TMA4267 Linear Statistical Models V2017 (L13)

Part 3: Hypothesis testing and analysis of variance
Hypothesis testing: why, how and be aware
Reproduciability
The universal F-test [F:3.3]

Mette Langaas

Department of Mathematical Sciences, NTNU

To be lectured: March 3, 2017

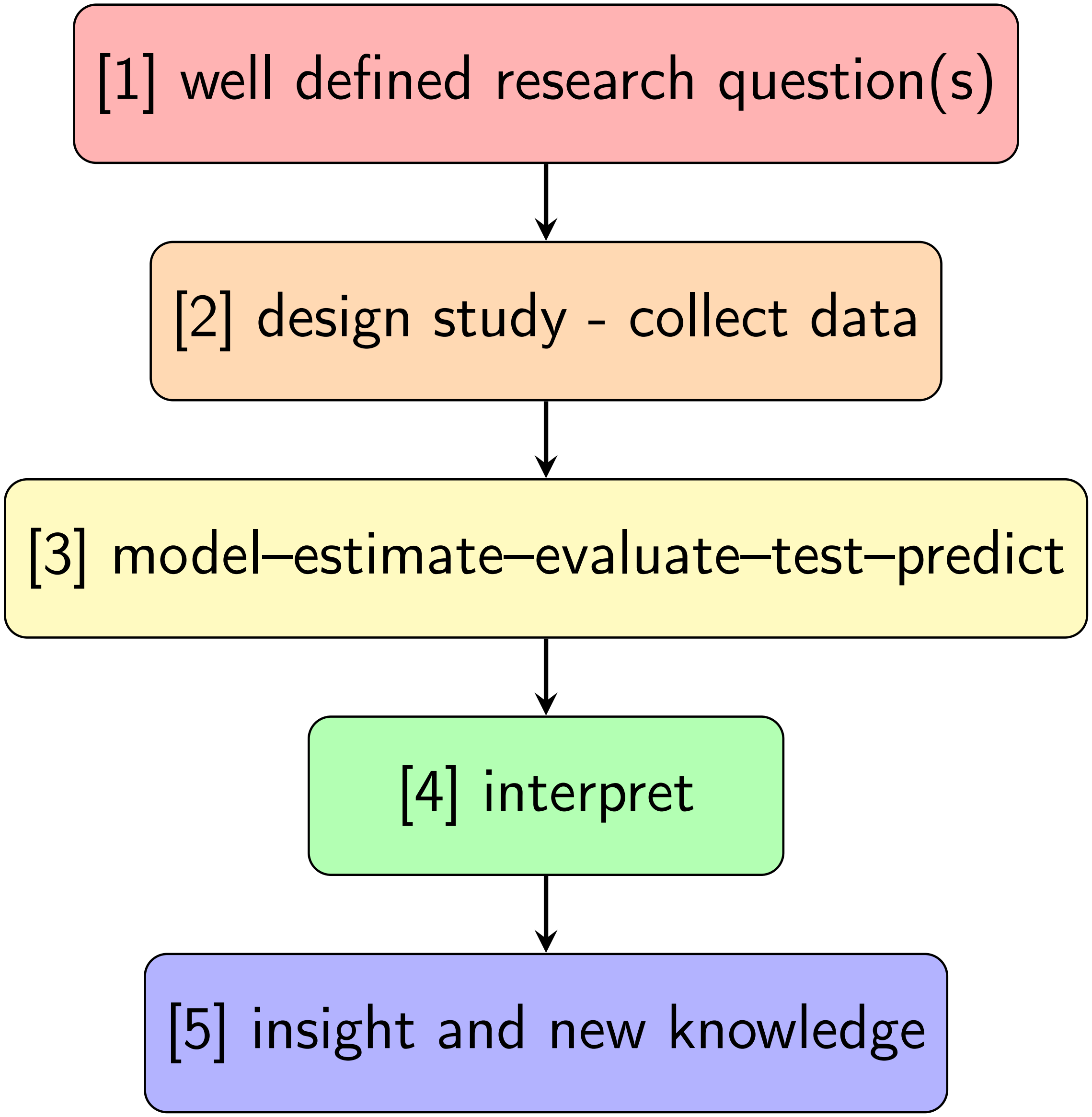
Today

- The scientific process.
- The basics of hypothesis testing and interpretation of p-value.
- The reproduciability "crisis".
- Properties of p-values.
- Linear hypotheses in regression vs. nested models.
- The universal F-test for linear hypotheses (nested models)

Basal metabolic rate and the FTO-gene

- The gene called FTO is known to be related to obesity
- The basal metabolic rate says how many calories you burn when you rest (hvilemetabolisme).
- Data has been collected for 101 patient from the obesity clinic at St. Olavs Hospital.
- Research question: is there an association between the variant of the FTO gene of the patient and the basal metabolic rate?
- Regression setting, other covariates include age, sex, weight, height, BMI, diet, exercise level, smoking, etc.

The scientific process



[1] well defined research question(s)

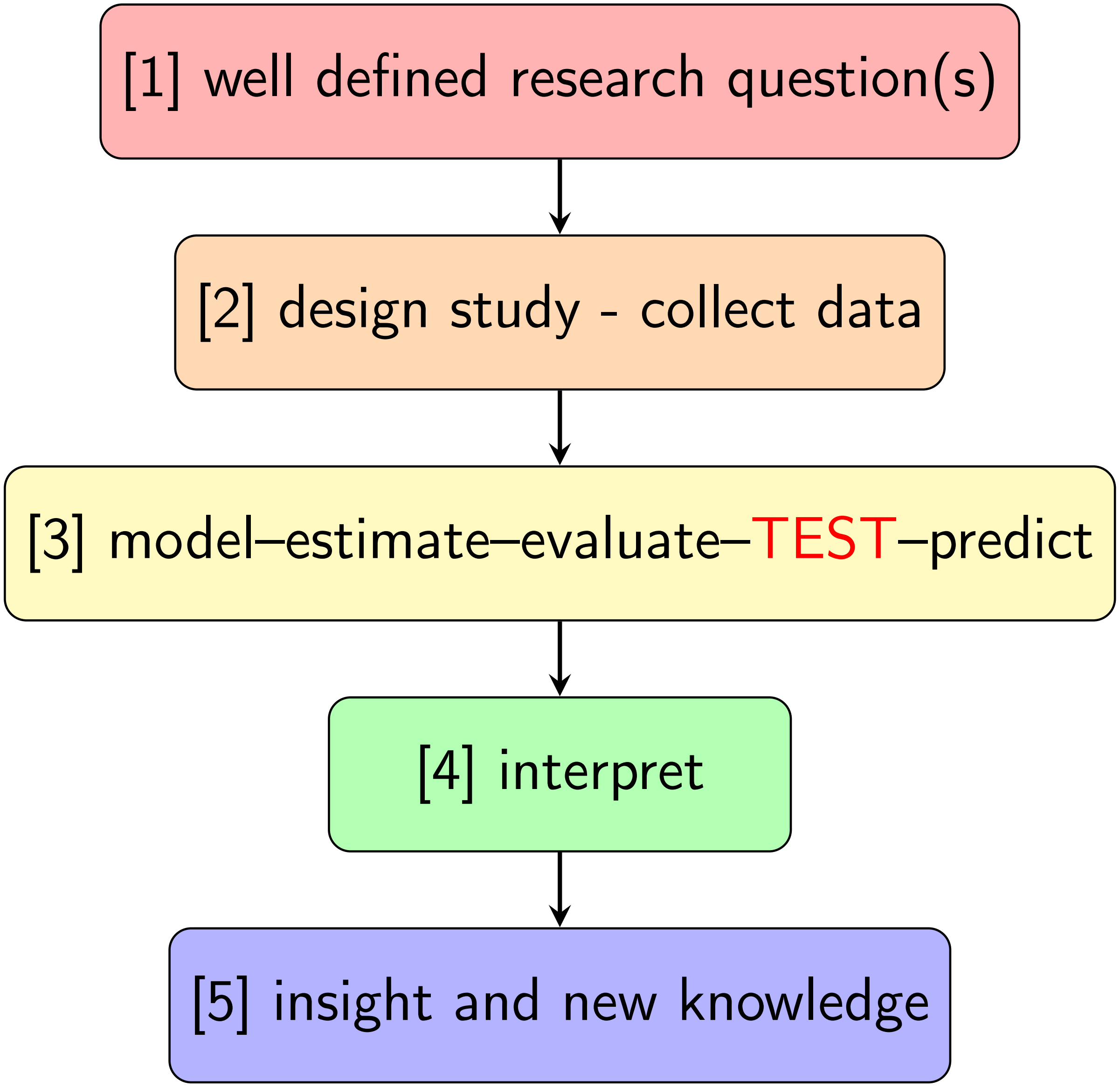
[2] design study - collect data

[3] model–estimate–evaluate–test–predict

[4] interpret

[5] insight and new knowledge

The scientific process



[1] well defined research question(s)

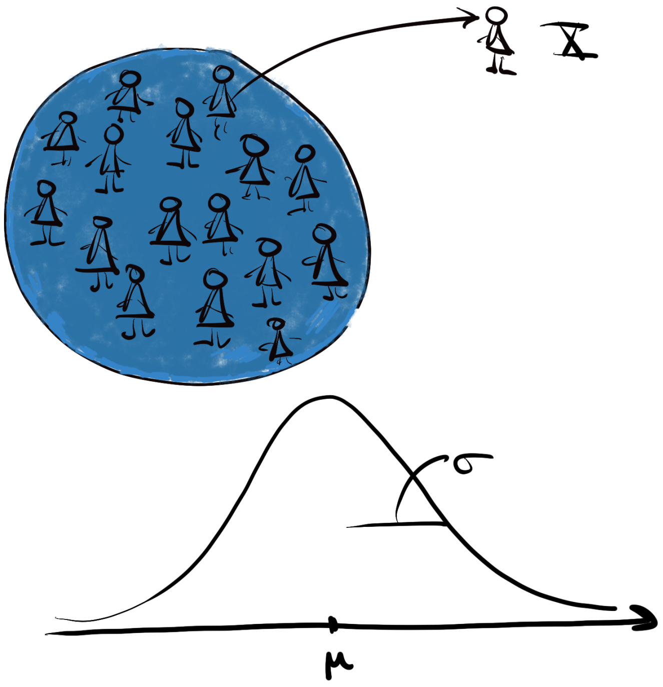
[2] design study - collect data

[3] model–estimate–evaluate–TEST–predict

[4] interpret

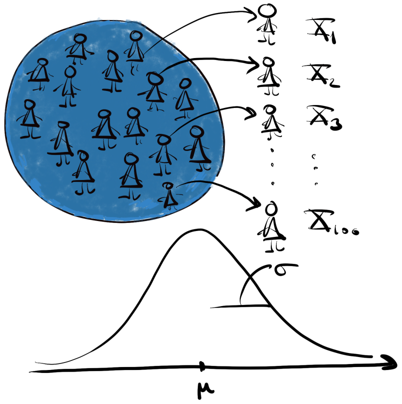
[5] insight and new knowledge

Hypothesis testing example



- It is known that in a population of women of age 20-29 years the systolic blood pressure is normally distributed with mean $\mu = 120$ mmHg.
- We study a population of women of age 20-29 that have a specific disease (blue population), and also here we assume that the systolic blood pressure is normally distributed (with standard deviation 10 mmHg), but here we don't know the mean in the population.
- In addition to estimating this unknown mean we want to investigate if the mean blood pressure of the blue population is larger than 120 mmHg (because if it is, we need to start more investigations into the cause of this).
- $H_0 : \mu = 120$ vs. $H_1 : \mu > 120$.

Hypothesis testing example (cont.)



- We draw a random sample of size $n = 100$ from the blue population and measure systolic blood pressure: $X_1, X_2, \dots, X_n$.
- Test statistic: $\bar{X} \sim N(120, 1)$ when H_0 is true.
- We find that $\bar{x} = 122$ mmHg.
- Data: $n = 100$, $\bar{x} = 122$, gives a p-verdi=0.02.

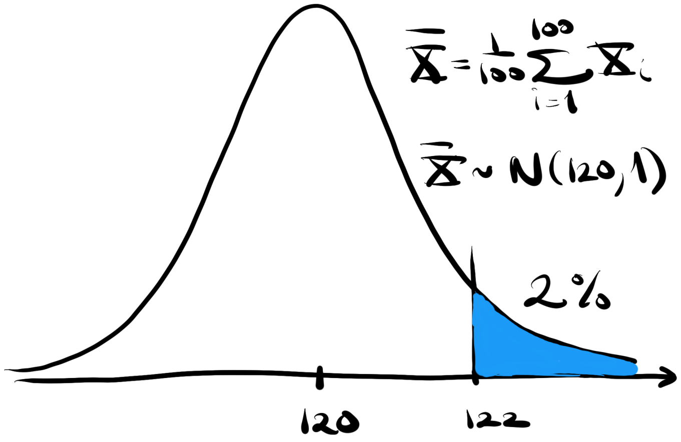
Questions:

- How have I calculated this p-value?
- Should I conclude that $\mu > 120$?

Q and A

- How have I calculated this p-value?
  $P(\bar{X} > 122 \mid H_0 \text{ true})$.
- Should I conclude that $\mu > 120$?
  Yes, if you choose significance level higher than 0.02. But, you should also report a (two-sided) confidence interval for μ:
  Here $[120.04, 123.96]$.

Hypothesis testing example (end)



- The *p*-value is often based on a test statistic, and can be found in many ways (known distribution, enumerations, asymptotic).

- Significance level: highest probability of miscarriage of justice that we would tolerate.
- We reject the null hypothesis - and say that we have a significant finding at significance level α if a/the *p*-value for the hypothesis test is below α.

What is a *p*-value

From *The research handbook of Carlsen & Staff (2014)*
. . . the *p*-value, the probability that the result could have occurred randomly, p=probability.

This is common, but not the correct definition of the *p*-value.
What is wrong? Discuss!

Slide reconstructed from talk by Kristoffer H. Hellton, NR

What is a *p*-value

A more correct definition so that:
the p-value is the probability of your result or a more extreme result, given that H_0 is true.

or

the probability of your result or a more extreme result, given that it occurred randomly.

This is different from: the probability of your result occurring randomly.

Slide reconstructed from talk by Kristoffer H. Hellton, NR

A simple example

- Null hypothesis: It is sunny outside.
- Data: I enter the room soaking wet.
- Wrong *p*-value: the probability that it is sunny outside.
- Impossible to calculate.
- Right *p*-value: the probability that I'm wet, given that it is sunny.
- Should be small.

Important! From Bayes theorem:
P(observation | hypothesis) $\neq$ P(hypothesis|observation)
The probability of observing a result given some hypothesis is true not equivalent to the probability that the hypothesis is true given that some result has be observed.
To be able to calculate the right hand side, we need P(hypothesis), the probability of the hypothesis. This is exactly what is introduced in Bayesian statistics through the so-called prior, and some see the Bayes factor as the replacement for *p*-values.

Slide reconstructed from talk by Kristoffer H. Hellton, NR

Statistical significance and p-values

On March 7, 2016, the American Statistical Association posted a statement on statistical significance and p-values - "clarifying several widely agreed upon principles underlying the proper use and interpretation of the p-value".

Statement on proper use and interpretation of the *p*-value

Why is this needed: (1)
American Statistical Association discussion forum, 2014.

- Q: Why do so many colleges and grad schools teach p = 0.05?
- A: Because that's still what the scientific community and journal editors use.
- Q: Why do so many people still use p = 0.05?
- A: Because that's what they were taught in college or grad school.

Problem?
Urban knowledge: Unless an hypothesis test results in a *p*-value below 0.05 there is no finding. So, in some journals a researcher will not be able to publish his paper unless the test performed has a *p*-value below 0.05.

Statement on proper use and interpretation of the *p*-value

Why is this needed: (2)

Hack your way to scientific glory

Ioannidis (2005): How many nonsignificant results have been studied before one research group has published its first significant finding?

Statement on proper use and interpretation of the *p*-value

Why is this needed: (3)

The journal *Basic and Applied Social Psychology* (editors Trafimow and Marks, 2015) put a *ban* on null hypothesis significance testing.

ASA Statement on Statistical Significance and *P*-values, March 2016

The ASA's statement on p-values: context, process, and purpose, Ronald L. Wasserstein & Nicole A. Lazar, The American Statistician, DOI:10.1080/00031305.2016.1154108.

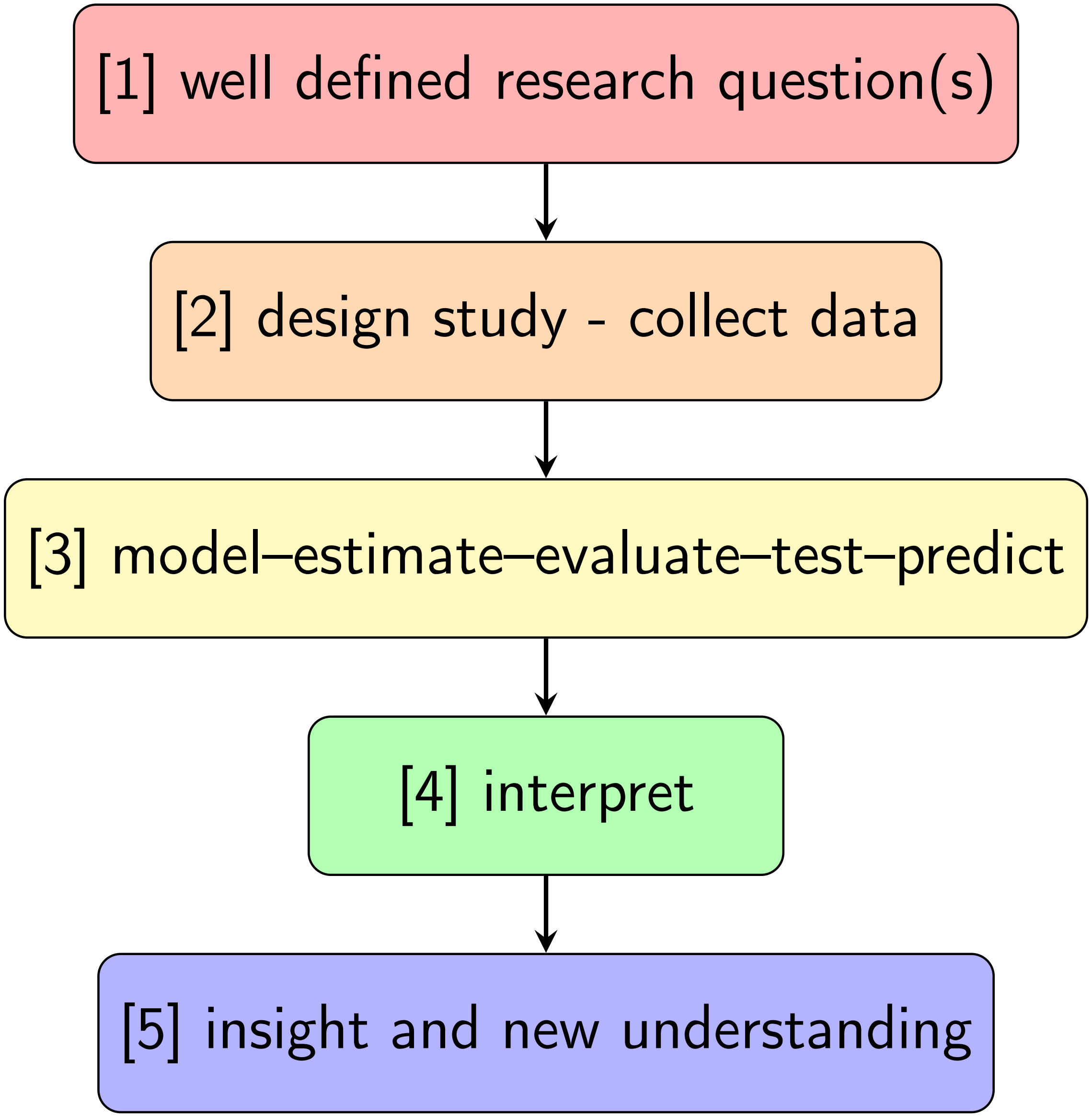
- While the *p*-value can be a useful statistical measure, it is commonly misused and misinterpreted.
- Informally, a *p*-value is the probability under a specified statistical model that a statistical summary of the data would be equal to or more extreme than its observed value.
- P1: *P*-values can indicate how incompatible the data are with a specified statistical model.
- P2: *P*-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.
- P3: Scientific conclusions and business or policy decisions should not be based only on whether at *p*-value passes a specific threshold.

ASA Statement on Statistical Significance and *P*-values

- P4: Proper inference requires full reporting and transparency.
- P5: A *p*-value, or statistical significance, does not measure the size of an effect or the importance of a result.
- P6: By itself, a *p*-value does not provide a good measure of evidence regarding a model or hypothesis.

Take home message: the *p*-value is a very risky tool ...
(Benjamini, 2016): but, replacing the *p*-value with other tools may lead to many of the same indeficiencies - so it would be better to instead focus on the appropriate use of statistical tools for addressing the crisis of reproducibility and replicability in science.

The scientific process



[1] well defined research question(s)

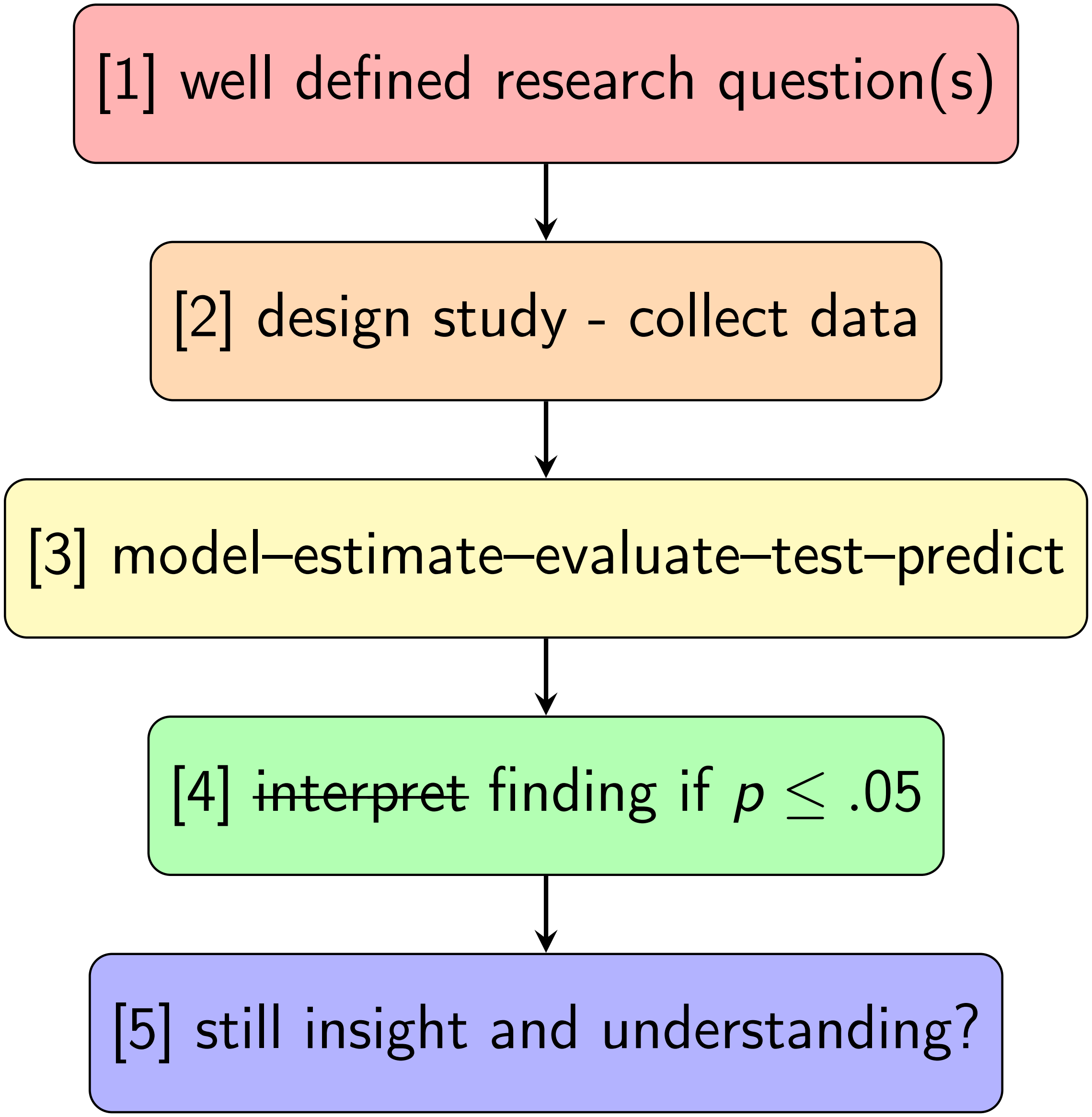
[2] design study - collect data

[3] model–estimate–evaluate–test–predict

[4] interpret

[5] insight and new understanding

Scenario: finding only for $p \leq 0.05$



[1] well defined research question(s)

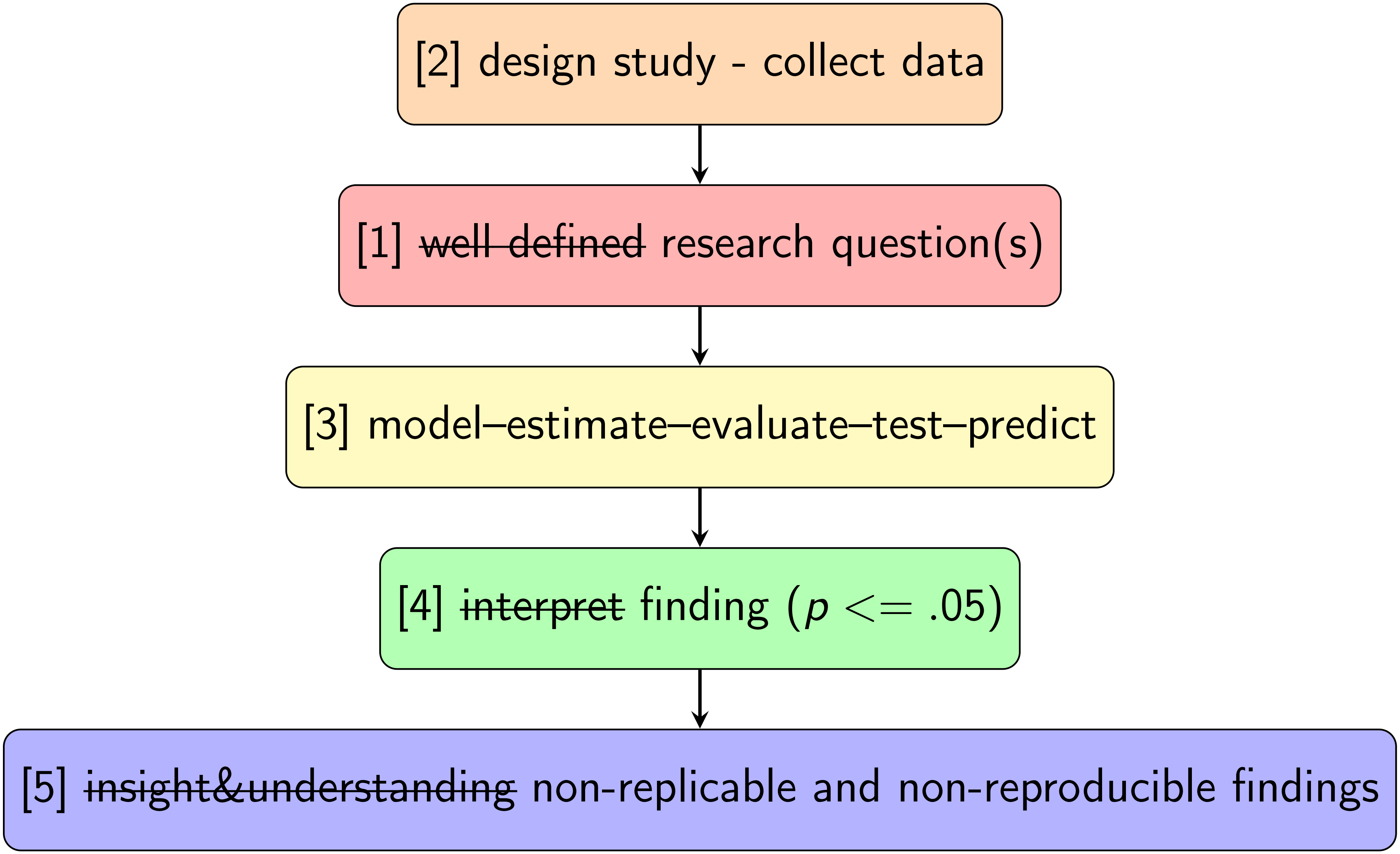
[2] design study - collect data

[3] model–estimate–evaluate–test–predict

[4] ~~interpret~~ finding if $p \leq .05$

[5] still insight and understanding?

Scenario: Cherry-picking aka Selective Inference aka *p*-hacking



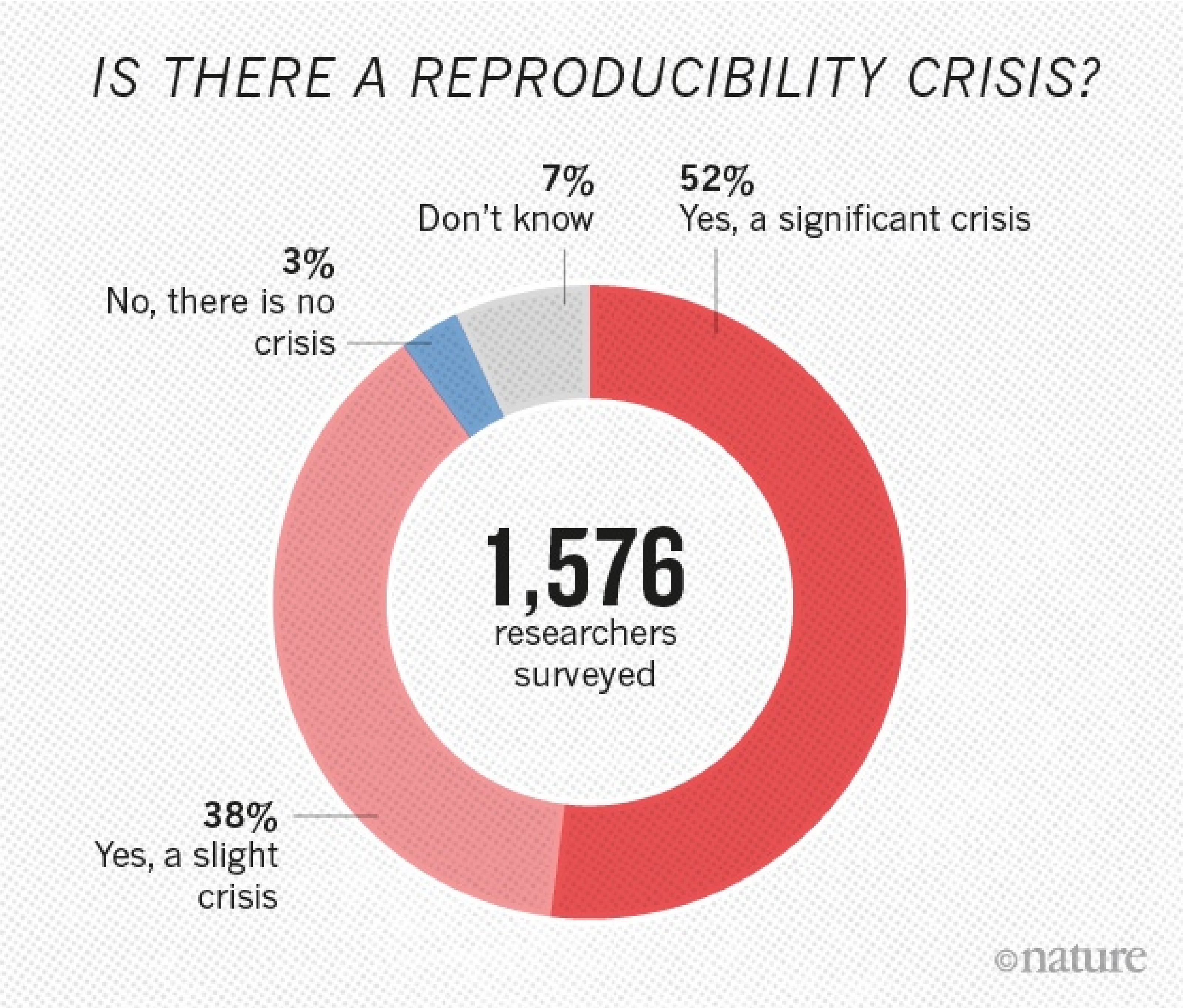
[2] design study - collect data

[1] ~~well defined~~ research question(s)

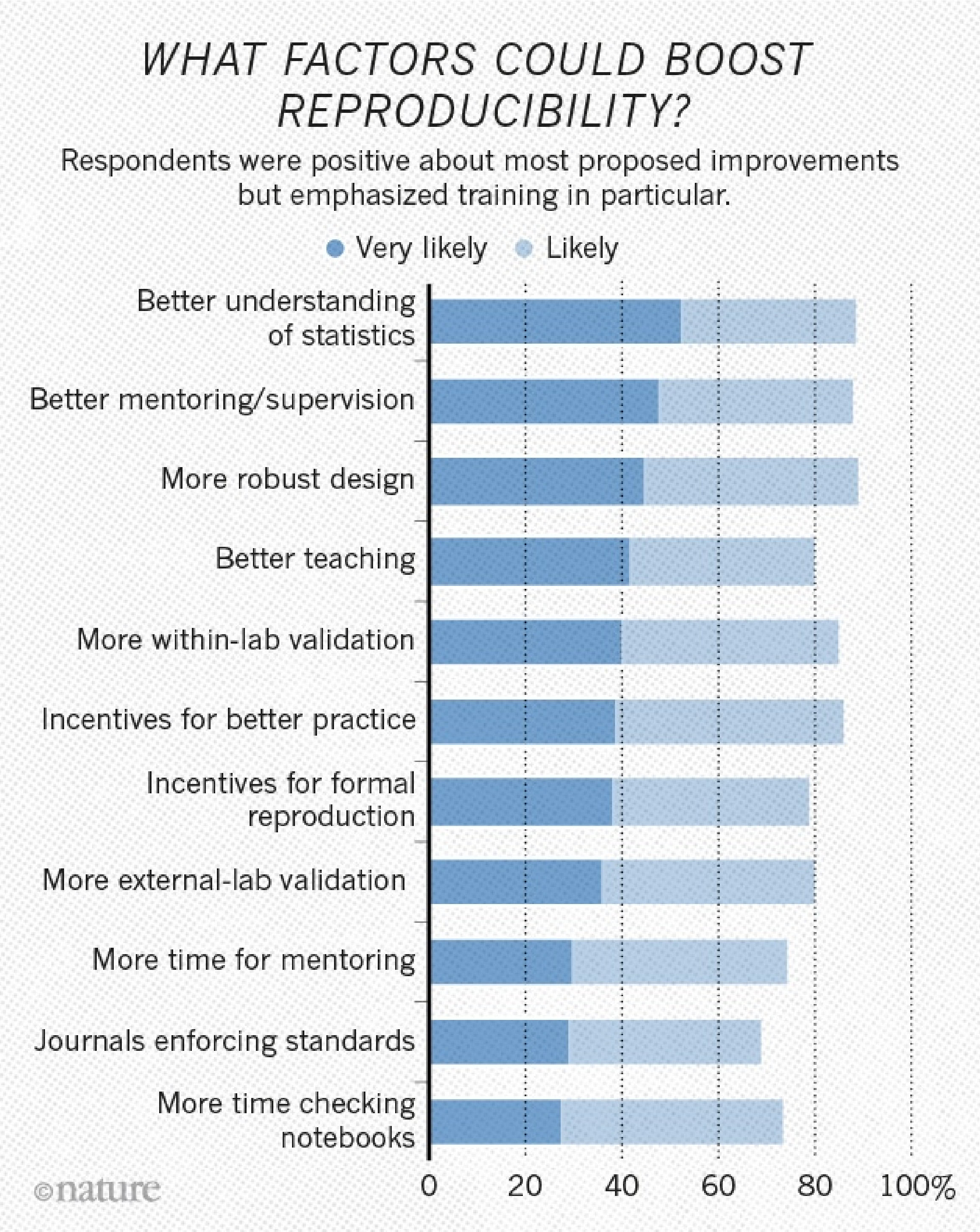
[3] model–estimate–evaluate–test–predict

[4] ~~interpret~~ finding ($p <= .05$)

[5] ~~insight&understanding~~ non-replicable and non-reproducible findings

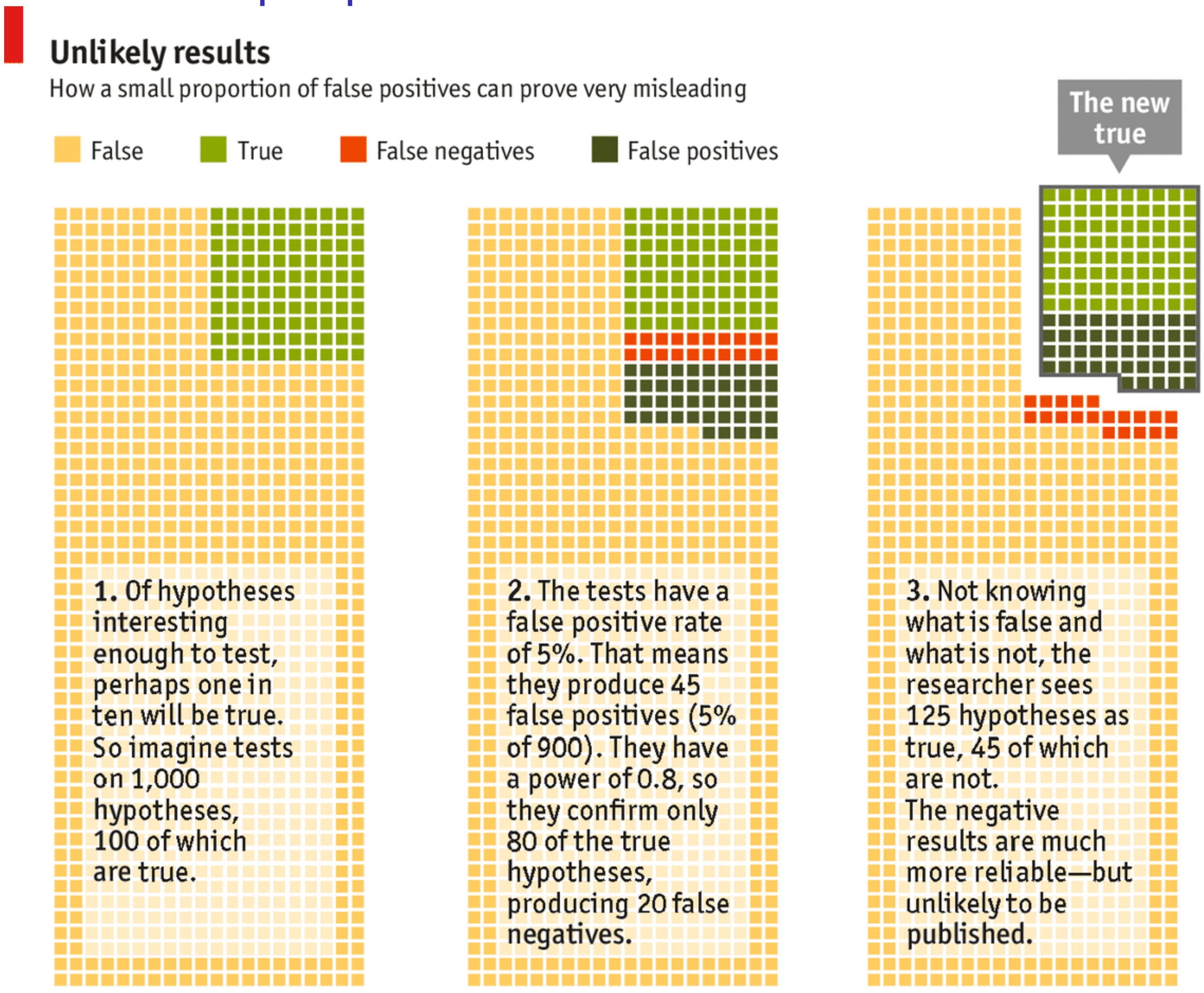


http://www.nature.com/news/1-500-scientists-lift-the-lid-on-reproducibility-1.19970



http://www.nature.com/news/1-500-scientists-lift-the-lid-on-reproducibility-1.19970

What is the proportion of fake news?



Unlikely results

How a small proportion of false positives can prove very misleading

1. Of hypotheses interesting enough to test, perhaps one in ten will be true. So imagine tests on 1,000 hypotheses, 100 of which are true.

2. The tests have a false positive rate of 5%. That means they produce 45 false positives (5% of 900). They have a power of 0.8, so they confirm only 80 of the true hypotheses, producing 20 false negatives.

3. Not knowing what is false and what is not, the researcher sees 125 hypotheses as true, 45 of which are not. The negative results are much more reliable—but unlikely to be published.

Source: *The Economist*

True=true H_1 (100 hypotheses) and False=false H_1 (900 hypotheses).

http://www.economist.com/news/briefing/21588057-scientists-think-science-self-correcting-alarming-degree-it-not-trouble

What is the proportion of fake news?

Color-coding for the far left figure:

- Yellow: all the hypotheses where H_0 is true (and H_1 is false), and H_0 is not rejected. All is good here, but this interesting(?) findings are very seldom published.
- Light green: all the hypotheses where H_0 is false (and H_1 is true) and the research reject the H_0 and make a correct discovery. This are our true news!
- Dark green: all the hypothesis where H_0 are true (and H_1 are false) but the researcher wrongly reject H_0. These are our fake news!
- Red: all the hypotheses where H_0 are false (and H_1 is true) but where the researcher fail to reject H_0 - let guilty criminal go free. These are called false negatives and are usually not reported (unless the researcher is report a negative finding).

So, not 5% of published results are false positives (fake news), but rather at substantially larger number - 40-90% has be hinted to in different publications.

Single hypothesis testing set-up

	H_0 true	H_0 false
Not reject H_0	Correct	Type II error
Reject H_0	Type I error	Correct

Two types of errors:

- False positives = type I error =miscarriage of justice. These are our *fake news*.
- False negatives = type II error= guilty criminal go free.

The significance level of the test is α.

We say that : Type I error is "controlled" at significance level α.

The probability of miscarriage of justice (Type I error) does not exceed α.

So far

- We (statisticians and other scientists) must focus on sound scientific process - and step away from cherry-picking and the "finding=p-value ≤ 0.05" urban truth.
- We must always report effect size.
- We must be aware that these two effects (selective inference and practical vs. statistical significance) are especially important for large than small data sets (both many samples and variables).
- Now, we move to hypothesis testing in linear regression and look at one unifying F-test can be used for all linear hypotheses.

Happiness ($n = 39$)

Are love and work the important factors determining happiness?

- y, `happiness`. 10-point scale, with 1 representing a suicidal state, 5 representing a feeling of «just muddling along», and 10 representing a euphoric state.
- x_1, `money`. Annual family income in thousands of dollars.
- x_2, `sex`. Sex was measured as the values 0 or 1, with 1 indicating a satisfactory level of sexual activity.
- x_3, `love`. 3-point scale, with 1 representing loneliness and isolation, 2 representing a set of secure relationships, and 3 representing a deep feeling of belonging and caring in the context of some family or community.
- x_4, `work`. 5-point scale, with 1 indicating that an individual is seeking other employment, 3 indicating the job is OK, and 5 indicating that the job is enjoyable.

Data taken from library faraway, data set happy.

What is $\boldsymbol{C}$ and $\boldsymbol{d}$?

Use the happiness data, with the four covariates x1=money, x2=sex, x3=love, x4=work, to construct the $\boldsymbol{C}$ and $\boldsymbol{d}$ to test $H_0 : \boldsymbol{C\beta} = \boldsymbol{d}$.

There is a linear effect in money? $H_0 : \beta_1 = 0$

$$\boldsymbol{C} = \left[\begin{array}{ccccc} 0 & 1 & 0 & 0 & 0 \end{array}\right], \boldsymbol{d} = 0$$

Is the regression significant?$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$

$$\boldsymbol{C} = \left[\begin{array}{ccccc} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{array}\right], \boldsymbol{d} = \left[\begin{array}{c} 0 \\ 0 \\ 0 \\ 0 \end{array}\right]$$

Is there a linear effect of money and/or sex? $H_0 : \beta_1 = \beta_2 = 0$

$$\boldsymbol{C} = \left[\begin{array}{ccccc} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{array}\right], \boldsymbol{d} = \left[\begin{array}{c} 0 \\ 0 \end{array}\right]$$

The Fisher distribution [F: B.1 Def 8.14], Exercise 2 Problem 5

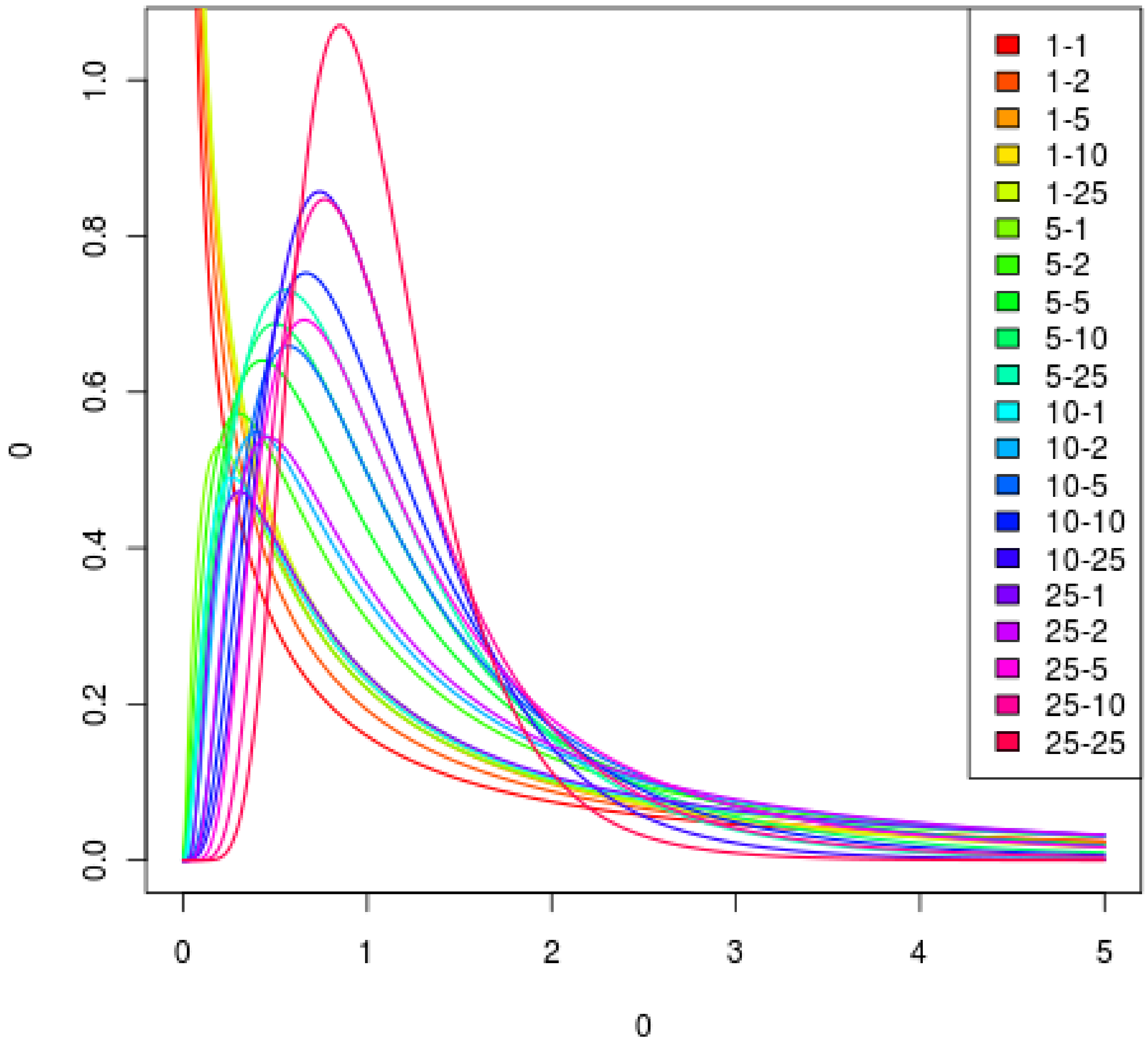
“Tabeller og formeler i statistikk”:
If Z_1 and Z_2 are independent and χ^2-distributed with ν_1 and ν_2 degrees of freedom, then

$$F = \frac{Z_1/\nu_1}{Z_2/\nu_2}$$

is F(isher)-distributed with ν_1 and ν_2 degrees of freedom.

- The expected value of F is $\mathrm{E}(F) = \frac{\nu_2}{\nu_2 - 2}$.
- The mode is at $\frac{\nu_1 - 2}{\nu_1}\frac{\nu_2}{\nu_2 + 2}$.
- Identity:

$$f_{1-\alpha,\nu_1,\nu_2} = \frac{1}{f_{\alpha,\nu_2,\nu_1}}$$



The Fisher distribution with different degrees of freedom ν_1 and ν_2 (given in the legend).

Unrestricted (Model A): all 4 covariates present

```
fitA <- lm(happy~.,data=happy)
summary(fitA)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.072081   0.852543  -0.085   0.9331
money        0.009578   0.005213   1.837   0.0749 .
sex         -0.149008   0.418525  -0.356   0.7240
love         1.919279   0.295451   6.496 1.97e-07 ***
work         0.476079   0.199389   2.388   0.0227 *
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’

Residual standard error: 1.058 on 34 degrees of freedom
Multiple R-squared:  0.7102,Adjusted R-squared:  0.6761
F-statistic: 20.83 on 4 and 34 DF,  p-value: 9.364e-09
```

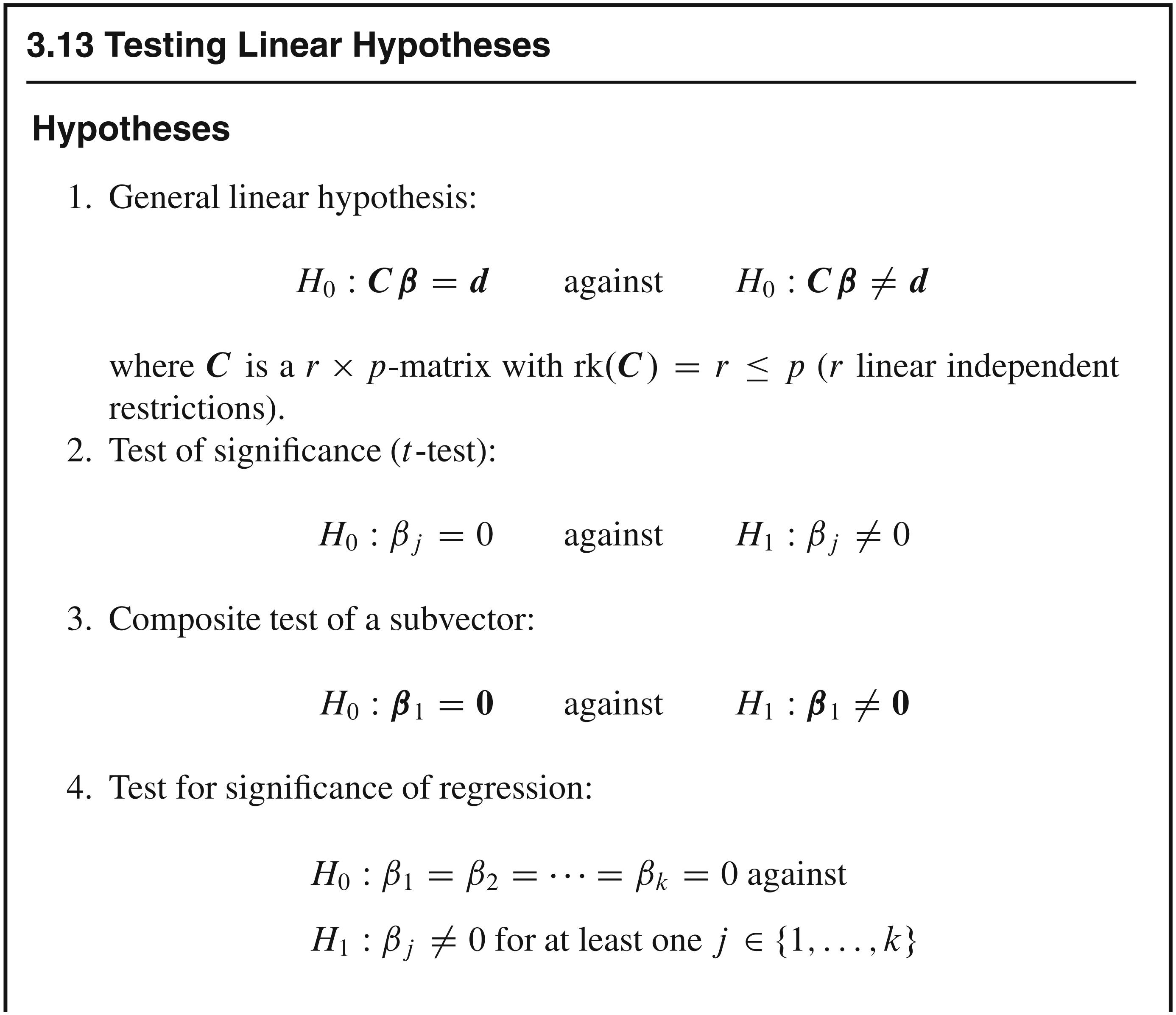
Restricted (Model B): only love and work

The estimate $\hat{\beta}_3$ (`love`) is 1.919 for model A and 1.959 for model B. Explain why these two estimates differ.

```
fitB <- lm(happy~love+work,data=happy)
summary(fitB)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.2057     0.7757   0.265  0.79241
love          1.9592     0.2954   6.633 9.99e-08 ***
work          0.5106     0.1874   2.725  0.00987 **
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’

Residual standard error: 1.08 on 36 degrees of freedom
Multiple R-squared:  0.6808,Adjusted R-squared:  0.6631
F-statistic: 38.39 on 2 and 36 DF,  p-value: 1.182e-09
```

Model A vs model B

```
> anova(fitA,fitB)
Analysis of Variance Table

Model 1: happy ~ money + sex + love + work
Model 2: happy ~ love + work
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     34 38.087
2     36 41.952 -2   -3.8651 1.7252 0.1934
```



3.13 Testing Linear Hypotheses

Hypotheses

1. General linear hypothesis:

$$H_0 : \boldsymbol{C}\boldsymbol{\beta} = \boldsymbol{d} \qquad \text{against} \qquad H_0 : \boldsymbol{C}\boldsymbol{\beta} \neq \boldsymbol{d}$$

   where $\boldsymbol{C}$ is a $r \times p$-matrix with $\text{rk}(\boldsymbol{C}) = r \leq p$ (r linear independent restrictions).
2. Test of significance (t-test):

$$H_0 : \beta_j = 0 \qquad \text{against} \qquad H_1 : \beta_j \neq 0$$

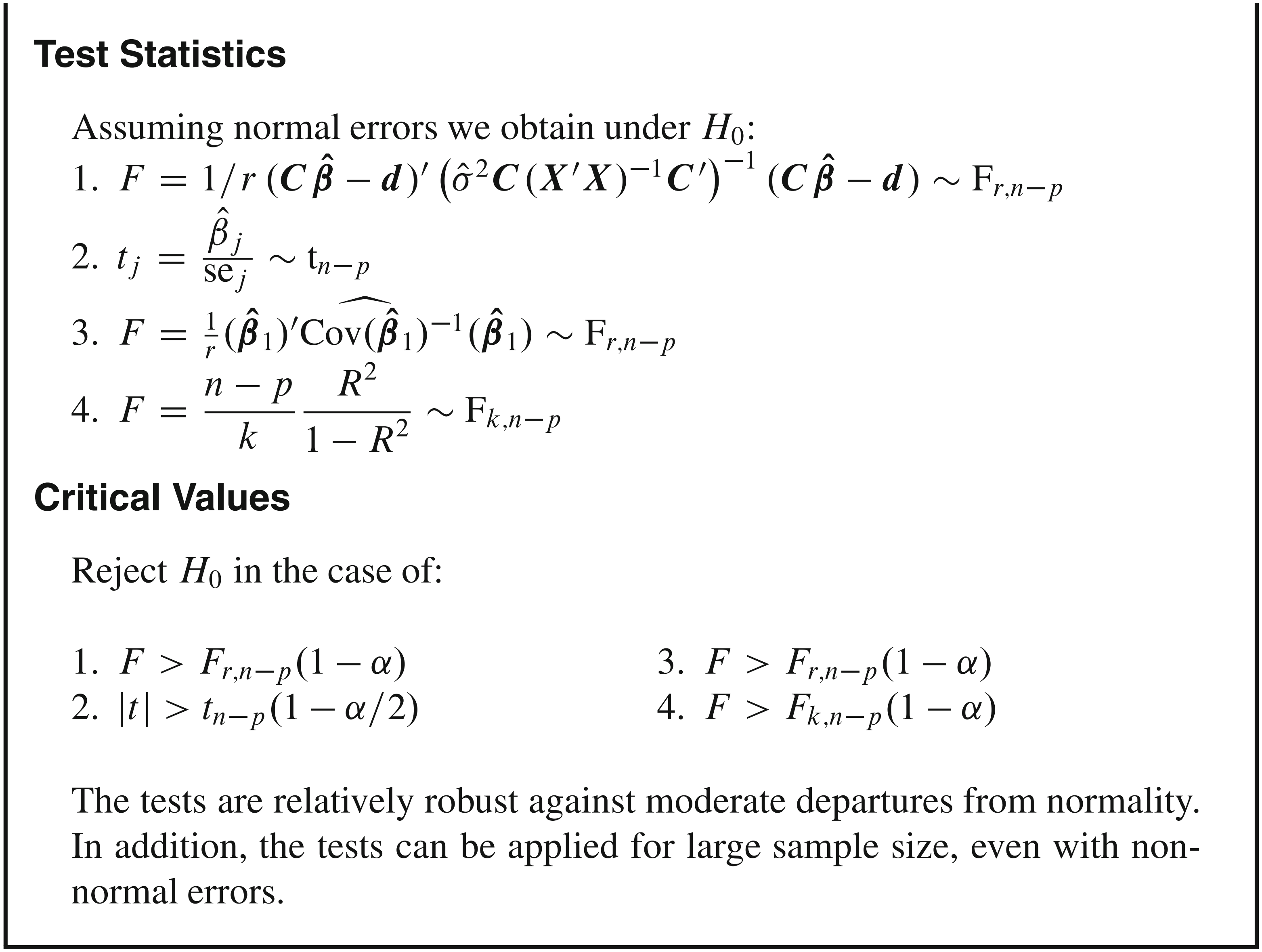
3. Composite test of a subvector:

$$H_0 : \boldsymbol{\beta}_1 = \boldsymbol{0} \qquad \text{against} \qquad H_1 : \boldsymbol{\beta}_1 \neq \boldsymbol{0}$$

4. Test for significance of regression:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0 \text{ against}$$
$$H_1 : \beta_j \neq 0 \text{ for at least one } j \in \{1, \ldots, k\}$$

Box from our text book: Fahrmeir et al (2013): Regression. Springer. (p.135)



Test Statistics

Assuming normal errors we obtain under H_0:

1. $F = 1/r\,(\boldsymbol{C}\hat{\boldsymbol{\beta}} - \boldsymbol{d})' \left(\hat{\sigma}^2 \boldsymbol{C}\,(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{C}'\right)^{-1} (\boldsymbol{C}\hat{\boldsymbol{\beta}} - \boldsymbol{d}) \sim \mathrm{F}_{r,n-p}$
2. $t_j = \dfrac{\hat{\beta}_j}{\mathrm{se}_j} \sim \mathrm{t}_{n-p}$
3. $F = \frac{1}{r}(\hat{\boldsymbol{\beta}}_1)'\widehat{\mathrm{Cov}(\hat{\boldsymbol{\beta}}_1)}^{-1}(\hat{\boldsymbol{\beta}}_1) \sim \mathrm{F}_{r,n-p}$
4. $F = \dfrac{n-p}{k}\dfrac{R^2}{1-R^2} \sim \mathrm{F}_{k,n-p}$

Critical Values

Reject H_0 in the case of:

1. $F > F_{r,n-p}(1-\alpha)$
2. $|t| > t_{n-p}(1-\alpha/2)$
3. $F > F_{r,n-p}(1-\alpha)$
4. $F > F_{k,n-p}(1-\alpha)$

The tests are relatively robust against moderate departures from normality. In addition, the tests can be applied for large sample size, even with non-normal errors.

Box from our text book: Fahrmeir et al (2013): Regression. Springer. (p.135)

Today

- Reproduciable research and the scientific method.
- Hypothesis testing and p-values in general.
- Type I errors=false positives=fake news.
- Linear hypotheses, and the F_{obs} test statistic.

PART 3: HYPOTHESIS TESTING AND ANALYSIS OF VARIANCE (ANOVA)

TMA4267 L13
07.03.2017

4 lectures + 1 RecEx + 1 ComalsoryEx

Hypothesis testing in linear regression [f. 3.7]

$Y = X\beta + \varepsilon \,, \quad \varepsilon \sim N_n(0, \sigma^2 I)$

So far we have looked at two types of hypotheses:

1) Test for significance of one β_j (Ex: β_{money} happiness)

$H_0: \beta_j = 0$ vs. $H_1: \beta_j \neq 0$

$\Rightarrow$ summary(lm-model) automatically added.

2) Is the regression significant?

$H_0: \beta_1 = \beta_2 = \dots = \beta_k = 0$ vs H_1: at least one $\neq 0$

in addition we migth want to

3) Test of equality (Ex Munich rent: top vs good location)

$H_0: \beta_j - \beta_r = 0$ vs $H_1: \beta_j - \beta_r \neq 0$

All of these situations can be written as a general linear hypothesis

$H_0: C\beta = d$ vs $H_1: C\beta \neq d$

$r \times p$ matrix

$r \times 1$ vector of constants

r linearly independent constraints under H_0

$\operatorname{rank}(C) = r \leq p$

restricted model

model B

unrestricted model

model A

Model B is a subset as model A.

Ex: Happiness, find C:

1) Is ther a linear effect of money? $\beta_0\ \beta_1\ \beta_2\ \beta_3\ \beta_4$

$H_0: \beta_1 = 0$ vs $H_1: \beta_1 \neq 0$

$\underset{1\times 5}{C} = [0\ 1\ 0\ 0\ 0]$ $\quad \underset{1\times 1}{d} = 0$

$r = 1$

$C\beta = d \Leftrightarrow \beta_1 = 0$

2) Is the regression significant?

$H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$ vs H_1: at leest one $\neq 0$

$$\underset{4\times5}{C} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \qquad d = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

$r = 4$

3) Is ther a lineer effect of money and/or sex?

$H_0: \beta_1 = \beta_2 = 0$ vs H_1: at leest one $\neq 0$

$$\underset{2\times5}{C} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{bmatrix} \qquad d = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$r = 2$

Procedure (for testing linear hypotheses)

Urestricted model vs Restricted model

$Y = X\beta + \varepsilon$, $\varepsilon \sim N_n(0,\sigma^2)$ $\nearrow C\beta = d$

Ex: "$\beta_1 = 0$" money exemple

Unrestricted $\overset{(A)}{}$: fit all covenates: x_1, x_2, x_3, x_4

Restricted $\overset{(B)}{}$: fit only: x_2, x_3, x_4

i) Fit the unrestricted model (A) end compute $SSE = \hat{\varepsilon}^T \hat{\varepsilon}$. Assume p regr. perem. fitted.

ii) Fit the restricted model (B) end compute $SSE_{H_0} = \hat{\varepsilon}_{H_0}^T \hat{\varepsilon}_{H_0}$

NB: the restricted model needs to be nested within the unrestricted.

iii) Calculate the test statistic:

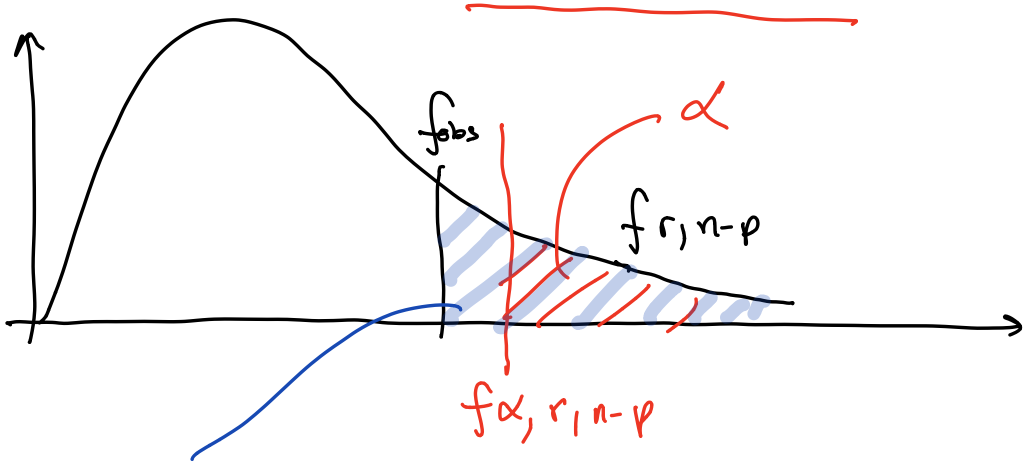
$$F_{obs} = \frac{\frac{1}{r}\Delta SSE}{\frac{1}{n-p} SSE} = \frac{\frac{1}{r}\left(\overbrace{SSE_{H_0} - SSE}^{\geq 0}\right)}{\frac{1}{n-p} SSE}$$

Q: What is the relationship between SSE_{H_0} and SSE ?

$$\underset{\substack{\uparrow \\ \text{unrestricted} \\ =\text{larger model}}}{SSE} \le SSE_{H_0}$$

iv) Under H_0 : $F_{obs} \sim F_{r,n-p}$

Reject H_0 when $\underline{f_{obs} > f_{\alpha,r,n-p}}$



p-value: $P(F_{r,n-p} > f_{obs})$

How to use this procedure? How to find F_{obs}?

a) Math formula for F_{obs} based on X, C, d, and $\hat{\beta}$, $\hat{\sigma}^2$ from the unrestricted model (A)
↖ on Tuesday!

b) If possible: fit unrestricted model and restricted model and read off SSE to get F_{obs}.

Ex: Happiness: $\overset{H_0:}{\beta_1=\beta_2=0}$ $\qquad \hat{\sigma}^2 = \frac{SSE}{n-p}$

A: Full model: $x_1\ x_2\ x_3\ x_4$

$$SSE = \hat{\sigma}^2\cdot(n-p) = (1.058)^2\cdot 34 = 38.087$$

B: Restricted model: $x_3\ x_4$

$$SSE_{H_0} = (1.08)^2\cdot 36 = 41.952$$

$$F_{obs} = \frac{\frac{1}{2}(41.952-38.087)}{\frac{1}{34}38.087} = 1.752$$

$$\text{p-value} = P(F_{2,34} > 1.752) = 0.1934$$

⇒ do not reject H_0: we prefer the smaller model "$x_3 + x_4$".

$H_0: \beta_1=\beta_2=0$ $\qquad H_1$: at least one $\neq 0$

TMA4267 Linear Statistical Models V2017 (L14)

Part 3: Hypothesis testing and analysis of variance
The universal F-test [F:3.3]
One-way ANOVA [H:8.1.1]

Mette Langaas

Department of Mathematical Sciences, NTNU

To be lectured: March 7, 2017

Today

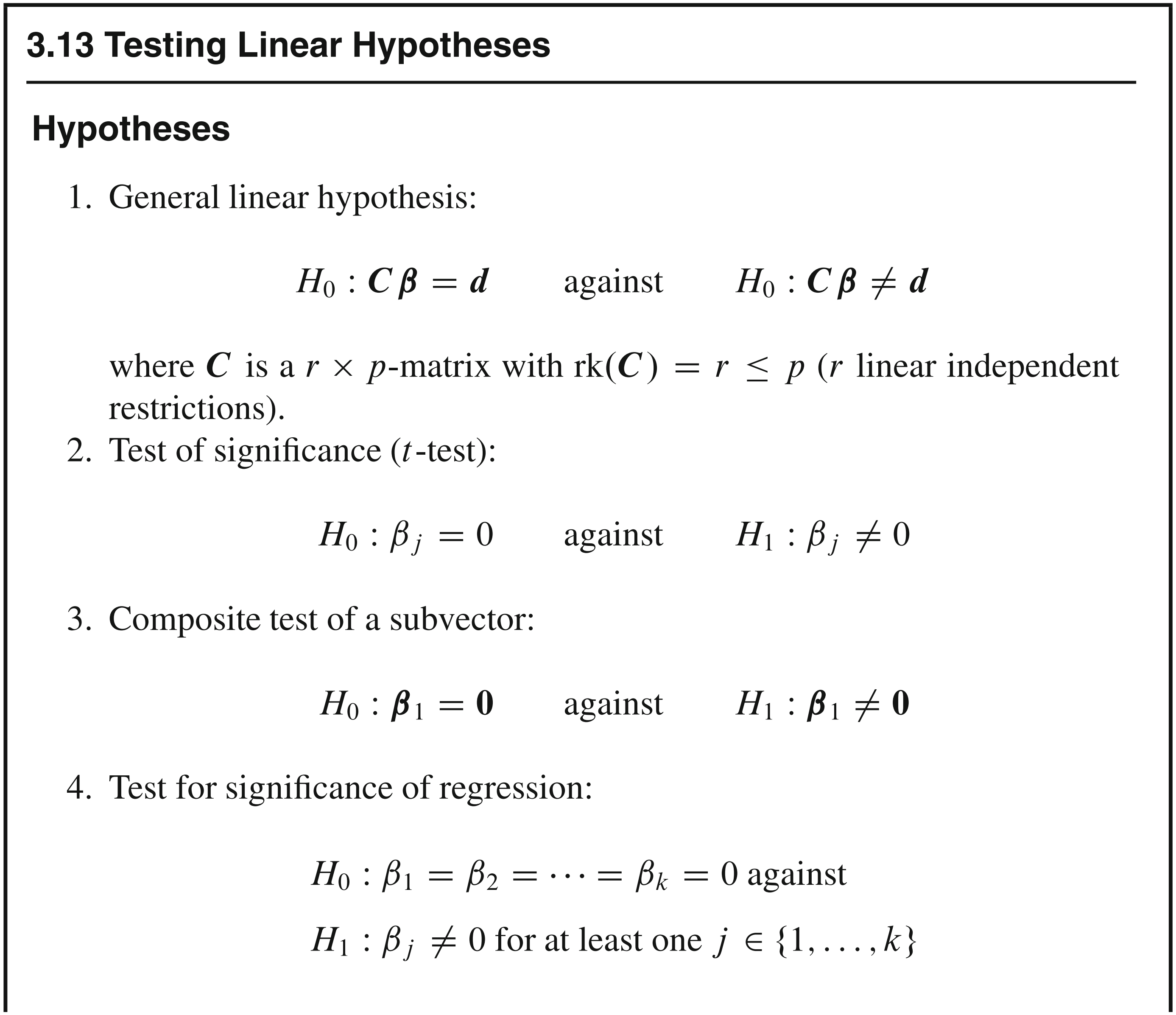
- Linear hypotheses in regression vs. nested models.
- The universal F-test for linear hypotheses: two formulas.
- The two formulas: one easy to use, one easy for proving F-distribution.
- Special cases of the universal F-test.
- New problem: categorical covariate with effect coding (for interpretation)

Happiness ($n = 39$)

Are love and work the important factors determining happiness?

- y, `happiness`. 10-point scale, with 1 representing a suicidal state, 5 representing a feeling of «just muddling along», and 10 representing a euphoric state.
- x_1, `money`. Annual family income in thousands of dollars.
- x_2, `sex`. Sex was measured as the values 0 or 1, with 1 indicating a satisfactory level of sexual activity.
- x_3, `love`. 3-point scale, with 1 representing loneliness and isolation, 2 representing a set of secure relationships, and 3 representing a deep feeling of belonging and caring in the context of some family or community.
- x_4, `work`. 5-point scale, with 1 indicating that an individual is seeking other employment, 3 indicating the job is OK, and 5 indicating that the job is enjoyable.

Data taken from library faraway, data set happy.



3.13 Testing Linear Hypotheses

Hypotheses

1. General linear hypothesis:

$$H_0 : \boldsymbol{C}\boldsymbol{\beta} = \boldsymbol{d} \qquad \text{against} \qquad H_0 : \boldsymbol{C}\boldsymbol{\beta} \neq \boldsymbol{d}$$

   where $\boldsymbol{C}$ is a $r \times p$-matrix with $\text{rk}(\boldsymbol{C}) = r \leq p$ (r linear independent restrictions).

2. Test of significance (t-test):

$$H_0 : \beta_j = 0 \qquad \text{against} \qquad H_1 : \beta_j \neq 0$$

3. Composite test of a subvector:

$$H_0 : \boldsymbol{\beta}_1 = \boldsymbol{0} \qquad \text{against} \qquad H_1 : \boldsymbol{\beta}_1 \neq \boldsymbol{0}$$

4. Test for significance of regression:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0 \text{ against}$$
$$H_1 : \beta_j \neq 0 \text{ for at least one } j \in \{1, \ldots, k\}$$

Box from our text book: Fahrmeir et al (2013): Regression. Springer. (p.135)

Constrained and unconstrained estimate

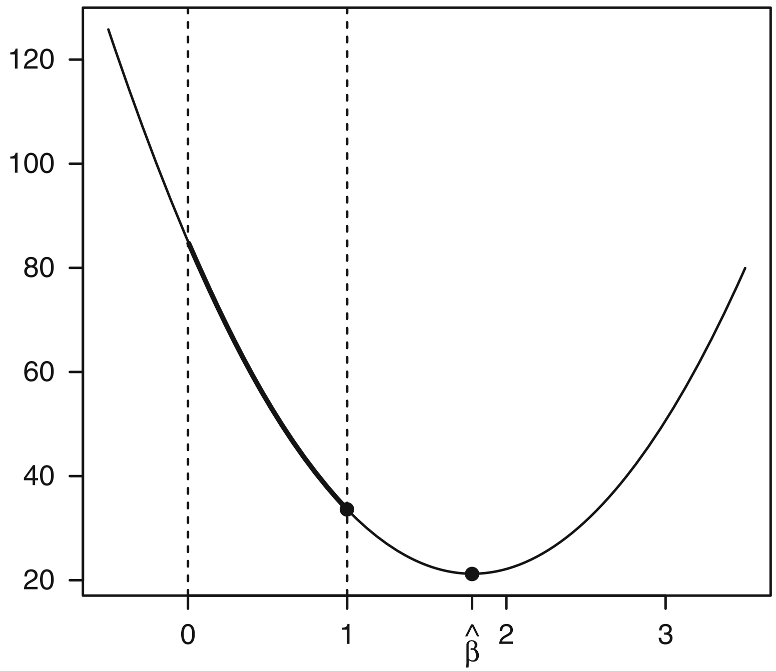
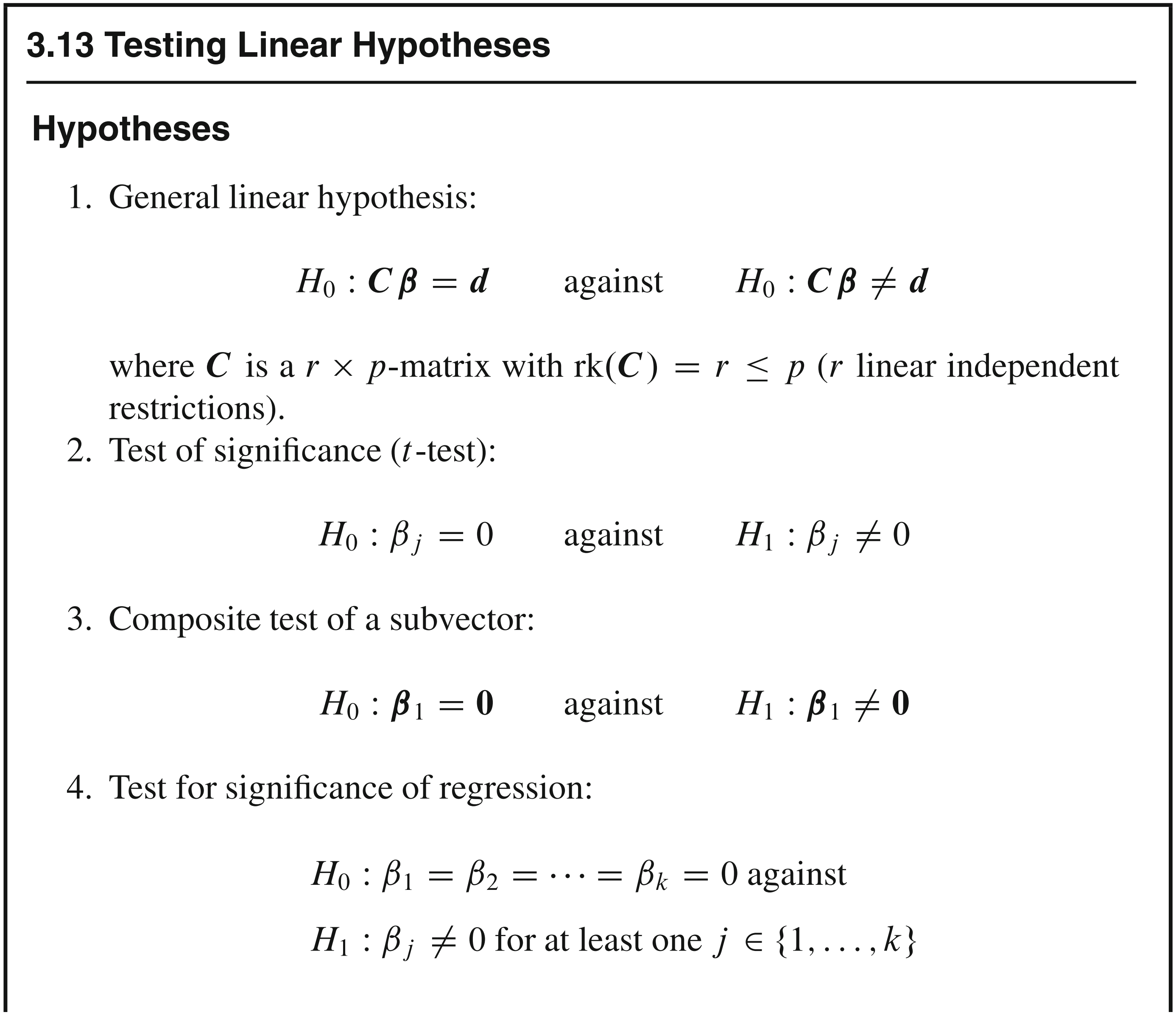


Fig. 3.15 Illustration of the difference in goodness of fit between the unconstrained least squares estimator and the estimator under the constraint $0 \leq \beta \leq 1$. The (unconstrained) least squares estimator is labeled as $\hat{\beta}$. For the constrained solution, we have $\hat{\beta} = 1$

Figure 3.15 from our text book: Fahrmeir et al (2013): Regression. Springer. (p.1329)



3.13 Testing Linear Hypotheses

Hypotheses

1. General linear hypothesis:

$$H_0 : \boldsymbol{C}\boldsymbol{\beta} = \boldsymbol{d} \qquad \text{against} \qquad H_0 : \boldsymbol{C}\boldsymbol{\beta} \neq \boldsymbol{d}$$

   where $\boldsymbol{C}$ is a $r \times p$-matrix with $\text{rk}(\boldsymbol{C}) = r \leq p$ (r linear independent restrictions).

2. Test of significance (t-test):

$$H_0 : \beta_j = 0 \qquad \text{against} \qquad H_1 : \beta_j \neq 0$$

3. Composite test of a subvector:

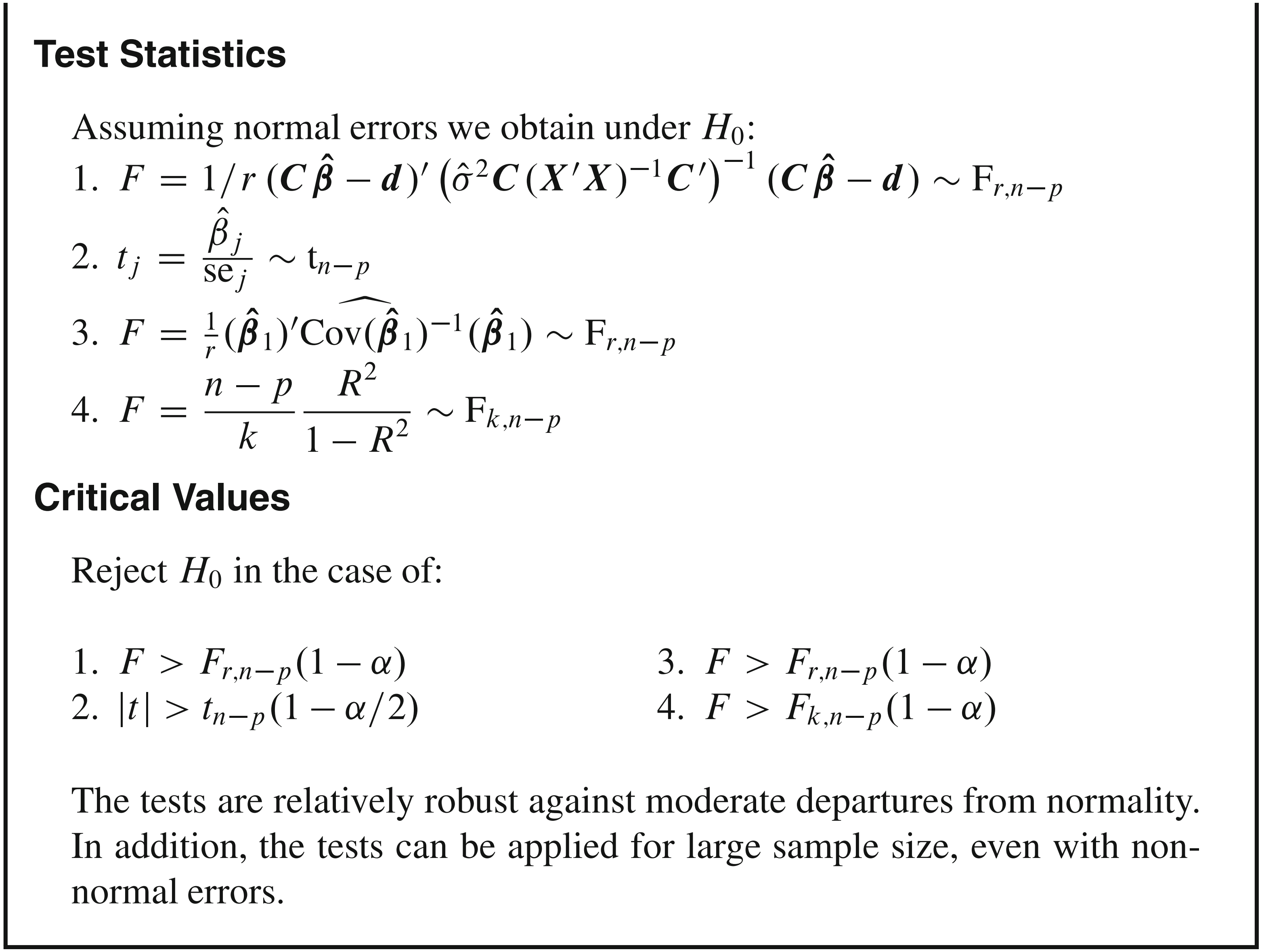
$$H_0 : \boldsymbol{\beta}_1 = \boldsymbol{0} \qquad \text{against} \qquad H_1 : \boldsymbol{\beta}_1 \neq \boldsymbol{0}$$

4. Test for significance of regression:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0 \text{ against}$$

$$H_1 : \beta_j \neq 0 \text{ for at least one } j \in \{1, \ldots, k\}$$

Box from our text book: Fahrmeir et al (2013): Regression. Springer. (p.135)



Test Statistics

Assuming normal errors we obtain under H_0:

1. $F = 1/r\,(\boldsymbol{C}\hat{\boldsymbol{\beta}} - \boldsymbol{d})' \left(\hat{\sigma}^2 \boldsymbol{C}\,(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{C}'\right)^{-1} (\boldsymbol{C}\hat{\boldsymbol{\beta}} - \boldsymbol{d}) \sim \mathrm{F}_{r,n-p}$
2. $t_j = \dfrac{\hat{\beta}_j}{\mathrm{se}_j} \sim \mathrm{t}_{n-p}$
3. $F = \frac{1}{r}(\hat{\boldsymbol{\beta}}_1)'\widehat{\mathrm{Cov}(\hat{\boldsymbol{\beta}}_1)}^{-1}(\hat{\boldsymbol{\beta}}_1) \sim \mathrm{F}_{r,n-p}$
4. $F = \dfrac{n-p}{k}\dfrac{R^2}{1-R^2} \sim \mathrm{F}_{k,n-p}$

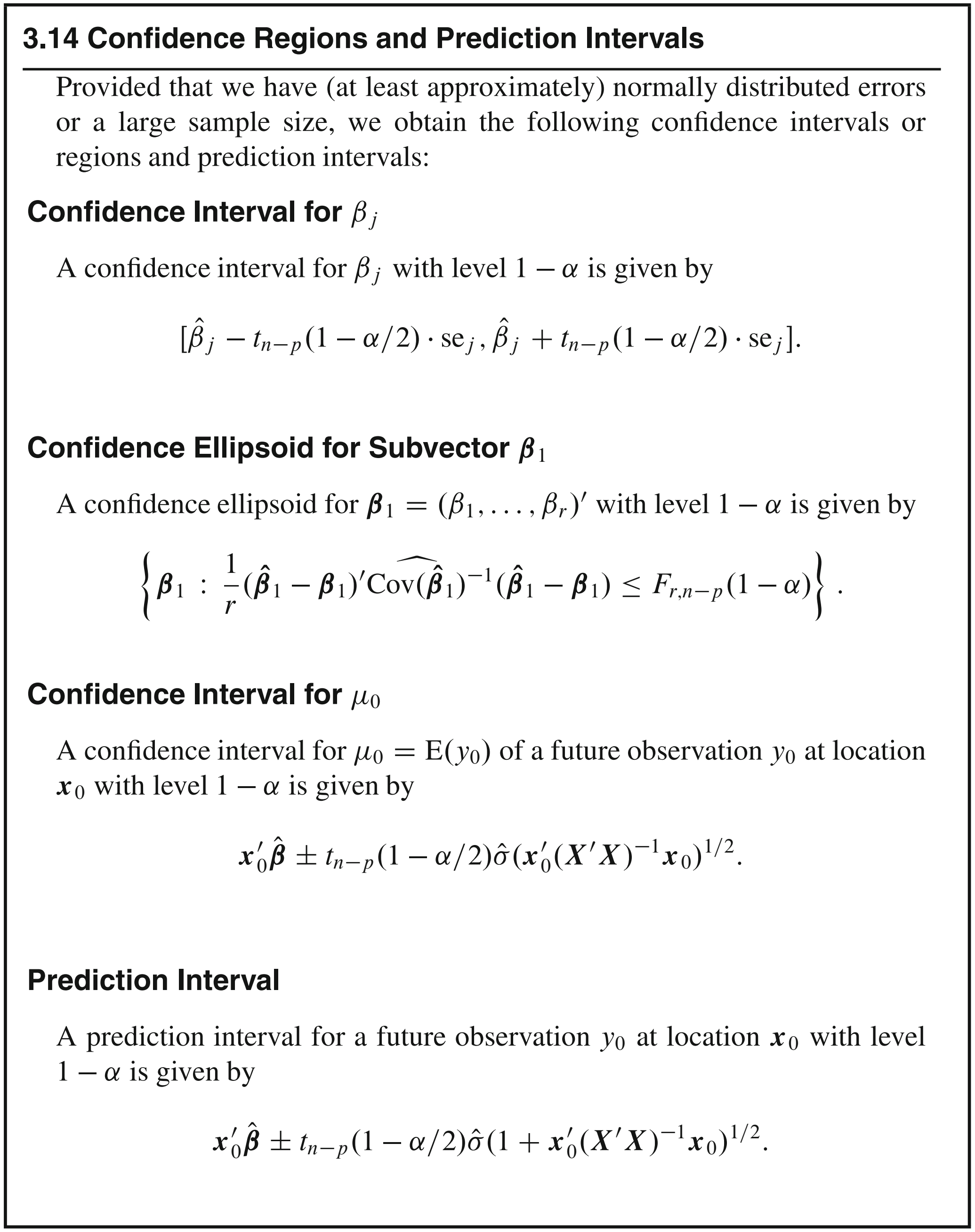
Critical Values

Reject H_0 in the case of:

1. $F > F_{r,n-p}(1-\alpha)$
2. $|t| > t_{n-p}(1-\alpha/2)$
3. $F > F_{r,n-p}(1-\alpha)$
4. $F > F_{k,n-p}(1-\alpha)$

The tests are relatively robust against moderate departures from normality. In addition, the tests can be applied for large sample size, even with non-normal errors.

Box from our text book: Fahrmeir et al (2013): Regression. Springer. (p.135)



3.14 Confidence Regions and Prediction Intervals

Provided that we have (at least approximately) normally distributed errors or a large sample size, we obtain the following confidence intervals or regions and prediction intervals:

Confidence Interval for β_j

A confidence interval for β_j with level $1-\alpha$ is given by

$$[\hat{\beta}_j - t_{n-p}(1-\alpha/2)\cdot \mathrm{se}_j\,, \hat{\beta}_j + t_{n-p}(1-\alpha/2)\cdot \mathrm{se}_j].$$

Confidence Ellipsoid for Subvector $\boldsymbol{\beta}_1$

A confidence ellipsoid for $\boldsymbol{\beta}_1 = (\beta_1, \ldots, \beta_r)'$ with level $1-\alpha$ is given by

$$\left\{\boldsymbol{\beta}_1 \,:\, \frac{1}{r}(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1)'\widehat{\mathrm{Cov}(\hat{\boldsymbol{\beta}}_1)}^{-1}(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1) \le F_{r,n-p}(1-\alpha)\right\}.$$

Confidence Interval for μ_0

A confidence interval for $\mu_0 = \mathrm{E}(y_0)$ of a future observation y_0 at location $\boldsymbol{x}_0$ with level $1-\alpha$ is given by

$$\boldsymbol{x}_0'\hat{\boldsymbol{\beta}} \pm t_{n-p}(1-\alpha/2)\hat{\sigma}(\boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0)^{1/2}.$$

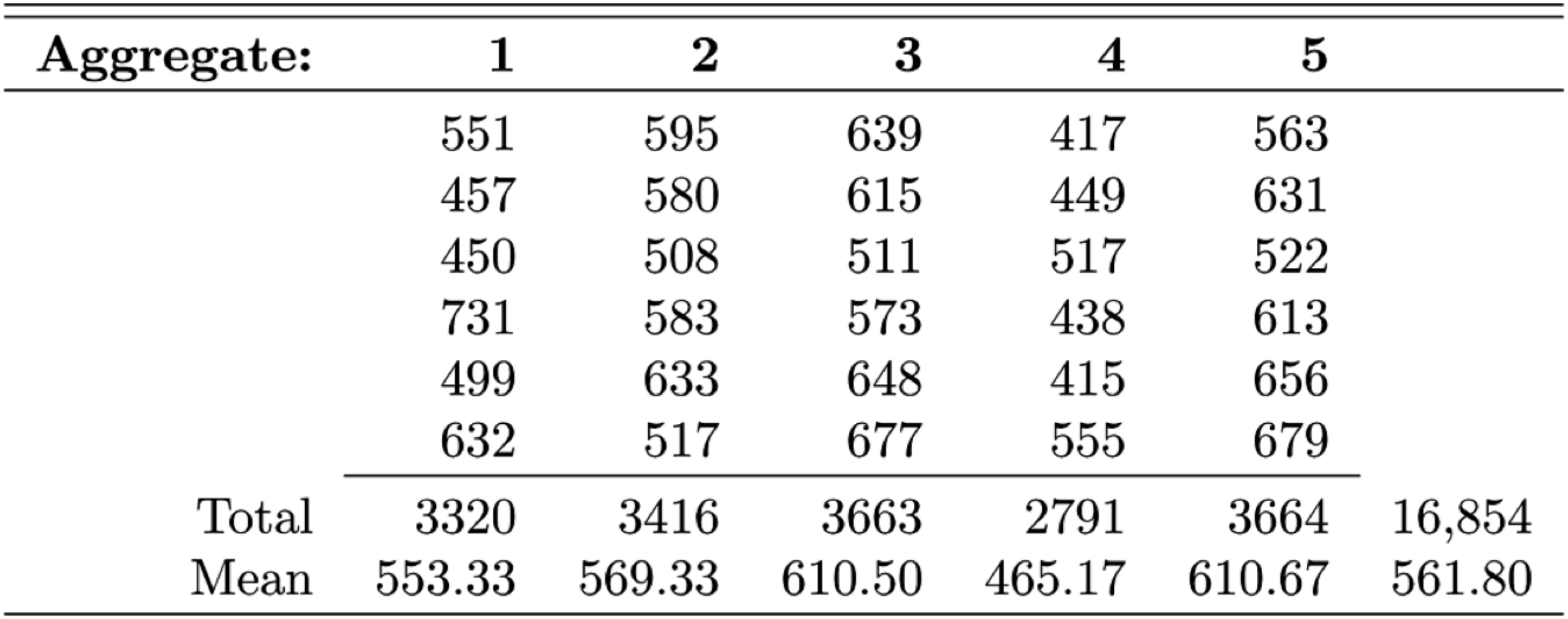
Prediction Interval

A prediction interval for a future observation y_0 at location $\boldsymbol{x}_0$ with level $1-\alpha$ is given by

$$\boldsymbol{x}_0'\hat{\boldsymbol{\beta}} \pm t_{n-p}(1-\alpha/2)\hat{\sigma}(1 + \boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0)^{1/2}.$$

Box from our text book: Fahrmeir et al (2013): Regression. Springer. (p.137)

Concrete aggregates data



Aggregate:	**1**	**2**	**3**	**4**	**5**	
	551	595	639	417	563	
	457	580	615	449	631	
	450	508	511	517	522	
	731	583	573	438	613	
	499	633	648	415	656	
	632	517	677	555	679	
Total	3320	3416	3663	2791	3664	16,854
Mean	553.33	569.33	610.50	465.17	610.67	561.80

Table 13.1 of Walepole, Myers, Myers, Ye: Statistics for Engineers and Scientists – our textbook from the introductory TMA4240/TMA4245 Statistics course.

Today

- Linear hypotheses in regression vs. nested models.
- The universal F-test for linear hypotheses: two formulas.
- The two formulas: one easy to use, one easy for proving F-distribution.
- Special cases of the universal F-test.
- Next time: categorical covariate with effect coding (for interpretation)

Testing linear hypotheses [F: 3.3]

TMA4267 L14
09.03.2017

① Regression model

$$\underset{n\times 1}{Y} = \underset{n\times p}{X}\,\underset{p\times 1}{\beta} + \underset{n\times 1}{\varepsilon}, \quad \varepsilon \sim N_n(0, \sigma^2 I)$$

unrestriced model (A)

$\hat{\beta}, \hat{\sigma}^2$ $\quad \hat{\varepsilon} = Y - \hat{Y} = Y - X\hat{\beta}$

$SSE = \hat{\varepsilon}^T \hat{\varepsilon}$

$$\frac{SSE}{\sigma^2} \sim \chi^2_{n-p}$$

Ex: Happiness

x_1, x_2, x_3, x_4

SSE

Want to test

$H_0: \beta_1 = \beta_2 = 0$ vs

H_1: at least one $\neq 0$

$$H_0: \underset{r\times p}{C}\beta = d$$

$$\begin{array}{ccccc} \beta_0 & \beta_1 & \beta_2 & \beta_3 & \beta_4 \end{array}$$

$$C = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{bmatrix}, \quad d = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$H_1: C\beta \neq d$

$x_3 + x_4$ as covariates →

② Restricted regression model (B)

Here "$C\beta = d$"

The restricted model is a special case of the unrestricted model (nested within)

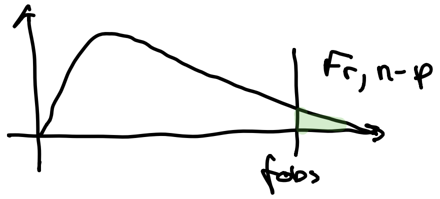
$$SSE_{H_0} = \hat{\varepsilon}_{H_0}^T \hat{\varepsilon}_{H_0}$$

$$\hat{\beta}_{H_0}, \quad \hat{\varepsilon}_{H_0} = Y - X\hat{\beta}_{H_0}$$

(3)

$$F_{obs} = \frac{\frac{1}{r}\left(SSE_{H_0} - SSE\right)}{\frac{1}{n-p}\ SSE} \sim F_{r,n-p}$$

when H_0 is true



understanding: when is $\underbrace{F_{obs}\ \text{large}}$ → when SSE_{H_0} is much larger than SSE

↑ reason to believe that we have "missed" something when fitting restriced model.
Therefore we want to reject H_0 for large Fobs

Now: - Why is $F_{obs} \sim F_{r,n-p}$ under H_0

- is it possible to write Fobs using $X, C, d, \hat{\beta}, \hat{\sigma}^2$?

We start with a new version of fobs:

$$F_{obs} \overset{\circledast}{=} \frac{1}{r}\left(C\hat{\beta} - d\right)^T\left[\hat{\sigma}^2 C\left(X^TX\right)^{-1}C^T\right]^{-1}\left(C\hat{\beta} - d\right)$$

Why is this a $F_{r,n-p}$ - distribution ?

1) $\hat{\beta} \sim N_p(\beta, (X^TX)^{-1}\sigma^2)$

$$Z = \underset{r\times p}{\underset{\nearrow}{C}}\,\underset{p\times 1}{\underset{\uparrow}{\hat{\beta}}} \sim N_r(C\beta, C\,Cov(\hat{\beta})C^T)$$

$$= N_r(\underset{d \text{ when } H_0 \text{ true}}{\underset{\|}{C\beta}}, \sigma^2 C(X^TX)^{-1}C^T)$$

2) $(C\hat{\beta} - d)^T[\sigma^2 C(X^TX)^{-1}C^T]^{-1}(C\hat{\beta}-d)$

$\sim \chi^2_r$ (Part 1)

3) σ^2 is unknown, but $\hat{\sigma}^2 = \frac{SSE}{n-p}$

and we know that $\frac{SSE}{\sigma^2} = \frac{(n-p)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-p}$

(Part 2)

4) And $\hat{\beta}$ and SSE are independent – known from Part 2.

5) So, finally:

$$\frac{\bullet\ \chi^2_r/r}{\bullet\ \chi^2_{n-p}/n-p} \overset{\text{definition}}{\underset{\swarrow}{\sim}} F_{r,n-p}$$

$$F_{obs} = \frac{\frac{1}{r}(C\hat\beta-d)^T[\cancel{\sigma^2}C(X^TX)^{-1}C^T]^{-1}(C\hat\beta-d)}{\frac{1}{\cancel{n-p}}\;\frac{(\cancel{n-p})\hat\sigma^2}{\cancel{\sigma^2}}}$$

$$(cA)^{-1} = \frac{1}{c}A^{-1}$$

$$= \underbrace{\frac{1}{r}(C\hat\beta-d)^T[\hat\sigma^2 C(X^TX)^{-1}C^T]^{-1}(C\hat\beta-d)}$$

Why is F_{obs} = ______ ?

1) For the unrestricted model $\hat\beta = (X^TX)^{-1}X^TY$ found by minimizing $LS(\beta) = (Y-X\beta)^T(Y-X\beta)$.

2) For the restricted model we minimize $LS(\beta)$ subject to "$C\beta = d$", by minimizing

$$LS(\beta) + 2\lambda^T(C\beta - d) = LSR(\beta)$$

$$\hat\beta^R = \hat\beta - (X^TX)^{-1}C^T[C(X^TX)^{-1}C^T](C\hat\beta-d)$$

see F: p 172-173

3) $\Delta SSE = \hat{\varepsilon}_{vto}^{T} \hat{\varepsilon}_{rto} - \hat{\varepsilon}^{T} \hat{\varepsilon}$

$$= (Y - \mathbb{X}\hat{\beta}^{R})^{T}(Y - \mathbb{X}\hat{\beta}^{R}) - (Y - \mathbb{X}\hat{\beta})^{T}(Y - \mathbb{X}\hat{\beta})$$

$$= \dots = \underline{(C\hat{\beta} - d)^{T}\left[C(\mathbb{X}^{T}\mathbb{X})^{-1}C^{T}\right]^{-1}(C\hat{\beta} - d)}$$

$\nearrow$

F: p 173-174

4) $\underline{SSE = \hat{\sigma}^{2}(n-p)}$

5) $$\frac{\frac{1}{r}\,\Delta SSE}{\frac{1}{n-p}\,\underline{SSE}} = \dots =$$

⊛ $$\boxed{\frac{1}{r}\left(C\hat{\beta} - d\right)^{T}\left[\hat{\sigma}^{2}C(\mathbb{X}^{T}\mathbb{X})^{-1}C^{T}\right]^{-1}(C\hat{\beta} - d)}$$

How can we use Fobs for hypothesis testing?

Problem: $Y = X\beta + \varepsilon$, $\varepsilon \sim N_n(0, \sigma^2 I)$

$\underbrace{H_0 : C\beta = d}_{\text{restricted model (B)}}$ vs $\underbrace{H_1 : C\beta \neq d}_{\text{unrestricted model (A)}}$

First: is the regression significant? Then maybe compare full and reduced model (Comp. Ex 3. P1) or test each $\beta_j = 0$?

Solution 1 :

a) Fit unrestricted model and get SSE. (Ex: $X_1 + X_2 + X_3 + X_4$) Happiness

b) Fit restricted model and get SSE_{H_0} (Ex: $X_3 + X_4$)

c) $$F_{obs} = \frac{\frac{1}{r} \Delta SSE}{\frac{SSE}{n-p}}$$

d) Calculate p-value, $P(F_{r,n-p} > f_{obs})$, and reject or not H_0.

Solution 2

a) Fit unrestricted model → $\hat{\beta}, \hat{\sigma}^2$

b) What is C and d

c) $F_{obs} = \circledast$ calculate this

d) Calculate the p-value.

⇒ Hands-on: CompEx 3. Problem 1.

Q: We had $H_0: \beta_j = 0$ vs $H_1: \beta_j \neq 0$ and used a t-test

$$T_j = \frac{\hat{\beta}_j - 0}{\widehat{SD}(\hat{\beta}_j)} \sim t_{n-p}$$

$\widehat{SD}(\hat{\beta}_j) \to \hat{\sigma}\sqrt{c_{jj}}$

and not an F-test. Is it still the same test as using F_{obs}?

A: $H_0: \beta_j = 0$ vs $H_1: \beta_j \neq 0$

where $\underset{1\times p}{C} = [0 \cdots 0 \ \underset{\overset{\uparrow}{j}}{1} \ 0 \cdots 0]$ and $d = 0$, $r = 1$

$p = \#\text{parameters} = \text{intercept} + \underset{k}{(p-1)}$ covariates

$n = \#\text{observation}$

$$C\hat{\beta} = \hat{\beta}_j$$

$$\Big[\underbrace{C\,(X^TX)^{-1}C^T}_{\underbrace{(X^TX)^{-1}_{[j,j]}}_{c_{jj}}}\Big]^{-1} = \frac{1}{c_{jj}}$$

$$\underbrace{\frac{1}{r}}_{1}\big(\underbrace{C\hat{\beta}-d}_{\hat{\beta}_j - 0}\big)^T\Big[\underbrace{\hat{\sigma}^2\, C(X^TX)^{-1}C'}_{\frac{1}{\hat{\sigma}^2 c_{jj}}}\Big]^{-1}\big(\underbrace{C\hat{\beta}-d}_{\hat{\beta}_j - 0}\big)$$

$$= \frac{(\hat{\beta}_j - 0)^2}{\hat{\sigma}^2 c_{jj}} = \underset{\substack{\uparrow \\ (t_{n-p})^2}}{T_j^{\,2}} \in F_{1,n-p}$$

From part 1: $$(T_j)^2 = \left(\frac{Z}{\sqrt{\frac{\chi^2_\nu}{\nu}}}\right)^2 = \frac{Z^2}{\frac{\chi^2}{\nu}} \sim F_{1,\nu} \qquad Z^2 \leftarrow \frac{\chi^2_1}{1}$$

⇒ all on, this is an F-test!

TMA4267 Linear Statistical Models V2017 (L15)

Part 3: Hypothesis testing and analysis of variance
One- and two-way ANOVA [H:8.1.1]

Mette Langaas

Department of Mathematical Sciences, NTNU

To be lectured: March 10, 2017

Today: Analysis of variance (ANOVA) and analysis of covariance (ANCOVA)

- Good news: really nothing new, just linear regression where we have one or more categorical covariates.
- Bad news: a bit technical with respect to coding the covariates in the design matrix.
- Bad or good news: also tell the story of ANOVA without linear regression since that is the classical way to do things - so you will be able to recognize that this is a problem that you can solve.
- Good news: we are taking one step toward the last topic Part 4: Design of experiments.

Rothamsted Experimental Station

- founded in 1843 by John Bennet Lawes on his inherited 16t century estate, Rothamsted Manor,
  - wanted to investigate the impact of inorganic and organic fertilizers on crop yield
  - had founded a fertilizer manufacturing company in 1842
- Lawes appointed the chemist Joseph Henry Gilbert to the directorship of the chemical laboratory
- the two began a series of field experiments to examine the effects of inorganic fertilizers and organic manures on the nutrition and yield of a number of important crops

http://www.stats.uwo.ca/faculty/bellhouse/stat499lecture13.pdf

The Broadbalk Field Trial at Rothamsted

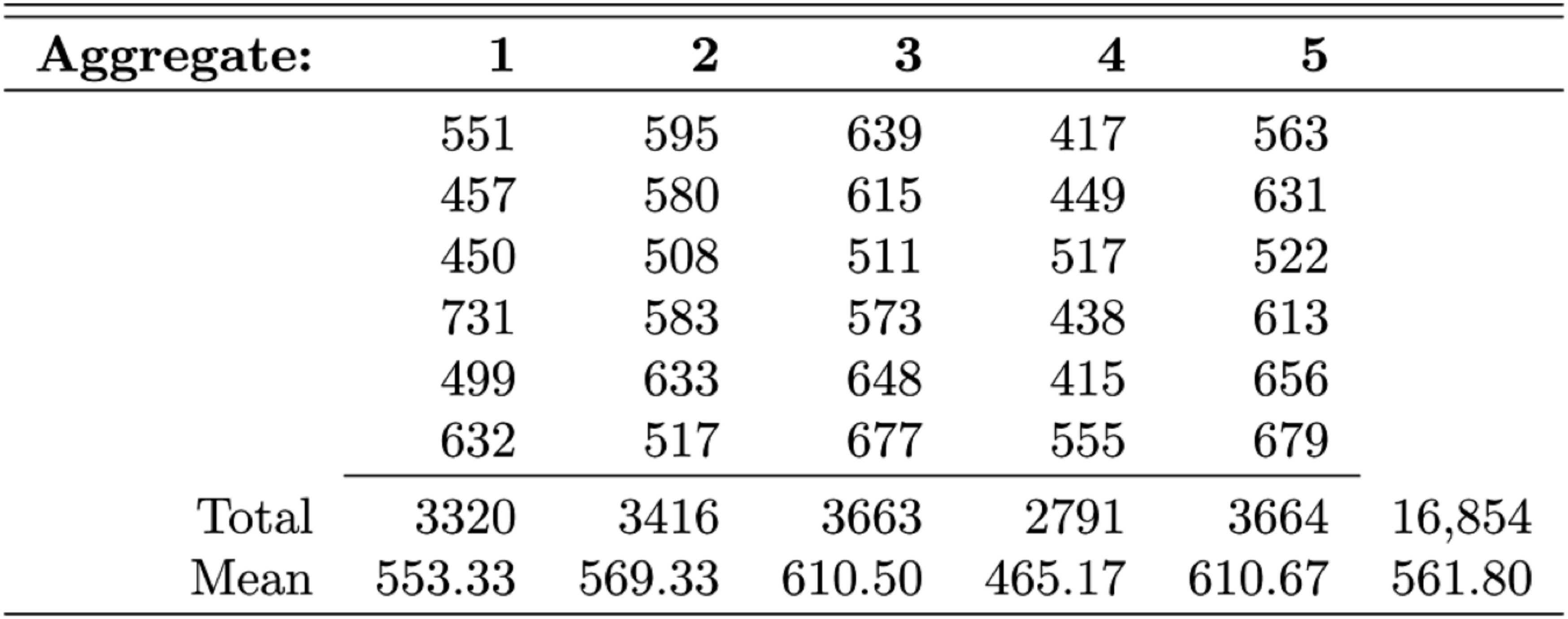
- this was the first field trial started by Lawes and Gilbert
- began in 1843
- purpose was to investigate the relative importance of different plant nutrients (N, P, K, Na, Mg) on grain yield of winter wheat
- weeds were controlled by hand hoeing and fallowing
  - now some herbicides are used
- The experiment continues to this day

http://www.stats.uwo.ca/faculty/bellhouse/stat499lecture13.pdf

Concrete aggregates example

- Aggregates are inert granular materials such as sand, gravel, or crushed stone that, along with water and portland cement, are an essential ingredient in concrete.
- For a good concrete mix, aggregates need to be clean, hard, strong particles free of absorbed chemicals or coatings of clay and other fine materials that could cause the deterioration of concrete.
- We could like to examine 5 different aggregates, and measure the absorption of moisture after 48hrs exposure (to moisture).
- A total of 6 samples are tested for each aggregate.
- Research question: Is there a difference between the aggregates with respect to absorption of moisture?

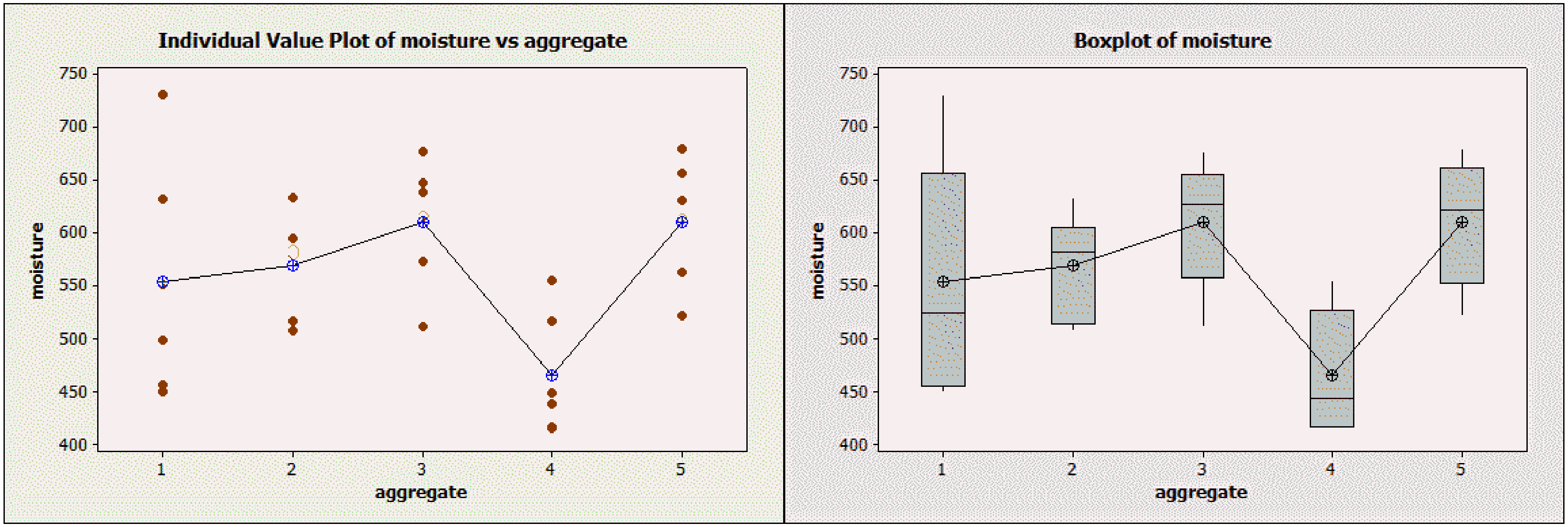
Concrete aggregates data



Aggregate:	**1**	**2**	**3**	**4**	**5**	
	551	595	639	417	563	
	457	580	615	449	631	
	450	508	511	517	522	
	731	583	573	438	613	
	499	633	648	415	656	
	632	517	677	555	679	
Total	3320	3416	3663	2791	3664	16,854
Mean	553.33	569.33	610.50	465.17	610.67	561.80

Table 13.1 of Walepole, Myers, Myers, Ye: Statistics for Engineers and Scientists – our textbook from the introductory TMA4240/TMA4245 Statistics course.

Concrete aggregates example



One-way Analysis of Variance (ANOVA)

Model

$$Y_{ij} = \mu_i + \varepsilon_{ij} \text{ for } i = 1, 2, ..., p \text{ and } j = 1, 2, ..., n_i$$

alternative parameterization

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

The sample sizes for each group, n_i may vary. $\varepsilon_{ij} \sim N(0, \sigma^2)$. Let $n = \sum_{i=1}^{p} n_i$ be the total number of observations.
Aim: look at parameter estimates and test if there is any difference between the groups.
How can that be done using our linear regression model?

Concrete aggregates data

```
# means for each recipe
> means=
    aggregate(ds,by=list(ds$aggregate),FUN=mean)$moisture
> grandmean=mean(ds$moisture)
> grandmean
[1] 561.8
> alphas=means-grandmean
> alphas
[1]  -8.466667   7.533333  48.700000 -96.633333  48.866667
```

Concrete aggregates data

```
# the same with regression
> options(contrasts=c("contr.sum","contr.sum"))
> obj <-lm(moisture~as.factor(aggregate),data=ds)
> summary(obj)
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)            561.800     12.859  43.688  < 2e-16 ***
as.factor(aggregate)1   -8.467     25.719  -0.329 0.744743
as.factor(aggregate)2    7.533     25.719   0.293 0.772005
as.factor(aggregate)3   48.700     25.719   1.894 0.069910 .
as.factor(aggregate)4  -96.633     25.719  -3.757 0.000921 ***
```

Concrete aggregates data

```
#comparing means and regression estimates
>cbind(c(grandmean,alphas),
   c(obj$coefficients,-sum(obj$coefficients[2:5])))
                            [,1]       [,2]
(Intercept)           561.800000 561.800000
as.factor(aggregate)1  -8.466667  -8.466667
as.factor(aggregate)2   7.533333   7.533333
as.factor(aggregate)3  48.700000  48.700000
as.factor(aggregate)4 -96.633333 -96.633333
                       48.866667  48.866667
```

Run R code from course lectures tab for model matrix.

Concrete aggregates data (1)

```
# checking manually with linear hypotheses
r=4
C=cbind(rep(0,r),diag(r))
d=matrix(rep(0,r),ncol=1)
betahat=matrix(obj$coefficients,ncol=1)
sigma2hat=summary(obj)$sigma^2
Fobs=(t(C%*%betahat-d)%*%
solve(C%*%solve(t(X)%*%X)%*%t(C))%*%
(C%*%betahat-d))/(r*sigma2hat)
> Fobs
         [,1]
[1,] 4.301536
> 1-pf(Fobs,r,n-r-1)
            [,1]
[1,] 0.008751641
```

Concrete aggregates data (2)

```
> fitA=obj
> fitB=lm(moisture~1,data=aggregates)
> anova(fitA,fitB)
Analysis of Variance Table

Model 1: moisture ~ as.factor(aggregate)
Model 2: moisture ~ 1
  Res.Df    RSS Df Sum of Sq      F   Pr(>F)
1     25 124020
2     29 209377 -4    -85356 4.3015 0.008752 **
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’
```

Concrete aggregates data (3)

```
# performing ANOVA using method anova -
> anova(obj)
Analysis of Variance Table

Response: moisture
                     Df Sum Sq Mean Sq F value   Pr(>F)
as.factor(aggregate)  4  85356 21339.1  4.3015 0.008752 **
Residuals            25 124020  4960.8
```

One factor: unequal sample sizes

Classical formulation with ANOVA decomposition

$$
\begin{aligned}
Y_{ij} - Y_{..} &= (Y_{ij} - Y_{i.}) + (Y_{i.} - Y_{..}) \\
\sum_{i=1}^{p}\sum_{j=1}^{n_i}(Y_{ij} - Y_{..})^2 &= \sum_{i=1}^{p}\sum_{j=1}^{n_i}(Y_{ij} - Y_{i.})^2 + \sum_{i=1}^{p}\sum_{j=1}^{n_i}(Y_{i.} - Y_{..})^2 \\
\sum_{i=1}^{p}\sum_{j=1}^{n_i}(Y_{ij} - Y_{..})^2 &= \sum_{i=1}^{p}\sum_{j=1}^{n_i}(Y_{ij} - Y_{i.})^2 + \sum_{i=1}^{p} n_i(Y_{i.} - Y_{..})^2 \\
\mathrm{SST} &= \mathrm{SSE} + \mathrm{SSA}
\end{aligned}
$$

One factor: unequal sample sizes

ANOVA decomposition: what happened to the cross-term?

$$2\sum_{i=1}^{p}\sum_{j=1}^{n_i}(Y_{ij} - Y_{i.})(Y_{i.} - Y_{..}) = 2\sum_{i=1}^{p}(Y_{i.} - Y_{..})\sum_{j=1}^{n_i}(Y_{ij} - Y_{i.}) = 0$$

$$\sum_{j=1}^{n_i}(Y_{ij} - Y_{i.}) = \sum_{j=1}^{n_i} Y_{ij} - \sum_{j=1}^{n_i} Y_{i.} = n_i Y_{i.} - n_i Y_{i.} = 0$$

One factor: unequal sample sizes

$H_0 : \mu_1 = \mu_2 = \cdots = \mu_p = 0$ vs. $H_1 :$ At least one pair of μ_i different

is then tested based on

$$F = \frac{\frac{\text{SSA}}{p-1}}{\frac{\text{SSE}}{n-p}}$$

Where H_0 is rejected if $f_{\text{obs}} > f_\alpha, (p-1), (n-p)$.

Machine example

- Response: time (s) spent to assemble a product.
- Factor: this is done by four different machines; M_1, M_2, M_3, M_4.
- Question: Do the machines perform at the same mean rate of speed?

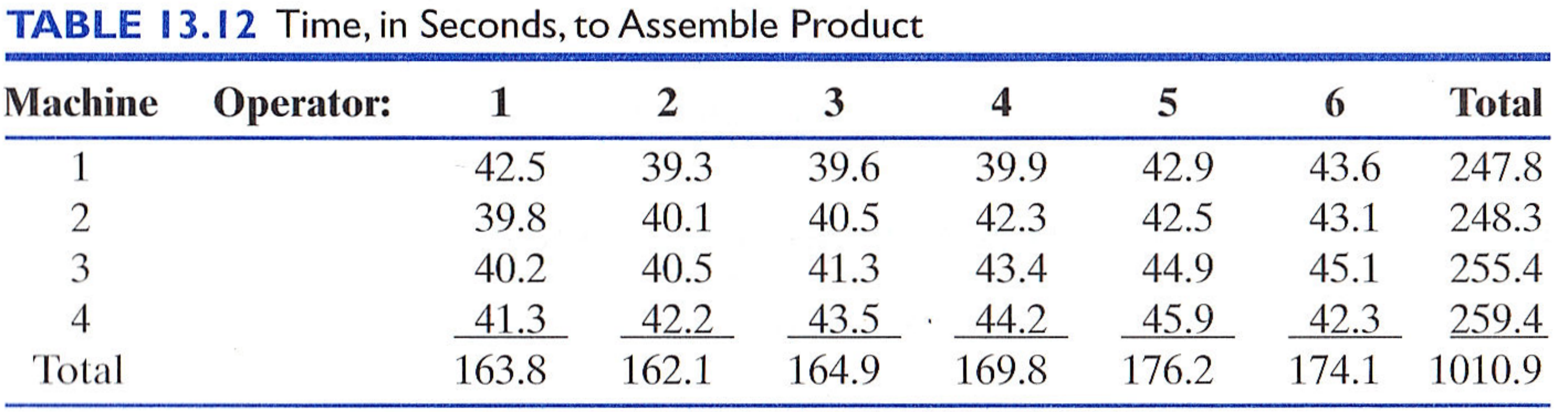
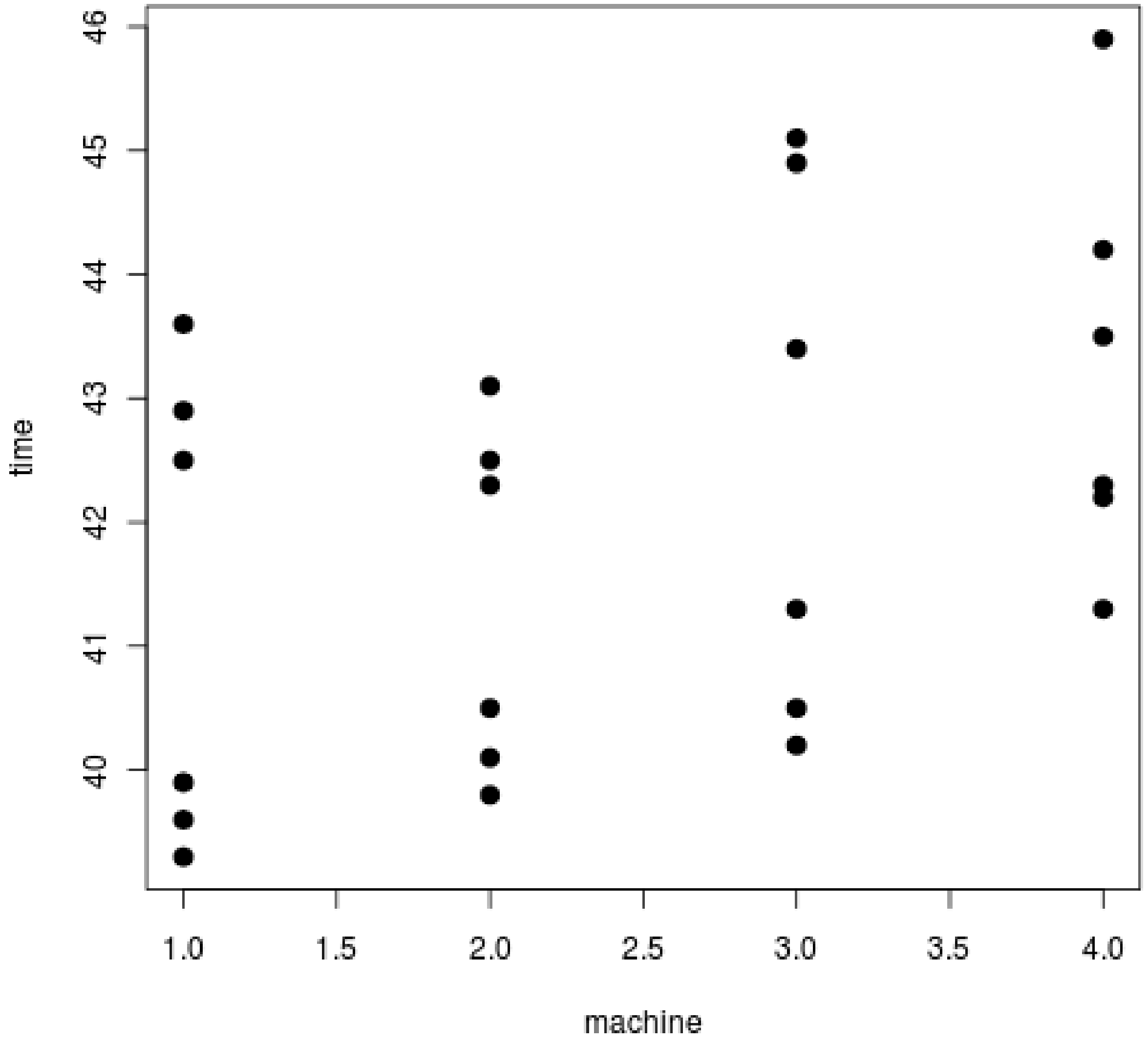


TABLE 13.12 Time, in Seconds, to Assemble Product

Machine	Operator:	1	2	3	4	5	6	Total
1		42.5	39.3	39.6	39.9	42.9	43.6	247.8
2		39.8	40.1	40.5	42.3	42.5	43.1	248.3
3		40.2	40.5	41.3	43.4	44.9	45.1	255.4
4		41.3	42.2	43.5	44.2	45.9	42.3	259.4
Total		163.8	162.1	164.9	169.8	176.2	174.1	1010.9

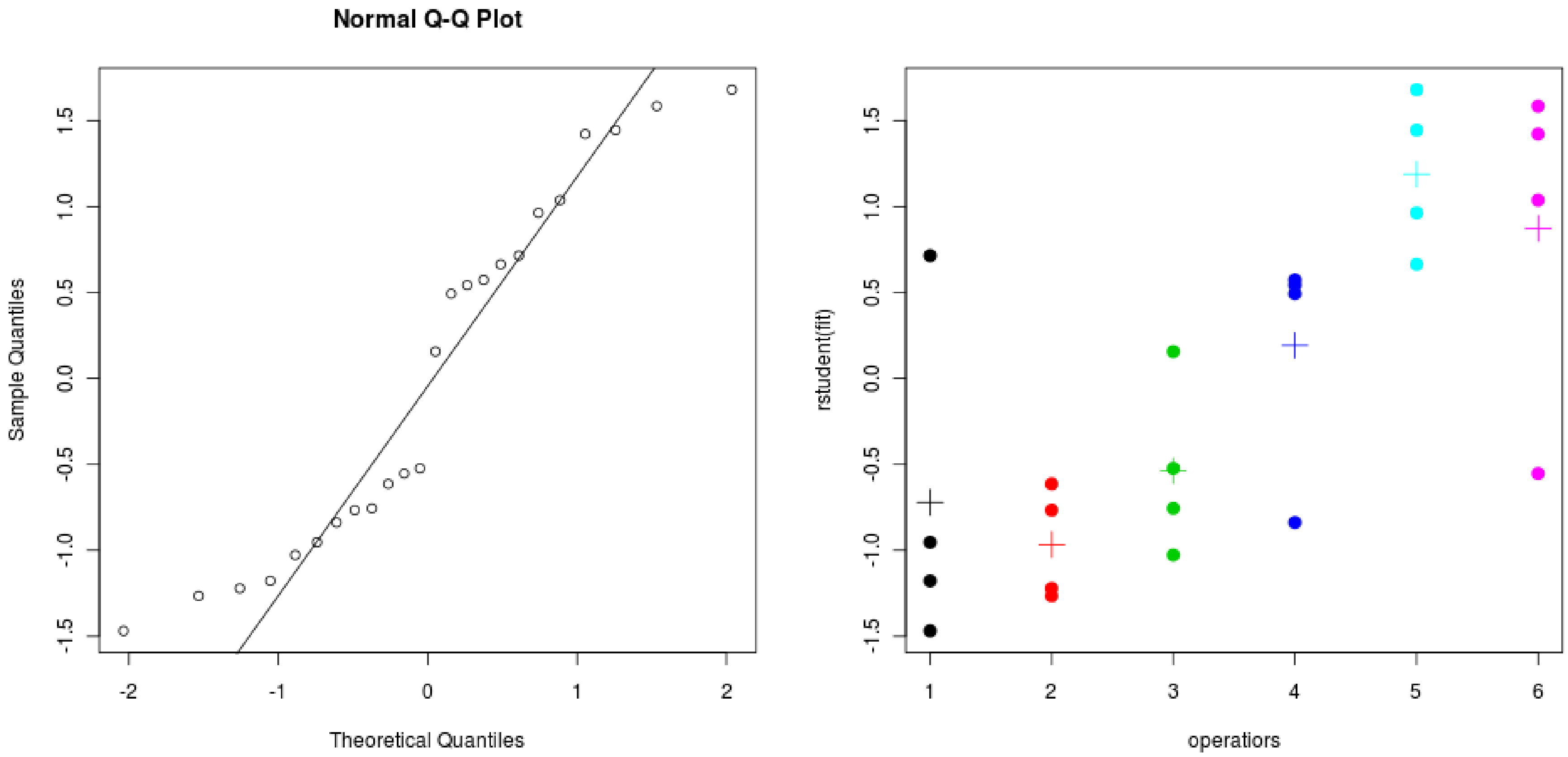
Data from Walepole, Myers, Myers, Ye: "Statistics for Engineers and Scientists", Example 13.6= our TMA4245/40 textbook.


time
40
41
42
43
44
45
46
machine
1.0
1.5
2.0
2.5
3.0
3.5
4.0

One factor ANOVA

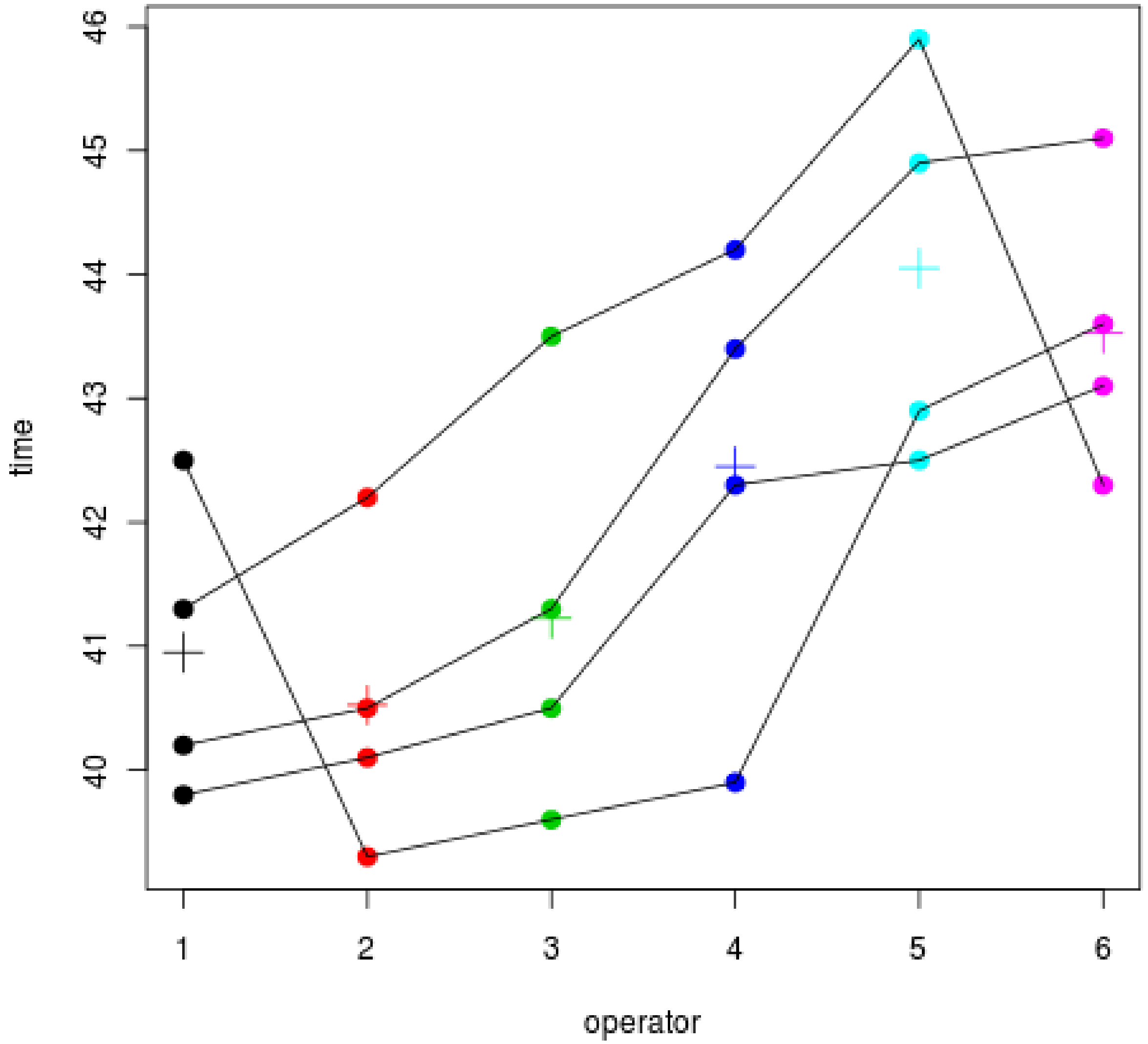
```
> options(contrasts=c("contr.sum","contr.sum"))
> fit <- lm(time~as.factor(machine),data=dsmat)
> summary(fit)
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)    
(Intercept)          42.1208     0.3706 113.647   <2e-16 ***
as.factor(machine)1  -0.8208     0.6419  -1.279    0.216    
as.factor(machine)2  -0.7375     0.6419  -1.149    0.264    
as.factor(machine)3   0.4458     0.6419   0.695    0.495    

Residual standard error: 1.816 on 20 degrees of freedom
Multiple R-squared:  0.1945,Adjusted R-squared:  0.07372
F-statistic:  1.61 on 3 and 20 DF,  p-value: 0.2186

> anova(fit)
Response: time
                   Df Sum Sq Mean Sq F value Pr(>F)
as.factor(machine)  3 15.925  5.3082  1.6101 0.2186
Residuals          20 65.935  3.2968
```

Residuals


Normal Q-Q Plot
Sample Quantiles
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
-2
-1
0
1
2
Theoretical Quantiles
rstudent(fit)
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
1
2
3
4
5
6
operators

Machine example: operators

- The 6 repeated measurements for each machine was in fact made by 6 different operators.
- The operation of the machines requires physical dexterity and differences among the operators in the speed with which they operate the machines is anticipated.
- All of the 6 operators have operated all the 4 machines, and the machines were assigned in random order to the operators= *randomized complete block design*.
- By including a blocking factor called Operator, we will reduce the variation in the experiment that is du to random error. Thus, we reduce variation due to *anticipated factors*.
- By randomizing the order the machines were assigned to the operators we aim to reduce the variation due to *unanticipated factors*.


time
46
45
44
43
42
41
40
1
2
3
4
5
6
operator

Model and Sums of squares

Model

$$Y_{ij} = \mu + \alpha_i + \gamma_j + \varepsilon_{ij} \text{ for } i = 1, 2, ..., r \text{ and } j = 1, 2, ..., s$$

Sums of Squares Identity

$$\begin{aligned}
Y_{ij} &= Y_{..} + (Y_{i.} - Y_{..}) + (Y_{.j} - Y_{..}) + (Y_{ij} - Y_{i.} - Y_{.j} + Y_{..}) \\
\sum_{i=1}^{r}\sum_{j=1}^{s}(Y_{ij} - Y_{..})^2 &= s\sum_{i=1}^{r}(Y_{i.} - Y_{..})^2 + r\sum_{j=1}^{s}(Y_{.j} - Y_{..})^2 \\
&\quad + \sum_{i=1}^{r}\sum_{j=1}^{s}(Y_{ij} - Y_{i.} - Y_{.j} + Y_{..})^2 \\
\text{SST} &= \text{SSA} + \text{SSB} + \text{SSE} \\
r \cdot s - 1 &= (r-1) + (s-1) + (r-1)(s-1)
\end{aligned}$$

Effect of factor A:

$H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_r = 0$ vs. $H_1 :$ At least one α_i different from 0

is then tested based on

$$F_1 = \frac{\frac{\text{SSA}}{r-1}}{\frac{\text{SSE}}{(r-1)(s-1)}}$$

Where H_0 is rejected if $f_1 > f_{\alpha}, (r-1), (r-1)(s-1)$.

Block effect present?

$H_0 : \gamma_1 = \gamma_2 = \cdots = \gamma_s = 0$ vs. $H_1 :$ At least one γ_j different from 0

is then tested based on

$$F_2 = \frac{\frac{\text{SSB}}{s-1}}{\frac{\text{SSE}}{(r-1)(s-1)}}$$

Where H_0 is rejected if $f_2 > f_{\alpha}, (s-1), (r-1)(s-1)$.

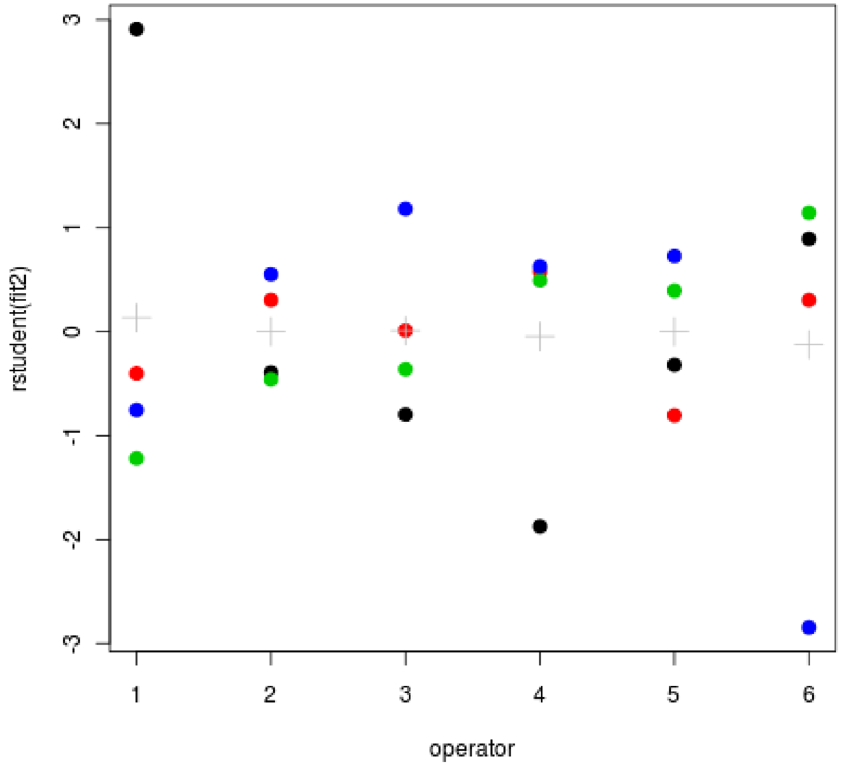
RCBD ANOVA

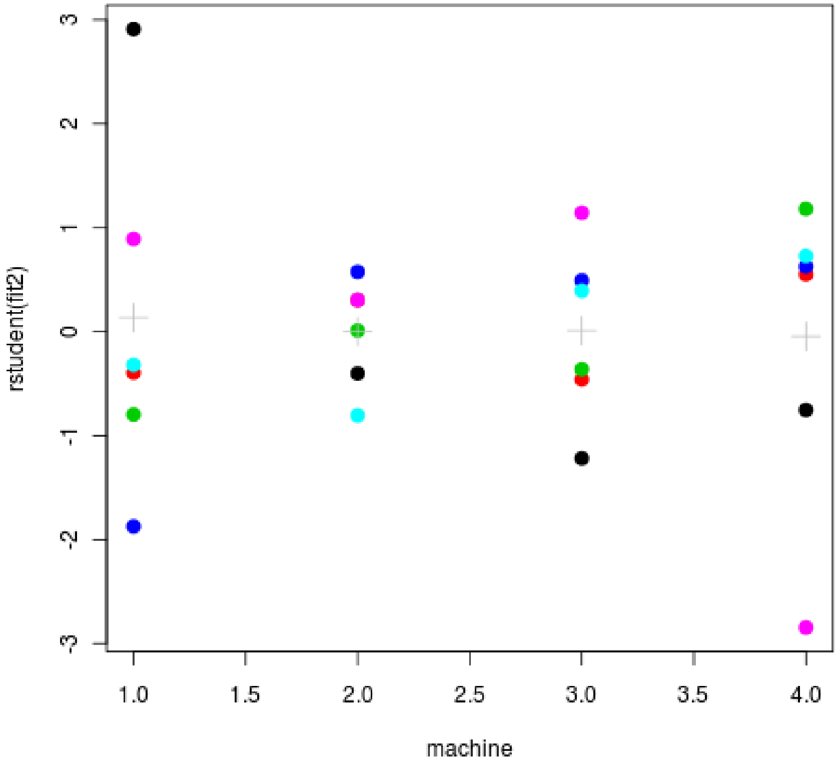
```
> fit2 <- lm(time~as.factor(machine)+as.factor(operator),
data=dsmat)

> anova(fit2)
                    Df Sum Sq Mean Sq F value   Pr(>F)
as.factor(machine)   3 15.925  5.3082  3.3388 0.047904 *
as.factor(operator)  5 42.087  8.4174  5.2944 0.005328 **
Residuals           15 23.848  1.5899
```

Effect of operator with linear hypotheses

```
fit2 <- lm(time~as.factor(machine)+as.factor(operator),
data=dsmat)
r=5
C=cbind(rep(0,5),rep(0,5),rep(0,5),rep(0,5),diag(5))
d=matrix(rep(0,r),ncol=1)
betahat=matrix(fit2$coefficients,ncol=1)
X=model.matrix(fit2)
sigma2hat=summary(fit2)$sigma^2
Fobs=(t(C%*%betahat-d)%*%solve(C%*%solve(t(X)%*%X)%*%t(C))
%*%(C%*%betahat-d))/(r*sigma2hat)
> Fobs
         [,1]
[1,] 5.294435
> 1-pf(Fobs,r,n-dim(C)[2])
            [,1]
[1,] 0.005327541
```

Residuals


rstudent(fit2)
3
2
1
0
-1
-2
-3
1
2
3
4
5
6
operator


rstudent(fit2)
3
2
1
0
-1
-2
-3
1.0
1.5
2.0
2.5
3.0
3.5
4.0
machine

A second look at the RCBD: additive effects

Previously, randomized complete block design (RCBD) with the machine example:

$$Y_{ij} = \mu + \alpha_i + \gamma_j + \varepsilon_{ij}$$

where $\sum_{i=1}^{r} \alpha_i = 0$ and $\sum_{j=0}^{s} \gamma_j = 0$.
This is called *additive effects of treatment and blocks*.

- This means that if we compare two operators there is a constant difference in time to assemble the product,
- or, if we compare machines, these are ranked in the same order of (wrt time) for each operator.

Estimates

$$\hat{\mu} = 42.1208$$

$$\hat{\alpha}_1 = -0.8208$$

$$\hat{\alpha}_2 = -0.7375$$

$$\hat{\alpha}_3 = 0.4458$$

$$\hat{\alpha}_4 = 1.1125$$

$$\hat{\gamma}_1 = -1.1708$$

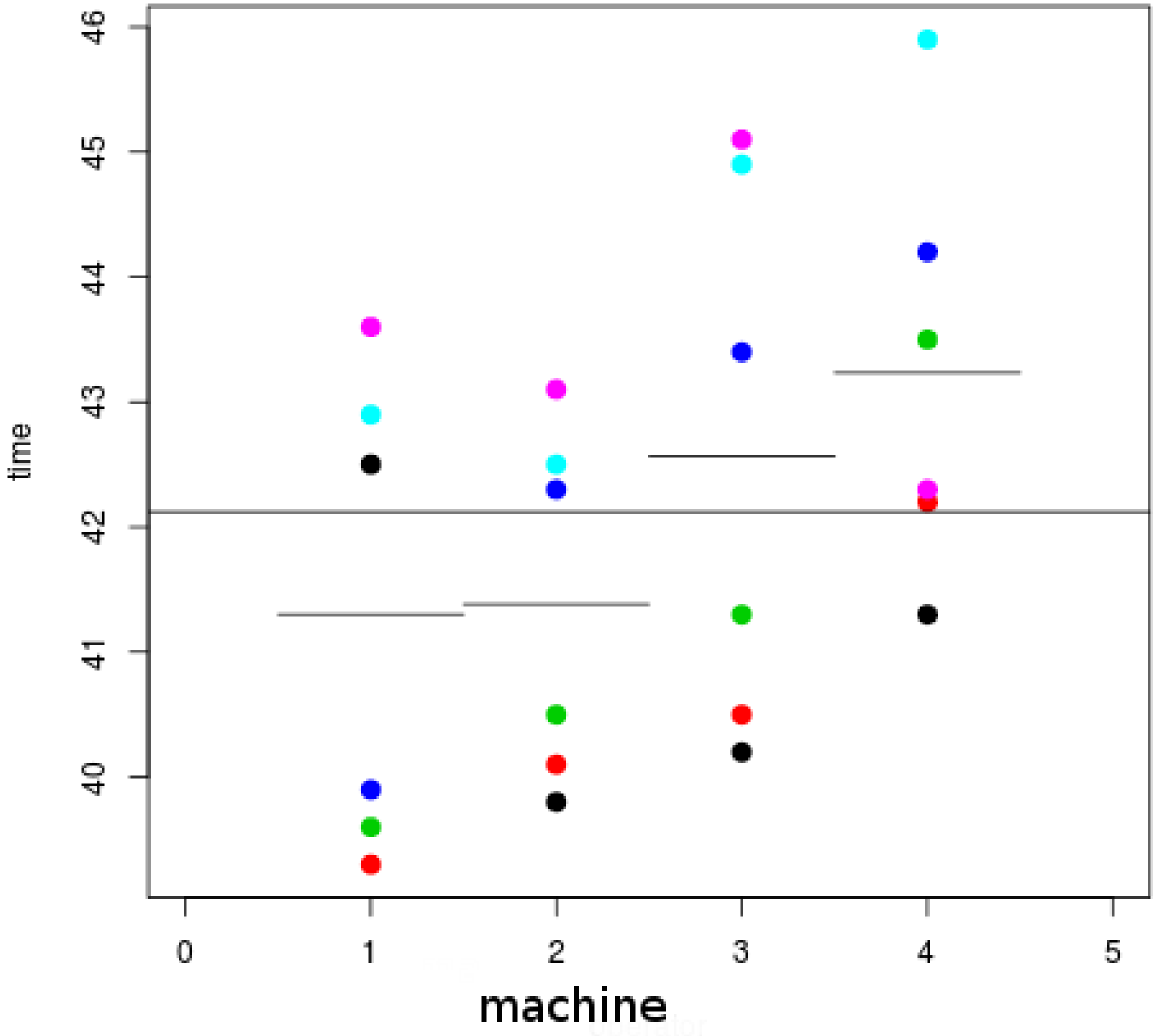
$$\hat{\gamma}_2 = -1.5958$$

$$\hat{\gamma}_3 = -0.8958$$

$$\hat{\gamma}_4 = 0.3292$$

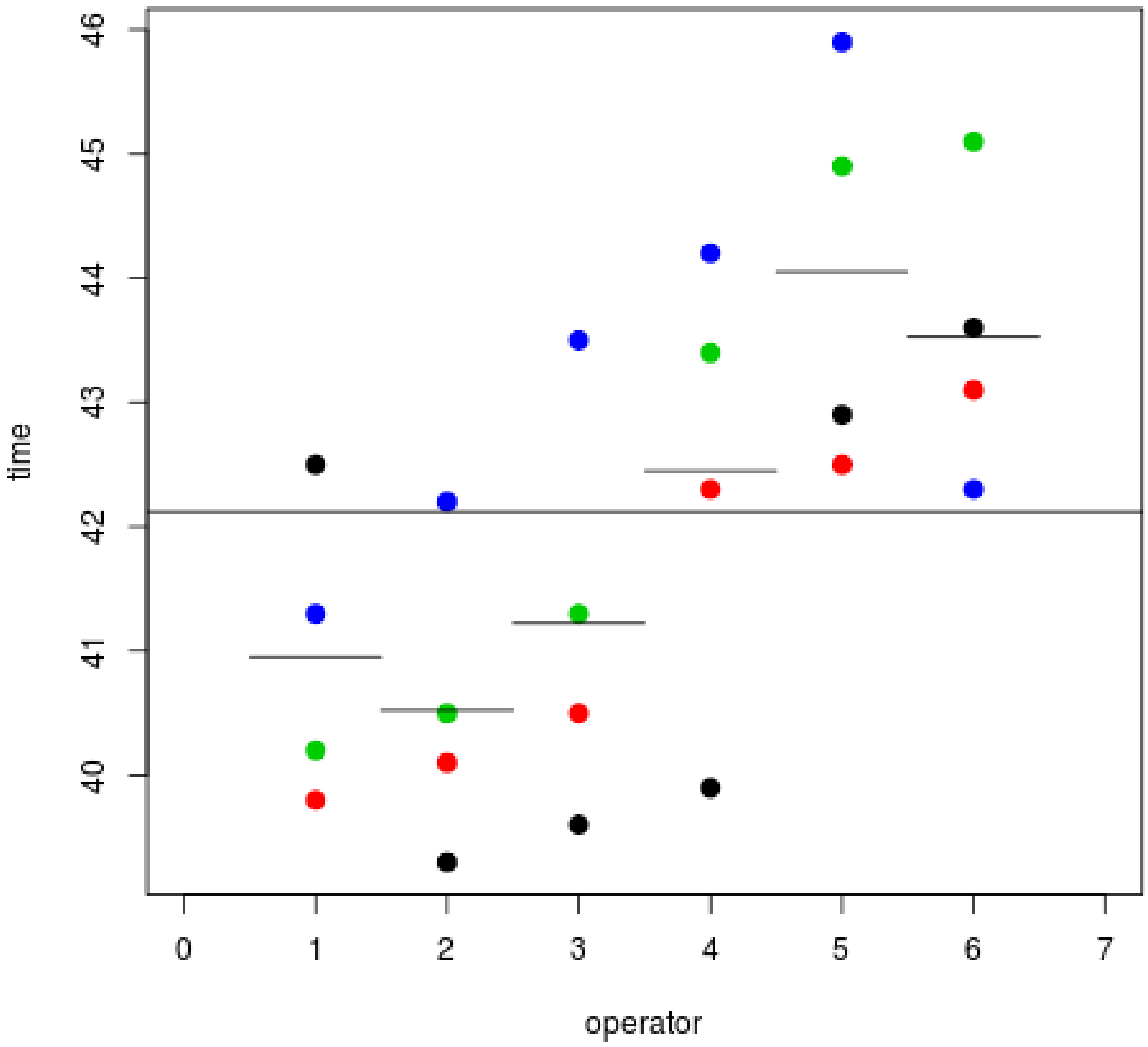
$$\hat{\gamma}_5 = 1.9292$$

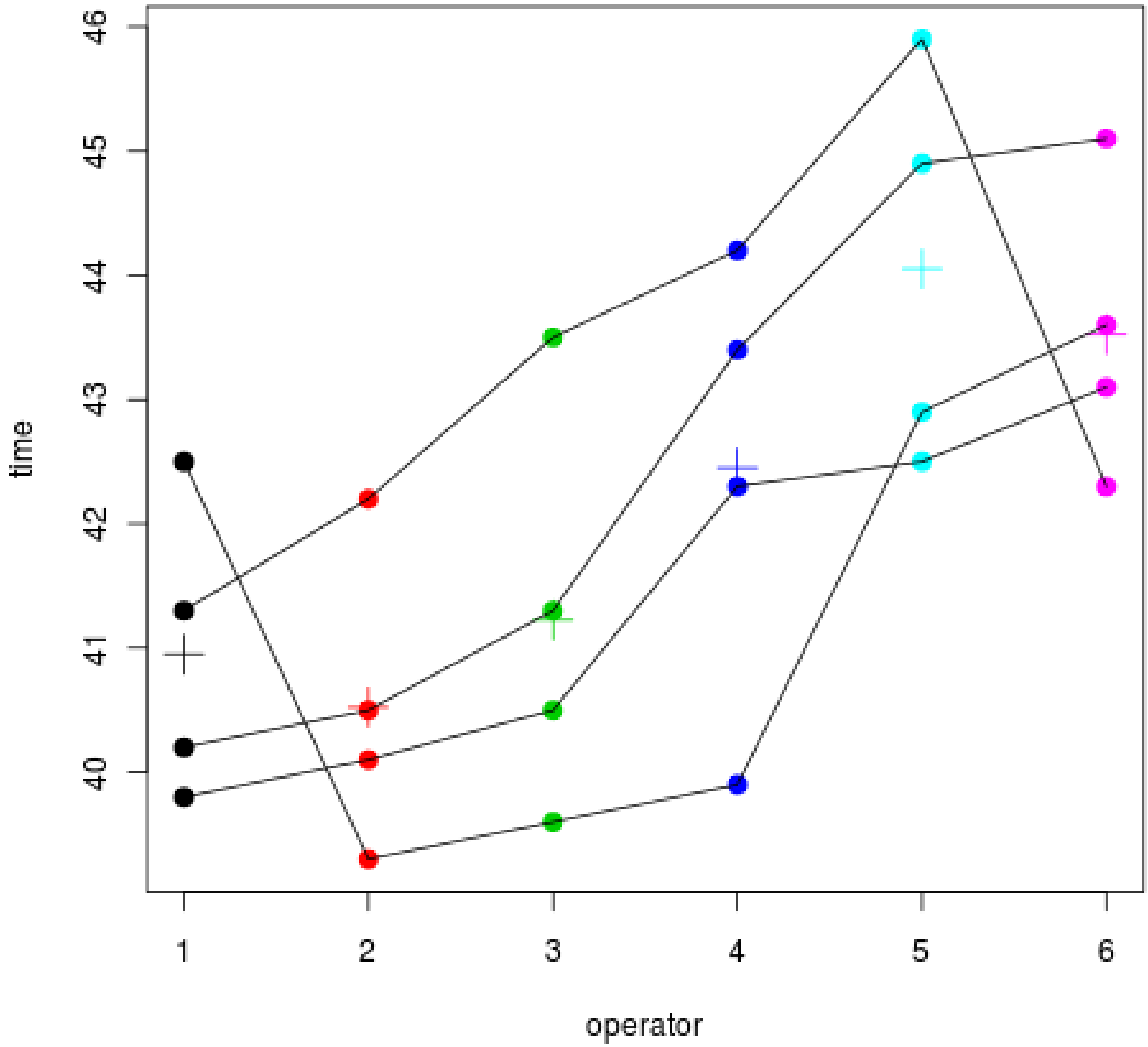
$$\hat{\gamma}_6 = 1.404167$$



Estimates

$$\hat{\mu} = 42.1208$$
$$\hat{\alpha}_1 = -0.8208$$
$$\hat{\alpha}_2 = -0.7375$$
$$\hat{\alpha}_3 = 0.4458$$
$$\hat{\alpha}_4 = 1.1125$$
$$\hat{\gamma}_1 = -1.1708$$
$$\hat{\gamma}_2 = -1.5958$$
$$\hat{\gamma}_3 = -0.8958$$
$$\hat{\gamma}_4 = 0.3292$$
$$\hat{\gamma}_5 = 1.9292$$
$$\hat{\gamma}_6 = 1.404167$$




time
40
41
42
43
44
45
46
operator
1
2
3
4
5
6

Interaction effect?

But, it could be interactions present. What if one of the operators really could not manage one of the machines?
Model with interaction between treatment and block:

$$Y_{ij} = \mu + \alpha_i + \gamma_j + (\alpha\gamma)_{ij} + \varepsilon_{ij}$$

where $\sum_{i=1}^{r}(\alpha\gamma)_{ij} = \sum_{j=1}^{s}(\alpha\gamma)_{ij} = 0$ (for all i and j) in addition to $\sum_{i=1}^{r}\alpha_i = 0$ and $\sum_{j=1}^{s}\gamma_j = 0$.
But, since we only have one observation for each combination of i and j, we can not separate $(\alpha\gamma)_{ij}$ and ε_{ij}.

Interaction effect?

$$SSE = \sum_{i=1}^{r} \sum_{j=1}^{s} (Y_{ij} - Y_{\cdot i} - Y_{j\cdot} + Y_{\cdot\cdot})^2$$

$$E(\frac{SSE}{(r-1)(s-1)}) = \sigma^2 + \frac{\sum_{i=1}^{r} \sum_{j=1}^{s} (\alpha\gamma)_{ij}^2}{(s-1)(r-1)}$$

A large value of SSE will either mean that we have an interaction term present, or that σ^2 is large. We can not assess interaction in a RCBD. We need more than one observation for each observation to distinguish between $(\alpha\gamma)_{ij}$ and ε_{ij}.

Age and memory

- Why do older people often seem not to remember things as well as younger people? Do they not pay attention? Do they just not process the material as thoroughly?
- One theory regarding memory is that verbal material is remembered as a function of the degree to which is was processed when it was initially presented.
- Eysenck (1974) randomly assigned 50 younger subjects and 50 older (between 55 and 65 years old) to one of five learning groups.
- After the subjects had gone through a list of 27 items three times they were asked to write down all the words they could remember.

Eysenck study of recall of older and younger subjects under conditions of differential processing, Eysenck (1974) and presented in Howell (1999).

The Age and Memory data set

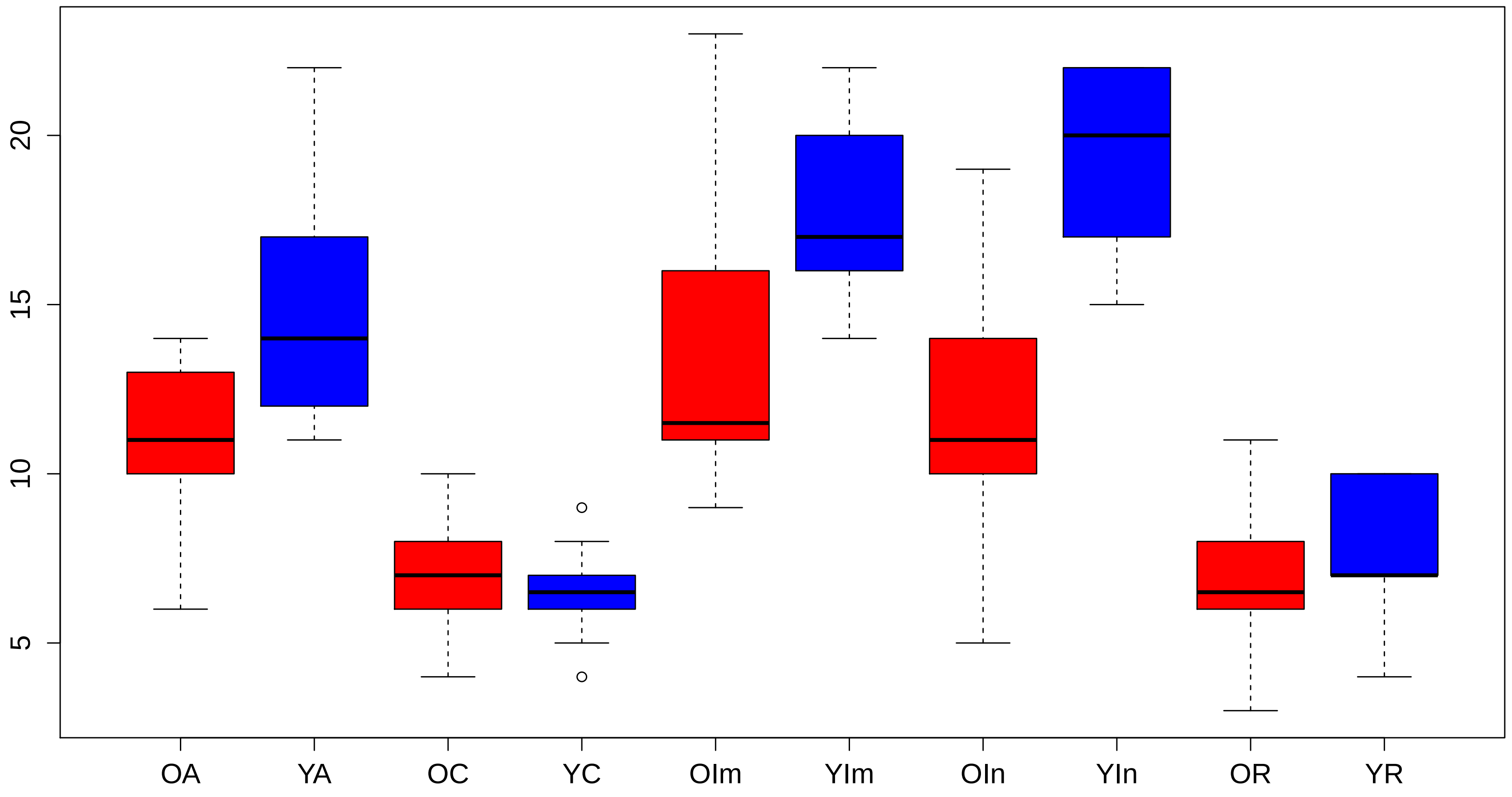
- Number of words recalled: After the subjects had gone through the list of 27 items three times they were asked to write down all the words they could remember.
- Age: Younger (18-30) and Older (55-65).

The Age and Memory data set: Process

- The Counting group was asked to read through a list of words and count the number of letters in each word. This involved the lowest level of processing.
- The Rhyming group was asked to read each word and think of a word that rhymed with it.
- The Adjective group was asked to give an adjective that could reasonably be used to modify each word in the list.
- The Imagery group was instructed to form vivid images of each word, and this was assumed to require the deepest level of processing.

  None of these four groups was told they would later be asked to recall the items.
- Finally, the Intentional group was asked to memorize the words for later recall.

Data taken from: http://www.statsci.org/data/general/eysenck.html



Y=younger (blue), O=older (red), A=adjective, C=counting, Im=Imagery, In=intentional, R=rythming.

Eysenck ANOVA

```
> res <- lm(Words˜Age*Process)
> summary(res)
Call:
lm(formula = Words ˜ Age * Process)

Residuals:
   Min     1Q Median     3Q    Max
  -7.0   -1.6   -0.5    2.0    9.6

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    11.6100     0.2833  40.982  < 2e-16 ***
Age1           -1.5500     0.2833  -5.471 3.98e-07 ***
Process1        1.2900     0.5666   2.277 0.025170 *
Process2       -4.8600     0.5666  -8.578 2.60e-13 ***
Process3        3.8900     0.5666   6.866 8.24e-10 ***
Process4        4.0400     0.5666   7.130 2.43e-10 ***
Age1:Process1  -0.3500     0.5666  -0.618 0.538312
Age1:Process2   1.8000     0.5666   3.177 0.002040 **
Age1:Process3  -0.5500     0.5666  -0.971 0.334288
Age1:Process4  -2.1000     0.5666  -3.706 0.000363 ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Residual standard error: 2.833 on 90 degrees of freedom
Multiple R-squared:  0.7293,Adjusted R-squared:  0.7022
F-statistic: 26.93 on 9 and 90 DF,  p-value: < 2.2e-16
```

Eysenck model matrix

```
> X=model.matrix(res)
> X[c(1,11,21,31,41,51,61,71,81,91),]
   (Intercept) Age1 Process1 Process2 Process3 Process4 Age1:Process1
1            1   -1        0        1        0        0             0
11           1   -1       -1       -1       -1       -1             1
21           1   -1        1        0        0        0            -1
31           1   -1        0        0        1        0             0
41           1   -1        0        0        0        1             0
51           1    1        0        1        0        0             0
61           1    1       -1       -1       -1       -1            -1
71           1    1        1        0        0        0             1
81           1    1        0        0        1        0             0
91           1    1        0        0        0        1             0
   Age1:Process2 Age1:Process3 Age1:Process4
1             -1             0             0
11             1             1             1
21             0             0             0
31             0            -1             0
41             0             0            -1
51             1             0             0
61            -1            -1            -1
71             0             0             0
81             0             1             0
91             0             0             1
```

Model and Sums of Squares

Model:

$$Y_{ijk} = \mu + \alpha_i + \gamma_j + (\alpha\gamma)_{ij} + \varepsilon_{ijk}$$

$$\text{for } i = 1, 2, ..., r \text{ and } j = 1, 2, ..., s \text{ and } k = 1, ..., m$$

$$\varepsilon_{ijk} \sim N(0, \sigma^2)$$

Two-way ANOVA questions

There are three main questions that we might ask in two-way ANOVA:

- Does the response variable depend on Factor A?
- Does the response variable depend on Factor B?
- Does the response variable depend on Factor A differently for different values of Factor B, and vice versa?

All of these questions can be answered using hypothesis tests, first we test the interaction.

Effect of interaction AB

$$H_0^A :(\alpha\gamma)_{11} = (\alpha\gamma)_{12} = \cdots = (\alpha\gamma)_{rs} = 0 \text{ vs.}$$
$$H_1 : \text{At least one } (\alpha\gamma)_{ij} \text{ different from } 0$$

is then tested based on

$$F_3 = \frac{\frac{\mathrm{SS(AB)}}{(r-1)(s-1)}}{\frac{\mathrm{SSE}}{rs(m-1)}}$$

Where H_0 is rejected if $f_3 > f_{\alpha}, (r-1)(s-1), rs(m-1)$.

What do we do after testing for interaction?

- If the interaction is significant (we reject H_0^{AB}).
  - Then it is not recommended to test for main effects (that is, the marginal contributions of the two factors A and B separately). This is since the interpretation of the marginal “main effect” is unclear in the presence of interaction. How can we “separate out” the effect of A from the interaction?
  - Instead, it is usually preferable to examine contrasts in the treatment combinations.
- If the interaction is not found to be significant (do not reject H_0^{AB}).
  - We are then interested in the main effects. These can now be tested within the complete model.

Effect of factor A:

$H_0^A : \alpha_1 = \alpha_2 = \cdots = \alpha_r = 0$ vs. $H_1 :$ At least one α_i different from 0

is then tested based on

$$F_1 = \frac{\frac{\mathsf{SSA}}{r-1}}{\frac{\mathsf{SSE}}{rs(m-1)}}$$

Where H_0^A is rejected if $f_1 > f_\alpha, (r-1), rs(m-1)$.

Effect of factor B:

$H_0^B : \gamma_1 = \gamma_2 = \cdots = \gamma_s = 0$ vs. $H_1 :$ At least one γ_i different from 0

is then tested based on

$$F_2 = \frac{\frac{\mathsf{SSB}}{s-1}}{\frac{\mathsf{SSE}}{rs(m-1)}}$$

Where H_0^B is rejected if $f_2 > f_\alpha, (s-1), sn(m-1)$.

Eysenck ANOVA

```
> res <- lm(Words~Age*Process)
> summary(res)
Call:
lm(formula = Words ~ Age * Process)

Residuals:
   Min     1Q Median     3Q    Max
  -7.0   -1.6   -0.5    2.0    9.6

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    11.6100     0.2833  40.982  < 2e-16 ***
Age1           -1.5500     0.2833  -5.471 3.98e-07 ***
Process1        1.2900     0.5666   2.277 0.025170 *
Process2       -4.8600     0.5666  -8.578 2.60e-13 ***
Process3        3.8900     0.5666   6.866 8.24e-10 ***
Process4        4.0400     0.5666   7.130 2.43e-10 ***
Age1:Process1  -0.3500     0.5666  -0.618 0.538312
Age1:Process2   1.8000     0.5666   3.177 0.002040 **
Age1:Process3  -0.5500     0.5666  -0.971 0.334288
Age1:Process4  -2.1000     0.5666  -3.706 0.000363 ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Residual standard error: 2.833 on 90 degrees of freedom
Multiple R-squared:  0.7293,Adjusted R-squared:  0.7022
F-statistic: 26.93 on 9 and 90 DF,  p-value: < 2.2e-16
```

Eysenck ANOVA

```
> res <- lm(Words~Age*Process)
> anova(res)
Analysis of Variance Table

Response: Words
            Df  Sum Sq Mean Sq F value    Pr(>F)    
Age          1  240.25  240.25 29.9356 3.981e-07 ***
Process      4 1514.94  378.74 47.1911 < 2.2e-16 ***
Age:Process  4  190.30   47.58  5.9279 0.0002793 ***
Residuals   90  722.30    8.03                      
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

Next: maybe want to compare different combinations of age and process? Then, easiest to just combine the two factors into a new joint factor and skip the intercept.

Summing up

Topic today: the one-way and two-way ANOVA models.

- Classical formulation has focus on comparing sums of squares.
- We don't have to prove the classical results because we instead fit the ANOVA model using linear regression with effect coding of covariates.
- It is important to plot results and to understand when an interaction term is needed.
- To test ANOVA hypotheses we use linear hypotheses in the regression – where we automatically have theoretical results for F-distributions.
- We will meet linear regression models with k factors with two levels each in Part 4: Design of Experiments (DOE).

Analysis of variance (ANOVA)

TMA4267 L15
10.03.2017

Ex: Concrete recipes : 5 recipes to produce concrete, tested on 6 samples each. Measured moisture. Q: is there a difference between the recipes wrt moisture.

$$\Updownarrow$$

is the variability between the recipes large compared variability within.

1) One-way ANOVA model

$$Y_{ij} = \mu_i + \varepsilon_{ij} \qquad \begin{matrix} i = 1,\dots,p \\ j = 1,\dots,n_i \end{matrix} \qquad \begin{matrix} \text{Ex: } p=5 \\ n_i = 6 \ \forall i \end{matrix}$$

$\varepsilon_{ij} \sim N(0,\sigma^2)$ and ε_{ij}'s independent

$$Y_{ij} = \underset{\text{grand mean}}{\mu} + \alpha_i + \varepsilon_{ij}$$

$$\mu_i = \mu + \alpha_i \iff \overset{\text{difference to grand mean}}{\alpha_i} = \mu_i - \mu$$

Q: how can we write this as a linear regression? $(Y = X\beta + \varepsilon)$

$$n = \sum_{i=1}^{p} n_i \ , \ \text{Ex: } 5 \cdot 6 = 30$$

Yes, the model can be fitted as a linear regression with parameters $(\underset{\beta_0}{\underset{\uparrow}{\mu}}, \underset{\beta_1}{\underset{\uparrow}{\alpha_1}}, \underset{\beta_2}{\underset{\uparrow}{\alpha_2}}, .. \dots, \alpha_p)$.

Previously: dummy variable coding. ← problem: we want $\hat{\mu}$ average of all measurements

Now: effect coding

Impose a restriction on the α's: sum-to-zero-constraint $\sum_{i=1}^{p} \alpha_i = 0$, in practice only use $\alpha_1, .., \alpha_{p-1}$ and let $\alpha_p = -\sum_{i=1}^{p-1} \alpha_i$. This gives effect-coding of design matrix

Ex:

$$\underset{30\times 5}{X} = \begin{array}{cccccl} \beta_0=\mu & \beta_1=\alpha_1 & \beta_2=\alpha_2 & \beta_3=\alpha_3 & \beta_4=\alpha_4 & \{\alpha_5=-(\alpha_1+..+\alpha_4)\} \\ 1 & 1 & 0 & 0 & 0 & \text{recipe 1} \\ & & & & & \text{6 identical rows} \\ 1 & 0 & 1 & 0 & 0 & \text{recipe 2} \\ 1 & 0 & 0 & 1 & 0 & \text{recipe 3} \\ 1 & 0 & 0 & 0 & 1 & \text{recipe 4} \\ 1 & -1 & -1 & -1 & -1 & \text{recipe 5} \end{array}$$

Then, use the "old" $Y = X\beta + \varepsilon$, $\varepsilon \sim N_n(0, \sigma^2 I)$

and $\beta = \begin{bmatrix} \mu \\ \alpha_1 \\ \vdots \\ \alpha_{p-1} \end{bmatrix}$, and $\hat{\beta} = (X^T X)^{-1} X^T Y$.

$\longleftrightarrow \hat{\alpha}_p = -(\hat{\alpha}_1 + \dots + \hat{\alpha}_{p-1})$

2) Hypothesis test

$H_0: \mu_1 = \mu_2 = \mu_2 = \dots = \mu_p$ vs H_1: at least one different

$H_0: \alpha_1 = \alpha_2 = \dots = \alpha_p = 0$ vs H_1: at least one $\neq 0$.

Q: How can we do this with linear hypotheses and Fobs from L14?

Solution a: Write as linear hypotheses:

$\mu, \alpha_1, .., \alpha_{p-1}$

number of parameters: p

$$\text{"}C\beta = d\text{"} \longleftarrow \beta = \begin{bmatrix} \mu \\ \alpha_1 \\ \vdots \\ \alpha_4 \end{bmatrix}$$

Ex:

$$\underset{\substack{r \times p \\ \shortparallel \\ 4 \times 5}}{C} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \qquad d = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

$$f_{obs} = 4.3, \quad p\text{-value} = P(F_{4, \overset{n-p = 30-5}{25}} > 4.3) = 0.00875$$

$\Rightarrow$ reject H_0 $\Rightarrow$ difference between recipes.

Solution b) : in R, fit full model and anova(fullmodel). ANOVA table

"SS"	df	SS	$\frac{SS}{df} = MS$	Fvalue	p-val
(SSR) Treatment (regression)	$r = p-1$	x	x	x	x
(SSE) Error	$n-p$	x	x		

The classical way

$Y_{..} = \overset{\text{total}}{\text{average}}$ of data

$Y_{i.} =$ average grp i

SSA = our SSRegression, SSE = as before

$\Rightarrow$ F-test.

Two factor experiments

Ex: machine α's machines

γs operator

$$Y_{ij} = \mu + \alpha_i + \gamma_j + \varepsilon_{ij}$$

$i = 1, .., r^*$, $j = 1, .., s = n_i \ \forall i$

$\uparrow$ same (not n_i as before)

$\varepsilon_{ij} \sim N(0, \sigma^2)$ independent

We use sum-zero-constraint both for α's and γ's

$H_0^A: \alpha_1 = \alpha_2 \cdots = \alpha_r = 0$ effect of machine

Ex: machine: $r^* = 4$, $n = 4 \cdot 6 = 24$, #param: $1+3+5 = 9$

$$\underset{3\times 9}{C_A} = \begin{bmatrix} 0 & 1 & 0 & 0 & \\ 0 & 0 & 1 & 0 & \underset{3\times 5}{0} \\ 0 & 0 & 0 & 1 & \end{bmatrix} \rightarrow \underset{\underset{3,15}{\uparrow}}{F_A} = 3.34, \text{ p-value} = 0.048$$

$H_0^B: \gamma_1 = \gamma_2 = \cdots = \gamma_s = 0$ effect of operator

Ex: $s = 6$

$$\underset{5\times 9}{C_B} = \begin{bmatrix} 0 & & 1 & 0 & 0 & 0 & 0 \\ 0 & & 0 & 1 & 0 & 0 & 0 \\ 0 & \underset{5\times 3}{0} & 0 & 0 & 1 & 0 & 0 \\ 0 & & 0 & 0 & 0 & 1 & 0 \\ 0 & & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \quad \underset{\underset{5,15}{\uparrow}}{F_B} = 5.29, \text{ p-value} = 0.005$$

We have used "$C\beta = d$" for hypothesis test, but in practise you may use ANOVA ← already implemented in R:

```
fit = lm(y~A+B)
anova(fit)
```

Additive effects and interactions

$$Y_{ij} = \mu + \alpha_i + \gamma_j + \varepsilon_{ij} \quad \text{additive}$$

$$Y_{ijk} = \mu + \alpha_i + \gamma_j + (\alpha\gamma)_{ij} + \varepsilon_{ijk}$$

$i = 1,.., r$ (Ex: 2)
$j = 1,..., s$ (Ex: 5)
$k = 1,.., n_{ij}$ (Ex: 10)

$$\sum_{i=1}^{r} \alpha_i = 0$$

$$\sum_{j=1}^{s} \gamma_j = 0$$

$$\sum_{i=1}^{r} (\alpha\gamma)_{ij} = 0 \quad \text{for all } j$$

$$\sum_{j=1}^{s} (\alpha\gamma)_{ij} = 0 \quad i$$

In R: Words ~ Age * Process

Age + Process + Age : Process

cAge = $\boxed{\begin{matrix} -1 \\ +1 \end{matrix}}$ $\begin{matrix} = \text{young} \\ = \text{old} \end{matrix}$

First test: Interaction present? Ex: Yes!

Yes ↓ compare combinations

No ↓ Check effect of A: α_i's
B: γ_j's

TMA4267 Linear Statistical Models V2017 (L16)

Part 3: Hypothesis testing and analysis of variance
Multiple testing [note]

Mette Langaas

Department of Mathematical Sciences, NTNU

To be lectured: March 14, 2017

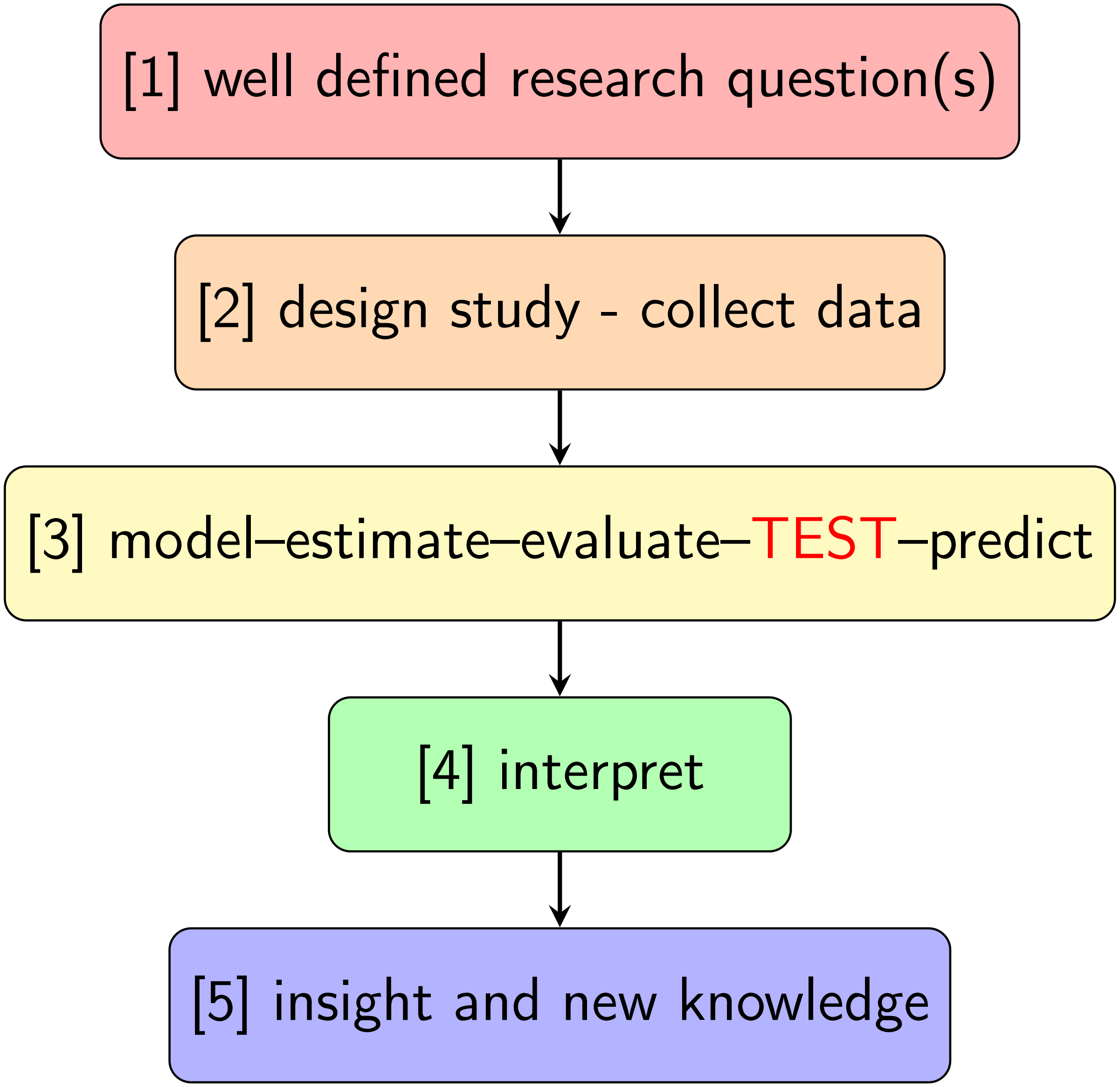
Today: Multiple testing

- Single hypothesis testing: H_0 and H_1, test statistic and p-value.
- Controlling Type I error (false positive findings) by selecting a significance level.
- Properties of p-values from true and false null hypotheses.
- Testing many hypotheses: why?
- Generalizing the type I error from single to multiple hypothesis testing: FWER and FDR.
- Two methods (Bonferroni and Šidák) that control the FWER
- Summarizing Part 3 with a quiz.

Basal metabolic rate and the FTO-gene

- The gene called FTO is known to be related to obesity
- The basal metabolic rate says how many calories you burn when you rest (hvilemetabolisme).
- Data has been collected for 101 patient from the obesity clinic at St. Olavs Hospital.
- Research question: is there an association between the variant of the FTO gene of the patient and the basal metabolic rate?
- Regression setting, other covariates include age, sex, weight, height, BMI, diet, exercise level, smoking, etc.

The scientific process



[1] well defined research question(s)

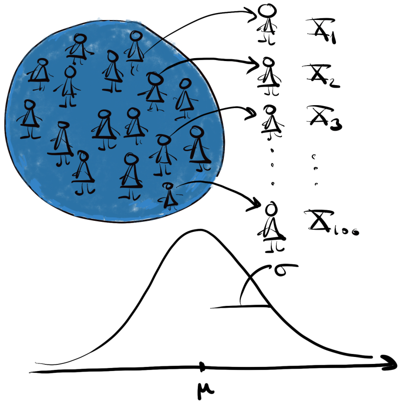
[2] design study - collect data

[3] model–estimate–evaluate–TEST–predict

[4] interpret

[5] insight and new knowledge

Hypothesis testing example (from L13)



- We draw a random sample of size $n = 100$ from the blue population and measure systolic blood pressure: $X_1, X_2, \dots, X_n$.
- Test statistic: $\bar{X} \sim N(120, 1)$ when H_0 is true.
- We find that $\bar{x} = 122$ mmHg.
- Data: $n = 100$, $\bar{x} = 122$, gives a p-verdi=0.02.

Hypothesis testing example (from L13)

Questions:

- How have I calculated this *p*-value?
  $P(\bar{X} > 122 \mid H_0 \text{ true})$.
- How can I interpret this *p*-value?
  Informally, a *p*-value is the probability under a specified statistical model that a statistical summary of the data would be equal to or more extreme than its observed value.
- Should I conclude that $\mu > 120$?
  Yes, if you choose significance level higher than 0.02. But, you should also report a (two-sided) confidence interval for μ: Here $[120.04, 123.96]$.

Single hypothesis testing set-up

	H_0 true	H_0 false
Not reject H_0	Correct	Type II error
Reject H_0	Type I error	Correct

Two types of errors:

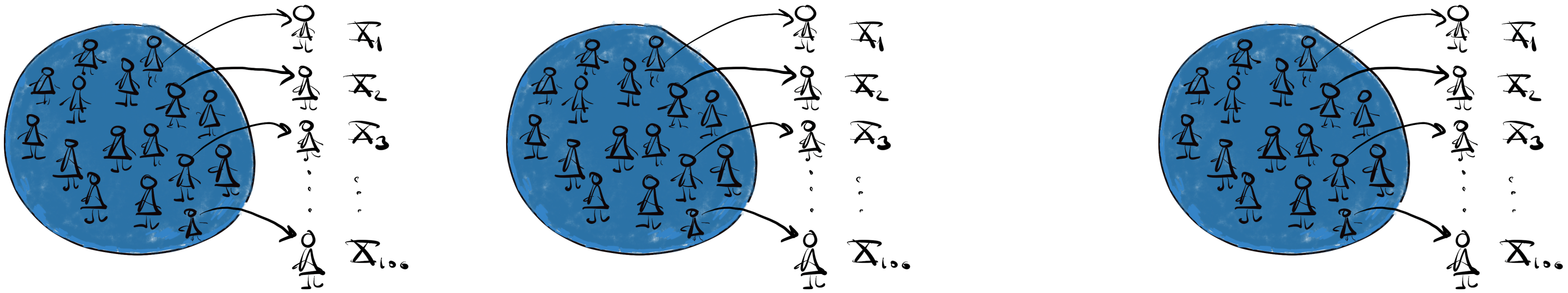
- False positives = type I error =miscarriage of justice.
- False negatives = type II error= guilty criminal go free.

The significance level of the test is α.
We reject the null hypothesis when the *p*-value is *below* α.

We say that : Type I error is "controlled" at significance level α.

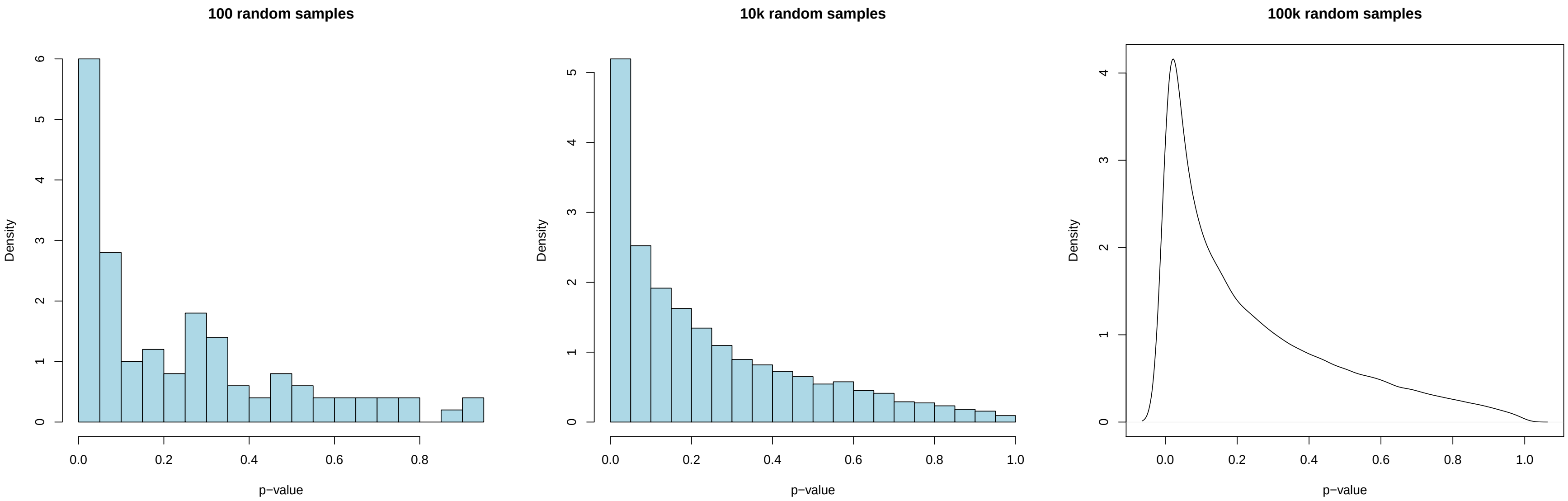
The probability of miscarriage of justice (Type I error) does not exceed α.

Repeating the blood pressure experiment



$\bar{x}$=120.9 $\bar{x}$ =118.9 $\cdots$ $\bar{x}$ = 121.2

p-value=0.18 *p*-value=0.86 $\cdots$ *p*-value=0.12



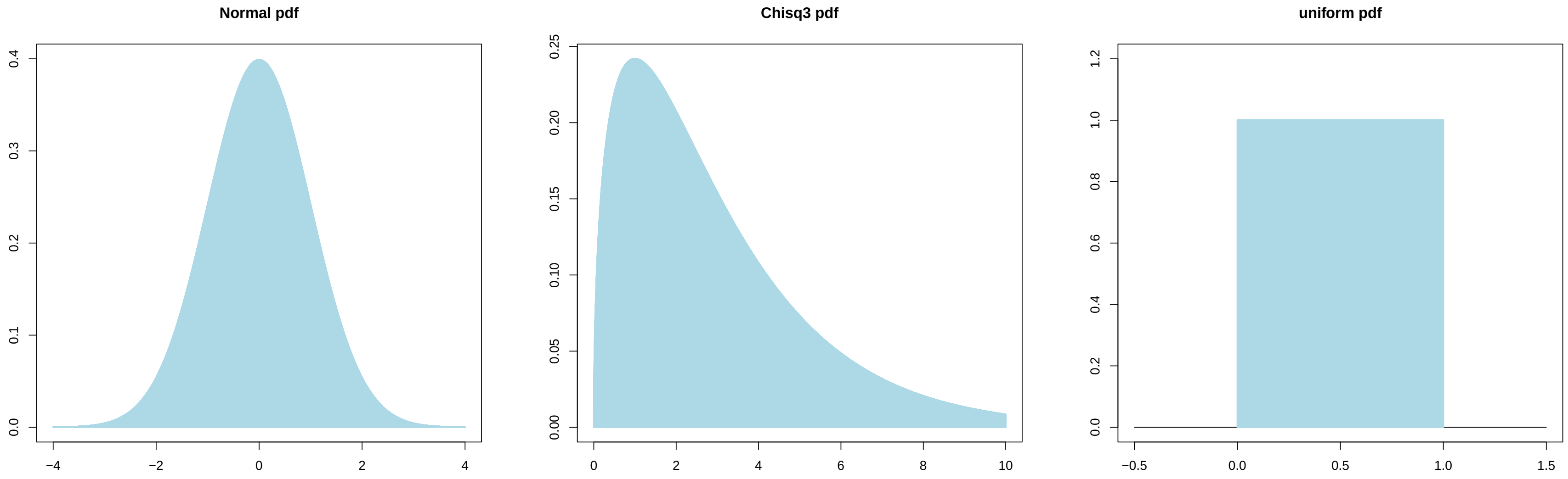
Histogram - and smoothed histogram of *p*-values.

More about the *p*-value

- The *p*-value is just a function of the random sample and can be regarded as a random variable.
  We had: $P(\bar{X} > \text{observed mean} \mid H_0 \text{ true})$.
- But, isn't the *p*-value a probability? A number?
- A random variable (like the *p*-value) has a *probability distribution*.
- What is the distribution of a *p*-value?

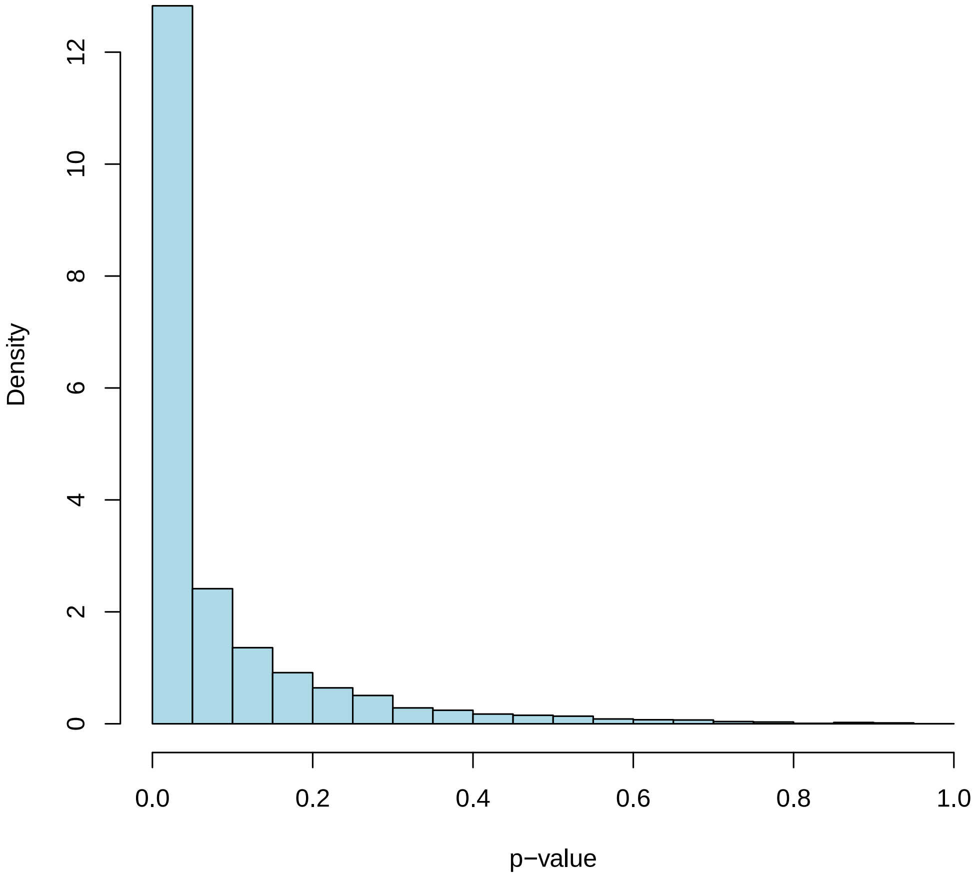
Probability distribution for random variable Y

- Continuous random variable Y (could be the p-value).
- Probability distribution function (pdf): $f(y)$.



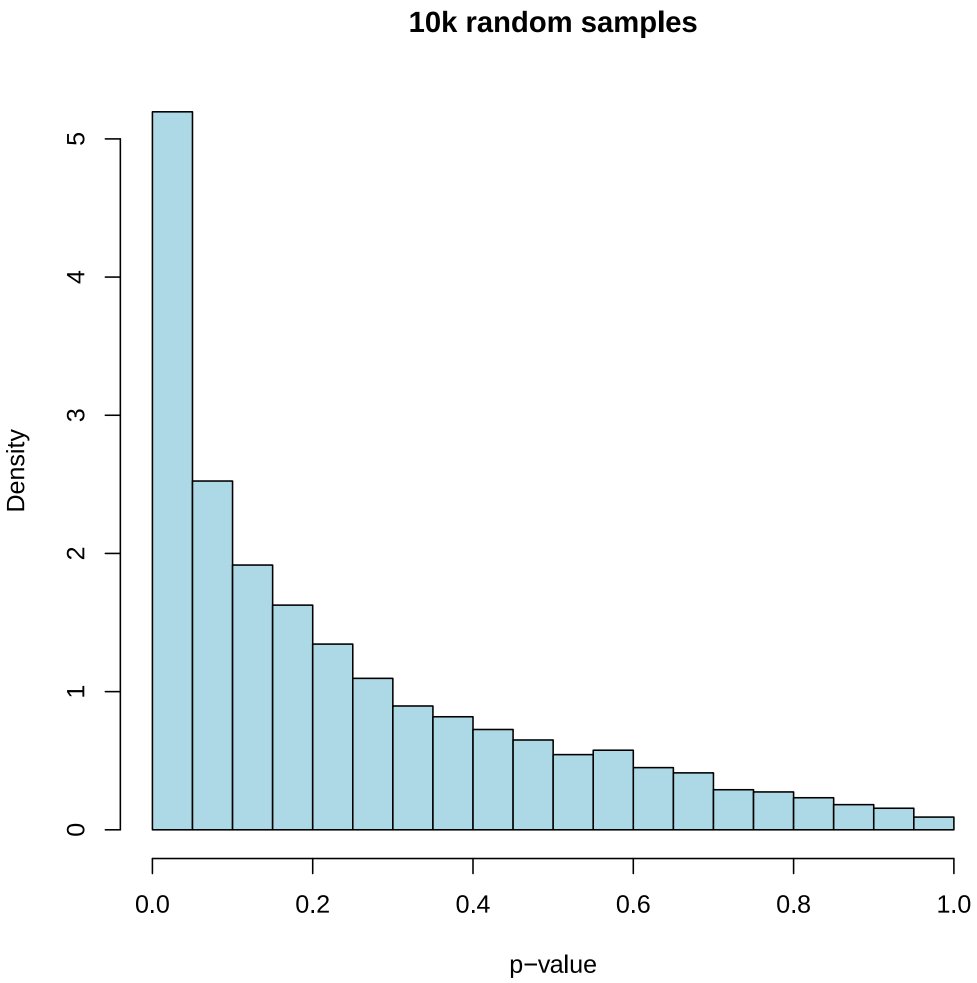
Distribution of *p*-values for false hypothesis?

Blood pressure example: Assume that $\mu = 122$ so that H_0 is false, and that we collect a random sample of size 100. What is then the distribution of the *p*-value?

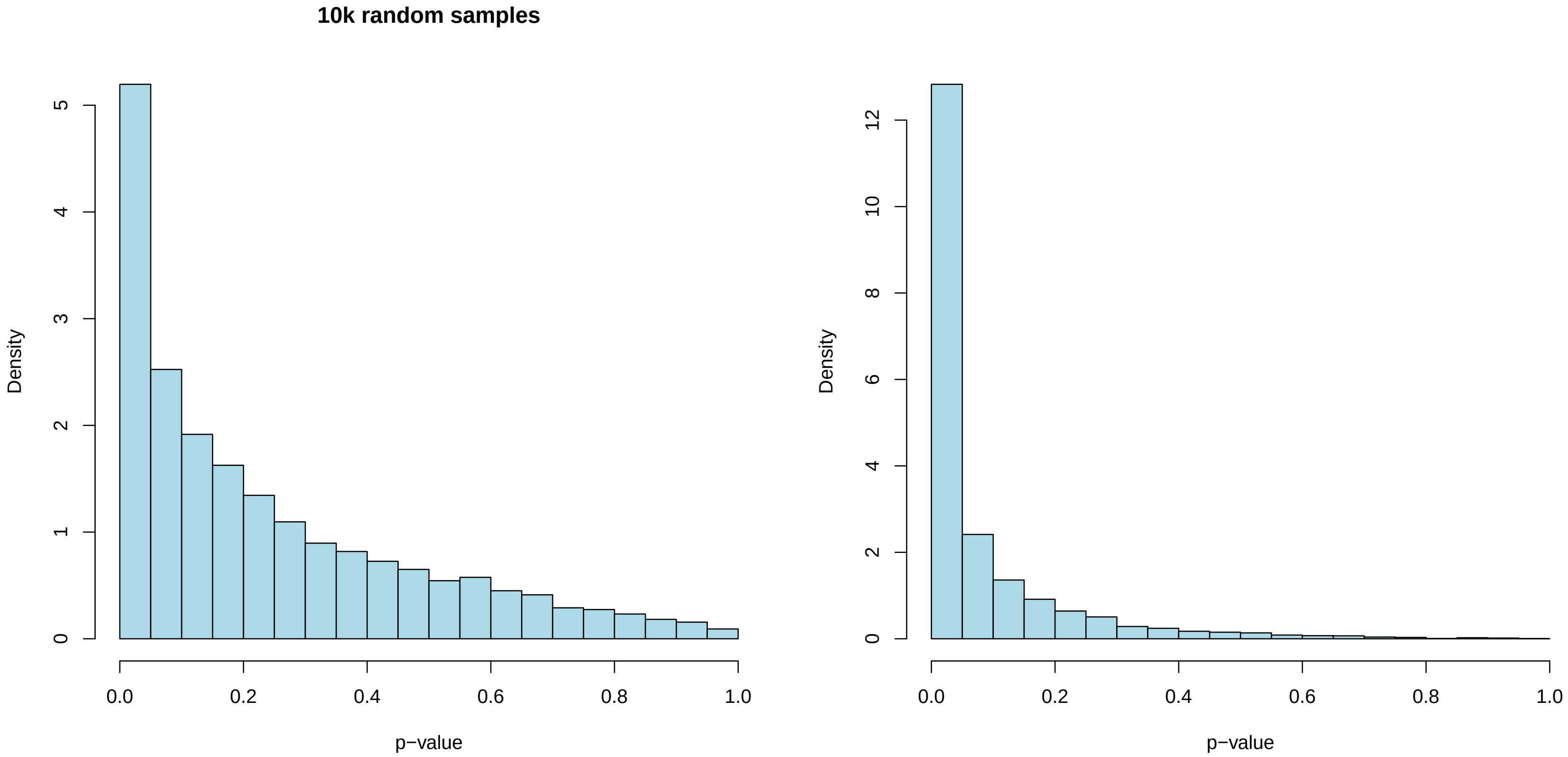


Distribution of *p*-values for false hypothesis?

Blood pressure example: Assume that $\mu = 121$ so that H_0 is false, and that we collect a random sample of size 100. What is then the distribution of the *p*-value?

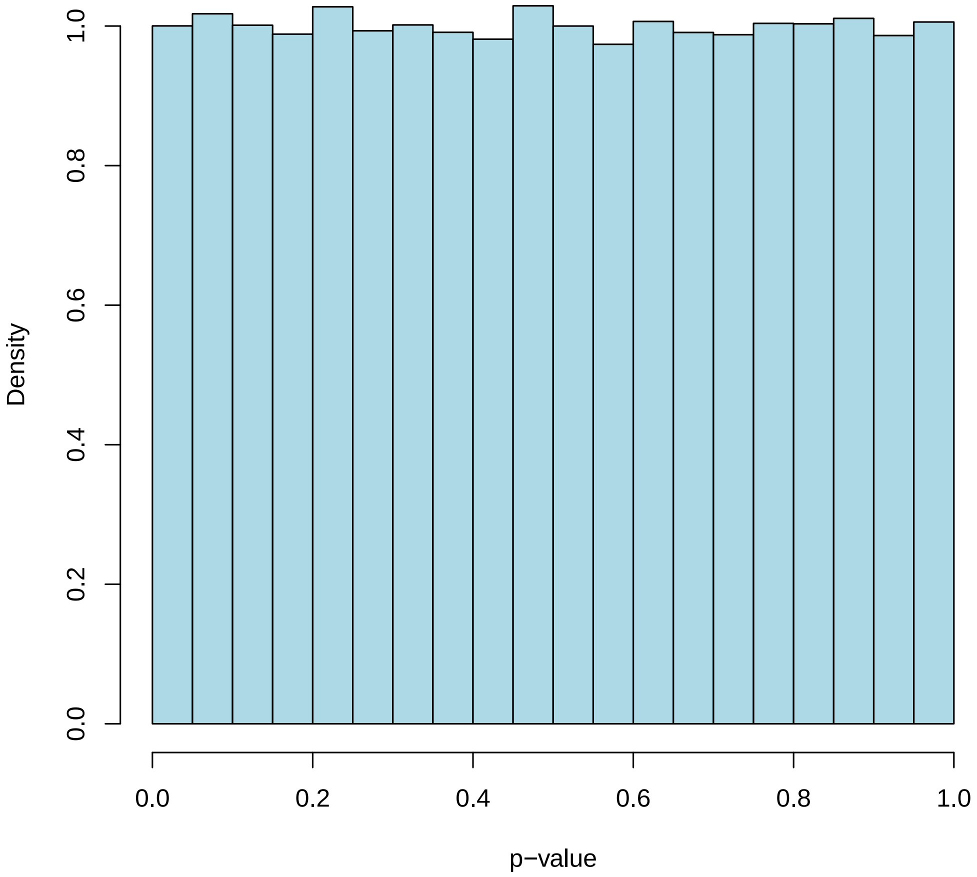


False null: $\mu = 121$ left, and $\mu = 122$ right, when $H_0 : \mu = 120$



Distribution of *p*-values for true hypothesis?

Blood pressure example: Assume that $\mu = 120$ so that H_0 is true, and that we collect a random sample of size 100. What is then the distribution of the *p*-value?



Urban myth: A *p*-value for a true null hypothesis is close to 1. No, all intervals of equal length are equally probable! =uniform distribution

p-values from true null hypothesis is uniformly distributed

Why is this important:

- so you don't believe the urban myth, and
- it might be useful to understand plots (pdf or cdf) of *p*-values, and these are often used for quality control of statistical models.

Assume that large values of the test statistic T leads to rejection of the null hypothesis, and that a value t of the test statistic T corresponds to a value w of the *p*-value W. This means that $P(T \geq t) = P(W \leq w)$. On the other hand the *p*-value is $P(W \leq w) = P(T \geq t) = w$ when H_0 is true.
This means that $P(W \leq w) = w$ when H_0 is true. If W is a continuous random variable taking values from 0 to 1, the the *p*-value W must be uniformly distributed over the interval from 0 to 1.
This is true when the *p*-value is continuous and exact.

Exact *p*-value

If $P(p(\boldsymbol{Y}) \leq \alpha) = \alpha$ for all α, $0 \leq \alpha \leq 1$, the *p*-value is called an *exact p*-value.

Valid *p*-value

A p-value $p(\boldsymbol{Y})$ is *valid* if

$$P(p(\boldsymbol{Y}) \leq \alpha) \leq \alpha$$

for all α, $0 \leq \alpha \leq 1$, whenever H_0 is true, that is, if the p-value is valid, rejection on the basis of the p-value ensures that the probability of type I error does not exceed α.

From single to multiple hypothesis testing

In many situations we are not interested in testing only one hypothesis, but instead m hypotheses.

- In a regression setting m might be the number of covariates in the regression model, and we would test $H_0 : \beta_j = 0$ vs $H_1 : \beta_j \neq 0$ for all $j = 1, \ldots, m$.
- If we have a linear regression with one categorical covariate with k levels, called a one-way analysis of variance model, we might first want to test $H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$ against the alternative hypothesis, H_1, that the means of at least two of the k levels are different from each other. If the null hypothesis is rejected we might want to continue to test which of all possible pairs of the means that are different – giving $m = \binom{k}{2}$ hypothesis tests, or compare the mean of all levels to a common reference level μ_1, giving $m = k - 1$ hypothesis tests.

But, can't we still use cut-off α on the p-values to detect significant findings?

Westfall & Young (1993): Multicenter Oat Bran Study

- At each of ten study centers a control vs treated experiment is performed with 20 subjects per group.
- It is common to analyze the data for each center separately, as well as to combine over center.
- T-statistics are computed for each center as

$$\frac{\bar{y}_T - \bar{y}_C}{\sqrt{(s_T^2 + s_C^2)/20}}$$

  with p-values calculated as lower tail probabilities from the t-distribution with 38 degrees of freedom.

FIRST Oat Bran Study

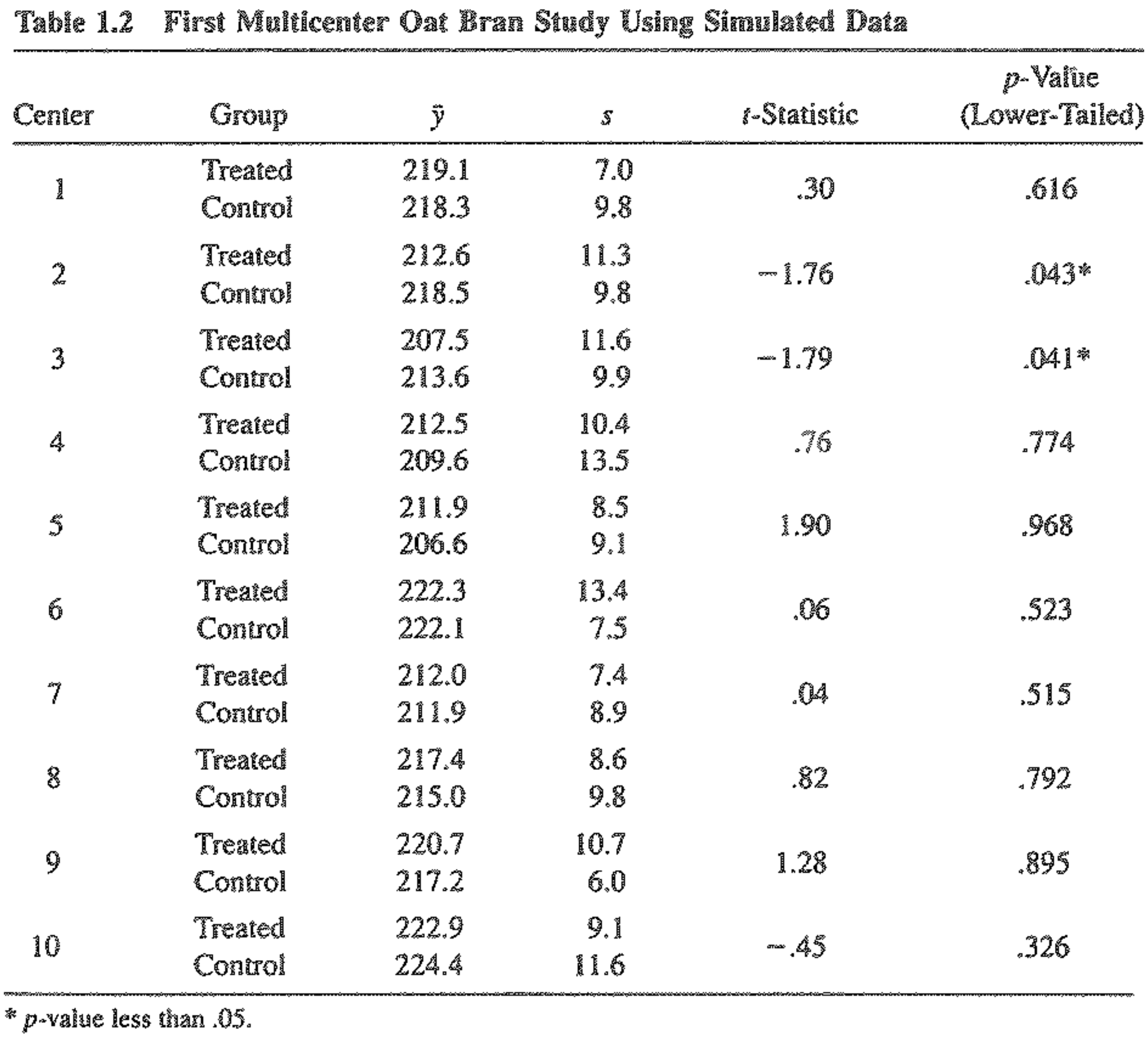


Table 1.2 First Multicenter Oat Bran Study Using Simulated Data

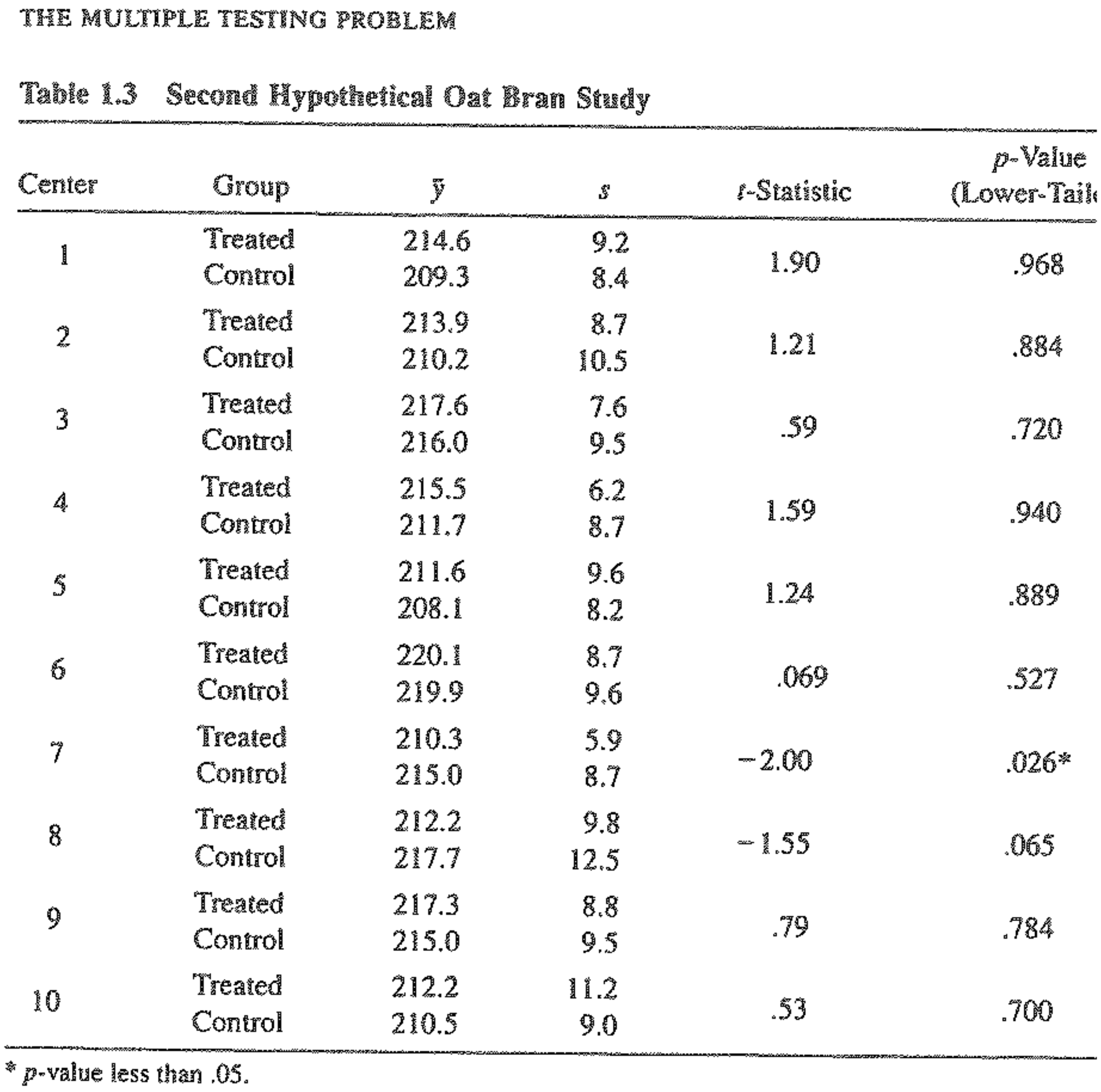
Center	Group	$\bar{y}$	s	t-Statistic	p-Value (Lower-Tailed)
1	Treated	219.1	7.0	.30	.616
	Control	218.3	9.8		
2	Treated	212.6	11.3	−1.76	.043*
	Control	218.5	9.8		
3	Treated	207.5	11.6	−1.79	.041*
	Control	213.6	9.9		
4	Treated	212.5	10.4	.76	.774
	Control	209.6	13.5		
5	Treated	211.9	8.5	1.90	.968
	Control	206.6	9.1		
6	Treated	222.3	13.4	.06	.523
	Control	222.1	7.5		
7	Treated	212.0	7.4	.04	.515
	Control	211.9	8.9		
8	Treated	217.4	8.6	.82	.792
	Control	215.0	9.8		
9	Treated	220.7	10.7	1.28	.895
	Control	217.2	6.0		
10	Treated	222.9	9.1	−.45	.326
	Control	224.4	11.6		

* p-value less than .05.

FIRST Oat Bran Study

- Centres 2 and 3 show significant reduction in blood cholestreol for the treatment group.
- Centre 5 happens to show a significant increase, but that is not "noticed" since one-sided tests are performed.
- If the studies were run as uncoordinated trials, it is likely that the two significant studies would be reported and perhaps published in reputable journals.
- The eight nonsignificant studies would go to the file drawer and a "true, confirmed" effect would be established for the two sites where significance is found.
- The centres with insignificant results may decide to collect fresh data, and analyse only the new data.

SECOND Oat Bran Study



THE MULTIPLE TESTING PROBLEM

Table 1.3 Second Hypothetical Oat Bran Study

Center	Group	$\bar{y}$	s	t-Statistic	p-Value (Lower-Tail
1	Treated Control	214.6 209.3	9.2 8.4	1.90	.968
2	Treated Control	213.9 210.2	8.7 10.5	1.21	.884
3	Treated Control	217.6 216.0	7.6 9.5	.59	.720
4	Treated Control	215.5 211.7	6.2 8.7	1.59	.940
5	Treated Control	211.6 208.1	9.6 8.2	1.24	.889
6	Treated Control	220.1 219.9	8.7 9.6	.069	.527
7	Treated Control	210.3 215.0	5.9 8.7	−2.00	.026*
8	Treated Control	212.2 217.7	9.8 12.5	−1.55	.065
9	Treated Control	217.3 215.0	8.8 9.5	.79	.784
10	Treated Control	212.2 210.5	11.2 9.0	.53	.700

* p-value less than .05.

Oat bran study: lessons to be learned

- These are SIMULATED data with equal means of the control and the treatment group, i.e. the truth is that there are no biological effects of the treatment.
- With simulated data: simple to point to the multiplicity issue as the *cause* for the small p-values for some centres.
- Real studies: not easy to determine if a seen effect is real or not.

Oat bran study: lessons to be learned

- Real studies: not easy to determine if a seen effect is real or not.
- At a particular centre showing significance: scientists would believe that the effect is real, because why should the existence of other centres in the study affect the outcome at the given centre?
- How should one verify that an unusual event is real or artificial?
- The possibility of false positive results is very real, and can lead to serious misinterpretation by analysts: it is human nature to rationalize any dramatic- statistically significant - change.

From single to multiple hypothesis testing

Set-up

- Let us assume that we perform m hypothesis tests,
- giving m p-values and then
- choose a cut-off on the p-values at some value α_{loc} (called a local significance level) to decide if we want to reject each null hypothesis.
- We then reject the null hypotheses where the p-value is smaller than α_{loc}, and this leads to rejection of R hypotheses.

Multiple hypothesis testing set-up

One hypothesis:

	Not reject H_0	Reject H_0
H_0 true	Correct	Type I error
H_0 false	Type II error	Correct

m **hypotheses:**

	Not reject H_0	Reject H_0	Total
H_0 true	U	V	m_0
H_0 false	T	S	$m - m_0$
Total	$m - R$	R	m

- R rejected null hypotheses
- V false positives (type I errors)
- T false negatives (type II errors)

Only m and R are observed. **What should we now control?**

Overall Type I error control (1)

- In some situation one expects that just a few null hypothesis are false,
- therefore a *strict* criterion for controlling an overall version of the Type I error is chosen.
- Family-Wise Error Rate (FWER) is controlled at level α.

  $$\text{FWER} = P(V \geq 1) = P(\text{the number of false positives is } \geq 1)$$

  (remark: V is not observed)
- The FWER can be controlled by defining a *local significance level* α_{LOC} for each test and reject the H_0 of that test if the p-value of the test is less than the α_{LOC}.

Basal metabolic rate and the FTO-gene: revisited

- The gene called FTO is known to be related to obesity
- The basal metabolic rate says how many calories you burn when you rest (hvilemetabolisme).
- Data has been collected for 101 patient from the obesity clinic at St. Olavs Hospital.
- Research question: is there an association between the variant of the FTO gene of the patient and the basal metabolic rate?
- Regression setting, other covariates include age, sex, weight, height, BMI, diet, exercise level, smoking, etc.

If we had not only collected data on this one gene, but instead for many (e.g. $m = 100000$) genetic markers positioned along the chromosome, and then wanted to test m hypotheses, we would not expect to find many true associations. This strategy is called a genome-wide association analysis and for this purpose FWER is usually controlled.

Overall Type I error control for GWA data: FWER control

- GWAS often use $\alpha_{\text{LOC}} = 5 \cdot 10^{-8}$.
- The most popular method controlling the FWER is the Bonferroni method, which can always be used.
- The Bonferroni method might be slightly conservative (too low α_{LOC}), since it is constructed to control FWER for all types of dependency structures between the test statistics for the different hypotheses- including independence.
- `https://arxiv.org/abs/1603.05938`: *Efficient and powerful familywise error control in genome-wide association studies using generalized linear models*, K. K. Halle, Ø. Bakke, S. Djurovic, A. Bye, E. Ryeng, U. Wisløff, O. A. Andreassen, M. Langaas.

Overall Type I error control (2)

- For other types of data one expects that many null hypotheses are false,
- and therefore a less strict criterion for controlling an overall version of the Type I error is chosen.
- The False Discovery Rate (FDR) by Benjamini & Hochberg (1995) is controlled at level α.
- Informally, the FDR is the expected proportion of Type I errors among the rejected hypotheses.

$$\text{FDR} = E(Q) \text{ where by definition}$$

$$Q = \begin{cases} V/R & \text{if } R > 0, \text{ or} \\ 0 & \text{if } R = 0 \end{cases}$$

Hedenfalk et al (2001) gene expression dataset

Available from library(qvalue) from Bioconductor

- The data from the breast cancer gene expression study of Hedenfalk et al. (2001) were obtained and analyzed.
- A comparison was made between 3,226 genes of two mutation types, BRCA1 (7 arrays) and BRCA2 (8 arrays).
- The data included here are p-values, test-statistics, and permutation null test-statistics obtained from a two-sample t-test analysis on a set of 3170 genes, as described in Storey and Tibshirani (2003).

For such gene expression data researchers expect to find may genes that are differently expressed between conditions and therefore the false discovery rate (FDR) is usually controlled. Hedenfalk I et al. (2001). Gene expression profiles in hereditary breast cancer. *New England Journal of Medicine*, 344: 539-548. Storey JD and Tibshirani R. (2003). Statistical significance for genome-wide studies. *Proceedings of the National Academy of Sciences*, 100: 9440-9445. http://www.pnas.org/content/100/16/9440.full

Overall Type I error control for gene expression data

- Popular algorithm for controlling the FDR: the Benjamini-Hochberg step-up procedure.
- Focus on minimal interesting biological effect: is possible that you don't want to test *difference between treatments*=0, but instead $\geq$ minimal biological interesting effect.

Multiple testing

- Note from course www-page.
- RecEx5.Problem 2.
- CompulsoryPart3 Problem 2.
- This topic is new on the reading list in 2017.
- It replaces the topics of regularization with the lasso and ridge regression, which will be covered in TMA4268 Statistical Learning.

Summarizing Part 3

with quiz in Kahoot!

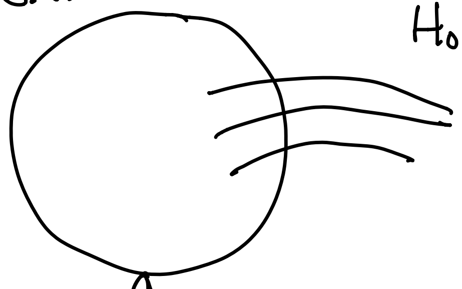
Multiple hypothesis testing
(note available from Bb)

Mlb, TMA4267
14.03.2017

First: single hypothesis testing

Ex:

$H_0: \mu = 120$ mmHg vs $H_1: \mu > 120$ mmHg



X_1
$\vdots$
X_n

$n = 100$

$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

$X \sim N(\mu, \sigma^2 = 10^2)$

$\bar{X} \sim N(\mu, \frac{\sigma^2}{n} = \frac{10^2}{100} = 1)$

$\bar{X} \sim N(120, 1)$ when H_0 true

Observed $\bar{x} = 122$.

p-value: $P(\bar{X} \geq 122) = P(\frac{\bar{X} - 120}{1} \geq \frac{122-120}{1})$

$= 1 - \Phi(2) \approx 0.02$

Informally: the p-value is the probability that our test statistic ($\bar{X}$) is observed to be $\bar{x} =$ "122 or a more extreme value" (that is $\bar{X} \geq 122$), when the truth is that $\mu = 120$ so that $\bar{X} \sim N(120, 1)$.

Then we choose if we have enough evidence against H_0 by looking at the p-value.

If the p-value is small then what we have observed (or more extreme obs.) is not very probable when H_0 is true. ⇒ so for small p-values we believe that H_0 must be false and reject H_0.

smaller than chosen significance level α

10%, 5%, 1%

⇒ So, the p-value can be seen as a probability?

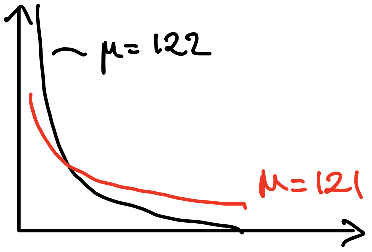
Q: What happens if I collect data on $n=100$ new persons from the population. We observe a new $\bar{x}$, and will get a new p-value.

⟹ The p-value is a random variable - and it has a probability distribution.

Ex: blood pressure. Easy to semple 100 from $N(\mu, \sigma^2 = 100)$, calculate $\bar{x}$ and p-value. ⇒ make histogram.

$P(\bar{X} \geq \_\_ \mid \mu = 120)$

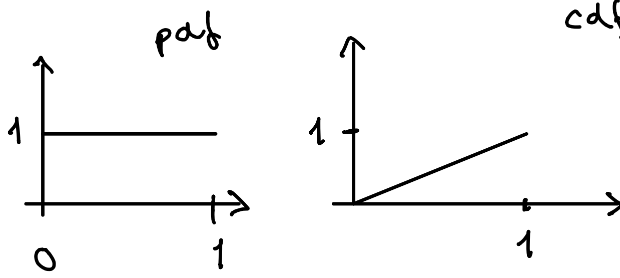
$N(122, 10^2)$



⇐ p-values from false H_0's.

When H_0 is true the p-values are uniformly distributed.

$f(u) = u$



$F(u) = P(U \le u) = u$

$\Rightarrow$ see note for R-code & proof!

This is (usually) non-intuitive to people... but rather useful to know...

See note: define valid and exact p-value.

Multiple hypotheses

we want to decide on value for α_{Loc} $\Downarrow$

$m = \#$ hypotheses

$R = \#$ hypotheses that we reject, where $p\text{-value} < \alpha_{Loc}$

	Not reject H_0	Reject H_0	Total
H_0 true	Correct	Type I error ← fake news V false positives	m_0
H_0 false	Type II errors	Correct	$m - m_0$
Total		R	m

⇖ we only know m and R

Generalization of type I error

$$\text{FWER} = P(V > 0) = P(V \geq 1)$$

↑ familywise error rate

↗ one or more fake news (false positive)

We want to control FWER – that means to find α_{Loc} so that $P(V>0) \leq \overset{\alpha}{\swarrow}\ 0.1$ / 0.05

Let $R_i = \{\text{reject } H_0 \text{ nr } i, \text{ i.e. } P_i \le \alpha_{loc}\}$

$\bar{R}_i = \{\text{not reject } H_0 \text{ nr } i, \ P_i > \alpha_{loc}\}$

$$P(V>0) = 1 - P(V=0) \overset{\text{assume all } H_0 \text{ true}}{=} 1 - P(\underbrace{\bar{R}_1 \cap \bar{R}_2 \cap \dots \cap \bar{R}_m}_{\text{need the joint distribution of the } m \text{ Test statistics } T_1, \dots, T_m})$$

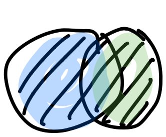
→ perform a multiple integral. Difficult to solve. See note on details.

Bonferroni's method: Assume all H_0 are true.

$$P(V>0) = P(R_1 \cup R_2 \cup R_3 \cup \dots \cup R_m)$$
$$\le P(R_1) + P(R_2) + \dots + P(R_m)$$

$$P(A \cup B) \le P(A) + P(B)$$

Boole's inequality



$$P(V>0) \le \underbrace{P(\overset{R_1}{\text{rejecting } H_0 \text{ nr } 1})}_{P(P_i \le \alpha_{Loc})} + \dots + P(\overset{R_m}{\text{rej. } H_0 \text{ nr } m})$$

$$\underset{\underset{\alpha}{\parallel}}{\underset{\text{FWER}}{\underset{\parallel}{P(V>0)}}} \le \underset{\uparrow}{\alpha_{LOC}} + \alpha_{LOC} + \cdots + \alpha_{LOC} = m \cdot \alpha_{LOC}$$

since p-values from true H_0's are uniform! $P(p_i \le \alpha_{LOC}) = \alpha_{LOC}$

Then I need to choose $\boxed{\alpha_{LOC} = \dfrac{\alpha}{m}}$

Rule: reject H_0 when p-value $< \dfrac{\overset{\text{FWER}}{\alpha}}{\underset{\# \text{ tests.}}{\underset{\uparrow}{m}}}$

if p-values are valid $P(p_i \le \alpha_{LOC}) \le \alpha_{LOC}$, so that is also ok here.

TMA4267 Linear statistical models

Part 3: Hypothesis testing and ANOVA

March 14, 2017

Happiness

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.072081   0.852543  -0.085   0.9331
money        0.009578   0.005213   1.837   0.0749
sex         -0.149008   0.418525  -0.356   0.7240
love         1.919279   0.295451   6.496 1.97e-07
work         0.476079   0.199389   2.388   0.0227
```

For which covariates would we reject the null hypothesis $\beta = 0$ at significance level 1%?

A money

B sex

C love

D work

Type I errors

What is a commonly used name for the type I errors?

A true positives

B false positives

C false negatives

D true negatives

Linear hypotheses

$H_0 : \boldsymbol{C}\beta = \boldsymbol{d}$ in a regression model $\boldsymbol{Y} = \boldsymbol{X}\beta + \varepsilon$.
n=number of observations,
p = number of estimated regression coefficients
r=number of linear hypotheses (rank of $\boldsymbol{C}$).

What is the distribution of F_{obs}
$=\frac{1}{r}(\boldsymbol{C}\hat{\beta} - \boldsymbol{d})^T(\hat{\sigma}^2\boldsymbol{C}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{C}^T)^{-1}(\boldsymbol{C}\hat{\beta} - \boldsymbol{d})$?

A $F_{r,n-p}$

B $F_{p,n-r}$

C $N(\beta, \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1})$

D $N(0, \sigma^2\boldsymbol{I})$

ANOVA

Which type of covariate coding is used in the one-way ANOVA model with design matrix given as:

$$\begin{matrix}
1 & 1 & 0 & 0 & 0 \\
1 & 0 & 1 & 0 & 0 \\
1 & 0 & 1 & 0 & 0 \\
1 & 0 & 0 & 1 & 0 \\
1 & 0 & 0 & 0 & 1 \\
1 & 0 & 0 & 0 & 1 \\
1 & -1 & -1 & -1 & -1 \\
1 & -1 & -1 & -1 & -1
\end{matrix}$$

A Continuous

B Effect coding

C Dummy variable coding

D Categorical

ANOVA

Is the interaction term significant at significance level 0.01?

```
> res <- lm(Words~Age*Process)
> anova(res)
            Df  Sum Sq Mean Sq F value    Pr(>F)
Age          1  240.25  240.25 29.9356 3.981e-07 ***
Process      4 1514.94  378.74 47.1911 < 2.2e-16 ***
Age:Process  4  190.30   47.58  5.9279 0.0002793 ***
Residuals   90  722.30    8.03
```

A Yes

B Not enough information to decide

C No

p-value from true null hypothesis

For a continuous test statistic that gives an exact *p*-value, what is the distribution the *p*-value when the null hypothesis is true?

A Normal

B Chisquared

C Exponential

D Uniform

FWER

V=number of false positives and
R=number of rejections.
The familywise error rate FWER is

A $E(V/R)$

B $E(V)$

C $P(V/R > 0.05)$

D $P(V > 0)$

Bonferroni

α=level for control of FWER.

α_{loc}=cut-off on p-value

m =number of tests.

What is the Bonferroni rule?

A $\alpha_{\text{loc}} = m\alpha$

B $\alpha_{\text{loc}} = \frac{\alpha}{m}$

C $\alpha_{\text{loc}} = \alpha^{m}$

D $\alpha_{\text{loc}} = (1 - \alpha)^{1/m}$

Correct?

Are you sure you want to read the correct answers? Maybe try first? The answers are explained on the next two slides.

Answers

1. C: only love is significant on level 1%, since this is the only *p*-value below 0.01 (last column).
2. B: type I errors are called false positive findings
3. A: linear hypotheses with $F_{r,n-p}$-distributed statistic.
4. B: Effect coding is used in ANOVA.

Answers

5. A: Interaction term has p-value below 0.01.
6. D: p-values from true nulls are uniform.
7. D: FWER is the probability of one or more false positives.
8. B: Bonferroni rule is α/m.