TMA4267 Linear Statistical Models V2017 (L22)

Summing up

Mette Langaas

Department of Mathematical Sciences, NTNU

To be lectured: April 28, 2017

Outline: Summing up

- Learning outcomes
- Overview, important concepts and exam problems V2016:
  - Part 1: Multivariate RVs and the multivariate normal distribution
  - Part 2: Linear regression
  - Part 3: Hypothesis testing and analysis of variance
  - Part 4: Design of experiments
- Final reading list
- Exam
- Activities before the exam

Part 1: Multivariate RVs and the multivariate normal distribution [L1-L6, CompEx1, RecEx1-2]

Curriculum:

- Härdle, Simar (2015): Applied Multivariate Statistical Analysis, Fourth edition, Springer.
  - Chapter 2
  - Chapter 3.3 (89-93)
  - Chapter 4.1-4.5 (117-149)
  - Chapter 5.1 (p.181-190)
  - Chapter 11.1-11.3 (p. 319-331)
- Fahrmeir, Kneib, Lang and Marx (2013): Regression, Springer.
  - Appendix B
- Slides and handouts ($\sim$ 200 pages): `https://www.math.ntnu.no/emner/TMA4267/2017v/TMA4267V2017Part1.pdf`

Part 2: Linear regression [L7-L12, CompEx2, RecEx3-4]

Curriculum:

- Fahrmeir, Kneib, Lang and Marx (2013): Regression, Springer.
  - Chapter 3.
  - Appendix B
- Slides and handouts: `https://www.math.ntnu.no/emner/TMA4267/2017v/TMA4267V2017Part2.pdf`

V2016 Problem 2: Plant stress (slightly modified to reflect changes to the reading list)

At the Department of Biology at NTNU researchers use the model plant *Arabidopsis thaliana* to study the response of a plant to different sources of stress. In an experiment *Arabidopsis thaliana* seedlings were subject to a stress situation. The following factors were fitted:

- D (damage): $D = 1$ means that the plant was damaged mechanically by cutting into the leaves of the plant by a pair of scissors. $D = -1$ means damage was not inflicted (no cutting).
- F (flagellin): $F = 1$ means that the pathogen-derived peptide flagellin was sprayed on the leaves of the plant. $F = -1$ means water (not flagellin) was sprayed.
- T (time): Plants were harvested at two different time points after the stress situation. $T = 1$ means that the plant was harvested 60 minutes after the stress situation and $T = -1$ means that the plant was harvested 30 minutes after the stress situation.

Thus, we have three factors, D, F and T, each at two levels. In the study experiments for all possible combinations of the three factors were performed four times yielding 32 experiments in total.

V2016 Problem 2: Plant stress

The response measured in the experiment, was the observed gene activity level (a continuous measurement) of each of around 40 000 genes. We will only focus on the gene activity level of one of these genes, the AT1G32920 gene, and we denote the gene activity level of this gene by Y. It is known that this gene is active in response to wounding.
For experiment number i (where $i = 1, \ldots, 32$): Y_i is the observed response, D_i is chosen value of D, F_i is chosen value of F, and T_i is chosen value of T. A multiple regression model with all main effects, and two- and three-way interactions, was considered,

$$Y_i = \beta_0 + \beta_D D_i + \beta_F F_i + \beta_T T_i + \beta_{D:F} D_i F_i + \beta_{D:T} D_i T_i + \beta_{F:T} F_i T_i + \beta_{D:F:T} D_i F_i T_i + \varepsilon_i,$$

where $i = 1, \ldots, 32$, and we assume ε_i independent and identically normally distributed with mean 0 and variance σ^2. We refer to this as the *full model*.
Note that the interactions are simply products of the factors. The vector of regression parameters is
$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 & \beta_D & \beta_F & \beta_T & \beta_{D:F} & \beta_{D:T} & \beta_{F:T} & \beta_{D:F:T} \end{pmatrix}^{\mathsf{T}}$, and the ith row of the design matrix $\boldsymbol{X}$ is $\begin{pmatrix} 1 & D_i & F_i & T_i & D_i F_i & D_i T_i & F_i T_i & D_i F_i T_i \end{pmatrix}$.

V2016 Problem 2a: Plant stress

In the print-out from `summary(fit)` *four* numerical values are replaced by question marks. Calculate numerical values for each of these, and explain what each of the values means.

```
> summary(fit)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 16.15942    0.04140  ?       < 2e-16
D            0.93739    0.04140  22.644  < 2e-16
F            0.28546    0.04140   6.896 3.93e-07
T           -0.52354    0.04140 -12.647 4.18e-12
D:F         -0.08878    0.04140  -2.145  0.04231
D:T         -0.00242    ?        -0.058  0.95386
F:T         -0.12614    0.04140  -3.047  0.00555
D:F:T        0.09099    0.04140   2.198  ?
Residual standard error: 0.2342 on 24 degrees of freedom
Multiple R-squared:      ?, Adjusted R-squared:  0.9594
F-statistic: 105.6 on 7 and 24 DF,  p-value: < 2.2e-16
```

V2016: Plant stress 2b

How would you, based on the figures given and the R-output, evaluate the fit of the model?

How would you explain to a biologist what the estimated main effect of damage means in practice? How would you explain the estimated interaction effect between damage and flagellin?

Let $\gamma = 2^{\beta_F - \beta_D}$ be a new parameter of interest.
Suggest an estimator, $\hat{\gamma}$, for γ. Use approximate methods to find the expected value and variance of this estimator, that is, $\mathrm{E}(\hat{\gamma})$ and $\mathrm{Var}(\hat{\gamma})$. Use results in the prinout to calculate numerical value for $\hat{\gamma}$, and estimated numerical values for $\mathrm{E}(\hat{\gamma})$ and $\mathrm{Var}(\hat{\gamma})$.
Hint: You may use that $2^x = \exp(x \ln 2)$, where ln is the natural logarithm.

Topic: choosing the "best" linear regression model!

- First, debunk popular strategies (based on simulations studies were we knew the "true" model):
  - Popular 1: fit all available covariates.
    Problem: overfitting (=fitting trends and noise).
  - Popular 2: fit all available covariates, then remove the insignificant ones (=those β_j where $H_0 : \beta_j = 0$ is rejected).
    Problem: may also remove important covariates that are correlated with unimportant ones - but insignificant because being masked by the unimportant ones.
- Study of irrelevant and missing covariates:

  Irrelevant : variables that are included in the regression but should not have been (IQ of lumberjack)

  missing : variables that are not included, but should have been (omitting height in the tree volum example)

  Conclusion in book: the model should not contain irrelevant covariates, and we should aim for a sparse model. Take home message is the "Law of parsimony": *If two models are not very different – then always choose the simplest one.*

Model selection

Want to choose the model that minimize the

$$\text{SPSE} = \sum_{j=1}^{J} \mathrm{E}((Y_j - \hat{Y}_{jM})^2)$$

Several solution based on test set or cross-validation. We have focused on using only the original data and a penalized criterion: a first term based on SSE (or R^2) for model M, and a second term penalizing the model complexity.

R^2 adjusted (corrected)

Mallows' C_p

Akaike Information Criterion (AIC)

Bayesian Information Criterion (BIC)

NB: there is no overall best choice for criterion - all of these are used.

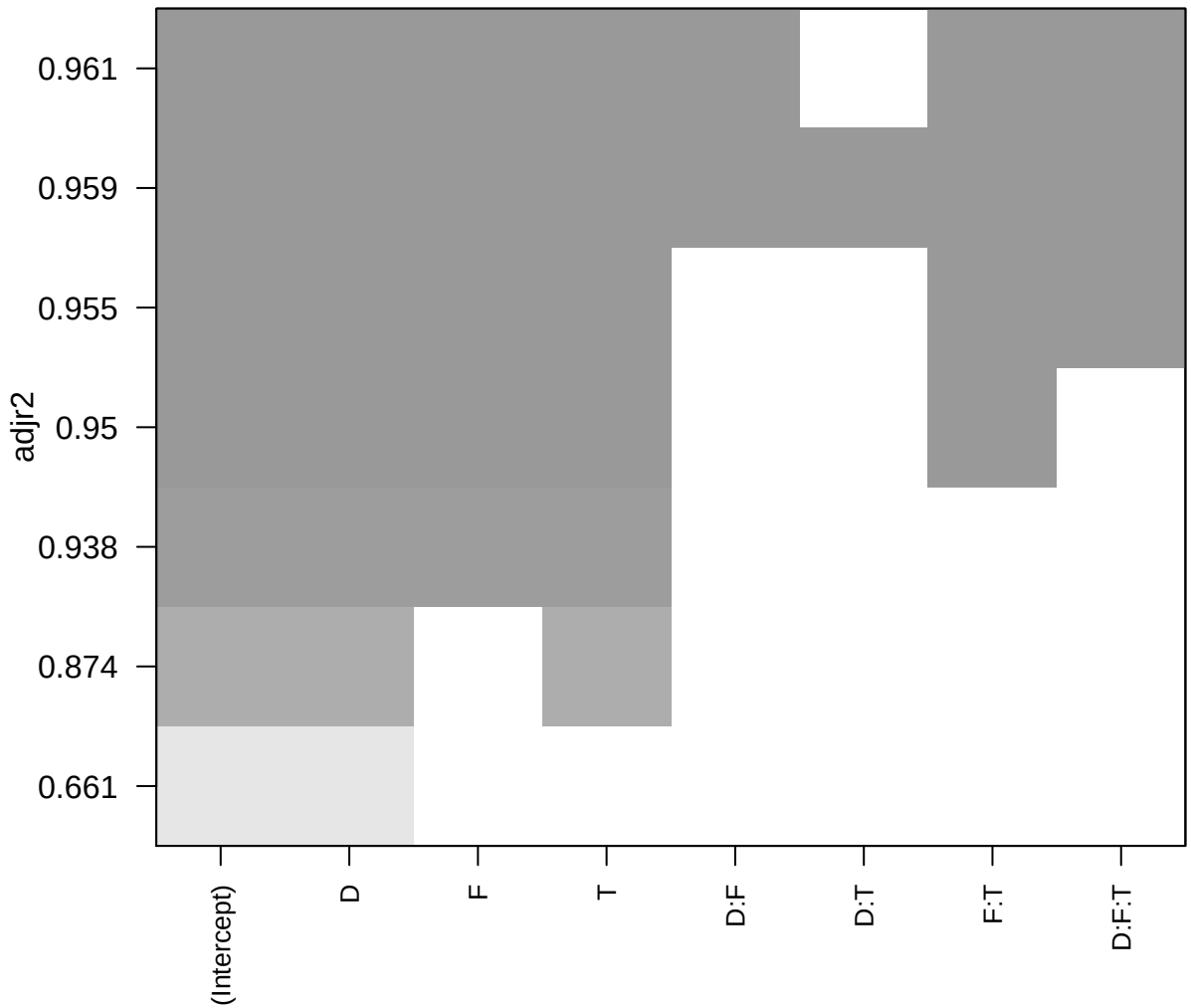
V2016 Plant stress 2d: model selection (missing part with lasso not on reading list)

The researchers want to use the data to fit a prediction model, and want to consider reduced versions of the full model, using best subset model selection.

Explain briefly what is done in the best subset model selection, and choose a good model based on the R^2_{adj}-criterion. Write down the fitted regression model for the model you choose.

```
> x <- model.matrix(fit)[,-1]; dim(x)
[1] 32  7
> y <- ds$Y
> library(leaps)
> bests <- regsubsets(x,y)
> sumbests=summary(bests)
> sumbests
1 subsets of each size up to 7
Selection Algorithm: exhaustive
         D   F   T   D:F D:T F:T D:F:T
1  ( 1 ) "*" " " " " " " " " " " " "
2  ( 1 ) "*" " " "*" " " " " " " " "
3  ( 1 ) "*" "*" "*" " " " " " " " "
4  ( 1 ) "*" "*" "*" " " " " "*" " "
5  ( 1 ) "*" "*" "*" " " " " "*" "*"
6  ( 1 ) "*" "*" "*" "*" " " "*" "*"
7  ( 1 ) "*" "*" "*" "*" "*" "*" "*"
> plot(bests,scale="adjr2",col=gray(seq(0.6,0.9,length=20)))
> round(sumbests$adjr2,3)
[1] 0.661 0.874 0.938 0.950 0.955 0.961 0.959
```

V2016: Plant stress 2d

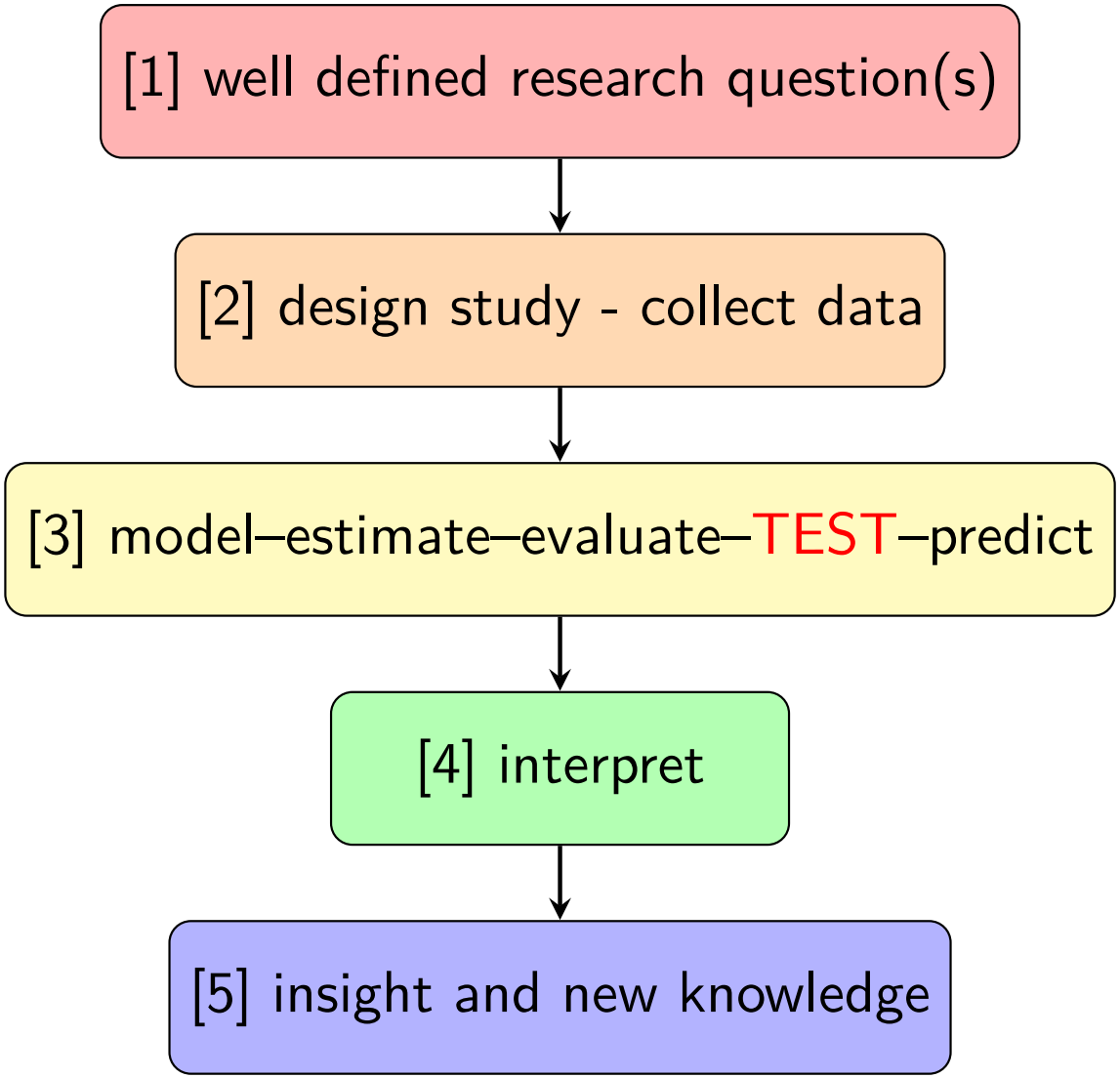


Part 3: Hypothesis testing and analysis of variance [L13-L16, CompEx3, RecEx5]

Curriculum:

- Fahrmeir, Kneib, Lang and Marx (2013): Regression, Springer.
  - Chapter 3.3
  - Appendix B
- Härdle, Simar (2015): Applied Multivariate Statistical Analysis, Fourth edition, Springer.
  - Chapter 8.1.1.
- Note: Multiple testing by Halle, Bakke and Langaas.
- Slides and handouts: `https://www.math.ntnu.no/emner/TMA4267/2017v/TMA4267V2017Part3.pdf`

The scientific process



Single hypothesis testing

Null- and alternative hypothesis:

$$H_0 : \beta_j = 0 \quad \text{vs.} \quad H_1 : \beta_j \neq 0.$$

Two types of errors:

	Not reject H_0	Reject H_0
H_0 true	Correct	Type I error
H_0 false	Type II 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 p-value

- A *p-value* $p(X)$ is a test statistic satisfying $0 \leq p(\boldsymbol{Y}) \leq 1$ for every vector of observations $\boldsymbol{Y}$.
- Small values give evidence that H_1 is true.
- In single hypothesis testing, if the p-value is less than the chosen significance level (chosen upper limit for the probability of committing a type I error), then we reject the null hypothesis, H_0. The chosen significance level is often referred to as α.
- A p-value is *valid* if

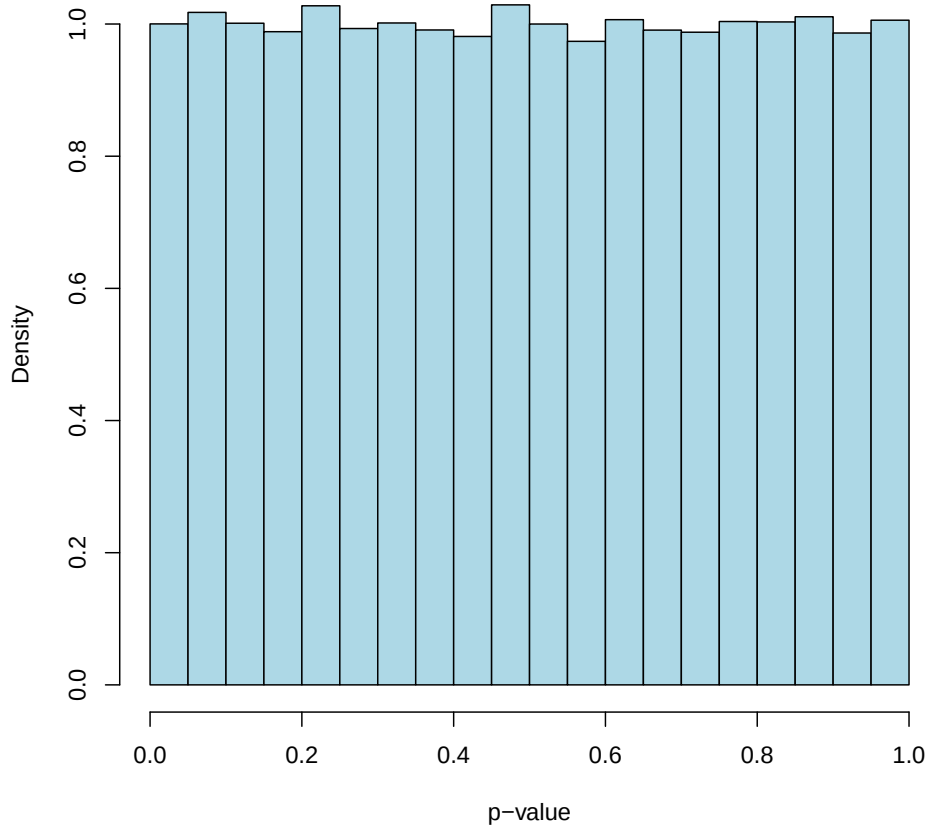
$$P(p(\boldsymbol{Y}) \leq \alpha) \leq \alpha$$

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.

- 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.

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

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.

ASA Statement on Statistical Significance and *P*-values

- P3: Scientific conclusions and business or policy decisions should not be based only on whether at *p*-value passes a specific threshold.
- 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.

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.

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}.
- Most popular method: Bonferroni - valid for all dependency structures - and $\alpha_{\text{LOC}} = \alpha/m$.

Overall Type I error control (2)

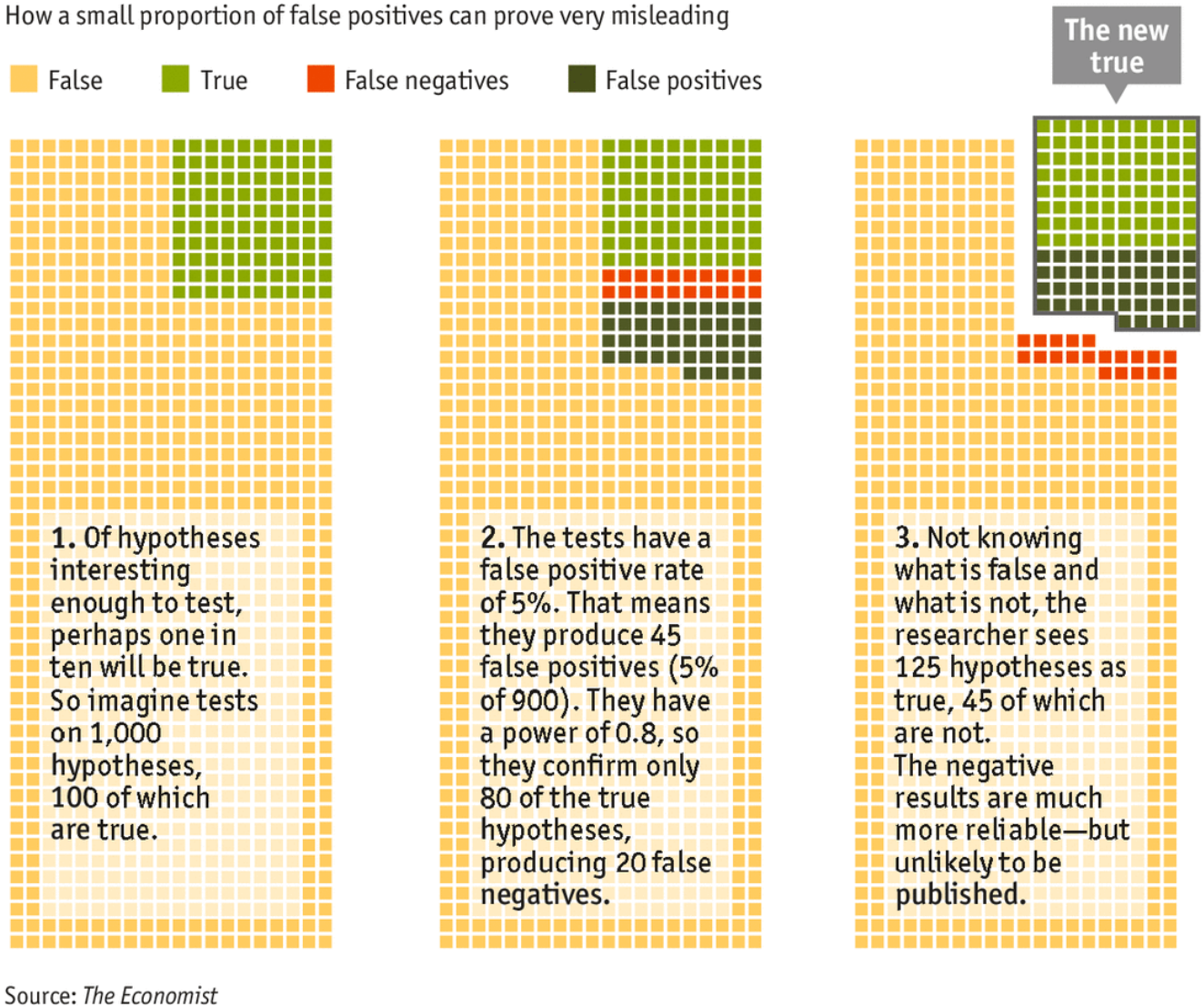
- For other types of data one expects that many null hypotheses are false,
- and therefore a less strict criterion for controlling an overall version of the Type I error is chosen.
- The False Discovery Rate (FDR) by Benjamini & Hochberg (1995) is controlled at level α.
- Informally, the FDR is the expected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

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

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

- Popular algorithm for controlling the FDR: the Benjamini-Hochberg step-up procedure.

What is the proportion of fake news?



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.

Exam questions?

Reproducible research and multiple testing is new on the reading list this year.
Questions on exercises: CompEx4.P2 and RecEx5.P2.

Testing linear hypotheses in regression

We study a normal linear regression model with $p = k + 1$ covariates, and refer to this as model A (the larger model). We then want to investigate the null and alternative hypotheses of the following type(s):

$$
\begin{aligned}
H_0 : \beta_j &= 0 \text{ vs. } H_1 : \beta_j \neq 0 \\
H_0 : \beta_1 &= \beta_2 = \beta_3 = 0 \text{ vs. } H_1 : \text{ at least one of theses } \neq 0 \\
H_0 : \beta_1 &= \beta_2 = \cdots = \beta_k = 0 \text{ vs. } H_1 : \text{ at least one of these } \neq 0
\end{aligned}
$$

We call the restricted model (when the null hypotesis is true) model B, or the smaller model.
These null hypotheses and alternative hypotheses can all be rewritten as a linear hypotheses

$$H_0 : \boldsymbol{C}\beta = \boldsymbol{d} \text{ vs. } \boldsymbol{C}\beta \neq \boldsymbol{d}$$

by specifying $\boldsymbol{C}$ to be a $r \times p$ matrix and $\boldsymbol{d}$ to be a column vector of length p.

Testing linear hypotheses in regression

The test statistic for performing the test is called F_{obs} and can be formulated in two ways:

$$F_{obs} = \frac{\frac{1}{r}(SSE_{H_0} - SSE)}{\frac{SSE}{n-p}} \quad (1)$$

$$F_{obs} = \frac{1}{r}(\boldsymbol{C}\hat{\beta} - \boldsymbol{d})^{\mathsf{T}}[\hat{\sigma}^2\boldsymbol{C}(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X})^{-1}\boldsymbol{C}^{\mathsf{T}}]^{-1}(\boldsymbol{C}\hat{\beta} - \boldsymbol{d}) \quad (2)$$

where SSE is from the larger model A, SSE_{H_0} from the smaller model A, and $\hat{\beta}$ and $\hat{\sigma}^2$ are estimators from the larger model A.

V2016: Plant stress 2c - with the full model

The researchers want to test the hypothesis

H_0: $\beta_{D:T} = \beta_{F:T} = \beta_{D:F:T} = 0$ vs.

H_1: at least one of $\beta_{D:T}$, $\beta_{F:T}$, $\beta_{D:F:T}$ is different from 0.

Perform the hypothesis test at a significance level of your own choice. All the numerical values you need for the calculations are found in the R-printout.

V2016 Problem 2a: Plant stress

In the print-out from `summary(fit)` *four* numerical values are replaced by question marks. Calculate numerical values for each of these, and explain what each of the values means.

```
> summary(fit)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 16.15942    0.04140  ?       < 2e-16
D            0.93739    0.04140  22.644  < 2e-16
F            0.28546    0.04140   6.896 3.93e-07
T           -0.52354    0.04140 -12.647 4.18e-12
D:F         -0.08878    0.04140  -2.145  0.04231
D:T         -0.00242    ?        -0.058  0.95386
F:T         -0.12614    0.04140  -3.047  0.00555
D:F:T        0.09099    0.04140   2.198  ?
Residual standard error: 0.2342 on 24 degrees of freedom
Multiple R-squared:      ?, Adjusted R-squared:  0.9594
F-statistic: 105.6 on 7 and 24 DF,  p-value: < 2.2e-16
```

V2016: 2d - full vs. reduced model

```
> fitRED=lm(Y~D+F+T+D:F,data=ds)
> summary(fitRED)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 16.15942    0.04919 328.528  < 2e-16
D            0.93739    0.04919  19.057  < 2e-16
F            0.28546    0.04919   5.804 3.56e-06
T           -0.52354    0.04919 -10.644 3.66e-11
D:F         -0.08878    0.04919  -1.805   0.0822
Residual standard error: 0.2782 on 27 degrees of freedom
Multiple R-squared:   0.95,Adjusted R-squared:  0.9426
F-statistic: 128.4 on 4 and 27 DF,  p-value: < 2.2e-16
> qt(0.025,32,lower.tail=FALSE)
[1] 2.036933
> qt(0.025,27,lower.tail=FALSE)
[1] 2.051831
> qt(0.025,24,lower.tail=FALSE)
[1] 2.063899
```

Grading V2016

Results for 53 students:

Grade	A	B	C	D	E	F
Frequency %	28	28	23	13	2	6

Item	1a	2a	2b	2c	2d	2e	3a	3b
Average score	9.25	8.47	7.42	4.37	5.97	4.68	8.82	6.18

One-way ANOVA

Classical formulation - one factor with k levels

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \quad \varepsilon_{ij} \text{ i.i.d } N(0, \sigma^2)$$

for $i = 1, \ldots, k$, $j = 1, \ldots, n_i$. with hypothesis of interest

$$\alpha_1 = \alpha_2 = \cdots = \alpha_k = 0 \text{ vs. } H_1 : \text{at least one different from 0}$$

Can be solved by fitting a linear regression model

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_{k-1} x_{k-1,i} + \varepsilon_i, \quad \varepsilon_i \text{ i.i.d } N(0, \sigma^2)$$

where $\beta_k = -\beta_1 - \beta_2 - \cdots - \beta_{k-1}$ (effect coding, aka sum-zero-constraint) and testing a linear hypothesis $\boldsymbol{C}\beta = \boldsymbol{d}$ where (if $k = 5$)

$$\boldsymbol{C} = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}, \quad \boldsymbol{d} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}.$$

Two-way ANOVA

Classical formulation - two factors with r and s levels, and interactions thereof

$$Y_{ijk} = \mu + \alpha_i + \gamma_j + (\alpha\gamma)_{ij} + \varepsilon_{ijk}, \quad \varepsilon_{ijk} \text{ i.i.d } N(0, \sigma^2)$$

for $i = 1, \ldots, r$, $j = 1, \ldots, s$, $k = 1, \ldots, n_{ij}$. with hypotheses of interest

$$\alpha_1 = \alpha_2 = \cdots = \alpha_r = 0 \text{ vs. } H_1 : \text{at least one different from 0}$$

$$\gamma_1 = \gamma_2 = \cdots = \gamma_s = 0 \text{ vs. } H_1 : \text{at least one different from 0}$$

$$(\alpha\gamma)_{11} = (\alpha\gamma)_{12} = \cdots = (\alpha\gamma)_{rs} = 0 \text{ vs. } H_1 : \text{at least one different 0}$$

Can be solved by fitting a linear regression model - with effect coding for main and interaction effects and then testing linear hypotheses.
Involves a lot more notation.

Exam questions?

ANOVA is on

- RecEx5.P4 and P5, and on
- exams K2013.P2, V2013.P2, V2012.P2 (all rather technical).

(V=spring, K=continuation, P=Problem)

Part 4: Design of experiments [L17-L20, CompEx4, RecEx6]

Curriculum:

- Note on Design of experiments by Tyssedal (not p 19: partial confounding and p28 fold-over).
- Slides and handouts: `https://www.math.ntnu.no/emner/TMA4267/2017v/TMA4267V2017Part4.pdf`

Design of experiments (DOE) terminology and results

- Move from observational studies to designed experiments (we decide on the design matrix).
- We still work with linear regression, but now with a specific special case.
- Covariates are called factors, and denoted A, B, C, ...
- We will only look at factors with two levels:
  - high, coded as $+1$ or just $+$, and,
  - low, coded as -1 or just $-$.
- In a full factorial experiment 2^k all possible combinations of the factors are performed, and
- the design matrix is coded so that the columns (containing -1 and 1 using *effect coding* aka sum-zero-contraint) are orthogonal, and thus
  - $\sum_{i=1}^{n} x_{ij}x_{ik} = 0$ for all combinations of the columns of the design matrix $\boldsymbol{X}$.
  - $\sum_{i=1}^{n} x_{ij}^2 = n$.

Full factorial 2^k designs

The orthogonal columns lead to that the following formulas are easy to interpret and calculate:

- $\boldsymbol{X}^T\boldsymbol{X}$ = diagonal matrix with n on the diagonal.
- $\hat{\beta}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij} Y_i$.
- $\mathrm{Var}(\hat{\beta}_j) = \frac{\sigma^2}{n}$.
- $\mathrm{Cov}(\hat{\beta}_j, \hat{\beta}_k) = 0$ for all $j \neq k$.
- SSR$=\sum_{j=1}^{p-1} n\hat{\beta}_j^2$.

Topic: choosing the "best" linear regression model!

- First, debunk popular strategies (based on simulations studies were we knew the "true" model):
  - Popular 1: fit all available covariates.
    Problem: overfitting (=fitting trends and noise).
  - Popular 2: fit all available covariates, then remove the insignificant ones (=those β_j where $H_0 : \beta_j = 0$ is rejected).
    Problem: may also remove important covariates that are correlated with unimportant ones - but insignificant because being masked by the unimportant ones.
- Study of irrelevant and missing covariates:

  Irrelevant : variables that are included in the regression but should not have been (IQ of lumberjack)

  missing : variables that are not included, but should have been (omitting height in the tree volum example)

  Conclusion in book: the model should not contain irrelevant covariates, and we should aim for a sparse model. Take home message is the "Law of parsimony": *If two models are not very different – then always choose the simplest one.*

V2016: Problem 1e - full vs. reduced model

The researchers choose to use the following *reduced model* for prediction:

$$Y_i = \beta_0 + \beta_D D_i + \beta_F F_i + \beta_T T_i + \beta_{D:F} D_i F_i + \varepsilon_i,$$

where $i = 1, \ldots, 32$, and we assume ε_i independent and identically normally distributed with mean 0 and variance σ^2.

Compare the estimated regression parameters and the estimated standard deviations of the estimated regression parameters for the full model and the reduced model, and explain what you observe.
Based on the reduced model, provide a prediction and a 95% prediction interval for the gene activity level for the factor combination $D = 1$, $F = 1$, $T = -1$.

Hint: In a multiple linear regression with $n \times p$ design matrix $\boldsymbol{X}$, estimated regression coefficients $\hat{\beta}$ and unbiased estimated error variance s^2, a $(1 - \alpha)100\%$ prediction interval at $\boldsymbol{x}_0$ is given as

$$\boldsymbol{x}_0^T \hat{\beta} \pm t_{\frac{\alpha}{2}, n-p}\, s \sqrt{1 + \boldsymbol{x}_0^T (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{x}_0},$$

where $t_{\alpha/2, n-p}$ denotes the value in the t-distribution with $n - p$ degrees of freedom that has area $\frac{\alpha}{2}$ to the right.

V2016: 2d - full vs. reduced model

```
> fitRED=lm(Y~D+F+T+D:F,data=ds)
> summary(fitRED)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 16.15942    0.04919 328.528  < 2e-16
D            0.93739    0.04919  19.057  < 2e-16
F            0.28546    0.04919   5.804 3.56e-06
T           -0.52354    0.04919 -10.644 3.66e-11
D:F         -0.08878    0.04919  -1.805   0.0822
Residual standard error: 0.2782 on 27 degrees of freedom
Multiple R-squared:   0.95,Adjusted R-squared:  0.9426
F-statistic: 128.4 on 4 and 27 DF,  p-value: < 2.2e-16
> qt(0.025,32,lower.tail=FALSE)
[1] 2.036933
> qt(0.025,27,lower.tail=FALSE)
[1] 2.051831
> qt(0.025,24,lower.tail=FALSE)
[1] 2.063899
```

From coefficients to effects

For each β_j in the model (except for the intercept) we define an effect to be

$$Effect_j = 2\beta_j$$

because β_j gives the change in the response when x_j increases by one, and we want to get the change when x_j goes from -1 to 1.

This implies that $\widehat{Effect}_j = 2\hat{\beta}_j$.

Lima beans example: full 2^3 factorial design

- A: depth of planting (0.5 inch or 1.5 inch)
- B: watering daily (once or twice)
- C: type of lima bean (baby or large)
- Y: yield

A	B	C	AB	AC	BC	ABC	Level code	Response
-	-	-	+	+	+	-	1	6
+	-	-	-	-	+	+	a	4
-	+	-	-	+	-	+	b	10
+	+	-	+	-	-	-	ab	7
-	-	+	+	-	-	+	c	4
+	-	+	-	+	-	-	ac	3
-	+	+	-	-	+	-	bc	8
+	+	+	+	+	+	+	abc	5
x_1	x_2	x_3	x_{12}	x_{13}	x_{23}	x_{123}		y

Main effects in DOE

For full 2^3 design with 8 observations.

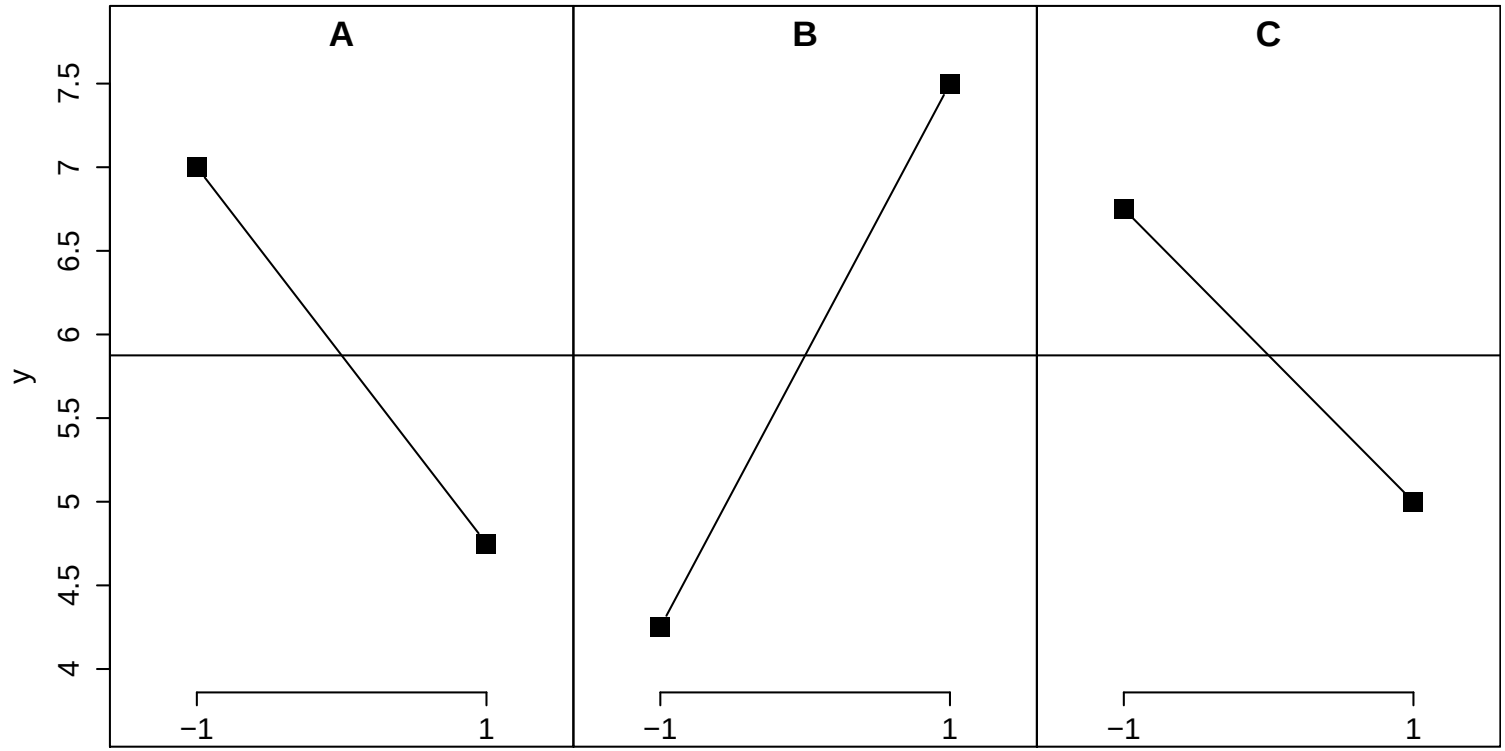
Main effect of A

$$\begin{aligned}\widehat{A} &= 2\hat{\beta}_1 \\ &= \frac{y_2 + y_4 + y_6 + y_8}{4} - \frac{y_1 + y_3 + y_5 + y_7}{4}\end{aligned}$$

Interpretation: mean response when A is high MINUS mean response when A is low.
Similarily, main effect of B

$$\begin{aligned}\widehat{B} &= 2\hat{\beta}_2 \\ &= \frac{y_3 + y_4 + y_7 + y_8}{4} - \frac{y_1 + y_2 + y_5 + y_6}{4}\end{aligned}$$

Interpretation: mean response when B is high MINUS mean response when B is low.




```
  A      B      C       A:B    A:C      B:C    A:B:C
-2.25   3.25  -1.75   -0.75   0.25    -0.25   -0.25
```

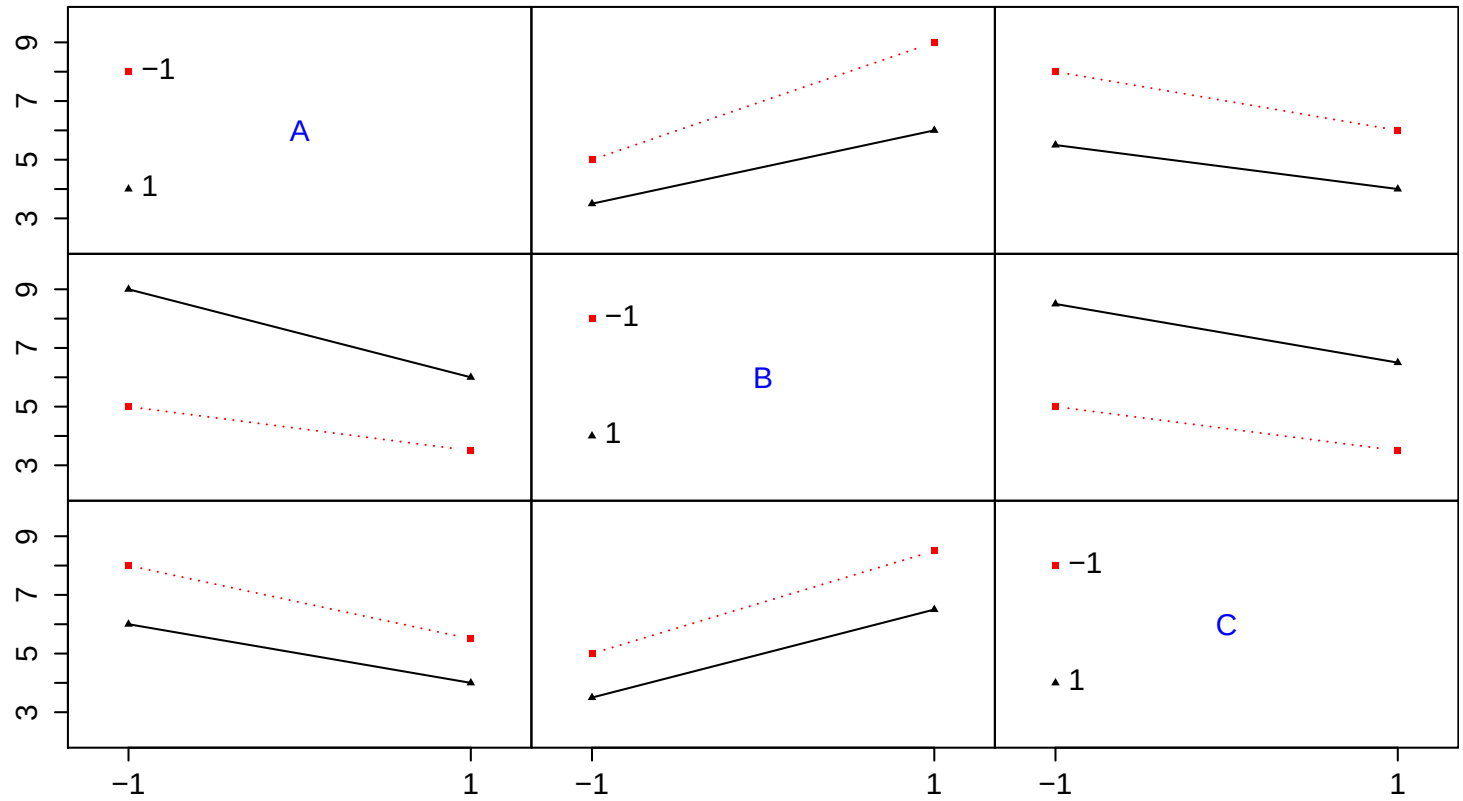
Explain the main effects in plain words!

A: depth (0.5 or 1), B: watering daily (once, twice), C: type (baby, large).

Interaction effect in DOE

- What is the terpretation in DOE associated with β_{12}?
- In DOE $2\hat{\beta}_{12}$ is denoted $\widehat{AB}$ and is called the *estimated interaction effect between A and B*.

$$
\begin{aligned}
\widehat{AB} &= 2\hat{\beta}_{12} \\
&= \frac{\text{estimated main effect of } A \text{ when } B \text{ is high}}{2} \\
&- \frac{\text{estimated main effect of } A \text{ when } B \text{ is low}}{2} \\
&= \frac{\text{estimated main effect of } B \text{ when } A \text{ is high}}{2} \\
&- \frac{\text{estimated main effect of } B \text{ when } A \text{ is low}}{2}
\end{aligned}
$$




```
  A       B       C         A:B     A:C       B:C     A:B:C
-2.25   3.25   -1.75    -0.75    0.25     -0.25    -0.25
```

See classnotes from L18 for explanation of figure!

Interpretation of $\widehat{ABC}$

- $\widehat{ABC} = \frac{1}{2}\widehat{AB}$ interaction when C is at the high level - $\frac{1}{2}\widehat{AB}$ interaction when C is at the low level.
- Or, two other possible interpretation with swapped placed for A, B and C.
- And remember that $\widehat{AB} = \frac{1}{2}\widehat{A}$ main effect when B is at the high level - $\frac{1}{2}\widehat{A}$ main effect when B is at the low level.

Estimation of σ^2

1. Perform replicates, estimate the full model and use s^2 from regression model.
2. Assuming specified higher order interactions are zero (changing the regression model).
3. ONLY if the two above is not possible: Lenth's Pseudo Standard Error (PSE).

Performing inference

Distribution of Effect:

$$\widehat{Effect}_j \sim N(Effect_j, \sigma^2_{Effect})$$

where $\sigma^2_{Effect} = \frac{4}{n}\sigma^2$.
(And, also $\sigma^2_{Effect} = 4\mathrm{Var}(\hat{\beta}_j)$, and $\sigma_{Effect} = 2\mathrm{SD}(\hat{\beta}_j)$.)

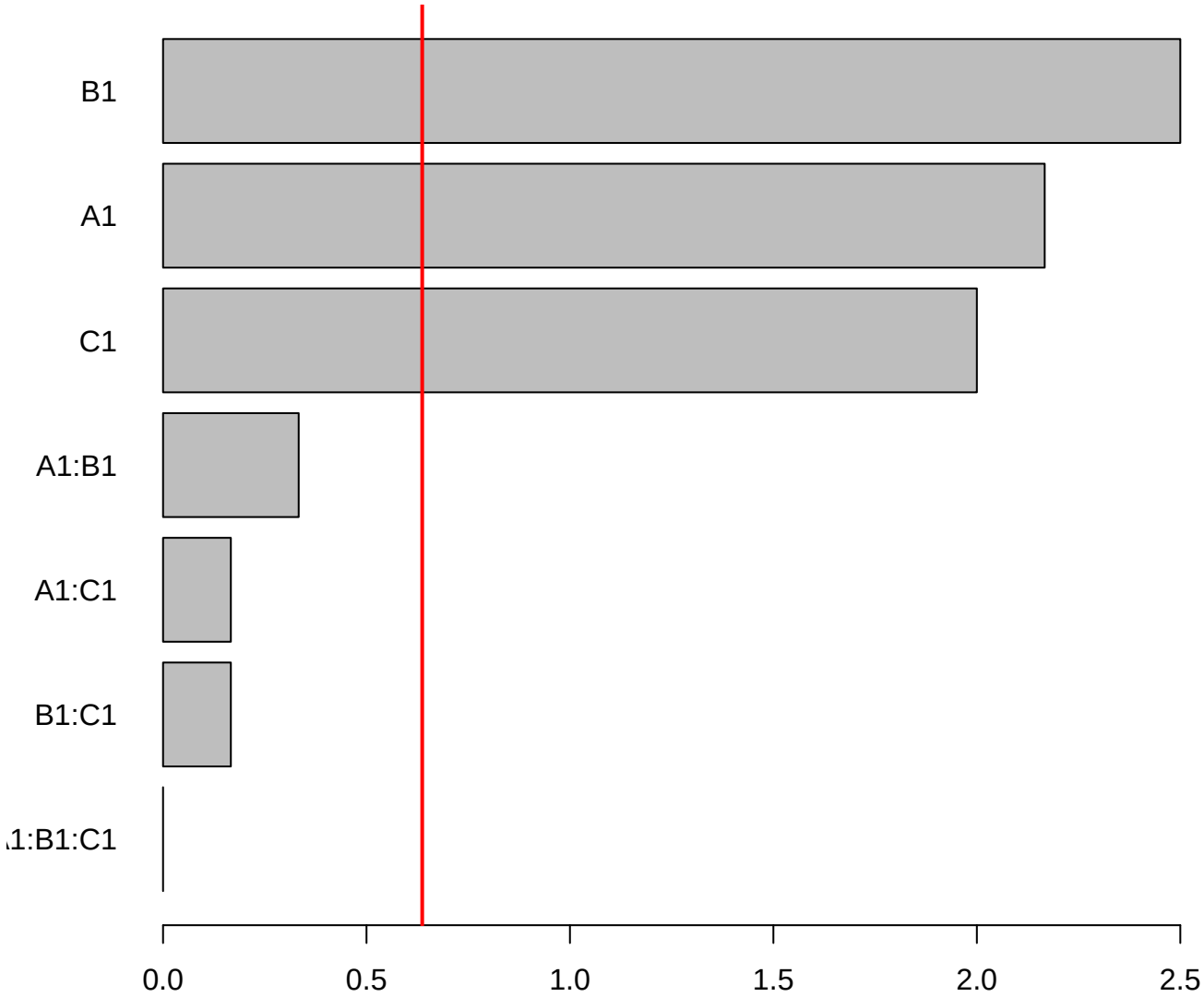
And, we know (previous page) how to estimate σ^2.

We then replace σ^2 by s^2 to insert into σ^2_{Effect} to make s^2_{Effect}.

This gives us s t-distributed statistic

$$T_j = \frac{\widehat{Effect}_j - Effect_j}{s_{Effect}} = \frac{2\hat{\beta}_j - 2\beta_j}{2\widehat{\mathrm{SD}}(\hat{\beta}_j)} = \frac{2\hat{\beta}_j - 2\beta_j}{2\sqrt{\frac{1}{n}s^2}}$$

with ν degrees of freedom - where ν=number of observations minus number of parameters estimated (except for Lenths method).



Red line at $t_{\alpha/2,\nu} s_{Effect} = t_{\alpha/2,\nu} \cdot 2 \cdot \widehat{\mathrm{SD}}(\hat{\beta}_j) = t_{\alpha/2,\nu} \cdot 2\sqrt{\frac{1}{n}} s.$

DOE workflow

1. Set up full factorial design with k factors in R, and
2. randomize the runs.
3. Perform experiments (genuine run replicates), and enter data into R.
4. Fit a full model (all interactions).
5. If you do not have replications, look at Pareto plots and, use this to suggest at reduced model (if possible). Refit the reduced model.
6. Assess model fit (residual plots, need transformations?).
7. Assess significance.
8. Interpret you results (main and interaction plots).

Blocking

- Why may experiments need to be performed in blocks? (Batches of raw material, performed on different days, different people performing the experiments.)
- Should we also add a "block" effect if we perform repeated experiments? (Sometimes. If done by different people, or external factors have changed.)
- Should then the block effect be a part of the regression model? (In most cases: yes!)

Fractional factorials

- Why don't we want to perform a full factorial experiment, but a instead a fractional factorial? (If we have many factors we maybe not need to be able to estimate all possible interactions, and may accept that effects are confounded.)
- What is the easiest way to design a half-fraction of a 2^k factorial experiment? (Perform all the experiments where the highest order interaction =-1 or +1. E.g. for k=4 we may do 16 different experiments, and now we only do the 8 possible experiments where ABCD=+1=defining relation. This is the same as thinking that D=ABC=generator).
- New words:
  - *generator(s)*=how to generate the design,
  - *defining relation(s)*, found from the generators,
  - *resolution*=length of shortest defining relation,
  - *alias structure*=confounding pattern, found by multiplying each effect of interest with the defining relation.

Exam question on fractional factorials (K2014)

In a pilot study with four factors A, B, C and D, the 8 experiments listed below were run.

	A	B	C	D
1	-1	-1	-1	1
2	1	-1	-1	-1
3	-1	1	-1	-1
4	1	1	-1	1
5	-1	-1	1	1
6	1	-1	1	-1
7	-1	1	1	-1
8	1	1	1	1

What type of experiment is this?
What is the generator and the defining relation for the experiment?
What is the resolution of the experiment?
Write down the alias structure of the experiment.

Exam

- 9.00-13.00, May 19, 2017.
- Written.
- Makes up 80% of the final grade, the remaining 20 % from the four compulsory exercises.
- Permitted aids: (Code C). One yellow A5 with own handwritten notes, Rottmann: Matematisk formelsamling, Tabeller og formler i statistikk, specified calculator.

Why one yellow A5 sheet?

- Force you to structure the course key concepts?
- Memorizing not needed?
- Security blanket.

Final grade in TMA4267

- 20% of final grade from the 4 compulsory exercises,
- and the remaining 80% on the 4hrs written exam.
- Written exam:
  - mostly focussed on the "knowledge learning outcome"
  - 8 "questions" each with maximum 10 points score
  - the plan is 3*Easy+3*Medium+2*Hard
  - the plan is 3 from Part 1, 3 from Part 2, 1 from Part 3 and 1 from Part 4.
- Remember that all answers must be *justified* and *correct notation and vocabulary* used to score well.
- The written exam must give at least 41% score (that is at least 32-33 out of 80 points) for a passing grade.

Final reading list

- Fairmeir et al (2013):
  Chapter 3 and Appendix B.
- Härdle et al (2015):
  Chapters 2, 3.3, 4.1-4.5, 5.1, 8.1.1 and 11.1-11.3.
- Multiple testing note by Halle, Bakke and Langaas.
- DOE-note by Tyssedal.
- BoxCox: from L12 in lecture notes/handouts (and on several exams).
- The 4 compulsory and 6 recommended exercises.

Comparison with reading list earlier years

Not on the reading list V2017, but on before:
Analysis of contingency tables.
The most complex parts of Design of experiments (folding, combining blocking and fractionating).
Random effects ANOVA.
More effort made earlier with quadratic forms for ANOVA (especially with idempotent centering matrices and sums-of-squares).
Hotelling T^2. Tukeys test.
Penalized regression: lasso and ridge (will be part of TMA4268 Statistical learning).

New(ish) on the reading list:
Replication crises, properties of p-values.
Multiple testing with FWER and FDR.
Testing of linear hypotheses (F:3.3) (new in 2016)
Effect coding in linear regression to see analysis of variance just as a special case of regression, and using the linear hypothesis F-test instead of much work with sums-of-squares (new in 2016)

Activities before the exam

- The exam is Friday, May 19, 9-13.
- Exam problems from earlier year is available from the course www-page (also outside Bb).
- Supervision - and you may sit and work - Smia (changed on May 3)
  - Before May 15 - just stop by the office of Jacob or Mette (better to stop by than sending email - difficult to give good answer on email).
  - Monday May 15: 12-14
  - Tuesday May 16: 10-12 (booked 10-14)
  - Thursday May 18: 10-12 (booked 10-14)
- I will try to monitor the discussion forum on Bb - you may start a thread!
- After the exam: tentative solutions posted
- and hopefully (if allowed) automatic feedback given together with exam grade.

Statistics courses

- Autumn semester
  - TMA4295 Statistical Inference
  - TMA4285 Time series
  - TMA4315 Generalized linear models
- Spring semester
  - TMA4250 Spatial statistics
  - TMA4268 Statistical learning
  - TMA4275 Survival analysis
  - TMA4300 Computational statistics

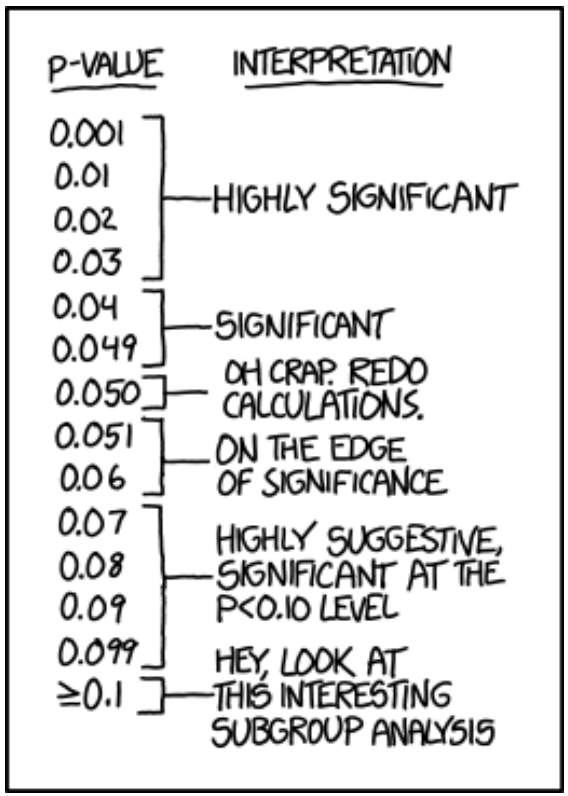
StudForsk

(Undergraduate research program in mathematics)

- Many new, and also some older, projects are available in StudForsk
- Completed projects are granted 10 000 NOK, but no credit points
- Can be started at any time (in agreement with supervisor)
- See homepages for more information and all available projects `https://wiki.math.ntnu.no/studforsk/start`
- Contact: Aslak Buan, Øyvind Bakke or Petter Bergh
- The project is financed by Olav Thon stiftelsen

Research ethics!

"If all else fails, use "significant at a p>0.05 level" and hope no one notices".



https://www.explainxkcd.com/wiki/index.php/1478:_P-Values and
https://www.explainxkcd.com/wiki/index.php/File:null_hypothesis.png

Future studies?

What is your current plan of topic for future studies?

- A: Statistics
- B: Mathematics
- C: Numerics
- D: Other
- E: Don't know

Use your smart phone, or other devise with internet access and go to **http://clicker.math.ntnu.no/**, and then select TMA4267 as classroom.