Statistisk inferens

Kunsten å trekke konklusjoner om en stor populasjon fra et mindre utvalg

Utfordringa er hvordan vi skal overføre informasjon om et utvalg til informasjon om hele populasjonen

Relevante konsepter fra kapittel 3

- Underliggende variable og kausalitet
- Innsamling av data
  - Planlagte eksperimenter (randomisering)
  - Utvalgsundersøkelser (SRS = Enkelt tilfeldig utvalg)

Vi bruker verdien til observatoren for å estimere (anslå) verdien til parameteren

En **parameter** er et tall som beskriver en egenskap til populasjonen. Verdien til en parameter er ofte ukjent. Dette er fordi vi stort sett i praksis ikke kan undersøke hele populasjonen.

En **observator** (eller **statistikk**) er et tall som beskriver en egenskap til et utvalg. Verdien til en observator kan regnes ut direkte fra utvalgsdataene, men den kan variere fra utvalg til utvalg. Vi bruker ofte en observator til å estimere en ukjent parameter.

Vi bruker de greske bokstavene μ (my) og σ (sigma) for henholdsvis populasjonsgjennomsnittet og standardavviket til populasjonen. Vi bruker $\bar{x}$ (x med strek over) for gjennomsnittet og s for standardavviket til utvalget.

Verdien av en observator varierer når man trekker repeterte tilfeldige utvalg

- Et enkelt tilfeldig utvalg gir forventningsrett observator,
  - men vi vil alltid ha variabilitet
  - Og variasjonen følger et forutsigbart mønster

Statistisk inferens baserer seg på forståelsen av hva som skjer hvis en prosedyre blir repetert mange ganger. Det sier også noe om hvor pålitelig prosedyren er

Hvordan kan vi lære om utvalgsvariabilitet?

1. Observere mange utvalg fra samme populasjon
2. Beregn $\hat{p}$ for hvert utvalg
3. Lag et histogram av $\hat{p}$
4. Undersøk histogrammets for form, senter og spredning

I praksis er det ofte for dyrt/tidkrevende å se på mange reelle utvalg

- Effektivt alternativ: Å simulere ved hjelp av en datamaskin, dvs *imiterer* repeterte (mange) utvalg

Utvalgsvariabilitet:
Simulering av utvalg, n= 100

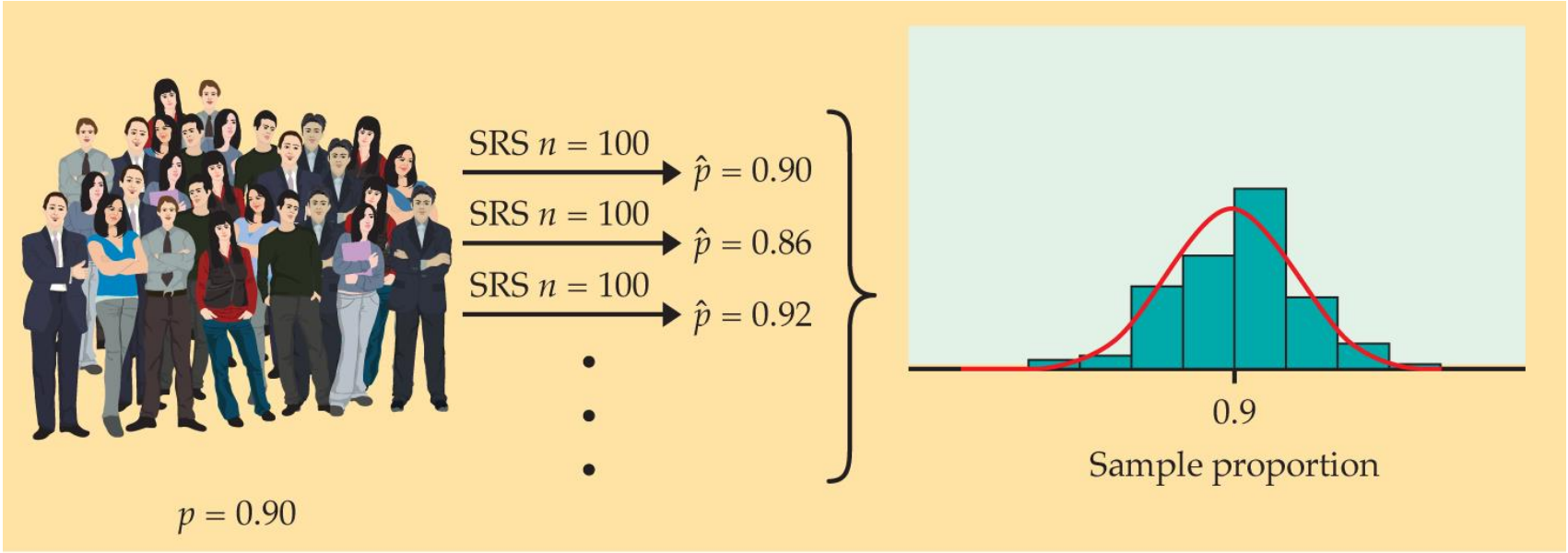


Figure 5.1
Moore/McCabe/Craig, *Introduction to the Practice of Statistics*, 9e, © 2017 W. H. Freeman and Company

Vi kan simulere repeterte utvalg for å lære om utvalgsvariabilitet

- Later som om vi vet at andelen i populasjonen som har en bestemt mening er f.eks. 90%, dvs p=0.9
- Trekker 1000 utvalg av størrelse *n*=100 fra en slik populasjon (trekker 100 personer 1000 ganger fra en populasjon med p=0.9)
- For hvert av de 1000 utvalgene beregnes $\hat{p}$ (dvs antall med den meningen for hvert utvalg, delt på *n*=100), lager så histogram av disse 1000 verdiene av $\hat{p}$
- Gjentas for utvalg av størrelse *n*=1200

Utvalgsvariabilitet:
Simulering av utvalg, n= 100

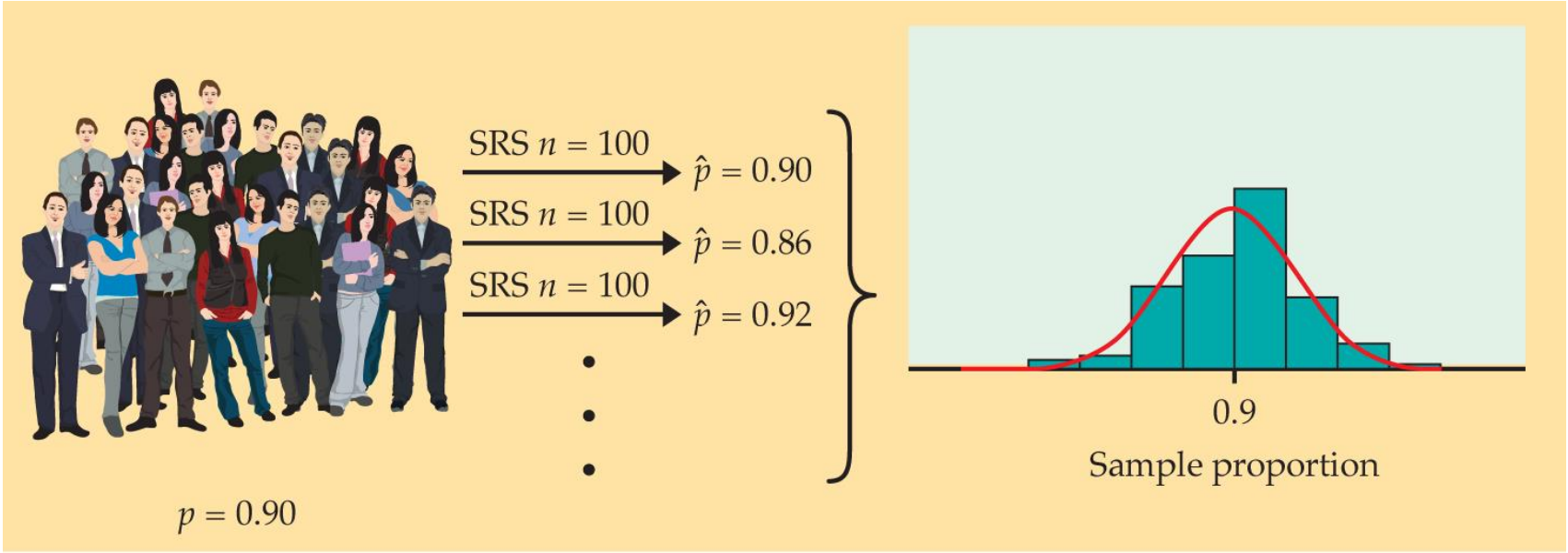


Figure 5.1
Moore/McCabe/Craig, *Introduction to the Practice of Statistics*, 9e, © 2017 W. H. Freeman and Company

Utvalgsvariabilitet:
Simulering av utvalg, n=1200

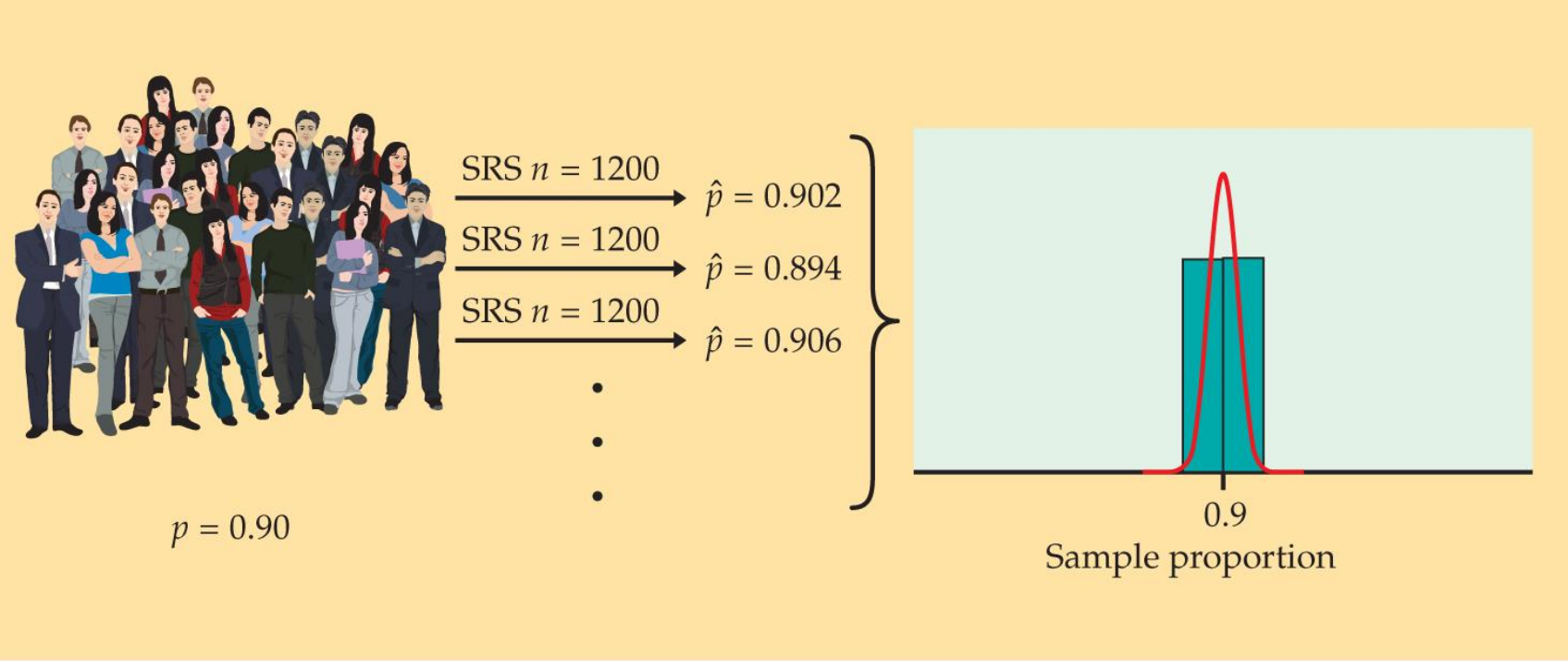


Figure 5.2
Moore/McCabe/Craig, *Introduction to the Practice of Statistics*, 9e, © 2017 W. H. Freeman and Company

De to histogrammene gir et bilde av utvalgsfordelinga til observatoren, for to ulike verdier av n

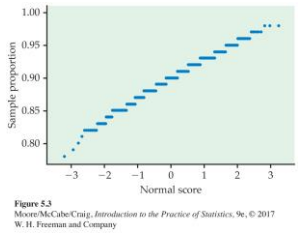


Figure 5.3
Moore/McCabe/Craig, *Introduction to the Practice of Statistics*, 9e, © 2017 W. H. Freeman and Company

- Form: Begge histogrammene ser ut til å stemme godt med normalfordeling (kvantilplott bekrefter dette (for *n*=100))
- Senter: Begge histogrammer er sentrert i 0.9, ingen tendens til at verdiene er høyere eller lavere enn 0.9, dvs $\hat{p}$ har ingen forventningsskjevhet som estimator for p
- Spredning: Mye lavere spredning for *n*=1200 sammenlignet med *n*=100

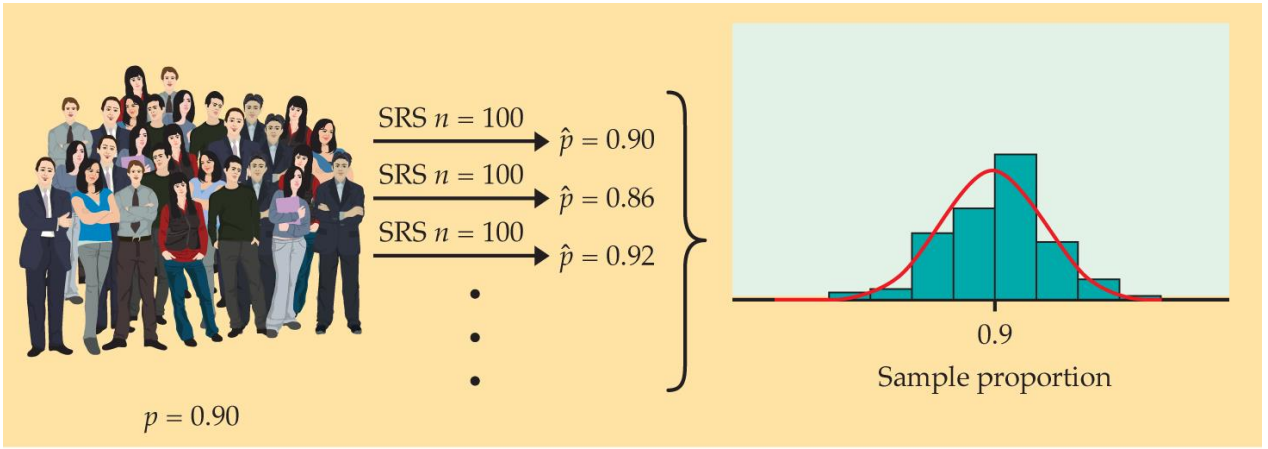


Figure 5.1
Moore/McCabe/Craig, *Introduction to the Practice of Statistics*, 9e, © 2017 W. H. Freeman and Company

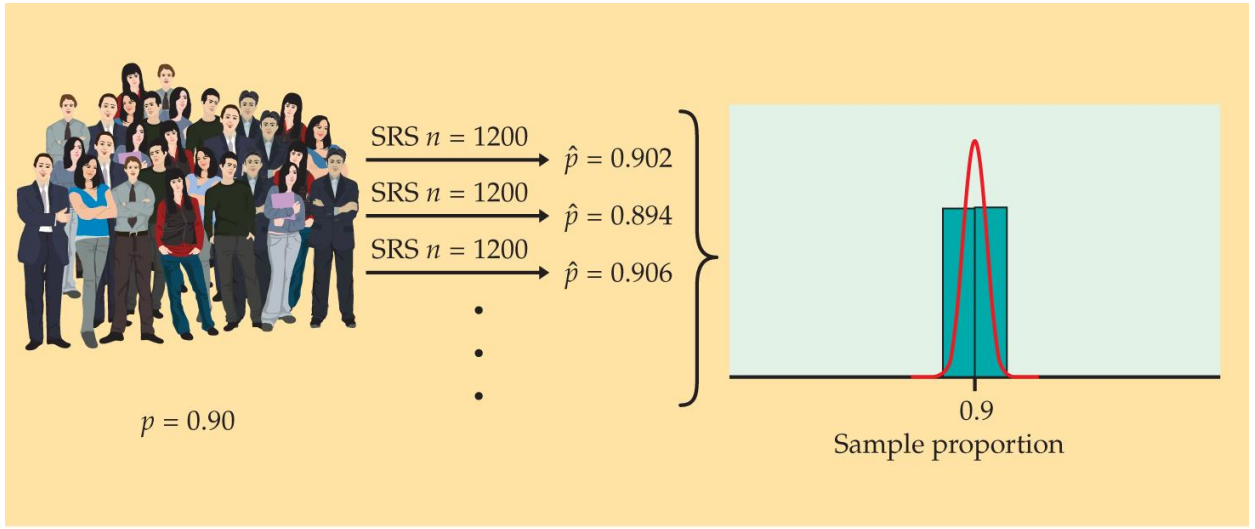


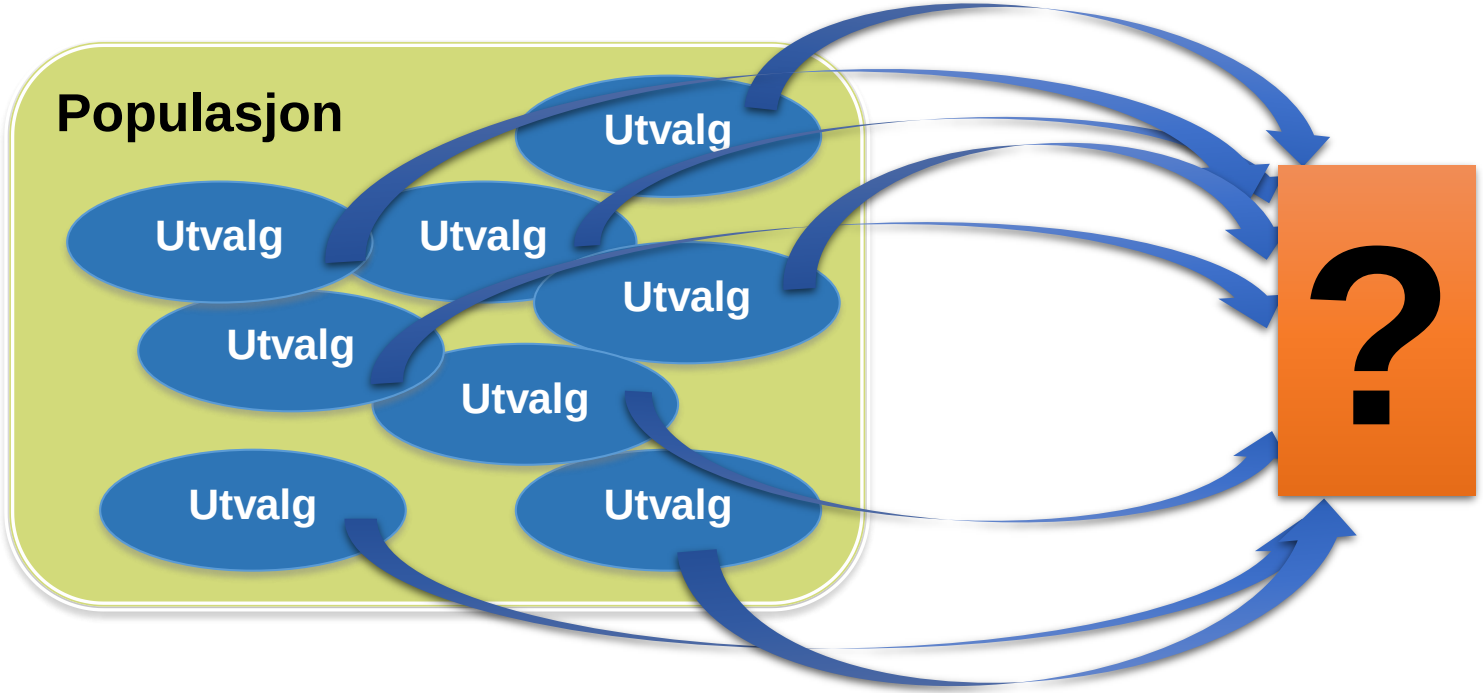
Figure 5.2
Moore/McCabe/Craig, *Introduction to the Practice of Statistics*, 9e, © 2017 W. H. Freeman and Company

Utvalgsfordelinga til en observator er fordelinga av verdier som observatoren kan ta, over alle mulige utvalg av samme størrelse n fra populasjonen

- Simulering gir en tilnærming av den sanne utvalgsfordelinga (eksempel: vi så kun på 1000 utvalg, *ikke alle mulige)*
- Sannsynlighetsteori kan noen ganger angi eksakt fordeling
- Uansett: Beskriver fordelinga ved form, senter og spredning

Begrepet **utvalgsvariabilitet** blir brukt til å beskrive at verdien til en observator varierer mellom gjentatte tilfeldige utvalg

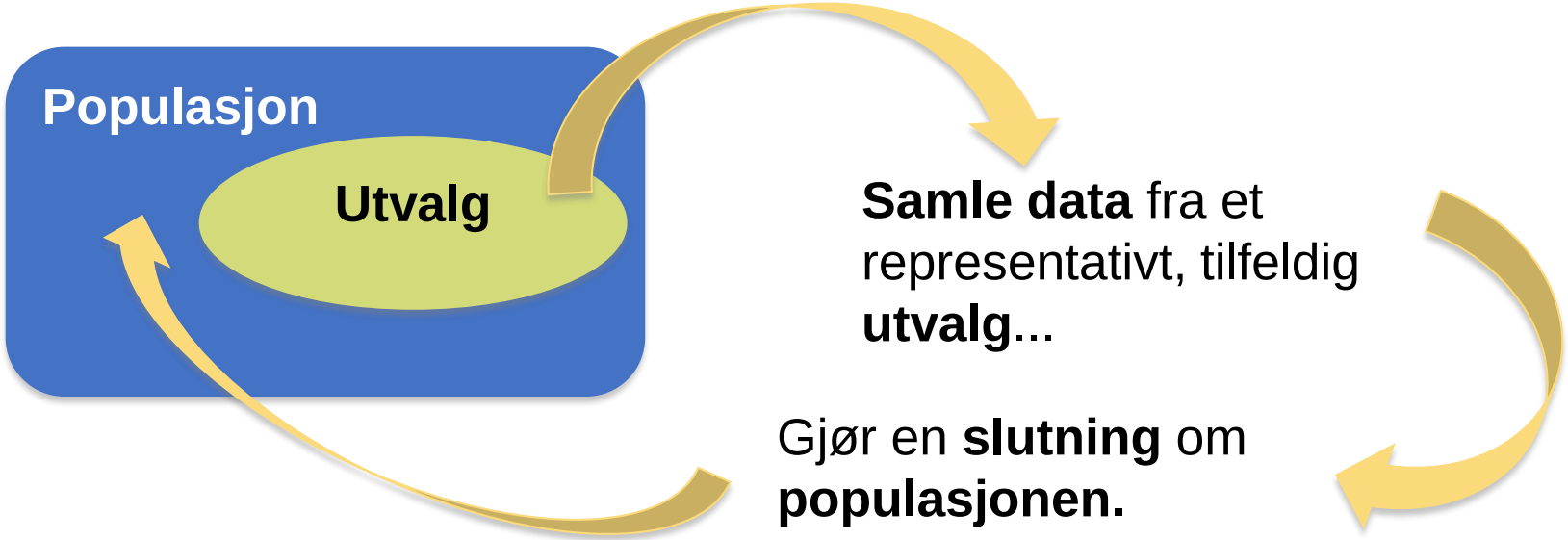
«Hva ville skje dersom vi trakk flere repeterte utvalg?»



Prosessen statistisk inferens innebærer å bruke informasjon fra et mindre utvalg til å komme med konklusjoner om en større populasjon

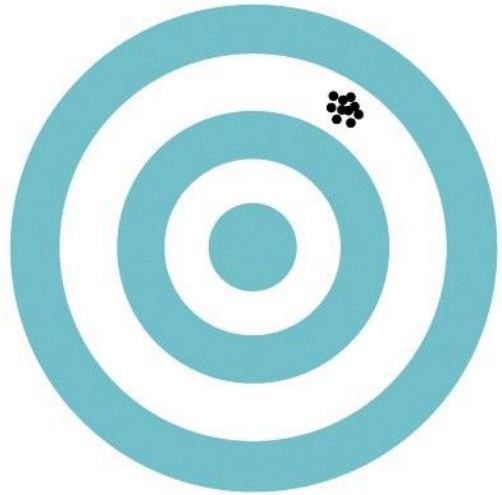
Utvalgsfordelinga til en observator: fordelinga av verdiene observatoren tar, over alle mulige utvalg med **en gitt størrelse** fra den **samme populasjonen**.

Ulike tilfeldige utvalg gir ulik verdi til observatoren.
For å kunne gjennomføre statistisk inferens, trenger vi å kunne beskrive **utvalgsfordelinga** til de mulige verdiene av en observator.

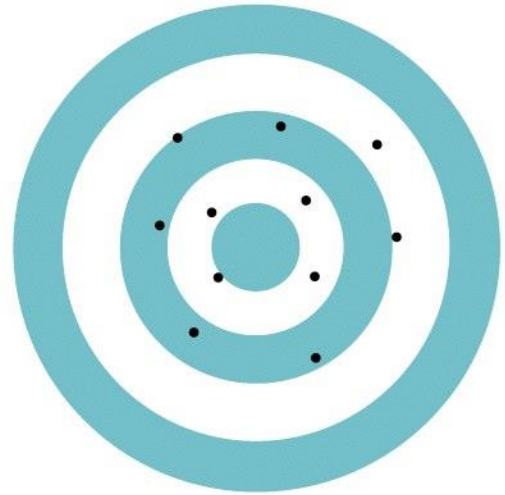


Vi kan illustrere **forventningsskjevhet** og **variabilitet** ved å tenke på den sanne parameterverdien for populasjonen som blinken på en skyteskive

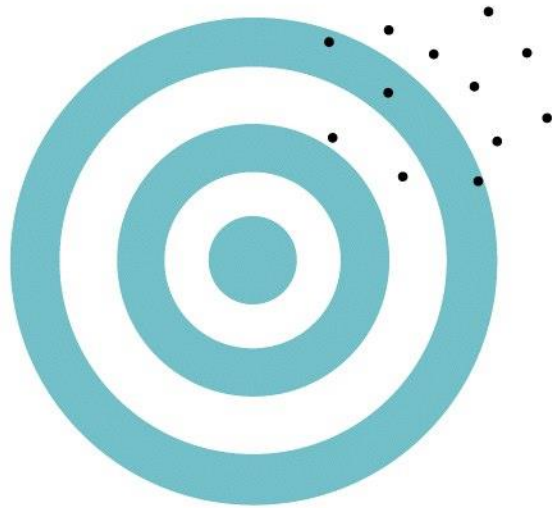
- Observatoren er et enkelt pilkast
- Forventningsskjevhet og variabilitet beskriver hva som skjer når man kaster mange ganger
- Forventningsskjevhet: Hvor langt fra blinken pilene *systematisk* treffer
- Variabilitet: Hvor spredt pilene treffer
- Et godt utvalgs-design har lav forventningsskjevhet og lav variabilitet, akkurat som en god pilkaster



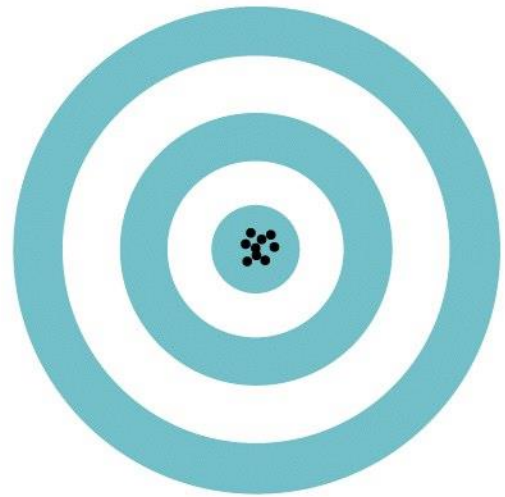
High bias, low variability
(a)



Low bias, high variability
(b)

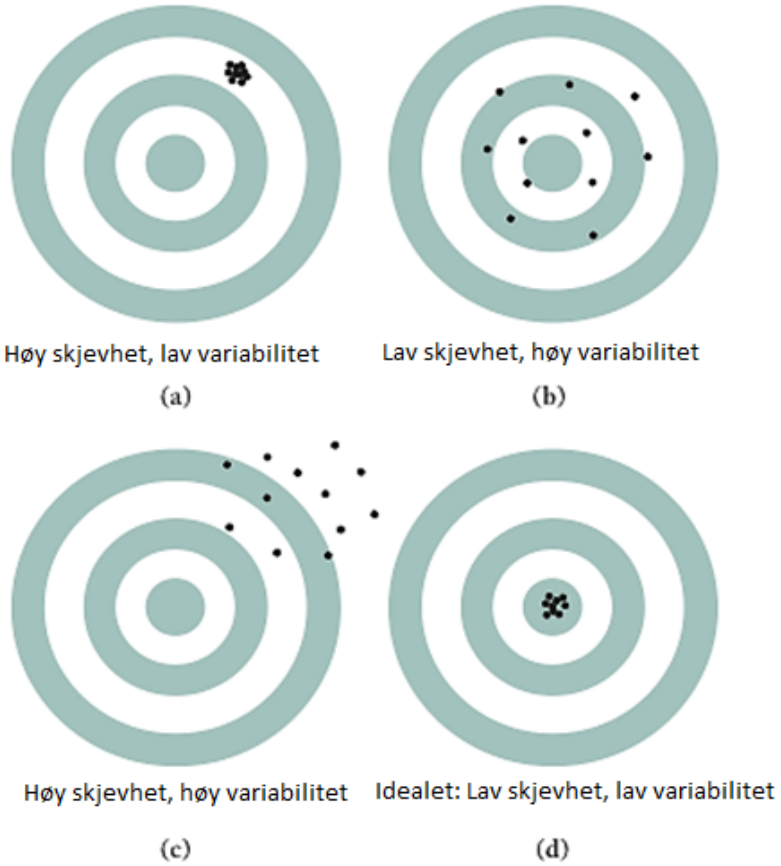


High bias, high variability
(c)



The ideal: low bias, low variability
(d)

Vi kan tenke på den ekte verdien til populasjonsparameteren som den innerste blinken på en dartskive, og observatoren fra et utvalg som en dartpil. Forventningsskjevhet og variabilitet beskriver hva som skjer når det kastes gjentatte piler mot målet.



Forventningsskjevhet (bias) handler om sentrum til utvalgsfordelinga:
En observator brukt til å estimere en parameter er **forventningsrett** hvis forventninga til utvalgsfordelinga har lik verdi som sanne parameterverdien.

Variabiliteten til en observator er beskrevet av spredninga til utvalgsfordelinga: Spredninga blir bestemt av utvalgsdesignet og utvalgsstørrelsen *n*. Observatorer fra større utvalg har mindre spredning.

Randomisering fjerner skjevheter

- SRS: senter i utvalgsfordelinga blir lik sann parameterverdi
- Lar oss bruke sannsynlighetsteori for å analysere data
- Form på utvalgsfordeling kjent, ofte (tilnærma) normalfordelt
- Spredninga kan reduseres ved å øke utvalgsstørrelsen
- Senere vil vi komme tilbake til mer tekniske detaljer for utvalgsfordelinger, og hvordan vi kan trekke statistiske konklusjoner basert på dem
- Statistisk teori baserer seg på:
  - *Hva skjer hvis eksperimentet repeteres mange ganger*

Et godt utvalgsdesign har både lav skjevhet og lav variabilitet

For å redusere forventningsskjevhet, bruk tilfeldige utvalg

For å redusere variabiliteten til en observator fra et enkelt tilfeldig utvalg, bruk et større utvalg.

Variabiliteten til en observator fra et tilfeldig utvalg avhenger ikke av størrelsen til populasjonen, så lenge populasjonen er minst 20 ganger større enn utvalget.

Godt design

- Redusere skjevhet: Tilfeldig utvalg
- Redusere variabilitet: Bruk stort utvalg
- Usikkerhetsmarginer:
  - Setter grenser for størrelse på feil i estimatet
  - Reflekterer utvalgsvariabiliteten
  - Avhenger av utvalgsstørrelsen: mindre for større utvalg
- Populasjonsstørrelse betyr ingenting, gitt at populasjonen er minst 20 ganger større enn utvalgsstørrelsen *n*

Husk dette: ☺

  - Utvalg på 2500 like bra for populasjon med 900.000 individer som for populasjon med 9.000.000 individer

2 flervalgsspørsmål

5.2 Fordelinga til gjennomsnittet i et utvalg

- Populasjonsfordeling
- Forventning og standardavvik til gjennomsnittet
- Fordelinga til et gjennomsnitt
- Sentralgrenseteoremet

Utvalgsfordelinger

Vi har sett at utvalgsfordelinga til en observator (statistic) er fordelinga av verdiene observatoren tar ved mange gjentatte utvalg av samme størrelse fra samme populasjon

Vi benytter utvalg som en tilfeldig mekanisme

Sannsynlighetsregning dreier seg om tilfeldige mekanismer

Så i dette kapitlet samler vi trådene, og bruker det vi kan om sannsynlighetsregning til å studere utvalgsfordelinga til de vanligste observatorene

For kvantitative data brukes observatorer som empirisk gjennomsnitt, andeler, persentiler og empirisk standardavvik

Alle disse er observatorer og har en utvalgsfordeling.

Vi skal i første omgang konsentrere oss om gjennomsnittet av et sett observasjoner.
Gjennomsnitt er den vanligste observatoren

Vi skal etterhvert se på andeler.

En observator fra et tilfeldig utvalg eller randomisert eksperiment er en tilfeldig variabel. Utvalgsfordelinga er sannsynlighetsfordelinga til observatoren

Utvalgene må trekkes fra en populasjon, og vi antar videre at også enhetene i populasjonen har en tilfeldig fordeling.

Populasjonsfordelinga er fordelinga til alle enhetene i populasjonen; den underliggende sannsynlighetsfordelinga når man trekker en enhet tilfeldig fra populasjonen.

Forventning til tilfeldige variable

- Forventning: Massesenteret (”Balansepunkt”) i fordelinga
- For diskrete tilfeldige variable: Vektet gjennomsnitt av verdiene x_i med sannsynlighetene p_i som vekter
- For kontinuerlige tilfeldige variable må vi integrere produktet av x og sannsynlighetstettheten til x

Varians til tilfeldige variable

- Teoretisk varians: Idealisert beskrivelse av lang-tids utfall av spredningen/variabiliteten
- For diskrete tilfeldige variable: Vektet gjennomsnitt av $(x_i-\mu_X)^2$-ene med sannsynlighetene som vekter
- For kontinuerlige tilfeldige variable må vi integrere produktet av $(x-\mu_X)^2$ og sannsynlighetstettheten til x

Empirisk korrelasjon r

mellom to observerte variable:

- Gjennomsnittet av produktet av de standardiserte observasjonene for hvert individ

Teoretisk korrelasjon ρ

mellom de stokastiske variablene X og Y:

- En type vektet gjennomsnitt av produktet av de standardiserte tilfeldige variablene
- Tall mellom -1 og 1
- Måler retning og styrke av lineær sammenheng mellom to variable
- Korrelasjonen mellom to uavhengige tilfeldige variable er 0 (men ikke nødvendigvis omvendt!)

Table 5.1 Length (in Minutes) of 60 Visits to a Statistics Help Room

60 tider for samtalelengder (i minutter) hos en «Statistikk-hjelpeservice» ved et universitet

TABLE 5.1	Length (in Minutes) of 60 Visits to a Statistics Help Room								
10	14	15	16	18	20	20	20	23	25
28	30	30	30	30	30	31	33	35	35
46	48	50	50	50	50	51	54	55	55
60	60	60	60	60	60	60	65	65	65
75	77	80	80	84	85	88	98	100	100
105	105	105	115	120	135	135	136	157	210

De 60 konsultasjonstidene er et utvalg fra en større populasjon med 1264 registreringer. Populasjonsfordelinga er svært skjev.

Fra tallene på side 295 kan vi altså få et gjennomsnitt for et utvalg på 60. Å trekke et utvalg på 60 fra populasjonen på 1264 og deretter finne gjennomsnittlig konsultasjonstid kan gjentas mange ganger og man kan lage et histogram av verdiene til de gjentatte gjennomsnittene.

Da ser man noe interessant: Spredninga til histogrammet til gjennomsnittene er mye mindre enn i fordelinga til de 1264 konsultasjonslengdene. Dessuten ser fordelinga symmetrisk ut, og et *kvantilplott* viser at normalfordelinga passer godt.

Fordelinga til lengden av 1264 konsultasjoner

Histogram over gjennomsnittlig lengde på konsultasjoner

Utvalgsgjennomsnitt hvert av 500 utvalg av størrelse 60 fra populasjonen på 1264

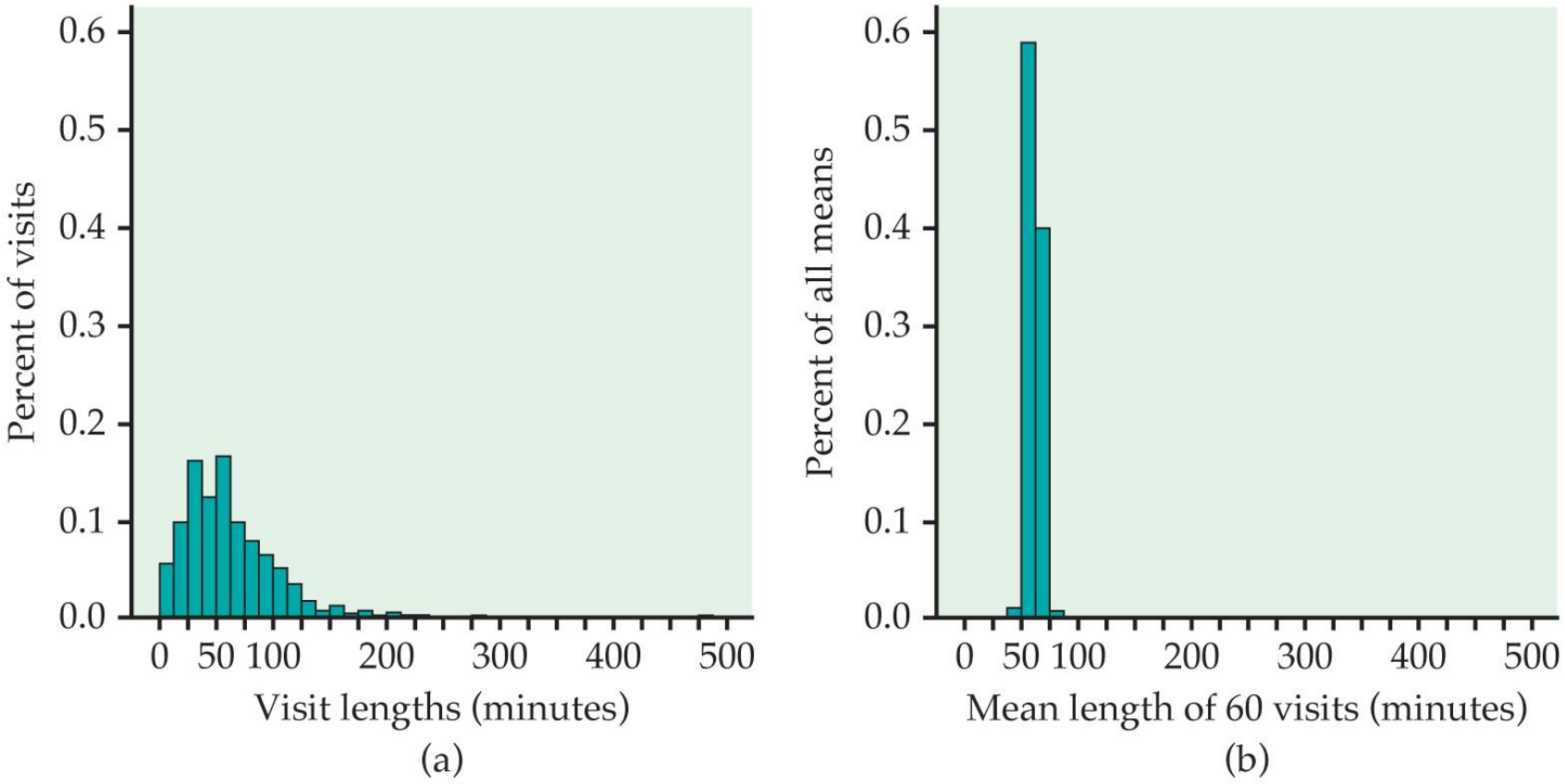


Figure 5.6

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

Kvantilplott for gjennomsnittlig lengde på konsultasjoner (500 gjennomsnitt for 500 utvalg av størrelse 60 fra populasjonen på 1264)

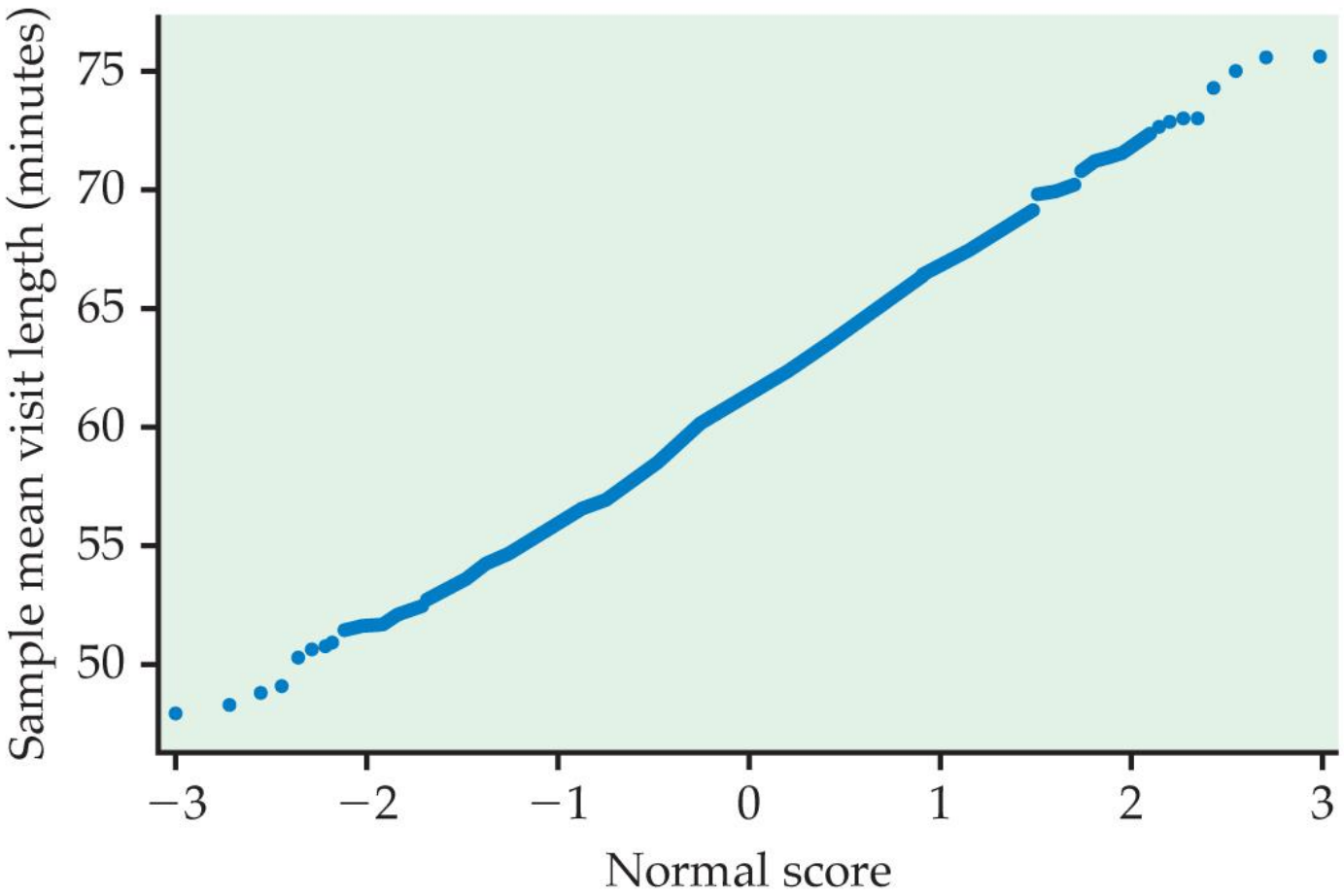


Figure 5.7
Moore/McCabe/Craig, *Introduction to the Practice of Statistics*, 9e, © 2017 W. H. Freeman and Company

Utvalgsfordelinga viser hva som skjer over mange utvalg av størrelse *n* fra den samme populasjonen

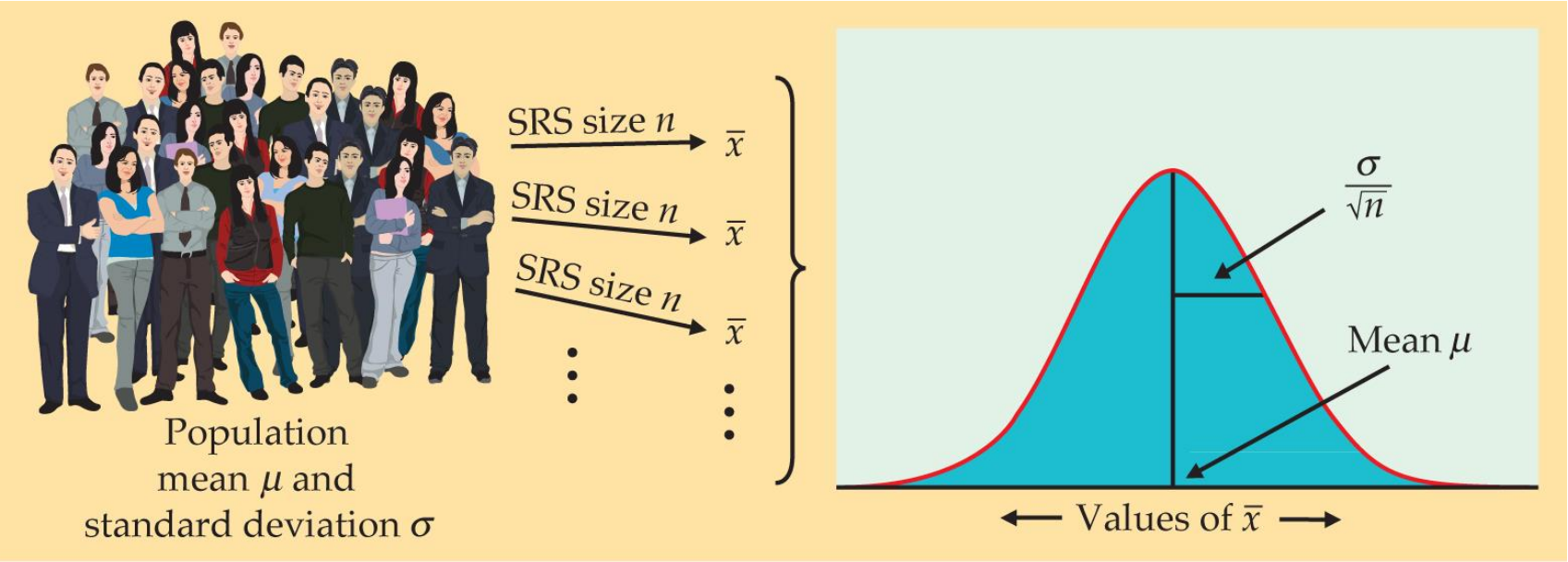


Figure 5.11
Moore/McCabe/Craig, *Introduction to the Practice of Statistics,* 9e, © 2017 W. H. Freeman and Company

Forventninga til gjennomsnittet har samme verdi som forventninga til individuelle observasjoner

Gjennomsnitt

- er mindre variable enn individuelle observasjoner
- er mer normalfordelte enn individuelle observasjoner

Forventning og standardavvik for gjennomsnittet

$$\bar{x} = \frac{1}{n}(X_1 + X_2 + X_3 + \ldots + X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Utgangspunkt: Trekker et enkelt tilfeldig utvalg (**SRS**) av størrelsen *n* fra en populasjon.

De enkelte observasjonene x_i kan sees på utfall av tilfeldige variable X_i . (Husk tilfeldige variable angis med store bokstaver)

Dermed blir gjennomsnittet $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} X_i$ også en tilfeldig variabel.

Gjennomsnittet får altså en fordeling og spesielt en forventning og et standardavvik.

De ulike trekningene X_i er fra samme populasjon og må ha identisk fordeling med samme forventning μ og standardavvik σ.

Dersom populasjonen er mye større enn utvalget så vil ikke verdien av en trekning X_i påvirke verdien av en annen trekning X_j. Vi kan derfor anta at X_1 , …. , X_n er **uavhengige** tilfeldige variable.

Dette er sannsynlighetsmodellen for målinger av hvert individ i et enkelt tilfeldige utvalg (SRS).

Vi skal regne på forventning og varians/standardavvik for gjennomsnittet $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} X_i$ under denne modellen.

Regneregler for forventning og varians

RULES FOR MEANS

Rule 1. If X is a random variable and a and b are fixed numbers, then

$$\mu_{a+bX} = a + b\mu_X$$

Rule 2. If X and Y are random variables, then

$$\mu_{X+Y} = \mu_X + \mu_Y$$

RULES FOR VARIANCES

Rule 1. If X is a random variable and a and b are fixed numbers, then

$$\sigma^2_{a+bX} = b^2\sigma^2_X$$

Rule 2. If X and Y are independent random variables, then

$$\sigma^2_{X+Y} = \sigma^2_X + \sigma^2_Y$$
$$\sigma^2_{X-Y} = \sigma^2_X + \sigma^2_Y$$

This is the **addition rule for variances of independent random variables.**

Rule 3. If X and Y have correlation ρ, then

$$\sigma^2_{X+Y} = \sigma^2_X + \sigma^2_Y + 2\rho\sigma_X\sigma_Y$$
$$\sigma^2_{X-Y} = \sigma^2_X + \sigma^2_Y - 2\rho\sigma_X\sigma_Y$$

This is the **general addition rule for variances of random variables**.

Vi benytter reglene for forventninger fra Kapittel 4. Spesielt hadde vi regelen $\mu_{X+Y} = \mu_X + \mu_Y$

Regelen utvides til $\mu_{X_1+X_2+\ldots+X_n} = \mu_{X_1} + \mu_{X_2} + \cdots + \mu_{X_n}$

I en SRS har alle X_i samme fordeling, derfor blir alle forventningene like, dvs. $\mu_{X_i} = \mu$ for alle i

I en SRS blir dermed $\mu_{X_1+X_2+\ldots+X_n} = n\mu$

Dessuten hadde vi at $\mu_{aX} = a\mu_X$ for en konstant a

Dette gir at forventninga til gjennomsnittet blir

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

altså den samme forventninga som for en enkelt observasjon

Siden gjennomsnittet $\bar{x}$ har forventning μ, sier vi at $\bar{x}$ er en forventningsrett (unbiased) estimator for μ.

Husk at μ er en (typisk ukjent) parameter og må anslås (estimeres).

Merk også at vi bare brukte at alle observasjonene X_i hadde samme forventning i denne utledningen, vi benyttet ikke at de er uavhengige i SRS.

For å finne variansen til gjennomsnittet bruker vi regelen at variansen til en sum *X+Y* gis ved $\sigma^2_{X+Y} = \sigma^2_X + \sigma^2_Y$

altså som summen av variansene – når *X* og *Y* er uavhengige.

Denne kan utvides til $s^2_{X_1+X_2+\ldots+X_n} = s^2_{X_1} + s^2_{X_2} + \cdots + s^2_{X_n}$ med uavhengige X_i.

Siden alle X_i i en en SRS har samme fordeling, blir også alle varianser like, dvs. vi kan skrive $\sigma^2_{X_i} = \sigma^2$.

Dermed blir $s^2_{X_1+X_2+\ldots+X_n} = n\, s^2$.

Vi hadde også regelen $\sigma^2_{aX} = a^2 \sigma^2_X$. Denne leder til at variansen til gjennomsnittet blir lik

$$\sigma^2_{\bar{x}} = \left(\frac{1}{n}\right)^2 n\, \sigma^2 = \frac{1}{n}\sigma^2$$

I en SRS blir dermed standardavviket til gjennomsnittet gitt som

$$\sigma_{\bar{x}} = \frac{1}{\sqrt{n}}\sigma$$

Dermed vil spredninga til gjennomsnittet bli vilkårlig liten når n blir stor nok.

Man refererer ofte til gjennomsnittets standardavvik som "standard error (of the mean)", ofte forkortet SE

Gjennomsnittets fordeling er dessuten sentrert rundt forventninga μ.

$$\bar{x} \to \mu \text{ når } n \to \infty.$$

Dette kjenner vi til fra før gjennom Store talls lov (Kap 4.4)

MEAN AND STANDARD DEVIATION OF A SAMPLE MEAN

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

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

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

Regneeksempel: Statistikk-konsultasjoner

- Forventet lengde på statistikk-konsultasjonene er μ=61.28 minutter og standardavvik σ=41.84 minutter
- Lengden på en gitt konsultasjon kan variere mye fra forventninga.
- Tar et tilfeldig utvalg på 15 konsultasjoner, standardavviket til gjennomsnittet av disse 15 er :

$$s_{\bar{x}} = \frac{1}{\sqrt{15}} s$$

= 10.80 minutter

- Tar så et tilfeldig utvalg på 60 konsultasjoner,

$$s_{\bar{x}} = \frac{1}{\sqrt{60}} s$$

= 5.40 minutter (halvert når 4 ganger så mange)

Fordeling til en lineærkombinasjon av normalfordelte variabler

- At gjennomsnittet av n observerte verdier av uavhengige normalfordelte variabler $X_1, X_2, \ldots, X_n$ også er normalfordelt, er et spesialtilfelle av en generell regel:
  - X er $N(\mu_X, \sigma_X)$-fordelt og Y er $N(\mu_Y, \sigma_Y)$-fordelt
  - X og Y er uavhengige (dvs korrelasjon=0)
  - Da er $V = aX+bY$ (der a og b er faste konstanter) også normalfordelt med
    - forventning $\mu_V = a\mu_X + b\mu_Y$
    - standardavvik $\sigma_V = \sqrt{(a^2\sigma_X^2 + b^2\sigma_Y^2)}$

Eksempel

- Du tar trikken til og fra universitetet. Antall minutter du sitter på trikken hver dag varierer, men
  - X: tiden det tar til UiO har fordeling *N(20,4)*
  - Y: tiden det tar tilbake fra UiO har fordeling *N(18,8)*
- Hvis X og Y er uavhengige, hva er sannsynligheten for at du vil sitte kortere tid på trikken til UiO enn hjem fra UiO? Dvs, hva er P(X < Y) = P(X-Y < 0)
- *X - Y er normalfordelt med forventning*

$$\mu_{X-Y} = \mu_X - \mu_Y = 20 - 18 = 2$$

$$\sigma_{X-Y} = \sqrt{(4^2+(-1)^2\,8^2)} = \sqrt{80} \approx 8.94$$

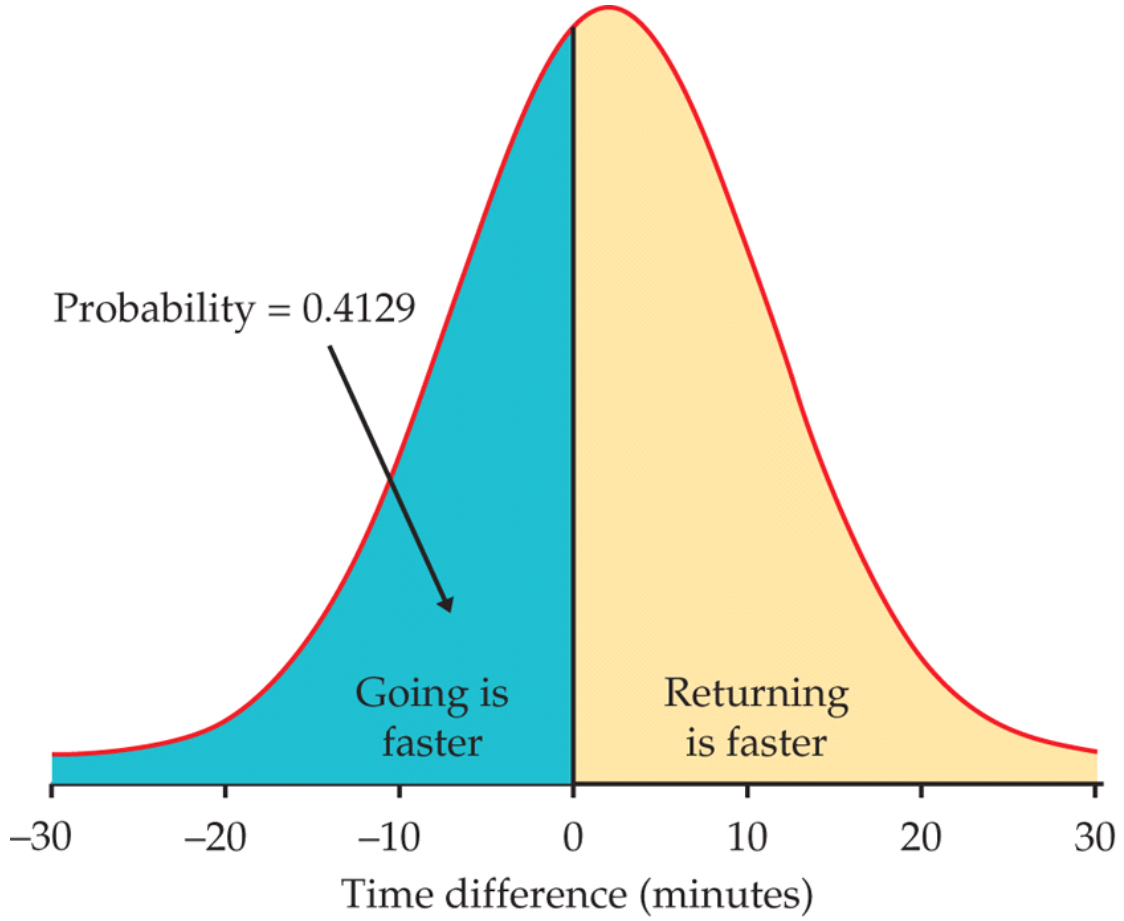
- X - Y har fordeling *N(2, 8.94)*

Eksempel forts.

- $P(X < Y) = P(X-Y < 0)$
- Standardiser

  $Z=(X-Y-2)/\sqrt{80}$
- $X-Y<0$ betyr $Z<-0.22$
- $P(Z < -0.22) = 0.4129$

  (fra tabell)



Fordeling til gjennomsnitt av normalfordelte variabler

- Vi har beskrevet senter (forventninga) og spredning (standardavviket) til fordelinga til gjennomsnittet for et utvalg på størrelse *n*, men ikke formen

- Normalfordelte variable
  - Anta X_i er uavhengige og N(μ,σ)-fordelte
  - Da er gjennomsnittet $\bar{X}$ N(μ,σ/√n)-fordelt (eksakt)

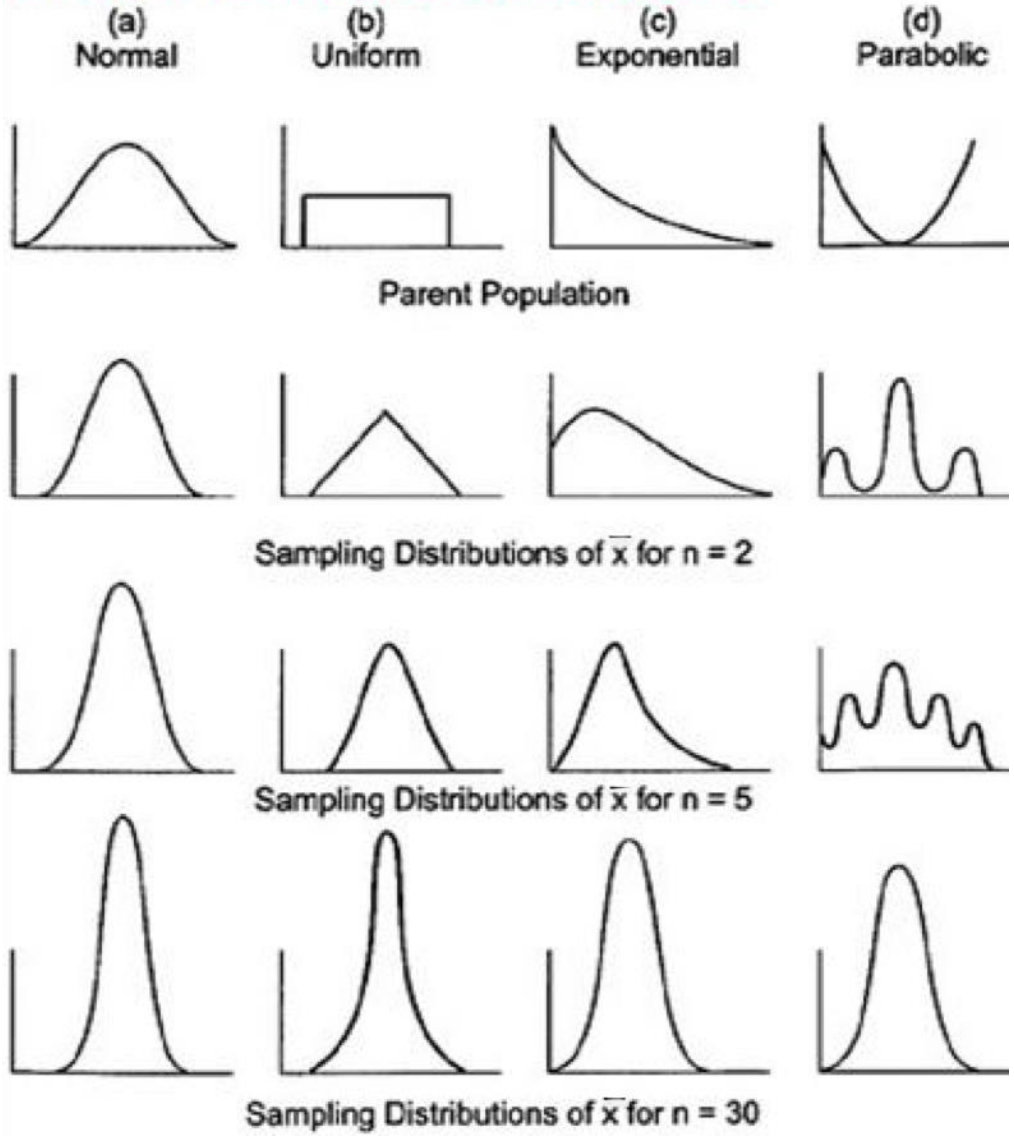
- Dette er et spesialtilfelle av at lineærkombinasjoner av uavhengige normalfordelte variable selv er normalfordelte

Sentralgrenseteoremet for SRS av størrelse n

- Trekk et tilfeldig utvalg av størrelse *n* fra en populasjon med forventning μ og endelig standardavvik σ, dvs.
  - Forventninga til hver X_i er μ
  - Standardavviket til hver X_i er σ
- Antar n stor
  - Da er $\bar{x}$ tilnærma N(μ,σ/√n)-fordelt

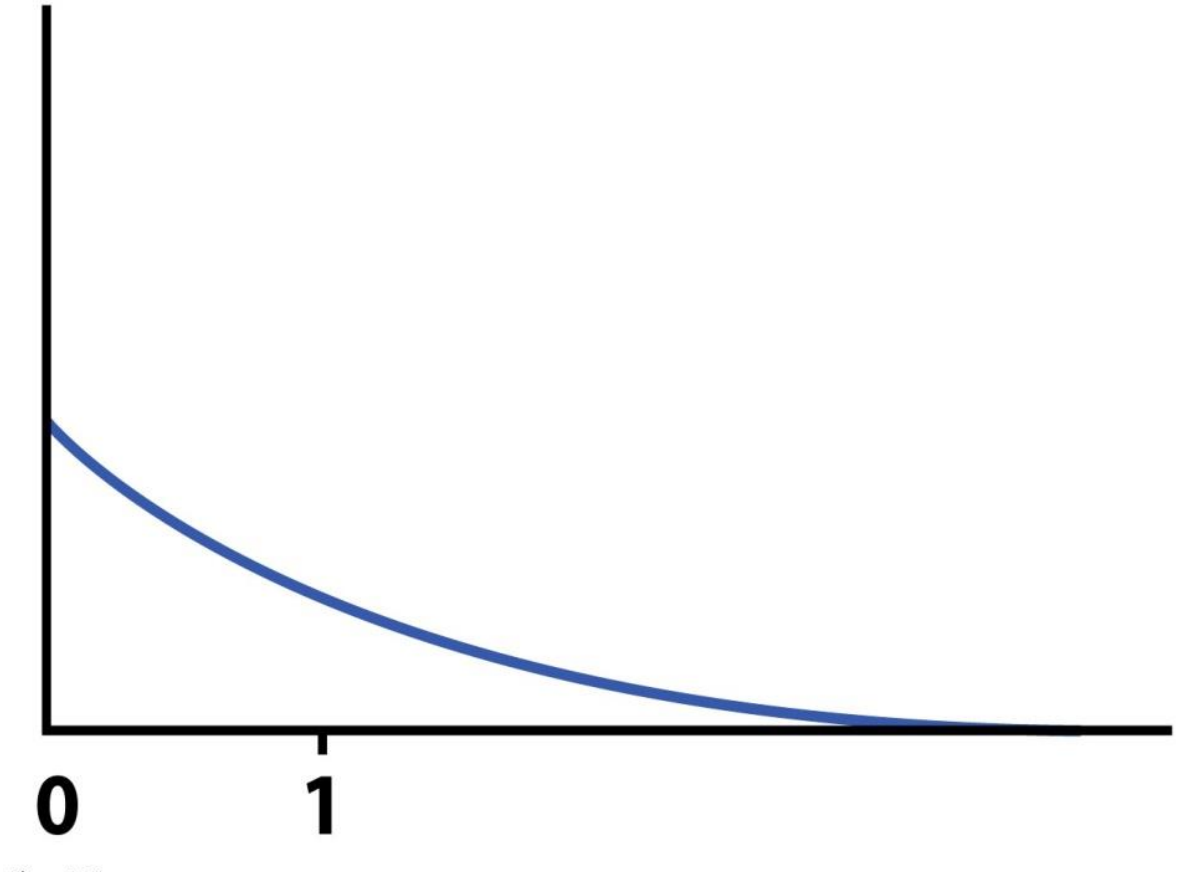
- Merk at det ikke er antatt noen spesiell form på fordelinga til de individuelle X_i :
  når den er forskjellig fra normalfordelinga, er $\bar{x}$ *tilnærma* normalfordelt for stor n

With increasing sample size the distribution of sample means approach the normal distribution irrespective of the distribution of the Parent Population



Eksempel: fordeling langt fra normal
Eksponensiell fordeling

- Mye brukt for ventetider
  - Ventetid i resept-skranken på apotek (?)
  - Lengde mellom påfølgende mutasjoner langs DNA



Fordelinga til gjennomsnittet av N observasjoner fra en eksponensialfordelt populasjon

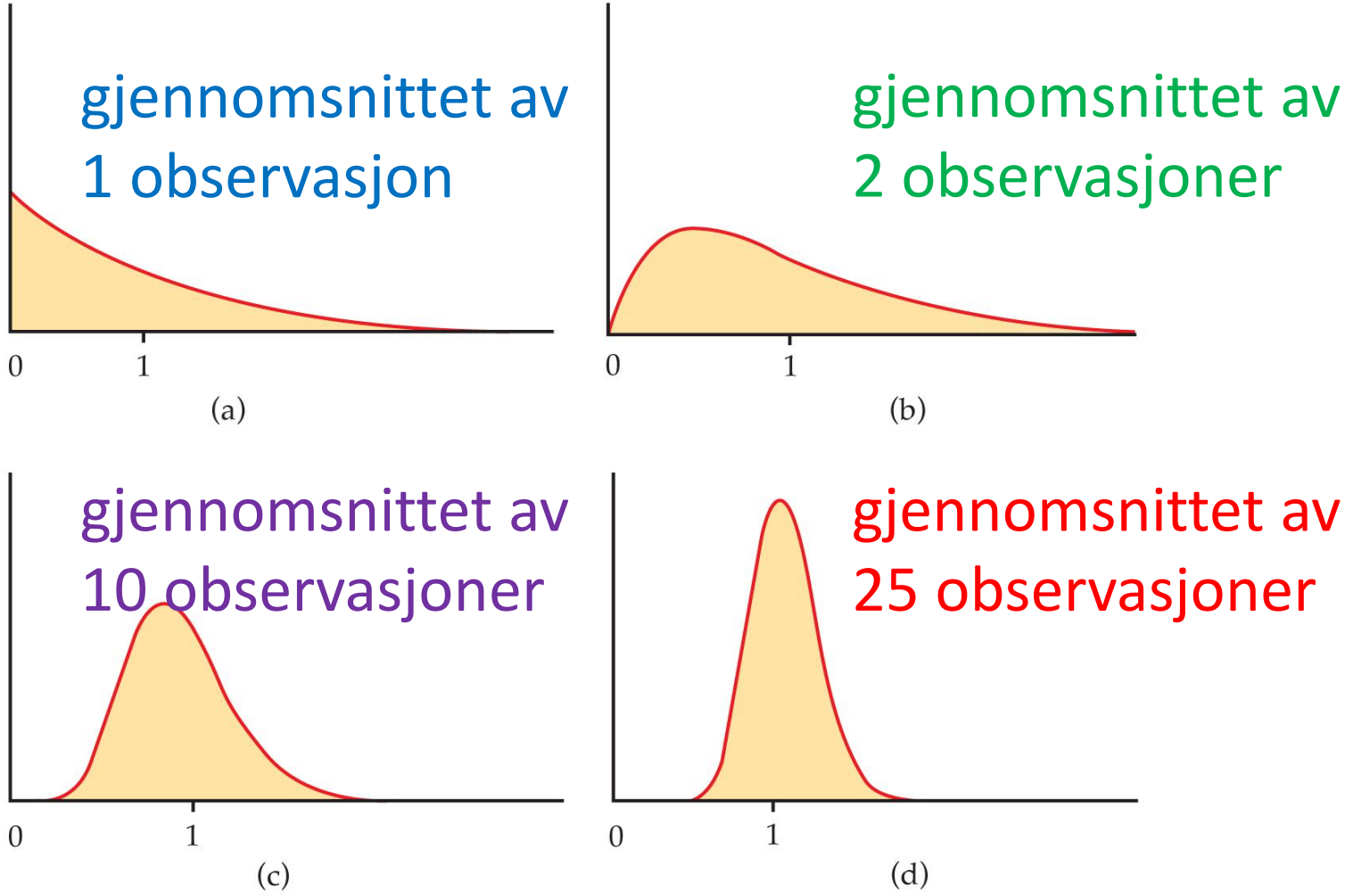


Figure 5.8
Moore/McCabe/Craig, *Introduction to the Practice of Statistics*, 9e, © 2017 W. H. Freeman and Company

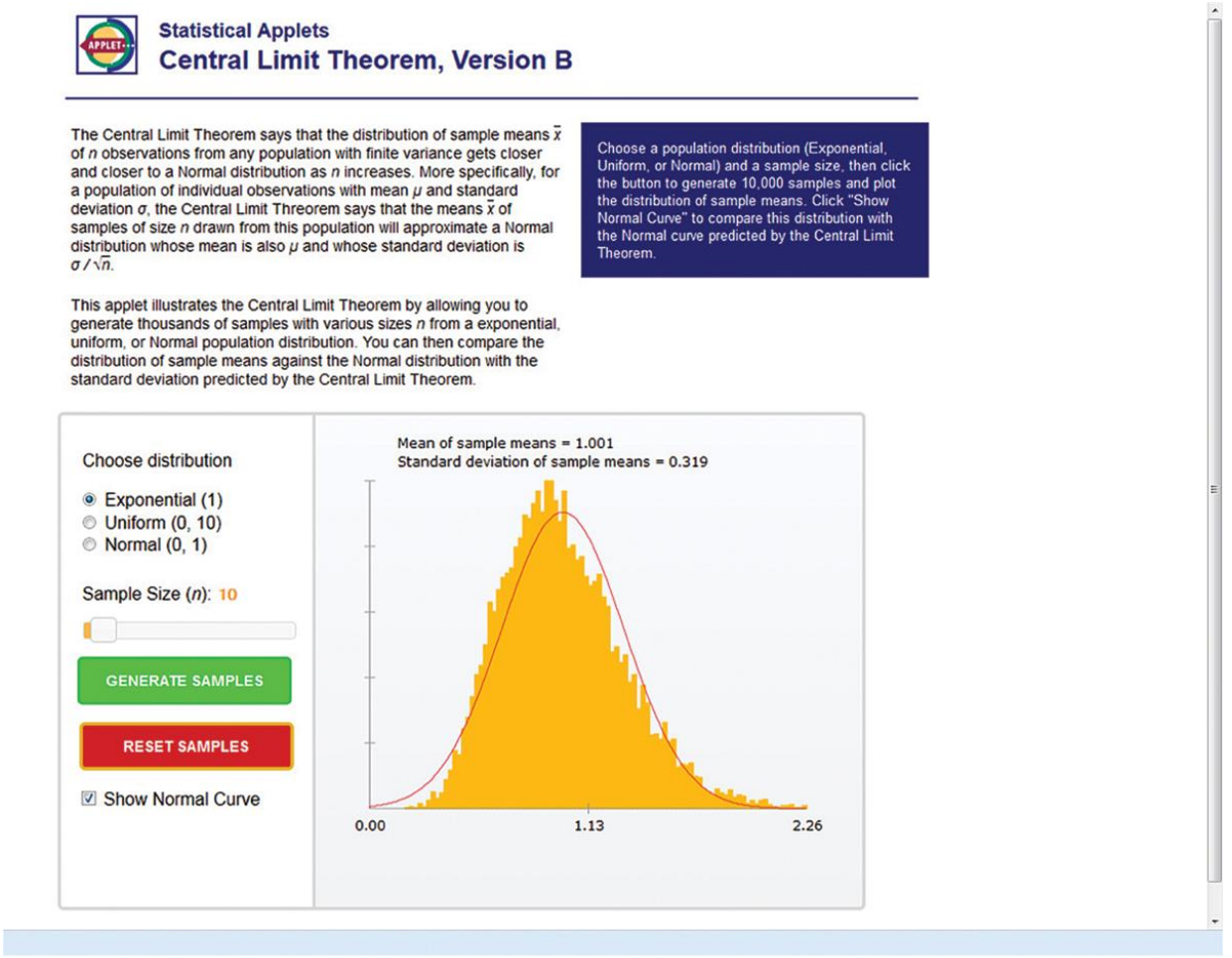


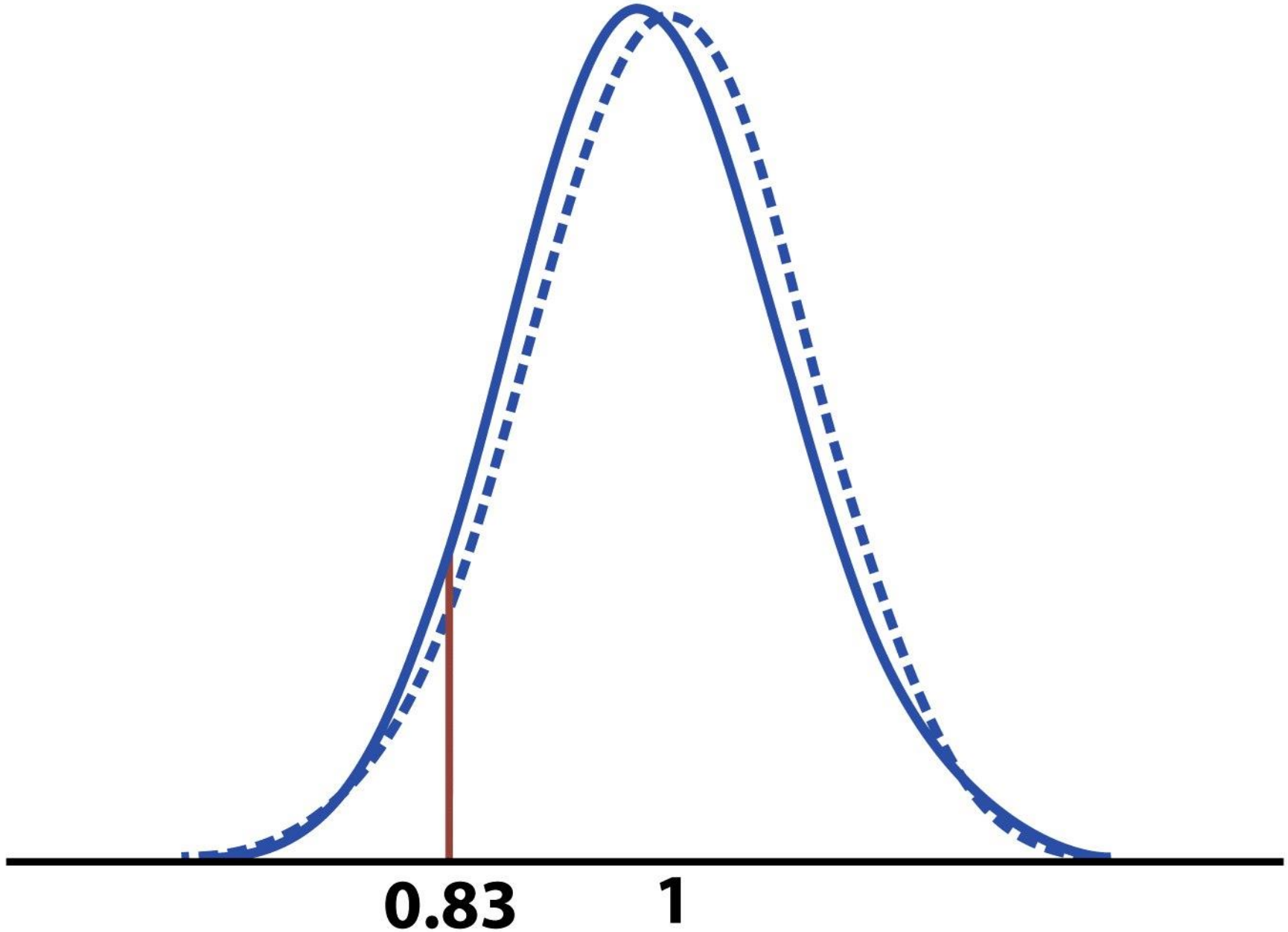
Figure 5.9
Moore/McCabe/Craig, *Introduction to the Practice of Statistics*, 9e, © 2017 W. H. Freeman and Company

Figure 5.9 Screenshot of the *Central Limit Theorem* applet for the exponential distribution when $n = 10$, Example 5.10.

Eksponensialfordeling - regneeksempel

- X: ventetid for å få oversendt elektronisk resept fra fastlege til apotek
- Antar at X har eksponensiell fordeling, forventet tid $\mu = 1$ time, standardavvik $\sigma = 1$ time
- Hva er sannsynligheten for at gjennomsnittlig ventetid for n=70 (uavhengige) resepter er mer enn 50 minutter?
- $P(\bar{x} > 50 \text{ min}) = P(\bar{x} > 0.83 \text{ timer}) = ?$
- Sentralgrenseteoremet: $\bar{x}$ tilnærma $N(1,1/\sqrt{70}) = N(1,0.12)$
- $P(\bar{x} > 0.83) \approx P(Z > (0.83-1)/0.12) \approx 0.92$

- Eksakt fra eksponensiell fordeling (ved bruk av dataprogram): 0.9294


0.83
1

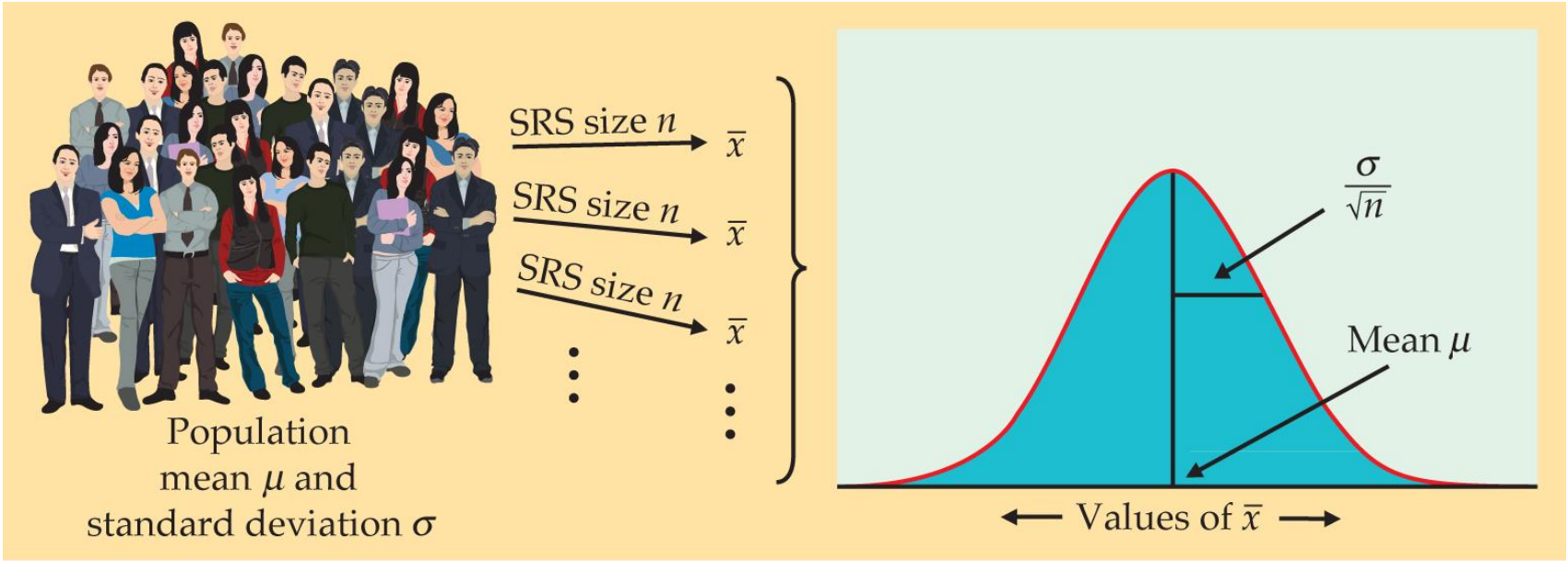


Figure 5.11
Moore/McCabe/Craig, *Introduction to the Practice of Statistics*, 9e, © 2017 W. H. Freeman and Company

1 flervalgsspørsmål

5.3 Utvalgsfordelinger ved tellevariabler og andeler

Antall og andeler

- Binære data (1/0, Ja/Nei, Suksess/Feil)
  - Utvalgsundersøkelser: Ja/Nei-spørsmål
  - Tilstedeværelse av arter: Tilstede/Ikke-tilstede (1/0)
  - Effekt av medisin: Ja/Nei
  - Observator X = Antall Ja (eller antall 1-ere) for utvalget av størrelse *n*
  - Observator $\hat{p}$ = X/n er andel i utvalget med Ja eller med 1-ere

Binomisk oppsett

- Fast antall observasjoner n
- De n observasjonene er uavhengige
- To mulige utfall av hver observasjon:
  - Kalles Suksess/Feil
  - Tilsvarer f.eks. Ja/Nei eller 1/0
- Sannsynlighet p for suksess for hver av de n observasjonene

Binomisk fordelte data

- Myntkast med idealisert mynt
  - Kaster en mynt n=10 ganger
  - To mulige utfall: Kron eller mynt
  - Sannsynlighet for kron er p=0.5

- Genetikk tilsier at barn av samme foreldre får gener fra foreldrene uavhengig av hverandre
  - To foreldre får n=5 barn sammen
  - Hvert barn av disse foreldrene har sannsynlighet p=0.25 for å få blodtype 0
  - Lager oss en binomisk setting med to muligheter: Blodtype 0(«suksess») eller ikke blodtype 0 («feil»)

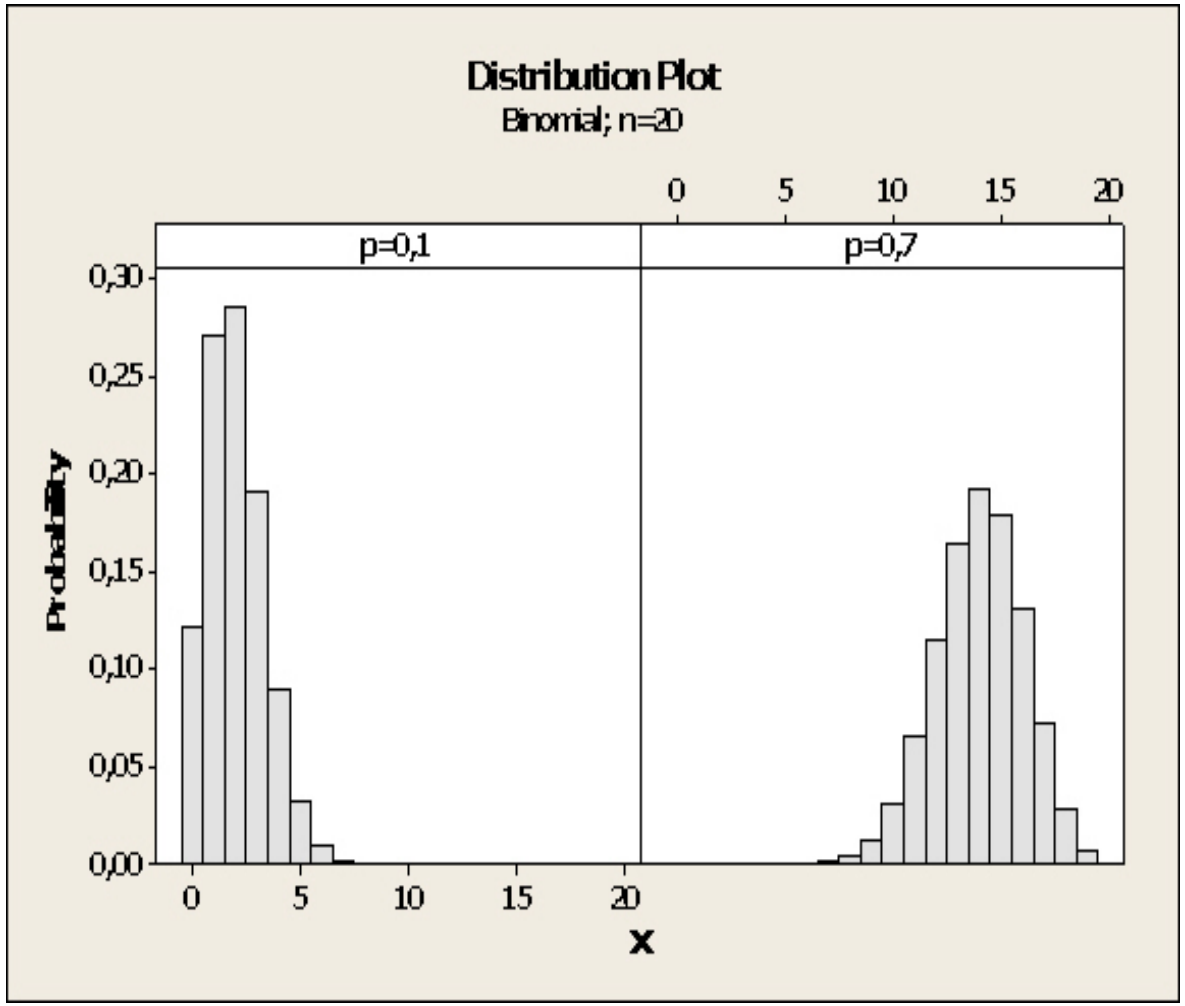
Binomisk fordeling

- Fordeling til antallet *X* av suksesser i en binomisk setting
  - Binomisk fordeling med parametre *n* (antall observasjoner) og *p* (sannsynligheten for suksess for hver observasjon)
  - Utfallsrom {*0,1,…,n*}
  - X er Bin(n,p)-fordelt
- Viktig *diskret* fordeling (sannsynlighetsfordeling for en diskret tilfeldig variabel X)

Binomisk fordelte data

- Myntkast med idealisert mynt
  - Kaster en mynt n=10 ganger
  - Sannsynlighet for kron er p=0.5
  - X = Antall kron i de 10 kastene (antallet suksesser)
  - X er Bin(10,0.5)-fordelt
- Genetikk tilsier at barn av samme foreldre får gener fra foreldrene uavhengig av hverandre
  - To foreldre får n=5 barn sammen
  - Hvert barn disse foreldrene har sannsynlighet p=0.25 for å få blodtype 0
  - X = Antall barn som får blodtype 0 (antallet suksesser)
  - X er Bin(5,0.25)-fordelt

Binomisk fordeling: Sannsynlighets-histogrammer



1 flervalgsspørsmål

Eksempel

Effekt av behandling

- *p* er sannsynligheten for at pasienter har effekt av behandling (populasjonsparameter)
- *n* pasienter
- *X* pasienter har effekt
- *observator* $\hat{p} = X/n$ er andel med effekt, estimat på *p*
- Anta sannsynlighet for å ha effekt av gammel behandling er $p_{gammel} = 0.5$
- Prøver ny behandling på 100 pasienter, 60 har effekt: $n=100$, $X=60$, $\hat{p} = 0.6$.
- Kan vi si at ny behandling er bedre, dvs at $p > p_{gammel}$?
- Bruker sannsynlighetsfordeling

Utvalgsfordeling for antall suksesser

- Populasjon av størrelse *N*, andel suksess i populasjonen *p*
- Utvalg av størrelse *n,* observatoren *X* er antall suksesser i utvalget
- Utfall av 2. observasjon avhenger av utfall av 1. observasjon
- Eksempel
  - *N*=52 kort, trekker *n*=2 kort
  - P(1. Rødt)=26/52=0.5, P(2. Rødt | 1. Rødt)=25/51<0.5
- Avhengighet mellom observasjonene
- MEN: Hvis *N* mye større enn *n (N>20n)*, kan man neglisjere slik avhengighet, og X er tilnærma Bin(n,p)-fordelt
- Presisjonen til denne tilnærminga er bedre jo større forholdet N/n er

Binomiske sannsynligheter

- Fordeling til antallet *X* av suksesser
  - Fordeling med parametre *n* (antall observasjoner) og *p* (sannsynligheten for suksess for hver observasjon)
  - Utfallsrom {*0,1,...,n*}
  - X er Bin(n,p)-fordelt
- Sannsynligheten for at X=i, for i=0,1,...,n kan finnes i tabell (Table C i boken) eller ved å bruke dataprogram som R
  - Avhenger kun av *n* og *p*, dvs for gitt *n* og *p* er sannsynligheten for at X=i bestemt
  - Eksempel: n=6, p=0.35, da er P(X=2)=0.3280

Binomiske sannsynligheter: Eksempel

- Genetikk tilsier at barn av samme foreldre får gener fra foreldrene uavhengig av hverandre
  - To foreldre får n=5 barn sammen
  - Hvert barn disse foreldrene får har sannsynlighet p=0.25 for å få blodtype 0
  - X=Antall barn som får blodtype 0 (antallet suksesser)
  - X er Bin(5,0.25)-fordelt
- Hva er sannsynligheten for at minst 2 av barna får blodtype 0?
- $P(X \geq 2)=1-P(X<2) = 1-P(X \leq 1) = 1 - (P(X=0)+P(X=1)) = ?$
- Bruk Tabell C eller R

TABLE C

Binomial probabilities (continued)

Entry is $P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$

		p								
n	k	.10	.15	.20	.25	.30	.35	.40	.45	.50
2	0	.8100	.7225	.6400	.5625	.4900	.4225	.3600	.3025	.2500
	1	.1800	.2550	.3200	.3750	.4200	.4550	.4800	.4950	.5000
	2	.0100	.0225	.0400	.0625	.0900	.1225	.1600	.2025	.2500
3	0	.7290	.6141	.5120	.4219	.3430	.2746	.2160	.1664	.1250
	1	.2430	.3251	.3840	.4219	.4410	.4436	.4320	.4084	.3750
	2	.0270	.0574	.0960	.1406	.1890	.2389	.2880	.3341	.3750
	3	.0010	.0034	.0080	.0156	.0270	.0429	.0640	.0911	.1250
4	0	.6561	.5220	.4096	.3164	.2401	.1785	.1296	.0915	.0625
	1	.2916	.3685	.4096	.4219	.4116	.3845	.3456	.2995	.2500
	2	.0486	.0975	.1536	.2109	.2646	.3105	.3456	.3675	.3750
	3	.0036	.0115	.0256	.0469	.0756	.1115	.1536	.2005	.2500
	4	.0001	.0005	.0016	.0039	.0081	.0150	.0256	.0410	.0625
5	0	.5905	.4437	.3277	.2373	.1681	.1160	.0778	.0503	.0313
	1	.3280	.3915	.4096	.3955	.3602	.3124	.2592	.2059	.1563
	2	.0729	.1382	.2048	.2637	.3087	.3364	.3456	.3369	.3125
	3	.0081	.0244	.0512	.0879	.1323	.1811	.2304	.2757	.3125
	4	.0004	.0022	.0064	.0146	.0284	.0488	.0768	.1128	.1562
	5		.0001	.0003	.0010	.0024	.0053	.0102	.0185	.0312
6	0	.5314	.3771	.2621	.1780	.1176	.0754	.0467	.0277	.0156
	1	.3543	.3993	.3932	.3560	.3025	.2437	.1866	.1359	.0938
	2	.0984	.1762	.2458	.2966	.3241	.3280	.3110	.2780	.2344
	3	.0146	.0415	.0819	.1318	.1852	.2355	.2765	.3032	.3125
	4	.0012	.0055	.0154	.0330	.0595	.0951	.1382	.1861	.2344
	5	.0001	.0004	.0015	.0044	.0102	.0205	.0369	.0609	.0937
	6			.0001	.0002	.0007	.0018	.0041	.0083	.0156
7	0	.4783	.3206	.2097	.1335	.0824	.0490	.0280	.0152	.0078
	1	.3720	.3960	.3670	.3115	.2471	.1848	.1306	.0872	.0547
	2	.1240	.2097	.2753	.3115	.3177	.2985	.2613	.2140	.1641
	3	.0230	.0617	.1147	.1730	.2269	.2679	.2903	.2918	.2734
	4	.0026	.0109	.0287	.0577	.0972	.1442	.1935	.2388	.2734
	5	.0002	.0012	.0043	.0115	.0250	.0466	.0774	.1172	.1641
	6		.0001	.0004	.0013	.0036	.0084	.0172	.0320	.0547
	7				.0001	.0002	.0006	.0016	.0037	.0078
8	0	.4305	.2725	.1678	.1001	.0576	.0319	.0168	.0084	.0039
	1	.3826	.3847	.3355	.2670	.1977	.1373	.0896	.0548	.0313
	2	.1488	.2376	.2936	.3115	.2965	.2587	.2090	.1569	.1094
	3	.0331	.0839	.1468	.2076	.2541	.2786	.2787	.2568	.2188
	4	.0046	.0185	.0459	.0865	.1361	.1875	.2322	.2627	.2734
	5	.0004	.0026	.0092	.0231	.0467	.0808	.1239	.1719	.2188
	6		.0002	.0011	.0038	.0100	.0217	.0413	.0703	.1094
	7			.0001	.0004	.0012	.0033	.0079	.0164	.0312
	8					.0001	.0002	.0007	.0017	.0039

(*Continued*)

Binomiske sannsynligheter: Eksempel

- X er Bin(5,0.25)-fordelt
- Hva er sannsynligheten for at minst 2 av barna får blodtype 0?
- $P(X \geq 2) = 1 - P(X < 2) = 1-P(X \leq 1) = 1 - (P(X=0)+P(X=1))$
  $= 1 – 0.2373 – 0.3955 = 0.3672$
- I R skriv:

```
> 1-pbinom(1,5,0.25)
[1] 0.3671875
```

- Eller:

```
> pbinom(1,5,0.25,lower.tail=F)
[1] 0.3671875
```

Binomisk fordeling i R, X Bin(n,p)

- Kommandoen `pbinom(k,n,p)` gir P(X≤k)
- Kommandoen `pbinom(k,n,p,lower.tail=F)` gir P(X>k)
- Kommandoen `dbinom(k,n,p)` gir P(X=k)

```
> dbinom(0:5,5,0.25)
[1] 0.2373 0.3955 0.2637 0.0879 0.0146 0.0010
```

(her for P(X=0), P(X=1), P(X=2), P(X=3), P(X=4) og P(X=5))

- Kommandoen `qbinom(q,n,p)` gir persentiler (kvantiler) i bin(n,p) fordelinga
- Kommandoen `rbinom(m,n,p)` trekker m observasjoner ”tilfeldig” fra bin(n,p) fordelinga

Table C: Bare for $p \leq 0.5$

- Dersom man ser etter sannsynlighetsfordelinga til X som er binomisk fordelt med $p>0.5$:
  - Snu om på situasjonen slik Y teller antallet feil (i stedet for suksesser)
  - Da blir $p<0.5$ for Y som teller antall feil
  - Eksempel:
    - Antall barn som ikke har blodtype 0 er Bin(5,0.75)-fordelt
    - Antall barn som har blodtype 0 er Bin(5,0.25)-fordelt
- Tenk alltid nøye igjennom hva man teller som suksess og hva den riktige p er da!

Forventning i binomisk fordeling

- Anta X er $Bin(n,p)$
- La S_i være en binær tilfeldig variabel som indikerer om observasjon *i* er en suksess ($S_i=1$) eller ikke ($S_i=0$)
- Da er $X=S_1+S_2+...+S_n$ (antallet suksesser)
- $P(S_i=1) = p = 1-P(S_i=0)$
- S_i-ene har samme fordeling, og forventninga til hver S_i er $\mu_S=1*p + 0*(1-p) = p$
- Forventninga til X er $\mu_X=\mu_S+\mu_S+...+\mu_S=np$

Varians og standardavvik i binomisk fordeling

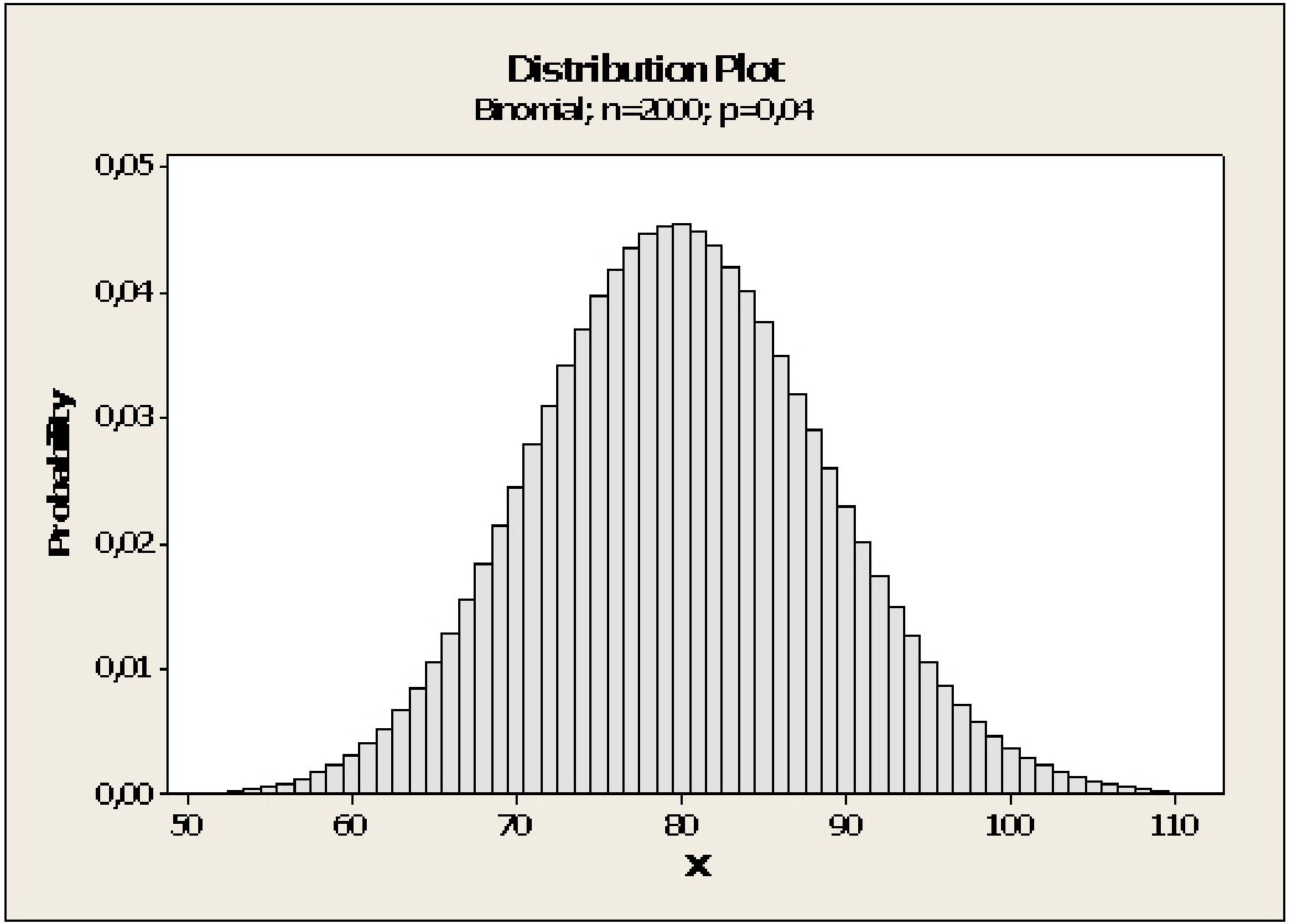
- S_i-ene har samme fordeling, og variansen og standardavviket til hver S_i er:

  $\sigma_S^2 = (1\text{-}p)^2\ p + (0\text{-}p)^2\ (1\text{-}p) = p(1\text{-}p)$

  $\sigma_S = \sqrt{[p(1\text{-}p)]}$
- $X = S_1 + S_2 + \ldots + S_n$
- S_i-ene er uavhengige av hverandre (binomisk setting), dvs alle parvise korrelasjoner er 0
- $\sigma_X^2 = \sigma_S^2 + \sigma_S^2 + \ldots + \sigma_S^2 = n\ p(1\text{-}p)$
- $\sigma_X = \sqrt{[np(1\text{-}p)]}$

Antikolesterol og hjerteinfarkt

- Menn i alder 40-55, p=0.04 for hjerteinfarkt
  - 2000 menn får legemiddelet Gemfibrozil
  - 2000 menn får placebo
- Hva er på forhånd forventet antall hjerteinfarkt blant 2000 menn dersom Gemfibrozil ikke har effekt?
- p=0.04: $\mu_X=np=80$ $\sigma_X^2=np(1-p)=76.8$, $\sigma_X=8.76$
- Observert i studie:
  - Placebo: x=84
  - Gemfibrozil: x=56
- Ser ut til at Gemfibrozil reduserer risikoen for hjerteinfarkt


Distribution Plot
Binomial; n=2000; p=0,04
Probability
0,05
0,04
0,03
0,02
0,01
0,00
50
60
70
80
90
100
110
X

2 flervalgsspørsmål

Andeler

- $\hat{p} = X/n$ = antall suksesser/størrelse av utvalg
  = andelen suksesser i utvalget
- $\hat{p}$ er estimator for andelen suksesser i populasjonen
- X tar heltallsverdier mellom 0 og *n* og er Bin(n,p)-fordelt
- $\hat{p}$ tar verdier i intervallet [0,1] og er *ikke* binomisk fordelt!
- Men kan bruke forventning og varians til X til å finne forventning og varians til $\hat{p}$:
  - $\mu_{\hat{p}} = \mu_{X/n} = (1/n)\,\mu_X = np/n = p$ - Forventningsrett estimator for p!
  - $\sigma_{\hat{p}}^2 = \sigma_{X/n} = (1/n)^2\,\sigma_X^2 = np(1-p)/n^2 = p(1-p)/n$
  - $\sigma_{\hat{p}} = \sqrt{[p(1-p)/n]}$
- Variasjonen (usikkerheten) minker med økende n!
- $\sqrt{n}$ i nevneren for standardavviket betyr at for å halvere standardavviket til $\hat{p}$, må vi firedoble utvalgsstørrelsen *n*

Antikolesterol og hjerteinfarkt

- Observert x = 56, $\hat{p}$ = 56/2000 = 0.028
- Anta p=0.04, n=2000. Hva er $\mu_{\hat{p}}$, $\sigma_{\hat{p}}^2$ og $\sigma_{\hat{p}}$?

  Med andre ord, hva er forventning, varians og standardavvik for andelen $\hat{p}$?

Dette spørsmålet er essensielt for å avgjøre om forskjellen mellom $\hat{p}$ og p=0.04 er statistisk signifikant

(Mer om dette i kapittel 6)

Antikolesterol og hjerteinfarkt

- Observert x=56, $\hat{p} = 0.028$
- Dersom vi antar at p=0.04:
  - $\mu_{\hat{p}} = 0.04$, $\sigma_{\hat{p}}^2 = 0.0000192$, $\sigma_{\hat{p}} = \sqrt{[p(1-p)/n]} = 0.0043$
- Hva hvis vi lurer på $P(\hat{p} \leq 0.028)$? Eller ekvivalent $P(X \leq 56)$?
- $\hat{p}$ er ikke binomisk fordelt, men vi kan utnytte at X er Bin(2000,0.04)-fordelt:

$$P(\hat{p} \leq 0.028) = P(X \leq 56) = P(X=0)+P(X=1)+\cdots+P(X=56)$$
$$= 0.002497$$

- Fullt mulig å gjøre, men litt tungvint

Tilnærming til normalfordeling

- *X* er Bin(*n,p*)-fordelt og *n* er stor. Da gjelder

$$\text{X} \ \text{ tilnærmet } \ N(np\ ,\ \sqrt{np(1-p)}\ )$$

$$\hat{p} \ \text{ tilnærmet } \ N(p\ ,\ \sqrt{\frac{p(1-p)}{n}}\ )$$

- Kan brukes for å beregne sannsynligheter
- Rimelig tilnærming når np>10 og n(1-p)>10

Kolesterol og hjerteinfarkt

- p=0.04, n=2000, X=56
- Her np = 80 og n(1-p)=1920, så normaltilnærmingen er OK å bruke
- X er tilnærmet N(np,√(np(1-p))-fordelt
- Vet fra før: $\mu_X = np = 80$ $\sigma_X = \sqrt{[np(1-p)]} = 8.76$
- Standardiserer: $Z = (X-np)/\sqrt{[np(1-p)]} = (X-80)/8.76$, da er Z tilnærmet N(0,1)-fordelt
- Dermed: $P(X \leq 56) \approx P(Z \leq (56-80)/8.76) = P(Z \leq -2.74) = 0.0031$
  - Dvs ganske nær eksakt verdi basert på binomisk fordeling for X, som var $P(X \leq 56) = 0.002497$

Begrunnelse: Andeler/antall og normalfordeling

$$X \text{ tilnærmet } N(np\ ,\ \sqrt{np(1-p)}\ )$$

$$\hat{p} \text{ tilnærmet } N(p\ ,\ \sqrt{\frac{p(1-p)}{n}}\ )$$

- S_i =1 hvis suksess, 0 ellers
- $X = S_1 + S_2 ... + S_n$ = antall suksess
- $\hat{p} = X/n = \overline{S}$ - altså gjennomsnittet av S_i'ene
- Så $\hat{p}$ tilnærmet normalfordelt følger av sentralgrenseteoremet
- $X = n\,\hat{p}$, og en konstant multiplisert med en normalfordelt variabel er også normalfordelt.

Utvalgsfordeling: Gir svaret på hva som ville skje dersom vi så på mange utvalg med størrelse *n* fra den samme populasjonen

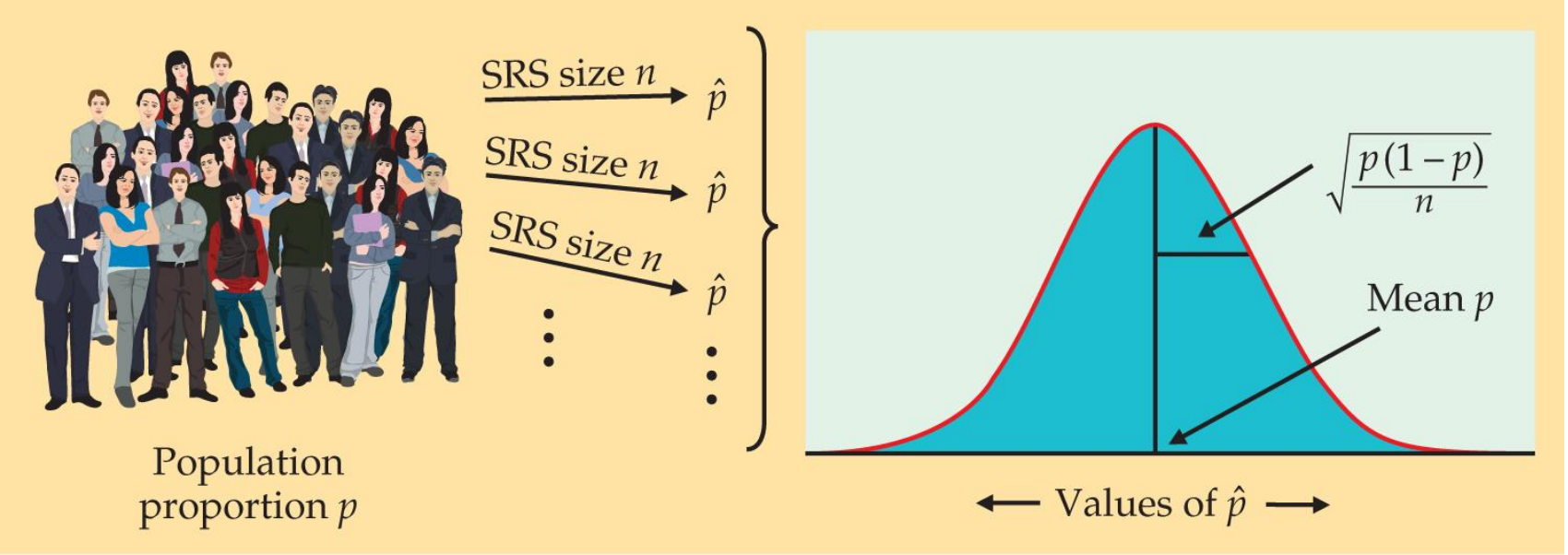


Figure 5.17
Moore/McCabe/Craig, *Introduction to the Practice of Statistics*, 9e, © 2017 W. H. Freeman and Company

Normalapproksimasjon for binomisk fordeling

- Normaltilnærmingen best når *p* nær 0.5, minst eksakt for *p* nær 0 eller 1

- Figur: Sannsynlighetshistogram og normalfordelings-tilnærmingen når X er Bin(150,0.08)-fordelt

- Sannsynlighetshistogrammet er litt høyreskjevt, noe normalfordelinga ikke kan fange opp (Her: np=12, n(1-p)=138)

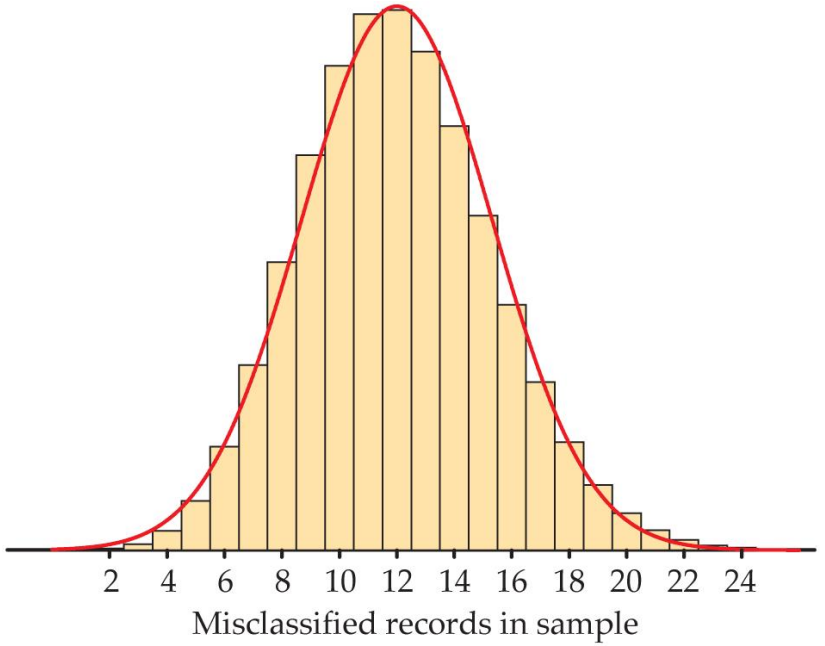


Figure 5.19
Moore/McCabe/Craig, *Introduction to the Practice of Statistics*, 9e, © 2017 W. H. Freeman and Company

Poisson-oppsett

Ganske ofte møter vi tellevariable som kan gå i det uendelige, i hvert fall i teoretisk forstand:

- Antallet kunder på en restaurant mellom kl. 19.00 og 22.00.
- Antallet elg som krysser en vei
- Antallet trafikkulykker i et år

Et **Poisson-oppsett** oppstår når vi ser på antallet suksesser som skjer innenfor en måleenhet. Eksempler på måleenheter kan være en tidsperiode eller et geografisk område.

Betingelser for Poissonfordeling

1. Antallet suksesser som skjer i to ikke-overlappende måleenheter er **uavhengige**.
2. Sannsynligheten for at en suksess inntreffer inne i en måleenhet er den samme for alle enheter av samme størrelse. Sannsynligheten er da også proporsjonal med størrelsen til enheten.
3. Sannsynligheten for at mer enn én hendelse skjer innenfor en enhet er neglisjerbar for veldig små enheter. Med andre ord, hendelsene skjer én av gangen.

Poissonfordeling

Poissonsannsynlighet

Fordelinga til antallet X av suksesser i et Poisson-oppsett er **Poissonfordelinga** med **forventning μ**. Parameteren μ er forventet antall suksesser per måleenhet.

De mulige verdiene til X er heltallene 0, 1, 2, 3, …

Hvis k er hvilket som helst ikke-negativt heltall er:

$$P(X = k) = \frac{e^{-m} m^{k}}{k!}.$$

Standardavviket til fordelinga er $\sqrt{\mu}$.

Poissonfordeling: Eksempel

Anta at antallet brutte samtaler på en mobiltelefon varierer, men med en forventning på 2.1 samtaler per dag. Hvis vi antar at det er rimelig at dette er et Poisson-forsøk, kan vi modellere det daglige antallet *X* ved å bruke en Poissonfordeling med *μ* = 2.1. Hva er sannsynligheten for å ikke få mer enn to brutte samtaler i morgen?

$$P(X=k)=\frac{e^{-m}m^{k}}{k!}.$$

$$P(X \pounds 2)=P(X=0)+P(X=1)+P(X=2)$$

$$=\frac{e^{-2.1}(2.1)^{0}}{0!}+\frac{e^{-2.1}(2.1)^{1}}{1!}+\frac{e^{-2.1}(2.1)^{2}}{2!}$$

$$=0.1225+0.2572+0.2700=0.6497$$

Det er cirka 65% sjanse for at det ikke vil være mer enn to brutte samtaler i morgen.

Oppsummering

- Repetisjon tilfeldige variable
- Gjennomsnitt som observator
  - Forventning, varians og utvalgsfordeling
- Sentralgrenseteoremet
- Binomisk setting:
  - Forventning, varians og utvalgsfordeling til antall
  - Forventning, varians og utvalgsfordeling til andel

Oversikt pensum: fortid, nåtid og fremtid

- Eksplorativ data-analyse (Kap 1, 2)
- Hvordan ”produsere” data (Kap 3)
- Teori om sannsynlighet (Kap 4)
- Utvalgsfordelinger til observatorer (Kap 5)
- Introduksjon til inferens om ukjente parametre (Kap 6)
  - Konfidensintervaller
  - Hypotesetesting
- Inferens om
  - Forventning (Kap 7)
  - Regresjonsmodeller (Kap 10,11,14)