# Statistisk inferens

- Tidligere i kurset har vi undersøkt data og trukket uformelle konklusjoner om verdiene:
- Oblig1 – deskriptiv dataanalyse:
  - Biologiversjonen: Er det forskjell i kortisolnivå hos ulv mellom de som utsettes for jakt, og de som ikke?
  - Geologiversjonen: Er det forskjell i forekomst av sandstein i grunnen på dypt eller grunt vann utenfor kysten av Puerto Rico?
- Nå skal vi over på formell statistisk inferens:
  - Mål: Trekke konklusjoner fra data
  - Basere konklusjoner på sannsynlighetsberegninger
  - Tar hensyn til usikkerhet/variasjon

# Statistisk inferens: Fra observerte data om utvalg til konklusjon om populasjonsparameter

- To viktige metoder:
  - Konfidensintervall
  - Signifikanstester
- Basert på *utvalgsfordelinga* til observator
  - Krever *sannsynlighetsmodell* for dataene
  - Statistisk inferens baserer seg på at dataene kommer fra et *tilfeldig utvalg* eller et *randomisert eksperiment*

# Nå i kapittel 6, antar vi at vi kjenner verdien av populasjons-standardavviket σ

- dette hjelper oss å beskrive tankegangen bak statistisk inferens, fordi noen detaljer blir enklere
- I bakhodet har vi med oss at det ofte er en **urealistisk antagelse** at vi kjenner det teoretiske standardavviket σ
- I kapittel 7 vil vi se på **mer realistiske** metoder, som kan brukes for de fleste typer av data vi har sett på tidligere

# Konfidensintervall [$\bar{x}$-m, $\bar{x}$+m]: estimat med feilmargin

- Konfidensintervall for forventningsverdien μ for populasjonen
- Estimat for μ gitt av observator $\bar{x}$ (gjennomsnitt)
- Konfidensnivå C (typisk 95% eller 99%)
- Feilmarginen $m = z \cdot \sigma / \sqrt{n}$ bestemmes av standardavviket σ til observasjonene, utvalgstørrelsen n og nivå C.

# Tolkning av konfidensintervall

- Parameterverdien vår er et fast, men ukjent tall og vil enten ligge innenfor konfidensintervallet eller ikke.
- *95% konfidensintervall betyr altså at hadde vi gjentatt forsøket mange ganger, ville den sanne parameterverdien ligget innenfor konfidensintervallet 95% av gangene.*

# Noen forsiktighetsregler for konfidensintervall

- Data bør være fra et *enkelt tilfeldig utvalg* (SRS) av populasjonen
  - Viktig med *uavhengige* observasjoner fra populasjonen
  - Det finnes korrigerte formler for mer kompliserte design
- Ingen fancy formel kan redde 'dårlige data': Konfidensintervall gjelder ikke for skjeve utvalg
- Formelen for konfidensintervall er sensitiv til *uteliggere*
- Lite robust for små $n$ (bør ha $n>15$ når data ikke er normalfordelte)

# Statistiske signifikanstester brukes når vi ønsker å ta stilling til et utsagn om en parameter i en populasjon

- Utsagnet kaller vi hypotesen
- Bruker observerte data til å teste hypotesen om populasjonen
- Overordnet, er prosedyren:
  - Beregn sannsynlighet for observert utfall av observator (eller noe mer ekstremt) gitt antatt hypotese
  - Hvis sannsynlighet liten, forkast hypotese

# Fra deskriptiv dataanalyse i oblig 1 til statistisk inferens: Et spørsmål om statistisk signifikans

- Er det en forskjell i cortisol-konsentrasjon mellom ulv som utsettes for tung jakt og lett jakt? Hårprøver fra 103 tungt jaktede og 45 lett jaktede ulv, cortisol måles.
- Gjennomsnittlig cortisol-konsentrasjon:
  - $\bar{x}_1$ = 17.07 pg/mg for tungt jaktede
  - $\bar{x}_2$ = 15.56 pg/mg for lett jaktede

Er den observerte differansen 1.51 pg/mg, reell eller tilfeldig?

# Hvordan gjennomføres en hypotesetest?

1. Formaliserer spørsmålet:
   Er det forskjell på verdi av populasjonsparameteren [forventningsverdi] mellom populasjonene?
2. Undersøker om observerte data er kompatible med hypotesen om at det ikke er noen forskjell:

Vi regner ut sannsynlighet for å observere forskjellen vi har sett mellom utvalgene **under antagelsen om at det ikke er noen forskjell mellom populasjonene**

• Hvis sannsynligheten er høy (eks 23 %), er det ikke grunnlag i data for å si det er en forskjell

•Hvis sannsynligheten er lav (eks 0.3 %): Forkast antagelse om det ikke forskjell mellom populasjonene, og konkluder med at det ER en forskjell

# Null-hypotesen $H_0$

- Utgangshypotese
- Beholdes til data tilsier at $H_0$ er urimelig

– Typisk nullhypotese: Det er ingen effekt/forskjell, eller populasjons-forventningene er like

  - Notasjon: $H_0$: $\mu = \mu_1 - \mu_2 = 0$, ekvivalent: $\mu_1 = \mu_2$

- Signifikanstesten er designet for å angi bevisstyrke *mot* $H_0$

# Alternativ hypotese $H_a$

- Det vi ønsker å konkludere med dersom dataene ikke er i samsvar med $H_0$
- Typisk alternativ hypotese: 'Det er en effekt', eller 'Det er en forskjell'
  - Notasjon: $H_a$: $\mu = \mu_1 - \mu_2 \neq 0$ , ekvivalent: $\mu 1 \neq \mu 2$
- Triks: Fordi $H_a$ utrykker effekten vi skal undersøke om eksisterer eller ikke, er det ofte praktisk å starte med å formulere $H_a$ og deretter sette opp $H_0$ som utsagnet om at den ønskede effekten ikke er tilstede

# Hypoteser

- Hypoteser er alltid utsagn/antagelser om populasjoner (eller modeller), ikke et spesielt utfall. Derfor må $H_0$ og $H_a$ alltid formuleres ut i fra de ukjente populasjonsparameterne

- Når vi skal formulere $H_a$, må vi tas stilling til om den skal være *ensidig* (f.eks. $H_a: \mu > 0$) eller *tosidig* (f.eks. $H_a: \mu \neq 0$)
  - Uttrykker om parameteren er forskjellig fra nullhypotese-verdien i en bestemt retning eller ikke
  - Metoden er ikke gyldig om man ser på dataene for å formulere $H_a$
  - Med mindre du har god begrunnelse for retning på effekten gjennom forhåndskunnskap, **velg tosidig $H_a$**.

# 1 flervalgsspørsmål

# Hypotesetesten er basert på en testobservator

•Observatoren estimerer parameteren vi er interessert i, og er ofte den samme vi ville brukt til et konfidensintervall for parameteren

– Eks.: $\bar{x}_1 - \bar{x}_2$ estimerer $\mu = \mu_1 - \mu_2$

• Verdier langt fra parameterverdi spesifisert av $H_0$ gir bevis mot $H_0$

• $H_a$ angir hvilken retning som teller:

– Ensidig $H_a$: $\mu_1 > \mu_2$ angir at vi må ha stor $\bar{x}_1 - \bar{x}_2$ som bevis mot $H_0$

– Ensidig $H_a$: $\mu_1 < \mu_2$ angir at vi må ha liten $\bar{x}_1 - \bar{x}_2$ (stor $\bar{x}_2 - \bar{x}_1$) som bevis mot $H_0$

– Tosidig $H_a$: $\mu \neq \mu_2$ angir at vi må ha stor $|\bar{x}_1 - \bar{x}_2|$ som bevis mot $H_0$

# Standardisert testobservator

• For å undersøke hvor langt estimatet er fra parameterverdien spesifisert av $H_0$, standardiserer vi estimatet

$$Z = \frac{Estimat - Parameterverdi\ under\ H_0}{Standardavvik\ estimat}$$

# Tenk tilbake på observert differanse på 1.51 pg/mg for tungt vs lett jaktet ulv

- $H_0$: Det er ingen forskjell  i de sanne forventningene
- $H_a$: Det er en forskjell i de sanne forventningene
- Formelt skriver vi:

$H_0: \mu_1 = \mu_2$ mot $H_a: \mu_1 \neq \mu_2$,
som er det samme som

- $H_0: \mu_1 - \mu_2 = 0$ mot $H_a: \mu_1 - \mu_2 \neq 0$
- Tosidig alternativ: Store verdier av både positive og negative forskjeller, dvs store (nok) verdier av absoluttverdien $|\bar{x}_1 - \bar{x}_2|$ teller som bevis mot $H_0$

# Standardisert testobservator i ulveeksempelet

Estimat for $\mu_1 - \mu_2$ : $\bar{x}_1 - \bar{x}_2 = 1.51$ pg/mg
Anta: Standardavvik for $\bar{x}_1 - \bar{x}_2 = 1.22$

$$z = \frac{Estimat - Parameterverdi\ under\ H_0}{Standardavvik\ estimat}$$

- Standardisert testobservator gitt $\mu_1 - \mu_2 = 0$:
  $z=(1.51 - 0)/1.22 = 1.24$
- Trenger en måte å avgjøre om z = 1.24 er stor (ekstrem) nok til å være **bevis mot $H_0$**

# Statistisk signifikanstest: P-verdi

- *P-verdi*: Sannsynligheten for at et utfall er like ekstremt eller mer ekstremt enn faktisk observerte estimatet. Sannsynligheten beregnes ved å anta at parameterverdien gitt av $H_0$ er sann.

  - Ekstremt definerer vi som langt fra hva vi ville forvente gitt at $H_0$ er sann. Retning på hva som regnes som ekstremt bestemmes av $H_a$ og $H_0$
  - Liten P-verdi: Sterk grad av **bevis mot $H_0$**

# Er det forskjell i cortisol-nivå hos tungt mot lett jaktet ulv ?

- $H_0$: $\mu_1 - \mu_2 = 0$ mot $H_a$: $\mu_1 - \mu_2 \neq 0$
- **Ekstremt betyr her**: Langt fra hva vi ville forvente hvis $H_0$ var sann, dvs positive og negative forskjeller like store eller større enn observert verdi av $\bar{x}_1 - \bar{x}_2$, dvs like store eller større enn observert verdi av absolouttverdien $|\bar{x}_1 - \bar{x}_2|$

- Standardisert testobservator:

$$z = \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{\sigma_{\bar{x}_1 - \bar{x}_2}} = \frac{1.51 - 0}{1.22} = 1.24$$

- **Ekstremt**: Like store eller større enn observert verdi av absoluttverdien $|\bar{x}_1 - \bar{x}_2|$, som betyr like store eller større enn observert verdi av $|z|$
- **Z tilnærmet N(0,1) fordi vi antar at $H_0$ gjelder når vi beregner P-verdien**
- P-verdi = P(Z > 1.24) + P(Z < -1.24) = 2*0.1075=0.215

- P-verdien er stor, så vi kan ikke forkaste nullhypotesen!

# Hvordan konkludere om vi har observert en statistisk signifikant forskjell?

*Vi må på forhånd ha bestemt signifikansnivå* α:

– **Typisk: α=0.05 (eller 0.01)**

– Grenseverdi for når vi forkaster

– Forkaster når P-verdi ≤ α

Signifikant på nivå α betyr at observert P-verdi var mindre enn α,

Med andre ord: det er ikke grunnlag for å forkaste når P-verdi > α

Signifikant betyr at bevisene mot nullhypotesen nådde standarden satt av α for datasettet som ble analysert

NB: Det er ingen skarp grense mellom signifikant og ikke-signifikant, men det er økende grad av bevis mot $H_0$ når p-verdien minker.
*Rapporter alltid p-verdien!*

# Hvordan gjennomføre en statistisk signifikanstest?

1. Formuler $H_0$ og $H_a$
2. Beregn test-observator
3. Finn P-verdi
4. Formuler en konklusjon

NB: Bruker ofte datamaskin til å finne P-verdi, men en datamaskin

- Kan ikke formulere $H_0$ og $H_a$
- Kan ikke tolke P-verdien for deg
- Kan ikke bedømme om forutsetningene for å bruke testen er oppfylt

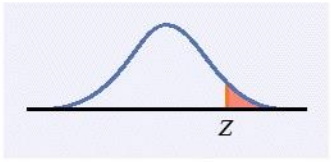
## *z* TEST FOR A POPULATION MEAN

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

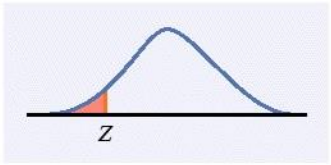
$$z = \frac{\overline{x} - \mu_0}{\sigma / \sqrt{n}}$$

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

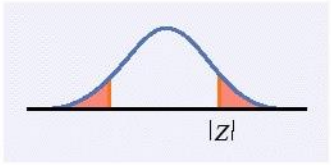
$H_a$: $\mu > \mu_0$ is $P(Z \geq z)$




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




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




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

# Blodtrykk for menn i alderen 35-44 år har forventning $\mu=128$ og standardavvik $\sigma=15$

*Kilde: National Center for Health Statistics*

En bedriftslege ønsker å undersøke: Har mannlige toppledere (alder 35-44) i bedriften annerledes forventet blodtrykk enn den generelle populasjonen av menn på samme alder?

Bedriftslegen har ingen grunn til å anta noen retning på forskjellen.

Målinger av blodtrykket til 72 toppledere har gjennomsnitt

$\bar{x} = 126.07$

Spørsmål vi må stille oss er:

- Ensidig eller tosidig test?
- Hva er verdien til standardisert testobservator?
- Hva er P-verdien?

# Blodtrykk for menn i alderen 35-44år har forventning μ=128 og standardavvik σ=15

*Kilde: National Center for Health Statistics*

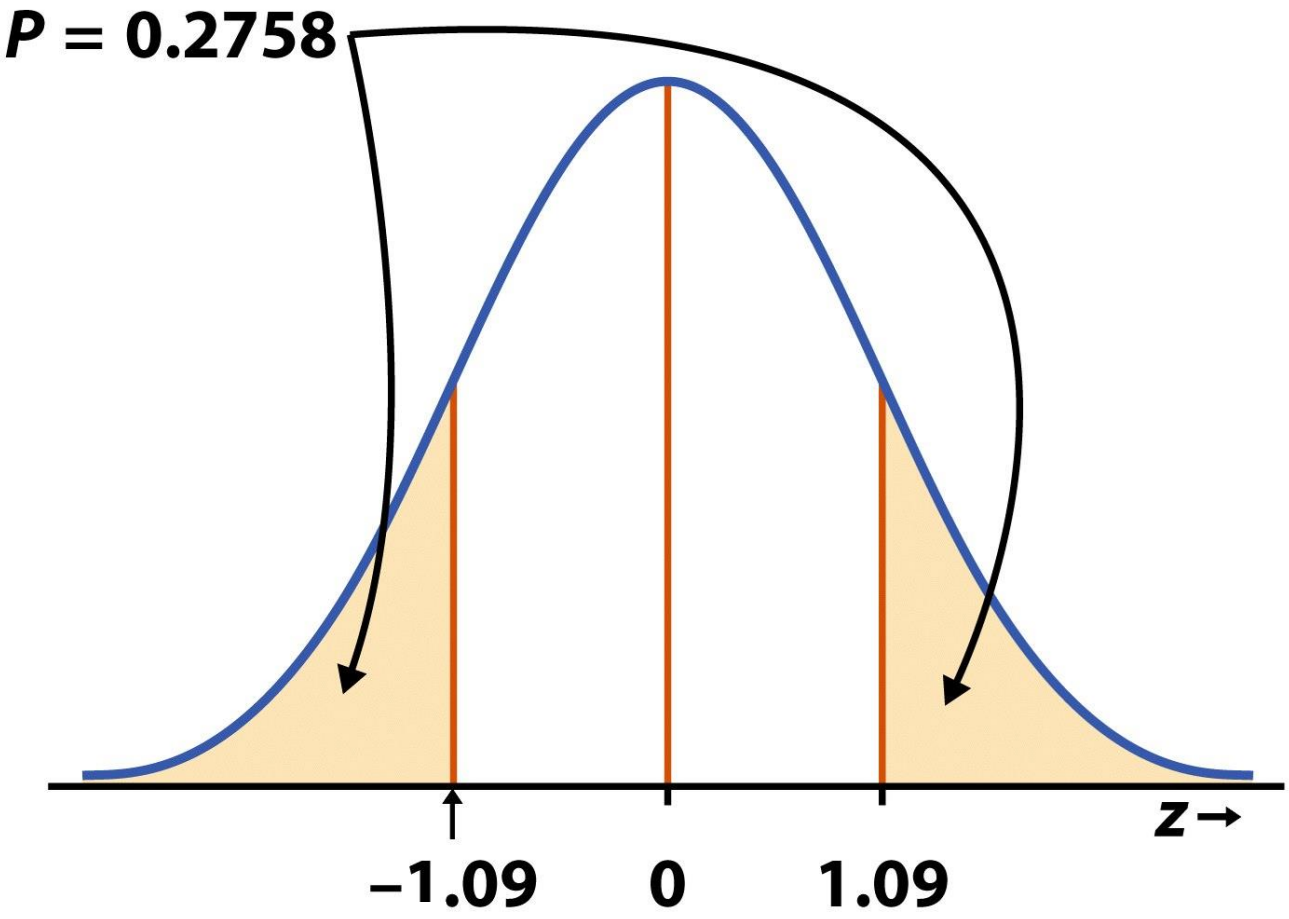
En bedriftslege ønsker å undersøke: Har mannlige toppledere (alder 35-44) i bedriften annerledes forventet blodtrykk enn den generelle populasjonen av menn på samme alder?

Bedriftslegen har ingen grunn til å anta noen retning på forskjellen.

Målinger av blodtrykket til 72 toppledere har gjennomsnitt

$\bar{x}$ =126.07

$$z = (126.07 - 128)/(15 / \sqrt{72}) = -1.09$$


P = 0.2758
z →
–1.09
0
1.09

# Eksempel: Blodtrykk

- P-verdi = 0.2758
- Tolkning: Et SRS med n=72 fra den generelle populasjonen av menn på samme alder vil 27% av gangene gi et gjennomsnittlig blodtrykk like langt eller lenger fra 128 som det observerte $\bar{x}_1$= 126.07.
- Konklusjon: Den observerte $\bar{x}$ gir altså ikke bevis for at topplederene er annerledes enn andre menn på samme alder. Ingen grunn til å forkaste $H_0$

- HUSK: Å ikke finne bevis mot $H_0$ betyr bare at data er forenlige med null-hypotesen, det betyr *ikke* at vi har bevis for at null-hypotesen er sann.

# To-sidige tester og konfidensintervall

Konfidensintervall med konfidensnivå C: $[\bar{x}-z^*\sigma/\sqrt{n}, \bar{x}+z^*\sigma/\sqrt{n}]$

Verdier av μ utenfor intervall er ikke kompatible med data

Mulig test-prosedyre:

– Forkaste $H_0$ hvis $\mu_0$ ikke i konfidensintervall

Kan vises: Ekvivalent med signifikanstest

*En tosidig signifikanstest med nivå α som forkaster $H_0$:$\mu=\mu_0$ er ekvivalent med at $\mu_0$ faller utenfor konfidensintervallet for μ med nivå C=(1-α)%*

Noen ganger enklere å konstruere konfidensintervaller

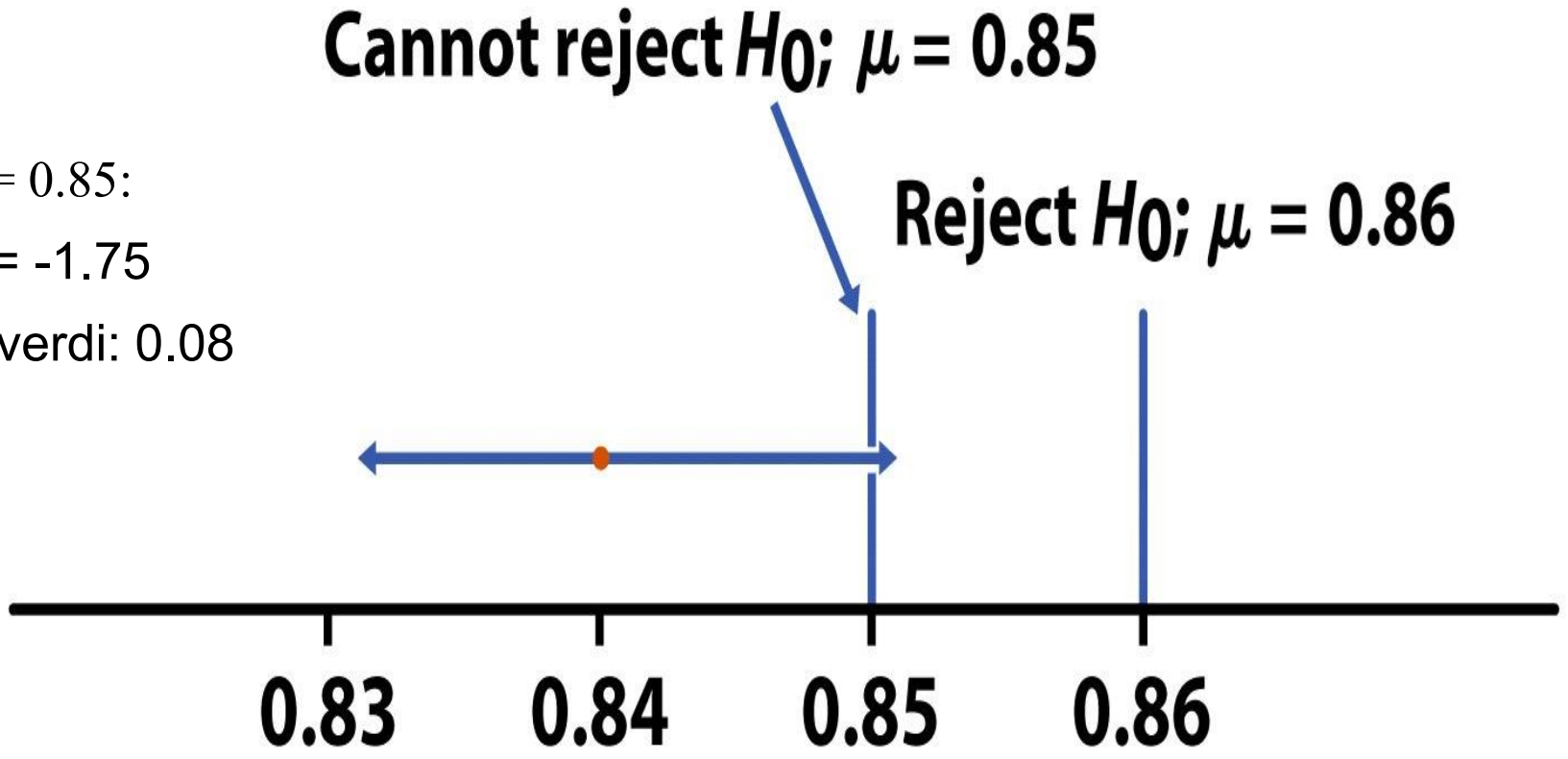
# Eksempel: To-sidige tester og konfidensintervall

- Analyse av konsentrasjonen av farmasøytisk produkt. Ikke helt presis målemetode, repeterte analyser av samme prøve vil gi litt forskjellig svar.
- Forhåndskunnskap: Konsentrasjon kan antas N(μ,σ=0.0095)-fordelt
- Laboratoriet har blitt spurt om å evaluere en påstand om at konsentrasjonen i en prøve er 0.86%, dvs $H_0$: $\mu = 0.86$ $H_a$: $\mu \neq 0.86$
  - Tre målinger av prøven ga resultatene 0.8403, 0.8363 og 0.8447
  - Gjennomsnitt $\bar{x} = 0.8404$
  - Testobservator $z = (0.8404 - 0.86)/(0.0095 / \sqrt{3}) = -3.57$
  - P-verdi < 0.0003*2 = 0.0006, forkaster $H_0$ med nivå $\alpha$=0.05
  - 95% konfidensintervall
    $0.8404 \pm 1.96 * 0.0095/\sqrt{3} = (0.830 , 0.851)$
    inneholder ikke $\mu = 0.86$:

$\mu = 0.85$:

Z = -1.75

P-verdi: 0.08



# Å kjenne p-verdien gir oss muligheten til å vurdere statistisk signifikans på ethvert nivå

P-verdien er laveste signifikansnivå som gir forkastning av nullhypotesen

*Oppgi derfor P-verdien i tillegg til resultatet av hypotesetesten.*

En konklusjon a la «$H_0$ er forkastet på nivå α» (dvs P-verdi < α) forteller ikke alt, og er ikke like informativt som å angi selve P-verdien!

Eksempel med testing av konsentrasjonen i farmasøytisk produkt ga P=0.00036

Dette er signifikant på 0.05 nivå, og på 0.01 nivå, og på 0.001 nivå

# KONFIDENSINTERVALLER OG TESTER FOR STATISTISK SIGNIFIKANS

- Konfidensintervaller er en utvidelse av å estimere en populasjonsparameter ved å beregne en enkeltverdi, og sier noe om hvilke verdier av en populasjonsparameter som er i overenstemmelse med data
- Signifikanstester er nyttige dersom vi ønsker å ta stilling til et utsagn (en hypotese) om en parameter i en populasjon
- Antagelsen om kjent populasjonsstandardavvik $\sigma$ lot oss fokusere på disse nye konseptene; i kapittel 7 skal vi lære hva som gjøres i t-prosedyrer (når vi ikke antar kjent $\sigma$)

# Bruk og misbruk av tester (6.3)

Det er ikke innafor å teste hypoteser
med data brukt til å formulere dem!

# Det er nesten for enkelt å utføre en statistisk hypotesetest med programvare (eks: R)

1. Å tolke og å forstå resultatet er vanskeligere
   - Testen kun gyldig under visse forutsetninger
   - Konfidensnivå: Ingen klar grense; vanlig og nyttig å rapportere p-verdi!

2. Ofte gjennomføres mange ulike tester
   - P-verdi relatert til å gjøre *en* test
   - Ved mange tester må justeringer gjøres for å kontrollere på falske positiv

# Å forkaste $H_0$ betyr at vi har observert en *statistisk signifikant* forskjell

- Ikke vektlegg statistisk signifikans alene, se også på de faktiske resultatene. God idé å rapportere konfidensintervall for parameteren vi undersøker i tillegg
  - På den ene siden: Kan ha forkastet nullhypotesen selv om effekten er ubetydelig liten; hvis *n* stor
  - På den andres siden: Å ikke kunne forkaste betyr ikke at $H_0$ er sann; dette er spesielt viktig å huske når *n* liten

# Statistisk signifikans betyr *ikke* at det er praktisk relevant!

Eksempel: Korrelasjon

- To variable X og Y, $H_0$: Ingen korrelasjon, $\rho=0$
- 400 observasjoner, $r=0.10$
- Statistisk signifikant med $\alpha=0.05$
- Klar indikasjon på sammenheng mellom X og Y
- Forklart variasjon: $r^2=0.01$,

  svak sammenheng!

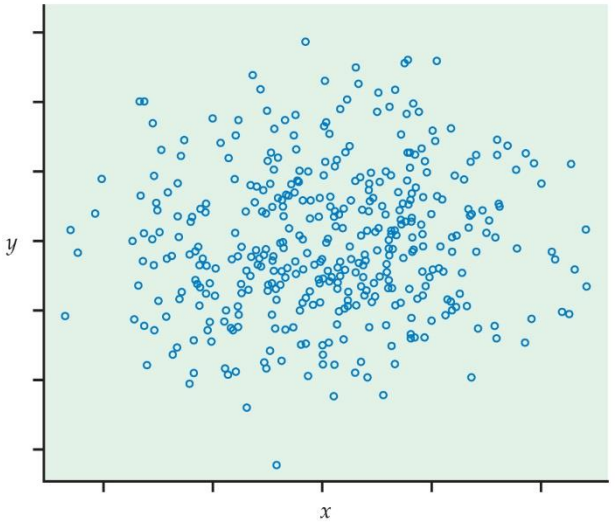



**Figure 6.15**
Moore/McCabe/Craig, *Introduction to the Practice of Statistics*, 9e, © 2017 W. H. Freeman and Company

# $H_0$ behøver ikke være sann selv om den ikke kan forkastes
## Eksempel: Hiv-behandling

- Intervensjonsstudie med behandlingsgruppe og kontrollgruppe for å undersøke mulig reduksjon av smitte-risiko
- Insidensrateforhold ble analysert
  I:Forhold mellom smitterate i behandlingsgruppa og kontrollgruppa
- $H_0$: I=1, og data ga 95% konfidensintervall [0.63,1.58]
- Fordi $H_0$ kunne ikke forkastes, ble det rapportert at insidensrateforholdet var I = 1

- Men det er misvisende å konkludere at behandling ikke påvirker smitterisko
- Hva som faktisk er sant bør undersøkes nærmere med mer data!

# Multippel testing
# F.eks. i genomiske eksperimenter

- Ønsker å finne gener som forklarer sykdom
- Undersøker samtidig ti-tusener av mulige gener, utfører en test på hvert gen
- 10 000 tester, $\alpha=0.05$
  - Forventer at 500 tester vil være signifikante kun pga. tilfeldigheter!
- Aktivt forskningsfelt:
  - Hvordan håndtere mange tester simultant
  - Må justere for multippel testing, False Discovery Rate