Statistik – Formelsamling

HA Almen, 1. semester

Statistik - Formelsamling

Indholdsfortegnelse

Hvordan kan formelsamlingen bruges?

Ud fra den enkelte opgave til eksamen kan man slå op i denne formelsamling for at finde fremgangsmåden til at løse opgaven.

Dette gøres således:
1) Find overemnet som opgaven omhandler - fx "Hypotesetest"
2) Find den specifikke opgavebeskrivelse - fx "Difference mellem 2 populationsmiddelværdier med kendt σ^2".
3) Følg den generelle fremgangsmåde for overemnet.
4) Benyt formlerne for den specifikke opgavebeskrivelse.

Værd at vide

"NCT" henviser til grundbogen i statistik:
"Statistics for Business and Economics" (af Paul Newbold, William Carlson & Betty Thorne)

Oversigt – Mest brugte symboler

Størrelse	Population	Stikprøve
Antal observationer	N	n
Gennemsnit	μ	x
Varians	σ2	s2
Standardafvigelse	σ	s
Variationskoefficienten	CV	CV
Kovarians	Cov(X,Y) = σxy	Cov(X,Y) = sxy
Korrelationskoefficient	p	r

Disclaimer

For at få udbytte af denne formelsamling kræver det et grundlæggende kendskab til faget statistik og forståelse for, hvordan man løser generelle problemstillinger. Er dét på plads, fungerer denne formelsamling som et godt værktøj til at spare tid til eksamen.

Uni Bazaar IVS tager forbehold for tastefejl og ændringer i pensum. Desuden skal det bemærkes, at den præcise brug af symboler kan variere i forhold til den enkelte underviser.

Konfidensintervaller

Generel fremgangmåde

1) Find formel, evt. ved hjælp af træet (bilag i fællesnoter)
2) Gør relevante antagelser
3) Konkluder at den sande populationsvariabel med $100(1-\alpha)\%$ sikkerhed er givet i intervallet.
4) Kommentér evt. på om 0 ligger i intervallet

ME (marginal error) er alt der efter $\pm$ i formlerne og bredden$= 2 * ME$

Populationsmiddelværdi μ , med kendt σ^2

Formel: $\overline{x} \pm z_{\frac{\alpha}{2}} * \frac{\sigma}{\sqrt{n}}$

$z_{\frac{\alpha}{2}}$ findes i NCT på side 738 (eller i den lille tabel på side 294)

Antagelser:

- Kendt populationsvarians
- Normalfordelt population
- Tilfældig udvalgt stikprøve

Difference mellem 2 populationsmiddelværdier med kendt σ^2

Formel: $(\overline{x} - \overline{y}) z_{\frac{\alpha}{2}} * \sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}$

$z_{\frac{\alpha}{2}}$ findes i NCT på side 738 (eller i den lille tabel på side 294)

Antagelser

- Kendte populationsvarianser
- Normalfordelt population
- Tilfældig udvalgt stikprøve

Populationsmiddelværdien, μ , med ukendt σ^2

Formel: $\overline{x} \pm t_{n-1,\frac{\alpha}{2}} * \frac{s}{\sqrt{n}}$

$t_{n-1;\frac{\alpha}{2}}$ findes i NCT på side 770

Antagelser:

- Ukendt populationsvarians
- Normalfordelt population
- Tilfældig udvalgt stikprøve

Differencen mellem 2 populationsmiddelværdier med samme, men ukendt σ^2

S-pooled beregner en fælles varians

$$s_p^2 = \frac{(n_x - 1) * s_x^2 + (n_y - 1) * s_y^2}{n_x + n_y - 2}$$

Formel:
$$(\bar{x} - \bar{y}) \pm t_{n_x+n_y-2;\frac{\alpha}{2}} * \sqrt{\frac{s_p^2}{n_x} + \frac{s_p^2}{n_y}}$$

$t_{n_x+n_y-2;\frac{\alpha}{2}}$ findes i NCT på side 770

Antagelser

- Ukendte men ens varianser
- Normalfordelte populationer
- Tilfældigt udvalgte stikprøver

Populationsvarians, σ^2

Formel:
$$Lower(L(\sigma^2)) = \frac{(n-1) * s^2}{\chi^2_{n-1;\frac{\alpha}{2}}}$$

$$Upper(L(\sigma^2)) = \frac{(n-1) * s^2}{\chi^2_{n-1;-(\frac{\alpha}{2})}}$$

χ^2-fordelingerne findes i NCT på side 768 og 769

Antagelser

- Normalfordelt population
- Populationsvarians der følger chi^2-fordeling
- Tilfældigt udvalgte stikprøver

Differencen mellem 2 populationsvarianser

Ikke en del af pensum

Populationsandel, P

Formel: $\hat{p} \pm Z_{\frac{\alpha}{2}} * \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

$\hat{p} = \frac{X}{n}$

$Z_{\frac{\alpha}{2}}$ findes i NCT på side 738 (eller i den lille tabel på side 294)

Antagelser

- Population er binomialfordelt $X \sim b(n, P)$: to mulige udfald, konstant P og stokastisk uafhængighed
- Den kan approksimeres til en normalfordeling, når $nP(1-P) > 5$ (variansen skal være større end 5)
- Tilfældigt udvalgte stikprøver

Differencen mellem to populationsandele, $P_x - P_y$

Formel: $(\hat{p}_x - \hat{p}_y) \pm Z_{\frac{\alpha}{2}} * \sqrt{\frac{\hat{p}_x(1-\hat{p}_x)}{n_x} + \frac{\hat{p}_y(1-\hat{p}_y)}{n_y}}$

$Z_{\frac{\alpha}{2}}$ findes i NCT på side 738 (eller i den lille tabel på side 294)

Antagelser

- Population er binomialfordelt $X \sim b(n, P_x)$ og $Y \sim b(n, P_y)$: to mulige udfald for X og Y, konstant P og stokastisk uafhængighed
- Den kan approksimeres til en normalfordeling, når $n_x P_x(1-P_x) > 5$ og $n_y P_y(1-P_y) > 5$
- Tilfældigt udvalgte stikprøver

Hypotesetest

Altid stærkere at lave en nulhypotese, der kan forkastes

Fremgangsmåde

- Opstil passende H_0 og H_1

For den højresidede test vil $H_0: \theta \leq \theta$ og $H_1: \theta > \theta$

For den venstresidede test vil $H_0: \theta \geq \theta$ og $H_1: \theta < \theta$

For den dobbeltsidede test vil $H_0: \theta = \theta$ og $H_1: \theta \neq \theta$

$\frac{\alpha}{2}$ benyttes ved den dobbeltsidede test.

- Vælg sikkerhedsniveau (α). Hvis intet er givet, brug 5 %.
- Find den passende formel, evt. ud fra brug af træet
- Gør de relevante antagelser
- Sæt teststørrelsen, T, overfor den kritiske værdi, K. Hvis T er mindre ekstrem end K medfører det, at vi ikke forkaster H_0. Det betyder desuden, at hvis T er mere ekstrem end K, skal vi forkaste H_0

Fortolkninger

Hvis vi beviser H_0 betyder det blot, at vi ikke kan forkaste den. Det betyder IKKE, at den er sand.

Hvis vi modbeviser H_0, kan vi forkaste med $100(1-\alpha)\%$ sikkerhed

P-værdi

Kræves ikke medmindre, der direkte bliver spurgt om det.

P-værdien er sandsynligheden for at observere en mere ekstrem værdi end teststørrelsen, når H_0 er sand. Er P-værdien mindre end α, så forkaster vi.

P-værdien kan især bruges ved grænsesignifikans, da sikkerhedsniveauet kan være afgørende i de tilfælde for om vi forkaster eller ej.

Type I og type II fejl

- Type 1 (α): risikoen for at forkaste en sand H_0
- Type 2 (β): risikoen for ikke at forkaste en falsk H_0

Der er risiko for fejl især ved grænsesignifikans.

Populationsmiddelværdi μ , med kendt σ^2

Teststørrelse: $$Z_{test} = \frac{\overline{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

Kritisk værdi: z_α findes i NCT på side 738 (eller i den lille tabel på side 294)

Antagelser:

- Kendt populationsvarians
- Normalfordelt population
- Tilfældig udvalgt stikprøve

Difference mellem 2 populationsmiddelværdier med kendt σ^2

Teststørrelse: $$Z_{test} = \frac{(\overline{x} - \overline{y}) - D_0}{\sqrt{\frac{\sigma_x^2}{n_x} er + \frac{\sigma_y^2}{n_y}}}$$

D_0 er det vi tester om differencen er $\left(\mu_x - \mu_y = D_0\right)$

Kritisk værdi: z_α findes i NCT på side 738 (eller i den lille tabel på side 294)

Antagelser:

- Kendte populationsvarianser
- Normalfordelt population
- Tilfældig udvalgt stikprøve

Populationsmiddelværdien, μ , med ukendt σ^2

Teststørrelse: $$t_{test} = \frac{\overline{x} - \mu_0}{\frac{s}{\sqrt{n}}}$$

Kritisk værdi: $t_{n-1,\alpha}$ findes i NCT på side 770

Antagelser:

- Ukendt populationsvarians
- Normalfordelt population
- Tilfældig udvalgt stikprøve

Differencen mellem 2 populationsmiddelværdier med ukendt σ^2

Teststørrelse: $$t_{test} = \frac{(\overline{x} - \overline{y}) - D_0}{\sqrt{\frac{s_p^2}{n_x} + \frac{s_p^2}{n_y}}}$$

$$s_p^2 = \frac{(n_x - 1) * s_x^2 + (n_y - 1) * s_y^2}{n_x + n_y - 2}$$

D_0 er det vi tester om differencen er $(\mu_x - \mu_y = D_0)$

Kritisk værdi: $t_{n_x+n_y-2;\alpha}$ findes i NCT på side 770

Antagelser:

- Ukendte men **ens** varianser
- Normalfordelte populationer
- Tilfældigt udvalgte stikprøver

Populationsvarians, σ^2

Teststørrelse: $$\chi^2_{test} = \frac{(n-1) * s^2}{\sigma_0^2}$$

Kritisk værdi (øvre): $\chi^2_{(n-1);\alpha}$

Kritisk værdi (nedre): $\chi^2_{(n-1);-\alpha}$

De kritiske værdier findes i NCT på side 768 og 769

Antagelser:

- Normalfordelt population
- Populationsvarians der følger χ^2-fordeling
- Tilfældigt udvalgte stikprøver

Forholdet mellem 2 populationsvarianser, $\frac{\sigma_x^2}{\sigma_y^2}$

Teststørrelse: $F_{test} = \frac{s_X^2}{s_y^2}$, hvor $s_x^2 > s_y^2$

Deler man to $\chi^2 -$ fordelinger med hinanden, så får man et F-test i stedet.

Kritisk værdi: $F_{n_x-1;n_y-1;\alpha}$ som findes i NCT på side 771-774

Antagelser:

- Ukendte populationsvarianser
- Normalfordelte populationer
- Tilfældigt udvalgte stikprøver

Populationsandel, P

Formel: $Z_{test} = \frac{\hat{p} - P_0}{\sqrt{\frac{P_o(1-P_0)}{n}}}$ hvor $\hat{p} = \frac{X}{n}$

Kritisk værdi: z_α findes i NCT på side 738 (eller i den lille tabel på side 294)

Antagelser

- Population er binomialfordelt $X \sim b(n, P)$: to mulige udfald, konstant P og stokastisk uafhængighed
- Den kan approksimeres til en normalfordeling, når $nP(1-P) > 5$ (variansen skal være større end 5)
- Tilfældigt udvalgte stikprøver

Differencen mellem to populationsandele, $P_x - P_y$

Teststørrelse: $Z_{test} = \frac{(\hat{p}_x - \hat{p}_y) - D_0}{\sqrt{\frac{\hat{p}_0(1-\hat{p}_0)}{n_x} + \frac{\hat{p}_0(1-\hat{p}_0)}{n_y}}}$

$$\hat{p}_0 = \frac{n_x * \hat{p}_x + n_y * \hat{p}_y}{n_x + n_y}$$

Kritisk værdi: z_α findes i NCT på side 738 (eller i den lille tabel på side 294)

Antagelser:

- Population er binomialfordelt $X \sim b(n, P_x)$ og $Y \sim b(n, P_y)$: to mulige udfald for X og Y, konstante P'er og stokastisk uafhængighed

- Den kan approksimeres til en normalfordeling, når $n_x P_x(1-P_x) > 5$ og $n_y P_y(1-P_y) > 5$. CLT er opfyldt når de to foregående formler er korrekte.

- Tilfældigt udvalgte stikprøver

Sandsynlighedsregning

TJEK OM DER ER ANTAGET UAFHÆNGIGHED – det ændrer det hele.

Uafhængighed er ikke det samme som disjoint events. Uafhængige events kan godt have fællesmængde.

Stokastisk uafhængighed

Uafhængighed når: $P(A \cap B) = P(A) * P(B)$ og $P(A|B) = P(A)$

Additionsreglen (Probability that event A or event B occurs)

$P(A \cup B) = P(A) + P(B) - P(A \cap B)$
$P(A \cap B) = P(A) + P(B) - P(A \cup B)$

Er de to events disjoint (ingen fællesmængde) så kan man nøjes med addere P(A) og P(B) for at finde den forenede mængde.

$\cup$ = forenet (union of events). Se side 112 for illustration.

$\cap$ = fælles (intersection)

Betinget sandsynlighed (når, givet, hvis)

$P(A|B)$ (siges som A givet B. "sandsynligheden for at være statistiklærer (A) givet man er kvinde (B)").

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Multiplikationsreglen (og)

$P(A \cap B) = P(A|B) * P(B)$

$P(A \cap B) = P(B|A) * P(A)$

Ved uafhængighed, da er de betingede sandsynligheder lig den oprindelige sandsynlighed: $P(A|B) = P(A)$ hvorfor multiplikationsreglen i stedet bliver $P(A \cap B) = P(A) * P(B)$.

Bayes sætning

Også en givet sandsynlighed (når, givet, hvis). Multiplikationsregel i tælleren for betinget sandsynlighed

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)} = \frac{P(A \cap B)}{P(B)}$$

Marginal sandsynlighed

$$P(A) = \sum (P(A \cap E_i))$$

$$P(A) = \sum (P(alle) * P(E_i))$$

Kobling til binomialfordeling

A, B, osv. Kunne være $P(X = x), P(X < x), P(X > x)$

Regressionsanalyse

Grundlæggende formler/forklaringer

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

Y = responsvariabel

X = kovariater/forklarende variabler

β_0 = intercept (skæring med y-aksen)

β_1 = hældning

ε = fejlled/residualer

Parameterestimater, Simpel regression (side 419)

$$\hat{\beta}_1 = \frac{cov(x,y)}{s_x^2}$$

$$\hat{\beta}_0 = \bar{y} - (\hat{\beta}_1 * \bar{x})$$

Fortolkning

Simpel eller lineær regression?

β_0 : Y har en **forventet** værdi på β_0 enheder(y). Det sker når alle kovariater (ved simpel bare den ene kovariat) er lig 0. Værdien for β_0 giver ikke altid mening i sig selv – så er det vigtigt at nævne! Fx hvis vi har negative værdier for noget, der ikke bør kunne være negativt.

β_1 : Y har en **forventet** stigning/fald på β_1 enheder(Y), når vi X_1 siger med en enhed(X_1).

β_j : Findes kun ved multipel lineær regression. Y har en **forventet** stigning/fald på β_j enheder(Y), når X_j enheder(X_j), hvis man holder de andre kovariater konstant.

Variationsstørrelser

SSR: Den del af variationen, som modellen forklarer.

SSE: Den del af variationen, som modellen ikke forklarer

SST: Den totale variation $(SST = SSR + SSE)$

SSR: $(\hat{\beta}_1)^2 * \sum_{i=1}^{n}(x_i - \bar{X})^2 \rightarrow (\hat{\beta}_1)^2 * s_x^2 * (n-1)$

SST: $\sum(y_i - \bar{y})^2 \rightarrow s_y^2 * (n-1)$

MSR: $\frac{SSR}{1}$

MSE: $\frac{SSE}{n-2} = s_e^2$

F-teststørrelse (ratio): $\frac{MSR}{MSE} = \frac{SSR}{S_e^2}$

Parameterestimater:

$$\hat{\beta}_1 = \frac{S_{xy}}{s_x^2} = r_{xy} * \frac{s_y}{s_x}$$

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

Det vil sige, at regressionslinjen går gennem punktet $(\bar{X}, \bar{Y})$

Determinationskoefficienten:

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SEE}{SST} = r_{xy}^2$$

Vigtigt at bemærke at formlen også kan skrives som $R^2 = \frac{(\hat{\beta}_1)^2 \sum_{i=1}^{n}(X_i - \bar{X})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$. Formlen viser, at forklaringskraften vokser med variabiliteten af kovariaterne om deres gennemsnit. Dvs. at R^2 er større når $\sum_{i=1}^{n}(X_i - \bar{X})^2$ er større. Man skal derfor forsøge at rbuge kovariater med så stor varians som muligt for på den måde at opnå den størst mulige forklaringskraft i regressionsmodellen.

Justeret determinationskoefficient:

$$\bar{R}^2 = 1 - \frac{\frac{SSE}{n-K-1}}{\frac{SST}{n-1}}$$

Fejlleddenes varians:

$$s_e^2 = \frac{SSE}{n - K - 1}$$

Std. Error på den enkelte β_j :

$$S_{\beta_j}^2 = \frac{s_e^2}{(n-1) * s_{xj}^2}$$

R^2 kan øges kunstigt, hvis man tilfører flere kovariater – selv hvis de ingen forklaringskraft har. R^2 adj. Er justeret for dette.

Antagelser og kontrol af disse

Antagelse om lineæritet

Der skal være lineær sammenhæng mellem responsvariablen og alle kovariater.

Kontrol:

- Lav en graf, der viser Y mod X'erne:
- Led efter lineære sammenhænge. Finder man en eksponentiel, kvadratisk eller anden sammenhæng, så kan der anbefales en transformation til en lineær sammenhæng. Vi kommer ikke selv til at skulle lave transformationen, men vi kan foreslå at gøre det.

Antagelse om normalfordelte og uafhængige fejlled

Residualerne ε_i er uafhængige af kovariaterne for alle i og de er normalfordelte med middelværdien 0. Vi antager at middelværdien er 0.

Kontrol:

- Lav e graf med de rå eller studentiserede fejlled mod X'erne. Der må ikke være nogen mønstre og de skal ligge omkring 0.
- Normalfordelingsplot af de rå eller studentiserede residualer. Punkterne må ikke ligge uden for båndende ved 95 %. Ligger der enkelte punkter udenfor båndene er det ok, hvis n er stor.
- Test for normalfordelte residualer på de studentiserede eller de rå (Jaque Bera testet). Forkastes nulhypotesen, betyder det, at fejlleddene ikke er normalfordelte.

Homoskedasticitet

Residualerne er homoskedastiske, hvilket vil sige, at de har konstant varians: $Var(\varepsilon_i) = \sigma^2$ for alle i .

Kontrol:

- Lav en graf med de studentiserede residualer mod hhv. row eller predicted Y.
- Der skal være ens varians (spredning) over hele x-aksen
- Test for homoskedasticitet. Forkastes nulhypotesen, betyder det at vi har heteroskedastiske fejlled.

Parvis uafhængighed

Residualerne er parvis uafhængige, dvs. at ved at have observeret en ε_i kan vi ikke sige noget om den næste ε_j. Der må ikke være systematik i fejlene.

Kontrol:

- Lav en graf med studentiserede residualer mod row eller predicted Y. Se efter mønstre i plottet.

Multikolinaritet (multipel regression)

En kovariat må ikke være en linearkombination af en anden kovariat, dvs. de ikke må forklare det samme om Y.

Kontrol:

- Lav et korrelationsmatrix med alle de numeriske kovariater
- Ingen må overstige 0,7 numerisk set (dvs. større end 0,7 og mindre end -0,7)

Der findes eksempler plots (uafhængighed og homoskedasticitet) tegnet på papir.

Jaque Bera-test

H_0 : Normalfordelte residualer

H_1: Ikke normalfordelte residualer (komplement til H_0)

Teststørrelse: $JB_{test} = n\left(\frac{(skewness)^2}{6} + \frac{(kurtosis-3)^2}{24}\right)$ OBS: JMP har allerede trukket de 3 fra.

Kritisk værdi: $JB_{n,\alpha}$ som findes i NCT på side 612. Forkast H_0 hvis teststørrelsen er større end den kritiske værdi.

Når testet har stort nok n kan den approksimeres til en χ^2-fordeling.

Test for homoskedasticitet (Whites test)

H_0 : Homoskedasticitet

H_1: Heteroskedasticitet (hvis der denne lineære sammenhæng: $\varepsilon^2 = \beta_0 + \beta_1 * \hat{Y} = R^2_{\hat{\varepsilon}^2}$)

Teststørrelse:

Kritisk værdi: $\chi^2_{1;\alpha}$

Forkast H_0 hvis teststørrelsen er større end den kritiske værdi.

Test for simpel regression

Test af ingen lineær sammenhæng, hvor

$H_0: \beta_j = 0$ og $H_1: \beta_j \neq 0$

Test 1

Teststørrelse: $t_{test} = \frac{\widehat{\beta_j}}{s_{\beta j}}$

Kritisk værdi: $t_{n-K-1;\frac{\alpha}{2}}$

Test 2

Teststørrelse: $F_{test} = \frac{\frac{SSR}{K}}{S_\varepsilon^2} = \frac{\frac{SSR}{K}}{\frac{SSE}{n} - K - 1} = \frac{MSR}{MSE}$

Kristik værdi: $F_{K;n-K-1;\alpha}$

Test for multipel regression

Test 1

Test af ingen marginaleffekt af den j'te kovariat

$H_0: \beta_j = 0$ og $H_1: \beta_j \neq 0$

Teststørrelse: $t_{test} = \frac{\widehat{\beta_j}}{s_{\beta j}}$

Kritisk værdi: $t_{n-K-1;\frac{\alpha}{2}}$

Test 2

Test af ingen simultaneffekt af K antal kovariater

$H_0: \beta_1, \quad \beta_2, \dots, \beta_K = 0$ og $H_1: mindt\ én\ \beta_j \neq 0$

Teststørrelse: $F_{test} = \frac{\frac{SSR}{K}}{S_\varepsilon^2} = \frac{\frac{SSR}{K}}{\frac{SSE}{n} - K - 1} = \frac{MSR}{MSE}$

Kristik værdi: $F_{K;n-K-1;\alpha}$

Test 3

$H_0: \alpha_1 = \alpha_2 = \alpha_K = 0$ og $H_1: mindst\ én\ \alpha \neq 0$

Test af ingen simultaneffekt af delmængden R ud af K kovariater. Findes der kun én ny kovariat (R=1) er det ikke simultan- men marginaleffekt. Vigtigt at notere.

Teststørrelse: $F_{test} = \frac{\frac{SSE(R) - SSE}{R}}{\frac{SSE}{n - K - 1}}$

Hvor SSE(R) er fra den gamle model

SSE er fra den nye model

K er antal kovariater i den model med færrest kovariater

R er antal tilføjede kovariater

Kritisk værdi: $F_{R;n-K-R-1;\alpha}$

Prediktionsinterval (PI)

$$\bar{y}_{n+1} \pm t_{n-2;\alpha/2} * \sqrt{\left(1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{X})^2}{\sum(x_i - \bar{x})^2}\right)} * S_\varepsilon$$

Nævneren kan også skrives som $(n-1) * s_x^2$

Nogle gange må x_{n-1} antage samme værdi som $\overline{X}$, hvorved hele det sidste led bortfalder.

Dette interval indeholder med $100(1-\alpha)\%$ sikkerhed værdien af en **ny** observation Y_{n+1}, når X antager værdien x_{n-1}.

Konfidensinterval (KI)

$$\bar{y}_{n+1} \pm t_{n-2;\alpha/2} * \sqrt{\left(\frac{1}{n} + \frac{(x_{n+1} - \bar{X})^2}{\sum(x_i - \bar{x})^2}\right)} * S_\varepsilon$$

Dette interval indeholder med $100(1-\alpha)\%$ sikkerhed værdien af Y_{n+1}, når X antager værdien x_{n-1}.

Konfidensintervallet er altid mindre bredt end prediktionsintervallet og derved mere "sikkert".

Konfidensinterval for β_j

$\hat{\beta}_j \pm t_{n-K-1;\frac{\alpha}{2}} * s_{\beta\square}$

Ikke-parametrisk statistik

Goodness of fit 1 faktor

1 faktor

Parametre:

O_i : antal observationer i kategori i

P_i: Sandsynligheden for at ende i kategori i

E_i: forventet antal i kategori i

$E_i = n * P_i$

H_0: der er Goodness of fit. Det kan også skrives som $P_1, P_2, \dots, P_K$ er korrekt specificeret.

H_1: $P_1, P_2, \dots, P_K$ er specificeret forkert.

Teststørrelse: $$\chi^2_{test} = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

Kritisk værdi: $\chi^2_{k-1;\alpha}$ som findes i NCT på side 768. K angiver antallet af kategoier.

Fokast H_0 hvis teststørrelsen er større end den kritiske værdi.

Antagelse

- n er tilstrækkelig stor, sp $n * P_i = E_i > 5$ for hver i

2 faktorer (kontingenstabeller)

Faktor A/Faktor B	1	2	…	c	Total
1	O_{11}	O_{12}	…	O_{1c}	R_1
2	O_{21}		…		R_2
…	…	…	…	…	…
r	O_{r1}		…		R_r
Total	C_1	C_2	…	C_c	n

Faktor A har r kategorier (rækker) hvilket vil sige at $i = \mathbf{1}$ til $i = r$

Faktor B har c kategorier (kolonner) hvilket vil sige at $j = \mathbf{1}$ til $j = c$

H_0: uafhængig mellem faktor A og B

H_1: Afhængighed

$$E_{ij} = \frac{R_i * C_j}{n}$$

Teststørrelse: $$\chi^2_{test} = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Kritisk værdi: $\chi^2_{(r-1)(c-1),\alpha}$

Vi forkaster, hvis teststørrelsen er større end den kritiske værdi.

Antagelse:

- R_i og C_j er tilstrækkeligt støre, så $E_{ij} > 5$ for hver i og j

Variansanalyse

Et-sidet variansanalyse

Antal populationer, hvor vi vil teste ens/forskellig middelværdi, men med ens varians. Kategoriske variable.

$H_0: \mu_1 = \mu_2 = \cdots = \mu_k$

$H_1: mindst\ 2\ af\ \mu\ er\ forskellige$

Teststørrelse: $$F_{test} = \frac{MSG}{MSW} = \frac{\frac{SSG}{K-1}}{\frac{SSW}{n-K}}$$

Hvor K er antallet af grupper

N er antallet af observationer

SSG er variationen mellem grupper

SSW er variationen indenfor grupperne

Kritisk værdi: $F_{(K-1)(n-K);\alpha}$ som findes i NCT på side 771-774

Forkast H_0 når teststørrelsen er større end den kritiske værdi.

Antagelser:

- Normalfordelte populationer
- Uafhængige stikprøver
- Varianshomogenitet

ANOVA-tabel

Variation	SS	frihedsgrader	MS	F-teststørrelse
Mellem grp.	SSG	$K-1$	$MSG = \frac{SSG}{K-1}$	$F = \frac{MSG}{MSW}$
Indenfor grp.	SSW	$n-K$	$MSW = \frac{SSW}{n-K}$	
Total	SST	$n-1$		

To-sidet variansanalyse

BLOK / GRUPPE	1	2	…	K	
1	$X_{111}, X_{112}, \dots, X_{11m}$		…	$X_{K11}, X_{K12}, \dots, X_{K1m}$	
2			…		
…	…	…	…	…	
H	$X_{1H1}, \dots X_{1Hm}$		…	$X_{KH1}, \dots, X_{KHm}$	

K er antallet af grupper i gruppefaktoren

H er antallet af grupper i blokfaktoren

M er antal observationer indenfor hvert niveau.

Hypotesetest

Forkaster man en af nedenstående hypoteser, så er der altså en effekt af en af faktorerne. For alle test gælder det, at vi forkaster H_0, hvis teststørrelsen er større end den kritiske værdi.

Test 1 – ingen gruppeeffekt

$H_0: G_1 = G_2 = \cdots = G_k = 0$

Teststørrelse: $F_{test} = \frac{MSG}{MSE}$

Kritisk værdi: $F_{(K-1);(KHm-1);\alpha}$

Test 2 – ingen blokeffekt

$H_0: B_1, B_2, \ldots, B_k = 0$

Teststørrelse: $F_{test} = \frac{MSB}{MSE}$

Kritisk værdi: $F_{(H-1);(KHm-1);\alpha}$

Test 3 – ingen vekselvirkningseffekt

$H_0: L_{ij} = 0 \; for \; alle \; ij$

Teststørrelse: $F_{test} = \frac{MSI}{MSE}$

Kritisk værdi: $F_{((K-1)(H-1));(KHm-1);\alpha}$

Antagelser:

- Normalfordelte populationer
- Uafhængige stikprøver
- Varianshomogenitet

ANOVA-tabel

Variation	SS	f.g.	MS	F-teststørrelse
Mellem grupper	SSG	$K-1$	$MSG = \frac{SSG}{K-1}$	$F = \frac{MSG}{MSE}$
Mellem blokke	SSB	$H-1$	$MSB = \frac{SSB}{H-1}$	$F = \frac{MSB}{MSE}$
Vekselvirkning	SSI	$(K-1)(H-1)$	$MSI = \frac{SSI}{(K-1)(H-1)}$	$F = \frac{MSI}{MSE}$
Fejl	SSE	$KH(m-1)$	$MSE = \frac{SSE}{KH(m-1)}$	
Total	SST	$KHm-1$		

Til hver af tabellens 3 første rækker svarer en nulhypotese, som forkastes på niveau α hvis rækkens F-teststørrelse opfylder $F > F_{v,KH(m-1),\alpha}$, hvor v er frihedsgraderne i rækkens 'f.g.'-kolonne.

Fordelinger

Binomialfordeling $X \sim b(n, P)$

To mulige udfald: Succes eller fiasko hvor succes er det, vi leder efter.

P: Succes-sandsynligheden

$(1-P)$: fiasko-sandsynligheden

n : antal uafhængige forsøg

SSH. For at få bestemt x: $E[x] = n * P$ og variansen af bestemt x: $Var(x) = n * P(1-P)$

$$P(X = x) = \frac{n!}{x!\,(n-x)!} * P^x * (1-P)^{n-x}$$

$$P(X < x) = P(X = 0) + P(X = 1) + \cdots + P(X = x)$$

$$P(X > x) = P(X = n) + P(X = n-1) + \cdots + P(X = n-x)$$

$P(X > x) = 1 - P(X < x)$ fordi sandsynligheden altid summer som 1

Standardiseret normalfordeling $Z \sim N(0, 1)$

Middelværdien er 0 og **variansen** og standardafvigelsen er 1

Der kan transformeres til standardnormalfordelingen: $z = \frac{x-\mu}{\sigma}$

$$P(X < x) = P(Z < z) = P\left(Z < \frac{x-\mu}{\sigma}\right)$$

$$P(X > x) = P(Z > z) = 1 - P\left(Z < \frac{x-\mu}{\sigma}\right)$$

$$P(a < X < b) = P(a < Z < b) = P\left(\frac{a-\mu}{\sigma} < Z < \frac{b-\mu}{\sigma}\right) = P(X < b) - P(X < a) = F(b) - F(a)$$

$\frac{x-\mu}{\sigma}$ giver en værdi i Z-fordelingen som svarer til en SSH. DETTE ER SVARET.

Hvis: Z er negativ $1 - F(|z|)$

Approksimation af binomialfordeling til normalfordeling

Fra kategorisk til numerisk. Har man mange observationer, så ligner binomialfordelingen næsten en kontinuert linje. Vi approksimerer, fordi binomialfordelingen fordi den er meget regnetung.

Må anvendes når $nP(1-P) > 5$ (altså: variansen skal være større end 5). Det gælder fra ca. $n = 50$ (tommelfingerregel).

Udføres vha. transformation: $Z = \frac{x - nP}{\sqrt{nP(1-P)}}$

Sandsynlighederne findes ved:

$$P(X < x) = P(Z < z) = P\left(Z < \frac{x - nP}{\sqrt{nP(1-P)}}\right)$$

$$P(X > x) = P(Z > z) = 1 - P\left(Z < \frac{x - nP}{\sqrt{nP(1-P)}}\right)$$

$$P(a < X < b) = P(a < Z < b) = P\left(\frac{a - nP}{\sqrt{nP(1-P)}} < Z < \frac{b - nP}{\sqrt{nP(1-P)}}\right) = P(X < b) - P(X < a) = F(b) - F(a)$$