HD Dataanalyse

Note 5: Analyse af én gruppe (konfidensinterval)

Copenhagen Business School

Emne i dette notesæt

Mange variable kan overordnet set (dvs. teoretisk) beskrives ved hjælp af enten en normalfordeling eller en binomialfordeling.

For at kunne lave konkret analyse af en variabel, skal vi først have estimeret parametrene i dens teoretiske fordeling (dvs. normal- eller binomialfordelingen).

For en variabel, der kan beskrives ved en ***normalfordeling***, er det først og fremmest middelværdien μ, vi er interesseret i at estimere. Et estimat af μ udtrykker vores gæt på, hvad værdien af variablens ukendte (teoretiske) middelværdi er. Men som ethvert gæt, er også ***vores gæt*** på værdien af μ ***behæftet med usikkerhed***.

Spørgsmålet er: Hvor sikre kan vi være på vores gæt på værdien af μ? Kunne vi ligeså godt gætte på en helt anden værdi? Eller kan vi føle os nogenlunde sikre på, at vores gæt er forholdsvis præcist?

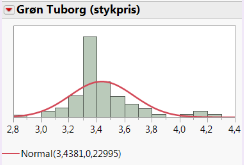
(tilsvarende overvejelser gælder for et estimat af sandsynligheden p i en binomialfordeling)

Note 5 beskæftiger sig med, hvordan vi kan måle præcisionen af de gæt på ukendte parametre i en teoretisk fordeling, vi har behov for at lave for at kunne bruge henholdsvis normal- eller binomialfordelingen til videre statistisk analyse.

EMNE I DETTE NOTESÆT

Eksempel: Ølsalg

Ser vi på prisen for 1 stk. Grøn Tuborg (33 cl glasflaske) i supermarkedskæden Føtex, får vi følgende histogram over variablens (empiriske) fordeling:



Datamaterialet består af 157 ugers priser og kan med en vis rimelighed beskrives af en normalfordeling med estimerede parametre $\hat{\mu} = 3,44$ og $\hat{\sigma} = 0,23$.

Den ukendte teoretiske middelværdi μ for prisen på Grøn Tuborg, gætter vi således på er 3,44 kr. Med andre ord, på baggrund af datamaterialet er vores bedste gæt, at vi skal forvente at prisen på Grøn Tuborg i Føtex er 3,44 kr.

Vores gæt på 3,44 kr. er baseret på samtlige 157 prisobservationer i datamaterialet (det er beregnet som gennemsnittet af alle 157 observationer), og er naturligvis kun et gæt, for vi ved jo godt, at prisen kan variere lidt fra uge til uge.

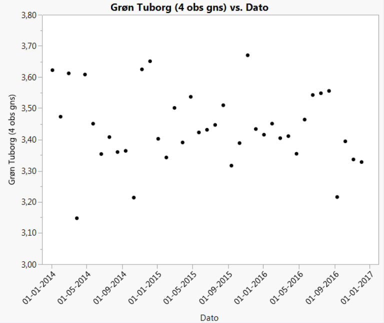
Hvis vi nu i stedet nøjes med at gætte på værdien af μ på baggrund af de første 4 prisobservationer (svarende til priserne i perioden 30/12 2013 til 26/1 2014), får vi i stedet et gæt på $\hat{\mu} = 3{,}62$. Gætter vi baggrund af de næstfølgende 4 prisobservationer (svarende til priserne i perioden 27/1 2014 til 23/2 2014), får vi et gæt på $\hat{\mu} = 3,47$.

Emne i dette notesæt

Eksempel: Ølsalg (fortsat)

Hver gang vi vælger et nyt datasæt at basere vores gæt på, får vi en anden værdi af $\hat{\mu}$. Det skyldes den almindelige variation i priserne fra uge til uge. Det betyder, at et gæt på værdien af μ altid vil være behæftet med en vis usikkerhed. Det gælder, uanset hvor få eller hvor mange observationer vi baserer vores gæt på.

Hvis vi, som illustrativt eksempel, går igennem datasættet og regner et gæt på værdien af μ ud for 4 ugers observationer ad gangen og herefter tegner de mange gæt op i en figur, så kommer det til at se således ud:



Der er helt tydeligt stor variation på de mange forskellige gæt på værdien af μ, vi kan producere ud fra 4 observationer i datamaterialet.

EMNE I DETTE NOTESÆT

Eksempel: Ølsalg (fortsat)

Selv hvis vi baserer vores gæt på alle 157 prisobservationer, vil det være behæftet med en vis usikkerhed.

Vi vil jo næppe forvente, at såfremt vi venter, til der er gået nye 157 uger, og vi dermed kan beregne et nyt gennemsnit af 157 ugers priser, at vi så vil få præcis samme gæt (3,44 kr.) én gang til.

Summa summarum:
Når vi beregner et gæt på $\hat{\mu} = 3,44$ på den ukendte middelværdi for prisen på Grøn Tuborg i Føtex, så er gættet behæftet med usikkerhed.

Det vi skal se på i denne note er, hvordan vi kan måle denne usikkerhed, dvs. hvordan vi kan afgøre, hvor præcist vores gæt på 3,44 kr. for 1 stk. Grøn Tuborg i Føtex egentlig er. Ku′ vi ligeså godt have gættet på en pris på 4 kr., eller er det mere rimeligt at tro, at den forventede pris ligger omkring 3,44 kr. plus/minus et par øre?

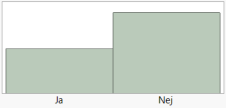
▶ JMP-video [Analyze -> Distribution ; Graph -> Graph Builder]

Emne i dette notesæt

Eksempel: Skat

I en spørgeskemaundersøgelse om velfærd og skat udsendt til 975 personer har vi set, at svarene på spørgsmålet *"Er topskatten for høj?"* kan beskrives ved hjælp af en binomialfordeling med parametre $n = 975$ og $\hat{p} = 35{,}5\%$.

Er topskatten for høj?



Frequencies

Level	Count	Prob
Ja	346	0,35487
Nej	629	0,64513
Total	975	1,00000

N Missing 0
2 Levels

Her beskriver p sandsynligheden for, at en given respondent svarer "Ja" på spørgsmålet, og på baggrund af datamaterialet estimeres den til 35,5%.

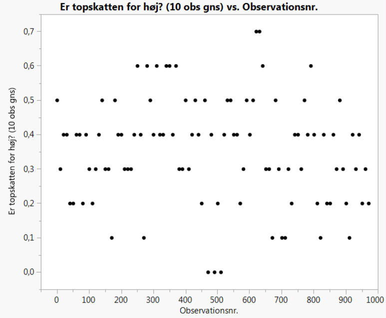
Vores gæt på $\hat{p} = 35{,}5\%$ er baseret på svarene fra alle datamaterialets 975 respondenter (det er beregnet som andelen af de 975 respondenter, der svarede "Ja"). Men eftersom det kun er et gæt, er det naturligvis behæftet med usikkerhed.

Hvis vi i stedet nøjes med at basere vores gæt på p på svarene fra de 10 første respondenter, så får vi en værdi på $\hat{p} = 0{,}5$. Ser vi i stedet på svarene fra de næste 10 respondenter, bliver værdien $\hat{p} = 0{,}3$ osv.

EMNE I DETTE NOTESÆT

Eksempel: Skat (fortsat)

Går vi igennem datasættet og beregner et gæt på p ved at se på svarene for 10 respondenter ad gangen, får vi nedenstående resultat:



Der er helt tydeligt stor variation på de mange forskellige gæt på værdien af p. Som i foregående eksempel er pointen også her, at ethvert gæt på den ukendte sandsynlighed p vil være behæftet med usikkerhed. Det gælder uanset hvor få eller hvor mange personers svar, vi baserer vores gæt på.

Spørgsmålet er derfor også her, hvor præcist vi kan regne med, at vores gæt på $\hat{p} = 35{,}5\%$ egentlig er?

JMP-video [Analyze -> Distribution ; Graph -> Graph Builder]

INDHOLDSFORTEGNELSE

Fordeling af $\hat{\mu}$

Når vi skal estimere (dvs. gætte på) en værdi af de ukendte parametre μ og σ i en normalfordeling, er det ikke på forhånd klart, hvordan vi gør det bedst muligt.

Der er imidlertid tungtvejende teoretiske argumenter for, at den bedste måde at estimere værdier for μ og σ på er som angivet i nedenstående resultat.

Resultat [Fordeling af $\hat{\mu}$]

Lad $X_1,\ldots,X_n$ være indbyrdes uafhængige observationer af en variabel, der er normalfordelt $N(\mu,\sigma)$.

Vi estimerer normalfordelingens parametre ved

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad \hat{\sigma} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \hat{\mu})^2}$$

Estimatet af μ bliver selv normalfordelt

$$\hat{\mu} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

Bemærk: Symbolet μ betegner en ukendt teoretisk værdi, som vi er interesseret i at estimere. Symbolet $\hat{\mu}$ betegner vores gæt på den ukendte størrelse μ.

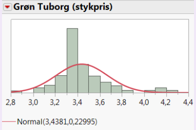
Forklaring af resultatet:

- Vi antager, at vi har et datamateriale med n observationer af en normalfordelt variabel $N(\mu, \sigma)$
- Den første observation i datamaterialet betegner vi med X_1, den næste med X_2 osv.
- Vi tænker på det, som om vi endnu ikke kender talværdierne af de enkelte observationer. Derfor skriver vi observationerne som store bogstaver (dvs. X_1, X_2 osv.). Det vi antager er, at alle observationer kan beskrives af ***den samme*** normalfordeling $N(\mu, \sigma)$
- Resultatet fortæller, at vores estimat $\hat{\mu}$ af den ukendte middelværdi μ i sig selv kan beskrives ved en normalfordeling
  (dvs. at hvis vi udregner en masse forskellige estimater, så vil estimaterne kunne beskrives ved en normalfordeling)
- Denne normalfordeling har parametre μ (middelværdi) og $\frac{\sigma}{\sqrt{n}}$ (standardafvigelse)
- Ideen er at udnytte det, til at sige noget om størrelsen af usikkerheden på vores estimat $\hat{\mu}$

FORDELING AF $\hat{\mu}$

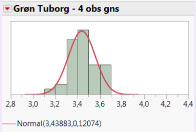
Eksempel: Ølsalg

Fordelingen af de ugentlige observationer af prisen for 1 stk. Grøn Tuborg (33 cl glasflaske) i supermarkedskæden Føtex fremgår af nedenstående histogram:



Den indtegnede normalfordelingskurve viser, at prisen på Grøn Tuborg med en vis rimelighed kan beskrives ved en normalfordeling $N(\mu, \sigma)$.

Hvis vi, for eksemplets skyld, vælger at estimere middelværdien μ ud fra $n = 4$ ugers prisobservationer, så får vi en lang række forskellige estimater afhængig af præcis hvilke 4 ugers priser, vi vælger. Tegner vi et histogram baseret på alle de forskellige estimater, får vi tegnet ***fordelingen af estimatet af*** μ baseret på 4 observationer:



Det fremgår af histogrammet, at fordelingen af 4 ugers-estimatet har samme centrum (μ) som den oprindelige variabel, men at det har en meget mindre standardafvigelse ($\sigma/\sqrt{4}$) end standardafvigelsen i fordelingen af den oprindelige variabel (σ).

JMP-video [Analyze -> Distribution]

TRANSFORMATION

Vi er interesseret i at finde ud af, hvor præcist vores estimat $\hat{\mu}$ af den ukendte middelværdi μ i normalfordelingen er.

Vi ved, at estimatet i sig selv kan beskrives ved hjælp af en normalfordeling $\hat{\mu} \sim N\left(\mu, \sigma/\sqrt{n}\right)$.

Det kan vi alternativt også skrive som $\dfrac{\hat{\mu} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$ (pga. normalfordelingens transformationsegenskab).

I størrelsen $\frac{\hat{\mu}-\mu}{\sigma/\sqrt{n}}$ indgår imidlertid ***TO*** ukendte størrelser – både μ og σ. Hvis vi vil sige noget om estimatet af μ, er vi nødt til først at fjerne den ukendte størrelse σ.

Det problem løser vi, ved at indsætte vores estimat $\hat{\sigma}$ for σ, dvs. ved i stedet at se på $\frac{\hat{\mu}-\mu}{\hat{\sigma}/\sqrt{n}}$.

Det har imidlertid som konsekvens, at resultatet ændrer sig til i stedet at blive $\frac{\hat{\mu}-\mu}{\hat{\sigma}/\sqrt{n}} \sim t(n-1)$, hvor $t(n-1)$ er den såkaldte **t-fordeling** *("t distribution")* med $n-1$ **frihedsgrader** *("degrees of freedom")*.

Transformation

Vi kan opskrive disse overvejelser mere formelt i form af nedenstående resultater.

Resultat [transformation fra $N(\mu,\sigma)$ til $N(0,1)$]

Hvis $X_1,\ldots,X_n$ er indbyrdes uafhængige normalfordelte variable $N(\mu,\sigma)$, så er

$$Z = \frac{\hat{\mu}-\mu}{\frac{\sigma}{\sqrt{n}}}$$

standardnormalfordelt $Z \sim N(0,1)$.

Resultat [transformation fra $N(\mu,\sigma)$ til t]

Hvis $X_1,\ldots,X_n$ er indbyrdes uafhængige normalfordelte variable $N(\mu,\sigma)$, så er

$$Z = \frac{\hat{\mu}-\mu}{\frac{\hat{\sigma}}{\sqrt{n}}}$$

t-fordelt med $n-1$ frihedsgrader $Z \sim t(n-1)$.

Forskellen på de to resultater er, at vi i det nederste resultat erstatter den ukendte teoretiske standardafvigelse σ med vores estimat $\hat{\sigma}$.Det har som konsekvens, at standardnormalfordelingen $N(0,1)$ bliver skiftet ud med t-fordelingen.

t-FORDELINGEN

t-fordelingen er vigtig, fordi den beskriver fordelingen af størrelsen $\frac{\hat{\mu}-\mu}{\hat{\sigma}/\sqrt{n}}$, og denne størrelse kan – som vi skal se om lidt – sammen med t-fordelingen bruges til at udtale sig om, hvor præcist et estimat $\hat{\mu}$ er.

Men først nogle facts om t-fordelingen. Fordelingen...

- er klokkeformet
- har centrum i 0
- er symmetrisk omkring 0
- har én parameter f, der kaldes "antal frihedsgrader"
- minder meget om standardnormalfordelingen $N(0, 1)$
  (begge fordelinger er symmetriske, klokkeformede og har centrum i 0)
- har lidt tungere haler end standardnormalfordelingen
- minder mere og mere om standardnormalfordelingen, desto større antallet af frihedsgrader f er
  (og for passende store værdier af f (> 30) vil t-fordelingen og standardnormalfordelingen for alle praktiske formål være ens)

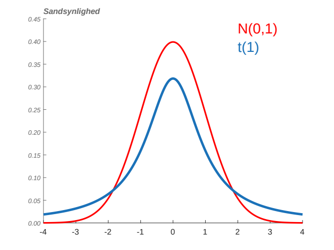
Sandsynligheder og fraktiler i t-fordelingen kan i øvrigt beregnes i JMP på samme måde som for normalfordelingen.

JMP-video [Help -> Sample Data -> Teaching Scripts -> Interactive Teaching Modules -> Distribution Calculator]

t-FORDELINGEN

Nedenstående animation viser, hvordan t-fordelingen (den blå kurve) minder mere og mere om standardnormalfordelingen (den røde kurve), i takt med at antallet af frihedsgrader f øges.

ANIMATION: t-fordelingen – ændring af antal frihedsgrader f



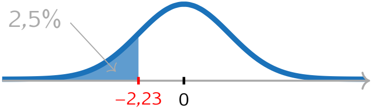
Start animation

Symmetri i t-fordelingen

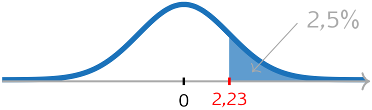
t-fordelingen er symmetrisk omkring 0, og det kan vi udnytte.

Ser vi eksempelvis på 2,5%-fraktilen i tilfældet $f = 10$, som vi ved beregning kan finde til $t_{2,5\%}(10) = -2{,}23$, så betyder det, at der per definition ligger...

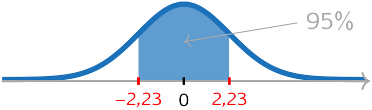
- 2,5% sandsynlighed *til venstre* for værdien −2,23



- og pga. symmetrien også 2,5% sandsynlighed *til højre* for værdien 2,23



- og dermed ligger der 95% sandsynlighed i intervallet [−2,23; 2,23]

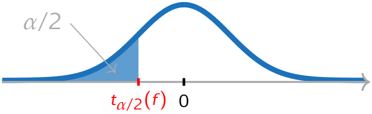


SYMMETRI I t-FORDELINGEN

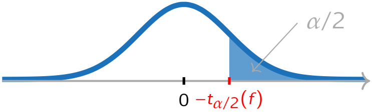
t-fordelingen er symmetrisk omkring 0 uanset antallet af frihedsgrader. Det leder til følgende generelle overvejelse (tilfældet på foregående side svarer til $\alpha = 5\%$).

For $0 < \alpha < 1$ ligger der per definition...

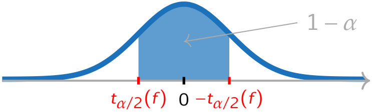
- $\alpha/2$ sandsynlighed *til venstre* for $\alpha/2$-fraktilen $t_{\alpha/2}(f)$



- og pga. symmetrien også $\alpha/2$ sandsynlighed *til højre* for værdien $-t_{\alpha/2}(f)$



- og dermed ligger der $1-\alpha$ sandsynlighed i intervallet $[t_{\alpha/2}(f);\ -t_{\alpha/2}(f)]$



Præcision af $\hat{\mu}$

Vi er nu endelig klar til at sige noget om præcisionen af vores middelværdiestimat $\hat{\mu}$.

Vi ved, at størrelsen $\frac{\hat{\mu}-\mu}{\hat{\sigma}/\sqrt{n}}$ er beskrevet ved en t-fordeling med $n-1$ frihedsgrader.

Vi har på de foregående sider set, hvordan vi i t-fordelingen kan konstruere et interval, som indeholder en vis mængde sandsynlighed.

Ved at bruge disse overvejelser på størrelsen $\frac{\hat{\mu}-\mu}{\hat{\sigma}/\sqrt{n}}$ kan vi konstruere et interval, som indeholder $\frac{\hat{\mu}-\mu}{\hat{\sigma}/\sqrt{n}}$ med en vis sandsynlighed.

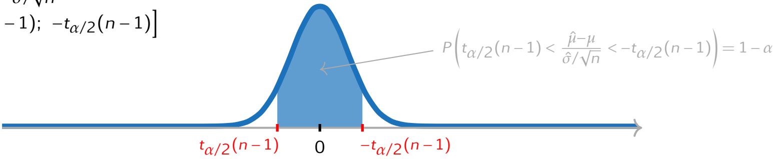
Ved at flytte lidt rundt på tingene (transformere) kan vi ændre det til et interval, som indeholder $\hat{\mu}$ med en vis sandsynlighed.

Konklusion: Vi får hermed konstrueret et ***interval som med en vis sandsynlighed indeholder vores middelværdiestimat*** $\hat{\mu}$. Intervallet siger dermed noget om, hvor meget eller lidt vi skal forvente, at middelværdiestimatet $\hat{\mu}$ vil variere, dvs. det siger noget om, hvor præcist vores estimat $\hat{\mu}$ er.

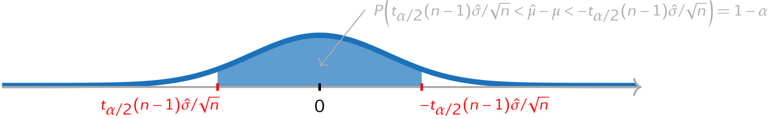
PRÆCISION AF $\hat{\mu}$

Mere formelt, så laver vi følgende overvejelse. For $0 < \alpha < 1$...

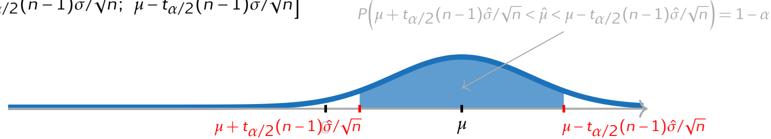
- ligger $\frac{\hat{\mu}-\mu}{\hat{\sigma}/\sqrt{n}}$ med sandsynlighed $1-\alpha$ i intervallet $\left[t_{\alpha/2}(n-1);\ -t_{\alpha/2}(n-1)\right]$



- og dermed ligger $\hat{\mu}-\mu$ med sandsynlighed $1-\alpha$ i intervallet $\left[t_{\alpha/2}(n-1)\hat{\sigma}/\sqrt{n};\ -t_{\alpha/2}(n-1)\hat{\sigma}/\sqrt{n}\right]$



- og dermed ligger $\hat{\mu}$ med sandsynlighed $1-\alpha$ i intervallet $\left[\mu+t_{\alpha/2}(n-1)\hat{\sigma}/\sqrt{n};\ \mu-t_{\alpha/2}(n-1)\hat{\sigma}/\sqrt{n}\right]$



Præcision af $\hat{\mu}$

Vi har nu fundet frem til det resultat, vi skal bruge til at sige noget om præcisionen af vores middelværdiestimat $\hat{\mu}$.

Udgangspunktet er, at vi gerne vil estimere middelværdien μ i en normalfordeling.

Resultatet siger, at med sandsynlighed $1-\alpha$ er forskellen mellem den ukendte størrelse μ og vores estimat $\hat{\mu}$ mindre end $-2t_{\alpha/2}(n-1)\hat{\sigma}/\sqrt{n}$ (= længden af intervallet $\left[\mu + t_{\alpha/2}(n-1)\hat{\sigma}/\sqrt{n};\ \mu - t_{\alpha/2}(n-1)\hat{\sigma}/\sqrt{n}\right]$).

Ved at vælge en værdi af α tæt på 0, bliver $1-\alpha$ tæt på 1, og dermed siger resultatet, at afstanden mellem den ukendte værdi μ og vores estimat $\hat{\mu}$ med stor sandsynlighed $(= 1-\alpha)$ ligger indenfor en afstand på $-2t_{\alpha/2}(n-1)\hat{\sigma}/\sqrt{n}$ af hinanden.

Beregning

Når vi skal estimere en ukendt middelværdi μ i en normalfordeling $N(\mu, \sigma)$, kan vi bruge det nedenstående resultat til at sige noget om, hvor præcist vores estimat $\hat{\mu}$ af den ukendte middelværdi μ er.

Det nedenstående interval kaldes et **$1-\alpha$ konfidensinterval for μ**. Ofte anvendes $\alpha = 5\%$ ved beregning af intervallet (som dermed bliver et 95% $(= 1-\alpha)$ konfidensinterval).

Resultat [Konfidensinterval for μ]

Hvis $X_1, \ldots, X_n$ er indbyrdes uafhængige observationer af en normalfordelt variabel $N(\mu, \sigma)$, så vil den ukendte middelværdi μ ligge i intervallet

$$\left[\hat{\mu} + t_{\alpha/2}(n-1) \cdot \frac{\hat{\sigma}}{\sqrt{n}}; \quad \hat{\mu} - t_{\alpha/2}(n-1) \cdot \frac{\hat{\sigma}}{\sqrt{n}}\right]$$

***med sandsynlighed** $1-\alpha$, hvor $t_{\alpha/2}(n-1)$ er $\alpha/2$-fraktilen i $t(n-1)$-fordelingen.*

Bemærk:

For ethvert $0 < \alpha < 1$ er $t_{\alpha/2}(n-1) < 0$ og dermed $-t_{\alpha/2}(n-1) > 0$, således at intervallet ovenfor altid er veldefineret.

Beregning

Idéen bag at beregne et konfidensinterval for den ukendte middelværdi μ er, at...

- uanset hvordan vi beregner et estimat af (= gæt på) værdien af den ukendte størrelse μ, så vil estimatet variere lidt fra gang til gang alene på grund af tilfældig variation. Derfor giver det ikke mening kun at gætte på én bestemt værdi af μ (et såkaldt **punktestimat** *("point estimate")*).
- et estimat af den ukendte værdi μ ikke er meget værd, hvis ikke vi ved, hvor præcist det er.
- vi beregner derfor et helt interval (= konfidensintervallet) af plausible værdier for μ (et såkaldt **intervalestimat** *("interval estimate")*).
- værdierne i konfidensintervallet er de værdier, vi vil anse som fornuftige gæt på den sande værdi af μ på baggrund af vores observationer i datamaterialet.

Det beregnede interval kaldes et "konfidensinterval", fordi det udtrykker, hvor sikre *("confident")* vi er på, at intervallet indeholder den ukendte størrelse μ.

Værdien $1-\alpha$ kaldes intervallets **konfidensniveau** *("confidence level")* og udtrykker med hvor stor sandsynlighed, intervallet indeholder den ukendte størrelse μ.

BEREGNING

Eksempel: Ølsalg

Vi ser igen på prisen på Grøn Tuborg i Føtex og antager som hidtil, at prisen kan beskrives ved en normalfordeling $N(\mu, \sigma)$. Hvis vi bruger de første $n = 4$ ugers prisobservationer som grundlag for at beregne et estimat af $\hat{\mu}$, finder vi at $\hat{\mu} = 3{,}62$ og $\hat{\sigma} = 0{,}36$.

Summary Statistics

Mean	3,6227695
Std Dev	0,3611782

Det er naturligvis klart, at vores gæt på en forventet pris på Grøn Tuborg på $\hat{\mu} = 3{,}62$ kr. er behæftet med en vis usikkerhed, fordi det kun er baseret på 4 ugers observationer.

Vi kan derfor også beregne et konfidensinterval hørende til estimatet $\hat{\mu} = 3{,}62$. Hvis vi eksempelvis sætter $\alpha = 20\%$, finder vi at et 80%-konfidensinterval for μ er givet ved

$$\left[\hat{\mu} + t_{\alpha/2}(n-1) \cdot \frac{\hat{\sigma}}{\sqrt{n}}; \quad \hat{\mu} - t_{\alpha/2}(n-1) \cdot \frac{\hat{\sigma}}{\sqrt{n}}\right]$$

$$= \left[3{,}62 - 1{,}64 \cdot 0{,}36/\sqrt{4}; \quad 3{,}62 + 1{,}64 \cdot 0{,}36/\sqrt{4}\right] = [3{,}33; \quad 3{,}92]$$

Confidence Intervals

Parameter	Estimate	Lower CI	Upper CI	1-Alpha
Mean	3,62277	3,327011	3,918528	0,800
Std Dev	0,361178	0,250204	0,818345	0,800

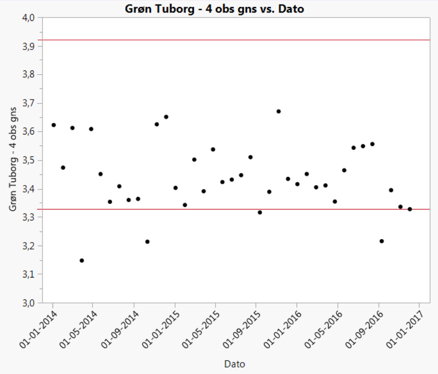
Med 80% sandsynlighed vil den sande (men for os ukendte) forventede pris på Grøn Tuborg dermed ligge mellem 3,33 kr. og 3,92 kr. Intervallets konfidensniveau på 80% betyder, at hvis vi gentagne gange beregner et estimat af μ baseret på 4 ugers observationer, så vil estimatet af μ i det lange løb ligge i intervallet fra 3,33 kr. til 3,92 kr. 80% af tiden (dvs. 4 ud af 5 gange).

Konfidensinterval for μ (praktisk anvendelse)

Beregning

Eksempel: Ølsalg (fortsat)

Tegner vi alle 4 ugers prisestimaterne op over tid og sammenligner med det fundne 80%-konfidensinterval (de røde linjer i figuren nedenfor) kan vi netop se, at intervallet indeholder de fleste – men ikke alle – af estimaterne.



I princippet burde intervallet indeholde ca. 80% af estimaterne, men vi kan se, at i praksis indeholder det faktisk lidt mere end 80% (helt præcist: 35 ud af 39 estimater). Udover tilfældig variation kan det skyldes, at normalfordelingen – som vi tidligere har set – ikke giver en 100% korrekt beskrivelse af prisen på Grøn Tuborg, og vores metode derfor ikke passer perfekt til data.

BEREGNING

Eksempel: Ølsalg (fortsat)

Eksemplet med kun at anvende 4 ugers observationer til estimation af μ er udelukkende medtaget for at vise, hvordan estimaterne ændrer sig i takt med, at der vælges data for nye 4 ugers perioder.

I praksis er der naturligvis ikke nogen grund til ikke at anvende hele datasættets 157 observationer til beregning af $\hat{\mu}$. I det tilfælde finder vi at $\hat{\mu} = 3{,}44$ og $\hat{\sigma} = 0{,}23$, og dermed at eksempelvis et 95%-konfidensinterval for μ er givet som

$$\left[\hat{\mu} + t_{\alpha/2}(n-1)\cdot\frac{\hat{\sigma}}{\sqrt{n}};\quad \hat{\mu} - t_{\alpha/2}(n-1)\cdot\frac{\hat{\sigma}}{\sqrt{n}}\right]$$

$$= \left[3{,}44 - 1{,}98\cdot 0{,}23/\sqrt{157};\quad 3{,}44 + 1{,}98\cdot 0{,}23/\sqrt{157}\right] = [3{,}40;\quad 3{,}47]$$

Summary Statistics

Mean	3,4381017
Std Dev	0,2299489
Std Err Mean	0,0183519
Upper 95% Mean	3,474352
Lower 95% Mean	3,4018514
N	157

Tilsvarende er eksempelvis et 99%-konfidensinterval givet som

$$\left[\hat{\mu} + t_{\alpha/2}(n-1)\cdot\frac{\hat{\sigma}}{\sqrt{n}};\quad \hat{\mu} - t_{\alpha/2}(n-1)\cdot\frac{\hat{\sigma}}{\sqrt{n}}\right]$$

$$= \left[3{,}44 - 2{,}61\cdot 0{,}23/\sqrt{157};\quad 3{,}44 + 2{,}61\cdot 0{,}23/\sqrt{157}\right] = [3{,}39;\quad 3{,}49]$$

Confidence Intervals

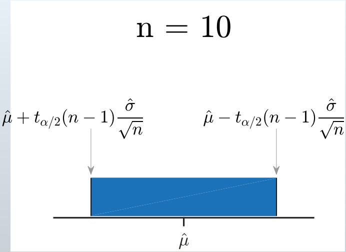
Parameter	Estimate	Lower CI	Upper CI	1-Alpha
Mean	3,438102	3,390245	3,485958	0,990
Std Dev	0,229949	0,200474	0,268688	0,990

JMP-video [Analyze -> Distribution]

Intuition

Nedenstående animation viser, hvordan konfidensintervallet for μ ændrer sig, i takt med at ***antallet af observationer*** n i datamaterialet øges.

ANIMATION: Konfidensinterval for μ - ændring af antal observationer n



Start animation

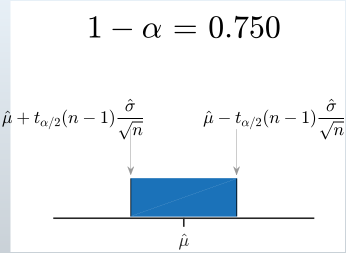
Jo ***flere*** observationer n, desto ***smallere*** bliver konfidensintervallet. Intuitionen er, at jo flere observationer (dvs. jo mere information) vi har til rådighed, desto mere præcist er vi i stand til at gætte på værdien af μ.

KONFIDENSINTERVAL FOR μ (PRAKTISK ANVENDELSE)

INTUITION

Nedenstående animation viser, hvordan konfidensintervallet for μ ændrer sig i takt med at ***konfidensniveauet*** $1-\alpha$ øges.

ANIMATION: Konfidensinterval for μ - ændring af konfidensniveau $1-\alpha$



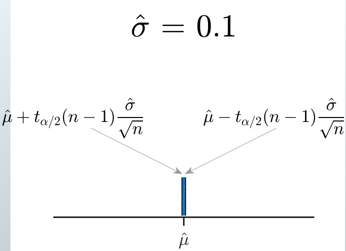
Start animation

Jo ***højere*** konfidensniveau $1-\alpha$, desto ***bredere*** bliver konfidensintervallet. Intuitionen er, at jo mere sikker vi vil være på, at intervallet indeholder den sande værdi μ, desto bredere er vi nødt til at gøre intervallet.

Intuition

Nedenstående animation viser, hvordan konfidensintervallet for μ ændrer sig i takt med at ***standardafvigelsen*** σ øges.

ANIMATION: Konfidensinterval for μ - ændring af standardafvigelse σ



Start animation

Jo ***højere*** standardafvigelse σ, desto ***bredere*** bliver konfidensintervallet. Intuitionen er, at jo mere usikkerhed, der er omkring hver enkelt observation i datamaterialet, desto mindre præcist er vi i stand til at gætte på værdien af μ.

Intuition

Eksempel: Ølsalg

Vi ser igen på prisen på Grøn Tuborg i Føtex beskrevet ved en normalfordeling $N(\mu, \sigma)$ med estimerede parametre $\hat{\mu} = 3{,}44$ og $\hat{\sigma} = 0{,}23$ baseret på et datamateriale med $n = 157$ observationer.

Vi har set, at den forventede pris med 95% sandsynlighed vil ligge i intervallet $[3{,}40;\ 3{,}47]$.

Hvis datamaterialet indeholdt 300 observationer, så ville 95%-konfidensintervallet i stedet blive $[3{,}41;\ 3{,}47]$.

Summary Information	
Sample Average	3,44
Sample Standard Deviation	0,23
Sample Size	300

Confidence Interval

Confidence Level 0,95

Result	Value
t multiplier	1,96793
Standard Error of the Mean	0,01328
Lower Limit	3,41387
Upper Limit	3,46613

Vi får således ikke en væsentligt mere præcis idé om den forventede pris på Grøn Tuborg i Føtex, selv hvis vi næsten fordobler vores datamateriale (fra 157 til 300 observationer).

Hvis variationen fra uge til uge i prisen på Grøn Tuborg var dobbelt så stor ($\hat{\sigma} = 0{,}46$ i stedet for $\hat{\sigma} = 0{,}23$), så ville 95%-konfidensintervallet i stedet blive $[3{,}37;\ 3{,}51]$.

Summary Information	
Sample Average	3,44
Sample Standard Deviation	0,46
Sample Size	157

Confidence Interval

Confidence Level 0,95

Result	Value
t multiplier	1,97529
Standard Error of the Mean	0,03671
Lower Limit	3,36748
Upper Limit	3,51252

Vi får således et dobbelt så bredt interval, dvs. usikkerheden om den forventede pris fordobles, på grund af den dobbelt så store variation i de ugentlige priser.

JMP-video [Help -> Sample Data -> Calculators -> Confidence Interval for One Mean]

Stikprøvestørrelse

Det er i mange sammenhænge værd at overveje, hvor stor en stikprøve man har behov for, for at opnå en given præcision (= bredde af konfidensintervallet) i sin analyse.

Hvis det eksempelvis er omkostningsfyldt (i kr. eller tid) at indsamle datamaterialet, er det relevant at vide, hvor stort et datamateriale det er nødvendigt at indsamle, allerede inden man begynder dataindsamlingen.

Spørgsmålet er derfor:
Hvis vi ønsker en given bredde af vores konfidensinterval for μ, hvor mange observationer n har vi så behov for at indsamle (dvs. hvor stor skal n være)?

Antallet af observationer n kaldes også for **stikprøvestørrelsen** *("sample size")*.

STIKPRØVESTØRRELSE

Resultat [Stikprøvestørrelse for $\hat{\mu}$]

Lad $X_1,\ldots,X_n$ være indbyrdes uafhængige observationer af en variabel, der er normalfordelt $N(\mu,\sigma)$.

Hvis et $1-\alpha$ konfidensinterval for μ maksimalt skal have en bredde på $2m$ kræver det mindst

$$n = \frac{z_{\alpha/2}^2 \hat{\sigma}^2}{m^2}$$

observationer, hvor $z_{\alpha/2}$ er $\alpha/2$-fraktilen i standardnormalfordelingen $N(0,1)$.

BEMÆRK:

- Der er en udfordring ved at anvende ovenstående resultat til at bestemme den nødvendige stikprøvestørrelse n forud for indsamlingen af data
- Man har behov for et gæt $\hat{\sigma}$ på værdien af standardafvigelsen σ for at kunne bestemme n (og det kender vi normalt først *efter*, vi har lavet vores dataindsamling)
- Et gæt $\hat{\sigma}$ kan man i stedet finde ved at se på data fra tidligere undersøgelser, eller ved først at lave en mini-dataindsamling og beregne $\hat{\sigma}$ ud fra den, inden man gennemfører den reelle dataindsamling

Stikprøvestørrelse

Forklaring af resultatet:

- Bredden af konfidensintervallet er

$$\hat{\mu} - t_{\alpha/2}(n-1)\frac{\hat{\sigma}}{\sqrt{n}} - \left(\hat{\mu} + t_{\alpha/2}(n-1)\frac{\hat{\sigma}}{\sqrt{n}}\right) = -2t_{\alpha/2}(n-1)\frac{\hat{\sigma}}{\sqrt{n}}$$

- For $n > 30$ kan vi approksimere t-fordelingen med standardnormalfordelingen $N(0,1)$, og dermed i stedet beregne bredden af konfidensintervallet v.hj.a. en fraktil i $N(0,1)$-fordelingen

$$-2t_{\alpha/2}(n-1)\frac{\hat{\sigma}}{\sqrt{n}} \approx -2z_{\alpha/2}\frac{\hat{\sigma}}{\sqrt{n}}$$

- Ved at sætte ovenstående lig den ønskede intervalbredde på $2m$

$$-2z_{\alpha/2}\frac{\hat{\sigma}}{\sqrt{n}} = 2m$$

og dernæst isolere n, finder vi at

$$n = \frac{z^2_{\alpha/2}\hat{\sigma}^2}{m^2}$$

KONFIDENSINTERVAL FOR μ (PRAKTISK ANVENDELSE)

STIKPRØVESTØRRELSE

Eksempel: Ølsalg

Vi ser igen på prisen på Grøn Tuborg i Føtex beskrevet ved en normalfordeling $N(\mu, \sigma)$. Vi er interesserede i at bestemme, hvor mange ugers data n vi har behov for, for at opnå en acceptabel præcision i vores estimat af den forventede pris μ på Grøn Tuborg.

Hvis vi antager, at vi har et gæt $\hat{\sigma}$ på værdien af standardafvigelsen, f.eks. ud fra tidligere lignende undersøgelser, ud fra generelt branchekendskab el.lign., så kan vi herefter fastlægge antal observationer n. Vi anvender her værdien $\hat{\sigma} = 0,23$, svarende til værdien fundet i datamaterialet.

Hvis vi ønsker at kunne bestemme den forventede pris ved et konfidensniveau på $1 - \alpha = 95\%$ med en usikkerhed på $\pm$ 0,10 kr. (dvs. $m = 0,10$), så kræver det, at vi har ca.

$$n = \frac{(-1,96)^2 0,23^2}{0,10^2} = 20,3 \approx 21 \text{ (rundet op)}$$

ugers prisobservationer til rådighed.

Sample Size for Estimation of the Mean

Inputs

Confidence Level	0,95
Population Std. Dev. (Planning Value)	0,23
Desired Margin of Error (C.I. 1/2-Width)	0,1

Results

Calculated Value	20,3213
Calculated Value (rounded up)	21

Med andre ord, hvis vi med 95% sandsynlighed vil kunne fastlægge den forventede pris på Grøn Tuborg i Føtex med en fejl på højst 0,10 kr., så kræver det, at vi indsamler prisinformation for 21 uger.

JMP-video [Help -> Sample Data -> Calculators -> Sample Size for Confidence Intervals]

Fordeling af $\hat{p}$

Binomialfordelingen består af to parametre n og p, hvor n angiver antalsparameteren, som vi sædvanligvis antager er kendt. Der er derfor kun én ukendt parameter i fordelingen, nemlig p.

Når vi skal estimere (dvs. gætte på) en værdi af den ukendte parameter p i en binomialfordeling, er det ikke på forhånd klart, hvordan vi gør det bedst muligt.

Der er imidlertid tungtvejende teoretiske argumenter for, at den bedste måde at estimere en værdi for p er som angivet på næste side.

Resultat [Fordeling af $\hat{p}$]

Lad $X_1, \ldots, X_n$ være indbyrdes uafhængige observationer af en variabel med to mulige udfald: 1 og 0, og lad p betegne sandsynligheden for at få udfaldet 1. Vi estimerer binomialfordelingens parameter p ved

$$\hat{p} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

Estimatet af p bliver omtrent normalfordelt

$$\hat{p} \overset{a}{\sim} N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$$

når både $n\hat{p} > 15$ og $n(1-\hat{p}) > 15$.

BEMÆRK:

- Symbolet p betegner en ukendt teoretisk værdi, som vi er interesseret i at estimere. Symbolet $\hat{p}$ betegner vores gæt på den ukendte størrelse p.
- Notationen $\overset{a}{\sim}$ angiver, at $\hat{p}$ sådan cirka (= **a**symptotisk) er beskrevet ved den angivne normalfordeling, når blot betingelserne $n\hat{p} > 15$ og $n(1-\hat{p}) > 15$ er opfyldt.

Fordeling af $\hat{p}$

Forklaring af resultatet:

- Vi antager, at vi har et datamateriale med n observationer af en variabel med to mulige udfald: 1 og 0
- Den første observation i datamaterialet betegner vi med X_1, den næste med X_2 osv.
- Vi tænker på det, som om vi endnu ikke kender talværdierne af de enkelte observationer. Derfor skriver vi observationerne som store bogstaver (dvs. X_1, X_2 osv.). Det vi antager er, at alle observationer har ***den samme*** sandsynlighed p for at give udfaldet 1
- Resultatet fortæller, at vores estimat $\hat{p}$ af den ukendte sandsynlighed p sådan cirka kan beskrives ved en normalfordeling
  (dvs. at hvis vi udregner en masse forskellige estimater, så vil estimaterne sådan cirka kunne beskrives ved en normalfordeling)
- Denne normalfordeling har parametre p (middelværdi) og $\sqrt{\frac{p(1-p)}{n}}$ (standardafvigelse)
- Ideen er at udnytte det, til at sige noget om størrelsen af usikkerheden på vores estimat $\hat{p}$

Eksempel: Skat

Fordelingen af de indkomne svar på spørgsmålet *"Er topskatten for høj?"* kan med en vis rimelighed beskrives ved en binomialfordeling $Bin(n, p)$. Der er registreret svar fra $n = 975$ personer, og vi lader her 1 repræsentere svarmuligheden "Ja" og 0 repræsentere svarmuligheden "Nej".

Hvis vi, for eksemplets skyld, vælger at estimere sandsynligheden p ud fra $n = 10$ respondenters svar, så får vi en lang række forskellige estimater afhængig af præcis hvilke 10 respondenters svar, vi vælger.

Tegner vi et histogram baseret på alle de forskellige estimater, får vi tegnet ***fordelingen af estimatet af*** p baseret på 10 observationer:

Det fremgår af histogrammet, at fordelingen af et estimat baseret på 10 observationer med en vis rimelighed kan beskrives ved en normalfordeling (pga. klokkeformen).

JMP-video [Analyze -> Distribution]

TRANSFORMATION

Vi er interesseret i at finde ud af, hvor præcist vores estimat $\hat{p}$ af den ukendte sandsynlighed p i binomialfordelingen er.
(NB: Vi anvender ordene "sandsynlighed" og "andel" synonymt som betegnelse for p)

Vi ved, at estimatet $\hat{p}$ i sig selv sådan cirka kan beskrives ved hjælp af en normalfordeling $\hat{p} \overset{a}{\sim} N\left(p, \sqrt{p(1-p)/n}\right)$.

Det kan vi alternativt også skrive som $\dfrac{\hat{p}-p}{\sqrt{\hat{p}(1-\hat{p})/n}} \overset{a}{\sim} N(0,1)$
(pga. normalfordelingens transformationsegenskab).

På de følgende sider gennemgår vi kort, hvordan vi kan udnytte det, til at sige noget om præcisionen af estimatet $\hat{p}$.

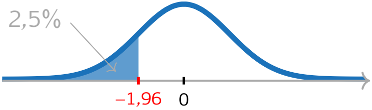
Overvejelserne følger samme fremgangsmåde som konstruktionen af konfidensintervallet for μ. Reelt er der tale om en gentagelse af de argumenter, som vi også har set på i note 4 i forbindelse med omtalen af normalfordelingens transformationsegenskab.

Symmetri i $N(0,1)$-fordelingen

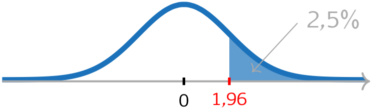
$N(0,1)$-fordelingen er symmetrisk omkring 0, og det kan vi udnytte.

Ser vi eksempelvis på 2,5%-fraktilen, som vi ved beregning kan finde til $z_{2,5\%} = -1{,}96$, så betyder det, at der per definition ligger...

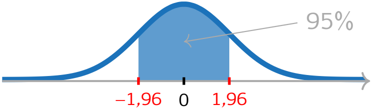
- 2,5% sandsynlighed *til venstre* for værdien $-1{,}96$



- og pga. symmetrien også 2,5% sandsynlighed *til højre* for værdien 1,96



- og dermed ligger der 95% sandsynlighed i intervallet $[-1{,}96;\ 1{,}96]$

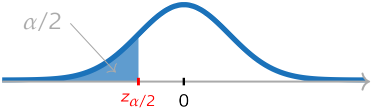


Symmetri i $N(0,1)$-fordelingen

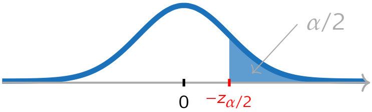
Det leder til følgende generelle overvejelse (tilfældet på foregående side svarer til $\alpha = 5\%$).

For $0 < \alpha < 1$ ligger der per definition...

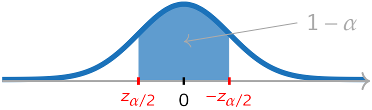
- $\alpha/2$ sandsynlighed *til venstre* for $\alpha/2$-fraktilen $z_{\alpha/2}$



- og pga. symmetrien også $\alpha/2$ sandsynlighed *til højre* for værdien $-z_{\alpha/2}$



- og dermed ligger der $1-\alpha$ sandsynlighed i intervallet $[z_{\alpha/2}; -z_{\alpha/2}]$



Præcision af $\hat{p}$

Vi er nu klar til at sige noget om præcisionen af vores estimat $\hat{p}$.

Vi ved, at størrelsen $\frac{\hat{p}-p}{\sqrt{\hat{p}(1-\hat{p})/n}}$ sådan cirka er beskrevet ved en $N(0,1)$-fordeling.

Vi har på de foregående sider set, hvordan vi i $N(0,1)$-fordelingen kan konstruere et interval, som indeholder en vis mængde sandsynlighed.

Ved at bruge disse overvejelser på størrelsen $\frac{\hat{p}-p}{\sqrt{\hat{p}(1-\hat{p})/n}}$ kan vi konstruere et interval, som indeholder $\frac{\hat{p}-p}{\sqrt{\hat{p}(1-\hat{p})n}}$ med en vis sandsynlighed.

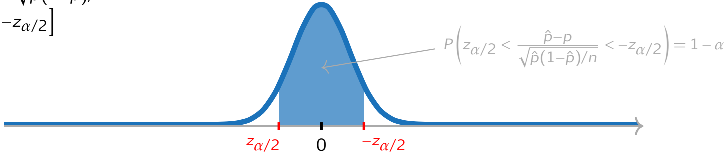
Ved at flytte lidt rundt på tingene (transformere) kan vi ændre det til et interval, som indeholder $\hat{p}$ med en vis sandsynlighed.

Konklusion: Vi får hermed konstrueret et ***interval som med en vis sandsynlighed indeholder vores estimat*** $\hat{p}$. Intervallet siger dermed noget om, hvor meget eller lidt vi skal forvente, at estimatet $\hat{p}$ vil variere, dvs. det siger noget om, hvor præcist vores estimat $\hat{p}$ er.

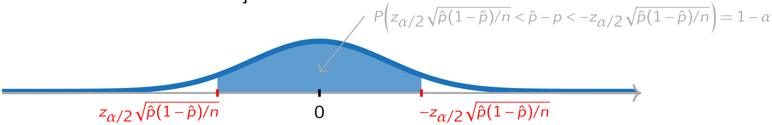
Præcision af $\hat{p}$

Mere formelt, så laver vi følgende overvejelse. For $0 < \alpha < 1$...

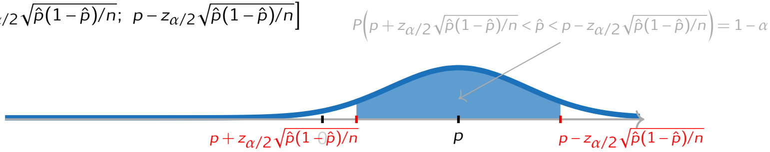
- ligger $\frac{\hat{p}-p}{\sqrt{\hat{p}(1-\hat{p})/n}}$ ca. med sandsynlighed $1-\alpha$ i intervallet $\left[z_{\alpha/2};\ -z_{\alpha/2}\right]$



- og dermed ligger $\hat{p}-p$ med sandsynlighed $1-\alpha$ i intervallet $\left[z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n};\ -z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}\right]$



- og dermed ligger $\hat{p}$ med sandsynlighed $1-\alpha$ i intervallet $\left[p+z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n};\ p-z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}\right]$



Præcision af $\hat{p}$

Vi har nu fundet frem til det resultat, vi skal bruge til at sige noget om præcisionen af vores estimat $\hat{p}$.

Udgangspunktet er, at vi gerne vil estimere sandsynligheden p i en binomialfordeling.

Resultatet siger, at med sandsynlighed $1-\alpha$ er forskellen mellem den ukendte størrelse p og vores estimat $\hat{p}$ mindre end $-2z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}$ (= længden af intervallet $\left[p + z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n};\ p - z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}\right]$).

Ved at vælge en værdi af α tæt på 0, bliver $1-\alpha$ tæt på 1, og dermed siger resultatet, at afstanden mellem den ukendte værdi p og vores estimat $\hat{p}$ med stor sandsynlighed $(= 1-\alpha)$ ligger indenfor en afstand på $-2z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}$ af hinanden.

BEREGNING

Når vi skal estimere en ukendt sandsynlighed p i en binomialfordeling $Bin(n,p)$, kan vi bruge det nedenstående resultat til at sige noget om, hvor præcist vores estimat $\hat{p}$ af den ukendte sandsynlighed p er.

Det nedenstående interval kaldes et **$1-\alpha$ konfidensinterval for p**. Ofte anvendes $\alpha = 5\%$ ved beregning af intervallet (som dermed bliver et 95% $(= 1-\alpha)$ konfidensinterval).

Resultat [Konfidensinterval for p]

Hvis $X_1,\ldots,X_n$ er indbyrdes uafhængige observationer af en variabel med to mulige udfald: 1 og 0, så vil den ukendte sandsynlighed p for at få udfaldet 1 ligge i intervallet

$$\left[\hat{p} + z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}};\quad \hat{p} - z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]$$

*ca. **med sandsynlighed** $1-\alpha$, hvor $z_{\alpha/2}$ er $\alpha/2$-fraktilen i standardnormalfordelingen $N(0,1)$.*

BEMÆRK:
For ethvert $0 < \alpha < 1$ er $z_{\alpha/2} < 0$ og dermed $-z_{\alpha/2} > 0$, således at intervallet ovenfor altid er veldefineret.

Beregning

Idéen bag at beregne et konfidensinterval for den ukendte sandsynlighed p er, at...

- uanset hvordan vi beregner et estimat af (= gæt på) værdien af den ukendte størrelse p, så vil estimatet variere lidt fra gang til gang alene på grund af tilfældig variation. Derfor giver det ikke mening kun at gætte på én bestemt værdi af p.
- et estimat af den ukendte værdi p ikke er meget værd, hvis ikke vi ved, hvor præcist det er.
- vi beregner derfor et helt interval (= konfidensintervallet) af plausible værdier for p.
- værdierne i konfidensintervallet er de værdier, vi vil anse som fornuftige gæt på den sande værdi af p på baggrund af vores observationer i datamaterialet.

BEREGNING

Fordeling af $\hat{\mu}$
Konf.int. μ (teori)
Konf.int. μ (praksis)
Fordeling af $\hat{p}$
Konf.int. p (teori)
Konf.int. p (praksis)
Beregning
Intuition
Stikprøvestørrelse
OPSUMMERING

Eksempel: Skat

Vi ser igen på svarene på spørgmålet *"Er topskatten for høj?"* og antager som hidtil, at de indkomne svar kan beskrives ved en binomialfordeling $Bin(n,p)$. Hvis vi bruger de første $n = 10$ personers svar som grundlag for at beregne et estimat af $\hat{p}$, finder vi at $\hat{p} = 0{,}5$.

Frequencies

Level	Count	Prob
Ja	5	0,50000
Nej	5	0,50000
Total	10	1,00000

Det er naturligvis klart, at vores gæt på $\hat{p} = 0{,}5$ som andelen af den voksne danske befolkning, der mener at topskatten er for høj, er behæftet med en vis usikkerhed, fordi det kun er baseret på svar fra 10 personer.

Vi kan derfor også beregne et konfidensinterval hørende til estimatet $\hat{p} = 0{,}5$. Hvis vi eksempelvis sætter $\alpha = 10\%$, finder vi at et 90%-konfidensinterval for p er givet ved

$$\left[\hat{p} + z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}};\quad \hat{p} - z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]$$

$$= \left[0{,}5 - 1{,}64 \cdot \sqrt{\frac{0{,}5 \cdot (1-0{,}5)}{10}};\quad 0{,}5 + 1{,}64 \cdot \sqrt{\frac{0{,}5 \cdot (1-0{,}5)}{10}}\right] = [0{,}27;\quad 0{,}73]$$

Confidence Intervals

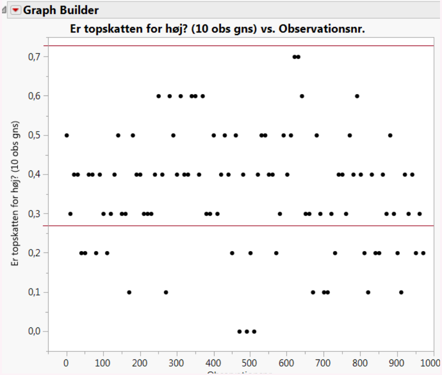
Level	Count	Prob	Lower CI	Upper CI	1-Alpha
Ja	5	0,50000	0,269272	0,730728	0,900
Nej	5	0,50000	0,269272	0,730728	0,900
Total	10				

Med 90% sandsynlighed vil den sande (men for os ukendte) andel af befolkningen, som mener topskatten er for høj, ligge mellem 0,27 og 0,73. Konfidensniveauet på 90% betyder, at hvis vi gentagne gange beregner et estimat af p baseret på 10 personers svar, så vil estimatet af p i det lange løb ligge i intervallet fra 0,27 til 0,73 90% af tiden (dvs. 9 ud af 10 gange).

BEREGNING

Eksempel: Skat (fortsat)

Tegner vi alle 10 personers estimaterne op over tid og sammenligner med det fundne 90%-konfidensinterval (de røde linjer i figuren nedenfor) kan vi netop se, at intervallet indeholder de fleste – men ikke alle – af estimaterne.



I princippet burde intervallet indeholde ca. 90% af estimaterne, men vi kan se, at i praksis indeholder det noget færre end 90% (helt præcist: 74 ud af 98 estimater). Udover tilfældig variation skyldes det, at betingelserne $n\hat{p} > 15$ og $n(1-\hat{p}) > 15$ ikke er opfyldt (idet $n\hat{p} = 10 \cdot 0{,}5 = 5$ og $n(1-\hat{p}) = 10 \cdot (1-0{,}5) = 5$), og vores metode derfor ikke fungerer, når vi kun baserer vores gæt på $n = 10$ observationer (NB: hvis n var større ville betingelserne formentlig være opfyldt).

Konfidensinterval for p (praktisk anvendelse)
Beregning

Eksempel: Skat (fortsat)

Eksemplet med kun at anvende svar fra 10 personer til estimation af p er udelukkende medtaget for at vise, hvordan estimaterne ændrer sig i takt med, at der vælges data for nye grupper af 10 personer.

I praksis er der naturligvis ikke nogen grund til ikke at anvende hele datasættets 975 observationer til beregning af $\hat{p}$. I det tilfælde finder vi at $\hat{p} = 346/975 = 0{,}355$, og dermed at eksempelvis et 95%-konfidensinterval for p er givet som

$$\left[\hat{p} + z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}; \quad \hat{p} - z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]$$

$$= \left[0{,}355 - 1{,}96 \cdot \sqrt{\frac{0{,}355(1-0{,}355)}{975}}; \quad 0{,}35 + 1{,}96 \cdot \sqrt{\frac{0{,}355(1-0{,}355)}{975}}\right] = [0{,}325; \quad 0{,}385]$$

Confidence Intervals

Level	Count	Prob	Lower CI	Upper CI	1-Alpha
Ja	346	0,35487	0,325462	0,385421	0,950
Nej	629	0,64513	0,614579	0,674538	0,950
Total	975				

Tilsvarende er eksempelvis et 99%-konfidensinterval givet som

$$\left[\hat{p} + z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}; \quad \hat{p} - z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]$$

$$= \left[0{,}355 - 2{,}58 \cdot \sqrt{\frac{0{,}355(1-0{,}355)}{975}}; \quad 0{,}355 + 2{,}58 \cdot \sqrt{\frac{0{,}355(1-0{,}355)}{975}}\right] = [0{,}317; \quad 0{,}395]$$

Confidence Intervals

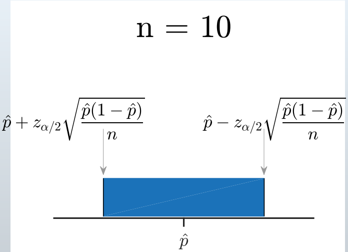
Level	Count	Prob	Lower CI	Upper CI	1-Alpha
Ja	346	0,35487	0,316504	0,395202	0,990
Nej	629	0,64513	0,604798	0,683496	0,990
Total	975				

JMP-video [Analyze -> Distribution]

INTUITION

Nedenstående animation viser, hvordan konfidensintervallet for p ændrer sig, i takt med at antallet af observationer n i datamaterialet øges.

ANIMATION: Konfidensinterval for p - ændring af antal observationer n



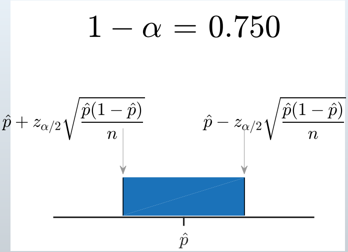
Start animation

Jo ***flere*** observationer n, desto ***smallere*** bliver konfidensintervallet. Intuitionen er, at jo flere observationer (dvs. jo mere information) vi har til rådighed, desto mere præcist er vi i stand til at gætte på værdien af p.

INTUITION

Nedenstående animation viser, hvordan konfidensintervallet for p ændrer sig i takt med at konfidensniveauet $1-\alpha$ øges.

ANIMATION: Konfidensinterval for p - ændring af konfidensniveau $1-\alpha$



Start animation

Jo ***højere*** konfidensniveau $1-\alpha$, desto ***bredere*** bliver konfidensintervallet. Intuitionen er, at jo mere sikker vi vil være på, at intervallet indeholder den sande værdi p, desto bredere er vi nødt til at gøre intervallet.

INTUITION

Eksempel: Skat

Vi ser igen på svarene på spørgsmålene vedr. topskat, der er beskrevet ved en binomialfordeling $Bin(n, p)$ med estimeret sandsynlighedsparameter $\hat{p} = 346/975 = 0{,}355$ baseret på et datamateriale med $n = 975$ observationer.

Vi har set, at den forventede andel, som mener topskatten er for høj, med 95% sandsynlighed vil ligge i intervallet $\big[0{,}325;\ 0{,}385\big]$.

Hvis datamaterialet i stedet var dobbelt så stort $(= 2 \cdot 975 = 1.950$ observationer), og hvor ligeså mange som hidtil mener, at topskatten er for høj $(= 2 \cdot 346 = 692)$, så ville 95%-konfidensintervallet i stedet blive $[0{,}334;\ 0{,}376]$.

Vi får således ikke en væsentligt mere præcis idé om den forventede andel af befolkningen, der mener topskatten er for høj, selv hvis vi fordobler vores datamateriale.

Number of Successes 692
Sample Size 1950

Confidence Interval

Confidence Level 0,95

Result	Value
Estimated Proportion	0,35487
Standard Error of Proportion	0,02124
Lower Limit	0,33363
Upper Limit	0,37611

▶ JMP-video [Help -> Sample Data -> Calculators -> Confidence Interval for One Proportion]

Stikprøvestørrelse

Det er i mange sammenhænge værd at overveje, hvor stor en stikprøve man har behov for, for at opnå en given præcision (= bredde af konfidensintervallet) i sin analyse.

Hvis det eksempelvis er omkostningsfyldt (i kr. eller tid) at indsamle datamaterialet, er det relevant at vide, hvor stort et datamateriale det er nødvendigt at indsamle, allerede inden man begynder dataindsamlingen.

Spørgsmålet er derfor:
Hvis vi ønsker en given bredde af vores konfidensinterval for p, hvor mange observationer n har vi så behov for at indsamle (dvs. hvor stor skal n være)?

Antallet af observationer n kaldes også for stikprøvestørrelsen.

Stikprøvestørrelse

Resultat [Stikprøvestørrelse for $\hat{p}$]

Lad $X_1, \ldots, X_n$ være indbyrdes uafhængige observationer af en variabel med to mulige udfald: 1 og 0. Lad p betegne den ukendte sandsynlighed for at få udfaldet 1.

Hvis et $1-\alpha$ konfidensinterval for p maksimalt skal have en bredde på $2m$ kræver det mindst

$$n = \frac{z_{\alpha/2}^2 \hat{p}(1-\hat{p})}{m^2}$$

observationer, hvor $z_{\alpha/2}$ er $\alpha/2$-fraktilen i standardnormalfordelingen $N(0,1)$.

Bemærk:

- Der er en udfordring ved at anvende ovenstående resultat til at bestemme den nødvendige stikprøvestørrelse n forud for indsamlingen af data
- Man har behov for et gæt $\hat{p}$ på værdien af p for at kunne bestemme n
  (og det kender vi normalt først *efter*, vi har lavet vores dataindsamling)
- Det konservative valg er at sætte $\hat{p} = 0{,}5$ i formlen ovenfor, omend det alt andet lige fører til en større værdi af n, end hvad der muligvis er nødvendigt
  (fordi $\hat{p} = 0{,}5$ giver en større værdi i udtrykket ovenfor end enhver anden værdi af $\hat{p}$)

Stikprøvestørrelse

Forklaring af resultatet:

- Bredden af konfidensintervallet er

$$\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} - \left(\hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) = -2z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

- Ved at sætte ovenstående lig den ønskede intervalbredde på $2m$

$$-2z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 2m$$

og dernæst isolere n, finder vi at

$$n = \frac{z_{\alpha/2}^2\hat{p}(1-\hat{p})}{m^2}$$

KONFIDENSINTERVAL FOR p (PRAKTISK ANVENDELSE)

STIKPRØVESTØRRELSE

Eksempel: Skat

Vi ser igen på svarene på spørgsmålet *"Er topskatten for høj?"*, der er beskrevet ved en binomialfordeling $Bin(n, p)$. Vi er interesserede i at bestemme, hvor mange svarpersoner vi har behov for, for at opnå en acceptabel præcision i vores estimat af andelen p af den voksne befolkning, som mener at topskatten er for høj.

Hvis vi antager, at vi har et gæt $\hat{p}$ på den ukendte sandsynlighed, f.eks. ud fra tidligere lignende undersøgelser, så kan vi herefter fastlægge antal observationer n. Vi anvender her værdien $\hat{p} = 0,355$, svarende til værdien fundet i datamaterialet.

Hvis vi ønsker at kunne bestemme andelen p ved et konfidensniveau på $1 - \alpha = 95\%$ med en usikkerhed på $\pm\, 0{,}01$ (dvs. $m = 0{,}01$), så kræver det, at vi har spørgeskemasvar fra ca.

$$n = \frac{(-1{,}96)^2 \cdot 0{,}355 \cdot (1 - 0{,}355)}{0{,}01^2} = 8.796 \text{ (rundet op)}$$

personer til rådighed.

Sample Size for Estimation of the Proportion

Inputs

Confidence Level	0,95
Expected Proportion (Planning Value)	0,355
Desired Margin of Error (C.I. 1/2-Width)	0,01

Results

Calculated Value	8795,98
Calculated Value (rounded up)	8796

Med andre ord, hvis vi med 95% sandsynlighed vil kunne fastlægge andelen af den voksne danske befolkning, som mener at topskatten er for høj, så kræver det, at vi indhenter svar fra 8.796 personer.

JMP-video [Help -> Sample Data -> Calculators -> Sample Size for Confidence Intervals]

Kort opsummering af dette notesæt:

Konfidensinterval for middelværdi μ

Et $1-\alpha$ konfidensinterval for middelværdien μ i en normalfordeling er givet ved

$$\left[\hat{\mu} + t_{\alpha/2}(n-1) \cdot \frac{\hat{\sigma}}{\sqrt{n}}; \quad \hat{\mu} - t_{\alpha/2}(n-1) \cdot \frac{\hat{\sigma}}{\sqrt{n}}\right]$$

hvor $t_{\alpha/2}(n-1)$ er $\alpha/2$-fraktilen i t-fordelingen med $n-1$ frihedsgrader.

Konfidensintervallet for μ bliver...

- smallere, når antallet af observationer n stiger
- bredere, når konfidensniveauet $1-\alpha$ stiger
- bredere, når standardafvigelsen σ stiger

Konfidensintervallet for μ har maksimalt en bredde på $2m$, hvis antallet af observationer n er større end

$$\frac{z_{\alpha/2}^2 \hat{\sigma}^2}{m^2}$$

OPSUMMERING

<u>Konfidensinterval for andel/sandsynlighed p</u>

Et $1-\alpha$ konfidensinterval for sandsynligheden p i en binomialfordeling er givet ved

$$\left[\hat{p} + z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}; \quad \hat{p} - z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]$$

hvor $z_{\alpha/2}$ er $\alpha/2$-fraktilen i $N(0,1)$-fordelingen.

Konfidensintervallet for p bliver...

- smallere, når antallet af observationer n stiger
- bredere, når konfidensniveauet $1-\alpha$ stiger

Konfidensintervallet for p har maksimalt en bredde på $2m$, hvis antallet af observationer n er større end

$$\frac{z_{\alpha/2}^2 \hat{p}(1-\hat{p})}{m^2}$$

INDEKS

Nye funktionaliteter i dette notesæt:

- *Analyze -> Distribution:*
  - Beregning af konfidensinterval for μ eller p (baseret på data)
- *Help -> Sample Data -> Calculators -> Confidence Interval for One Mean:*
  - Beregning af konfidensinterval for μ (baseret på input $n, \hat{\mu}, \hat{\sigma}, 1-\alpha$)
- *Help -> Sample Data -> Calculators -> Confidence Interval for One Proportion:*
  - Beregning af konfidensinterval for p (baseret på input $n, \hat{p}, 1-\alpha$)
- *Help -> Sample Data -> Calculators -> Sample Size for Confidence Intervals:*
  - Beregning af stikprøvestørrelse for μ eller p
- *Help -> Sample Data -> Teaching Scripts -> Interactive Teaching Modules -> Distribution Calculator:*
  - t-fordelingen – beregning af sandsynligheder og fraktiler

JMP-videoer:

s. 4: ▶ [Analyze -> Distribution ; Graph -> Graph Builder] ($\hat{\mu}$ varierer)
s. 6: ▶ [Analyze -> Distribution ; Graph -> Graph Builder] ($\hat{p}$ varierer)
s. 10: ▶ [Analyze -> Distribution] (fordeling af $\hat{\mu}$)
s. 14: ▶ [Help -> Sample Data -> Teaching Scripts -> Interactive Teaching Modules -> Distribution Calculator] (t-fordelingen)
s. 26: ▶ [Analyze -> Distribution] (konfidensinterval for μ)
s. 30: ▶ [Help -> Sample Data -> Calculators -> Confidence Interval for One Mean] (konfidensinterval for μ)
s. 34: ▶ [Help -> Sample Data -> Calculators -> Sample Size for Confidence Intervals] (stikprøvestørrelse for μ)
s. 39: ▶ [Analyze -> Distribution] (fordeling af $\hat{p}$)
s. 52: ▶ [Analyze -> Distribution] (konfidensinterval for p)
s. 55: ▶ [Help -> Sample Data -> Calculators -> Confidence Interval for One Proportion] (konfidensinterval for p)
s. 59: ▶ [Help -> Sample Data -> Calculators -> Sample Size for Confidence Intervals] (stikprøvestørrelse for p)