

02402 Statistik (Polyteknisk grundlag)

Uge 4: Konfidensintervaller

DTU Compute
Danmarks Tekniske Universitet
2800 Kgs. Lyngby

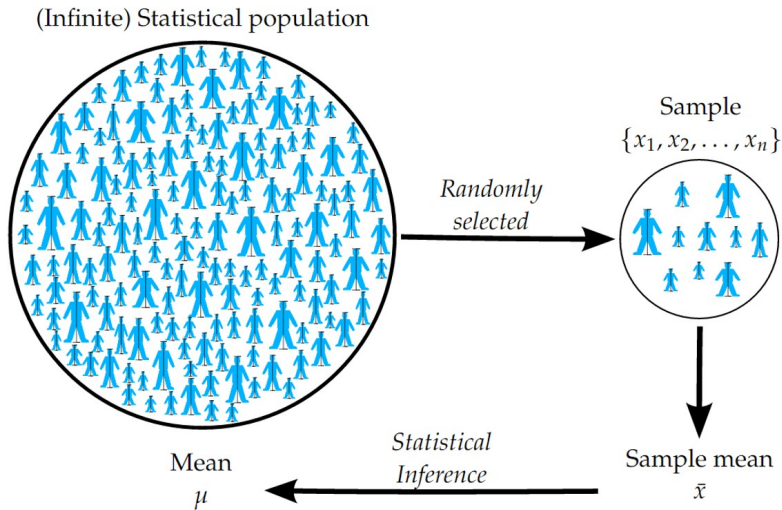
- 1 Intro til Statistisk Inferens
- 2 Når data er normalfordelt
 - Gennemsnittet ($\bar{X}$)
 - Varians og spredning
- 3 Når data ikke er normalfordelt
 - Central Limit Theorem (CLT)
- 4 Eksempler på eksamensopgaver

02402 Statistik (Polyteknisk grundlag)

Intro til Statistisk Inferens

DTU Compute
Danmarks Tekniske Universitet
2800 Kgs. Lyngby

Populationen og stikprøven



Statistisk inferens

Indtil nu har vi mest beskæftiget os med teoretiske fordelinger og beregnet sandsynligheder for diverse data-udfald *givet* at vi kender populationens fordeling (Sansynlighedsregning).

Vi har talt om stokastiske variable, der repræsenterer data *før de bliver målt*. Vi har også talt om stokastiske variable der afledes fra andre stokastiske variable (eller events), fx en lineær skalering af ét tilfældigt udfald ($Y = aX + b$) og lineære kombinationer af flere tilfældige udfald ($Y = aX + bZ + cW + \dots$).

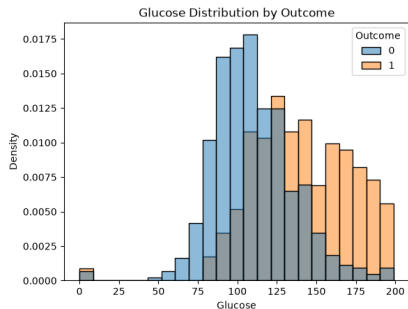
Nu vil vi gerne gå "baglæns":

Hvordan kan vi lære noget om populationen hvis vi kun kender stikprøven?

Hvad kan vi lære af vores data?

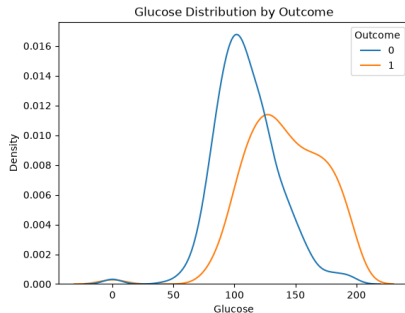
Eksempel med data

Her ses data for blodsukker-målinger lavet i forbindelse med en glucose tolerance test:



Blå: 500 raske personer

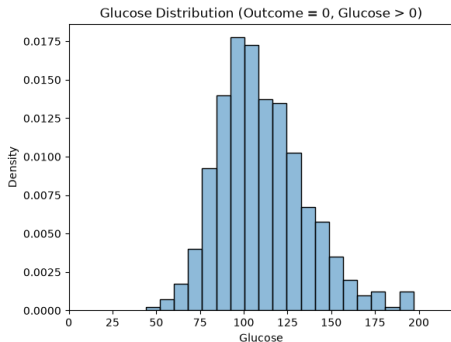
Gul: 268 personer med sukkersyge



Hvad kan vi lære fra disse data?

Vi starter med at tale om én stikprøve (og fortsætter med 2 stikprøver i uge 6).

Eksempel med data fra én stikprøve

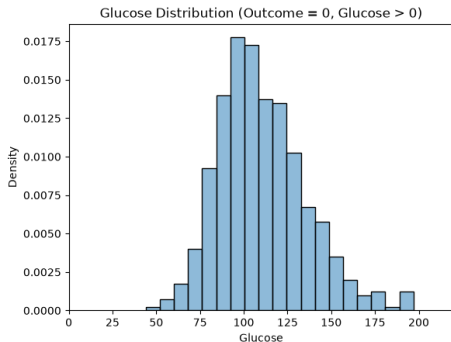


Fra stikprøven:

$$\bar{x} = 110.6$$

$$s = 24.8$$

Eksempel med data fra én stikprøve



Estimer:

$$\hat{\mu} = 110.6$$

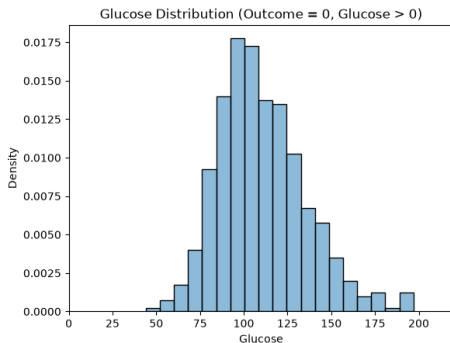
$$\hat{\sigma} = 24.8$$

Fra stikprøven:

$$\bar{x} = 110.6$$

$$s = 24.8$$

Eksempel med data fra én stikprøve



Estimer:

$$\hat{\mu} = 110.6$$

$$\hat{\sigma} = 24.8$$

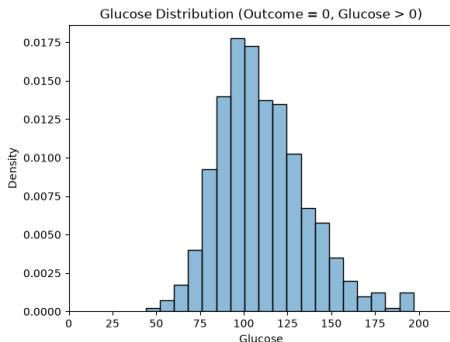
Fra stikprøven:

$$\bar{x} = 110.6$$

$$s = 24.8$$

Men hvor godt kan vi stole på disse estimater?

Eksempel med data fra én stikprøve



Estimater:

$$\hat{\mu} = 110.6$$

$$\hat{\sigma} = 24.8$$

Fra stikprøven:

$$\bar{x} = 110.6$$

$$s = 24.8$$

Men hvor godt kan vi stole på disse estimater?

NYT: Konfidensintervaller:

$$\hat{\mu} = 110.6 \quad [108; 112]$$

$$\hat{\sigma} = 24.8 \quad [23.3; 26.4]$$

Statistisk inferens: Læring fra data

- Vi *antager* at vores observationer (data) udgør en *tilfældig stikprøve* (en repræsentativ stikprøve).
- Populationen har nogle *parametre*, fx: μ og σ
- Vi ønsker nu at *estimere* parametrene ud fra data i stikprøven
 - $\hat{\mu} = \bar{x}$ er *estimatet* for μ (konkret udfaldsværdi)
 - $\bar{X}$ er *estimatoren* for μ (nu set som stokastisk variabel)
 - $\hat{\sigma}^2 = s^2$ er *estimatet* for σ^2 (konkret udfaldsværdi)
 - S^2 er *estimatoren* for σ^2 (nu set som stokastisk variabel)
- For at kvantificere præcisionen af vores estimater må vi *antage* en *model* for populationen, fx: $X \sim N(\mu, \sigma^2)$

En tilfældig stikprøve

||| Definition 3.12 Random sample

A random sample from an (infinite) population: A set of observations $X_1, \dots, X_n$ constitutes a random sample of size n from the infinite population $f(x)$ if:

1. Each X_i is a random variable whose distribution is given by $f(x)$
2. The n random variables are independent

Hvad betyder det?

- 1 Alle observationer skal komme fra den samme population
- 2 De må IKKE være afhængige af hinanden (fx hvis man havde udtaget hele familier i stedet for enkeltindivider)

En statistisk model

Før vi kan beregne konfidensintervaller, må vi antage en lidt mere konkret model.

Ofte vil vi antage at data følger en Normalfordeling.
Dvs vores **statistiske model** er:

$$X \sim N(\mu, \sigma^2)$$

En anden måde at skrive modellen er:

$$X_i = \mu + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2)$$

(dvs den enkelte observation beskrives som en gennemsnitlig værdi plus en "fejl" / "afvigelse" / "residual" beskrevet ved den stokastiske variabel ε)

Når vi har antaget en model kan vi beregne sandsynligheder for udfald af data ("forlæns"). Herefter vil vi fra de rigtige data forsøge at konkludere noget om den model (dvs den population) data kom fra (inferens / "baglæns").

02402 Statistik (Polyteknisk grundlag)

Når data er normalfordelt

- Gennemsnittet ($\bar{X}$)
- Varians og spredning

$\bar{X}$ og S^2 er stokastiske variable

I det følgende vil vi betragte en stikprøve, $\{X_1, X_2, \dots, X_n\}$, af **normalfordelt data**.

Den **enkelte observation** beskrives med en **stokastisk variabel**:
 $X_i \sim N(\mu, \sigma^2)$.

Stikprøvegennemsnittet: $\bar{X} = \frac{1}{n}(X_1 + \dots + X_n)$
 $\bar{X}$ er selv en **stokastisk variabel** (med sin egen fordeling).

Stikprøvevariansen: $S^2 = \frac{1}{n-1}((X_1 - \mu)^2 + \dots + (X_n - \mu)^2)$
 S^2 er selv en **stokastisk variabel** (med sin egen fordeling).

- Gå Python notebook
 "week4_python.ipynb" i VS Code



Visual Studio Code

02402 Statistik (Polyteknisk grundlag)

- Gennemsnittet ($\bar{X}$)
- Varians og spredning

Gennemsnittet er en lineær kombination

Stikprøvegennemsnittet er en lineær kombination af n tilfældige tal.

Den stokastiske variabel $\bar{X}$ er en lineær kombination af normalfordelte stokastiske variable:

$$\bar{X} = \frac{1}{n}X_1 + \frac{1}{n}X_2 + \dots + \frac{1}{n}X_n$$

En lineær kombination af normalfordelte stokastiske variable vil også selv være **normalfordelt**!

$$\bar{X} \sim N(\mu_{\bar{X}}, \sigma_{\bar{X}}^2)$$

OBS: bemærk og pas på de små detaljer i notationen

$E[\bar{X}]$ og $V[\bar{X}]$ følger af regneregler (2.56)

$$\bar{X} = \frac{1}{n}X_1 + \frac{1}{n}X_2 + \dots + \frac{1}{n}X_n$$

Hvis: $X_i \sim N(\mu, \sigma^2)$

Så ved vi fra regneregler:

$$\begin{aligned} E[\bar{X}] &= \mu_{\bar{X}} = \frac{1}{n}E[X_1] + \frac{1}{n}E[X_2] + \dots + \frac{1}{n}E[X_n] &= n \cdot \frac{1}{n}\mu &= \mu \\ V[\bar{X}] &= \sigma_{\bar{X}}^2 = \frac{1}{n^2}V[X_1] + \frac{1}{n^2}V[X_2] + \dots + \frac{1}{n^2}V[X_n] &= n \cdot \frac{1}{n^2}\sigma^2 &= \frac{\sigma^2}{n} \end{aligned}$$

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

Fordeling for stikprøvegennemsnittet af normalfordelte variable

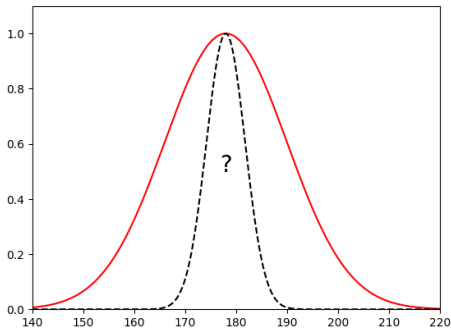
||| Theorem 3.3 The distribution of the mean of normal random variables

Assume that $X_1, \dots, X_n$ are independent and identically normally distributed random variables, $X_i \sim N(\mu, \sigma^2), i = 1, \dots, n$, then

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \sim N\left(\mu, \frac{\sigma^2}{n}\right). \quad (3-2)$$

(Teoretisk) fordeling for stikprøvegennemsnittet

Hvis data følger den røde normalfordeling, vil stikprøvegennemsnittet følge den sorte - **smallere** - normalfordeling.



Kahoot!
(x3)

Hvad er typiske værdier af $\bar{X}$ (stikprøvegennemsnit) i ovenstående eksempel?

Hvad er typiske værdier af X (enkelte observationer) i ovenstående eksempel?

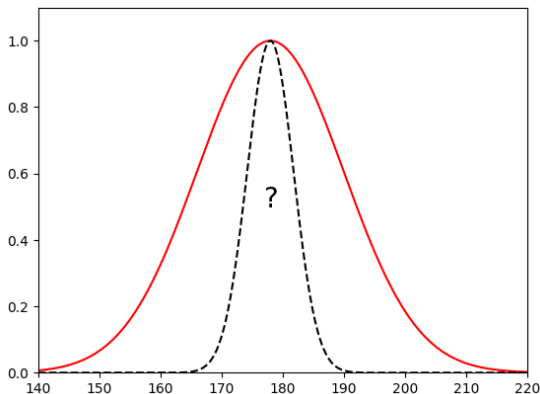
Konfidensinterval for gennemsnittet

Hvad kan vi forvente om værdien af $\bar{X}$?

I praksis har vi kun én stikprøve og derfor kun én observation af $\bar{X}$.

Men ud fra vores teori ved vi at $\bar{X}$ med 95% sandsynlighed vil ligge i intervallet:

$$\left[\mu - 1.96 \frac{\sigma}{\sqrt{n}}; \mu + 1.96 \frac{\sigma}{\sqrt{n}}\right]$$



Standardiseret version af de samme ting

||| Theorem 3.4 The distribution of the σ -standardized mean of normal random variables

Assume that $X_1, \dots, X_n$ are independent and identically normally distributed random variables, $X_i \sim N(\mu, \sigma^2)$ where $i = 1, \dots, n$, then

$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \sim N(0, 1^2). \quad (3-7)$$

That is, the standardized sample mean Z follows a standard normal distribution.

Sandsynligheden for at Z ligger i intervallet $[-1.96; 1.96]$ er 95%.

Disse tal (-1.96 og $+1.96$) er fraktiler i standardnormalfordelingen:

$$P(Z \leq -1.96) = 0.025$$

$$P(Z \leq +1.96) = 0.975$$

Hvad er "ekstreme værdier" af $|Z|$?

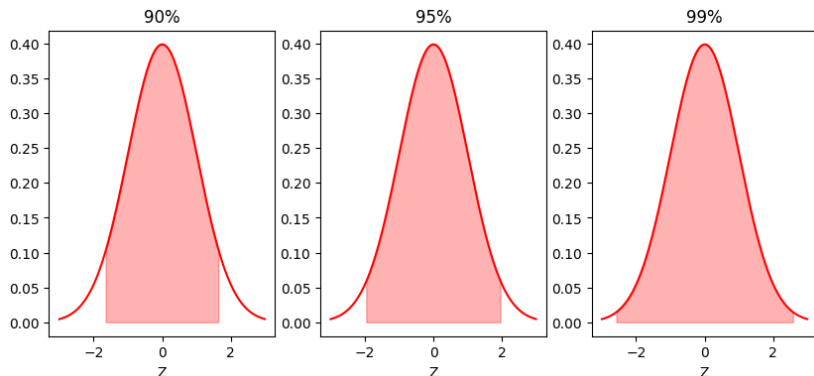
$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1^2)$$

Store værdier af $|Z|$ (også kaldet "**ekstreme værdier**"), svarer til situationer hvor $\bar{X}$ er meget langt fra μ - dvs. man har været "uheldig" med stikprøven i den forstand at estimatet ($\hat{\mu} = \bar{x}$) ligger langt fra den sande værdi μ .

I det følgende vil vi bygge på en antagelse om at man nok ikke har været "alt for uheldig", med den konkrete stikprøve. Præcis hvor uheldig(/heldig) kvantificeres med et "**signifiansniveau**".

Z-intervaller for forskellige signifikansniveau'er

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1^2)$$



De "ekstreme" tilfælde (vi kunne også sige "ekstremt uheldige" tilfælde) ligger udenfor det røde interval.

Interval for μ

Man skal vælge et ”**signifikansniveau**” α , svarende til det interval man vil fokusere på.

Herefter kan man ud fra intervalgrænser på Z *regne baglæns* for at få interval for μ .

Eksempel med interval der dækker 95% ($\alpha = 0.05$):

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} > z_{0.025} \quad \text{og} \quad Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < z_{0.975}$$

Omregn de to uligheder til et tilsvarende interval for μ :

(Bemærk at $z_{0.025} = -z_{0.975}$)

Praktisk problem i alt dette!

Nu har vi et interval for μ :-)

MEN: Der er et problem! σ indgår i udtrykkene.

Oplagt løsning:

Anvend estimeret s i stedet for σ i formlerne!

MEN:

Så bryder den givne teori faktisk sammen!

HELDIGVIS:

Findes der en udvidet teori, der kan klare det!

Mere anvendeligt resultat:

Vi betragter i stedet den stokastiske variabel: $T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$

I udtrykket for T indgår to stokastiske variable: $\bar{X}$ og S^2

||| Theorem 3.5 The distribution of the S -standardized mean of normal random variables

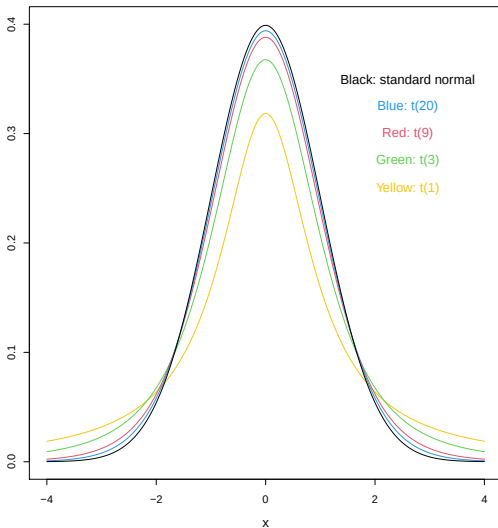
Assume that $X_1, \dots, X_n$ are independent and identically normally distributed random variables, where $X_i \sim N(\mu, \sigma^2)$ and $i = 1, \dots, n$, then

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t(n-1), \quad (3-8)$$

where $t(n-1)$ is the t -distribution with $n-1$ degrees of freedom.

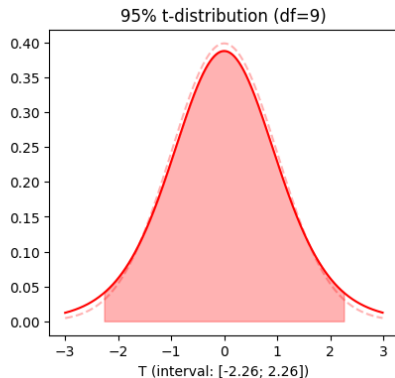
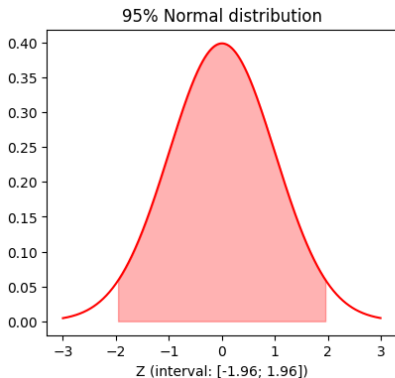
T følger en bestemt t -fordeling (afhænger af parametren $\nu = n-1$, se formelsamling)

t -fordelingen:



Jo, mindre antal af frihedsgrader (mindre $n - 1$), jo tungere haler.

t -fordelingen med 9 frihedsgrader og standardnormalfordelingen:



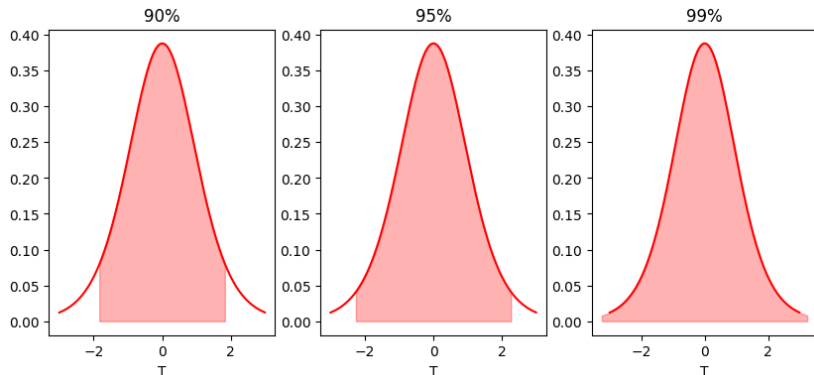
t -fordelingen ligner Normalfordelingen.

I t -fordelingen er sandsynlighedsmassen lidt bredere fordelt.

Bredden på t -fordelingen afhænger af antallet af frihedsgrader (df).

T-intervaller

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t(n-1)$$



Store værdier af $|T|$ ("ekstreme værdier"), svarer til situationer hvor man har været "uheldig" med stikprøven i den forstand at estimatet $\hat{\mu} = \bar{x}$ ligger langt fra den sande værdi μ og/eller s er meget lille.

Interval for μ

Igen vælges et ”**signifikans-niveau**” α , og vi *regner baglæns* for at få et interval for μ .

Eksempel med interval der dækker 95%:

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} > t_{0.025} \quad \text{og} \quad T = \frac{\bar{X} - \mu}{S/\sqrt{n}} < t_{0.975}$$

Omregn de to uligheder til et tilsvarende interval for μ :

(Bemærk at $t_{0.025} = -t_{0.975}$)

Konfidensinterval for μ

||| Method 3.9 The one sample confidence interval for μ

For a sample $x_1, \dots, x_n$ the $100(1 - \alpha)\%$ confidence interval is given by

$$\bar{x} \pm t_{1-\alpha/2} \cdot \frac{s}{\sqrt{n}}, \quad (3-10)$$

where $t_{1-\alpha/2}$ is the $(1 - \alpha/2)$ quantile from the t -distribution with $n - 1$ degrees of freedom.^a

Most commonly used is the 95%-confidence interval:

$$\bar{x} \pm t_{0.975} \cdot \frac{s}{\sqrt{n}}. \quad (3-11)$$

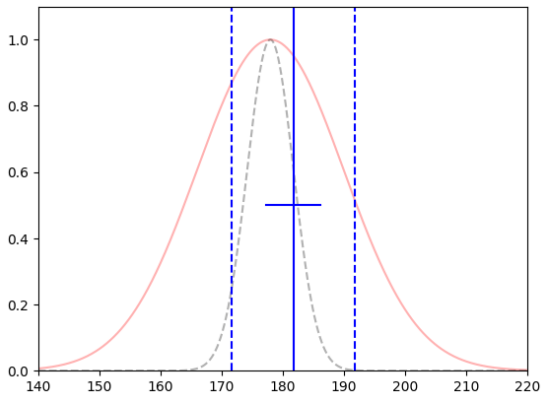
Nu kan vi udtale os om hvad vi har lært om populationen (specifikt μ) ud fra vores observerede data (et eksempel på statistisk inferens).

Visualisering af det enkelte tilfælde

Vi har nu estimeret $\bar{x}$ og s^2 fra vores stikprøve data.

Vi kender ikke de "rigtige" μ og σ^2 , vi kan angive et **konfidensinterval** for μ .

Sammen med et konfidensinterval bør man altid oplyse signifikansniveauet α .



'Repeated sampling' fortolkning

Hvis $\alpha = 0.05$:

$$P\left(\bar{X} - t_{0.975} \frac{S}{\sqrt{n}} < \mu < \bar{X} + t_{0.975} \frac{S}{\sqrt{n}}\right) = 0.95$$

I det lange løb fanger vi den sande værdi i 95% af tilfældene:

Konfidensintervallet vil variere i både bredde (s) og position ($\bar{x}$), hvis man gentager sit studie.

Kahoot!

(x3)

Signifikans-niveau'et: α

Signifikans-niveau " α ":

Et 95% konfidensinterval svarer til $\alpha = 0.05$.

Et 99% konfidensinterval svarer til $\alpha = 0.01$, osv.

Ofte vælges $\alpha = 0.05$.

Fraktiler i t -fordelingen skal findes med Python

Når vi skal beregne konfidensintervaller har vi brug for fraktiler fra t -fordelingen.

For at finde disse skal vi kende α (for 95% konfidensinterval er $\alpha = 0.05$) og vi skal kende antallet af frihedsgrader $\nu = n - 1$ (hvor n er stikprøvestørrelsen).

Vi skal da finde $(1 - \alpha/2)$ -fraktilen i en $t(n - 1)$ -fordeling (denne fraktil kaldes " $t_{1-\alpha/2}$ ").

Eksempel med $t_{0.975}$ (dvs $\alpha = 0.05$) og $n = 30$ (en stikprøve på 30 observationer):

```
stats.t.ppf(q=0.975, df=29)
```

Se også formelsamling af t -fordelinger.

Standard error of the mean, SEM

Hvor stor fejl ($\bar{X} - \mu$) ?

Vi har set at bredden på fordelingen for $\bar{X}$ er givet ved $\sigma/\sqrt{n}$.

Den siger dermed også noget om den "fejl" vi laver når vi estimerer μ med stikprøvegennemsnittet:

$$fejl = \bar{X} - \mu$$

$$fejl \sim N(0, \sigma^2/n)$$

$\sigma/\sqrt{n}$ er standardafvigelsen på "fejlen"

Standard Error of the Mean

Da vi ikke kender den sande σ estimerer vi $\frac{\sigma}{\sqrt{n}}$ med: $\frac{s}{\sqrt{n}}$. Sidstnævnte har sit eget navn, nemlig "**Standardfejlen på middelværdien**":

||| Definition 3.7 Standard Error of the mean

Given a sample $X_1, \dots, X_n$, the *Standard Error of the Mean* is defined as

$$\sigma_{\bar{x}} = \frac{S}{\sqrt{n}}. \quad (3-9)$$

It can also be read as the *Sampling Error* of the mean, and can be called the standard deviation of the *sampling distribution* of the mean.

Nogle gange ses notationen: $SE_{\bar{x}} = \frac{s}{\sqrt{n}}$, hvor SE står for Standard Error.

02402 Statistik (Polyteknisk grundlag)

- Gennemsnittet ($\bar{X}$)
- Varians og spredning

Python: (Empirisk) Fordeling af stikprøvevariansen

Nu har vi snakket meget om stikprøve**gennemsnittet** (μ).

Men hvad med stikprøve**variansen** (σ^2)?

- Gå Python notebook
"week4_python.ipynb" i VS Code
(genbesøg simulering af stikprøvevariens)



Visual Studio Code

Fordelingen for stikprøvevariansen

Vi betragter nu den stokastiske variabel: $\chi^2 = \frac{(n-1)S^2}{\sigma^2}$

I udtrykket for χ^2 indgår den stokastiske variabel S^2

||| Theorem 2.81

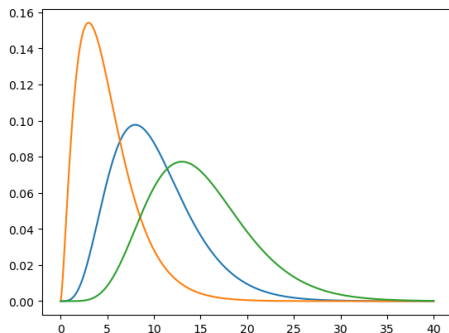
Given a sample of size n from the normal distributed random variables X_i with variance σ^2 , then the sample variance S^2 (viewed as random variable) can be transformed into

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2}, \quad (2-95)$$

which follows the χ^2 -distribution with degrees of freedom $\nu = n - 1$.

Bemærk notationen: Den stokastiske variabel " χ^2 " følger en χ^2 -fordeling.

χ^2 -fordelingen



$$\chi^2 = \frac{(n-1)S^2}{\sigma^2}$$

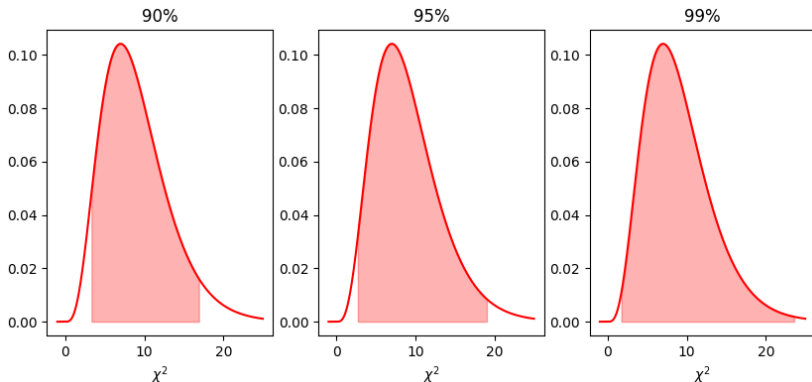
χ^2 -fordelingen er ikke symmetrisk.

Kan χ^2 blive negativ? Hvornår er χ^2 stor/lille? Hvornår er $\chi^2 = n - 1$?

Konfidensinterval for variansen

χ^2 -intervaller

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2}$$



Store/små værdier af χ^2 ("ekstreme værdier"), svarer til situationer hvor man har været "uheldig" med stikprøven i den forstand at estimatet $\hat{\sigma}^2 = S^2$ er meget større/mindre end den sande værdi σ^2 .

Interval for σ^2

Igen vælges et ”**signifikans-niveau**” α , og vi *regner baglæns* for at få et interval for σ^2 .

Eksempel med interval der dækker 95%:

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2} > \chi_{0.025}^2 \quad \text{og} \quad \frac{(n-1)S^2}{\sigma^2} > \chi_{0.975}^2$$

Omregn de to uligheder til et tilsvarende interval for σ^2 :

Konfidensinterval for σ^2 (og σ)

||| Method 3.19 Confidence interval for the variance/standard deviation

A $100(1 - \alpha)\%$ confidence interval for the variance σ^2 is

$$\left[\frac{(n-1)s^2}{\chi_{1-\alpha/2}^2}, \frac{(n-1)s^2}{\chi_{\alpha/2}^2} \right], \quad (3-18)$$

where the quantiles come from a χ^2 -distribution with $\nu = n - 1$ degrees of freedom.

A $100(1 - \alpha)\%$ confidence interval for the standard deviation σ is

$$\left[\sqrt{\frac{(n-1)s^2}{\chi_{1-\alpha/2}^2}}, \sqrt{\frac{(n-1)s^2}{\chi_{\alpha/2}^2}} \right]. \quad (3-19)$$

Fraktiler i χ^2 -fordelingen skal findes med Python

Når vi skal beregne konfidensintervaller for σ^2 har vi brug for fraktiler fra χ^2 -fordelingen.

For at finde disse skal vi kende α og antallet af frihedsgrader $\nu = n - 1$ (n er stikprøvestørrelsen).

$(1 - \alpha/2)$ -fraktilen i en $\chi^2(n - 1)$ -fordeling kaldes " $\chi^2_{1-\alpha/2}$ ".

$(\alpha/2)$ -fraktilen i en $\chi^2(n - 1)$ -fordeling kaldes " $\chi^2_{\alpha/2}$ ".

Eksempel med $\chi^2_{0.975}$ og $\chi^2_{0.025}$ (dvs $\alpha = 0.05$) og $n = 30$ (en stikprøve på 30 observationer):

```
stats.chi2.ppf(q=0.975, df=29)
```

```
stats.chi2.ppf(q=0.025, df=29)
```

Kahoot!

(x2)

Se også formelsamling.

Eksempel med konfidensinterval for s

En konditor skal bruge en masse ensartede jordbær til en kage. Han ønsker at købe jordbær fra den leverandør der kan sælge mest ensartede bær - dvs. han ønsker spredningen på jordbærenes diameter skal være så lille som mulig.

Som test køber han en bakke med 23 jordbær og måler disses diameter på det bredeste sted. Jordbærenes diameter har gennemsnit 3 cm og standard afvigelse 0.5 cm.

Konditoren ønsker nu at beregne et 95% konfidensinterval på stadardafvigelsen.

Det oplyses at:

```
stats.chi2.ppf(0.025, df=22) = 3.31396
```

```
stats.chi2.ppf(0.975, df=22) = 6.06471
```

Hvad er signifikansniveauet?

Beregn konfidensintervallet for s . Er det de rigtige fraktiler der er oplyst?

Hvad har vi antaget om fordelingen af jordbærenes diameter? Er denne antagelse nødvendig?

02402 Statistik (Polyteknisk grundlag)

Når data ikke er normalfordelt

- Central Limit Theorem (CLT)

Middelværdien og variansen følger af regneregler (2.56)

Stikprøvegennemsnittet er en lineær kombination af stokastiske variable:

$$\bar{X} = \frac{1}{n}X_1 + \frac{1}{n}X_2 + \dots + \frac{1}{n}X_n$$

Middelværdien af $\bar{X}$:

$$E[\bar{X}] = \frac{1}{n}E[X_1] + \frac{1}{n}E[X_2] + \dots + \frac{1}{n}E[X_n] = n\frac{1}{n}\mu = \mu$$

Variansen for $\bar{X}$:

$$V[\bar{X}] = \frac{1}{n^2}V[X_1] + \frac{1}{n^2}V[X_2] + \dots + \frac{1}{n^2}V[X_n] = n\frac{1}{n^2}\sigma^2 = \frac{\sigma^2}{n}$$

Fordelingen af $\bar{X}$:

Det ved vi ikke!

Python: CLT in action

Lad os prøve at simulere nogle forskellige situationer.

- Gå til dagens Python notebook i VS Code
 - "CLT in action"



Visual Studio Code

02402 Statistik (Polyteknisk grundlag)

- Central Limit Theorem (CLT)

Den centrale grænseværdisætning (CLT)

||| Theorem 3.14 Central Limit Theorem (CLT)

Let $\bar{X}$ be the sample mean of a random sample of size n taken from a population with mean μ and variance σ^2 , then

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}, \quad (3-12)$$

is a random variable which distribution function approaches that of the standard normal distribution, $N(0, 1^2)$, as $n \rightarrow \infty$. In other words, for large enough n , it holds approximately that

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1^2). \quad (3-13)$$

Konsekvens af den centrale grænseværdisætning:

Konfidensintervallet for μ gælder også for ikke-normale data:

Konfidensintervaller for gennemsnit kan beregnes baseret på t -fordelingen i stort set alle situationer, blot n er "stor nok".

Hvornår er n "stor nok"?

Faktisk svært at svare præcist på, MEN:

- **Tommelfingerregel:** $n \geq 30$
- Selv for mindre n kan formelen være (næsten)gyldig for ikke-normale data (især hvis data følger en nogenlunde symmetrisk fordeling).

OBS: CLT gælder for **gennemsnit** - ikke varians!

02402 Statistik (Polyteknisk grundlag)

Eksempler på eksamensopgaver

DTU Compute
Danmarks Tekniske Universitet
2800 Kgs. Lyngby

Eksempel: Eksamensopgave fra 2016

Opgave IX

Et kursus på en højere læreranstalt bliver udbudt hvert semester, typisk med flere end 300 studerende, der går til eksamen. Eksamensresultaterne, for 280 studerende, der har bestået kurset ved den sidste eksamen, er gengivet i nedenstående tabel. Eksempelvis ses, at 24 studerende fik karakteren 12. Fordelingen af de 280 karakterer indgår i de 4 næste spørgsmål.

Karakter	02	4	7	10	12	I alt
Antal	22	78	84	72	24	280

Data kan indlæses i R ved:

```
karakterer = rep(x=c(2,4,7,10,12), times=c(22,78,84,72,24))
```

Spørgsmål IX.1 (12)

Benyt den centrale grænseværdisætning til at bestemme et 95% konfidensinterval for middelværdien baseret på de studerende, der har bestået eksamen (Det er vigtigt i dette spørgsmål, at karakterene opfattes numerisk, fx. svarer 02 til tallet 2 osv.).

Eksempel: Maj 2020, Ex III

Vi regner Maj 2020, ex III

Tjekliste

Efter i dag skal du kunne:

- Forklare hvornår en stikprøve er repræsentativ, og vurdere om specifikke stikprøver er repræsentative for deres relevante population.
- Estimere populationens μ (estimeres med $\bar{x}$) og σ (estimeres med s) ud fra stikprøvedata. Forstå og forklar princippet med estimater som stokastiske variable.
- Beregne SEM ud fra stikprøvedata.
- Beregne konfidensintervaller for μ og σ ud fra (normalfordelt) stikprøve data og et givent konfidens niveau.
- Beskrive teorien der ligger til grund for bregning af konfidensintervaller, herunder de stokastiske variable $T = (\bar{X} - \mu)/(s/\sqrt{n})$ og $\chi^2 = (n-1)S^2/\sigma^2$ og deres fordelinger.
- Forklare konceptet "signifikans niveau" (α), inkl. hvordan dette indgår i beregninger for fx. konfidensintervaller.
- Beskrive konsekvensen af CLT og anvende dette i praksis (kend tommelfingerreglen om $n \geq 30$)
- PYTHON: Finde fraktiler i t - og χ^2 -fordelinger

Øvelser

Tirsdage 10-12, Bygning 116 (hjælpelærere ankommer 10.15)

- "Klasseundervisning":
 - **010**: Elektroteknologi, Softwareteknologi, Cyberteknologi
 - **011**: Bæredygtigt energidesign, Miljøteknologi, Design og innovation
 - **012**: Kemi og teknologi, Medicin og teknologi, Teknisk biomedicin, Bioteknologi, Life science
 - **013**: Byggeteknologi, Bygningsdesign, Produktion og konstruktion, General Engineering
 - **015**: Fysik og ingeniørvidenskab, Geofysik og rumteknologi, Anvendt matematik
 - **016**: Kunstig intelligens og data
- Arbejd selv og få hjælp efter behov:
 - **Foyer areas** (with circulating TA's): Blandet