

# Inf2 - Foundations of Data Science: Hypothesis testing



THE UNIVERSITY *of* EDINBURGH  
**informatics**

**FOUNDATIONS**  
**OF**  
**DATA**  
**SCIENCE**

# Plan for statistical inference

1. Randomness, sampling and simulations (S2 Week 1)
2. Estimation, including confidence intervals (S2 Week 2)
3. Hypothesis testing (S2 Week 3)
4. A/B testing (S2 Week 3)

Onwards to Logistic regression (S2 Week 4)

# Today

1. Principle of hypothesis testing (using statistical simulations)
2. p-values (using statistical simulations)
3. Issues in hypothesis testing
4. Theoretical methods
5. Practical applications
6. Example: testing for goodness of fit to a model

# Inf2 - 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

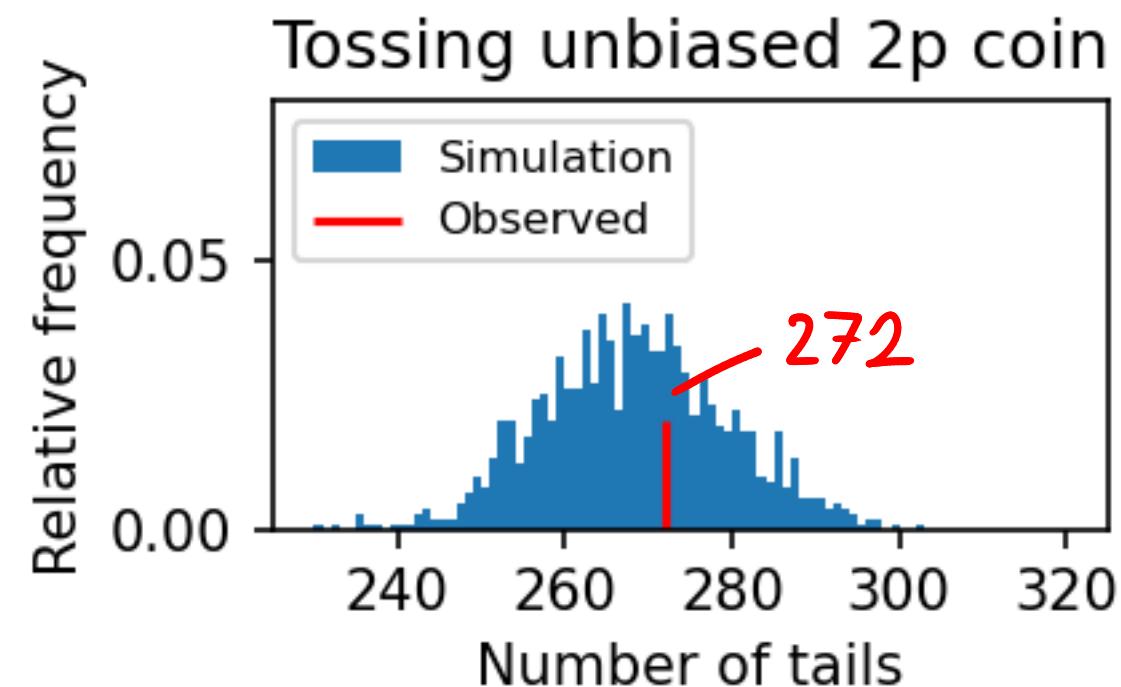
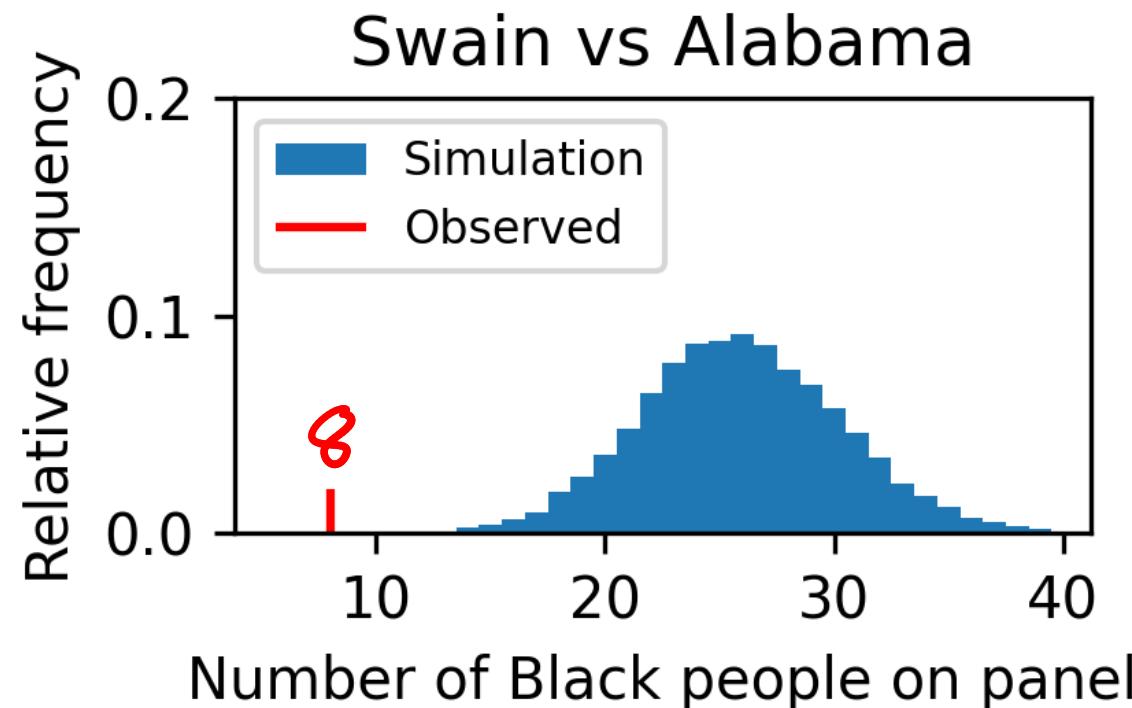
selection  
procedure

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

# Statistical simulation versus observations

Simulate unbiased procedures

Compare with observations



# Method of hypothesis testing

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

e.g.  $H_0$ : The jury panel was chosen by random selection from the population in the district.

e.g.  $H_0$  : The coin was unbiased

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

e.g.  $H_a$ : The jury was chosen by some other, unspecified, method that was unfavourable to Black people

e.g.  $H_a$  : The coin is biased (either towards heads or tails)

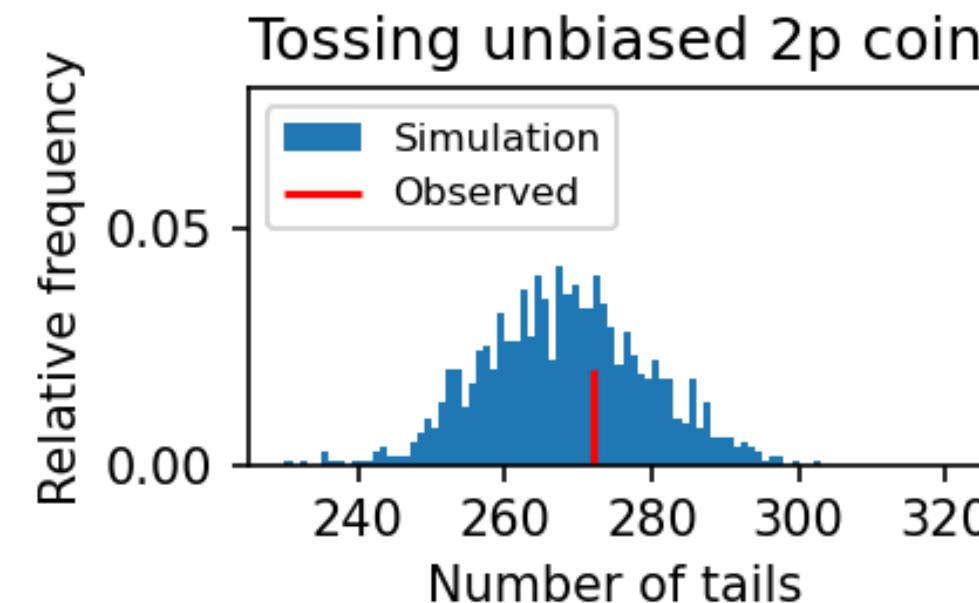
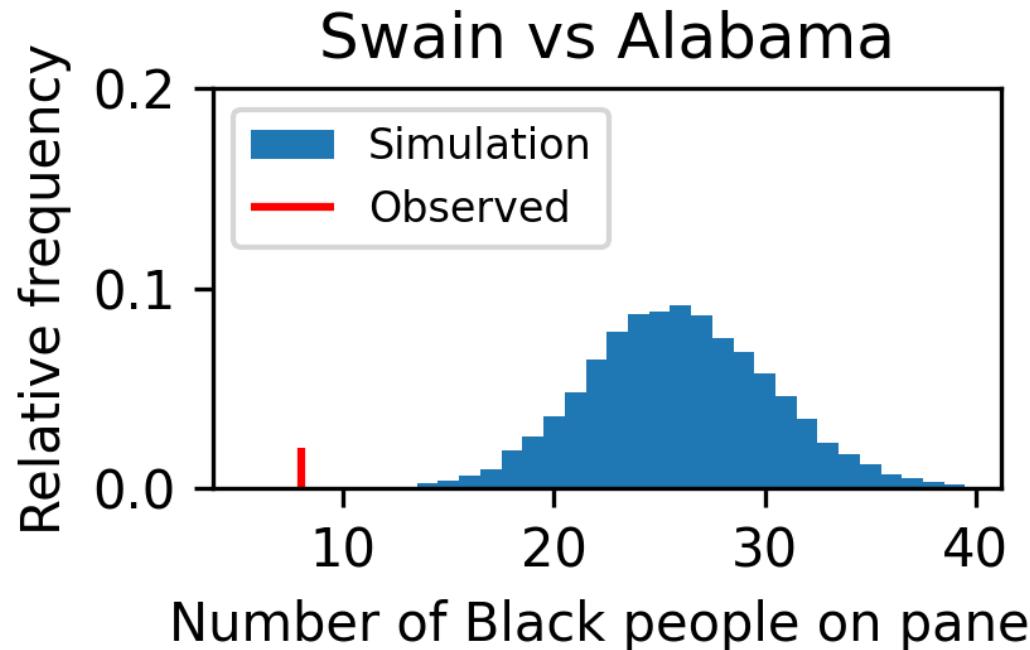
AIM: Reject or not reject  $H_0$

# Test procedure

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

$$t_0 = 8 \text{ (observed)}$$

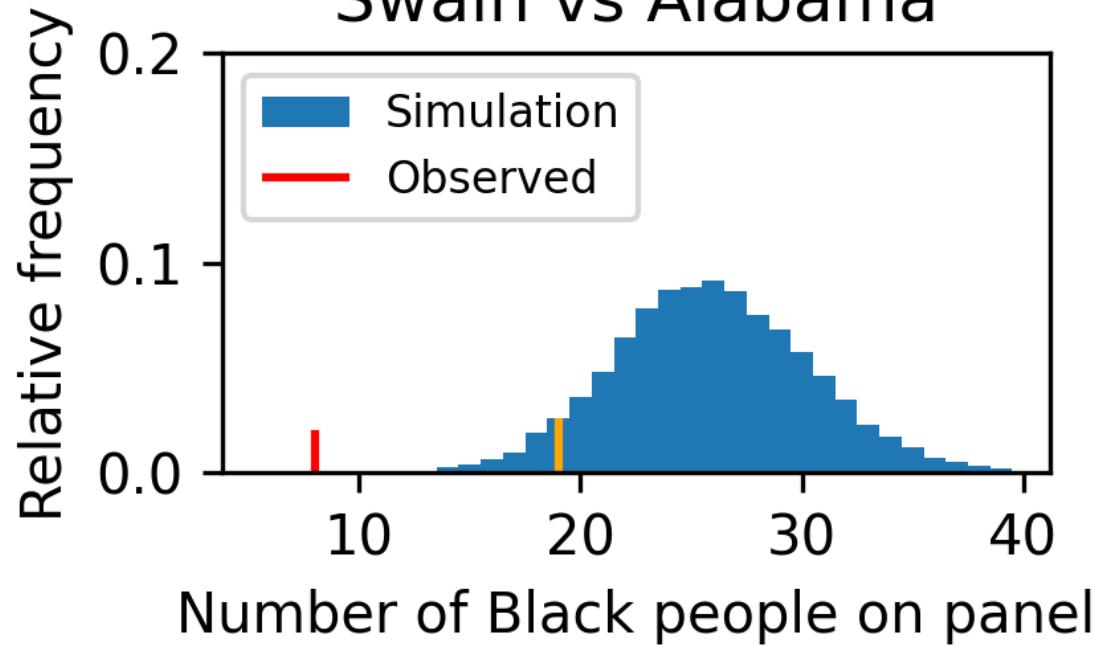
2. Distribution of the test statistic under  $H_0$



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

# One-tailed rejection regions

Swain vs Alabama



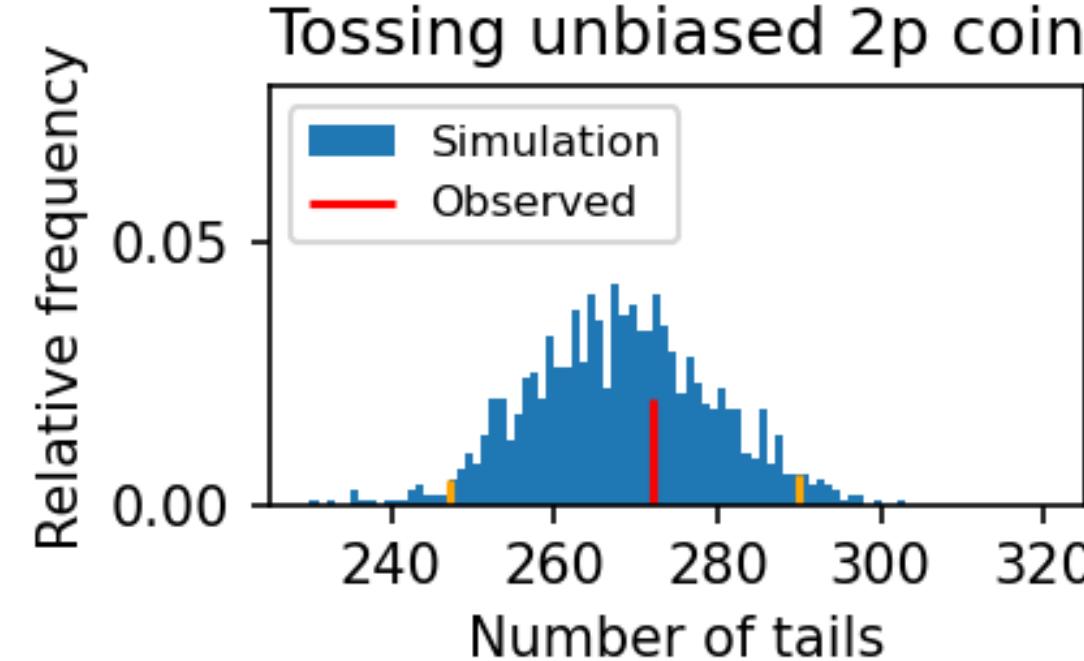
$H_a$ : Number of black people is below the number expected by chance

Observation in rejection region => reject, otherwise do not reject

Reject at 5% level?

# Two-tailed rejection regions

Tossing unbiased 2p coin



$H_a$ : Number of tails is different from the number expected by chance

Reject at 5% level?

# Inf2 - Foundations of Data Science: Hypothesis testing - p-values

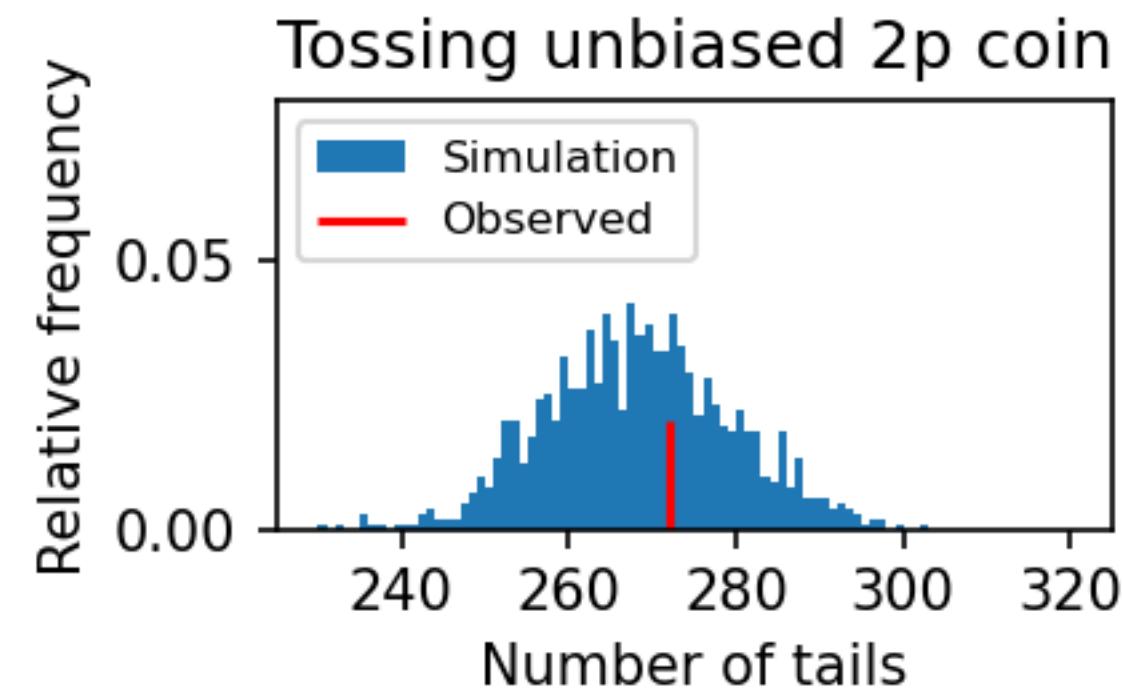
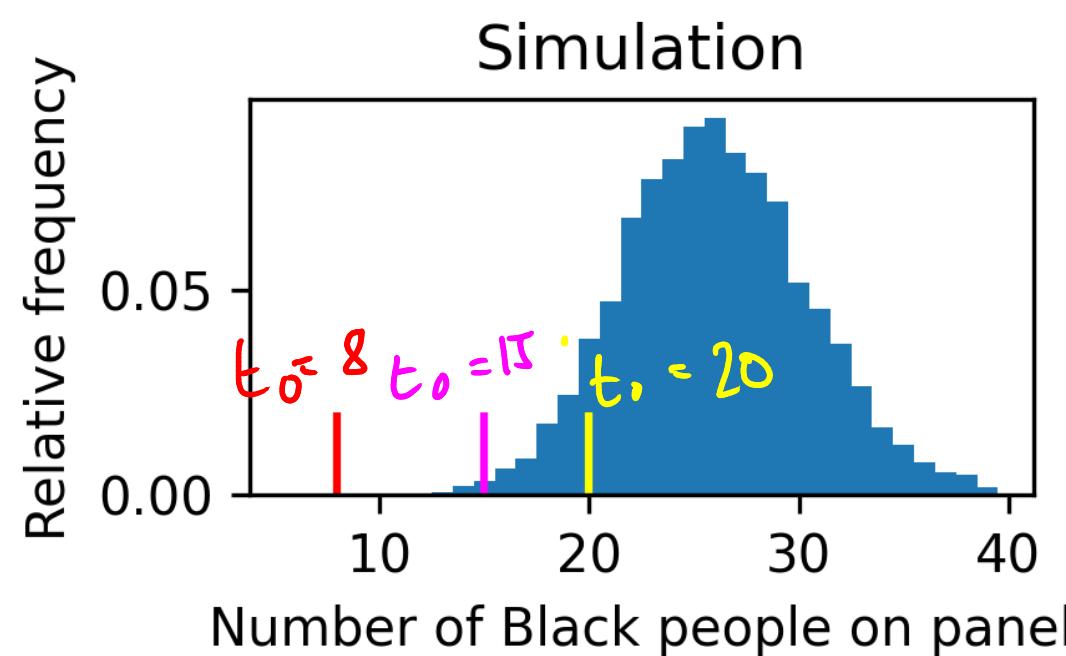


THE UNIVERSITY *of* EDINBURGH  
**informatics**

**FOUNDATIONS**  
**OF**  
**DATA**  
**SCIENCE**

# Principle of p-values

Observed data is boundary of rejection region



0.367

# Definitions of the p-value

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

Informally, a *p*-value is the probability under a specified statistical model that a statistical summary of the data (e.g., the sample mean difference between two compared groups) would be equal to or more extreme than its observed 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)

## Question

1. In the hypothetical case of 8 black people on the jury, which has a p-value of 0.10, is the null hypothesis true?
2. For the coin tossing, is the probability that 2p coins are unbiased equal to

$$p=0.118$$

or

$$1-p = 0.882 ?$$

# What p-values are and are not (ASA Statement on Statistical Significance and P-values)

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

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.

# Role of hypothesis testing

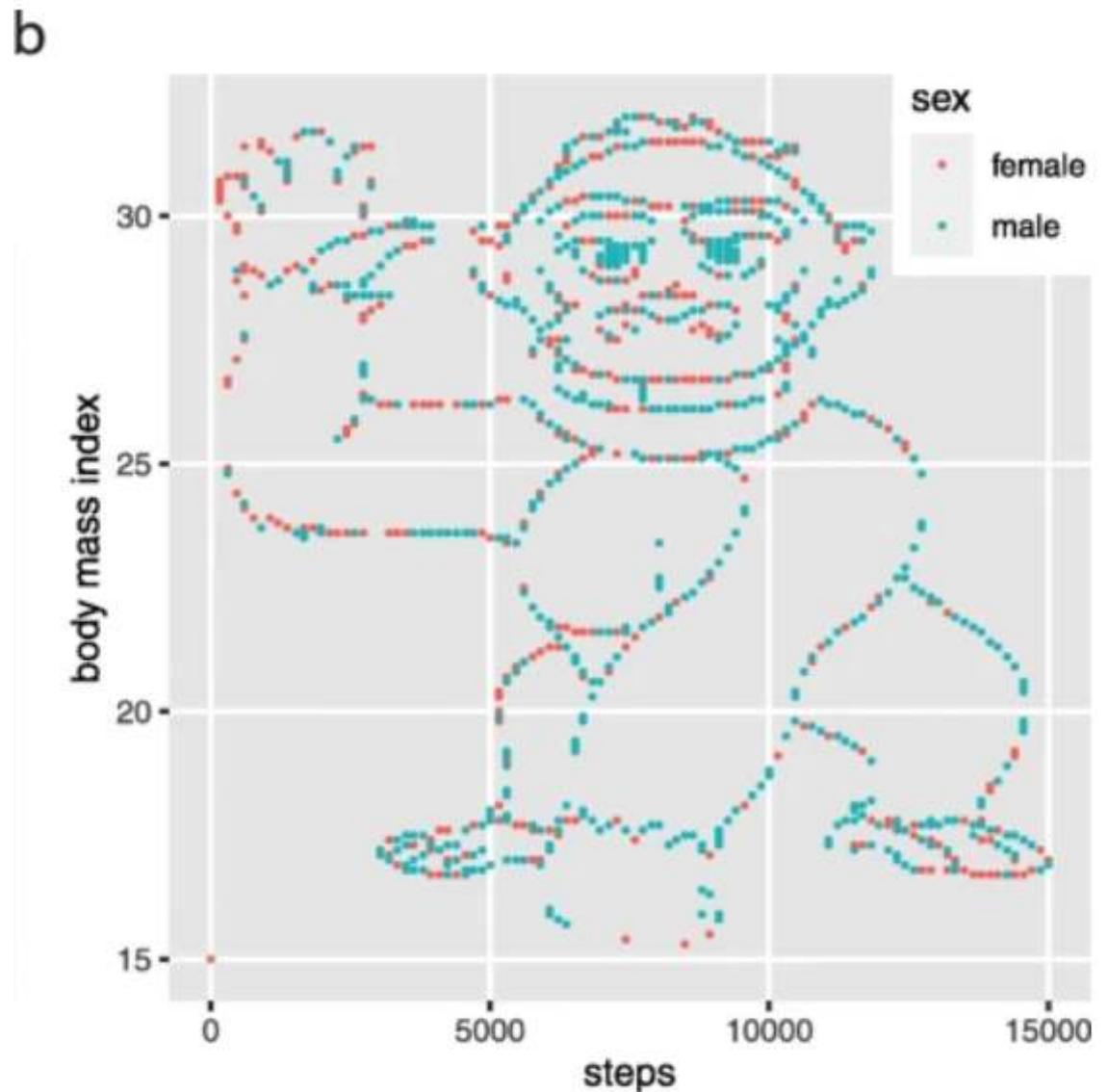
1. Decide whether a hypothesis or model is compatible with data from observational studies or randomised experiments
2. Investigate mechanisms specific to data

# Question

Suppose you are asked to investigate if there is a relationship between BMI (body mass index) and steps walked.

This data visualised apparently shows the relationship between BMI and steps walked each day by men and women.

Having seen this visualisation would you use this data to test if there is a relationship between BMI and number of steps walked?



# Inf2 - Foundations of Data Science: Hypothesis testing - Issues in hypothesis testing



THE UNIVERSITY *of* EDINBURGH  
**informatics**

FOUNDATIONS  
OF  
DATA  
SCIENCE

## "Statistical significance"

$p < 0.05 \Rightarrow$  "statistically significant"

\* significant at the  $p < 0.05$  level

\*\* " " "  $p < 0.01$  "

\*\*\* " " "  $p < 0.001$  "

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  $p=0.05$ ?

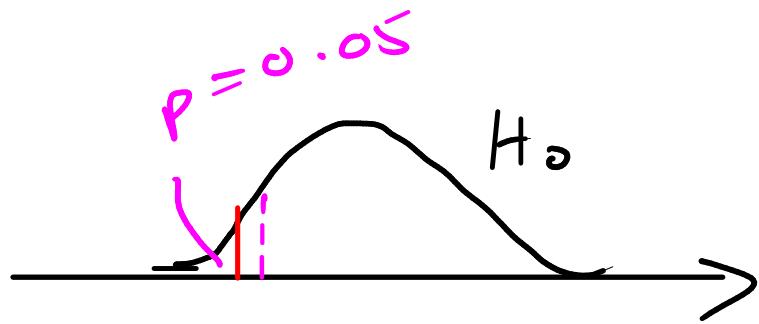
A: Because that's what they were taught at grad school.

- George Cobb, ASA Statement on p-values

# Question

In the coin tossing experiment, imagine that we repeat the experiment 1000 times and that we demand statistical significance at the 0.01 level. Assuming the null hypothesis is true (unbiased coin), on how many experiments do we expect to reject the null hypothesis?

# Type I and Type II Errors



Type I error: Rejecting  $H_0$  when it is true  
- control by setting  $\alpha$ -size of rejection region

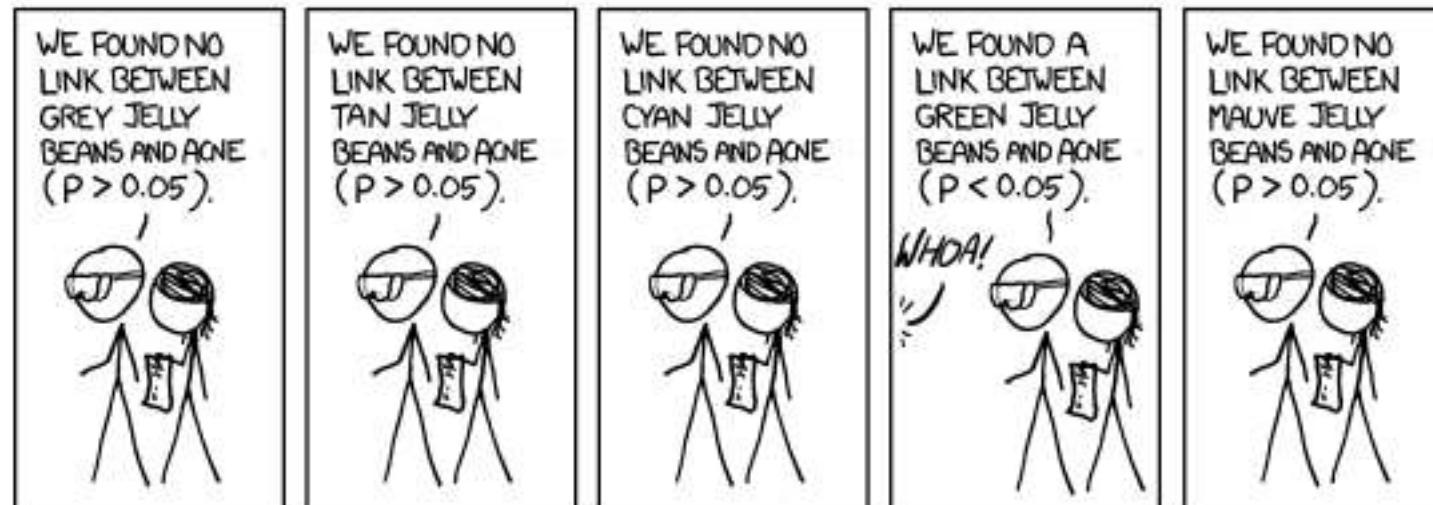
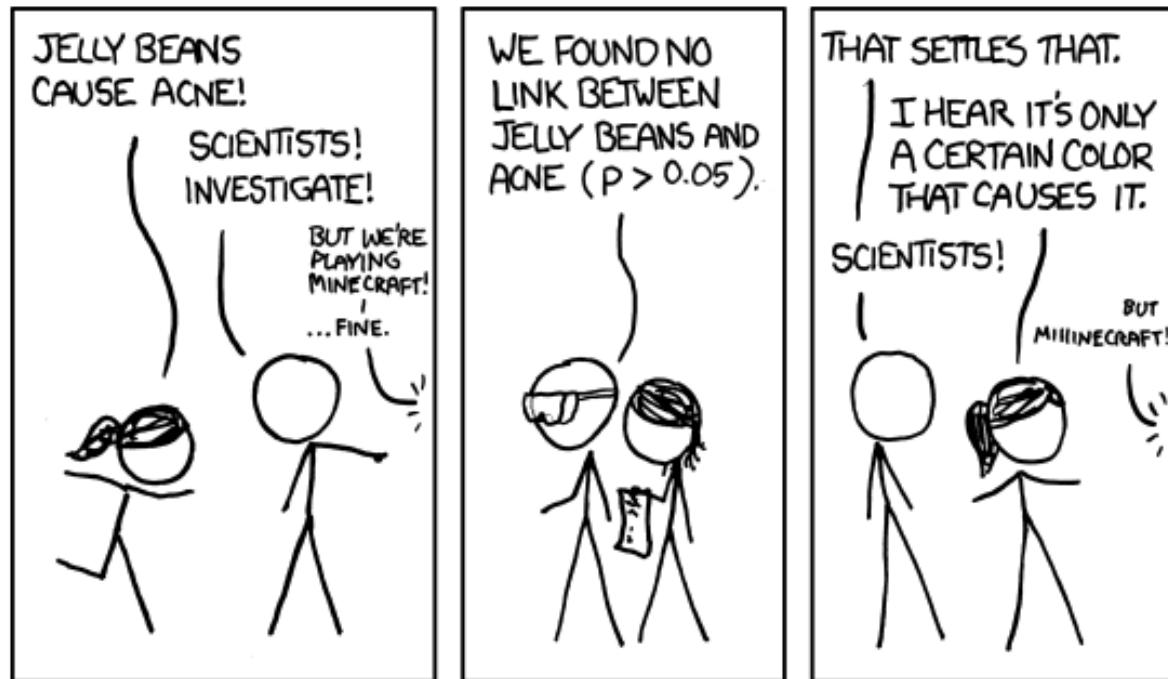
Type II error: not rejecting  $H_0$  when it is false  
- more difficult to control for

# Proper inference requires full reporting and transparency

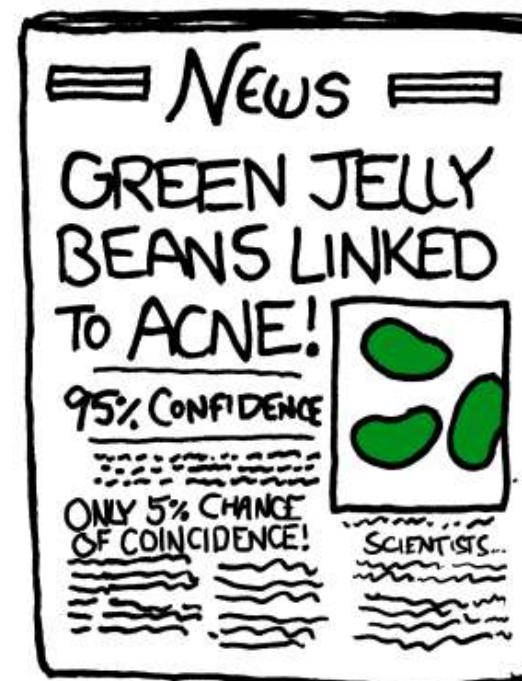
- P-values and related analyses should not be reported selectively.
- Conducting multiple analyses of the data and reporting only those with certain p-values (typically those passing a significance threshold) renders the reported p-values essentially uninterpretable.
- Cherry-picking promising findings, 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)

# Multiple testing



20 tests at with 5% rejection region  
=> 0.64 chance of at least one Type I error



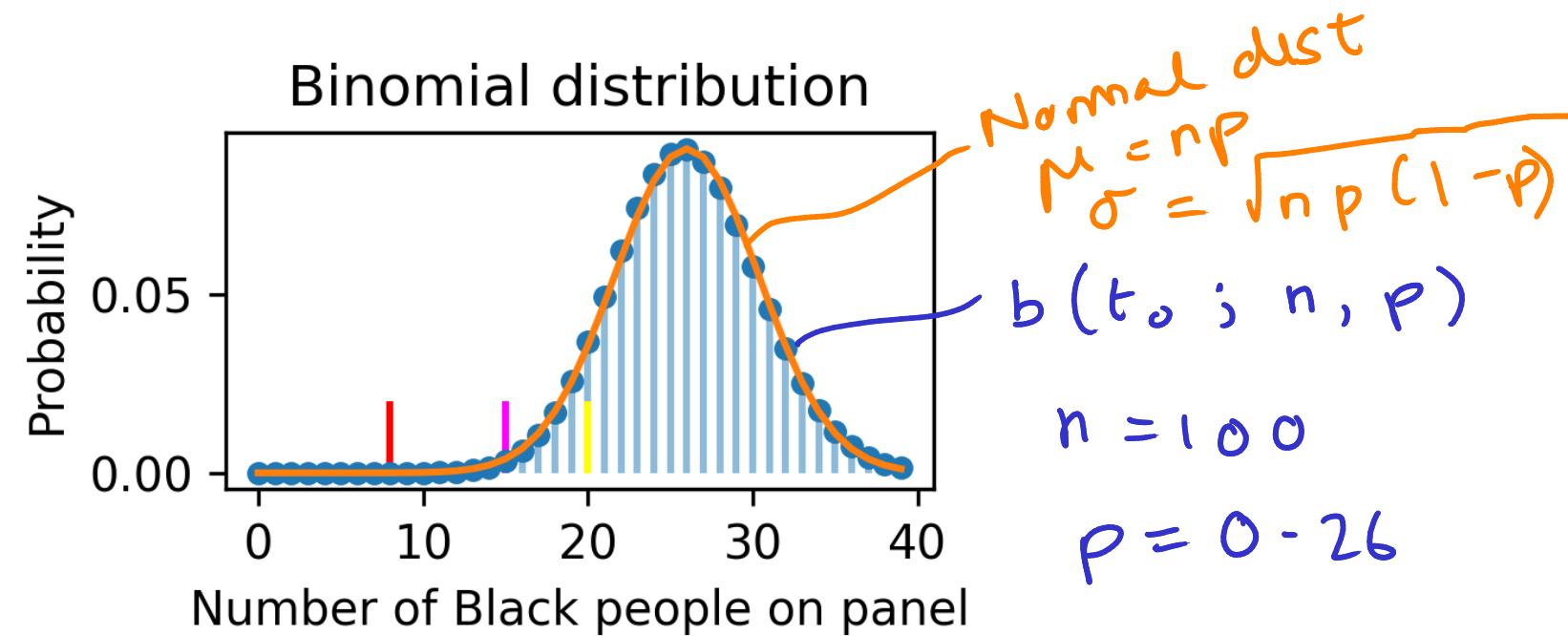
# Inf2 - Foundations of Data Science: Hypothesis testing - Theoretical methods



THE UNIVERSITY *of* EDINBURGH  
**informatics**

FOUNDATIONS  
OF  
DATA  
SCIENCE

# Determining p-values from probability dists



Binomial

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

cumulative dist.

Normal approximation

$$p\text{-value} = \Phi\left(\frac{t_0 - \mu}{\sigma}\right)$$
 where  $\Phi(z)$  cumulative dist. function of  $z$ -distribution

## Normal approximation to the binomial distribution

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

$$\mu = np \text{ and } \sigma^2 = np(1-p) = 100 \times 0.26 \times (1 - 0.26)$$

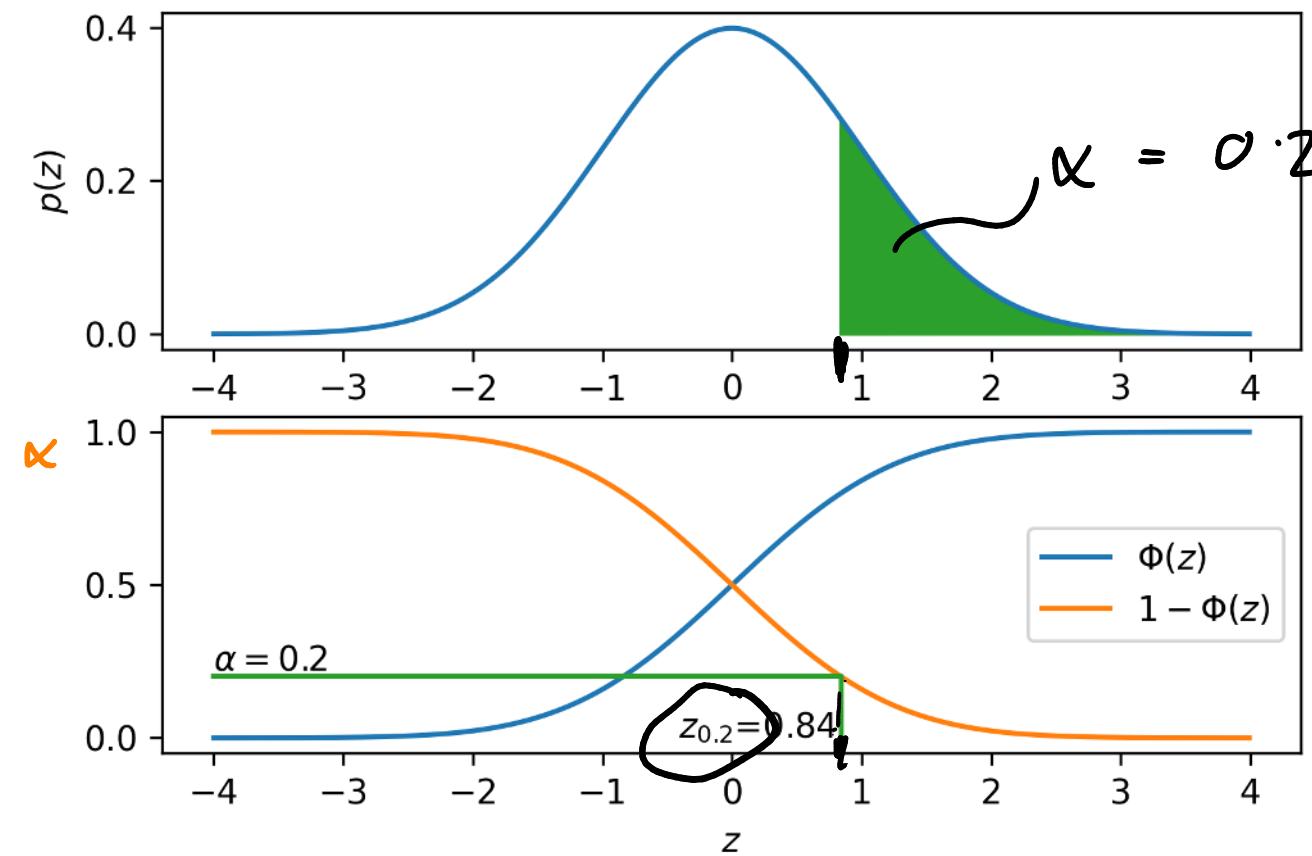
$\Rightarrow \frac{T_0 - \mu}{\sigma}$  has a  $z$ -distribution

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



area to  
right

## P-values computed by various methods for Swain versus Alabama

$t_0$	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

# Inf2 - Foundations of Data Science: Hypothesis testing - Testing for goodness-of-fit



THE UNIVERSITY *of* EDINBURGH  
**informatics**

FOUNDATIONS  
OF  
DATA  
SCIENCE

# 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
<u>(Observed-Expected)<sup>2</sup></u> 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  $i$ th group

$n_i$  - observed number in  $i$ th group

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

$np_i$  - 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.

$$\begin{array}{ll} np_i = 100 & n_i = 95 \\ np_i = 10 & n_i = 5 \end{array}$$

5

50% 50%

$$\chi^2 = 357.36$$

## 2. $H_0$ formulated as a statistical model

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

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

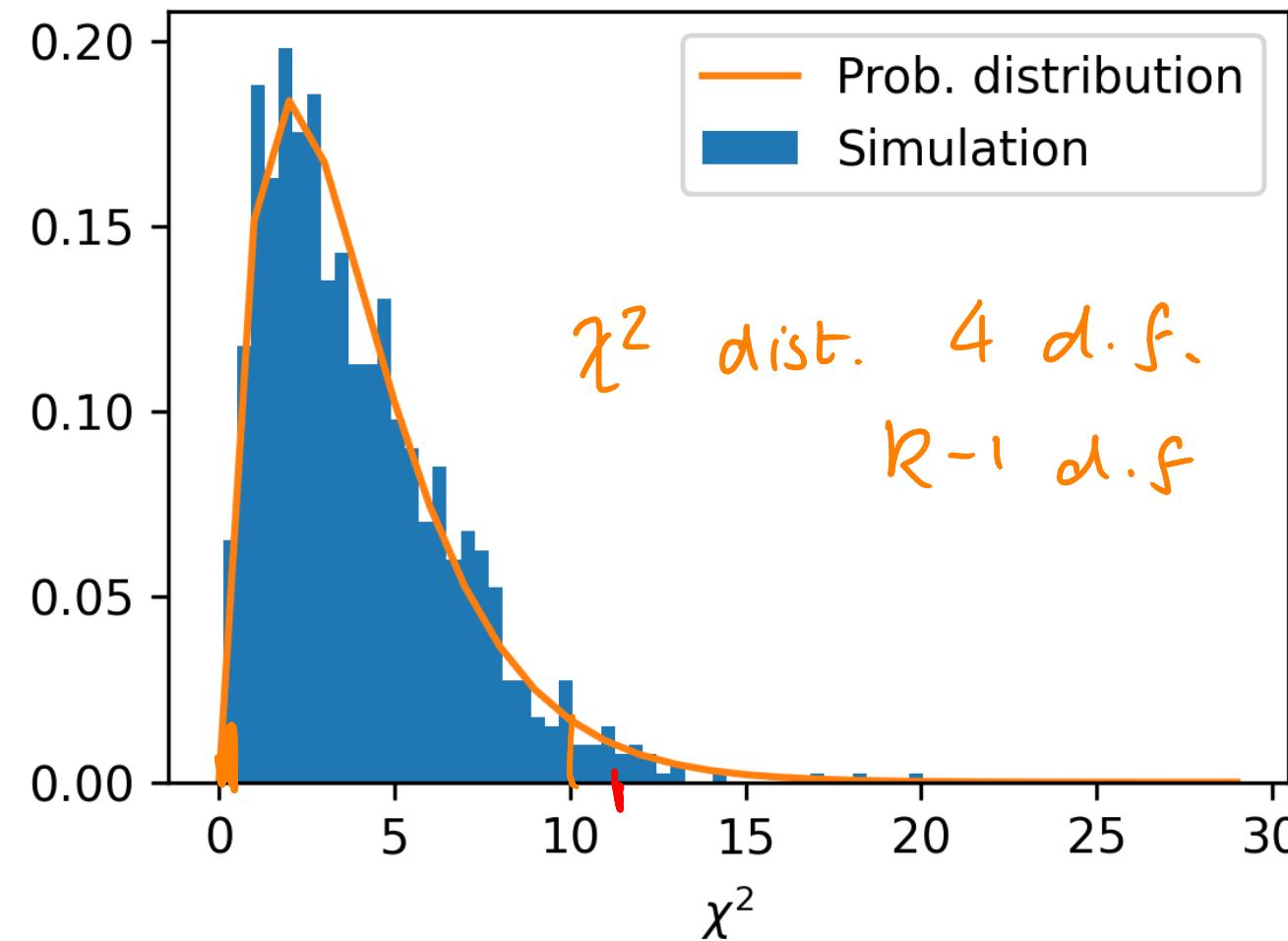
but constrained so that  $\sum_{i=1}^k n_i = n$

$\Rightarrow k-1$  degrees of freedom.

CODE

## 2. Distribution of test statistic under $H_0$

$n=1543$



$p \approx 0$

**Table A.10** Chi-Squared Curve Tail Areas

Upper-Tail Area	$\nu = 1$	$\nu = 2$	$\nu = 3$	$\nu = 4$	$\nu = 5$
> .100	< 2.70	< 4.60	< 6.25	< 7.77	< 9.23
.100	2.70	4.60	6.25	7.77	9.23
.095	2.78	4.70	6.36	7.90	9.37
.090	2.87	4.81	6.49	8.04	9.52
.085	2.96	4.93	6.62	8.18	9.67
.080	3.06	5.05	6.75	8.33	9.83
.075	3.17	5.18	6.90	8.49	10.00
.070	3.28	5.31	7.06	8.66	10.19
.065	3.40	5.46	7.22	8.84	10.38
.060	3.53	5.62	7.40	9.04	10.59
.055	3.68	5.80	7.60	9.25	10.82
.050	3.84	5.99	7.81	9.48	11.07
.045	4.01	6.20	8.04	9.74	11.34
.040	4.21	6.43	8.31	10.02	11.64
.035	4.44	6.70	8.60	10.34	11.98
.030	4.70	7.01	8.94	10.71	12.37
.025	5.02	7.37	9.34	11.14	12.83
.020	5.41	7.82	9.83	11.66	13.38
.015	5.91	8.39	10.46	12.33	14.09
.010	6.63	9.21	11.34	13.27	15.08
.005	7.87	10.59	12.83	14.86	16.74
.001	10.82	13.81	16.26	18.46	20.51
< .001	> 10.82	> 13.81	> 16.26	> 18.46	> 20.51
Upper-Tail Area	$\nu = 6$	$\nu = 7$	$\nu = 8$	$\nu = 9$	$\nu = 10$
> .100	< 10.64	< 12.01	< 13.36	< 14.68	< 15.98
.100	10.64	12.01	13.36	14.68	15.98

Upper-Tail Area	$\nu = 1$	$\nu = 2$	$\nu = 3$	$\nu = 4$	$\nu = 5$
> .100	< 10.64	< 12.01	< 13.36	< 14.68	< 15.98
.100	10.64	12.01	13.36	14.68	15.98
.095	11.07	12.49	13.88	15.22	16.52
.090	11.53	12.92	14.30	15.64	16.92
.085	12.01	13.37	14.73	16.06	17.35
.080	12.50	13.83	15.18	16.50	17.77
.075	12.99	14.27	15.61	16.91	18.17
.070	13.47	14.70	16.03	17.31	18.60
.065	13.94	15.12	16.44	17.72	19.00
.060	14.40	15.53	16.84	18.02	19.37
.055	14.85	15.93	17.23	18.31	19.72
.050	15.30	16.32	17.61	18.59	19.99
.045	15.74	16.70	17.98	18.86	20.25
.040	16.17	17.07	18.34	19.12	20.50
.035	16.60	17.43	18.69	19.37	20.74
.030	17.02	17.78	19.03	19.61	20.97
.025	17.43	18.12	19.36	19.84	21.19
.020	17.84	18.45	19.68	20.16	21.40
.015	18.24	18.77	20.00	20.47	21.60
.010	18.64	19.08	20.29	20.75	21.79
.005	19.03	19.38	20.57	21.01	22.17
.001	20.53	21.87	23.14	24.40	25.64
< .001	> 20.53	> 21.87	> 23.14	> 24.40	> 25.64

# Statistical tables

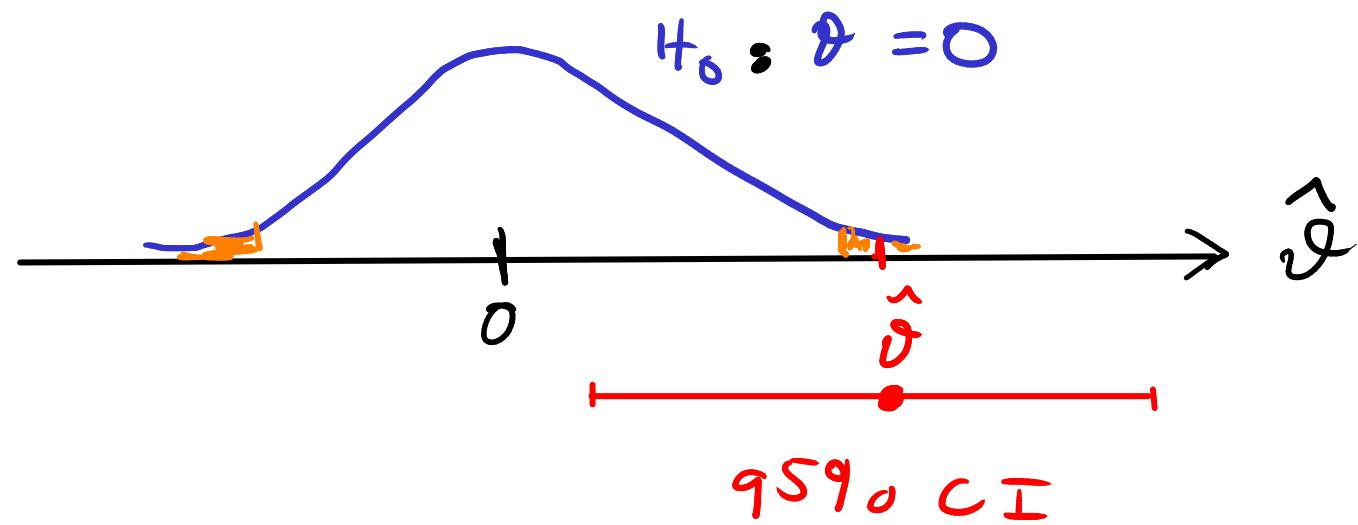
# Inf2 - Foundations of Data Science: Hypothesis testing - Confidence intervals and p-values



THE UNIVERSITY *of* EDINBURGH  
**informatics**

FOUNDATIONS  
OF  
DATA  
SCIENCE

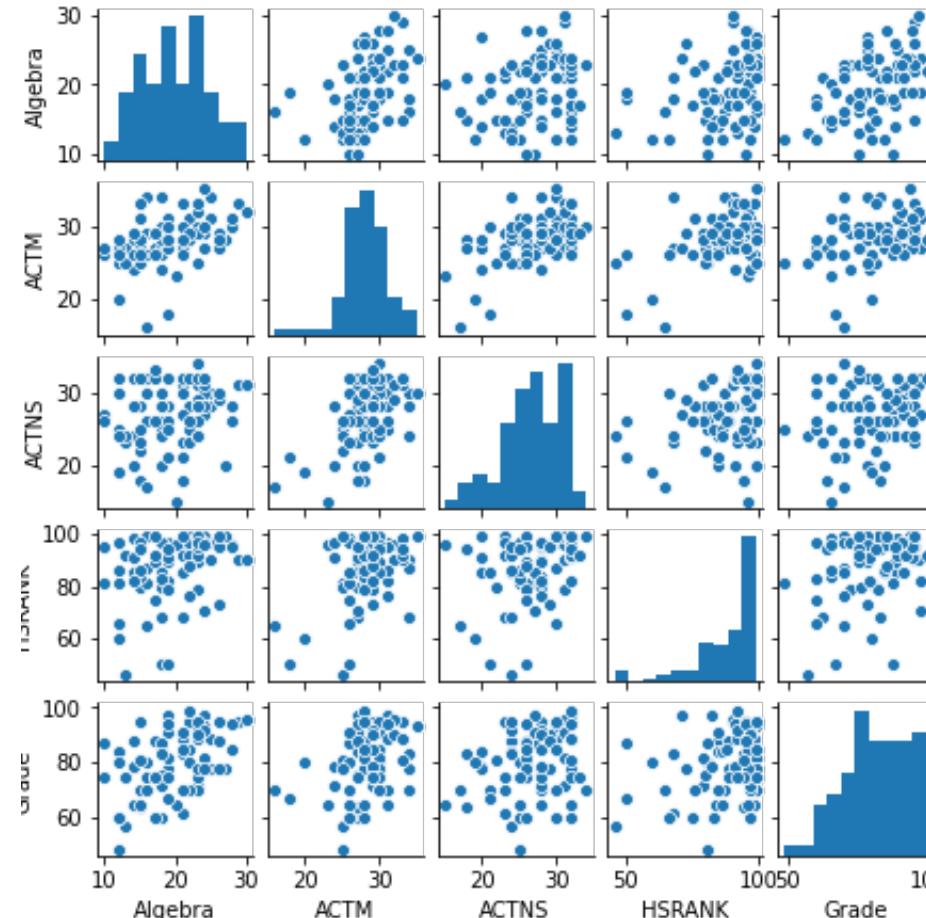
# Confidence intervals and p-values



Approx relation: if 95% CI for parameter doesn't contain 0, reject that

# p-values in Regression output

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)

# Summary

1. Principle of Hypothesis testing
  - (a) Rejection method
  - (b) p-values
2. Hypothesis testing applied to problems involving testing if observed numbers are consistent with expected proportions
  - Many other uses
3. Uses and limitations of hypothesis testing and p-values

## Working for multiple-testing example

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

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

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

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