02402 Statistik (Polyteknisk grundlag)

Uge 6: Analyser med to stikprøver

DTU Compute
Danmarks Tekniske Universitet
2800 Kgs. Lyngby

Dagsorden

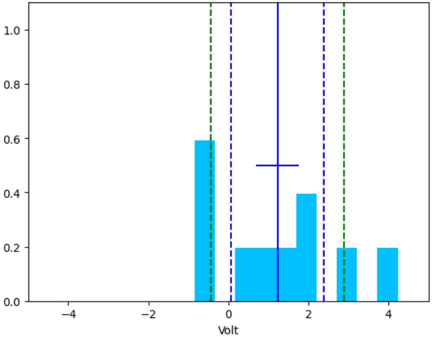
1. Opsummering fra sidste uge
2. Statistik med 2 stikprøver
   - t-test for differens mellem stikprøver
   - t-test med sammenvejet ("pooled") varians – Et alternativ
   - Konfidensintervallet for differensen
   - t-test med to parrede stikprøver (parret t-test)
3. Styrke og stikprøvestørrelse – Forsøgsdesign
   - Krav til præcision
   - Styrke og stikprøvestørrelse – En stikprøve
   - Styrke og stikprøvestørrelse – To stikprøver

02402 Statistik (Polyteknisk grundlag)

Opsummering fra sidste uge

DTU Compute
Danmarks Tekniske Universitet
2800 Kgs. Lyngby

Hypotesetest



t-test "cook-book":

- Hulhypotese $H_0 : \mu = \mu_0 (= 0?)$
- Beregn: $\bar{x}$, s, and $s.e._{\bar{x}} = s/\sqrt{n}$
- Beregn teststørrelse: $t_{obs} = \frac{\bar{x} - \mu_0}{s.e._{\bar{x}}}$
- Beregn p-værdi: $2 \cdot P(T > |t_{obs}|)$ (opslag i t-fordeling, lav tegning)
- Sammenhold p-værdi med ønsket α

\- Ligger værdien μ_0 *udenfor* $(1 - \alpha) \cdot 100\%$-konfidensintervallet? (fx 95%-CI)
\- Er den beregnede $|t_{obs}|$ *større* end $t_{1-\alpha/2}$? (fx $t_{0.975}$)
\- Er p-værdien *mindre* end α? (fx 0.05)
Ja? - så afviser vi nulhypotesen ved signifikansniveau α.

p-værdien svarer til det signifikansnivau (α) der lige netop ville indeholde nulhypotesen i konfidensintervallet.

02402 Statistik (Polyteknisk grundlag)

Statistik med 2 stikprøver

- t-test for differens mellem stikprøver
- t-test med sammenvejet ("pooled") varians – Et alternativ
- Konfidensintervallet for differensen
- t-test med to parrede stikprøver (parret t-test)

DTU Compute
Danmarks Tekniske Universitet
2800 Kgs. Lyngby

I dag

I dag

- Vi vil ofte gerne vide om der er ("signifikant") forskel på to forskellige ting/grupper/interventioner.
- Vi har set eksempler hvor data afspejler målte forskelle - og vi tester om denne forskel i gennemsnit er nul.
- Men hvad hvis vi fx vil vide om der er forskel på to forskellige grupper, hvor vi ikke kan opsamle data for individuelle forskelle? (forskellen kan ikke måles på én statistisk enhed, grupperne kan også være forskellig størrelse)
- Så skal vi se på t-test for **to** stikprøver.

Motiverende eksempel: Ernæringsstudie

Forskel på energiforbrug?

I et ernæringsstudie ønsker man at undersøge, om der er en forskel i energiforbruget for forskellige typer (moderat fysisk krævende) arbejde.

I studiet har 9 sygeplejersker fra hospital A og 9 (andre) sygeplejersker fra hospital B fået målt deres energiforbrug. Målingerne ses i følgende tabel (i enheden megajoule MJ):

Stikprøve fra hvert hospital: $n_1 = n_2 = 9$:

Hospital A	Hospital B
7.53	9.21
7.48	11.51
8.08	12.79
8.09	11.85
10.15	9.97
8.40	8.79
10.88	9.69
6.13	9.68
7.90	9.19

Python

- Gå til Python notebook "nutrition_study.ipynb" (1) i VS Code

Visual Studio Code

Kahoot!
(x1)

02402 Statistik (Polyteknisk grundlag)

- t-test for differens mellem stikprøver
- t-test med sammenvejet ("pooled") varians – Et alternativ
- Konfidensintervallet for differensen
- t-test med to parrede stikprøver (parret t-test)

DTU Compute
Danmarks Tekniske Universitet
2800 Kgs. Lyngby

Forskel på to grupper

Hvad er nulhypotesen? Hvordan ville man formulere den? (diskutér 2 min)

(tavle)

Forskel på to grupper

Overvejelser til nulhypotesen:

- Der er ingen forskel på grupperne "A" og "B"
- De to stikprøver kommer fra *præcis samme* underliggende fordeling (?)
- **De underliggende fordelinger har samme gennemsnit:** $\mu_A = \mu_B$
- De underliggende fordelinger har samme *varians*: $\sigma_A = \sigma_B$ (?)

Eksempel: Ernæringsstudie

Hypotesen om ingen forskel (i det gennemsnitlige energiforbrug) ønskes undersøgt:

$$H_0: \ \mu_A = \mu_B \quad \text{eller} \quad H_0: \ \mu_A - \mu_B = 0$$

Eksempel: Ernæringsstudie

Hypotesen om ingen forskel (i det gennemsnitlige energiforbrug) ønskes undersøgt:

$$H_0: \ \mu_A = \mu_B \quad \text{eller} \quad H_0: \ \mu_A - \mu_B = 0$$

Stikprøvegennemsnit og -standardafvigelser:

$$\hat{\mu}_A = \bar{x}_A = 8.293 \ (s_A = 1.428)$$
$$\hat{\mu}_B = \bar{x}_B = 10.298 \ (s_B = 1.398)$$

Eksempel: Ernæringsstudie

Hypotesen om ingen forskel (i det gennemsnitlige energiforbrug) ønskes undersøgt:

$$H_0: \ \mu_A = \mu_B \quad \text{eller} \quad H_0: \ \mu_A - \mu_B = 0$$

Stikprøvegennemsnit og -standardafvigelser:

$$\hat{\mu}_A = \bar{x}_A = 8.293 \ (s_A = 1.428)$$

$$\hat{\mu}_B = \bar{x}_B = 10.298 \ (s_B = 1.398)$$

Er data i overensstemmelse med nulhypotesen H_0?

$$\text{Data: } \bar{x}_B - \bar{x}_A = 2.005$$

$$\text{Nulhypotese: } H_0: \ \mu_B - \mu_A = 0$$

Eksempel: Ernæringsstudie

Hypotesen om ingen forskel (i det gennemsnitlige energiforbrug) ønskes undersøgt:

$$H_0: \ \mu_A = \mu_B \quad \text{eller} \quad H_0: \ \mu_A - \mu_B = 0$$

Stikprøvegennemsnit og -standardafvigelser:

$$\hat{\mu}_A = \bar{x}_A = 8.293 \ (s_A = 1.428)$$
$$\hat{\mu}_B = \bar{x}_B = 10.298 \ (s_B = 1.398)$$

Er data i overensstemmelse med nulhypotesen H_0?

Data: $\bar{x}_B - \bar{x}_A = 2.005$

Nulhypotese: $H_0: \ \mu_B - \mu_A = 0$

NYT: **p-værdi for forskel:**

$$p = 0.0083$$

(Beregnet under antagelsen, at H_0 er sand.)

NYT: **Konfidensinterval for forskellen**:

$$2.005 \pm 1.412 = [0.59; \ 3.42]$$

Nulhypotese for en forskel mellem to grupper

Hvad er vores nulhypotese?

$$H_0: \quad \delta = \mu_X - \mu_Y = \delta_0$$

Notation: δ er den "sande" forskel
(μ_X og μ_Y er de "sande" middelværdier)

Ofte er nulhypotesen:

$$H_0: \quad \delta = \mu_X - \mu_Y = 0$$

Der er "ingen forskel" (på populations-gennemsnittene μ_X og μ_Y).

Differensen mellem to gennemsnit

Vi beskriver **differensen** med den en stokastisk variabel:

$$D = \bar{X} - \bar{Y}$$

Middelværdi af D:

$$\mathbf{E}[D] = \mathbf{E}[\bar{X} - \bar{Y}] = \mathbf{E}[\bar{X}] - \mathbf{E}[\bar{Y}] = \mu_X - \mu_Y$$

Variansen af D:

$$\mathbf{V}[D] = \mathbf{V}[\bar{X} - \bar{Y}] = 1^2\mathbf{V}[\bar{X}] + (-1)^2\mathbf{V}[\bar{Y}] = \frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}$$

Standardafvigelse for D:

$$\sigma_D = \sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}$$

Kahoot!
(x1)

Standard error for en **differens**

Standard Error for differensen:

$$\sigma_D = \sqrt{\frac{S_X^2}{n_X} + \frac{S_Y^2}{n_Y}} = \sqrt{(\sigma_{\bar{X}})^2 + (\sigma_{\bar{Y}})^2}$$

eller

$$SE_d = \sqrt{(SE_{\bar{x}})^2 + (SE_{\bar{y}})^2}$$

(Denne er ikke defineret i bogen)

Teststørrelsen t_{obs} for ("Welch") test med to stikprøver

For t-test med én stikprøve:

Vi betragter $T = \frac{\bar{X}-\mu}{S/\sqrt{n}}$, der følger en t-fordeling med $\nu = n-1$ frihedsgrader.

Fra stikprøvedata beregnes $t_{obs} = \frac{\bar{x}-\mu_0}{s/\sqrt{n}}$ og denne bruges til at udføre t-testen.

$$t_{obs} = \frac{\bar{x} - \mu_0}{SE_{\bar{x}}}$$

For t-test med to stikprøver (denne version kaldes en Welch t-test):

Vi betragter $T = \frac{(\bar{X}-\bar{Y})-\delta}{\sqrt{S_X^2/n_X+S_Y^2/n_Y}}$, der følger en ??-fordeling (se næste slide).

Fra stikprøvedata beregnes $t_{obs} = \frac{(\bar{x}-\bar{y})-\delta_0}{\sqrt{s_X^2/n_X+s_Y^2/n_Y}}$ og denne bruges til at udføre t-testen.

$$t_{obs} = \frac{d - \delta_0}{SE_d}$$

Fordeling for T ("Welch" test med to stikprøver)

Theorem 3.50 The distribution of the (Welch) two-sample statistic

The (Welch) two-sample statistic seen as a random variable

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - \delta_0}{\sqrt{S_1^2/n_1 + S_2^2/n_2}}, \tag{3-49}$$

approximately, under the null hypothesis, follows a t-distribution with ν degrees of freedom, where

$$\nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{(s_1^2/n_1)^2}{n_1-1} + \frac{(s_2^2/n_2)^2}{n_2-1}}, \tag{3-50}$$

if the two population distributions are normal or if the two sample sizes are large enough.

("Welch") t-test for differens mellem to stikprøver

Method 3.51 The level α two-sample t-test

1. Compute the test statistic using Equation (3-48) and ν from Equation (3-50)

$$t_{\text{obs}} = \frac{(\bar{x}_1 - \bar{x}_2) - \delta_0}{\sqrt{s_1^2/n_1 + s_2^2/n_2}} \text{ and } \nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{(s_1^2/n_1)^2}{n_1-1} + \frac{(s_2^2/n_2)^2}{n_2-1}}$$

2. Compute the evidence against the *null hypothesis*[a]

$$H_0: \quad \mu_1 - \mu_2 = \delta_0,$$

vs. the *alternative hypothesis*

$$H_1: \quad \mu_1 - \mu_2 \neq \delta_0,$$

by the

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

where the t-distribution with ν degrees of freedom is used

3. If p-value $< \alpha$: we reject H_0, otherwise we accept H_0,

or

The rejection/acceptance conclusion can equivalently be 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

$$t_{obs} = \frac{d - \delta_0}{SE_d} = \frac{(\bar{x} - \bar{y}) - \delta_0}{\sqrt{s_X^2/n_X + s_Y^2/n_Y}}$$

Forudsætninger:

Testen gælder når begge stikprøver er store (CLT), eller når begge stikprøver kommer fra normalfordelte populationer.

Eksempel: Ernæringsstudie - Welch t-test

Udfør en Welch t-test:

- Opstil nulhypotese $H_0 : \delta = \delta_0 (= 0?)$
- Beregn: $d = \bar{x} - \bar{y}$ og SE_d
- Beregn teststørrelse: $t_{obs} = \frac{d - \delta_0}{SE_d}$
- Beregn p-værdi: $2 \cdot P(T > |t_{obs}|)$
  (opslag i t-fordeling med korrekt ν)
- Sammenhold p-værdi med ønsket α

Man kan bruge funktionen:

- `stats.ttest_ind(A, B, equal_var=False)` (fra `Scipy.stats` pakken)

- Gå til Python notebook "nutrition_study.ipynb" (2) i VS Code

Visual Studio Code

02402 Statistik (Polyteknisk grundlag)

- t-test for differens mellem stikprøver
- t-test med sammenvejet ("pooled") varians – Et alternativ
- Konfidensintervallet for differensen
- t-test med to parrede stikprøver (parret t-test)

DTU Compute
Danmarks Tekniske Universitet
2800 Kgs. Lyngby

Sammenvejet (pooled) varians

Forskel fra før: Nu laver vi den *ekstra antagelse* at variansen i de to underliggende fordelinger er ens: $\sigma_X = \sigma_Y$

Vi estimerer variansen i de underliggende fordelinger med en "pooled variance" (sammenvejet varians), s_p^2:

Method 3.52 The pooled two-sample estimate of variance

Under the assumption that $\sigma_1^2 = \sigma_2^2$ the *pooled* estimate of variance is the weighted average of the two sample variances

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}. \tag{3-51}$$

Fordeling for T ("pooled" test med to stikprøver)

Under antagelsen $\sigma_X^2 = \sigma_Y^2 = \sigma_p^2$ har vi $SE_d = \sqrt{s_p^2/n_X + s_p^2/n_Y}$.

Nu betragter vi derfor $T = \frac{(\bar{X}-\bar{Y})-\delta}{\sqrt{S_X^2/n_X+S_Y^2/n_Y}} = \frac{(\bar{X}-\bar{Y})-\delta}{\sqrt{S_p^2/n_X+S_p^2/n_Y}}$

Theorem 3.54 The distribution of the pooled two-sample t-test statistic

The pooled two-sample statistic seen as a random variable:

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - \delta_0}{\sqrt{S_p^2/n_1 + S_p^2/n_2}}. \tag{3-54}$$

follows, under the null hypothesis and under the assumption that $\sigma_1^2 = \sigma_2^2$, a t-distribution with $n_1 + n_2 - 2$ degrees of freedom if the two population distributions are normal.

Teststørrelsen t_{obs} for "pooled" test med to stikprøver

Method 3.53 The pooled two-sample t-test statistic

When considering the null hypothesis about the difference between the means of two *independent* samples

$$\begin{aligned} \delta &= \mu_1 - \mu_2, \\ H_0 :\ \delta &= \delta_0. \end{aligned} \tag{3-52}$$

the pooled two-sample t-test statistic is

$$t_{\text{obs}} = \frac{(\bar{x}_1 - \bar{x}_2) - \delta_0}{\sqrt{s_p^2/n_1 + s_p^2/n_2}}. \tag{3-53}$$

Eksempel: Ernæringsstudie - pooled t-test

Gentag t-test men denne gang under antagelsen $\sigma_X^2 = \sigma_Y^2$ ("pooled" t-test)

- Opstil nulhypotese $H_0 : \delta = \delta_0 (= 0?)$
- Beregn: $d = \bar{x} - \bar{y}$ og SE_d - kræver beregning af s_p^2!
- Beregn teststørrelse: $t_{obs} = \frac{d - \delta_0}{SE_d}$
- Beregn p-værdi: $2 \cdot P(T > |t_{obs}|)$
  (opslag i t-fordeling med korrekt ν. OBS: ν er ikke det samme som før)
- Sammenhold p-værdi med ønsket α

Man kan bruge funktionen:

- `stats.ttest_ind(A, B, equal_var=True)` (fra `Scipy.stats` pakken)

- Gå til Python notebook
  "nutrition_study.ipynb" (3) i VS Code

Visual Studio Code

Vi bruger helst "Welch"-versionen

- Jo færre antagelser jo bedre - brug derfor "Welch" versionen.
- Hvis $s_1^2 = s_2^2$, så er de to test ens. Hvis det er tilfældet, så foretrækker vi ikke nødvendigvis testen med den sammenvejede varians, da antagelsen om ens varianser kan være højst tvivlsom.
- Kun hvis de to varianser er meget forskellige, kan det ske, at de to test giver meget forskellige resultater. Hvis varianserne virker meget forskellige, brydes antagelsen om ens varianser formodentligt.
- I tilfælde med en lille stikprøvestørrelse i mindst en af grupperne, er Welch-versionen en mere "forsigtig" tilgang.
- Senere i kurset (når vi skal sammenligne flere grupper) vil vi dog antage ens varians i alle grupperne.

02402 Statistik (Polyteknisk grundlag)

- t-test for differens mellem stikprøver
- t-test med sammenvejet ("pooled") varians – Et alternativ
- Konfidensintervallet for differensen
- t-test med to parrede stikprøver (parret t-test)

DTU Compute
Danmarks Tekniske Universitet
2800 Kgs. Lyngby

Konfidensinterval for differens mellem to stikprøver

$$d \pm t_{1-\alpha/2} \cdot SE_d$$

Det er samme "opskrift" som sidste gang.

Vær opmærksom på at beregne ν til at indsætte som antal frihedsgrader i t-fordelingen (brug "Welch" versionen).

Konfidensinterval for differens mellem to stikprøver

Method 3.47 The two-sample confidence interval for $\mu_1 - \mu_2$

For two samples $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$ the $100(1-\alpha)\%$ confidence interval for $\mu_1 - \mu_2$ is given by

$$\bar{x} - \bar{y} \pm t_{1-\alpha/2} \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}, \tag{3-45}$$

where $t_{1-\alpha/2}$ is the $(1-\alpha/2)$-quantile from the t-distribution with ν degrees of freedom given from Equation (3-50)

$$\nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{(s_1^2/n_1)^2}{n_1-1} + \frac{(s_2^2/n_2)^2}{n_2-1}}. \tag{3-46}$$

Eksempel: Ernæringsstudie

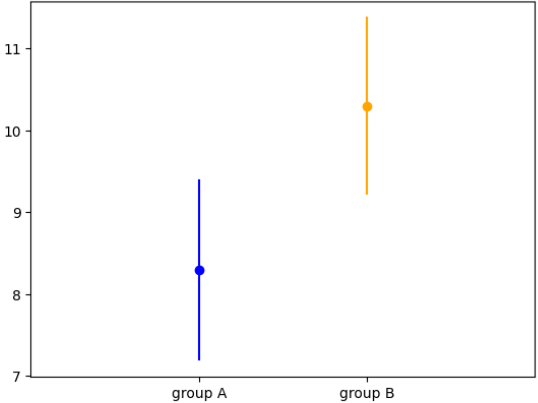
Man kan også bruge:

- `stats.ttest_ind(A, B, equal_var=False).confidence_interval(...)`

- Gå til Python notebook
  "nutrition_study.ipynb" (4) i VS Code

Overlappende konfidensintervaller

Kan I huske de overlappende konfidensintervaller? (for individuelle stikprøvegennemsnit)

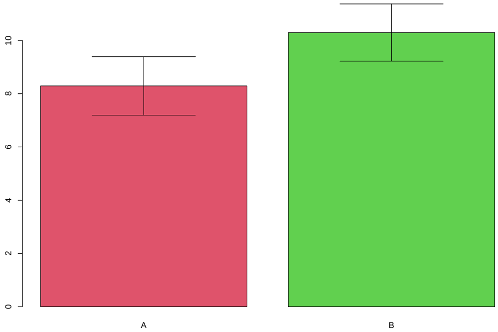


Selvom de to konfidensintervaller overlapper, endte vi med at konkludere at der ER signifikant forskel.

Alternativ præsentation fra bogen

Søjlediagrammer med *fejlbjælker* ses ofte:

Et søjlediagram med nogle fejlbjælker (error bars): Herunder vises 95%-konfidensintervallerne for hver gruppe:



Vær varsom

Når to konfidensintervaller IKKE overlapper: De to grupper er signifikant forskellige.

Når to konfidensintervaller overlapper: Ingen konklusion kan drages uden at undersøge konfidensintervallet for *forskellen* mellem grupperne.

Remark 3.59

When interpreting two (and multi-) independent samples mean bar plots with added confidence intervals:

When two CIs do NOT overlap: The two groups are significantly different

When two CIs DO overlap: We do not know from this what the conclusion is (but then we can use the presented two-sample test method)

02402 Statistik (Polyteknisk grundlag)

- t-test for differens mellem stikprøver
- t-test med sammenvejet ("pooled") varians – Et alternativ
- Konfidensintervallet for differensen
- t-test med to parrede stikprøver (parret t-test)

DTU Compute
Danmarks Tekniske Universitet
2800 Kgs. Lyngby

Motiverende eksempel: Sovemedicin

Forskel på sovemedicin?

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

Stikprøve med $n = 10$:

Person	A	B	$D = B - A$
1	+0.7	+1.9	+1.2
2	-1.6	+0.8	+2.4
3	-0.2	+1.1	+1.3
4	-1.2	+0.1	+1.3
5	-1.0	-0.1	+0.9
6	+3.4	+4.4	+1.0
7	+3.7	+5.5	+1.8
8	+0.8	+1.6	+0.8
9	0.0	+4.6	+4.6
10	+2.0	+3.4	+1.4

Motiverende eksempel: Sovemedicin

Forskel på sovemedicin?

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

Stikprøve med $n = 10$:

Person	A	B	$D = B - A$
1	+0.7	+1.9	+1.2
2	-1.6	+0.8	+2.4
3	-0.2	+1.1	+1.3
4	-1.2	+0.1	+1.3
5	-1.0	-0.1	+0.9
6	+3.4	+4.4	+1.0
7	+3.7	+5.5	+1.8
8	+0.8	+1.6	+0.8
9	0.0	+4.6	+4.6
10	+2.0	+3.4	+1.4

$\bar{x} = 1.67$
$s = 1.13$

t-test for parret data

Når man har to "parrede" stikprøver, kan disse *oversættes* til én stikprøve med målte forskelle. Data analyseres derfor med en almindelig *t*-test for én stikprøve.

Man kan bruge:

- `stats.ttest_1samp(diff_data, pop_mean=0)`

Eller:

- `stats.ttest_rel(A, B)`

(*_rel* står for *relative*, dvs parret data)

- Gå til Python notebook "sleep_medicine.ipynb" i VS Code

Visual Studio Code

Forsøgsopsætning: Parrede og uafhængige stikprøver

Fuldstændigt tilfældigt (uafhængige stikprøver):

Vi har 20 patienter, som tilfældigt fordeles på to grupper (normalt lige mange i hver gruppe). Dvs. at der er forskellige (uafhængige) patienter i de to grupper.

Parrede observationer (afhængige stikprøver):

Vi har 10 patienter, som alle får begge behandlinger (typisk med noget tid imellem og med tilfældig rækkefølge af behandlingerne).
Dvs. de samme patienter fremgår i de to grupper.

02402 Statistik (Polyteknisk grundlag)

Styrke og stikprøvestørrelse – Forsøgsdesign

- Krav til præcision
- Styrke og stikprøvestørrelse – En stikprøve
- Styrke og stikprøvestørrelse – To stikprøver

DTU Compute
Danmarks Tekniske Universitet
2800 Kgs. Lyngby

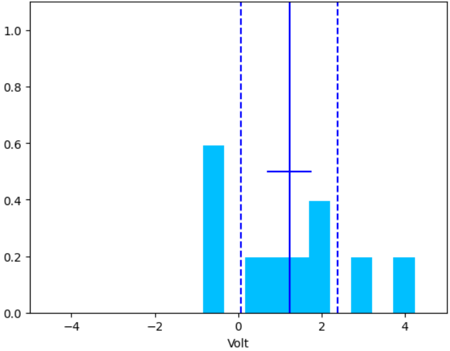
02402 Statistik (Polyteknisk grundlag)

- Krav til præcision
- Styrke og stikprøvestørrelse – En stikprøve
- Styrke og stikprøvestørrelse – To stikprøver

DTU Compute
Danmarks Tekniske Universitet
2800 Kgs. Lyngby

Eksempel - letbane

Lad os se på data fra letbane-eksemplet igen. Data måler spændingsfald henover et sted hvor spændingsfaldet bør være nul i gennemsnit.



Fra stikprøven (n=10):
$\bar{x} = 1.23$
$s = 1.62$

konfidensinterval:
$\bar{x} \pm t_{1-\alpha/2} \cdot (s/\sqrt{n}) =$
1.23 ± 1.16
$[0.07\,;\,2.9]$

I dette tilfælde har vi: $\hat{\mu} = 1.23 \pm 1.16$ volt ($\alpha = 0.05$).
Vi siger at **"Margin of Error" (ME)** er 1.16 (volt).

Eksempel - letbane

Konklusionen fra stikprøven blev at vi afviser nulhypotesen - dermed konkluderer vi at spændingsfaldet IKKE er nul (i gennemsnit).

Men chefen er ikke tilfreds og siger at vi må undersøge sagen bedre. Vores fejlmargin på ± 1.16 volt er for stor.

Vi planlægger derfor at lave en ny stikprøve.

Hvordan kan vi gøre det bedre end første gang?

Hvad kan vi gøre for at forbedre vores estimat af μ?

Kahoot!

(x1)

Forsøgsplanlægning med krav til præcisionen

Fejlmarginen (margin of error - ME) er defineret som

$$ME = t_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$$

For store nok stikprøver kan vi approximere $t_{1-\alpha/2}$ (fra t-fordeling) med $z_{1-\alpha/2}$ (fra normalfordeling).

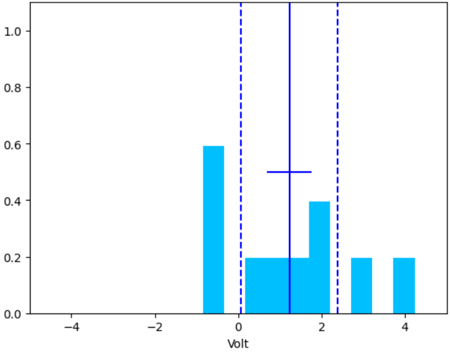
Method 3.63 The one-sample CI sample size formula

When σ is known or guessed at some value, we can calculate the sample size n needed to achieve a given margin of error, ME, with probability $1-\alpha$ as

$$n = \left(\frac{z_{1-\alpha/2}\cdot\sigma}{ME}\right)^2. \qquad (3\text{-}59)$$

Eksempel - letbane

Lad os estimere størrelsen på stikprøven hvis vi ønsker ME på højest ± 1 volt.
Vi antager at $s = 1.62$ er et udemærket gæt på den sande σ.



$$n = \left(\frac{z_{1-\alpha/2} \cdot \sigma}{ME}\right)^2 = \left(\frac{1.96 \cdot 1.62}{1}\right)^2 = 10.08$$

Da vi ikke kan lave 10.08 målinger runder vi op til 11
(i den nye stikprøve er $n = 11$).

02402 Statistik (Polyteknisk grundlag)

- Krav til præcision
- Styrke og stikprøvestørrelse – En stikprøve
- Styrke og stikprøvestørrelse – To stikprøver

DTU Compute
Danmarks Tekniske Universitet
2800 Kgs. Lyngby

Planlægning (Styrke og stikprøvestørrelse)

Vi skal nu snakke om forsøgsplanlægning og "styrke".

Til dette skal vi overveje de to typer af fejl, som vi introducerede sidste gang:

Type I: Afvisning af H_0, når H_0 er sand.

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

Type II: Accept (ikke afvisning) af H_0, når H_1 er sand.

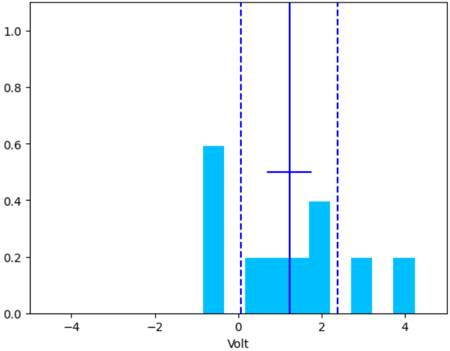
$$P(\text{Type II fejl}) = \beta$$

Eksempel: letbane

Chefen bliver nu lidt mere krævende. Hun siger at hvis det sande gennemsnitlige spændingsfald er ± 1 volt (eller mere), så vil vi gerne være mere sikre på at vi opdager det, med vores nye stikprøve.

Hvordan skal vi bære os ad?

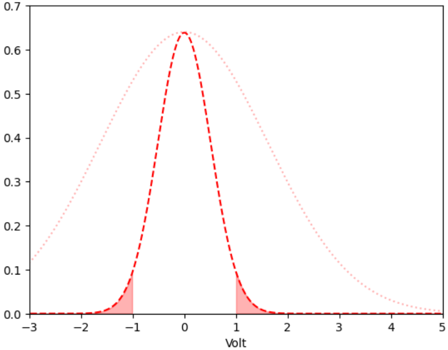
Vi gennemgår nu nogle forskellige **scenarier**.



Eksempel: letbane

Scenarie 1: Nulhypotesen er sand

- *Hvis* det sande gennemsnitlige spændingsfald *er* nul
- og vi laver en ny omgang målinger - denne gang med større n
- da vil risikoen for at forkaste nulhypotesen ved en fejl være α (fx 0.05, eller 5%)



De værdier af $\bar{x}$ som vi potentielt accepterer som værende "ikke significant forskellige fra nul" ligger i området $0 \pm ME$. Bredden af dette område afhænger af størrelsen på stikprøven (n).

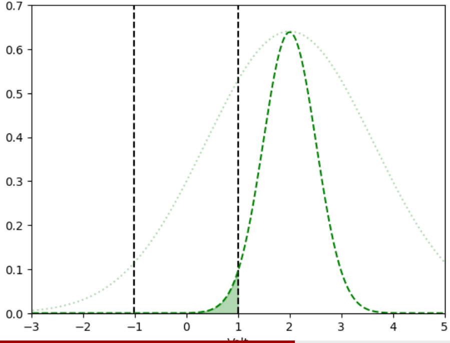
Kahoot!

(x2)

Eksempel: letbane

Scenarie 2: Modhypotesen er sand

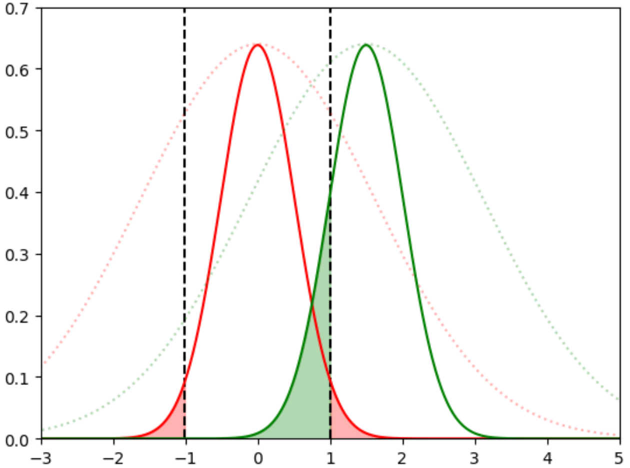
- *Hvis* det sande gennemsnitlige spændingsfald i virkeligheden er noget andet end nul - den sande værdi er μ_1
- og vi laver en ny omgang målinger - denne gang med større n
- Hvad er da risikoen (sandsynligheden β) for at acceptere (ikke afvise) nulhypotesen?



Ja, det kommer an på...
...hvor stor μ_1 egentlig er

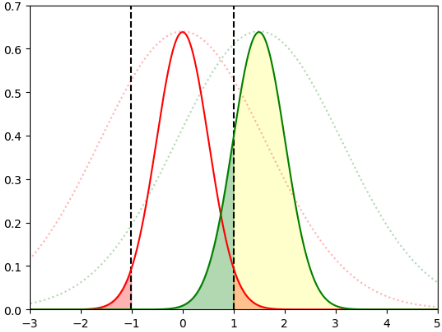
Kahoot!
(x2)

Type I og Type II fejl



α: det røde areal β: det grønne areal $(1-\beta)$ kaldes **styrken** (eng: **power**)

Forsøgsplanlægning: Styrke



- **Styrke** $= 1 - \beta$
- $P(\text{Type II fejl}) = \beta$
- Styrke = "Sandsynligheden for en korrekt afvisning of H_0".
- Styrke = "Sandsynligheden for at detektere en (påstået) effekt".
- Typiske værdier for ønsket styrke er 80%, 90%, .. (noget relativt stort)
- I praksis: Brug en scenarie-baseret tilgang

Kahoot! (x3)

Forsøgsplanlægning: Stikprøvestørrelsen n

Det store spørgsmål: Hvor stort skal n være?

Vi skal have nok observationer til at kunne detektere en relevant effekt med høj styrke $1-\beta$ (typisk mindst 80%):

Method 3.65 The one-sample sample size formula

For the one-sample t-test for given α, β and σ

$$n = \left(\sigma \frac{z_{1-\beta} + z_{1-\alpha/2}}{(\mu_0 - \mu_1)}\right)^2,$$

where $\mu_0 - \mu_1$ is the difference in means that we would want to detect and $z_{1-\beta}$, $z_{1-\alpha/2}$ are quantiles of the standard normal distribution.

Forsøgsplanlægning: Styrke

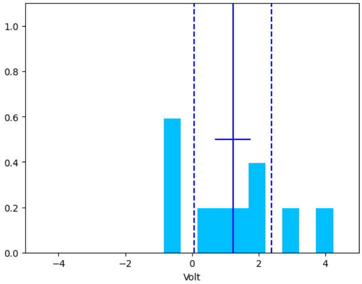
Hvis vi kender (eller antager) fire ud af de fem følgende størrelser, så kan vi finde den manglende:

- Stikprøvestørrelsen (sample size), n.
- Signifikansniveauet, α, som vi tester på.
- Forskellen i middelværdi (effekt-størrelsen), $\mu_0 - \mu_1$.
- Populationsstandardafvigelsen, σ.
- Styrken (power), $1 - \beta$.

Eksempel: letbane

Vi siger nu til chefen at vi kan designe den næste stikprøve således at vi vil have en 80% sandsynlighed for at "detektere" et spændingsfald på 1 volt (eller mere), ved et signifikansniveau $\alpha = 0.05$.

For at beregne den nødvendige stikprøvestørrelse, bruger vi (som informeret gæt): $\sigma = 1.62$, svarende til den spredning vi observerede i vores tidligere målinger.



Vi kommer da frem til at den næste stikprøve skal have 20.6 målinger:

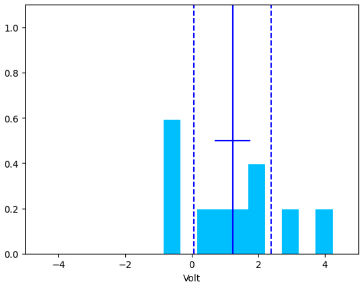
$$
\begin{aligned}
n &= \left(\sigma \frac{z_{1-\beta} + z_{1-\alpha/2}}{\mu_0 - \mu_1}\right)^2 \\
&= \left(1.62 \frac{z_{0.80} + z_{0.975}}{1}\right)^2 \\
&= 20.6
\end{aligned}
$$

Eksempel: letbane

"21 målinger bliver for dyrt", siger chefen.

"Kan vi nøjes med 15 målinger?"

Ja.. - men så får vores experiment en mindre styrke.



Ny styrkeberegning:

$$n = \left(\sigma \frac{z_{1-\beta} + z_{1-\alpha/2}}{\mu_0 - \mu_1}\right)^2$$

$$\Rightarrow$$

$$z_{1-\beta} = \sqrt{n \frac{(\mu_0 - \mu_1)^2}{\sigma^2}} - z_{0.975}$$

$$= 0.43$$

svarende til en styrke på 0.67

Python

Vi skal nu bruge nogle funktioner fra pakken `statsmodels.stats.power`

(`import statsmodels.stats.power as smp`)

Man kan bruge:

- `smp.TTestPower().solve_power(effect_size=delta/sd, alpha=0.05, power=0.80)`

- Gå til Python notebook "power_calculations.ipynb" i VS Code

02402 Statistik (Polyteknisk grundlag)

- Krav til præcision
- Styrke og stikprøvestørrelse – En stikprøve
- Styrke og stikprøvestørrelse – To stikprøver

DTU Compute
Danmarks Tekniske Universitet
2800 Kgs. Lyngby

Forsøgsplanlægning: To stikprøver

Det store spørgsmål: Hvor store skal n_1 og n_2 være?

Vi kan generalisere forsøgsplanlægningen til den situation hvor vi sammenligner to stikprøver. Stikprøvestørrelserne er nu givet ved n_1 og n_2. Vi vil antage at variansen (σ^2) er ens i de to populationer.

Formel for stikprøvestørrelse med to stikprøver

For en t-test med to stikprøver, hvor α, β og σ er givet:

$$n_1 = (k+1)\left(\sigma \frac{z_{1-\beta} + z_{1-\alpha/2}}{\mu_1 - \mu_2}\right)^2$$

Her er $\mu_1 - \mu_2$ den forskel i middelværdier, som vi ønsker at måle, medens $z_{1-\beta}$ og $z_{1-\alpha/2}$ er fraktiler i standardnormalfordelingen. Desuden er $k = n_1/n_2$ forholdet mellem det to stikprøvestørrelser, således at vi har $n_2 = n_1/k$.

(Denne formel er ikke med i bogen)

Python

Vi skal bruge nogle funktioner fra pakken `statsmodels.stats.power`

(`import statsmodels.stats.power as smp`)

Til stikprøvestørrelser for test med to stikprøver kan kan bruge:

- `smp.TTestIndPower().solve_power(effect_size=delta/sd, alpha=0.05, power=0.90, ratio=k)`

- Gå til Python notebook "power_calculations.ipynb" i VS Code

Dagsorden

Kahoot! (x4)

1. Opsummering fra sidste uge
2. Statistik med 2 stikprøver
   - t-test for differens mellem stikprøver
   - t-test med sammenvejet ("pooled") varians – Et alternativ
   - Konfidensintervallet for differensen
   - t-test med to parrede stikprøver (parret t-test)
3. Styrke og stikprøvestørrelse – Forsøgsdesign
   - Krav til præcision
   - Styrke og stikprøvestørrelse – En stikprøve
   - Styrke og stikprøvestørrelse – To stikprøver

Tjekliste

Efter i dag skal du kunne:

- Beregne standard error for en differens
- Udføre t-test for to stikprøver - både Welch og pooled
- Redegøre for forskellen mellem Welch og pooled t-test
- Beregne konfidensinterval for en differens
- Redegøre for hvornår man bør udføre en parret t-test og kunne udføre denne
- Beskrive hvad der menes med Type 1 og Type 2 fejl.
- Beskrive hvad der menes med statistisk styrke (power)
- Beregne den nødvendige stikprøvestørrelse n, ud fra α, β, σ og $\mu_0 - \mu_A$
- Python: Benytte funktioner til at udføre diverse t-tests
- Python: Benytte funktioner til at beregne stikprøvestørrelse, power, mm.