Marit Helene Gladhaug

MET4: Pensumoversikt

Del I	Del II	Del III
Beskrivende statistikk Utvalg og estimering Inferens om en og to populasjoner Enveis og toveis variansanalyse Fordelingsfrie tester Kjikvadrattester	Enkel regresjon Multippel regresjon Modellbygging Paneldataøkonometri OLS	Tidsserieanalyse Logistisk regresjon

DEL I

I. Beskrivende statistikk

Populasjon: Alle enhetene som en problemstilling gjelder for

Utvalg: Den delmengden av populasjonen som blir undersøkt. Vi forutsetter tilfeldige utvalg: dvs at alle har samme sannsynlighet for å komme med i utvalget. Et tilfeldig utvalg sikrer representativitet og muligheter for å generalisere fra utvalget til populasjonen.

Marit Helene Glad Rang

Observator

Ukjent

Populasjon → Utvalg n → Kjent: Utfører en undersøkelse

Konklusjonen gjelder for hele populasjonen

μ σ^2 **Parameter**: En tallstørrelse som beskriver en eller annen egenskap ved en populasjon, f.eks forventningen eller variansen. Dette er sanne, men ukjente størrelser.

$\bar{X}$ S^2 **Observator**: En observator er en tallstørrelse som beskriver en egenskap ved utvalget, eks gjennomsnittet og utvalgsvarians.

Utvalgsplan:

- Enkel tilfeldig trekning
- Proporsjonal stratifisering
- Disproporsjonal stratifisering
- Klyngeutvelgelse

NB! Det er den absolutte størrelsen på utvalget som avgjør presisjonen av estimatorene, ikke den relative.

Marit Helene Gladhaug

Beskrivende statistikk: Gir informasjon om
- fordelinger ⇒ sentraltendens og variasjon
- sammenhenger ⇒ korrelasjon og regresjon

Variablers målenivå: To hovedkategorier;

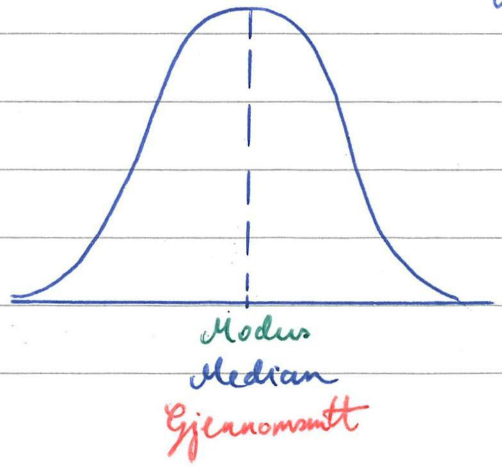
(1) Kategoridata (inkl. rangeringsdata)
- Dikotome (dummy, indikator)
  - to kategorier 0/1
- Nominale variabler
  - kategorier som ikke kan rangeres
- Ordinale data (rangeringsdata)

(2) Målevariabler
- Intervalldata
- Forholdsdata

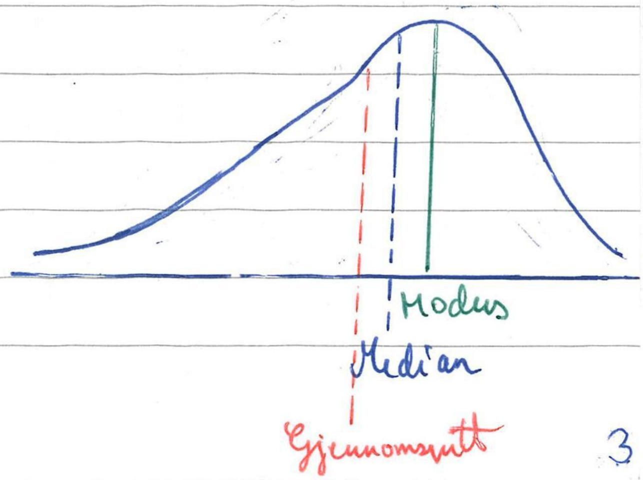
Sentraltendens:

(1) Gjennomsnitt $\overline{X}$ → gjelder kun for målvariabler
(2) Median → målvariabler og ordinale variabler
(3) Typetall (modus) → den vanligste verdien

Symmetrisk fordeling



Skewed fordeling



Marit Helene Gladhaug

Spredning:

Variasjonsbredde

Varians og standardavvik

→ et mål på det gjennomsnittlige avviket fra gjennomsnittet

Variasjonskoeffisienten (CV): standardavviket delt på gjennomsnittet

- Variansen til populasjonen:

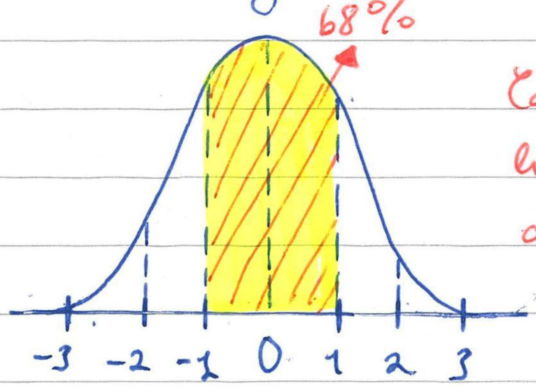
$$\sigma^2 = \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{N}$$

- Variansen til utvalget:

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

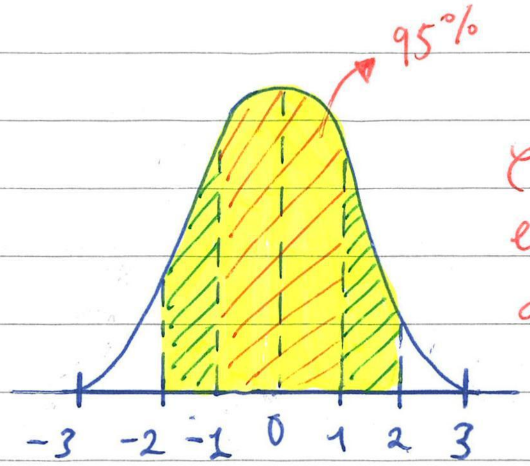
- Variasjonskoeffisientene: $\frac{\sigma}{\mu}$ og $\frac{S}{\bar{x}}$

- Tolkning av standardavvik:

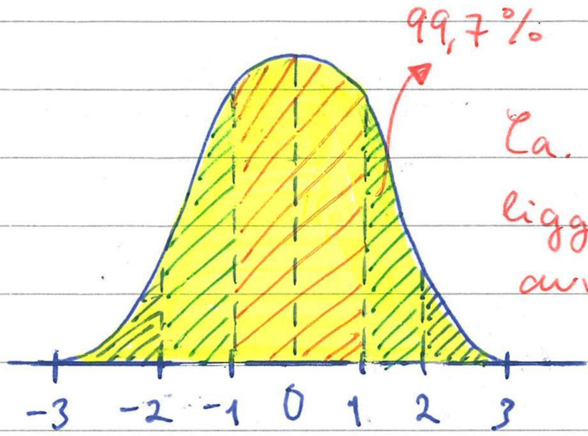


Ca. 68% av observasjonene ligger mindre enn 1 standardavvik fra gjennomsnittet

Marit Helene Gladhaug



Ca. 95 % av observasjonene ligger mindre enn 2 standard-avvik fra gjennomsnittet.



Ca. 99,7 % av observasjonene ligger mindre enn 3 standard-avvik fra gjennomsnittet.

- Persentiler : Angir den verdien som har p prosent av observasjonene under seg

  - Merk: Median = 50 % persentilen

- Korrelasjon :

Korrelasjonskoeffisienten r forteller oss hvor nær vi er en lineær sammenheng

- $r \in [-1, 1]$
- Merk ! r = 0 sier ikke at det ikke er en sammenheng, bare at sammenhengen ikke er lineær !

$$r = \sum_{i=1}^{n} \frac{(x_i - \overline{x})(y_i - \overline{y})}{(n-1) S_x S_y} = \frac{Cov(X,Y)}{S_x S_y}$$

Marit Helene Gladhaug

NB! Med $r = 0$ kan det likevel være en sammenheng.

→ Ukorrelerthet medfører ikke uavhengighet

→ Men uavhengighet medfører ukorrelerthet

II. Utvalg og estimering

Stokastisk variabel: En funksjon (eller regel) som tilordner en tallverdi, x, til ethvert utfall u, i en sannsynlighetsmodell.

Forventning: $E(X) = \sum x \cdot p(x)$

Varians:
$$Var(X) = E(X - E(X))^2$$
$$\sigma^2 = E(X^2) - [E(X)]^2$$

Standardavvik: $\sigma = \sqrt{\sigma^2} = \sqrt{Var(X)}$

Kovarians:
$$Cov(X,Y) = E[(X-E(X)) \cdot (Y-E(Y))]$$
$$= E(X \cdot Y) - E(X) \cdot E(Y)$$

Regneregler for forventning:

(1) $E(k) = k$

(2) $E(k + X) = E(k) + E(X) = k + E(X)$

(3) $E(kX) = k \cdot E(X)$

(4) $E(X+Y) = E(X) + E(Y)$

(5) $E(X \cdot Y) = E(X) \cdot E(Y)$ hvis uavhengige

(6) $E(X \cdot Y) = E(X) \cdot E(Y) + COV(X,Y)$ hvis avhengige

Marit Helene Gladhaug

Regneregler for varians:

(1) $Var(X) = E(X^2) - [E(X)]^2$

(2) $Var(k) = 0$

(3) $Var(k+X) = Var(k) + Var(X) = Var(X)$

(4) $Var(kX) = k^2 \cdot Var(X)$

(5) $Var(X+Y) = Var(X) + Var(Y)$ hvis uavhengige

NB! (6) $Var(X-Y) = Var(X) + Var(Y)$ hvis uavhengige

avhengig:

(7) $Var(X+Y) = Var(X) + Var(Y) + 2\,COV(X,Y)$

(8) $Var(X-Y) = Var(X) + Var(Y) - 2\,COV(X,Y)$

Forventning og varians til et gjennomsnitt:

- $X_1, X_2, X_3, \ldots, X_n$ uavhengige stokastisk variabler
- utvalg: n observasjoner
- hver X har fordelingen $X \sim N(\mu, \sigma^2)$
- vi vil finne fordelingen til gjennomsnittet; $\bar{X}$

$$\bar{X} = \frac{1}{n}(X_1 + X_2 + \ldots + X_n)$$

$$E(\bar{X}) = E\left(\frac{1}{n}(X_1 + X_2 + \ldots + X_n)\right)$$

$$= \frac{1}{n} E(X_1 + X_2 + \ldots + X_n)$$

$$= \frac{1}{n}\left(\underbrace{E(X_1)}_{\mu} + \underbrace{E(X_2)}_{\mu} + \ldots + \underbrace{E(X_n)}_{\mu}\right)$$

$$= \frac{1}{n}(\mu + \mu + \ldots + \mu)$$

$$= \frac{1}{n}(\mu \cdot n) = \frac{\mu \cdot \cancel{n}}{\cancel{n}} = \underline{\underline{\mu}}$$

Marit Helene Gladhaug

$$\mathrm{Var}(\overline{X}) = \mathrm{Var}\left(\frac{1}{n}(X_1 + X_2 + \ldots + X_n)\right)$$

$$= \left(\frac{1}{n}\right)^2 \cdot \mathrm{Var}(X_1 + \ldots + X_n)$$

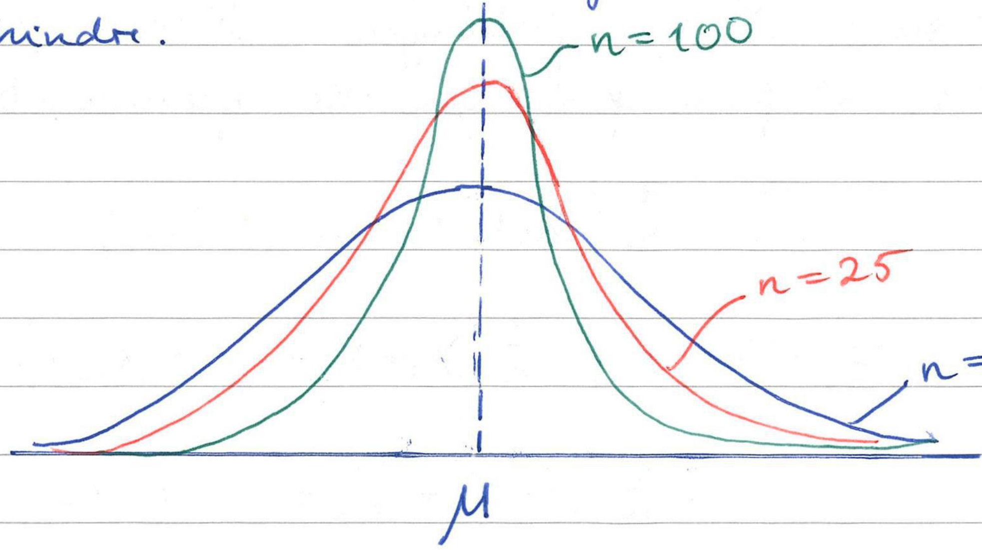
$$= \left(\frac{1}{n}\right)^2 \cdot \underbrace{\mathrm{Var}(X_1)}_{\sigma^2} + \underbrace{\mathrm{Var}(X_2)}_{\sigma^2} + \ldots + \underbrace{\mathrm{Var}(X_n)}_{\sigma^2}$$

$$= \left(\frac{1}{n}\right)^2 \cdot n \cdot \sigma^2 = \frac{\cancel{n} \cdot \sigma^2}{n^{\cancel{2}}} = \underline{\underline{\frac{\sigma^2}{n}}}$$

Konklusjon: $X \sim N(\mu, \sigma^2)$
$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$

$$\sigma(\overline{X}) = \sqrt{\mathrm{Var}(\overline{X})} = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$$

Merk: når vi øker utvalgsstørrelsen, blir variansen mindre.



Variansen (spredningen mellom observasjonene) reduseres når n øker. Halene trekker seg sammen slik at de fleste observasjonene befinner seg i et intervall som sentrerer seg rundt gjennomsnittet i populasjonen (μ).

Marit Helene Gladhaug

Sentralgrenseteoremet:

Et gjennomsnitt vil alltid nærme seg normalfordelingen når vi har mange observasjoner, uansett hvilken fordeling X har.

$$S_n = X_1 + X_2 + \dots + X_n \qquad X_i = \text{enkelt-observasjoner}$$

$$E(S_n) = E(X_1) + E(X_2) + \dots + E(X_n) = n \cdot E(X)$$

$$\begin{aligned} Var(S_n) &= Var(X_1 + X_2 + \dots + X_n) \\ &= \sum Var(X_i) + \sum Cov(X_i, X_j) \\ &= \sum Var(X_i) \quad \text{hvis uavhengige} \\ &= n \cdot Var(X) \quad \hookrightarrow Cov(X_i, X_j) = 0 \end{aligned}$$

$$\text{Når } n \to \infty \Rightarrow S_n \sim N(n \cdot \mu, n \cdot \sigma^2)$$

Forventning og varians til en andel:

- Binomisk forsøksrekke: $I_j = 1 + 0 + 1 + \dots + 1_j$
- $E(I_j) = p \cdot 1 + (1-p) \cdot 0 = p$
- $Var(I_j) = E(I^2) - E(I)^2 = p - p^2 = p(1-p)$
- Vi er interessert i antall suksesser i et utvalg på n:

$$X_n = \sum^n I_j = I_1 + I_2 + \dots + I_n$$

$$\begin{aligned} E(X_n) &= E(I_1 + I_2 + \dots + I_n) \\ &= \underbrace{E(I_1)}_{p} + \dots + \underbrace{E(I_n)}_{p} = \underline{n \cdot p} \end{aligned}$$

Marit Helene Gladhaug

$$\mathrm{Var}(X_n) = \mathrm{Var}(I_1 + \ldots + I_n)$$
$$= \mathrm{Var}(I_1) + \ldots + \mathrm{Var}(I_n)$$
$$= p(1-p) + \ldots + p(1-p)$$
$$= \underline{\underline{np(1-p)}}$$

$X_n \sim \mathrm{Bin}(n,p)$ der $E(X_n) = np$
$\mathrm{Var}(X_n) = np(1-p)$

- Vi kan nå regne ut forventning og varians for andelen suksesser : $\dfrac{X_n}{n}$

$$E\left(\frac{X_n}{n}\right) = \frac{1}{n}E(X_n) = \frac{1}{\cancel{n}} \cdot \cancel{n}p = \underline{\underline{p}}$$

$$\mathrm{Var}\left(\frac{X_n}{p}\right) = \left(\frac{1}{n}\right)^2 \mathrm{Var}(X_n) = \frac{1}{n^{\cancel{2}}} \cdot \cancel{n}p(1-p) = \underline{\underline{\frac{p(1-p)}{n}}}$$

Merk: $E(X_n) = n \cdot p$ $\quad E\left(\frac{X_n}{n}\right) = p$

$\mathrm{Var}(X_n) = np(1-p)$ $\quad \mathrm{Var}\left(\frac{X_n}{n}\right) = \frac{p(1-p)}{n}$

$$\boxed{\frac{X_n}{n} \sim N\left(p, \frac{p(1-p)}{n}\right) \text{ når } n \text{ er stor}}$$

$$\boxed{\frac{X_n}{n} = \hat{p} \sim N\left(p, \frac{p(1-p)}{n}\right)}$$

$$E(\hat{p}) = \mu = p$$
$$\mathrm{Var}(\hat{p}) = \sigma^2 = \frac{p(1-p)}{n}$$

Marit Helene Gladhaug

Inferens: Vi trekker konklusjoner om populasjonens parametre basert på det vi observerer i utvalg.

Ukjente populasjonsparametre		Estimatorer
μ	$\longleftarrow$	$\overline{X}$
p	$\longleftarrow$	$\hat{p}$
σ^2	$\longleftarrow$	S^2

En god estimator er forventningsrett og konsistent:

Forventningsrett: $E(\hat{\theta}) = \theta$

Konsistent: $S^2 \to \sigma^2$ når $n \to \infty$

Merk: En estimator kan være konsistent (lav varians) uten å være forventningsrett, og forventningsrett uten å være konsistent.

Viktigst: at estimatoren er konsistent!

Hypotesetesting: $H_0: \mu = \ldots$
$H_A: \mu \neq \ldots, \mu <, \mu > \ldots$

P-verdi: den maksimale sannsynligheten for å observere det vi fikk, eller høyere enn det vi fikk, gitt at forutsetningene under H_0 stemmer.

Marit Helene Gladhaug

$\Rightarrow P_{H_0}(\overline{X} > \text{observert verdi}) = \text{P-verdi}$

Tosidig test: $2 \cdot P_{H_0}(\overline{X} > \text{observert verdi}) = \text{P-verdi}$

To typer feil:

(1) Forhastningsfeil: Sannsynligheten for å forkaste en hypotese som egentlig er sann: α. Type 1-feil

(2) Godkjenningsfeil: Sannsynligheten for å godta en hypotese som egentlig er feil. β. Type 2-feil.

	Naturens ukjente sannhet	
	H_0 er sann	H_A er sann
Vår beslutning: Beholde H_0	Riktig konklusjon	Gal konklusjon Type 2-feil: β
Forkaste H_0	Gal konklusjon Type 1-feil: α	Riktig konklusjon
	↓ Forhastningsfeil	↓ Godkjenningsfeil

Styrken til en test $= 1 - \beta$

$\Rightarrow$ sannsynligheten for at vi forkaster en nullhypotese som er feil

Marit Helene Gladhaug

Inferens om én populasjon med ukjent standardavvik:

I. Inferens om gjennomsnittet $\bar{X}$

II. Inferens om standardavviket / variansen σ^2

III. Inferens om en andel $\hat{p}$

I. Inferens om gjennomsnittet — t-test

$$\hat{\mu} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad \hat{\mu} \rightarrow \mu$$

$$E(\hat{\mu}) = E(\bar{X}) = \mu \qquad \text{Forventningsrett}$$

$$Var(\hat{\mu}) = Var(\bar{X}) = \frac{\sigma^2}{n}$$

Problem: vi kjenner ikke σ^2. Må estimeres med utgangspunkt i observasjonene; med utvalgsvariansen.

$$S^2 = \frac{1}{n-1}\sum (X_i - \bar{X})^2 \qquad S^2 \rightarrow \sigma^2$$

$$E(S^2) = \sigma^2$$

$$\hat{\sigma}(\bar{X}) = S(\bar{X}) = \frac{S}{\sqrt{n}}$$

Når X-ene er $N \sim (\mu, \sigma^2)$ er

1. $Z = \dfrac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$ når $n > 30$

2. $T = \dfrac{\bar{X}-\mu}{S/\sqrt{n}} \sim t_{n-1}$ når $n < 30$

t-test: vi kjenner ikke variansen (utvalget er lite)

Marit Helene Gladhaug

- Når σ er kjent (eller n stor) $\Rightarrow$ bruk Z som testobservator.

- Når σ er ukjent (eller n liten) $\Rightarrow$ bruk T som testobservator.

II. Inferens om standardavvik

- Når vi har ukjent standardavvik / varians kan det være aktuelt å teste hypoteser om σ^2
- Slike tester er basert på kjikvadratfordelingen (χ^2)
- Anta at $X_1, X_2, \ldots, X_n \sim N(0,1)$. Da er

$$Q = \sum_{i=1}^{n} X_i^2 \sim \chi_n^2$$

$$E(Q) = n$$

- Testobservator for variansen: $\boxed{\frac{n-1}{\sigma^2} S^2 \sim \chi_{n-1}^2}$

III. Inferens om en andel

- X_n = antall suksesser i n forsøk

- Estimator for den sanne (ukjente) andelen p:

$$\hat{p} = \frac{X_n}{n} \sim N\left(p, \frac{p(1-p)}{n}\right)$$

Marit Helene Gladhaug

- Testobservator: $Z = \dfrac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0,1)$ under H_0

- Konfidensintervall: $\hat{p} \pm k \cdot \hat{\sigma}(\hat{p})$

$\Rightarrow \hat{p} \pm k \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

$\Rightarrow \hat{p} \pm \underbrace{B}_{\text{feilmargin}} = \hat{p} \pm \underbrace{k \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}_{\text{Feilmargin}}$

$\Rightarrow$ n som en funksjon av feilmarginen/presisjonen:

$$n = \left(\frac{k \cdot \sqrt{\hat{p}(1-\hat{p})}}{B} \right)^2$$

Inferens om to populasjoner:

I. Sammenligning av to gjennomsnitt fra normalfordelte populasjoner

→ Uavhengig utvalg med lik/ulik varians
→ Matchede par

Toutvalgsmodellen: $n_1, n_2, X_1, X_2, \overline{X}_1, \overline{X}_2$

- Vi har to populasjoner med ukjente forventninger μ_1 og μ_2.

Merk: $\overline{X}_1 \sim N(\mu_1, \frac{\sigma_1^2}{n_1})$ $\overline{X}_2 \sim N(\mu_2, \frac{\sigma_2^2}{n_2})$

- n_1 observasjoner fra populasjon 1
- n_2 observasjoner fra populasjon 2
- gjennomsnitt : $\overline{X}_1$ og $\overline{X}_2$

- Vi gjør inferens om $(\overline{X}_1 - \overline{X}_2)$:

$$E(\overline{X}_1 - \overline{X}_2) = \mu_1 - \mu_2$$

$$Var(\overline{X}_1 - \overline{X}_2) = Var(\overline{X}_1) + Var(\overline{X}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

antar uavhengige

- Testobservator :

$$Z = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

under H_0

<u>Merk</u>: standardavviken σ_1^2 og σ_2^2 er ukjent. Bruker den empiriske variansen S^2; OBS! Må bruk t- tabellen og ikke normalfordelingstabellen.

- Vi kan anta at de to populasjonene har lik varians og estimere denne som

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

$\downarrow$

"Pooled" varians

Marit Helene Gladhaug

Vi setter S_p^2 inn for σ^2 og får en observator som er t-fordelt med $v = n_1 + n_2 - 2$ frihetsgrader.

$$T = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim t_{n_1+n_2-2}$$

- Vi kan anta at de to populasjonene har ulik varians og setter de estimerte standardavvikene S_1 og S_2 for σ_1 og σ_2 og får testobservatoren

$$T = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \sim t_v$$

som er tilnærmet t-fordelt med antall frihetsgrader:

$$v = \frac{\left(S_1^2/n_1 + S_2^2/n_2\right)^2}{\frac{(S_1^2/n_1)^2}{(n_1-1)} + \frac{(S_2^2/n_2)^2}{(n_2-1)}}$$

Merk: Når vi har flere frihetsgrader får vi lettere forkastning.

Marit Helene Gladhaug

II. Sammenligning av to varianser

- Når vi sammenligner to varianser ser vi på forholdet S_1^2 / S_2^2 og ikke differansen slik som for gjennomsnittet.

$\Rightarrow$ Varianser : $\frac{S_1^2}{S_2^2}$

$\Rightarrow$ Gjennomsnitt : $\overline{X}_1 - \overline{X}_2$

- $H_0 : \sigma_1^2 = \sigma_2^2$
  $H_A : \sigma_1^2 \neq \sigma_2^2$

- Testobservatoren :

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} = \frac{S_1^2}{S_2^2} \sim F_{n_1-1,\, n_2-1}$$

III. Sammenligning av to andeler

- To populasjoner n_1 og n_2
- Observerer $\hat{p}_1 = \frac{X_1}{n_1}$ og $\hat{p}_2 = \frac{X_2}{n_2}$

- $X_1 \sim Bin(n,p)$ med $E(X_1) = np$
  $Var(X_1) = np(1-p)$

  $X_2 \sim Bin(n,p)$ med $E(X_2) = np$
  $Var(X_2) = np(1-p)$

- $E(\underbrace{\frac{X_{n_1}}{n_1}}_{\hat{p}_1}) = p_1 \qquad Var(\underbrace{\frac{X_{n_1}}{n_1}}_{\hat{p}_1}) = \frac{p(1-p)}{n_1}$

Marit Helene Gladhaug

- $E\left(\overset{\hat{p}_2}{\frac{X_{n2}}{n_2}}\right) = p_2 \qquad Var\left(\overset{\hat{p}_2}{\frac{X_{n2}}{n_2}}\right) \quad \frac{p(1-p)}{n_2}$

- $E(\hat{p}_1 - \hat{p}_2) = p_1 - p_2$

- $Var(\hat{p}_1 - \hat{p}_2) = \left(\frac{1}{n_1} + \frac{1}{n_2}\right) p(1-p)$

- Testobservator:

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) p(1-p)}} \sim N(0,1)$$

der $p = p_1 = p_2$

- ANOVA: Enveis variansanalyse

- Variansanalyse er en utvidelse av toutvalgs-modeller til situasjoner der vi ønsker å sammenligne to eller flere populasjoner (grupper)

- Antar normalfordelt responsvariabel med lik varians for alle grupper

- H_0: Alle gruppene har lik forventning
  H_A: Minst én er forskjellig

$\Rightarrow$ $H_0: \mu_1 = \mu_2 = \mu_3 = \ldots\ldots = \mu_n$
H_A: Minst én forskjellig

Marit Helene Gladhaug

- X_{ij} = utfallet for observasjon i tilhørende gruppe j
- $E(X_{ij}) = \mu_j \qquad Var(X_{ij}) = \sigma^2$
- $\hat{\mu}_j = \overline{X}_j = \frac{1}{n_j} \sum X_{ij}$ → gruppeforventning
- Samlet forventningsverdi som er lik for hver gruppe under H_0:

$$\hat{\mu}_j = \overline{\overline{X}} = \frac{1}{n} \sum_{j=1}^{k} \sum_{i=1}^{n_j} X_{ij}$$

- Variasjon mellom gruppene er

$$SST = \sum_{j=1}^{k} n_j \left(\overline{X}_j - \overline{\overline{X}} \right)^2$$

- Variasjon innad i gruppene er

$$SSE = \sum_{j=1}^{k} \sum_{i=1}^{n_j} \left(X_{ij} - \overline{X} \right)^2$$

- Totalvariasjon : $SS(Total) = SST + SSE$

$$SS(Total) = \sum_{j=1}^{k} \sum_{i=1}^{n_j} \left(X_{ij} - \overline{\overline{X}} \right)^2$$

- Testobservator :

$$F = \frac{SST / (k-1)}{SSE / (n-k)} \sim F_{(k-1),(n-k)}$$

Marit Helene Gladhaug

Merk: Dersom variasjonen mellom gruppene (SST) er stor relativt til variasjonen innen gruppene (SSE) er det lite sannsynlig at de ulike gruppene har lik forventning

$\Rightarrow$ SST $\uparrow$ og SSE $\downarrow$: vi deler et stort tall på et lite tall $\Rightarrow$ vi får en stor F-verdi
$\Rightarrow$ Stor F-verdi forkaster H_0 om lik forventning

Merk ANOVA-testen sier ikke hvilken gruppe som skiller seg ut. For å finne den avvikende gruppen kan vi bruke Tukey-testen.

- ANOVA: toveis variansanalyse

- $E(X_{ij}) = \mu_{ij} = \mu + \alpha_i + \beta_j$

- Antar felles varians for alle grupper

- $\hat{\mu} = \bar{\bar{X}} = \frac{1}{ab} \sum_{i=1}^{a} \sum_{j=1}^{b} X_{ij}$

- Testobservatorer:

$$F_A = \frac{SS(A) \,/\, (a-1)}{SSE \,/\, (a-1)(b-1)} \sim F_{(a-1),\,(a-1),\,(b-1)}$$

Marit Helene Gladhaug

- $F_B = \dfrac{SS(B)/(b-1)}{SSE/(a-1)(b-1)} \sim F_{(b-1),(a-1),(b-1)}$

- Fordelingsfrie tester → Når vi ikke har normalfordelte observasjoner

I. Wilcoxons test for to utvalg

$$\mu_1 = \frac{n_1(n_1+n_2+1)}{2}$$

$$\sigma = \sqrt{\frac{n_1 n_2 (n_1+n_2+1)}{12}}$$

H_0: Mediandforskjellen er lik null
H_A: Medianforskjellen er ulik null

II. Wilcoxons test for parrede observasjoner

$$\mu = \frac{n(n+1)}{4}$$

$$\sigma = \sqrt{\frac{n(n+1)(2n+1)}{24}}$$

III. Kruskal-Wallis-testen: Fordelingsfri enveis-variansanalyse

IV. Friedmans-test: Fordelingsfri toveis variansanalyse

Marit Helene Gladhaug

- Kjikvadrattester

I. Kjikvadrattest for modelltilpasning

II. Kjikvadrattest for uavhengighet

H_0: kjennetegnene er uavhengige
H_A: kjennetegnene er ikke uavhengige

$$X^2 = \frac{(X_1 - \overline{X}_2)^2}{\overline{X}_1} + \frac{(X_2 - \overline{X}_2)^2}{\overline{X}_2} + \dots + \frac{(X_n - \overline{X}_n)^2}{\overline{X}_n}$$

DEL II

Enkel regresjon

$$Y = \beta_0 + \beta_1 X_1 + \varepsilon$$

$$E(Y|X) = \beta_0 + \beta_1 X$$

der $\hat{\beta}_1 = \dfrac{Cov(X,Y)}{Var(X)}$

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}$$

Forutsetningene for OLS

(i) $E(\varepsilon) = 0$

Marit Helene Gladhaug

(ii) σ_ϵ er konstant for alle X homoskedastisitet

(iii) ϵ_i er uavhengig av ϵ_j for alle i og j
Brudd på denne forutsetningen kalles autokorrelasjon

(iv) Dersom X-variablene ikke er deterministisk, men stokastisk, må den være uavhengig av ϵ.
Avhengighet mellom X og ϵ kalles endogenitet.

(v) Normalitet: ϵ er normalfordelt.

Standardavvikene

$S_{\hat{\epsilon}}$	Estimert standardavvik til feilleddet
$S(\hat{\beta}_1)$	Estimert standardavvik til stigningstallet
$S(\hat{Y})$	Estimert standardavvik til estimert forventet Y
$S(Y-\hat{Y})$	Estimert standardavvik til prediksjonsfeilen

Marit Helene Gladhaug

(1) Hvorfor en koeffisient kan endre fortegn når en annen forklaringsvariabel inkluderes

Tenk deg at du har en åker, og du måler avlingen Y. Så ønsker du å forklare avlingen med nedbør i en lineær regresjon, og finner ut at koeffisienten blir negativ. Altså, dersom nedbøren øker med en enhet, blir avlingen beta enheter dårligere. Så inkluderer du temperatur som forklaringsvariabel, og får at begge er positive. Det kan skje dersom temperatur og nedbør er *negativt* korrelerte.

Dersom du bare har nedbør, kan vi tenke oss at mye nedbør ofte henger sammen med kaldt vær, og at det ikke er bra for avlingen. Men for en **gitt** temperatur, vil mye nedbør være bra. Derfor kan nedbørskoeffisienten bli positiv når vi kontrollerer for temperatur.

(2) Hvorfor en koeffisient kan gå fra å være signifikant til ikke signifikant (og omvendt) når en annen forklaringsvariabel inkluderes i modellen

En koeffisient er signifikant dersom den er stor i forhold til sitt standardavvik, sjekk formel på s. 35 i forelesning 11. den kan miste sin signifikans pga multikolinearitet (andre faktor) aller få signifikans dersom den nye variabelen forklarer mye variasjon, slik at S_{epsilon} blir mindre (første faktor).

S_ϵ

Marit Helene Gladhaug

(1) White-test : tester heteroskedastisitet

$H_0 : \sigma_1^2 = \sigma_2^2$ for alle i (homoskedastisitet)

$H_A : \sigma_1^2 \neq \sigma_2^2$ for alle i (heteroskedastisitet)

Høy P-verdi ⇒ Behold H_0 ⇒ homoskedastisitet ☺

(2) Kjikvadrattest : tester uavhengighet

H_0 : kjennetegn er uavhengige

H_A : kjennetegn er avhengige

Høy P-verdi ⇒ Behold H_0 ⇒ uavhengighet ☺

- Heteroskedastisitet : estimatene er forventningsrette, men inferens er ikke gyldig.
- Autokorrelasjon : estimatene er forventningsrette, men inferens er ikke gyldig.
- Ikke normalfordelte feilledd : inferens er ikke gyldig i små utvalg.
- Endogenitet : forventningsskjeve estimatorer