Kursus 02323: Introducerende Statistik

Forelæsning 12: Forsøgsplanlægning

Peder Bacher

DTU Compute, Dynamiske Systemer
Bygning 303B, Rum 010
Danmarks Tekniske Universitet
2800 Lyngby – Danmark
e-mail: pbac@dtu.dk

Forår 2021

Afsnit 3.3 og 7.2.2: Forsøgsplanlægning

Grundlæggende koncepter for forsøgsplanlægning:

- Testens styrke er $1-\beta$ *(hvor β er sandsynligheden for at begå Type II fejl)*

Specifikke metoder, forsøgsplanlægning
(middelværdi, både one og two sample setup):

- Stikprøvestørrelse n for ønsket præcision af konfidensintervaller
- Stikprøvestørrelse n for ønsket styrke af tests

Specifikke metoder, forsøgsplanlægning
(andel, one sample setup):

- Stikprøvestørrelse n for ønsket præcision af konfidensintervaller

Section 3.3 and 7.2.2: Design of experiments

General concepts for design of experiments:

- Power of a test is $1-\beta$ *(where β is the probability of making a Type II error)*

Specific methods, design of experiments (mean, both one and two sample setup):

- Sample size n for wanted precision of confidence intervals
- Sample size n for wanted power of tests

Specific methods, design of experiments (proportion, one-sample setup):

- Sample size n for wanted precision of confidence intervals

Oversigt

1. Planlægning af studie med krav til præcision
2. Planlægning: Power og sample size
3. Andele: Bestemmelse af stikprøvestørrelse

Planlægning af studie med krav til præcision

Man vil gerne tænke sig om inden eksperimentet udføres:

- Brug for at vide hvor præcise resultater (f.eks. konfidensinterval) forventes at blive med et fremtidigt eksperiment
- Hvor stor en effekt forventes at kunne opdages (e.g. hvis sovemedicinen virker 2 timer bedre, hvad er sandsynligheden for at det opdages?)
- Spørgsmål om at optimere økonomiske ressourcer og etik!

Method 3.63: The one-sample CI sample size formula

Antal observationer og den forventede (halve) bredde af konfidensintervallet

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

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

- Margin of error ME er den *halve bredde* af konfidensintervallets forventede bredde

Eksempel på to mulige fejl ved hypotesetest – sovemedicin

To mulige sandheder vs. to mulige konklusioner:

Test udfald / *Virkelighed*	**Afvis** H_0	**Afvis ikke** H_0
Sand H_0: Medicinen virker ikke	Type I fejl (α)	Korrekt: Ingen effekt
Falsk H_0**:** Medicinen virker	Korrekt: Påvist effekt	Type II fejl (β)

Eksempel på to mulige fejl ved hypotesetest (repetition)

To mulige sandheder vs. to mulige konklusioner:

Test udfald / *Virkelighed*	**Afvis** H_0	**Afvis ikke** H_0
Sand H_0	Type I fejl (α)	Korrekt accept af H_0
Falsk H_0	Korrekt afvisning af H_0	Type II fejl (β)

Mulige fejl ved hypotesetests (repetition)

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

Type I: Rejection of H_0 when H_0 is true
(Medicinen virker ikke, men vi kommer til at tro den virker)

Type II: Non-rejection of H_0 when H_1 is true
(Medicinen virker, men vi opdager det ikke)

Risikoen for de to typer fejl kaldes sædvanligvis:

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

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

Testens styrke (power)

Hvad er styrken $1-\beta$ for et kommende studie/eksperiment:

- Sandsynligheden for at opdage en effekt (af en vis størrelse $|\mu_0 - \mu_1|$)
- Probability of correct rejection of H_0
- $P($"Accept af H_0"$)$ når H_1 er sand

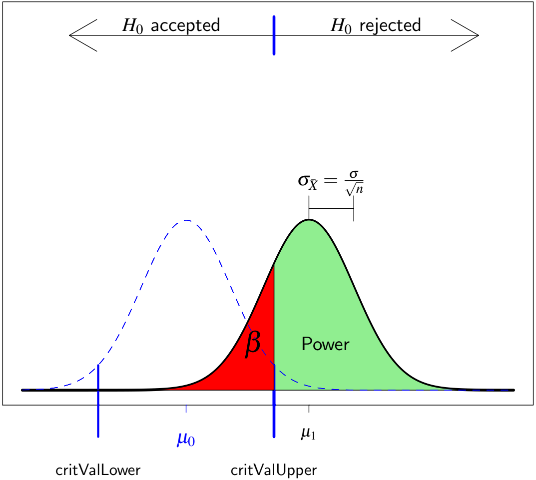
Udfordring: Nulhypotesen kan være forkert på mange måder!

I praksis, scenarie-baseret approach:

- F.eks. "Hvis nu sovemedicinen *rent faktisk virker 2 timer bedre*, hvor godt vil mit studie være til at opdage dette? "
- Eller, jeg vil gerne *hvis min sovemedicin virker 3 timer bedre*, opdage dette med en bestemt sandsynlighed (power)

$\bar{X} \sim N(\mu_1, \frac{\sigma^2}{n})$ er den antagede fordeling (dvs. om μ_1 og $ME = |\mu_1 - \mu_0|$)
$H_0 : \mu = \mu_0$ er nulhypotesen

Vi kan se hvad β er: $P(H_0$ accepteres forkert$) = P($Type II$) = \beta$

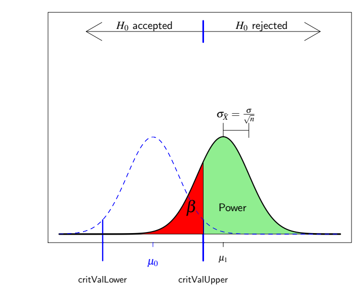


Spørgsmål om power (socrative.com - ROOM:PBAC)

Vi antager at $\bar{X} \sim N(\mu_1, \frac{\sigma^2}{n})$ (dvs. fordelt om μ_1, som er effekten vi vil kunne påvise), $H_0 : \mu = \mu_0$ er nulhypotesen

Hvordan kan vi opnå en større power *uden at kompromitere noget ved testen (dvs. ikke ændre på hypotesen eller risikoen for type I fejl)*?

- A: Mindske μ_0 så $ME = |\mu_0 - \mu_1|$ øges
- B: Øge α (på figur vil 'critvalUpper' mindskes)
- C: Øge n antallet af observationer
- D: Desværre kan det ikke lade sig gøre
- E: Ved ikke



Planlægning, find sample size n

Det store spørgsmål i praksis: HVAD skal n være?

Forsøget skal være stort nok til at kunne opdage en relevant effekt med stor power (som regel mindst 80%):

Metode 3.65: Tilnærmet svar for en one-sample t-test:

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.

Planlægning, sæt 4 prm. og beregn den sidste

Når man har fastlagt hvilket test, der skal bruges:

Kender man (eller fastlægger/gætter på) fire ud af følgende fem oplysninger, kan man regne sig frem til den femte:

- n Sample size
- α Significance level of the test
- δ A difference in mean that you would want to detect (effect size) (*dvs. μ_2 er her den middelværdi med afstand til μ_1 som vi "mindst" vil kunne påvise)*
- σ The population standard deviation
- $1-\beta$ The power

Eksempel - The sample size for power$= 0.80$

```r
## Stikprøvestørrelse for t-test
power.t.test(power = .80, delta = 4, sd = 12.21, sig.level=0.05,
     type = "one.sample")

##
##      One-sample t test power calculation
##
##               n = 75.08
##           delta = 4
##              sd = 12.21
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
```

Svar: $n = 76$ *(husk at runde op)*

Metode 3.65: *Tilnærmet* svar for en one-sample t-test:

Formlen giver lidt lavere resultat, fordi normalfordelingen bruges i stedet for t-fordelingen:

$$
\begin{aligned}
n &= \left(\sigma \frac{z_{1-\beta} + z_{1-\alpha/2}}{(\mu_0 - \mu_1)}\right)^2 \\
&= \left(12.21 \frac{0.84 + 1.96}{4}\right)^2 \\
&= 73.05
\end{aligned}
$$

I opgaver bruges resultatet fra `power.t.test()`, *hvis der ikke henvises konkret til formlen fra bogen.*

Eksempel - The power for $n = 40$

```
## Beregn power for t-test
power.t.test(n = 40, delta = 4, sd = 12.21, sig.level=0.05,
      type = "one.sample")

##
##      One-sample t test power calculation
##
##               n = 40
##           delta = 4
##              sd = 12.21
##       sig.level = 0.05
##           power = 0.5242
##     alternative = two.sided
```

Styrke og stikprøvestørrelse - two-sample

Finding the sample size for detecting a group difference of 2 with $\sigma = 1$ and power$= 0.9$:

```
## Beregn stikprøvestørrelsen
power.t.test(power = 0.90, delta = 2, sd = 1, sig.level = 0.05)

##
##      Two-sample t test power calculation
##
##               n = 6.387
##           delta = 2
##              sd = 1
##       sig.level = 0.05
##           power = 0.9
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

Svar: $n = 7$ *(husk at runde op)*

Styrke og stikprøvestørrelse - two-sample

Finding the power of detecting a group difference of 2 with $\sigma = 1$ for $n = 10$:

```
## Power beregning
power.t.test(n = 10, delta = 2, sd = 1, sig.level = 0.05)

##
##      Two-sample t test power calculation
##
##               n = 10
##           delta = 2
##              sd = 1
##       sig.level = 0.05
##           power = 0.9882
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

Styrke og stikprøvestørrelse - two-sample

Finding the detectable effect size (delta) with $\sigma = 1$, $n = 10$ and power$= 0.9$:

```
## Beregn margin of error
power.t.test(power = 0.90, n = 10, sd = 1, sig.level = 0.05)

##
##      Two-sample t test power calculation
##
##               n = 10
##           delta = 1.534
##              sd = 1
##       sig.level = 0.05
##           power = 0.9
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

Andele: Stikprøvestørrelse: "Margin of Error" (ME):

Margin of Error på estimat kan siges at være:

- Forventningsværdi af *"halvdelen af konfidensintervallets bredde"*
- *"Den forskel i middelværdi"* man gerne vil være i stand til at påvise
- Under H_0: Forventningsværdi af afstanden mellem middelværdien og det kritiske niveau

Margin of Error

med $(1-\alpha)\%$ konfidens bliver

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

hvor et estimat af p fås ved $p = \frac{x}{n}$

Spørgsmål om Margin of Error (socrative.com, ROOM: pbac)

Hvad er "Margin of Error"(ME) hvis man vil have et konfidensinterval med bredde på 0.2?

- A: 0.1
- B: -0.15
- C: 0.2
- D: 0.4
- E: Ved ikke

Spørgsmål om forskel i middel (socrative.com, ROOM: pbac)

Hvad er "Margin of Error"(ME) hvis man vil være i stand til at påvise forskel i middelværdi på 0.2?

- A: 0.1
- B: -0.15
- C: 0.2
- D: 0.4
- E: Ved ikke

Bestemmelse af stikprøvestørrelse

Method 7.13

Såfremt man højst vil tillade en Margin of Error (ME) med $(1-\alpha)\%$ konfidens, bestemmes den nødvendige stikprøvestørrelse ved

$$n = p(1-p)\left[\frac{z_{1-\alpha/2}}{ME}\right]^2$$

Method 7.13

Såfremt man højst vil tillade en Margin of Error ME med $(1-\alpha)\%$ konfidens, og p ikke kendes, bestemmes den nødvendige stikprøvestørrelse ved

$$n = \frac{1}{4}\left[\frac{z_{1-\alpha/2}}{ME}\right]^2$$

idet man får den mest konservative stikprøvestørrelse ved at vælge $p = \frac{1}{2}$

Eksempel 1 - fortsat

Venstrehåndede:

Antag vi ønsker $ME = 0.01$ (med $\alpha = 0.05$) - hvad skal n være?

Antag $p \approx 0.10$:

$$n = 0.1 \cdot 0.9 \left(\frac{1.96}{0.01}\right)^2 = 3467.4 \approx 3468$$

UDEN antagelse om størrelsen af p:

$$n = \frac{1}{4}\left(\frac{1.96}{0.01}\right)^2 = 9604$$

Spørgsmål om stikprøvestørrelse (`socrative.com`, ROOM: pbac)

Ved test af hvilken af følgende nulhypoteser skal bruges den største stikprøvestørrelse (n) ved samme α signifikansniveau?

- A: $H_0 : p = 0.2$
- B: $H_0 : p = 0.1$
- C: $H_0 : p = 0.4$
- D: $H_0 : p = 0.95$
- E: Ved ikke

Spørgsmål om stikprøvestørrelse (socrative.com, ROOM: pbac)

Kan I nu beregne hvor mange gange man skal slå med en terning for at teste om den har sandsynlighed $1/6$ indenfor 0.01 for at slå en sekser?

- A: Ja
- B: Nej
- C: Ved ikke

```
## Andel (sandsynlighed) vi vil teste for
p <- 1/6
## Signifikansniveau
## (hvor ofte vil vi lave denne fejl: Terningen er fair, men
##  vi konkluderer den ikke er fair)
alpha <- 0.05
## Fejlmargen vi vil tillade
ME <- 0.01
## Beregn antal gange vi skal slå med terningen
p * (1-p) * (qnorm(1-alpha/2)/ME)^2

## [1] 5335
```

Husk også at sige at Exercise 3.10 spørgsmål c) er ret abstrakt og man kan springe den over.