



WIRTSCHAFTS  
UNIVERSITÄT  
WIEN VIENNA  
UNIVERSITY OF  
ECONOMICS  
AND BUSINESS

**Thomas Salzberger, PD Dr.**  
**thomas.salzberger@wu.ac.at**



## Quantitative Research Methods

<Analysis>

Hypothesis Testing

**Conditional** Probabilities

*& Missing Data*

# Missing data

# Missing data

- **Missing completely at random (MCAR)**
  - Missing values are neither related to unobservable nor to observable data
  - E.g., whether a study participant leaves a question on satisfaction empty is entirely due to chance; it is not related to any person characteristic, nor to the level of satisfaction (or any other latent property)
  - No imputation required
  - Unrealistic assumption in most cases
  - With multi-item scales, imputation of single missing values may still be beneficial
    - Whole case might be missing otherwise

# Missing data

- **Missing at random (MAR)**
  - Missingness is due to a variable that has been observed
  - E.g., whether a study participant leaves a question on satisfaction empty is entirely a function of gender; it is not related to any unobserved person characteristic, nor to the level of satisfaction (or any other latent property)
  - Full Information Maximum Likelihood Estimation applicable
    - Estimates parameters for which observed values are most plausible accounting for the fact that there are missing values (no imputation required)

# Missing data

- **Missing not at random (MNAR)**

- Missingness is due to a variable that has not been observed
- E.g., whether a study participant leaves a question on satisfaction empty is a function of the level of satisfaction or any other personal property that is related to satisfaction
- High risk of bias
- Listwise deletion (i.e. persons with missing values are excluded completely) reduces statistical power (smaller sample size) and may not overcome risk of bias
- Pairwise deletion (e.g. correlations as input to factor analysis) uses all available data, but problem of unequal sample sizes for correlations, factor score computation problematic

- *Note: in measurement, Item Response Theory (IRT) and Rasch Measurement Theory/Methods (RMT) do not require complete data*

# Missing data

- **Missing not at random (MNAR)**

- Imputation

- Replace by mean score (mean across sample) reduces variance
- Replace by mean score across the same person (scores on other variables in a multi-item scale) is better
- Predicting value based on other observable data (regression analysis)
- Multiple imputation (impute missing value by a distribution of possible values rather than a single value), yields multiple samples

Whole field of scientific enquiry

# Hypothesis Testing

- Exemplified by a two group mean comparison (t-Test)
- Weather forecast
  - For a generally sunny place, a metrological institute predicts rain for tomorrow.
  - At the place, there is no rain (“sun”) on 9 out of 10 days.
  - The institute has the following track record:
    - When there is sun, the institute has predicted sun in 90% of the cases.
    - When there is rain, the institute has predicted rain in 80% of the cases.
  - How likely will there be rain tomorrow?

# Hypothesis Testing

Prediction	
Sun	Null hypothesis (no signal)
Rain	Alternative hypothesis

- Rain corresponds to “correct theory”
- Here, we also know that there is only 1 in 10 days where there is rain. (Unconditional, or a priori, probability of rain is just 10%.)



# Hypothesis Testing

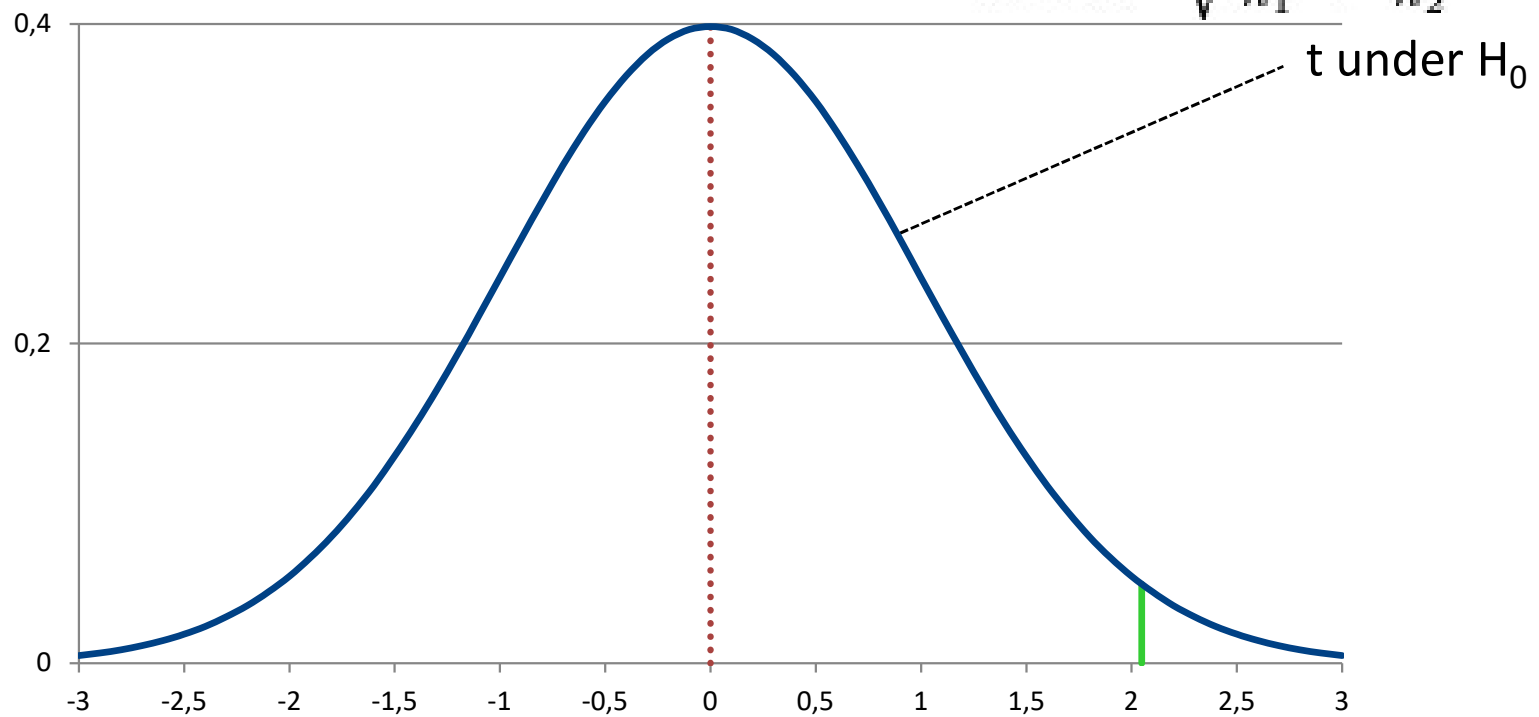
- Null hypothesis  $H_0$ : no difference, no relationship in the total population
  - *could, in principle, also be any specific value*
  - *but in the vast majority of cases no justifiable value other than 0*
- Alternative hypothesis  $H_A$  ( $H_1$ ): difference/relationship in the total population
- E.g., mean comparison, two groups (t-test)
  - $H_0: \mu_1 = \mu_2$  or  $\mu_1 - \mu_2 = 0$
  - $H_A: \mu_1 \neq \mu_2$  or  $\mu_1 - \mu_2 \neq 0$
- Special case one-tailed hypothesis testing
  - $H_0: \mu_1 \leq \mu_2$  or  $\mu_1 - \mu_2 \leq 0$
  - $H_A: \mu_1 > \mu_2$  or  $\mu_1 - \mu_2 > 0$

$$t = \frac{\bar{X}_1 - \bar{X}_2}{S_{X_1 X_2} \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

## Distribution of t under $H_0$

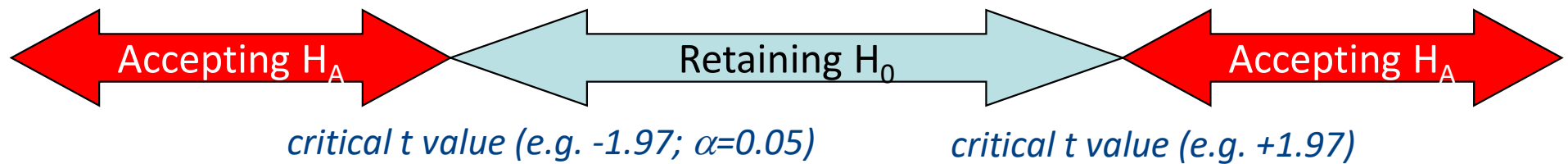
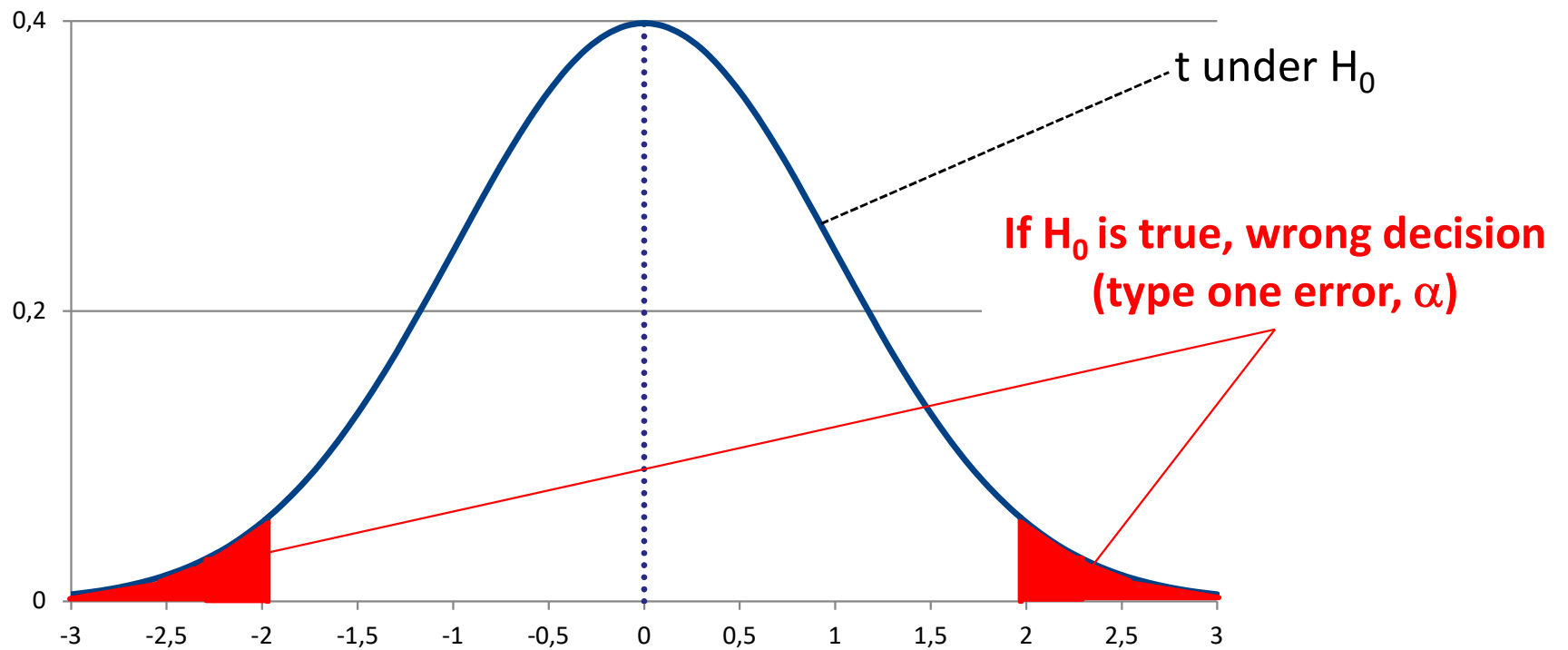
*Two groups (male, female)  
Measurement of satisfaction*

$$t = \frac{\bar{X}_1 - \bar{X}_2}{S_{X_1 X_2} \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

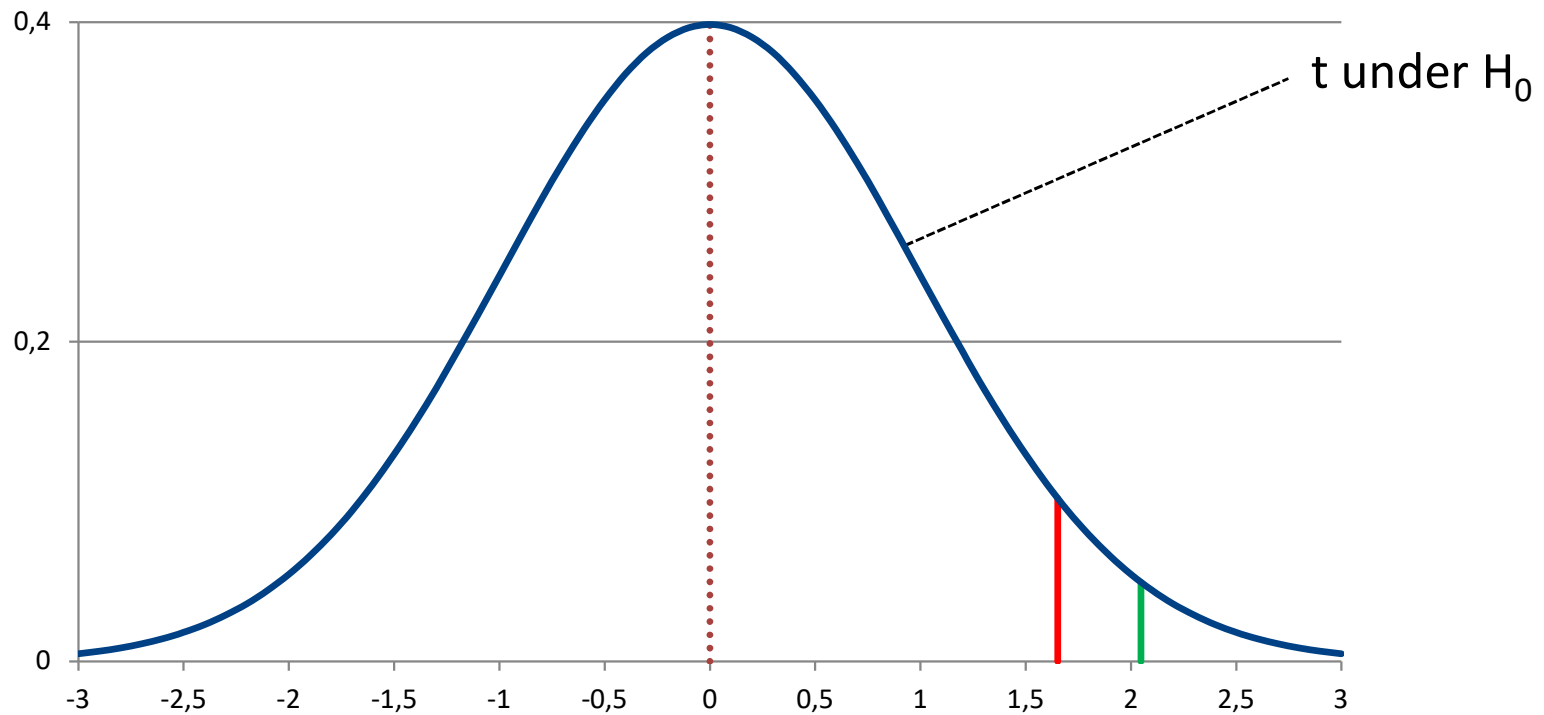


empirical  $t = 2.048$  (example)

# Decision: Null or Alternative Hypothesis (two tailed; $\alpha=5\%$ )

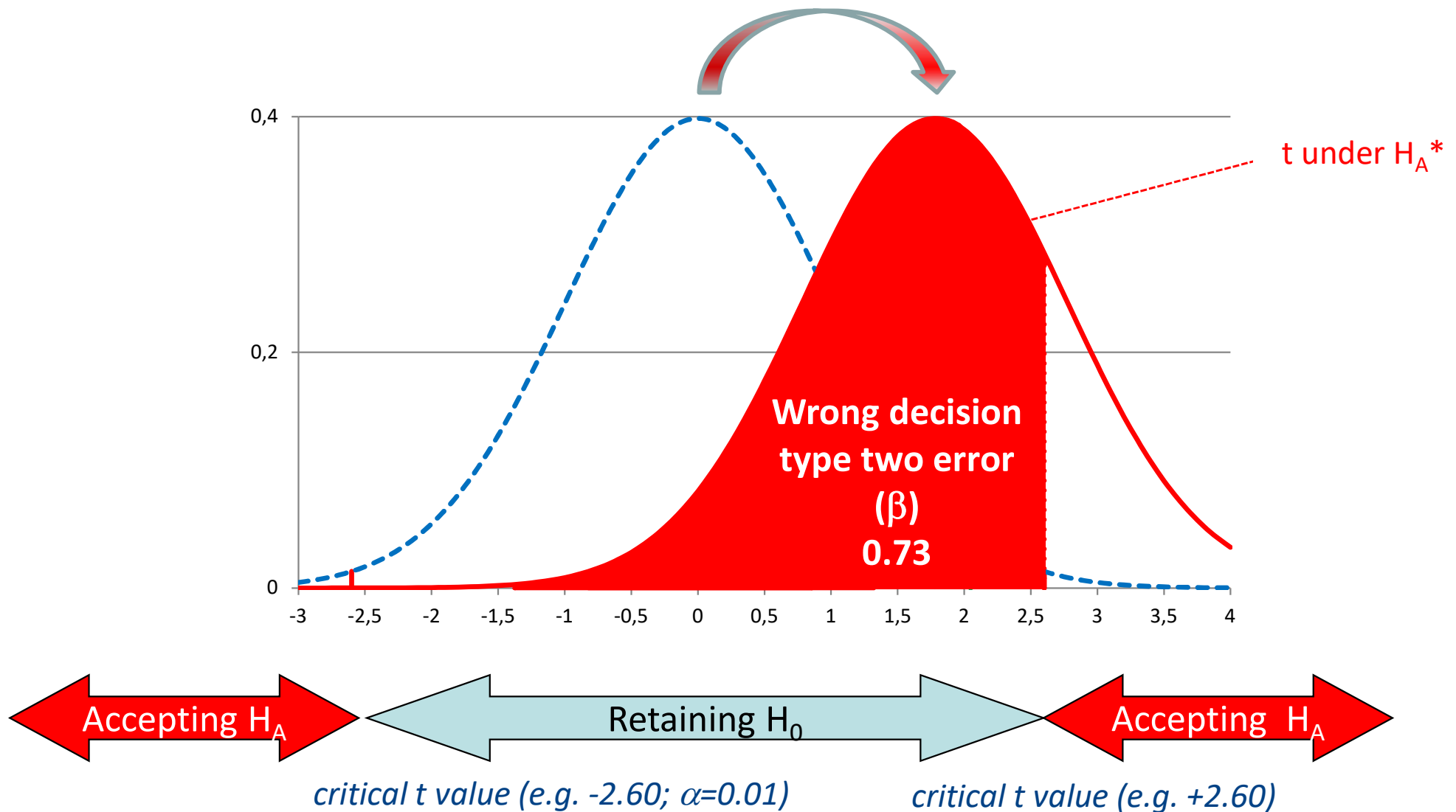


# Decision: Null or Alternative Hypothesis (one tailed; $\alpha=5\%$ )



# Decision: Null or Alternative Hypothesis (two tailed; $\alpha=1\%$ )

- Scenario:  $H_A$  is true \*



\* for a specific difference (here half a scale point); The expected t value of 1.78 is based on the difference of 0.5 divided by the standard error of the difference (0.281 in this case)

# Hypothesis Testing

- Decision to reject  $H_0$  and accept  $H_A$  depends on:
  - $\Delta$ : True mean difference\* (or true strength of relationship)
  - $\sigma$ : Standard deviation (incl. measurement error)
  - $n$ : Sample size
  - $\alpha$ : type one error ( $\alpha$ )
  - $\rightarrow/\leftrightarrow$ : one-tailed hypothesis (versus two-tailed)

*bigger – more likely*

*smaller – more likely*

*bigger – more likely*

*bigger – more likely*

*one-tailed – more likely*

- $\beta$ : type two error ( $\beta$ ) depends on  $\Delta$ ,  $\sigma$ ,  $\alpha$  and  $n$  (and  $\rightarrow/\leftrightarrow$ )

- Assume  $\Delta$ ,  $\sigma$ ,  $\alpha$  and  $\beta$  (and  $\rightarrow/\leftrightarrow$ )
  - optimal sample size  $n$

\* *Practically meaningful difference?*  
*Relevance versus significance.*  
*Effect size! (Cohen's  $d$ )*

$$\epsilon_{\text{soll}} = \frac{(\mu_1 - \mu_0)}{\sigma_0}$$

$$n_{\text{opt}} = \frac{(z_{(1-\alpha)} - z_{\beta})^2}{\epsilon_{\text{soll}}^2} \cdot 2$$

*n per group*

# Hypothesis Testing

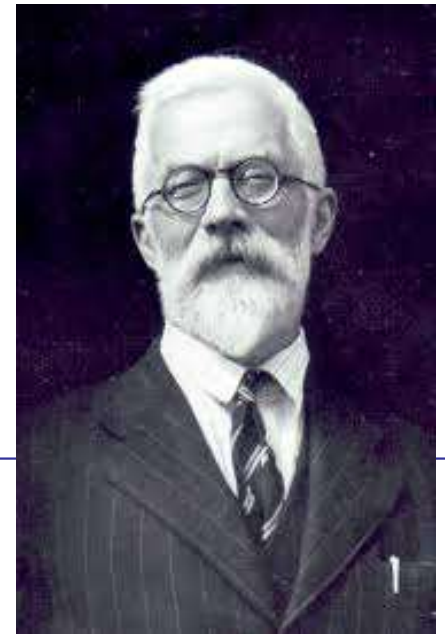
Decision \ "Reality"		$H_0$ is true	$H_A$ is true
		Retaining $H_0$	<b>Correct decision</b> $P=1-\alpha$ ✓
Accepting $H_A$	<b>Type 1 error</b> $P=\alpha$	<b>Correct decision</b> $P=1-\beta$ ✓	
		$P = 1-\alpha + \alpha = 100\%$	$P = 1-\beta + \beta = 100\%$

## Worth noting ...

- p-value proposed by Sir Ronald Fisher as a measure of congruence of data with hypothesis of no difference/no relationship
- Originally, no decision based on some (arbitrary) threshold (implied by type 1 error rate) intended
- A p-value of, say, 0.05 (or smaller) means the data are relatively unlikely under the null hypothesis
- But does this mean that the null hypothesis is wrong and the alternative hypothesis is true?
- What is the probability of the alternative hypothesis being true?
- Arguably, there is no real difference between a p value of 0.04 and 0.06.

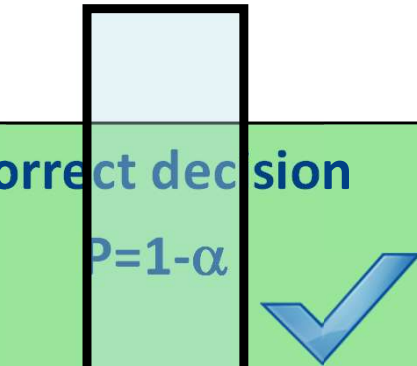
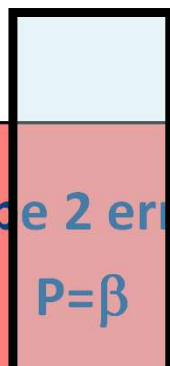
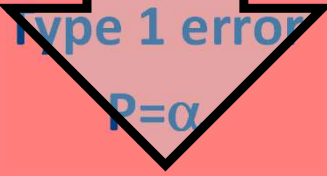
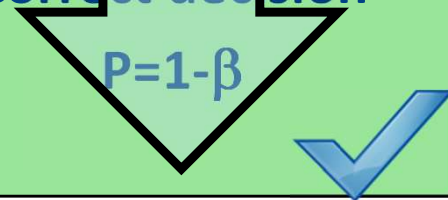


*Sir Ronald Aylmer Fisher  
1890 London  
– 1962 Adelaide*





# Hypothesis Testing

Decision \ "Reality"		"Reality"	
		$H_0$ is true	$H_A$ is true
Decision	Retaining $H_0$	Correct decision $P=1-\alpha$ 	Type 2 error $P=\beta$ 
	Accepting $H_A$	Type 1 error $P=\alpha$ 	Correct decision $P=1-\beta$ 
		$P = 1-\alpha + \alpha = 100\%$	$P = 1-\beta + \beta = 100\%$

Conditional probability: p-value (also  $\alpha$ )= how likely is the difference in the data GIVEN the  $H_0$  is true

# How Likely is the Hypothesis True?

Decision \ "Reality"		$H_0$ is true	$H_A$ is true
		negative	positive
Decision	Retaining $H_0$	Correct decision $P=1-\alpha$ Specificity (correct neg) ✓	Type 2 error $P=\beta$ false neg
	Accepting $H_A$	Type 1 error $P=\alpha$ false pos	Correct decision $P=1-\beta$ Sensitivity (correct pos) ✓
Conditional probabilities add up to 1 in each column but NOT in each row		$P = 1-\alpha + \alpha = 100\%$	$P = 1-\beta + \beta = 100\%$

Conditional probability: how likely is the  $H_0$  (or the  $H_A$ ) GIVEN our decision based on the difference in the data

# Hypothesis Testing

Prediction		There is sun	There is rain
Sun	Null hypothesis (no signal)	90%	20% ( $\beta$ )
Rain	Alternative hypothesis	10% ( $\alpha$ )	80%
		100% (of sunny days)	100% (of rainy days)

- Rain corresponds to “correct theory”
- Here, we also know that there is only 1 in 10 days where there is rain. (Unconditional, or a priori, probability of rain is just 10%.)

# Hypothesis Testing

Prediction		There is sun	There is rain	
Sun	Null hypothesis (no signal)	81	2 ( $\beta=20\%$ )	83 predictions 81 correct (98%)
Rain	Alternative hypothesis	9 ( $\alpha=10\%$ )	8	17 predictions 8 correct (47%)
		90 sunny days	10 rainy days	

- Rain corresponds to “correct theory”
- Here, we also know that there is only 1 in 10 days where there is rain.
- We need the a priori probability of our theory being correct.

## Epidemiology (*all concrete numbers are illustrative only*)

- COVID-19 Testing (e.g. antigen or antibody)
- **Sensitivity:** true positive rate (how many actually positive cases are tested positive)
  - Testing positive GIVEN one has antibodies (=evidence in favour of theory if correct)
- **Specificity:** true negative rate (how many truly negative cases are tested negative)
  - Testing negative GIVEN one has no antibodies (=evidence against theory if incorrect)
- Say, a test has a sensitivity of 90% (i.e.  $\beta=10\%$ ) and a specificity of 95% (i.e.  $\alpha=5\%$ ).
  
- If you are tested positive, how likely are you truly positive (have antibodies)?
- If you are tested negative, how likely are you negative (have no antibodies)?
- Quick answer perhaps is 90% and 95%, respectively.
- In fact, it depends.
  
- Say, we have a town with 2000 people (for illustration only, number does not matter).
- Assume, **5% have had the disease**, i.e. they have antibodies (or have the disease, if we use a test for current infection). Therefore, prevalence is 5%. As we will see, THIS does matter.

# Epidemiology

- 2000 people
- Prevalence: 5%: 100 people
- Sensitivity of 90% and specificity of 95% (test characteristics)

	Truly positive	Truly negative	Total
Positive test result	90	95	
Negative test result	10	1805	
	<b>100</b>	1900	<b>2000</b>
<i>Sensitivity</i>	<i>90/100=90%</i>		<i>Prevalence: 5%</i>
<i>Specificity</i>		1805/1900=95%	

- If you are tested positive, how likely are you truly positive (have antibodies)?
- If you are tested negative, how likely are you negative (have no antibodies)?

# Epidemiology

- If you are tested positive, how likely are you truly positive (have antibodies)?
- If you are tested negative, how likely are you negative (have no antibodies)?

	Truly positive	Truly negative	Total	Positive predicted value/ negative predicted value
Positive test result	90	95	185	90/185=49% (even higher spec. needed)
Negative test result	10	1805	1815	1805/1815=99%
	<b>100</b>	1900	<b>2000</b>	
<i>Sensitivity</i>	<i>90/100=90%</i>			<i>Prevalence: 5%</i>
<i>Specificity</i>		1805/1900=95%		<i>Rule out when prev. low and spec. high</i>

# Epidemiology

- If you are tested positive, how likely are you truly positive (have antibodies)?
- If you are tested negative, how likely are you negative (have no antibodies)?

	Truly positive	Truly negative	Total	Positive predicted value/ negative predicted value
Positive test result	80	19	99	80/99=81%
Negative test result	20	1881	1901	1881/1901=99%
	<b>100</b>	1900	<b>2000</b>	
<i>Sensitivity</i>	<i>80/100=80%</i>			<i>Prevalence: 5%</i>
<i>Specificity</i>		1881/1900=99%		<i>Increasing specificity might decrease sensitivity</i>



# Epidemiology

- If you are tested positive, how likely are you truly positive (have antibodies)?
- If you are tested negative, how likely are you negative (have no antibodies)?

	Truly positive	Truly negative	Total	Positive predicted value/ negative predicted value
Positive test result	60	19	79	60/79=76%
Negative test result	40	1881	1921	1881/1921=98%
	<b>100</b>	1900	<b>2000</b>	
<i>Sensitivity</i>	<i>60/100=60%</i>			<i>Prevalence: 5%</i>
<i>Specificity</i>		1881/1900=99%		<i>Increasing specificity might decrease sensitivity</i>

# Epidemiology

- What if a lot of people have antibodies, say 50%?

	Truly positive	Truly negative	Total	Positive predicted value/ negative predicted value
Positive test result	900	50	950	$900/950=95\%$
Negative test result	100	950	1050	$950/1050=90\%$
	<b>1000</b>	1000	<b>2000</b>	
<i>Sensitivity</i>	$90/100=90\%$			<i>Prevalence: 50%</i>
<i>Specificity</i>		$950/1000=95\%$		<b>Rule in when prev. high and sens. high</b>

# Epidemiology

- What if most people have antibodies, say 80%?

	Truly positive	Truly negative	Total	Positive predicted value/ negative predicted value
<b>Positive test result</b>	1440	20	1460	1440/1460=99%
<b>Negative test result</b>	160	380	540	380/540=70%
	<b>1600</b>	400	<b>2000</b>	
<i>Sensitivity</i>	<i>90/100=90%</i>			<i>Prevalence: 50%</i>
<i>Specificity</i>		950/1000=95%		

# Conclusions

The probability of antibodies (or infection) given a positive test and the probability of no antibodies (or no infection) given a negative test depend on

- the sensitivity of the test
- the specificity of the test
- **and, most importantly, on the prevalence!**
  
- A low prevalence means we need very high specificity (because there is a lot of room for false positive results)
- A high prevalence means we need very high sensitivity (to decrease false negative results).
  
- The same is true for hypothesis testing.

## Conclusions for hypothesis testing

Null hypothesis = truly negative

Alternative hypothesis = truly positive

non significant = negative test result

significant = positive test result

The probability of an alternative hypothesis to be true given a significant test result depends on

- the type-one error rate ( $\alpha$ )
- the type-two error rate ( $\beta$ )
- and on the a priori probability of the hypothesis!  $\sim$  prevalence
- Note that here the notion of probability is not frequentist. A theory (or its trueness) is not a random variable that is sometimes true, sometimes not.
- Rather, it is a subjective probability. What we believe based on previous knowledge.
- Bayes

# Bayes' Theorem

Hypothesis is true (A)  
given decision (B)

Decision given Hypothesis is true

Hypothesis is true

$$P(A | B) = \frac{P(B | A) \cdot P(A)}{P(B)}$$

P (hyp true and we decide true)

P (we decide true)

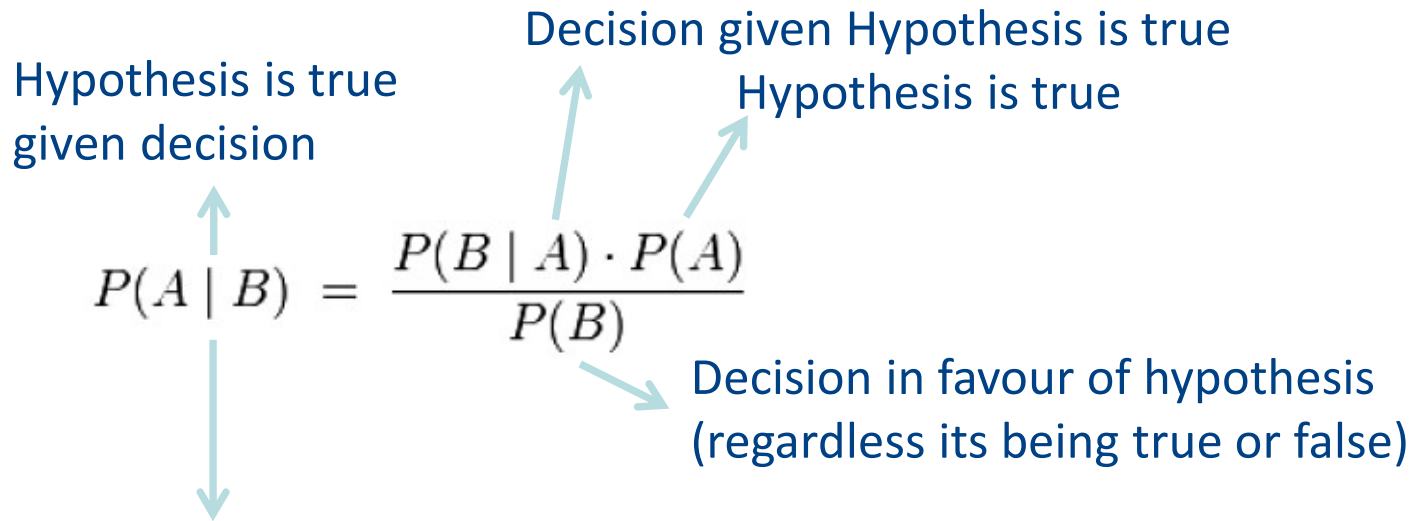
Decision in favour of hypothesis  
(regardless of its being true or false)



Thomas Bayes (1701-1761)

- $P(A | B) = P(H_A \text{ is true} | \text{Accepting } H_A)$  *this is what we want to know*
- $P(B | A) = P(\text{Accepting } H_A | H_A \text{ is true})$  *that is 1-β (power)*
- $P(A)$  *a priori probability, subjective (~ prevalence; before the test)*
- $P(B) = P(\text{Accepting } H_A)$  *can be computed based on P(A)*
- $P(B|A) * P(A) + P(B|\neg A) * P(\neg A)$ 
  - or:  $(1-\beta) * P(A) + (\alpha) * \{1- P(A)\}$  *deciding  $H_A$  for the right and for the wrong reason*

# Bayes' Theorem: if a priori prob. is 0.10 (~ low prevalence)



Thomas Bayes (1701-1761)

- $P(A | B) = P(H_A \text{ is true} | \text{Accepting } H_A) ?$
- $P(B | A) = P(\text{Accepting } H_A | H_A \text{ is true}) = 1 - \beta = 0.80$
- $P(A) = 0.10$
- $P(B) = P(\text{Accepting } H_A) \text{ can be computed based on } P(A)$ 
  - $= (1 - \beta) * P(A) + (\alpha) * \{1 - P(A)\}$
  - $0.80 * 0.10 + 0.05 * 0.90 = 0.08 + 0.045 = 0.125$
- $P(A | B) = P(H_A \text{ is true} | \text{Accepting } H_A) = 0.80 * 0.10 / 0.125 = 0.64$

$\alpha = 5\%$ $\beta = 20\%$ $1 - \beta = 80\%$
--

# Bayes' Theorem

	<i>Reality</i>		
<i>Decision</i>	Theory wrong [H <sub>0</sub> ]	Theory correct [H <sub>A</sub> ]	Σ
Retaining H <sub>0</sub>	95%	20%	n.a.
Accepting H <sub>A</sub>	5%	80%	n.a.
	$\alpha = 5\%$	$\beta = 20\%$	

<b>A priori H<sub>A</sub>=10%</b>	<i>Reality</i>		
<i>Decision</i>	Theory wrong [H <sub>0</sub> ]	Theory correct [H <sub>A</sub> ]	Σ
Retaining H <sub>0</sub>	85.5%	2%	87.5%
Accepting H <sub>A</sub>	4.5%	8%	12.5%
	90%	10%	100%

→ P (B) = P(Accepting H<sub>A</sub>)

→ P (A) = 0.10

$8 / 12.5 = 0.64$

$P (B | A) * P (A) = P(\text{Accepting } H_A | H_A \text{ is true}) * P (A) = 0.80 * 0.10$



## Conclusions for hypothesis testing

The p-value is just one part of the decision making.

It determines whether we retain the null hypothesis or reject it.

It is obvious that error rates are important.

The quality of the test is also important. For example, the t-test has the highest power possible (for a two-group mean comparison) if its requirements are met.

That's why we should use the best possible tests.

(Note: if prerequisites are not met, this no longer applies. Fancy statistics with poor data can be deceptive. Problems with structural equation modelling?)

It is (perhaps) less obvious that the *a priori probability* also matters (existing knowledge).

With very unlikely theories, specificity should be high (very small type-one error alpha).

With very likely theories, sensitivity should be high (power, low type-two error beta).

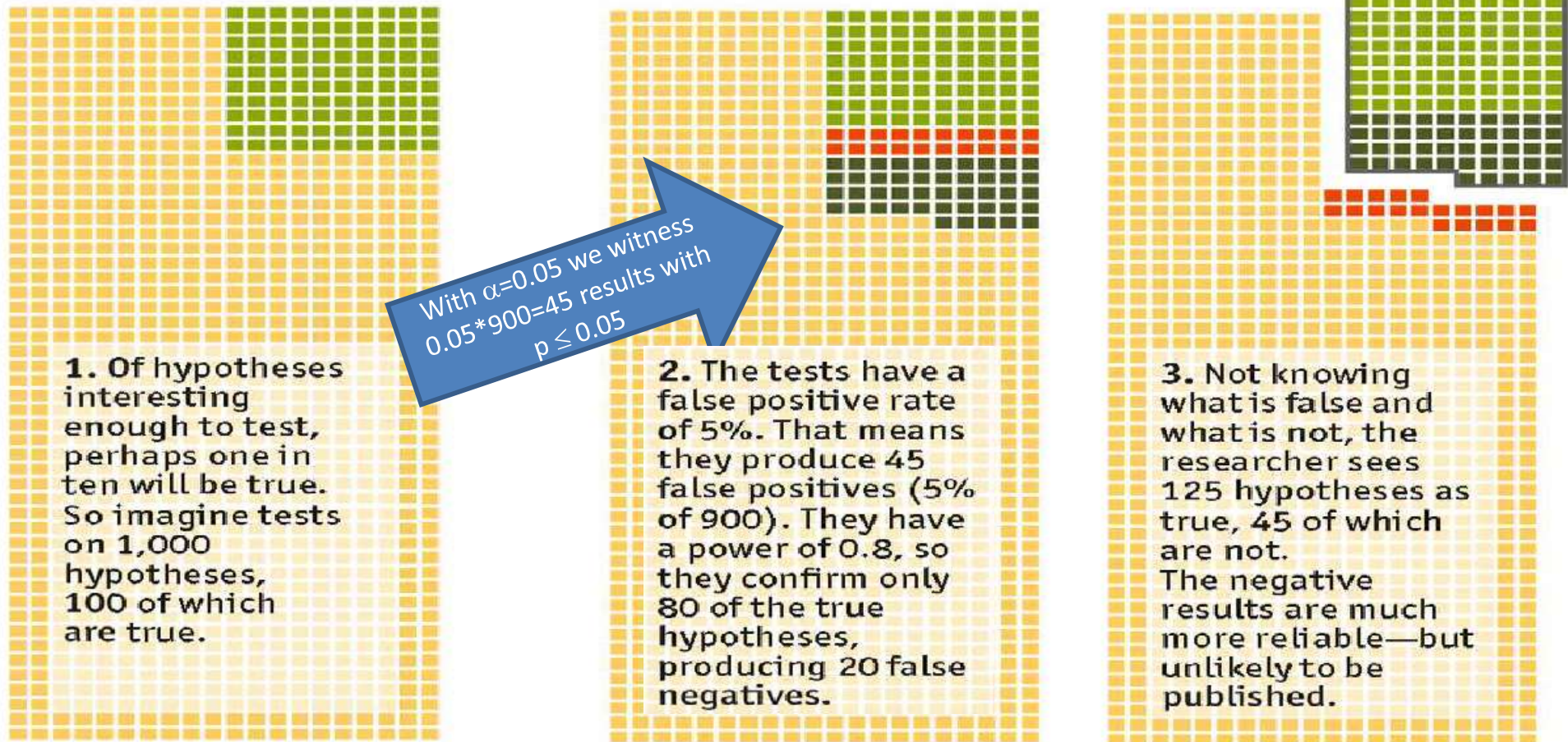
(Consequences of wrong decisions also matter, of course.)

# Rather daring hypotheses (Prob. of 10%)

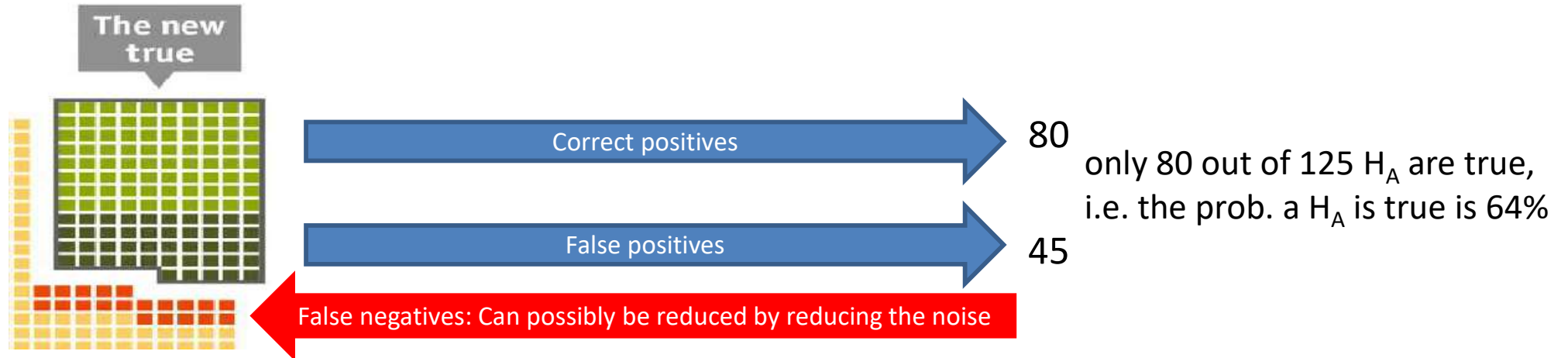
## Unlikely results

How a small proportion of false positives can prove very misleading

False True False negatives False positives



## Rather daring hypotheses (Prob. of 10%)



- Reducing false negatives by 50% (power of 90% rather 80%) increases the 80 to 90 (and the sum from 125 to 135)
  - Still only  $90/135 = 67\%$  of all significant  $H_A$  are true
- Should we look at the exact p value?
  - Remember, its inventor, Sir Roland Fisher suggested p as a measure of how likely the result is based on chance alone.

# The p value

p-Value:

- Associated with the assumption that the  $H_0$  is true
- Probability that the result (relationship, difference, etc.) is due to chance alone
- Idea: we need to do (much) better than chance ...

A small p-value implies that the data are very unlikely given the  $H_0$  is true

- By implication, the smaller the p-value is, the less likely  $H_0$  is.