Kursus 02402/02323 Introducerende Statistik

Forelæsning 4: Konfidensinterval for middelværdi (og spredning)

Per Bruun Brockhoff

DTU Compute, Statistik og Dataanalyse
Bygning 324, Rum 220
Danmarks Tekniske Universitet
2800 Lyngby – Danmark

e-mail: perbb@dtu.dk

Oversigt

Eksempel - Højde af 10 studerende:

Stikprøve, $n = 10$:

168 161 167 179 184 166 198 187 191 179

Sample mean og standard deviation:

$$\bar{x} = 178$$
$$s = 12.21$$

Estimerer population mean og standard deviation:

$$\hat{\mu} = 178$$
$$\hat{\sigma} = 12.21$$

NYT:**Konfidensinterval, μ:**

$$178 \pm 2.26 \cdot \frac{12.21}{\sqrt{10}} \Leftrightarrow [169.3;\ 186.7]$$

NYT:**Konfidensinterval, σ:**

$$[8.4;\ 22.3]$$

Theorem 3.2: Fordeling for gennemsnit af normalfordelinger

(Stikprøve-) fordelingen/ The (sampling) distribution for $\bar{X}$

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

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

Middelværdi og varians følger af regneregler

Middelværdien af $\bar{X}$

$$E(\bar{X}) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n}\sum_{i=1}^{n} \mu = \frac{1}{n} n\mu = \mu$$

Variansen for $\bar{X}$

$$\text{Var}(\bar{X}) = \frac{1}{n^2}\sum_{i=1}^{n} \text{Var}(X_i) = \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}$$

Vi kender nu fordelingen af den fejl vi begår:

(Når vi bruger $\bar{x}$ som estimat for μ)

Spredningen af $\bar{X}$

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

Spredningen af $(\bar{X} - \mu)$

$$\sigma_{(\bar{X}-\mu)} = \frac{\sigma}{\sqrt{n}}$$

Standardiseret version af de samme ting, Corollary 3.3:

Fordelingen for den standardiserede fejl vi begår:

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

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

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

Praktisk problem i alt dette, so far:

Hvordan skal alt dette omsættes til et konkret interval for μ?

Når nu populationsspredningen σ indgår i alle formlerne?

Oplagt løsning:

Anvend estimtatet s i stedet for σ i formlerne!

MEN MEN:

Så bryder den givne teori faktisk sammen!!

HELDIGVIS:

Der findes en udvidet teori, der kan klare det!!

Theorem 3.4: More applicable extension of the same stuff: (kopi af Theorem 2.49)

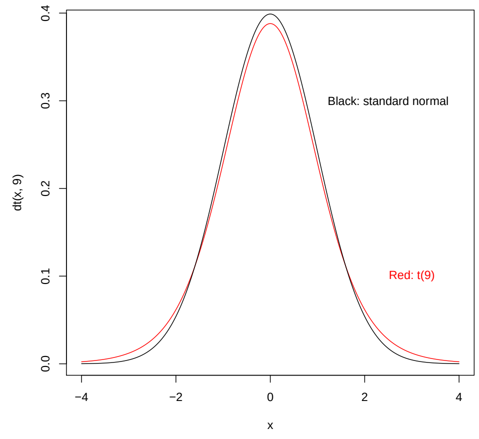
t-fordelingen tager højde for usikkerheden i at bruge s:

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

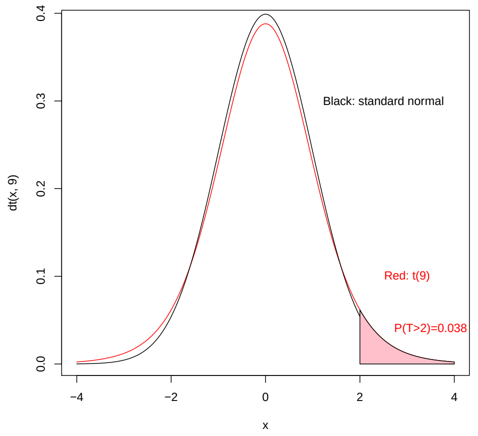
$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t$$

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

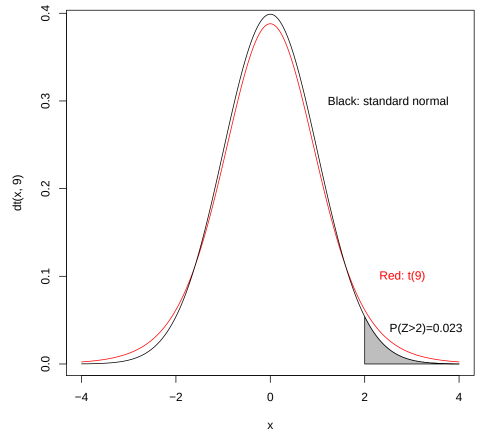
t-fordelingen med 9 frihedsgrader ($n = 10$):



t-fordelingen med 9 frihedsgrader og standardnormalfordelingen:



t-fordelingen med 9 frihedsgrader og standardnormalfordelingen:



Metodeboks 3.8: One-sample konfidensinterval for μ

Brug den rigtige t-fordeling til at lave konfidensintervallet:

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

$$\bar{x} \pm t_{1-\alpha/2} \cdot \frac{s}{\sqrt{n}}$$

where $t_{1-\alpha/2}$ is the $100(1-\alpha)\%$ quantile from the t-distribution with $n-1$ degrees of freedom.

Mest almindeligt med $\alpha = 0.05$:

The most commonly used is the 95%-confidence interval:

$$\bar{x} \pm t_{0.975} \cdot \frac{s}{\sqrt{n}}$$

Højde-eksempel

```
## The t-quantiles for n=10:
qt(0.975,9)
```

[1] 2.2622

Og vi kan genkende det allerede angivne resultat:

$$178 \pm 2.26 \cdot \frac{12.21}{\sqrt{10}}$$

which is:

$$178 \pm 8.74 = [169.3;\ 186.7]$$

Højde-eksempel, 99% Konfidensinterval (CI)

```
qt(0.995,9)
```

[1] 3.2498

$$178 \pm 3.25 \cdot \frac{12.21}{\sqrt{10}}$$

som giver

$$178 \pm 12.55 = [165.4;\ 190.6]$$

Der findes en R-funktion, der kan gøre det hele (og lidt mere til):

```
x <- c(168,161,167,179,184,166,198,187,191,179)
t.test(x,conf.level=0.99)

##
##  One Sample t-test
##
## data:  x
## t = 46.096, df = 9, p-value = 5.326e-12
## alternative hypothesis: true mean is not equal to 0
## 99 percent confidence interval:
##  165.45 190.55
## sample estimates:
## mean of x
##       178
```

Den formelle ramme for *statistisk inferens*

Fra eNote, Chapter 1:

- An *observational unit* is the single entity/level about which information is sought (e.g. a person) (**Observationsenhed**)
- The *statistical population* consists of all possible "measurements" on each *observational unit* (**Population**)
- The *sample* from a statistical population is the actual set of data collected. (**Stikprøve**)

Sprogbrug og koncepter:

- μ og σ er parametre, som beskriver populationen
- $\bar{x}$ er *estimatet* for μ (konkret udfald)
- $\bar{X}$ er *estimatoren* for μ (nu set som stokastisk variabel)
- Begrebet *'statistic(s)'* er en fællesbetegnelse for begge

Den formelle ramme for *statistisk inferens* - Eksempel

Fra eNote, Chapter 1, højdeeksempel

Vi måler højden for 10 tilfældige personer i Danmark

Stikprøven/The sample:

De 10 konkrete talværdier: $x_1, \ldots, x_{10}$

Populationen:

Højderne for alle mennesker i Danmark.

Observationsenheden:

En person

Statistisk inferens = Learning from data

Learning from data:

Is learning about parameters of distributions that describe populations.

Vigtigt i den forbindelse:

Stikprøven skal på meningsfuld vis være repræsentativ for en eller anden veldefineret population

Hvordan sikrer man det:

F.eks. ved at sikre at stikprøven er fuldstændig tilfældig udtaget

Tilfældig stikprøveudtagning

Definition 3.11:

- A random sample from an (infinite) population: A set of observations $X_1, X_2, ..., X_n$ 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. These n random variables are independent

Hvad betyder det????

1. Alle observationer skal komme fra den samme population
2. De må IKKE dele information med hinanden (f.eks. hvis man havde udtaget hele familier i stedet for enkeltindivider)

Theorem 3.13: The Central Limit Theorem

Uanset hvad bliver fordelingen for et gennemsnit en normalfordeling:

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

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

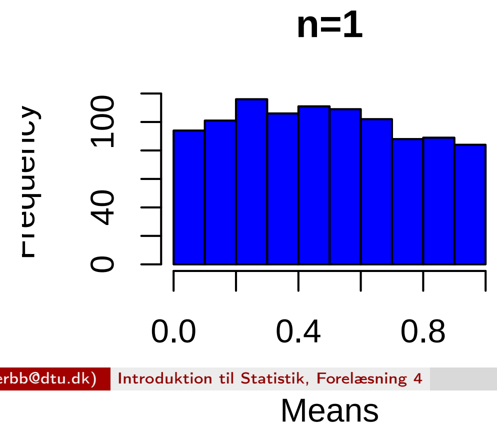
is a random variable whose distribution function approaches that of the standard normal distribution, $N(0, 1^2)$, as $n \to \infty$

Dvs., hvis n er stor nok, kan vi (tilnærmelsesvist) antage:

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

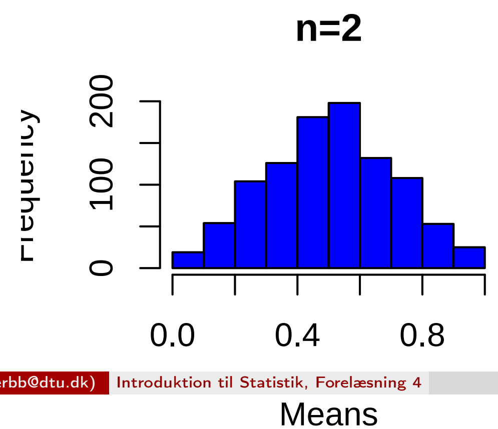
CLT in action - gennemsnit af Uniform fordelte observationer

```
n=1
k=1000
u=matrix(runif(k*n),ncol=n)
hist(apply(u,1,mean),col="blue",main="n=1",xlab="Means")
```



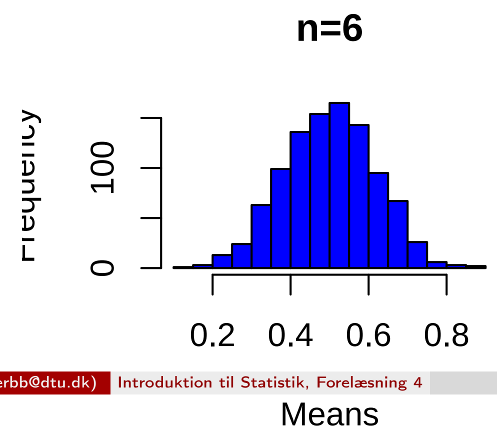
CLT in action - gennemsnit af Uniform fordelte observationer

```
n=2
k=1000
u=matrix(runif(k*n),ncol=n)
hist(apply(u,1,mean),col="blue",main="n=2",xlab="Means")
```



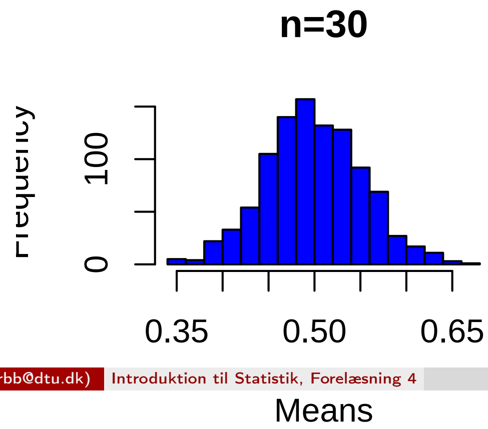
CLT in action - gennemsnit af Uniform fordelte observationer

```
n=6
k=1000
u=matrix(runif(k*n),ncol=n)
hist(apply(u,1,mean),col="blue",main="n=6",xlab="Means")
```



CLT in action - gennemsnit af Uniform fordelte observationer

```
n=30
k=1000
u=matrix(runif(k*n),ncol=n)
hist(apply(u,1,mean),col="blue",main="n=30",xlab="Means", nclass=15)
```



Konsekvens af CLT:

Vores CI-metode virker OGSÅ for ikke-normale data:

Vi kan bruge konfidens-interval baseret på t-fordelingen i stort set alle situationer, blot n er "stor nok"

Hvad er "stor nok"?

Faktisk svært at svare præcist på, MEN:

- Tommelfingerregel: $n \geq 30$
- Selv for mindre n kan formlen være (næsten)gyldig for ikke-normale data.

'Repeated sampling' fortolkning

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.

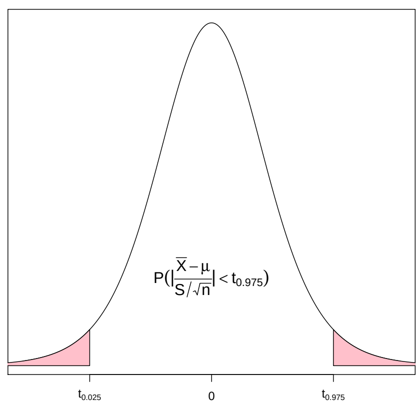
Mere formelt udtrykt (Theorem 3.4 og 2.49):

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

Som er ækvivalent med:

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

'Repeated sampling' fortolkning



Eksempel

Produktion af tabletter

Vi producere pulverblanding og tabletter deraf, så koncenttrationen af det aktive stof i tabletterne skal være 1 mg/g med den mindst mulige spredning. En tilfældig stikprøve udtages, hvor vi måler mængden af aktivt stof.

Stikprøvefordelingen for varians-estimatet (Theorem 2.53)

Variansestimater opfører sig som en χ^2-fordeling:

Let

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

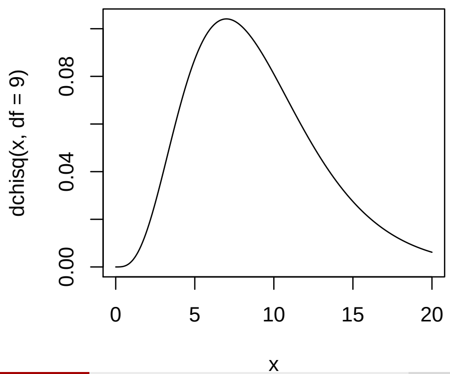
then:

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

is a stochastic variable following the χ^2-distribution with $v = n - 1$ degrees of freedom.

χ^2-fordelingen med $\nu = 9$ frihedsgrader

```
x <- seq(0, 20, by = 0.1)
plot(x, dchisq(x, df = 9), type = "l")
```



Metode 3.18: Konfidensinterval for stikprøvevarians og -spredning

Variansen:

A $100(1-\alpha)\%$ confidence interval for a sample variance $\hat{\sigma}^2$ is:

$$\left[\frac{(n-1)s^2}{\chi^2_{1-\alpha/2}};\ \frac{(n-1)s^2}{\chi^2_{\alpha/2}}\right]$$

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

Spredningen:

A $100(1-\alpha)\%$ confidence interval for the sample standard deviation $\hat{\sigma}$ is:

$$\left[\sqrt{\frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}};\ \sqrt{\frac{(n-1)s^2}{\chi^2_{\alpha/2}}}\right]$$

Eksempel

Data:
En tilfældig stikprøve med $n = 20$ tabletter er udtaget og fra denne får man:

$$\hat{\mu} = \bar{x} = 1.01,\ \hat{\sigma}^2 = s^2 = 0.07^2$$

95%-konfidensinterval for variansen - vi skal bruge χ^2-fraktilerne:

$$\chi^2_{0.025} = 8.9065,\ \chi^2_{0.975} = 32.8523$$

```
qchisq(c(0.025, 0.975), df = 19)
```

[1] 8.9065 32.8523

Eksempel

Så konfidensintervallet for variansen σ^2 bliver:

$$\left[\frac{19 \cdot 0.7^2}{32.85};\ \frac{19 \cdot 0.7^2}{8.907}\right] = [0.002834;\ 0.01045]$$

Og konfidensintervallet for spredningen σ bliver:

$$\left[\sqrt{0.002834};\ \sqrt{0.01045}\right] = [0.053;\ 0.102]$$

Højdeeksempel

Vi skal bruge χ^2-fraktilerne med $\nu = 9$ frihedsgrader:

$$\chi^2_{0.025} = 2.700389,\ \chi^2_{0.975} = 19.022768$$

```
qchisq(c(0.025, 0.975), df = 9)
```

[1] 2.7004 19.0228

Så konfidensintervallet for højdespredningen σ bliver:

$$\left[\sqrt{\frac{9 \cdot 12.21^2}{19.022768}};\ \sqrt{\frac{9 \cdot 12.21^2}{2.700389}}\right] = [8.4;\ 22.3]$$

Eksempel - Højde af 10 studerende - recap:

Stikprøve, $n = 10$:

168 161 167 179 184 166 198 187 191 179

Sample mean og standard deviation:

$$\bar{x} = 178$$
$$s = 12.21$$

Estimerer population mean og standard deviation:

$$\hat{\mu} = 178$$
$$\hat{\sigma} = 12.21$$

NYT:**Konfidensinterval, μ:**

$$178 \pm 2.26 \cdot \frac{12.21}{\sqrt{10}} \Leftrightarrow [169.3;\ 186.7]$$

NYT:**Konfidensinterval, σ:**

$$[8.4;\ 22.3]$$

Oversigt