Foundations of Data Science: Hypothesis testing

THE UNIVERSITY *of* EDINBURGH
informatics

FOUNDATIONS OF DATA SCIENCE

Today

1. Principle of hypothesis testing

2. p-values

3. Testing for goodness of fit to a model

4. Issues in hypothesis testing

Foundations of Data Science: Hypothesis testing - Principle of hypothesis testing

THE UNIVERSITY of EDINBURGH
informatics
FOUNDATIONS OF DATA SCIENCE

Inferential statistics tasks: Hypothesis testing

Yes/no questions: E.g. 1: "Is Chocolate good for you"

E.g. 2: Is a coin biased?

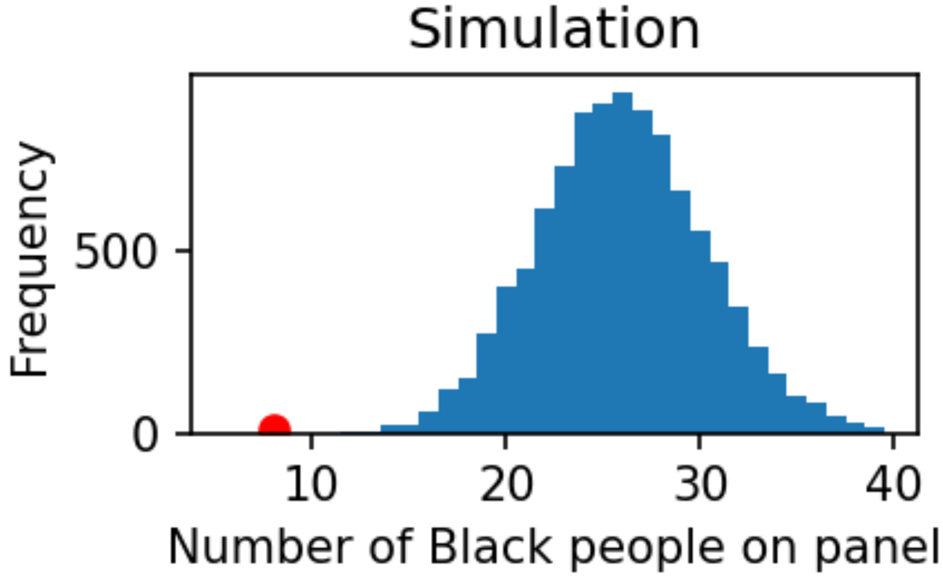
E.g. 3: Swain versus Alabama (1965).
Is this jury selection procedure biased?

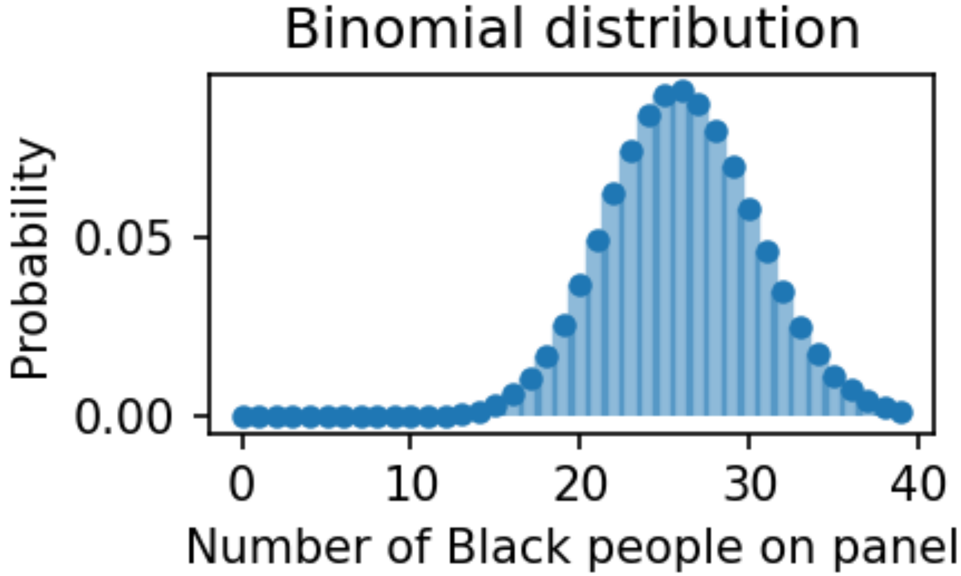
Population of Alabama
26% Black
74% Non-black

selection procedure

Jury panel of 100:
8 Black and
92 Non-black

Swain versus Alabama simulation results





Method of hypothesis testing

Null hypothesis H_0: Claim initially assumed to be true, formalised as a statistical model

E.g.: The jury panel was chosen by random selection from the population in the district.

Alternative hypothesis H_a: Claim contradictory to H_0 typically not formalised as a statistical model

E.g.: The jury was chosen by some other, unspecified, method

AIM: Reject or not reject H_0

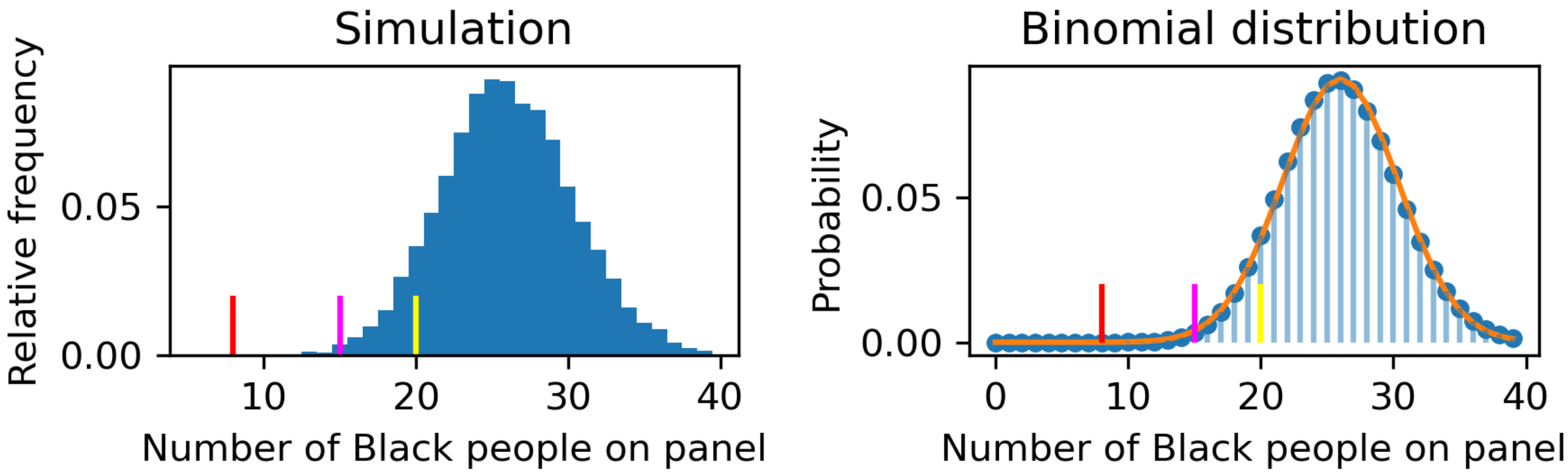
Test procedure

1. Test statistic: e.g. number of black people on a jury panel

$t_0 = 8$ (observed)

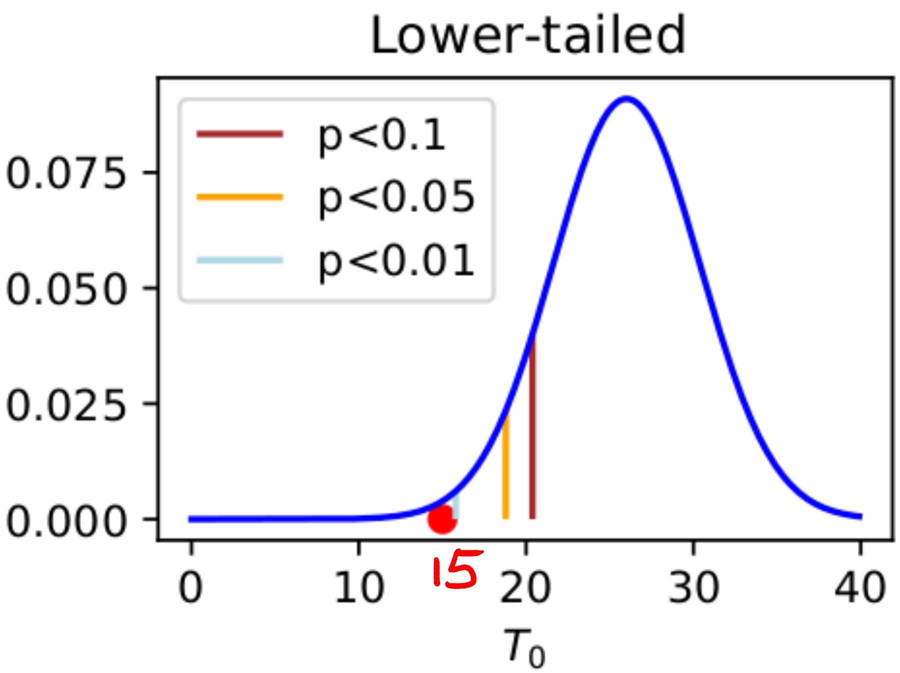
2. Distribution of the test statistic under H_0

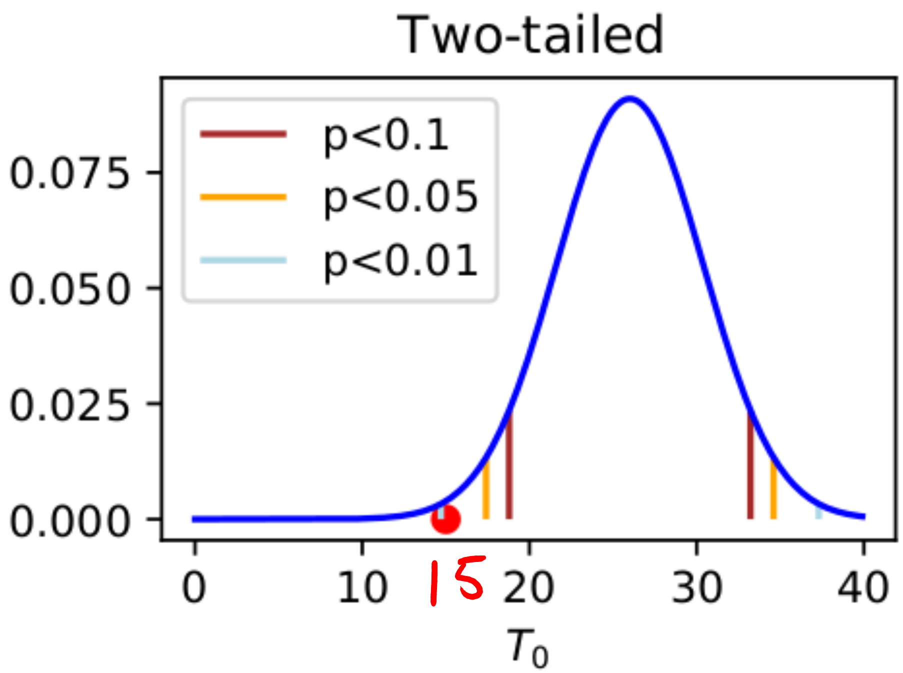
T_0 - random variable



3. (a) Rejection region
   (b) Return a p-value

Rejection regions





Number of black people is below the number expected by chance

Number of black people is different from the number expected by chance

Normal approximation to the binomial distribution

n large $\Rightarrow$ binomial dist is approx normal with

$\mu = np$ and $\sigma^2 = np(1-p)$ $= 100 \times 0 \cdot 26 \times (1 - 0 \cdot 26)$

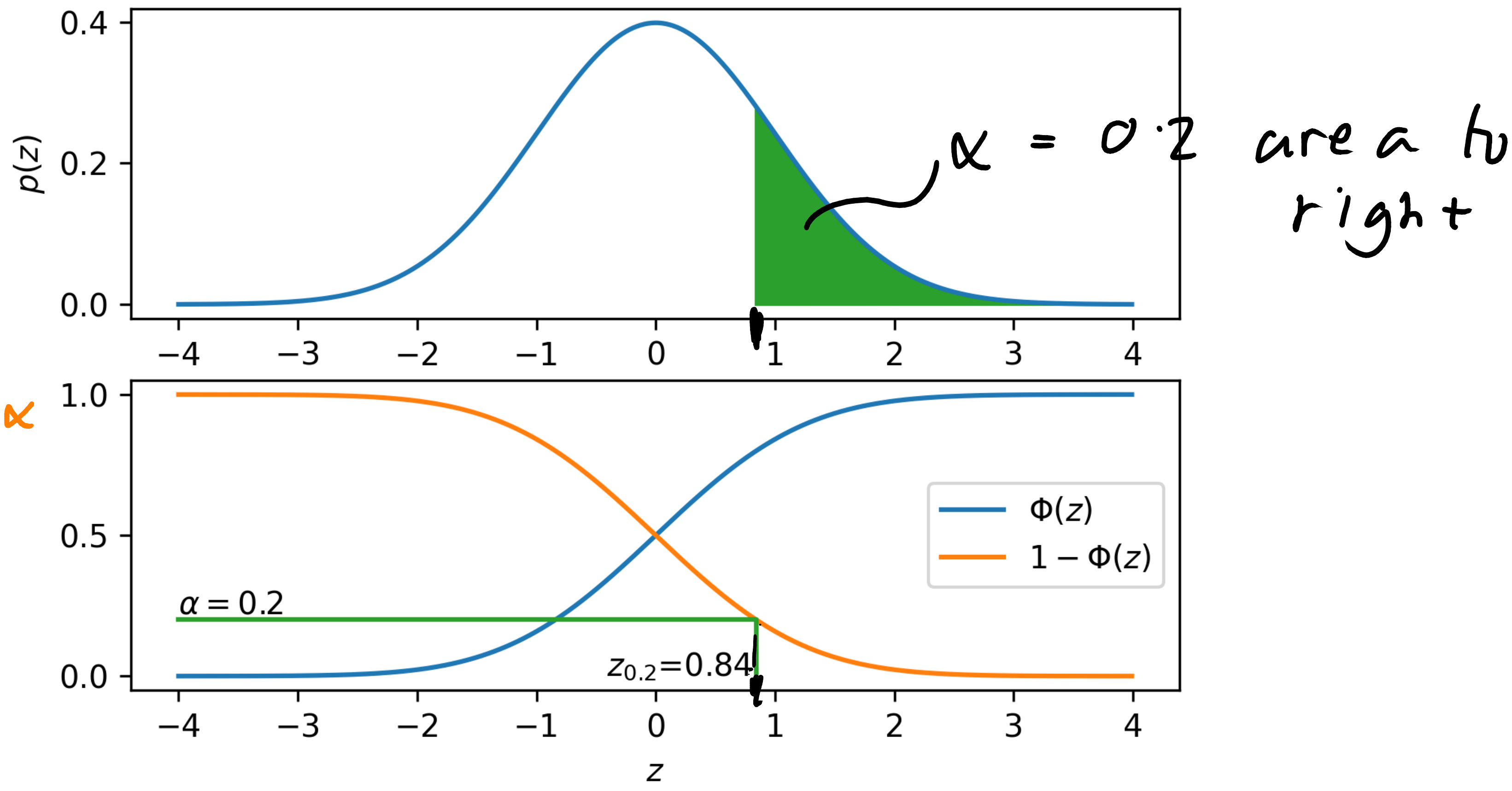
$\Rightarrow$ $Z = \frac{T_0 - \mu}{\sigma}$ has a z- distribution

1% rejection region has 99% of weight to its right $\Rightarrow$

At boundary of 1% rejection region

$$Z = z_{0.99} = \frac{T_0 - \mu}{\sigma} \Rightarrow T_0 = \mu + \sigma z_{0.99}$$

z-critical values



Aspects of hypothesis testing

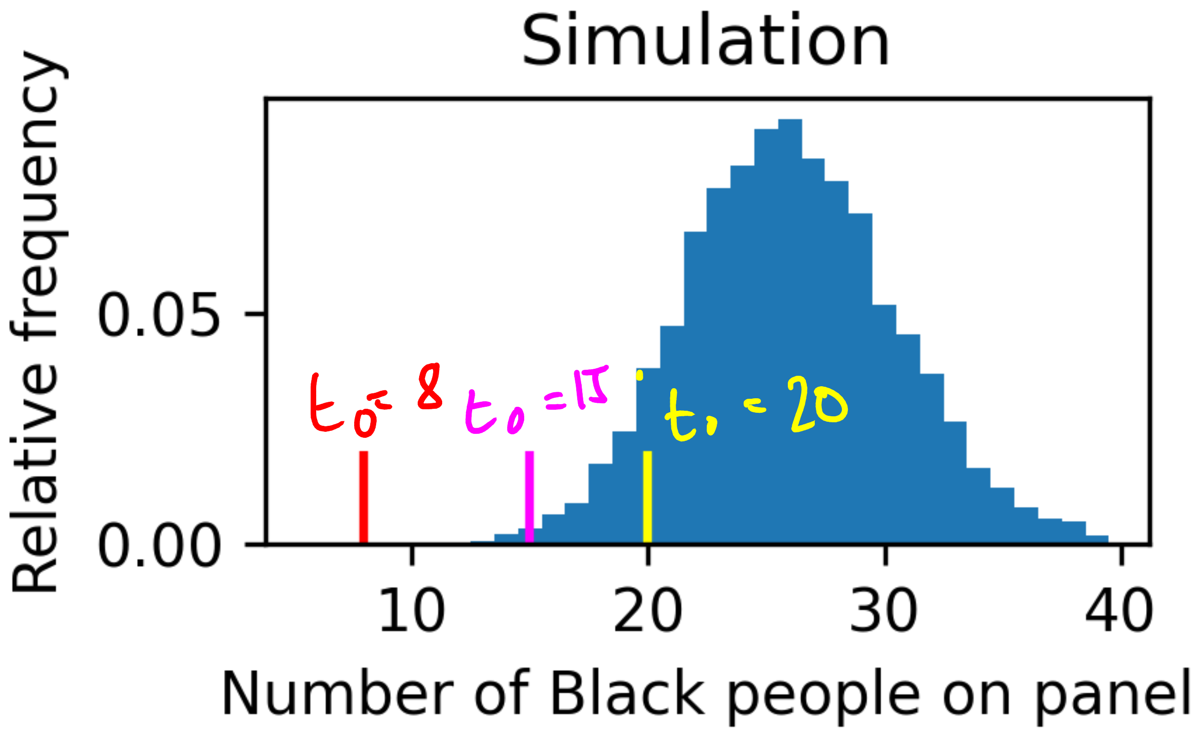
1. Decide whether a hypothesis or model is compatible with data from observaitional studies or randomised experiments

2. Investigate mechanisms specific to data

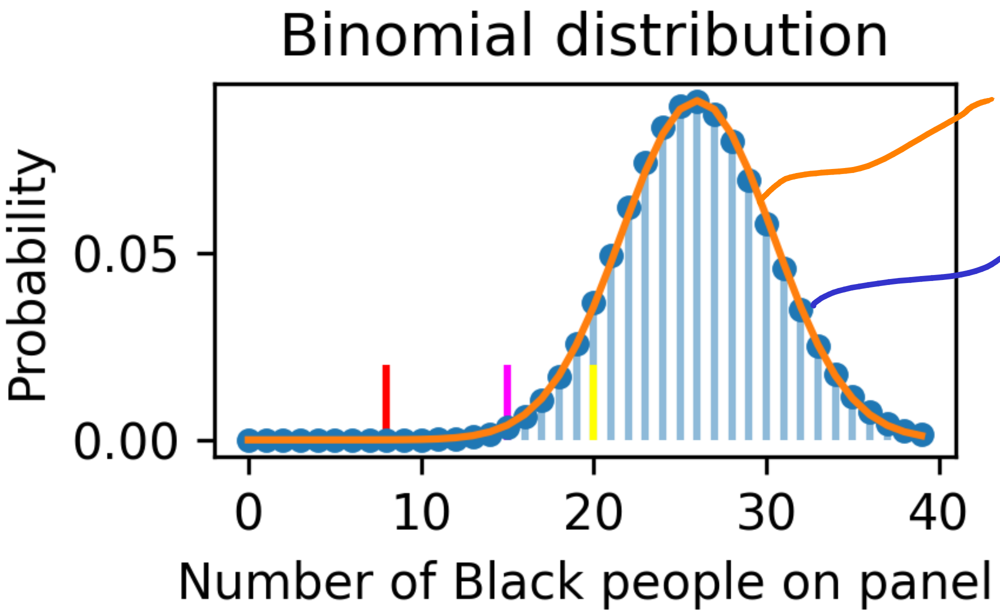
Foundations of Data Science: Hypothesis testing - p-values

Principle of p-values

Observed data ⟹ boundary of rejection region



Determining p-values from probability dists



Normal dist

$\mu = np$

$\sigma = \sqrt{np(1-p)}$

$b(t_0; n, p)$

$n = 100$

$p = 0.26$

Binomial

cumulative dist.

$$\text{p-value} = P(T_0 \leq t_0) = B(t_0; n, p) = \sum_{t=0}^{t_0} b(t; n, p)$$

Normal approximation

$$\text{p-value} = \Phi\left(\frac{t_0 - \mu}{\sigma}\right)$$

where $\Phi(z)$ cumulative dist function of z-distribution

P-values computed by various methods

t_0	Simulation	Binomial	Normal
8	0	4.73e-06	2.03e-05
15	0.0067	0.0061	0.0061
20	0.1020	0.1030	0.0857

Definition of p-value

The p-value is the probability, calculated assuming the null hypothesis is true, of obtaining a value of the test statistic at least as contradictory to H_0 as the value calculated from the available sample. (*Modern Mathematical Statistics with Applications*, p. 456)

THE AMERICAN STATISTICIAN
2016, VOL. 70, NO. 2, 129–133
http://dx.doi.org/10.1080/00031305.2016.1154108

Taylor & Francis
Taylor & Francis Group

EDITORIAL

The ASA's Statement on *p*-Values: Context, Process, and Purpose

What p-values are

P-**values can indicate how incompatible the data are with a specified statistical model**...

The smaller the *p*-value, the greater the statistical incompatibility of the data with the null hypothesis, if the underlying assumptions used to calculate the *p*-value hold. This incompatibility can be interpreted as casting doubt on or providing evidence against the null hypothesis or the underlying assumptions. (*ASA Statement on Statistical Significance and P-values*)

Question

In the hypothetical case of 20 black people on the jury, which has a p-value of 0.10, would the null hypothesis be true?

Why?

What p-values are not

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.

Researchers often wish to turn a p-value into a statement about the truth of a null hypothesis, or about the probability that random chance produced the observed data. The p-value is neither. It is a statement about data in relation to a specified hypothetical explanation, and is not a statement about the explanation itself. (*ASA Statement on Statistical Significance and P-values*)

.

"Statistical significance"

$p < 0.05 \Rightarrow$ "statistically significant"

*	significant at the $p < 0.05$ level
**	" " " $p < 0.01$ "
***	" " " $p < 0.001$ "

Statistical significance

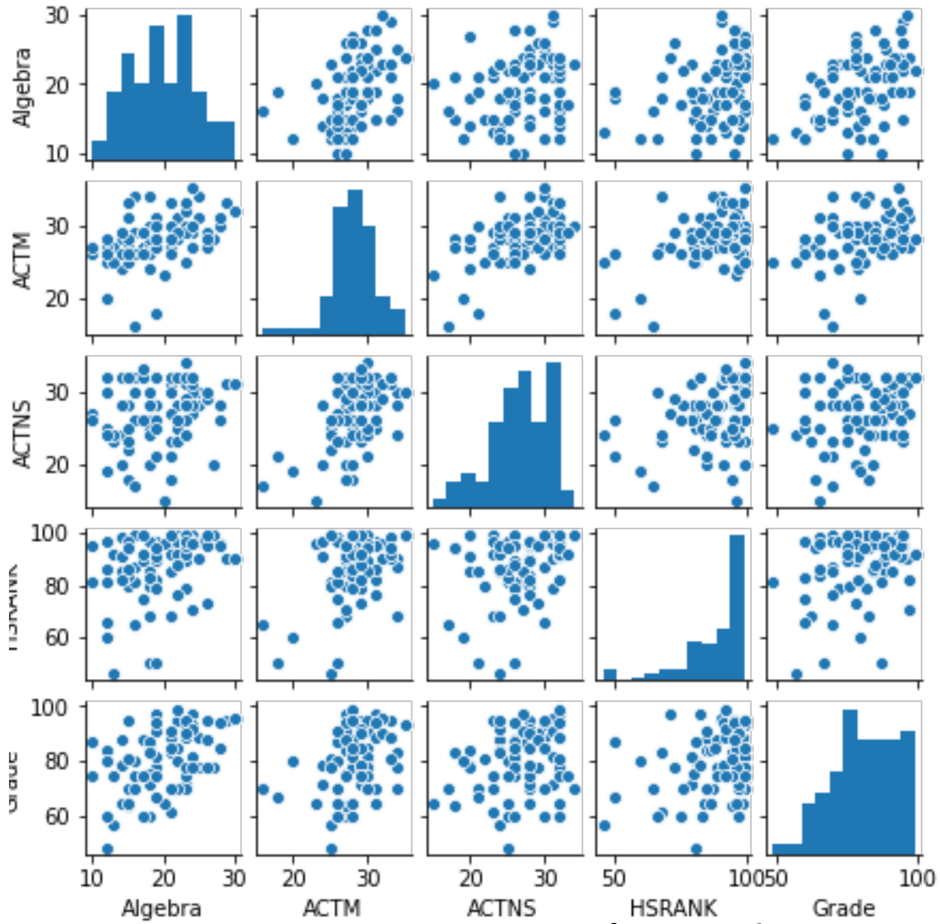
In February 2014, George Cobb, Professor Emeritus of Mathematics and Statistics at Mount Holyoke College, posed these questions to an ASA discussion forum:

> Q: Why do so many colleges and grad schools teach $p = 0.05$?
> A: Because that's still what the scientific community and journal editors use.
> Q: Why do so many people still use $p = 0.05$?
> A: Because that's what they were taught in college or grad school.

Confidence intervals and p-values

Dep. Variable:	Grade	**R-squared:**	0.289
Model:	OLS	**Adj. R-squared:**	0.251
Method:	Least Squares	**F-statistic:**	7.622
Date:	Wed, 26 Oct 2022	**Prob (F-statistic):**	3.30e-05
Time:	09:42:47	**Log-Likelihood:**	-294.31
No. Observations:	80	**AIC:**	598.6
Df Residuals:	75	**BIC:**	610.5
Df Model:	4		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	36.1215	10.752	3.360	0.001	14.703	57.540
Algebra	0.9610	0.264	3.640	0.000	0.435	1.487
ACTM	0.2718	0.454	0.599	0.551	-0.632	1.175
ACTNS	0.2161	0.313	0.690	0.492	-0.408	0.840
HSRANK	0.1353	0.104	1.306	0.196	-0.071	0.342



$x^{(1)}$ $x^{(2)}$ $x^{(3)}$ $x^{(4)}$ y

Edge and Friedberg (1984)

Foundations of Data Science: Hypothesis testing - Testing for goodness-of-fit

Multiple categories

American Civil Liberties Union investigation into jury selection in Alameda County, CA

	Caucasian	Black/AA	Hispanic	Asian/PI	Other	Total
Population %	54	18	12	15	1	100
Observed panel numbers	780	117	114	384	58	1453
Expected panel numbers	784.62	261.54	174.36	217.95	14.53	1453.00
$\frac{(\text{Observed}-\text{Expected})^2}{\text{Expected}}$	0.03	79.88	20.90	126.51	130.05	357.36

H_0: The panels were chosen by random selection from the population

H_a: The panels were chosen by some other, unspecified method.

1. Test statistic

k - groups

p_i - population proportion in the ith group

n_i - observed number in ith group

n - total size of population $\quad n = \sum_{i=1}^{k} n_i$

np_i - expected number in each group.

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i}$$

"chi-squared"

Generally used to measure goodness-of-fit.

e.g.

$np_i = 100 \quad n_i = 95$ — 5, 5%

$np_i = 10 \quad n_i = 5$ — 5, 50%

$\chi^2 = 357.36$

2. H_0 formulated as a statistical model

Draw $n_1, \dots, n_k$ from Multinomial distribution

$$p(n_1, \dots n_k) = n! \frac{p_1^{n_1} \cdot \dots \cdot p_k^{n_k}}{(n_1!) \dots (n_k!)}$$

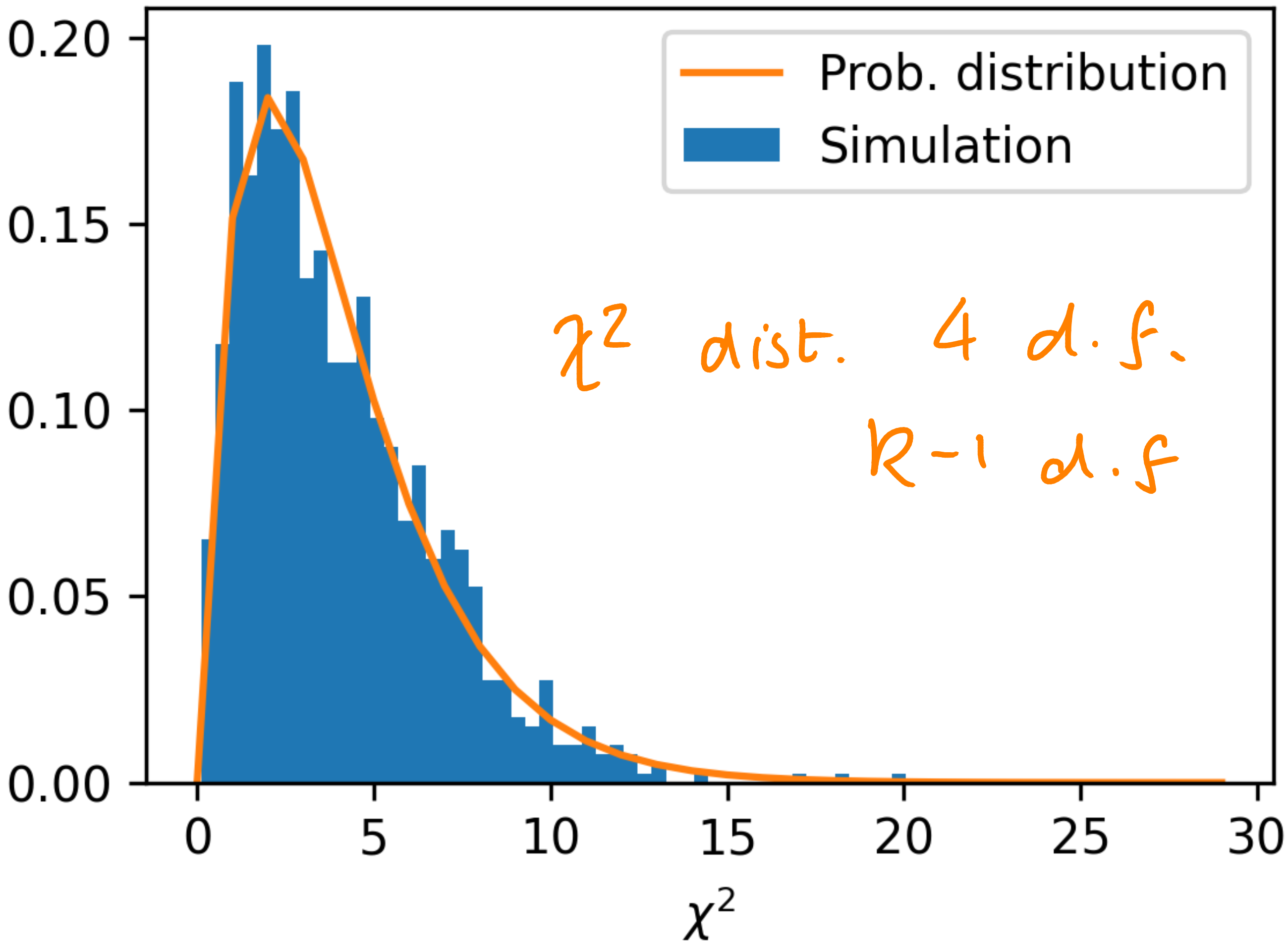
but constrained so that $\sum_{i=1}^{k} n_i = n$

$\Rightarrow$ $k-1$ degrees of freedom.

CODE

2. Distribution of test statistic under H_0

$n = 1543$



χ^2 dist. 4 d.f.

$k-1$ d.f

$\rightarrow 357$

$p \approx 0$

2-way contingency tables

	Female	Male	Total
Depressed	30	12	42
Not depressed	2048	1663	3711
Total	2078	1675	3753

	Population 1	Population 2	Total
Category 1	n_{11}	n_{12}	$n_{1\bullet}$
Category 2	n_{21}	n_{22}	$n_{2\bullet}$
Total	$n_{\bullet 1}$	$n_{\bullet 2}$	$n_{\bullet\bullet}$

H_0 = Being severely depressed is independent of being female or male

H_a = Some other, unspecified hypothes.

$$H_0 = P(X = x \mid Y = y) = P(X = x)\, P(Y = y)$$

$$p_{i\bullet} = \frac{n_{i\bullet}}{n_{\bullet\bullet}} \qquad p_{\bullet j} = \frac{n_{\bullet j}}{n_{\bullet\bullet}}$$

$$\Rightarrow \quad \hat{e}_{ij} = n_{\bullet\bullet}\, p_{i\bullet}\, p_{\bullet j} = \frac{n_{i\bullet}\, n_{\bullet j}}{n_{\bullet\bullet}}$$

Multiway contingency tables

$J=2$

$I=2$

	Female	Male	Total
Depressed	30	12	42
Not depressed	2048	1663	3711
Total	2078	1675	3753

	Population 1	Population 2	Total
Category 1	n_{11}	n_{12}	$n_{1\bullet}$
Category 2	n_{21}	n_{22}	$n_{2\bullet}$
Total	$n_{\bullet 1}$	$n_{\bullet 2}$	$n_{\bullet\bullet}$

	Female	Male
Depressed	23.25	18.75
Not depressed	2054.75	1656.25

	Population 1	Population 2
Category 1	$\hat{e}_{11}$	$\hat{e}_{12}$
Category 2	$\hat{e}_{21}$	$\hat{e}_{22}$

$$\chi^2 = \sum \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}} = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(n_{ij} - \hat{e}_{ij})^2}{\hat{e}_{ij}}$$

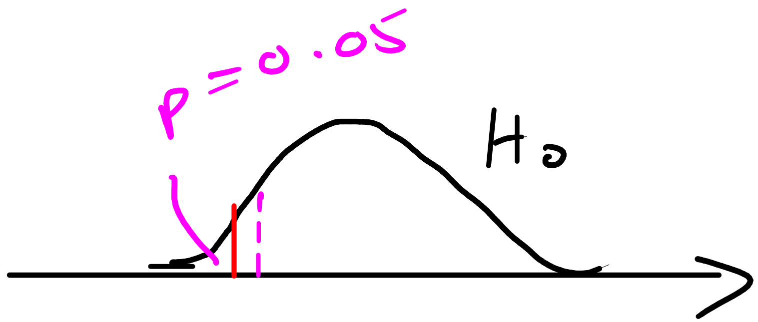
$$= 4.433$$

$$\# \text{ d.f} = (I-1)(J-1)+1 = 1$$

$$p = 0.035$$

Foundations of Data Science: Hypothesis testing - Issues in hypothesis testing

Type I and Type II Errors



Type I error: Rejecting H_0 when it is true

Control by setting α – size of rejection region

Type II error: not rejecting H_0 when it is false.

"Cherry-picking", "Data dredging", "p-value hacking"

Proper inference requires full reporting and transparency.

P-values and related analyses should not be reported selectively. Conducting multiple analyses of the data and reporting only those with certain *p*-values (typically those passing a significance threshold) renders the reported *p*-values essentially uninterpretable. Cherry-picking promising findings, also known by such terms as data dredging, significance chasing, significance questing, selective inference, and "p-hacking," leads to a spurious excess of statistically significant results in the published literature and should be vigorously avoided... (*ASA Statement on Statistical Significance and P-values*)

Type I error = prob. rejecting H_0 when it is true

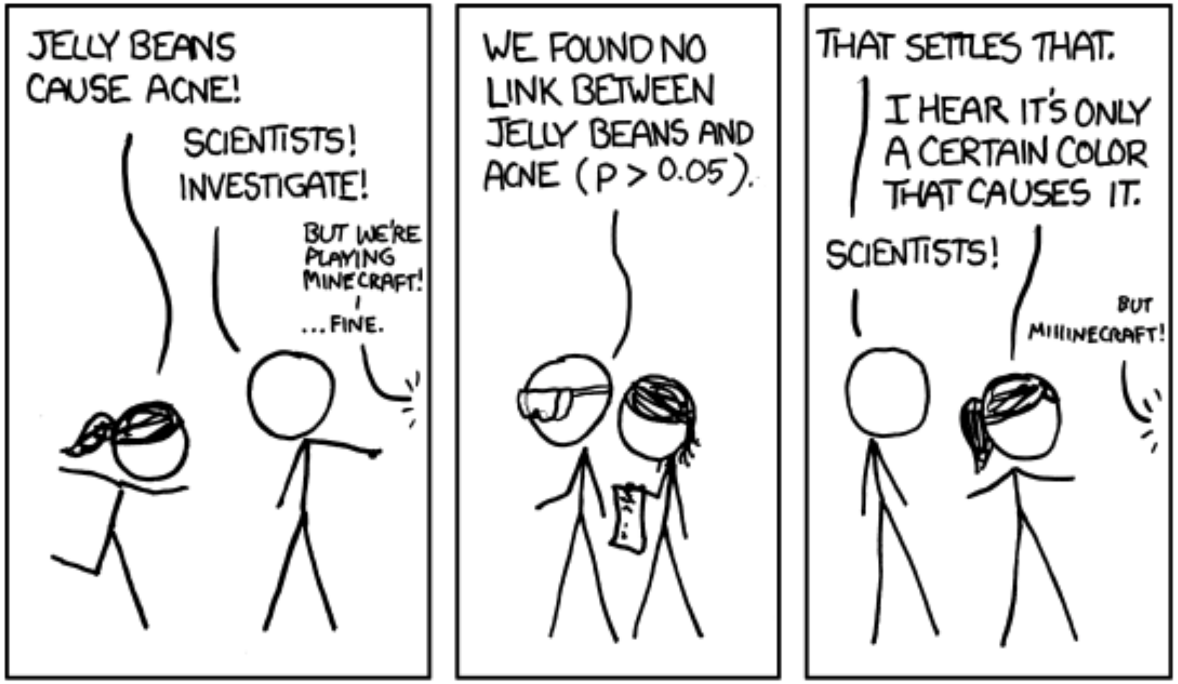
Multiple testing

Suppose 20 tests ; 0.05 chance Type I error on each test

$\Rightarrow$ 0.95 chance of no type I error one test

$\Rightarrow$ 0.95^{20} chance no type I errors overall

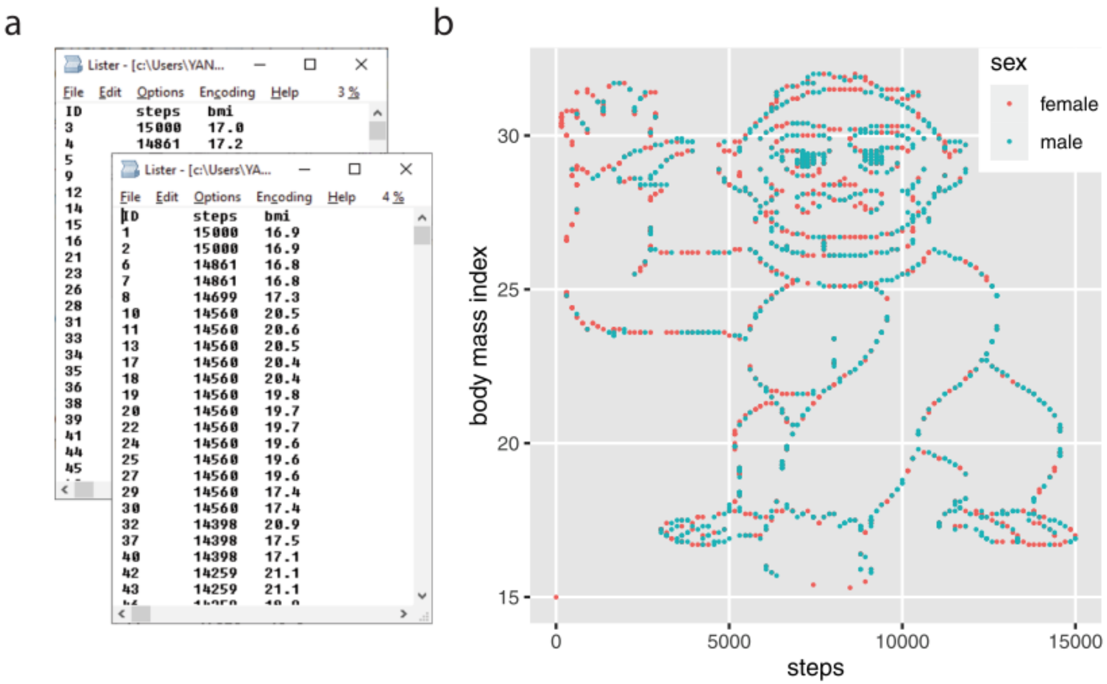
$\Rightarrow$ $1 - 0.95^{20} = 0.64$ chance type I error



Read the whole cartoon at:

https://www.explainxkcd.com/wiki/index.php/882:_Significant

"A hypothesis is a liability"



c

	Gorilla not discovered	Gorilla discovered
Hypothesis-focused	14	5
Hypothesis-free	5	9

Fig. 1 a An artificial dataset given to students with and without explicit hypotheses on the relationship between BMI and the steps taken on a particular day, for men and women. **b** A plot of the dataset. **c** The contingency table for students in the two groups ("hypothesis-focused," "hypothesis-free") that discovered the gorilla or not [6]

Yanai and Lercher (2020)

Summary

1. Principle of Hypothesis testing
   (a) Rejection method
   (b) p-values

2. Hypothesis testing applied to 3 problems involving testing if observed numbers are consistent with expected proportions
   - Many other uses

3. Limitations of hypothesis testing and p-values