Statistics and How to Interpret It

Descriptive Statistics

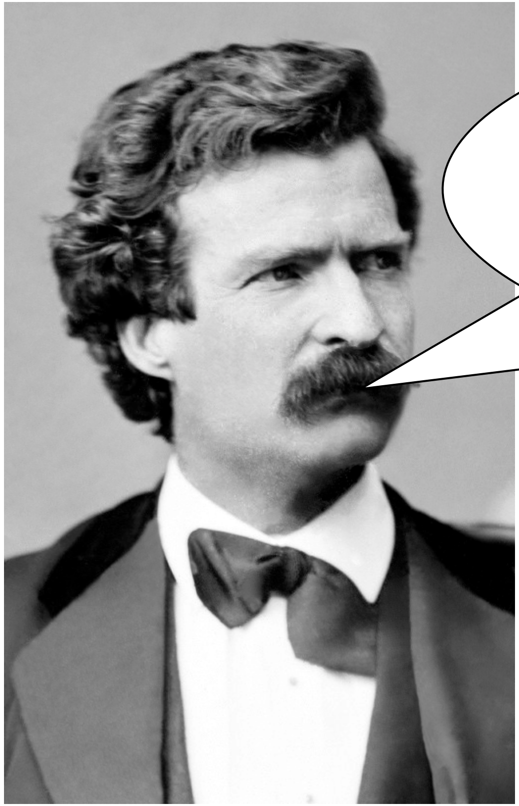
…describes

….summarizes

Inferential Statistics

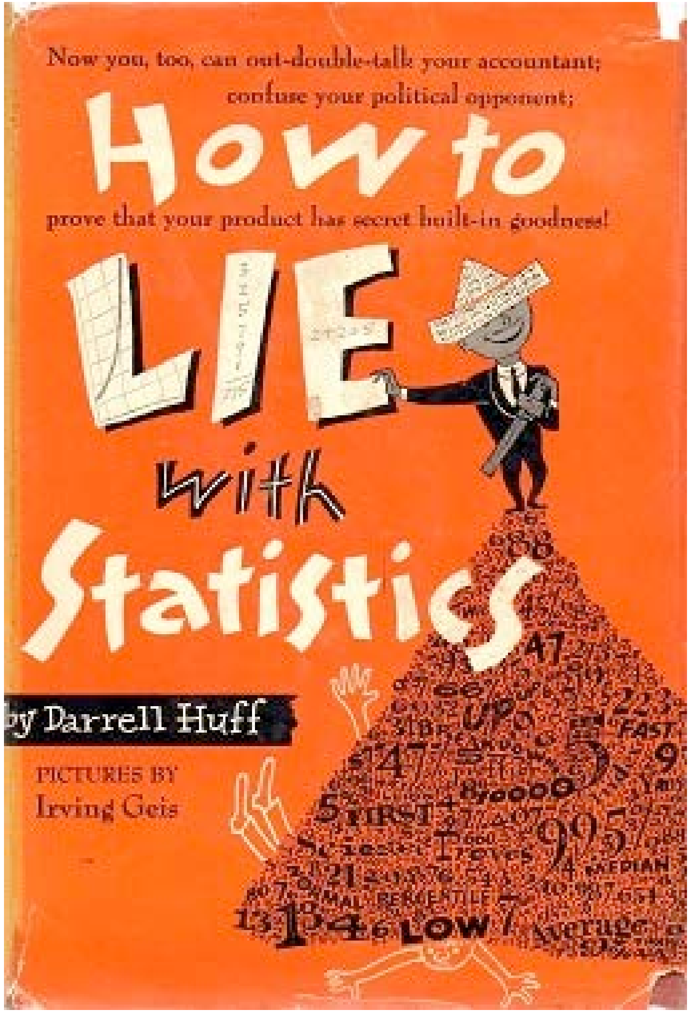
…infers to population

Data



There are three kinds of lies: lies, damned lies, and statistics.

Mark Twain (1835-1910)



“The most widely read statistics book in the history of the world ”

J. M. Steele. "Darrell Huff and Fifty Years of How to Lie with Statistics. Statistical Science, 20 (3), 2005, 205–209.

Lie: an intentionally false statement.

Lie: an intentionally false statement.

Lying with statistics:

- based on genuine, good quality data.
- employs proper statistical tools
- yield claims that are false - or at least what most people understand them to say is false

Statistics offers a toolbox with many tools, applicable to the same data, but yielding different results

Example: what is the AVERAGE income in your immediate family?

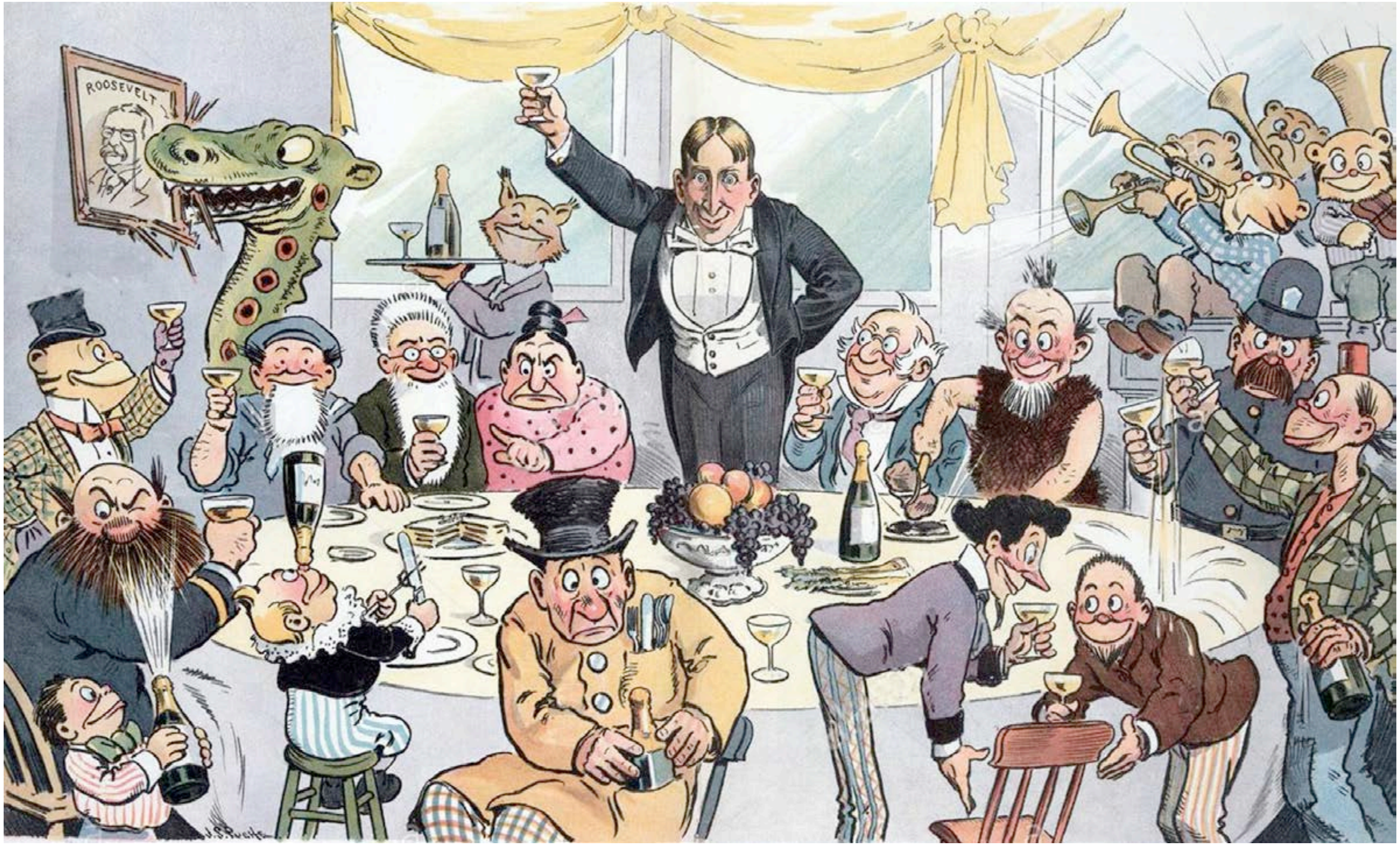
5 people, who earn:

€200.000, €91.000, €39.000, €37.000, €25.000

Mean: €78.400

Median: €39.000

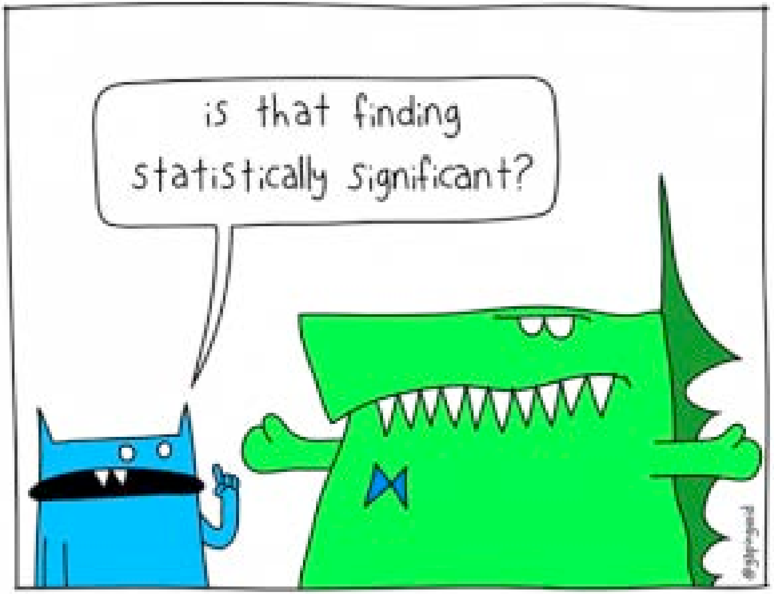
“The average income in *our* family is €78.400”



Statistics offers various tools that yield different results when applied to the same kind of data.

Inferential statistics

- e.g. assess hypotheses – by:
    - *Significance test*



Statistics offers various tools that yield different results when applied to the same kind of data.

Inferential statistics

- e.g. assess hypotheses – by:
    - *Significance test*
    - *Error probability*
    - *Bayesian inference*

Statistics offers various **methods**

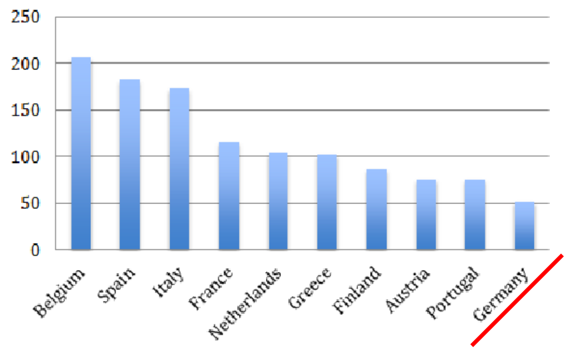
Methodology: choose the right tool for one’s purpose, and justify why this tool is the right one, rather some other, equally applicable one

Justified Choice of *Descriptive* Statistical Tools

Ambiguous Concepts

Press reports on ECB 2013 survey: “*On average*, German households hold lowest wealth of all countries in the Eurozone”

Figure 1. Net wealth of median households (1000€)



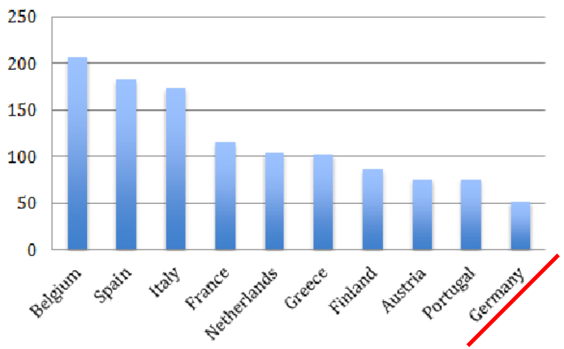
Source: European Central Bank (2013).

Inference: Germans on average are poorer than other Europeans, and therefore should not be required to contribute more than others (see, e.g., *Wall Street Journal* 2013, *Financial Times* 2013, *Frankfurter Allgemeine* 2013).

Ambiguous Concepts

Press reports on ECB 2013 survey: “*On average*, German households hold lowest wealth of all countries in the Eurozone”

Figure 1. Net wealth of median households (1000€)



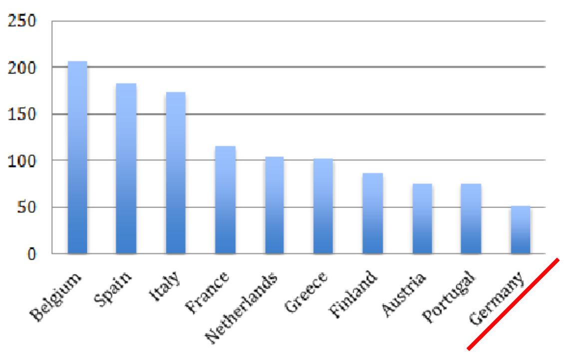
Source: European Central Bank (2013).

Inference: Germans on average are poorer than other Europeans, and therefore should not be required to contribute more than others (see, e.g., *Wall Street Journal* 2013, *Financial Times* 2013, *Frankfurter Allgemeine* 2013).

Ambiguous Concepts

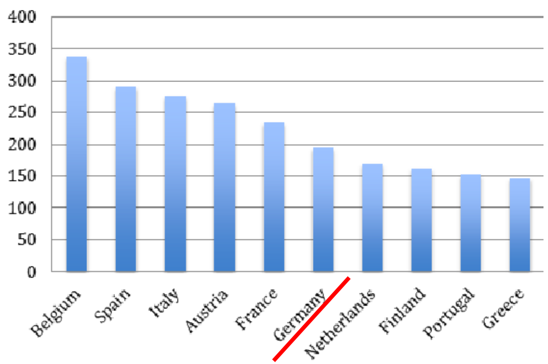
Press reports on ECB 2013 survey: “*On average*, German households hold lowest wealth of all countries in the Eurozone”

Figure 1. Net wealth of median households (1000€)



Source: European Central Bank (2013).

Figure 2. Mean household net wealth (1000€)



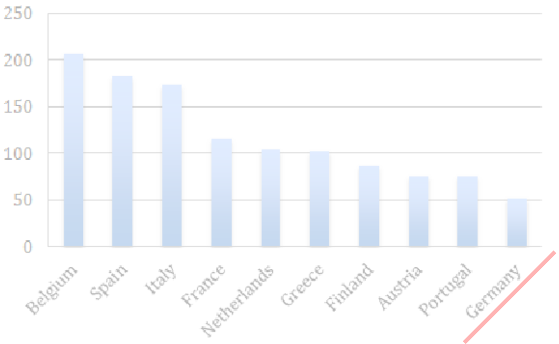
Source: European Central Bank (2013).

Inference: Germans on average are poorer than other Europeans, and therefore should not be required to contribute more than others (see, e.g., *Wall Street Journal* 2013, *Financial Times* 2013, *Frankfurter Allgemeine* 2013).

CONCEPTEST 1

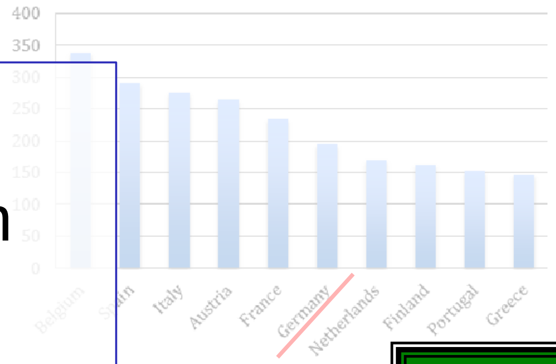
Which notion of average is the right one for this purpose?

Figure 1. Net wealth of median households (1000€)



Source: European Central Bank (2013).

Figure 2. Mean household net wealth (1000€)



Source: European Central Bank (2013).

A. Mean
B. Median
C. Mode
D. other

Time allowed: 1’30

Practical advice:

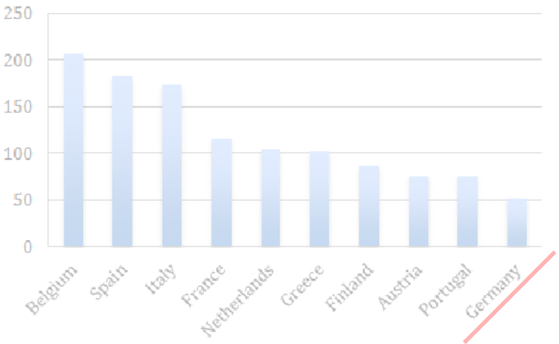
1. Log into your CANVAS account
2. Go to *Theory and Methodology of Science*
3. Select the correct week and lecture
4. Type in the access code and answer the question shown on the slide

From the next lecture onwards, answering all questions in a lecture will give you 0.5 bonus points for the exam

CONCEPTEST 1

Which notion of average is the right one for this purpose?

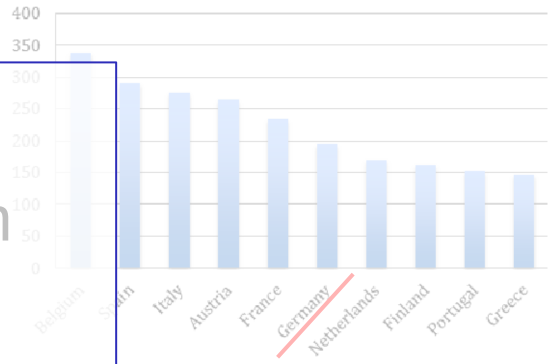
Figure 1. Net wealth of median households (1000€)



Source: European Central Bank (2013).

A. Mean
B. Median
C. Mode
D. other

Figure 2. Mean household net wealth (1000€)

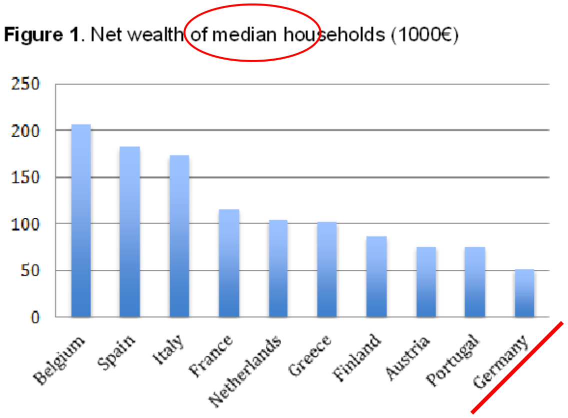


Source: European Central Bank (2013).

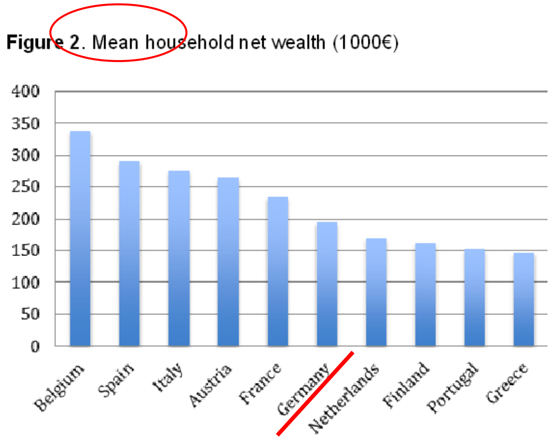
Inference: Germans on average are poorer than other Europeans, and therefore should not be required to contribute more than others (see, e.g., *Wall Street Journal* 2013, *Financial Times* 2013, *Frankfurter Allgemeine* 2013).

Ambiguous Concepts

Press reports on ECB 2013 survey: “*On average*, German households hold lowest wealth of all countries in the Eurozone”



Source: European Central Bank (2013).



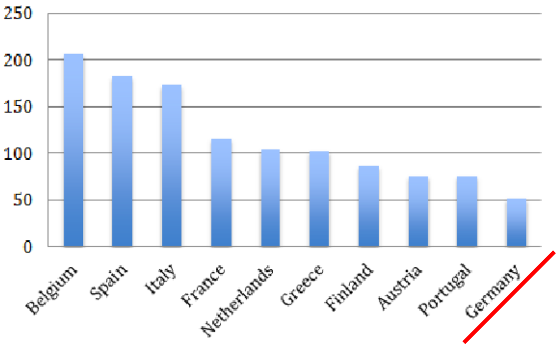
Source: European Central Bank (2013).

Alternative Purpose: Arguing for claim that Germany is the country with the most *unequal wealth distribution* in the Eurozone.

Additionally: Measurement Problem

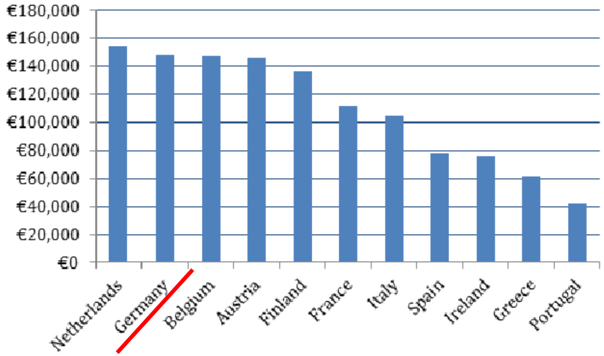
Press reports on ECB 2013 survey: “*On average*, German households hold lowest wealth of all countries in the Eurozone”

Figure 1. Net wealth of median households (1000€)

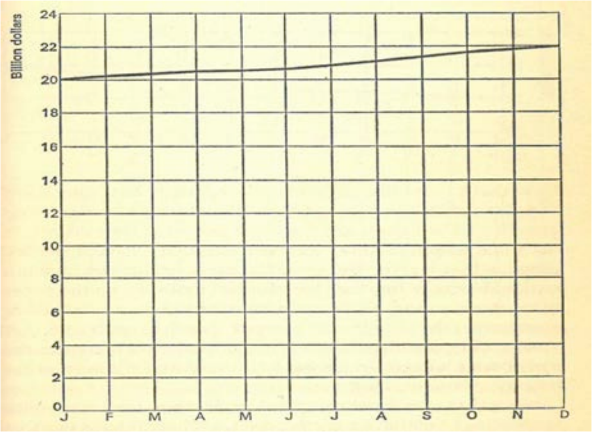


Source: European Central Bank (2013).

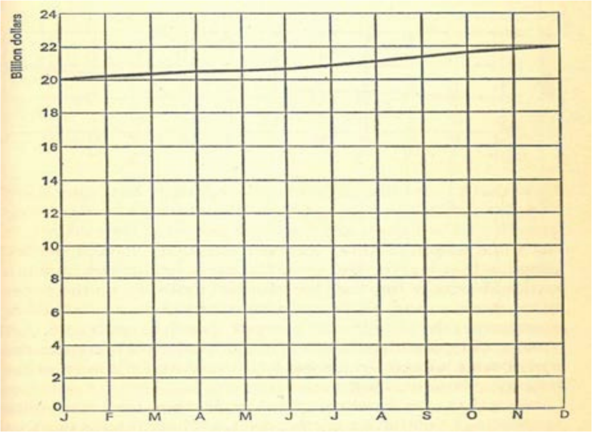
Figure 6. Total capital stock per capita (euro)

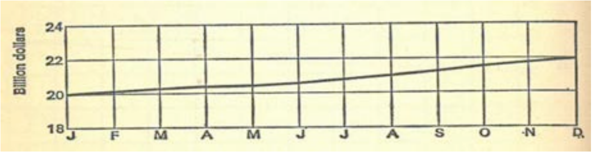


A graph showing how national income changed in a year.

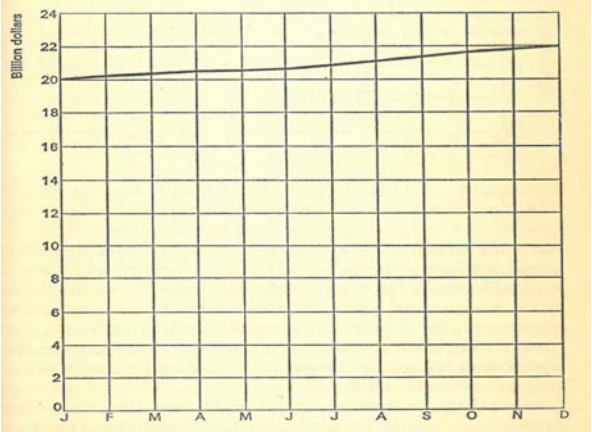


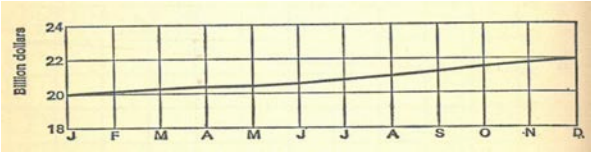
A graph showing how national income changed in a year.

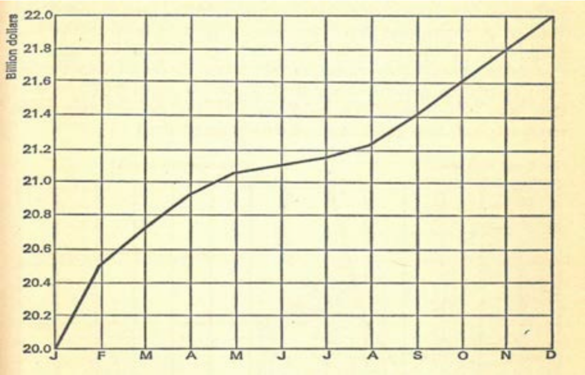




A graph showing how national income changed in a year.







Which treatment would you prefer?

A	B	C
"Our medical intervention results in a 33% relative decrease in the incidence of fatal and nonfatal myocardial infarction"	“Our medical intervention results in a 1.3 %-point decrease in the incidence of fatal and nonfatal myocardial infarction—from 3.9% to 2.6%”	“We must treat 77 persons to prevent 1 fatal or nonfatal myocardial infarction”

A
relative risk reduction
"Our medical intervention results in a 33% relative decrease in the incidence of fatal and nonfatal myocardial infarction"

A	B
relative risk reduction	**absolute risk reduction**
"Our medical intervention results in a 33% relative decrease in the incidence of fatal and nonfatal myocardial infarction"	“Our medical intervention results in a 1.3 %-point decrease in the incidence of fatal and nonfatal myocardial infarction—from 3.9% to 2.6%”

A	B
relative risk reduction	**absolute risk reduction**
"Our medical intervention results in a 33% relative decrease in the incidence of fatal and nonfatal myocardial infarction"	“Our medical intervention results in a 1.3 %-point decrease in the incidence of fatal and nonfatal myocardial infarction—from 3.9% to 2.6%”

A and B identical: a decrease of 1.3%-points from 3.9% is a 33% decrease

B	C
absolute risk reduction	**number needed to treat format**
“Our medical intervention results in a 1.3 %-point decrease in the incidence of fatal and nonfatal myocardial infarction—from 3.9% to 2.6%”	“We must treat 77 persons to prevent 1 fatal or nonfatal myocardial infarction”

B	C
absolute risk reduction	**number needed to treat format**
“Our medical intervention results in a 1.3 %-point decrease in the incidence of fatal and nonfatal myocardial infarction—from 3.9% to 2.6%”	“We must treat 77 persons to prevent 1 fatal or nonfatal myocardial infarction”

B and C identical: 1 less out of 77 is a ≈1.3%-point decrease.

A	B	C
relative risk reduction	**absolute risk reduction**	**number needed to treat format**
"Our medical intervention results in a 33% relative decrease in the incidence of fatal and nonfatal myocardial infarction"	“Our medical intervention results in a 1.3 %-point decrease in the incidence of fatal and nonfatal myocardial infarction—from 3.9% to 2.6%”	“We must treat 77 persons to prevent 1 fatal or nonfatal myocardial infarction”

A and B identical: a decrease of 1.3%-points from 3.9% is a 33% decrease

B and C identical: 1 less out of 77 is a ≈1.3%-point decrease.

Biasing Presentation

The 1995 Contraceptive Pill Scare

U.K. Committee on Safety of Medicines, 1995, to 190.000 doctors:

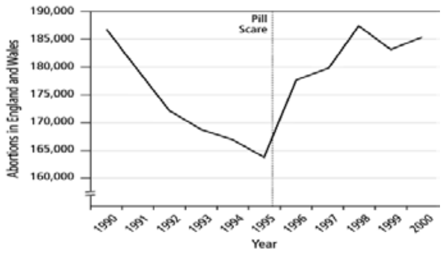
> “third-generation oral contraceptive pills increase the risk of potentially life-threatening blood clots in the legs or lungs twofold—that is, by 100%.”

Biasing Presentation

The 1995 Contraceptive Pill Scare

U.K. Committee on Safety of Medicines, 1995, to 190.000 doctors:

> "third-generation oral contraceptive pills increase the risk of potentially life-threatening blood clots in the legs or lungs twofold—that is, by 100%."



care.

- 13,000 additional abortions
- 13,000 additional births, including 800 additional pregnancies in under-16 year olds
- Increased risk of blood clots from pregnancies
- NHS additional costs £46 million

Biasing Presentation

The 1995 Contraceptive Pill Scare

U.K. Committee on Safety of Medicines, 1995, to 190.000 doctors:

> “third-generation oral contraceptive pills increase the risk of potentially life-threatening blood clots in the legs or lungs twofold—that is, by 100%.”

Warning based on studies that show:

- of every 7,000 women who took the earlier, second-generation oral contraceptive pills, about 1 had a thrombosis; this number increased to 2 among women who took third-generation pills.
- ***absolute risk*** increase 1 in 7,000, ***relative risk*** increase 100%.

Choosing Descriptive Statistical Tools

- Statistics offers a different precise notions to disambiguate commonsense concepts: median, mean, mode,…
- Statistics offers different formats for representing uncertainty: relative risk, absolute risk, natural frequencies,…
- Statistics provides a *toolbox* of descriptive concepts
- You have to *justify* the choice from the toolbox
- This depends on what you want to use the concept for

=> Statistical Reasoning

Statistical Inference:

Why Evaluate Hypotheses with Statistical Tools?

Hypothesis Testing

Inference from observable properties (in sample) to general unobservable properties

$$H \rightarrow c$$

$$\underline{\textit{observe}\ \ \text{not c}}$$

$$\textit{reject}\ \ H$$

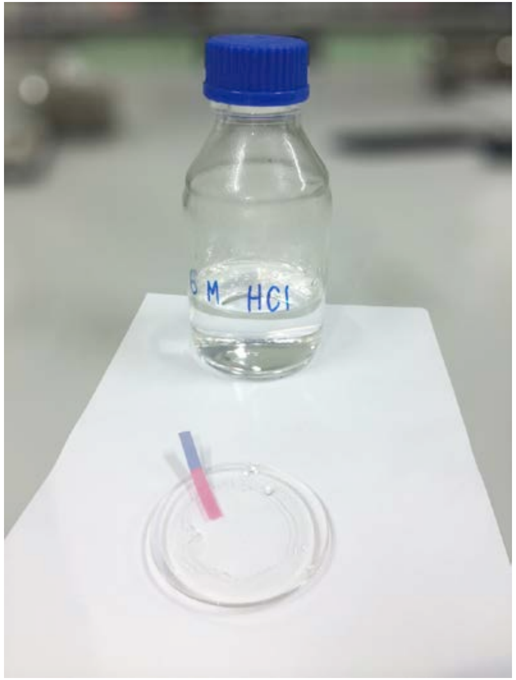
Hypothesis Testing

Inference from observable properties (in sample) to general unobservable properties

H→c *observe* not c *reject* H	H→c *observe* c *accept* H

Hypothesis Testing

Inference from observable properties (in sample) to general unobservable properties



$$H \to c$$

<u>*observe* not c</u>

reject H

$$H \to c$$

<u>*observe* c</u>

accept H

Hypothesis Testing

Inference from observable properties (in sample) to general unobservable properties

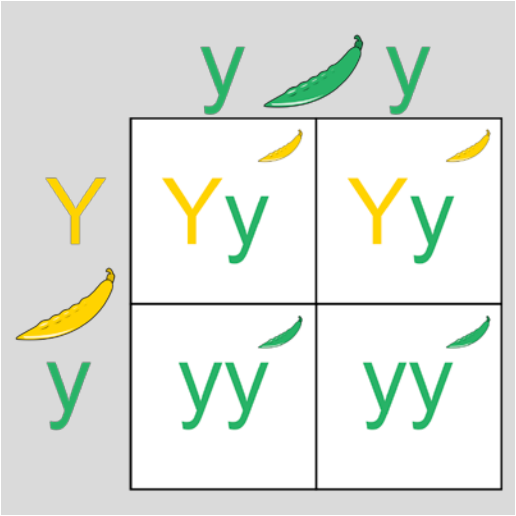
$H \rightarrow c$ *observe* <u>not c</u> *reject* H	$H \rightarrow c$ *observe* <u>c</u> *accept* H

Why evaluate a hypothesis *statistically*?

1st Reason: Stochastic Implications of H

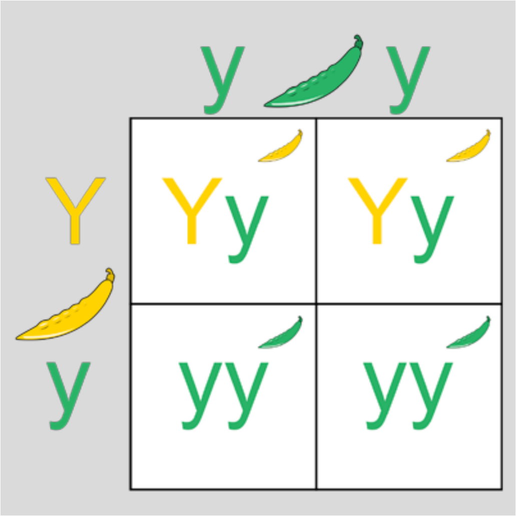
1^{st} Reason: Stochastic Implications of H

E.g. Mendelian Genetics:



1st Reason: Stochastic Implications of H

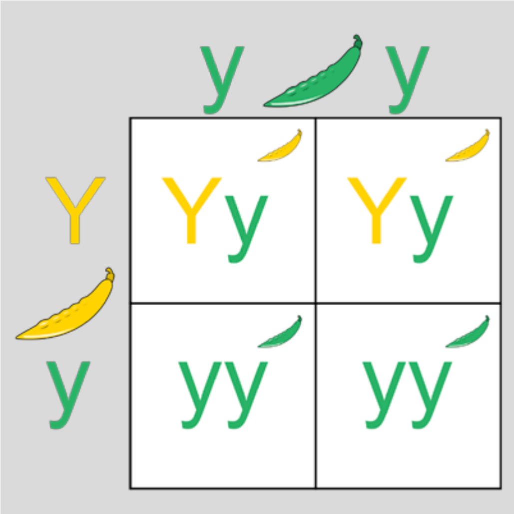
E.g. Mendelian Genetics:



	Exp 1	Exp 2	Exp 3	Exp 4	Exp 5	Exp 6	Exp 7
Parents	Yy/yy	Yy/yy	Yy/yy	Yy/yy	Yy/yy	Yy/yy	…
Offspring	Yy	yy	yy	Yy	yy	Yy	…

1st Reason: Stochastic Implications of H

E.g. Mendelian Genetics:



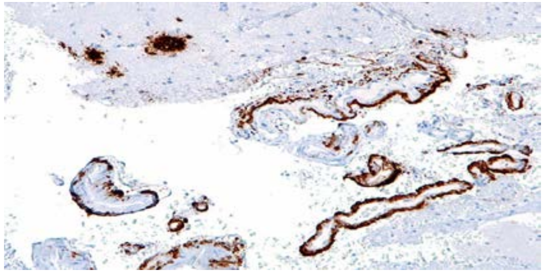
	Exp 1	Exp 2	Exp 3	Exp 4	Exp 5	Exp 6	Exp 7
Parents	Yy/yy	Yy/yy	Yy/yy	Yy/yy	Yy/yy	Yy/yy	…
Offspring	Yy	yy	yy	Yy	yy	Yy	…

Probabilistic hypotheses have distributions as observable implications. But we only observe certain instances, not distributions. Thus we need statistical tools to link single observations to distributions.

2nd Reason: Quantifying Error

2^{nd} Reason: Quantifying Error

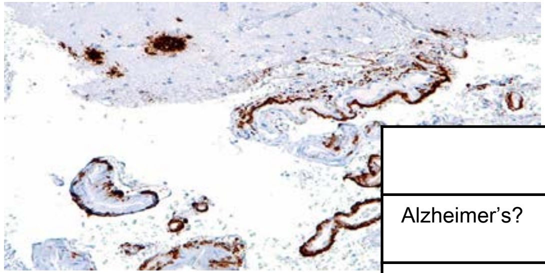
Deterministic hypotheses, e.g. “Amyloid plaque is the only cause of Alzheimer’s disease”



2nd Reason: Quantifying Error

Deterministic hypotheses, e.g. “Amyloid plaque is the only cause of Alzheimer’s disease”

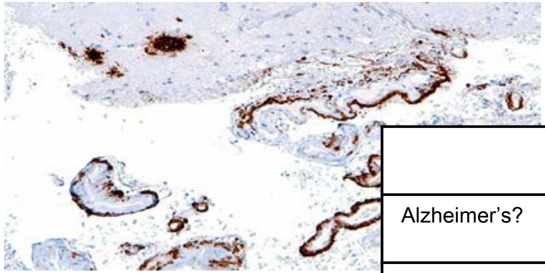
Amyloid plaque identified through tissue staining with enzymes. These enzymes are capable of catalyzing reactions that give a colored product. This is detectable by light microscopy.



	Obs 1	**Obs 2**	**Obs 3**	**Obs 4**	**…**	**Obs n**
Alzheimer’s?	Yes	No	No	Yes	…	No
Amyloid plaque?	Yes	No	Yes	No	…	No

2^{nd} Reason: Quantifying Error

Deterministic hypotheses, e.g. "Amyloid plaque is the only cause of Alzheimer's disease"

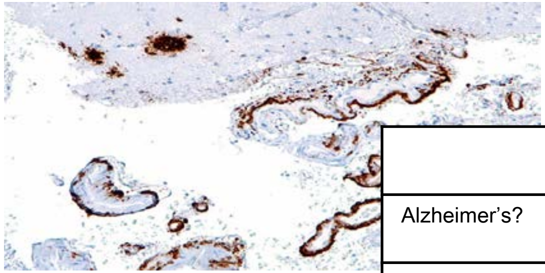


Amyloid plaque identified through tissue staining with enzymes. These enzymes are capable of catalyzing reactions that give a colored product. This is detectable by light microscopy.

	Obs 1	Obs 2	Obs 3	Obs 4	…	Obs n
Alzheimer's?	Yes	No	No	Yes	…	No
Amyloid plaque?	Yes	No	Yes	No	…	No

2nd Reason: Quantifying Error

Deterministic hypotheses, e.g. “Amyloid plaque is the only cause of Alzheimer’s disease”



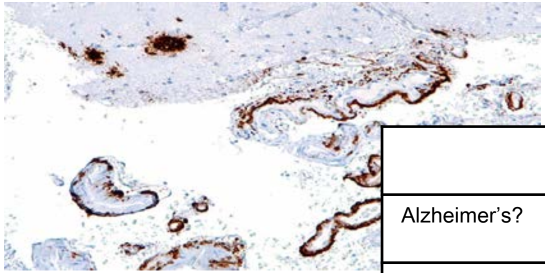
Amyloid plaque identified through tissue staining with enzymes. These enzymes are capable of catalyzing reactions that give a colored product. This is detectable by light microscopy.

	Obs 1	**Obs 2**	**Obs 3**	**Obs 4**	**…**	**Obs n**
Alzheimer’s?	Yes	No	No	Yes	…	No
Amyloid plaque?	Yes	No	Yes	No	…	No

Observations contain random measurement error. A single (seemingly) falsifying observation might thus not be a good reason to reject hypothesis

2nd Reason: Quantifying Error

Deterministic hypotheses, e.g. “Amyloid plaque is the only cause of Alzheimer’s disease”



Amyloid plaque identified through tissue staining with enzymes. These enzymes are capable of catalyzing reactions that give a colored product. This is detectable by light microscopy.

	Obs 1	Obs 2	Obs 3	Obs 4	…	Obs n
Alzheimer’s?	Yes	No	No	Yes	…	No
Amyloid plaque?	Yes	No	Yes	No	…	No

Observations contain random measurement error. A single (seemingly) falsifying observation might thus not be a good reason to reject hypothesis

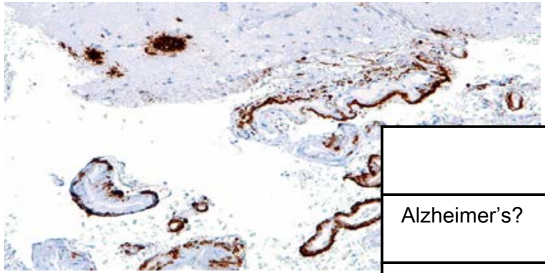
Deduce observable consequences C from H, in conjunction with auxiliary hypothesis AH

$$H \ \& \ AH \rightarrow C$$

2^{nd} Reason: Quantifying Error

<u>Deterministic hypotheses, e.g. “Amyloid plaque is the only cause of Alzheimer’s disease”</u>

Amyloid plaque identified through tissue staining with enzymes. These enzymes are capable of catalyzing reactions that give a colored product. This is detectable by light microscopy.



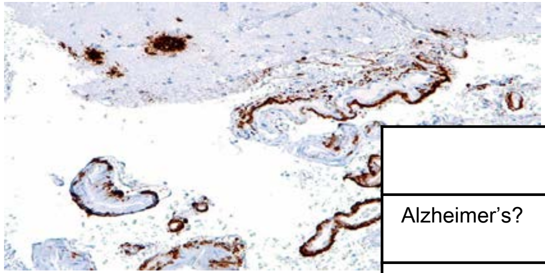
	Obs 1	**Obs 2**	**Obs 3**	**Obs 4**	**…**	**Obs n**
Alzheimer’s?	Yes	No	No	Yes	…	No
Amyloid plaque?	Yes	No	Yes	No	…	No

- Quantifying error:
  - how probable is it to *not* observe plaque, even if there is Amyloid plaque?
  - how probable is it to observe plaque, even if there is *no* Amyloid plaque?

2^{nd} Reason: Quantifying Error

Deterministic hypotheses, e.g. “Amyloid plaque is the only cause of Alzheimer’s disease”

Amyloid plaque identified through tissue staining with enzymes. These enzymes are capable of catalyzing reactions that give a colored product. This is detectable by light microscopy.

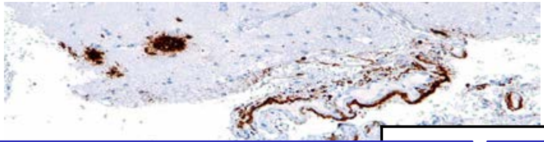


	Obs 1	**Obs 2**	**Obs 3**	**Obs 4**	**…**	**Obs n**
Alzheimer’s?	Yes	No	No	Yes	…	No
Amyloid plaque?	Yes	No	Yes	No	…	No

- Is the probability of error sufficiently small?
  - Yes: Reject H
  - No: Do not reject H

2nd Reason: Quantifying Error

<u>Deterministic hypotheses, e.g. “Amyloid plaque is the only cause of Alzheimer’s disease”</u>



Amyloid plaque identified through tissue staining with enzymes. These enzymes are capable of catalyzing reactions that give a colored product. This is detectable by light microscopy.

Obs 4	…	Obs n
Yes	…	No
No	…	No

… e Amyloid plaque, even if

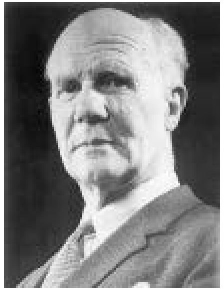
… | *No Alzheimer's) < c*

Fisher’s Null Hypothesis Testing

Ronald A. Fisher 1890-1962

Neyman-Pearson Decision Theory

Jerzy Neyman 1894-1981



Egon Pearson 1895-1980

3rd Reason: Quantifying Confidence

Observing truth of any implication of H does not justify concluding that H is true, but only that we are *more confident* in H

- How probable is the hypothesis given the observed data?
- That depends on the difference between the probability of observing the data, and the probability of observing the data *given* H is true

3rd Reason: Quantifying Confidence

Observing truth of any implication of H does not justify concluding that H is true, but only that we are *more confident* in H

- pothesis given the observed data?
- *imer's | Amyloid plaque observed)*
- fference between the probability of observing bility of observing the data *given* H is true

Bayesian Statistics

Leonard Jimmie Savage 1917-71

Statistical Inferences

Statistical Inferences

Statistical tests of deterministic hypotheses a ***weakening*** of (non-statistical) accounts of confirmation and falsification

- not every observation compatible with hypotheses yields full confirmation
- not every observation contradicting hypothesis gives reason for rejection

Statistical Inferences

Statistical tests of deterministic hypotheses a ***weakening*** of (non-statistical) accounts of confirmation and falsification

- not every observation if instances of hypotheses yields full confirmation
- not every observation contradicting hypothesis gives reason for rejection

Often for *legitimate reasons*:

- Random error argument against immediate rejection
- Degrees of confirmation

But weakening also opens new possibilities of *abuse*!

Statistical Inferences

Statistical tests of deterministic hypotheses a ***weakening*** of (non-statistical) accounts of confirmation and falsification

Austin Bradford Hill
(1897 –1991)

Yet I cannot find anywhere I thought it necessary to use a test of significance. The evidence was so clear cut, the differences between the groups were mainly so large, the contrast between respiratory and non-respiratory causes of illness so specific, that no formal tests could really contribute anything of value to the argument. So why use them?

Hill 1965, The Environment and Disease: Association or Causation? 299


Ronald A. Fisher
1890-1962

Fisher’s Significance Testing

Fisher’s Significance Testing - Example

1. Specify the main hypothesis *H*

    e.g. "This is a fair coin"

Fisher's Significance Testing - Example

1. Specify the main hypothesis *H*
2. Devise an experiment to test *H*, and specify its possible outcomes (the so-called "test statistic").

e.g. Tossing the coin 20 times.

Possible outcomes:

TTTTTTTTTTTTTTTTTTTT, } *0 Heads:* 1 possibility

HTTTTTTTTTTTTTTTTTTT,
THTTTTTTTTTTTTTTTTTT
TTHTTTTTTTTTTTTTTTTT
…
TTTTTTTTTTTTTTTTTTTH, } *1 Heads:* 20 possibilities

HHTTTTTTTTTTTTTTTTTT,
HTHTTTTTTTTTTTTTTTTT,
…
TTTTTTTTTTTTTTTTTTHH, } *2 Heads:* 190 possibilities

…

…

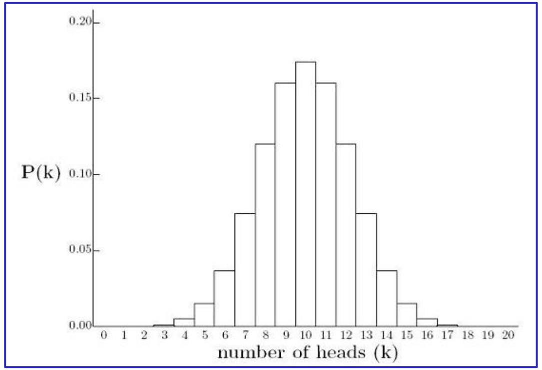
HHHHHHHHHHHHHHHHHHHH } *20 Heads:* 1 possibility

Fisher’s Significance Testing - Example

1. Specify the main hypothesis *H*
2. Devise an experiment to test *H*, and specify its possible outcomes (the so-called “test statistic”).

e.g. Tossing the coin 20 times.

Possible outcomes:



TTTTTTTTTTTTTTTTTTTT, } *0 Heads:* 1 possibility

HTTTTTTTTTTTTTTTTTTT,
THTTTTTTTTTTTTTTTTTT
TTHTTTTTTTTTTTTTTTTT
…
TTTTTTTTTTTTTTTTTTTH, } *1 Heads:* 20 possibilities

HHTTTTTTTTTTTTTTTTTT,
HTHTTTTTTTTTTTTTTTTT,
…
TTTTTTTTTTTTTTTTTTHH, } *2 Heads:* 190 possibilities

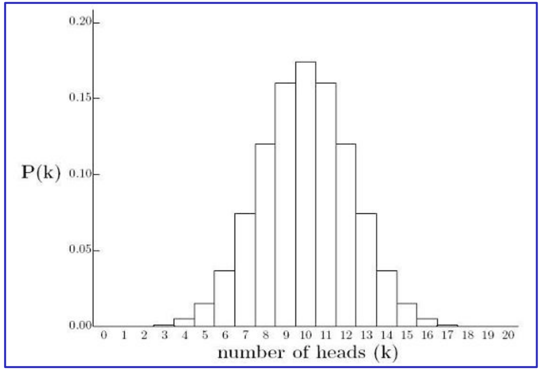
…

…

HHHHHHHHHHHHHHHHHHHH } *20 Heads:* 1 possibility

Fisher’s Significance Testing - Example

1. Specify the main hypothesis *H*
2. Devise an experiment to test *H*, and specify its possible outcomes (the so-called “test statistic”).
3. Determine the distribution of the test statistic, under the assumption that *H* is true.



The Probabilities of Obtaining *r* Heads in a Trial consisting of 20 Tosses of a Fair Coin

Number of Heads (r)	*Probability*	*Number of Heads* (r)	*Probability*
0	9×10^{-7}	11	0.1602
1	1.9×10^{-5}	12	0.1201
2	2×10^{-4}	13	0.0739
3	0.0011	14	0.0370
4	0.0046	15	0.0148
5	0.0148	16	0.0046
6	0.0370	17	0.0011
7	0.0739	18	2×10^{-4}
8	0.1201	19	1.9×10^{-5}
9	0.1602	20	9×10^{-7}
10	0.1762		

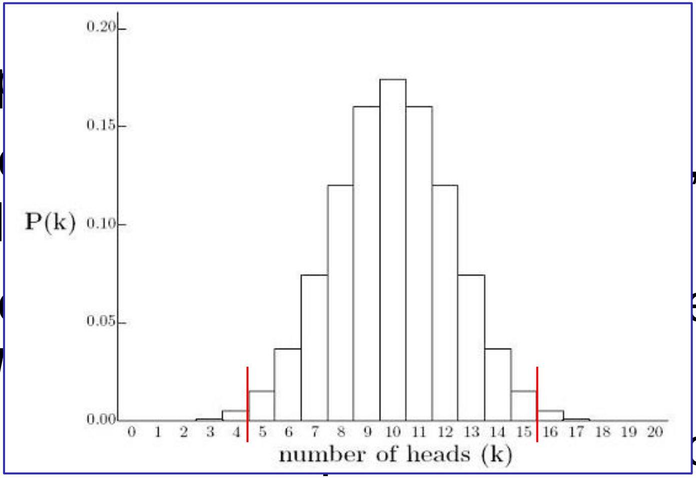
Fisher’s Significance Testing - Example

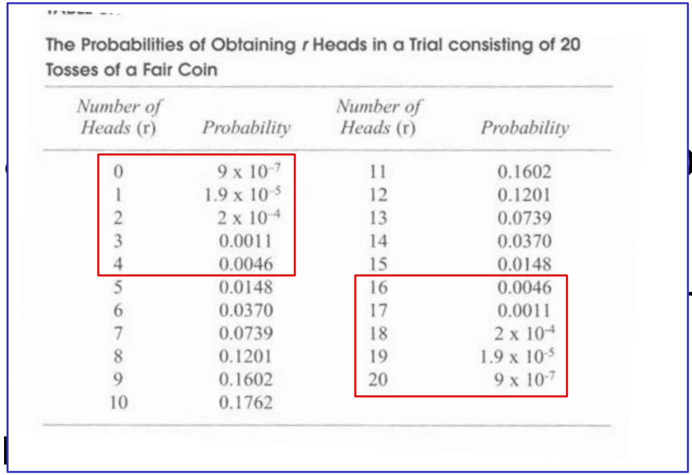
1. Specify the main hypothesis *H*
2. Devise an experiment to test *H*, and specify its possible outcomes (the so-called “test statistic”).
3. Determine the distribution of the test statistic, under the assumption that *H* is true.
4. Observe the experimental outcome.

e.g. 4 heads, 16 tails

Fisher’s Significance Testing - Example

1. Sp
2. D ... , ... mes (the so-cal
3. D ... e ... mption that *H*
4. O ... o





The Probabilities of Obtaining *r* Heads in a Trial consisting of 20 Tosses of a Fair Coin

Number of Heads (r)	Probability	Number of Heads (r)	Probability
0	9×10^{-7}	11	0.1602
1	1.9×10^{-5}	12	0.1201
2	2×10^{-4}	13	0.0739
3	0.0011	14	0.0370
4	0.0046	15	0.0148
5	0.0148	16	0.0046
6	0.0370	17	0.0011
7	0.0739	18	2×10^{-4}
8	0.1201	19	1.9×10^{-5}
9	0.1602	20	9×10^{-7}
10	0.1762		

5. Calculate the *p-value*. This is the probability of observing a result at least as extreme as the one observed, given the hypothesis is true.

   Results at least as extreme as “4 heads, 16 tails” are r = 4,3,2,1,0 and r = 16,17,18,19,20

   The probability of any of them occurring (sum respective probabilities) is p^* = 0.012

Fisher’s Significance Testing - Example

1. Specify the main hypothesis *H*
2. Devise an experiment to test *H*, and specify its possible outcomes (the so-called “test statistic”).
3. Determine the distribution of the test statistic, under the assumption that *H* is true.
4. Observe the experimental outcome.
5. Calculate the *p-value*. This is the probability of observing a result at least as extreme as the one observed, given the hypothesis is true.
6. If the p-value is smaller than a conventionally set *significance level* (typically 0.05, but sometimes also 0.01 or 0.001), reject *H*

p^* = 0.012 < 0.05, hence experimental result is significant. Reject H

How to Lie with Significance Testing

1. Specify the main hypothesis *H*
2. Devise an experiment to test *H*, and specify its possible outcomes (the so-called "test statistic").
3. Determine the distribution of the test statistic, under the assumption that *H* is true.
4. Observe the experimental outcome.
5. Calculate the *p-value*. This is the probability of observing a result at least as extreme as the one observed, given the hypothesis is true.
6. If the p-value is smaller than a conventionally set *significance level* (typically 0.05, but sometimes also 0.01 or 0.001), reject *H*

How to Lie with Significance Testing

Which hypothesis?

1. Specify the main hypothesis *H*
2. Devise an experiment to test *H*, and specify its possible outcomes (the so-called “test statistic”).
3. Determine the distribution of the test statistic, under the assumption that *H* is true.
4. Observe the experimental outcome.
5. Calculate the *p-value*. This is the probability of observing a result at least as extreme as the one observed, given the hypothesis is true.
6. If the p-value is smaller than a conventionally set *significance level* (typically 0.05, but sometimes also 0.01 or 0.001), reject *H*

How to Lie with Significance Testing

Which hypothesis?

What experiment?

1. Specify the main hypothesis *H*
2. Devise an experiment to test *H*, and specify its possible outcomes (the so-called “test statistic”).
3. Determine the distribution of the test statistic, under the assumption that *H* is true.
4. Observe the experimental outcome.
5. Calculate the *p-value*. This is the probability of observing a result at least as extreme as the one observed, given the hypothesis is true.
6. If the p-value is smaller than a conventionally set *significance level* (typically 0.05, but sometimes also 0.01 or 0.001), reject *H*

How to Lie with Significance Testing

Which hypothesis?

1. Specify the main hypothesis *H*

What experiment?

2. Devise an experiment to test *H*, and specify its possible outcomes (the so-called “test statistic”).

How to partition possibilities?

3. Determine the distribution of the test statistic, under the assumption that *H* is true.
4. Observe the experimental outcome.
5. Calculate the *p-value*. This is the probability of observing a result at least as extreme as the one observed, given the hypothesis is true.
6. If the p-value is smaller than a conventionally set *significance level* (typically 0.05, but sometimes also 0.01 or 0.001), reject *H*

How to Lie with Significance Testing

Which hypothesis?

1. Specify the main hypothesis *H*

What experiment?

2. Devise an experiment to test *H*, and specify its possible outcomes (the so-called “test statistic”).

How to partition possibilities?

3. Determine the distribution of the test statistic, under the assumption that *H* is true.

What sample distribution?

4. Observe the experimental outcome.
5. Calculate the *p-value*. This is the probability of observing a result at least as extreme as the one observed, given the hypothesis is true.
6. If the p-value is smaller than a conventionally set *significance level* (typically 0.05, but sometimes also 0.01 or 0.001), reject *H*

How to Lie with Significance Testing

Which hypothesis?

1. Specify the main hypothesis *H*

What experiment?

2. Devise an experiment to test *H*, and specify its possible outcomes (the so-called “test statistic”).

How to partition possibilities?

3. Determine the distribution of the test statistic, under the assumption that *H* is true.

What sample distribution?

4. Observe the experimental outcome.

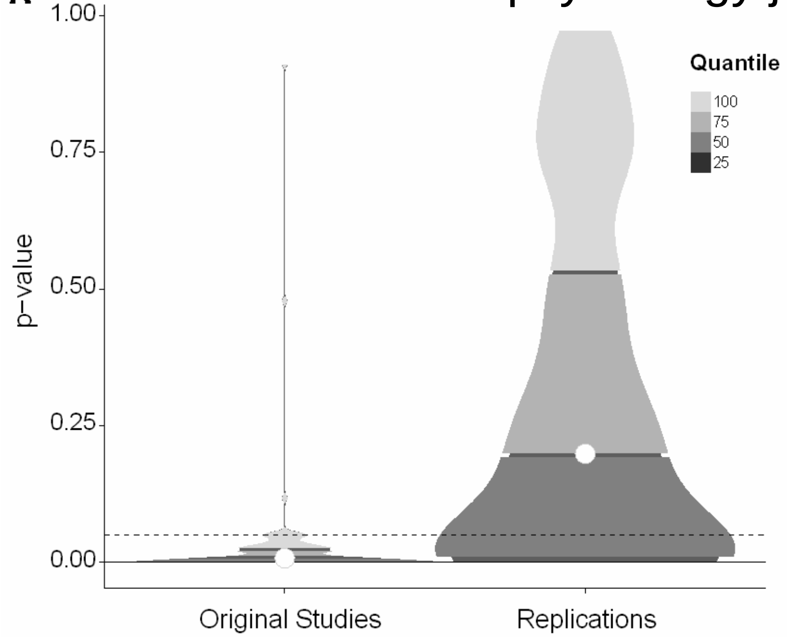
5. Calculate the *p-value*. This is the probability of observing a result at least as extreme as the one observed, given the hypothesis is true.

6. If the p-value is smaller than a conventionally set *significance level* (typically 0.05, but sometimes also 0.01 or 0.001), reject *H*

What significance level?

Evidence for p-value Abuse

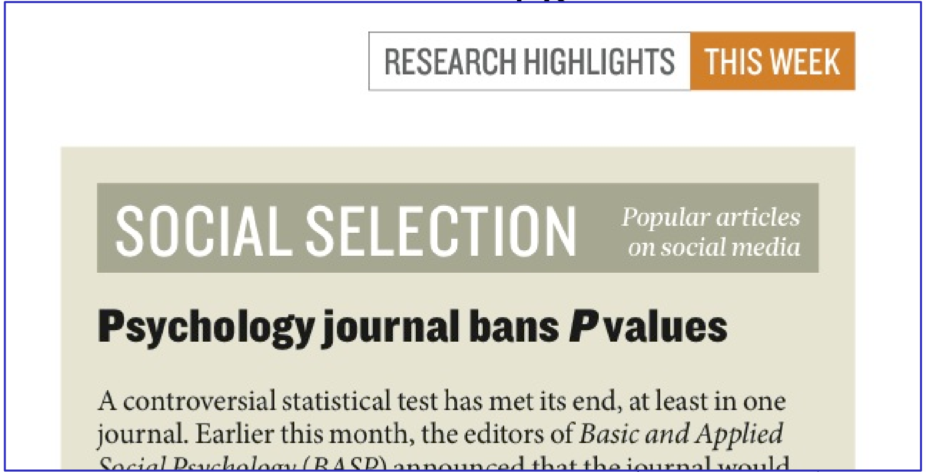
Reproductions of 100 experimental and correlational studies published in three main psychology journals.



Open Science Collaboration. (2015). *Science*.

Evidence for p-value Abuse

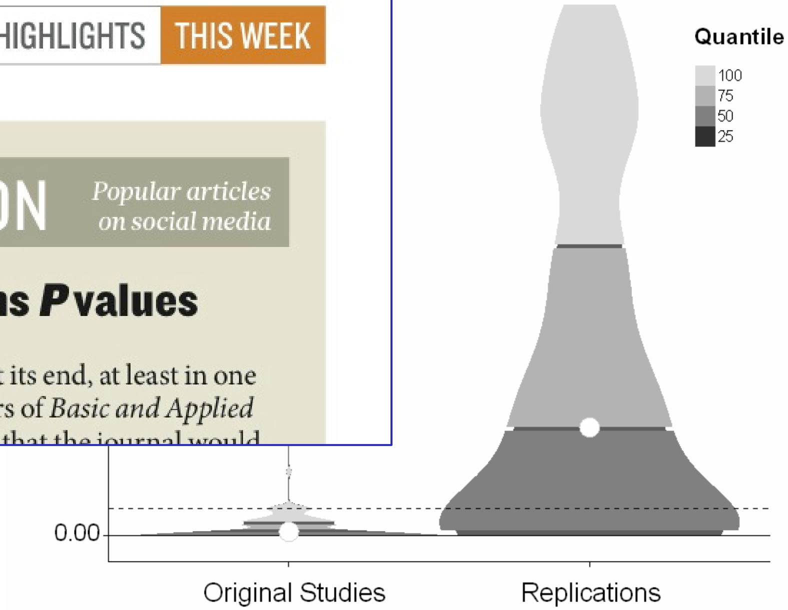
Reproductions of 100 experimental and correlational studies published in three main psychology journals.


RESEARCH HIGHLIGHTS THIS WEEK

SOCIAL SELECTION *Popular articles on social media*

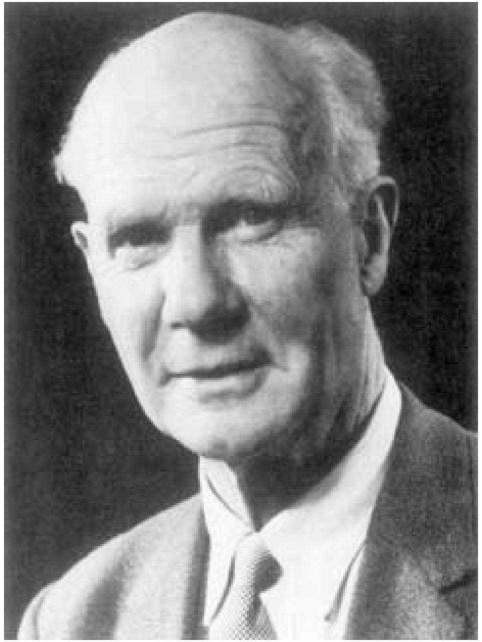
Psychology journal bans *P* values

A controversial statistical test has met its end, at least in one journal. Earlier this month, the editors of *Basic and Applied Social Psychology* (*BASP*) announced that the journal would



Open Science Collaboration. (2015). *Science*.

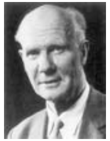
Neyman-Pearson Hypothesis Testing



Egon Pearson 1895-1980

Jerzy Neyman 1894-1981

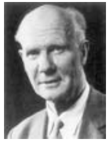
Neyman-Pearson Hypothesis Testing

H_0 : “The coin is not fair”

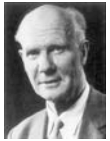
H_a: “The coin is fair”

Neyman-Pearson Hypothesis Testing



	H_0 true	**H_0 false**
Accept H_0	Correct decision	Type II error
Reject H_0	Type I error	Correct decision

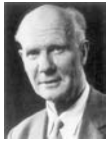
Neyman-Pearson Hypothesis Testing



	H_0 true	**H_0 false**
Accept H_0	Correct decision	Type II error
Reject H_0	Type I error	Correct decision

Power of a test: the probability that the test correctly rejects the null hypothesis (H_0) when a specific alternative hypothesis (H_a) is true

Neyman-Pearson Hypothesis Testing



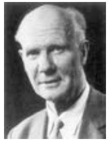

	H_0 true	H_0 false
Accept H_0	Correct decision	Type II error
Reject H_0	Type I error	Correct decision

Power of a test: the probability that the test correctly rejects the null hypothesis (H_0) when a specific alternative hypothesis (H_a) is true

Depends on:

- the type-I error rate set for the test
- the magnitude of the effect of interest in the population
- the sample size used to detect the effect

Summary

- Fisher’s Significance Testing
- p-value, significance level
- Ways how to manipulate testing procedure
- Evidence for p-value abuse
- Neyman-Pearson Hypothesis Testing
- Type-I and Type-II errors

Leonard J. Savage
1917-71

Bayesian Statistics

Bayesian Statistics - Example

1. Determine the set of competing hypotheses ($H_1, \ldots, H_n$).

   e.g. H_1: *"Coin is fair"*, H_2: *"Coin is not fair"*

Bayesian Statistics - Example

1. Determine the set of competing hypotheses ($H_1, \ldots, H_n$).
2. Determine *prior probabilities* for each H_i

   e.g. p(*"coin is fair"*) = *0.7* , *p*(H_2) = *0.3*

Bayesian Statistics - Example

1. Determine the set of competing hypotheses ($H_1,\ldots,H_n$).
2. Determine *prior probabilities* for each H_i
3. Interpret probabilities as “degrees of belief”, “subjective confidence in hypothesis”

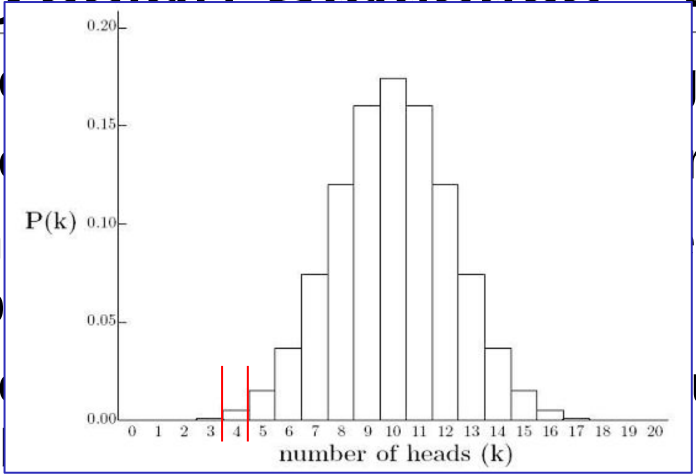
Bayesian Statistics - Example

1. Determine the set of competing hypotheses ($H_1,\ldots,H_n$).
2. Determine *prior probabilities* for each H_i
3. Interpret probabilities as “degrees of belief”, “subjective confidence in hypothesis”
4. Collect new data that were not used in computing the prior probabilities.

e..g. throw coin 20 times, observe 4 heads, 16 tails

Bayesian Statistics - Example

1. D… l
2. D… r
3. In… e… dence in hyp
4. C… u… probal



The Probabilities of Obtaining *r* Heads in a Trial consisting of 20 Tosses of a Fair Coin

Number of Heads (r)	*Probability*	*Number of Heads (r)*	*Probability*
0	9×10^{-7}	11	0.1602
1	1.9×10^{-5}	12	0.1201
2	2×10^{-4}	13	0.0739
3	0.0011	14	0.0370
4	0.0046	15	0.0148
5	0.0148	16	0.0046
6	0.0370	17	0.0011
7	0.0739	18	2×10^{-4}
8	0.1201	19	1.9×10^{-5}
9	0.1602	20	9×10^{-7}
10	0.1762		

5. Determine the likelihood of the data conditional on hypotheses, *$p(E|H_i)$.*

The probability of “4 heads, 16 tails”, given that H_0 is true, is 0.0046.

Bayesian Statistics - Example

1. Determine the set of competing hypotheses ($H_1,\ldots,H_n$).
2. Determine *prior probabilities* for each H_i
3. Interpret probabilities as “degrees of belief”, “subjective confidence in hypothesis”
4. Collect new data that were not used in computing the prior probabilities.
5. Determine the likelihood of the data conditional on hypotheses, $p(E|H_i)$.
6. Calculate the posterior probability $p(H_i|E)$ for each hypothesis, using Bayes Rule:

$$p(H_1|E) = \frac{p(E|H_1)*p(H_1)}{p(E)} = \frac{p(E|H_1)*p(H_1)}{p(E|H_1)*p(H_1)+p(E|-H_1)*p(-H_1)} = \frac{0.0046*0.7}{0.0046*0.7+0.53*0.3}$$

$$\approx 0.02$$

Bayesian Statistics - Example

1. Determine the set of competing hypotheses ($H_1,\ldots,H_n$).
2. Determine *prior probabilities* for each H_i
3. Interpret probabilities as “degrees of belief”, “subjective confidence in hypothesis”
4. Collect new data that were not used in computing the prior probabilities.
5. Determine the likelihood of the data conditional on hypotheses, $p(E|H_i)$.
6. Calculate the posterior probability $p(H_i|E)$ for each hypothesis, using Bayes Rule

7. Update prior probability: $p'(\text{“coin is fair”}) = p'(H_1) = p(H_1|E) \approx 0.02$.

Bayesian Statistics - Example

1. Determine the set of competing hypotheses ($H_1, \ldots, H_n$).

Many hypotheses

2. Determine *prior probabilities* for each H_i
3. Interpret probabilities as “degrees of belief”, “subjective confidence in hypothesis”
4. Collect new data that were not used in computing the prior probabilities.
5. Determine the likelihood of the data conditional on hypotheses, $p(E|H_i)$.
6. Calculate the posterior probability $p(H_i|E)$ for each hypothesis, using Bayes Rule

7. Update prior probability: $p'(\text{“coin is fair”}) = p'(H_1) = p(H_1|E) \approx 0.02$.

Bayesian Statistics - Example

1. Determine the set of competing hypotheses ($H_1,\ldots,H_n$).
2. Determine *prior probabilities* for each H_i
3. Interpret probabilities as “degrees of belief”, “subjective confidence in hypothesis”
4. Collect new data that were not used in computing the [illegible] probabilities.
5. Determine the likelihood of the data conditional on hypotheses, $p(E|H_i)$.
6. Calculate the posterior probability $p(H_i|E)$ for each hypothesis, using Bayes Rule

7. Update prior probability: $p'(\text{“coin is fair”}) = p'(H_1) = p(H_1|E) \approx 0.02$.

Many hypotheses

Probabilities as personal beliefs

Bayesian Statistics - Example

1. Determine the set of competing hypotheses ($H_1, \ldots, H_n$).
2. Determine *prior probabilities* for each H_i
3. Interpret probabilities as "degrees of belief", "subjective confidence in hypothesis"
4. Collect new data that were not used in computing the probabilities.
5. Determine the likelihood of the data conditional on hypotheses, $p(E|H_i)$.
6. Calculate the posterior probability $p(H_i|E)$ for each hypothesis, using Bayes Rule
7. Update prior probability: $p'(\text{"coin is fair"}) = p'(H_1)$ =

Many hypotheses

Probabilities as personal beliefs

Main goal: assigning probability to hypotheses

Problems of Bayesian Statistics

Problem of determining priors

Solutions:

1. Subjectivist: Any prior other than 0 or 1 is ok because in the limit of iterative updating all agents will converge on the same probability anyway.

1. Objectivist: Before any evidence is gathered we should divide our belief equally among the mutually exclusive but jointly exhaustive outcomes.

Problems of Bayesian Statistics

<u>Problem of old evidence:</u>

E has a confirming (or disconfirming) effect on a hypothesis only when *E* is first determined to be true.

What if we already know that *E* is true for a while, and then learn that *E* is evidence for a hypothesis?

Problems of Bayesian Statistics

Problem of uncertain evidence:

Bayes theorem works via P(E) going to 1, but it is implausible that we are ever certain of anything. In particular, sometimes we find out that our evidence is false. Standard Bayesianism does not allow for this.

Statistical Toolbox & Statistical Reasoning

Toolbox Content

- Average concepts
- Graphs
- Uncertainty formats
- Significance tests
- Error statistics
- Bayesian inference

Statistical Reasoning Requirements

- Avoid equivocations
- Be mindful of possible biases in presentation
- Be clear about purposes of your hypothesis assessment
- Don’t exploit ambiguities
- Take a coherent view on probabilities