Kursus 02403 Introduktion til Matematisk Statistik

Forelæsning 5: One sample situationer

Jan Kloppenborg Møller

DTU Compute, Dynamiske Systemer
Bygning 303B, Rum 016
Danmarks Tekniske Universitet
2800 Lyngby – Danmark
e-mail: jkmo@dtu.dk

Oversigt

Oversigt

Theorem 3.14: The Central Limit Theorem

Gennemsnittet af en tilfældig stikprøve følger uanset hvad 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:

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

Konsekvens af CLT:

Vores normal fordelings baserede metoder 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.

Oversigt

Theorem 3.3: 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:

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

Standardiseret version af de samme ting, Theorem 3.4:

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:

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

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

Theorem 3.5: More applicable extension of the same stuff:

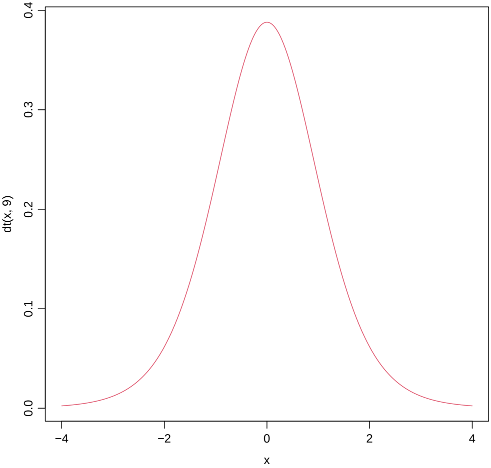
t-fordelingen tager højde for usikkerheden i at bruge s:

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:

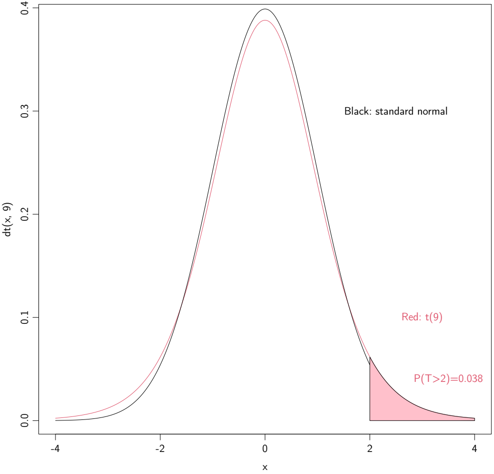
$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t$$

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

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

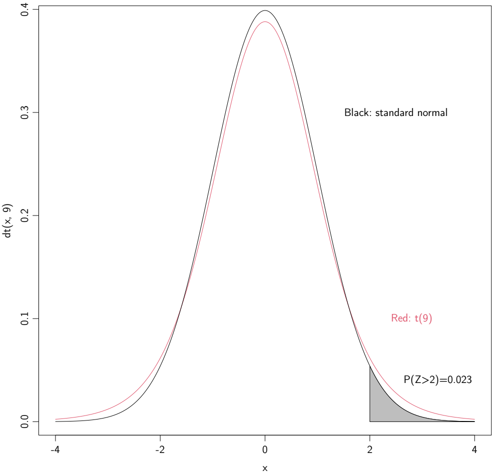


t-fordelingen med 9 frihedsgrader og standardnormalfordelingen:



DTU Compute
Department of Applied Mathematics and Computer Science

t-fordelingen med 9 frihedsgrader og standardnormalfordelingen:



Oversigt

DTU Compute
Department of Applied Mathematics and Computer Science

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 3.12:

- 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)

Oversigt

DTU Compute
Department of Applied Mathematics and Computer Science

'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 3.5):

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

Som er ækvivalent med:

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

Metodeboks 3.9: 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:

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

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

```
## Angiv data
x <- c(180 ,180 ,184 ,182 ,178 ,190 ,175)/100
## Beregn 99% konfidensinterval
t.test(x, conf.level=0.99)

##
##  One Sample t-test
##
## data:  x
## t = 100, df = 6, p-value = 6.7e-11
## alternative hypothesis: true mean is not equal to 0
## 99 percent confidence interval:
##  1.7458 1.8799
## sample estimates:
## mean of x
##    1.8129
```

Oversigt

Eksempel

Produktion af tabletter

Vi producere pulverblanding og tabletter deraf, så koncentrationen 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 2.81)

Variansestimatorer 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 random variable following the χ^2-distribution with $v = n - 1$ degrees of freedom.

Metode 3.19: Konfidensinterval for stikprøvevarians og -spredning

Variansen:

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

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

```r
## 2.5% og 97.5% fraktilerne i chi^2 fordelingen for n=20
qchisq(c(0.025, 0.975), df = 19)

## [1]  8.9065 32.8523
```

Eksempel

Så konfidensintervallet for variansen σ^2 bliver:

$$\left[\frac{19 \cdot 0.07^2}{32.85};\ \frac{19 \cdot 0.07^2}{8.907}\right] = [0.002834;\ 0.01045]$$

Og konfidensintervallet for spredningen σ bliver:

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

Oversigt

Motiverende eksempel - sovemedicin

Forskel på sovemedicin?

I et studie er man interesseret i at sammenligne 2 sovemidler A og B. For 10 testpersoner har man fået følgende resultater, der er givet i forlænget søvntid (i timer) (Forskellen på effekten af de to midler er angivet):

Stikprøve, $n = 10$:

person	$x =$ Beffect - Aeffect
1	1.2
2	2.4
3	1.3
4	1.3
5	0.9
6	1.0
7	1.8
8	0.8
9	4.6
10	1.4

Eksempel - sovemedicin

Hypotesen om ingen forskel ønskes undersøgt:

$$H_0: \; \mu = 0$$

Sample mean og standard deviation:

$$\bar{x} = 1.670 = \hat{\mu}$$
$$s = 1.13 = \hat{\sigma}$$

Er data i overenstemmelse med nulhyposen H_0?

Data: $\bar{x} = 1.67, \; H_0: \; \mu = 0$

NYT:p-**værdi:**

$$p - \text{værdi} = 0.00117$$

(Beregnet under det scenarie, at H_0 er sand)

NYT:**Konklusion**:

Idet data er usandsynligt under H_0, så **forkaster** vi H_0 - vi har påvist en **signifikant effekt** af middel B ift. middel A.

Oversigt

Metode 3.23: One-sample t-test og p-værdi

Hvordan beregner man p-værdien?

For a (quantitative) one sample situation, the (non-directional) p-value is given by:

$$p\text{-value} = 2 \cdot P(T > |t_{\text{obs}}|)$$

where T follows a t-distribution with $(n-1)$ degrees of freedom.
The observed value of the test statistics to be computed is

$$t_{\text{obs}} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

where μ_0 is the value of μ under the null hypothesis:

$$H_0: \ \mu = \mu_0$$

Definition og fortolkning af p-værdien (HELT generelt)

Definition 3.22 af p-værdien:

The p-value is the probability of obtaining a test statistic that is at least as extreme as the test statistic that was actually observed. This probability is calculated under the assumption that the null hypothesis is true.

p-værdien udtrykker *evidence* imod nulhypotesen – Tabel 3.1:

$p < 0.001$	Very strong evidence against H_0
$0.001 \leq p < 0.01$	Strong evidence against H_0
$0.01 \leq p < 0.05$	Some evidence against H_0
$0.05 \leq p < 0.1$	Weak evidence against H_0
$p \geq 0.1$	Little or no evidence against H_0

Eksempel - sovemedicin

Hypotesen om ingen forskel ønskes undersøgt:

$$H_0: \ \mu = 0$$

Beregne test-størrelsen:

$$t_{\text{obs}} = \frac{1.67 - 0}{1.13/\sqrt{10}} = 4.67$$

Beregne p-værdien:

$$2P(T > 4.67) = 0.00117$$

```
2 * (1-pt(4.67, 9))
```

Fortolkning af p-værdi i lyset af Tabel 3.1:

Der er stærk evidens imod nulhypotesen.

Eksempel - sovemedicin - i R - manuelt

```
## Angiv data
x <- c(1.2, 2.4, 1.3, 1.3, 0.9, 1.0, 1.8, 0.8, 4.6, 1.4)
n <- length(x)
## Beregn den observerede t værdi - den observerede test statistik
tobs <- (mean(x) - 0) / (sd(x) / sqrt(n))
## Beregn p-værdien, som sandsynligheden for at få tobs eller mere e
pvalue <- 2 * (1-pt(abs(tobs), df=n-1))
pvalue

## [1] 0.0011659
```

Eksempel - sovemedicin - i R - med indbygget funktion

```
## Kald funktionen med data x
t.test(x)

##
##  One Sample t-test
##
## data:  x
## t = 4.67, df = 9, p-value = 0.0012
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  0.86133 2.47867
## sample estimates:
## mean of x
##      1.67
```

Definition af hypotesetest og signifikans (HELT generelt)

Definition 3.24. Hypotesetest:

We say that we carry out a hypothesis test when we decide against a null hypothesis or not, using the data.

A null hypothesis is *rejected* if the p-value, calculated after the data has been observed, is less than some α, that is if the p-value $< \alpha$, where α is some pre-specifed (so-called) *significance level*. And if not, then the null hypothesis is said to be *accepted*.

Definition 3.29. Statistisk signifikans:

An *effect* is said to be *(statistically) significant* if the p-value is less than the significance level α.
(OFTE bruges $\alpha = 0.05$)

Eksempel - sovemedicin

Med $\alpha = 0.05$ kan vi konkludere:

Idet p-værdien er mindre end α så **forkaster** vi nulhypotesen.

Og dermed:

Vi har påvist en **signifikant effekt** af middel B ift. middel A. (Og dermed at B virker bedre end A)

Kritisk værdi

Definition 3.31 - de kritiske værdier for t-testet:

The $(1-\alpha)100\%$ critical values for the (non-directional) one-sample t-test are the $(\alpha/2)100\%$ and $(1-\alpha/2)100\%$ quantiles of the t-distribution with $n-1$ degrees of freedom:

$$t_{\alpha/2} \text{ and } t_{1-\alpha/2}$$

Metode 3.32: One-sample t-test vha. kritisk værdi:

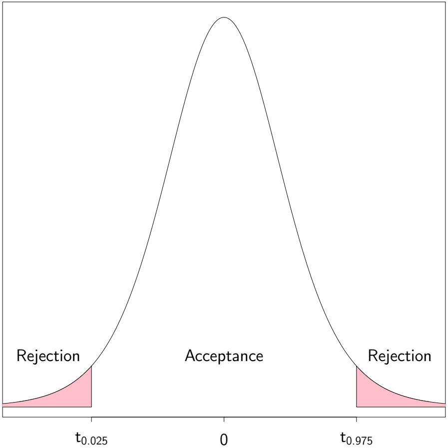
A null hypothesis is *rejected* if the observed test-statistic is more extreme than the critical values:

$$\text{If } |t_{\text{obs}}| > t_{1-\alpha/2} \text{ then } \textit{reject } H_0$$

otherwise *accept*.

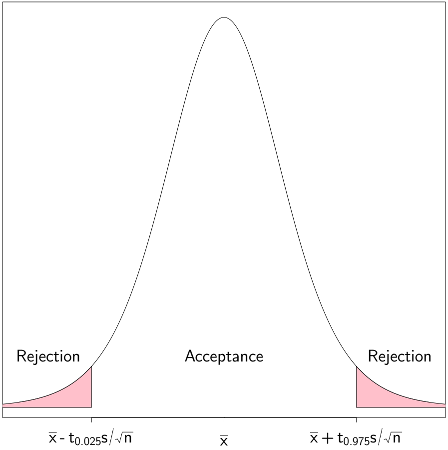
Kritisk værdi og hypotesetest

Acceptområdet er de mulige værdier for μ som ikke ligger for langt væk fra data - her på den standardiserede skala:



Kritisk værdi og hypotesetest

Acceptområdet er de mulige værdier for μ som ikke ligger for langt væk fra data - nu på den egentlige skala:



Kritisk værdi, konfidensinterval og hypotesetest

Theorem 3.33:
Kritisk-værdi-metode ækvivalent med Konfidensinterval-metode

We consider a $(1-\alpha) \cdot 100\%$ confidence interval for μ:

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

The confidence interval corresponds to the acceptance region for H_0 when testing the (non-directional) hypothesis

$$H_0 : \quad \mu = \mu_0$$

(Ny) fortolkning af konfidensintervallet:

De (hypotetiske) værdier for μ, som vi accepterer ved det tilsvarende hypotesetest.

Bevis:

Remark 3.34

A μ_0 inside the confidence interval will fullfill that

$$|\bar{x} - \mu_0| < t_{1-\alpha/2} \cdot \frac{s}{\sqrt{n}}$$

which is equivalent to

$$\frac{|\bar{x} - \mu_0|}{\frac{s}{\sqrt{n}}} < t_{1-\alpha/2}$$

and again to

$$|t_{\text{obs}}| < t_{1-\alpha/2}$$

which then exactly states that μ_0 is accepted, since the t_{obs} is within the critical values.

Oversigt

DTU Compute
Department of Applied Mathematics and Computer Science

Hypotese-test med alternativer

Indtil nu - underforstået: (= non-directional)

Alternativet til $H_0: \ \mu = \mu_0$ er : $H_1: \ \mu \neq \mu_0$

MEN der kan være andre settings, e.g. one-sided (=directional), "less":

Alternativet til $H_0: \ \mu = \mu_0$ er : $H_1: \ \mu < \mu_0$

Eller one-sided (=directional), "greater":

Alternativet til $H_0: \ \mu = \mu_0$ er : $H_1: \ \mu > \mu_0$

Metode 3.36. Steps ved hypotesetests - et overblik

Helt generelt består et hypotesetest af følgende trin:

1. Formulate the hypotheses and choose the level of significance α (choose the "risk-level")
2. Calculate, using the data, the value of the test statistic
3. Calculate the p-value using the test statistic and the relevant sampling distribution, and compare the p-value and the significance level α and make a conclusion
4. (Alternatively, make a conclusion based on the relevant critical value(s))

Det tosidede (non-directional) one-sample t-test igen

Et level α test er:

1. Compute t_{obs} as before
2. Compute the evidence against the *null hypothesis* $H_0: \ \mu = \mu_0$ vs. the *alternative hypothesis* $H_1: \ \mu \neq \mu_0$ by the

$$p\text{–value} = 2 \cdot P(T > |t_{\text{obs}}|)$$

   where the t-distribution with $n-1$ degrees of freedom is used.
3. If p–value $< \alpha$: We reject H_0, otherwise we accept H_0.
4. The rejection/acceptance conclusion could alternatively, but equivalently, be made based on the critical value(s) $\pm t_{1-\alpha/2}$:
   If $|t_{\text{obs}}| > t_{1-\alpha/2}$ we reject H_0, otherwise we accept H_0.

Mulige fejl ved hypotesetests

Der findes to slags fejl (dog kun een af gangen!)

Type I: Rejection of H_0 when H_0 is true
Type II: Non-rejection of H_0 when H_1 is true

Risikoen for de to typer fejl kaldes sædvanligvis:

$$P(\text{Type I error}) = \alpha$$

$$P(\text{Type II error}) = \beta$$

Retsalsanalogi

En person står stillet for en domstol:

A man is standing in a court of law accused of criminal activity.
The null- and the the alternative hypotheses are:

H_0 : The man is not guilty

H_1 : The man is guilty

At man ikke kan bevises skyldig er ikke det samme som at man er bevist uskyldig:

Absence of evidence is NOT evidence of absence!

Or differently put:
Accepting a null hypothesis is NOT a statistical proof of the null hypothesis being true!

Mulige fejl ved hypotesetests

To mulige sandheder vs. to mulige konklusioner:

	Reject H_0	Fail to reject H_0
H_0 is true	Type I error (α)	Correct acceptance of H_0
H_0 is false	Correct rejection of H_0 (Power)	Type II error (β)

Theorem 3.39: Signifikansniveauet = Risikoen for Type I fejl

The significance level α in hypothesis testing is the overall Type I risk:

$$P(\text{Type I error}) = P(\text{Rejection of } H_0 \text{ when } H_0 \text{ is true}) = \alpha$$

Eksempel: Sommervandtemperatur i Skive fjord

Hvad er:

- Den forventede vandtemperatur i Skive fjord i juli måned?
- Variationen i vandtemperaturen i juli måned i Skive fjord fra år til år?

Oversigt