STATISTIKNOTER

Mindre matematisk-statistisk opslagsværk, indeholdende bl.a. ordforklaringer, resuméer og tabeller

Jørgen Larsen

IMFUFA
Roskilde Universitetscenter

Februar 1999

IMFUFA, Roskilde Universitetscenter, Postboks 260, DK-4000 Roskilde.
Jørgen Larsen: **STATISTIKNOTER: Mindre matematisk-statistisk opslagsværk, indeholdende bl.a. ordforklaringer, resuméer og tabeller**
IMFUFA tekst nr. 304e/1995 55 sider ISSN 0106-6242

Dette hæfte er en del af undervisningsmaterialet til et kursus i statistik og statistiske modeller. Undervisningsmaterialet omfatter blandt andet følgende titler:

a. Simple binomialfordelingsmodeller
b. Simple normalfordelingsmodeller
c. Simple Poissonfordelingsmodeller
d. Simple multinomialfordelingsmodeller
e. Mindre matematisk-statistisk opslagsværk, indeholdende bl.a. ordforklaringer, resuméer og tabeller

•

Om kurset og kursusmaterialet kan blandt andet siges at

- når det er et gennemgående tema at påpege at likelihoodmetoden kan benyttes som et overordnet princip for valg af estimatorer og teststørrelser, er det blandt andet begrundet i *at* likelihoodmetoden har mange egenskaber der fra et matematisk-statistisk synspunkt anses for ønskelige, *at* likelihoodmetoden er meget udbredt og nyder stor anerkendelse (ikke mindst i Danmark), og *at* det i al almindelighed er værd at gøre opmærksom på at man også inden for faget statistik har overordnede og strukturerende begreber og metoder;
- når kursusmaterialet er skrevet på dansk (og ikke for eksempel på 'scientific English'), er det for at bidrage til at vedligeholde traditionerne for *hvordan* og *at* man kan tale om slige emner på dansk, og så sandelig også fordi dansk er det sprog som forfatteren – og vel også den forventede læser – er bedst til;
- når hæfterne foruden de sædvanlige simple modeller, metoder og eksempler også indeholder eksempler der er væsentligt sværere, er det for at antyde nogle af de retninger man kan arbejde videre i, og for at der kan være lidt udfordringer til den krævende læser.

Indhold

1 Symboler og notation

I statistik og sandsynlighedsregning har man, som i alle andre matematikbaserede fag, visse hævdvundne traditioner for valg af symboler og notation.

1.1 Nogle vigtige symboler

$\mathrm{E}(\dots)$: middelværdien af ...

$\mathbf{N}$: de naturlige tal ($\{1, 2, 3, \dots\}$).

$\mathcal{N}(\mu, \sigma^2)$: normalfordelingen med middelværdi μ og varians σ^2.

$\mathrm{P}(\dots)$: sandsynligheden for ...

$\mathbf{R}$: de reelle tal.

$\mathbf{Z}$: de hele tal.

$\mathrm{Var}(\dots)$: variansen af ...

$\binom{n}{k}$: binomialkoefficient (s.d.).

$\sum$ (summationstegn): Hvis $z_1, z_2, \dots, z_n$ er nogle tal, så er

$$\sum_{i=1}^{n} z_i = z_1 + z_2 + \dots + z_n$$

og mere generelt

$$\sum_{i=a}^{b} z_i = z_a + z_{a+1} + \dots + z_b$$

I disse udtryk er i en såkaldt summationsvariabel der benyttes til at indicere de forskellige led (z-erne). I eksemplerne har vi kaldt summationsvariablen i, men den kunne lige så godt hedde k eller p eller hvad man nu har lyst til.

Hvis man har med en dobbelt-indiceret størrelse at gøre, så kan man summere over enten det ene eller det andet eller begge indices:

$$\begin{aligned}
\sum_{j=1}^{s} z_{ij} &= \underbrace{z_{i1} + z_{i2} + \ldots + z_{is}}_{s \text{ led}} \\
\sum_{i=1}^{r} z_{ij} &= \underbrace{z_{1j} + z_{2j} + \ldots + z_{rj}}_{r \text{ led}} \\
\sum_{i=1}^{r} \sum_{j=1}^{s} z_{ij} &= \underbrace{z_{11} + z_{12} + \ldots + z_{rs}}_{rs \text{ led}}
\end{aligned}$$

$\prod$ (produkttegn): Hvis $z_1, z_2, \ldots, z_n$ er nogle tal, så er

$$\prod_{i=1}^{n} z_i = z_1 \cdot z_2 \cdot \ldots \cdot z_n$$

og mere generelt

$$\prod_{i=a}^{b} z_i = z_a \cdot z_{a+1} \cdot \ldots \cdot z_b$$

I øvrigt svarer konventionerne for produkttegn nøje til dem for summationstegn.

1.2 Hatte, streger, punkter mm.

1.2.1 Hatte

Et parameternavn med en *hat* over er en betegnelse for et *estimat* over den pågældende parameter, for eksempel er $\hat{\theta}$ en betegnelse for et estimat over θ. (For at øge forvirringen kan det også være betegnelse for en *estimator* over parameteren.) – Hvis man estimerer parameteren under forskellige hypoteser, kan man give de nye estimater nogle flere hatte for at kunne skelne dem fra hinanden, f.eks. $\hat{\hat{\theta}}$.

Hatten bruges næsten altid til at betegne *maksimaliseringsestimatet*. Hvis man benytter alternative estimationsmetoder, kan man betegne estimaterne med f.eks. en tilde ($\tilde{\theta}$).

Hatten kan desuden placeres over et observationsnavn for at betegne den fittede værdi, f.eks. $\hat{y}_i$.

1.2.2 Streger

Et variabelsymbol med en *streg* over er en betegnelse for *gennemsnittet* af nogle værdier af den pågældende variabel. For eksempel kan man betegne gennemsnittet af $y_1, y_2, \ldots, y_n$ med $\bar{y}$.

Hvis man har en dobbelt-indiceret størrelse, f.eks. y_{ij} hvor $i = 1, 2, \ldots, k$ og $j = 1, 2, \ldots, n_i$, så betegner $\bar{y}_{i\bullet}$ gennemsnittet over j af y-erne:

$$\bar{y}_{i\bullet} = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij}$$

1.2.3 Punkter

Hvis man har med med en indiceret størrelse at gøre, f.eks. y_i hvor $i = 1, 2, \ldots, n$, så betegner størrelsen hvor der står et *punkt* på indexets plads, *summen* af de oprindelige størrelser:

$$y_{\bullet} = \sum_{i=1}^{n} y_i$$

Hvis man har med en dobbelt-indiceret størrelse at gøre, så kan man summere over enten det ene eller det andet eller begge indices, og i alle tilfælde kan man markere det med et punkt på det pågældende index's plads.

Eksempel: Hvis man har størrelserne y_{ij}, hvor $i = 1, 2, \ldots, r$ og $j = 1, 2, \ldots, s$, så er

$$\begin{aligned} y_{i\bullet} &= \sum_{j=1}^{s} y_{ij} \\ y_{\bullet j} &= \sum_{i=1}^{r} y_{ij} \\ y_{\bullet\bullet} &= \sum_{i=1}^{r} \sum_{j=1}^{s} y_{ij} \end{aligned}$$

1.3 Græske/romerske bogstaver

Parametre i statistiske modeller navngives ofte med *græske* bogstaver så som μ, σ^2, λ, θ (men der er dog undtagelser, for eksempel betegnes sandsynligheds-parameteren i binomialfordelingen ofte p).

Konstanter og observationer navngives altid med *romerske* bogstaver.

1.4 Store/små bogstaver

I en del sammenhænge betegner man stokastiske variable med store bogstaver (f.eks. $Y_1, Y_2, \ldots, Y_n$) og de observerede værdier af dem med de tilsvarende små bogstaver (f.eks. $y_1, y_2, \ldots, y_n$).

Denne regel har dog mange undtagelser, for eksempel en lang række etablerede standardnavne for stikprøvefunktioner og teststørrelser (så som t, F, s^2, SS, Q, $-2 \ln Q$) der betegnes på samme måde hvad enten de opfattes som stokastiske variable eller som observerede værdier. (Undertiden kan man give den observerede værdi fodtegnet »obs« for at kunne skelne (eksempel: t_{obs}).)

1.5 Bogstaver med en speciel betydning

1.5.1 Græske bogstaver

α (alfa):

1. Anvendes undertiden i en linies ligning som betegnelse for skæringen med y-aksen (Eksempel: $y = \alpha + \beta x$).
2. Når man taler om fraktiler, bruges α undertiden som betegnelse for et typisk tal mellem 0 og 1.

β (beta): Betegner ofte en regressionskoefficient.

Γ (stort gamma): Gamma-funktionen.

δ (delta): Betegner ofte en parameter som angiver en differens.

Δ (stort delta): Benyttes ofte som »præfiks« til en størrelse for at angive en lille tilvækst af den pågældende størrelse, for eksempel Δt.

ε (epsilon): Betegner undertiden testsandsynligheden.

θ (theta): Betegner ofte en generisk parameter.

λ (lambda): Betegner ofte en intensitet.

μ (my): Betegner ofte en middelværdiparameter.

π (pi): Forholdet mellem en cirkels omkreds og dens diameter.

ρ (ro): Betegner ofte en korrelationskoefficient.

σ (sigma): Betegner ofte standardafvigelsen i en fordeling.

σ^2 : Betegner ofte variansen i en fordeling.

φ (phi): Tæthedsfunktionen for $\mathcal{N}(0,1)$-fordelingen.

Φ (stort phi): Fordelingsfunktionen for $\mathcal{N}(0,1)$-fordelingen.

χ^2 (khi i anden): Benyttes i forbindelser som χ^2-fordeling.

χ^2_f : Betegner ofte en stokastisk variabel som er χ^2-fordelt med f frihedsgrader.

1.5.2 Romerske bogstaver

E : Symbol der står for »middelværdien af«.

f :

1. Betegner antal frihedsgrader.
2. Navnet på den typiske modelfunktion.

F : F-teststørrelse.

F_{f_1,f_2} : Betegner ofte en stokastisk variabel som er F-fordelt med frihedsgrader f_1 og f_2.

H : En hypotese.

L : Likelihoodfunktionen.

$\mathbf{N}$: De naturlige tal.

p :

1. Et tal mellem 0 og 1.
2. Ofte sandsynlighedsparameteren i binomialfordelingen.

P : Symbol der står for »sandsynligheden for«

Q : Kvotientteststørrelse.

$\mathbf{R}$: De reelle tal.

s^2 : Variansskøn (som i normalfordelingsmodeller).

SS : Sum af kvadratiske afvigelser (engelsk: *Sum of Squared deviations*).

SP : Sum af produkter af afvigelser (engelsk: *Sum of Products of deviations*).

t : t-teststørrelse.

t_f : Betegner ofte en stokastisk variabel som er t-fordelt med f frihedsgrader.

$X, Y, Z, \ldots$: »typiske« navne på stokastiske variable.

$\mathcal{X}$: Betegner ofte et udfaldsrum (udfaldene betegnes så x).

x : I regressionssammenhænge det typiske navn på en forklarende variabel (baggrundsvariabel).

y : Det typiske navn på en observation.

2 Ordforklaringer og definitioner

Nedenfor er en alfabetisk ordnet liste med definitioner og forklaringer på en del centrale statistiske begreber.

antalsvariabel: En variabel der angiver *antal*, og som derfor kun antager hele ikke-negative værdier.

baggrundsvariabel (også kaldet forklarende variabel eller x-variabel) er i regressionssammenhænge en variabel der inddrages i modellen med håbet/formodningen om at man derved kan få »forklaret« noget af variationen i den variabel der skal modelleres.

Bartletts test for varianshomogenitet er et test for om et antal forskellige normalfordelte observationssæt kan antages at have samme varians, se Resumé 3.11.

binomialfordeling: Binomialfordelingen med antalsparameter n og sandsynlighedsparameter p er den sandsynlighedsfordeling der har punktsandsynligheder $\binom{n}{y} p^y (1-p)^{n-y}, \quad y \in \{0, 1, 2, \ldots, n\}$.

Middelværdien i fordelingen er np, og variansen er $np(1-p)$.

binomialformlen er formlen $(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k}$ der gælder for $n = 0, 1, 2, \ldots$ og for vilkårlige a og b.

binomialkoefficient: Tallet $\binom{n}{k} = \frac{n!}{k!\,(n-k)!}$ kaldes en binomialkoefficient og er lig antallet af måder hvorpå man kan udvælge k forskellige elementer fra en mængde med n elementer.

central estimator (engelsk: *unbiased estimator*): En estimator $\hat{\theta}$ over parameteren θ er central hvis middelværdien af $\hat{\theta}$ er lig θ.

determinationskoefficienten (eller kvadratet på den multiple korrelationskoefficient) R^2 er en størrelse som kan udregnes i forbindelse med lineær regressionsanalyse, og som er kvadratet på korrelationskoefficienten mellem de observerede og de fittede værdier (eller den del af den samlede variation omkring totalgennemsnittet der beskrives af regressionsmodellen),

$$R^2 = \frac{\left(\sum_{i=1}^{n}(y_i-\bar{y})(\hat{y}_i-\bar{y})\right)^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2 \cdot \sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2} = \frac{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}.$$

R^2 kan kun benyttes når der er et konstantled med i regressionen.

Determinationskoefficienten er i almindelighed et dårligt mål for hvor godt modellen passer.

diskret stokastisk variabel: En stokastisk variabel X er diskret hvis der findes en endelig eller tællelig mængde $\mathcal{X} = \{x_1, x_2, x_3, \dots\} \subseteq \mathbf{R}$ med den egenskab at $\mathrm{P}(X \in \mathcal{X}) = 1$.

Fordelingsfunktionen for X er stykkevis konstant og har spring af størrelsen $\mathrm{P}(X = x_i)$ i punktet x_i.

dispersionstest: Et test for om en stikprøve $y_1, y_2, \dots, y_n$ kan antages at stamme fra en Poissonfordeling. Teststørrelsen er $d = s^2/\bar{y}$ der skal være cirka 1 hvis modellen er rigtig; d er asymptotisk χ^2/f-fordelt med $f = n-1$ frihedsgrader.

eksakt test: En testmetode hvor testsandsynligheden udregnes eksakt (i modsætning til testmetoder hvor bestemmelsen af testsandsynligheden baseres på approksimationer så som $-2\ln Q$'s asymptotiske χ^2-fordelthed).

Eksakte tests foretages ofte i en betinget model hvor der betinges med de stikprøvefunktioner der er sufficiente under den testede hypotese. *Fishers eksakte test* for uafhængighed i en 2×2-tabel (eller til sammenligning af to binomialfordelinger) er et sådant betinget test hvor der betinges med række- og søjlesummerne.

ensidet variansanalyse (engelsk: *one-way analysis of variance*): En metode til sammenligning af *middelværdierne* i forskellige normalfordelte stikprøver, se Resumé 3.10.

estimat: et estimat over en bestemt ukendt parameter er den talværdi man på baggrund af et givet sæt observationer udregner som skøn over den pågældende parameter.

estimator: en estimator for en bestemt ukendt parameter er en funktion der til hvert muligt sæt observationer knytter det tilsvarende estimat over parameteren.

***F*-fordeling:** F-fordelingen med frihedsgrader f_1 og f_2 er den kontinuerte sandsynlighedsfordeling på den positive halvakse som har tæthedsfunktion

$$x \mapsto \frac{\Gamma(\frac{f_1+f_2}{2})}{\Gamma(\frac{f_1}{2})\Gamma(\frac{f_2}{2})} f_1^{\frac{f_1}{2}} f_2^{\frac{f_2}{2}} \frac{x^{\frac{f_1}{2}-1}}{(f_2+f_1x)^{\frac{f_1+f_2}{2}}}, \quad x > 0.$$

I normalfordelingsmodeller optræder F-fordelingen som fordelingen af forholdet mellem to uafhængige variansskøn (f.eks. s_1^2/s_0^2); F-fordelingens to frihedsgradstal er da antallene af frihedsgrader for henholdsvis tælleren og nævneren.

fakultetsfunktionen er defineret ved

$$n! = \begin{cases} 1 & \text{hvis } x = 0 \\ n \cdot (n-1)! & \text{hvis } x = 1, 2, 3, \ldots \end{cases}$$

dvs. $n! = 1 \cdot 2 \cdot 3 \cdot \ldots \cdot n$. (Se også *Gammafunktionen*.)

fittede værdier: Hvis man har en statistisk model for et antal observationer $y_1, y_2, \ldots, y_n$, så kan man ud fra disse observationer estimere de indgående parametre; derefter kan man udregne de *fittede* værdier $\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n$, dvs. de y-værdier man skulle forvente hvis de sande parameterværdier var lig med de estimerede.

fordelingsfunktion (eller *kumuleret fordelingsfunktion*) for en sandsynlighedsfordeling: En funktion F således at $F(x)$ er sandsynligheden for at få værdier som er mindre end eller lig x.

forklarende variabel: se *baggrundsvariabel*.

forventet værdi er det samme som *middelværdi*.

fraktil (engelsk: *quantile*): En fraktil i en sandsynlighedsfordeling er et tal x med den egenskab at der er en vis foreskreven sandsynlighed for at få værdier $\leq x$. Eksempelvis er 90%-fraktilen et tal x således at der er sandsynlighed 90% for at få værdier $\leq x$.

Mere generelt er en α-fraktil i en fordeling med fordelingsfunktion F en løsning x_α, så godt den nu findes, til ligningen $F(x) = \alpha$. Hvis F er kontinuert og strengt voksende, så er der en entydig løsning $x_\alpha = F^{-1}(\alpha)$; generelt gælder at en α-fraktil er et tal x_α med den egenskab at $F(x_\alpha) \geq \alpha$ og $F(x_\alpha - \varepsilon) \leq \alpha$ for alle $\varepsilon > 0$.

fraktildiagram: en særlig slags tegning der i visse tilfælde kan benyttes for at kontrollere om et sæt observationer $y_1, y_2, \ldots, y_n$ stammer fra en bestemt type fordeling, for eksempel en normalfordeling eller en eksponentialfordeling.

Lad $y_{(1)}, y_{(2)}, \ldots, y_{(n)}$ være de ordnede observationer. Antag at y-erne stammer fra en fordeling med fordelingsfunktion af formen $F(y) = F_0(\frac{y-\mu}{\sigma})$ hvor F_0 er en kendt funktion. Fraktildiagrammet er da et diagram indeholdende punkterne $\left(y_{(i)}, F_0^{-1}(\frac{i-0.5}{n})\right)$, $i = 1, 2, \ldots, n$. Hvis modelantagelsen er rigtig, skal disse punkter ligge omkring den rette linie som går gennem punktet $(\mu, 0)$ og som har hældning $1/\sigma$.

frihedsgrader (engelsk: *degrees of freedom*): En del sandsynlighedsfordelinger har som attribut et »frihedsgradsantal« der bestemmer deres nærmere udseende; det gælder blandt andet χ^2-fordelingen og t-fordelingen, samt F-fordelingen der har hele to frihedsgradsantal. Derudover optræder »frihedsgrader« som en attribut for visse stikprøvefunktioner, som for eksempel variansskøn og teststørrelser (strengt taget er det stikprøvefunktionernes fordelinger der har disse attributter).

En forholdsvis generel tommelfingerregel for bestemmelse af antallet af frihedsgrader er *antal frit varierende størrelser før* minus *antal frit varierende størrelser efter*. Eksempler: I normalfordelingsmodeller har variansskønnet et antal frihedsgrader som er antal observationer minus antal estimerede frie parametre. Den asymptotiske χ^2-fordeling af $-2\ln Q$ har et antal frihedsgrader som er antal frie parametre i grundmodellen minus antal frie parametre i den testede model.

Gammafordeling: Gammafordelingen med formparameter $\lambda > 0$ og skalaparameter $\beta > 0$ er den kontinuerte sandsynlighedsfordeling på den positive halvakse som har tæthedsfunktion

$$x \mapsto \frac{1}{\Gamma(\lambda)\beta^\lambda}\, x^{\lambda-1}\exp(-x/\beta), \quad x > 0.$$

Middelværdien i gammafordelingen er $\lambda\beta$ og variansen er $\lambda\beta^2$.

χ^2-fordelingen med f frihedsgrader er det samme som gammafordelingen med formparameter $f/2$ og skalaparameter 2.

Gammafunktionen er defineret ved

$$\Gamma(x) = \int_0^{+\infty} t^{x-1}\exp(-t)\,dt, \quad x \in \mathbf{R}\setminus\{0,-1,-2,-3,\dots\}.$$

Gammafunktionen har den egenskab at $\Gamma(x+1) = x\,\Gamma(x)$ for alle x.

Hvis n er et naturligt tal, så er $\Gamma(n+1) = n!$

Der gælder at $\Gamma(1/2) = \sqrt{\pi}$.

gennemsnit (engelsk: *mean*): Gennemsnittet (eller mere præcist det aritmetiske gennemsnit) af tallene $y_1, y_2, \dots, y_n$ er tallet

$$\bar{y} = (y_1 + y_2 + \ldots + y_n)/n.$$

Det *vægtede* gennemsnit af tallene $y_1, y_2, \dots, y_n$ med vægtene $w_1, w_2, \dots, w_n$ er tallet

$$\frac{w_1 y_1 + w_2 y_2 + \ldots + w_n y_n}{w_1 + w_2 + \ldots + w_n}.$$

Det *geometriske* gennemsnit af tallene $y_1, y_2, \dots, y_n$ er tallet $\sqrt[n]{y_1 \cdot y_2 \cdot \ldots \cdot y_n}$.

histogram: Et histogram er en art empirisk tæthedsfunktion.

Et histogram tegnes på den måde at man inddeler observationsaksen i et antal delintervaller, og for hvert af disse intervaller tegnes et rektangel med intervallet som nederste side og med et areal som er lig med den relative hyppighed af observationer i det pågældende interval.

hypotese: En *statistisk hypotese* er et udsagn om at de ukendte parametre i den statistiske model opfylder visse betingelser (f.eks. at nogle af dem er ens, eller lig 0, etc.).

hyppighed (engelsk: *frequency*): Hyppigheden (eller mere præcist *den absolutte hyppighed*) af en hændelse A er det antal gange A indtræffer når man udfører et bestemt tilfældighedseksperiment et fastsat antal gange.

*Den relative hyppighed*z af A er den absolutte hyppighed divideret med antallet af gange tilfældighedseksperimentet er udført.

indikatorvariabel: En indikatorvariabel for en hændelse A er en stokastisk variabel der antager værdien 1 når A indtræffer og 0 ellers.

Indikatorvariable kaldes også 01-*variable*.

kontingenstabel: Antag at nogle individer klassificeres efter to kriterier hvoraf det første har r og det andet s niveauer. På den måde placeres hvert individ i én af de rs forskellige klasser. Den tabel med r rækker og s søjler som indeholder *antallene* af individer i hver af de rs klasser, er en tosidet kontingenstabel.

kontinuert stokastisk variabel: En stokastisk variabel X er kontinuert hvis den har en *tæthedsfunktion*, dvs. en ikke-negativ funktion f med den egenskab at for vilkårlige reelle tal $a \leq b$ er

$$\mathrm{P}(a < X \leq b) \;=\; \int_a^b f(x)\,dx.$$

Hvis Δx er et lille positivt tal, så er $f(x)\Delta x \;\approx\; \mathrm{P}(x < X \leq x + \Delta x)$.

Fordelingsfunktionen for en kontinuert stokastisk variabel er kontinuert i alle punkter, og hvis x er et kontinuitetspunkt for f, så er F differentiabel i x med $F'(x) = f(x) = \lim\limits_{h \downarrow 0} \dfrac{1}{h}\mathrm{P}(|X - x| < h)$.

korrelationskoefficient: Korrelationskoefficienten mellem to stokastiske variable X og Y er tallet

$$\rho \;=\; \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}X}\,\sqrt{\mathrm{Var}Y}}.$$

Der gælder at $-1 \leq \rho \leq 1$.

Den *empiriske* korrelationskoefficient mellem talsættene $x_1, x_2, \ldots, x_n$ og $y_1, y_2, \ldots, y_n$ er tallet

$$r \;=\; \frac{\sum\limits_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum\limits_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum\limits_{i=1}^{n}(y_i - \bar{y})^2}}.$$

Der gælder at $-1 \leq r \leq 1$.

kovariansen mellem to stokastiske variable X_1 og X_2 er tallet

$$\mathrm{Cov}(X_1, X_2) \quad = \quad \mathrm{E}(X_1 - \mathrm{E}\,X_1)(X_2 - \mathrm{E}\,X_2).$$

kvotientteststørrelse: Kvotientteststørrelsen Q for en bestemt hypotese er kvotienten mellem den maksimale likelihoodfunktion under hypotesen og den maksimale likelihoodfunktion under den aktuelle grundmodel.

likelihoodfunktion: Likelihoodfunktionen svarende til et bestemt sæt observationer fås ved at man i modelfunktionen indsætter dette sæt observationer og betragter udtrykket som en funktion af de ukendte parametre.

Eksempel: Hvis modelfunktionen er $f(y;\theta)$, så er likelihoodfunktionen hørende til y_{obs} funktionen $L(\theta) = f(y_{\mathrm{obs}};\theta)$.

likelihoodmetoden: se side 39.

logistisk transformation: Funktionen $\mathrm{logit}(p) = \ln \dfrac{p}{1-p}$ kaldes den logistiske transformation eller blot »logit«.

Funktionen afbilder intervallet $]0, 1[$ bijektivt på den reelle akse, og den omvendte funktion er $p = \exp(z)/(1 + \exp(z))$.

maksimaliseringsestimat: Den værdi der maksimaliserer likelihoodfunktionen (eller log-likelihoodfunktionen).

median: det samme som 50%-fraktilen, dvs. halvdelen af sandsynlighedsmassen ligger til venstre for medianen og halvdelen til højre for.

middelfejl (engelsk: *standard error*): Middelfejlen på en estimator er det samme som estimatorens standardafvigelse, s.d.

middelværdi (engelsk: *expected value*) af en stokastisk variabel X er tallet $\mathrm{E}X$ defineret ved henholdsvis

$$\mathrm{E}X \quad = \quad \sum_x x\,p(x) \qquad \text{eller}$$
$$\mathrm{E}X \quad = \quad \int_{-\infty}^{+\infty} x\,f(x)\,dx,$$

alt efter som X er diskret (med sandsynlighedsfunktion p) eller kontinuert (med tæthedsfunktion f). – Middelværdien af X kaldes også *den forventede værdi* af X.

Hvis g er en funktion defineret på værdimængden for X, så kan middelværdien af den stokastiske variabel $g(X)$ fås som henholdsvis

$$\mathrm{E}\,g(X) \quad = \quad \sum_x g(x)\,p(x) \qquad \text{eller}$$
$$\mathrm{E}\,g(X) \quad = \quad \int_{-\infty}^{+\infty} g(x)\,f(x)\,dx,$$

Regneregler for middelværdier:

1. Hvis $X \geq 0$, så er $\mathrm{E}\, X \geq 0$.
2. $\mathrm{E}\, a = a$.
3. $\mathrm{E}(X_1 + X_2) = \mathrm{E}\, X_1 + \mathrm{E}\, X_2$.
4. $\mathrm{E}(aX) = a\, \mathrm{E}\, X$.
5. Hvis X_1 og X_2 er *stokastisk uafhængige*, så er $\mathrm{E}(X_1 \cdot X_2) = \mathrm{E}\, X_1 \cdot \mathrm{E}\, X_2$.

Her betegner a en konstant og X, X_1 og X_2 stokastiske variable.

modelfunktion: Modelfunktionen er sandsynlighedsfunktionen opfattet som en funktion af såvel observationer som parametre. Modelfunktionen angiver sandsynligheden for at få et bestemt udfald når de ukendte parametre har en bestemt værdi.

multinomialfordeling: Multinomialfordelingen (med r klasser) med antalsparameter n og sandsynlighedsparametre $p_1, p_2, \ldots, p_r$ (hvor $p_1 + p_2 + \ldots + p_r = 1$) er den r-dimensionale sandsynlighedsfordeling som har sandsynlighedsfunktion

$$f(y_1, y_2, \ldots, y_r; p_1, p_2, \ldots, p_r) = \frac{n!}{y_1! \cdot y_2! \cdot \ldots \cdot y_r!} p_1^{y_1} \cdot p_2^{y_2} \cdot \ldots \cdot p_r^{y_r}$$

når $y_1, y_2, \ldots, y_r$ er ikke-negative heltal der summerer til n.

(Multinomialfordelingen med $r = 2$ er det samme som binomialfordelingen.)

negativ binomialfordeling: Den negative binomialfordeling med formparameter $\kappa > 0$ og sandsynlighedsparameter $p \in]0,1[$ er den sandsynlighedsfordeling på $\{0, 1, 2, \ldots\}$ som har sandsynlighedsfunktion

$$f(y; \kappa, p) = \binom{y + \kappa - 1}{y} p^{\kappa} (1-p)^y, \quad y \in \{0, 1, 2, \ldots\}.$$

Middelværdien i fordelingen er $\kappa(1-p)/p$, og variansen er $\kappa(1-p)/p^2$.

normalfordeling: Normalfordelingen med parametre μ og σ^2, kort $\mathcal{N}(\mu, \sigma^2)$, er den sandsynlighedsfordeling på den reelle talakse $\mathbf{R}$ som har tæthedsfunktionen

$$f(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\tfrac{1}{2} \frac{(x-\mu)^2}{\sigma^2} \right).$$

Her kan parameteren μ være et vilkårligt reelt tal og parameteren σ^2 et vilkårligt positivt tal.

Middelværdien i $\mathcal{N}(\mu, \sigma^2)$-fordelingen er μ, og variansen er σ^2.

ordnede observationer (engelsk: *order statistics*): Hvis $y_1, y_2, \ldots, y_n$ er et sæt observationer, så er $y_{(1)}, y_{(2)}, \ldots, y_{(n)}$ en traditionel standardbetegnelse for de tilsvarende ordnede observationer, dvs. y-erne ordnet i rækkefølge sådan at $y_{(1)}$ = det mindste af y-erne, $y_{(2)}$ = det næstmindste af y-erne, osv.

outlier: en observation der ligger usædvanlig langt fra det store flertal af tilsvarende observationer.

Pascals trekant: en rekursiv algoritme til udregning af binomialkoefficienter:

$$\binom{n}{k} = \binom{n-1}{k} + \binom{n-1}{k-1}$$

Poissonfordeling: Poissonfordelingen med parameter $\mu \geq 0$ er den sandsynlighedsfordeling på $\{0, 1, 2, \dots\}$ som har sandsynlighedsfunktion

$$f(y; \mu) = \frac{\mu^y}{y!} \exp(-\mu), \quad y \in \{0, 1, 2, \dots\}.$$

Middelværdien i fordelingen er μ, og variansen er også μ.

probit (‘*prob*ability un*it*’): Hvis man lægger 5 til α-fraktilen i $\mathcal{N}(0,1)$-fordelingen, får man $\text{probit}(\alpha)$, dvs.

$$\text{probit}(\alpha) = \Phi^{-1}(\alpha) + 5,$$

hvor Φ er den kumulerede fordelingsfunktion for $\mathcal{N}(0,1)$-fordelingen.

R^2**:** se *determinationskoefficient*

regressionsanalyse: En generel betegnelse for en lang række analysemetoder der går ud på at beskrive (middelværdien af) $y_1, y_2, \dots, y_n$ ved hjælp af en eller flere forklarende variable.

regressionskoefficient: I den simple lineære regressionsmodel $\mathrm{E}Y_i = \alpha + \beta x_i$ kaldes parameteren β for regressionskoefficienten.

residual: Residualerne er forskellene $y_i - \hat{y}_i$ mellem de observerede og de fittede værdier.

sample: et engelsk ord der kan oversættes til *stikprøve* (eller hvis det er et udsagnsord til *at indsamle en stikprøve*).

sandsynlighedspapir (normalfordelingspapir): En særlig type papir med fortrykte skalaer der letter tegningen af fraktildiagrammer for normalfordelte observationer.

Sandsynlighedspapirets ordinatakse har to skalaer, dels en probit-skala, dels den tilsvarende sandsynlighedsskala.

scatterplot: en diagram hvor man i al enkelhed indtegner punkter (x_i, y_i).

signifikans: Udtrykket benyttes i forbindelse med test af en statistisk hypotese. Man siger at der er *signifikans på niveau* α hvis testsandsynligheden er udregnet til α (og man forkaster hypotesen).

Hvis man har den strategi at forkaste hypotesen hver gang testsandsynligheden er mindre end α, så siger man at man benytter et *signifikansniveau* på α, og man kan fortolke α som sandsynligheden for at forkaste hypotesen selv om den er rigtig. – Man benytter ofte $\alpha = 5\%$.

simpel hypotese: en hypotese der ikke indeholder nogen ukendte parametre (det modsatte er en sammensat hypotese).

standardafvigelse (engelsk: *standard deviation*) er kvadratroden af variansen (s.d.).

statistisk hypotese: se *hypotese*

statistisk model: se side 39.

stikprøve (engelsk: *sample*): At et sæt observationer $y_1, y_2, \ldots, y_n$ er en *stikprøve* fra en bestemt fordeling, betyder at y-erne er observationer af uafhængige identisk fordelte stokastiske variable $Y_1, Y_2, \ldots, Y_n$ som følger den pågældende fordeling.

stikprøvefunktion (engelsk: *statistic*): en funktion af observationerne.

stokastisk variabel (engelsk: *random variable*): et symbol der repræsenterer det tilfældige udfald af et bestemt tilfældighedseksperiment, eller sagt på en anden måde: en værdi valgt tilfældigt fra en bestemt sandsynlighedsfordeling.

Det er ofte underforstået at stokastiske variable antager værdier på den reelle talakse **R** eller en delmængde heraf.

Stokastiske variable betegnes ofte med symboler som X, Y, Z, osv.

Der findes også en stringent matematisk definition af »stokastisk variabel«.

t**-fordeling:** t-fordelingen med f frihedsgrader er den kontinuerte sandsynlighedsfordeling på den reelle akse som har tæthedsfunktion

$$x \mapsto \frac{\Gamma(\frac{f+1}{2})}{\sqrt{\pi f}\,\Gamma(\frac{f}{2})} \left(1 + \frac{x^2}{f}\right)^{-\frac{f+1}{2}}, \quad x \in \mathbf{R}.$$

test: Et statistisk test er en metode hvormed man sammenholder et foreliggende sæt observationer og en påstået statistisk hypotese/model for disse observationer.

Ofte består testet i at man udregner en velvalgt *teststørrelse* (en funktion af observationerne) og sammenligner den observerede værdi af teststørrelsen med en vis referencefordeling. Hvis værdien af teststørrelsen ligger meget ekstremt i forhold til sin referencefordeling, så *forkaster* man hypotesen. *Testsandsynligheden* er sandsynligheden (udregnet efter referencefordelingen) for at få noget der er mere ekstremt end den observerede værdi af teststørrelsen; man forkaster hypotesen hvis testsandsynligheden er lille (f.eks. mindre end 5%).

Se også side 39.

testsandsynlighed: se *test*.

teststørrelse (engelsk: *test statistic*): se *test*.

tæthedsfunktion: se *kontinuert stokastisk variabel*.

uafhængighed: Stokastiske variable $X_1, X_2, \dots, X_n$ er *uafhængige* (eller mere præcist *stokastisk uafhængige*) hvis det er sådan at kendskab til udfaldene af nogle af de variable ikke påvirker de sandsynlighedsudsagn man kan fremsætte om de øvrige variable; en mere formel definition er at der skal gælde at

$$\begin{aligned} &\mathrm{P}(X_1 \in A_1 \text{ og } X_2 \in A_2 \text{ og } \dots \text{ og } X_n \in A_n) \\ &= \mathrm{P}(X_1 \in A_1) \cdot \mathrm{P}(X_2 \in A_2) \cdot \ldots \cdot \mathrm{P}(X_n \in A_n) \end{aligned}$$

for alle valg af hændelser $A_1, A_2, \dots, A_n$.

variansen på en stokastisk variabel X er tallet $\mathrm{Var} X = \mathrm{E}\left((X - \mathrm{E}X)^2\right)$.

Regneregler for varianser:

1. $\mathrm{Var}\, X \geq 0$.
2. $\mathrm{Var}\, a = 0$.
3. $\mathrm{Var}(aX) = a^2 \,\mathrm{Var}\, X$.
4. $\mathrm{Var}(a + X) = \mathrm{Var}\, X$.
5. $\mathrm{Var}\, X = \mathrm{E}\, X^2 - (\mathrm{E}\, X)^2$.
6. $\mathrm{Var}\, X = \mathrm{E}\, X(X-1) - (\mathrm{E}\, X)(\mathrm{E}\, X - 1)$.
7. $\mathrm{Var}(X_1 + X_2) = \mathrm{Var}\, X_1 + \mathrm{Var}\, X_2 + 2\,\mathrm{Cov}(X_1, X_2)$.
8. Hvis X_1 og X_2 er *stokastisk uafhængige*, så er $\mathrm{Var}(X_1 + X_2) = \mathrm{Var}\, X_1 + \mathrm{Var}\, X_2$.

Her betegner a en konstant og X, X_1 og X_2 stokastiske variable.

χ^2-fordeling: χ^2-fordelingen med f frihedsgrader ($f > 0$) er den kontinuerte fordeling på den positive halvakse som har tæthedsfunktion

$$x \mapsto \frac{1}{\Gamma(\frac{f}{2})\, 2^{\frac{f}{2}}} x^{\frac{f}{2}-1} \exp(-x/2), \quad x > 0.$$

χ^2-fordelingen er en gammafordeling med formparameter $f/2$ og skalaparameter 2.

Mange teststørrelser har den egenskab at deres fordeling (når den testede hypotese er rigtig) kan approksimeres med en χ^2-fordeling med et vist antal frihedsgrader. Frihedsgradsantallet er forskellen mellem det faktiske antal parametre i grundmodellen og det faktiske antal parametre under hypotesen.

χ^2/f**-fordeling:** χ^2/f-fordelingen med f frihedsgrader ($f > 0$) er den kontinuerte fordeling på den positive halvakse som har tæthedsfunktion

$$x \;\mapsto\; \frac{1}{\Gamma(\frac{f}{2})\,(\frac{2}{f})^{\frac{f}{2}}}\,x^{\frac{f}{2}-1}\exp(-x/\tfrac{2}{f}), \quad x > 0.$$

χ^2/f-fordelingen er en gammafordeling med formparameter $f/2$ og skalaparameter $2/f$.

Hvis X er χ^2-fordelt med f frihedsgrader, så følger X/f en χ^2/f-fordeling med f frihedsgrader.

3 Resuméer

3.1 Den simple binomialfordelingsmodel

Situation: n individer er klassificeret efter om de gav udfaldet 1 eller 0:

udfald	obs. antal
1	y
0	$n-y$
i alt	n

Model: y er en observation fra en binomialfordeling med parametre n og p, hvoraf p er ukendt.

Estimation: p estimeres ved $\hat{p} = y/n$. Middelværdien af estimatoren $\hat{p}$ er p. Standardafvigelsen på $\hat{p}$ estimeres til $\sqrt{\hat{p}(1-\hat{p})/n}$.

Hypotese: Man ønsker at teste den simple statistiske hypotese $H_0 : p = p_0$ hvor p_0 er et på forhånd givet tal.

Teststørrelse: Under H_0 er situationen

udfald	obs. antal	»forventet« antal
1	y	$\hat{y} = np_0$
0	$n-y$	$n-\hat{y} = n(1-p_0)$
i alt	n	n

Kvotientteststørrelsen er $-2\ln Q \;=\; 2\left(y\ln\frac{y}{\hat{y}} + (n-y)\ln\frac{n-y}{n-\hat{y}}\right)$.

Testsandsynlighed: Testsandsynligheden ε bestemmes således:

1. Hvis begge de »forventede« antal er mindst 5, kan ε med god tilnærmelse findes som sandsynligheden for at få en værdi større end $-2\ln Q_{\text{obs}}$ i χ^2-fordelingen med 1 frihedsgrad:
$$\varepsilon \;=\; \mathrm{P}\left(\chi^2_1 \geq -2\ln Q_{\text{obs}}\right).$$

2. I modsat fald må man udregne den eksakte testsandsynlighed
$$\varepsilon \;=\; \sum_{y\,:\,-2\ln Q(y)\geq -2\ln Q_{\text{obs}}} \binom{n}{y} p_0^y (1-p_0)^{n-y}.$$

Konklusion: Hvis ε er meget lille, så er der en signifikant afvigelse mellem det observerede og det som H_0 foreskriver: man må da forkaste H_0.

Hvis ε ikke er meget lille, er H_0 forenelig med det observerede: man kan ikke forkaste H_0.

3.2 Sammenligning af binomialfordelinger

Situation: n_j individer fra gruppe j er klassificeret efter om de gav udfaldet 1 eller 0, $j = 1, 2, \dots, s$:

	gruppe nr.				
udfald	1	2	...	s	sum
1	y_1	y_2	...	y_s	$y_\bullet$
0	$n_1 - y_1$	$n_2 - y_2$	...	$n_s - y_s$	$n_\bullet - y_\bullet$
i alt	n_1	n_2	...	n_s	$n_\bullet$

Model: $y_1, y_2, \dots, y_s$ er uafhængige observationer fra binomialfordelinger således at y_j stammer fra en binomialfordeling med parametre n_j og p_j. De ukendte parametre er $p_1, p_2, \dots, p_s$.

Estimation: p_j estimeres ved $\hat{p}_j = y_j / n_j$, $j = 1, 2, \dots, s$.

Hypotese: Man ønsker at teste den statistiske hypotese

$$H_0 : p_1 = p_2 = \dots = p_s \ \ (= p)$$

om at der ikke er forskel på de s grupper.

Teststørrelse: Under H_0 er den »forventede« situation

	gruppe nr.				
udfald	1	2	...	s	sum
1	$\hat{y}_1$	$\hat{y}_2$	...	$\hat{y}_s$	$y_\bullet$
0	$n_1 - \hat{y}_1$	$n_2 - \hat{y}_2$	...	$n_s - \hat{y}_s$	$n_\bullet - y_\bullet$
i alt	n_1	n_2	...	n_s	$n_\bullet$

hvor $\hat{y}_j = n_j \hat{p}$ og $\hat{p} = y_\bullet / n_\bullet$.

Kvotientteststørrelsen er

$$-2 \ln Q \quad = \quad 2 \sum_{j=1}^{s} \left(y_j \ln \frac{y_j}{\hat{y}_j} + (n_j - y_j) \ln \frac{n_j - y_j}{n_j - \hat{y}_j} \right) .$$

Testsandsynlighed:

1. Hvis alle de $2s$ »forventede« antal $\hat{y}_j$ og $n_j - \hat{y}_j$ er mindst 5, kan testsandsynligheden ε med god tilnærmelse findes som sandsynligheden for at få en værdi større end $-2 \ln Q_{\text{obs}}$ i χ^2-fordelingen med $s - 1$ frihedsgrader:

$$\varepsilon \quad = \quad \mathrm{P}\left(\chi^2_{s-1} \geq -2 \ln Q_{\text{obs}} \right) .$$

2. I modsat fald må man prøve at slå nogle smågrupper sammen, således at 1. bliver opfyldt. Eller hvis $s = 2$ kan man benytte det eksakte test.

Det eksakte test når $s = 2$: Udregn størrelsen $g(z) = \binom{n_1}{z}\binom{n_2}{y_\bullet - z}$ for $z = 0, 1, 2, \ldots, y_\bullet$, og udregn summen $S = g(0) + g(1) + \ldots + g(y_\bullet)$ (der i øvrigt er lig $\binom{n_1+n_2}{y_\bullet}$). Testsandsynligheden er da

$$\varepsilon = \frac{1}{S} \sum_{z\,:\,g(z) \leq g(y_1)} g(z)\,.$$

Konklusion: Hvis ε er meget lille, så er der en signifikant afvigelse mellem det observerede og det som H_0 foreskriver, og man vil da forkaste H_0. I modsat fald er H_0 forenelig med det observerede, og man kan ikke forkaste H_0.

3.3 Betingelser for en Poissonmodel

1. Det der observeres, er *antallet* af en bestemt slags begivenheder i et bestemt tidsinterval.
2. Der indtræffer aldrig to (eller flere) begivenheder samtidigt.
3. Begivenhederne indtræffer uafhængigt af hinanden således at forstå at det der indtræffer i for eksempel tidsintervallet $]a, b]$, er stokastisk uafhængigt af hvad der sker uden for dette interval.
4. Sandsynligheden for at der indtræffer en begivenhed i et tidsinterval $]t, t + \Delta t]$ af længde Δt, afhænger *ikke* af hvor på tidsaksen intervallet er beliggende, dvs. den afhænger ikke af t (sålænge vi befinder os inden for det overordnede tidsinterval der i det hele taget er tale om).
5. Hvis $p(\Delta t)$ betegner sandsynligheden for at der indtræffer mindst en begivenhed i et tidsinterval af længde Δt, så vil

$$\lim_{\Delta t \to 0} \frac{p(\Delta t)}{\Delta t} = \lambda,$$

hvor λ er en positiv konstant kaldet *intensiteten*.

Under disse omstændigheder vil antallet af begivenheder i tidsintervallet $]a, b]$ være Poissonfordelt med parameter $\mu = \lambda\,(b - a)$.

3.4 Sammenligning af Poissonfordelinger

Situation: Man har optalt antal begivenheder af en bestemt slags i k forskellige grupper. I gruppe i er

$$\begin{aligned} y_i &= \text{observeret antal begivenheder} \\ t_i &= \text{observations-»perioden«} \\ r_i &= \text{størrelsen af »risikogruppen«} \end{aligned}$$

Model: Observationerne $y_1, y_2, \ldots, y_k$ er observerede værdier af uafhængige Poissonfordelte stokastiske variable $Y_1, Y_2, \ldots, Y_k$ således at Y_i har parameter $\lambda_i r_i t_i$ hvor λ_i er en ukendt parameter der beskriver intensiteten af begivenhederne pr. tid og pr. individ for gruppe i.

Estimation: Parameteren λ_i estimeres ved antal begivenheder divideret med den samlede »gennemlevede« tid i gruppe i: $\hat{\lambda}_i = \dfrac{y_i}{r_i t_i}$.

Hypotese: Man ønsker at teste hypotesen $H_0 : \lambda_1 = \lambda_2 = \ldots = \lambda_k$ om at intensiteterne er ens i de k grupper.

Teststørrelse: Den eventuelle fælles værdi af λ estimeres ved $\hat{\lambda} = y_{\bullet} \Big/ \sum_{l=1}^{k} r_l t_l$, og det »forventede« antal begivenheder i gruppe i under H_0 er $\hat{y}_i = \hat{\lambda}\, r_i t_i = \dfrac{r_i t_i}{\sum_{l=1}^{k} r_l t_l}\, y_{\bullet}$.

Kvotientteststørrelsen er $-2 \ln Q = 2 \sum_{i=1}^{k} y_i \ln \dfrac{y_i}{\hat{y}_i}$.

Testsandsynlighed: Hvis alle de »forventede« antal $\hat{y}_i$ er mindst fem, kan testsandsynligheden ε med god tilnærmelse bestemmes som sandsynligheden for at få en større værdi end $-2 \ln Q_{\text{obs}}$ i χ^2-fordelingen med $f = k - 1$ frihedsgrader:

$$\varepsilon = \mathrm{P}\left(\chi^2_{k-1} \geq -2 \ln Q_{\text{obs}}\right).$$

Konklusion: Hvis ε er meget lille, så er der en signifikant afvigelse mellem det observerede og det som modellen foreskriver, og man vil da forkaste H_0. I modsat fald er H_0 forenelig med det observerede, og man kan da ikke forkaste H_0.

3.5 Sammenligning af multinomialfordelinger

Situation: n_j individer fra gruppe j er klassificeret i r klasser:

		gruppe nr.				sum
	A_1	y_{11}	y_{12}	$\cdots$	y_{1s}	$y_{1\bullet}$
	A_2	y_{21}	y_{22}	$\cdots$	y_{2s}	$y_{2\bullet}$
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
klasse	A_i	y_{i1}	y_{i2}	$\cdots$	y_{is}	$y_{i\bullet}$
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
	A_r	y_{r1}	y_{r2}	$\cdots$	y_{rs}	$y_{r\bullet}$
	i alt	n_1	n_2	$\ldots$	n_s	$n_{\bullet}$

Model: De enkelte søjler af y-er er uafhængige observationer fra multinomialfordelinger således at multinomialfordelingen hørende til søjle nr. j har antalsparameter n_j og sandsynlighedsparametre $p_{1j}, p_{2j}, \dots, p_{rj}$. Her er p_{ij}-erne ukendte parametre med $p_{1j} + p_{2j} + \dots + p_{rj} = 1$ for alle j.

Estimation: p_{ij} estimeres ved den observerede relative hyppighed af A_i i gruppe j, dvs. $\hat{p}_{ij} \;=\; \dfrac{y_{ij}}{n_j}$.

Hypotese: Man ønsker at teste hypotesen

$$H_0 : \begin{pmatrix} p_{11} \\ p_{21} \\ \vdots \\ p_{r1} \end{pmatrix} = \begin{pmatrix} p_{12} \\ p_{22} \\ \vdots \\ p_{r2} \end{pmatrix} = \dots = \begin{pmatrix} p_{1s} \\ p_{2s} \\ \vdots \\ p_{rs} \end{pmatrix}$$

om at der ikke er forskel på søjlerne.

Teststørrelse: Under H_0 er den »forventede« situation

		gruppe nr.				sum
	A_1	$\hat{y}_{11}$	$\hat{y}_{12}$	$\dots$	$\hat{y}_{1s}$	$y_{1\bullet}$
	A_2	$\hat{y}_{21}$	$\hat{y}_{22}$	$\dots$	$\hat{y}_{2s}$	$y_{2\bullet}$
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
klasse	A_i	$\hat{y}_{i1}$	$\hat{y}_{i2}$	$\dots$	$\hat{y}_{is}$	$y_{i\bullet}$
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
	A_r	$\hat{y}_{r1}$	$\hat{y}_{r2}$	$\dots$	$\hat{y}_{rs}$	$y_{r\bullet}$
	i alt	n_1	n_2	$\dots$	n_s	$n_\bullet$

hvor $\hat{y}_{ij} = y_{i\bullet} n_j / n_{\bullet\bullet}$.

Kvotientteststørrelsen er

$$-2 \ln Q \;=\; 2 \sum_{j=1}^{s} \sum_{i=1}^{r} y_{ij} \ln \frac{y_{ij}}{\hat{y}_{ij}} \;.$$

Testsandsynlighed:

1. Hvis alle de rs »forventede« antal er mindst fem, kan testsandsynligheden ε med god tilnærmelse findes som sandsynligheden for at få en større værdi end $-2 \ln Q_{\mathrm{obs}}$ i χ^2-fordelingen med $(r-1)(s-1)$ frihedsgrader:

$$\varepsilon \;=\; \mathrm{P}\Big(\chi^2_{(r-1)(s-1)} \geq -2 \ln Q_{\mathrm{obs}}\Big) \,.$$

2. I modsat fald må man prøve at slå nogle smågrupper sammen eller at slå nogle klasser sammen el.lgn. for at opnå at 1 bliver opfyldt.

Konklusion: Hvis ε er meget lille, så er der en signifikant afvigelse mellem det observerede og det som H_0 foreskriver, og man vil da forkaste H_0. I modsat fald er H_0 forenelig med det observerede, og man kan ikke forkaste H_0.

3.6 Uafhængighedstest i en $r \times s$-tabel

Situation: n individer er klassificeret efter to kriterier:

		kriterium 2				
	klasse	B_1	B_2	$\dots$	B_s	sum
	A_1	y_{11}	y_{12}	$\cdots$	y_{1s}	$y_{1\bullet}$
kriterium 1	A_2	y_{21}	y_{22}	$\cdots$	y_{2s}	$y_{2\bullet}$
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
	A_r	y_{r1}	y_{r2}	$\cdots$	y_{rs}	$y_{r\bullet}$
	sum	$y_{\bullet 1}$	$y_{\bullet 2}$	$\cdots$	$y_{\bullet s}$	n

hvor

$$y_{ij} \quad = \quad \text{antal individer i klassen } A_i B_j \; (= A_i \cap B_j).$$

Model: De rs værdier y_{ij} udgør tilsammen en observation fra en multinomialfordeling med rs klasser, med antalsparameter n og med sandsynlighedsparametre p_{ij}, $i = 1, 2, \dots, r$, $j = 1, 2, \dots, s$.

Sandsynlighedsparametrene p_{ij} er ukendte parametre der summerer til 1: $p_{11} + p_{12} + \dots + p_{rs} = 1$.

Estimation: p_{ij} estimeres ved den observerede relative hyppighed af $A_i B_j$, dvs.

$$\hat{p}_{ij} \quad = \quad \frac{y_{ij}}{n}.$$

Hypotese: Man ønsker at teste uafhængighedshypotesen (hypotesen om forsvindende vekselvirkning)

$$H_0: \; p_{ij} = \alpha_i \, \beta_j \quad \text{for alle } i \text{ og } j,$$

hvor de ukendte parametre $(\alpha_1, \alpha_2, \dots, \alpha_r)$ og $(\beta_1, \beta_2, \dots, \beta_s)$ er ikke-negative talsæt der hver især summerer til 1.

Teststørrelse: Under H_0 er den »forventede« situation

		kriterium 2				
	klasse	B_1	B_2	$\dots$	B_s	sum
	A_1	$\hat{y}_{11}$	$\hat{y}_{12}$	$\cdots$	$\hat{y}_{1s}$	$y_{1\bullet}$
kriterium 1	A_2	$\hat{y}_{21}$	$\hat{y}_{22}$	$\cdots$	$\hat{y}_{2s}$	$y_{2\bullet}$
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
	A_r	$\hat{y}_{r1}$	$\hat{y}_{r2}$	$\cdots$	$\hat{y}_{rs}$	$y_{r\bullet}$
	sum	$y_{\bullet 1}$	$y_{\bullet 2}$	$\cdots$	$y_{\bullet s}$	n

hvor $\hat{y}_{ij} = y_{i\bullet} y_{\bullet j} / y_{\bullet\bullet}$.

Kvotientteststørrelsen er

$$-2 \ln Q \quad = \quad 2 \sum_{j=1}^{s} \sum_{i=1}^{r} y_{ij} \ln \frac{y_{ij}}{\hat{y}_{ij}} \, .$$

Testsandsynlighed:

1. Hvis alle de rs »forventede« antal er mindst fem, kan testsandsynligheden ε med god tilnærmelse findes som sandsynligheden for at få en større værdi end $-2\ln Q_{\text{obs}}$ i χ^2-fordelingen med $(r-1)(s-1)$ frihedsgrader:

$$\varepsilon \quad = \quad \mathrm{P}\Big(\chi^2_{(r-1)(s-1)} \geq -2\ln Q_{\text{obs}}\Big)\,.$$

2. I modsat fald må man prøve at slå nogle rækker eller søjler sammen for at opnå at 1. bliver opfyldt.

Konklusion: Hvis ε er meget lille, så er der en signifikant afvigelse mellem det observerede og det som H_0 foreskriver, og man vil da forkaste H_0. I modsat fald er H_0 forenelig med det observerede, og man kan ikke forkaste H_0.

3.7 Enstikprøveproblemet i normalfordelingen

Situation: Der foreligger observationer $y_1, y_2, \ldots, y_n$ på en kontinuert måleskala.

Model: Det antages at $y_1, y_2, \ldots, y_n$ er uafhængige observationer fra normalfordelingen $\mathcal{N}(\mu, \sigma^2)$ hvor μ og σ^2 er ukendte parametre.

Estimation: Middelværdiparameteren μ estimeres ved gennemsnittet $\bar{y} = (y_1 + y_2 + \ldots + y_n)/n$. Variansparameteren σ^2 estimeres ved summen af de kvadratiske afvigelser divideret med antallet af frihedsgrader, dvs. ved

$$s^2 \quad = \quad \frac{1}{n-1}\sum_{j=1}^{n}(y_j - \bar{y})^2$$

med $f = n - 1$ frihedsgrader.

Modelkontrol: Man kan tegne et *histogram* (samt den fittede normalfordelingstæthed) og/eller et *fraktildiagram* (samt den rette linie svarende til den fittede normalfordeling).

Hypotese: Man ønsker at teste den simple statistiske hypotese $H_0 : \mu = \mu_0$ hvor μ_0 er et på forhånd givet tal.

Teststørrelse: Kvotienttestet kan omformes til t-testet som er baseret på t-teststørrelsen

$$t \quad = \quad \frac{\bar{y} - \mu_0}{\sqrt{s^2/n}}\,.$$

Testsandsynlighed: Testsandsynligheden er sandsynligheden for at få en større $|t|$-værdi i t-fordelingen med $f = n - 1$ frihedsgrader, altså

$$\varepsilon \;=\; \mathrm{P}\Big(|t_f| > |t_{\text{obs}}|\Big) \;=\; 2 \cdot \mathrm{P}\Big(t_f > |t_{\text{obs}}|\Big)\,.$$

Denne sandsynlighed bestemmes let ved brug af tabeller over fraktiler i t-fordelingen.

3.8 Tostikprøveproblemet i normalfordelingen, uparrede observationer

Situation: Der foreligger nogle observationer y som er målt på en kontinuert måleskala, og som er inddelt i to grupper:

gruppe	observationer				
1	y_{11}	y_{12}	$\cdots$	y_{1j}	$\cdots$	y_{1n_1}
2	y_{21}	y_{22}	$\cdots$	y_{2j}	$\cdots$	y_{2n_2}

Her betegner y_{ij} observation nr. j i gruppe nr. i, $i = 1, 2$. Grupperne har henholdsvis n_1 og n_2 observationer.

Model: Det antages at y_{ij}-erne er observerede værdier af uafhængige stokastiske variable Y_{ij} som er normalfordelte med samme varians σ^2 og med middelværdier henholdsvis μ_1 og μ_2, kort

$$\begin{aligned} Y_{1j} &\sim \mathcal{N}(\mu_1, \sigma^2) \\ Y_{2j} &\sim \mathcal{N}(\mu_2, \sigma^2)\,. \end{aligned}$$

Her beskriver de to middelværdiparametre μ_1 og μ_2 den systematiske variation, dvs. de to gruppers niveauer, medens variansparameteren σ^2 (samt normalfordelingen) beskriver den tilfældige variation.

Estimation: Middelværdiparametrene μ_1 og μ_2 estimeres ved de tilsvarende gruppegennemsnit $\bar{y}_1$ og $\bar{y}_2$.

Variansparameteren σ^2 estimeres ved

$$s_0^2 \;=\; \frac{1}{n_1 + n_2 - 2}\left(\sum_{j=1}^{n_1}(y_{1j} - \bar{y}_1)^2 + \sum_{j=1}^{n_2}(y_{2j} - \bar{y}_2)^2\right),$$

med $n - 2 \;=\; n_1 + n_2 - 2$ frihedsgrader.

Modelkontrol: Hvis der er tilstrækkeligt mange observationer i grupperne kan man for hver gruppe tegne et histogram (samt den fittede normalfordelingstæthed) og/eller et fraktildiagram (samt den rette linie svarende til den fittede normalfordeling).

Antagelsen om varianshomogenitet kan eventuelt testes med Bartletts test for varianshomogenitet (tilfældet $k = 2$).

Hypotese: Man ønsker at teste den statistiske hypotese

$$H_0 \ : \ \mu_1 = \mu_2$$

om at de to grupper har samme middelværdiparameter.

Teststørrelse: Som teststørrelse benyttes

$$t \quad = \quad \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{s_0^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}.$$

Store værdier af $|t|$ er signifikante.

Testsandsynlighed: Testsandsynligheden ε bestemmes som sandsynligheden for at få en værdi uden for intervallet med endepunkter $-t_{\text{obs}}$ og t_{obs} i t-fordelingen med $n-2$ frihedsgrader,

$$\varepsilon \quad = \quad 2 \cdot \mathrm{P}\big(t_{n-2} > |t_{\text{obs}}|\big) \, .$$

Konklusion: Hvis ε er meget lille, så er der en signifikant forskel mellem de to grupper, dvs. hypotesen forkastes. Hvis ε ikke er meget lille, kan man ikke på det foreliggende grundlag forkaste hypotesen.

3.9 Tostikprøveproblemet i normalfordelingen, parrede observationer

Situation: Der foreligger r par af uafhængige observationer y som er målt på en kontinuert måleskala. Parrene er inddelt i to grupper:

	gruppe nr. 1	gruppe nr. 2
par nr. 1	y_{11}	y_{12}
par nr. 2	y_{21}	y_{22}
$\vdots$	$\vdots$	$\vdots$
par nr. i	y_{i1}	y_{i2}
$\vdots$	$\vdots$	$\vdots$
par nr. r	y_{r1}	y_{r2}

Model: Det antages at y_{ij}-erne er observerede værdier af stokastiske variable Y_{ij} med den egenskab at differenserne $D_i = Y_{i2} - Y_{i1}$ er uafhængige identisk normalfordelte:

$$Y_{i2} - Y_{i1} \quad \sim \quad \mathcal{N}(\delta, \sigma^2),$$

hvor δ og σ^2 er ukendte parametre.

Estimation: Kald differensen mellem elementerne i par nr. i for d_i, altså

$$d_i = y_{i2} - y_{i1} .$$

Parameteren δ estimeres ved gennemsnittet af differenserne eller differensen af gennemsnittet:

$$\begin{aligned} \hat{\delta} &= \bar{y}_2 - \bar{y}_1 \\ &= \bar{d} . \end{aligned}$$

Variansparameteren σ^2 estimeres ved

$$s^2 = \frac{1}{r-1} \sum_{i=1}^{r} \left(d_i - \bar{d}\right)^2$$

med $r - 1$ frihedsgrader.

Modelkontrol: For at checke om differenserne er normalfordelte kan man (hvis r er tilstrækkeligt stor) over differenserne $d_1, d_2, \ldots, d_r$ tegne et *histogram* (samt den fittede normalfordelingstæthed) og/eller et *fraktildiagram* (samt den rette linie svarende til den fittede normalfordeling).

Hypotese: Man ønsker at teste den simple statistiske hypotese

$$H_0 : \delta = 0$$

om at der ikke er nogen signifikant forskel mellem de to elementer i et par.

Teststørrelse: Som teststørrelse benyttes

$$t = \frac{\bar{d}}{\sqrt{s^2/r}} .$$

Store værdier af $|t|$ er signifikante.

Testsandsynlighed: Testsandsynligheden ε bestemmes som sandsynligheden for at få en værdi uden for intervallet med endepunkter $-t_{\text{obs}}$ og t_{obs} i t-fordelingen med $r - 1$ frihedsgrader,

$$\varepsilon = 2 \cdot \mathrm{P}\left(t_{r-1} > |t_{\text{obs}}|\right) .$$

Konklusion: Hvis ε er meget lille, så er der en signifikant forskel mellem de to grupper, dvs. hypotesen forkastes. Hvis ε ikke er meget lille, kan man ikke på det foreliggende grundlag forkaste hypotesen.

3.10 Ensidet variansanalyse

Situation: Der foreligger nogle observationer y som er målt på en kontinuert måleskala, og som er inddelt i k grupper således at der er n_i observationer i gruppe nr. i, $i = 1, 2, \ldots, k$. Observation nr. j fra gruppe nr. i betegnes y_{ij}. Skematisk ser det således ud:

gruppe	observationer					
1	y_{11}	y_{12}	$\cdots$	y_{1j}	$\cdots$	y_{1n_1}
2	y_{21}	y_{22}	$\cdots$	y_{2j}	$\cdots$	y_{2n_2}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
i	y_{i1}	y_{i2}	$\cdots$	y_{ij}	$\cdots$	y_{in_i}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
k	y_{k1}	y_{k2}	$\cdots$	y_{kj}	$\cdots$	y_{kn_k}

Model: Det antages at y_{ij}-erne er observerede værdier af uafhængige stokastiske variable Y_{ij} således at Y_{ij} er normalfordelt med middelværdi μ_i og varians σ^2, kort

$$Y_{ij} \quad \sim \quad \mathcal{N}(\mu_i, \sigma^2),$$

hvor $\mu_1, \mu_2, \ldots, \mu_k$ og σ^2 er ukendte parametre. Herved beskriver middelværdiparametrene $\mu_1, \mu_2, \ldots, \mu_k$ den systematiske variation, nemlig de enkelte gruppers niveauer, medens variansparameteren σ^2 (samt normalfordelingen) beskriver den tilfældige variation inden for grupperne; den tilfældige variation antages at være den samme i alle grupper.

Estimation: Middelværdiparametrene $\mu_1, \mu_2, \ldots, \mu_k$ estimeres ved gruppegennemsnittene:

$$\begin{aligned} \hat{\mu}_i &= \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij} \\ &= \bar{y}_i \,. \end{aligned}$$

Variansparameteren σ^2 estimeres ved residualkvadratsummen divideret med dens frihedsgradsantal, nemlig

$$s_0^2 \quad = \quad \frac{1}{n-k} \sum_{i=1}^{k} \sum_{j=1}^{n} (y_{ij} - \bar{y}_i)^2$$

hvor $f = n - k$ er antallet af frihedsgrader ($n = n_1 + n_2 + \ldots + n_k$).

Modelkontrol: Hvis der er tilstrækkeligt mange observationer i grupperne, kan man for hver gruppe tegne et histogram (samt den fittede normalfordelingstæthed) og/eller et fraktildiagram (samt den rette linie svarende til den fittede normalfordeling).

Antagelsen om varianshomogenitet kan eventuelt testes med *Bartletts test for varianshomogenitet*.

Hypotese: Man ønsker at teste hypotesen om homogenitet mellem grupper, dvs. den statistiske hypotese

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

om at der ikke er nogen signifikant forskel mellem grupperne.

Teststørrelse: Udregn størrelsen

$$s_1^2 = \frac{1}{k-1} \sum_{i=1}^{k} n_i (\bar{y}_i - \bar{y})^2$$

der beskriver variationen mellem grupper og har $k-1$ frihedsgrader. Som teststørrelse benyttes F-teststørrelsen

$$F = \frac{s_1^2}{s_0^2},$$

dvs. forholdet mellem variationen mellem grupper og variationen inden for grupper. – Store F-værdier er signifikante.

Testsandsynlighed: Testsandsynligheden er sandsynligheden for at få en større F-værdi i F-fordelingen med frihedsgradsantal $k-1$ og $n-k$, altså

$$\varepsilon = \mathrm{P}\big(F_{k-1,\,n-k} > F_{\mathrm{obs}}\big),$$

der let bestemmes ved brug af tabeller over fraktiler i F-fordelingen.

3.11 Bartletts test for varianshomogenitet

Situation: Der foreligger k grupper af normalfordelte observationer således at observationer i samme gruppe stammer fra samme normalfordeling. I hver gruppe er udregnet et variansestimat; estimatet fra gruppe i betegnes s_i^2 og har f_i frihedsgrader.

Hypotese: Man ønsker at teste den statistiske hypotese om at de k grupper har samme varians.

Estimation: Hvis hypotesen er rigtig, så skal den fælles varians estimeres som et vægtet gennemsnit af de enkelte gruppers varianser, nemlig ved $s_0^2 = \dfrac{1}{f} \sum_{i=1}^{k} f_i s_i^2$ der har $f = f_1 + f_2 + \ldots + f_k$ frihedsgrader.

Teststørrelse: Som teststørrelse benyttes $B = -\sum_{i=1}^{k} f_i \ln \dfrac{s_i^2}{s_0^2}$. Store værdier af B er signifikante.

Testsandsynlighed: Når $f_1, f_2, \ldots, f_k$ alle er mindst 5, så er B med god tilnærmelse χ^2-fordelt med $k-1$ frihedsgrader hvis hypotesen er rigtig, og testsandsynligheden kan da findes som

$$\varepsilon = \mathrm{P}\Big(\chi^2_{k-1} \geq B_{\mathrm{obs}}\Big).$$

Specialtilfældet $k = 2$: Når der kun er to variansestimater s_1^2 og s_2^2 der skal sammenlignes, kan det ske ved at man udregner kvotienten

$$R^\star = \frac{\max(s_1^2, s_2^2)}{\min(s_1^2, s_2^2)}.$$

$R^\star$-værdier langt fra 1 er signifikante, og man plejer at bestemme testsandsynligheden som

$$\varepsilon = \mathrm{P}\left(F_{f_1,f_2} \geq R^\star_{\text{obs}}\right) + \mathrm{P}\left(F_{f_2,f_1} \geq R^\star_{\text{obs}}\right).$$

3.12 Simpel lineær regressionsanalyse

Situation: Der foreligger n sammenhørende par (x_i, y_i) af en observation y og en baggrundsvariabel x; skematisk ser det sådan ud:

baggrundsvariabel	observation
x_1	y_1
x_2	y_2
$\vdots$	$\vdots$
x_n	y_n

Model: Det antages at x_i-erne er givne konstanter og at y_i-erne er observerede værdier af uafhængige stokastiske variable Y_i således at Y_i er normalfordelt med middelværdi $\alpha + \beta x_i$ og varians σ^2, kort

$$Y_i \sim \mathcal{N}(\alpha + \beta x_i, \sigma^2),$$

hvor α, β og σ^2 er ukendte parametre. Herved beskriver parametrene α og β den systematiske variation, nemlig y-s lineære afhængighed af x, og variansparameteren σ^2 (samt normalfordelingen) beskriver den tilfældige variation omkring regressionslinien; den tilfældige variation antages at være den samme for alle x.

Estimation: Udregn hjælpestørrelserne

$$\begin{aligned}
\bar{x} &= \frac{1}{n}\sum_{i=1}^n x_i\,, \\
\bar{y} &= \frac{1}{n}\sum_{i=1}^n y_i\,, \\
SP_{xy} &= \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})\,, \\
SS_x &= \sum_{i=1}^n (x_i - \bar{x})^2\,, \\
SS_y &= \sum_{i=1}^n (y_i - \bar{y})^2\,.
\end{aligned}$$

Så estimeres α og β ved

$$\begin{aligned}\hat{\beta} &= \frac{SP_{xy}}{SS_x}\,,\\ \hat{\alpha} &= \bar{y} - \hat{\beta}\bar{x}\,.\end{aligned}$$

Variansparameteren σ^2 estimeres ved

$$\begin{aligned}s_{02}^2 &= \frac{1}{n-2}\sum_{i=1}^{n}\big(y_i - (\hat{\alpha} + \hat{\beta}x_i)\big)^2\\ &= \frac{1}{n-2}\left(SS_y - \frac{SP_{xy}^2}{SS_x}\right)\end{aligned}$$

der har $n-2$ frihedsgrader.

Modelkontrol: Man kan lave en tegning hvor man i et koordinatsystem afsætter sammenhørende værdier af x og y og desuden indtegner den estimerede regressionslinie. Punkterne skal da fordele sig nogenlunde tilfældigt omkring linien.

Hvis der til hver x-værdi er flere y-observationer, kan man foretage et numerisk test for antagelsen om linearitet, se Resumé 3.13.

Hypoteser:

1. Hypotesen $H_3 : \beta = 0$ om at y ikke afhænger signifikant af x, testes med t-teststørrelsen

$$t = \frac{\hat{\beta}}{\sqrt{s_{02}^2/SS_x}}\,.$$

Store værdier af $|t|$ er signifikante.

2. Hypotesen $H_4 : \alpha = 0$ om at regressionslinien går gennem $(0,0)$, testes med t-teststørrelsen

$$t = \frac{\hat{\alpha}}{\sqrt{s_{02}^2\left(\frac{1}{n} + \frac{\bar{x}^2}{SS_x}\right)}}\,.$$

Store værdier at $|t|$ er signifikante.

I begge tilfælde bestemmes *testsandsynligheden* ε som sandsynligheden for at få en værdi uden for intervallet med endepunkter $-t_{\text{obs}}$ og t_{obs} i t-fordelingen med $n-2$ frihedsgrader,

$$\varepsilon = 2\,\mathrm{P}\big(t_{n-2} > |t_{\text{obs}}|\big)\,.$$

Hvis ε er meget lille, så er *konklusionen* at den pågældende parameter er signifikant forskellig fra 0, dvs. hypotesen forkastes. Hvis ε ikke er meget lille, kan man ikke på det foreliggende grundlag forkaste den pågældende hypotese.

3.13 Test for linearitet i den lineære regressionsmodel

Dette resumé skal læses i sammenhæng med Resumé 3.12 og handler om det tilfælde hvor der er flere y-værdier til hvert x.

Situation: Til hver af k forskellige værdier af baggrundsvariablen x foreligger et antal observationer y som det fremgår af nedenstående skema:

baggrundsvariabel	observationer			
x_1	y_{11}	y_{12}	$\cdots$	y_{1n_1}
x_2	y_{21}	y_{22}	$\cdots$	y_{2n_2}
x_3	y_{31}	y_{32}	$\cdots$	y_{3n_3}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
x_k	y_{k1}	y_{k2}	$\cdots$	y_{kn_k}

Det totale antal y-observationer er $n = n_1 + n_2 + \ldots + n_k$.

Model: Det antages at $x_1, x_2, \ldots, x_k$ er k forskellige givne konstanter, og at y_{ij}-erne er observerede værdier af uafhængige stokastiske variable Y_{ij} således at Y_{ij} er normalfordelt med middelværdi $\alpha + \beta x_i$ og varians σ^2, kort

$$Y_{ij} \sim \mathcal{N}(\alpha + \beta x_i, \sigma^2),$$

hvor α, β og σ^2 er ukendte parametre. Herved beskriver parametrene α og β den systematiske variation, nemlig y-s lineære afhængighed af x, og variansparameteren σ^2 (samt normalfordelingen) beskriver den tilfældige variation omkring regressionslinien; den tilfældige variation antages at være den samme for alle x.

Estimation: Udregn hjælpestørrelserne

$$\begin{aligned}
\bar{x} &= \frac{1}{n}\sum_{i=1}^{k} n_i x_i\,, \\
\bar{y}_i &= \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}\,, \\
\bar{y} &= \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}\,, \\
SP_{xy} &= \sum_{i=1}^{k} n_i (x_i - \bar{x})(\bar{y}_i - \bar{y}), \\
SS_x &= \sum_{i=1}^{k} n_i (x_i - \bar{x})^2, \\
SS_y &= \sum_{i=1}^{k}\sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2.
\end{aligned}$$

Desuden udregnes $\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_i)^2$.

Så estimeres α og β ved

$$\begin{aligned}\hat{\beta} &= \frac{SP_{xy}}{SS_x}\,,\\ \hat{\alpha} &= \bar{y}-\hat{\beta}\bar{x}\,.\end{aligned}$$

Variansparameteren σ^2 estimeres

$$\begin{aligned}s_{02}^2 &= \frac{1}{n-2}\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij}-(\hat{\alpha}+\hat{\beta}x_i)\right)^2\\ &= \frac{1}{n-2}\left(SS_y-\frac{SP_{xy}^2}{SS_x}\right)\end{aligned}$$

der har $n-2$ frihedsgrader.

Modelkontrol: Man ønsker at teste linearitetsantagelsen med et numerisk test, og derfor formuleres en ny model, nemlig den ensidede variansanalysemodel

$$Y_{ij} \sim \mathcal{N}(\mu_i,\sigma^2),$$

og i forhold hertil testes den oprindelige model som en hypotese. Teststørrelsen er

$$F = \frac{s_2^2}{s_0^2}$$

hvor

$$\begin{aligned}s_0^2 &= \frac{1}{n-k}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_i)^2 \quad \text{og}\\ s_2^2 &= \frac{1}{k-2}\sum_{i=1}^{k}n_i\left(\bar{y}_i-(\hat{\alpha}+\hat{\beta}x_i)\right)^2\\ &= \frac{1}{k-2}\left(SS_y-\frac{SP_{xy}^2}{SS_x}-\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_i)^2\right).\end{aligned}$$

Store værdier af F er signifikante.

Hvis linearitetshypotesen er rigtig, vil F være F-fordelt med $k-2$ og $n-k$ frihedsgrader, så *testsandsynligheden* ε bestemmes som

$$\varepsilon = \mathrm{P}\big(F_{k-2,n-k} > F_{\text{obs}}\big)\,.$$

Hvis ε ikke er meget lille, så er der ikke nogen signifikant afvigelse fra linearitet, dvs. linearitetsmodellen kan opretholdes.

3.14 Nogle begreber og principper for statistisk inferens

En statistisk model for nogle observationer er et udsagn om at observationerne opfattes som observerede værdier af en bestemt slags stokastiske variable, dvs. at observationerne opfattes som værende fremkommet som tilfældige tal fra en bestemt slags sandsynlighedsfordelinger.

Parametre. I den fuldstændige angivelse af de stokastiske variables fordelinger indgår også nogle ukendte parametre.

Modelfunktionen er sandsynlighedsfunktionen/tæthedsfunktionen opfattet som en funktion af observationer såvel som parametre. Modelfunktionen angiver sandsynligheden for at få et bestemt udfald når de ukendte parametre har en bestemt værdi.

Likelihoodfunktionen svarende til et bestemt sæt observationer fremkommer ved at man i modelfunktionen indsætter dette sæt observationer og derved får en funktion af de ukendte parametre.

En statistisk hypotese er et udsagn om at de ukendte parametre opfylder visse betingelser (f.eks. at nogle af dem er ens, eller lig 0, etc.).

Maksimaliseringsestimatet (under en vis hypotese) over de ukendte parametre er den værdi af parametrene som maksimaliserer likelihoodfunktionen eller log-likelihoodfunktionen (og som opfylder de betingelser som hypotesen angiver).

Kvotientteststørrelsen Q for en bestemt hypotese er kvotienten mellem den maksimale likelihoodfunktion under hypotesen og den maksimale likelihoodfunktion under den aktuelle grundmodel.

Testsandsynligheden ε er sandsynligheden for at $Q \leq Q_{\text{obs}}$ (eller $-2\ln Q \geq -2\ln Q_{\text{obs}}$) under antagelse af den hypotese der testes.

Man kan bevise en matematisk sætning der fortæller at i visse nærmere angivne situationer er $-2\ln Q$ med god tilnærmelse χ^2-fordelt med et antal frihedsgrader der udregnes som *antal frie parametre i den aktuelle grundmodel* minus *antal parametre under den hypotese der testes*.

Det var nu nogle matematiske definitioner og en enkelt sætning. Disse henter deres interesse fra nogle *statistiske principper* (som *ikke* er matematiske definitioner eller sætninger):

Likelihoodmetodens grundprincip går ud på at når man har lagt sig fast på sin statistiske model, så gælder at al den information vedrørende de ukendte parametre som observationerne kan give, kan man hente ud af likelihoodfunktionen (dvs. man kan så at sige smide observationerne væk og nøjes med at gemme likelihoodfunktionen).

Dernæst nogle principper om hvordan man benytter den information der således er i likelihoodfunktionen.

Om sammenligning: At et sæt parameterværdier er *bedre end* et andet til at beskrive de foreliggende observationer, betyder at det første sæt giver en højere værdi af likelihoodfunktionen end det andet.

Maximum likelihood princippet: Det *bedste* estimat over de ukendte parametre er det sæt parameterværdier der maksimaliserer likelihoodfunktionen, altså maksimaliseringsestimatet.

Princippet om kvotienttest: Når man ønsker at teste en bestemt statistisk hypotese, skal man som teststørrelse benytte kvotientteststørrelsen Q. En Q-værdi nær 1 betyder at hypotesen giver en næsten lige så god beskrivelse af det observerede som grundmodellen gør. En Q-værdi langt fra 1 betyder at hypotesen ikke er særlig forenelig med det observerede.

Om vurdering af Q: En Q-værdi Q_{obs} *ligger langt fra 1* hvis det er usandsynligt at få en værdi der ligger endnu længere fra 1 end Q_{obs} gør. Det man egentlig har brug for at vide, er derfor ikke værdien af Q_{obs}, men den tilsvarende værdi af testsandsynligheden.

Om signifikansgrænser. Hvis man skal have nogen fornøjelse af det sidste princip, så må man have nogle *retningslinier* for hvornår man skal sige at testsandsynligheden ε er så lille at det er usandsynligt at få en værre Q-værdi, altså at Q er signifikant. Man vil ofte sige at hvis ε er væsentlig mindre end 5%, f.eks. 2.5%, så er Q *signifikant*, dvs. hypotesen er uforenelig med det observerede; hvis omvendt ε er pænt større end 5%, f.eks. 10%, så er Q *ikke-signifikant*, dvs. hypotesen er pænt forenelig med det observerede; hvis endelig ε er tæt på 5%, så er vi havnet i et »gråt område« hvor der ikke er noget klart svar. – Det må dog siges at være god tone altid at angive værdien af ε og ikke blot notere om teststørrelsen er signifikant eller ikke-signifikant.

4 Tabeller

Det gælder for de allerfleste af de fordelinger som den praktisk arbejdende statistiker benytter, at hverken fordelingsfunktionen eller den inverse fordelingsfunktion (der leverer fraktilerne i fordelingen) er lette at beregne når hjælpemidlerne er papir og blyant og »almindelige« matematiske funktioner (så som addition, multiplikation, division, kvadratrod, kvadrering, logaritmefunktion, eksponentialfunktion osv.). I tidligere tider var det et betydeligt regnearbejde at udregne pålidelige numeriske approksimationer til de almindelige fordelinger og deres fraktiler, og man nøjedes derfor med at udregne funktionsværdierne for udvalgte værdier af argumenterne (her er en af forklaringerne på de magiske fem procent!), og statistiske tabeller var noget meget dyrebart (og copyright-belagt). – I vore dage er det anderledes. Enhver kan nu på en almindelig PCer på ingen tid udregne de almindeligt brugte fordelingsfunktioner og fraktiler med stor præcision.

Tabellerne på de følgende sider er udregnet ved brug af Tue Tjurs unit `distr` til Turbo Pascal.

4.1 Fordelingsfunktionen for den normerede normalfordeling

$$\text{Tabel over } \Phi(u) = \int_{-\infty}^{u} \frac{1}{\sqrt{2\pi}} \exp(-x^2/2)\, dx$$

u	$u+5$	$\Phi(u)$	u	$u+5$	$\Phi(u)$
			0.00	5.00	0.5000
-3.75	1.25	0.0001	0.10	5.10	0.5398
-3.50	1.50	0.0002	0.20	5.20	0.5793
-3.25	1.75	0.0006	0.30	5.30	0.6179
-3.00	2.00	0.0013	0.40	5.40	0.6554
-2.80	2.20	0.0026	0.50	5.50	0.6915
-2.60	2.40	0.0047	0.60	5.60	0.7257
-2.40	2.60	0.0082	0.70	5.70	0.7580
-2.20	2.80	0.0139	0.80	5.80	0.7881
-2.00	3.00	0.0228	0.90	5.90	0.8159
-1.90	3.10	0.0287	1.00	6.00	0.8413
-1.80	3.20	0.0359	1.10	6.10	0.8643
-1.70	3.30	0.0446	1.20	6.20	0.8849
-1.60	3.40	0.0548	1.30	6.30	0.9032
-1.50	3.50	0.0668	1.40	6.40	0.9192
-1.40	3.60	0.0808	1.50	6.50	0.9332
-1.30	3.70	0.0968	1.60	6.60	0.9452
-1.20	3.80	0.1151	1.70	6.70	0.9554
-1.10	3.90	0.1357	1.80	6.80	0.9641
-1.00	4.00	0.1587	1.90	6.90	0.9713
-0.90	4.10	0.1841	2.00	7.00	0.9772
-0.80	4.20	0.2119	2.20	7.20	0.9861
-0.70	4.30	0.2420	2.40	7.40	0.9918
-0.60	4.40	0.2743	2.60	7.60	0.9953
-0.50	4.50	0.3085	2.80	7.80	0.9974
-0.40	4.60	0.3446	3.00	8.00	0.9987
-0.30	4.70	0.3821	3.25	8.25	0.9994
-0.20	4.80	0.4207	3.50	8.50	0.9998
-0.10	4.90	0.4602	3.75	8.75	0.9999

4.2 Fraktiler i den normerede normalfordeling

α	$u_\alpha = \Phi^{-1}(\alpha)$	$u_\alpha + 5$	α	$u_\alpha = \Phi^{-1}(\alpha)$	$u_\alpha + 5$
			0.500	0.000	5.000
0.001	-3.090	1.910	0.520	0.050	5.050
0.002	-2.878	2.122	0.540	0.100	5.100
0.005	-2.576	2.424	0.560	0.151	5.151
0.010	-2.326	2.674	0.580	0.202	5.202
0.015	-2.170	2.830	0.600	0.253	5.253
0.020	-2.054	2.946	0.620	0.305	5.305
0.025	-1.960	3.040	0.640	0.358	5.358
0.030	-1.881	3.119	0.660	0.412	5.412
0.035	-1.812	3.188	0.680	0.468	5.468
0.040	-1.751	3.249	0.700	0.524	5.524
0.045	-1.695	3.305	0.720	0.583	5.583
0.050	-1.645	3.355	0.740	0.643	5.643
0.055	-1.598	3.402	0.750	0.674	5.674
0.060	-1.555	3.445	0.760	0.706	5.706
0.070	-1.476	3.524	0.780	0.772	5.772
0.080	-1.405	3.595	0.800	0.842	5.842
0.090	-1.341	3.659	0.825	0.935	5.935
0.100	-1.282	3.718	0.850	1.036	6.036
0.125	-1.150	3.850	0.875	1.150	6.150
0.150	-1.036	3.964	0.900	1.282	6.282
0.175	-0.935	4.065	0.910	1.341	6.341
0.200	-0.842	4.158	0.920	1.405	6.405
0.220	-0.772	4.228	0.930	1.476	6.476
0.240	-0.706	4.294	0.940	1.555	6.555
0.250	-0.674	4.326	0.945	1.598	6.598
0.260	-0.643	4.357	0.950	1.645	6.645
0.280	-0.583	4.417	0.955	1.695	6.695
0.300	-0.524	4.476	0.960	1.751	6.751
0.320	-0.468	4.532	0.965	1.812	6.812
0.340	-0.412	4.588	0.970	1.881	6.881
0.360	-0.358	4.642	0.975	1.960	6.960
0.380	-0.305	4.695	0.980	2.054	7.054
0.400	-0.253	4.747	0.985	2.170	7.170
0.420	-0.202	4.798	0.990	2.326	7.326
0.440	-0.151	4.849	0.995	2.576	7.576
0.460	-0.100	4.900	0.998	2.878	7.878
0.480	-0.050	4.950	0.999	3.090	8.090

4.3 Fraktiler i χ^2-fordelingen

Fraktiler i χ^2-fordelingen med f frihedsgrader

	Sandsynlighed i procent						
f	50	90	95	97.5	99	99.5	99.9
1	0.45	2.71	3.84	5.02	6.63	7.88	10.83
2	1.39	4.61	5.99	7.38	9.21	10.60	13.82
3	2.37	6.25	7.81	9.35	11.34	12.84	16.27
4	3.36	7.78	9.49	11.14	13.28	14.86	18.47
5	4.35	9.24	11.07	12.83	15.09	16.75	20.52
6	5.35	10.64	12.59	14.45	16.81	18.55	22.46
7	6.35	12.02	14.07	16.01	18.48	20.28	24.32
8	7.34	13.36	15.51	17.53	20.09	21.95	26.12
9	8.34	14.68	16.92	19.02	21.67	23.59	27.88
10	9.34	15.99	18.31	20.48	23.21	25.19	29.59
11	10.34	17.28	19.68	21.92	24.72	26.76	31.26
12	11.34	18.55	21.03	23.34	26.22	28.30	32.91
13	12.34	19.81	22.36	24.74	27.69	29.82	34.53
14	13.34	21.06	23.68	26.12	29.14	31.32	36.12
15	14.34	22.31	25.00	27.49	30.58	32.80	37.70
16	15.34	23.54	26.30	28.85	32.00	34.27	39.25
17	16.34	24.77	27.59	30.19	33.41	35.72	40.79
18	17.34	25.99	28.87	31.53	34.81	37.16	42.31
19	18.34	27.20	30.14	32.85	36.19	38.58	43.82
20	19.34	28.41	31.41	34.17	37.57	40.00	45.31
21	20.34	29.62	32.67	35.48	38.93	41.40	46.80
22	21.34	30.81	33.92	36.78	40.29	42.80	48.27
23	22.34	32.01	35.17	38.08	41.64	44.18	49.73
24	23.34	33.20	36.42	39.36	42.98	45.56	51.18
25	24.34	34.38	37.65	40.65	44.31	46.93	52.62
26	25.34	35.56	38.89	41.92	45.64	48.29	54.05
27	26.34	36.74	40.11	43.19	46.96	49.64	55.48
28	27.34	37.92	41.34	44.46	48.28	50.99	56.89
29	28.34	39.09	42.56	45.72	49.59	52.34	58.30
30	29.34	40.26	43.77	46.98	50.89	53.67	59.70
31	30.34	41.42	44.99	48.23	52.19	55.00	61.10
32	31.34	42.58	46.19	49.48	53.49	56.33	62.49
33	32.34	43.75	47.40	50.73	54.78	57.65	63.87
34	33.34	44.90	48.60	51.97	56.06	58.96	65.25
35	34.34	46.06	49.80	53.20	57.34	60.27	66.62
36	35.34	47.21	51.00	54.44	58.62	61.58	67.99
37	36.34	48.36	52.19	55.67	59.89	62.88	69.35
38	37.34	49.51	53.38	56.90	61.16	64.18	70.70
39	38.34	50.66	54.57	58.12	62.43	65.48	72.05
40	39.34	51.81	55.76	59.34	63.69	66.77	73.40

Fraktiler i χ^2-fordelingen med f frihedsgrader

	Sandsynlighed i procent						
f	50	90	95	97.5	99	99.5	99.9
41	40.34	52.95	56.94	60.56	64.95	68.05	74.74
42	41.34	54.09	58.12	61.78	66.21	69.34	76.08
43	42.34	55.23	59.30	62.99	67.46	70.62	77.42
44	43.34	56.37	60.48	64.20	68.71	71.89	78.75
45	44.34	57.51	61.66	65.41	69.96	73.17	80.08
46	45.34	58.64	62.83	66.62	71.20	74.44	81.40
47	46.34	59.77	64.00	67.82	72.44	75.70	82.72
48	47.34	60.91	65.17	69.02	73.68	76.97	84.04
49	48.33	62.04	66.34	70.22	74.92	78.23	85.35
50	49.33	63.17	67.50	71.42	76.15	79.49	86.66
51	50.33	64.30	68.67	72.62	77.39	80.75	87.97
52	51.33	65.42	69.83	73.81	78.62	82.00	89.27
53	52.33	66.55	70.99	75.00	79.84	83.25	90.57
54	53.33	67.67	72.15	76.19	81.07	84.50	91.87
55	54.33	68.80	73.31	77.38	82.29	85.75	93.17
56	55.33	69.92	74.47	78.57	83.51	86.99	94.46
57	56.33	71.04	75.62	79.75	84.73	88.24	95.75
58	57.33	72.16	76.78	80.94	85.95	89.48	97.04
59	58.33	73.28	77.93	82.12	87.17	90.72	98.32
60	59.33	74.40	79.08	83.30	88.38	91.95	99.61
61	60.33	75.51	80.23	84.48	89.59	93.19	100.89
62	61.33	76.63	81.38	85.65	90.80	94.42	102.17
63	62.33	77.75	82.53	86.83	92.01	95.65	103.44
64	63.33	78.86	83.68	88.00	93.22	96.88	104.72
65	64.33	79.97	84.82	89.18	94.42	98.11	105.99
66	65.33	81.09	85.96	90.35	95.63	99.33	107.26
67	66.33	82.20	87.11	91.52	96.83	100.55	108.53
68	67.33	83.31	88.25	92.69	98.03	101.78	109.79
69	68.33	84.42	89.39	93.86	99.23	103.00	111.06
70	69.33	85.53	90.53	95.02	100.43	104.21	112.32
71	70.33	86.64	91.67	96.19	101.62	105.43	113.58
72	71.33	87.74	92.81	97.35	102.82	106.65	114.84
73	72.33	88.85	93.95	98.52	104.01	107.86	116.09
74	73.33	89.96	95.08	99.68	105.20	109.07	117.35
75	74.33	91.06	96.22	100.84	106.39	110.29	118.60
76	75.33	92.17	97.35	102.00	107.58	111.50	119.85
77	76.33	93.27	98.48	103.16	108.77	112.70	121.10
78	77.33	94.37	99.62	104.32	109.96	113.91	122.35
79	78.33	95.48	100.75	105.47	111.14	115.12	123.59
80	79.33	96.58	101.88	106.63	112.33	116.32	124.84

4.4 Fraktiler i F-fordelingen

90% fraktiler i F-fordelingen.

f_1 er antal frihedsgrader for tælleren, f_2 er antal frihedsgrader for nævneren.

	f_1										
f_2	1	2	3	4	5	6	7	8	9	10	15
1	39.86	49.50	53.59	55.83	57.24	58.20	58.91	59.44	59.86	60.19	61.22
2	8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.38	9.39	9.42
3	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.24	5.23	5.20
4	4.54	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.94	3.92	3.87
5	4.06	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.32	3.30	3.24
6	3.78	3.46	3.29	3.18	3.11	3.05	3.01	2.98	2.96	2.94	2.87
7	3.59	3.26	3.07	2.96	2.88	2.83	2.78	2.75	2.72	2.70	2.63
8	3.46	3.11	2.92	2.81	2.73	2.67	2.62	2.59	2.56	2.54	2.46
9	3.36	3.01	2.81	2.69	2.61	2.55	2.51	2.47	2.44	2.42	2.34
10	3.29	2.92	2.73	2.61	2.52	2.46	2.41	2.38	2.35	2.32	2.24
11	3.23	2.86	2.66	2.54	2.45	2.39	2.34	2.30	2.27	2.25	2.17
12	3.18	2.81	2.61	2.48	2.39	2.33	2.28	2.24	2.21	2.19	2.10
13	3.14	2.76	2.56	2.43	2.35	2.28	2.23	2.20	2.16	2.14	2.05
14	3.10	2.73	2.52	2.39	2.31	2.24	2.19	2.15	2.12	2.10	2.01
15	3.07	2.70	2.49	2.36	2.27	2.21	2.16	2.12	2.09	2.06	1.97
16	3.05	2.67	2.46	2.33	2.24	2.18	2.13	2.09	2.06	2.03	1.94
17	3.03	2.64	2.44	2.31	2.22	2.15	2.10	2.06	2.03	2.00	1.91
18	3.01	2.62	2.42	2.29	2.20	2.13	2.08	2.04	2.00	1.98	1.89
19	2.99	2.61	2.40	2.27	2.18	2.11	2.06	2.02	1.98	1.96	1.86
20	2.97	2.59	2.38	2.25	2.16	2.09	2.04	2.00	1.96	1.94	1.84
22	2.95	2.56	2.35	2.22	2.13	2.06	2.01	1.97	1.93	1.90	1.81
24	2.93	2.54	2.33	2.19	2.10	2.04	1.98	1.94	1.91	1.88	1.78
26	2.91	2.52	2.31	2.17	2.08	2.01	1.96	1.92	1.88	1.86	1.76
28	2.89	2.50	2.29	2.16	2.06	2.00	1.94	1.90	1.87	1.84	1.74
30	2.88	2.49	2.28	2.14	2.05	1.98	1.93	1.88	1.85	1.82	1.72
40	2.84	2.44	2.23	2.09	2.00	1.93	1.87	1.83	1.79	1.76	1.66
50	2.81	2.41	2.20	2.06	1.97	1.90	1.84	1.80	1.76	1.73	1.63
75	2.77	2.37	2.16	2.02	1.93	1.85	1.80	1.75	1.72	1.69	1.58
100	2.76	2.36	2.14	2.00	1.91	1.83	1.78	1.73	1.69	1.66	1.56
150	2.74	2.34	2.12	1.98	1.89	1.81	1.76	1.71	1.67	1.64	1.53
200	2.73	2.33	2.11	1.97	1.88	1.80	1.75	1.70	1.66	1.63	1.52
300	2.72	2.32	2.10	1.96	1.87	1.79	1.74	1.69	1.65	1.62	1.51
400	2.72	2.32	2.10	1.96	1.86	1.79	1.73	1.69	1.65	1.61	1.50
500	2.72	2.31	2.09	1.96	1.86	1.79	1.73	1.68	1.64	1.61	1.50

95% fraktiler i F-fordelingen.

f_1 er antal frihedsgrader for tælleren, f_2 er antal frihedsgrader for nævneren.

	f_1										
f_2	1	2	3	4	5	6	7	8	9	10	15
1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54	241.88	245.95
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.43
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.70
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.86
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.62
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	3.94
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.51
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.22
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.01
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.85
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.72
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.62
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.53
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.46
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.40
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.35
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.31
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.27
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.23
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.20
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.15
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.11
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.07
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.04
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.01
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	1.92
50	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07	2.03	1.87
75	3.97	3.12	2.73	2.49	2.34	2.22	2.13	2.06	2.01	1.96	1.80
100	3.94	3.09	2.70	2.46	2.31	2.19	2.10	2.03	1.97	1.93	1.77
150	3.90	3.06	2.66	2.43	2.27	2.16	2.07	2.00	1.94	1.89	1.73
200	3.89	3.04	2.65	2.42	2.26	2.14	2.06	1.98	1.93	1.88	1.72
300	3.87	3.03	2.63	2.40	2.24	2.13	2.04	1.97	1.91	1.86	1.70
400	3.86	3.02	2.63	2.39	2.24	2.12	2.03	1.96	1.90	1.85	1.69
500	3.86	3.01	2.62	2.39	2.23	2.12	2.03	1.96	1.90	1.85	1.69

97.5% fraktiler i F-fordelingen.

f_1 er antal frihedsgrader for tælleren, f_2 er antal frihedsgrader for nævneren.

	f_1										
f_2	1	2	3	4	5	6	7	8	9	10	15
1	647.79	799.50	864.16	899.58	921.85	937.11	948.22	956.66	963.28	968.63	984.87
2	38.51	39.00	39.17	39.25	39.30	39.33	39.36	39.37	39.39	39.40	39.43
3	17.44	16.04	15.44	15.10	14.88	14.73	14.62	14.54	14.47	14.42	14.25
4	12.22	10.65	9.98	9.60	9.36	9.20	9.07	8.98	8.90	8.84	8.66
5	10.01	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68	6.62	6.43
6	8.81	7.26	6.60	6.23	5.99	5.82	5.70	5.60	5.52	5.46	5.27
7	8.07	6.54	5.89	5.52	5.29	5.12	4.99	4.90	4.82	4.76	4.57
8	7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.36	4.30	4.10
9	7.21	5.71	5.08	4.72	4.48	4.32	4.20	4.10	4.03	3.96	3.77
10	6.94	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.78	3.72	3.52
11	6.72	5.26	4.63	4.28	4.04	3.88	3.76	3.66	3.59	3.53	3.33
12	6.55	5.10	4.47	4.12	3.89	3.73	3.61	3.51	3.44	3.37	3.18
13	6.41	4.97	4.35	4.00	3.77	3.60	3.48	3.39	3.31	3.25	3.05
14	6.30	4.86	4.24	3.89	3.66	3.50	3.38	3.29	3.21	3.15	2.95
15	6.20	4.77	4.15	3.80	3.58	3.41	3.29	3.20	3.12	3.06	2.86
16	6.12	4.69	4.08	3.73	3.50	3.34	3.22	3.12	3.05	2.99	2.79
17	6.04	4.62	4.01	3.66	3.44	3.28	3.16	3.06	2.98	2.92	2.72
18	5.98	4.56	3.95	3.61	3.38	3.22	3.10	3.01	2.93	2.87	2.67
19	5.92	4.51	3.90	3.56	3.33	3.17	3.05	2.96	2.88	2.82	2.62
20	5.87	4.46	3.86	3.51	3.29	3.13	3.01	2.91	2.84	2.77	2.57
22	5.79	4.38	3.78	3.44	3.22	3.05	2.93	2.84	2.76	2.70	2.50
24	5.72	4.32	3.72	3.38	3.15	2.99	2.87	2.78	2.70	2.64	2.44
26	5.66	4.27	3.67	3.33	3.10	2.94	2.82	2.73	2.65	2.59	2.39
28	5.61	4.22	3.63	3.29	3.06	2.90	2.78	2.69	2.61	2.55	2.34
30	5.57	4.18	3.59	3.25	3.03	2.87	2.75	2.65	2.57	2.51	2.31
40	5.42	4.05	3.46	3.13	2.90	2.74	2.62	2.53	2.45	2.39	2.18
50	5.34	3.97	3.39	3.05	2.83	2.67	2.55	2.46	2.38	2.32	2.11
75	5.23	3.88	3.30	2.96	2.74	2.58	2.46	2.37	2.29	2.22	2.01
100	5.18	3.83	3.25	2.92	2.70	2.54	2.42	2.32	2.24	2.18	1.97
150	5.13	3.78	3.20	2.87	2.65	2.49	2.37	2.28	2.20	2.13	1.92
200	5.10	3.76	3.18	2.85	2.63	2.47	2.35	2.26	2.18	2.11	1.90
300	5.07	3.73	3.16	2.83	2.61	2.45	2.33	2.23	2.16	2.09	1.88
400	5.06	3.72	3.15	2.82	2.60	2.44	2.32	2.22	2.15	2.08	1.87
500	5.05	3.72	3.14	2.81	2.59	2.43	2.31	2.22	2.14	2.07	1.86

99% fraktiler i F-fordelingen.

f_1 er antal frihedsgrader for tælleren, f_2 er antal frihedsgrader for nævneren.

	f_1										
f_2	1	2	3	4	5	6	7	8	9	10	15
1	4052.18	4999.50	5403.35	5624.58	5763.65	5858.99	5928.36	5981.07	6022.47	6055.85	6157.28
2	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39	99.40	99.43
3	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35	27.23	26.87
4	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55	14.20
5	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05	9.72
6	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.56
7	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.31
8	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.52
9	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	4.96
10	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.56
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.25
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.01
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	3.82
14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94	3.66
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.52
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.41
17	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.31
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.23
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.15
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.09
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	2.98
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	2.89
26	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18	3.09	2.81
28	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12	3.03	2.75
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.70
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.52
50	7.17	5.06	4.20	3.72	3.41	3.19	3.02	2.89	2.78	2.70	2.42
75	6.99	4.90	4.05	3.58	3.27	3.05	2.89	2.76	2.65	2.57	2.29
100	6.90	4.82	3.98	3.51	3.21	2.99	2.82	2.69	2.59	2.50	2.22
150	6.81	4.75	3.91	3.45	3.14	2.92	2.76	2.63	2.53	2.44	2.16
200	6.76	4.71	3.88	3.41	3.11	2.89	2.73	2.60	2.50	2.41	2.13
300	6.72	4.68	3.85	3.38	3.08	2.86	2.70	2.57	2.47	2.38	2.10
400	6.70	4.66	3.83	3.37	3.06	2.85	2.68	2.56	2.45	2.37	2.08
500	6.69	4.65	3.82	3.36	3.05	2.84	2.68	2.55	2.44	2.36	2.07

4.5 Fraktiler i t-fordelingen

Fraktiler i t-fordelingen med f frihedsgrader

	Sandsynlighed i procent						
f	90	95	97.5	99	99.5	99.9	f
1	3.078	6.314	12.706	31.821	63.657	318.309	1
2	1.886	2.920	4.303	6.965	9.925	22.327	2
3	1.638	2.353	3.182	4.541	5.841	10.215	3
4	1.533	2.132	2.776	3.747	4.604	7.173	4
5	1.476	2.015	2.571	3.365	4.032	5.893	5
6	1.440	1.943	2.447	3.143	3.707	5.208	6
7	1.415	1.895	2.365	2.998	3.499	4.785	7
8	1.397	1.860	2.306	2.896	3.355	4.501	8
9	1.383	1.833	2.262	2.821	3.250	4.297	9
10	1.372	1.812	2.228	2.764	3.169	4.144	10
11	1.363	1.796	2.201	2.718	3.106	4.025	11
12	1.356	1.782	2.179	2.681	3.055	3.930	12
13	1.350	1.771	2.160	2.650	3.012	3.852	13
14	1.345	1.761	2.145	2.624	2.977	3.787	14
15	1.341	1.753	2.131	2.602	2.947	3.733	15
16	1.337	1.746	2.120	2.583	2.921	3.686	16
17	1.333	1.740	2.110	2.567	2.898	3.646	17
18	1.330	1.734	2.101	2.552	2.878	3.610	18
19	1.328	1.729	2.093	2.539	2.861	3.579	19
20	1.325	1.725	2.086	2.528	2.845	3.552	20
21	1.323	1.721	2.080	2.518	2.831	3.527	21
22	1.321	1.717	2.074	2.508	2.819	3.505	22
23	1.319	1.714	2.069	2.500	2.807	3.485	23
24	1.318	1.711	2.064	2.492	2.797	3.467	24
25	1.316	1.708	2.060	2.485	2.787	3.450	25
30	1.310	1.697	2.042	2.457	2.750	3.385	30
50	1.299	1.676	2.009	2.403	2.678	3.261	50
75	1.293	1.665	1.992	2.377	2.643	3.202	75
100	1.290	1.660	1.984	2.364	2.626	3.174	100
150	1.287	1.655	1.976	2.351	2.609	3.145	150
200	1.286	1.653	1.972	2.345	2.601	3.131	200
400	1.284	1.649	1.966	2.336	2.588	3.111	400

4.6 Fraktiler i χ^2/f-fordelingen

Fraktiler i χ^2/f-fordelingen med f frihedsgrader

	Sandsynlighed i procent						
f	1	2.5	5	50	95	97.5	99
1	0.000	0.001	0.004	0.455	3.841	5.024	6.635
2	0.010	0.025	0.051	0.693	2.996	3.689	4.605
3	0.038	0.072	0.117	0.789	2.605	3.116	3.782
4	0.074	0.121	0.178	0.839	2.372	2.786	3.319
5	0.111	0.166	0.229	0.870	2.214	2.567	3.017
6	0.145	0.206	0.273	0.891	2.099	2.408	2.802
7	0.177	0.241	0.310	0.907	2.010	2.288	2.639
8	0.206	0.272	0.342	0.918	1.938	2.192	2.511
9	0.232	0.300	0.369	0.927	1.880	2.114	2.407
10	0.256	0.325	0.394	0.934	1.831	2.048	2.321
11	0.278	0.347	0.416	0.940	1.789	1.993	2.248
12	0.298	0.367	0.436	0.945	1.752	1.945	2.185
13	0.316	0.385	0.453	0.949	1.720	1.903	2.130
14	0.333	0.402	0.469	0.953	1.692	1.866	2.082
15	0.349	0.417	0.484	0.956	1.666	1.833	2.039
16	0.363	0.432	0.498	0.959	1.644	1.803	2.000
17	0.377	0.445	0.510	0.961	1.623	1.776	1.965
18	0.390	0.457	0.522	0.963	1.604	1.751	1.934
19	0.402	0.469	0.532	0.965	1.587	1.729	1.905
20	0.413	0.480	0.543	0.967	1.571	1.708	1.878
22	0.434	0.499	0.561	0.970	1.542	1.672	1.831
24	0.452	0.517	0.577	0.972	1.517	1.640	1.791
26	0.469	0.532	0.592	0.974	1.496	1.612	1.755
28	0.484	0.547	0.605	0.976	1.476	1.588	1.724
30	0.498	0.560	0.616	0.978	1.459	1.566	1.696
40	0.554	0.611	0.663	0.983	1.394	1.484	1.592
50	0.594	0.647	0.695	0.987	1.350	1.428	1.523
75	0.660	0.706	0.747	0.991	1.283	1.345	1.419
100	0.701	0.742	0.779	0.993	1.243	1.296	1.358
150	0.751	0.787	0.818	0.996	1.197	1.239	1.288
200	0.782	0.814	0.841	0.997	1.170	1.205	1.247
300	0.820	0.846	0.870	0.998	1.138	1.166	1.200
400	0.843	0.866	0.887	0.998	1.119	1.143	1.172
800	0.887	0.904	0.919	0.999	1.084	1.100	1.120

5 Stikord