Hypotesetesting, p-verdier, p-hacking, falske funn — og multippel testing

ST1201 Statistiske metoder 2018

Mette Langaas

Foreleses 9. oktober, med oppgaveregning 12.oktober

Introduksjon

Innen forskning setter vi opp en hypotese, vi samler inn data, analysere dataene, fortolker resultater og kommer frem til en konklusjon.

Ofte konstruerer vi et konfidensintervall, utfører en hypotesetest og beregner en p-verdi.

Vi bruker tid på å forstå hva en p-verdi sier oss og hva et falsk positivt resultat er.

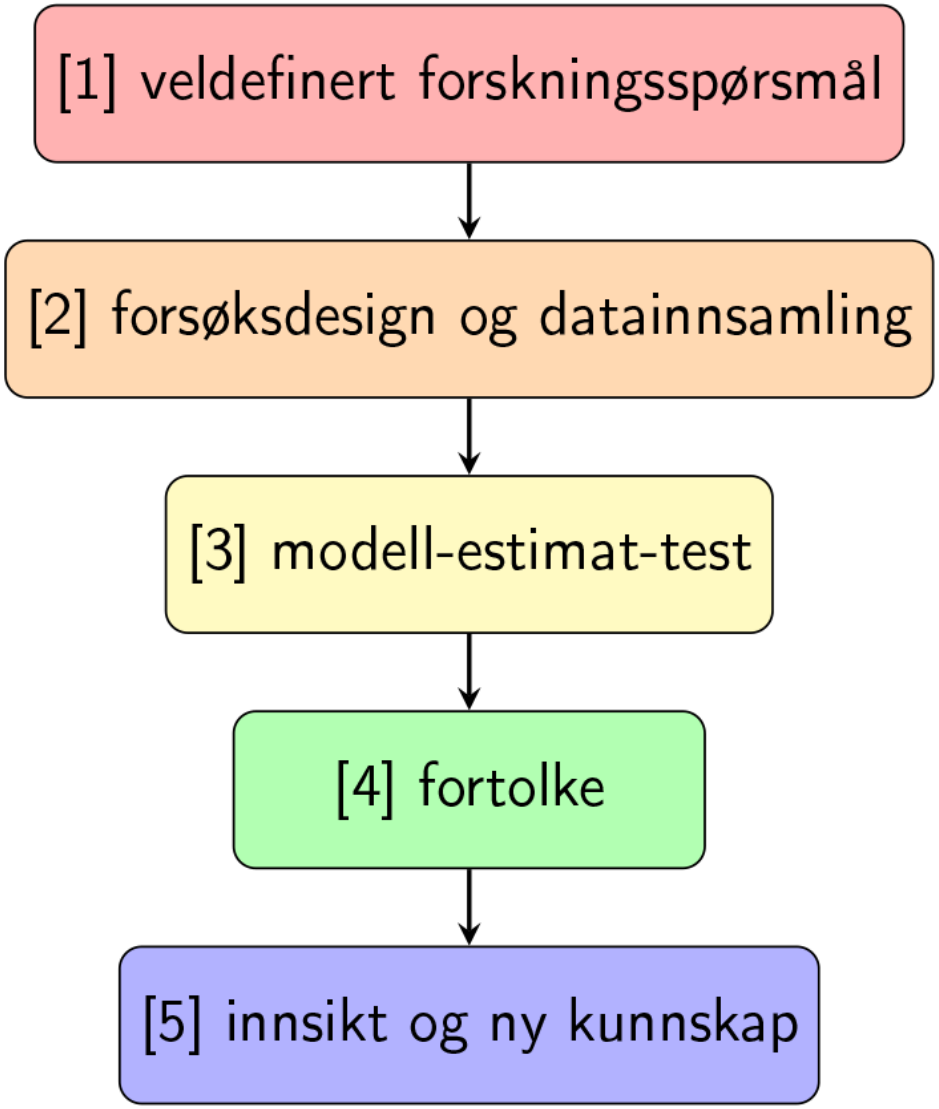
Dessverre har vi i de siste årene vært vitne til at flere resultater innen forskning ikke kan reproduseres, og det snakkes om en reproduserbarhetskrise. Vi ser på hvordan det kan kobles til at vi kan gjøre mange hypotesetester og ser på begrepet p-hacking.

Deretter går vi over til matematisk teori for hvordan vi se hvordan fordelingen til sanne og falske p-verdier er og avslutter med å se hvordan vi kan generalisere Type I feil til mer enn en hypotese.

Innhold

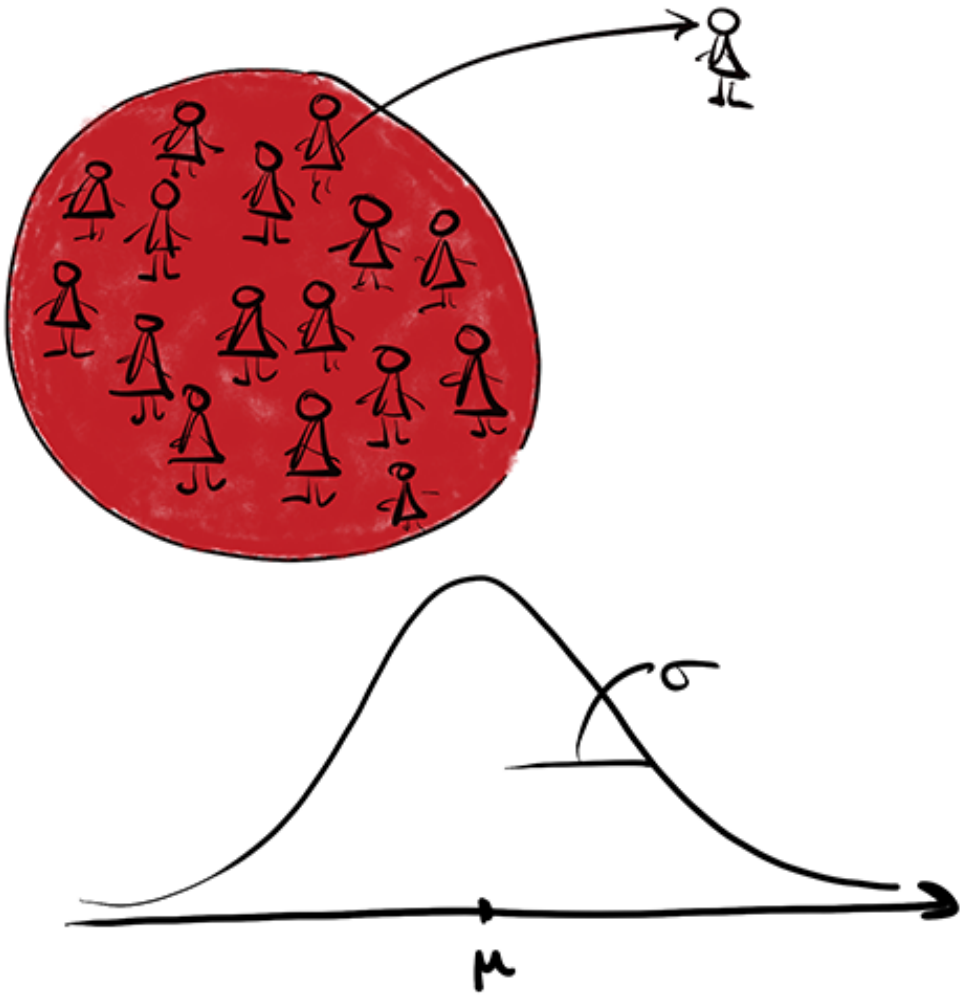
- Den vitenskaplige prosessen
- Hypotesetest
- *P*-verdi - hva er det egentlig?
- Reproduserbarhetskrisen og *p*-hacking
- Fra en til flere hypotester: STUV-tabell
- FWER og FDR
- Oppsummering

Den vitenskaplige prosessen



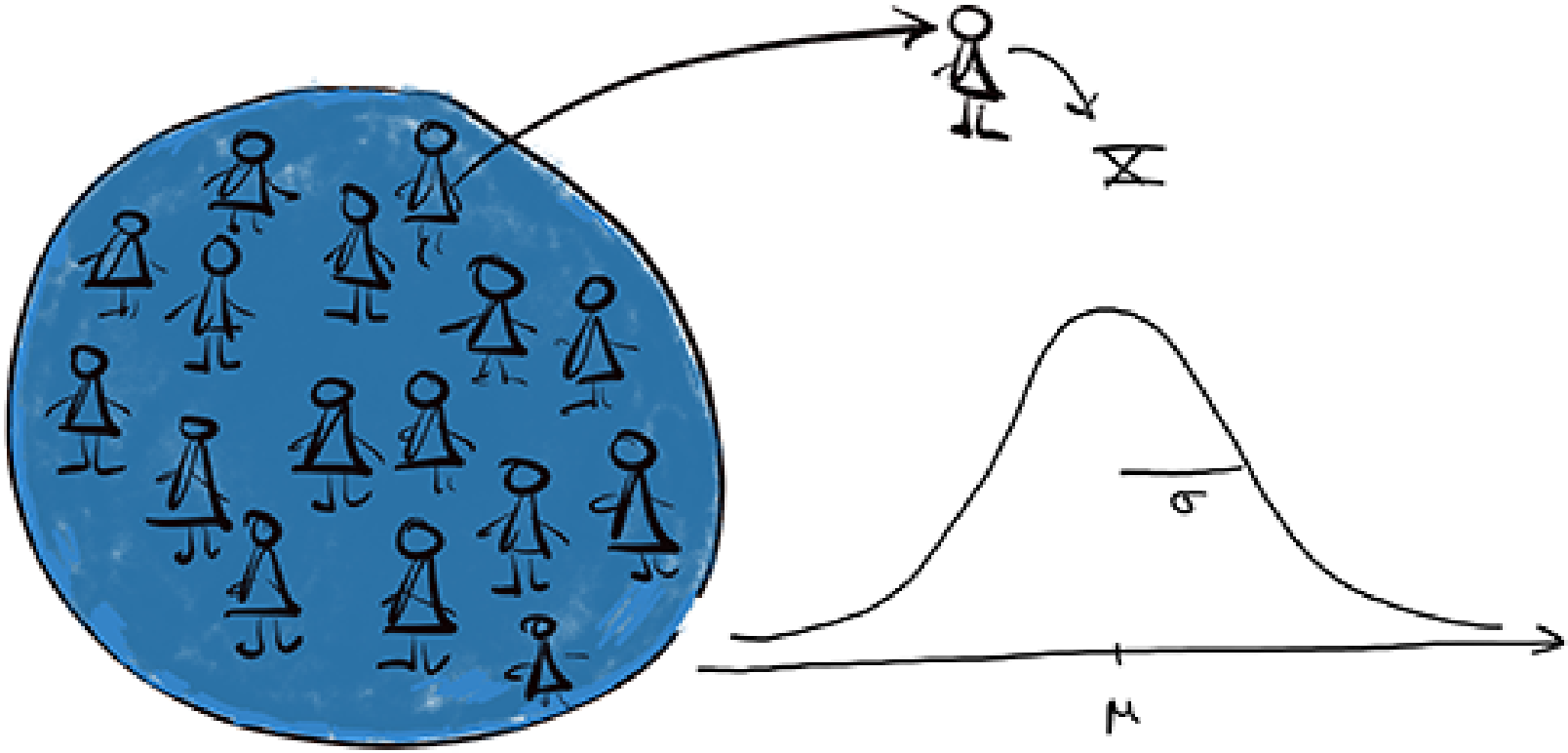
Illustrasjon: Systolisk blodtrykk

- I en normalbefolkning av kvinner i alderen 20-29 år er det kjent at det systoliske blodtrykket er normalfordelt med gjennomsnitt (forventingsverdi) $\mu = 120$ mmHg.
- Hva betyr det?

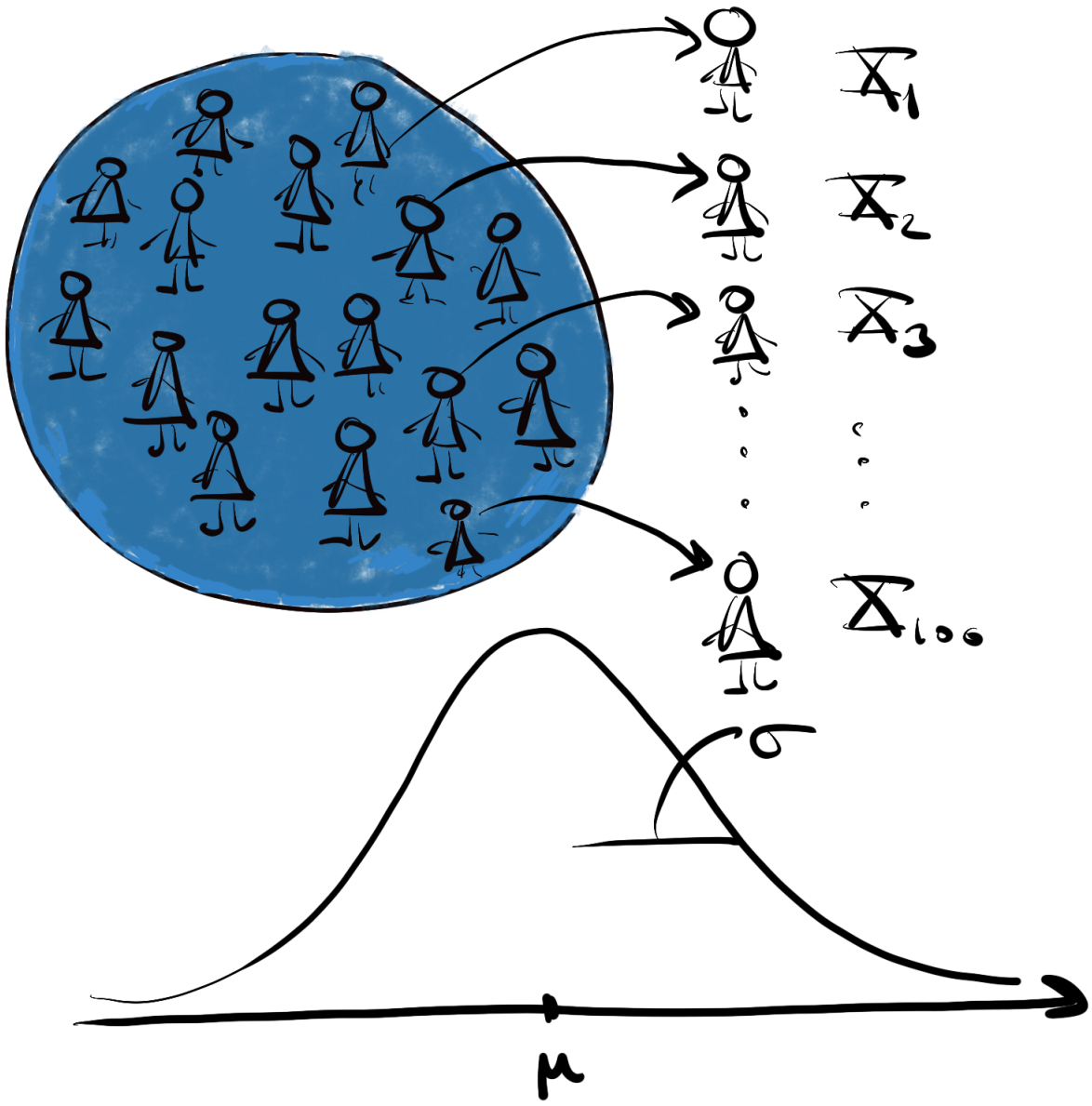


Estimering og konfidensintervall

- Vi studerer en populasjon av kvinner i alderen 20-29 år som *har en spesiell sykdom* (blå).
- Vi antar at systolisk blodtrykk i denne populasjonen er normalfordelt med ukjent gjennomsnitt μ (men kjent standard avvik 10 mmHg).
- Først: vi ønsker å estimere den ukjente forventningsverdien (punktestimat og konfidensintervall). Hvordan skal vi gå frem?



- Vi trekker et *tilfeldig utvalg* av størrelse $n = 100$ fra en blå populasjonen, og måler blodtrykk $X_1, X_2, \ldots, X_n$.
- Vi finner at $\bar{x} = 122$ mgHg.



Hypoteser - feil og signifikans

- Vi ønsker å teste om gjennomsnittlig systolisk blodtrykk i den blå populasjonen er *større enn 120 mmHg*.
- Sett opp null- og alternativ hypotese.
- Hva er type I og type II feil her?
- Hvor mye type I feil tåler vi?

Set-up - en hypotese

	H_0 sann	H_0 gal
Ikke forkast H_0	Korrekt	Type II feil
Forkast H_0	Type I feil	Korrekt

To typer feil

- Falske positive = type I feil =justismord. Dette er våre *fake news*.
- Falske negative = type II feil= skyldig går fri.

Signifikansnivået kalles α.

Vi sier at: type I-feilen er "kontrollert" på nivå α. Da overgår ikke "justismordsannsynligheten" α.

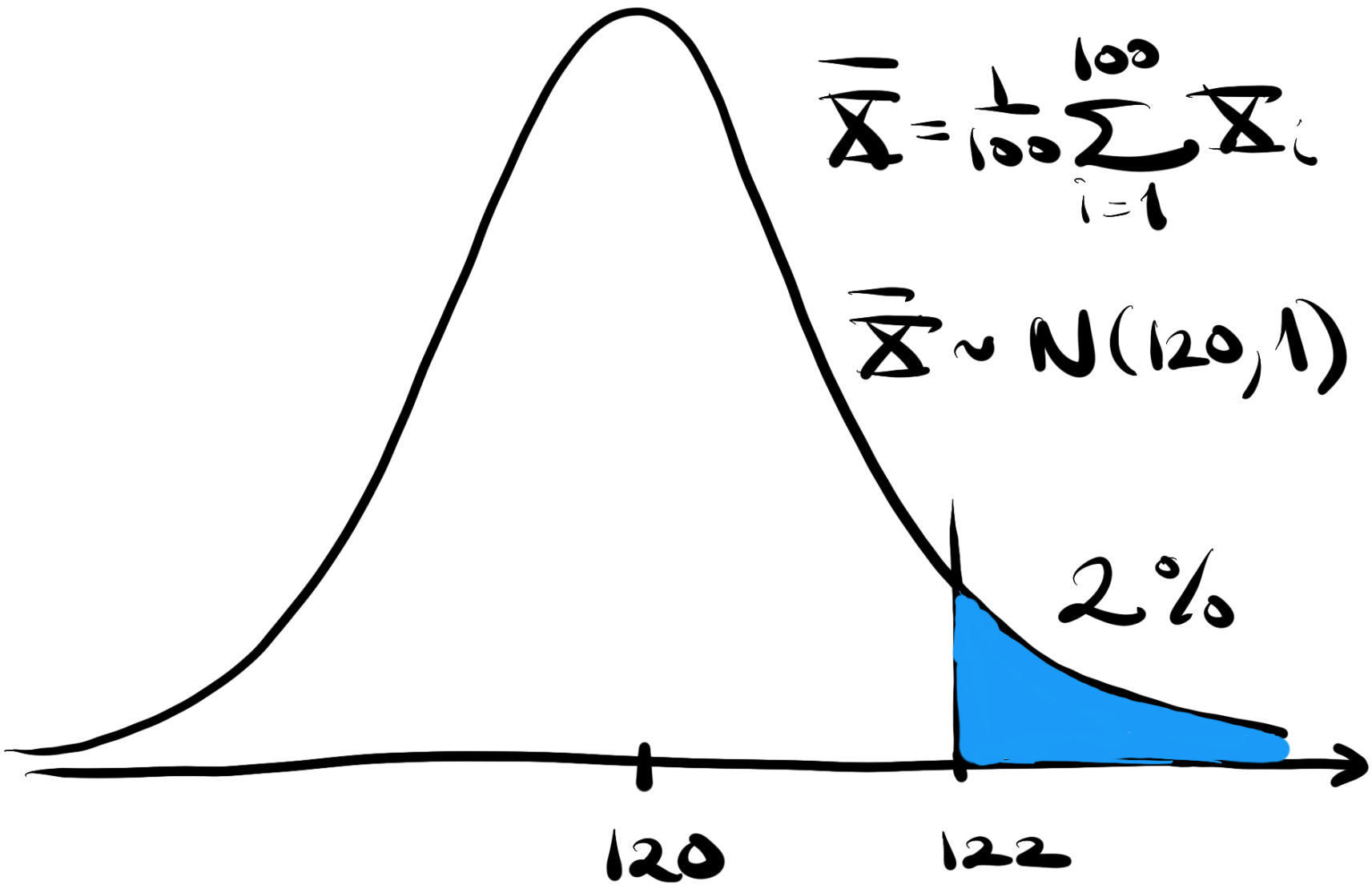
Rettssak-analogien

- Lille my (μ) er anklaget for å være større enn 120 - dette er H_1.
- Men, når lille my μ er uskylding er $\mu = 120$ (eller mindre). Dette er H_0.
- Type-I-feilen er justismordet: forkaste nullhypotesen når den er sann.
- Type-II-filen er å la forbryter gå fri: ikke forkaste nullhypotesen

P-verdi

- Sannsynligheten for det vi har observert eller noe mer ekstremt, gitt at nullhypotesen er korrekt.
- Hvis p-verdien er mindre enn vårt valgte signifikansnivå så forkaster vi nullhypotesen.

- Skal vi forkaste nullhypotesen eller ikke?
- Hva konkluderer vi da med?



Dualitet mellom konfidensintervall og hypotesetest

P-verdi

Uformelt sagt: p-verdien er sannsynligheten - under en spesifisert statistisk modell - at en testobservator fra data er lik eller mer ekstrem enn det vi har observert.

Eksempel

- Nullhypotese: det er sol ute.
- Data: jeg kommer inn i rommet, jeg er våt i håret og det drypper fra paraplyen min.
- Gal p-verdi: sannsynligheten for at det er sol ute.
- Umulig å regne ut.
- Riktig p-verdi: sannsyligheten for at jeg er våt i håret og det drypper fra paraplyen min gitt at det er sol ute.
- Her burde p-verdien være liten.

P-verdi - statement

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".

Hvorfor trengte/trenger man det?

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.

Hack your way to scientific glory

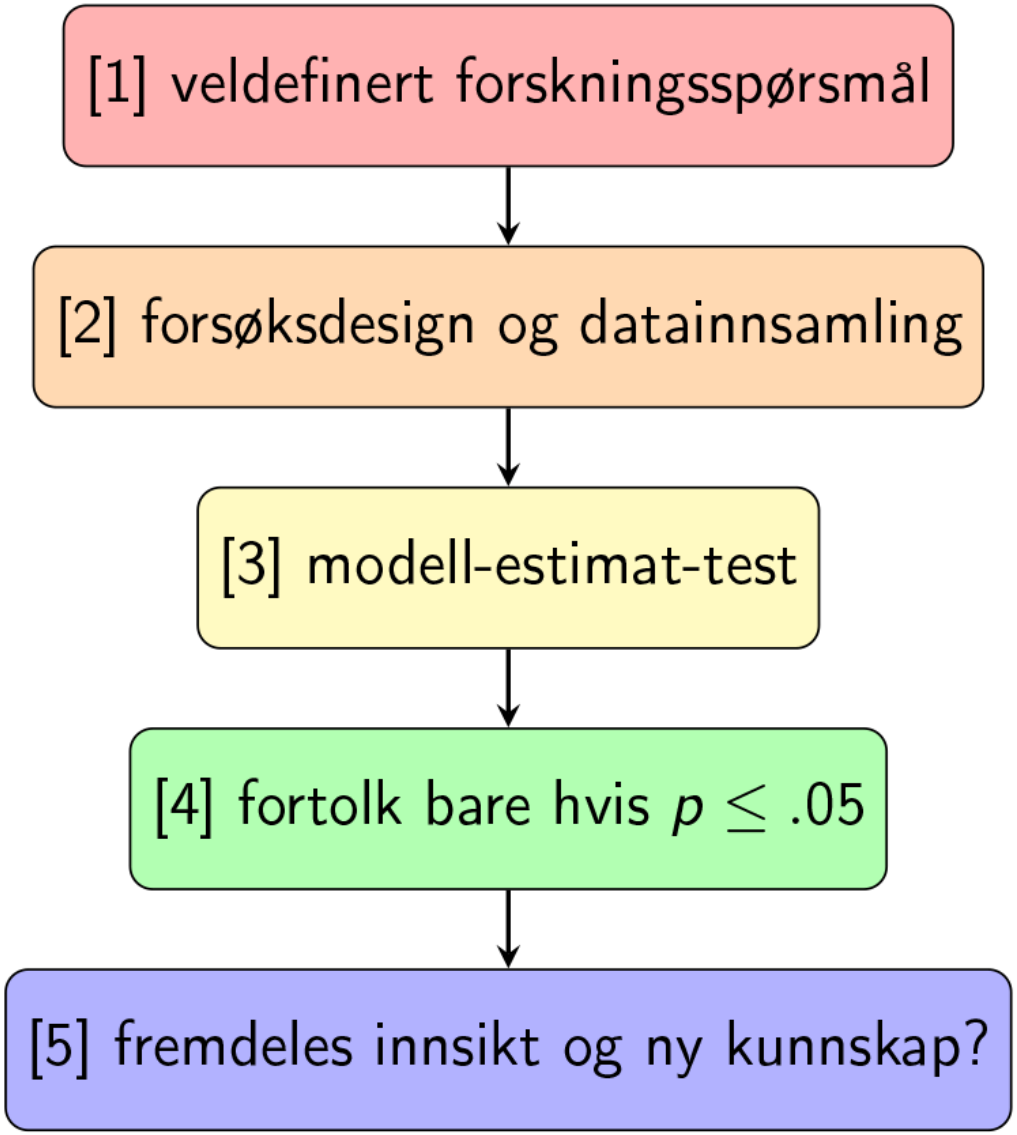
Ioannidis (2005): How many nonsignificant results have been studied before one research group has published its first significant finding?

What is and is not the *p*-value?

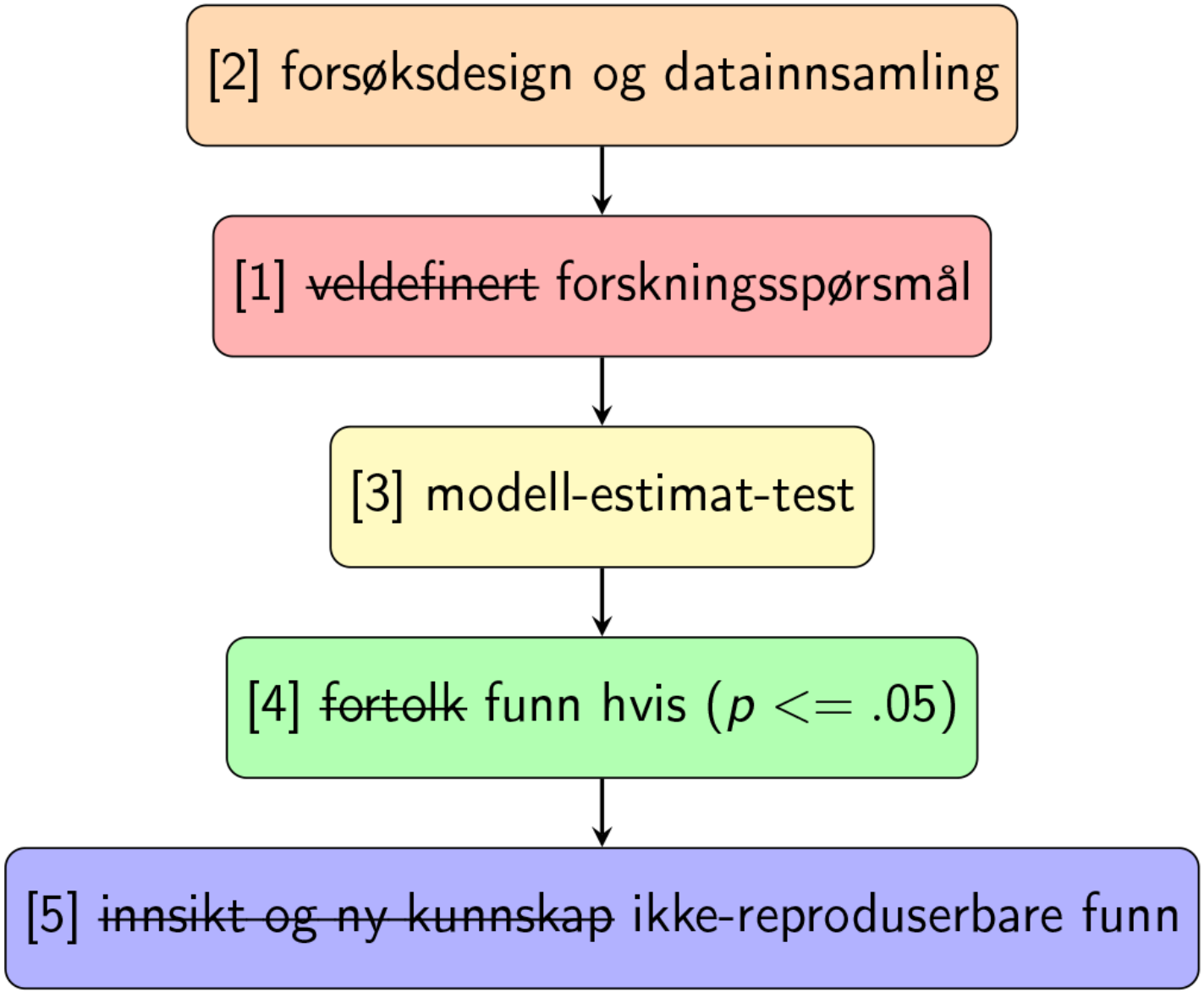
While the *p*-value can be a useful statistical measure, it is commonly misused and misinterpreted.

- 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.
- 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.

Bare funn for $p \leq 0.05$



P-hacking


[2] forsøksdesign og datainnsamling
[1] ~~veldefinert~~ forskningsspørsmål
[3] modell-estimat-test
[4] ~~fortolk~~ funn hvis ($p <= .05$)
[5] ~~innsikt og ny kunnskap~~ ikke-reproduserbare funn

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.

En reproduserbarhetskrise?

The reprocibility project in psychology

- Reproducibility is a defining feature of science, but the extent to which it characterizes current research is unknown.
- We conducted replications of 100 experimental and correlational studies published in three psychology journals using high-powered designs and original materials when available.

Kilde

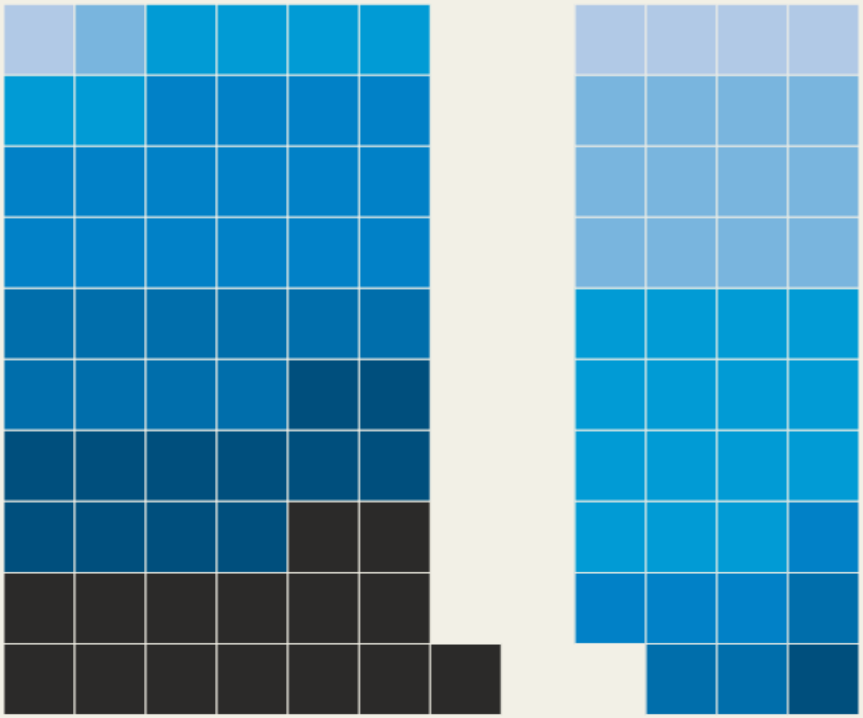
RELIABILITY TEST

An effort to reproduce 100 psychology findings found that only 39 held up.* But some of the 61 non-replications reported similar findings to those of their original papers.

Did replicate match original's results?

NO: 61

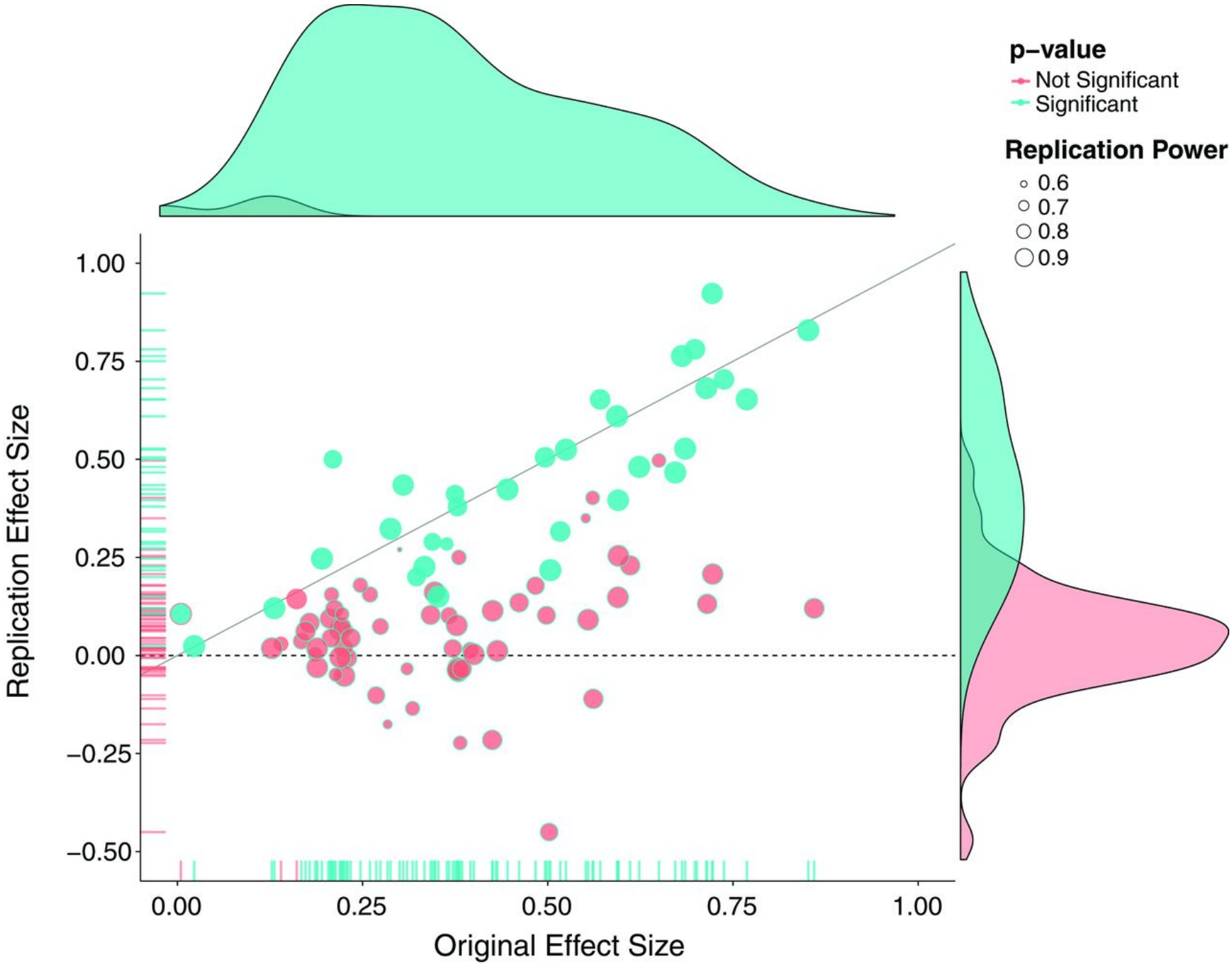
YES: 39



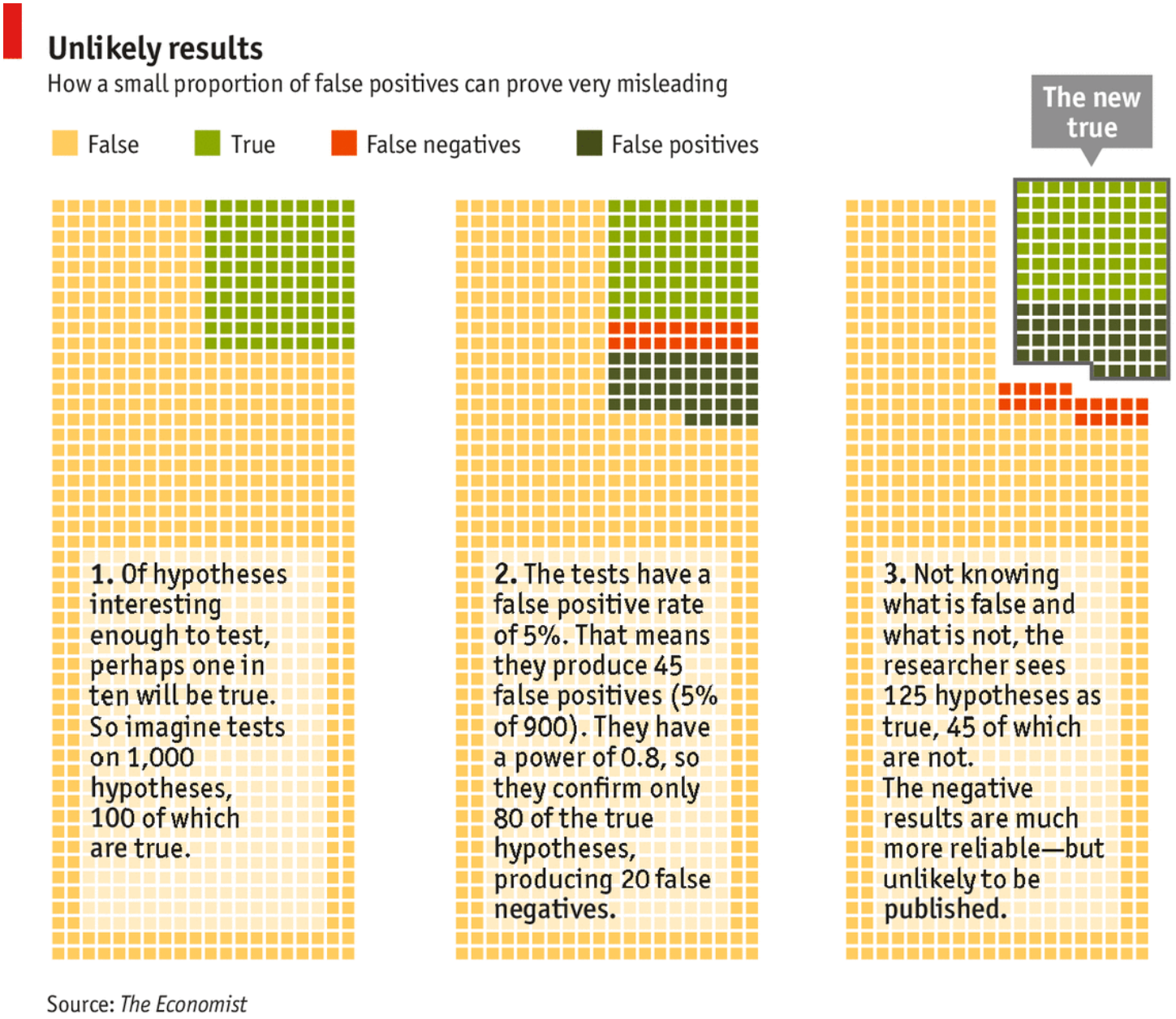
Replicator's opinion: How closely did findings resemble the original study:

- Virtually identical
- Extremely similar
- Very similar
- Moderately similar
- Somewhat similar
- Slightly similar
- Not at all similar

* based on criteria set at the start of each study


p–value
Not Significant
Significant
Replication Power
0.6
0.7
0.8
0.9
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
Replication Effect Size
0.00
0.25
0.50
0.75
1.00
Original Effect Size

Hva er andelen falske funn (fake news) i forskning?



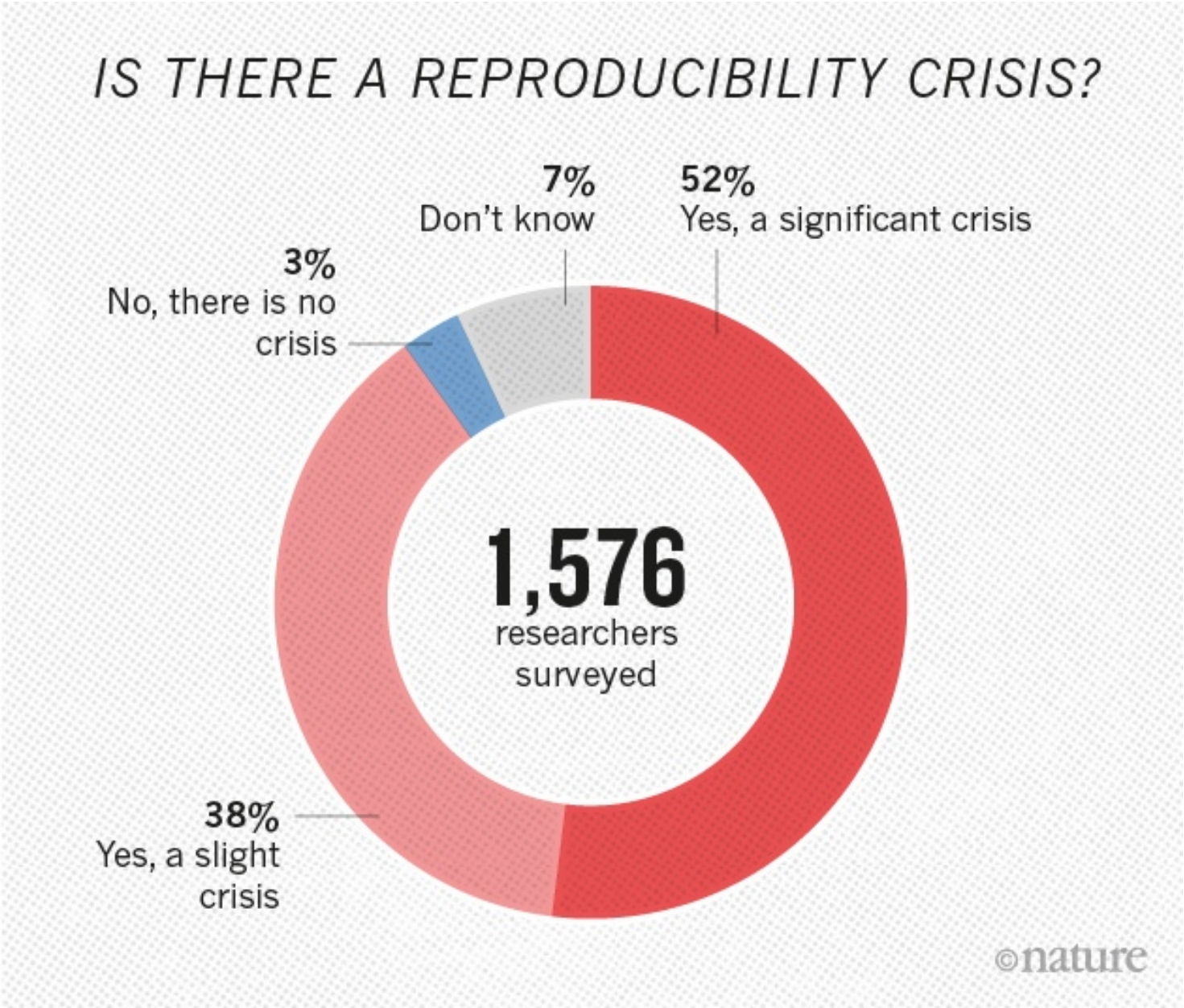
True=true H_1 (100 hypotheses) and False=false H_1 (900 hypotheses).
Kilde

Forklaring til figuren

- 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).

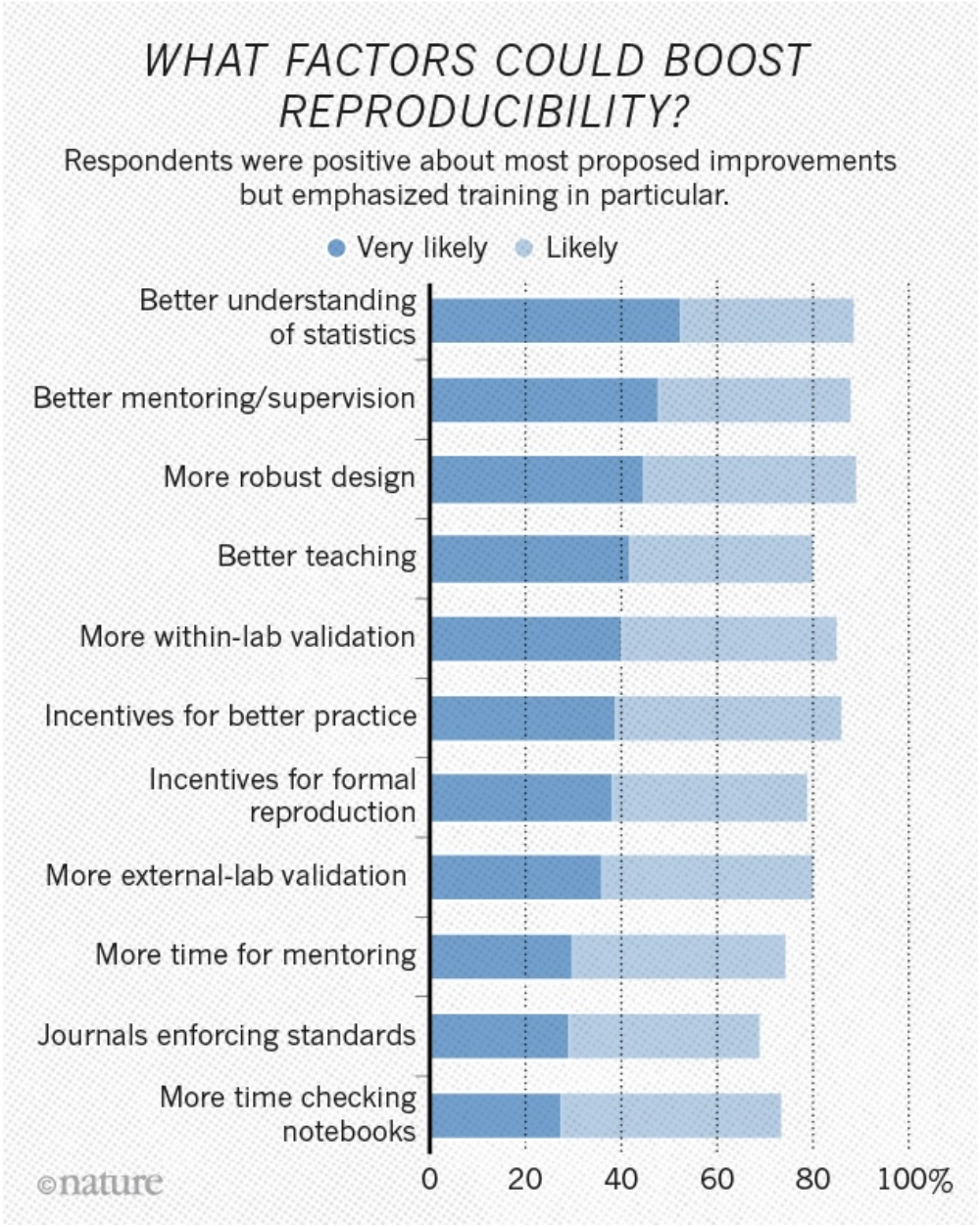
So, not 5% of published results are false positives (fake news), but rather at substantially larger number - 40-90% has be hinted to in different publications.

Er det en reproduserbarhetskrise i vitenskapen?



Kilde

Hvilke faktorer "could boost reproduciability""



Kilde

Om ikke det var nok

The spread of true and false news online. Soroush Vosoughi, Deb Roy, Sinan Aral, Science 09 Mar 2018: Vol. 359, Issue 6380, pp. 1146-1151. DOI: 10.1126/science.aap9559

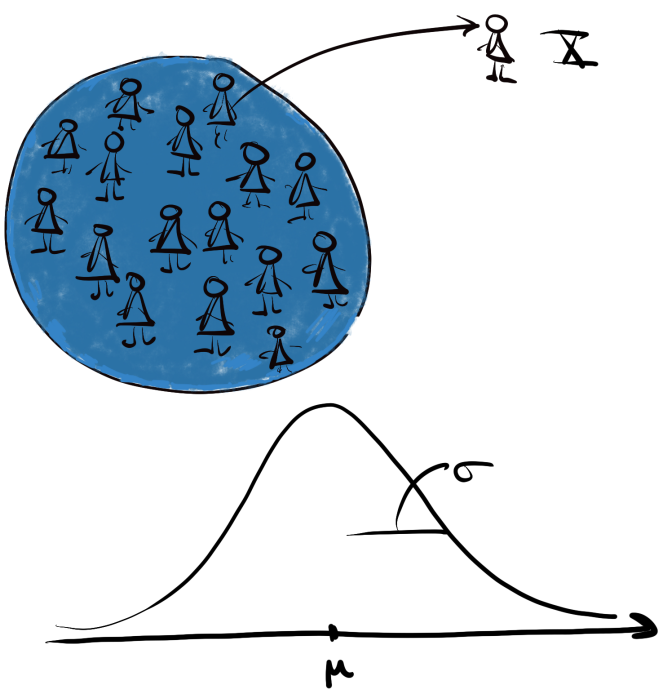
Lies spread faster than the truth

- There is worldwide concern over false news and the possibility that it can influence political, economic, and social well-being.
- To understand how false news spreads, Vosoughi et al. used a data set of rumor cascades on Twitter from 2006 to 2017.
- About 126,000 rumors were spread by approx. 3 million people.
- False news reached more people than the truth; the top 1% of false news cascades diffused to between 1000 and 100,000 people, whereas the truth rarely diffused to more than 1000 people.
- Falsehood also diffused faster than the truth. The degree of novelty and the emotional reactions of recipients may be responsible for the differences observed.

Videre

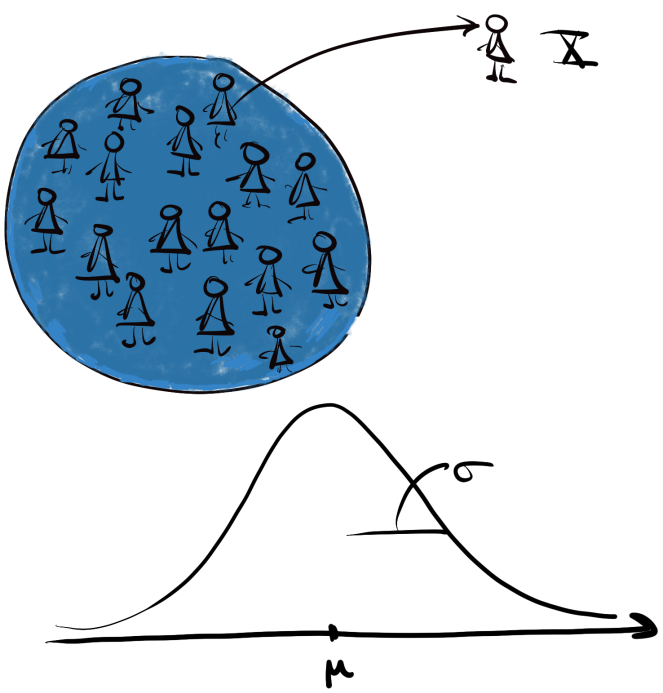
- Selv om et resultat er *statistisk signifikant* kan det være at avviket fra H_0 er så lite at det ikke er *praktisk interessant*.
- Hvis man gjør mer enn en hypotesetest bør man justere signifikansnivået for å ta hensyn til dette.
- Det finnes mange måter å kontrollere en generalisering av type-I-feilen når man gjør m hypotesetester - vi skal nå straks se på to metoder!

Hypothesis testing example



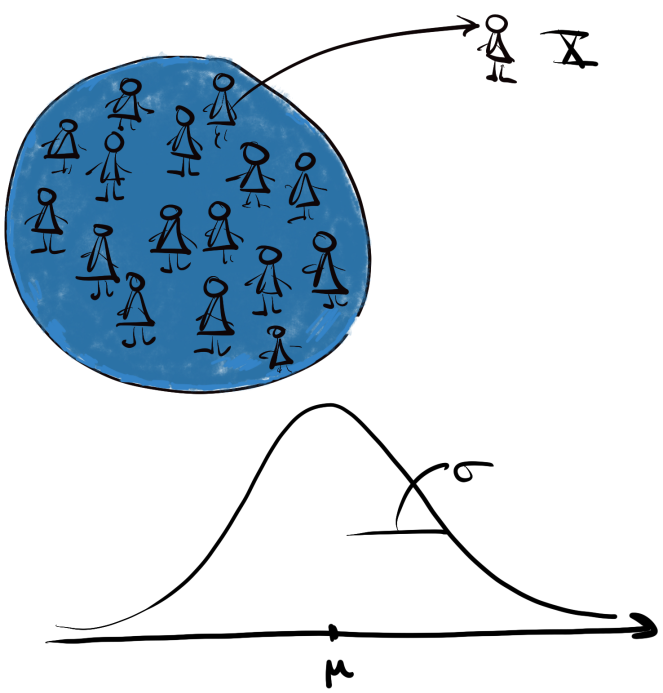
- It is known that in a population of women of age 20-29 years the systolic blood pressure is normally distributed with mean $\mu = 120$ mmHg.

Hypothesis testing example



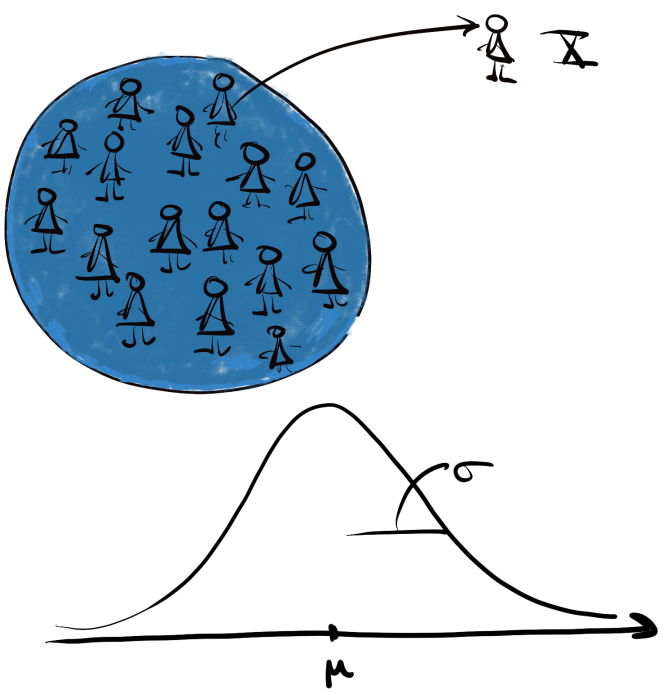
- It is known that in a population of women of age 20-29 years the systolic blood pressure is normally distributed with mean $\mu = 120$ mmHg.
- We study a population of women of age 20-29 that have a specific disease (blue population), and also here we assume that the systolic blood pressure is normally distributed (with standard deviation 10 mmHg), but here we don't know the mean in the population.

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.

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. $H_1 : \mu > 120$.

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:

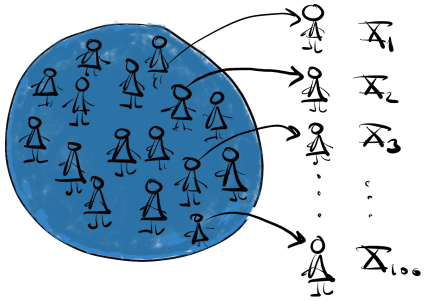
- False positives = type I error =miscarriage of justice. These are our *fake news*.
- False negatives = type II error= guilty criminal go free.

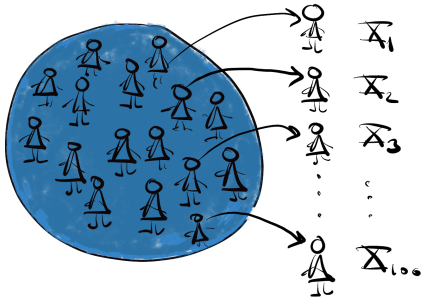
The significance level of the test is α.

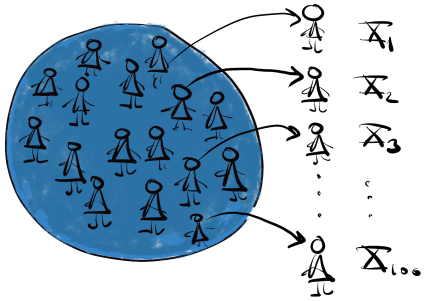
We say that : Type I error is "controlled" at significance level α.

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

Repeating the blood pressure experiment



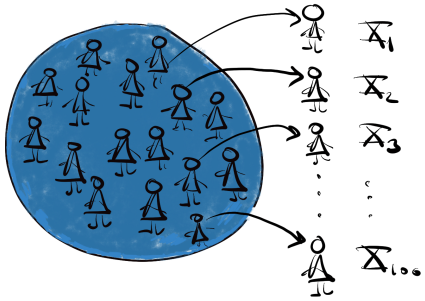


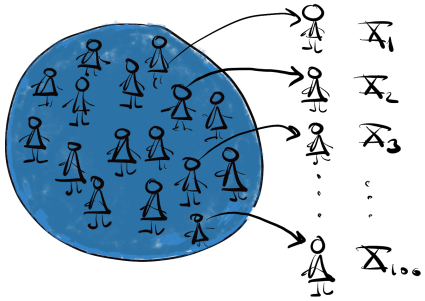


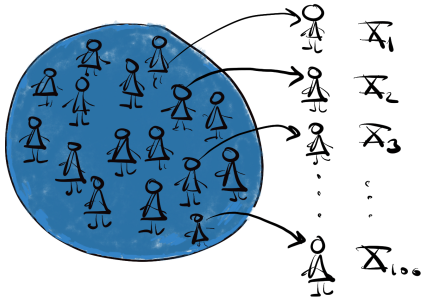
$\bar{x}$=120.9 $\bar{x}$ =118.9 $\cdots$ $\bar{x}$ = 121.2

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

Repeating the blood pressure experiment

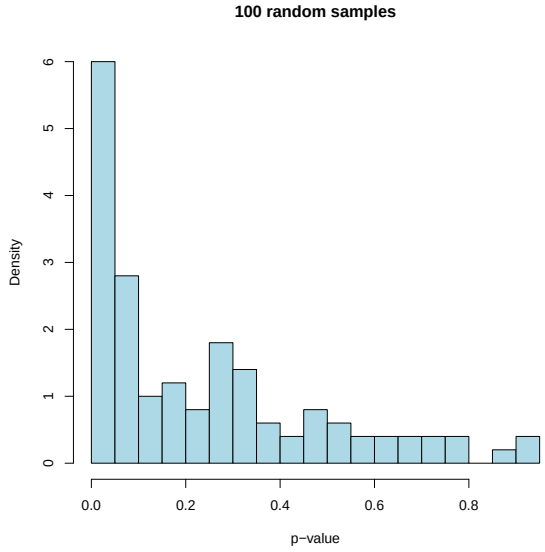




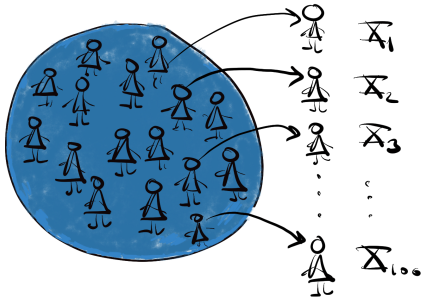


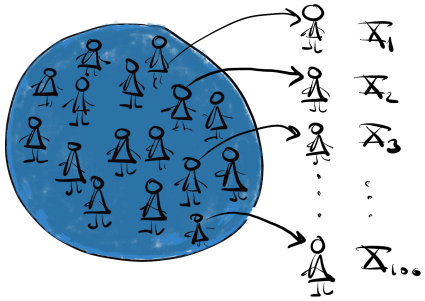
$\bar{x}$=120.9 $\bar{x}$ =118.9 $\cdots$ $\bar{x}$ = 121.2

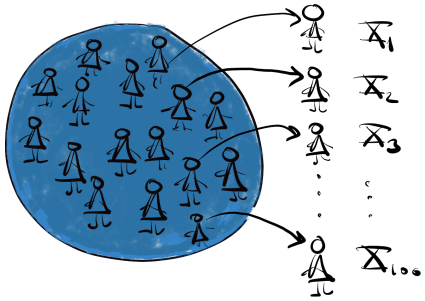
p-value=0.18 p-value=0.86 $\cdots$ p-value=0.12



Repeating the blood pressure experiment

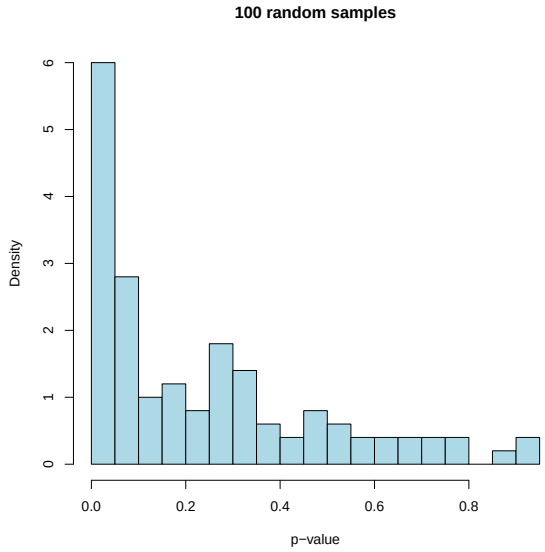


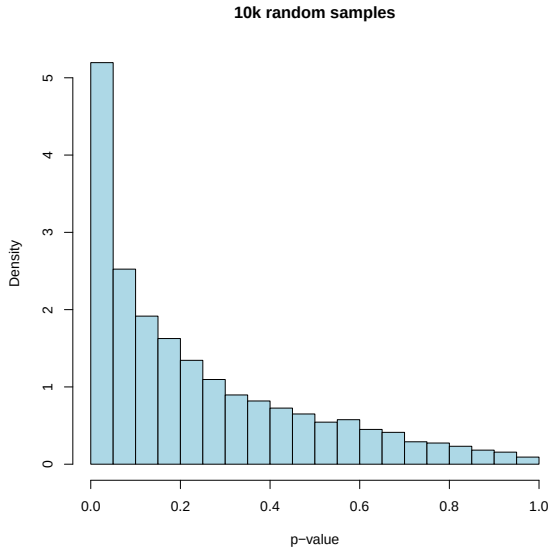




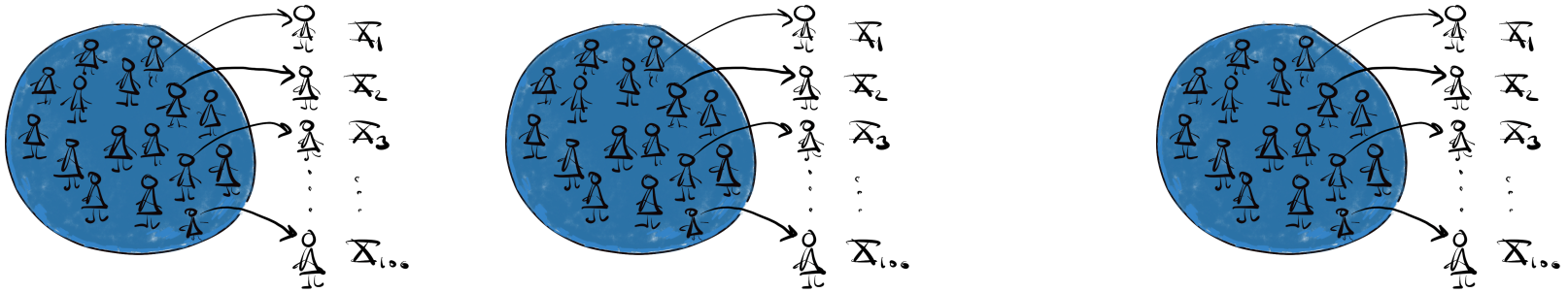
$\bar{x}$=120.9 $\bar{x}$ =118.9 $\cdots$ $\bar{x}$ = 121.2

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



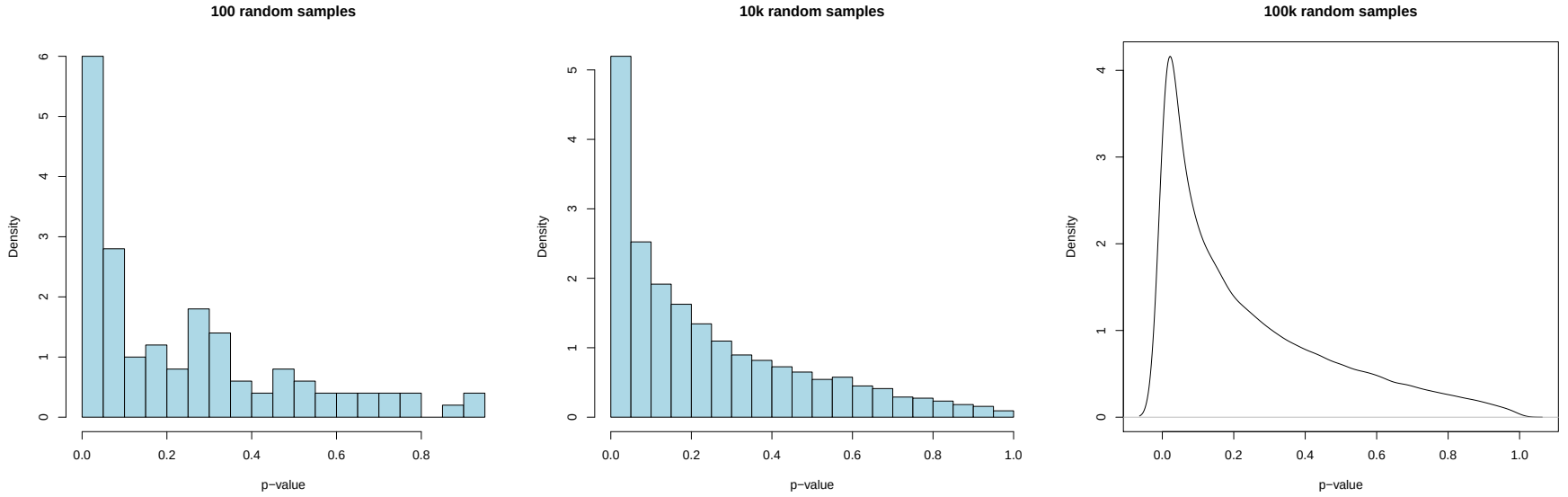


Repeating the blood pressure experiment



$\bar{x}$=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: $P(\bar{X} > \text{observed mean} \mid H_0 \text{ true})$.

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?

More about the p-value

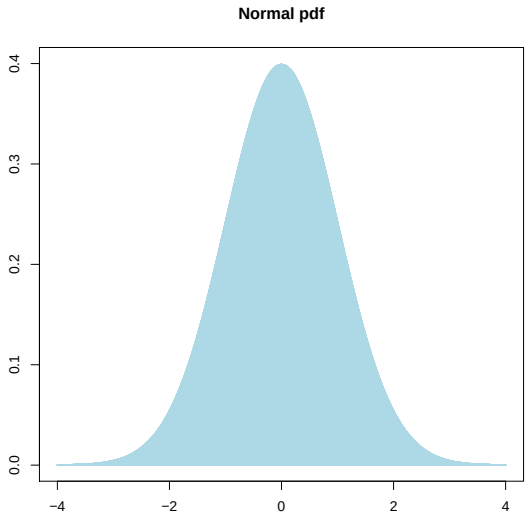
- The p-value is just a function of the random sample and can be regarded as a random variable.
  We had: $P(\bar{X} > \text{observed mean} \mid H_0 \text{ true})$.
- But, isn't the p-value a probability? A number?
- A random variable (like the p-value) has a *probability distribution*.

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?

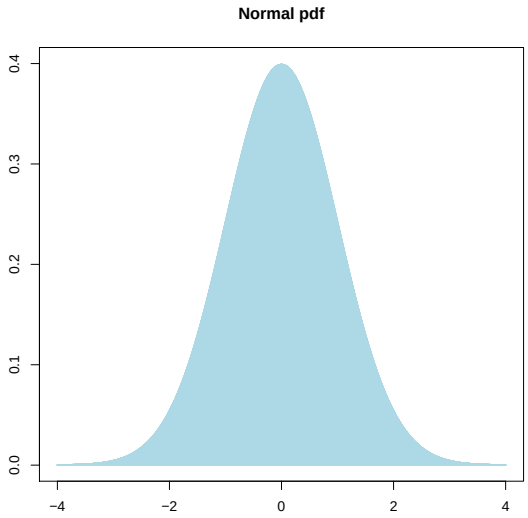
Probability distribution for random variable Y

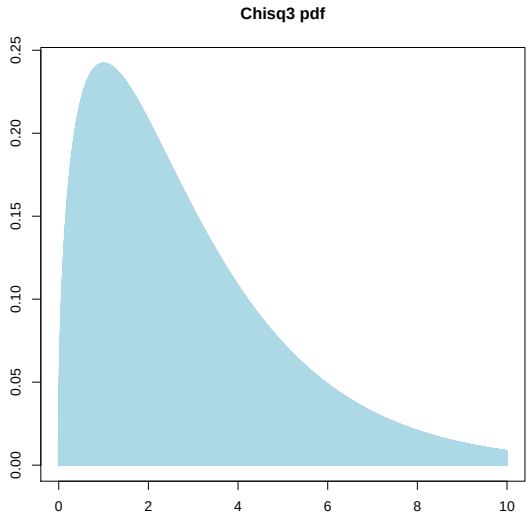
- Continuous random variable Y (could be the p-value).
- Probability distribution function (pdf): $f(y)$.



Probability distribution for random variable Y

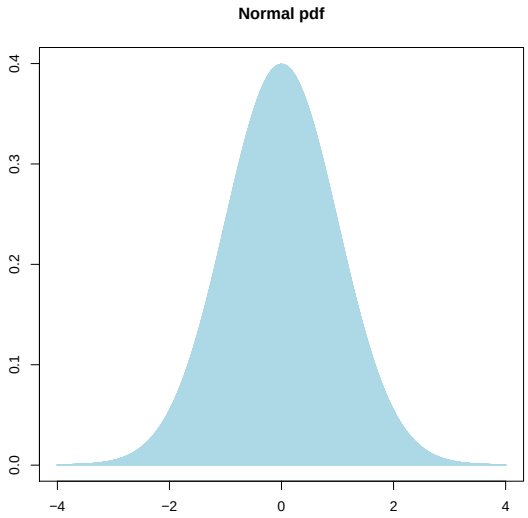
- Continuous random variable Y (could be the p-value).
- Probability distribution function (pdf): $f(y)$.

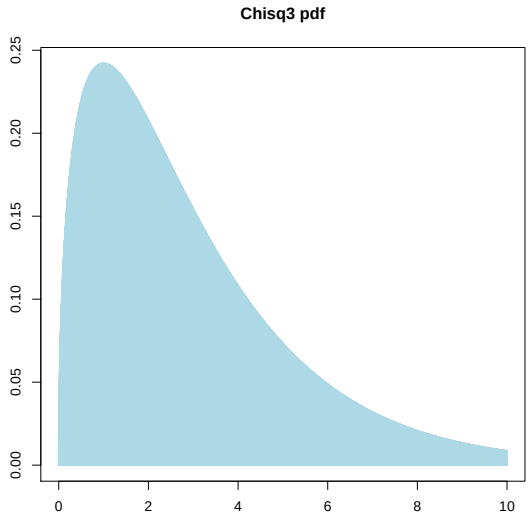


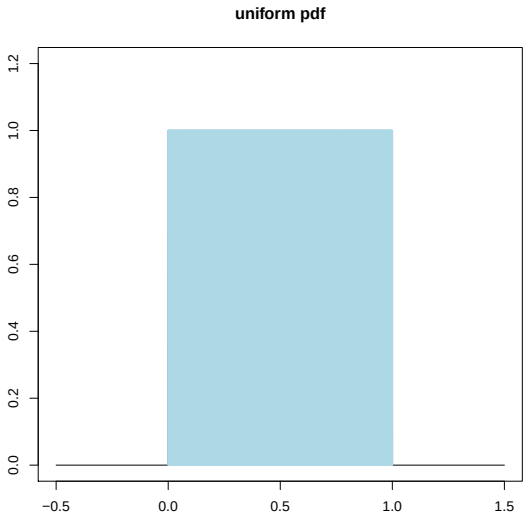


Probability distribution for random variable Y

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





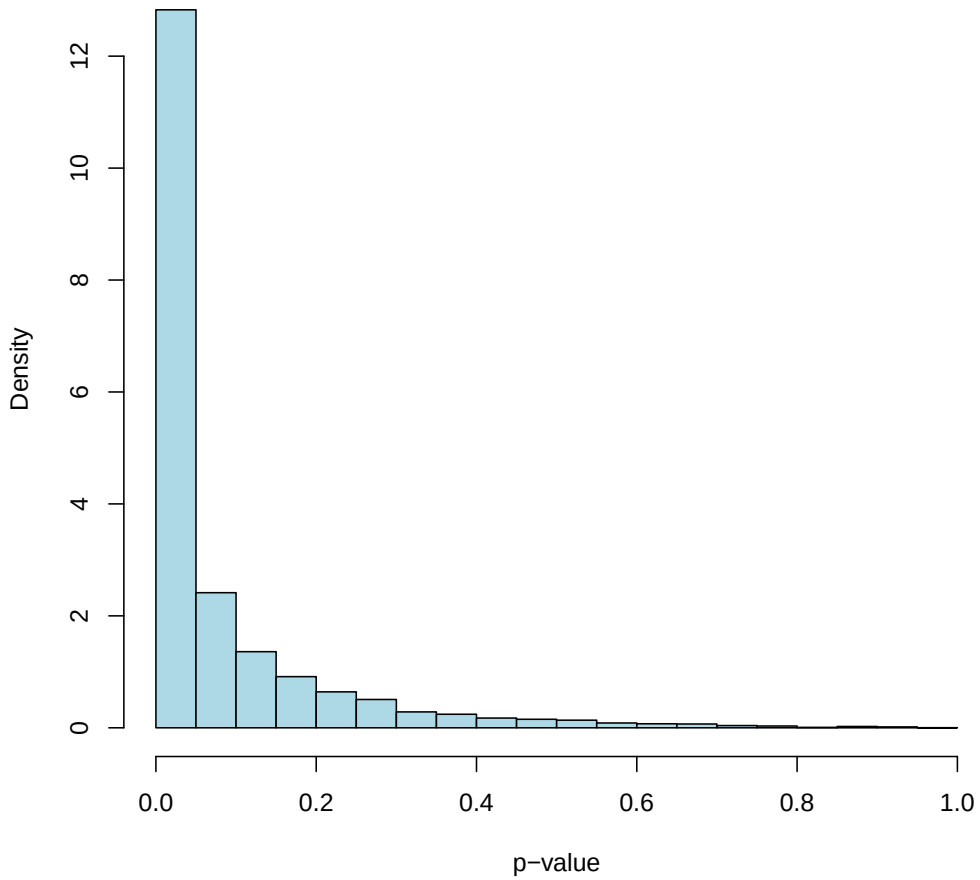


Distribution of p-values for false hypothesis?

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 $\mu = 122$ 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?

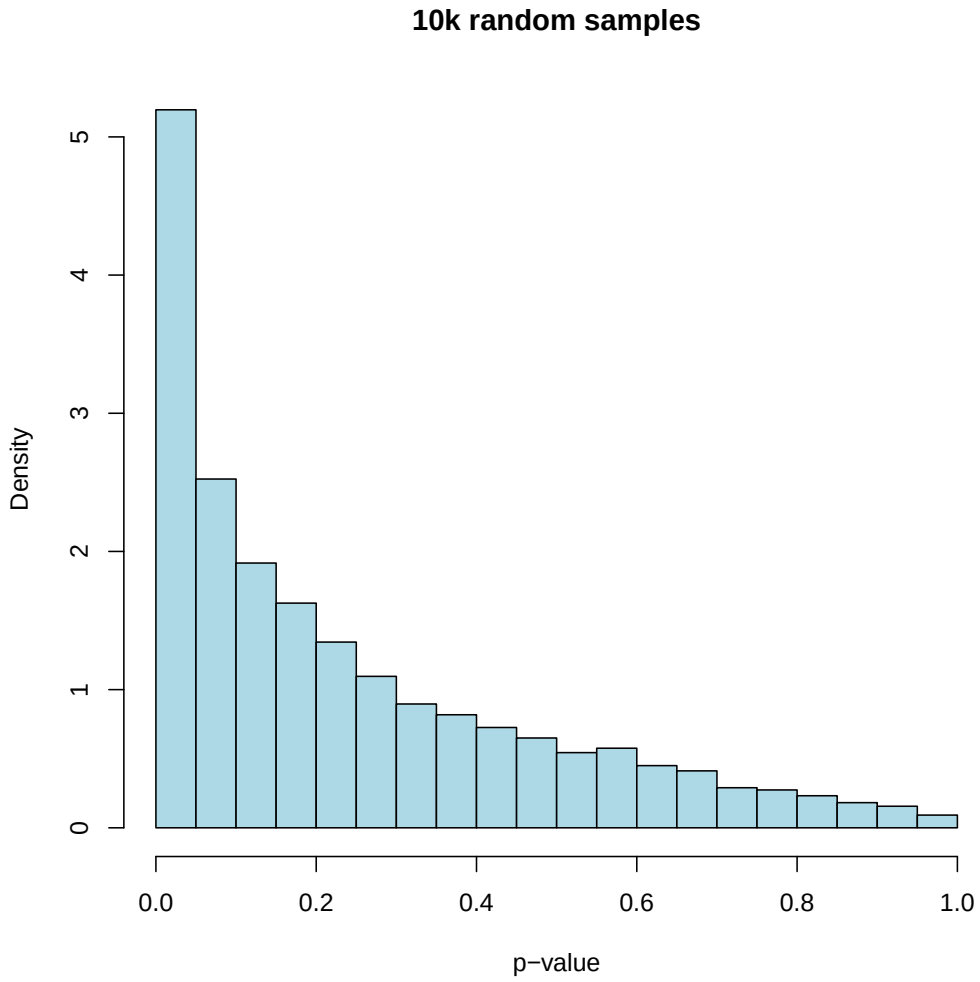


Distribution of p-values for false hypothesis?

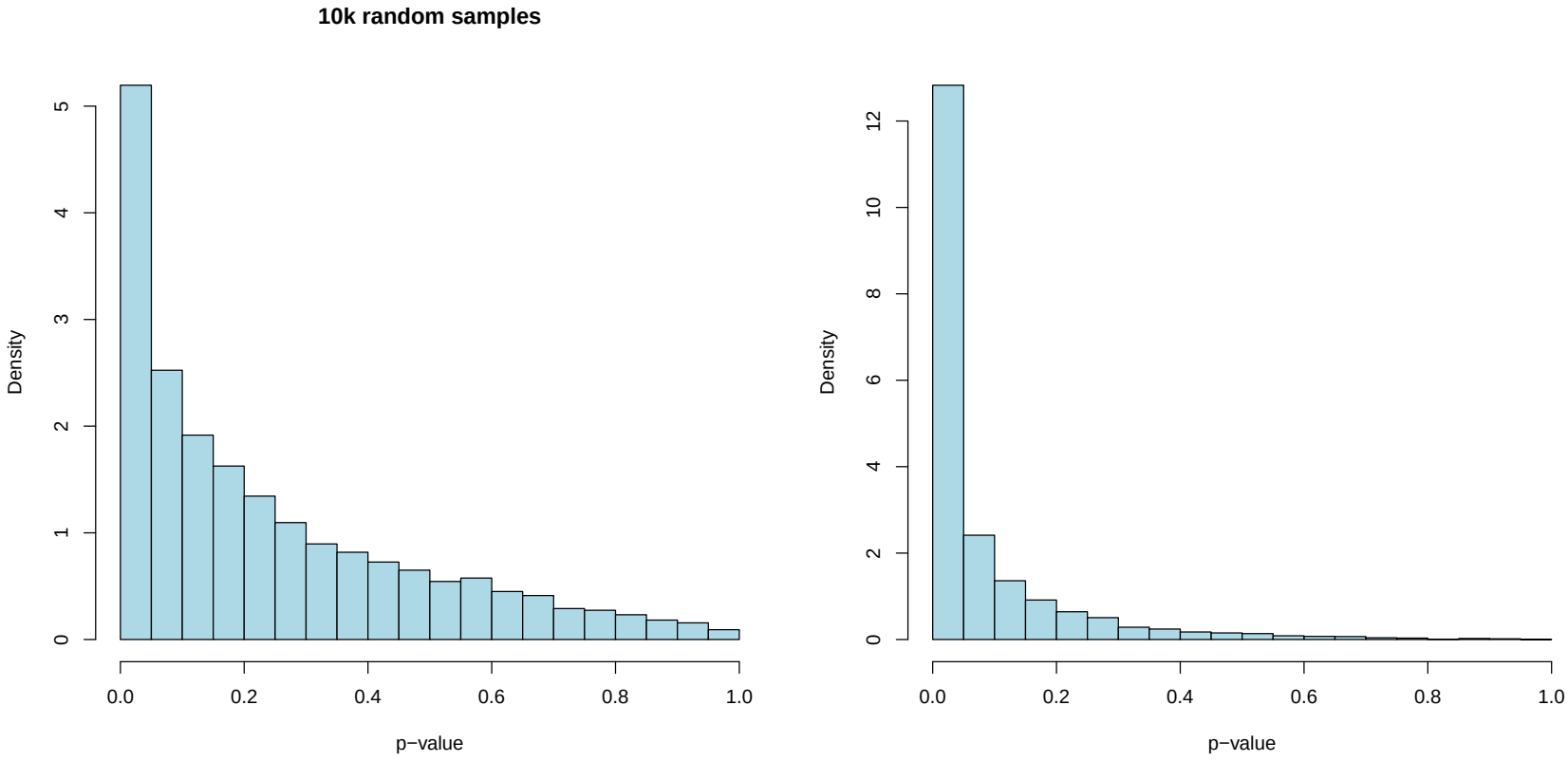
Blood pressure example:
Assume that $\mu = 121$ so that H_0 is false, and that we collect a random sample of size 100. What is then the distribution of the p-value?

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 $H_0 : \mu = 120$

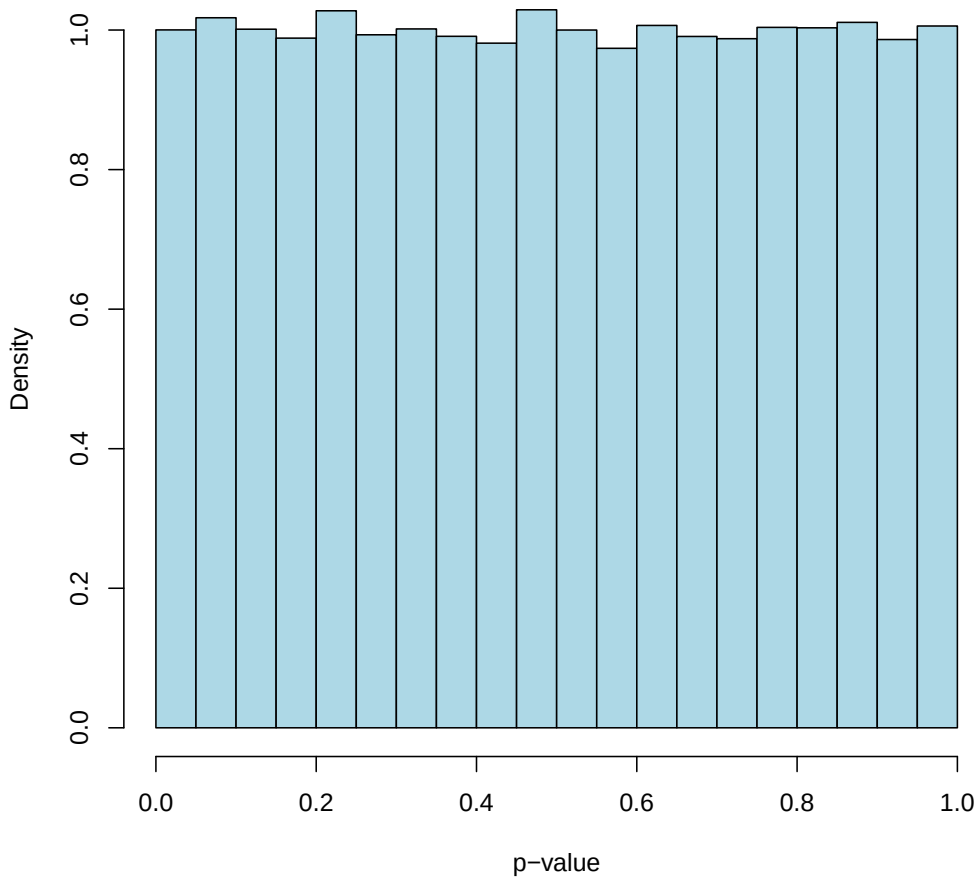


Distribution of p-values for true hypothesis?

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?

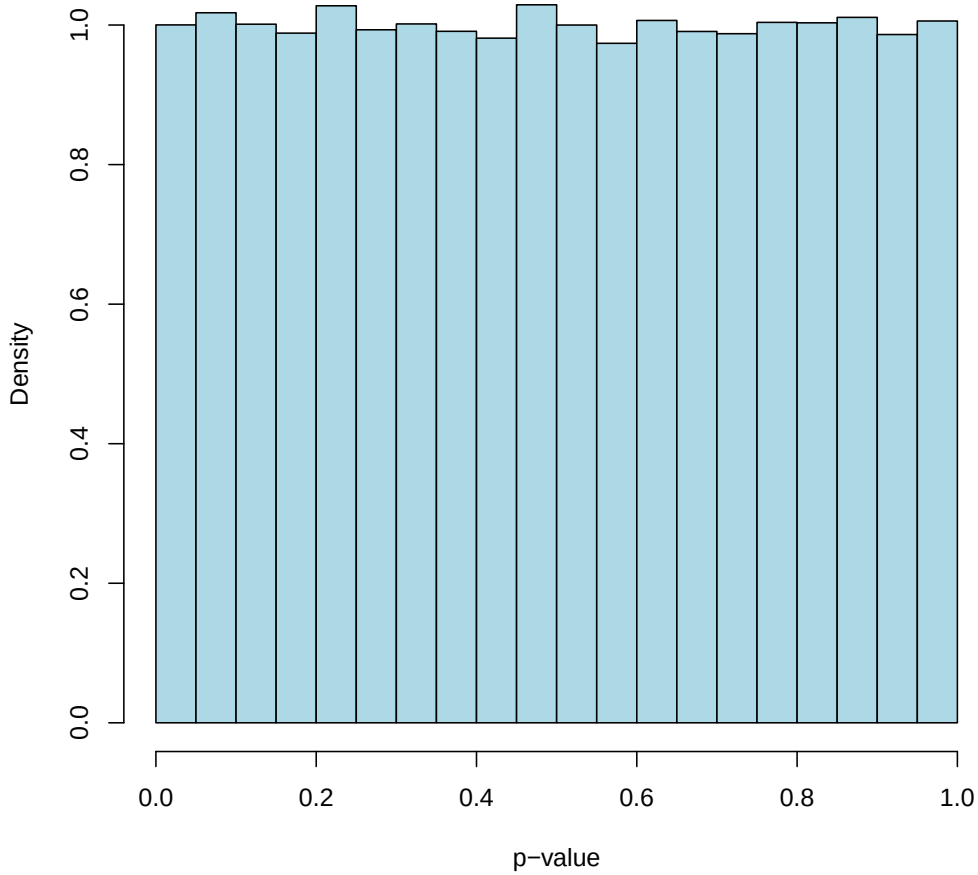
Distribution of p-values for true hypothesis?

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?



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.

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.

- 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? Sadly, no.

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

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,

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.

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.

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

  (remark: V is not observed)

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?

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 $\alpha_{\mathsf{LOC}} = 5 \cdot 10^{-8}$.

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.

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 sets $\alpha_{\text{LOC}} = \alpha/m$.

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 sets $\alpha_{\text{LOC}} = \alpha/m$.
- 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.

Overall Type I error control (2)

- For other types of data one expects that many null hypotheses are false,

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.

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.

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.

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

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

Multiple testing error control

More about FWER and FDR in TMA4267 Linear statistical method.
And, next week you learn about one specific method to control FWER for the ANOVA situation!

Oppsummering