02402 Statistik (Polyteknisk grundlag)

Uge 7: Simulering

DTU Compute
Danmarks Tekniske Universitet
2800 Kgs. Lyngby

Dagsorden

1. Til formelsamlingen/appendix
2. Introduktion til simulation
3. Fejlophobning
4. Introduktion til den generelle lineære model

02402 Statistik (Polyteknisk grundlag)

Til formelsamlingen/appendix

DTU Compute
Danmarks Tekniske Universitet
2800 Kgs. Lyngby

Python kommandoer for t-test

t-test med én stikprøve:

- `stats.ttest_1samp(data, pop_mean = mu0)`

Welch t-test med to stikprøve:

- `stats.ttest_ind(dataA, dataB, equal_var = False)`

Pooled t-test med to stikprøve:

- `stats.ttest_ind(dataA, dataB, equal_var = True)`

Parret t-test med to stikprøver:

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

Kahoot!
(x5)

Desuden kan konfidens interval beregnes med (her vises 95% CI):

- `stats.ttest_****(...).confidence_interval(0.95)`

Python kommandoer til styrke-beregninger

Én stikprøve:

- `smp.TTestPower().solve_power(...)`

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

To stikprøver:

- `smp.TTestIndPower().solve_power(...)`
- angiv desuden:
  `nobs1` $= n_1$
  og
  `ratio` $= n_2/n_1 (= 1/k)$

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

$k = n_1/n_2$

(denne formel er ikke i bogen)

OBS: Python kommandoerne beregner ikke præcis det samme som formlerne!

Python kommandoer til styrke-beregninger

Én stikprøve (udklip fra dokumentation):

```
TTestPower.solve_power(
    effect_size=None,
    nobs=None,
    alpha=None,
    power=None,
    alternative='two-sided'
) ¶                                                    [source]
```

solve for any one parameter of the power of a one sample t-test

for the one sample t-test the keywords are:

effect_size, nobs, alpha, power

Exactly one needs to be `None`, all others need numeric values.

Python kommandoer til styrke-beregninger

To stikprøver (udklip fra dokumentation):

```
TTestIndPower.solve_power(
    effect_size=None,
    nobs1=None,
    alpha=None,
    power=None,
    ratio=1.0,
    alternative='two-sided'
) ¶                                                    [source]
```

solve for any one parameter of the power of a two sample t-test

for t-test the keywords are:

effect_size, nobs1, alpha, power, ratio

exactly one needs to be `None`, all others need numeric values

Kahoot!

(x1)

02402 Statistik (Polyteknisk grundlag)

Introduktion til simulation

DTU Compute
Danmarks Tekniske Universitet
2800 Kgs. Lyngby

Motivation

Vi har allerede brugt simulering:

- Til at simulere ternigekast
- Til at motivere CLT (Central Limit Theorem)
- Til at afprøve diverse scenarier relateret til stikprøver

Vi har de sidste par uger beskæftiget os med hypotesetest for middelværdi(er).

Med simulering kan vi undersøge mange andre størrelser - også tilfældige størrelser, som følger mere komplicerede fordelinger.

Hvad er simulering egentlig?

- **Stokastisk simulering** - "sandsynligheds calculus værktøj"
- (Pseudo)-tilfældige tal genereret af en computer.
- En *tilfældighedsgenerator* er en algoritme, der kan generere pseudo-tilfældige tal.
- Algoritmen kræver et såkaldt *seed* (ofte bruges klokkeslet, hvis andet seed ikke er angivet).

Simulering fra forskellige fordelinger

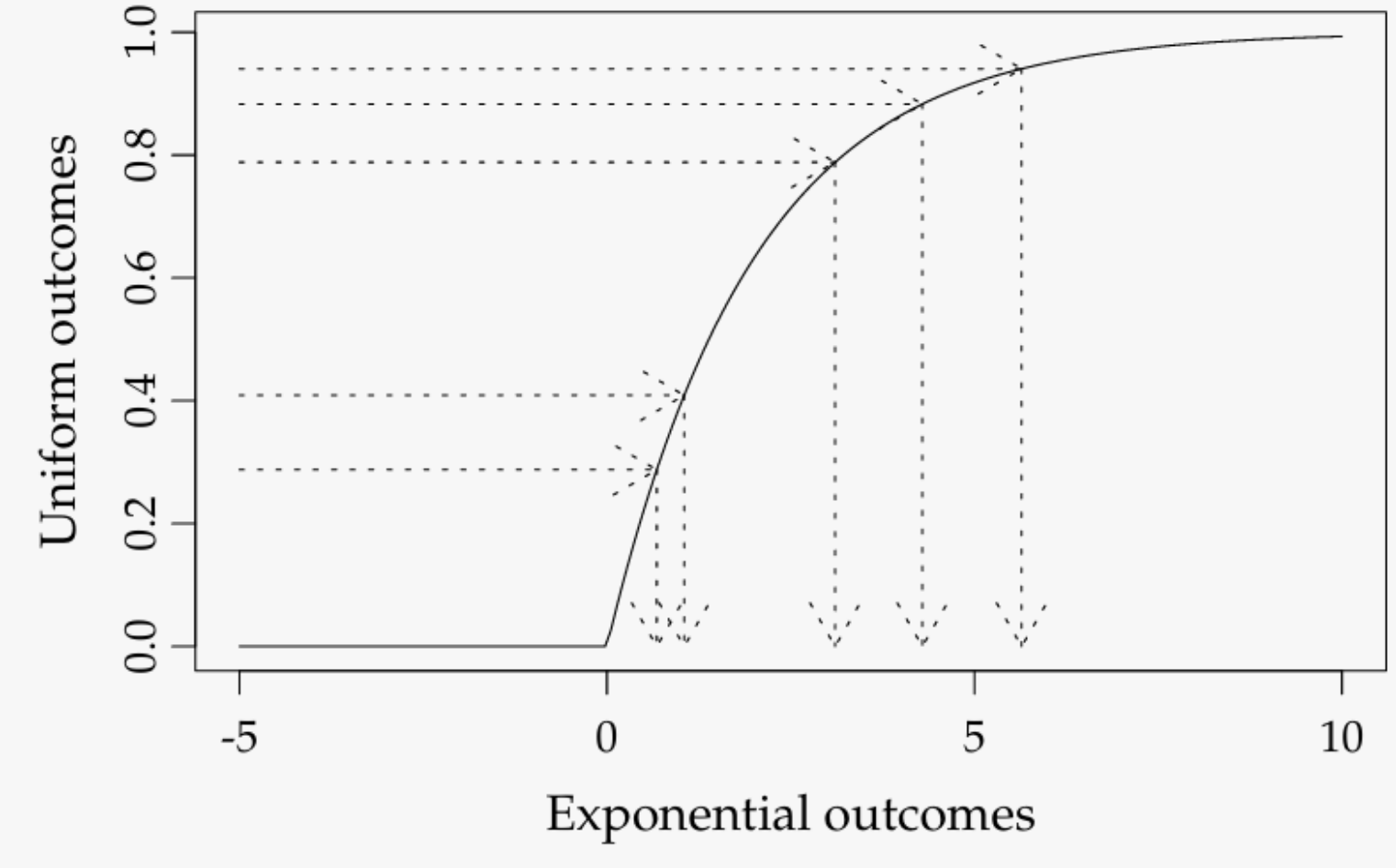
- Man kan simulere fra (næsten) alle fordelinger ved at invertere fordelingsfunktionen F (cdf), og indsætte uniformt fordelt tal:

Theorem 2.51

If $U \sim \text{Uniform}(0,1)$ and F is a distribution function for any probability distribution, then $F^{-1}(U)$ follow the distribution given by F

Eksempel: Eksponentialfordelingen med $\lambda = 0.5$

$$F(x) = \int_0^x f(t)dt = 1 - e^{-0.5x}$$



I praksis i Python

Mange fordelinger er tilgængelige via `scipy.stats` pakken og man kan simulere tilfældige tal med `.rvs()`-funktionen (står for *random variates*):

`scipy.stats.binom.rvs(...)`	Binomialfordelingen
`scipy.stats.poisson.rvs(...)`	Poissonfordelingen
`scipy.stats.hypergeom.rvs(...)`	Den hypergeometriske fordeling
`scipy.stats.norm.rvs(...)`	Normalfordelingen
`scipy.stats.lognorm.rvs(...)`	Lognormalfordelingen
`scipy.stats.expon.rvs(...)`	Eksponentialfordelingen
`scipy.stats.uniform.rvs(...)`	Den uniforme fordeling
`scipy.stats.t.rvs(...)`	t-fordelingen
`scipy.stats.chi2.rvs(...)`	χ^2-fordelingen
`scipy.stats.f.rvs(...)`	F-fordelingen
Husk også:	
`np.random.choice(...)`	en diskret fordelingen (fx terning)

Eksempel: Areal af plader

En virksomhed producerer rektangulære plader.

Længden af pladerne (i meter), X, antages at kunne beskrives ved normalfordelingen $N(2, 0.01^2)$, medens bredden af pladerne (i meter), Y, antages at kunne beskrives ved normalfordelingen $N(3, 0.02^2)$. Man kan antage, at pladernes længder og bredder er uafhængige.

Man er interesseret i arealet, A, som jo så givet ved $A = XY$.

- Hvad er middelarealet?
- Hvad er spredningen i arealet fra plade til plade?
- Hvor ofte har sådanne plader et areal, der afviger mere end 0.1 m^2 fra de angivne 6 m^2?
- Sandsynligheder for andre hændelser.
- Generelt: Hvad er fordelingen for den stokastiske variabel A?

Eksempel: Areal af plader (simulation)

Kahoot!

(x1)

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

(+ Kahoot x1)

02402 Statistik (Polyteknisk grundlag)

Fejlophobning

DTU Compute
Danmarks Tekniske Universitet
2800 Kgs. Lyngby

Fejlophobning (eng: "Error propagation")

Man ønsker at finde:

$$\sigma^2_{f(X_1,\ldots,X_n)} = \mathrm{Var}(f(X_1,\ldots,X_n))$$

Lineærkombination af uafhængige variable:

$$\sigma^2_{f(X_1,\ldots,X_n)} = \sum_{i=1}^{n} a_i^2\sigma_i^2 \quad \text{når} \quad f(X_1,\ldots,X_n) = \sum_{i=1}^{n} a_i X_i \ (\text{med uafhængighed})$$

Method 4.3 The non-linear approximative error propagation rule

If $X_1,\ldots,X_n$ are independent random variables with variances $\sigma_1^2,\ldots,\sigma_n^2$ and f is a (potentially non-linear) function of n variables, then the variance of the f-transformed variables can be approximated linearly by

$$\sigma^2_{f(X_1,\ldots,X_n)} = \sum_{i=1}^{n} \left(\frac{\partial f}{\partial x_i}\right)^2 \sigma_i^2, \tag{4-2}$$

where $\frac{\partial f}{\partial x_i}$ is the partial derivative of f with respect to the i'th variable

Eksempel: Areal af plader (fejlophobning)

I første del af forelæsningen brugte vi simulation til at undersøge variansen af arealet af plader.

Vi kan også beregne variansen af $A = XY$ med fejlophobningsloven.

Til denne beregning skal vi kende følgende størrelser:

$$\mathbf{E}[X] = 2$$
$$\mathbf{E}[Y] = 3$$
$$\mathbf{V}[X] = 0.01^2$$
$$\mathbf{V}[Y] = 0.02^2$$

Kahoot!

(x2)

Eksempel: Areal af plader (fejlophobning)

Varianserne er:

$$\sigma_X^2 = \mathbf{V}[X] = 0.01^2 \text{ og } \sigma_Y^2 = \mathbf{V}[Y] = 0.02^2.$$

Eksempel: Areal af plader (fejlophobning)

Varianserne er:

$$\sigma_X^2 = \mathbf{V}[X] = 0.01^2 \text{ og } \sigma_Y^2 = \mathbf{V}[Y] = 0.02^2.$$

Funktionen og dens partielt afledte er:

$$f(x,y) = xy, \ \frac{\partial f}{\partial x} = y, \ \frac{\partial f}{\partial y} = x.$$

Eksempel: Areal af plader (fejlophobning)

Varianserne er:

$$\sigma_X^2 = \mathbf{V}[X] = 0.01^2 \text{ og } \sigma_Y^2 = \mathbf{V}[Y] = 0.02^2.$$

Funktionen og dens partielt afledte er:

$$f(x,y) = xy, \ \frac{\partial f}{\partial x} = y, \ \frac{\partial f}{\partial y} = x.$$

Så resultatet bliver:

$$\begin{aligned} \mathbf{V}[A] = \sigma_A^2 &\approx \left(\frac{\partial f}{\partial x}\right)^2 \sigma_X^2 + \left(\frac{\partial f}{\partial y}\right)^2 \sigma_Y^2 \\ &= y^2\sigma_X^2 + x^2\sigma_Y^2 \\ &= 3.00^2 \cdot 0.01^2 + 2.00^2 \cdot 0.02^2 = 0.0025 \end{aligned}$$

og derfor:

$$\sigma_A = 0.05 \quad \text{(passer med simuleringen)}$$

Eksempel: Areal af plader (fejlophobning)

Hvad hvis vi i stedet betragter størrelsen:

$Thickness^2/Area$ (den vi kaldte "value" i simuleringen)

Eksempel: Areal af plader (fejlophobning)

Hvad hvis vi i stedet betragter størrelsen:

$Thickness^2/Area$ (den vi kaldte "value" i simuleringen)

Funktionen og dens partielt afledte er:

$$f(x,y,t) = t^2/(xy),\ \frac{\partial f}{\partial x} = -t^2/(x^2y),\ \frac{\partial f}{\partial y} = -t^2/(xy^2),\ \frac{\partial f}{\partial t} = 2t/(xy)$$

Eksempel: Areal af plader (fejlophobning)

Hvad hvis vi i stedet betragter størrelsen:

$Thickness^2/Area$ (den vi kaldte "value" i simuleringen)

Funktionen og dens partielt afledte er:

$$f(x,y,t) = t^2/(xy),\ \frac{\partial f}{\partial x} = -t^2/(x^2y),\ \frac{\partial f}{\partial y} = -t^2/(xy^2),\ \frac{\partial f}{\partial t} = 2t/(xy)$$

Så resultatet bliver:

$$\begin{aligned}
\mathbf{V}[f] = \sigma_f^2 &\approx \left(\frac{\partial f}{\partial x}\right)^2 \sigma_X^2 + \left(\frac{\partial f}{\partial y}\right)^2 \sigma_Y^2 + \left(\frac{\partial f}{\partial t}\right)^2 \sigma_T^2 \\
&= (-t^2/(x^2y))^2\sigma_X^2 + (-t^2/(xy^2))^2\sigma_Y^2 + (2t/(xy))^2\sigma_T^2 \\
&= (-0.01^2/(2^2\cdot 3))^2\sigma_X^2 + (-0.01^2/(2\cdot 3^2))^2\sigma_Y^2 + (2\cdot 0.01/(2\cdot 3))^2\sigma_T^2 \\
\sigma_f^2 &= 9.3\cdot 10^{-13} \quad \text{(passer med simuleringen)}
\end{aligned}$$

Fejlophobning – ved simulation

Method 4.4 Non-linear error propagation by simulation

Assume we have actual measurements $x_1, \ldots, x_n$ with known/assumed error variances $\sigma_1^2, \ldots, \sigma_n^2$:

1. Simulate k outcomes of all n measurements from assumed error distributions, e.g. $N(x_i, \sigma_i^2)$: $X_i^{(j)}, j = 1 \ldots, k$.

2. Calculate the standard deviation directly as the observed standard deviation of the k simulated values of f:

$$s^{\text{sim}}_{f(X_1,\ldots,X_n)} = \sqrt{\frac{1}{k-1}\sum_{j=1}^{k}(f_j - \bar{f})^2}, \tag{4-3}$$

where

$$f_j = f(X_1^{(j)}, \ldots, X_n^{(j)}). \tag{4-4}$$

Eksempel: Areal af plader (teoretisk udledning)

Faktisk kan vi i dette eksempel udlede variansen for $A = XY$ teoretisk (fra ex. 4.5):

following fundamental relationship

$$\mathrm{V}(X) = \mathrm{E}(X - \mathrm{E}(X))^2 = \mathrm{E}(X^2) - \mathrm{E}(X)^2. \tag{4-5}$$

So, one can actually deduce the variance of A theoretically, it is only necessary to know in addition that for independent random variables: $\mathrm{E}(XY) = \mathrm{E}(X)\,\mathrm{E}(Y)$ (which by the way then also tells us that $\mathrm{E}(A) = \mathrm{E}(X)\,\mathrm{E}(Y) = 6$)

$$\begin{aligned}
\mathrm{V}(XY) &= \mathrm{E}\left[(XY)^2\right] - \mathrm{E}(XY)^2 \\
&= \mathrm{E}(X^2)\,\mathrm{E}(Y^2) - \mathrm{E}(X)^2\,\mathrm{E}(Y)^2 \\
&= \left[\mathrm{V}(X) + \mathrm{E}(X)^2\right]\left[\mathrm{V}(Y) + \mathrm{E}(Y)^2\right] - \mathrm{E}(X)^2\,\mathrm{E}(Y)^2 \\
&= \mathrm{V}(X)\,\mathrm{V}(Y) + \mathrm{V}(X)\,\mathrm{E}(Y)^2 + \mathrm{V}(Y)\,\mathrm{E}(X)^2 \\
&= 0.01^2 \cdot 0.02^2 + 0.01^2 \cdot 3^2 + 0.02^2 \cdot 2^2 \\
&= 0.00000004 + 0.0009 + 0.0016 \\
&= 0.00250004.
\end{aligned}$$

Note, how the approximate error propagation rule actually corresponds to the two latter terms in the correct variance, while the first term – the product of the two variances is ignored. Fortunately, this term is the smallest of the three in this case. It does not always have to be like that. If you want to learn how to make a theoretical derivation of the density function for $A = XY$ then take a course in probability calculation.

Forskellige tilgange

Tre forskellige tilgange:

1. Simulation
2. Den approksimative *fejlophobningslov*
3. Teoretisk udledning

Forskellige tilgange

Tre forskellige tilgange:

1. Simulation
2. Den approksimative *fejlophobningslov*
3. Teoretisk udledning

Simulation har nogle vigtige fordele:

1. Nem måde at beregne andre størrelser end σ.
2. Nem måde at bruge andre fordelinger end normalfordelingen.
3. Afhænger ikke af en lineær tilnærmelse af den underliggende ikke-lineære funktion (i modsætning til fejlophobningsloven).

Eksamensopgave

Spørgsmål (20) fra Maj 2024 eksamen:

Opgave XI

I et laboratorie ønsker man at fortynde en lagerbeholdningen af opløsning A med koncentration C_A til en opløsning B med koncentration C_B. Den følgende regel anvendes:

$$C_B = \frac{C_A V_A}{V_B}$$

Her angiver V_A og V_B volumenerne af hhv. opløsning A og opløsning B. C_A, C_B, V_A og V_B er alle tilfældige variable.

Spørgsmål XI.1 (20)

Hvilken af de følgende udtryk kan bruges til at approximere standard afvigelsen af opløsning B's koncentration (σ_{C_B}) via den ikke-lineære regel for udbredning af fejl?

Eksamensopgave

$$C_B = \frac{C_A V_A}{V_B}$$

1 ☐ $\sigma_{C_B} = \sqrt{\frac{\left(\frac{\partial C_B}{\partial C_A}\sigma_{C_A}\right)^2 + \left(\frac{\partial C_B}{\partial V_A}\sigma_{V_A}\right)^2}{\left(\frac{\partial C_B}{\partial V_B}\sigma_{V_B}\right)^2}}$

2 ☐ $\sigma_{C_B} = \left(\frac{\partial C_B}{\partial C_A}\sigma_{C_A}\right)^2 + \left(\frac{\partial C_B}{\partial V_A}\sigma_{V_A}\right)^2 + \left(\frac{\partial C_B}{\partial V_B}\sigma_{V_B}\right)^2$

3 ☐ $\sigma_{C_B} = \sqrt{\left(\frac{\partial C_B}{\partial C_A}\sigma_{C_A}\right)^2 + \left(\frac{\partial C_B}{\partial V_A}\sigma_{V_A}\right)^2 + \left(\frac{\partial C_B}{\partial V_B}\sigma_{V_B}\right)^2}$

4 ☐ $\sigma_{C_B} = \frac{\left(\frac{\partial C_B}{\partial C_A}\sigma_{C_A}\right)^2 + \left(\frac{\partial C_B}{\partial V_A}\sigma_{V_A}\right)^2}{\left(\frac{\partial C_B}{\partial V_B}\sigma_{V_B}\right)^2}$

5 ☐ $\sigma_{C_B} = \left(\frac{\partial C_B}{\partial C_A}\sigma_{C_A}\right) + \left(\frac{\partial C_B}{\partial V_A}\sigma_{V_A}\right) + \left(\frac{\partial C_B}{\partial V_B}\sigma_{V_B}\right)$

Kahoot!

(x1)

Eksamensopgave

Hvad ville man faktisk sætte ind i formlen?

$$C_B = \frac{C_A V_A}{V_B}$$

1 ☐ $\sigma_{C_B} = \sqrt{\frac{\left(\frac{\partial C_B}{\partial C_A}\sigma_{C_A}\right)^2 + \left(\frac{\partial C_B}{\partial V_A}\sigma_{V_A}\right)^2}{\left(\frac{\partial C_B}{\partial V_B}\sigma_{V_B}\right)^2}}$

2 ☐ $\sigma_{C_B} = \left(\frac{\partial C_B}{\partial C_A}\sigma_{C_A}\right)^2 + \left(\frac{\partial C_B}{\partial V_A}\sigma_{V_A}\right)^2 + \left(\frac{\partial C_B}{\partial V_B}\sigma_{V_B}\right)^2$

3 ☐ $\sigma_{C_B} = \sqrt{\left(\frac{\partial C_B}{\partial C_A}\sigma_{C_A}\right)^2 + \left(\frac{\partial C_B}{\partial V_A}\sigma_{V_A}\right)^2 + \left(\frac{\partial C_B}{\partial V_B}\sigma_{V_B}\right)^2}$

4 ☐ $\sigma_{C_B} = \frac{\left(\frac{\partial C_B}{\partial C_A}\sigma_{C_A}\right)^2 + \left(\frac{\partial C_B}{\partial V_A}\sigma_{V_A}\right)^2}{\left(\frac{\partial C_B}{\partial V_B}\sigma_{V_B}\right)^2}$

5 ☐ $\sigma_{C_B} = \left(\frac{\partial C_B}{\partial C_A}\sigma_{C_A}\right) + \left(\frac{\partial C_B}{\partial V_A}\sigma_{V_A}\right) + \left(\frac{\partial C_B}{\partial V_B}\sigma_{V_B}\right)$

Eksamensopgave

Hvordan kunne man løse samme opgave med simulering?

Opgave XI

I et laboratorie ønsker man at fortynde en lagerbeholdningen af opløsning A med koncentration C_A til en opløsning B med koncentration C_B. Den følgende regel anvendes:

$$C_B = \frac{C_A V_A}{V_B}$$

Her angiver V_A og V_B volumenerne af hhv. opløsning A og opløsning B. C_A, C_B, V_A og V_B er alle tilfældige variable.

Spørgsmål XI.1 (20)

Hvilken af de følgende udtryk kan bruges til at approximere standard afvigelsen af opløsning B's koncentration (σ_{C_B}) via den ikke-lineære regel for udbredning af fejl?

02402 Statistik (Polyteknisk grundlag)

Introduktion til den generelle lineære model

DTU Compute
Danmarks Tekniske Universitet
2800 Kgs. Lyngby

En statistisk model

Remark 3.2 How to write a statistical model

In all statistical analysis there must be an assumption of a *model*, which should be stated clearly in the presentation of the analysis. The model expressing that the sample was taken randomly from the population, which is normal distributed, can be written by

$$X_i \sim N(\mu, \sigma^2) \text{ and i.i.d., where } i = 1, \ldots, n. \tag{3-1}$$

Hence we n random variables representing the sample and they are *independent and identically distributed* (i.i.d).

En statistisk model for normalfordelt data

Vi opskriver en model for data der følger en Normalfordeling.
Dvs vores (ret simple) **statistiske model** er:

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

En anden måde at skrive modellen er:

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

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

Den generelle lineære model

I de kommende uger vil vi lave mere komplicerede modeller - med det til fælles at den stokastiske del beskrives med en normalfordeling:

$$Y_i = \ldots + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2)$$

Her er "..." en **lineær** funktion af modellens parametre

For én stikprøve er "..." kun gennemsnittet μ (μ er en parameter)

I de kommende uger skal vi lave mere avancerede modeller

Motivation for mere avancerede modeller

I virkelige undersøgelser/eksperimenter vil man ofte gerne finde mere *avancerede sammenhænge i data*.

Eksempel med højder på studerende:
Lad os sige vi har indsamlet data på en masse studerende.
Vi er interesserede i hvad der afgør personers højde.
Vi har derfor målt højden på hver studerende, men vi har også indsamlet data på en række størrelser vi *tror kan være relevante* til højde.
Fx har vi noteret **køn** og **vægt** for hver studerende.

Hvordan laver vi en model for hvordan højde afhænger af både køn og vægt (og måske endnu flere faktorer)?

Før vi kan svare på dette vil vi opstille et mere generelt setup (der lige kræver lidt notation).

Den generelle lineære model for to stikprøver

Faktisk har vi allerede set på en lidt mere avanceret model:
Data kan komme fra to forskellige grupper (med forskellige gennemsnit).

Vi kan opskrive en model for tilfældet med to stikprøver:

$$Y_{ij} = \mu_i + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim N(0, \sigma^2)$$

Her er $i = 1$ (for gruppe 1) eller $i = 2$ (for gruppe 2).
Dvs det ovenstående model er egentlig to udtryk (én model for *hver* gruppe):

$$Y_{1j} = \mu_1 + \varepsilon_{1j}, \quad \varepsilon_{1j} \sim N(0, \sigma^2), \quad j = 1, 2, ..., n_1$$

$$Y_{2j} = \mu_2 + \varepsilon_{2j}, \quad \varepsilon_{2j} \sim N(0, \sigma^2), \quad j = 1, 2, ..., n_2$$

Den generelle lineære model for to stikprøver

For situationen med to grupper kan vi opskrive én *samlet* model er ved at introducere en *forklarende variabel* x_i:

- x_i er enten 0 eller 1
- x_i angiver om enhed i tilhører gruppe 1 $(x_i = 0)$ eller gruppe 2 $(x_i = 1)$

Vi kan da opskrive en model for tilfældet med to stikprøver:

$$Y_i = \mu_1 + \delta \cdot x_i + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2)$$

$\delta = \mu_2 - \mu_1$ er **differensen** på de to populationers middelværdier.

For alle enheder i gruppe 1 $(x_i = 0)$ har vi: $Y_i = \mu_1 + \varepsilon_i$

For alle enheder i gruppe 2 $(x_i = 1)$ har vi: $Y_i = \mu_2 + \varepsilon_i$

Den generelle lineære model for to stikprøver

Modellen for tilfældet med to stikprøver:

$$Y_i = \mu_1 + \delta \cdot x_i + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2)$$

er en **lineær** funktion af modellens parametre: μ_1 og δ

Bemærk at modellen her antager ens varians (samme σ) i de to grupper!

Matrice notation for to stikprøver

Modellen for tilfældet med to stikprøver:

$$Y_i = \mu_1 + \delta \cdot x_i + \varepsilon_i$$

kan også skrives på matriceform:

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon$$

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} \qquad \mathbf{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \qquad \beta = \begin{bmatrix} \mu_1 \\ \delta \end{bmatrix} \qquad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Her er $\{x_1, x_2, ..., x_n\}$ en binær forklarende variabel $(\{0, 0, ..., 1, 1, ...\})$

Matrice notation for den generelle lineære model

Hvorfor alt dette? $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} \qquad \mathbf{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \qquad \beta = \begin{bmatrix} \mu_1 \\ \delta \end{bmatrix} \qquad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Med denne notation bliver det nemt at udvide modellen til at indeholde *flere forklarende variable*:

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} \qquad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{21} & \dots \\ 1 & x_{12} & x_{22} & \dots \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1n} & x_{2n} & \dots \end{bmatrix} \qquad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \end{bmatrix} \qquad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Design matricen

Matricen $\mathbf{X}$ kaldes også for "design matricen".

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{21} & \dots \\ 1 & x_{12} & x_{22} & \dots \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1n} & x_{2n} & \dots \end{bmatrix}$$

Matricen $\mathbf{X}$ indeholder information fra alle de *forklarende variable* der indgår i modellen.

Eksempel med højder på studerende:
For eksemplet med højder kunne de forklarende variable være køn (0/1) samt vægt (målt i kg)

$$\mathbf{X} = \begin{bmatrix} 1 & 0 & 83.2 \\ 1 & 1 & 67.5 \\ \vdots & \vdots & \vdots \end{bmatrix}$$

Kahoot!
(x3)

De kommende uger

I eksemplet med højder på studerende har vi to forklarende variabler af *forskellig type*: Den ene forklarende variabel er kvalitativ (gruppering, oversat til "0/1") og den anden forklarende variabel er kvantitativ (målt i kg).

I de kommende uger skal vi se på modeller med én eller flere kvantitative forklarende variabler (lineær regression).

Herefter skal vi se på modeller kun med én eller flere kvalitattive forklarende variabler (dvs grupperinger) (ANOVA).

Endelig skal vi se på simple tilfælde hvor *interessevariablen* Y_i er kvalitativ (antal og andele).

Typen af model hænger sammen med typen af data.

Dagsorden

1. Til formelsamlingen/appendix
2. Introduktion til simulation
3. Fejlophobning
4. Introduktion til den generelle lineære model

Tjekliste

Efter i dag skal du kunne:

- Simulere tilfældige tal fra diverse sandsynlighedfordelinger
- Opsætte en simulering der kan bruges til at undersøge en kontekst med tilfældige udfald
- Simulere afledte størrelser fra stokastiske variable og udføre diverse beregner på disse afledte størrelser
- Benytte fejlophobningsloven til at estimere usikkerheder på afledte størrelser af mere simple stokastiske variable
- Udpege interessevariabel og forklarende variable ud fra en kontekst
- Opskrive en simpel statistisk model i form af *den generelle lineære model*
- Beskrive hvad en design matrice er
- Opskrive en design matrice til en simpel lineær model (givet et datasæt med relevante oplysninger).

Øvelser Uge 7

4.1 Brug Python - øvelse i at lave egen simulering.

5.5 a-c uden Python - Brug Python til at simulere til sidst (d).

5.6 Forsøg at løse uden Python (men brug evt Python til at teste at du har forstået hvad der foregår).

Project Sidste chance til at få hjælp. Project 1 afleveres i dag (kl 23.59). Du aflverer på Learn.