better than the one now in use (H_0 true) and yet, in this particular randomly selected group of individuals, more than 8 surpass the 2-year period without contracting the virus. We would be committing an error by rejecting H_0 in favor of H_1 when, in fact, H_0 is true. Such an error is called a **type I error**.

Definition 10.2: Rejection of the null hypothesis when it is true is called a **type I error**.

A second kind of error is committed if 8 or fewer of the group surpass the 2-year period successfully and we are unable to conclude that the vaccine is better when it actually is better (H_1 true). Thus, in this case, we fail to reject H_0 when in fact H_0 is false. This is called a **type II error**.

Definition 10.3: Nonrejection of the null hypothesis when it is false is called a **type II error**.

In testing any statistical hypothesis, there are four possible situations that determine whether our decision is correct or in error. These four situations are

summarized in Table 10.1.

Table 10.1: Possible Situations for Testing a Statistical Hypothesis

	H_0 **is true**	H_0 **is false**
Do not reject H_0	Correct decision	Type II error
Reject H_0	Type I error	Correct decision

The probability of committing a type I error, also called the **level of significance**, is denoted by the Greek letter α. In our illustration, a type I error will occur when more than 8 individuals inoculated with the new vaccine surpass the 2-year period without contracting the virus and researchers conclude that the new vaccine is better when it is actually equivalent to the one in use. Hence, if X is the number of individuals who remain free of the virus for at least 2 years,

$$\alpha = P(\text{type I error}) = P\left(X > 8 \text{ when } p = \frac{1}{4}\right) = \sum_{x=9}^{20} b\left(x; 20, \frac{1}{4}\right)$$

$$= 1 - \sum_{x=0}^{8} b\left(x; 20, \frac{1}{4}\right) = 1 - 0.9591 = 0.0409.$$

We say that the null hypothesis, $p = 1/4$, is being tested at the $\alpha = 0.0409$ level of significance. Sometimes the level of significance is called the **size of the test**. A critical region of size 0.0409 is very small, and therefore it is unlikely that a type I error will be committed. Consequently, it would be most unusual for more than 8 individuals to remain immune to a virus for a 2-year period using a new vaccine that is essentially equivalent to the one now on the market.

The Probability of a Type II Error

The probability of committing a type II error, denoted by β, is impossible to compute unless we have a specific alternative hypothesis. If we test the null hypothesis that $p = 1/4$ against the alternative hypothesis that $p = 1/2$, then we are able to compute the probability of not rejecting H_0 when it is false. We simply find the probability of obtaining 8 or fewer in the group that surpass the 2-year period when $p = 1/2$. In this case,

$$\beta = P(\text{type II error}) = P\left(X \leq 8 \text{ when } p = \frac{1}{2}\right)$$

$$= \sum_{x=0}^{8} b\left(x; 20, \frac{1}{2}\right) = 0.2517.$$

This is a rather high probability, indicating a test procedure in which it is quite likely that we shall reject the new vaccine when, in fact, it is superior to what is now in use. Ideally, we like to use a test procedure for which the type I and type II error probabilities are both small.

It is possible that the director of the testing program is willing to make a type II error if the more expensive vaccine is not significantly superior. In fact, the only

time he wishes to guard against the type II error is when the true value of p is at least 0.7. If $p = 0.7$, this test procedure gives

$$\beta = P(\text{type II error}) = P(X \leq 8 \text{ when } p = 0.7)$$
$$= \sum_{x=0}^{8} b(x; 20, 0.7) = 0.0051.$$

With such a small probability of committing a type II error, it is extremely unlikely that the new vaccine would be rejected when it was 70% effective after a period of 2 years. As the alternative hypothesis approaches unity, the value of β diminishes to zero.

The Role of α, β, and Sample Size

Let us assume that the director of the testing program is unwilling to commit a type II error when the alternative hypothesis $p = 1/2$ is true, even though we have found the probability of such an error to be $\beta = 0.2517$. It is always possible to reduce β by increasing the size of the critical region. For example, consider what happens to the values of α and β when we change our critical value to 7 so that all scores greater than 7 fall in the critical region and those less than or equal to 7 fall in the nonrejection region. Now, in testing $p = 1/4$ against the alternative hypothesis that $p = 1/2$, we find that

$$\alpha = \sum_{x=8}^{20} b\left(x; 20, \frac{1}{4}\right) = 1 - \sum_{x=0}^{7} b\left(x; 20, \frac{1}{4}\right) = 1 - 0.8982 = 0.1018$$

and

$$\beta = \sum_{x=0}^{7} b\left(x; 20, \frac{1}{2}\right) = 0.1316.$$

By adopting a new decision procedure, we have reduced the probability of committing a type II error at the expense of increasing the probability of committing a type I error. For a fixed sample size, a decrease in the probability of one error will usually result in an increase in the probability of the other error. Fortunately, **the probability of committing both types of error can be reduced by increasing the sample size**. Consider the same problem using a random sample of 100 individuals. If more than 36 of the group surpass the 2-year period, we reject the null hypothesis that $p = 1/4$ and accept the alternative hypothesis that $p > 1/4$. The critical value is now 36. All possible scores above 36 constitute the critical region, and all possible scores less than or equal to 36 fall in the acceptance region.

To determine the probability of committing a type I error, we shall use the normal curve approximation with

$$\mu = np = (100)\left(\frac{1}{4}\right) = 25 \quad \text{and} \quad \sigma = \sqrt{npq} = \sqrt{(100)(1/4)(3/4)} = 4.33.$$

Referring to Figure 10.2, we need the area under the normal curve to the right of $x = 36.5$. The corresponding z-value is

$$z = \frac{36.5 - 25}{4.33} = 2.66.$$

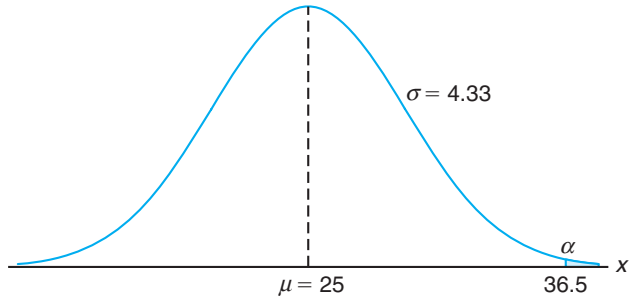


Figure 10.2: Probability of a type I error.

From Table A.3 we find that

$$\alpha = P(\text{type I error}) = P\left(X > 36 \text{ when } p = \frac{1}{4}\right) \approx P(Z > 2.66)$$
$$= 1 - P(Z < 2.66) = 1 - 0.9961 = 0.0039.$$

If H_0 is false and the true value of H_1 is $p = 1/2$, we can determine the probability of a type II error using the normal curve approximation with

$$\mu = np = (100)(1/2) = 50 \quad \text{and} \quad \sigma = \sqrt{npq} = \sqrt{(100)(1/2)(1/2)} = 5.$$

The probability of a value falling in the nonrejection region when H_0 is true is given by the area of the shaded region to the left of $x = 36.5$ in Figure 10.3. The z-value corresponding to $x = 36.5$ is

$$z = \frac{36.5 - 50}{5} = -2.7.$$

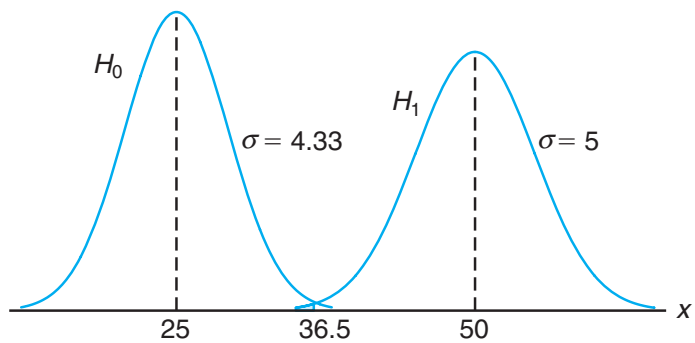


Figure 10.3: Probability of a type II error.

Therefore,

$$\beta = P(\text{type II error}) = P\left(X \leq 36 \text{ when } p = \frac{1}{2}\right) \approx P(Z < -2.7) = 0.0035.$$

Obviously, the type I and type II errors will rarely occur if the experiment consists of 100 individuals.

The illustration above underscores the strategy of the scientist in hypothesis testing. After the null and alternative hypotheses are stated, it is important to consider the sensitivity of the test procedure. By this we mean that there should be a determination, for a fixed α, of a reasonable value for the probability of wrongly accepting H_0 (i.e., the value of β) when the true situation represents some *important deviation from* H_0. A value for the sample size can usually be determined for which there is a reasonable balance between the values of α and β computed in this fashion. The vaccine problem provides an illustration.

Illustration with a Continuous Random Variable

The concepts discussed here for a discrete population can be applied equally well to continuous random variables. Consider the null hypothesis that the average weight of male students in a certain college is 68 kilograms against the alternative hypothesis that it is unequal to 68. That is, we wish to test

$$H_0: \ \mu = 68,$$
$$H_1: \ \mu \neq 68.$$

The alternative hypothesis allows for the possibility that $\mu < 68$ or $\mu > 68$.

A sample mean that falls close to the hypothesized value of 68 would be considered evidence in favor of H_0. On the other hand, a sample mean that is considerably less than or more than 68 would be evidence inconsistent with H_0 and therefore favoring H_1. The sample mean is the test statistic in this case. A critical region for the test statistic might arbitrarily be chosen to be the two intervals $\bar{x} < 67$ and $\bar{x} > 69$. The nonrejection region will then be the interval $67 \leq \bar{x} \leq 69$. This decision criterion is illustrated in Figure 10.4.

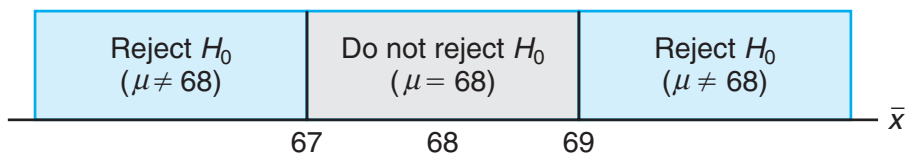


Figure 10.4: Critical region (in blue).

Let us now use the decision criterion of Figure 10.4 to calculate the probabilities of committing type I and type II errors when testing the null hypothesis that $\mu = 68$ kilograms against the alternative that $\mu \neq 68$ kilograms.

Assume the standard deviation of the population of weights to be $\sigma = 3.6$. For large samples, we may substitute s for σ if no other estimate of σ is available. Our decision statistic, based on a random sample of size $n = 36$, will be $\bar{X}$, the most efficient estimator of μ. From the Central Limit Theorem, we know that the sampling distribution of $\bar{X}$ is approximately normal with standard deviation $\sigma_{\bar{X}} = \sigma/\sqrt{n} = 3.6/6 = 0.6$.

The probability of committing a type I error, or the level of significance of our test, is equal to the sum of the areas that have been shaded in each tail of the distribution in Figure 10.5. Therefore,

$$\alpha = P(\bar{X} < 67 \text{ when } \mu = 68) + P(\bar{X} > 69 \text{ when } \mu = 68).$$

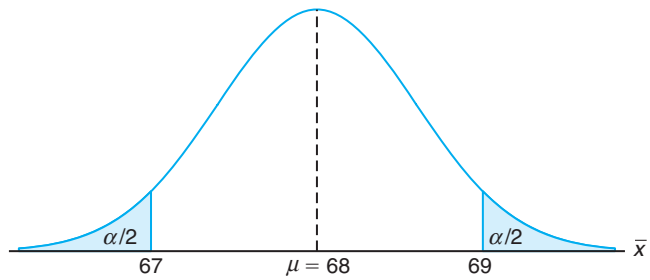


Figure 10.5: Critical region for testing $\mu = 68$ versus $\mu \neq 68$.

The z-values corresponding to $\bar{x}_1 = 67$ and $\bar{x}_2 = 69$ when H_0 is true are

$$z_1 = \frac{67 - 68}{0.6} = -1.67 \quad \text{and} \quad z_2 = \frac{69 - 68}{0.6} = 1.67.$$

Therefore,

$$\alpha = P(Z < -1.67) + P(Z > 1.67) = 2P(Z < -1.67) = 0.0950.$$

Thus, 9.5% of all samples of size 36 would lead us to reject $\mu = 68$ kilograms when, in fact, it is true. To reduce α, we have a choice of increasing the sample size or widening the fail-to-reject region. Suppose that we increase the sample size to $n = 64$. Then $\sigma_{\bar{X}} = 3.6/8 = 0.45$. Now

$$z_1 = \frac{67 - 68}{0.45} = -2.22 \quad \text{and} \quad z_2 = \frac{69 - 68}{0.45} = 2.22.$$

Hence,

$$\alpha = P(Z < -2.22) + P(Z > 2.22) = 2P(Z < -2.22) = 0.0264.$$

The reduction in α is not sufficient by itself to guarantee a good testing procedure. We must also evaluate β for various alternative hypotheses. If it is important to reject H_0 when the true mean is some value $\mu \geq 70$ or $\mu \leq 66$, then the probability of committing a type II error should be computed and examined for the alternatives $\mu = 66$ and $\mu = 70$. Because of symmetry, it is only necessary to consider the probability of not rejecting the null hypothesis that $\mu = 68$ when the alternative $\mu = 70$ is true. A type II error will result when the sample mean $\bar{x}$ falls between 67 and 69 when H_1 is true. Therefore, referring to Figure 10.6, we find that

$$\beta = P(67 \leq \bar{X} \leq 69 \text{ when } \mu = 70).$$

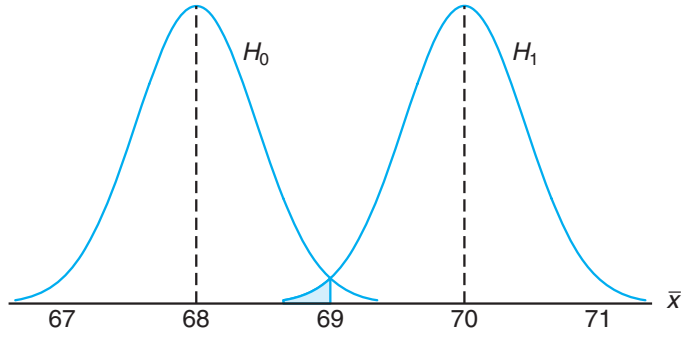


Figure 10.6: Probability of type II error for testing $\mu = 68$ versus $\mu = 70$.

The z-values corresponding to $\bar{x}_1 = 67$ and $\bar{x}_2 = 69$ when H_1 is true are

$$z_1 = \frac{67 - 70}{0.45} = -6.67 \quad \text{and} \quad z_2 = \frac{69 - 70}{0.45} = -2.22.$$

Therefore,

$$\begin{aligned}\beta = P(-6.67 < Z < -2.22) &= P(Z < -2.22) - P(Z < -6.67) \\ &= 0.0132 - 0.0000 = 0.0132.\end{aligned}$$

If the true value of μ is the alternative $\mu = 66$, the value of β will again be 0.0132. For all possible values of $\mu < 66$ or $\mu > 70$, the value of β will be even smaller when $n = 64$, and consequently there would be little chance of not rejecting H_0 when it is false.

The probability of committing a type II error increases rapidly when the true value of μ approaches, but is not equal to, the hypothesized value. Of course, this is usually the situation where we do not mind making a type II error. For example, if the alternative hypothesis $\mu = 68.5$ is true, we do not mind committing a type II error by concluding that the true answer is $\mu = 68$. The probability of making such an error will be high when $n = 64$. Referring to Figure 10.7, we have

$$\beta = P(67 \leq \bar{X} \leq 69 \text{ when } \mu = 68.5).$$

The z-values corresponding to $\bar{x}_1 = 67$ and $\bar{x}_2 = 69$ when $\mu = 68.5$ are

$$z_1 = \frac{67 - 68.5}{0.45} = -3.33 \quad \text{and} \quad z_2 = \frac{69 - 68.5}{0.45} = 1.11.$$

Therefore,

$$\begin{aligned}\beta = P(-3.33 < Z < 1.11) &= P(Z < 1.11) - P(Z < -3.33) \\ &= 0.8665 - 0.0004 = 0.8661.\end{aligned}$$

The preceding examples illustrate the following important properties:

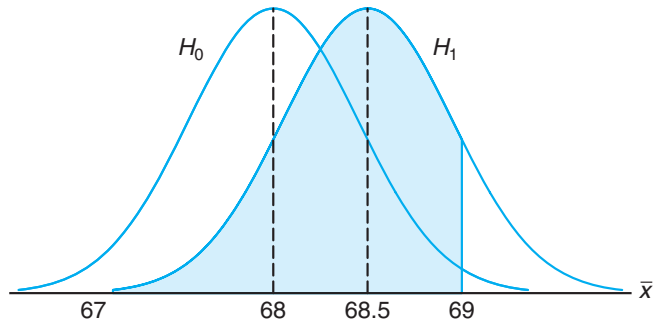


Figure 10.7: Type II error for testing $\mu = 68$ versus $\mu = 68.5$.

Important Properties of a Test of Hypothesis

1. The type I error and type II error are related. A decrease in the probability of one generally results in an increase in the probability of the other.
2. The size of the critical region, and therefore the probability of committing a type I error, can always be reduced by adjusting the critical value(s).
3. An increase in the sample size n will reduce α and β simultaneously.
4. If the null hypothesis is false, β is a maximum when the true value of a parameter approaches the hypothesized value. The greater the distance between the true value and the hypothesized value, the smaller β will be.

One very important concept that relates to error probabilities is the notion of the **power** of a test.

Definition 10.4:

The **power** of a test is the probability of rejecting H_0 given that a specific alternative is true.

The power of a test can be computed as $1 - \beta$. Often **different types of tests are compared by contrasting power properties**. Consider the previous illustration, in which we were testing H_0: $\mu = 68$ and H_1: $\mu \neq 68$. As before, suppose we are interested in assessing the sensitivity of the test. The test is governed by the rule that we do not reject H_0 if $67 \leq \bar{x} \leq 69$. We seek the capability of the test to properly reject H_0 when indeed $\mu = 68.5$. We have seen that the probability of a type II error is given by $\beta = 0.8661$. Thus, the **power** of the test is $1 - 0.8661 = 0.1339$. In a sense, the power is a more succinct measure of how sensitive the test is for detecting differences between a mean of 68 and a mean of 68.5. In this case, if μ is truly 68.5, the test as described will *properly reject H_0 only 13.39% of the time*. As a result, the test would not be a good one if it was important that the analyst have a reasonable chance of truly distinguishing between a mean of 68.0 (specified by H_0) and a mean of 68.5. From the foregoing, it is clear that to produce a desirable power (say, greater than 0.8), one must either increase α or increase the sample size.

So far in this chapter, much of the discussion of hypothesis testing has focused on foundations and definitions. In the sections that follow, we get more specific

and put hypotheses in categories as well as discuss tests of hypotheses on various parameters of interest. We begin by drawing the distinction between a one-sided and a two-sided hypothesis.

One- and Two-Tailed Tests

A test of any statistical hypothesis where the alternative is **one sided**, such as

$$H_0\colon\ \theta = \theta_0,$$
$$H_1\colon\ \theta > \theta_0$$

or perhaps

$$H_0\colon\ \theta = \theta_0,$$
$$H_1\colon\ \theta < \theta_0,$$

is called a **one-tailed test**. Earlier in this section, we referred to the **test statistic** for a hypothesis. Generally, the critical region for the alternative hypothesis $\theta > \theta_0$ lies in the right tail of the distribution of the test statistic, while the critical region for the alternative hypothesis $\theta < \theta_0$ lies entirely in the left tail. (In a sense, the inequality symbol points in the direction of the critical region.) A one-tailed test was used in the vaccine experiment to test the hypothesis $p = 1/4$ against the one-sided alternative $p > 1/4$ for the binomial distribution. The one-tailed critical region is usually obvious; the reader should visualize the behavior of the test statistic and notice the obvious *signal* that would produce evidence supporting the alternative hypothesis.

A test of any statistical hypothesis where the alternative is **two sided**, such as

$$H_0\colon\ \theta = \theta_0,$$
$$H_1\colon\ \theta \neq \theta_0,$$

is called a **two-tailed test**, since the critical region is split into two parts, often having equal probabilities, in each tail of the distribution of the test statistic. The alternative hypothesis $\theta \neq \theta_0$ states that either $\theta < \theta_0$ or $\theta > \theta_0$. A two-tailed test was used to test the null hypothesis that $\mu = 68$ kilograms against the two-sided alternative $\mu \neq 68$ kilograms in the example of the continuous population of student weights.

How Are the Null and Alternative Hypotheses Chosen?

The null hypothesis H_0 will often be stated using the *equality sign*. With this approach, it is clear how the probability of type I error is controlled. However, there are situations in which "do not reject H_0" implies that the parameter θ might be any value defined by the natural complement to the alternative hypothesis. For example, in the vaccine example, where the alternative hypothesis is $H_1\colon p > 1/4$, it is quite possible that nonrejection of H_0 cannot rule out a value of p less than $1/4$. Clearly though, in the case of one-tailed tests, the statement of the alternative is the most important consideration.

Whether one sets up a one-tailed or a two-tailed test will depend on the conclusion to be drawn if H_0 is rejected. The location of the critical region can be determined only after H_1 has been stated. For example, in testing a new drug, one sets up the hypothesis that it is no better than similar drugs now on the market and tests this against the alternative hypothesis that the new drug is superior. Such an alternative hypothesis will result in a one-tailed test with the critical region in the right tail. However, if we wish to compare a new teaching technique with the conventional classroom procedure, the alternative hypothesis should allow for the new approach to be either inferior or superior to the conventional procedure. Hence, the test is two-tailed with the critical region divided equally so as to fall in the extreme left and right tails of the distribution of our statistic.

Example 10.1: A manufacturer of a certain brand of rice cereal claims that the average saturated fat content does not exceed 1.5 grams per serving. State the null and alternative hypotheses to be used in testing this claim and determine where the critical region is located.

***Solution*:** The manufacturer's claim should be rejected only if μ is greater than 1.5 milligrams and should not be rejected if μ is less than or equal to 1.5 milligrams. We test

$$H_0: \ \mu = 1.5,$$
$$H_1: \ \mu > 1.5.$$

Nonrejection of H_0 does not rule out values less than 1.5 milligrams. Since we have a one-tailed test, the greater than symbol indicates that the critical region lies entirely in the right tail of the distribution of our test statistic $\bar{X}$. ■

Example 10.2: A real estate agent claims that 60% of all private residences being built today are 3-bedroom homes. To test this claim, a large sample of new residences is inspected; the proportion of these homes with 3 bedrooms is recorded and used as the test statistic. State the null and alternative hypotheses to be used in this test and determine the location of the critical region.

***Solution*:** If the test statistic were substantially higher or lower than $p = 0.6$, we would reject the agent's claim. Hence, we should make the hypothesis

$$H_0: \ p = 0.6,$$
$$H_1: \ p \neq 0.6.$$

The alternative hypothesis implies a two-tailed test with the critical region divided equally in both tails of the distribution of $\hat{P}$, our test statistic. ■

10.3 The Use of *P*-Values for Decision Making in Testing Hypotheses

In testing hypotheses in which the test statistic is discrete, the critical region may be chosen arbitrarily and its size determined. If α is too large, it can be reduced by making an adjustment in the critical value. It may be necessary to increase the

sample size to offset the decrease that occurs automatically in the power of the test.

Over a number of generations of statistical analysis, it had become customary to choose an α of 0.05 or 0.01 and select the critical region accordingly. Then, of course, strict rejection or nonrejection of H_0 would depend on that critical region. For example, if the test is two tailed and α is set at the 0.05 level of significance and the test statistic involves, say, the standard normal distribution, then a z-value is observed from the data and the critical region is

$$z > 1.96 \qquad \text{or} \qquad z < -1.96,$$

where the value 1.96 is found as $z_{0.025}$ in Table A.3. A value of z in the critical region prompts the statement "The value of the test statistic is significant," which we can then translate into the user's language. For example, if the hypothesis is given by

$$H_0\colon\ \mu = 10,$$
$$H_1\colon\ \mu \neq 10,$$

one might say, "The mean differs significantly from the value 10."

Preselection of a Significance Level

This preselection of a significance level α has its roots in the philosophy that the maximum risk of making a type I error should be controlled. However, this approach does not account for values of test statistics that are "close" to the critical region. Suppose, for example, in the illustration with $H_0\colon \mu = 10$ versus $H_1\colon \mu \neq 10$, a value of $z = 1.87$ is observed; strictly speaking, with $\alpha = 0.05$, the value is not significant. But the risk of committing a type I error if one rejects H_0 in this case could hardly be considered severe. In fact, in a two-tailed scenario, one can quantify this risk as

$$P = 2P(Z > 1.87 \text{ when } \mu = 10) = 2(0.0307) = 0.0614.$$

As a result, 0.0614 is the probability of obtaining a value of z as large as or larger (in magnitude) than 1.87 when in fact $\mu = 10$. Although this evidence against H_0 is not as strong as that which would result from rejection at an $\alpha = 0.05$ level, it is important information to the user. Indeed, continued use of $\alpha = 0.05$ or 0.01 is only a result of what standards have been passed down through the generations. The ***P*-value approach has been adopted extensively by users of applied statistics**. The approach is designed to give the user an alternative (in terms of a probability) to a mere "reject" or "do not reject" conclusion. The P-value computation also gives the user important information when the z-value falls well *into the ordinary critical region*. For example, if z is 2.73, it is informative for the user to observe that

$$P = 2(0.0032) = 0.0064,$$

and thus the z-value is significant at a level considerably less than 0.05. It is important to know that under the condition of H_0, a value of $z = 2.73$ is an extremely rare event. That is, a value at least that large in magnitude would only occur 64 times in 10,000 experiments.

A Graphical Demonstration of a *P*-Value

One very simple way of explaining a P-value graphically is to consider two distinct samples. Suppose that two materials are being considered for coating a particular type of metal in order to inhibit corrosion. Specimens are obtained, and one collection is coated with material 1 and one collection coated with material 2. The sample sizes are $n_1 = n_2 = 10$, and corrosion is measured in percent of surface area affected. The hypothesis is that the samples came from common distributions with mean $\mu = 10$. Let us assume that the population variance is 1.0. Then we are testing

$$H_0\colon\ \mu_1 = \mu_2 = 10.$$

Let Figure 10.8 represent a point plot of the data; the data are placed on the distribution stated by the null hypothesis. Let us assume that the "×" data refer to material 1 and the "○" data refer to material 2. Now it seems clear that the data do refute the null hypothesis. But how can this be summarized in one number? **The *P*-value can be viewed as simply the probability of obtaining these data given that both samples come from the same distribution**. Clearly, this probability is quite small, say 0.00000001! Thus, the small P-value clearly refutes H_0, and the conclusion is that the population means are significantly different.

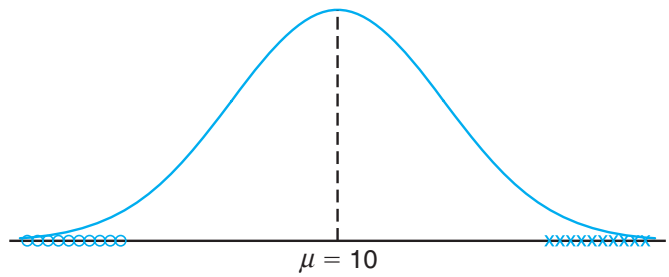


Figure 10.8: Data that are likely generated from populations having two different means.

Use of the P-value approach as an aid in decision-making is quite natural, and nearly all computer packages that provide hypothesis-testing computation print out P-values along with values of the appropriate test statistic. The following is a formal definition of a P-value.

Definition 10.5: A ***P*-value** is the lowest level (of significance) at which the observed value of the test statistic is significant.

How Does the Use of *P*-Values Differ from Classic Hypothesis Testing?

It is tempting at this point to summarize the procedures associated with testing, say, $H_0\colon \theta = \theta_0$. However, the student who is a novice in this area should understand that there are differences in approach and philosophy between the classic

fixed α approach that is climaxed with either a "reject H_0" or a "do not reject H_0" conclusion and the P-value approach. In the latter, no fixed α is determined and conclusions are drawn on the basis of the size of the P-value in harmony with the subjective judgment of the engineer or scientist. While modern computer software will output P-values, nevertheless it is important that readers understand both approaches in order to appreciate the totality of the concepts. Thus, we offer a brief list of procedural steps for both the classical and the P-value approach.

Approach to Hypothesis Testing with Fixed Probability of Type I Error

1. State the null and alternative hypotheses.
2. Choose a fixed significance level α.
3. Choose an appropriate test statistic and establish the critical region based on α.
4. Reject H_0 if the computed test statistic is in the critical region. Otherwise, do not reject.
5. Draw scientific or engineering conclusions.

Significance Testing (P-Value Approach)

1. State null and alternative hypotheses.
2. Choose an appropriate test statistic.
3. Compute the P-value based on the computed value of the test statistic.
4. Use judgment based on the P-value and knowledge of the scientific system.

In later sections of this chapter and chapters that follow, many examples and exercises emphasize the P-value approach to drawing scientific conclusions.

Exercises

10.1 Suppose that an allergist wishes to test the hypothesis that at least 30% of the public is allergic to some cheese products. Explain how the allergist could commit

(a) a type I error;

(b) a type II error.

10.2 A sociologist is concerned about the effectiveness of a training course designed to get more drivers to use seat belts in automobiles.

(a) What hypothesis is she testing if she commits a type I error by erroneously concluding that the training course is ineffective?

(b) What hypothesis is she testing if she commits a type II error by erroneously concluding that the training course is effective?

10.3 A large manufacturing firm is being charged with discrimination in its hiring practices.

(a) What hypothesis is being tested if a jury commits a type I error by finding the firm guilty?

(b) What hypothesis is being tested if a jury commits a type II error by finding the firm guilty?

10.4 A fabric manufacturer believes that the proportion of orders for raw material arriving late is $p = 0.6$. If a random sample of 10 orders shows that 3 or fewer arrived late, the hypothesis that $p = 0.6$ should be rejected in favor of the alternative $p < 0.6$. Use the binomial distribution.

(a) Find the probability of committing a type I error if the true proportion is $p = 0.6$.

(b) Find the probability of committing a type II error for the alternatives $p = 0.3$, $p = 0.4$, and $p = 0.5$.

10.5 Repeat Exercise 10.4 but assume that 50 orders are selected and the critical region is defined to be $x \leq 24$, where x is the number of orders in the sample that arrived late. Use the normal approximation.

10.6 The proportion of adults living in a small town who are college graduates is estimated to be $p = 0.6$. To test this hypothesis, a random sample of 15 adults is selected. If the number of college graduates in the sample is anywhere from 6 to 12, we shall not reject the null hypothesis that $p = 0.6$; otherwise, we shall conclude that $p \neq 0.6$.

(a) Evaluate α assuming that $p = 0.6$. Use the binomial distribution.

(b) Evaluate β for the alternatives $p = 0.5$ and $p = 0.7$.

(c) Is this a good test procedure?

10.7 Repeat Exercise 10.6 but assume that 200 adults are selected and the fail-to-reject region is defined to be $110 \leq x \leq 130$, where x is the number of college graduates in our sample. Use the normal approximation.

10.8 In *Relief from Arthritis* published by Thorsons Publishers, Ltd., John E. Croft claims that over 40% of those who suffer from osteoarthritis receive measurable relief from an ingredient produced by a particular species of mussel found off the coast of New Zealand. To test this claim, the mussel extract is to be given to a group of 7 osteoarthritic patients. If 3 or more of the patients receive relief, we shall not reject the null hypothesis that $p = 0.4$; otherwise, we conclude that $p < 0.4$.

(a) Evaluate α, assuming that $p = 0.4$.

(b) Evaluate β for the alternative $p = 0.3$.

10.9 A dry cleaning establishment claims that a new spot remover will remove more than 70% of the spots to which it is applied. To check this claim, the spot remover will be used on 12 spots chosen at random. If fewer than 11 of the spots are removed, we shall not reject the null hypothesis that $p = 0.7$; otherwise, we conclude that $p > 0.7$.

(a) Evaluate α, assuming that $p = 0.7$.

(b) Evaluate β for the alternative $p = 0.9$.

10.10 Repeat Exercise 10.9 but assume that 100 spots are treated and the critical region is defined to be $x > 82$, where x is the number of spots removed.

10.11 Repeat Exercise 10.8 but assume that 70 patients are given the mussel extract and the critical region is defined to be $x < 24$, where x is the number of osteoarthritic patients who receive relief.

10.12 A random sample of 400 voters in a certain city are asked if they favor an additional 4% gasoline sales tax to provide badly needed revenues for street repairs. If more than 220 but fewer than 260 favor the sales tax, we shall conclude that 60% of the voters are for it.

(a) Find the probability of committing a type I error if 60% of the voters favor the increased tax.

(b) What is the probability of committing a type II error using this test procedure if actually only 48% of the voters are in favor of the additional gasoline tax?

10.13 Suppose, in Exercise 10.12, we conclude that 60% of the voters favor the gasoline sales tax if more than 214 but fewer than 266 voters in our sample favor it. Show that this new critical region results in a smaller value for α at the expense of increasing β.

10.14 A manufacturer has developed a new fishing line, which the company claims has a mean breaking strength of 15 kilograms with a standard deviation of 0.5 kilogram. To test the hypothesis that $\mu = 15$ kilograms against the alternative that $\mu < 15$ kilograms, a random sample of 50 lines will be tested. The critical region is defined to be $\bar{x} < 14.9$.

(a) Find the probability of committing a type I error when H_0 is true.

(b) Evaluate β for the alternatives $\mu = 14.8$ and $\mu = 14.9$ kilograms.

10.15 A soft-drink machine at a steak house is regulated so that the amount of drink dispensed is approximately normally distributed with a mean of 200 milliliters and a standard deviation of 15 milliliters. The machine is checked periodically by taking a sample of 9 drinks and computing the average content. If $\bar{x}$ falls in the interval $191 < \bar{x} < 209$, the machine is thought to be operating satisfactorily; otherwise, we conclude that $\mu \neq 200$ milliliters.

(a) Find the probability of committing a type I error when $\mu = 200$ milliliters.

(b) Find the probability of committing a type II error when $\mu = 215$ milliliters.

10.16 Repeat Exercise 10.15 for samples of size $n = 25$. Use the same critical region.

10.17 A new curing process developed for a certain type of cement results in a mean compressive strength of 5000 kilograms per square centimeter with a standard deviation of 120 kilograms. To test the hypothesis that $\mu = 5000$ against the alternative that $\mu < 5000$, a random sample of 50 pieces of cement is tested. The critical region is defined to be $\bar{x} < 4970$.

(a) Find the probability of committing a type I error when H_0 is true.

(b) Evaluate β for the alternatives $\mu = 4970$ and $\mu = 4960$.

10.18 If we plot the probabilities of failing to reject H_0 corresponding to various alternatives for μ (including the value specified by H_0) and connect all the points by a smooth curve, we obtain the **operating characteristic curve** of the test criterion, or simply the OC curve. Note that the probability of failing to reject H_0 when it is true is simply $1 - \alpha$. Operating characteristic curves are widely used in industrial applications to provide a visual display of the merits of the test criterion. With reference to Exercise 10.15, find the probabilities of failing to reject H_0 for the following 9 values of μ and plot the OC curve: 184, 188, 192, 196, 200, 204, 208, 212, and 216.

⊙ **Guided Practice 4.17** Use the data in Guided Practice 4.16 to create a 90% confidence interval for the average days active per week of all YRBSS students.[13]

4.2.5 Interpreting confidence intervals

A careful eye might have observed the somewhat awkward language used to describe confidence intervals. Correct interpretation:

> We are XX% confident that the population parameter is between...

Incorrect language might try to describe the confidence interval as capturing the population parameter with a certain probability. This is a common error: while it might be useful to think of it as a probability, the confidence level only quantifies how plausible it is that the parameter is in the interval.

Another important consideration of confidence intervals is that they *only try to capture the population parameter*. A confidence interval says nothing about the confidence of capturing individual observations, a proportion of the observations, or about capturing point estimates. Confidence intervals only attempt to capture population parameters.

4.3 Hypothesis testing

Are students lifting weights or performing other strength training exercises more or less often than they have in the past? We'll compare data from students from the 2011 YRBSS survey to our sample of 100 students from the 2013 YRBSS survey.

We'll also consider sleep behavior. A recent study found that college students average about 7 hours of sleep per night.[14] However, researchers at a rural college are interested in showing that their students sleep longer than seven hours on average. We investigate this topic in Section 4.3.4.

4.3.1 Hypothesis testing framework

Students from the 2011 YRBSS lifted weights (or performed other strength training exercises) 3.09 days per week on average. We want to determine if the `yrbss_samp` data set provides strong evidence that YRBSS students selected in 2013 are lifting more or less than the 2011 YRBSS students, versus the other possibility that there has been no change.[15] We simplify these three options into two competing **hypotheses**:

H_0: The average days per week that YRBSS students lifted weights was the same for 2011 and 2013.

H_A: The average days per week that YRBSS students lifted weights was *different* for 2013 than in 2011.

We call H_0 the null hypothesis and H_A the alternative hypothesis.

H_0
null hypothesis

H_A
alternative hypothesis

[13] We first find $z^\star$ such that 90% of the distribution falls between -$z^\star$ and $z^\star$ in the standard normal model, $N(\mu = 0, \sigma = 1)$. We can look up -$z^\star$ in the normal probability table by looking for a lower tail of 5% (the other 5% is in the upper tail): $z^\star = 1.65$. The 90% confidence interval can then be computed as $\bar{x}_{active} \pm 1.65 \times SE_{\bar{x}} \rightarrow (3.32, 4.18)$. (We had already verified conditions for normality and the standard error.) That is, we are 90% confident the average days active per week is between 3.32 and 4.18 days.

[14] *Poll shows college students get least amount of sleep.* theloquitur.com/?p=1161

[15] While we could answer this question by examining the entire YRBSS data set from 2013 (`yrbss`), we only consider the sample data (`yrbss_samp`), which is more realistic since we rarely have access to population data.

Null and alternative hypotheses

The **null hypothesis** (H_0) often represents either a skeptical perspective or a claim to be tested. The **alternative hypothesis** (H_A) represents an alternative claim under consideration and is often represented by a range of possible parameter values.

The null hypothesis often represents a skeptical position or a perspective of no difference. The alternative hypothesis often represents a new perspective, such as the possibility that there has been a change.

TIP: Hypothesis testing framework

The skeptic will not reject the null hypothesis (H_0), unless the evidence in favor of the alternative hypothesis (H_A) is so strong that she rejects H_0 in favor of H_A.

The hypothesis testing framework is a very general tool, and we often use it without a second thought. If a person makes a somewhat unbelievable claim, we are initially skeptical. However, if there is sufficient evidence that supports the claim, we set aside our skepticism and reject the null hypothesis in favor of the alternative. The hallmarks of hypothesis testing are also found in the US court system.

⊙ **Guided Practice 4.18** A US court considers two possible claims about a defendant: she is either innocent or guilty. If we set these claims up in a hypothesis framework, which would be the null hypothesis and which the alternative?[16]

Jurors examine the evidence to see whether it convincingly shows a defendant is guilty. Even if the jurors leave unconvinced of guilt beyond a reasonable doubt, this does not mean they believe the defendant is innocent. This is also the case with hypothesis testing: *even if we fail to reject the null hypothesis, we typically do not accept the null hypothesis as true.* Failing to find strong evidence for the alternative hypothesis is not equivalent to accepting the null hypothesis.

In the example with the YRBSS, the null hypothesis represents no difference in the average days per week of weight lifting in 2011 and 2013. The alternative hypothesis represents something new or more interesting: there was a difference, either an increase or a decrease. These hypotheses can be described in mathematical notation using μ_{13} as the average days of weight lifting for 2013:

H_0: $\mu_{13} = 3.09$

H_A: $\mu_{13} \neq 3.09$

where 3.09 is the average number of days per week that students from the 2011 YRBSS lifted weights. Using the mathematical notation, the hypotheses can more easily be evaluated using statistical tools. We call 3.09 the **null value** since it represents the value of the parameter if the null hypothesis is true.

[16] The jury considers whether the evidence is so convincing (strong) that there is no reasonable doubt regarding the person's guilt; in such a case, the jury rejects innocence (the null hypothesis) and concludes the defendant is guilty (alternative hypothesis).

4.3.2 Testing hypotheses using confidence intervals

We will use the `yrbss_samp` data set to evaluate the hypothesis test, and we start by comparing the 2013 point estimate of the number of days per week that students lifted weights: $\bar{x}_{13} = 2.78$ days. This estimate suggests that students from the 2013 YRBSS were lifting weights less than students in the 2011 YRBSS. However, to evaluate whether this provides strong evidence that there has been a change, we must consider the uncertainty associated with $\bar{x}_{13}$.

We learned in Section 4.1 that there is fluctuation from one sample to another, and it is unlikely that the sample mean will be exactly equal to the parameter; we should not expect $\bar{x}_{13}$ to exactly equal μ_{13}. Given that $\bar{x}_{13} = 2.78$, it might still be possible that the average of all students from the 2013 YRBSS survey is the same as the average from the 2011 YRBSS survey. The difference between $\bar{x}_{13}$ and 3.09 could be due to *sampling variation*, i.e. the variability associated with the point estimate when we take a random sample.

In Section 4.2, confidence intervals were introduced as a way to find a range of plausible values for the population mean.

Example 4.19 In the sample of 100 students from the 2013 YRBSS survey, the average number of days per week that students lifted weights was 2.78 days with a standard deviation of 2.56 days (coincidentally the same as days active). Compute a 95% confidence interval for the average for all students from the 2013 YRBSS survey. You can assume the conditions for the normal model are met.

The general formula for the confidence interval based on the normal distribution is

$$\bar{x} \pm z^\star SE_{\bar{x}}$$

We are given $\bar{x}_{13} = 2.78$, we use $z^\star = 1.96$ for a 95% confidence level, and we can compute the standard error using the standard deviation divided by the square root of the sample size:

$$SE_{\bar{x}} = \frac{s_{13}}{\sqrt{n}} = \frac{2.56}{\sqrt{100}} = 0.256$$

Entering the sample mean, $z^\star$, and the standard error into the confidence interval formula results in (2.27, 3.29). We are 95% confident that the average number of days per week that all students from the 2013 YRBSS lifted weights was between 2.27 and 3.29 days.

Because the average of all students from the 2011 YRBSS survey is 3.09, which falls within the range of plausible values from the confidence interval, we cannot say the null hypothesis is implausible. That is, we fail to reject the null hypothesis, H_0.

TIP: Double negatives can sometimes be used in statistics

In many statistical explanations, we use double negatives. For instance, we might say that the null hypothesis is *not implausible* or we *failed to reject* the null hypothesis. Double negatives are used to communicate that while we are not rejecting a position, we are also not saying it is correct.

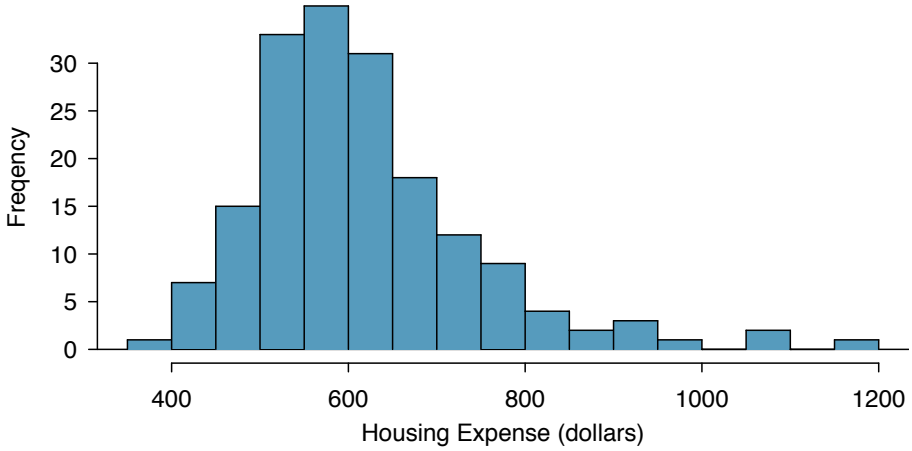


Figure 4.11: Sample distribution of student housing expense. These data are strongly skewed, which we can see by the long right tail with a few notable outliers.

⊙ **Guided Practice 4.20** Colleges frequently provide estimates of student expenses such as housing. A consultant hired by a community college claimed that the average student housing expense was $650 per month. What are the null and alternative hypotheses to test whether this claim is accurate?[17]

⊙ **Guided Practice 4.21** The community college decides to collect data to evaluate the $650 per month claim. They take a random sample of 175 students at their school and obtain the data represented in Figure 4.11. Can we apply the normal model to the sample mean?[18]

Evaluating the skew condition is challenging

Don't despair if checking the skew condition is difficult or confusing. You aren't alone – nearly all students get frustrated when checking skew. Properly assessing skew takes practice, and you won't be a pro, even at the end of this book.

But this doesn't mean you should give up. Checking skew and the other conditions is extremely important for a responsible data analysis. However, rest assured that evaluating skew isn't something you need to be a master of by the end of the book, though by that time you should be able to properly assess clear cut cases.

[17] H_0: The average cost is $650 per month, $\mu = \$650$.
H_A: The average cost is different than $650 per month, $\mu \neq \$650$.

[18] Applying the normal model requires that certain conditions are met. Because the data are a simple random sample and the sample (presumably) represents no more than 10% of all students at the college, the observations are independent. The sample size is also sufficiently large ($n = 175$) and the data exhibit strong skew. While the data are strongly skewed, the sample is sufficiently large that this is acceptable, and the normal model may be applied to the sample mean.

Example 4.22 The sample mean for student housing is \$616.91 and the sample standard deviation is \$128.65. Construct a 95% confidence interval for the population mean and evaluate the hypotheses of Guided Practice 4.20.

The standard error associated with the mean may be estimated using the sample standard deviation divided by the square root of the sample size. Recall that $n = 175$ students were sampled.

$$SE = \frac{s}{\sqrt{n}} = \frac{128.65}{\sqrt{175}} = 9.73$$

You showed in Guided Practice 4.21 that the normal model may be applied to the sample mean. This ensures a 95% confidence interval may be accurately constructed:

$$\bar{x} \pm z^\star SE \quad \rightarrow \quad 616.91 \pm 1.96 \times 9.73 \quad \rightarrow \quad (597.84, 635.98)$$

Because the null value \$650 is not in the confidence interval, a true mean of \$650 is implausible and we reject the null hypothesis. The data provide statistically significant evidence that the actual average housing expense is less than \$650 per month.

4.3.3 Decision errors

Hypothesis tests are not flawless, since we can make a wrong decision in statistical hypothesis tests based on the data. For example, in the court system innocent people are sometimes wrongly convicted and the guilty sometimes walk free. However, the difference is that in statistical hypothesis tests, we have the tools necessary to quantify how often we make such errors.

There are two competing hypotheses: the null and the alternative. In a hypothesis test, we make a statement about which one might be true, but we might choose incorrectly. There are four possible scenarios, which are summarized in Table 4.12.

		Test conclusion	
		do not reject H_0	reject H_0 in favor of H_A
Truth	H_0 true	okay	Type 1 Error
	H_A true	Type 2 Error	okay

Table 4.12: Four different scenarios for hypothesis tests.

A **Type 1 Error** is rejecting the null hypothesis when H_0 is actually true. A **Type 2 Error** is failing to reject the null hypothesis when the alternative is actually true.

Guided Practice 4.23 In a US court, the defendant is either innocent (H_0) or guilty (H_A). What does a Type 1 Error represent in this context? What does a Type 2 Error represent? Table 4.12 may be useful.[19]

Guided Practice 4.24 How could we reduce the Type 1 Error rate in US courts? What influence would this have on the Type 2 Error rate?[20]

[19] If the court makes a Type 1 Error, this means the defendant is innocent (H_0 true) but wrongly convicted. A Type 2 Error means the court failed to reject H_0 (i.e. failed to convict the person) when she was in fact guilty (H_A true).

[20] To lower the Type 1 Error rate, we might raise our standard for conviction from "beyond a reasonable doubt" to "beyond a conceivable doubt" so fewer people would be wrongly convicted. However, this would also make it more difficult to convict the people who are actually guilty, so we would make more Type 2 Errors.

⊙ **Guided Practice 4.25** How could we reduce the Type 2 Error rate in US courts? What influence would this have on the Type 1 Error rate?[21]

Exercises 4.23-4.25 provide an important lesson: if we reduce how often we make one type of error, we generally make more of the other type.

Hypothesis testing is built around rejecting or failing to reject the null hypothesis. That is, we do not reject H_0 unless we have strong evidence. But what precisely does *strong evidence* mean? As a general rule of thumb, for those cases where the null hypothesis is actually true, we do not want to incorrectly reject H_0 more than 5% of the time. This corresponds to a **significance level** of 0.05. We often write the significance level using α (the Greek letter *alpha*): $\alpha = 0.05$. We discuss the appropriateness of different significance levels in Section 4.3.6.

α
significance level of a hypothesis test

If we use a 95% confidence interval to evaluate a hypothesis test where the null hypothesis is true, we will make an error whenever the point estimate is at least 1.96 standard errors away from the population parameter. This happens about 5% of the time (2.5% in each tail). Similarly, using a 99% confidence interval to evaluate a hypothesis is equivalent to a significance level of $\alpha = 0.01$.

A confidence interval is, in one sense, simplistic in the world of hypothesis tests. Consider the following two scenarios:

- The null value (the parameter value under the null hypothesis) is in the 95% confidence interval but just barely, so we would not reject H_0. However, we might like to somehow say, quantitatively, that it was a close decision.
- The null value is very far outside of the interval, so we reject H_0. However, we want to communicate that, not only did we reject the null hypothesis, but it wasn't even close. Such a case is depicted in Figure 4.13.

In Section 4.3.4, we introduce a tool called the *p-value* that will be helpful in these cases. The p-value method also extends to hypothesis tests where confidence intervals cannot be easily constructed or applied.

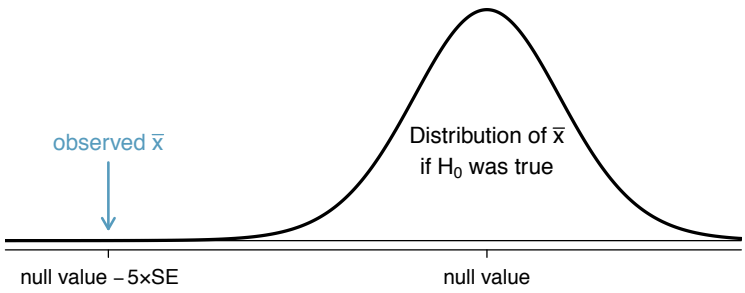


Figure 4.13: It would be helpful to quantify the strength of the evidence against the null hypothesis. In this case, the evidence is extremely strong.

[21] To lower the Type 2 Error rate, we want to convict more guilty people. We could lower the standards for conviction from "beyond a reasonable doubt" to "beyond a little doubt". Lowering the bar for guilt will also result in more wrongful convictions, raising the Type 1 Error rate.

4.3.4 Formal testing using p-values

The p-value is a way of quantifying the strength of the evidence against the null hypothesis and in favor of the alternative. Formally the *p-value* is a conditional probability.

p-value

The **p-value** is the probability of observing data at least as favorable to the alternative hypothesis as our current data set, if the null hypothesis is true. We typically use a summary statistic of the data, in this chapter the sample mean, to help compute the p-value and evaluate the hypotheses.

⊙ **Guided Practice 4.26** A poll by the National Sleep Foundation found that college students average about 7 hours of sleep per night. Researchers at a rural school are interested in showing that students at their school sleep longer than seven hours on average, and they would like to demonstrate this using a sample of students. What would be an appropriate skeptical position for this research?[22]

We can set up the null hypothesis for this test as a skeptical perspective: the students at this school average 7 hours of sleep per night. The alternative hypothesis takes a new form reflecting the interests of the research: the students average more than 7 hours of sleep. We can write these hypotheses as

H_0: $\mu = 7$.

H_A: $\mu > 7$.

Using $\mu > 7$ as the alternative is an example of a **one-sided** hypothesis test. In this investigation, there is no apparent interest in learning whether the mean is less than 7 hours.[23] Earlier we encountered a **two-sided** hypothesis where we looked for any clear difference, greater than or less than the null value.

Always use a two-sided test unless it was made clear prior to data collection that the test should be one-sided. Switching a two-sided test to a one-sided test after observing the data is dangerous because it can inflate the Type 1 Error rate.

TIP: One-sided and two-sided tests

When you are interested in checking for an increase or a decrease, but not both, use a one-sided test. When you are interested in any difference from the null value – an increase or decrease – then the test should be two-sided.

TIP: Always write the null hypothesis as an equality

We will find it most useful if we always list the null hypothesis as an equality (e.g. $\mu = 7$) while the alternative always uses an inequality (e.g. $\mu \neq 7$, $\mu > 7$, or $\mu < 7$).

[22] A skeptic would have no reason to believe that sleep patterns at this school are different than the sleep patterns at another school.

[23] This is entirely based on the interests of the researchers. Had they been only interested in the opposite case – showing that their students were actually averaging fewer than seven hours of sleep but not interested in showing more than 7 hours – then our setup would have set the alternative as $\mu < 7$.

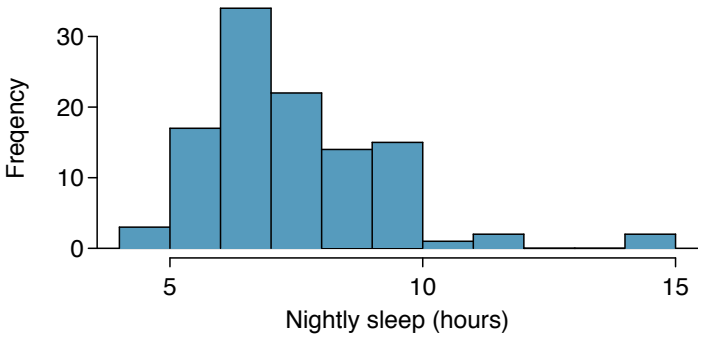


Figure 4.14: Distribution of a night of sleep for 110 college students. These data are strongly skewed.

The researchers at the rural school conducted a simple random sample of $n = 110$ students on campus. They found that these students averaged 7.42 hours of sleep and the standard deviation of the amount of sleep for the students was 1.75 hours. A histogram of the sample is shown in Figure 4.14.

Before we can use a normal model for the sample mean or compute the standard error of the sample mean, we must verify conditions. (1) Because this is a simple random sample from less than 10% of the student body, the observations are independent. (2) The sample size in the sleep study is sufficiently large since it is greater than 30. (3) The data show strong skew in Figure 4.14 and the presence of a couple of outliers. This skew and the outliers are acceptable for a sample size of $n = 110$. With these conditions verified, the normal model can be safely applied to $\bar{x}$ and we can reasonably calculate the standard error.

⊙ **Guided Practice 4.27** In the sleep study, the sample standard deviation was 1.75 hours and the sample size is 110. Calculate the standard error of $\bar{x}$.[24]

The hypothesis test for the sleep study will be evaluated using a significance level of $\alpha = 0.05$. We want to consider the data under the scenario that the null hypothesis is true. In this case, the sample mean is from a distribution that is nearly normal and has mean 7 and standard deviation of about $SE_{\bar{x}} = 0.17$. Such a distribution is shown in Figure 4.15.

The shaded tail in Figure 4.15 represents the chance of observing such a large mean, conditional on the null hypothesis being true. That is, the shaded tail represents the p-value. We shade all means larger than our sample mean, $\bar{x} = 7.42$, because they are more favorable to the alternative hypothesis than the observed mean.

We compute the p-value by finding the tail area of this normal distribution, which we learned to do in Section 3.1. First compute the Z-score of the sample mean, $\bar{x} = 7.42$:

$$Z = \frac{\bar{x} - \text{null value}}{SE_{\bar{x}}} = \frac{7.42 - 7}{0.17} = 2.47$$

Using the normal probability table, the lower unshaded area is found to be 0.993. Thus the shaded area is $1 - 0.993 = 0.007$. *If the null hypothesis is true, the probability of observing a sample mean at least as large as 7.42 hours for a sample of 110 students is only 0.007.* That is, if the null hypothesis is true, we would not often see such a large mean.

[24] The standard error can be estimated from the sample standard deviation and the sample size: $SE_{\bar{x}} = \frac{s_x}{\sqrt{n}} = \frac{1.75}{\sqrt{110}} = 0.17$.

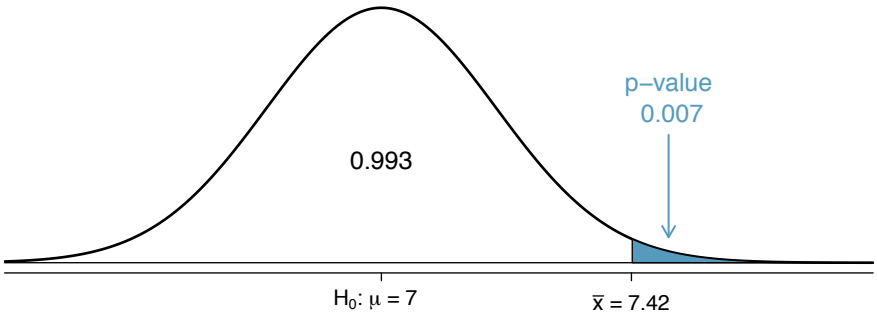


Figure 4.15: If the null hypothesis is true, then the sample mean $\bar{x}$ came from this nearly normal distribution. The right tail describes the probability of observing such a large sample mean if the null hypothesis is true.

We evaluate the hypotheses by comparing the p-value to the significance level. Because the p-value is less than the significance level (p-value $= 0.007 < 0.05 = \alpha$), we reject the null hypothesis. What we observed is so unusual with respect to the null hypothesis that it casts serious doubt on H_0 and provides strong evidence favoring H_A.

p-value as a tool in hypothesis testing

The smaller the p-value, the stronger the data favor H_A over H_0. A small p-value (usually < 0.05) corresponds to sufficient evidence to reject H_0 in favor of H_A.

TIP: It is useful to first draw a picture to find the p-value

It is useful to draw a picture of the distribution of $\bar{x}$ as though H_0 was true (i.e. μ equals the null value), and shade the region (or regions) of sample means that are at least as favorable to the alternative hypothesis. These shaded regions represent the p-value.

The ideas below review the process of evaluating hypothesis tests with p-values:

- The null hypothesis represents a skeptic's position or a position of no difference. We reject this position only if the evidence strongly favors H_A.
- A small p-value means that if the null hypothesis is true, there is a low probability of seeing a point estimate at least as extreme as the one we saw. We interpret this as strong evidence in favor of the alternative.
- We reject the null hypothesis if the p-value is smaller than the significance level, α, which is usually 0.05. Otherwise, we fail to reject H_0.
- We should always state the conclusion of the hypothesis test in plain language so non-statisticians can also understand the results.

The p-value is constructed in such a way that we can directly compare it to the significance level (α) to determine whether or not to reject H_0. This method ensures that the Type 1 Error rate does not exceed the significance level standard.

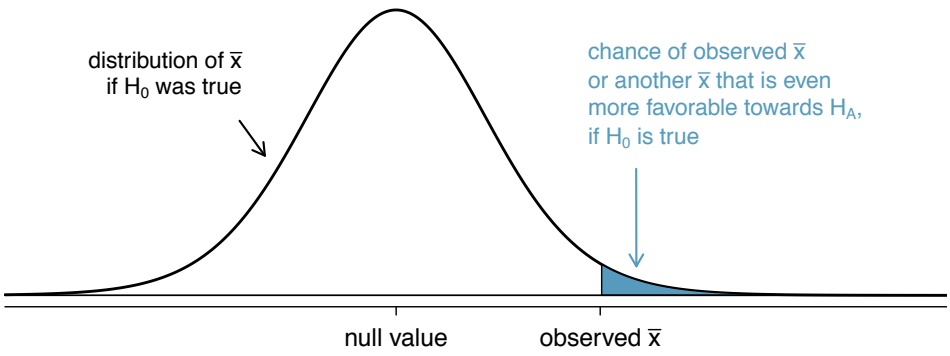


Figure 4.16: To identify the p-value, the distribution of the sample mean is considered as if the null hypothesis was true. Then the p-value is defined and computed as the probability of the observed $\bar{x}$ or an $\bar{x}$ even more favorable to H_A under this distribution.

⊙ **Guided Practice 4.28** If the null hypothesis is true, how often should the p-value be less than 0.05?[25]

⊙ **Guided Practice 4.29** Suppose we had used a significance level of 0.01 in the sleep study. Would the evidence have been strong enough to reject the null hypothesis? (The p-value was 0.007.) What if the significance level was $\alpha = 0.001$? [26]

⊙ **Guided Practice 4.30** Ebay might be interested in showing that buyers on its site tend to pay less than they would for the corresponding new item on Amazon. We'll research this topic for one particular product: a video game called *Mario Kart* for the Nintendo Wii. During early October 2009, Amazon sold this game for $46.99. Set up an appropriate (one-sided!) hypothesis test to check the claim that Ebay buyers pay less during auctions at this same time.[27]

⊙ **Guided Practice 4.31** During early October 2009, 52 Ebay auctions were recorded for *Mario Kart*.[28] The total prices for the auctions are presented using a histogram in Figure 4.17, and we may like to apply the normal model to the sample mean. Check the three conditions required for applying the normal model: (1) independence, (2) at least 30 observations, and (3) the data are not strongly skewed.[29]

[25] About 5% of the time. If the null hypothesis is true, then the data only has a 5% chance of being in the 5% of data most favorable to H_A.

[26] We reject the null hypothesis whenever $p\text{-}value < \alpha$. Thus, we would still reject the null hypothesis if $\alpha = 0.01$ but not if the significance level had been $\alpha = 0.001$.

[27] The skeptic would say the average is the same on Ebay, and we are interested in showing the average price is lower.

H_0: The average auction price on Ebay is equal to (or more than) the price on Amazon. We write only the equality in the statistical notation: $\mu_{ebay} = 46.99$.

H_A: The average price on Ebay is less than the price on Amazon, $\mu_{ebay} < 46.99$.

[28] These data were collected by OpenIntro staff.

[29] (1) The independence condition is unclear. *We will make the assumption that the observations are independent, which we should report with any final results.* (2) The sample size is sufficiently large: $n = 52 \geq 30$. (3) The data distribution is not strongly skewed; it is approximately symmetric.

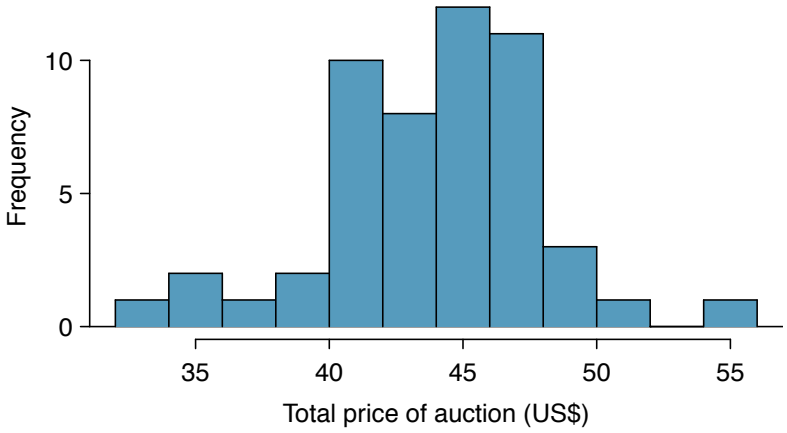
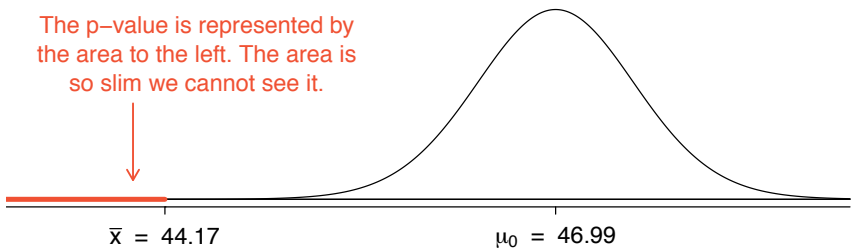


Figure 4.17: A histogram of the total auction prices for 52 Ebay auctions.

Example 4.32 The average sale price of the 52 Ebay auctions for *Wii Mario Kart* was \$44.17 with a standard deviation of \$4.15. Does this provide sufficient evidence to reject the null hypothesis in Guided Practice 4.30? Use a significance level of $\alpha = 0.01$.

The hypotheses were set up and the conditions were checked in Exercises 4.30 and 4.31. The next step is to find the standard error of the sample mean and produce a sketch to help find the p-value.

$$SE_{\bar{x}} = s/\sqrt{n} = 4.15/\sqrt{52} = 0.5755$$



Because the alternative hypothesis says we are looking for a smaller mean, we shade the lower tail. We find this shaded area by using the Z-score and normal probability table: $Z = \frac{44.17-46.99}{0.5755} = -4.90$, which has area less than 0.0002. The area is so small we cannot really see it on the picture. This lower tail area corresponds to the p-value.

Because the p-value is so small – specifically, smaller than $\alpha = 0.01$ – this provides sufficiently strong evidence to reject the null hypothesis in favor of the alternative. The data provide statistically significant evidence that the average price on Ebay is lower than Amazon's asking price.

What's so special about 0.05?

It's common to use a threshold of 0.05 to determine whether a result is statistically significant, but why is the most common value 0.05? Maybe the standard significance level should be bigger, or maybe it should be smaller. If you're a little puzzled, that probably means you're reading with a critical eye – good job! We've made a 5-minute task to help clarify *why 0.05*:

www.openintro.org/why05

Sometimes it's also a good idea to deviate from the standard. We'll discuss when to choose a threshold different than 0.05 in Section 4.3.6.

4.3.5 Two-sided hypothesis testing with p-values

We now consider how to compute a p-value for a two-sided test. In one-sided tests, we shade the single tail in the direction of the alternative hypothesis. For example, when the alternative had the form $\mu > 7$, then the p-value was represented by the upper tail (Figure 4.16). When the alternative was $\mu < 46.99$, the p-value was the lower tail (Guided Practice 4.30). In a two-sided test, *we shade two tails* since evidence in either direction is favorable to H_A.

Guided Practice 4.33 Earlier we talked about a research group investigating whether the students at their school slept longer than 7 hours each night. Let's consider a second group of researchers who want to evaluate whether the students at their college differ from the norm of 7 hours. Write the null and alternative hypotheses for this investigation.[30]

Example 4.34 The second college randomly samples 122 students and finds a mean of $\bar{x} = 6.83$ hours and a standard deviation of $s = 1.8$ hours. Does this provide strong evidence against H_0 in Guided Practice 4.33? Use a significance level of $\alpha = 0.05$.

First, we must verify assumptions. (1) A simple random sample of less than 10% of the student body means the observations are independent. (2) The sample size is 122, which is greater than 30. (3) Based on the earlier distribution and what we already know about college student sleep habits, the sample size will be acceptable.

Next we can compute the standard error ($SE_{\bar{x}} = \frac{s}{\sqrt{n}} = 0.16$) of the estimate and create a picture to represent the p-value, shown in Figure 4.18. Both tails are shaded. An estimate of 7.17 or more provides at least as strong of evidence against the null hypothesis and in favor of the alternative as the observed estimate, $\bar{x} = 6.83$.

We can calculate the tail areas by first finding the lower tail corresponding to $\bar{x}$:

$$Z = \frac{6.83 - 7.00}{0.16} = -1.06 \quad \overset{table}{\rightarrow} \quad \text{left tail} = 0.1446$$

Because the normal model is symmetric, the right tail will have the same area as the left tail. The p-value is found as the sum of the two shaded tails:

$$\text{p-value} = \text{left tail} + \text{right tail} = 2 \times (\text{left tail}) = 0.2892$$

[30] Because the researchers are interested in any difference, they should use a two-sided setup: $H_0 : \mu = 7$, $H_A : \mu \neq 7$.

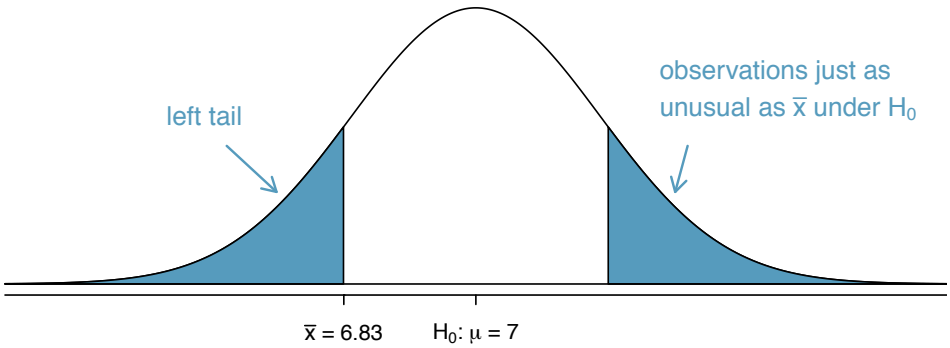


Figure 4.18: H_A is two-sided, so *both* tails must be counted for the p-value.

This p-value is relatively large (larger than $\alpha = 0.05$), so we should not reject H_0. That is, if H_0 is true, it would not be very unusual to see a sample mean this far from 7 hours simply due to sampling variation. Thus, we do not have sufficient evidence to conclude that the mean is different than 7 hours.

Example 4.35 It is never okay to change two-sided tests to one-sided tests after observing the data. In this example we explore the consequences of ignoring this advice. Using $\alpha = 0.05$, we show that freely switching from two-sided tests to one-sided tests will cause us to make twice as many Type 1 Errors as intended.

Suppose the sample mean was larger than the null value, μ_0 (e.g. μ_0 would represent 7 if H_0: $\mu = 7$). Then if we can flip to a one-sided test, we would use H_A: $\mu > \mu_0$. Now if we obtain any observation with a Z-score greater than 1.65, we would reject H_0. If the null hypothesis is true, we incorrectly reject the null hypothesis about 5% of the time when the sample mean is above the null value, as shown in Figure 4.19.

Suppose the sample mean was smaller than the null value. Then if we change to a one-sided test, we would use H_A: $\mu < \mu_0$. If $\bar{x}$ had a Z-score smaller than -1.65, we would reject H_0. If the null hypothesis is true, then we would observe such a case about 5% of the time.

By examining these two scenarios, we can determine that we will make a Type 1 Error $5\% + 5\% = 10\%$ of the time if we are allowed to swap to the "best" one-sided test for the data. This is twice the error rate we prescribed with our significance level: $\alpha = 0.05$ (!).

Caution: One-sided hypotheses are allowed only *before* seeing data

After observing data, it is tempting to turn a two-sided test into a one-sided test. Avoid this temptation. Hypotheses must be set up *before* observing the data. If they are not, the test should be two-sided.

4.3.6 Choosing a significance level

Choosing a significance level for a test is important in many contexts, and the traditional level is 0.05. However, it is often helpful to adjust the significance level based on the

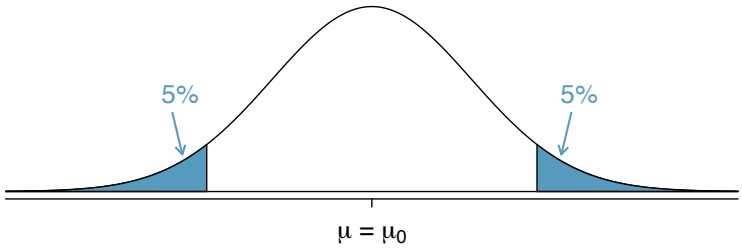


Figure 4.19: The shaded regions represent areas where we would reject H_0 under the bad practices considered in Example 4.35 when $\alpha = 0.05$.

application. We may select a level that is smaller or larger than 0.05 depending on the consequences of any conclusions reached from the test.

If making a Type 1 Error is dangerous or especially costly, we should choose a small significance level (e.g. 0.01). Under this scenario we want to be very cautious about rejecting the null hypothesis, so we demand very strong evidence favoring H_A before we would reject H_0.

If a Type 2 Error is relatively more dangerous or much more costly than a Type 1 Error, then we should choose a higher significance level (e.g. 0.10). Here we want to be cautious about failing to reject H_0 when the null is actually false.

Significance levels should reflect consequences of errors
The significance level selected for a test should reflect the consequences associated with Type 1 and Type 2 Errors.

Example 4.36 A car manufacturer is considering a higher quality but more expensive supplier for window parts in its vehicles. They sample a number of parts from their current supplier and also parts from the new supplier. They decide that if the high quality parts will last more than 12% longer, it makes financial sense to switch to this more expensive supplier. Is there good reason to modify the significance level in such a hypothesis test?

The null hypothesis is that the more expensive parts last no more than 12% longer while the alternative is that they do last more than 12% longer. This decision is just one of the many regular factors that have a marginal impact on the car and company. A significance level of 0.05 seems reasonable since neither a Type 1 or Type 2 Error should be dangerous or (relatively) much more expensive.

Example 4.37 The same car manufacturer is considering a slightly more expensive supplier for parts related to safety, not windows. If the durability of these safety components is shown to be better than the current supplier, they will switch manufacturers. Is there good reason to modify the significance level in such an evaluation?

The null hypothesis would be that the suppliers' parts are equally reliable. Because safety is involved, the car company should be eager to switch to the slightly more expensive manufacturer (reject H_0) even if the evidence of increased safety is only moderately strong. A slightly larger significance level, such as $\alpha = 0.10$, might be appropriate.

To obtain a 95% confidence interval, we choose $\alpha = 0.05$. Then, using Table A.5 with $v = 9$ degrees of freedom, we find $\chi^2_{0.025} = 19.023$ and $\chi^2_{0.975} = 2.700$. Therefore, the 95% confidence interval for σ^2 is

$$\frac{(9)(0.286)}{19.023} < \sigma^2 < \frac{(9)(0.286)}{2.700},$$

or simply $0.135 < \sigma^2 < 0.953$.

9.13 Two Samples: Estimating the Ratio of Two Variances

A point estimate of the ratio of two population variances σ_1^2/σ_2^2 is given by the ratio s_1^2/s_2^2 of the sample variances. Hence, the statistic S_1^2/S_2^2 is called an estimator of σ_1^2/σ_2^2.

If σ_1^2 and σ_2^2 are the variances of normal populations, we can establish an interval estimate of σ_1^2/σ_2^2 by using the statistic

$$F = \frac{\sigma_2^2 S_1^2}{\sigma_1^2 S_2^2}.$$

According to Theorem 8.8, the random variable F has an F-distribution with $v_1 = n_1 - 1$ and $v_2 = n_2 - 1$ degrees of freedom. Therefore, we may write (see Figure 9.8)

$$P[f_{1-\alpha/2}(v_1, v_2) < F < f_{\alpha/2}(v_1, v_2)] = 1 - \alpha,$$

where $f_{1-\alpha/2}(v_1, v_2)$ and $f_{\alpha/2}(v_1, v_2)$ are the values of the F-distribution with v_1 and v_2 degrees of freedom, leaving areas of $1 - \alpha/2$ and $\alpha/2$, respectively, to the right.

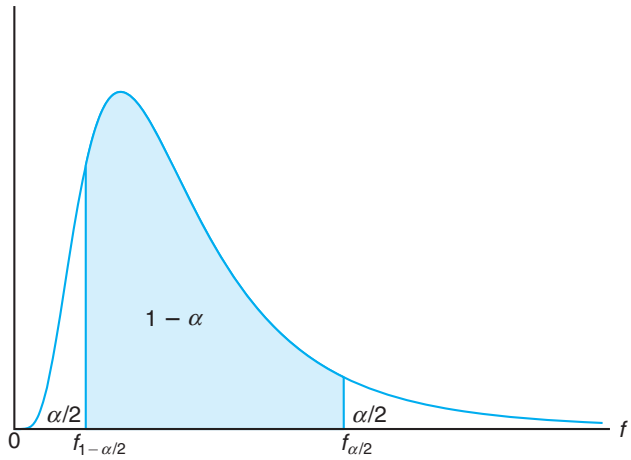


Figure 9.8: $P[f_{1-\alpha/2}(v_1, v_2) < F < f_{\alpha/2}(v_1, v_2)] = 1 - \alpha$.

Substituting for F, we write

$$P\left[f_{1-\alpha/2}(v_1,v_2) < \frac{\sigma_2^2 S_1^2}{\sigma_1^2 S_2^2} < f_{\alpha/2}(v_1,v_2)\right] = 1-\alpha.$$

Multiplying each term in the inequality by S_2^2/S_1^2 and then inverting each term, we obtain

$$P\left[\frac{S_1^2}{S_2^2}\frac{1}{f_{\alpha/2}(v_1,v_2)} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{S_1^2}{S_2^2}\frac{1}{f_{1-\alpha/2}(v_1,v_2)}\right] = 1-\alpha.$$

The results of Theorem 8.7 enable us to replace the quantity $f_{1-\alpha/2}(v_1,v_2)$ by $1/f_{\alpha/2}(v_2,v_1)$. Therefore,

$$P\left[\frac{S_1^2}{S_2^2}\frac{1}{f_{\alpha/2}(v_1,v_2)} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{S_1^2}{S_2^2} f_{\alpha/2}(v_2,v_1)\right] = 1-\alpha.$$

For any two independent random samples of sizes n_1 and n_2 selected from two normal populations, the ratio of the sample variances s_1^2/s_2^2 is computed, and the following $100(1-\alpha)\%$ confidence interval for σ_1^2/σ_2^2 is obtained.

Confidence Interval for σ_1^2/σ_2^2

If s_1^2 and s_2^2 are the variances of independent samples of sizes n_1 and n_2, respectively, from normal populations, then a $100(1-\alpha)\%$ confidence interval for σ_1^2/σ_2^2 is

$$\frac{s_1^2}{s_2^2}\frac{1}{f_{\alpha/2}(v_1,v_2)} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{s_1^2}{s_2^2} f_{\alpha/2}(v_2,v_1),$$

where $f_{\alpha/2}(v_1,v_2)$ is an f-value with $v_1 = n_1 - 1$ and $v_2 = n_2 - 1$ degrees of freedom, leaving an area of $\alpha/2$ to the right, and $f_{\alpha/2}(v_2,v_1)$ is a similar f-value with $v_2 = n_2 - 1$ and $v_1 = n_1 - 1$ degrees of freedom.

As in Section 9.12, an approximate $100(1-\alpha)\%$ confidence interval for σ_1/σ_2 is obtained by taking the square root of each endpoint of the interval for σ_1^2/σ_2^2.

Example 9.19: A confidence interval for the difference in the mean orthophosphorus contents, measured in milligrams per liter, at two stations on the James River was constructed in Example 9.12 on page 290 by assuming the normal population variance to be unequal. Justify this assumption by constructing 98% confidence intervals for σ_1^2/σ_2^2 and for σ_1/σ_2, where σ_1^2 and σ_2^2 are the variances of the populations of orthophosphorus contents at station 1 and station 2, respectively.

***Solution*:** From Example 9.12, we have $n_1 = 15$, $n_2 = 12$, $s_1 = 3.07$, and $s_2 = 0.80$. For a 98% confidence interval, $\alpha = 0.02$. Interpolating in Table A.6, we find $f_{0.01}(14,11) \approx 4.30$ and $f_{0.01}(11,14) \approx 3.87$. Therefore, the 98% confidence interval for σ_1^2/σ_2^2 is

$$\left(\frac{3.07^2}{0.80^2}\right)\left(\frac{1}{4.30}\right) < \frac{\sigma_1^2}{\sigma_2^2} < \left(\frac{3.07^2}{0.80^2}\right)(3.87),$$

which simplifies to $3.425 < \frac{\sigma_1^2}{\sigma_2^2} < 56.991$. Taking square roots of the confidence limits, we find that a 98% confidence interval for σ_1/σ_2 is

$$1.851 < \frac{\sigma_1}{\sigma_2} < 7.549.$$

Since this interval does not allow for the possibility of σ_1/σ_2 being equal to 1, we were correct in assuming that $\sigma_1 \neq \sigma_2$ or $\sigma_1^2 \neq \sigma_2^2$ in Example 9.12.

Exercises

9.71 A manufacturer of car batteries claims that the batteries will last, on average, 3 years with a variance of 1 year. If 5 of these batteries have lifetimes of 1.9, 2.4, 3.0, 3.5, and 4.2 years, construct a 95% confidence interval for σ^2 and decide if the manufacturer's claim that $\sigma^2 = 1$ is valid. Assume the population of battery lives to be approximately normally distributed.

9.72 A random sample of 20 students yielded a mean of $\bar{x} = 72$ and a variance of $s^2 = 16$ for scores on a college placement test in mathematics. Assuming the scores to be normally distributed, construct a 98% confidence interval for σ^2.

9.73 Construct a 95% confidence interval for σ^2 in Exercise 9.9 on page 283.

9.74 Construct a 99% confidence interval for σ^2 in Exercise 9.11 on page 283.

9.75 Construct a 99% confidence interval for σ in Exercise 9.12 on page 283.

9.76 Construct a 90% confidence interval for σ in Exercise 9.13 on page 283.

9.77 Construct a 98% confidence interval for σ_1/σ_2 in Exercise 9.42 on page 295, where σ_1 and σ_2 are, respectively, the standard deviations for the distances traveled per liter of fuel by the Volkswagen and Toyota mini-trucks.

9.78 Construct a 90% confidence interval for σ_1^2/σ_2^2 in Exercise 9.43 on page 295. Were we justified in assuming that $\sigma_1^2 \neq \sigma_2^2$ when we constructed the confidence interval for $\mu_1 - \mu_2$?

9.79 Construct a 90% confidence interval for σ_1^2/σ_2^2 in Exercise 9.46 on page 295. Should we have assumed $\sigma_1^2 = \sigma_2^2$ in constructing our confidence interval for $\mu_I - \mu_{II}$?

9.80 Construct a 95% confidence interval for σ_A^2/σ_B^2 in Exercise 9.49 on page 295. Should the equal-variance assumption be used?

9.14 Maximum Likelihood Estimation (Optional)

Often the estimators of parameters have been those that appeal to intuition. The estimator $\bar{X}$ certainly seems reasonable as an estimator of a population mean μ. The virtue of S^2 as an estimator of σ^2 is underscored through the discussion of unbiasedness in Section 9.3. The estimator for a binomial parameter p is merely a sample proportion, which, of course, is an *average* and appeals to common sense. But there are many situations in which it is not at all obvious what the proper estimator should be. As a result, there is much to be learned by the student of statistics concerning different philosophies that produce different methods of estimation. In this section, we deal with the **method of maximum likelihood**.

Maximum likelihood estimation is one of the most important approaches to estimation in all of statistical inference. We will not give a thorough development of the method. Rather, we will attempt to communicate the philosophy of maximum likelihood and illustrate with examples that relate to other estimation problems discussed in this chapter.

10.4 Single Sample: Tests Concerning a Single Mean

In this section, we formally consider tests of hypotheses on a single population mean. Many of the illustrations from previous sections involved tests on the mean, so the reader should already have insight into some of the details that are outlined here.

Tests on a Single Mean (Variance Known)

We should first describe the assumptions on which the experiment is based. The model for the underlying situation centers around an experiment with $X_1, X_2, \ldots, X_n$ representing a random sample from a distribution with mean μ and variance $\sigma^2 > 0$. Consider first the hypothesis

$$H_0: \mu = \mu_0,$$
$$H_1: \mu \neq \mu_0.$$

The appropriate test statistic should be based on the random variable $\bar{X}$. In Chapter 8, the Central Limit Theorem was introduced, which essentially states that despite the distribution of X, the random variable $\bar{X}$ has approximately a normal distribution with mean μ and variance σ^2/n for reasonably large sample sizes. So, $\mu_{\bar{X}} = \mu$ and $\sigma^2_{\bar{X}} = \sigma^2/n$. We can then determine a critical region based on the computed sample average, $\bar{x}$. It should be clear to the reader by now that there will be a two-tailed critical region for the test.

Standardization of $\bar{X}$

It is convenient to standardize $\bar{X}$ and formally involve the **standard normal** random variable Z, where

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}.$$

We know that *under* H_0, that is, if $\mu = \mu_0$, $\sqrt{n}(\bar{X} - \mu_0)/\sigma$ follows an $n(x; 0, 1)$ distribution, and hence the expression

$$P\left(-z_{\alpha/2} < \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} < z_{\alpha/2}\right) = 1 - \alpha$$

can be used to write an appropriate nonrejection region. The reader should keep in mind that, formally, the critical region is designed to control α, the probability of type I error. It should be obvious that a *two-tailed signal* of evidence is needed to support H_1. Thus, given a computed value $\bar{x}$, the formal test involves rejecting H_0 if the computed *test statistic* z falls in the critical region described next.

Test Procedure for a Single Mean (Variance Known)

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} > z_{\alpha/2} \quad \text{or} \quad z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} < -z_{\alpha/2}$$

If $-z_{\alpha/2} < z < z_{\alpha/2}$, do not reject H_0. Rejection of H_0, of course, implies acceptance of the alternative hypothesis $\mu \neq \mu_0$. With this definition of the critical region, it should be clear that there will be probability α of rejecting H_0 (falling into the critical region) when, indeed, $\mu = \mu_0$.

Although it is easier to understand the critical region written in terms of z, we can write the same critical region in terms of the computed average $\bar{x}$. The following can be written as an identical decision procedure:

$$\textit{reject } H_0 \text{ if } \bar{x} < a \text{ or } \bar{x} > b,$$

where

$$a = \mu_0 - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \qquad b = \mu_0 + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

Hence, for a significance level α, the critical values of the random variable z and $\bar{x}$ are both depicted in Figure 10.9.

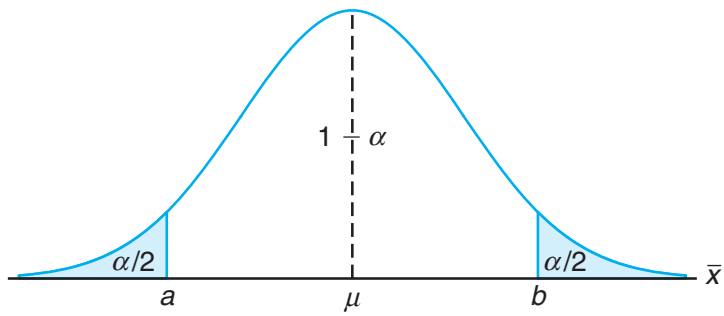


Figure 10.9: Critical region for the alternative hypothesis $\mu \neq \mu_0$.

Tests of one-sided hypotheses on the mean involve the same statistic described in the two-sided case. The difference, of course, is that the critical region is only in one tail of the standard normal distribution. For example, suppose that we seek to test

$$H_0\colon\ \mu = \mu_0,$$
$$H_1\colon\ \mu > \mu_0.$$

The signal that favors H_1 comes from *large values* of z. Thus, rejection of H_0 results when the computed $z > z_\alpha$. Obviously, if the alternative is $H_1\colon \mu < \mu_0$, the critical region is entirely in the lower tail and thus rejection results from $z < -z_\alpha$. Although in a one-sided testing case the null hypothesis can be written as $H_0\colon \mu \leq \mu_0$ or $H_0\colon \mu \geq \mu_0$, it is usually written as $H_0\colon \mu = \mu_0$.

The following two examples illustrate tests on means for the case in which σ is known.

Example 10.3: A random sample of 100 recorded deaths in the United States during the past year showed an average life span of 71.8 years. Assuming a population standard deviation of 8.9 years, does this seem to indicate that the mean life span today is greater than 70 years? Use a 0.05 level of significance.

Solution:

1. H_0: $\mu = 70$ years.
2. H_1: $\mu > 70$ years.
3. $\alpha = 0.05$.
4. Critical region: $z > 1.645$, where $z = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}$.
5. Computations: $\bar{x} = 71.8$ years, $\sigma = 8.9$ years, and hence $z = \frac{71.8-70}{8.9/\sqrt{100}} = 2.02$.
6. Decision: Reject H_0 and conclude that the mean life span today is greater than 70 years.

The P-value corresponding to $z = 2.02$ is given by the area of the shaded region in Figure 10.10.

Using Table A.3, we have

$$P = P(Z > 2.02) = 0.0217.$$

As a result, the evidence in favor of H_1 is even stronger than that suggested by a 0.05 level of significance. ∎

Example 10.4: A manufacturer of sports equipment has developed a new synthetic fishing line that the company claims has a mean breaking strength of 8 kilograms with a standard deviation of 0.5 kilogram. Test the hypothesis that $\mu = 8$ kilograms against the alternative that $\mu \neq 8$ kilograms if a random sample of 50 lines is tested and found to have a mean breaking strength of 7.8 kilograms. Use a 0.01 level of significance.

Solution:

1. H_0: $\mu = 8$ kilograms.
2. H_1: $\mu \neq 8$ kilograms.
3. $\alpha = 0.01$.
4. Critical region: $z < -2.575$ and $z > 2.575$, where $z = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}$.
5. Computations: $\bar{x} = 7.8$ kilograms, $n = 50$, and hence $z = \frac{7.8-8}{0.5/\sqrt{50}} = -2.83$.
6. Decision: Reject H_0 and conclude that the average breaking strength is not equal to 8 but is, in fact, less than 8 kilograms.

Since the test in this example is two tailed, the desired P-value is twice the area of the shaded region in Figure 10.11 to the left of $z = -2.83$. Therefore, using Table A.3, we have

$$P = P(|Z| > 2.83) = 2P(Z < -2.83) = 0.0046,$$

which allows us to reject the null hypothesis that $\mu = 8$ kilograms at a level of significance smaller than 0.01. ∎

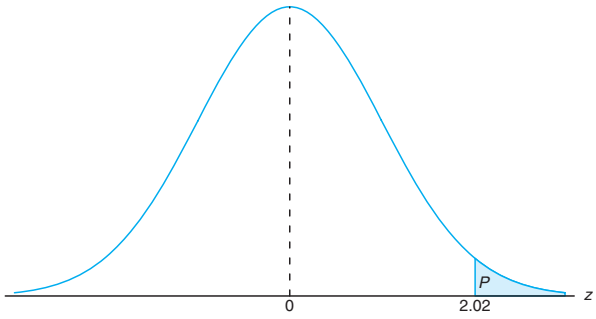


Figure 10.10: P-value for Example 10.3.

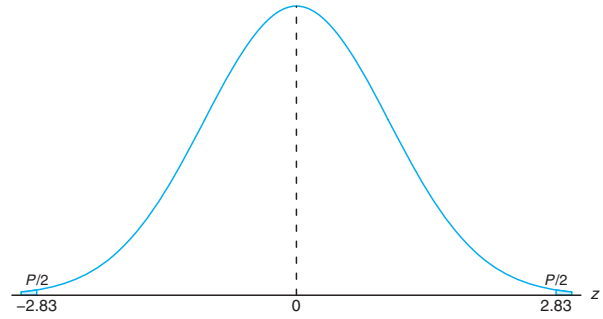


Figure 10.11: P-value for Example 10.4.

Relationship to Confidence Interval Estimation

The reader should realize by now that the hypothesis-testing approach to statistical inference in this chapter is very closely related to the confidence interval approach in Chapter 9. Confidence interval estimation involves computation of bounds within which it is "reasonable" for the parameter in question to lie. For the case of a single population mean μ with σ^2 known, the structure of both hypothesis testing and confidence interval estimation is based on the random variable

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}.$$

It turns out that the testing of H_0: $\mu = \mu_0$ against H_1: $\mu \neq \mu_0$ at a significance level α *is equivalent to computing a* $100(1-\alpha)\%$ *confidence interval on* μ *and rejecting* H_0 *if* μ_0 *is outside the confidence interval.* If μ_0 is inside the confidence interval, the hypothesis is not rejected. The equivalence is very intuitive and quite simple to illustrate. Recall that with an observed value $\bar{x}$, failure to reject H_0 at significance level α implies that

$$-z_{\alpha/2} \leq \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \leq z_{\alpha/2},$$

which is equivalent to

$$\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu_0 \leq \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

The equivalence of confidence interval estimation to hypothesis testing extends to differences between two means, variances, ratios of variances, and so on. As a result, the student of statistics should not consider confidence interval estimation and hypothesis testing as separate forms of statistical inference. For example, consider Example 9.2 on page 271. The 95% confidence interval on the mean is given by the bounds (2.50, 2.70). Thus, with the same sample information, a two-sided hypothesis on μ involving any hypothesized value between 2.50 and 2.70 will not be rejected. As we turn to different areas of hypothesis testing, the equivalence to the confidence interval estimation will continue to be exploited.

Tests on a Single Sample (Variance Unknown)

One would certainly suspect that tests on a population mean μ with σ^2 unknown, like confidence interval estimation, should involve the use of Student t-distribution. Strictly speaking, the application of Student t for both confidence intervals and hypothesis testing is developed under the following assumptions. The random variables $X_1, X_2, \ldots, X_n$ represent a random sample from a normal distribution with unknown μ and σ^2. Then the random variable $\sqrt{n}(\bar{X} - \mu)/S$ has a Student t-distribution with $n-1$ degrees of freedom. The structure of the test is identical to that for the case of σ known, with the exception that the value σ in the test statistic is replaced by the computed estimate S and the standard normal distribution is replaced by a t-distribution.

The t-Statistic for a Test on a Single Mean (Variance Unknown)

For the two-sided hypothesis

$$H_0: \ \mu = \mu_0,$$
$$H_1: \ \mu \neq \mu_0,$$

we reject H_0 at significance level α when the computed t-statistic

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

exceeds $t_{\alpha/2,n-1}$ or is less than $-t_{\alpha/2,n-1}$.

The reader should recall from Chapters 8 and 9 that the t-distribution is symmetric around the value zero. Thus, this two-tailed critical region applies in a fashion similar to that for the case of known σ. For the two-sided hypothesis at significance level α, the two-tailed critical regions apply. For H_1: $\mu > \mu_0$, rejection results when $t > t_{\alpha,n-1}$. For H_1: $\mu < \mu_0$, the critical region is given by $t < -t_{\alpha,n-1}$.

Example 10.5: The Edison Electric Institute has published figures on the number of kilowatt hours used annually by various home appliances. It is claimed that a vacuum cleaner uses an average of 46 kilowatt hours per year. If a random sample of 12 homes included in a planned study indicates that vacuum cleaners use an average of 42 kilowatt hours per year with a standard deviation of 11.9 kilowatt hours, does this suggest at the 0.05 level of significance that vacuum cleaners use, on average, less than 46 kilowatt hours annually? Assume the population of kilowatt hours to be normal.

Solution:

1. H_0: $\mu = 46$ kilowatt hours.
2. H_1: $\mu < 46$ kilowatt hours.
3. $\alpha = 0.05$.
4. Critical region: $t < -1.796$, where $t = \frac{\bar{x}-\mu_0}{s/\sqrt{n}}$ with 11 degrees of freedom.
5. Computations: $\bar{x} = 42$ kilowatt hours, $s = 11.9$ kilowatt hours, and $n = 12$. Hence,

$$t = \frac{42 - 46}{11.9/\sqrt{12}} = -1.16, \qquad P = P(T < -1.16) \approx 0.135.$$

6. Decision: Do not reject H_0 and conclude that the average number of kilowatt hours used annually by home vacuum cleaners is not significantly less than 46.

Comment on the Single-Sample *t*-Test

The reader has probably noticed that the equivalence of the two-tailed t-test for a single mean and the computation of a confidence interval on μ with σ replaced by s is maintained. For example, consider Example 9.5 on page 275. Essentially, we can view that computation as one in which we have found all values of μ_0, the hypothesized mean volume of containers of sulfuric acid, for which the hypothesis H_0: $\mu = \mu_0$ *will not be rejected* at $\alpha = 0.05$. Again, this is consistent with the statement "Based on the sample information, values of the population mean volume between 9.74 and 10.26 liters are not unreasonable."

Comments regarding the normality assumption are worth emphasizing at this point. We have indicated that when σ is known, the Central Limit Theorem allows for the use of a test statistic or a confidence interval which is based on Z, the standard normal random variable. Strictly speaking, of course, the Central Limit Theorem, and thus the use of the standard normal distribution, does not apply unless σ is known. In Chapter 8, the development of the t-distribution was given. There we pointed out that normality on $X_1, X_2, \ldots, X_n$ was an underlying assumption. Thus, *strictly speaking*, the Student's t-tables of percentage points for tests or confidence intervals should not be used unless it is known that the sample comes from a normal population. In practice, σ can rarely be assumed to be known. However, a very good estimate may be available from previous experiments. Many statistics textbooks suggest that one can safely replace σ by s in the test statistic

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

when $n \geq 30$ with a bell-shaped population and still use the Z-tables for the appropriate critical region. The implication here is that the Central Limit Theorem is indeed being invoked and one is relying on the fact that $s \approx \sigma$. Obviously, when this is done, the results must be viewed as approximate. Thus, a computed P-value (from the Z-distribution) of 0.15 may be 0.12 or perhaps 0.17, or a computed confidence interval may be a 93% confidence interval rather than a 95% interval as desired. Now what about situations where $n \leq 30$? The user cannot rely on s being close to σ, and in order to take into account the inaccuracy of the estimate, the confidence interval should be wider or the critical value larger in magnitude. The t-distribution percentage points accomplish this but are correct only when the sample is from a normal distribution. Of course, normal probability plots can be used to ascertain some sense of the deviation of normality in a data set.

For small samples, it is often difficult to detect deviations from a normal distribution. (Goodness-of-fit tests are discussed in a later section of this chapter.) For bell-shaped distributions of the random variables $X_1, X_2, \ldots, X_n$, the use of the t-distribution for tests or confidence intervals is likely to produce quite good results. When in doubt, the user should resort to nonparametric procedures, which are presented in Chapter 16.

Annotated Computer Printout for Single-Sample *t*-Test

It should be of interest for the reader to see an annotated computer printout showing the result of a single-sample t-test. Suppose that an engineer is interested in testing the bias in a pH meter. Data are collected on a neutral substance (pH = 7.0). A sample of the measurements were taken with the data as follows:

$$7.07 \quad 7.00 \quad 7.10 \quad 6.97 \quad 7.00 \quad 7.03 \quad 7.01 \quad 7.01 \quad 6.98 \quad 7.08$$

It is, then, of interest to test

$$H_0\text{: } \mu = 7.0,$$
$$H_1\text{: } \mu \neq 7.0.$$

In this illustration, we use the computer package *MINITAB* to illustrate the analysis of the data set above. Notice the key components of the printout shown in Figure 10.12. Of course, the mean $\bar{y}$ is 7.0250, StDev is simply the sample standard deviation $s = 0.044$, and SE Mean is the estimated standard error of the mean and is computed as $s/\sqrt{n} = 0.0139$. The t-value is the ratio

$$(7.0250 - 7)/0.0139 = 1.80.$$

```
pH-meter
  7.07   7.00   7.10   6.97   7.00   7.03   7.01   7.01   6.98   7.08
MTB > Onet 'pH-meter'; SUBC>   Test 7.

One-Sample T: pH-meter Test of mu = 7 vs not = 7
Variable  N    Mean   StDev  SE Mean      95% CI          T      P
pH-meter 10 7.02500 0.04403  0.01392 (6.99350, 7.05650) 1.80 0.106
```

Figure 10.12: *MINITAB* printout for one sample t-test for pH meter.

The P-value of 0.106 suggests results that are inconclusive. There is no evidence suggesting a strong rejection of H_0 (based on an α of 0.05 or 0.10), **yet one certainly cannot truly conclude that the pH meter is unbiased**. Notice that the sample size of 10 is rather small. An increase in sample size (perhaps another experiment) may sort things out. A discussion regarding appropriate sample size appears in Section 10.6.

10.5 Two Samples: Tests on Two Means

The reader should now understand the relationship between tests and confidence intervals, and can only heavily rely on details supplied by the confidence interval material in Chapter 9. Tests concerning two means represent a set of very important analytical tools for the scientist or engineer. The experimental setting is very much like that described in Section 9.8. Two independent random samples of sizes

n_1 and n_2, respectively, are drawn from two populations with means μ_1 and μ_2 and variances σ_1^2 and σ_2^2. We know that the random variable

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

has a standard normal distribution. Here we are assuming that n_1 and n_2 are sufficiently large that the Central Limit Theorem applies. Of course, if the two populations are normal, the statistic above has a standard normal distribution even for small n_1 and n_2. Obviously, if we can assume that $\sigma_1 = \sigma_2 = \sigma$, the statistic above reduces to

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sigma\sqrt{1/n_1 + 1/n_2}}.$$

The two statistics above serve as a basis for the development of the test procedures involving two means. The equivalence between tests and confidence intervals, along with the technical detail involving tests on one mean, allow a simple transition to tests on two means.

The two-sided hypothesis on two means can be written generally as

$$H_0\colon \mu_1 - \mu_2 = d_0.$$

Obviously, the alternative can be two sided or one sided. Again, the distribution used is the distribution of the test statistic under H_0. Values $\bar{x}_1$ and $\bar{x}_2$ are computed and, for σ_1 and σ_2 known, the test statistic is given by

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - d_0}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}},$$

with a two-tailed critical region in the case of a two-sided alternative. That is, reject H_0 in favor of H_1: $\mu_1 - \mu_2 \neq d_0$ if $z > z_{\alpha/2}$ or $z < -z_{\alpha/2}$. One-tailed critical regions are used in the case of the one-sided alternatives. The reader should, as before, study the test statistic and be satisfied that for, say, H_1: $\mu_1 - \mu_2 > d_0$, the signal favoring H_1 comes from large values of z. Thus, the upper-tailed critical region applies.

Unknown But Equal Variances

The more prevalent situations involving tests on two means are those in which variances are unknown. If the scientist involved is willing to assume that both distributions are normal and that $\sigma_1 = \sigma_2 = \sigma$, the *pooled t-test* (often called the two-sample t-test) may be used. The test statistic (see Section 9.8) is given by the following test procedure.

Two-Sample Pooled t-Test

For the two-sided hypothesis

$$H_0: \ \mu_1 = \mu_2,$$
$$H_1: \ \mu_1 \neq \mu_2,$$

we reject H_0 at significance level α when the computed t-statistic

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - d_0}{s_p\sqrt{1/n_1 + 1/n_2}},$$

where

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

exceeds $t_{\alpha/2,n_1+n_2-2}$ or is less than $-t_{\alpha/2,n_1+n_2-2}$.

Recall from Chapter 9 that the degrees of freedom for the t-distribution are a result of pooling of information from the two samples to estimate σ^2. One-sided alternatives suggest one-sided critical regions, as one might expect. For example, for H_1: $\mu_1 - \mu_2 > d_0$, reject H_1: $\mu_1 - \mu_2 = d_0$ when $t > t_{\alpha,n_1+n_2-2}$.

Example 10.6: An experiment was performed to compare the abrasive wear of two different laminated materials. Twelve pieces of material 1 were tested by exposing each piece to a machine measuring wear. Ten pieces of material 2 were similarly tested. In each case, the depth of wear was observed. The samples of material 1 gave an average (coded) wear of 85 units with a sample standard deviation of 4, while the samples of material 2 gave an average of 81 with a sample standard deviation of 5. Can we conclude at the 0.05 level of significance that the abrasive wear of material 1 exceeds that of material 2 by more than 2 units? Assume the populations to be approximately normal with equal variances.

***Solution*:** Let μ_1 and μ_2 represent the population means of the abrasive wear for material 1 and material 2, respectively.

1. H_0: $\mu_1 - \mu_2 = 2$.
2. H_1: $\mu_1 - \mu_2 > 2$.
3. $\alpha = 0.05$.
4. Critical region: $t > 1.725$, where $t = \frac{(\bar{x}_1-\bar{x}_2)-d_0}{s_p\sqrt{1/n_1+1/n_2}}$ with $v = 20$ degrees of freedom.
5. Computations:

$$\bar{x}_1 = 85, \quad s_1 = 4, \quad n_1 = 12,$$
$$\bar{x}_2 = 81, \quad s_2 = 5, \quad n_2 = 10.$$

Hence

$$s_p = \sqrt{\frac{(11)(16) + (9)(25)}{12 + 10 - 2}} = 4.478,$$
$$t = \frac{(85 - 81) - 2}{4.478\sqrt{1/12 + 1/10}} = 1.04,$$
$$P = P(T > 1.04) \approx 0.16. \qquad \text{(See Table A.4.)}$$

6. Decision: Do not reject H_0. We are unable to conclude that the abrasive wear of material 1 exceeds that of material 2 by more than 2 units. ∎

Unknown But Unequal Variances

There are situations where the analyst is **not** able to assume that $\sigma_1 = \sigma_2$. Recall from Section 9.8 that, if the populations are normal, the statistic

$$T' = \frac{(\bar{X}_1 - \bar{X}_2) - d_0}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$$

has an approximate t-distribution with approximate degrees of freedom

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

As a result, the test procedure is to *not reject* H_0 when

$$-t_{\alpha/2,v} < t' < t_{\alpha/2,v},$$

with v given as above. Again, as in the case of the pooled t-test, one-sided alternatives suggest one-sided critical regions.

Paired Observations

A study of the two-sample t-test or confidence interval on the difference between means should suggest the need for experimental design. Recall the discussion of experimental units in Chapter 9, where it was suggested that the conditions of the two populations (often referred to as the two treatments) should be assigned randomly to the experimental units. This is done to avoid biased results due to systematic differences between experimental units. In other words, in hypothesis-testing jargon, it is important that any significant difference found between means be due to the different conditions of the populations and not due to the experimental units in the study. For example, consider Exercise 9.40 in Section 9.9. The 20 seedlings play the role of the experimental units. Ten of them are to be treated with nitrogen and 10 with no nitrogen. It may be very important that this assignment to the "nitrogen" and "no-nitrogen" treatments be random to ensure that systematic differences between the seedlings do not interfere with a valid comparison between the means.

In Example 10.6, time of measurement is the most likely choice for the experimental unit. The 22 pieces of material should be measured in random order. We

need to guard against the possibility that wear measurements made close together in time might tend to give similar results. **Systematic** (nonrandom) **differences in experimental units** *are not expected*. However, random assignments guard against the problem.

References to planning of experiments, randomization, choice of sample size, and so on, will continue to influence much of the development in Chapters 13, 14, and 15. Any scientist or engineer whose interest lies in analysis of real data should study this material. The pooled t-test is extended in Chapter 13 to cover more than two means.

Testing of two means can be accomplished when data are in the form of paired observations, as discussed in Chapter 9. In this pairing structure, the conditions of the two populations (treatments) are assigned randomly within homogeneous units. Computation of the confidence interval for $\mu_1 - \mu_2$ in the situation with paired observations is based on the random variable

$$T = \frac{\bar{D} - \mu_D}{S_d/\sqrt{n}},$$

where $\bar{D}$ and S_d are random variables representing the sample mean and standard deviation of the differences of the observations in the experimental units. As in the case of the *pooled t-test*, the assumption is that the observations from each population are normal. This two-sample problem is essentially reduced to a one-sample problem by using the computed differences $d_1, d_2, \ldots, d_n$. Thus, the hypothesis reduces to

$$H_0\colon\ \mu_D = d_0.$$

The computed test statistic is then given by

$$t = \frac{\bar{d} - d_0}{s_d/\sqrt{n}}.$$

Critical regions are constructed using the t-distribution with $n - 1$ degrees of freedom.

Problem of Interaction in a Paired *t*-Test

Not only will the case study that follows illustrate the use of the paired t-test but the discussion will shed considerable light on the difficulties that arise when there is an interaction between the treatments and the experimental units in the paired t structure. Recall that interaction between factors was introduced in Section 1.7 in a discussion of general types of statistical studies. The concept of interaction will be an important issue from Chapter 13 through Chapter 15.

There are some types of statistical tests in which the existence of interaction results in difficulty. The paired t-test is one such example. In Section 9.9, the paired structure was used in the computation of a confidence interval on the difference between two means, and the advantage in pairing was revealed for situations in which the experimental units are homogeneous. The pairing results in a reduction in σ_D, the standard deviation of a difference $D_i = X_{1i} - X_{2i}$, as discussed in

Section 9.9. If interaction exists between treatments and experimental units, the advantage gained in pairing may be substantially reduced. Thus, in Example 9.13 on page 293, the no interaction assumption allowed the difference in mean TCDD levels (plasma vs. fat tissue) to be the same across veterans. A quick glance at the data would suggest that there is no significant violation of the assumption of no interaction.

In order to demonstrate how interaction influences $\text{Var}(D)$ and hence the quality of the paired t-test, it is instructive to revisit the ith difference given by $D_i = X_{1i} - X_{2i} = (\mu_1 - \mu_2) + (\epsilon_1 - \epsilon_2)$, where X_{1i} and X_{2i} are taken on the ith experimental unit. If the pairing unit is homogeneous, the errors in X_{1i} and in X_{2i} should be similar and not independent. We noted in Chapter 9 that the positive covariance between the errors results in a reduced $\text{Var}(D)$. Thus, the size of the difference in the treatments and the relationship between the errors in X_{1i} and X_{2i} contributed by the experimental unit will tend to allow a significant difference to be detected.

What Conditions Result in Interaction?

Let us consider a situation in which the experimental units are not homogeneous. Rather, consider the ith experimental unit with random variables X_{1i} and X_{2i} that are not similar. Let ϵ_{1i} and ϵ_{2i} be random variables representing the errors in the values X_{1i} and X_{2i}, respectively, at the ith unit. Thus, we may write

$$X_{1i} = \mu_1 + \epsilon_{1i} \text{ and } X_{2i} = \mu_2 + \epsilon_{2i}.$$

The errors with expectation zero may tend to cause the response values X_{1i} and X_{2i} to move in opposite directions, resulting in a negative value for $\text{Cov}(\epsilon_{1i}, \epsilon_{2i})$ and hence negative $\text{Cov}(X_{1i}, X_{2i})$. In fact, the model may be complicated even more by the fact that $\sigma_1^2 = \text{Var}(\epsilon_{1i}) \neq \sigma_2^2 = \text{Var}(\epsilon_{2i})$. The variance and covariance parameters may vary among the n experimental units. Thus, unlike in the homogeneous case, D_i will tend to be quite different across experimental units due to the heterogeneous nature of the difference in $\epsilon_1 - \epsilon_2$ among the units. This produces the interaction between treatments and units. In addition, for a specific experimental unit (see Theorem 4.9),

$$\sigma_D^2 = \text{Var}(D) = \text{Var}(\epsilon_1) + \text{Var}(\epsilon_2) - 2\ \text{Cov}(\epsilon_1, \epsilon_2)$$

is inflated by the negative covariance term, and thus the advantage gained in pairing in the homogeneous unit case is lost in the case described here. While the inflation in $\text{Var}(D)$ will vary from case to case, there is a danger in some cases that the increase in variance may neutralize any difference that exists between μ_1 and μ_2. Of course, a large value of $\bar{d}$ in the t-statistic may reflect a treatment difference that overcomes the inflated variance estimate, s_d^2.

Case Study 10.1: **Blood Sample Data**: In a study conducted in the Forestry and Wildlife Department at Virginia Tech, J. A. Wesson examined the influence of the drug succinylcholine on the circulation levels of androgens in the blood. Blood samples were taken from wild, free-ranging deer immediately after they had received an intramuscular injection of succinylcholine administered using darts and a capture gun. A second blood sample was obtained from each deer 30 minutes after the

first sample, after which the deer was released. The levels of androgens at time of capture and 30 minutes later, measured in nanograms per milliliter (ng/mL), for 15 deer are given in Table 10.2.

Assuming that the populations of androgen levels at time of injection and 30 minutes later are normally distributed, test at the 0.05 level of significance whether the androgen concentrations are altered after 30 minutes.

Table 10.2: Data for Case Study 10.1

	Androgen (ng/mL)		
Deer	**At Time of Injection**	**30 Minutes after Injection**	d_i
1	2.76	7.02	4.26
2	5.18	3.10	−2.08
3	2.68	5.44	2.76
4	3.05	3.99	0.94
5	4.10	5.21	1.11
6	7.05	10.26	3.21
7	6.60	13.91	7.31
8	4.79	18.53	13.74
9	7.39	7.91	0.52
10	7.30	4.85	−2.45
11	11.78	11.10	−0.68
12	3.90	3.74	−0.16
13	26.00	94.03	68.03
14	67.48	94.03	26.55
15	17.04	41.70	24.66

***Solution*:** Let μ_1 and μ_2 be the average androgen concentration at the time of injection and 30 minutes later, respectively. We proceed as follows:

1. H_0: $\mu_1 = \mu_2$ or $\mu_D = \mu_1 - \mu_2 = 0$.
2. H_1: $\mu_1 \neq \mu_2$ or $\mu_D = \mu_1 - \mu_2 \neq 0$.
3. $\alpha = 0.05$.
4. Critical region: $t < -2.145$ and $t > 2.145$, where $t = \frac{\overline{d} - d_0}{s_D/\sqrt{n}}$ with $v = 14$ degrees of freedom.
5. Computations: The sample mean and standard deviation for the d_i are

$$\overline{d} = 9.848 \qquad \text{and} \qquad s_d = 18.474.$$

Therefore,

$$t = \frac{9.848 - 0}{18.474/\sqrt{15}} = 2.06.$$

6. Though the t-statistic is not significant at the 0.05 level, from Table A.4,

$$P = P(|T| > 2.06) \approx 0.06.$$

As a result, there is some evidence that there is a difference in mean circulating levels of androgen.

The assumption of no interaction would imply that the effect on androgen levels of the deer is roughly the same in the data for both treatments, i.e., at the time of injection of succinylcholine and 30 minutes following injection. This can be expressed with the two factors switching roles; for example, the difference in treatments is roughly the same across the units (i.e., the deer). There certainly are some deer/treatment combinations for which the no interaction assumption seems to hold, but there is hardly any strong evidence that the experimental units are homogeneous. However, the nature of the interaction and the resulting increase in $\text{Var}(\bar{D})$ appear to be dominated by a substantial difference in the treatments. This is further demonstrated by the fact that 11 of the 15 deer exhibited positive signs for the computed d_i and the negative d_i (for deer 2, 10, 11, and 12) are small in magnitude compared to the 12 positive ones. Thus, it appears that the mean level of androgen is significantly higher 30 minutes following injection than at injection, and the conclusions may be stronger than $p = 0.06$ would suggest.

Annotated Computer Printout for Paired *t*-Test

Figure 10.13 displays a *SAS* computer printout for a paired t-test using the data of Case Study 10.1. Notice that the printout looks like that for a single sample t-test and, of course, that is exactly what is accomplished, since the test seeks to determine if $\overline{d}$ is significantly different from zero.

Analysis Variable : Diff

N	Mean	Std Error	t Value	Pr > |t|
15	9.8480000	4.7698699	2.06	0.0580

Figure 10.13: *SAS* printout of paired t-test for data of Case Study 10.1.

Summary of Test Procedures

As we complete the formal development of tests on population means, we offer Table 10.3, which summarizes the test procedure for the cases of a single mean and two means. Notice the approximate procedure when distributions are normal and variances are unknown but not assumed to be equal. This statistic was introduced in Chapter 9.

10.6 Choice of Sample Size for Testing Means

In Section 10.2, we demonstrated how the analyst can exploit relationships among the sample size, the significance level α, and the power of the test to achieve a certain standard of quality. In most practical circumstances, the experiment should be planned, with a choice of sample size made prior to the data-taking process if possible. The sample size is usually determined to achieve good power for a fixed α and fixed specific alternative. This fixed alternative may be in the

Table 10.3: Tests Concerning Means

H_0	Value of Test Statistic	H_1	Critical Region
$\mu = \mu_0$	$z = \dfrac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$; σ known	$\mu < \mu_0$ $\mu > \mu_0$ $\mu \neq \mu_0$	$z < -z_\alpha$ $z > z_\alpha$ $z < -z_{\alpha/2}$ or $z > z_{\alpha/2}$
$\mu = \mu_0$	$t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}}$; $v = n-1$, σ unknown	$\mu < \mu_0$ $\mu > \mu_0$ $\mu \neq \mu_0$	$t < -t_\alpha$ $t > t_\alpha$ $t < -t_{\alpha/2}$ or $t > t_{\alpha/2}$
$\mu_1 - \mu_2 = d_0$	$z = \dfrac{(\bar{x}_1 - \bar{x}_2) - d_0}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$; σ_1 and σ_2 known	$\mu_1 - \mu_2 < d_0$ $\mu_1 - \mu_2 > d_0$ $\mu_1 - \mu_2 \neq d_0$	$z < -z_\alpha$ $z > z_\alpha$ $z < -z_{\alpha/2}$ or $z > z_{\alpha/2}$
$\mu_1 - \mu_2 = d_0$	$t = \dfrac{(\bar{x}_1 - \bar{x}_2) - d_0}{s_p\sqrt{1/n_1 + 1/n_2}}$; $v = n_1 + n_2 - 2$, $\sigma_1 = \sigma_2$ but unknown, $s_p^2 = \dfrac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$	$\mu_1 - \mu_2 < d_0$ $\mu_1 - \mu_2 > d_0$ $\mu_1 - \mu_2 \neq d_0$	$t < -t_\alpha$ $t > t_\alpha$ $t < -t_{\alpha/2}$ or $t > t_{\alpha/2}$
$\mu_1 - \mu_2 = d_0$	$t' = \dfrac{(\bar{x}_1 - \bar{x}_2) - d_0}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$; $v = \dfrac{(s_1^2/n_1 + s_2^2/n_2)^2}{\frac{(s_1^2/n_1)^2}{n_1-1} + \frac{(s_2^2/n_2)^2}{n_2-1}}$, $\sigma_1 \neq \sigma_2$ and unknown	$\mu_1 - \mu_2 < d_0$ $\mu_1 - \mu_2 > d_0$ $\mu_1 - \mu_2 \neq d_0$	$t' < -t_\alpha$ $t' > t_\alpha$ $t' < -t_{\alpha/2}$ or $t' > t_{\alpha/2}$
$\mu_D = d_0$ paired observations	$t = \dfrac{\bar{d} - d_0}{s_d/\sqrt{n}}$; $v = n-1$	$\mu_D < d_0$ $\mu_D > d_0$ $\mu_D \neq d_0$	$t < -t_\alpha$ $t > t_\alpha$ $t < -t_{\alpha/2}$ or $t > t_{\alpha/2}$

form of $\mu - \mu_0$ in the case of a hypothesis involving a single mean or $\mu_1 - \mu_2$ in the case of a problem involving two means. Specific cases will provide illustrations.

Suppose that we wish to test the hypothesis

$$H_0\colon\ \mu = \mu_0,$$
$$H_1\colon\ \mu > \mu_0,$$

with a significance level α, when the variance σ^2 is known. For a specific alternative, say $\mu = \mu_0 + \delta$, the power of our test is shown in Figure 10.14 to be

$$1 - \beta = P(\bar{X} > a \text{ when } \mu = \mu_0 + \delta).$$

Therefore,

$$\begin{aligned}\beta &= P(\bar{X} < a \text{ when } \mu = \mu_0 + \delta)\\ &= P\left[\frac{\bar{X} - (\mu_0 + \delta)}{\sigma/\sqrt{n}} < \frac{a - (\mu_0 + \delta)}{\sigma/\sqrt{n}} \text{ when } \mu = \mu_0 + \delta\right].\end{aligned}$$

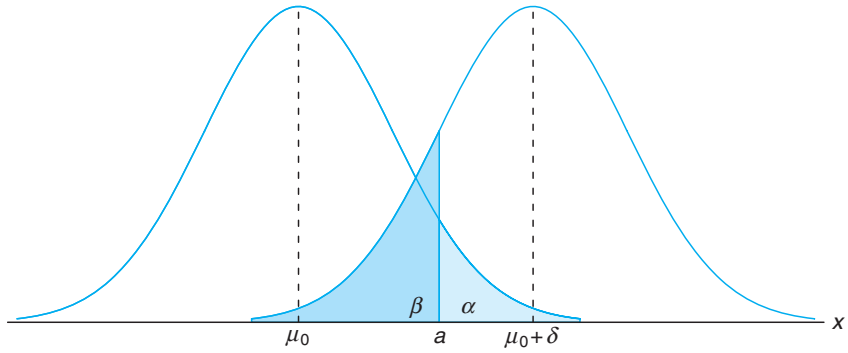


Figure 10.14: Testing $\mu = \mu_0$ versus $\mu = \mu_0 + \delta$.

Under the alternative hypothesis $\mu = \mu_0 + \delta$, the statistic

$$\frac{\bar{X} - (\mu_0 + \delta)}{\sigma/\sqrt{n}}$$

is the standard normal variable Z. So

$$\beta = P\left(Z < \frac{a - \mu_0}{\sigma/\sqrt{n}} - \frac{\delta}{\sigma/\sqrt{n}}\right) = P\left(Z < z_\alpha - \frac{\delta}{\sigma/\sqrt{n}}\right),$$

from which we conclude that

$$-z_\beta = z_\alpha - \frac{\delta\sqrt{n}}{\sigma},$$

and hence

$$\text{Choice of sample size:} \quad n = \frac{(z_\alpha + z_\beta)^2\sigma^2}{\delta^2},$$

a result that is also true when the alternative hypothesis is $\mu < \mu_0$.

In the case of a two-tailed test, we obtain the power $1 - \beta$ for a specified alternative when

$$n \approx \frac{(z_{\alpha/2} + z_\beta)^2\sigma^2}{\delta^2}.$$

Example 10.7: Suppose that we wish to test the hypothesis

$$H_0\text{: } \mu = 68 \text{ kilograms},$$
$$H_1\text{: } \mu > 68 \text{ kilograms}$$

for the weights of male students at a certain college, using an $\alpha = 0.05$ level of significance, when it is known that $\sigma = 5$. Find the sample size required if the power of our test is to be 0.95 when the true mean is 69 kilograms.

***Solution*:** Since $\alpha = \beta = 0.05$, we have $z_\alpha = z_\beta = 1.645$. For the alternative $\beta = 69$, we take $\delta = 1$ and then

$$n = \frac{(1.645 + 1.645)^2(25)}{1} = 270.6.$$

Therefore, 271 observations are required if the test is to reject the null hypothesis 95% of the time when, in fact, μ is as large as 69 kilograms.

Two-Sample Case

A similar procedure can be used to determine the sample size $n = n_1 = n_2$ required for a specific power of the test in which two population means are being compared. For example, suppose that we wish to test the hypothesis

$$H_0: \ \mu_1 - \mu_2 = d_0,$$
$$H_1: \mu_1 - \mu_2 \neq d_0,$$

when σ_1 and σ_2 are known. For a specific alternative, say $\mu_1 - \mu_2 = d_0 + \delta$, the power of our test is shown in Figure 10.15 to be

$$1 - \beta = P(|\bar{X}_1 - \bar{X}_2| > a \text{ when } \mu_1 - \mu_2 = d_0 + \delta).$$

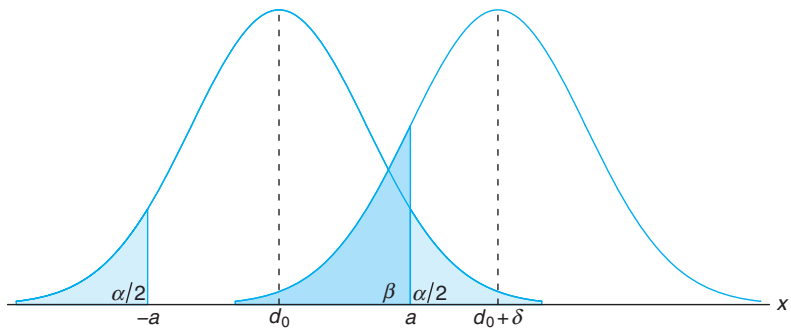


Figure 10.15: Testing $\mu_1 - \mu_2 = d_0$ versus $\mu_1 - \mu_2 = d_0 + \delta$.

Therefore,

$$\begin{aligned}\beta &= P(-a < \bar{X}_1 - \bar{X}_2 < a \text{ when } \mu_1 - \mu_2 = d_0 + \delta) \\ &= P\left[\frac{-a - (d_0 + \delta)}{\sqrt{(\sigma_1^2 + \sigma_2^2)/n}} < \frac{(\bar{X}_1 - \bar{X}_2) - (d_0 + \delta)}{\sqrt{(\sigma_1^2 + \sigma_2^2)/n}}\right. \\ &\qquad \left. < \frac{a - (d_0 + \delta)}{\sqrt{(\sigma_1^2 + \sigma_2^2)/n}} \text{ when } \mu_1 - \mu_2 = d_0 + \delta\right].\end{aligned}$$

Under the alternative hypothesis $\mu_1 - \mu_2 = d_0 + \delta$, the statistic

$$\frac{\bar{X}_1 - \bar{X}_2 - (d_0 + \delta)}{\sqrt{(\sigma_1^2 + \sigma_2^2)/n}}$$

is the standard normal variable Z. Now, writing

$$-z_{\alpha/2} = \frac{-a - d_0}{\sqrt{(\sigma_1^2 + \sigma_2^2)/n}} \quad \text{and} \quad z_{\alpha/2} = \frac{a - d_0}{\sqrt{(\sigma_1^2 + \sigma_2^2)/n}},$$

we have

$$\beta = P\left[-z_{\alpha/2} - \frac{\delta}{\sqrt{(\sigma_1^2 + \sigma_2^2)/n}} < Z < z_{\alpha/2} - \frac{\delta}{\sqrt{(\sigma_1^2 + \sigma_2^2)/n}}\right],$$

from which we conclude that

$$-z_\beta \approx z_{\alpha/2} - \frac{\delta}{\sqrt{(\sigma_1^2 + \sigma_2^2)/n}},$$

and hence

$$n \approx \frac{(z_{\alpha/2} + z_\beta)^2(\sigma_1^2 + \sigma_2^2)}{\delta^2}.$$

For the one-tailed test, the expression for the required sample size when $n = n_1 = n_2$ is

$$\text{Choice of sample size:} \quad n = \frac{(z_\alpha + z_\beta)^2(\sigma_1^2 + \sigma_2^2)}{\delta^2}.$$

When the population variance (or variances, in the two-sample situation) is unknown, the choice of sample size is not straightforward. In testing the hypothesis $\mu = \mu_0$ when the true value is $\mu = \mu_0 + \delta$, the statistic

$$\frac{\bar{X} - (\mu_0 + \delta)}{S/\sqrt{n}}$$

does not follow the t-distribution, as one might expect, but instead follows the **noncentral *t*-distribution**. However, tables or charts based on the noncentral t-distribution do exist for determining the appropriate sample size if some estimate of σ is available or if δ is a multiple of σ. Table A.8 gives the sample sizes needed to control the values of α and β for various values of

$$\Delta = \frac{|\delta|}{\sigma} = \frac{|\mu - \mu_0|}{\sigma}$$

for both one- and two-tailed tests. In the case of the two-sample t-test in which the variances are unknown but assumed equal, we obtain the sample sizes $n = n_1 = n_2$ needed to control the values of α and β for various values of

$$\Delta = \frac{|\delta|}{\sigma} = \frac{|\mu_1 - \mu_2 - d_0|}{\sigma}$$

from Table A.9.

Example 10.8: In comparing the performance of two catalysts on the effect of a reaction yield, a two-sample t-test is to be conducted with $\alpha = 0.05$. The variances in the yields

are considered to be the same for the two catalysts. How large a sample for each catalyst is needed to test the hypothesis

$$H_0: \ \mu_1 = \mu_2,$$
$$H_1: \ \mu_1 \neq \mu_2$$

if it is essential to detect a difference of 0.8σ between the catalysts with probability 0.9?

***Solution*:** From Table A.9, with $\alpha = 0.05$ for a two-tailed test, $\beta = 0.1$, and

$$\Delta = \frac{|0.8\sigma|}{\sigma} = 0.8,$$

we find the required sample size to be $n = 34$. ∎

In practical situations, it might be difficult to force a scientist or engineer to make a commitment on information from which a value of Δ can be found. The reader is reminded that the Δ-value quantifies the kind of difference between the means that the scientist considers important, that is, a difference considered *significant* from a scientific, not a statistical, point of view. Example 10.8 illustrates how this choice is often made, namely, by selecting a fraction of σ. Obviously, if the sample size is based on a choice of $|\delta|$ that is a small fraction of σ, the resulting sample size may be quite large compared to what the study allows.

10.7 Graphical Methods for Comparing Means

In Chapter 1, considerable attention was directed to displaying data in graphical form, such as stem-and-leaf plots and box-and-whisker plots. In Section 8.8, quantile plots and quantile-quantile normal plots were used to provide a "picture" to summarize a set of experimental data. Many computer software packages produce graphical displays. As we proceed to other forms of data analysis (e.g., regression analysis and analysis of variance), graphical methods become even more informative.

Graphical aids cannot be used as a replacement for the test procedure itself. Certainly, the value of the test statistic indicates the proper type of evidence in support of H_0 or H_1. However, a pictorial display provides a good illustration and is often a better communicator of evidence to the beneficiary of the analysis. Also, a picture will often clarify why a significant difference was found. Failure of an important assumption may be exposed by a summary type of graphical tool.

For the comparison of means, side-by-side box-and-whisker plots provide a telling display. The reader should recall that these plots display the 25th percentile, 75th percentile, and the median in a data set. In addition, the whiskers display the extremes in a data set. Consider Exercise 10.40 at the end of this section. Plasma ascorbic acid levels were measured in two groups of pregnant women, smokers and nonsmokers. Figure 10.16 shows the box-and-whisker plots for both groups of women. Two things are very apparent. Taking into account variability, there appears to be a negligible difference in the sample means. In addition, the variability in the two groups appears to be somewhat different. Of course, the analyst must keep in mind the rather sizable differences between the sample sizes in this case.

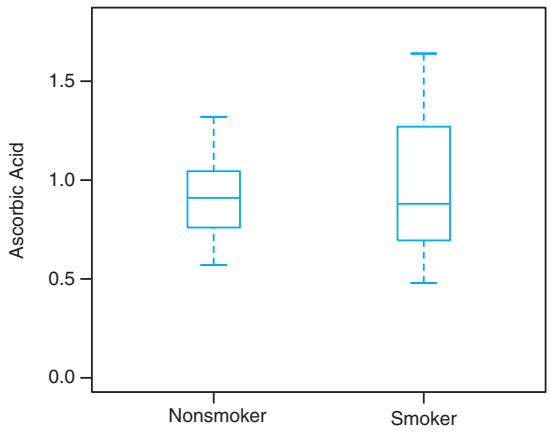


Figure 10.16: Two box-and-whisker plots of plasma ascorbic acid in smokers and nonsmokers.

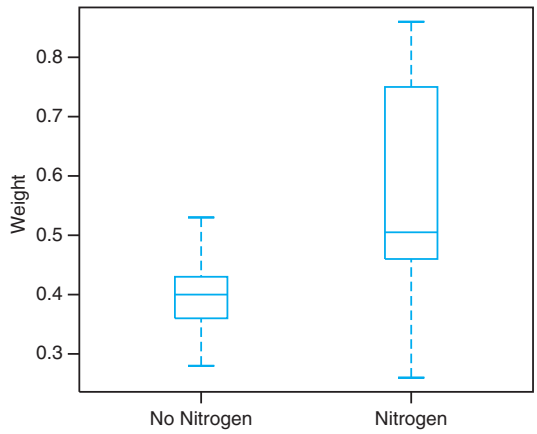


Figure 10.17: Two box-and-whisker plots of seedling data.

Consider Exercise 9.40 in Section 9.9. Figure 10.17 shows the multiple box-and-whisker plot for the data on 10 seedlings, half given nitrogen and half given no nitrogen. The display reveals a smaller variability for the group containing no nitrogen. In addition, the lack of overlap of the box plots suggests a significant difference between the mean stem weights for the two groups. It would appear that the presence of nitrogen increases the stem weights and perhaps increases the variability in the weights.

There are no certain rules of thumb regarding when two box-and-whisker plots give evidence of significant difference between the means. However, a rough guideline is that if the 25th percentile line for one sample exceeds the median line for the other sample, there is strong evidence of a difference between means.

More emphasis is placed on graphical methods in a real-life case study presented later in this chapter.

Annotated Computer Printout for Two-Sample *t*-Test

Consider once again Exercise 9.40 on page 294, where seedling data under conditions of nitrogen and no nitrogen were collected. Test

$$H_0: \ \mu_{\text{NIT}} = \mu_{\text{NON}},$$
$$H_1: \ \mu_{\text{NIT}} > \mu_{\text{NON}},$$

where the population means indicate mean weights. Figure 10.18 is an annotated computer printout generated using the *SAS* package. Notice that sample standard deviation and standard error are shown for both samples. The t-statistics under the assumption of equal variance and unequal variance are both given. From the box-and-whisker plot of Figure 10.17 it would certainly appear that the equal variance assumption is violated. A P-value of 0.0229 suggests a conclusion of unequal means. This concurs with the diagnostic information given in Figure 10.18. Incidentally, notice that t and t' are equal in this case, since $n_1 = n_2$.

```
                TTEST Procedure
Variable Weight
    Mineral      N    Mean   Std Dev Std Err
No nitrogen     10  0.3990    0.0728  0.0230
   Nitrogen     10  0.5650    0.1867  0.0591

Variances       DF    t Value    Pr > |t|
    Equal       18       2.62      0.0174
  Unequal     11.7       2.62      0.0229

          Test the Equality of Variances
Variable   Num DF    Den DF    F Value    Pr > F
  Weight        9         9       6.58    0.0098
```

Figure 10.18: *SAS* printout for two-sample t-test.

Exercises

10.19 In a research report, Richard H. Weindruch of the UCLA Medical School claims that mice with an average life span of 32 months will live to be about 40 months old when 40% of the calories in their diet are replaced by vitamins and protein. Is there any reason to believe that $\mu < 40$ if 64 mice that are placed on this diet have an average life of 38 months with a standard deviation of 5.8 months? Use a P-value in your conclusion.

10.20 A random sample of 64 bags of white cheddar popcorn weighed, on average, 5.23 ounces with a standard deviation of 0.24 ounce. Test the hypothesis that $\mu = 5.5$ ounces against the alternative hypothesis, $\mu < 5.5$ ounces, at the 0.05 level of significance.

10.21 An electrical firm manufactures light bulbs that have a lifetime that is approximately normally distributed with a mean of 800 hours and a standard deviation of 40 hours. Test the hypothesis that $\mu = 800$ hours against the alternative, $\mu \neq 800$ hours, if a random sample of 30 bulbs has an average life of 788 hours. Use a P-value in your answer.

10.22 In the American Heart Association journal *Hypertension*, researchers report that individuals who practice Transcendental Meditation (TM) lower their blood pressure significantly. If a random sample of 225 male TM practitioners meditate for 8.5 hours per week with a standard deviation of 2.25 hours, does that suggest that, on average, men who use TM meditate more than 8 hours per week? Quote a P-value in your conclusion.

10.23 Test the hypothesis that the average content of containers of a particular lubricant is 10 liters if the contents of a random sample of 10 containers are 10.2, 9.7, 10.1, 10.3, 10.1, 9.8, 9.9, 10.4, 10.3, and 9.8 liters. Use a 0.01 level of significance and assume that the distribution of contents is normal.

10.24 The average height of females in the freshman class of a certain college has historically been 162.5 centimeters with a standard deviation of 6.9 centimeters. Is there reason to believe that there has been a change in the average height if a random sample of 50 females in the present freshman class has an average height of 165.2 centimeters? Use a P-value in your conclusion. Assume the standard deviation remains the same.

10.25 It is claimed that automobiles are driven on average more than 20,000 kilometers per year. To test this claim, 100 randomly selected automobile owners are asked to keep a record of the kilometers they travel. Would you agree with this claim if the random sample showed an average of 23,500 kilometers and a standard deviation of 3900 kilometers? Use a P-value in your conclusion.

10.26 According to a dietary study, high sodium intake may be related to ulcers, stomach cancer, and migraine headaches. The human requirement for salt is only 220 milligrams per day, which is surpassed in most single servings of ready-to-eat cereals. If a random sample of 20 similar servings of a certain cereal has a mean sodium content of 244 milligrams and a standard deviation of 24.5 milligrams, does this suggest at the 0.05 level of significance that the average sodium content for a single serving of such cereal is greater than 220 milligrams? Assume the distribution of sodium contents to be normal.

10.27 A study at the University of Colorado at Boulder shows that running increases the percent resting metabolic rate (RMR) in older women. The average RMR of 30 elderly women runners was 34.0% higher than the average RMR of 30 sedentary elderly women, and the standard deviations were reported to be 10.5 and 10.2%, respectively. Was there a significant increase in RMR of the women runners over the sedentary women? Assume the populations to be approximately normally distributed with equal variances. Use a P-value in your conclusions.

10.28 According to *Chemical Engineering*, an important property of fiber is its water absorbency. The average percent absorbency of 25 randomly selected pieces of cotton fiber was found to be 20 with a standard deviation of 1.5. A random sample of 25 pieces of acetate yielded an average percent of 12 with a standard deviation of 1.25. Is there strong evidence that the population mean percent absorbency is significantly higher for cotton fiber than for acetate? Assume that the percent absorbency is approximately normally distributed and that the population variances in percent absorbency for the two fibers are the same. Use a significance level of 0.05.

10.29 Past experience indicates that the time required for high school seniors to complete a standardized test is a normal random variable with a mean of 35 minutes. If a random sample of 20 high school seniors took an average of 33.1 minutes to complete this test with a standard deviation of 4.3 minutes, test the hypothesis, at the 0.05 level of significance, that $\mu = 35$ minutes against the alternative that $\mu < 35$ minutes.

10.30 A random sample of size $n_1 = 25$, taken from a normal population with a standard deviation $\sigma_1 = 5.2$, has a mean $\bar{x}_1 = 81$. A second random sample of size $n_2 = 36$, taken from a different normal population with a standard deviation $\sigma_2 = 3.4$, has a mean $\bar{x}_2 = 76$. Test the hypothesis that $\mu_1 = \mu_2$ against the alternative, $\mu_1 \neq \mu_2$. Quote a P-value in your conclusion.

10.31 A manufacturer claims that the average tensile strength of thread A exceeds the average tensile strength of thread B by at least 12 kilograms. To test this claim, 50 pieces of each type of thread were tested under similar conditions. Type A thread had an average tensile strength of 86.7 kilograms with a standard deviation of 6.28 kilograms, while type B thread had an average tensile strength of 77.8 kilograms with a standard deviation of 5.61 kilograms. Test the manufacturer's claim using a 0.05 level of significance.

10.32 *Amstat News* (December 2004) lists median salaries for associate professors of statistics at research institutions and at liberal arts and other institutions in the United States. Assume that a sample of 200 associate professors from research institutions has an average salary of \$70,750 per year with a standard deviation of \$6000. Assume also that a sample of 200 associate professors from other types of institutions has an average salary of \$65,200 with a standard deviation of \$5000. Test the hypothesis that the mean salary for associate professors in research institutions is \$2000 higher than for those in other institutions. Use a 0.01 level of significance.

10.33 A study was conducted to see if increasing the substrate concentration has an appreciable effect on the velocity of a chemical reaction. With a substrate concentration of 1.5 moles per liter, the reaction was run 15 times, with an average velocity of 7.5 micromoles per 30 minutes and a standard deviation of 1.5. With a substrate concentration of 2.0 moles per liter, 12 runs were made, yielding an average velocity of 8.8 micromoles per 30 minutes and a sample standard deviation of 1.2. Is there any reason to believe that this increase in substrate concentration causes an increase in the mean velocity of the reaction of more than 0.5 micromole per 30 minutes? Use a 0.01 level of significance and assume the populations to be approximately normally distributed with equal variances.

10.34 A study was made to determine if the subject matter in a physics course is better understood when a lab constitutes part of the course. Students were randomly selected to participate in either a 3-semester-hour course without labs or a 4-semester-hour course with labs. In the section with labs, 11 students made an average grade of 85 with a standard deviation of 4.7, and in the section without labs, 17 students made an average grade of 79 with a standard deviation of 6.1. Would you say that the laboratory course increases the average grade by as much as 8 points? Use a P-value in your conclusion and assume the populations to be approximately normally distributed with equal variances.

10.35 To find out whether a new serum will arrest leukemia, 9 mice, all with an advanced stage of the disease, are selected. Five mice receive the treatment and 4 do not. Survival times, in years, from the time the experiment commenced are as follows:

Treatment	2.1	5.3	1.4	4.6	0.9
No Treatment	1.9	0.5	2.8	3.1	

At the 0.05 level of significance, can the serum be said to be effective? Assume the two populations to be normally distributed with equal variances.

10.36 Engineers at a large automobile manufacturing company are trying to decide whether to purchase brand A or brand B tires for the company's new models. To help them arrive at a decision, an experiment is conducted using 12 of each brand. The tires are run

until they wear out. The results are as follows:

$$\begin{aligned} \text{Brand } A: &\quad \bar{x}_1 = 37{,}900 \text{ kilometers}, \\ &\quad s_1 = 5100 \text{ kilometers}. \\ \text{Brand } B: &\quad \bar{x}_1 = 39{,}800 \text{ kilometers}, \\ &\quad s_2 = 5900 \text{ kilometers}. \end{aligned}$$

Test the hypothesis that there is no difference in the average wear of the two brands of tires. Assume the populations to be approximately normally distributed with equal variances. Use a P-value.

10.37 In Exercise 9.42 on page 295, test the hypothesis that the fuel economy of Volkswagen mini-trucks, on average, exceeds that of similarly equipped Toyota mini-trucks by 4 kilometers per liter. Use a 0.10 level of significance.

10.38 A UCLA researcher claims that the average life span of mice can be extended by as much as 8 months when the calories in their diet are reduced by approximately 40% from the time they are weaned. The restricted diets are enriched to normal levels by vitamins and protein. Suppose that a random sample of 10 mice is fed a normal diet and has an average life span of 32.1 months with a standard deviation of 3.2 months, while a random sample of 15 mice is fed the restricted diet and has an average life span of 37.6 months with a standard deviation of 2.8 months. Test the hypothesis, at the 0.05 level of significance, that the average life span of mice on this restricted diet is increased by 8 months against the alternative that the increase is less than 8 months. Assume the distributions of life spans for the regular and restricted diets are approximately normal with equal variances.

10.39 The following data represent the running times of films produced by two motion-picture companies:

Company	Time (minutes)						
1	102	86	98	109	92		
2	81	165	97	134	92	87	114

Test the hypothesis that the average running time of films produced by company 2 exceeds the average running time of films produced by company 1 by 10 minutes against the one-sided alternative that the difference is less than 10 minutes. Use a 0.1 level of significance and assume the distributions of times to be approximately normal with unequal variances.

10.40 In a study conducted at Virginia Tech, the plasma ascorbic acid levels of pregnant women were compared for smokers versus nonsmokers. Thirty-two women in the last three months of pregnancy, free of major health disorders and ranging in age from 15 to 32 years, were selected for the study. Prior to the collection of 20 ml of blood, the participants were told to avoid breakfast, forgo their vitamin supplements, and avoid foods high in ascorbic acid content. From the blood samples, the following plasma ascorbic acid values were determined, in milligrams per 100 milliliters:

Plasma Ascorbic Acid Values

Nonsmokers		Smokers
0.97	1.16	0.48
0.72	0.86	0.71
1.00	0.85	0.98
0.81	0.58	0.68
0.62	0.57	1.18
1.32	0.64	1.36
1.24	0.98	0.78
0.99	1.09	1.64
0.90	0.92	
0.74	0.78	
0.88	1.24	
0.94	1.18	

Is there sufficient evidence to conclude that there is a difference between plasma ascorbic acid levels of smokers and nonsmokers? Assume that the two sets of data came from normal populations with unequal variances. Use a P-value.

10.41 A study was conducted by the Department of Zoology at Virginia Tech to determine if there is a significant difference in the density of organisms at two different stations located on Cedar Run, a secondary stream in the Roanoke River drainage basin. Sewage from a sewage treatment plant and overflow from the Federal Mogul Corporation settling pond enter the stream near its headwaters. The following data give the density measurements, in number of organisms per square meter, at the two collecting stations:

Number of Organisms per Square Meter

Station 1		Station 2	
5030	4980	2800	2810
13,700	11,910	4670	1330
10,730	8130	6890	3320
11,400	26,850	7720	1230
860	17,660	7030	2130
2200	22,800	7330	2190
4250	1130		
15,040	1690		

Can we conclude, at the 0.05 level of significance, that the average densities at the two stations are equal? Assume that the observations come from normal populations with different variances.

10.42 Five samples of a ferrous-type substance were used to determine if there is a difference between a laboratory chemical analysis and an X-ray fluorescence analysis of the iron content. Each sample was split into two subsamples and the two types of analysis were applied. Following are the coded data showing the iron content analysis:

	Sample				
Analysis	**1**	**2**	**3**	**4**	**5**
X-ray	2.0	2.0	2.3	2.1	2.4
Chemical	2.2	1.9	2.5	2.3	2.4

Assuming that the populations are normal, test at the 0.05 level of significance whether the two methods of analysis give, on the average, the same result.

10.43 According to published reports, practice under fatigued conditions distorts mechanisms that govern performance. An experiment was conducted using 15 college males, who were trained to make a continuous horizontal right-to-left arm movement from a microswitch to a barrier, knocking over the barrier coincident with the arrival of a clock sweephand to the 6 o'clock position. The absolute value of the difference between the time, in milliseconds, that it took to knock over the barrier and the time for the sweephand to reach the 6 o'clock position (500 msec) was recorded. Each participant performed the task five times under prefatigue and postfatigue conditions, and the sums of the absolute differences for the five performances were recorded.

	Absolute Time Differences	
Subject	**Prefatigue**	**Postfatigue**
1	158	91
2	92	59
3	65	215
4	98	226
5	33	223
6	89	91
7	148	92
8	58	177
9	142	134
10	117	116
11	74	153
12	66	219
13	109	143
14	57	164
15	85	100

An increase in the mean absolute time difference when the task is performed under postfatigue conditions would support the claim that practice under fatigued conditions distorts mechanisms that govern performance. Assuming the populations to be normally distributed, test this claim.

10.44 In a study conducted by the Department of Human Nutrition and Foods at Virginia Tech, the following data were recorded on sorbic acid residuals, in parts per million, in ham immediately after dipping in a sorbate solution and after 60 days of storage:

	Sorbic Acid Residuals in Ham	
Slice	**Before Storage**	**After Storage**
1	224	116
2	270	96
3	400	239
4	444	329
5	590	437
6	660	597
7	1400	689
8	680	576

Assuming the populations to be normally distributed, is there sufficient evidence, at the 0.05 level of significance, to say that the length of storage influences sorbic acid residual concentrations?

10.45 A taxi company manager is trying to decide whether the use of radial tires instead of regular belted tires improves fuel economy. Twelve cars were equipped with radial tires and driven over a prescribed test course. Without changing drivers, the same cars were then equipped with regular belted tires and driven once again over the test course. The gasoline consumption, in kilometers per liter, was recorded as follows:

	Kilometers per Liter	
Car	**Radial Tires**	**Belted Tires**
1	4.2	4.1
2	4.7	4.9
3	6.6	6.2
4	7.0	6.9
5	6.7	6.8
6	4.5	4.4
7	5.7	5.7
8	6.0	5.8
9	7.4	6.9
10	4.9	4.7
11	6.1	6.0
12	5.2	4.9

Can we conclude that cars equipped with radial tires give better fuel economy than those equipped with belted tires? Assume the populations to be normally distributed. Use a P-value in your conclusion.

10.46 In Review Exercise 9.91 on page 313, use the t-distribution to test the hypothesis that the diet reduces a woman's weight by 4.5 kilograms on average against the alternative hypothesis that the mean difference in weight is less than 4.5 kilograms. Use a P-value.

10.47 How large a sample is required in Exercise 10.20 if the power of the test is to be 0.90 when the true mean is 5.20? Assume that $\sigma = 0.24$.

10.48 If the distribution of life spans in Exercise 10.19 is approximately normal, how large a sample is required in order that the probability of committing a type II error be 0.1 when the true mean is 35.9 months? Assume that $\sigma = 5.8$ months.

10.49 How large a sample is required in Exercise 10.24 if the power of the test is to be 0.95 when the true average height differs from 162.5 by 3.1 centimeters? Use $\alpha = 0.02$.

10.50 How large should the samples be in Exercise 10.31 if the power of the test is to be 0.95 when the true difference between thread types A and B is 8 kilograms?

10.51 How large a sample is required in Exercise 10.22 if the power of the test is to be 0.8 when the true mean meditation time exceeds the hypothesized value by 1.2σ? Use $\alpha = 0.05$.

10.52 For testing

$$H_0: \ \mu = 14,$$
$$H_1: \ \mu \neq 14,$$

an $\alpha = 0.05$ level t-test is being considered. What sample size is necessary in order for the probability to be 0.1 of falsely failing to reject H_0 when the true population mean differs from 14 by 0.5? From a preliminary sample we estimate σ to be 1.25.

10.53 A study was conducted at the Department of Veterinary Medicine at Virginia Tech to determine if the "strength" of a wound from surgical incision is affected by the temperature of the knife. Eight dogs were used in the experiment. "Hot" and "cold" incisions were made on the abdomen of each dog, and the strength was measured. The resulting data appear below.

Dog	Knife	Strength
1	Hot	5120
1	Cold	8200
2	Hot	10,000
2	Cold	8600
3	Hot	10,000
3	Cold	9200
4	Hot	10,000
4	Cold	6200
5	Hot	10,000
5	Cold	10,000
6	Hot	7900
6	Cold	5200
7	Hot	510
7	Cold	885
8	Hot	1020
8	Cold	460

(a) Write an appropriate hypothesis to determine if there is a significant difference in strength between the hot and cold incisions.

(b) Test the hypothesis using a paired t-test. Use a P-value in your conclusion.

10.54 Nine subjects were used in an experiment to determine if exposure to carbon monoxide has an impact on breathing capability. The data were collected by personnel in the Health and Physical Education Department at Virginia Tech and were analyzed in the Statistics Consulting Center at Hokie Land. The subjects were exposed to breathing chambers, one of which contained a high concentration of CO. Breathing frequency measures were made for each subject for each chamber. The subjects were exposed to the breathing chambers in random sequence. The data give the breathing frequency, in number of breaths taken per minute. Make a one-sided test of the hypothesis that mean breathing frequency is the same for the two environments. Use $\alpha = 0.05$. Assume that breathing frequency is approximately normal.

Subject	With CO	Without CO
1	30	30
2	45	40
3	26	25
4	25	23
5	34	30
6	51	49
7	46	41
8	32	35
9	30	28

10.8 One Sample: Test on a Single Proportion

Tests of hypotheses concerning proportions are required in many areas. Politicians are certainly interested in knowing what fraction of the voters will favor them in the next election. All manufacturing firms are concerned about the proportion of defective items when a shipment is made. Gamblers depend on a knowledge of the proportion of outcomes that they consider favorable.

We shall consider the problem of testing the hypothesis that the proportion of successes in a binomial experiment equals some specified value. That is, we are testing the null hypothesis H_0 that $p = p_0$, where p is the parameter of the binomial distribution. The alternative hypothesis may be one of the usual one-sided

10.10 One- and Two-Sample Tests Concerning Variances

In this section, we are concerned with testing hypotheses concerning population variances or standard deviations. Applications of one- and two-sample tests on variances are certainly not difficult to motivate. Engineers and scientists are confronted with studies in which they are required to demonstrate that measurements involving products or processes adhere to specifications set by consumers. The specifications are often met if the process variance is sufficiently small. Attention is also focused on comparative experiments between methods or processes, where inherent reproducibility or variability must formally be compared. In addition, to determine if the equal variance assumption is violated, a test comparing two variances is often applied prior to conducting a t-test on two means.

Let us first consider the problem of testing the null hypothesis H_0 that the population variance σ^2 equals a specified value σ_0^2 against one of the usual alternatives $\sigma^2 < \sigma_0^2$, $\sigma^2 > \sigma_0^2$, or $\sigma^2 \neq \sigma_0^2$. The appropriate statistic on which to base our decision is the chi-squared statistic of Theorem 8.4, which was used in Chapter 9 to construct a confidence interval for σ^2. Therefore, if we assume that the distribution of the population being sampled is normal, the chi-squared value for testing $\sigma^2 = \sigma_0^2$ is given by

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2},$$

where n is the sample size, s^2 is the sample variance, and σ_0^2 is the value of σ^2 given by the null hypothesis. If H_0 is true, χ^2 is a value of the chi-squared distribution with $v = n - 1$ degrees of freedom. Hence, for a two-tailed test at the α-level of significance, the critical region is $\chi^2 < \chi^2_{1-\alpha/2}$ or $\chi^2 > \chi^2_{\alpha/2}$. For the one-sided alternative $\sigma^2 < \sigma_0^2$, the critical region is $\chi^2 < \chi^2_{1-\alpha}$, and for the one-sided alternative $\sigma^2 > \sigma_0^2$, the critical region is $\chi^2 > \chi^2_{\alpha}$.

Robustness of χ^2-Test to Assumption of Normality

The reader may have discerned that various tests depend, at least theoretically, on the assumption of normality. In general, many procedures in applied statistics have theoretical underpinnings that depend on the normal distribution. These procedures vary in the degree of their dependency on the assumption of normality. A procedure that is reasonably insensitive to the assumption is called a **robust procedure** (i.e., robust to normality). The χ^2-test on a single variance is very nonrobust to normality (i.e., the practical success of the procedure depends on normality). As a result, the P-value computed may be appreciably different from the actual P-value if the population sampled is not normal. Indeed, it is quite feasible that a statistically significant P-value may not truly signal H_1: $\sigma \neq \sigma_0$; rather, a significant value may be a result of the violation of the normality assumptions. Therefore, the analyst should approach the use of this particular χ^2-test with caution.

Example 10.12: A manufacturer of car batteries claims that the life of the company's batteries is approximately normally distributed with a standard deviation equal to 0.9 year.

If a random sample of 10 of these batteries has a standard deviation of 1.2 years, do you think that $\sigma > 0.9$ year? Use a 0.05 level of significance.

Solution: 1. H_0: $\sigma^2 = 0.81$.

2. H_1: $\sigma^2 > 0.81$.

3. $\alpha = 0.05$.

4. Critical region: From Figure 10.19 we see that the null hypothesis is rejected when $\chi^2 > 16.919$, where $\chi^2 = \frac{(n-1)s^2}{\sigma_0^2}$, with $v = 9$ degrees of freedom.

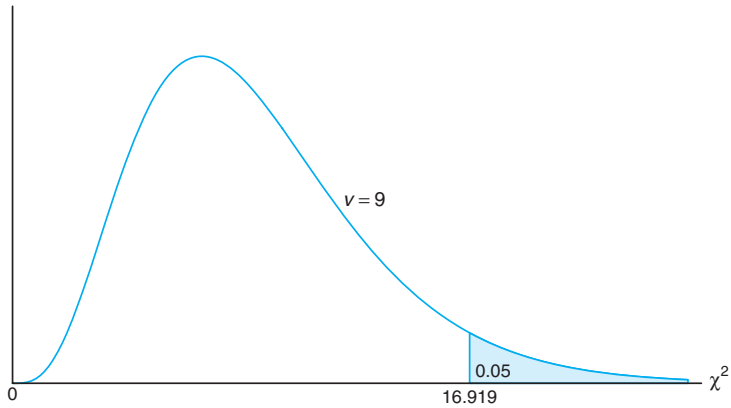


Figure 10.19: Critical region for the alternative hypothesis $\sigma > 0.9$.

5. Computations: $s^2 = 1.44$, $n = 10$, and

$$\chi^2 = \frac{(9)(1.44)}{0.81} = 16.0, \qquad P \approx 0.07.$$

6. Decision: The χ^2-statistic is not significant at the 0.05 level. However, based on the P-value 0.07, there is evidence that $\sigma > 0.9$.

Now let us consider the problem of testing the equality of the variances σ_1^2 and σ_2^2 of two populations. That is, we shall test the null hypothesis H_0 that $\sigma_1^2 = \sigma_2^2$ against one of the usual alternatives

$$\sigma_1^2 < \sigma_2^2, \quad \sigma_1^2 > \sigma_2^2, \quad \text{or} \quad \sigma_1^2 \neq \sigma_2^2.$$

For independent random samples of sizes n_1 and n_2, respectively, from the two populations, the ***f*-value for testing $\sigma_1^2 = \sigma_2^2$** is the ratio

$$f = \frac{s_1^2}{s_2^2},$$

where s_1^2 and s_2^2 are the variances computed from the two samples. If the two populations are approximately normally distributed and the null hypothesis is true, according to Theorem 8.8 the ratio $f = s_1^2/s_2^2$ is a value of the F-distribution with $v_1 = n_1 - 1$ and $v_2 = n_2 - 1$ degrees of freedom. Therefore, the critical regions

of size α corresponding to the one-sided alternatives $\sigma_1^2 < \sigma_2^2$ and $\sigma_1^2 > \sigma_2^2$ are, respectively, $f < f_{1-\alpha}(v_1, v_2)$ and $f > f_{\alpha}(v_1, v_2)$. For the two-sided alternative $\sigma_1^2 \neq \sigma_2^2$, the critical region is $f < f_{1-\alpha/2}(v_1, v_2)$ or $f > f_{\alpha/2}(v_1, v_2)$.

Example 10.13: In testing for the difference in the abrasive wear of the two materials in Example 10.6, we assumed that the two unknown population variances were equal. Were we justified in making this assumption? Use a 0.10 level of significance.

***Solution*:** Let σ_1^2 and σ_2^2 be the population variances for the abrasive wear of material 1 and material 2, respectively.

1. H_0: $\sigma_1^2 = \sigma_2^2$.
2. H_1: $\sigma_1^2 \neq \sigma_2^2$.
3. $\alpha = 0.10$.
4. Critical region: From Figure 10.20, we see that $f_{0.05}(11, 9) = 3.11$, and, by using Theorem 8.7, we find

$$f_{0.95}(11, 9) = \frac{1}{f_{0.05}(9, 11)} = 0.34.$$

   Therefore, the null hypothesis is rejected when $f < 0.34$ or $f > 3.11$, where $f = s_1^2/s_2^2$ with $v_1 = 11$ and $v_2 = 9$ degrees of freedom.
5. Computations: $s_1^2 = 16$, $s_2^2 = 25$, and hence $f = \frac{16}{25} = 0.64$.
6. Decision: Do not reject H_0. Conclude that there is insufficient evidence that the variances differ. ∎

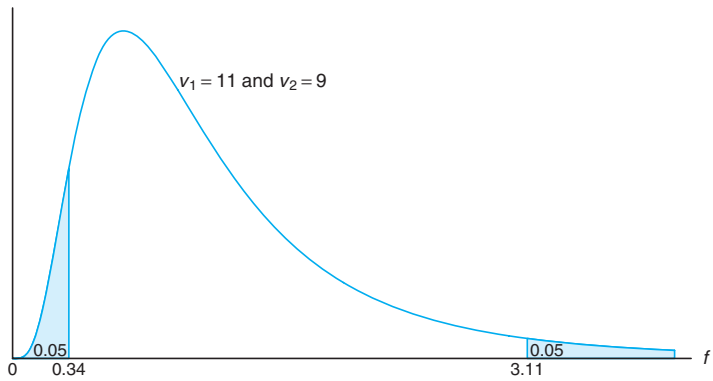


Figure 10.20: Critical region for the alternative hypothesis $\sigma_1^2 \neq \sigma_2^2$.

F-Test for Testing Variances in *SAS*

Figure 10.18 on page 356 displays the printout of a two-sample t-test where two means from the seedling data in Exercise 9.40 were compared. Box-and-whisker plots in Figure 10.17 on page 355 suggest that variances are not homogeneous, and thus the t'-statistic and its corresponding P-value are relevant. Note also that

the printout displays the F-statistic for H_0: $\sigma_1 = \sigma_2$ with a P-value of 0.0098, additional evidence that more variability is to be expected when nitrogen is used than under the no-nitrogen condition.

Exercises

10.67 The content of containers of a particular lubricant is known to be normally distributed with a variance of 0.03 liter. Test the hypothesis that $\sigma^2 = 0.03$ against the alternative that $\sigma^2 \neq 0.03$ for the random sample of 10 containers in Exercise 10.23 on page 356. Use a P-value in your conclusion.

10.68 Past experience indicates that the time required for high school seniors to complete a standardized test is a normal random variable with a standard deviation of 6 minutes. Test the hypothesis that $\sigma = 6$ against the alternative that $\sigma < 6$ if a random sample of the test times of 20 high school seniors has a standard deviation $s = 4.51$. Use a 0.05 level of significance.

10.69 Aflotoxins produced by mold on peanut crops in Virginia must be monitored. A sample of 64 batches of peanuts reveals levels of 24.17 ppm, on average, with a variance of 4.25 ppm. Test the hypothesis that $\sigma^2 = 4.2$ ppm against the alternative that $\sigma^2 \neq 4.2$ ppm. Use a P-value in your conclusion.

10.70 Past data indicate that the amount of money contributed by the working residents of a large city to a volunteer rescue squad is a normal random variable with a standard deviation of \$1.40. It has been suggested that the contributions to the rescue squad from just the employees of the sanitation department are much more variable. If the contributions of a random sample of 12 employees from the sanitation department have a standard deviation of \$1.75, can we conclude at the 0.01 level of significance that the standard deviation of the contributions of all sanitation workers is greater than that of all workers living in the city?

10.71 A soft-drink dispensing machine is said to be out of control if the variance of the contents exceeds 1.15 deciliters. If a random sample of 25 drinks from this machine has a variance of 2.03 deciliters, does this indicate at the 0.05 level of significance that the machine is out of control? Assume that the contents are approximately normally distributed.

10.72 Large-Sample Test of $\sigma^2 = \sigma_0^2$: When $n \geq 30$, we can test the null hypothesis that $\sigma^2 = \sigma_0^2$, or $\sigma = \sigma_0$, by computing

$$z = \frac{s - \sigma_0}{\sigma_0/\sqrt{2n}},$$

which is a value of a random variable whose sampling distribution is approximately the standard normal distribution.

(a) With reference to Example 10.4, test at the 0.05 level of significance whether $\sigma = 10.0$ years against the alternative that $\sigma \neq 10.0$ years.

(b) It is suspected that the variance of the distribution of distances in kilometers traveled on 5 liters of fuel by a new automobile model equipped with a diesel engine is less than the variance of the distribution of distances traveled by the same model equipped with a six-cylinder gasoline engine, which is known to be $\sigma^2 = 6.25$. If 72 test runs of the diesel model have a variance of 4.41, can we conclude at the 0.05 level of significance that the variance of the distances traveled by the diesel model is less than that of the gasoline model?

10.73 A study is conducted to compare the lengths of time required by men and women to assemble a certain product. Past experience indicates that the distribution of times for both men and women is approximately normal but the variance of the times for women is less than that for men. A random sample of times for 11 men and 14 women produced the following data:

Men	**Women**
$n_1 = 11$	$n_2 = 14$
$s_1 = 6.1$	$s_2 = 5.3$

Test the hypothesis that $\sigma_1^2 = \sigma_2^2$ against the alternative that $\sigma_1^2 > \sigma_2^2$. Use a P-value in your conclusion.

10.74 For Exercise 10.41 on page 358, test the hypothesis at the 0.05 level of significance that $\sigma_1^2 = \sigma_2^2$ against the alternative that $\sigma_1^2 \neq \sigma_2^2$, where σ_1^2 and σ_2^2 are the variances of the number of organisms per square meter of water at the two different locations on Cedar Run.

10.75 With reference to Exercise 10.39 on page 358, test the hypothesis that $\sigma_1^2 = \sigma_2^2$ against the alternative that $\sigma_1^2 \neq \sigma_2^2$, where σ_1^2 and σ_2^2 are the variances for the running times of films produced by company 1 and company 2, respectively. Use a P-value.

10.76 Two types of instruments for measuring the amount of sulfur monoxide in the atmosphere are being compared in an air-pollution experiment. Researchers

wish to determine whether the two types of instruments yield measurements having the same variability. The readings in the following table were recorded for the two instruments.

Sulfur Monoxide

Instrument *A*	Instrument *B*
0.86	0.87
0.82	0.74
0.75	0.63
0.61	0.55
0.89	0.76
0.64	0.70
0.81	0.69
0.68	0.57
0.65	0.53

Assuming the populations of measurements to be approximately normally distributed, test the hypothesis that $\sigma_A = \sigma_B$ against the alternative that $\sigma_A \neq \sigma_B$. Use a P-value.

10.77 An experiment was conducted to compare the alcohol content of soy sauce on two different production lines. Production was monitored eight times a day. The data are shown here.

```
Production line 1:
    0.48 0.39 0.42 0.52 0.40 0.48 0.52 0.52
Production line 2:
    0.38 0.37 0.39 0.41 0.38 0.39 0.40 0.39
```

Assume both populations are normal. It is suspected that production line 1 is not producing as consistently as production line 2 in terms of alcohol content. Test the hypothesis that $\sigma_1 = \sigma_2$ against the alternative that $\sigma_1 \neq \sigma_2$. Use a P-value.

10.78 Hydrocarbon emissions from cars are known to have decreased dramatically during the 1980s. A study was conducted to compare the hydrocarbon emissions at idling speed, in parts per million (ppm), for automobiles from 1980 and 1990. Twenty cars of each model year were randomly selected, and their hydrocarbon emission levels were recorded. The data are as follows:

```
1980 models:
141 359 247 940 882 494 306 210 105 880
200 223 188 940 241 190 300 435 241 380
1990 models:
140 160  20 20 223  60  20  95 360 70
220 400 217 58 235 380 200 175  85 65
```

Test the hypothesis that $\sigma_1 = \sigma_2$ against the alternative that $\sigma_1 \neq \sigma_2$. Assume both populations are normal. Use a P-value.

10.11 Goodness-of-Fit Test

Throughout this chapter, we have been concerned with the testing of statistical hypotheses about single population parameters such as μ, σ^2, and p. Now we shall consider a test to determine if a population has a specified theoretical distribution. The test is based on how good a fit we have between the frequency of occurrence of observations in an observed sample and the expected frequencies obtained from the hypothesized distribution.

To illustrate, we consider the tossing of a die. We hypothesize that the die is honest, which is equivalent to testing the hypothesis that the distribution of outcomes is the discrete uniform distribution

$$f(x) = \frac{1}{6}, \quad x = 1, 2, \ldots, 6.$$

Suppose that the die is tossed 120 times and each outcome is recorded. Theoretically, if the die is balanced, we would expect each face to occur 20 times. The results are given in Table 10.4.

Table 10.4: Observed and Expected Frequencies of 120 Tosses of a Die

Face:	**1**	**2**	**3**	**4**	**5**	**6**
Observed	20	22	17	18	19	24
Expected	20	20	20	20	20	20

10.49 How large a sample is required in Exercise 10.24 if the power of the test is to be 0.95 when the true average height differs from 162.5 by 3.1 centimeters? Use $\alpha = 0.02$.

10.50 How large should the samples be in Exercise 10.31 if the power of the test is to be 0.95 when the true difference between thread types A and B is 8 kilograms?

10.51 How large a sample is required in Exercise 10.22 if the power of the test is to be 0.8 when the true mean meditation time exceeds the hypothesized value by 1.2σ? Use $\alpha = 0.05$.

10.52 For testing

$$H_0: \ \mu = 14,$$
$$H_1: \ \mu \neq 14,$$

an $\alpha = 0.05$ level t-test is being considered. What sample size is necessary in order for the probability to be 0.1 of falsely failing to reject H_0 when the true population mean differs from 14 by 0.5? From a preliminary sample we estimate σ to be 1.25.

10.53 A study was conducted at the Department of Veterinary Medicine at Virginia Tech to determine if the "strength" of a wound from surgical incision is affected by the temperature of the knife. Eight dogs were used in the experiment. "Hot" and "cold" incisions were made on the abdomen of each dog, and the strength was measured. The resulting data appear below.

Dog	Knife	Strength
1	Hot	5120
1	Cold	8200
2	Hot	10,000
2	Cold	8600
3	Hot	10,000
3	Cold	9200
4	Hot	10,000
4	Cold	6200
5	Hot	10,000
5	Cold	10,000
6	Hot	7900
6	Cold	5200
7	Hot	510
7	Cold	885
8	Hot	1020
8	Cold	460

(a) Write an appropriate hypothesis to determine if there is a significant difference in strength between the hot and cold incisions.

(b) Test the hypothesis using a paired t-test. Use a P-value in your conclusion.

10.54 Nine subjects were used in an experiment to determine if exposure to carbon monoxide has an impact on breathing capability. The data were collected by personnel in the Health and Physical Education Department at Virginia Tech and were analyzed in the Statistics Consulting Center at Hokie Land. The subjects were exposed to breathing chambers, one of which contained a high concentration of CO. Breathing frequency measures were made for each subject for each chamber. The subjects were exposed to the breathing chambers in random sequence. The data give the breathing frequency, in number of breaths taken per minute. Make a one-sided test of the hypothesis that mean breathing frequency is the same for the two environments. Use $\alpha = 0.05$. Assume that breathing frequency is approximately normal.

Subject	With CO	Without CO
1	30	30
2	45	40
3	26	25
4	25	23
5	34	30
6	51	49
7	46	41
8	32	35
9	30	28

10.8 One Sample: Test on a Single Proportion

Tests of hypotheses concerning proportions are required in many areas. Politicians are certainly interested in knowing what fraction of the voters will favor them in the next election. All manufacturing firms are concerned about the proportion of defective items when a shipment is made. Gamblers depend on a knowledge of the proportion of outcomes that they consider favorable.

We shall consider the problem of testing the hypothesis that the proportion of successes in a binomial experiment equals some specified value. That is, we are testing the null hypothesis H_0 that $p = p_0$, where p is the parameter of the binomial distribution. The alternative hypothesis may be one of the usual one-sided

or two-sided alternatives:

$$p < p_0, \quad p > p_0, \quad \text{or} \quad p \neq p_0.$$

The appropriate random variable on which we base our decision criterion is the binomial random variable X, although we could just as well use the statistic $\hat{p} = X/n$. Values of X that are far from the mean $\mu = np_0$ will lead to the rejection of the null hypothesis. Because X is a discrete binomial variable, it is unlikely that a critical region can be established whose size is *exactly* equal to a prespecified value of α. For this reason it is preferable, in dealing with small samples, to base our decisions on P-values. To test the hypothesis

$$H_0\text{: } p = p_0,$$
$$H_1\text{: } p < p_0,$$

we use the binomial distribution to compute the P-value

$$P = P(X \leq x \text{ when } p = p_0).$$

The value x is the number of successes in our sample of size n. If this P-value is less than or equal to α, our test is significant at the α level and we reject H_0 in favor of H_1. Similarly, to test the hypothesis

$$H_0\text{: } p = p_0,$$
$$H_1\text{: } p > p_0,$$

at the α-level of significance, we compute

$$P = P(X \geq x \text{ when } p = p_0)$$

and reject H_0 in favor of H_1 if this P-value is less than or equal to α. Finally, to test the hypothesis

$$H_0\text{: } p = p_0,$$
$$H_1\text{: } p \neq p_0,$$

at the α-level of significance, we compute

$$P = 2P(X \leq x \text{ when } p = p_0) \qquad \text{if } x < np_0$$

or

$$P = 2P(X \geq x \text{ when } p = p_0) \qquad \text{if } x > np_0$$

and reject H_0 in favor of H_1 if the computed P-value is less than or equal to α.

The steps for testing a null hypothesis about a proportion against various alternatives using the binomial probabilities of Table A.1 are as follows:

Testing a Proportion (Small Samples)

1. H_0: $p = p_0$.
2. One of the alternatives H_1: $p < p_0$, $p > p_0$, or $p \neq p_0$.
3. Choose a level of significance equal to α.
4. Test statistic: Binomial variable X with $p = p_0$.
5. Computations: Find x, the number of successes, and compute the appropriate P-value.
6. Decision: Draw appropriate conclusions based on the P-value.

Example 10.9: A builder claims that heat pumps are installed in 70% of all homes being constructed today in the city of Richmond, Virginia. Would you agree with this claim if a random survey of new homes in this city showed that 8 out of 15 had heat pumps installed? Use a 0.10 level of significance.

Solution:

1. H_0: $p = 0.7$.
2. H_1: $p \neq 0.7$.
3. $\alpha = 0.10$.
4. Test statistic: Binomial variable X with $p = 0.7$ and $n = 15$.
5. Computations: $x = 8$ and $np_0 = (15)(0.7) = 10.5$. Therefore, from Table A.1, the computed P-value is

$$P = 2P(X \leq 8 \text{ when } p = 0.7) = 2\sum_{x=0}^{8} b(x; 15, 0.7) = 0.2622 > 0.10.$$

6. Decision: Do not reject H_0. Conclude that there is insufficient reason to doubt the builder's claim.

In Section 5.2, we saw that binomial probabilities can be obtained from the actual binomial formula or from Table A.1 when n is small. For large n, approximation procedures are required. When the hypothesized value p_0 is very close to 0 or 1, the Poisson distribution, with parameter $\mu = np_0$, may be used. However, the normal curve approximation, with parameters $\mu = np_0$ and $\sigma^2 = np_0q_0$, is usually preferred for large n and is very accurate as long as p_0 is not extremely close to 0 or to 1. If we use the normal approximation, the ***z*-value for testing $p = p_0$** is given by

$$z = \frac{x - np_0}{\sqrt{np_0q_0}} = \frac{\hat{p} - p_0}{\sqrt{p_0q_0/n}},$$

which is a value of the standard normal variable Z. Hence, for a two-tailed test at the α-level of significance, the critical region is $z < -z_{\alpha/2}$ or $z > z_{\alpha/2}$. For the one-sided alternative $p < p_0$, the critical region is $z < -z_\alpha$, and for the alternative $p > p_0$, the critical region is $z > z_\alpha$.

Example 10.10: A commonly prescribed drug for relieving nervous tension is believed to be only 60% effective. Experimental results with a new drug administered to a random sample of 100 adults who were suffering from nervous tension show that 70 received relief. Is this sufficient evidence to conclude that the new drug is superior to the one commonly prescribed? Use a 0.05 level of significance.

Solution:

1. H_0: $p = 0.6$.
2. H_1: $p > 0.6$.
3. $\alpha = 0.05$.
4. Critical region: $z > 1.645$.

5. Computations: $x = 70$, $n = 100$, $\hat{p} = 70/100 = 0.7$, and

$$z = \frac{0.7 - 0.6}{\sqrt{(0.6)(0.4)/100}} = 2.04, \quad P = P(Z > 2.04) < 0.0207.$$

6. Decision: Reject H_0 and conclude that the new drug is superior. ◢

10.9 Two Samples: Tests on Two Proportions

Situations often arise where we wish to test the hypothesis that two proportions are equal. For example, we might want to show evidence that the proportion of doctors who are pediatricians in one state is equal to the proportion in another state. A person may decide to give up smoking only if he or she is convinced that the proportion of smokers with lung cancer exceeds the proportion of nonsmokers with lung cancer.

In general, we wish to test the null hypothesis that two proportions, or binomial parameters, are equal. That is, we are testing $p_1 = p_2$ against one of the alternatives $p_1 < p_2$, $p_1 > p_2$, or $p_1 \neq p_2$. Of course, this is equivalent to testing the null hypothesis that $p_1 - p_2 = 0$ against one of the alternatives $p_1 - p_2 < 0$, $p_1 - p_2 > 0$, or $p_1 - p_2 \neq 0$. The statistic on which we base our decision is the random variable $\widehat{P}_1 - \widehat{P}_2$. Independent samples of sizes n_1 and n_2 are selected at random from two binomial populations and the proportions of successes $\widehat{P}_1$ and $\widehat{P}_2$ for the two samples are computed.

In our construction of confidence intervals for p_1 and p_2 we noted, for n_1 and n_2 sufficiently large, that the point estimator $\widehat{P}_1$ minus $\widehat{P}_2$ was approximately normally distributed with mean

$$\mu_{\widehat{P}_1 - \widehat{P}_2} = p_1 - p_2$$

and variance

$$\sigma^2_{\widehat{P}_1 - \widehat{P}_2} = \frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}.$$

Therefore, our critical region(s) can be established by using the standard normal variable

$$Z = \frac{(\widehat{P}_1 - \widehat{P}_2) - (p_1 - p_2)}{\sqrt{p_1 q_1/n_1 + p_2 q_2/n_2}}.$$

When H_0 is true, we can substitute $p_1 = p_2 = p$ and $q_1 = q_2 = q$ (where p and q are the common values) in the preceding formula for Z to give the form

$$Z = \frac{\widehat{P}_1 - \widehat{P}_2}{\sqrt{pq(1/n_1 + 1/n_2)}}.$$

To compute a value of Z, however, we must estimate the parameters p and q that appear in the radical. Upon pooling the data from both samples, the **pooled estimate of the proportion p** is

$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2},$$

where x_1 and x_2 are the numbers of successes in each of the two samples. Substituting $\hat{p}$ for p and $\hat{q} = 1 - \hat{p}$ for q, the ***z*-value for testing $p_1 = p_2$** is determined from the formula

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}\hat{q}(1/n_1 + 1/n_2)}}.$$

The critical regions for the appropriate alternative hypotheses are set up as before, using critical points of the standard normal curve. Hence, for the alternative $p_1 \neq p_2$ at the α-level of significance, the critical region is $z < -z_{\alpha/2}$ or $z > z_{\alpha/2}$. For a test where the alternative is $p_1 < p_2$, the critical region is $z < -z_\alpha$, and when the alternative is $p_1 > p_2$, the critical region is $z > z_\alpha$.

Example 10.11: A vote is to be taken among the residents of a town and the surrounding county to determine whether a proposed chemical plant should be constructed. The construction site is within the town limits, and for this reason many voters in the county believe that the proposal will pass because of the large proportion of town voters who favor the construction. To determine if there is a significant difference in the proportions of town voters and county voters favoring the proposal, a poll is taken. If 120 of 200 town voters favor the proposal and 240 of 500 county residents favor it, would you agree that the proportion of town voters favoring the proposal is higher than the proportion of county voters? Use an $\alpha = 0.05$ level of significance.

***Solution*:** Let p_1 and p_2 be the true proportions of voters in the town and county, respectively, favoring the proposal.

1. H_0: $p_1 = p_2$.
2. H_1: $p_1 > p_2$.
3. $\alpha = 0.05$.
4. Critical region: $z > 1.645$.
5. Computations:

$$\hat{p}_1 = \frac{x_1}{n_1} = \frac{120}{200} = 0.60, \quad \hat{p}_2 = \frac{x_2}{n_2} = \frac{240}{500} = 0.48, \quad \text{and}$$
$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{120 + 240}{200 + 500} = 0.51.$$

Therefore,

$$z = \frac{0.60 - 0.48}{\sqrt{(0.51)(0.49)(1/200 + 1/500)}} = 2.9,$$
$$P = P(Z > 2.9) = 0.0019.$$

6. Decision: Reject H_0 and agree that the proportion of town voters favoring the proposal is higher than the proportion of county voters.

Exercises

10.55 A marketing expert for a pasta-making company believes that 40% of pasta lovers prefer lasagna. If 9 out of 20 pasta lovers choose lasagna over other pastas, what can be concluded about the expert's claim? Use a 0.05 level of significance.

10.56 Suppose that, in the past, 40% of all adults favored capital punishment. Do we have reason to believe that the proportion of adults favoring capital punishment has increased if, in a random sample of 15 adults, 8 favor capital punishment? Use a 0.05 level of significance.

10.57 A new radar device is being considered for a certain missile defense system. The system is checked by experimenting with aircraft in which a kill or a no kill is simulated. If, in 300 trials, 250 kills occur, accept or reject, at the 0.04 level of significance, the claim that the probability of a kill with the new system does not exceed the 0.8 probability of the existing device.

10.58 It is believed that at least 60% of the residents in a certain area favor an annexation suit by a neighboring city. What conclusion would you draw if only 110 in a sample of 200 voters favored the suit? Use a 0.05 level of significance.

10.59 A fuel oil company claims that one-fifth of the homes in a certain city are heated by oil. Do we have reason to believe that fewer than one-fifth are heated by oil if, in a random sample of 1000 homes in this city, 136 are heated by oil? Use a P-value in your conclusion.

10.60 At a certain college, it is estimated that at most 25% of the students ride bicycles to class. Does this seem to be a valid estimate if, in a random sample of 90 college students, 28 are found to ride bicycles to class? Use a 0.05 level of significance.

10.61 In a winter of an epidemic flu, the parents of 2000 babies were surveyed by researchers at a well-known pharmaceutical company to determine if the company's new medicine was effective after two days. Among 120 babies who had the flu and were given the medicine, 29 were cured within two days. Among 280 babies who had the flu but were not given the medicine, 56 recovered within two days. Is there any significant indication that supports the company's claim of the effectiveness of the medicine?

10.62 In a controlled laboratory experiment, scientists at the University of Minnesota discovered that 25% of a certain strain of rats subjected to a 20% coffee bean diet and then force-fed a powerful cancer-causing chemical later developed cancerous tumors. Would we have reason to believe that the proportion of rats developing tumors when subjected to this diet has increased if the experiment were repeated and 16 of 48 rats developed tumors? Use a 0.05 level of significance.

10.63 In a study to estimate the proportion of residents in a certain city and its suburbs who favor the construction of a nuclear power plant, it is found that 63 of 100 urban residents favor the construction while only 59 of 125 suburban residents are in favor. Is there a significant difference between the proportions of urban and suburban residents who favor construction of the nuclear plant? Make use of a P-value.

10.64 In a study on the fertility of married women conducted by Martin O'Connell and Carolyn C. Rogers for the Census Bureau in 1979, two groups of childless wives aged 25 to 29 were selected at random, and each was asked if she eventually planned to have a child. One group was selected from among wives married less than two years and the other from among wives married five years. Suppose that 240 of the 300 wives married less than two years planned to have children some day compared to 288 of the 400 wives married five years. Can we conclude that the proportion of wives married less than two years who planned to have children is significantly higher than the proportion of wives married five years? Make use of a P-value.

10.65 An urban community would like to show that the incidence of breast cancer is higher in their area than in a nearby rural area. (PCB levels were found to be higher in the soil of the urban community.) If it is found that 20 of 200 adult women in the urban community have breast cancer and 10 of 150 adult women in the rural community have breast cancer, can we conclude at the 0.05 level of significance that breast cancer is more prevalent in the urban community?

10.66 Group Project: The class should be divided into pairs of students for this project. Suppose it is conjectured that at least 25% of students at your university exercise for more than two hours a week. Collect data from a random sample of 50 students. Ask each student if he or she works out for at least two hours per week. Then do the computations that allow either rejection or nonrejection of the above conjecture. Show all work and quote a P-value in your conclusion.

wish to determine whether the two types of instruments yield measurements having the same variability. The readings in the following table were recorded for the two instruments.

Sulfur Monoxide

Instrument *A*	Instrument *B*
0.86	0.87
0.82	0.74
0.75	0.63
0.61	0.55
0.89	0.76
0.64	0.70
0.81	0.69
0.68	0.57
0.65	0.53

Assuming the populations of measurements to be approximately normally distributed, test the hypothesis that $\sigma_A = \sigma_B$ against the alternative that $\sigma_A \neq \sigma_B$. Use a P-value.

10.77 An experiment was conducted to compare the alcohol content of soy sauce on two different production lines. Production was monitored eight times a day. The data are shown here.

```
Production line 1:
    0.48 0.39 0.42 0.52 0.40 0.48 0.52 0.52
Production line 2:
    0.38 0.37 0.39 0.41 0.38 0.39 0.40 0.39
```

Assume both populations are normal. It is suspected that production line 1 is not producing as consistently as production line 2 in terms of alcohol content. Test the hypothesis that $\sigma_1 = \sigma_2$ against the alternative that $\sigma_1 \neq \sigma_2$. Use a P-value.

10.78 Hydrocarbon emissions from cars are known to have decreased dramatically during the 1980s. A study was conducted to compare the hydrocarbon emissions at idling speed, in parts per million (ppm), for automobiles from 1980 and 1990. Twenty cars of each model year were randomly selected, and their hydrocarbon emission levels were recorded. The data are as follows:

```
1980 models:
141 359 247 940 882 494 306 210 105 880
200 223 188 940 241 190 300 435 241 380
1990 models:
140 160  20 20 223  60  20  95 360 70
220 400 217 58 235 380 200 175  85 65
```

Test the hypothesis that $\sigma_1 = \sigma_2$ against the alternative that $\sigma_1 \neq \sigma_2$. Assume both populations are normal. Use a P-value.

10.11 Goodness-of-Fit Test

Throughout this chapter, we have been concerned with the testing of statistical hypotheses about single population parameters such as μ, σ^2, and p. Now we shall consider a test to determine if a population has a specified theoretical distribution. The test is based on how good a fit we have between the frequency of occurrence of observations in an observed sample and the expected frequencies obtained from the hypothesized distribution.

To illustrate, we consider the tossing of a die. We hypothesize that the die is honest, which is equivalent to testing the hypothesis that the distribution of outcomes is the discrete uniform distribution

$$f(x) = \frac{1}{6}, \quad x = 1, 2, \ldots, 6.$$

Suppose that the die is tossed 120 times and each outcome is recorded. Theoretically, if the die is balanced, we would expect each face to occur 20 times. The results are given in Table 10.4.

Table 10.4: Observed and Expected Frequencies of 120 Tosses of a Die

Face:	**1**	**2**	**3**	**4**	**5**	**6**
Observed	20	22	17	18	19	24
Expected	20	20	20	20	20	20

By comparing the observed frequencies with the corresponding expected frequencies, we must decide whether these discrepancies are likely to occur as a result of sampling fluctuations and the die is balanced or whether the die is not honest and the distribution of outcomes is not uniform. It is common practice to refer to each possible outcome of an experiment as a cell. In our illustration, we have 6 cells. The appropriate statistic on which we base our decision criterion for an experiment involving k cells is defined by the following.

A **goodness-of-fit test** between observed and expected frequencies is based on the quantity

Goodness-of-Fit Test

$$\chi^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i},$$

where χ^2 is a value of a random variable whose sampling distribution is approximated very closely by the chi-squared distribution with $v = k - 1$ degrees of freedom. The symbols o_i and e_i represent the observed and expected frequencies, respectively, for the ith cell.

The number of degrees of freedom associated with the chi-squared distribution used here is equal to $k - 1$, since there are only $k - 1$ freely determined cell frequencies. That is, once $k - 1$ cell frequencies are determined, so is the frequency for the kth cell.

If the observed frequencies are close to the corresponding expected frequencies, the χ^2-value will be small, indicating a good fit. If the observed frequencies differ considerably from the expected frequencies, the χ^2-value will be large and the fit is poor. A good fit leads to the acceptance of H_0, whereas a poor fit leads to its rejection. The critical region will, therefore, fall in the right tail of the chi-squared distribution. For a level of significance equal to α, we find the critical value χ^2_α from Table A.5, and then $\chi^2 > \chi^2_\alpha$ constitutes the critical region. **The decision criterion described here should not be used unless each of the expected frequencies is at least equal to 5**. This restriction may require the combining of adjacent cells, resulting in a reduction in the number of degrees of freedom.

From Table 10.4, we find the χ^2-value to be

$$\chi^2 = \frac{(20-20)^2}{20} + \frac{(22-20)^2}{20} + \frac{(17-20)^2}{20} + \frac{(18-20)^2}{20} + \frac{(19-20)^2}{20} + \frac{(24-20)^2}{20} = 1.7.$$

Using Table A.5, we find $\chi^2_{0.05} = 11.070$ for $v = 5$ degrees of freedom. Since 1.7 is less than the critical value, we fail to reject H_0. We conclude that there is insufficient evidence that the die is not balanced.

As a second illustration, let us test the hypothesis that the frequency distribution of battery lives given in Table 1.7 on page 23 may be approximated by a normal distribution with mean $\mu = 3.5$ and standard deviation $\sigma = 0.7$. The expected frequencies for the 7 classes (cells), listed in Table 10.5, are obtained by computing the areas under the hypothesized normal curve that fall between the various class boundaries.

Table 10.5: Observed and Expected Frequencies of Battery Lives, Assuming Normality

Class Boundaries	o_i		e_i	
1.45–1.95	2	7	0.5	8.5
1.95–2.45	1		2.1	
2.45–2.95	4		5.9	
2.95–3.45	15		10.3	
3.45–3.95	10		10.7	
3.95–4.45	5	8	7.0	10.5
4.45–4.95	3		3.5	

For example, the z-values corresponding to the boundaries of the fourth class are

$$z_1 = \frac{2.95 - 3.5}{0.7} = -0.79 \quad \text{and} \quad z_2 = \frac{3.45 - 3.5}{0.7} = -0.07.$$

From Table A.3 we find the area between $z_1 = -0.79$ and $z_2 = -0.07$ to be

$$\begin{aligned} \text{area} = P(-0.79 < Z < -0.07) &= P(Z < -0.07) - P(Z < -0.79) \\ &= 0.4721 - 0.2148 = 0.2573. \end{aligned}$$

Hence, the expected frequency for the fourth class is

$$e_4 = (0.2573)(40) = 10.3.$$

It is customary to round these frequencies to one decimal.

The expected frequency for the first class interval is obtained by using the total area under the normal curve to the left of the boundary 1.95. For the last class interval, we use the total area to the right of the boundary 4.45. All other expected frequencies are determined by the method described for the fourth class. Note that we have combined adjacent classes in Table 10.5 where the expected frequencies are less than 5 (a rule of thumb in the goodness-of-fit test). Consequently, the total number of intervals is reduced from 7 to 4, resulting in $v = 3$ degrees of freedom. The χ^2-value is then given by

$$\chi^2 = \frac{(7 - 8.5)^2}{8.5} + \frac{(15 - 10.3)^2}{10.3} + \frac{(10 - 10.7)^2}{10.7} + \frac{(8 - 10.5)^2}{10.5} = 3.05.$$

Since the computed χ^2-value is less than $\chi^2_{0.05} = 7.815$ for 3 degrees of freedom, we have no reason to reject the null hypothesis and conclude that the normal distribution with $\mu = 3.5$ and $\sigma = 0.7$ provides a good fit for the distribution of battery lives.

The chi-squared goodness-of-fit test is an important resource, particularly since so many statistical procedures in practice depend, in a theoretical sense, on the assumption that the data gathered come from a specific type of distribution. As we have already seen, the normality assumption is often made. In the chapters that follow, we shall continue to make normality assumptions in order to provide a theoretical basis for certain tests and confidence intervals.

There are tests in the literature that are more powerful than the chi-squared test for testing normality. One such test is called **Geary's test**. This test is based on a very simple statistic which is a ratio of two estimators of the population standard deviation σ. Suppose that a random sample $X_1, X_2, \ldots, X_n$ is taken from a normal distribution, $N(\mu, \sigma)$. Consider the ratio

$$U = \frac{\sqrt{\pi/2}\sum_{i=1}^{n}|X_i - \bar{X}|/n}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2/n}}.$$

The reader should recognize that the denominator is a reasonable estimator of σ whether the distribution is normal or not. The numerator is a good estimator of σ if the distribution is normal but may overestimate or underestimate σ when there are departures from normality. Thus, values of U differing considerably from 1.0 represent the signal that the hypothesis of normality should be rejected.

For large samples, a reasonable test is based on approximate normality of U. The test statistic is then a standardization of U, given by

$$Z = \frac{U - 1}{0.2661/\sqrt{n}}.$$

Of course, the test procedure involves the two-sided critical region. We compute a value of z from the data and do not reject the hypothesis of normality when

$$-z_{\alpha/2} < Z < z_{\alpha/2}.$$

A paper dealing with Geary's test is cited in the Bibliography (Geary, 1947).

10.12 Test for Independence (Categorical Data)

The chi-squared test procedure discussed in Section 10.11 can also be used to test the hypothesis of independence of two variables of classification. Suppose that we wish to determine whether the opinions of the voting residents of the state of Illinois concerning a new tax reform are independent of their levels of income. Members of a random sample of 1000 registered voters from the state of Illinois are classified as to whether they are in a low, medium, or high income bracket and whether or not they favor the tax reform. The observed frequencies are presented in Table 10.6, which is known as a **contingency table**.

Table 10.6: 2 × 3 Contingency Table

	Income Level			
Tax Reform	**Low**	**Medium**	**High**	**Total**
For	182	213	203	598
Against	154	138	110	402
Total	336	351	313	1000

A contingency table with r rows and c columns is referred to as an $r \times c$ table ("$r \times c$" is read "r by c"). The row and column totals in Table 10.6 are called **marginal frequencies**. Our decision to accept or reject the null hypothesis, H_0, of independence between a voter's opinion concerning the tax reform and his or her level of income is based upon how good a fit we have between the observed frequencies in each of the 6 cells of Table 10.6 and the frequencies that we would expect for each cell under the assumption that H_0 is true. To find these expected frequencies, let us define the following events:

L: A person selected is in the low-income level.
M: A person selected is in the medium-income level.
H: A person selected is in the high-income level.
F: A person selected is for the tax reform.
A: A person selected is against the tax reform.

By using the marginal frequencies, we can list the following probability estimates:

$$P(L) = \frac{336}{1000}, \qquad P(M) = \frac{351}{1000}, \qquad P(H) = \frac{313}{1000},$$
$$P(F) = \frac{598}{1000}, \qquad P(A) = \frac{402}{1000}.$$

Now, if H_0 is true and the two variables are independent, we should have

$$P(L \cap F) = P(L)P(F) = \left(\frac{336}{1000}\right)\left(\frac{598}{1000}\right),$$
$$P(L \cap A) = P(L)P(A) = \left(\frac{336}{1000}\right)\left(\frac{402}{1000}\right),$$

$$P(M \cap F) = P(M)P(F) = \left(\frac{351}{1000}\right)\left(\frac{598}{1000}\right),$$
$$P(M \cap A) = P(M)P(A) = \left(\frac{351}{1000}\right)\left(\frac{402}{1000}\right),$$
$$P(H \cap F) = P(H)P(F) = \left(\frac{313}{1000}\right)\left(\frac{598}{1000}\right),$$
$$P(H \cap A) = P(H)P(A) = \left(\frac{313}{1000}\right)\left(\frac{402}{1000}\right).$$

The expected frequencies are obtained by multiplying each cell probability by the total number of observations. As before, we round these frequencies to one decimal. Thus, the expected number of low-income voters in our sample who favor the tax reform is estimated to be

$$\left(\frac{336}{1000}\right)\left(\frac{598}{1000}\right)(1000) = \frac{(336)(598)}{1000} = 200.9$$

when H_0 is true. The general rule for obtaining the expected frequency of any cell is given by the following formula:

$$\text{expected frequency} = \frac{(\text{column total}) \times (\text{row total})}{\text{grand total}}.$$

The expected frequency for each cell is recorded in parentheses beside the actual observed value in Table 10.7. Note that the expected frequencies in any row or column add up to the appropriate marginal total. In our example, we need to compute only two expected frequencies in the top row of Table 10.7 and then find the others by subtraction. The number of degrees of freedom associated with the chi-squared test used here is equal to the number of cell frequencies that may be filled in freely when we are given the marginal totals and the grand total, and in this illustration that number is 2. A simple formula providing the correct number of degrees of freedom is

$$v = (r-1)(c-1).$$

Table 10.7: Observed and Expected Frequencies

	Income Level			
Tax Reform	**Low**	**Medium**	**High**	**Total**
For	182 (200.9)	213 (209.9)	203 (187.2)	598
Against	154 (135.1)	138 (141.1)	110 (125.8)	402
Total	336	351	313	1000

Hence, for our example, $v = (2-1)(3-1) = 2$ degrees of freedom. To test the null hypothesis of independence, we use the following decision criterion.

Test for Independence

Calculate

$$\chi^2 = \sum_i \frac{(o_i - e_i)^2}{e_i},$$

where the summation extends over all rc cells in the $r \times c$ contingency table.

If $\chi^2 > \chi^2_\alpha$ with $v = (r-1)(c-1)$ degrees of freedom, reject the null hypothesis of independence at the α-level of significance; otherwise, fail to reject the null hypothesis.

Applying this criterion to our example, we find that

$$\chi^2 = \frac{(182-200.9)^2}{200.9} + \frac{(213-209.9)^2}{209.9} + \frac{(203-187.2)^2}{187.2} + \frac{(154-135.1)^2}{135.1} + \frac{(138-141.1)^2}{141.1} + \frac{(110-125.8)^2}{125.8} = 7.85,$$

$P \approx 0.02$.

From Table A.5 we find that $\chi^2_{0.05} = 5.991$ for $v = (2-1)(3-1) = 2$ degrees of freedom. The null hypothesis is rejected and we conclude that a voter's opinion concerning the tax reform and his or her level of income are not independent.

It is important to remember that the statistic on which we base our decision has a distribution that is only approximated by the chi-squared distribution. The computed χ^2-values depend on the cell frequencies and consequently are discrete. The continuous chi-squared distribution seems to approximate the discrete sampling distribution of χ^2 very well, provided that the number of degrees of freedom is greater than 1. In a 2×2 contingency table, where we have only 1 degree of freedom, a correction called **Yates' correction for continuity** is applied. The corrected formula then becomes

$$\chi^2(\text{corrected}) = \sum_i \frac{(|o_i - e_i| - 0.5)^2}{e_i}.$$

If the expected cell frequencies are large, the corrected and uncorrected results are almost the same. When the expected frequencies are between 5 and 10, Yates' correction should be applied. For expected frequencies less than 5, the Fisher-Irwin exact test should be used. A discussion of this test may be found in *Basic Concepts of Probability and Statistics* by Hodges and Lehmann (2005; see the Bibliography). The Fisher-Irwin test may be avoided, however, by choosing a larger sample.

10.13 Test for Homogeneity

When we tested for independence in Section 10.12, a random sample of 1000 voters was selected and the row and column totals for our contingency table were determined by chance. Another type of problem for which the method of Section 10.12 applies is one in which either the row or column totals are predetermined. Suppose, for example, that we decide in advance to select 200 Democrats, 150 Republicans, and 150 Independents from the voters of the state of North Carolina and record whether they are for a proposed abortion law, against it, or undecided. The observed responses are given in Table 10.8.

Table 10.8: Observed Frequencies

	Political Affiliation			
Abortion Law	**Democrat**	**Republican**	**Independent**	**Total**
For	82	70	62	214
Against	93	62	67	222
Undecided	25	18	21	64
Total	200	150	150	500

Now, rather than test for independence, we test the hypothesis that the population proportions within each row are the same. That is, we test the hypothesis that the proportions of Democrats, Republicans, and Independents favoring the abortion law are the same; the proportions of each political affiliation against the law are the same; and the proportions of each political affiliation that are undecided are the same. We are basically interested in determining whether the three categories of voters are **homogeneous** with respect to their opinions concerning the proposed abortion law. Such a test is called a test for homogeneity.

Assuming homogeneity, we again find the expected cell frequencies by multiplying the corresponding row and column totals and then dividing by the grand

total. The analysis then proceeds using the same chi-squared statistic as before. We illustrate this process for the data of Table 10.8 in the following example.

Example 10.14: Referring to the data of Table 10.8, test the hypothesis that opinions concerning the proposed abortion law are the same within each political affiliation. Use a 0.05 level of significance.

***Solution*:**

1. H_0: For each opinion, the proportions of Democrats, Republicans, and Independents are the same.
2. H_1: For at least one opinion, the proportions of Democrats, Republicans, and Independents are not the same.
3. $\alpha = 0.05$.
4. Critical region: $\chi^2 > 9.488$ with $v = 4$ degrees of freedom.
5. Computations: Using the expected cell frequency formula on page 375, we need to compute 4 cell frequencies. All other frequencies are found by subtraction. The observed and expected cell frequencies are displayed in Table 10.9.

Table 10.9: Observed and Expected Frequencies

	Political Affiliation			
Abortion Law	**Democrat**	**Republican**	**Independent**	**Total**
For	82 (85.6)	70 (64.2)	62 (64.2)	214
Against	93 (88.8)	62 (66.6)	67 (66.6)	222
Undecided	25 (25.6)	18 (19.2)	21 (19.2)	64
Total	200	150	150	500

Now,

$$\begin{aligned}\chi^2 &= \frac{(82-85.6)^2}{85.6} + \frac{(70-64.2)^2}{64.2} + \frac{(62-64.2)^2}{64.2} \\ &\quad + \frac{(93-88.8)^2}{88.8} + \frac{(62-66.6)^2}{66.6} + \frac{(67-66.6)^2}{66.6} \\ &\quad + \frac{(25-25.6)^2}{25.6} + \frac{(18-19.2)^2}{19.2} + \frac{(21-19.2)^2}{19.2} \\ &= 1.53.\end{aligned}$$

6. Decision: Do not reject H_0. There is insufficient evidence to conclude that the proportions of Democrats, Republicans, and Independents differ for each stated opinion.

Testing for Several Proportions

The chi-squared statistic for testing for homogeneity is also applicable when testing the hypothesis that k binomial parameters have the same value. This is, therefore, an extension of the test presented in Section 10.9 for determining differences between two proportions to a test for determining differences among k proportions. Hence, we are interested in testing the null hypothesis

$$H_0 : \ p_1 = p_2 = \cdots = p_k$$

against the alternative hypothesis, H_1, that the population proportions are *not all equal*. To perform this test, we first observe independent random samples of size $n_1, n_2, \ldots, n_k$ from the k populations and arrange the data in a $2 \times k$ contingency table, Table 10.10.

Table 10.10: k Independent Binomial Samples

Sample:	**1**	**2**	$\cdots$	$\boldsymbol{k}$
Successes	x_1	x_2	$\cdots$	x_k
Failures	$n_1 - x_1$	$n_2 - x_2$	$\cdots$	$n_k - x_k$

Depending on whether the sizes of the random samples were predetermined or occurred at random, the test procedure is identical to the test for homogeneity or the test for independence. Therefore, the expected cell frequencies are calculated as before and substituted, together with the observed frequencies, into the chi-squared statistic

$$\chi^2 = \sum_i \frac{(o_i - e_i)^2}{e_i},$$

with

$$v = (2-1)(k-1) = k-1$$

degrees of freedom.

By selecting the appropriate upper-tail critical region of the form $\chi^2 > \chi^2_\alpha$, we can now reach a decision concerning H_0.

Example 10.15: In a shop study, a set of data was collected to determine whether or not the proportion of defectives produced was the same for workers on the day, evening, and night shifts. The data collected are shown in Table 10.11.

Table 10.11: Data for Example 10.15

Shift:	**Day**	**Evening**	**Night**
Defectives	45	55	70
Nondefectives	905	890	870

Use a 0.025 level of significance to determine if the proportion of defectives is the same for all three shifts.

***Solution*:** Let p_1, p_2, and p_3 represent the true proportions of defectives for the day, evening, and night shifts, respectively.

1. H_0: $p_1 = p_2 = p_3$.
2. H_1: p_1, p_2, and p_3 are not all equal.
3. $\alpha = 0.025$.
4. Critical region: $\chi^2 > 7.378$ for $v = 2$ degrees of freedom.

5. Computations: Corresponding to the observed frequencies $o_1 = 45$ and $o_2 = 55$, we find

$$e_1 = \frac{(950)(170)}{2835} = 57.0 \quad \text{and} \quad e_2 = \frac{(945)(170)}{2835} = 56.7.$$

All other expected frequencies are found by subtraction and are displayed in Table 10.12.

Table 10.12: Observed and Expected Frequencies

Shift:	**Day**	**Evening**	**Night**	**Total**
Defectives	45 (57.0)	55 (56.7)	70 (56.3)	170
Nondefectives	905 (893.0)	890 (888.3)	870 (883.7)	2665
Total	950	945	940	2835

Now

$$\chi^2 = \frac{(45-57.0)^2}{57.0} + \frac{(55-56.7)^2}{56.7} + \frac{(70-56.3)^2}{56.3} + \frac{(905-893.0)^2}{893.0} + \frac{(890-888.3)^2}{888.3} + \frac{(870-883.7)^2}{883.7} = 6.29,$$

$P \approx 0.04$.

6. Decision: We do not reject H_0 at $\alpha = 0.025$. Nevertheless, with the above P-value computed, it would certainly be dangerous to conclude that the proportion of defectives produced is the same for all shifts. ◢

Often a complete study involving the use of statistical methods in hypothesis testing can be illustrated for the scientist or engineer using both test statistics, complete with P-values and statistical graphics. The graphics supplement the numerical diagnostics with pictures that show intuitively why the P-values appear as they do, as well as how reasonable (or not) the operative assumptions are.

10.14 Two-Sample Case Study

In this section, we consider a study involving a thorough graphical and formal analysis, along with annotated computer printout and conclusions. In a data analysis study conducted by personnel at the Statistics Consulting Center at Virginia Tech, two different materials, alloy A and alloy B, were compared in terms of breaking strength. Alloy B is more expensive, but it should certainly be adopted if it can be shown to be stronger than alloy A. The consistency of performance of the two alloys should also be taken into account.

Random samples of beams made from each alloy were selected, and strength was measured in units of 0.001-inch deflection as a fixed force was applied at both ends of the beam. Twenty specimens were used for each of the two alloys. The data are given in Table 10.13.

It is important that the engineer compare the two alloys. Of concern is average strength and reproducibility. It is of interest to determine if there is a severe