02402 Statistik (Polyteknisk grundlag)

Uge 12: Kategorisk data og andele

DTU Compute
Danmarks Tekniske Universitet
2800 Kgs. Lyngby

Dagsorden

- Praktiske informationer
- Opsummering / genopfriskning

1. Introduktion til andele og kategorisk data
   - Stokastisk variabel for en andel
2. Statistik for en andel
   - Konfidensinterval for en estimeret andel
   - Stikprøvestørrelse og forsøgsplanlægning
   - Hypotesetest for en andel
3. Sammenligning af andele i to grupper
   - Konfidensinterval for forskellen mellem to andele
   - Hypotesetest for forskellen mellem to andele
4. Sammenligning af andele i flere grupper
   - Hypotesetest for sammenligning af andele i flere grupper
5. Statistik for antalstabeller

02402 Statistik (Polyteknisk grundlag)

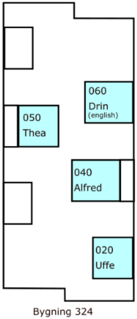
- Praktiske informationer
- Opsummering / genopfriskning

DTU Compute
Danmarks Tekniske Universitet
2800 Kgs. Lyngby

Praktiske informationer

Øvelser:

Vi fortsætter "klasseundervisning" i samme lokaler.



Eksamen:

Lørdag d. 20. december.

Tidspunkt og lokaler kommer først ca en uge før.

www.eksamensplan.dtu.dk

Næste gang ser vi på hvordan man udfylder **svarark/answer sheet**.

02402 Statistik (Polyteknisk grundlag)

- Praktiske informationer
- Opsummering / genopfriskning

DTU Compute
Danmarks Tekniske Universitet
2800 Kgs. Lyngby

Eksempel med diskret stokastisk variabel (uge 1)

En stokastisk variabel Y har følgende udfaldsrum og fordeling:

y	0	1
$f(y)$	0.30	0.70

Hvad er sandsynlighederne: $P(Y=0)$, $P(Y \leq 1)$, $P(Y=3)$?
(Hvad er *udfaldsummet* for Y?)

Tegn pdf(y) og cdf(y):

Hvad er $E[Y]$ og $V[Y]$?

Kahoot! (x4)

Stikprøvegennemsnittet, $\bar{Y}$, som stokastisk variabel (uge 4)

Vi betragter en stikprøve, $\{Y_1, Y_2, ..., Y_n\}$.

Den **enkelte observation** beskrives med en **stokastisk variabel**: Y_i.

Stikprøvegennemsnittet: $\bar{Y} = \frac{1}{n}(Y_i + ... + Y_n)$
$\bar{Y}$ er selv en **stokastisk variabel** (med sin egen fordeling).

Stikprøvegennemsnittet er en *lineær kombination* af stokastiske variable:

$$\bar{Y} = \frac{1}{n}Y_1 + \frac{1}{n}Y_2 + ... + \frac{1}{n}Y_n$$

Fordeling for stikprøvegennemsnittet, $\bar{Y}$ (uge 4)

Stikprøvegennemsnittet er en lineær kombination af normalfordelte stokastiske variable:

$$\bar{Y} = \frac{1}{n}Y_1 + \frac{1}{n}Y_2 + \ldots + \frac{1}{n}Y_n$$

Regneregel for Middelværdien af $\bar{Y}$:

$$\mathsf{E}[\bar{Y}] = \frac{1}{n}\mathsf{E}[Y_1] + \frac{1}{n}\mathsf{E}[Y_2] + \ldots + \frac{1}{n}\mathsf{Y}[X_n] = n\frac{1}{n}\mu = \mu$$

Regneregel for Variansen for $\bar{Y}$:

$$\mathsf{V}[\bar{Y}] = \frac{1}{n^2}\mathsf{V}[Y_1] + \frac{1}{n^2}\mathsf{V}[Y_2] + \ldots + \frac{1}{n^2}\mathsf{V}[Y_n] = n\frac{1}{n^2}\sigma^2 = \frac{\sigma^2}{n}$$

Men hvad med fordelingen af $\bar{Y}$:

Hvordan var det nu..?

Kahoot!

(x2)

Introduktion til andele og kategorisk data

- Stokastisk variabel for en andel

DTU Compute
Danmarks Tekniske Universitet
2800 Kgs. Lyngby

Forskellige analyser og data-typer

Middelværdier i kvantitative data

- Hypotesetest for én middelværdi baseret på én stikprøve
- Hypotesetest for to middelværdier baseret på to stikprøver
- Hypotesetest for flere middelværdier baseret på flere stikprøver (ANOVA).

Forskellige analyser og data-typer

Middelværdier i kvantitative data

- Hypotesetest for én middelværdi baseret på én stikprøve
- Hypotesetest for to middelværdier baseret på to stikprøver
- Hypotesetest for flere middelværdier baseret på flere stikprøver (ANOVA).

I dag: Andele i kvalitative data

- Hypotesetest for én andel baseret på én stikprøve.
- Hypotesetest for to andele baseret på to stikprøver.
- Hypotesetest for flere andele baseret på flere stikprøver.

Eksempel på andele

Venstrehåndede:

Andelen af venstrehåndede
(i Danmark vs Sverige)

eller:

Kvindelige ingeniørstuderende:

Andelen af kvindelige ingeniørstuderende
(på forskellige studieretninger)

Kahoot!
(x2)

02402 Statistik (Polyteknisk grundlag)

- Stokastisk variabel for en andel

DTU Compute
Danmarks Tekniske Universitet
2800 Kgs. Lyngby

Estimation af andele

- Vi definerer den stokastiske variabel P som antallet af "succes'er" (X) ud af et totalt antal (n):

$$P = \frac{X}{n}$$

- Ud fra stikprøve-data med x "succes'er" (stikprøve-størrelse n) estimerer vi andelen med:

$$\hat{p} = \frac{x}{n}$$

Bemærk:

- $P \in [0;1]$.
- p er den "sande" populations-sandsynlighed for at få en "succes".

Binomialfordelingen

Antallet af "succes'er" (X) følger en binomial fordeling med tæthedsfunktion:

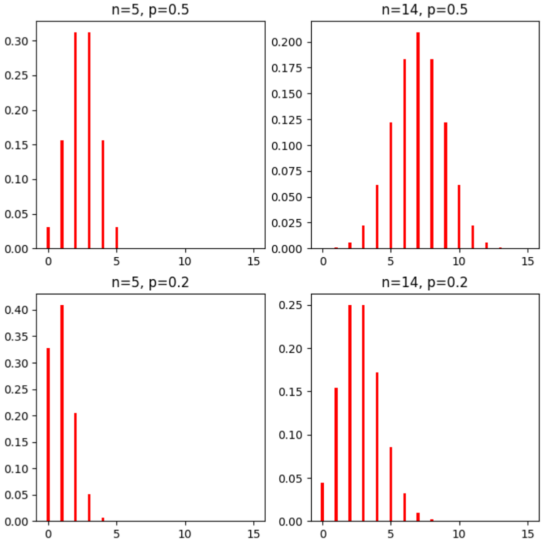
$$f(x;n,p) = \binom{n}{x} p^x (1-p)^{n-x}$$

Middelværdi og varians i binomialfordelingen, sektion 2.21

$$\begin{aligned} \mathbf{E}[X] &= np \\ \mathbf{V}[X] &= np(1-p) \end{aligned}$$

Se evt. Appendix

Eksempler på binomialfordelinger (fra uge 2)



Middelværdi og varians for andele

Middelværdi og varians for andelen P:

$$\begin{aligned} \mathbf{E}[P] = \mathbf{E}\left[\frac{X}{n}\right] &= \frac{np}{n} = p \\ \mathbf{V}[P] = \mathbf{V}\left[\frac{X}{n}\right] &= \frac{1}{n^2}\mathbf{V}[X] = \frac{p(1-p)}{n} \end{aligned}$$

Vi kan dermed definere:

$$\sigma_P = \sqrt{\frac{p(1-p)}{n}}$$

Bemærk:

σ_P er størst når $p = 1/2$.

Sammenhæng med $\bar{Y}$

Hvis Y er en stokastisk variabel:

y	0	1
$f(y)$	(1-p)	p

Så er den stokastiske variabel for en andel, P, det samme som den stokastiske variabel $\bar{Y}$.

Andel $=$ *gennemsnit af nuller of ét-taller*.

02402 Statistik (Polyteknisk grundlag)

Statistik for en andel

- Konfidensinterval for en estimeret andel
- Stikprøvestørrelse og forsøgsplanlægning
- Hypotesetest for en andel

DTU Compute
Danmarks Tekniske Universitet
2800 Kgs. Lyngby

Stor stikprøve: Central Limit Theorem

Hvis **n er stor** gælder *Central Limit Theorem* og andelen $P\ (=\bar{Y})$ følger en **normalfordeling**, med parametre:

$$\begin{aligned} \mu_P &= \mathbf{E}[P] = p \\ \sigma_P^2 &= \mathbf{V}[P] = \frac{p(1-p)}{n} \\ \sigma_P &= \sqrt{\frac{p(1-p)}{n}} \end{aligned}$$

> **Remark 7.4**
>
> As a rule of thumb the normal distribution is a good approximation of the binomial distribution if np and $n(1-p)$ are both greater than 15.

OBS: Ny tommelfingerregel

Normalfordelingsantagelsen for en andel

- Gå til Python notebook "proportions.ipynb" i VS Code
    1) Normal approximation of binomialdistribution

02402 Statistik (Polyteknisk grundlag)

- Konfidensinterval for en estimeret andel
- Stikprøvestørrelse og forsøgsplanlægning
- Hypotesetest for en andel

DTU Compute
Danmarks Tekniske Universitet
2800 Kgs. Lyngby

Konfidensinterval for en andel, når stikprøven er **stor**

Method 7.3 Proportion estimate and confidence interval

The best estimate of the probability p of belonging to a category (the population proportion) is the sample proportion

$$\hat{p} = \frac{x}{n}, \tag{7-8}$$

where x is the number of observations in the category and n is the total number of observations.

A large sample $(1-\alpha)100\%$ confidence interval for p is given as

$$\hat{p} \pm z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}. \tag{7-9}$$

Remark 7.4

As a rule of thumb the normal distribution is a good approximation of the binomial distribution if np and $n(1-p)$ are both greater than 15.

Konfidensinterval for en andel, når stikprøven er stor

Lidt notation:

$$SE_{\hat{p}} = \hat{\sigma}_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

(vi bruger $\hat{p}$ i stedet for p)

Konfidensinterval:

$$\hat{p} \pm z_{1-\alpha/2} SE_{\hat{p}}$$

$z_{1-\alpha/2}$ er en fraktil fra **standard normalfordelingen**.
For $\alpha = 0.05$ er konfidensintervallet:

$$\hat{p} \pm 1.96 SE_{\hat{p}}$$

Konfidensinterval for en andel, når stikprøven er lille

Remark 7.7 What about small samples then?

There exist several ways of expressing a valid confidence interval for p in small sample cases, that is, when either $np \leq 15$ or $n(1-p) \leq 15$. We mention three of these here - only for the last one we give the explicit formula:

Continuity correction

The so-called *continuity correction* is a general approach to making the best approximation of discrete probabilities (in this case the binomial probabilities) using a continuous distribution, (in this case the normal distribution). We do not give any details here.

Exact intervals

Probably the most well known of such small sample ways of obtaining a valid confidence interval for a proportion is the so-called *exact* method based on actual binomial probabilities rather than a normal approximation. It is not possible to give a simple formula for these confidence limits, and we will not explain the details here, but simply note that they can be obtained by the Python function `stats.binomtest`. These will be valid no matter the size of n and p.

"Plus 2"-approach

Finally, a simple approach to a good small sample confidence interval for a proportion, will be to us the simple formula given above in Method 7.3, but applied to $\tilde{x} = x + 2$ and $\tilde{n} = n + 4$.

"Plus 2"-tilgangen

"Plus 2"-tilgangen

Hvis stikprøven ikke er stor, anvendes $\tilde{x} = x + 2$ og $\tilde{n} = n + 4$.

$$\tilde{p} = \tilde{x}/\tilde{n}$$

I konfidensintervallet indsættes:

$$\tilde{p} \pm z_{1-\alpha/2}\,\sqrt{\tilde{p}(1-\tilde{p})/\tilde{n}}$$

Eksempel: Antal venstrehåndede, konfidensinterval

I en stikprøve på 100 personer observeres det at 15 er venstrehåndede og 85 er højrehåndede.

1) Beregn et 95% konfidensinterval for andelen af venstrehåndede.

2) Hvad bliver svaret, hvis man i stikprøven havde observeret 3 ventrehåndede og 97 højrehåndede?

- Gå til dagens Python notebook i VS Code
  - "Example: Confidence-interval of proportion for left-handed"

02402 Statistik (Polyteknisk grundlag)

- Konfidensinterval for en estimeret andel
- Stikprøvestørrelse og forsøgsplanlægning
- Hypotesetest for en andel

DTU Compute
Danmarks Tekniske Universitet
2800 Kgs. Lyngby

Fejlmarginen (ME: Margin of Error)

Fejlmarginen, *Margin of Error*, ME

ved et $(1-\alpha)$-konfidensniveau er:

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

hvor vi estimerer p med $\hat{p} = \frac{x}{n}$.

Fejlmarginen:

- Svarer til den halve bredde af $(1-\alpha)$-konfidensintervallet.
- Beskriver den forventede *præcision* (mindst ønskede præcision) på estimatet $\hat{p}$.

Præcision og stikprøvestørrelse

Method 7.13 Sample size formula for the CI of a proportion

Given some "guess" (scenario) of the size of the unknown p, and given some requirement to the ME-value (required expected precision) the necessary sample size is then

$$n = p(1-p)\left(\frac{z_{1-\alpha/2}}{ME}\right)^2. \tag{7-24}$$

If p is unknown, a worst case scenario with $p = 1/2$ is applied and necessary sample size is

$$n = \frac{1}{4}\left(\frac{z_{1-\alpha/2}}{ME}\right)^2. \tag{7-25}$$

Eksempel: Antal venstrehåndede, stikprøvestørrelse

I en stikprøve på 100 personer observeres det at 15 er venstrehåndede og 85 er højrehåndede.

1) Hvis man ønsker at planlægge en ny stikprøve og gerne vil have ME $=$ 0.01, hvor stor en stikprøve bør man så planlægge?

2) Hvad bliver svaret, hvis man ikke havde den tidligere stikprøve (intet gæt på p)

- Gå til dagens Python notebook i VS Code
  - "Example: Confidence-interval of proportion for left-handed"

Visual Studio Code

02402 Statistik (Polyteknisk grundlag)

- Konfidensinterval for en estimeret andel
- Stikprøvestørrelse og forsøgsplanlægning
- Hypotesetest for en andel

DTU Compute
Danmarks Tekniske Universitet
2800 Kgs. Lyngby

Trin i en hypotesetest – Overblik (repetition)

1. Opstil nulhypotesen og vælg et signifikansniveau α
2. Beregn den observerede teststørrelse
3. Beregn p-værdien ud fra den observerede teststørrelse og den relevante fordeling
4. Sammenlign p-værdien med signifikansniveauet α og konkludér

Alternativt: Sammenlign den observerede teststørrelse med kritiske værdier og konkludér.

Hypotesetest for en andel

Vi betragter en nul- og modhypotese for en andel p og vælger et signifikansniveau α:

$$\begin{aligned} H_0: &\quad p = p_0, \\ H_1: &\quad p \neq p_0. \end{aligned}$$

Som sædvanligt afvises H_0 eller accepteres H_0.

Teststørrelsen Z

Theorem 7.10

In the large sample case the random variable Z follows approximately a standard normal distribution

$$Z = \frac{X - np_0}{\sqrt{np_0(1-p_0)}} \sim N(0,1), \tag{7-18}$$

when the null hypothesis is true. As a rule of thumb, the result will be valid when both $np_0 > 15$ and $n(1-p_0) > 15$.

$$z_{\text{obs}} = \frac{x - np_0}{\sqrt{np_0(1-p_0)}} = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}$$

Hypotesetest for en andel

Method 7.11 One sample proportion hypothesis test

1. Compute the test statistic using Equation (7-16)

$$z_{\text{obs}} = \frac{x - np_0}{\sqrt{np_0(1-p_0)}}$$

2. Compute evidence against the *null hypothesis*

$$H_0 : p = p_0, \tag{7-19}$$

vs. the *the alternative hypothesis*

$$H_1 : p \neq p_0, \tag{7-20}$$

by the

$$p\text{-value} = 2 \cdot P(Z > |z_{\text{obs}}|). \tag{7-21}$$

where the standard normal distribution $Z \sim N(0, 1^2)$ is used

3. If the 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) $\pm z_{1-\alpha/2}$:
if $|z_{\text{obs}}| > z_{1-\alpha/2}$ we reject H_0, otherwise we accept H_0

Eksempel: Antal venstrehåndede, hypotesetest

I en stikprøve på 100 personer observeres det at 15 er venstrehåndede og 85 er højrehåndede.

1) Udfør en hypotesetest for nulhypotesen $H_0: \quad p = 0.5$

- Vi har brug for **Python** (eller lign) for at finde fraktiler i normalfordelingen (kritiske værdier, p-værdier)
- Gå til dagens Python notebook i VS Code
  - "Example: Hypothesis test for proportion of left-handed"

Python kommando du skal kende (!):

```
z_obs,p_value = smprop.proportions_ztest(count=15, nobs=100,
value=0.5, prop_var=0.5)
```

Sammenligning af andele i to grupper

- Konfidensinterval for forskellen mellem to andele
- Hypotesetest for forskellen mellem to andele

DTU Compute
Danmarks Tekniske Universitet
2800 Kgs. Lyngby

(2 x 2)-tabeller

	Group 1	Group 2
Success	x_1	x_2
Failure	$n_1 - x_1$	$n_2 - x_2$
Total	n_1	n_2

Estimat af andel i gruppe 1: $\hat{p}_1$

Estimat af andel i gruppe 2: $\hat{p}_2$

Hvordan sammenligner vi $\hat{p}_1$ og $\hat{p}_2$?

Se på **differensen**: $(\hat{p}_1 - \hat{p}_2)$

Standardfejl for **differens** mellem to andele

Remark 7.16

The standard error in Method 7.15 can be calculated by

$$V(\hat{p}_1 - \hat{p}_2) = V(\hat{p}_1) + V(\hat{p}_2) = \hat{\sigma}^2_{\hat{p}_1} + \hat{\sigma}^2_{\hat{p}_2}, \tag{7-31}$$

$$\hat{\sigma}_{\hat{p}_1 - \hat{p}_2} = \sqrt{V(\hat{p}_1 - \hat{p}_2)} = \sqrt{\hat{\sigma}^2_{\hat{p}_1} + \hat{\sigma}^2_{\hat{p}_2}}. \tag{7-32}$$

Notice, that the standard errors are added (before the square root) such that the standard error of the difference is larger than the standard error for the observed proportions alone. Therefore in practice the estimate of the difference $\hat{p}_1 - \hat{p}_2$ will often be further from the true difference $p_1 - p_2$ than $\hat{p}_1$ will be from p_1 or $\hat{p}_2$ will be from p_2.

02402 Statistik (Polyteknisk grundlag)

- Konfidensinterval for forskellen mellem to andele
- Hypotesetest for forskellen mellem to andele

DTU Compute
Danmarks Tekniske Universitet
2800 Kgs. Lyngby

Konfidensinterval for forskellen mellem to andele

Method 7.15

An estimate of the standard error of the estimator $\hat{p}_1 - \hat{p}_2$ is

$$\hat{\sigma}_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}. \tag{7-29}$$

The $(1-\alpha)100\%$ confidence interval for the difference $p_1 - p_2$ is

$$(\hat{p}_1 - \hat{p}_2) \pm z_{1-\alpha/2} \cdot \hat{\sigma}_{\hat{p}_1 - \hat{p}_2}. \tag{7-30}$$

This confidence interval requires independent random samples for the two groups and large enough sample sizes n_1 and n_2. A rule of thumb is that $n_i p_i \geq 10$ and $n_i(1-p_i) \geq 10$ for $i = 1, 2$, must be satisfied.

OBS: Ny tommelfingerregel

02402 Statistik (Polyteknisk grundlag)

- Konfidensinterval for forskellen mellem to andele
- Hypotesetest for forskellen mellem to andele

DTU Compute
Danmarks Tekniske Universitet
2800 Kgs. Lyngby

Hypotesetest for forskellen mellem to andele

Nulhypotesen

De to stikprøver kommer fra en underliggende population med **samme andel** af succes'er

$$\begin{aligned} H_0: &\quad p_1 = p_2, \\ H_1: &\quad p_1 \neq p_2, \end{aligned}$$

Teststørrelsen

$$z_{\text{obs}} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}, \quad hvor \quad \hat{p} = \frac{x_1 + x_2}{n_1 + n_2}$$

Hypotesetest for forskellen mellem to andele

Method 7.18 Two sample proportions hypothesis test

The two-sample hypothesis test for comparing two proportions is given by the following procedure:

1. Compute, with $\hat{p} = \frac{x_1+x_2}{n_1+n_2}$, the test statistic

$$z_{\text{obs}} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \tag{7-37}$$

2. Compute evidence against the *null hypothesis*

$$H_0 : p_1 = p_2, \tag{7-38}$$

vs. the *the alternative hypothesis*

$$H_1 : p_1 \neq p_2, \tag{7-39}$$

by the

$$p\text{-value} = 2 \cdot P(Z > |z_{\text{obs}}|). \tag{7-40}$$

where the standard normal distribution $Z \sim N(0, 1^2)$ is used

3. If the 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) $\pm z_{1-\alpha/2}$:
if $|z_{\text{obs}}| > z_{1-\alpha/2}$ we reject H_0, otherwise we accept H_0

Eksempel (2 x 2)-tabel

Er der en sammenhæng mellem brugen af p-piller og risikoen for blodpropper i hjertet?

I et studie (USA, 1975) undersøgtes sammenhængen mellem p-piller og risikoen for blodpropper i hjertet.

	Contraceptive pill	No pill
Blood clot	23	35
No blood clot	34	132
Total	57	167

Undersøg om der er sammenhæng mellem brug af p-piller og risiko for blodpropper i hjertet. Anvend signifikansniveauet $\alpha = 5\%$.

Kahoot!

(x5)

Eksempel (2 x 2)-tabel

I et studie (USA, 1975) undersøgtes sammenhængen mellem p-piller og risikoen for blodpropper i hjertet.

	Contraceptive pill	No pill
Blood clot	23	35
No blood clot	34	132
Total	57	167

Estimater i hver stikprøve

$$\hat{p}_1 = \frac{23}{57} = 0.4035, \;\; \hat{p}_2 = \frac{35}{167} = 0.2096$$

Fælles estimat:

$$\hat{p} = \frac{23+35}{57+167} = \frac{58}{224} = 0.2589$$

Eksempel (2 x 2)-tabel, hypotesetest

	Contraceptive pill	No pill
Blood clot	23	35
No blood clot	34	132
Total	57	167

- Gå til dagens Python notebook i VS Code
  - "Example: Contraceptive pills and risk of blood clots"

Visual Studio Code

Python kommando for 2x2 tabel:

```
z_obs,p_value = smprop.proportions_ztest(count = [23, 35],
nobs = [57, 167], value=0, prop_var=0)
```

Sammenligning af andele i flere grupper

- Hypotesetest for sammenligning af andele i flere grupper

DTU Compute
Danmarks Tekniske Universitet
2800 Kgs. Lyngby

(2 x c)-tabeller

	Group 1	Group 2	...	Group c	*Total*
Success	x_1	x_2	...	x_c	x
Failure	$n_1 - x_1$	$n_2 - x_2$	...	$n_c - x_c$	$n - x$
Total	n_1	n_2	...	n_c	n

(2 x c)-tabel (c = antal grupper)

Hvordan tester vi om der er forskel på grupperne?

Vi må opstille en nulhypotese der reflekterer situationen "ingen forskel", dvs. alle grupper har samme andel.

02402 Statistik (Polyteknisk grundlag)

- Hypotesetest for sammenligning af andele i flere grupper

DTU Compute
Danmarks Tekniske Universitet
2800 Kgs. Lyngby

Nulhypotesen

	Group 1	Group 2	...	Group c	*Total*
Success	x_1	x_2	...	x_c	x
Failure	$n_1 - x_1$	$n_2 - x_2$	...	$n_c - x_c$	$n - x$
Total	n_1	n_2	...	n_c	n

Nulhypotesen

Man er interesseret i at teste nulhypotesen:

$$H_0 : \quad p_1 = p_2 = ... = p_c = p$$

mod den alternative hypotese om at disse andele ikke er ens (dvs. mindst én er anderledes).

Ud fra nulhypotesen estimeres $\hat{p}$

Fælles (gennemsnitligt) estimat:

Under nulhypotesen er estimatet for p:

$$\hat{p} = \frac{x}{n}$$

Brug $\hat{p}$ til at beregne *forventet antal* i hver celle:

Hvis nulhypotesen er sand, så forventer vi at den j'te gruppe har $n_j \cdot \hat{p}$ *successer* og $n_j \cdot (1 - \hat{p})$ *fiaskoer*

Ud fra nulhypotesen beregnes forventede antal

Tabel med det *forventede* antal i de c stikprøver:

e_{ij}	Stikprøve 1	Stikprøve 2	...	Stikprøve c	Total
Succes	e_{11}	e_{12}	...	e_{1c}	x
Fiasko	e_{21}	e_{22}	...	e_{2c}	$n-x$
Total	n_1	n_2	...	n_c	n

Generel formel for beregning af forventede værdier i antalstabeller:

$$e_{ij} = \frac{(\text{Rækketotal } i) \cdot (\text{Kolonnetotal } j)}{\text{total}}$$

Teststørrelsen

Vi opstiller en teststørrelse der summerer de kvadrerede afvigelser fra det forventede antal:

Teststørrelsen bliver

$$\chi^2_{\text{obs}} = \sum_{i=1}^{2}\sum_{j=1}^{c}\frac{(o_{ij}-e_{ij})^2}{e_{ij}}$$

hvor o_{ij} er det *observerede* antal i celle (i,j) og e_{ij} er det *forventede* (*expected*) antal i celle (i,j).

Teststørrelsen følger en χ^2-fordeling

Stikprøvefordeling for teststørrelsen (under H_0):

χ^2-fordeling med $(c-1)$ frihedsgrader (tilnærmelsesvis)

Metode med kritiske værdier:

Hvis $\chi^2_{\text{obs}} > \chi^2_{1-\alpha}(c-1)$, så afvises nulhypotesen.

Tommelfingerregel for om testen er valid:

Alle forventede værdier $e_{ij} \geq 5$.

Hypotesetest for sammenligning af andele i flere grupper

Method 7.20 The multi-sample proportions χ^2-test

The hypothesis

$$H_0 : p_1 = p_2 = \ldots = p_c = p, \tag{7-45}$$

can be tested using the test statistic

$$\chi^2_{\text{obs}} = \sum_{i=1}^{2} \sum_{j=1}^{c} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}, \tag{7-46}$$

where o_{ij} is the observed number in cell (i, j) and e_{ij} is the expected number in cell (i, j).

The test statistic χ^2_{obs} should be compared with the χ^2-distribution with $c - 1$ degrees of freedom.

The χ^2-distribution is approximately the sampling distribution of the statistics under the null hypothesis. The rule of thumb is that it is valid when all the computed expected values are at least 5: $e_{ij} \geq 5$.

OBS: Ny tommelfingerregel

Eksempel - χ^2-test

De *observerede* værdier o_{ij}

	Contraceptive pill	No pill	*Total*
Blood clot	23	35	58
No blood clot	34	132	166
Total	57	167	224

Beregn de *forventede* værdier e_{ij}

	Contraceptive pill	No pill	*Total*
Blood clot	Kahoot!		58
No blood clot			166
Total	57	167	224

Eksempel - χ^2-test

De *forventede* værdier e_{ij}:

	Contraceptive pill	No pill	*Total*
Blood clot	14.76	42.24	58
No blood clot	43.24	123.76	166
Total	57	167	224

Tjek at alle de forventede værdier er ≥ 5!

Teststørrelsen (husk at inkludere alle celler):

$$\chi^2_{\text{obs}} = \frac{(23-14.76)^2}{14.76} + \frac{(34-42.24)^2}{42.24} + \frac{(35-43.24)^2}{43.24} + \frac{(132-123.76)^2}{123.76}$$
$$= 8.33$$

Eksempel - χ^2-test

Beregnet teststørrelse:

$\chi^2_{\text{obs}} = 8.33$

Den kritiske værdi:

$\chi^2_{1-\alpha}(c-1)$ for $\alpha = 0.05$ og $c = 2$ (2 stikprøver): 3.841

```
stats.chi2.ppf(0.95, df = (2-1))
```

p-værdi:

$P(\chi^2 \geq 8.33) = 0.0039$

```
1 - stats.chi2.cdf(8.33, df = (2-1))
```

Konklusion:

Vi afviser nulhypotesen.

02402 Statistik (Polyteknisk grundlag)

Statistik for antalstabeller

DTU Compute
Danmarks Tekniske Universitet
2800 Kgs. Lyngby

Eksempel: Stemmefordeling, (r x c)-tabel

(3x3)-tabel for stemmefordeling på kandidater:

	4 weeks before	2 weeks before	1 week before	*Row total*
Candidate 1	79	91	93	263
Candidate 2	84	66	60	210
Undecided	37	43	47	127
Column total	200	200	200	600

Er stemmefordelingen ens i de 3 surveys?

Vi må opstille en nulhypotese der reflekterer situationen "ingen forskel", dvs. alle stikprøver (surveys) har samme fordeling af stemmer:

$$H_0: \quad p_{i1} = p_{i2} = p_{i3}, \quad i = 1, 2, 3.$$

χ^2-test – uanset typen af tabel

Method 7.22 The $r \times c$ frequency table χ^2-test

For an $r \times c$ table the hypothesis

$$H_0 : p_{i1} = p_{i2} = \ldots = p_{ic} = p_i, \text{ for all rows } i = 1, 2, \ldots, r, \tag{7-54}$$

is tested using the test statistic

$$\chi^2_{\text{obs}} = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}. \tag{7-55}$$

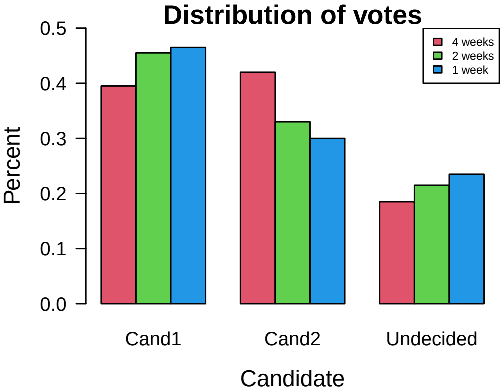
where o_{ij} is the observed number in cell (i, j) and e_{ij} is the expected number in cell (i, j). This test statistic should be compared with the χ^2-distribution with $(r-1)(c-1)$ degrees of freedom and the hypothesis is rejected at significance level α if

$$\chi^2_{\text{obs}} > \chi^2_{1-\alpha}\big((r-1)(c-1)\big). \tag{7-56}$$

Nu er antallet af frihedsgrader $(r-1)\cdot(c-1)$.

Der gælder stadig tommelfingerregel: $e_{ij} \geq 5$ for alle forventede værdier.

Eksempel: Stemmefordeling, (r x c)-tabel



Ændrer fordelingen sig "signifikant" over tid?

For at svare på dette udfører vi en χ^2-test. Vi skal beregne alle de forventede værdier e_{ij}, for at beregne en værdi χ^2_{obs}.

Eksempel: Stemmefordeling, (r x c)-tabel

	4 weeks before	2 weeks before	1 week before	*Row total*
Candidate 1	79	91	93	263
Candidate 2	84	66	60	210
Undecided	37	43	47	127
Column total	200	200	200	600

Kahoot! (x2)

- Gå til dagens Python notebook i VS Code
    - "Example: Example: Candidate votes over time"

Visual Studio Code

Python:

```
poll = np.array([[79, 91, 93], [84, 66, 60], [37, 43, 47]])

chi2, p_val, dof, expected = stats.chi2_contingency(poll,
correction=False)
```

χ^2-test – alternativ formulering af H_0

Samme teststørrelse og test - anden formulering af H_0:

$$\begin{aligned} &H_0 : \text{"The two variables are independent"}, \\ &H_1 : \text{"The two variables are not independent (they are associated)"}. \end{aligned} \tag{7-59}$$

Theorem 7.24

To test if two categorical variables are independent the null hypothesis

$$H_0 : p_{ij} = p_{i.}p_{.j} \text{ for all } i, j, \tag{7-60}$$

where $p_{i.} = \sum_{j=1}^{c} p_{ij}$ is the proportion of row i and $p_{.j} = \sum_{i=1}^{r} p_{ij}$ is the proportion of column j, is tested.

The p-value for the observed result under this null hypothesis is calculated using the χ^2 test statistic from Method 7.22.

Overview

- Praktiske informationer
- Opsummering / genopfriskning

1. Introduktion til andele og kategorisk data
    - Stokastisk variabel for en andel
2. Statistik for en andel
    - Konfidensinterval for en estimeret andel
    - Stikprøvestørrelse og forsøgsplanlægning
    - Hypotesetest for en andel
3. Sammenligning af andele i to grupper
    - Konfidensinterval for forskellen mellem to andele
    - Hypotesetest for forskellen mellem to andele
4. Sammenligning af andele i flere grupper
    - Hypotesetest for sammenligning af andele i flere grupper
5. Statistik for antalstabeller