

# DATA7002

## Statistical Significance

Slava Vaisman

Semester 2 2021



# Our objectives

- ▶ Understand the rigorous statistical framework for hypothesis testing.
- ▶ Understand the concept of statistical significance.
- ▶ Gain the ability to apply statistical tests in practice.
- ▶ Understand the limitation of statistical significance.

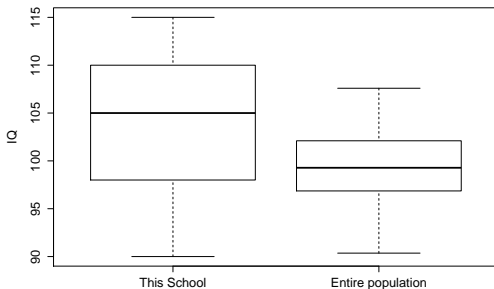
# Part I

## Mathematical framework

## Example (Choosing a school)

A certain **(and not very cheap)** private school claims that its students have a higher IQ. The entire student population is known to have an IQ that is Gaussian distributed with mean 100 and variance 16.

- ▶ Should we try to place our child in this school?
- ▶ Is the observed result *significant* **(can be trusted?)**, or due to a *chance*?



## Example (Medical treatment)

Consider an experimental medical treatment, in which 14 subjects were randomly assigned to control or treatment group. The survival times (in days) are shown in the table below.

	Data	Mean
Treatment group	91, 140, 16, 32, 101, 138, 24	77.428
Control group	3, 115, 8, 45, 102, 12, 18	43.285

- ▶ Did the treatment prolong the survival?
- ▶ Is the observed result *significant*, or due to a *chance*?

Making an error in this example, can have much more serious consequences when placing a child in an average school.

### Example (Tossing a coin)

I take a coin, toss it 10 times, and tell you the number of heads, (say 7).

- ▶ Is this a fair coin?
- ▶ Is the observed result *significant*, or due to a *chance*?

### Example (Testing an Improved Battery)

A manufacturer claims that its new improved batteries have a much longer lifetime. The old batteries are known to have a lifetime that is Normally distributed with mean 150 and variance 16. We measure the lifetime of nine batteries and obtain a sample mean of 155 hours.

- ▶ Is this new battery superior to the previous version?
- ▶ Is the observed result *significant*, or due to a *chance*?

# The framework

- ▶ Note that all the above examples are some-what similar.
- ▶ Specifically, we observed a system (school, or medical treatment, or coin toss, or electric battery),
- ▶ and asked ourself the following questions:
  1. Is the observed data is due to **chance**, or,
  2. due to **effect**?

For example,

1. Is the observed IQ in the school is due to “chance”, or
2. the observed IQ in the school is due to “effect”; in this case, one should definitely prefer this school!

# The framework

- ▶ To conclude, regardless the nature of our experiment, we always ask the same question:

## The question

Is the observed *data* is due to **chance**, or due to **effect**?

- ▶ This question brings us to a formulation of hypothesis. Specifically, given a data, our first task is to formulate two hypotheses.

## The research hypotheses

1. The *null hypothesis*  $H_0$ , which stands for our initial assumption about the data (due to **chance**).
2. The *alternative hypothesis*  $H_1$ , (sometimes called  $H_A$ ), (the data is due to **effect**).



# Setting the Hypothesis

Note that the *null* and the *alternative* hypotheses are two mutually exclusive statements!

## Example (Criminal Trial)

- ▶  $H_0$  : Defendant is **not guilty**.
- ▶  $H_1$  : Defendant is **guilty**.

## Example (Choosing a school)

1.  $H_0$  : The observed IQ in the school is due to “chance”.
2.  $H_1$  : The observed IQ in the school is due to “effect”. (One should definitely prefer this school!)

## Example (Medical treatment)

1.  $H_0$  : The observed data is due to “chance”, that is, the treatment does not prolong the survival.
2.  $H_1$  : The observed data is due to “effect”. (One should definitely consider this treatment!)

# Hypothesis testing

## Hypothesis testing

The general idea of hypothesis testing involves the following steps.

1. Collecting data.
2. Formulating the  $H_0$  and the  $H_1$  hypotheses.
3. Based on the data, decide whether to reject or not reject the initial hypothesis  $H_0$ .

- ▶ Sometimes, we alternate the first and the second steps.
- ▶ The first and the second steps look manageable.
- ▶ The third step looks like the most interesting (critical) one.

At this stage, suppose that we performed a test and made a decision regarding the **null** hypothesis.

## Making an error

Regardless of the procedure in the third step, we either

1. reject  $H_0$ , or
2. do not reject  $H_0$ .

This, can lead to an error, which is summarized in the table below.

True state	Decision	
	Retain $H_0$	Reject $H_0$
$H_0$ true	OK	Type I error (false positive)
$H_1$ true	Type II error (false negative)	OK

### Definition (Significance level of the statistical test)

The probability of a type I error is called the **significance level of the test** and is denoted by  $\alpha$ . (It is common to set the significance level to 0.05, that is, accepting to have a 5% probability of incorrectly rejecting the null hypothesis.)

## A short summary

1. We saw a few examples of experiments. Specifically, the school, the medical treatment, the coin toss, the new battery, and the criminal trial examples.
2. For each example, we formulated the null and the alternative hypotheses, namely:
  - 2.1  $H_0$  : the observed data is due to “chance”, and
  - 2.2  $H_1$  : the observed data is due to “effect”.
3. Since  $H_0$  and  $H_1$  are mutually exclusive, our decision to retain or to reject  $H_0$  can introduce an error. Specifically:
  - ▶ Type I error (false positive) — rejecting  $H_0$  when it is in fact true.
  - ▶ Type II error (false negative) — retaining  $H_0$  when it is in fact false.
4. A *significance level of the statistical test* is the probability of (making) the type I error. This probability is denoted by  $\alpha$ .

## Part II

Gaussian distribution,  
*critical* values, and  $p$ -values

## Normal, or Gaussian, Distribution

The normal (or Gaussian) distribution is the most important distribution in the study of statistics, engineering, and biology.

We say that a random variable has a normal distribution with parameters  $\mu$  and  $\sigma^2$  if its density function  $f$  is given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \quad x \in \mathbb{R}.$$

- ▶ We write  $X \sim N(\mu, \sigma^2)$ .
- ▶ The parameters  $\mu$  and  $\sigma^2$  turn out to be the expectation and variance of the distribution, respectively.
- ▶ If  $\mu = 0$  and  $\sigma = 1$  then

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}, \quad x \in \mathbb{R},$$

and the distribution is known as a **standard normal distribution**.

# Properties of Normal Distribution

- ▶ If  $X \sim N(\mu, \sigma^2)$ , then

$$\frac{X - \mu}{\sigma} \sim N(0, 1).$$

Thus by subtracting the mean and dividing by the standard deviation we obtain a standard normal distribution. This procedure is called **standardisation**.

- ▶ Standardisation enables us to express the cdf of any normal distribution in terms of the cdf of the standard normal distribution.
- ▶ A trivial rewriting of the standardisation formula gives the following important result: If  $X \sim N(\mu, \sigma^2)$ , then

$$X = \mu + \sigma Z, \quad Z \sim N(0, 1).$$

- ▶ In other words, any Gaussian (normal) random variable can be viewed as a so-called affine (linear + constant) transformation of a standard normal random variable.

# Central Limit Theorem

The Central Limit Theorem states roughly that:

**“The sum of a large number of iid random variables has approximately a Gaussian distribution.”**

More precisely, it states that, for all  $x$ ,

$$\mathbb{P}\left(\frac{S_n - n\mu}{\sqrt{n}\sigma} \leq x\right) = \Phi(x).$$

where  $\Phi$  is the cdf of the standard normal distribution.

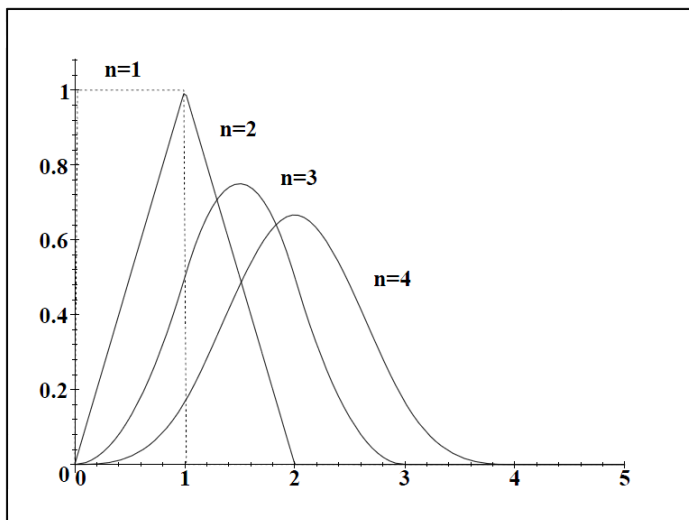
Regardless of  $X_i$ 's distribution, the sum behaves (approximately) as the Gaussian random variable!

Let us see the amazing CLT in action.



# Central Limit Theorem

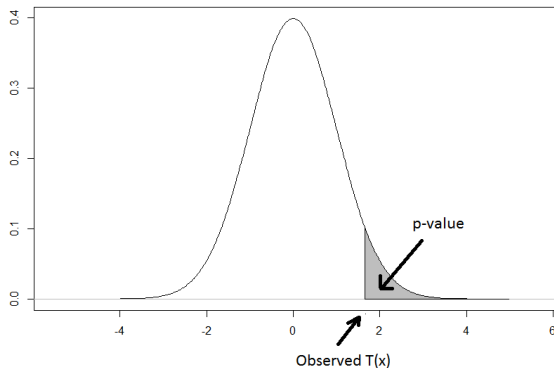
The next picture shows the pdf's of  $S_1, \dots, S_4$  for the case where the  $X_i$  have a  $U[0, 1]$  distribution.



# A central concept - the $p$ -value

## Definition ( $p$ -value)

The  $p$ -value is the probability that under the null hypothesis, the random test statistic takes a value as extreme as or more extreme than the one observed.



## A central concept - the $p$ -value

The general statistical test procedures using  $p$ -values is as follows.

1. Formulate a statistical model for the data.
2. Give the null and alternative hypotheses ( $H_0$  and  $H_1$ ).
3. Choose an appropriate test statistic.
4. Determine the distribution of the test statistic under  $H_0$ .
5. Evaluate the outcome of the test statistic.
6. Calculate the  $p$ -value.
7. Accept or reject  $H_0$  based on the  $p$ -value.

## A central concept - the $p$ -value

In the last step, if we reject  $H_0$  for  $p$ -value less than  $\alpha$ , we are back again to the statistical significance!

An easy to remember rule is:

$p$ -value low  $\Rightarrow H_0$  must go!

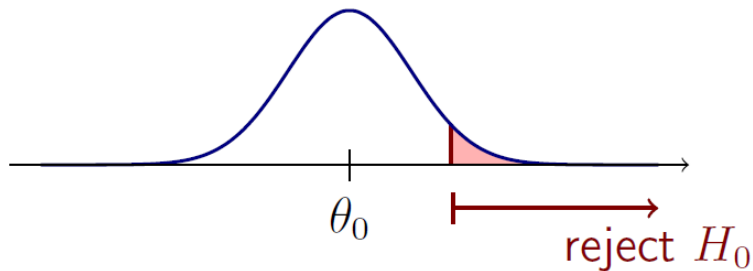
$p$ -value	evidence
$< 0.01$	very strong evidence against $H_0$
$0.01 - 0.05$	moderate evidence against $H_0$
$0.05 - 0.10$	suggestive evidence against $H_0$
$> 0.1$	little or no evidence against $H_0$

## Types of tests

- ▶ **Right one-sided test:** where  $H_0$  is rejected for the  $p$ -value defined by  $\mathbb{P}_{H_0}(T \geq t)$ .

- ▶  $H_0 : \theta = \theta_0$

- ▶  $H_1 : \theta > \theta_0$

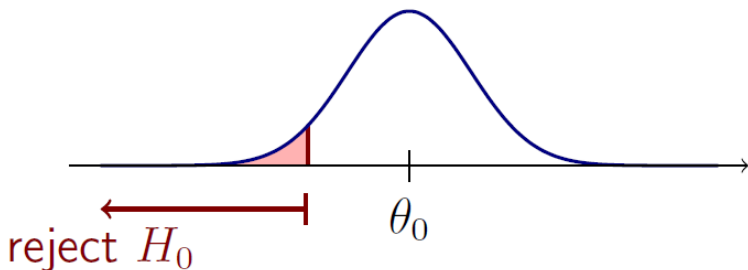


# Types of tests

- ▶ **Left one-sided test:** where  $H_0$  is rejected for the  $p$ -value defined by  $\mathbb{P}_{H_0}(T \leq t)$ .

- ▶  $H_0 : \theta = \theta_0$

- ▶  $H_1 : \theta < \theta_0$

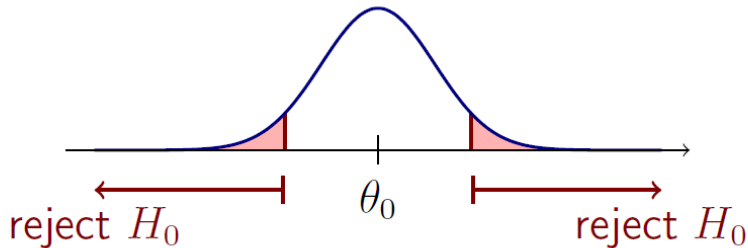


## Types of tests

- ▶ **Two-sided test** where  $H_0$  is rejected for the  $p$ -value defined by  $\mathbb{P}_{H_0}(T \geq |t|) + \mathbb{P}_{H_0}(T \leq -|t|) = 2\mathbb{P}_{H_0}(T \geq |t|)$ .

- ▶  $H_0 : \theta = \theta_0$

- ▶  $H_1 : \theta \neq \theta_0$



## A short summary

1. We saw a few examples of experiments.
2. For each example, we formulated the null and the alternative hypotheses, namely:
  - 2.1  $H_0$  : the observed data is due to “chance”, and
  - 2.2  $H_1$  : the observed data is due to “effect”.
3. We discussed
  - ▶ Type I error (false positive) — rejecting  $H_0$  when it is in fact true.
  - ▶ Type II error (false negative) — retaining  $H_0$  when it is in fact false.
4. We defined the probability of making Type I error as a significance level of a test.
5. We defined the test statistics, the critical value and the rejection region.
6. We introduced an equivalent approach for hypothesis testing using  $p$ -values.

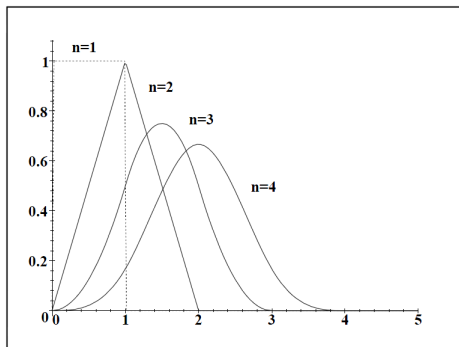


## Part III

# Basic statistical tests

# The Z-test

- ▶ A Z-test is any statistical test for which the distribution of the test statistic under the null hypothesis can be approximated by a normal distribution.
- ▶ Thanks to the central limit theorem, many test statistics are approximately **normally distributed** for large enough samples.



# The Z-test

- ▶ We will generally require the sample size to be at least 30.
- ▶ Let  $H_0 : \mu = \mu_0$ , and

$$H_1 : \begin{cases} \mu > \mu_0 & \text{right one sided test, or} \\ \mu < \mu_0 & \text{left one sided test, or} \\ \mu \neq \mu_0 & \text{two sided test} \end{cases}$$

- ▶ Let the test statistics be the average —  $\bar{X}$ .
- ▶ Recall that (CLT)

$$\left( \frac{\frac{1}{n} \sum_{i=1}^n X_i - \mu}{\sqrt{\frac{\sigma}{n}}} \right) = \sqrt{n} \left( \frac{\bar{X} - \mu}{\sigma} \right) \xrightarrow{d} N(0, 1).$$

# The Z-test

- ▶ That is,

$$\mathbb{P}_{H_0} \left( \underbrace{T(X)}_{\text{test statistics}} > \underbrace{\bar{X}}_{\text{observed}} \right) = \mathbb{P}_{H_0} \left( \underbrace{\frac{T(X) - \mu_0}{\sigma/\sqrt{n}}}_{Z \sim N(0,1)} > \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \right),$$

- ▶ or

$$\mathbb{P}_{H_0} (T(X) < \bar{X}) = \mathbb{P}_{H_0} \left( \frac{T(X) - \mu_0}{\sigma/\sqrt{n}} < \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \right),$$

- ▶ or

$$\begin{aligned} & \mathbb{P}_{H_0} (T(X) > |\bar{X}|) + \mathbb{P}_{H_0} (T(X) < -|\bar{X}|) \\ &= 2\mathbb{P}_{H_0} \left( \frac{T(X) - \mu_0}{\sigma/\sqrt{n}} > \left| \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \right| \right). \end{aligned}$$

# The Z-test

So we define the Z-score, to be:

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}},$$

With Z-score in hand, we can test our hypothesis.

```
# Zschool.py
# try to change n and mu_sc to explore sensitivity
import numpy as np
from scipy.stats import norm

mu_0 = 100
scsd = 4
mu_sc = 101
n = 40

IQschool = np.random.normal(mu_sc,scsd,n)
Zscore = (np.mean(IQschool) - mu_0)/(scsd/np.sqrt(len(IQschool)))
p_val = 1-norm.cdf(Zscore)
print(Zscore, p_val)
```

## Two sided $Z$ -test

- ▶ Suppose that  $X_1, \dots, X_n \sim N(\mu, \sigma^2)$ , ( $\sigma$  is known).
- ▶ We would like to test  $H_0 : \mu = \mu_0$ ,  $H_1 : \mu \neq \mu_0$ .
- ▶ Reject  $H_0$  at  $\alpha = 0.05$  if  $|Z| > 1.96$ .
- ▶ The  $p$ -value of the test is  $2\Phi(-|Z|)$ .

**Table:** A table with standard normal quantiles corresponding to common choices for  $\alpha$ .

$\alpha$	$Z_{1-\alpha/2}$	R
0.10	1.64	<code>qnorm(.95)</code>
0.05	1.96	<code>qnorm(.975)</code>
0.01	2.58	<code>qnorm(.995)</code>

## Z-test's assumptions

- ▶ *Nuisance parameters* should be known, or estimated with high accuracy (standard deviation).
- ▶ In particular, when the sample size  $n$  is large you may use

$$S = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2},$$

instead of  $\sigma$ .

- ▶ The test statistic should follow a normal distribution. If the variation of the test statistic is strongly non-normal, a Z-test should not be used.

# Two-Sample z-Test for the Difference Between Means

## Definition (two-sample hypothesis test)

For a two-sample hypothesis test,

- ▶  $H_0$  is a statistical hypothesis that usually states there is no difference between the parameters of two populations.
- ▶  $H_1$ , the alternative hypothesis, is a statistical hypothesis that must be true when  $H_0$  is false. That is, there exists a difference between two populations.

A two-sample z-test can be used to test the difference between two population means  $\mu_1$  and  $\mu_2$  when

1. a large sample (at least 30) is randomly selected from each population, and
2. the samples are independent.



## Two-Sample z-Test for the Difference Between Means

The test statistics is

$$\bar{x}_1 - \bar{x}_2,$$

and the z-score is

$$Z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sigma_{\bar{x}_1 - \bar{x}_2}},$$

where

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}.$$

## The $t$ -test

- ▶ The  $t$ -statistic was introduced in 1908 by William Sealy Gosset, a chemist working for the Guinness brewery in Dublin, Ireland.



- ▶ It can happen that we do not know the standard deviation, or
- ▶ the number of samples is less than 30.

## The $t$ -test

In this case, use the  $t$ -test. The  $t$  statistics with  $n - 1$  degrees of freedom is

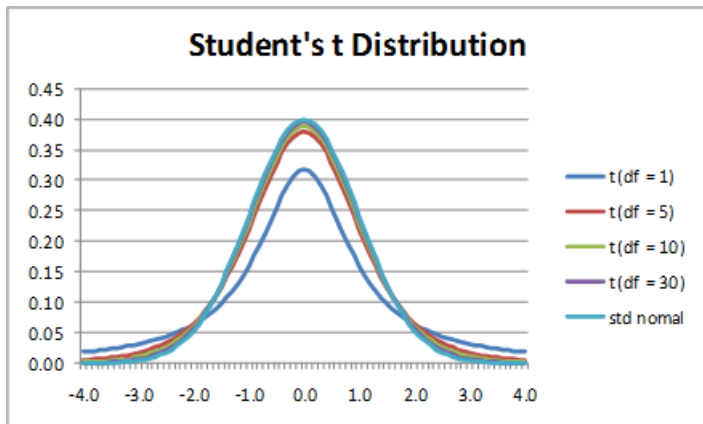
$$t_{n-1} = \frac{\bar{X} - \mu_0}{S/\sqrt{n}},$$

where  $S$  is the estimated standard deviation:

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2.$$

- ▶ Use the  $t$ -test when the data is approximately normally distributed.
- ▶ For large  $n$ ,  $t$ -test is indistinguishable from the  $z$ -test.

# The $t$ -distribution



# The $t$ -test example

```
# tTestSchool.py
import numpy as np
from scipy.stats import ttest_1samp
from scipy.stats import t

IQ = [110,105,97,104,98,112,115,108,90]
# t-test
onesample_results = ttest_1samp(IQ, 100)
print("sample mean=",np.mean(IQ), " t = ",
onesample_results[0], ", p-val = ", onesample_results[1])

# try t-test for another set
IQ2 = [105,105,104,102,99,108,105,108,103]
# t-test
onesample_results = ttest_1samp(IQ2, 100)
print("sample mean=",np.mean(IQ2), " t = ",
onesample_results[0], ", p-val = ", onesample_results[1])
```

Note that the sample mean in both cases is equal!

## The $t$ -test example 2

Table 2.1. *The mouse data. Sixteen mice were randomly assigned to a treatment group or a control group. Shown are their survival times, in days, following a test surgery. Did the treatment prolong survival?*

Group	Data			(Sample Size)	Mean	Estimated Standard Error
Treatment:	94	197	16	(7)	86.86	25.24
	38	99	141			
	23					
Control:	52	104	146	(9)	56.22	14.14
	10	51	30			
	40	27	46			
Difference:					30.63	28.93

## The *t*-test example 2

```
# twoSampleTtest.py
import numpy as np
from scipy.stats import ttest_ind

Treatment = np.array([94, 197, 16, 38, 99, 141, 23])
Control = np.array([52, 104, 146, 10, 51, 30, 40, 27, 46])

onesample_results = ttest_ind(Treatment, Control)
print("sample means=(,np.mean(Treatment),",",
      np.mean(Control) ,")",
      " t = ", onesample_results[0], ",
      p-val = ", onesample_results[1])
```

# Pearson's $\chi^2$ Test For Multinomial Data (background)

Pearson's  $\chi^2$  Test For Multinomial Data is one of the most interesting (and useful) tests, however we will need some background.

## Definition

We say that  $X$  has a Bernoulli distribution with success probability  $p$  if  $X$  can only assume the values 0 and 1, with probabilities

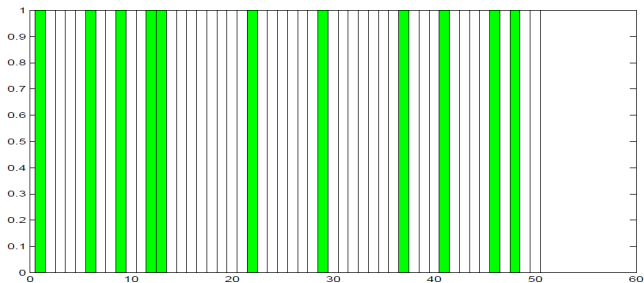
$$P(X = 1) = p = 1 - P(X = 0).$$

- ▶ We write  $X \sim \text{Ber}(p)$ .
- ▶ Despite its simplicity, this is one of the most important distributions in probability!
- ▶ It models for example:
  - ▶ a single coin toss experiment,
  - ▶ a success or a failure of message passing,
  - ▶ a success of a certain drug,
  - ▶ or, randomly selecting a person from a large population, and ask if she votes for a certain political party.



# Pearson's $\chi^2$ Test For Multinomial Data (background)

- ▶ Often, we have a sequence of independent Bernoulli trials.
- ▶ That is, we sequentially perform Bernoulli experiments, such that the outcome (success or failure) of each experiment does not depend on the other experiments.
- ▶ Here is a way to graphically show the outcomes (white – failure, green – success):



# Binomial Distribution

- ▶ Consider a sequence of  $n$  coin tosses.
- ▶ If  $X$  is the random variable which counts the total number of heads and the probability of “head” is  $p$  then we say  $X$  has a binomial distribution with parameters  $n$  and  $p$
- ▶ and write  $X \sim \text{Bin}(n, p)$ .
- ▶ The probability mass function  $X$  is given by

$$f(x, p) = \binom{n}{x} p^x (1 - p)^{n-x}, \quad x = 0, 1, 2, \dots, n.$$

- ▶ The expected value of  $X \sim \text{Bin}(n, p)$  is equal to  $np$ . This is sort of intuitive, since our success probability in a single trial is  $p$  and we perform  $n$  experiments overall.
- ▶ Moreover, a reasonable estimator for  $p$  is

$$\hat{p} = \bar{X}.$$

# Multinomial distribution

- ▶ The multinomial distribution is a generalization of the binomial distribution.
- ▶ It models the probability of counts for rolling a  $k$ -sided die  $n$  times.
- ▶ For  $n$  independent trials each of which leads to a success for exactly one of  $k$  categories, with each category having a given fixed success probability,  $p_1, \dots, p_k$ ; note that

$$p_1 + \dots + p_k = 1.$$

- ▶ The multinomial distribution gives the probability of any particular combination of numbers of successes for the various categories.
- ▶ Note that the expected number of times the outcome  $i$  was observed over  $n$  trials ( $X_i$ ) is  $\mathbb{E}[X_i] = n p_i$ , and therefore, a reasonable estimator for  $p_i$  is  $\hat{p}_i = \frac{X_i}{n}$ .

# Pearson's $\chi^2$ Test For Multinomial Data

- ▶ Suppose that

$$\mathbf{X} = (X_1, \dots, X_k),$$

has a multinomial distribution.

- ▶ We can estimate  $\hat{p}_1, \dots, \hat{p}_k$  via

$$(\hat{p}_1, \dots, \hat{p}_k) = \left( \frac{X_1}{n}, \dots, \frac{X_k}{n} \right).$$

- ▶ Let  $p_0 = (p_{01}, \dots, p_{0k})$  be some fixed vector and suppose we want to test

$$H_0 : p = p_0 \text{ versus } H_1 : p \neq p_0.$$

# Pearson's $\chi^2$ Test For Multinomial Data

Definition (Pearson's  $\chi^2$  statistic is)

$$T = \sum_{j=1}^k \frac{(X_j - np_{0j})^2}{np_{0j}} = \sum_{j=1}^k \frac{(X_j - E_j)^2}{E_j},$$

where  $E_j = \mathbb{E}[X_j] = np_{0j}$  is the expected value of  $X_j$  under  $H_0$ .

## Theorem

Under  $H_0$ ,  $T \sim \chi_{k-1}^2$  (approximately).

Hence, the p-value is

$$\mathbb{P}(\chi_{k-1}^2 > t),$$

where  $t$  is the observed value of the test statistic.

## The $\chi^2$ distribution

- ▶ Let  $Z_1, \dots, Z_k$  be independent, standard Normal random variables.
- ▶ Let

$$V = \sum_{i=1}^k Z_i^2.$$

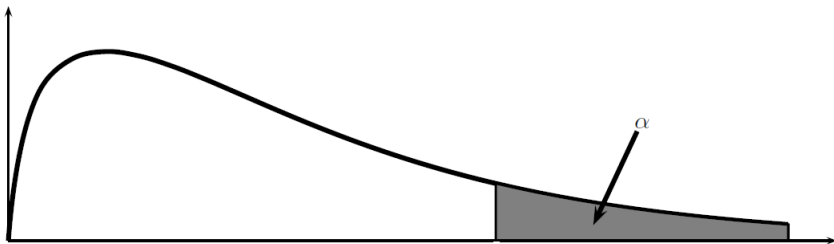
- ▶ Then we say that  $V$  has a  $\chi^2$  distribution with  $k$  degrees of freedom.
- ▶ We write

$$V \sim \chi_k^2.$$

## The $\chi^2$ distribution

We define the upper  $\alpha$  quantile  $\chi_{k,\alpha}^2 = F^{-1}(1 - \alpha)$  where  $F$  is the cumulative distribution function (cdf). That is

$$\mathbb{P}(\chi_k^2 > \chi_{k,\alpha}^2) = \alpha.$$



# Pearson's $\chi^2$ Test For Multinomial Data

## Example (Mendel's peas)

- ▶ Mendel bred peas with round yellow seeds and wrinkled green seeds.
- ▶ There are four types of progeny: round yellow, wrinkled yellow, round green, and wrinkled green.
- ▶ The number of each type is multinomial with probability  $p = (p_1, p_2, p_3, p_4)$ . His theory of inheritance predicts that  $p$  is equal to

$$p_0 = \left( \frac{9}{16}, \frac{3}{16}, \frac{3}{16}, \frac{1}{16} \right).$$

- ▶ In  $n = 556$  trials he observed  $X = (315, 101, 108, 32)$ .
- ▶ We will test  $H_0 : p = p_0$  versus  $H_1 : p \neq p_0$ .



# Pearson's $\chi^2$ Test For Multinomial Data (Mendel's peas )

- Note that

$$np_{01} = 556 \frac{9}{16} = 312.75, \quad np_{02} = np_{03} = 556 \frac{3}{16} 104.25, \quad \text{and} \\ np_{04} = 56 \frac{1}{16} = 34.75.$$

- Therefore, the test statistic is

$$T = \sum_{j=1}^4 \frac{(X_j - np_{0j})^2}{np_{0j}} = \frac{(315 - 312.75)^2}{312.75} + \frac{(101 - 104.25)^2}{104.25} \\ + \frac{(108 - 104.25)^2}{104.25} + \frac{(32 - 34.75)^2}{104.25} = 0.47.$$




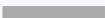





The  $p$ -value is  $\mathbb{P}(\chi_3^2 > 0.47) = 0.93$ . This is not evidence against  $H_0$ . Hence, the data do not contradict Mendel's theory.\*\*\*

# Benford's law

## Definition (Benford's law)

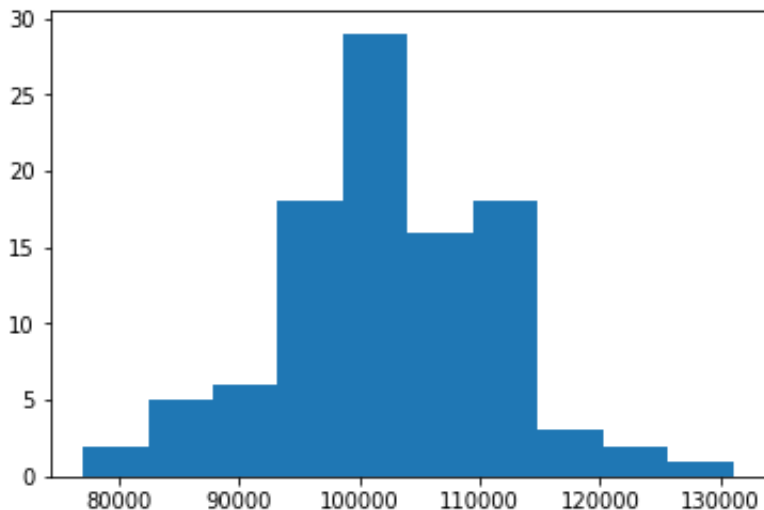
A set of numbers is said to satisfy Benford's law if the leading digit  $d$ , where  $(d \in \{1, \dots, 9\})$  occurs with probability

$$\mathbb{P}(D = d) = \log_{10}(d + 1) - \log_{10}(d) = \log_{10} \left( 1 + \frac{1}{d} \right).$$

$d$	$P(d)$	Relative size of $P(d)$
1	30.1%	
2	17.6%	
3	12.5%	
4	9.7%	
5	7.9%	
6	6.7%	
7	5.8%	
8	5.1%	
9	4.6%	

## Benford's law example

You are exploring a certain company salaries.



# Benford's law example

You are planning to invest, so this company sent you a comprehensive report.

```
# BenfordLaw.py
import numpy as np
import matplotlib.pyplot as plt
import math
from scipy.stats import chisquare

def first_n_digits(num, n):
    return num // 10 ** (int(math.log(num, 10)) - n + 1)

def BenfordTest(data):
    plt.hist(data, bins=10)
    plt.show()
    expected = np.zeros(9)
    observed = np.zeros(9)
    for i in range(9):
        expected[i] = np.log10(1 + (1/(i+1)))

    for num in data:
        digit = int(first_n_digits(num, 1))
        observed[digit - 1] = observed[digit - 1] + 1
    expected = expected * len(data)
    result = chisquare(observed, expected)
    print("statistic = ", result[0], "p-value = ", result[1])

np.random.seed(54321)
n = 100

data = np.random.normal(100000, 10000, n)
BenfordTest(data)
```

# Benford's law

- ▶ Benford's law can be expected to apply for:
  - ▶ Credit card transactions
  - ▶ Purchase orders
  - ▶ Loan data
  - ▶ Customer balances
  - ▶ Journal entries
  - ▶ Stock prices
  - ▶ Accounts payable transactions
  - ▶ Inventory prices
  - ▶ Customer refunds
- ▶ Examples of data sets that are not likely to be suitable for Benford's Law:
  - ▶ Airline passenger counts per plane
  - ▶ Telephone numbers
  - ▶ Data sets with 500 or fewer transactions
  - ▶ Data generated by formulas (e.g., YYMMXXXX as an insurance policy number)
  - ▶ Data restricted by a maximum or minimum number (e.g., hourly wage rate)

## A short summary

- ▶ We saw some basic and important statistical tests:
  - ▶  $Z$ -test,
  - ▶  $t$ -test, and
  - ▶  $\chi^2$ -test.
- ▶ We discussed Benford's law.
- ▶ All these tests assumed asymptotic, such as
  - ▶ Normality, and
  - ▶  $\chi^2$  distribution.

## Part IV

# Advanced hypothesis testing

# Statistical Test

In general, a statistical test involves the following steps.

1. Formulate a statistical model for the data.
2. Give the null and alternative hypotheses ( $H_0$  and  $H_1$ ).
3. Choose an appropriate test statistic.
4. Determine the distribution of the test statistic under  $H_0$ .
5. Evaluate the outcome of the test statistic.
6. Calculate the p-value.
7. Accept or reject  $H_0$  based on the p-value.

Do we always need asymptotic assumptions?



# Permutation Test

- ▶ The permutation test is a non-parametric method for testing whether two distributions are the same.
- ▶ This test is “exact”, meaning that it is not based on large sample theory approximations.
- ▶ Suppose that

$$X_1, \dots, X_m \sim F_X, \quad Y_1, \dots, Y_n \sim F_Y,$$

are two independent samples and  $H_0$  is the hypothesis that the two samples are identically distributed.

- ▶ Note that this is the type of hypothesis we would consider when testing whether a treatment differs from a placebo.
- ▶ Let  $T(x_1, \dots, x_m, y_1, \dots, y_n)$  be some test statistic, such as,

$$|\bar{X}_m - \bar{Y}_n|.$$

- ▶ What can we conclude if  $T$  is large? What will happen if it is close to 0?

# Permutation Test

- ▶ Let  $N = m + n$  and consider forming all  $N!$  permutations of the data

$$X_1, \dots, X_m, Y_1, \dots, Y_n.$$

- ▶ For each permutation, compute the test statistic  $T$ . Denote these values by  $T_1, \dots, T_{N!}$ .
- ▶ Under the null hypothesis, each of these values is equally likely. More precisely, under the null hypothesis, given the ordered data values,  $X_1, \dots, X_m, Y_1, \dots, Y_n$  is uniformly distributed over the  $N!$  permutations of the data.
- ▶ The distribution  $\mathbb{P}_0$  that puts mass  $1/N!$  on each  $T_j$  is called the permutation distribution of  $T$ .
- ▶ Let  $t_{obs}$  be the observed value of the test statistic. Assuming we reject when  $T$  is large, the p-value is

$$\text{p-value} = \mathbb{P}_0(T > t_{obs}) = \frac{1}{N!} \sum_{i=1}^{N!} I\{T_i > t_{obs}\}.$$

## Permutation Test Example

Let  $(X_1, X_2, Y_1) = (1, 9, 3)$ . So,  $|\bar{X} - \bar{Y}| = 5 - 3 = 2$ .

permutation	value of $T$	probability
(1,9,3)	2	1/6
(9,1,3)	2	1/6
(1,3,9)	7	1/6
(3,1,9)	7	1/6
(3,9,1)	5	1/6
(9,3,1)	5	1/6

The p-value is 4/6,

p-value	evidence
$< 0.01$	very strong evidence against $H_0$
$0.01 - 0.05$	strong evidence against $H_0$
$0.05 - 0.10$	weak evidence against $H_0$
$> 0.1$	little or no evidence against $H_0$
<b>That is, little or no evidence against <math>H_0</math>.</b>	

## An issue with the permutation test

- Recall that,

$$\text{p-value} = \mathbb{P}_0(T > t_{obs}) = \frac{1}{N!} \sum_{i=1}^{N!} I\{T_i > t_{obs}\},$$

where  $N = m+n$ .

- The problem is that we will need to perform  $N!$  operations.
- The solution is to use Monte Carlo!

# Monte Carlo

- ▶ Monte Carlo is about *generating random variables and calculating expectations*. Consider a random variable  $X \sim f(x)$  taking values in  $\mathcal{X}$ . A general objective of Monte Carlo simulation is to calculate

$$\ell = \mathbb{E}[H(X)],$$

where  $H : \mathcal{X} \rightarrow \mathbb{R}$  is a real-valued function.

- ▶ The Crude Monte Carlo (CMC) procedure for obtaining estimator  $\hat{\ell}$  of  $\ell$  is given by:

$$\hat{\ell} = \frac{1}{N} \sum_{i=1}^N H(X_i),$$

where  $X_i$  for  $i = 1, \dots, N$ , are independent copies of random variable generated from  $f(x)$ .

# Permutation Test via Monte Carlo

## Algorithm for Permutation Test

1. Compute the observed value of the test statistic

$$t_{\text{obs}} = T(X_1, \dots, X_m, Y_1, \dots, Y_n).$$

2. Randomly permute the data. Compute the statistic again using the permuted data.
3. Repeat the previous step  $B$  times and let  $T_1, \dots, T_B$  denote the resulting values.
4. The approximate p-value is

$$\frac{1}{B} \sum_{j=1}^B I(T_j > t_{\text{obs}}).$$

```

# mousePermTest.py
import numpy as np
Treatment = np.array([94, 197, 16, 38, 99, 141, 23])
Control = np.array([52, 104, 146, 10, 51, 30, 40, 27, 46])

t_obs = np.abs(np.mean(Treatment) - np.mean(Control))

B = 1000

combined = np.append(Treatment,Control)

ell = np.zeros(B)

for i in range(0,B):
    tmp = np.random.permutation(combined)
    t_tmp = tmp[0:Treatment.shape[0]]
    c_tmp = tmp[Treatment.shape[0]:combined.shape[0]]
    dif = np.abs(np.mean(t_tmp) - np.mean(c_tmp))
    if(dif>t_obs):
        ell[i]=1

ell_mean = np.mean(ell)
ell_std = np.std(ell)
print("p-val=",ell_mean, " p_value 95% CI = [",ell_mean -
    1.96*ell_std/np.sqrt(B), " , ", ell_mean + 1.96*ell_std/np.sqrt(B),"]")

```

Is the statistical significance  
really a silver bullet?  
(*P*-hacking)



## A small reminder about the today's lecture

- ▶ To draw conclusions from data, many scientists usually rely on significance testing.
- ▶ This means calculating the  $p$ -value, which is the probability of the observed result if there really is no effect.
- ▶ We saw that if the  $p$ -value is sufficiently small, the result is considered to be statistically significant.

Let us consider the following hypothetical example.

1. Alex works for a data-science company.
2. His first experiment does not work out very well, but he quickly refines the procedures and runs a second study.
3. This new procedure looks very promising, but unfortunately, he does not get the  $p$ -value of 0.05.
4. However, Alex really believes in this idea, so he gathers more data, drops some of the clear outliers,
5. performs a few more tweaks, and, identifies a slightly surprising but really interesting effect that achieves  $p < 0.05$ .

# What happened?

1. Alex collected additional data.
2. He dropped some data that seemed like outliers.
3. Alex dropped some of his measures and focused on the most promising observations.
4. He analysed the data in a different **fashion** and performed a few additional tweaks.

The major issue is that all his choices were made after observing the data.

A researcher can (**unconsciously**) select and **tweak** the data until he obtains the desired  $p$ -value. **Even when there is no effect.**

Statisticians say: “ if you torture the data enough, they will confess” .

To ensure a fair research, we should not be tempted to manipulate the data as above.

```

# mousePermTestPHack.py
import numpy as np

def PermTest(Treatment, Control):
    t_obs = np.abs(np.mean(Treatment) - np.mean(Control))

    B = 10000

    combined = np.append(Treatment,Control)

    ell = np.zeros(B)

    for i in range(0,B):
        tmp = np.random.permutation(combined)
        t_tmp = tmp[0:Treatment.shape[0]]
        c_tmp = tmp[Treatment.shape[0]:combined.shape[0]]
        dif = np.abs(np.mean(t_tmp) - np.mean(c_tmp))
        if(dif>t_obs):
            ell[i]=1

    ell_mean = np.mean(ell)
    ell_std = np.std(ell)
    print("p-val=",ell_mean, " p_value 95% CI = [",ell_mean - 1.96*ell_std/np.sqrt(B),
          " , ", ell_mean + 1.96*ell_std/np.sqrt(B),"]")

Treatment = np.array([94, 197, 16, 38, 99, 141, 23])
Control = np.array([52, 104, 146, 10, 51, 30, 40, 27, 46])
t_obs = np.abs(np.mean(Treatment) - np.mean(Control))
PermTest(Treatment, Control)

# The management feels that the variance in the Treatment group
# is too big. Using small adjustments, we repeat the experiment to obtain
Treatment2 = np.array([73, 69, 115, 110, 90, 75, 80, 100, 77,83, 68,97])
t_obs2 = np.abs(np.mean(Treatment2) - np.mean(Control))
print("The new differences in the observed statistic is ",t_obs2-t_obs)
PermTest(Treatment2, Control)

```

## Green-colored coins are biased!

1. We make a hypothesis that green-colored coins are loaded, and that they will yield more heads than tails.
2. We take 100 green-colored coins, toss them, and count how many times a head appears.
3. In our experiment, heads appeared 58 times out of 100.

The p-value is:

$$\mathbb{P}(X \geq 58) \approx 0.044,$$

where  $X \sim \text{Bin}(100, 0.5)$ .

## Some minor thing I forgot to tell you...

1. I took 100 coins, each colored with a different color; (20 colors overall).
2. I got the following results while tossing these coins: (see `coins1.py`).

```

#coins1.py
import numpy as np
from numpy.random import randint
from scipy import stats
import pandas as pd

np.random.seed(12345)

color = ['Purple', 'Brown', 'Pink', 'Blue', 'Teal',
         'Salmon', 'Red', 'Turquoise', 'Magenta', 'Yellow',
         'Tan', 'Green', 'Grey', 'Cyan', 'Mauve',
         'Beige', 'Lilac', 'Black', 'Peach', 'Orange']

# number of experiments
n = 100

df = pd.DataFrame(index=color)
for col in color:
    result = randint(0,1+1,n)
    df.loc[col,'Heads'] = np.sum(result)

tmp = df[df["Heads"]==max(df["Heads"])]

num_heads = tmp.values[0][0]

# 1-cdf
p_value = stats.binom(100, 1/2).sf(num_heads)

print("p-value = ",p_value)

```

## My experiment is as follows

1. I started with 100 coins each color among 20 colors.
2. I rolled each coin 1 time.
3. I collected the number of times a Head appeared for each color.
4. I found that green had the highest number of heads.
5. I computed the p-value of the green coin result.
6. I then claimed that the p-value for this experiment is about 4%.

Is this OK?

No - the last step is wrong.

## The 4% p-value corresponds to a different experiment

1. I started with 100 coins each color among 20 colors.
2. I decide to watch a green coin in particular.
3. I rolled each coin 1 time.
4. I collected the number of times a Head appeared for each color.
5. I found that green had the highest number of heads.
6. I computed the p-value of the green coin result.
7. I then claimed that the p-value for this experiment is about 4%.

The main difference between the experiments is as follows. In the first experiment, we selected the coin color after the experiment was run. In the second experiment, we selected the coin color before running the experiment..



## Are my results in the first experiment are reasonable?

- ▶ Of course, all we need to do is to calculate the probability that at least one type (color) of coins give 58 or more heads in 100 runs!
- ▶ Can you propose an easy method to do so?

```
#coins2.py
import numpy as np
from numpy.random import randint
from scipy import stats
import pandas as pd

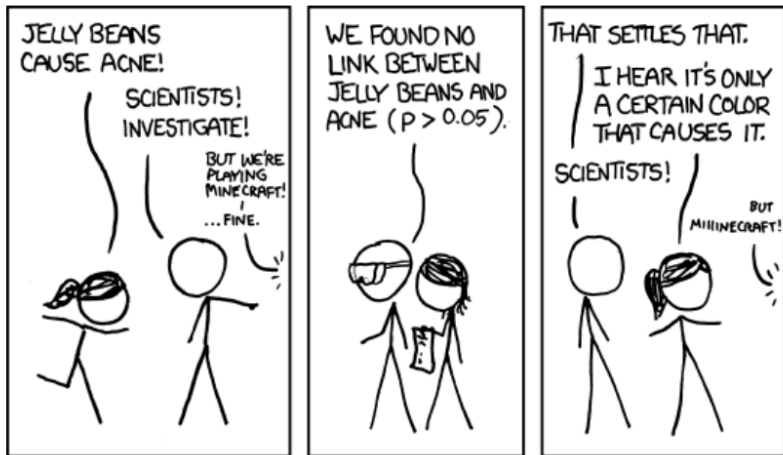
np.random.seed(12345)

N = 100000
ell = np.zeros(N)

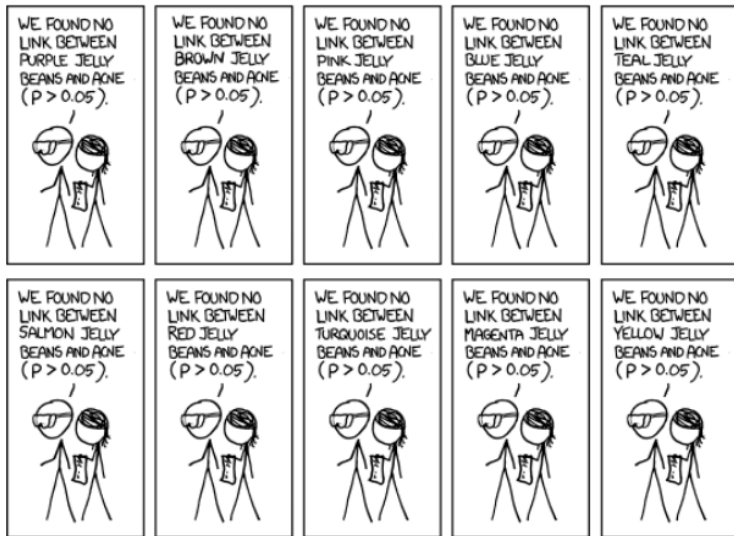
for i in range(N):
    exp = np.random.binomial(100,0.5,20)
    if(max(exp)>=58):
        ell[i]=1

print(np.mean(ell))
```

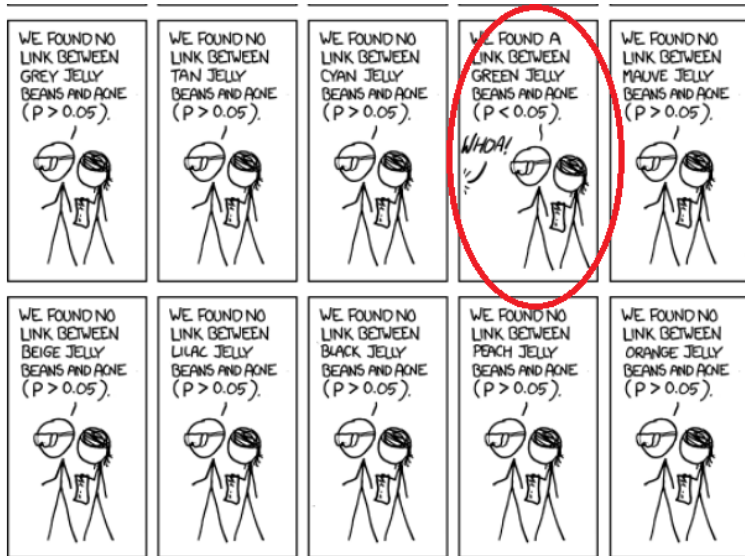
So, what we did?



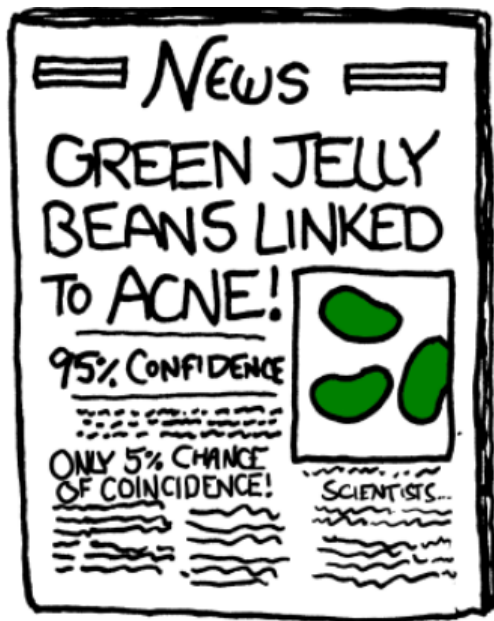
## So, what we did?



So, what we did?



So, what we did?



## How (not) to p-hack?

- ▶ Stop the data collection once  $p - \text{value} \leq 0.05$ .
- ▶ Analyze many measures with a view to report only these with  $p - \text{value} \leq 0.05$ .
- ▶ Analyze many conditions (different hypothesis) with a view to report only these with  $p - \text{value} \leq 0.05$ .
- ▶ Exclude participants such that  $p - \text{value} \leq 0.05$ .
- ▶ Apply an appropriate data transformation, such that  $p - \text{value} \leq 0.05$ .
- ▶ You can probably think about additional methods...

# Case study 1

Students data from  
[https://stluc.manta.  
uqcloud.net/  
mdatascience/public/  
datasets/StudentData/  
student\\_data.csv](https://stluc.manta.uqcloud.net/mdatascience/public/datasets/StudentData/student_data.csv)



## Student case study - studentcasestudy.py

1. Load the student data-set.
2. Are there any privacy issues with the data-set?
3. We will consider the following theory — gender affects the GPA.
4. Create a new data-set with contains gender and GPA only.
5. Check types and make sure that gender is set to be categorical.
6. Generate a descriptive statistics of GPA (for each gender) and plot the GPA as a function of gender.
7. Apply statistical tests to determine the significance of the above theory (gender affects the GPA). Specifically, apply the 2-sample t-test and the permutation test.
8. Summarize your conclusions.

# Case study 2

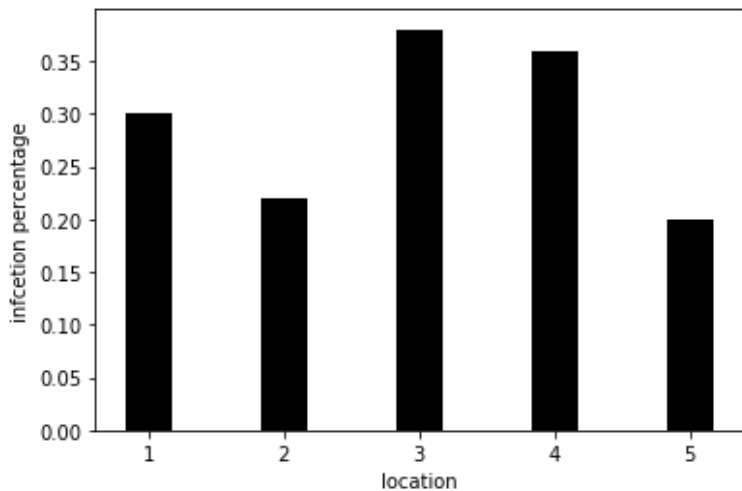
## Medicine distribution

## Medicine distribution - MedicalCaseStudy.py

- ▶ You are taking a part in a certain humanitarian mission and you are in charge of a life-saving drug distribution among 5 geographical locations (sites).
- ▶ Unfortunately, the drug inventory is limited.
- ▶ All local site manager indicated that the corresponding medical problem is severe and therefore they request the maximum possible amount of drug delivered.

# Medicine distribution

- ▶ What will be the most ethical decision?
- ▶ A reasonable (ethical?) approach is to distribute the drugs evenly between the 5 sites.
- ▶ Can we do better?
- ▶ It will be reasonable to conduct a statistical research so you request a sample of 50 people from each site and you receive the following graph. Will your decision change based on the below figure?



## Medicine distribution

You decision will have serious consequences so you decide to perform a rigorous statistical analysis using the Friedman chis-square test.

```
stat, p = friedmanchisquare(data1, data2,data3,data4,
                             data5)
print('Statistics=%.3f, p=%.3f' % (stat, p))
# interpret
alpha = 0.05
if p > alpha:
    print('Same distributions (fail to reject H0)')
else:
    print('Different distributions (reject H0)')

>>
Statistics=6.392, p=0.172
Same distributions (fail to reject H0)
```

## Medicine distribution

- ▶ You decide to increase the sample size to 100, 200,300,400 ,500, and 600.
- ▶ Only for the 600 sample size, we will manage to reject the  $H_0$  hypothesis.
- ▶ For the 600 sample size, we get the following percentage estimators:

$$p = (p_1, p_2, p_3, p_4, p_5) = (0.258, 0.223, 0.262, 0.312, 0.307).$$

- ▶ This is close to the true means:

$$p_{true} = (0.25, 0.23, 0.27, 0.31, 0.28).$$

- ▶ The reasonable percentage distribution percentage of drugs for location I is therefore:

$$p_i / (p_1 + p_2 + p_3 + p_4 + p_5).$$

*The End*