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

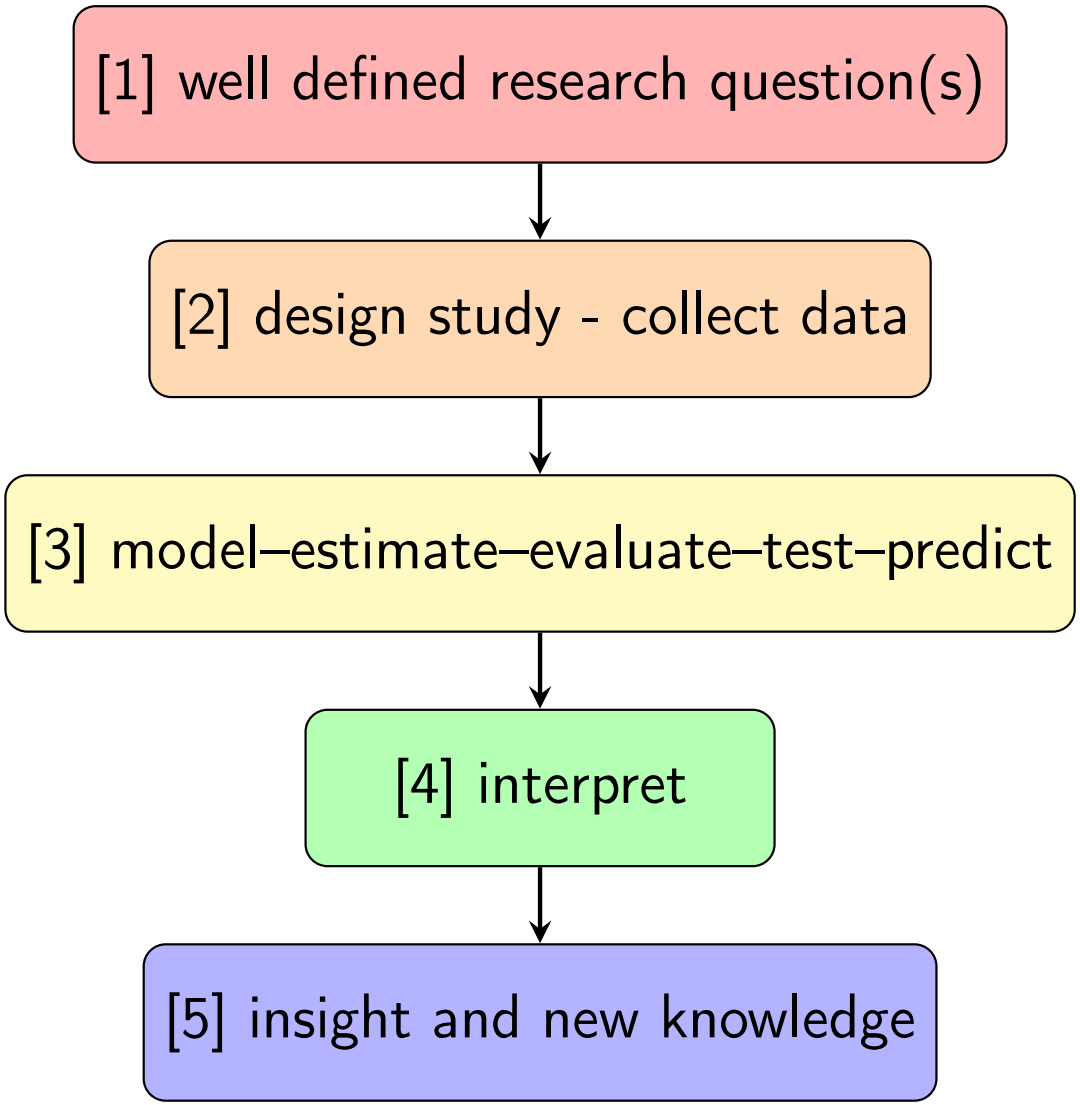
Today

- 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)

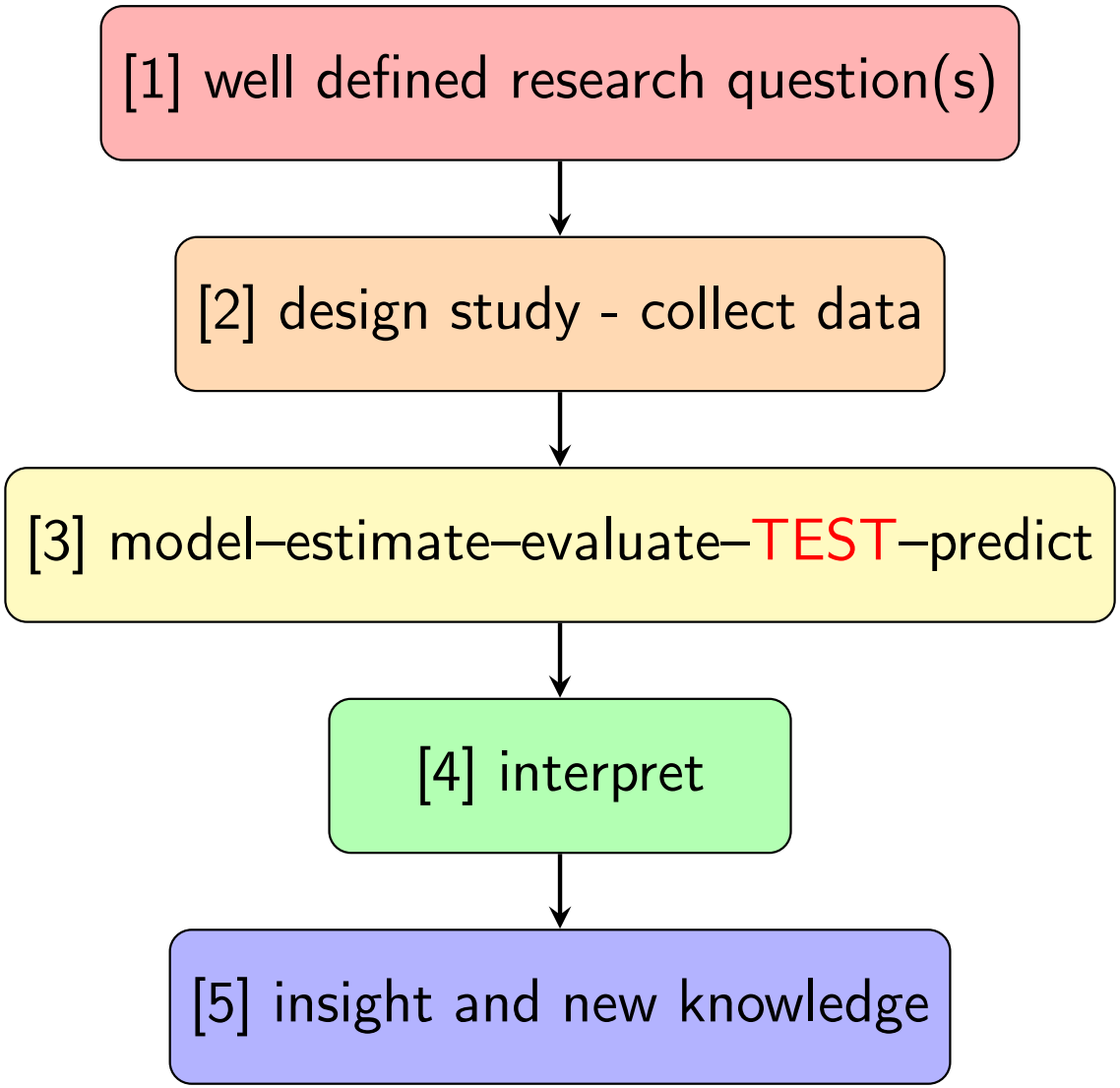
Basal metabolic rate and the FTO-gene

- 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.

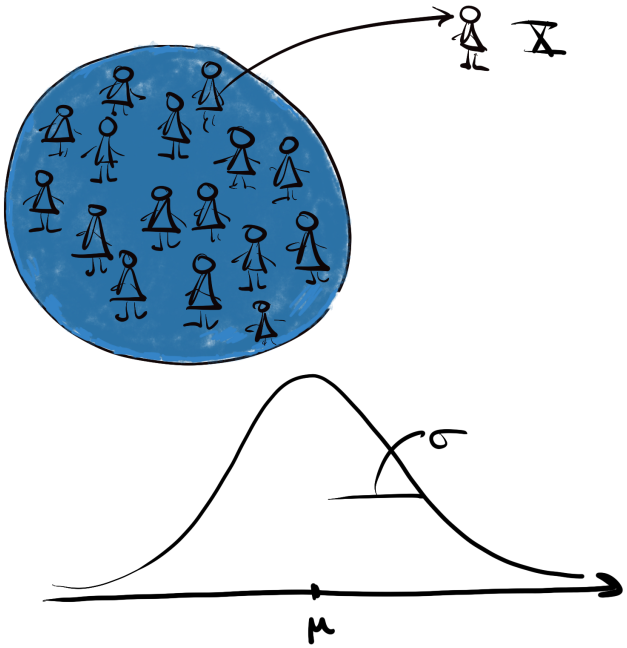
The scientific process



The scientific process

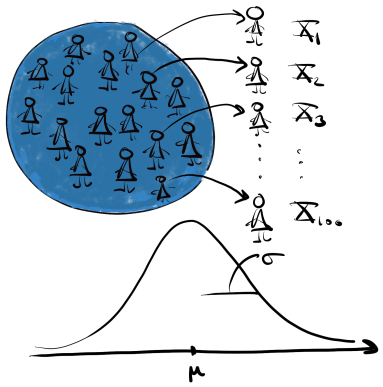


Hypothesis testing example



- 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: $X_1, X_2, \ldots, X_n$.
- 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.

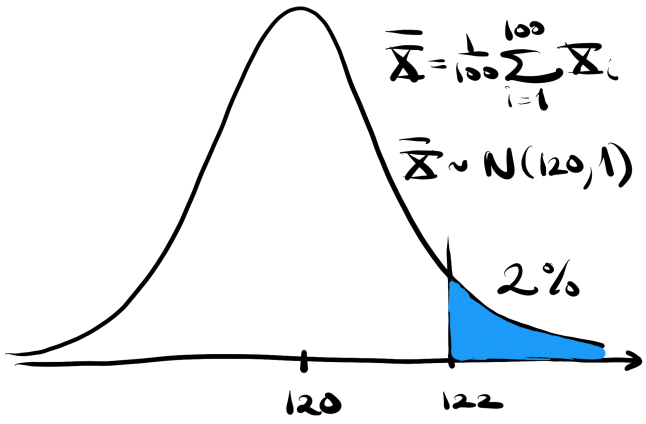
Questions:

- 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.

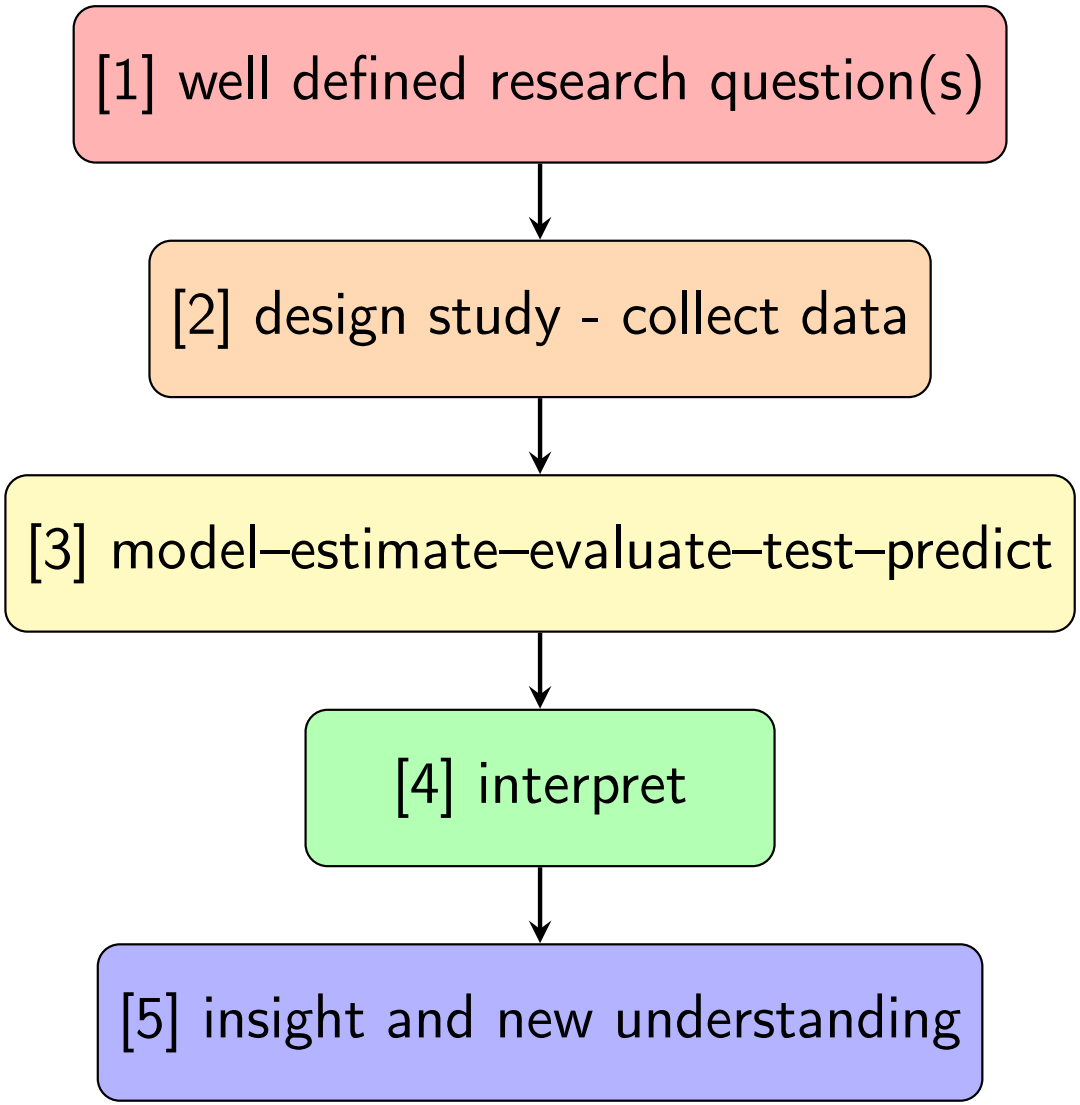
- While the *p*-value can be a useful statistical measure, it is commonly misused and misinterpreted.
- Informally, a *p*-value is the probability under a specified statistical model that a statistical summary of the data would be equal to or more extreme than its observed value.
- P1: *P*-values can 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

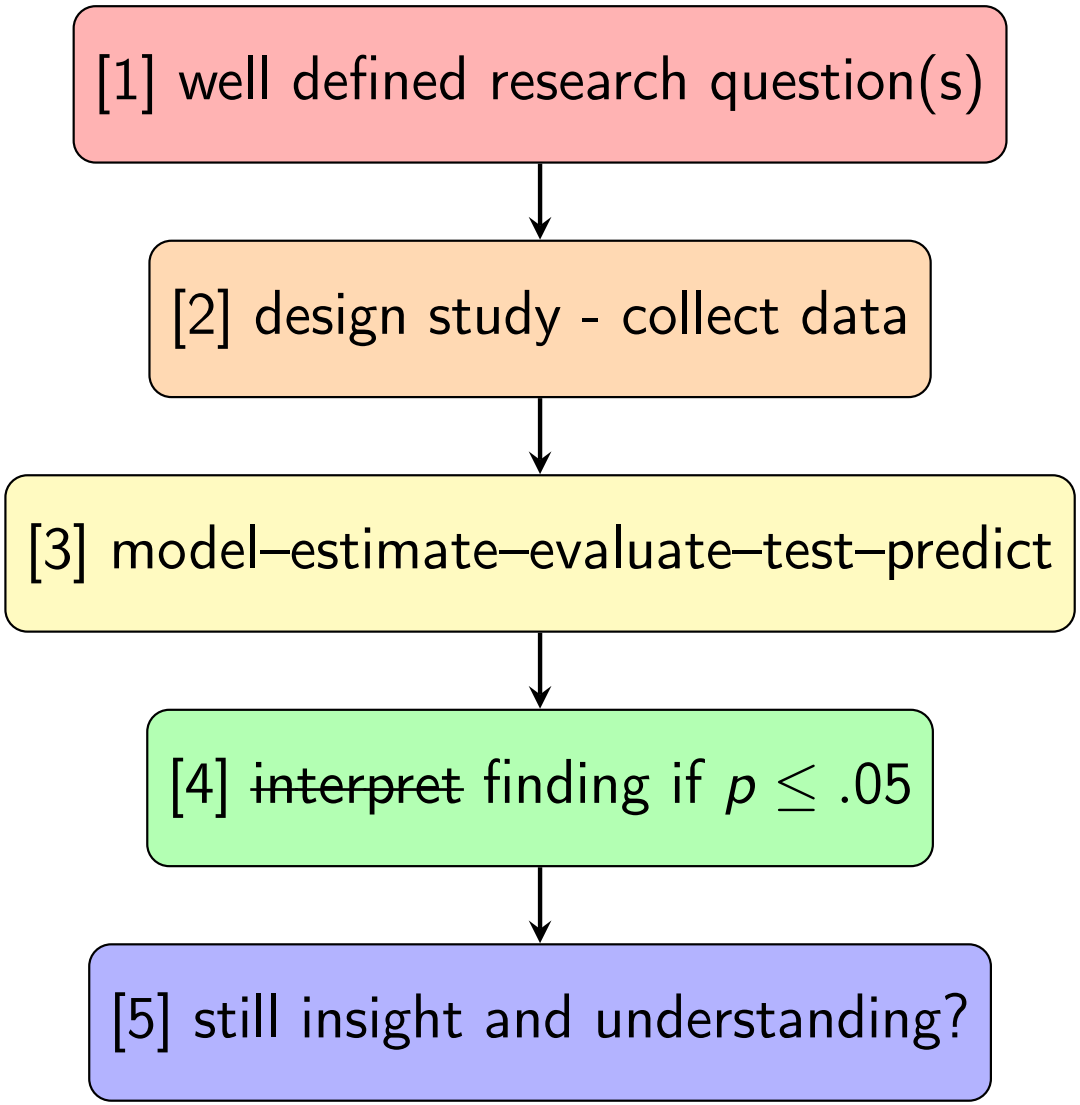
- P4: Proper inference requires full reporting and transparency.
- P5: A *p*-value, or statistical significance, does not measure the size of an effect or the importance of a result.
- P6: By itself, a *p*-value does not provide a good measure of evidence regarding a model or hypothesis.

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.

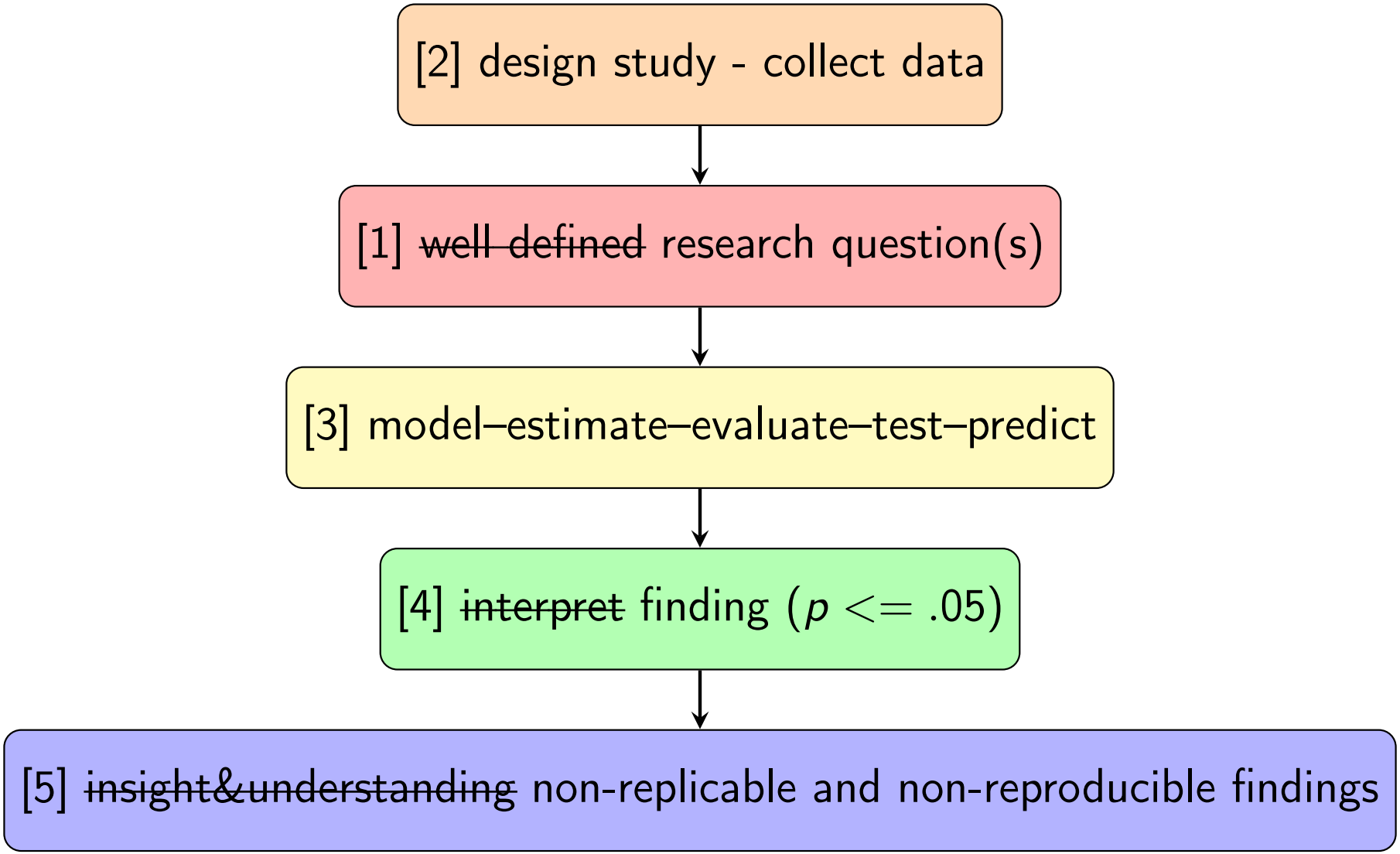
The scientific process

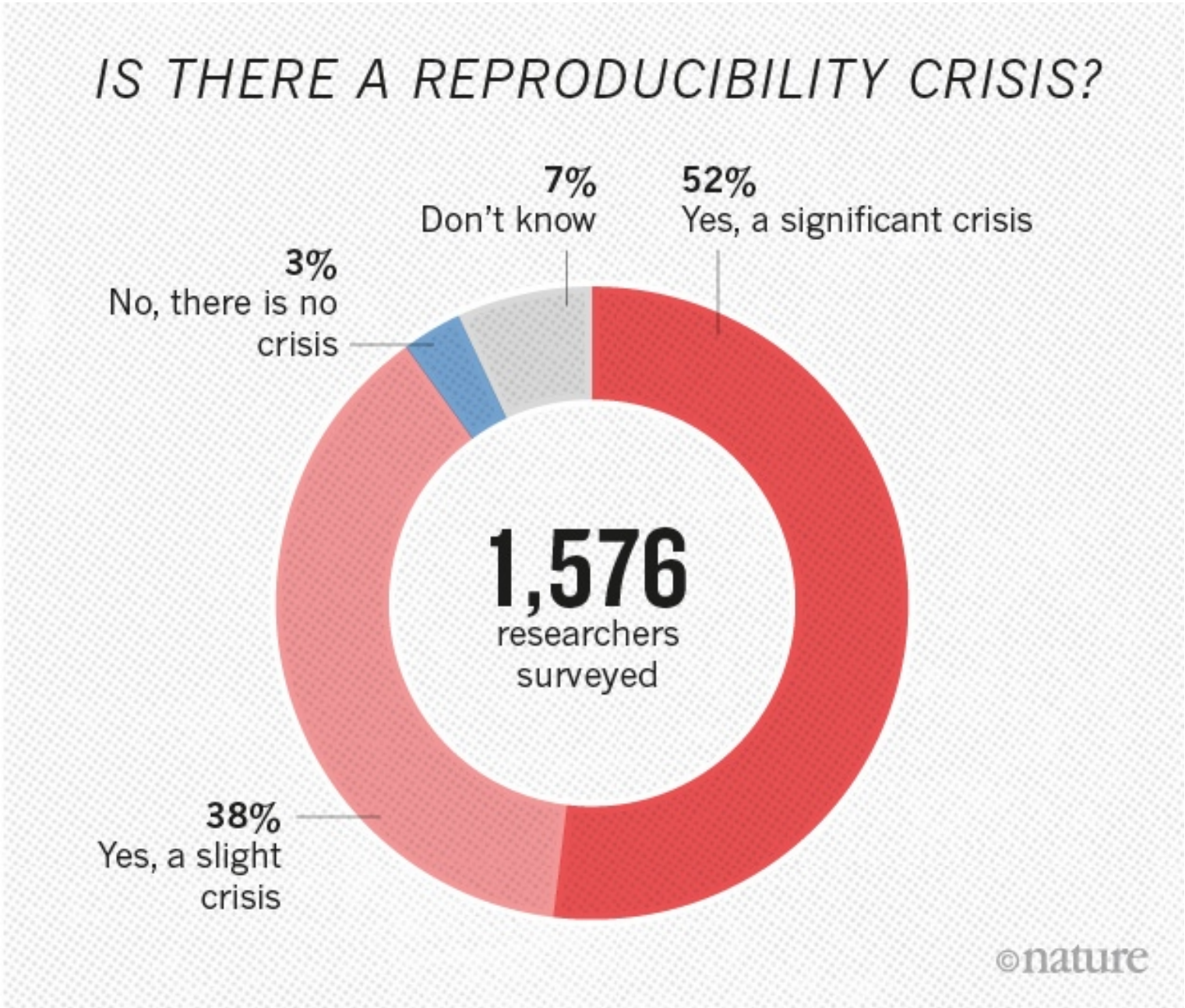


Scenario: finding only for $p \leq 0.05$

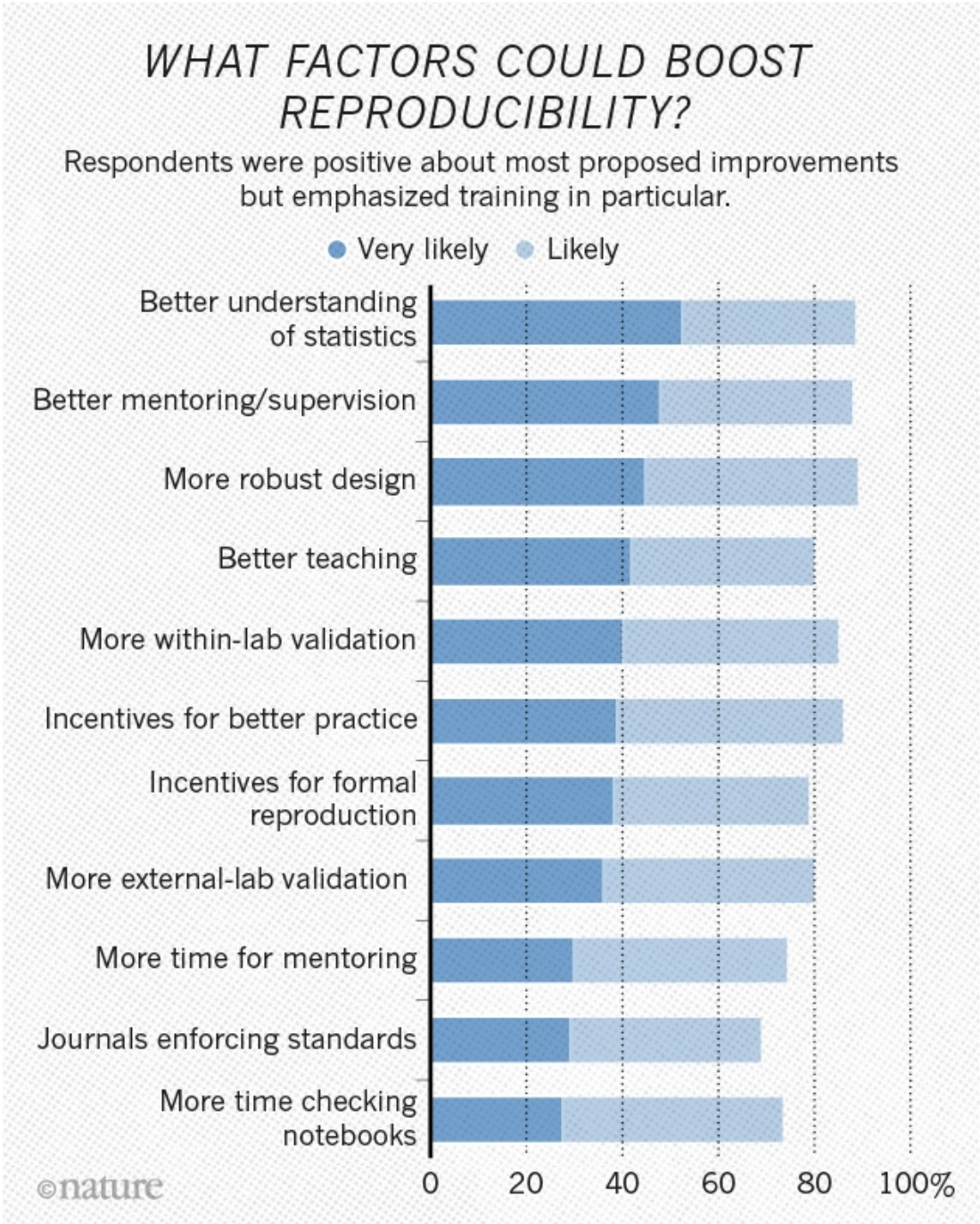


Scenario: Cherry-picking aka Selective Inference aka *p*-hacking



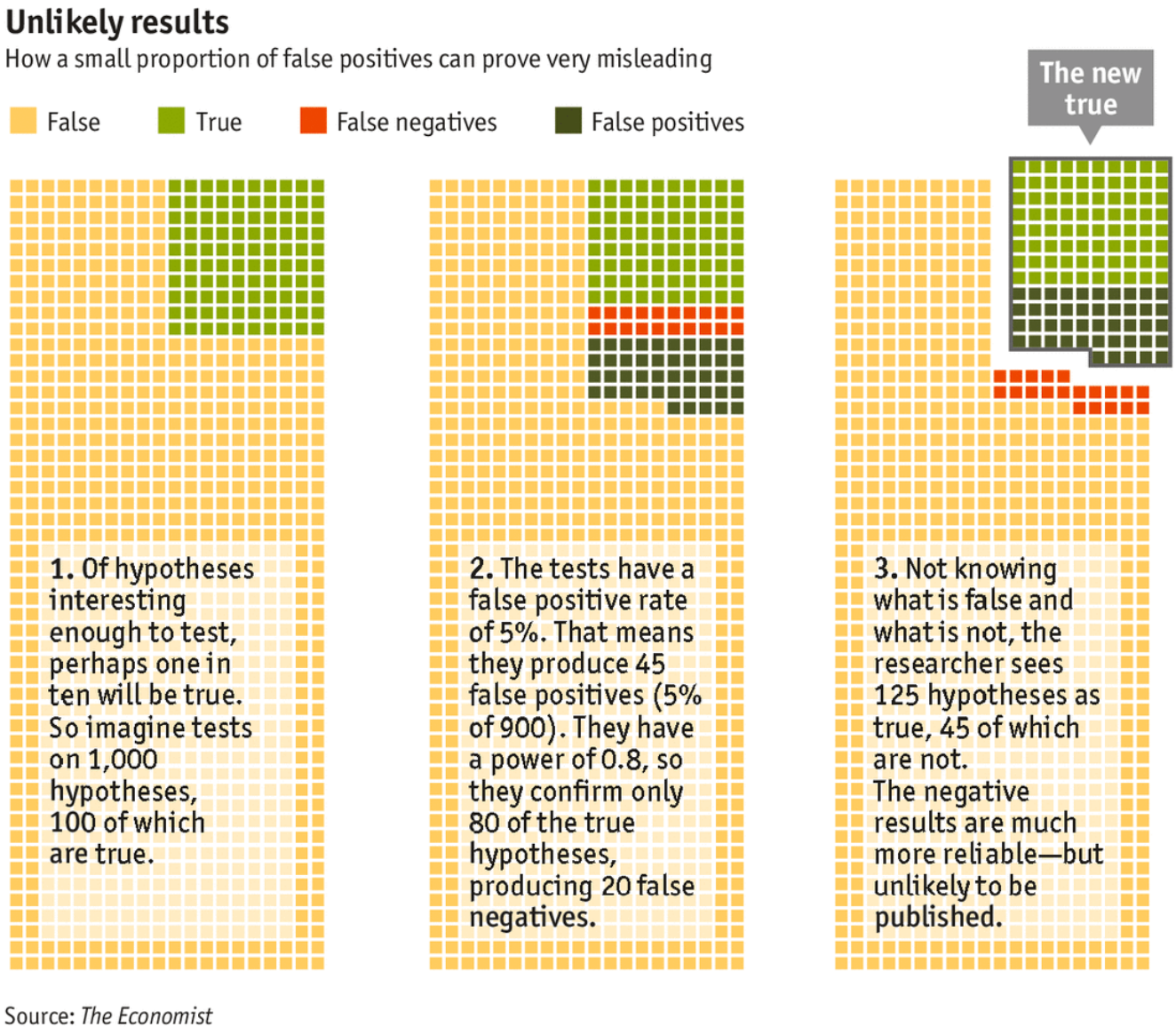


http://www.nature.com/news/1-500-scientists-lift-the-lid-on-reproducibility-1.19970



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

What is the proportion of fake news?



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

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

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

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

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

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

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

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

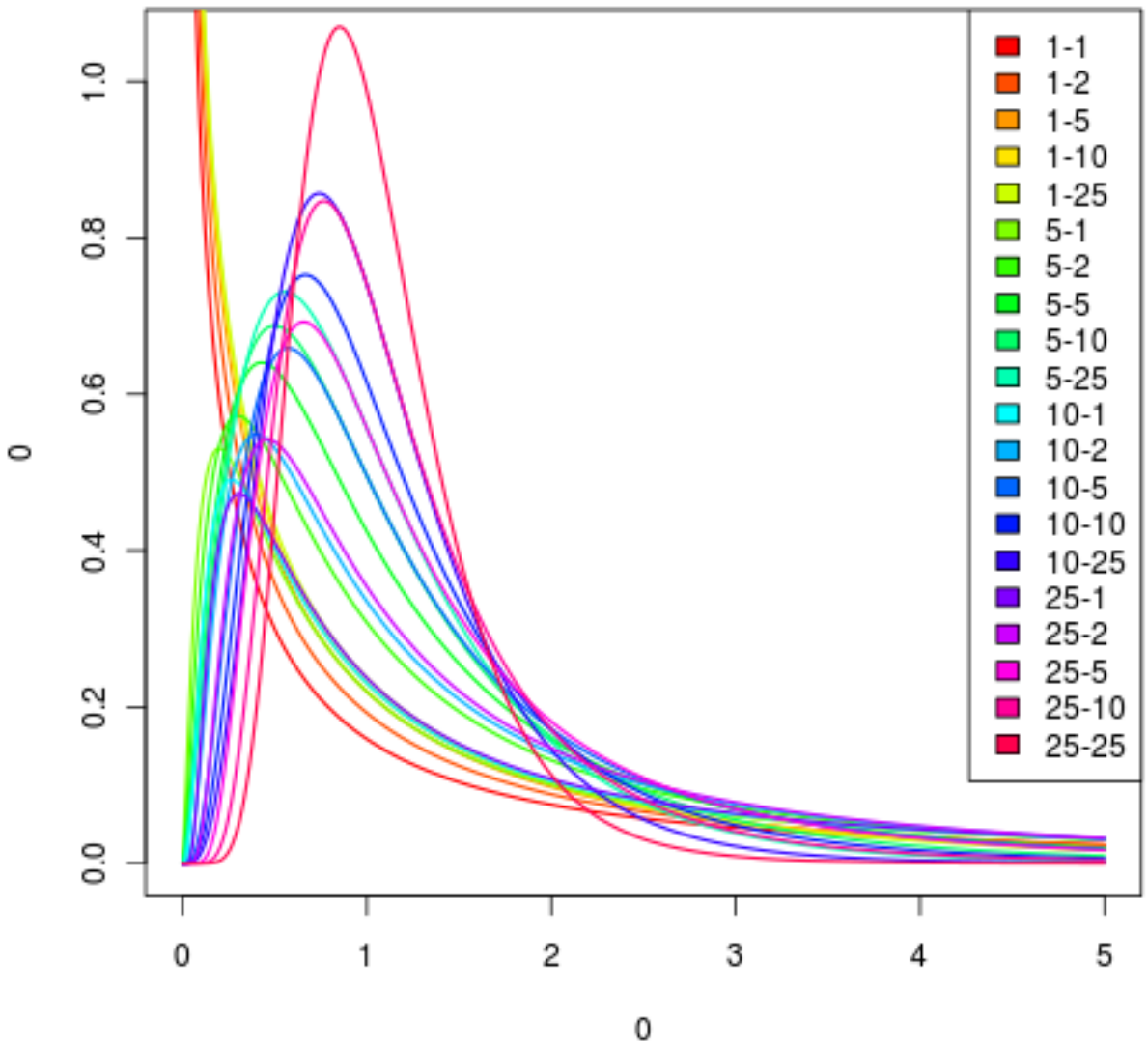
"Tabeller og formeler i statistikk":
If Z_1 and Z_2 are independent and χ^2-distributed with ν_1 and ν_2 degrees of freedom, then

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

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:

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

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.