P-values and statistical tests
1. Introduction

Marek Gierliński
Division of Computational Biology

Hand-outs available at http://is.gd/statlec

1. Introduction

Null hypothesis, statistical test, p-value
Fisher's test

2. Contingency tables

Chi-square test
G-test

3. T-test

One- and two-sample
Paired
One-sample variance test

4. ANOVA

One-way
Two-way

5. Non-parametric methods 1

Mann-Whitney
Wilcoxon signed-rank
Kruskal-Wallis

6. Non-parametric methods 2

Kolmogorov-Smirnov
Permutation
Bootstrap

7. Statistical power

Effect size
Power in t-test
Power in ANOVA

8. Multiple test corrections

Family-wise error rate
False discovery rate
Holm-Bonferroni limit
Benjamini-Hochberg limit
Storey method

9. What's wrong with p-values?

A lot

How Johnny the Biologist understands p-values

Experiment

Statistics

$$E\{\widehat{\mathrm{pFDR}}_\lambda(\gamma)\} - \mathrm{pFDR}(\gamma) \geqslant E\left[\frac{\{W(\lambda)/(1-\lambda)\}\gamma - V(\gamma)}{\{R(\gamma) \vee 1\}\Pr\{R(\gamma) > 0\}}\right],$$

$> 0\} \geqslant 1 - (1-\gamma)^m$ under independence. Conditioning on $R(\gamma)$, it follows th

$$\frac{W(\lambda)/(1-\lambda)\}\gamma - V(\gamma)}{R(\gamma) \vee 1\}\Pr\{R(\gamma) > 0\}}\,\Bigg|\, R(\gamma)\Bigg] = \frac{[E\{W(\lambda)|R(\gamma)\}/(1-\lambda)]\gamma - E\{V(\gamma)|R(\gamma)}{\{R(\gamma) \vee 1\}\Pr\{R(\gamma) > 0\}}$$

e, $E\{W(\lambda)|R(\gamma)\}$ is a linear non-increasing function of $R(\gamma)$, and $E\{V(\gamma)|R(\gamma$
function of $R(\gamma)$. Thus, by Jensen's inequality on $R(\gamma)$ it follows that

$$\frac{W(\lambda)/(1-\lambda)\}\gamma - V(\gamma)}{R(\gamma)\Pr\{R(\gamma) > 0\}}\,\Bigg|\, R(\gamma) > 0\Bigg] \geqslant \frac{E[\{W(\lambda)/(1-\lambda)\}\gamma - V(\gamma)|R(\gamma) > 0]}{E\{R(\gamma)|R(\gamma) > 0\}\Pr\{R(\gamma) > 0\}}$$

$= E\{R(\gamma)|R(\gamma) > 0\}\Pr\{R(\gamma) > 0\}$, it follows that

$$\frac{W(\lambda)/(1-\lambda)\}\gamma - V(\gamma)|R(\gamma) > 0]}{\{R(\gamma)|R(\gamma) > 0\}\Pr\{R(\gamma) > 0\}} = \frac{E[\{W(\lambda)/(1-\lambda)\}\gamma - V(\gamma)|R(\gamma) > 0]}{E\{R(\gamma)\}}.$$

$$p < 0.05$$

P-Values: Misunderstood and Misused

Bertie Vidgen and Taha Yasseri *

frontiers in Physics

MINI REVIEW
published: 04 March 2016
doi: 10.3389/fphy.2016.00006

The fickle *P* value generates irreproducible results

Lewis G Halsey, Douglas Curran-Everett, Sarah L Vowler & Gordon B Drummond

NATURE METHODS | VOL.12 NO.3 | MARCH 2015 | 179

Open access, freely available online

Essay

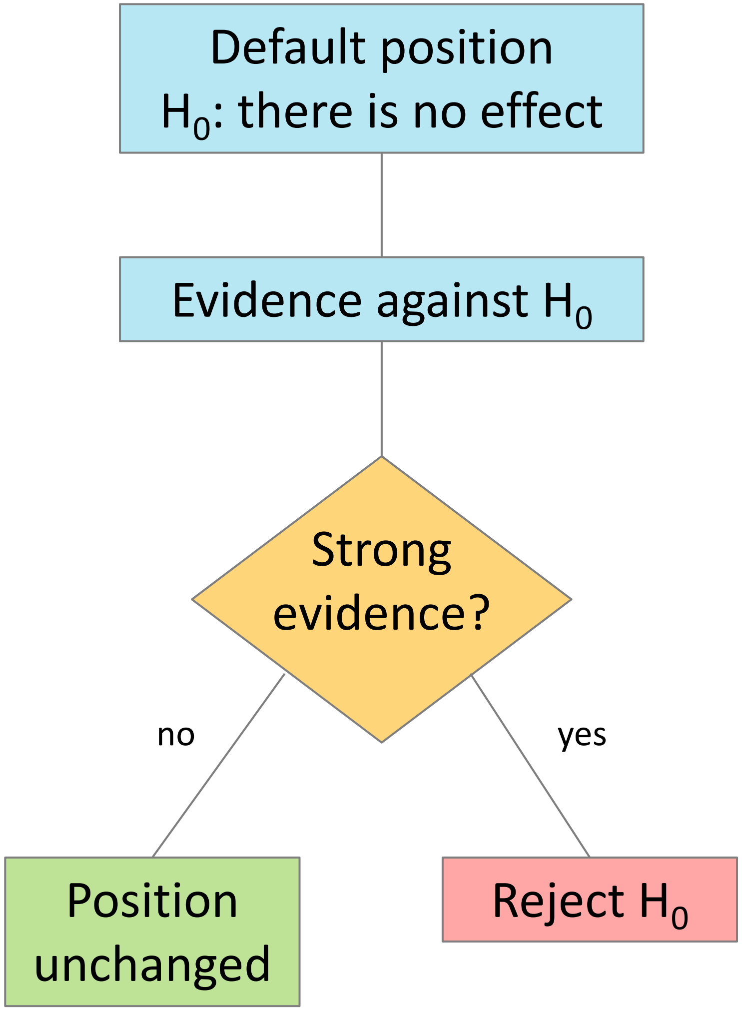
Why Most Published Research Findings Are False

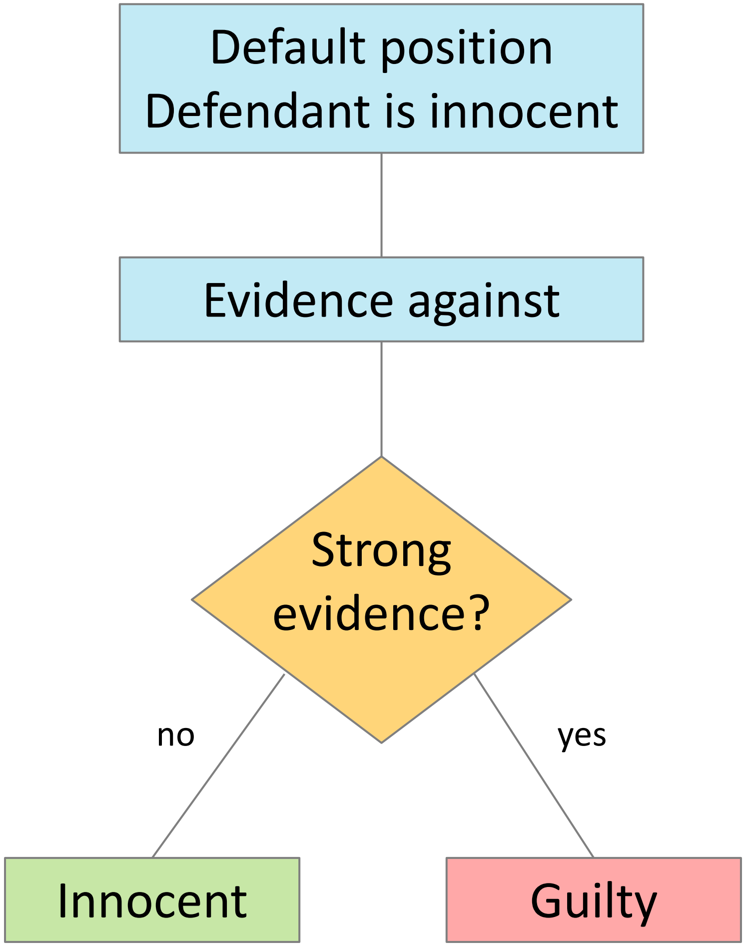
John P. A. Ioannidis

PLoS Medicine | www.plosmedicine.org 0696 August 2005 | Volume 2 | Issue 8 | e124

Null hypothesis

Null hypothesis





Evidence against H_o

- Two samples of mice
  - 12 English mice
  - 9 Scottish mice

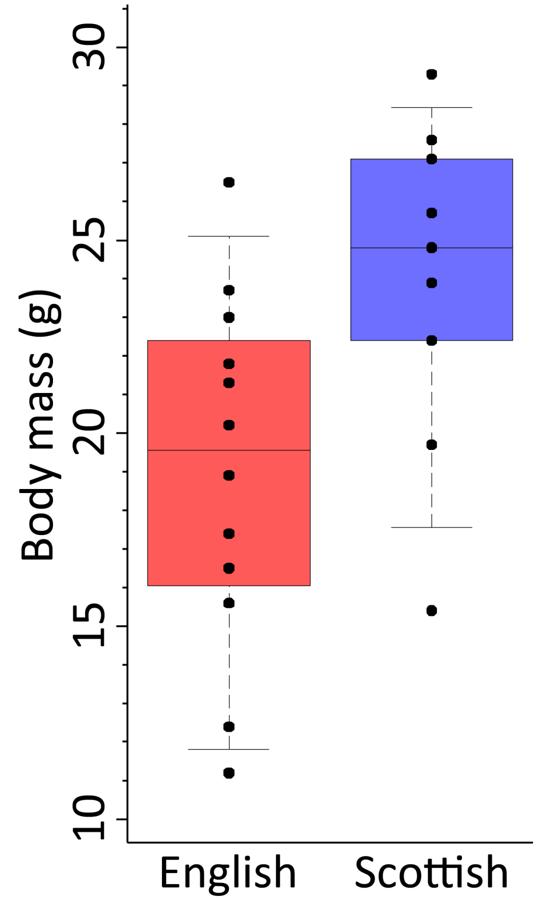
- Body mass difference:

  $\Delta M = M_S - M_E = 5.0$ g

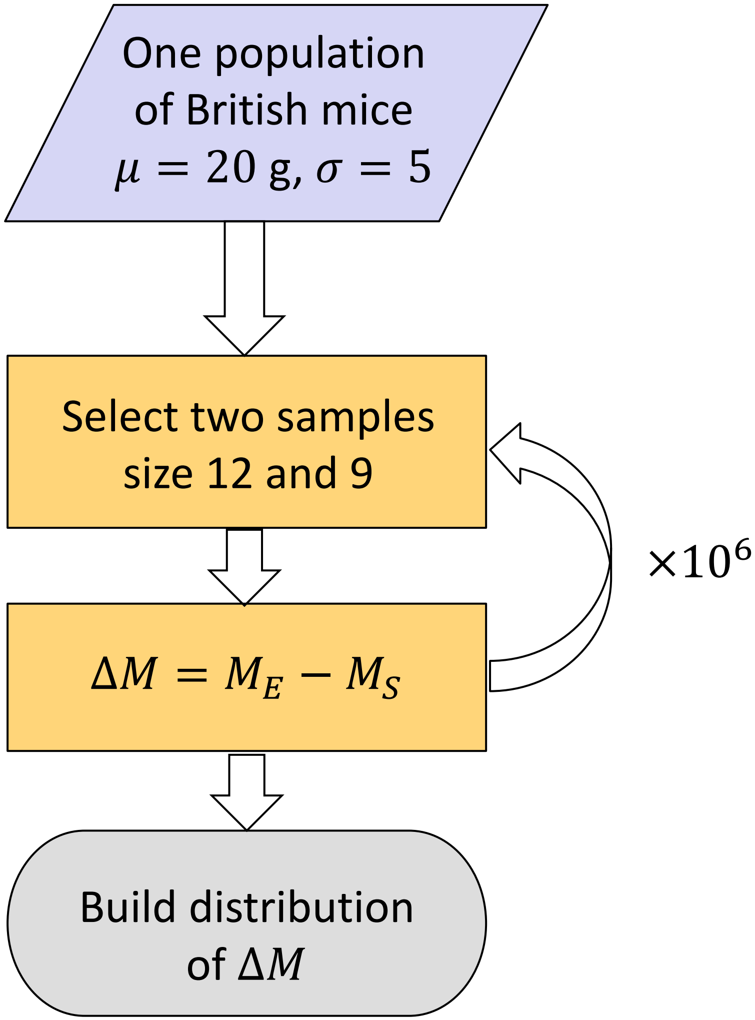
- Two possibilities
  - real difference
  - fluke

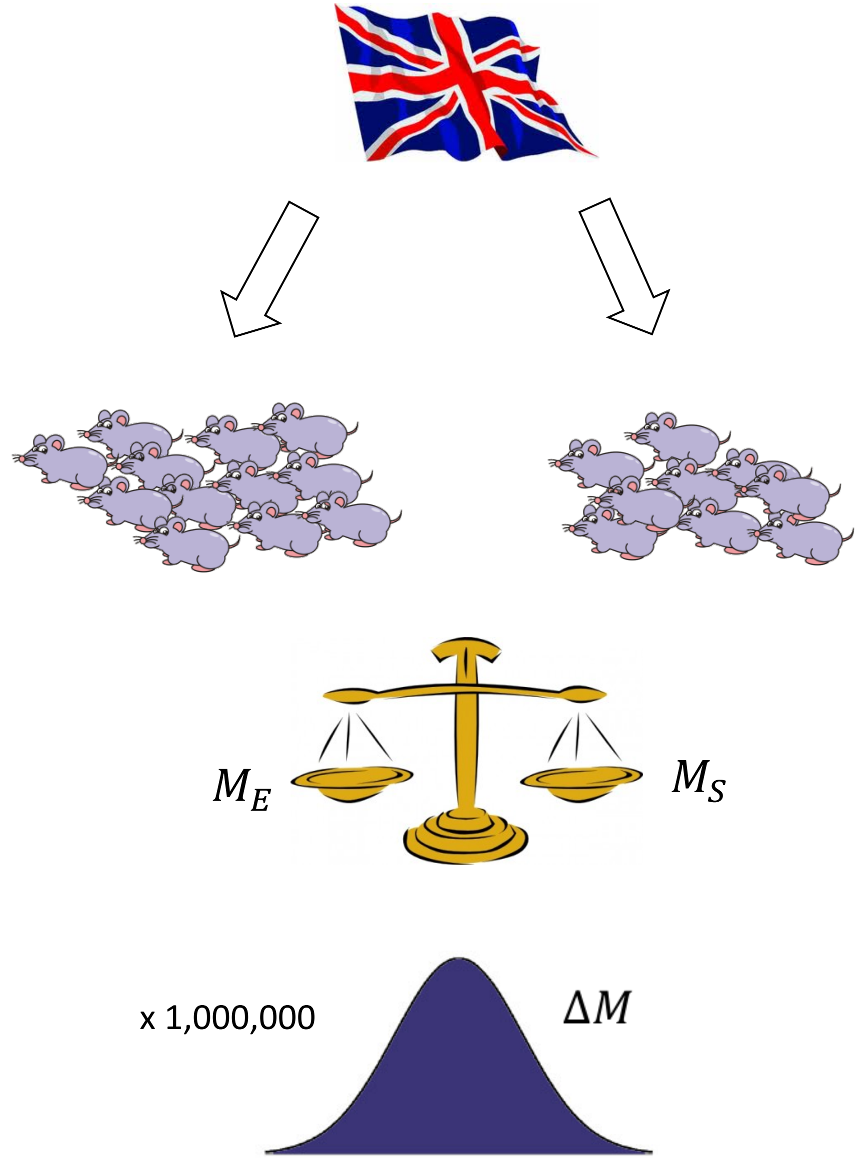
- What are the chances of the fluke?

$M_E = 19.0$ g $\quad M_S = 24.0$ g

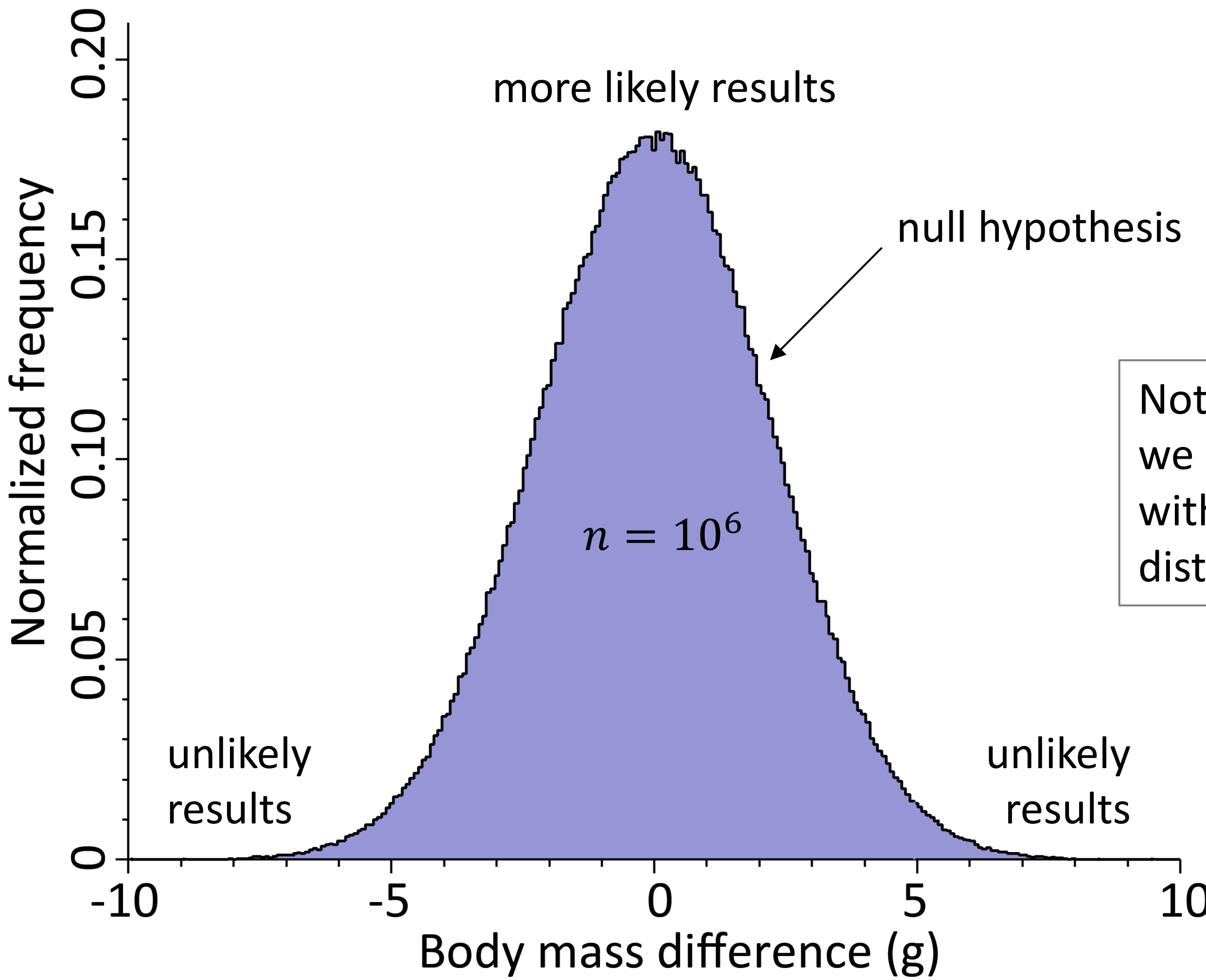


Gedankenexperiment under the null hypothesis



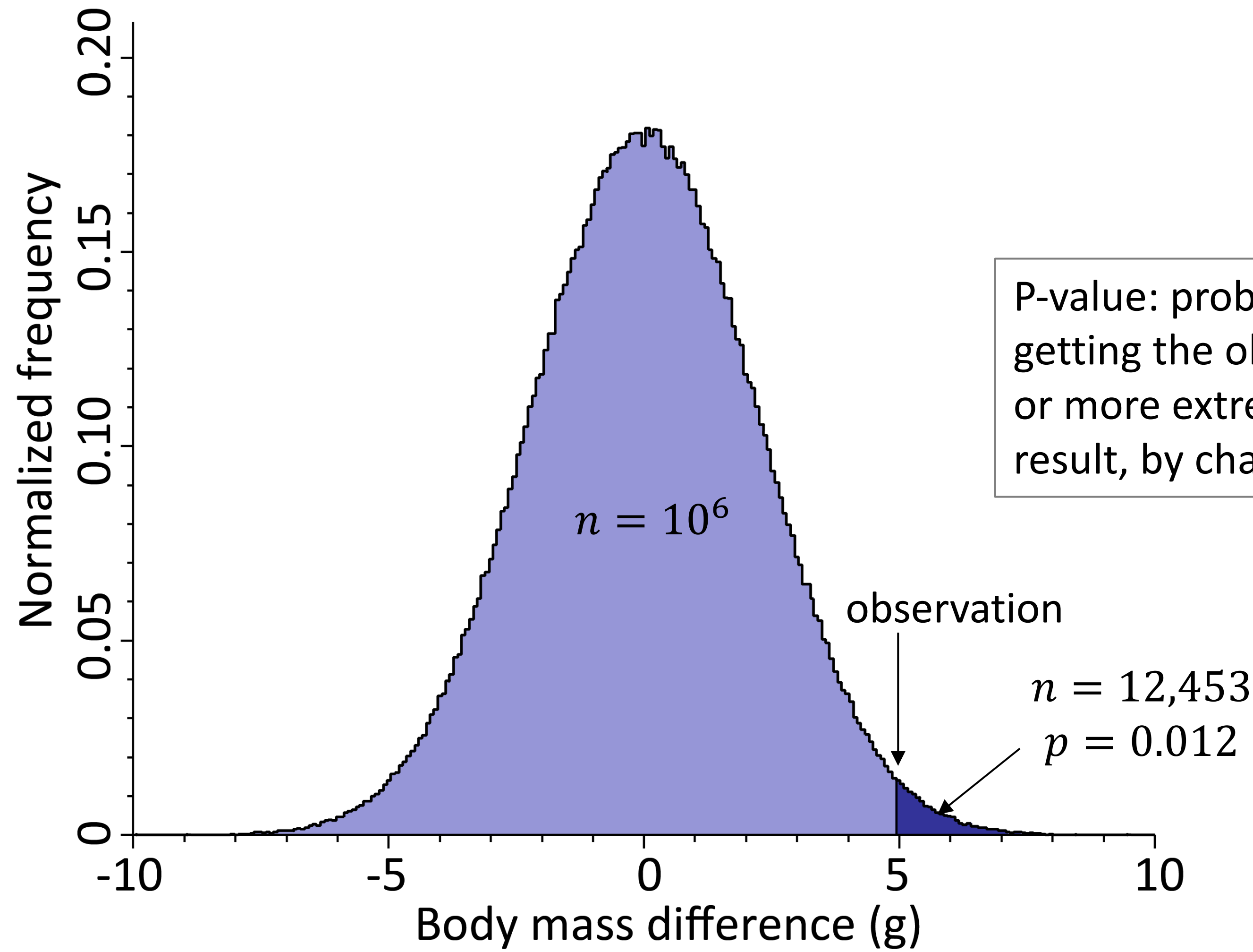


Gedankenexperiment: result under null hypothesis



Note: in real life we use a statistic with known distribution

Gedankenexperiment: p-value



P-value: probability of getting the observed, or more extreme result, by chance

Null hypothesis and p-value

null hypothesis

If

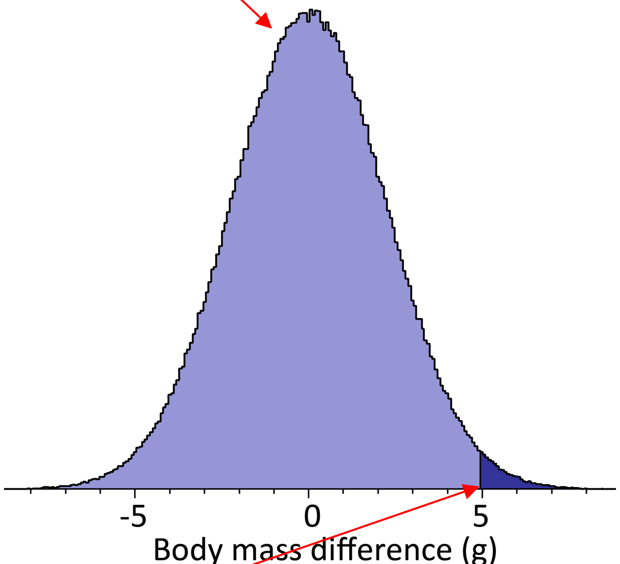
both samples were taken from the same population,

then

the probability of observing the difference in mean body mass of 5 g, **or more**,

by chance (due to random sampling)

would be 1.2%



p-value

We observe an effect, but it will occur by chance in 1.2% of repeated experiments (1 in 80)

You have 1.2% chance of making a fool of yourself (if you publish this result)

P-value is the probability of making a fool of yourself

Two approaches

Fisher

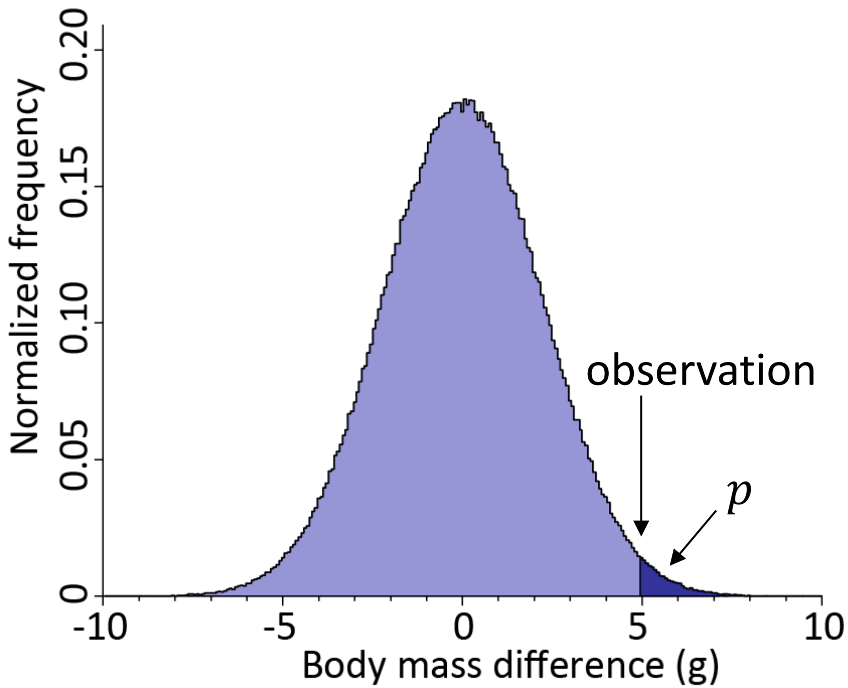
$H_0: \mu_E = \mu_S$

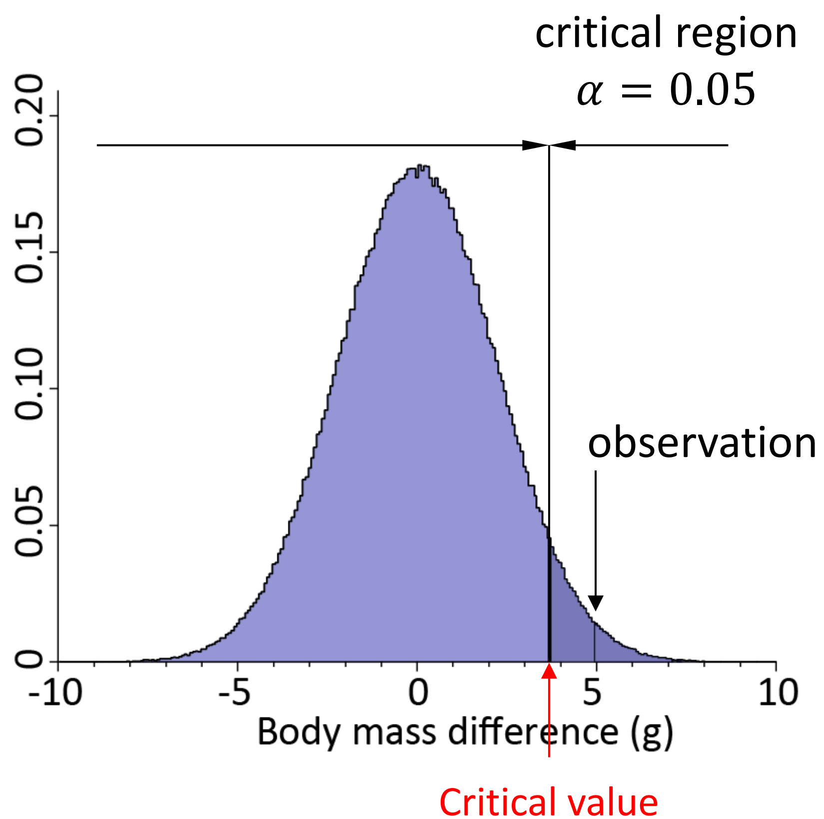
Neyman-Pearson

$H_0: \mu_E = \mu_S$

$H_1: \mu_E < \mu_S$

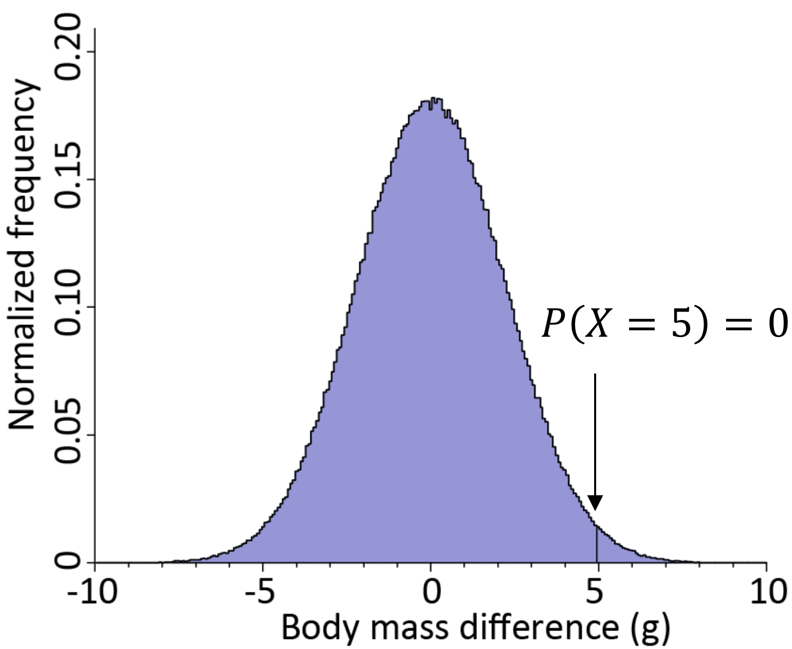
$\alpha = 0.05$



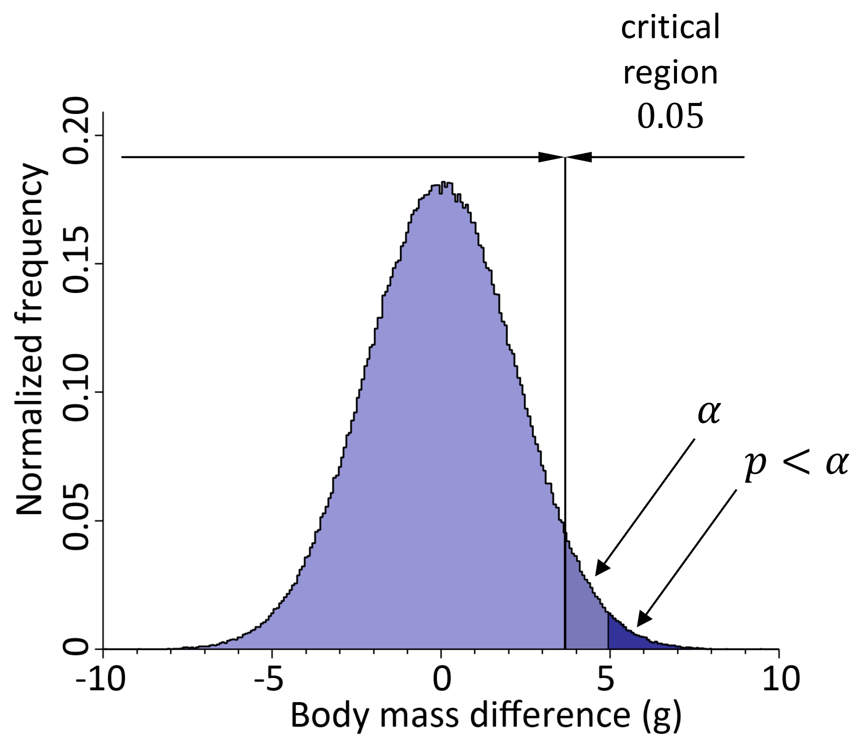


Why “more extreme”?

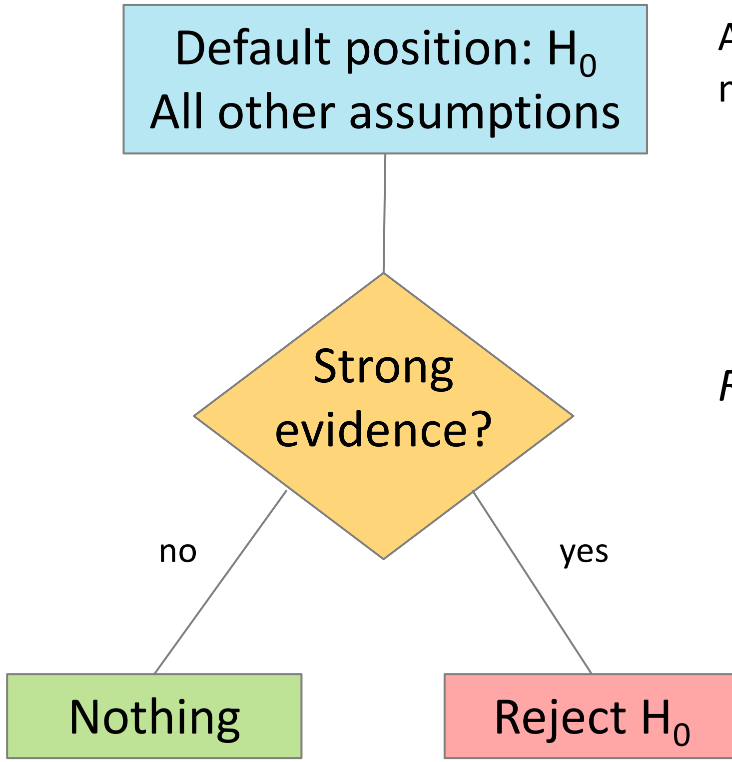
Zero probability



You need to select your rejection region before test (Neyman-Pearson)



Null hypothesis: reject or what?



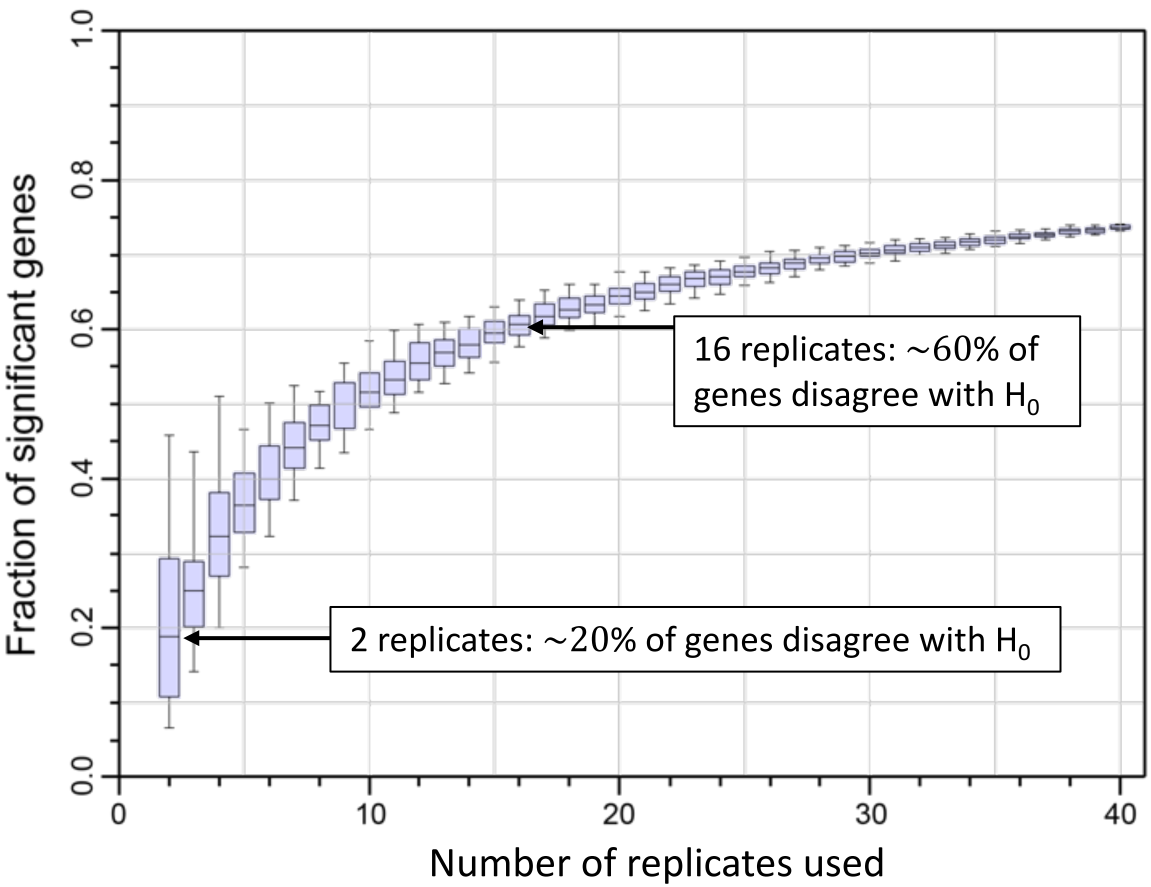
All other assumptions must be true!

Reductio ad absurdum

- absence of evidence is not evidence of absence!
- evidence too week?

- data are incompatible with H_0...
- ...or any of the other assumptions
- reject H_0 at your own risk

You cannot confirm the null hypothesis



Schurch et al. 2016

Differential gene expression between WT and a mutant

Genes that are "not different" from 2 replicates...

...are "significantly different" when using 16 replicates

$$p \geq \alpha$$

✗ No effect

✓ Insufficient evidence

You cannot prove the null hypothesis

Statistical testing

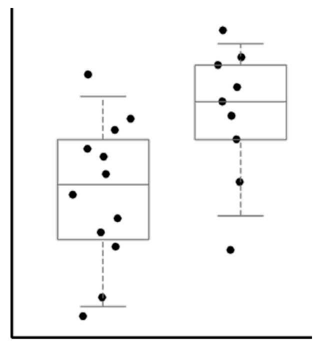
Statistical model

Null hypothesis
H_0: no effect

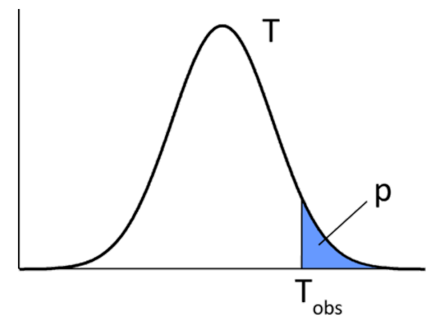
All other assumptions

Significance level
$\alpha = 0.05$

Data



Statistical test against H_0



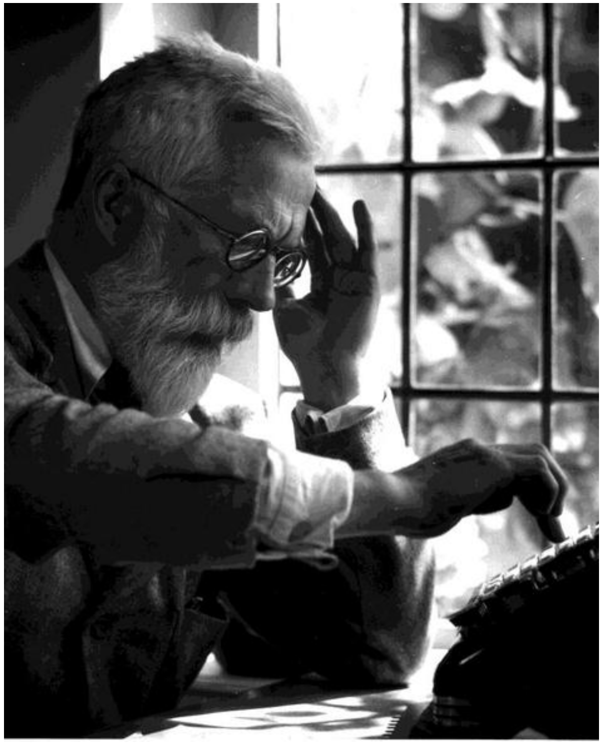
p-value: probability that the observed effect is random

$p < \alpha$
Reject H_0
(at your own risk)
Effect is real

$p \geq \alpha$
Insufficient evidence

Fisher’s exact test

Ronald Fisher



Sir Ronald Aylmer Fisher
(1890-1962)

Rothamsted Experimental Station
(Hertfordshire)

The appreciation of tea

Milk first





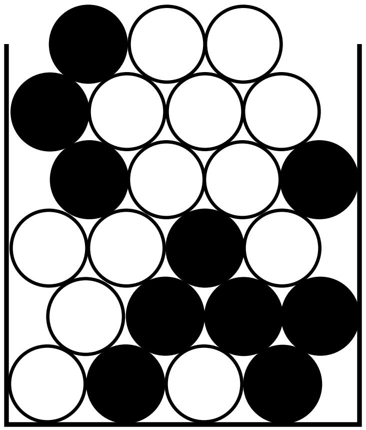


Tea first

Let’s draw some balls

Draw n balls without replacement

removing balls changes probability!



Urn with N balls
m of them white

What is the probability
of finding exactly k white balls?

Binomial coefficient

- “n chose k”

$$\binom{n}{k} = \frac{n!}{k!\,(n-1)!}$$

Set of 5 elements

1 2 3 4 5

- In *combinatorics* it is the number of possible *k*-element subsets of an *n*-element set

All possible 3-element subsets

1 2 3	1 4 5
1 2 4	2 3 4
1 2 5	2 3 5
1 3 4	2 4 5
1 3 5	3 4 5

- From a 5-element set there are 10 possible 3-element subsets

$$\binom{5}{3} = \frac{5!}{3!\,2!} = \frac{120}{6\times 2} = 10$$

Count all the possibilities

$$\binom{5}{3} = 10$$

① ② ③ ❹ ❺ → ○ ○ ●

$$\binom{3}{2} = 3 \qquad \binom{2}{1} = 2$$

Draw 3 balls. What is the probability of finding exactly 2 whites among them?

① ②

① ③ × ❹

② ③   ❺

$$P = \frac{\binom{2}{1} \times \binom{3}{2}}{\binom{5}{3}} = \frac{6}{10} = 0.6$$

Hypergeometric probability

- $N = 36$ balls
- $m = 20$ are white
- $n = 10$ balls drawn

- What is the probability of finding exactly $k = 8$ white balls in the draw?

$$P(X = 8) = \frac{\binom{20}{8}\binom{16}{2}}{\binom{36}{10}}$$

$$= \frac{125{,}970 \times 120}{254{,}186{,}856} = \frac{15{,}116{,}400}{254{,}186{,}856} \approx 0.059$$

	Drawn	Not drawn	Total
White	**8**	12	**20**
Black	2	14	16
Total	**10**	26	**36**

Contingency table

Contingency table contains counts

Hypergeometric probability

- N balls
- m are white
- n drawn

- What is the probability of finding exactly k white balls in the draw?

$$P(X = k) = \frac{\binom{m}{k}\binom{N-m}{n-k}}{\binom{N}{n}}$$

	Drawn	Not drawn	Total
White	k	$m-k$	m
Black	$n-k$	$N+k-n-m$	$N-m$
Total	n	$N-n$	N

Contingency table

Hypergeometric distribution

- If sums are fixed (blue fields), the cells in the table follow hypergeometric distribution

	Drawn	Not drawn	Total
White	k	$20-k$	20
Black	$10-k$	$6+k$	16
Total	10	26	36

$$P\begin{bmatrix}0 & 20\\10 & 6\end{bmatrix} = 3.2\times10^{-5}$$

$$P\begin{bmatrix}1 & 19\\9 & 7\end{bmatrix} = 0.00090$$

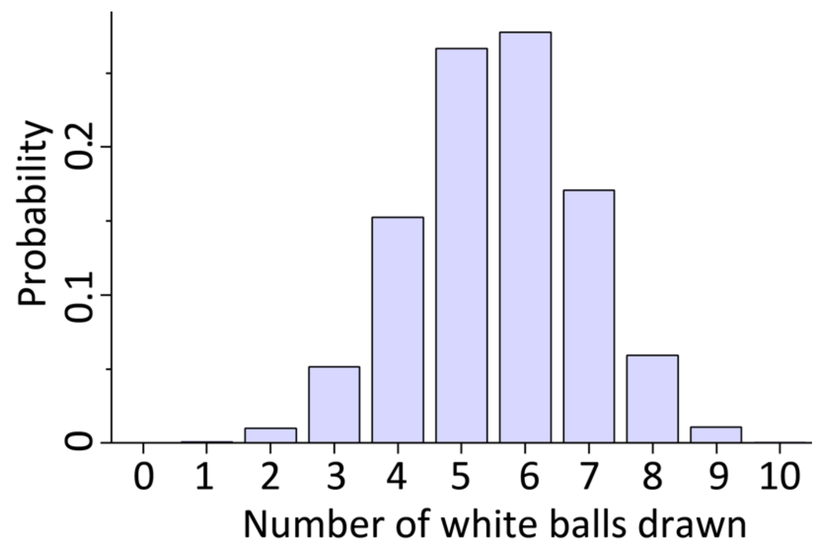
$$P\begin{bmatrix}2 & 18\\8 & 8\end{bmatrix} = 0.0096$$

…

$$P\begin{bmatrix}8 & 12\\2 & 14\end{bmatrix} = 0.059$$

$$P\begin{bmatrix}9 & 11\\1 & 15\end{bmatrix} = 0.011$$

$$P\begin{bmatrix}10 & 10\\0 & 16\end{bmatrix} = 0.00073$$



Hypergeometric distribution

- If sums are fixed (blue fields), the cells in the table follow hypergeometric distribution

$$P\begin{bmatrix}0 & 20\\10 & 6\end{bmatrix} = 3.2\times10^{-5}$$

$$P\begin{bmatrix}1 & 19\\9 & 7\end{bmatrix} = 0.00090$$

$$P\begin{bmatrix}2 & 18\\8 & 8\end{bmatrix} = 0.0096$$

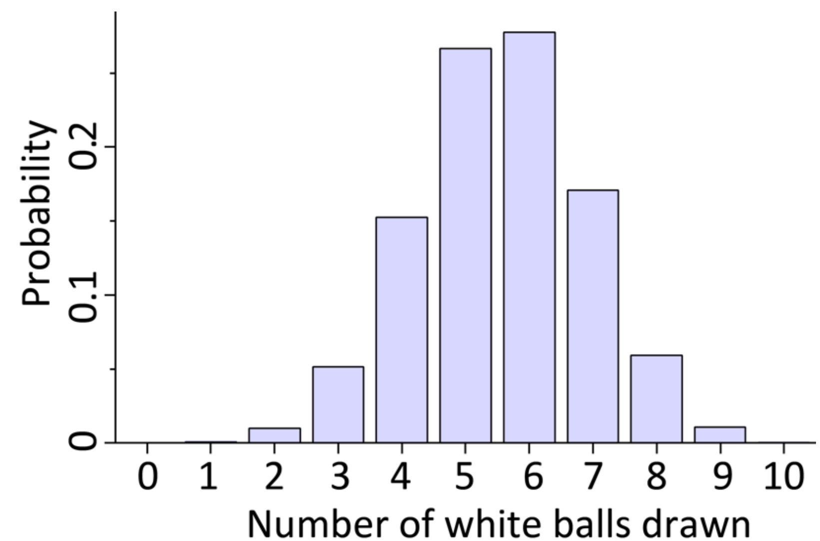
…

$$P\begin{bmatrix}8 & 12\\2 & 14\end{bmatrix} = 0.059$$

$$P\begin{bmatrix}9 & 11\\1 & 15\end{bmatrix} = 0.011$$

$$P\begin{bmatrix}10 & 10\\0 & 16\end{bmatrix} = 0.00073$$

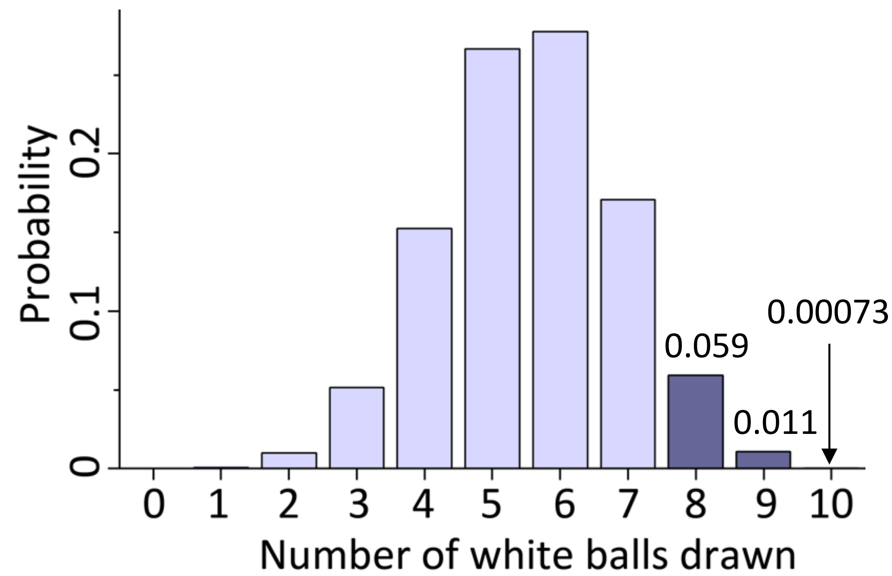
	Drawn	Not drawn	Total
White	10	10	20
Black	0	16	16
Total	10	26	36



One-sided test

- What is the probability of drawing **8 or more** white balls?

$$P(X \geq 8) = 0.059 + 0.011 + 0.00073$$
$$= 0.071$$



- *Enrichment*: do we have more than random? (right-sided test)

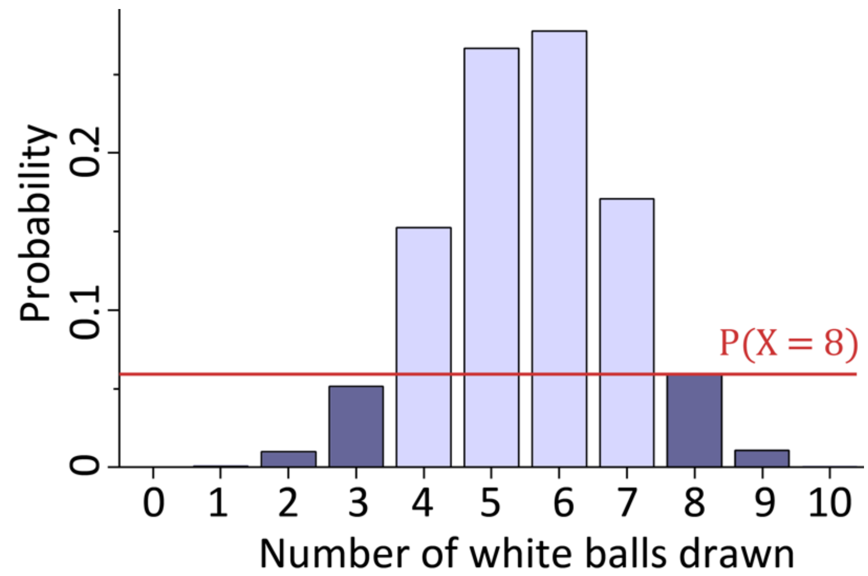
- *Depletion*: do we have fewer than random? (left-sided test)

Two-sided test

- One-sided test: do we observed too many white balls?
- Two-sided test: do we observe too many or too few white balls?
- Is my result extreme in any way?

- Add all probabilities less or equal $P(X = 8)$ on both sides

$$P(X \leq 3 \cup X \geq 8) = 0.13$$



Tea tasting by Muriel Bristol

Milk first

Tea first

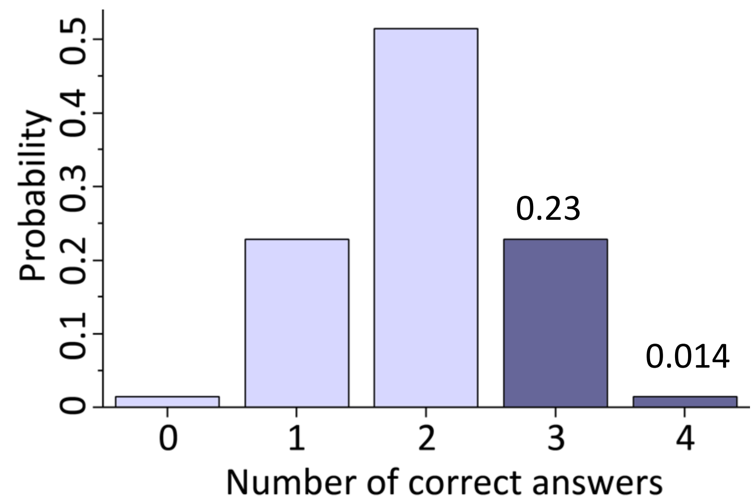
Tea tasting test

- Null hypothesis: Ms Bristol has no ability to tell the difference
- One-sided probability of getting this or more extreme result by chance is

  $$P(X \geq 3) = 0.229 + 0.014 \approx 0.24$$

- The null hypothesis cannot be rejected
- Insufficient data!

	Tea first	Milk first	Total
Ms Bristol says "tea first"	3	1	4
Ms Bristol says "milk first"	1	3	4
Total	4	4	8



Contingency table

- Two variables (in columns and rows)
- E.g. treatments vs outcomes
- Contingency = association

	Treatment 1	Treatment 2	Total
Success	a	b	$a + b$
Failure	c	d	$c + d$
Total	$a + c$	$b + d$	$a+b+c+d$

2x2 contingency table

Test of independence

- Two variables (in columns and rows)
- E.g. treatments vs outcomes
- H_0: variables are independent
- Ms Bristol's answers do not depend on whether she got milk or tea first; they are random

Columns / Rows

	Treatment 1	Treatment 2	Total
Success	a	b	$a + b$
Failure	c	d	$c + d$
Total	$a + c$	$b + d$	$a+b+c+d$

2x2 contingency table

Tea served	T	T	M	T	T	M	T	M	T	T	M	M
Ms. Bristol	T	M	M	M	T	T	T	T	T	M	T	T

$p = 0.58$

Tea served	T	T	M	T	T	M	T	M	T	T	M	M
Ms. Bristol	T	T	M	M	T	M	M	M	T	T	M	M

$p = 0.03$

Test of proportion

Tea served	T	T	M	T	T	M	T	M	T	T	M	M
Ms. Bristol	T	M	M	M	T	T	T	T	T	M	T	T

4	5
2	1

4:5
2:1

$p = 0.58$

Tea served	T	T	M	T	T	M	T	M	T	T	M	M
Ms. Bristol	T	T	M	M	T	M	M	M	T	T	M	M

5	2
0	5

5:2
0:5

$p = 0.03$

Proteomics example

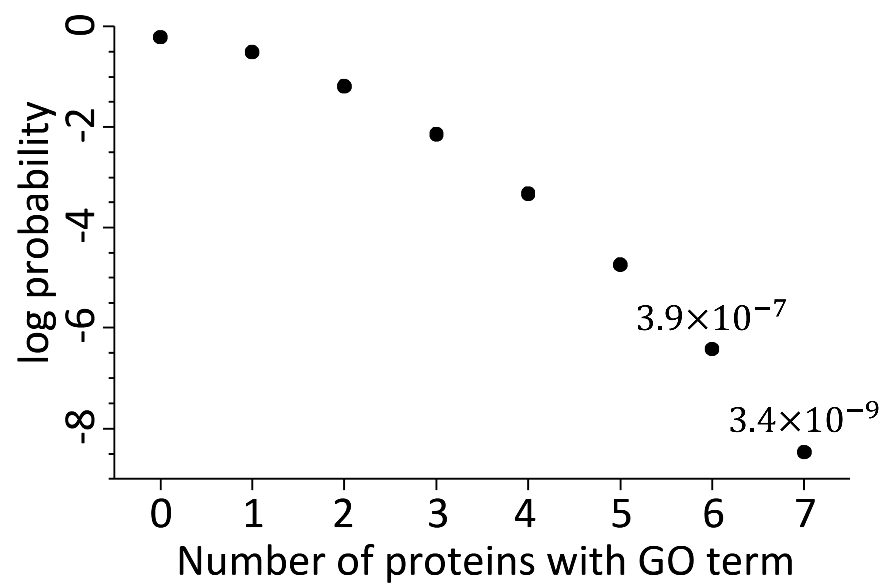
- There are 668 proteins in an experiment
- 7 of them have an associated Gene Ontology term (GO:00301174, regulation of DNA replication initiation)

	In cluster	Outside cluster	Total
With GO-term	6	1	7
Without GO-term	38	623	661
Total	44	624	668

- We have a cluster of 44 proteins with similar properties
- 6 of them have this GO term

- Is it significantly enriched?

$$P(X \geq 6) \approx 4 \times 10^{-7}$$



Absolute numbers are important

- A newspaper reports clinical tests on a new cancer drug
- 15% of patients treated with drug A survived
- 30% of patients treated with drug B survived
- So, drug B is 100% better than drug A!

Absolute numbers are important

- A newspaper reports clinical tests on a new cancer drug
- 15% of patients treated with drug A survived
- 30% of patients treated with drug B survived
- So, drug B is 100% better than drug A!
- Actual numbers: 20 and 10 patients
- $P(X \leq 3) = 0.31$

	Alive	Dead	Total
Drug A	3	17	20
Drug B	3	7	10
Total	6	24	30

p = 0.31

Absolute numbers are important

- A newspaper reports clinical tests on a new cancer drug
- 15% of patients treated with drug A survived
- 30% of patients treated with drug B survived
- So, drug B is 100% better than drug A!
- Actual numbers: 20 and 10 patients
- $P(X \leq 3) = 0.31$

	Alive	Dead	Total
Drug A	3	17	20
Drug B	3	7	10
Total	6	24	30

p = 0.31

- If we had 80 and 100 patients and the same proportions
- $P(X \leq 12) = 0.013$

	Alive	Dead	Total
Drug A	12	68	80
Drug B	30	70	100
Total	42	138	180

p = 0.013

- Moral 1: don’t trust newspapers
- Moral 2: estimate the size of your sample before you do your experiment, so the result is more significant

Never, ever use percentages in Fisher's test!

	Alive	Dead	Total
Drug A	15%	85%	
Drug B	30%	70%	
Total			

Fisher’s exact test: summary

Input	2×2 contingency table typically rows = groups, columns = conditions table contains counts
Usage	Examine if there is an association (contingency) between two variables; whether the proportions in one variable depend on the proportions in the other variable; if there is enrichment
Null hypothesis	The proportions in one variable do not depend on the proportions in the other variable
Comments	Exact test – count all possible combinations Use when you have small numbers For large numbers (hundreds) use chi-square test

How to do it in R?

```
# Tea tasting
> fisher.test(rbind(c(3, 1), c(1, 3)), alternative="greater")
          Fisher's Exact Test for Count Data
data:  rbind(c(3, 1), c(1, 3))
p-value = 0.2429
alternative hypothesis: true odds ratio is greater than 1
95 percent confidence interval:
 0.3135693       Inf


# GO enrichment
> fisher.test(rbind(c(6, 1), c(38, 623)), alternative="greater")
          Fisher's Exact Test for Count Data
data:  rbind(c(6, 1), c(38, 623))
p-value = 3.894e-07
alternative hypothesis: true odds ratio is greater than 1
95 percent confidence interval: 14.29724      Inf
```

Hand-outs available at http://is.gd/statlec