

# Utvalgsfordelinger

Vi har sett at utvalgsfordelinger til en statistikk (observator) er fordelingen av verdiene statistikken tar ved mange gjentatte utvalg av samme størrelse fra samme populasjon.

Utvalg er 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 utvalgsfordelingen til de vanligste statistikkene.

Å trekke tilfeldige utvalg fra endelige populasjonen som eksisterer kan utføres i praksis.

Men det er ikke alltid populasjonen faktisk eksisterer. F. eks. finnes populasjonen av alle STK1000 – H15 studenter, men populasjonen av alle STK1000 studenter, tidligere, nåværende og framtidige finnes ikke.

Hvis jeg for eksempel var interessert i å finne ut hvor mange timer STK1000 studenter vanligvis arbeider med ukeoppgavene og spør dere, er det naturlig å se svarene som svar fra tilfeldige representanter for denne hypotetiske populasjonen. Fordelingen av svarene vil da følge en sannsynlighetsfordeling.

**Fordeling til en statistikk:** En statistikk fra et tilfeldig utvalg eller randomisert eksperiment er en tilfeldig variabel. Sannsynlighetsfordelingen til statistikken er utvalgsfordelingen.

Men utvalgene må trekkes fra en populasjon, og nå antar vi at også enhetene i populasjonen har en tilfeldig fordeling.

**Populasjonsfordeling:** Populasjonsfordelingen til en variabel er fordelingen av verdien til alle enhetene i populasjonen. Den er også sannsynlighetsfordelingen til variabelen når man trekker en enhet tilfeldig fra populasjonen.

# Fordeling til gjennomsnitt i et utvalg

For kvantitative data brukes statistikker som (empirisk) gjennomsnitt, andeler, persentiler og (empirisk) standardavvik.

Alle disse er statistikker og har en utvalgsfordeling. Vi skal i første omgang konsentrere oss om gjennomsnittet av et sett observasjoner.

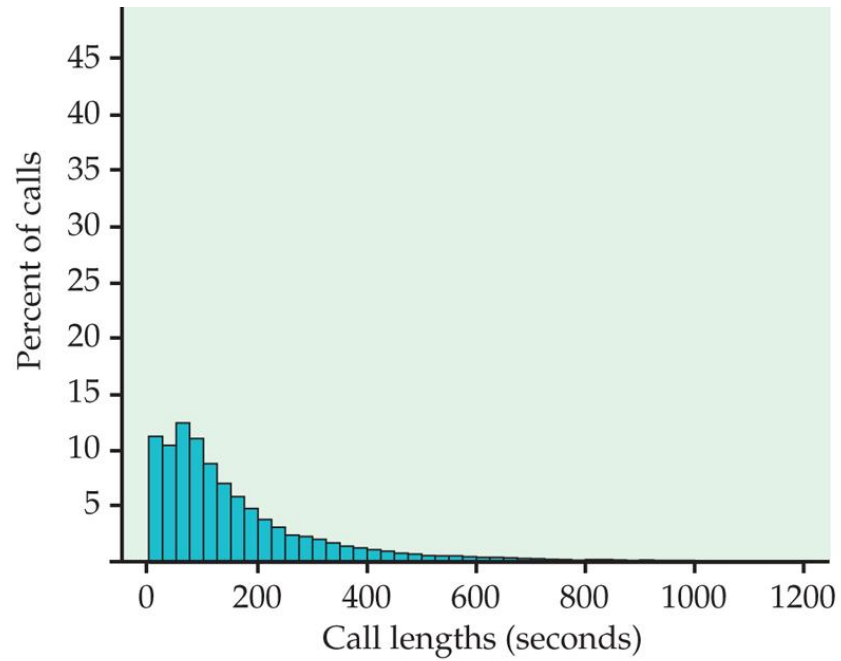
Gjennomsnitt den vanligste statistikken.

## **Eksempel:** Samtalelengder

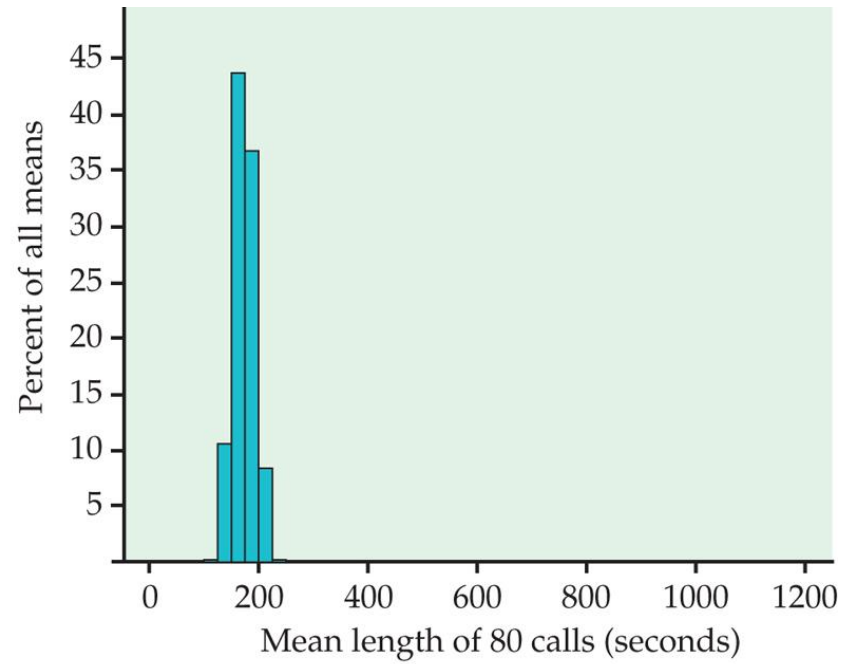
På side 19 i læreboka er 80 tider for samtalelengder ved et «callcenter» gjengitt. Samtalelengdene er et utvalg fra et større datamateriale med 30 000 enheter. Et histogram for denne fordelingen er svært skjev.

Fra tallene på side 19 har vi altså et gjennomsnitt i et utvalg på 80. Dette kan gjentas og man kan lage et histogram av verdiene til de gjentatte gjennomsnittene.

Da ser man noe interessant: Spredningen til histogrammet til gjennomsnittene er mye mindre enn i histogrammet basert på de 30 000 samtalelengdene.

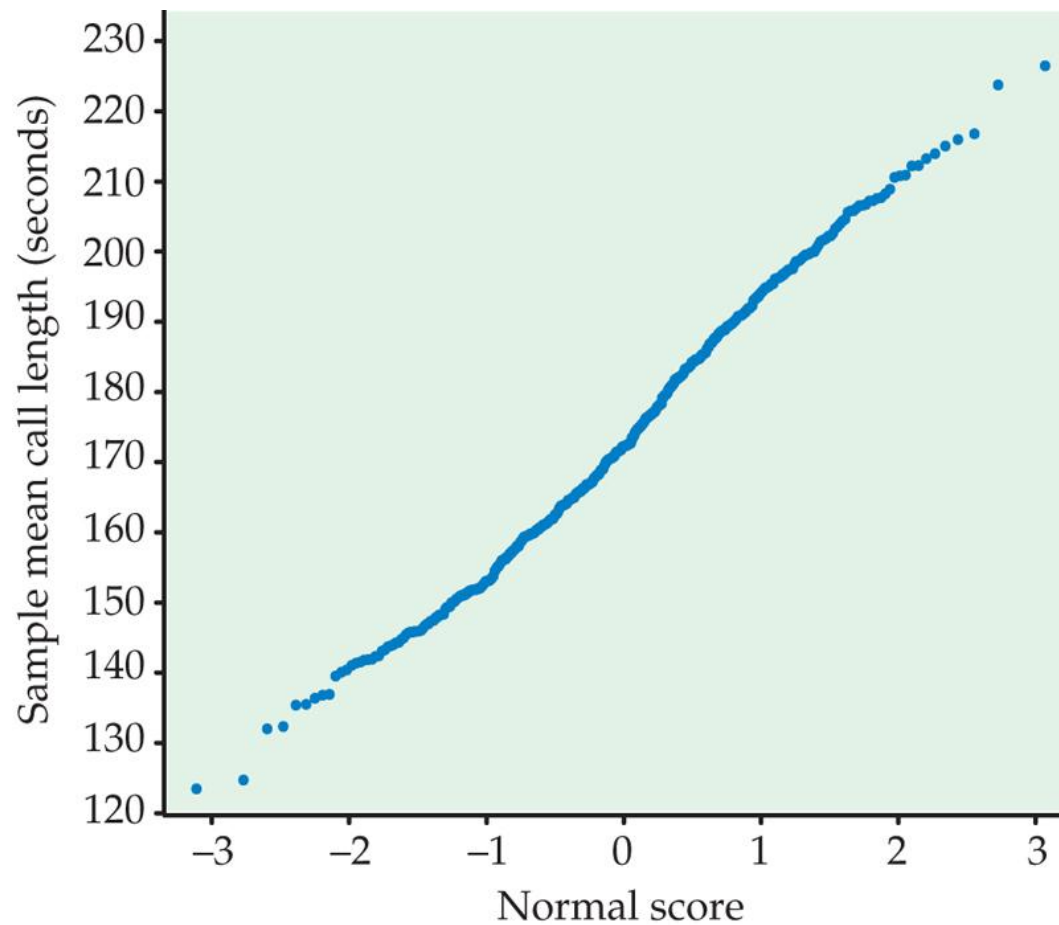


(a)



(b)

Dessuten: Et normalfordelingsplott illustrerer at histogrammet over gjennomsnittene et tilnærmet normal.



Eksemplet anskueliggjør to viktige punkter:

- Gjennomsnitt mindre variable enn enkeltobservasjoner.
- Gjennomsnitt nærmere normalfordelt enn enkeltobservasjoner.

Vi har dessuten at forventning (tyngdepunktet) i fordelingen av enkeltobservasjonene er den samme som forventningen av gjennomsnittenes fordeling.



## Forventning og standardavvik for gjennomsnittet

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

Utgangspunkt: Trekker et enkelt tilfeldig utvalg (simple random sample, **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 samme (**identiske**) fordeling.

Dessuten hvis 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  $X_1, \dots, X_n$  er **uavhengige**.

Slike uavhengige og identiske utvalg er modellen for trekning av enkle 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.

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

Denne kan utvides til  $\mu_{X_1+X_2+\dots+X_n} = \mu_{X_1} + \mu_{X_2} + \dots + \mu_{X_n}$ .

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

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

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

Dette gir at forventningen til gjennomsnittet blir

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

altså den samme forventningen som for en enkelt observasjon.

Siden gjennomsnittet  $\bar{x}$  har forventningen  $\mu$  sier vi at  $\bar{x}$  er en **forventningsrett** (unbiased) estimator for  $\mu$ .

Husk at  $\mu$  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  $\sigma_{X+Y}^2 = \sigma_X^2 + \sigma_Y^2$

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

Denne kan utvides til  $\sigma_{X_1+X_2+\dots+X_n}^2 = \sigma_{X_1}^2 + \sigma_{X_2}^2 + \dots + \sigma_{X_n}^2$  med uavhengige  $X_j$ .

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

Dermed blir  $\sigma_{X_1+X_2+\dots+X_n}^2 = n \sigma_X^2$ .

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

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

I en SRS blir dermed standardavviket til gjennomsnittet gitt som

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

Dermed vil spredningen til gjennomsnittet bli vilkårlig liten når  $n$  blir stor.

Gjennomsnittets fordeling er dessuten sentrert rundt forventningen  $\mu$ .

Dette er nettopp slik det skal være ut fra **store talls lov** (Kap 4.4)

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

## Eksempel: Samtalelengder.

Histogrammet til venstre har et standardavvik på  $\sigma = 184.81$  sekunder. Lengden på en tilfeldig samtale kan derfor variere mye i forhold til forventningen. Men et utvalg på 20 har standardavvik

$$\sigma_{\bar{x}} = 184.81/\sqrt{20} = 41.32 \text{ sekunder}$$

og et utvalg på 80 har standardavvik

$$\sigma_{\bar{x}} = 184.81/\sqrt{80} = 20.66 \text{ sekunder}$$

Hittil har vi bare sett på forventning og standardavvik til  $\bar{x}$ . For å bestemme selve fordelingen, må man vite noe om fordelingen til de enkelte målingene, altså fordelingen i populasjonen.

**Viktig tilfelle:** Populasjonsfordelingen normal.

Da gjelder: Hvis populasjonsfordelingen er  $N(\mu, \sigma)$  er fordelingen til gjennomsnittet av  $n$  uavhengige observasjoner,  $\bar{x}$ ,  $N(\mu, \sigma / \sqrt{n})$ .

Dette er en konsekvens av følgende: Enhver lineær kombinasjon av uavhengige normalfordelte variable er selv normalfordelt. Sum er en lineær kombinasjon.



Selv om det er viktig er resultatet på forrige slide et spesialtilfelle.

Mange populasjonsfordelinger er ikke normale.

Hva skjer med gjennomsnittet for et SRS og populasjonsfordelingen ikke er normalfordelt?

Til tross for at gjennomsnittet for et SRS ikke er eksakt normalfordelt, kan utvalgsfordelingen tilnærmes med en normalfordeling.

Og tilnærmelsen blir bedre jo større utvalget er. Dette, som kalles **sentralgrenseteoremet**, gjelder alle populasjonsfordelinger såsant standardavviket  $\sigma$  er endelig.

**Sentralgrenseteoremet:** Trekk et SRS på størrelse  $n$  fra en populasjon med forventning  $\mu$  og standardavvik  $\sigma$ . Da gjelder tilnærmet når  $n$  er stor at gjennomsnittet  $\bar{x}$  er tilnærmet normalfordelt med forventning  $\mu$  og standardavvik  $\sigma/\sqrt{n}$ ,

$$\bar{x} \sim N(\mu, \sigma/\sqrt{n})$$

Eksempel Samtalelengder: Standardavvik i populasjonen  $s = 185$ .

Med  $n=80$  blir  $\sigma_{\bar{x}} = \frac{1}{\sqrt{n}} \sigma = 20.66$

og med 68-95-99.7 regelen vil  $\bar{x}$  i 95% av de hypotetiske utvalgene ligge i intervallet  $(\mu - 2 \sigma_{\bar{x}}, \mu + 2 \sigma_{\bar{x}}) = (\mu - 41.3, \mu + 41.3)$ .

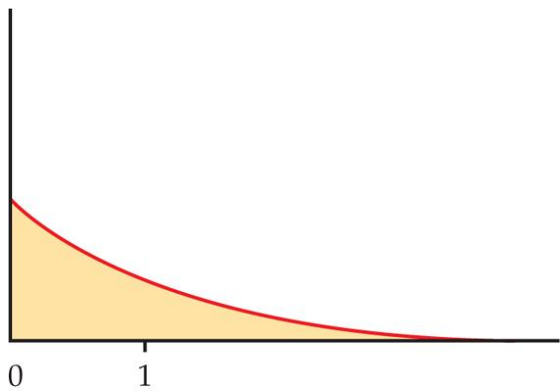
Hvis  $n$  økes til  $80 \times 16 = 1280$  blir  $\sigma_{\bar{x}} = \frac{1}{\sqrt{n}} \sigma = 5.17$

og intervallet blir  $(\mu - 10.3, \mu + 10.3)$ .

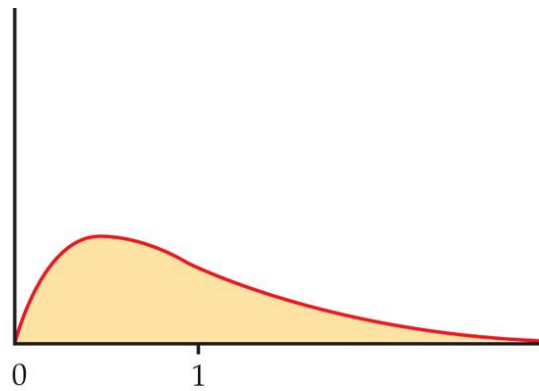
Ettersom  $n$  vokser, blir fordelingen til  $\bar{x}$  bedre og bedre tilnærmet med en normalfordeling.

Hvor god tilnærmelsen er avhenger av fordelingen til  $X_1, \dots, X_n$ . Flere observasjoner trengs hvis denne fordelingen avviker mye fra normalfordelingen.

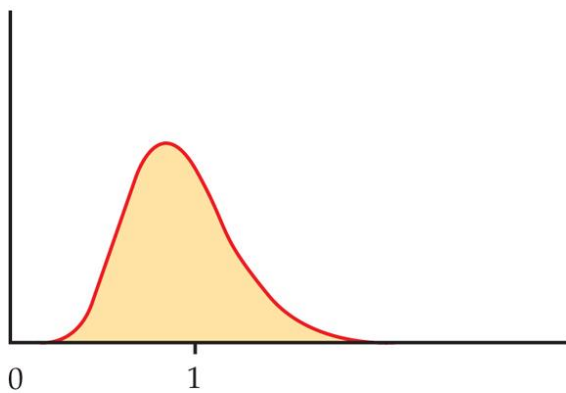
**Eksempel** La populasjonsfordelingen være gitt ved tetthetskurven  $\exp(-x)$ ,  $x \geq 0$ . Da er  $\mu=1$  og også  $\sigma=1$ . Men fordelingen er svært skjev. Figurene nedenfor viser hvordan tetthetskurvene til  $\bar{x}$  ser ut for utvalgsstørrelsene  $n=1, 2, 10, 25$ .



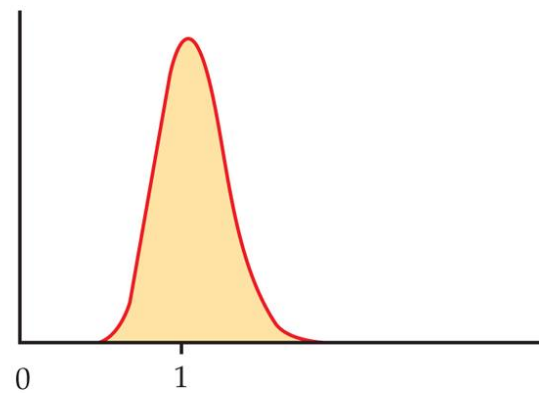
(a)



(b)



(c)



(d)

Husk at forventning er  $\mu_{\bar{x}} = 1$  og  $\sigma_{\bar{x}} = 1/\sqrt{n}$  så forventningene i fordelingene er de samme, men fordelingene blir mer konsentrert rundt  $\mu=1$  når  $n$  vokser.

Det ser vi av figurene, men i tillegg ser vi at allerede for  $n=10$  begynner tetthetskurven å se normal ut, og det blir ennå tydeligere for  $n=25$ .

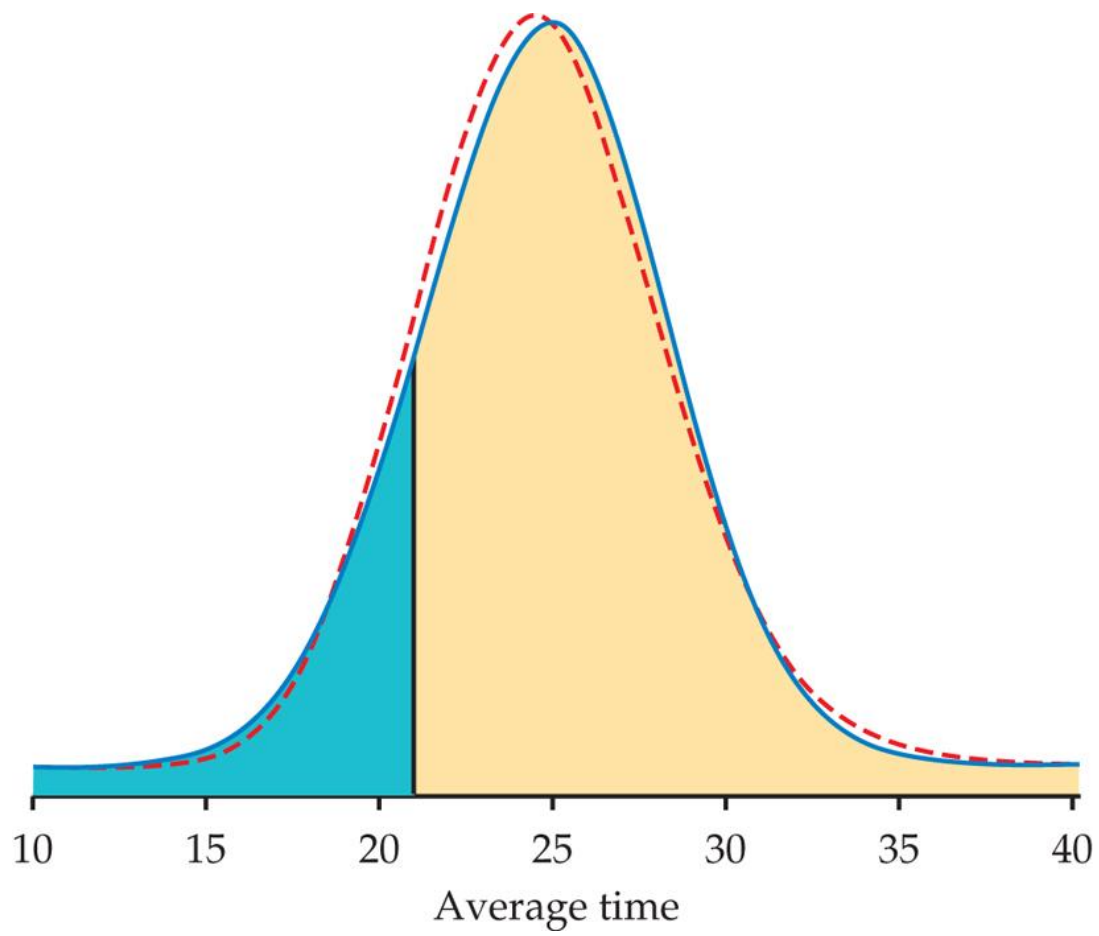
Anta at tiden mellom tekstmeldinger for unge voksne er eksponensielt fordelt med forventning  $\mu=25$  og standardavvik  $\sigma=25$  minutter.

Hva er sannsynligheten for at gjennomsnittet av 50 slike tidsrom er mer enn 21 minutter?

Fra sentralgrenseteoremet vet vi at  $\bar{x} \sim N(\mu, \sigma/\sqrt{n})$   
og  $\sigma/\sqrt{n} = 25 / \sqrt{50} = 3.54$ . Dermed blir

$$P(\bar{x} > 21) = P\left(\frac{\bar{x}-25}{3.54} > \frac{21-25}{3.54}\right) = P(Z > -1.13) = 0.8708$$

Dette svarer til arealet under kurven på neste side.



I figuren svarer den stiplede kurven til den tetthetskurven for  $\bar{x}$  når populasjonsfordelingen er eksponensiell.

Denne tetthetskurven kan beregnes ved en eksakt formel noe som er gjort for den stiplede kurven. Den eksakte verdien for  $P(\bar{x} > 21)$  kan da beregnes til 0.8750.

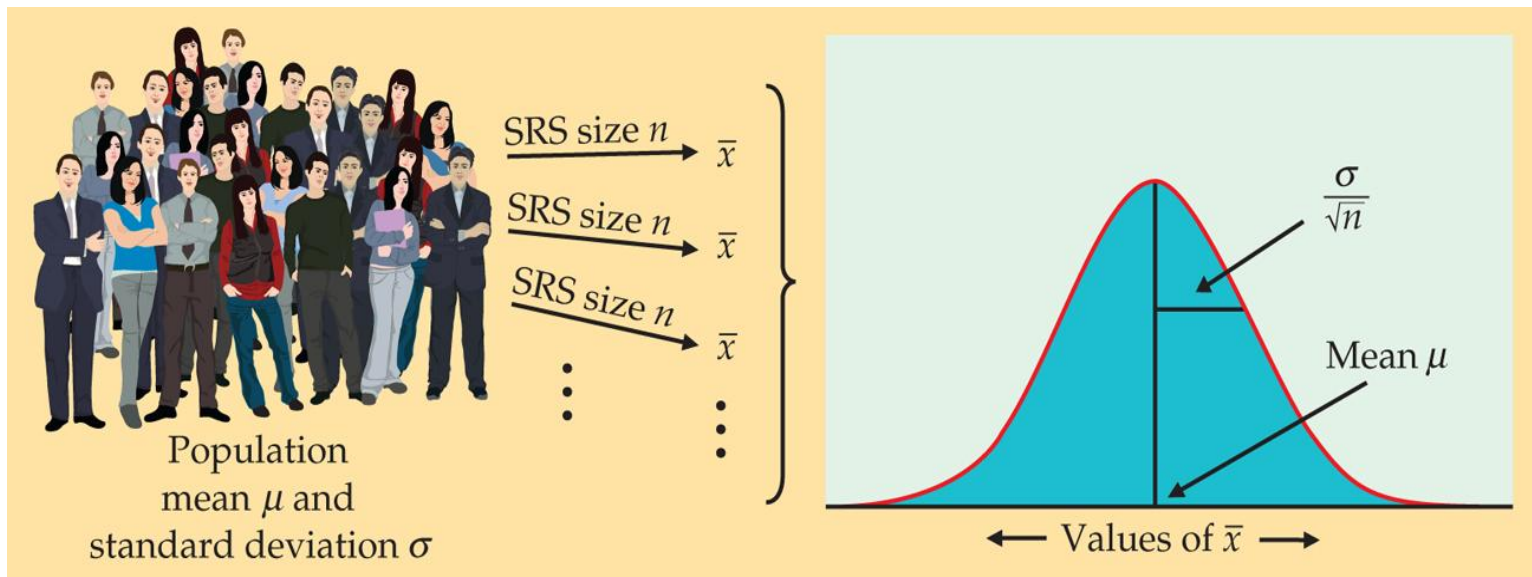
Avviket mellom den normaltilnærmede og den eksakte sannsynligheten blir da på kun  $0.8750 - 0.8708 = 0.0042$ .

Legg også merke til at  $P(\bar{x} > 21) = P(X_1 + \dots + X_{50} > 50 * 21)$ , så vi får også sannsynligheter svarende til totalsummen.



## Hovedpunktene ved utvalgsfordelingen til $\bar{x}$ :

- Trekk mange utvalg av størrelse  $n$  fra en populasjon med forventning  $\mu$  og standardavvik  $\sigma$ .
- Finn gjennomsnittet  $\bar{x}$  for hvert utvalg.
- Saml alle  $\bar{x}$  'ene og fremstill fordelingen.



Det kan være verdt å merke seg at parametrene  $\mu$  og  $\sigma$  inngår både i fordelingen til utvalget og populasjonsfordelingen.

Vi har sett tidligere: Enhver lineær kombinasjon av uavhengige normalfordelte variable er selv normalfordelt.

## Eksempel Reisetider

Anta at reisetiden for en besemt student til universitet,  $X$ , er  $N(20,4)$  og reisetiden fra universitet,  $Y$ , er  $N(18,8)$  .

Anta også at  $X$  og  $Y$  er uavhengige.

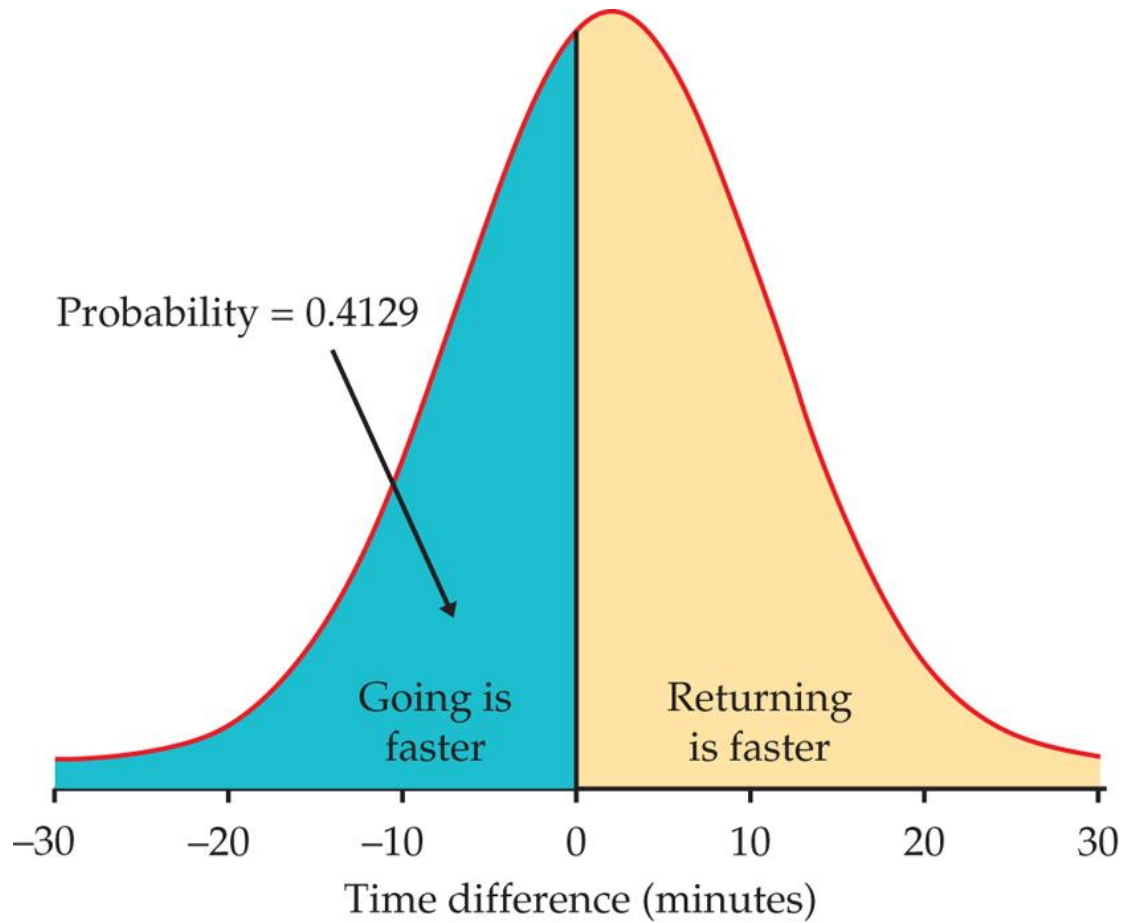
Forskjellen i reisetider er  $X-Y$ , og  $X<Y$  eller  $X-Y<0$  svarer til at reisetiden til universitet er minst. Hva er sannsynligheten?

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

$$\sigma_{X-Y}^2 = \sigma_X^2 + \sigma_Y^2 = 4^2 + 8^2 = 80, \quad \sigma_{X-Y} = \sqrt{80} = 8.94$$

så  $X-Y$  er  $N(2,8.94)$  og

$$P(X<Y) = P\left(\frac{(X-Y)-2}{8.94} < \frac{0-2}{8.94}\right) = P(Z < -0.22) = 0.4129.$$



Det finnes mer generelle versjoner av sentralgrenseteoremet. Variablene trenger ikke være uavhengige så lenge avhengigheten ikke er for stor. De trenger heller ikke ha samme fordeling så lenge ikke en enkelt variabel dominerer.

Vi vil snart se at resultatet også gjelder om variablene er diskret fordelt. Gjennomsnitt av diskret variable er også diskret fordelt. Men sentralgrenseteoremet er en tilnærming og tilnærmelsen gjelder også for gjennomsnitt av diskret fordelte variable.

# Fordeling til andeler og tellevariable

Vi vil nå se kategoriske tilfeldige variable, vi kan også si **diskrete** tilfeldige variable.

Eks. Spørreundersøkelse: Man spør et tilfeldig utvalg på  $n=1000$  stemmeberettigede personer om holdning til EØS.

Da er antall positivt innstilte  $X$  en diskret tilfeldig variabel.

I tilfeller som dette hvor individer bare kan angi to verdier (for/mot) kan andelen i utvalget  $\hat{p} = X / n$  brukes som en oppsummering av resultatet i spørreundersøkelsen.

Når populasjonen består av enheter som beskrives av to mulige verdier er fordelingen til summen med et av kjennetegnene i det tilfeldige utvalget en diskret tilfeldig variabel. Slike sannsynlighetsfordelinger behandles litt annerledes en kontinuerlige tilfeldige variable.

$X$  er summen av variable som bare kan anta to kjennetegn.

En vanlig sannsynlighetsmodell for å beskrive slike variable er følgende:

# Binomial oppsett

Det typiske eksemplet er myntkast.

Det er fire karakteristiske kjennetegn:

1. Det er et **fast** antall,  $n$ , observasjoner.
2. Observasjonene er **uavhengige**.
3. Observasjonen faller i **to kategorier** populært kalt ``suksess'' og ``fiasko''.
4. **Sannsynligheten** for suksess,  $p$ , er den **samme** for alle observasjonene.

$X$  = antallet «suksesser» blant de  $n$  forsøkene.

Ofte kalles oppsettet også en (binomisk) forsøksrekke.



## **Binomial fordeling:**

Fordelingen til den tilfeldige variabelen,  $X$ , i situasjonen beskrevet ovenfor, kalles **binomial fordelingen** med parametre  $n$  og  $p$ .

Parameteren  $n$  er antallet observasjoner (forsøk)

Parameteren  $p$  er sannsynligheten for suksess for hver observasjon (i hvert forsøk).

De mulige verdiene for  $X$  er alle hele tall mellom 0 og  $n$ .

Kort uttrykt sier man at  $X$  er  $\text{Bin}(n,p)$  –fordelt.

## Eksempler

- a) myntkast,  $n=10$  ,  $p=0.5$  ,  $X$  er  $\text{Bin}(10, 0.5)$  .
- b) utdeling av kort, ``suksess``: rødt. Ikke uavhengighet hvis uten tilbakelegging.
- c) genetikk, ``suksess``: blodtype O,  $n= 5$  barn,  $p=0.40$  ,  
 $X$  er  $\text{Bin}(5, 0.40)$  .
- d) pålitelighetsanalyse,  $n= 350$  fly, «suksess»: feilfrie i bestemt periode  $p=0.999$ ,  $X$  er  $\text{Bin}(350, 0.999)$  .

To typer trekking fra en endelig populasjon.

Det som brukes ved stikkprøver og spørreundersøkelser er **uten tilbakelegging**. Da blir ikke observasjonene uavhengige.

**Med tilbakelegging** gir at antallet  $X$  er binomialfordelt.

Når populasjonen er stor, og utvalget ikke utgjør en for stor andel, er forskjellen ikke stor.

**Eksempel:** Revisjon,  $N=10000$  poster, utvalg  $n=150$  poster

Andel feilposterings  $p=800/10000$ .

Fordi populasjonen er stor vil  $X$  være tilnærmet  $\text{Bin}(150, 0.08)$  fordelt.

Trekking uten tilbakelegging det aktuelle her.

Hvis populasjonen er mye større enn utvalget, vil antallet i et enkelt tilfeldig utvalg , SRS, være tilnærmet  $\text{Bin}(n, p)$  , der  $p$  er sannsynligheten for ``suksess''.

Tommelfingerregel: Hvis utvalget mindre enn 5% av populasjonen, kan vi anta at fordelingen til  $X$  er tilnærmet  $\text{Bin}(n, p)$ .

$P(X=k)$  for binomisk  $X$  er tabulert i tabell C i boka for  $n=1, \dots, 10, 12, 15, 20$  og  $p=0.01, \dots, 0.1, 0.10, 0.15, \dots, 0.45, 0.50$ .

Alternativt: Statistikkpakker

Minitab: ``Calc-Probability Distributions-Binomial``

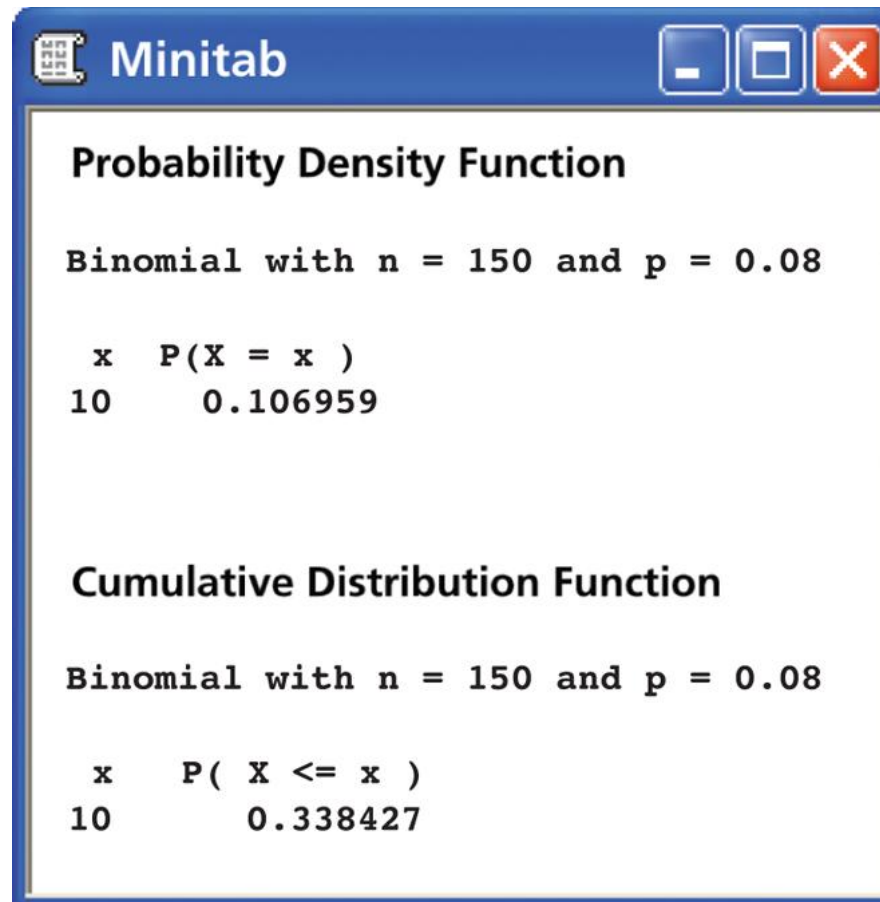
Input: ``number of trials``:  $n$ , ``Probability of success``:  $p$ ,  $x$

Output: ``Probability``:  $P(X = x)$ ,  
``Cumulative probability``:  $P(X \leq x)$ ,  
``Inverse cumulative probability``: Persentiler,  
``bruker tabellen baklengs``

**Eksempel**  $n=150$ ,  $p=0.08$ ,

$$P(X=10) = 0.10699$$

$$P(X \leq 10) = 0.338427$$



The image shows a screenshot of the Minitab software interface. The window title is "Minitab". The main content area displays the results of a binomial distribution analysis. It is divided into two sections: "Probability Density Function" and "Cumulative Distribution Function". Both sections specify a binomial distribution with  $n = 150$  and  $p = 0.08$ . The first section shows the probability  $P(X = 10) = 0.106959$ . The second section shows the cumulative probability  $P(X \leq 10) = 0.338427$ .

Probability Density Function	
Binomial with $n = 150$ and $p = 0.08$	
x	P(X = x)
10	0.106959

Cumulative Distribution Function	
Binomial with $n = 150$ and $p = 0.08$	
x	P(X ≤ x)
10	0.338427

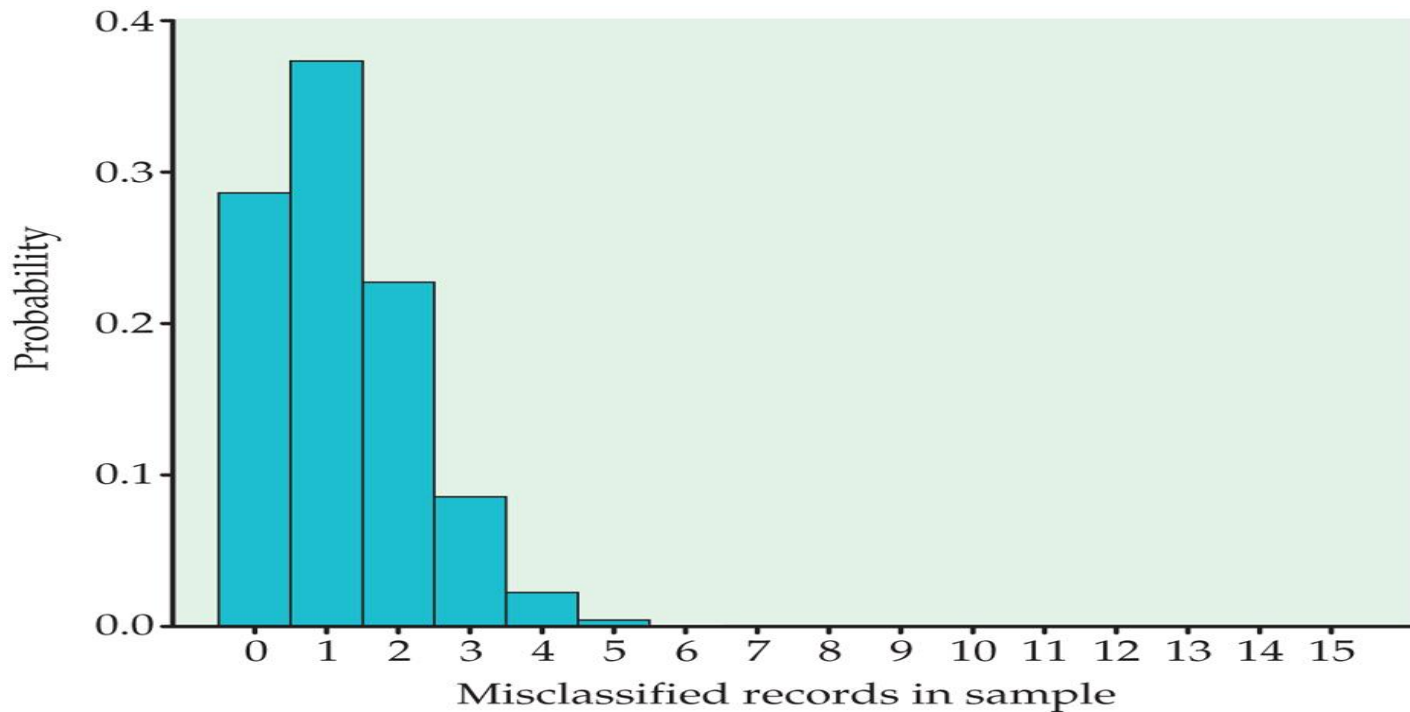
**Eksempel :**  $n=15$ ,  $p=0.08$

$$P(X \leq 1) = P(X=0) + P(X=1)$$

Fra tabell C , eller MINITAB

$P(X=0) = 0.2863$ ,  $P(X=1) = 0.3734$  slik at

$P(X \leq 1) = 0.6596$ .



X: antall suksesser i n forsøk eller blant n observasjoner

Definer

$S_i = 1$  hvis suksess,  $S_i = 0$  hvis fiasko

Da er

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

Kan nå bruke reglene for å finne  $\mu_X$  og  $\sigma_X$ .

$$\mu_S = 0 \times P(S=0) + 1 \times P(S=1) = 0 \times (1-p) + 1 \times p = p$$

$$\begin{aligned} \sigma_S^2 &= (0 - \mu_S)^2 \times P(S=0) + (1 - \mu_S)^2 \times P(S=1) = \\ &= p^2 \times (1-p) + (1-p)^2 \times p = (p+1-p) \times p \times (1-p) = p \times (1-p). \end{aligned}$$



$$\mu_X = \mu_{S_1} + \mu_{S_2} + \dots + \mu_{S_n} = n\mu_S = np$$

og siden  $S_1, S_2, \dots, S_n$  er uavhengige

$$\sigma_X^2 = \sigma_{S_1}^2 + \sigma_{S_2}^2 + \dots + \sigma_{S_n}^2 = n p \times (1-p).$$

Derfor gjelder at hvis  $X$  er  $\text{Bin}(n,p)$  fordelt, er

forventningen  $\mu_X = np$

og standardavviket  $\sigma_X = \sqrt{n p \times (1-p)}.$

Sannsynligheten for suksess i populasjonen er  $p$ , som er parameteren vi ønsker å estimere.

Andelen suksesser i et utvalg på  $n$  er

$$\hat{p} = \frac{\text{Antall suksesser i utvalg}}{\text{størrelse på utvalg}} = \frac{X}{n}$$

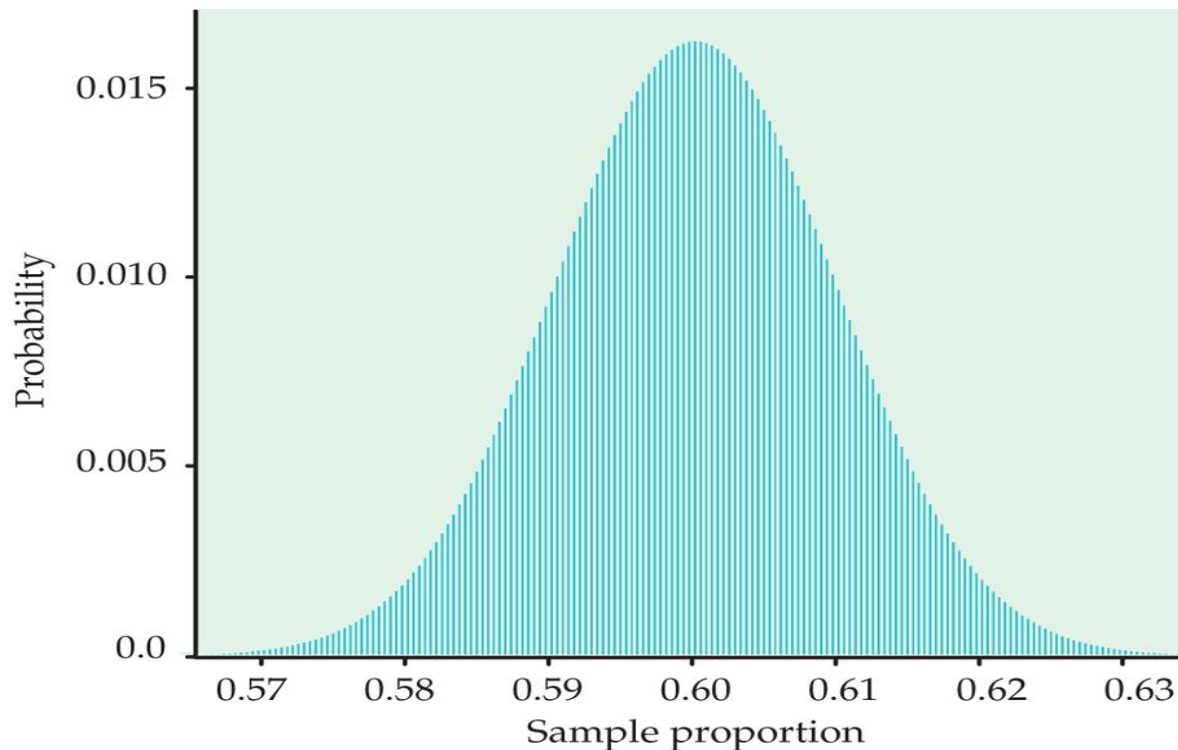
Å bruke at identiteter som

$$P(\hat{p} \geq 0.58) = P(X \geq 2500 \times 0.58 = 1450) = P(X=1450) + \dots + P(X=2500)$$

der  $X$  er  $\text{Bin}(2500, 0.60)$  er ikke spesielt tiltalende, så for store  $n$  brukes programmer for å begrense sannsynlighetene.

Men vi så at andeler kan oppfattes som gjennomstitt av suksesser i utvalget, og derfor vil det også eksistere en normaltilnærming i dette tilfellet. Til å bestemme tilnærmelsen trenger vi forventning og standardavvik.

**Eksempel: Sannsynlighets-histogram for  $B(2500,0.6)$ .**



La  $\hat{p} = X/n$  være andelen suksesser i et SRS (enkelt tilfeldig utvalg) fra en stor populasjon.

Da gjelder: Forventningen til  $\hat{p}$  er  $\mu_{\hat{p}} = p$  (forventningsrett).

Hvis  $X$  er binomialfordelt er standardavviket til  $\hat{p}$  **eksakt** lik

$$\sigma_{\hat{p}} = \sqrt{p(1-p)/n}$$

Hvis det er trekning uten tilbakelegging gjelder denne formelen for standardavviket **tilnærmet**.

Tilnærmingen er god hvis populasjonen er 20 ganger så stor som utvalget.

I dette tilfellet blir normaltilnærmelsen:

La  $X$  være antall suksesser i et utvalg fra en stor populasjon der sannsynligheten for suksess er  $p$ . La  $\hat{p} = X/n$  være andelen suksesser. Da gjelder:

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

$\hat{p}$  er tilnærmet  $N(p, \sqrt{p(1-p)/n})$ .

Tommelfingerregel. Hvis  $np \geq 10$  og  $n(1-p) \geq 10$  er tilnærmelsen tilfresstillende.



**Eksempel:** Anta X er Bin( 2500,0.6)

Da er  $P(\hat{p} \geq 0.58) = P(X \geq 2500 \times 0.58 = 1450) = 0.9802$  exact (Minitab).

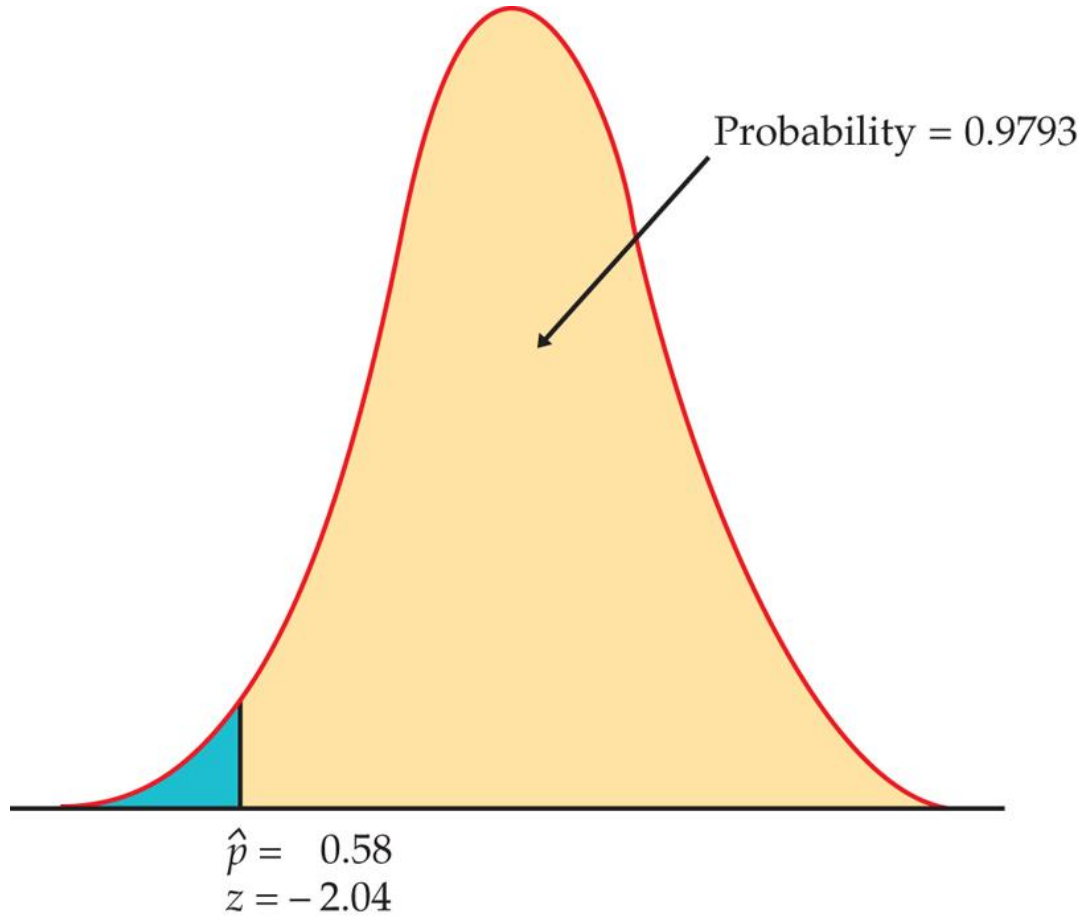
$$\mu_{\hat{p}} = p$$

$$\sigma_{\hat{p}} = \sqrt{p(1-p)/n} = 0.0098$$

Så, tilnærmelsen blir

$$P(\hat{p} \geq 0.58) = P\left(\frac{\hat{p} - 0.6}{0.0098} \geq \frac{0.58 - 0.6}{0.0098}\right) = P(Z \geq -2.04) = 0.9793.$$

Forskjellen er 0.0009.





**Eksempel:** Revisjon,  $N=10000$  poster, utvalg  $n=150$  poster

Andel feilposterings  $p=800/10000$ .

Fordi populasjonen er stor vil  $X$  være tilnærmet  $\text{Bin}(150,0.08)$  fordelt.

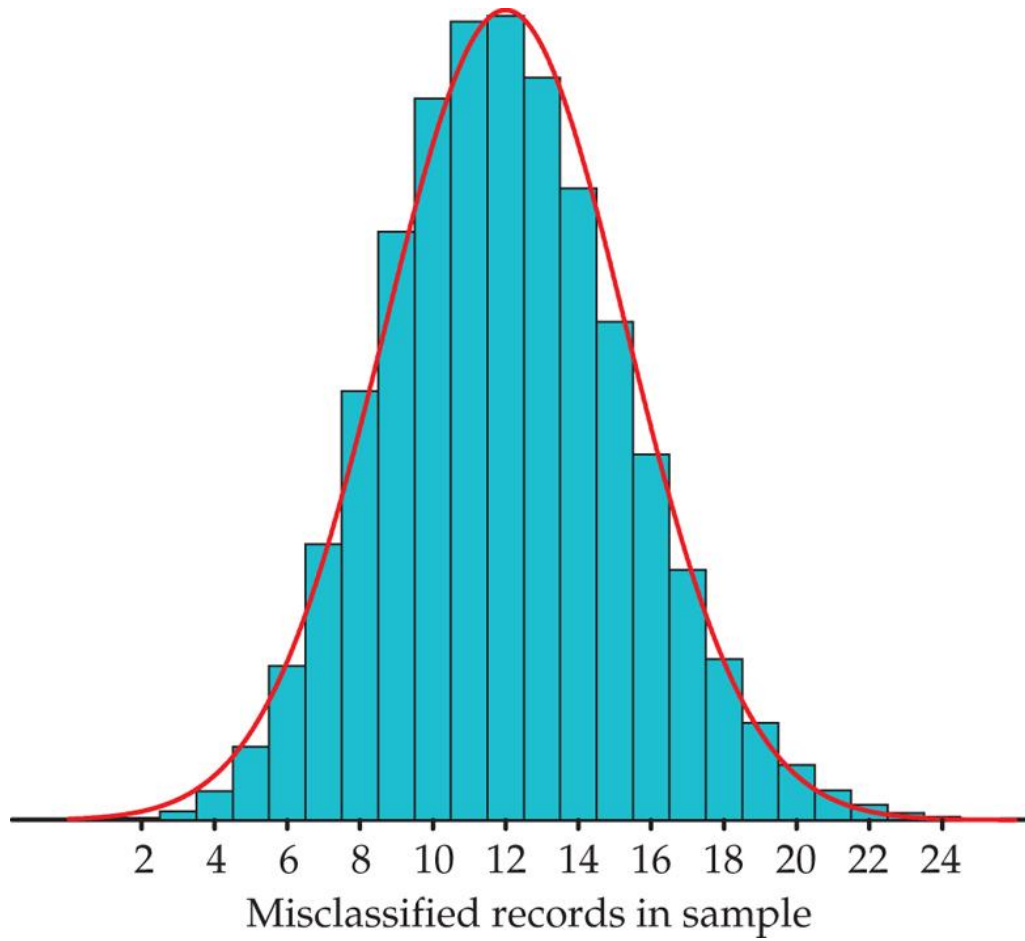
$$\mu_X = np = 150 \times 0.08 = 12$$

$$\sigma_X = \sqrt{np(1-p)} = \sqrt{150 \times 0.08 \times 0.92} = 3.3226.$$

$$P(X \leq 10) = P\left(\frac{X-12}{3.3226} \leq \frac{10-12}{3.3226}\right) = P(Z \leq -0.60) = 0.2743.$$

Men eksakt  $P(X \leq 10) = 0.3384$  ( Minitab)

Tilnærmelsen ikke spesielt god,  $np=150 \times 0.08=12$



**Eksempel** Da vi introduserte kontinuerlige tilfeldige variable, ble det påstått, side 259.

«Det kan vises at hvis parameteren  $p=0.26$  i en populasjon og man trekker et utvalg på  $n=500$ , er utvalgsfordelingen til  $\hat{p}$ , andelen i utvalget, tilnærmet  $N(0.26, 0.0196)$ »

Nå kan vi fylle ut detaljene:

$$\mu_{\hat{p}} = p = 0.26$$

$$\sigma_{\hat{p}} = \sqrt{p(1-p)/n} = \sqrt{0.26 \times 0.74/500} = 0.01961632$$