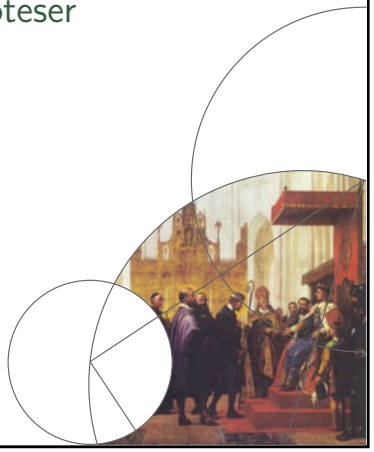
F-test og mere om test af hypoteser

Anders Tolver
Institut for Matematiske Fag



I dag

Vi gennemgår først slides 1-24 om

- Kap. 6.3: F-test for sammenligning af tre eller flere grupper
- Kap. 6.4: F-test mere generelt
- Tankegang i hypotesetest
- Kap. 6.5: Fejl af type I og type II

I det omfang tiden tillader snakker vi også om

- Bonferronikorrektion (slides 25-30)
- Kap. 5: (slides 31-41)
  Statistiske modeller med samme struktur, repetition af t-test

Overblik

Vi skal have "'udfyldt"' følgende skema over modeller (rækker) og statistiske begreber (søjler):

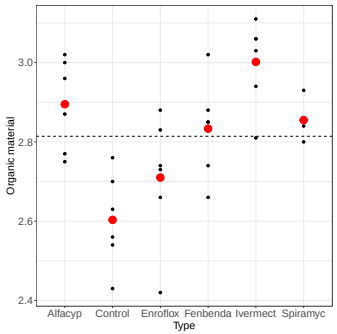
	Intro	Model	Est.+SE	KI	Test	Kontrol	Præd.
En stikprøve	✓	✓	✓	✓	✓	✓	
Ensidet ANOVA	✓	✓	✓	✓	nu		
Lineær regr.	✓	✓	✓	✓	(✓)		
To stikprøver	✓	✓	✓	✓	✓		
Multipel regr.							
Tosidet ANOVA							

F-testet for sammenligning af tre eller flere grupper

Ensidet ANOVA: antibio-datasættet

```
library(isdals)
data(antibio)
head(antibio, n = 7)

##        type  org
## 1 Ivermect 3.03
## 2 Ivermect 2.81
## 3 Ivermect 3.06
## 4 Ivermect 3.11
## 5 Ivermect 2.94
## 6 Ivermect 3.06
## 7  Alfacyp 3.00
```



- Respons: Mængden af organisk materiale efter otte uger
- En kategorisk forklarende var.: fodergruppe
- H_0 : samme mængde organisk stof i alle fodergrupper?

Statistisk model og hypoteser

Data: $y_1, \ldots, y_n$: mængde af organisk materiale fra k grupper med n_j observationer i gruppe j.

Statistisk model:

- $y_1, \ldots, y_n$ uafhængige og normalfordelte $\sim N(\mu_i, \sigma^2)$
- middelværdi $\rightarrow \mu_i = \alpha_{g(i)}$ afhænger af gruppe
- spredning $\rightarrow \sigma$ ens i alle grupper

Hypoteser:

- $\alpha_1 = 5$ (middelværdien har en bestemt værdi i en given gruppe)
- $\alpha_1 - \alpha_2 = 0$ (to grupper har samme middelværdi)
- $\alpha_1 = \alpha_2 = \ldots = \alpha_k$ (alle grupper har samme middelværdi)

De to første kan klares med t-test.

Til sidste hypotese har vi brug for en ny teststørrelse, som vi kalder ***F*-teststørrelsen**.

R: sammenligning af flere grupper

```
model1 <- lm(org ~ type, data = antibio)
summary(model1)

##
## Call:
## lm(formula = org ~ type, data = antibio)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.29000 -0.06000  0.01833  0.07250  0.18667
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.89500    0.04970  58.248  < 2e-16 ***
## typeControl  -0.29167    0.07029  -4.150 0.000281 ***
## typeEnroflox -0.18500    0.07029  -2.632 0.013653 *
## typeFenbenda -0.06167    0.07029  -0.877 0.387770
## typeIvermect  0.10667    0.07029   1.518 0.140338
## typeSpiramyc -0.04000    0.07858  -0.509 0.614738
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1217 on 28 degrees of freedom
## Multiple R-squared:  0.5874,Adjusted R-squared:  0.5137
## F-statistic: 7.973 on 5 and 28 DF,  p-value: 8.953e-05
```


R: sammenligning af flere grupper

```
drop1(model1, test = "F")

## Single term deletions
##
## Model:
## org ~ type
##        Df Sum of Sq    RSS    AIC F value    Pr(>F)
## <none>              0.4150 -137.8
## type    5   0.59082 1.0058 -117.7  7.9726 8.953e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

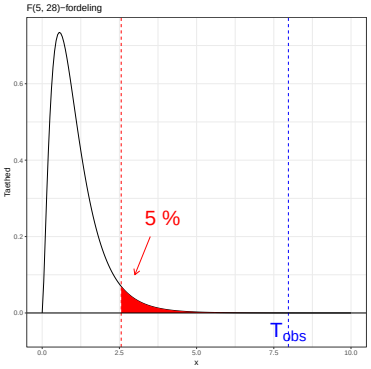
Bemærk:

- Vi kan **ikke** bruge modellen `lm(tid ~ type - 1, data=sudokoData)` til dette test!
- Bogen bruger `anova` snarere end `drop1`. Ikke forkert, men jeg foretrækker `drop1`.

F-fordelingen (opkaldt efter R.A. Fisher)

Hvis nul hypotesen er sand, så er F-teststørrelsen **F-fordelt med $(k-1, n-k)$ frihedsgrader.**

$$p\text{-værdi} = P(F \geq F_{\text{obs}})$$



Antibio: $n = 34$, $k = 6, F_{\text{obs}} = 7.97$

Beregning af p-værdi:
Sandsynlighed til højre for 7.97

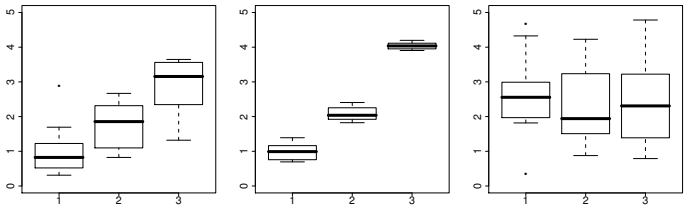
```
1-pf(7.97, df1 = 6-1, df2 = 34-6)

## [1] 8.975507e-05
```

Konklusion: Hypotesen om 6 ens middelværdier bliver klart afvist!

Fortolkninger af F-teststørrelsen

Opgave 3.1: Between-group, within-group variation



F-teststørrelsen forsøger at fange følgende:

- Variation mellem grupper **stor** ift. variation indenfor grupper: Tegn på forskel mellem grupperne
- Variation mellem grupper **lille** ift. variation indenfor grupper: Tegn på at der ikke er forskel mellem grupperne

Variation indenfor og mellem grupper

Variation mellem grupper: Gruppegennemsnit vs totalgennemsnit,

$$\text{SS}_{\text{between}} = \sum_{j=1}^{k} n_j(\bar{y}_j - \bar{y})^2$$

Variation indenfor grupper: Observationer vs gruppegennemsnit,

$$\text{SS}_{\text{within}} = \sum_{i=1}^{n} \left(y_i - \bar{y}_{g(i)}\right)^2$$

F-teststørrelse ser på forholdet mellem dem, passende normeret:

$$F = \frac{\text{MS}_{\text{between}}}{\text{MS}_{\text{within}}} = \frac{\text{SS}_{\text{between}}/(k-1)}{\text{SS}_{\text{within}}/(n-k)}$$

antibio-data: $F_{\text{obs}} = 7.97$.

Kan F bruges som teststørrelse?

De tre kriterier:

- Det er en **talværdi**, som kan beregnes udfra data ✓
- Den skal være et godt mål for hvor godt data stemmer med hypotesen ✓

  F er altid positiv. Små værdier passer godt med hypotesen, store værdier passer skidt. Siger at store værdier er kritiske.
- Under forudsætning af at hypotesen er sand, skal teststørrelsens sandsynlighedsfordeling kunne beregnes ✓

  Hvis H_0 er sand *kan* vi faktisk sige hvordan F vil opføre sig...

Tilsammen: Vi kan nu beregne **ssh. for at få en F-værdi der passer dårligere med hypotesen end den vi fik fra vores data.**

F-test mere generelt: fuld model og nulmodel

Husk at hypotesen beskriver en restriktion på modellen. Modellen hvor hypotesen er opfyldt, kaldes nulmodellen (null model).

Har altså to modeller hvor den ene er en delmodel af den anden:

- **Fuld model:** $y_i = \alpha_{\mathbf{g(i)}} + e_i$, e_i'erne uafhængige $N(0, \sigma^2)$.
- **Nulmodel:** $y_i = \alpha + e_i$, e_i'erne uafhængige $N(0, \sigma_0^2)$.

Residualkvadratsum — kan beregnes i hver af de to modeller:

$$\mathrm{SS} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

hvor $\hat{y}_i$ er de estimerede middelværdier (=fittede værdier).

- Mål for hvor godt modellerne passer til data.
- $\mathrm{SS}_0 > \mathrm{SS}_{\mathrm{full}}$: Vi kan tilpasse $\hat{y}_i$ bedre til data i fuld model
- $\mathrm{df}_0 > \mathrm{df}_{\mathrm{full}}$: Der er flere parametre i fuld model

En alternativt fortokning af F-teststørrelsen ...

Nyt forslag til teststørrelse:

$$F = \frac{(\mathrm{SS}_0 - \mathrm{SS}_{\mathrm{full}})/(\mathrm{df}_0 - \mathrm{df}_{\mathrm{full}})}{\mathrm{SS}_{\mathrm{full}}/\mathrm{df}_{\mathrm{full}}}$$

- Måler hvor meget større SS er under hypotesen ift. i den fulde model. Passende normeret med frihedsgrader.
- Store værdier er kritiske, dvs. passer dårligt med hypotesen

Kan vise at det er **præcis den samme F-teststørrelse som før!**

- Før: Variation mellem grupper ift. variation indenfor grupper
- Nu: Sammenligning af modellen med og uden restriktionen givet ved hypotesen

Den alternative fortolkning af F-teststørrelsen kan bruges til at teste andre typer af hypoteser.

R: sammenligning af flere grupper

```
fullModel <- lm(org ~ type, data = antibio)
nullModel <- lm(org ~ 1, data = antibio)
anova(nullModel, fullModel)

## Analysis of Variance Table
##
## Model 1: org ~ 1
## Model 2: org ~ type
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     33 1.0058
## 2     28 0.4150  5   0.59082 7.9726 8.953e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Bemærk: Jeg bruger `anova` når jeg **sammenligner to modelfit.**

Eksempel: ny brug af F-teststørrelsen!

Man kunne være interesset i at undersøge, om kontrolgruppen ('Control') er den eneste gruppe, hvor indholdet af organgisk stof afviger fra de øvrige.

Hypotesen er derfor,

$$H_0 : \alpha_{\text{Alfacyp}} = \alpha_{\text{Enroflox}} = \alpha_{\text{Fenbenda}} = \alpha_{\text{Ivermectin}} = \alpha_{\text{Spiramyc}}$$

Under hypotesen (hvis hypotesen er sand) er der kun to grupper. Vi kan udføre testet på følgende måde:

- Fit den fulde model, dvs. modellen med skes grupper
- Lav variabel med to grupper, fit nulmodellen med denne var.
- Sammenlign de to modeller med `anova`

Konklusion?

R

```
antibio$typeControl <- (antibio$type == "Control")
fullModel <- lm(org ~ type, data = antibio)
nullModel2 <- lm(org ~ typeControl, data = antibio)
anova(nullModel2, fullModel)

## Analysis of Variance Table
##
## Model 1: org ~ typeControl
## Model 2: org ~ type
##   Res.Df     RSS Df Sum of Sq      F   Pr(>F)
## 1     32 0.68212
## 2     28 0.41500  4   0.26712 4.5056 0.006171 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Hypotesetest, Fejl af type I og II

Tankegang i hypotesetest

Hypotestest er baseret på teststørrelser som opfører sig forskelligt alt efter om hypotesen er sand eller falsk.

- Hvis stat. model er OK og hypotesen er sand, så ved vi hvilke værdier af teststørrelsen der er sandsynlige/usandsynlige.
- Hvis den værdi vi får for data er (meget) usandsynlig, er det evidens for at hypotesen er falsk.

Falsificering

Vi kan **falsificere hypoteser:** Hvis hypotesen er sand, er det nogle meget mærkelige data vi har fået → hypotesen afvises.

Men vi **kan ikke påvise at hypoteser er sande.**

Der kan være flere grunde til at en hypotese ikke kan afvises:

- Hypotesen er sand
- Hypotesen er falsk, men datagrundlaget er for småt
- Hypotesen er falsk, men afvigelsen fra hypotesen er for lille
- Hypotesen er falsk, men der er for stor biologisk variation

Hypoteser

Statistiske hypoteser formuleres typisk "omvendt" af videnskabelige hypoteser.

Eksempel:

- Videnskabelig hypotese: Behandlingen **har** en effekt
- Statistisk hypotese: Behandlingen **har ikke** en effekt

Når vi afviser den statistiske hypotese, er det altså (typisk) evidens for den videnskabelige hypotese.

Forresten: **En forskel kan sagtens være statistisk signifikant uden at være relevant eller interessant!** Angiv estimater og KI.

Konklusioner i hypotesetest

Hvis vi bruger signifikansniveau 5%

- **afviser** vi hypotesen, hvis $p < 0.05$
- **accepterer** vi hypotesen, hvis $p > 0.05$

Accept/afvisning af en hypotese betyder desværre ikke nødvendigvis at hypotesen er sand/falsk.

Fejl af Type I og Type II

Fire muligheder:

	Hypotesen accepteres	Hypotesen afvises
H_0 sand	OK	fejl af **type I**
H_0 falsk	fejl af **type II**	OK

- Fejl af type I: **Falsk positiv.** Konkluderer at der er effekt/sammenhæng selvom der ikke er.
- Fejl af type II: **Falsk negativ.** Konkluderer at der ikke er effekt/sammenhæng selvom der faktisk er.

Hvis signifikansgrænsen på 5% bruges, så er **ssh. for en fejl af type I netop 5%.** Fejl at type II har vi ikke kontrol over.

Multiple testning og Bonferronikorrektion

Fejl af Type I og Type II

Vi minder om de fire muligheder:

	Hypotesen accepteres	Hypotesen afvises
H_0 sand	OK	fejl af **type I**
H_0 falsk	fejl af **type II**	OK

- Fejl af type I: **Falsk positiv.** Konkluderer at der er effekt/sammenhæng selvom der ikke er.
- Fejl af type II: **Falsk negativ.** Konkluderer at der ikke er effekt/sammenhæng selvom der faktisk er.

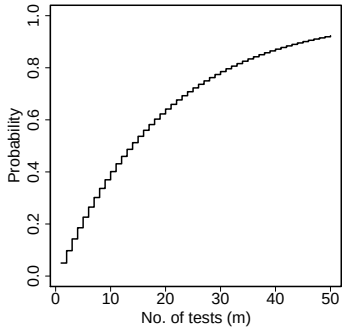
Signifikansgrænse på 5% $\rightarrow$ **ssh. for en fejl af type I netop 5%.**

Men hvad hvis vi laver **flere test**?

Multiple tests

Hvis vi bruger 5% som signifikansniveau:

- Ved et test: Risiko for fejl (falsk positiv) = 5%
- Ved m tests er risikoen for **mindst en type I fejl** $1 - 0.95^m$
- **Vokser hurtigt** når vi laver mange tests!



Bonferroni korrektion

Man bør overveje at **korrigere p-værdierne** hvis man udfører mange tests.

Den simpleste metode er **Bonferronikorrektion:**

- Beregn p-værdier som sædvanlig for hvert test
- Gang p-værdierne med antallet af tests $\rightarrow$ justerede p-værdier
- For hver hypotese: Sammenlign den justerede p-værdier med sign.-niveauet (5%) for at afgøre om hypotesen skal afvises

Hvis man fx tester fem hypoteser:

- Hver af de oprindelige p-værdier skal ganges med 5 før sammenligning med 5%
- Altså: Hypoteser forkastes hvis den oprindelige p-værdi er 1%

Bonferroni korrektion

Hvis man bruger Bonferroni, kan man vise at **risikoen for at lave mindst en type I fejl** (mindst en falsk positiv) er $< 5\%$**.**

Man afviser færre hypoteser efter korrektionen.

Faktisk er Bonferroni meget **konservativ** (streng) og man får typisk for få signifikante tests.

Bedre metode: **Holm's metode** (se evt. opgave 6.15).

Bonferroni-korrektion: antibio datasættet

```
antibio$myType <- relevel(antibio$type, ref = "Control")
model2 <- lm(org ~ myType, data = antibio)
summary(model2)$coefficients
```

```
##                 Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)    2.6033333 0.04970149 52.379382 1.731916e-29
## myTypeAlfacyp  0.2916667 0.07028852  4.149563 2.810352e-04
## myTypeEnroflox 0.1066667 0.07028852  1.517555 1.403375e-01
## myTypeFenbenda 0.2300000 0.07028852  3.272227 2.834419e-03
## myTypeIvermect 0.3983333 0.07028852  5.667118 4.498656e-06
## myTypeSpiramyc 0.2516667 0.07858496  3.202479 3.383831e-03
```

- Hvilke fodertyper giver signifikant højere indhold af organisk materiale end kontrolgruppen?
- Betyder det noget om vi anvender Bonferroni-korrektion af de fem p-værdier?

De statistiske modeller har samme struktur

Fællestræk / generelle statistisk begreber

Lærebogens Kapitel 5.1-5.3 forsøger at opsummere og fremhæve fællestræk ved de tre statistiske modeller, som vi har set på

- enstikprøveproblemet / model for en (enkelt) stikprøve
- ensidet variansanalyse (ANOVA)
- lineær regression

Kapitlerne er bevidst skrevet abstrakt fordi man vil undgå at bruge konkrete formler som knytter sig til en bestemt model.

Håbet er at nogle studerende får en bedre intuition omkring centrale begreber af at læse kapitlerne.

På de følgende slides forsøger jeg at fremhæve nogle af pointerne ..

De statistiske modeller

Data: $y_1, \ldots, y_n$ samt evt. forklarende variabel

Har specificeret statistisk model således: $y_1, \ldots, y_n$ **uafhængige** og y_i **normalfordelt** med **middelværdi ***** og **spredning σ.**

Middelværdierne:

- En enkelt stikprøve: Middelværdi μ
- Ensidet ANOVA: Middelværdi $\alpha_{g(i)}$
- Lineær regression: Middelværdi $\alpha + \beta x_i$

De statistiske modeller har samme struktur

Ækvivalent måde at skrive modellerne på:

$$y_i = \text{middelværdi} + e_i$$

hvor **restleddene** (residualerne) $e_1, \ldots, e_n$ er uafhængige og allesammen normalfordelt med middelværdi 0 og spredning σ.

σ kaldes også **residualspredningen**.

- En enkelt stikprøve: $y_i = \mu + e_i$
- Ensidet ANOVA: $y_i = \alpha_{g(i)} + e_i$
- Lineær regression: $y_i = \alpha + \beta x_i + e_i$

iid

Antager at $e_1, \ldots, e_n$ er uafhængige og allesammen normalfordelt med middelværdi 0 og spredning σ.

Vi siger også at $e_1, \ldots, e_n$ er **iid** $N(0, \sigma^2)$-fordelt:

$$\begin{aligned} \text{iid} &= \text{independent with identical distributions} \\ &= \text{uafhængige og identisk fordelt} \end{aligned}$$

Parametre der bestemmer middelværdien

Lidt mere generelt kan vi skrive

$$y_i = \mu_i + e_i = f(x_i; \theta_1, \ldots, \theta_p) + e_i$$

hvor $\theta_1, \ldots, \theta_p$ er de ukendte **parametre** der bruges til at beskrive middelværdierne, og $e_1, \ldots, e_n$ er **iid** $N(0, \sigma^2)$

- En enkelt stikprøve: $y_i = \mu + e_i$. Parameteren er μ
- Ensidet ANOVA: $y_i = \alpha_{g(i)} + e_i$. Parametreme er $\alpha_1, \ldots, \alpha_k$
- Lineær regression: $y_i = \alpha + \beta x_i + e_i$. Parametrene er α og β

Estimater (teori)

Mindste kvadraters metode $\rightarrow$ **estimater** $\hat{\theta}_1, \ldots, \hat{\theta}_p$

Fittede værdier

$$\hat{y}_i = f(x_i; \hat{\theta}_1, \ldots, \hat{\theta}_p)$$

Residualer

$$r_i = \text{observeret} - \text{fittet} = y_i - \hat{y}_i$$

Estimat for residualspredningen σ:

$$s = \hat{\sigma} = \sqrt{\frac{1}{n-p}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Standard error og konfidensinterval

Standard error (estimeret spredning) for hvert θ_j har formen

$$\text{SE}(\hat{\theta}_j) = s\sqrt{k_j}$$

hvor k_j er en konstant som ikke afhænger af y_i'erne, men kun af modellen (kan beregne den før man har set data).

95% konfidensinterval for hver parameter θ_j:

$$\text{estimat} \ \pm \ t_{0.975,n-p} \ \cdot \text{standard error}$$

$$\hat{\theta}_j \ \pm \ t_{0.975,n-p} \cdot \text{SE}(\hat{\theta}_j)$$

Hypotesetest for et enkelt θ_j

Hypotese, $H_0 : \theta_j = \theta_0$ for en **præ-specificeret værdi** θ_0.

***T*-teststørrelsen**

$$T_{\text{obs}} = \frac{\text{estimat} - \text{hypoteseværdi}}{\text{SE(estimat)}} = \frac{\hat{\theta}_j - \theta_0}{\text{SE}(\hat{\theta})}$$

***p*-værdi** bereges i t-fordelingen med $\text{df} = n - p$:

$$p = P\big(|T| \geq |T_{\text{obs}}|\big) = 2 \cdot P\big(T \geq |T_{\text{obs}}|\big).$$

Hypotesen forkastes hvis *p*-værdien er lille, nemlig $< 5\%$.
Angiv altid selve p-værdien, ikke blot om den er $< 5\%$ eller ej.

Vi siger at vi bruger **5% signifikansniveau.**

R

Modeller fittes med `lm` og undersøges derefter med `summary`, `confint` og — som vi skal se — `drop1` og `anova`.

Output fra `summary`

- En linie for hver parameter, dvs. for hvert θ_j
- Alle tal i linien hører til denne parameter!
- t-teststørrelse og p-værdi er for hypotesen $H_0 : \theta_j = 0$.
  Derfor er t-værdien altid defineret som $\frac{\text{estimat}}{\text{SE}}$
- Residualspredning $s = \hat{\sigma}$ er angivet under tabellen

t-test i lineær regression

Middelværdi (af y):

$$\alpha + \beta \cdot x$$

Hypotese, $H_0 : \beta = 0$ (ingen sammenhæng mellem x og y)

Eks: 144 kattes hjertevægt (`y = Hwt`) og kropsvægt (`x = Bwt`)

```
##               Estimate Std. Error    t value     Pr(>|t|)
## (Intercept) -0.3566624  0.6922770 -0.5152019 6.072131e-01
## Bwt          4.0340627  0.2502615 16.1193908 6.969045e-34
```

Estimat: $\hat{\beta} = 4.0341$, **Standard Error**: $SE(\hat{\beta}) = 0.2503$

T-test: $T_{\text{obs}} = \frac{\text{estimat} - \text{hypoteseværdi}}{\text{SE(estimat)}} = \frac{\hat{\theta}_j - \theta_0}{\text{SE}(\hat{\theta})} = \frac{4.0341 - 0}{0.2503} = 16.119$

P-værdi: $p = P(T > |T_{\text{obs}}|) = 0$