INTRODUCTION

To Data Science

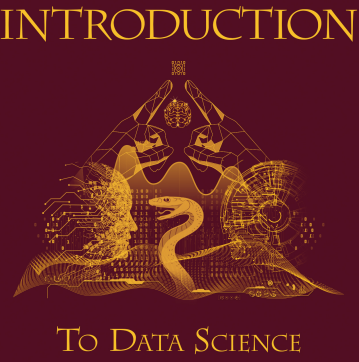
Introduction to Data Science

Inference - Part A - Class 7

Giora Simchoni

gsimchoni@gmail.com and add #intro2ds in subject

Stat. and OR Department, TAU

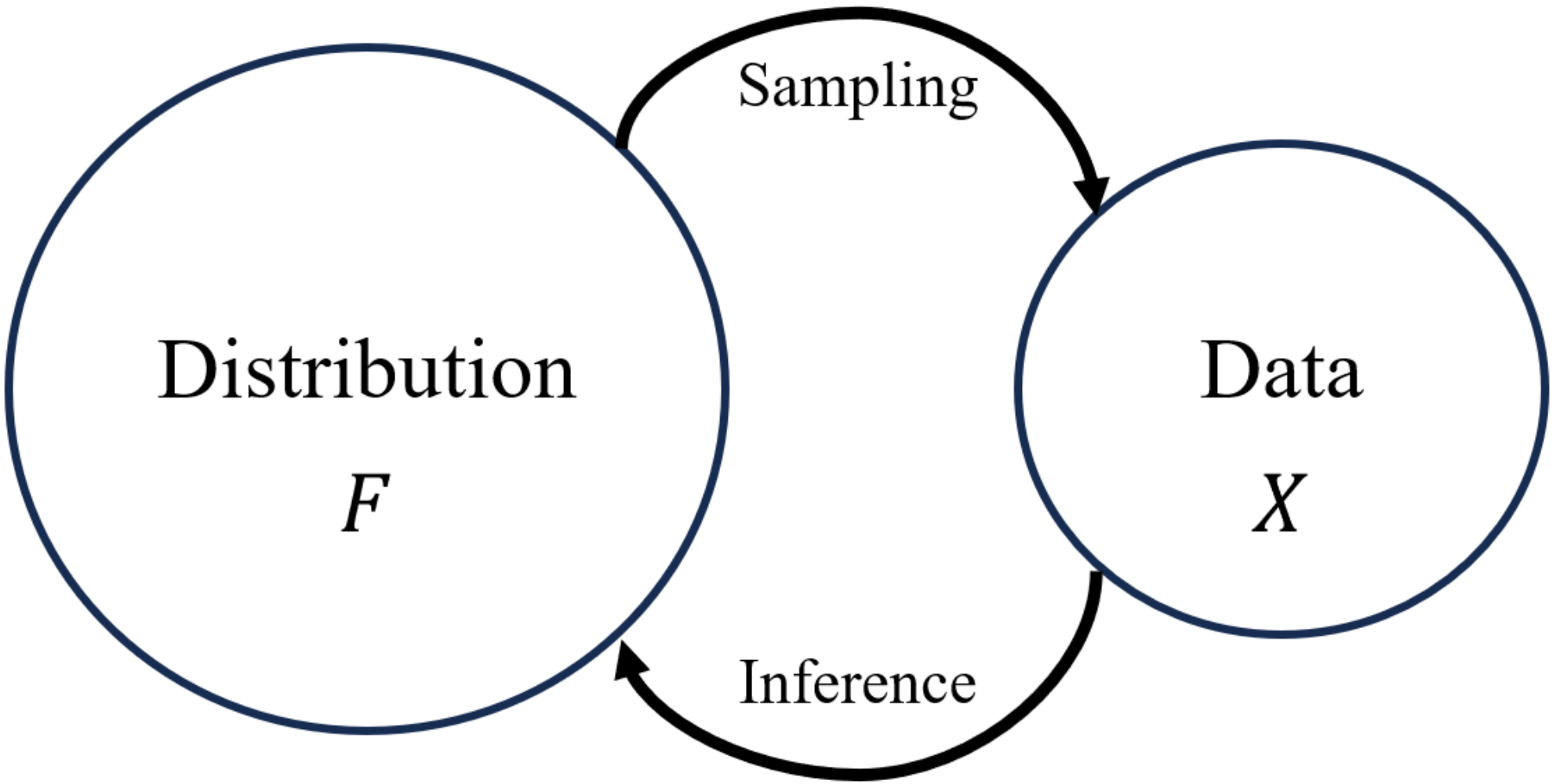


The Big Picture

INTRODUCTION

TO DATA SCIENCE

Basic idea of statistical inference



Basic idea of statistical inference

The *distribution* is something we want to learn about:

- Which candidate has more support in the population?
- Do impressionist paintings have more red than realist paintings?

We are given a *sample* of data X and want to use it to learn about the population:

- An election survey
- 30 impressionist paintings and 30 realist paintings

Hypothesis testing

We have a *null* world we believe in unless convinced otherwise:

- The candidates have equal support
- There is no difference in red level between impressionist and realist paintings

We want to use the sample to determine whether to reject the null:

- Does the sample **convincingly indicate** that candidate 1 has higher support?
- Does the sample contain **clear evidence** of more red in impressionist paintings?

This is often indicated through the p-value, which *calculates* how consistent our data is with the null hypothesis

Another view: the p-value measures how *surprising* the data we see is, if the null holds

Conceptual example: Criminal trial

- In a criminal trial, suspects get convicted only if their guilt is proven *beyond reasonable doubt*
- This is a hypothesis test with null hypothesis: the suspect is innocent
- Data: the evidence the sides bring in trial
- Beyond reasonable doubt: the evidence is not consistent with the null of innocence
- Difference: the decision is based on the judge's intuition, whereas formal hypothesis testing is based on calculating probabilities

Key element: The two hypotheses are not symmetric!

The importance of (formal) hypothesis testing in the world

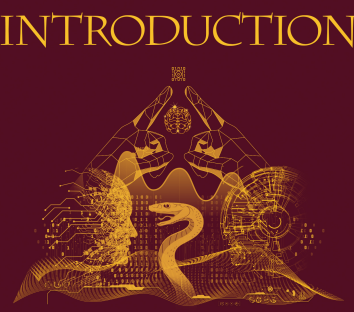
All scientific discovery is done through the hypothesis testing formalism:

- Null hypothesis: We did not discover something new (like a new particle, or a new genetic influence on disease)
- Examples: Higgs boson search, studies for finding genes that cause disease
- P value: strength of evidence that what we found is indeed new and different

All the formal processes of testing medications, food etc.

- Null hypothesis: the new medicine does not reduce cholesterol
- Example: study with people who got the medicine or placebo
- P value: how convincing is the evidence that the medicine is more effective than placebo?

Example: Red Paintings



Red Paintings Example

- We want to examine whether there is more red in impressionist paintings vs realist paintings
- The “world” is the 16K paintings we have: 8K realist, 8K impressionist
- Imagine we can’t check all of them, but can only sample a few of each kind and see the difference
- Our challenge: to determine if it is *convincing* evidence that impressionist paintings are redder overall
- Use hypothesis testing approach

```
real_sample = get_images_matrix(folder + 'realism_train.csv', folder + 'realism', n = 3
impr_sample = get_images_matrix(folder + 'impressionism_train.csv', folder + 'impressio

real_red = real_sample[:, :, :, 0].mean(axis = (1, 2))
impr_red = impr_sample[:, :, :, 0].mean(axis = (1, 2))

print(real_red[:10])
print(impr_red[:10])
```

```
[161.3162 147.0798 140.2261 122.5448 191.0334  52.9117  96.4099 110.3566
 171.9048  78.3864]
[ 51.7256  99.2127  95.8073 105.3173  81.8901 125.3137 128.208   78.7658
  81.1359 201.9115]
```

```
print(f'Realist paintings mean red value: {real_red.mean():.2f}')
print(f'Impressionist paintings mean red value: {impr_red.mean():.2f}')
print(f'Means difference: {impr_red.mean() - real_red.mean(): .2f}')
```

```
Realist paintings mean red value: 115.47
Impressionist paintings mean red value: 130.75
Means difference:  15.28
```

It looks as you expected, impressionist paintings average red pixel is higher by about 15 points, but if you do it again, results would be different, wouldn't they?

```
real_sample2 = get_images_matrix(folder + 'realism_train.csv', folder + 'realism', n =
impr_sample2 = get_images_matrix(folder + 'impressionism_train.csv', folder + 'impressi

real_red2 = real_sample2[:, :, :, 0].mean(axis = (1, 2))
impr_red2 = impr_sample2[:, :, :, 0].mean(axis = (1, 2))

print(f'Realist paintings mean red value: {real_red2.mean():.2f}')
print(f'Impressionist paintings mean red value: {impr_red2.mean():.2f}')
print(f'Means difference: {impr_red2.mean() - real_red2.mean(): .2f}')
```

```
Realist paintings mean red value: 125.98
Impressionist paintings mean red value: 134.96
Means difference:  8.98
```

Assume sampling is expensive. You have the capacity for 60 paintings.

How will you know, that the difference you're seeing is of significance? That it will "stick"? That what everyone is thinking, the null hypothesis, should be rejected, and your alternative hypothesis is more likely?

The Null Distribution by Simulation

INTRODUCTION

TO DATA SCIENCE

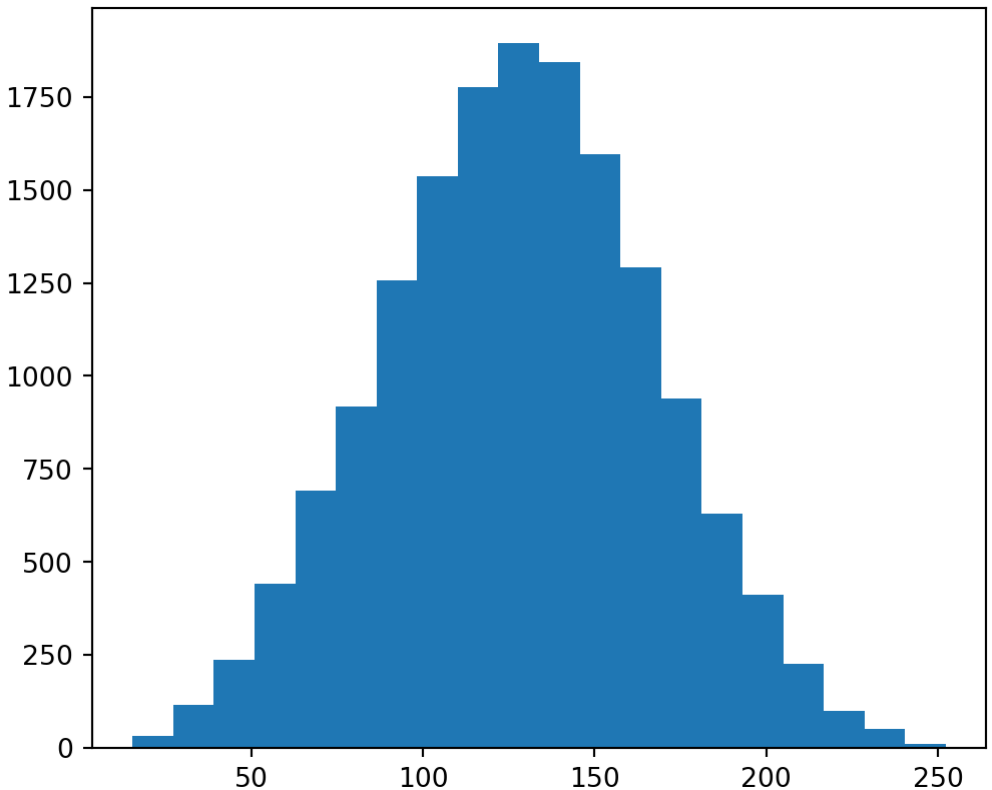
The Null Distribution by Simulation

- Under the null hypothesis, impressionist and realist paintings come from the same homogenous population.
- To illustrate this we will create an **artificial null** world, made of 16K paintings images in our training dataset. Then we can randomly assign half as impressionist and half as realist
- In this world we **know** that impressionist paintings and realist paintings have about the same amount of red

```
real_all = get_images_matrix(folder + 'realism_train.csv', folder + 'realism')
impr_all = get_images_matrix(folder + 'impressionism_train.csv', folder + 'impressionis

real_red_all = real_all[:, :, :, 0].mean(axis = (1, 2))
impr_red_all = impr_all[:, :, :, 0].mean(axis = (1, 2))

population = np.concatenate([real_red_all, impr_red_all])

print(population.shape)
```

```
(16000,)
```

```
plt.hist(population, bins=20)
plt.show()
```



- We can sample two random samples of so-called “impressionist” and so-called “realist” paintings to prove to ourselves that the difference between their means should be about zero:

```
real_red_null = np.random.choice(population, 30, replace=False)
impr_red_null = np.random.choice(population, 30, replace=False)
print(f'Means difference: {impr_red_null.mean() - real_red_null.mean(): .2f}')
```

```
Means difference:  4.54
```

- We got a mean difference which is different than zero, *by random*. And again and again:

```
real_red_null = np.random.choice(population, 30, replace=False)
impr_red_null = np.random.choice(population, 30, replace=False)
print(f'Means difference: {impr_red_null.mean() - real_red_null.mean(): .2f}')

real_red_null = np.random.choice(population, 30, replace=False)
impr_red_null = np.random.choice(population, 30, replace=False)
print(f'Means difference: {impr_red_null.mean() - real_red_null.mean(): .2f}')
```

```
Means difference: -18.38
Means difference: -13.45
```

The Null Distribution

- We want to know how is *our* original average difference of about 15 points is in comparison to these **null** average differences between groups coming from the same population.
- So we'll make a lot of the them and look at their distribution, the null distribution of the means difference:

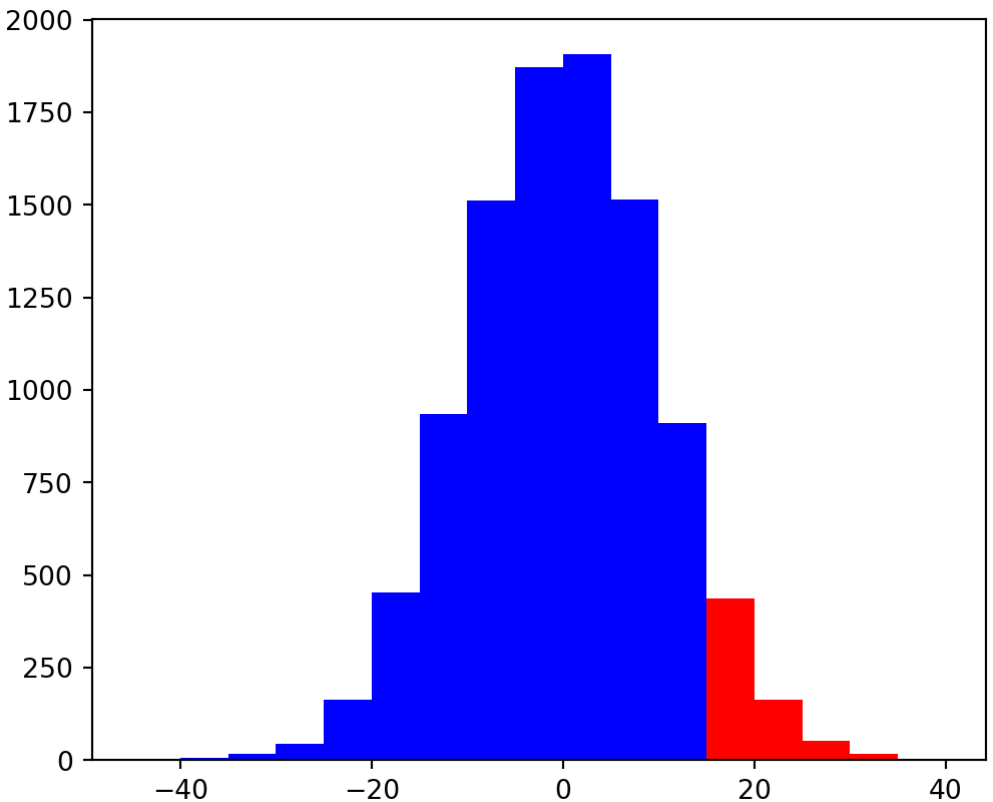
```
1  def sample_null_mean_diff(n = 30):
2      real_red_null = np.random.choice(population, n, replace=False)
3      impr_red_null = np.random.choice(population, n, replace=False)
4      return impr_red_null.mean() - real_red_null.mean()
5
6  null_mean_diffs = np.array([sample_null_mean_diff() for i in range(10000)])
7
8  print(f'Max null mean diff: {max(null_mean_diffs): .2f}')
9  print(f'Min null mean diff: {min(null_mean_diffs): .2f}')
```

```
Max null mean diff:  36.21
Min null mean diff: -37.46
```

- We can see that the max null mean differences is actually much higher than our original 15 points. So at random, when there is *no difference*, you can get mean differences of over 30!

Let’s look at the null distribution histogram:

▶ Code



It seems like our original value of 15 points difference is not that extreme. There’s a measure for that:

P-Value

- How extreme is our original 15 points result?

- What is the probability under the null distribution, where there is no difference between “realist” and “impressionist”, of getting 15 or higher?

```
one_sided_p_value = np.mean(null_mean_diffs >= 15)

print(f'P(mean_diff >= 15 | H0) = {one_sided_p_value: .2f}')
```

```
P(mean_diff >= 15 | H0) =  0.07
```

- It looks like the chance of getting a difference of 15 points or higher, when there is no difference, is ~6-7%. Does that convince you that there actually is a difference, that indeed the realist and impressionist samples came from two different, separate, distributions?

- It is a standard in both academia and industry to not be persuaded by a p-value larger than a threshold α of 1% or 5% (a.k.a Type I Error, see soon).

Two-sided Hypothesis

- If the original alternative hypothesis were “impressionist paintings images’ red level is *different* than realist”, the p-value should have been two sided.
- Because the probability of getting our original value or “more extreme” would have meant “more extreme in absolute value”:

```
two_sided_p_value = np.mean(np.abs(null_mean_diffs) > 15)

print(f'P(|mean_diff| >= 15 | H0) = {two_sided_p_value: .2f}')
```

```
P(|mean_diff| >= 15 | H0) =  0.13
```

- 13-14% chance of observing a result like 15 points or more extreme, at random, when there is no difference. 15 points doesn’t look convincing.
- But in real life we only have that one hard-earned sample. We don’t have the population. And from here, the rest is mathematical approximation for getting that p-value and other measures, with what we have.

Binomial Example

INTRODUCTION

TO DATA SCIENCE

Simpler conceptual example

- I am given a coin and I want to know if it is fair (heads/tails equally likely)
- I am only allowed to throw it 10 times, and I get 8 heads
- Mark the null hypothesis $H_0 : P(\text{head}) = \frac{1}{2}$
- Can I reject H_0?

- If I had a null distribution, I could get a p-value: see what % of the time I would get 8 or more heads if the coin was fair
- Can I get this null distribution? Easily!

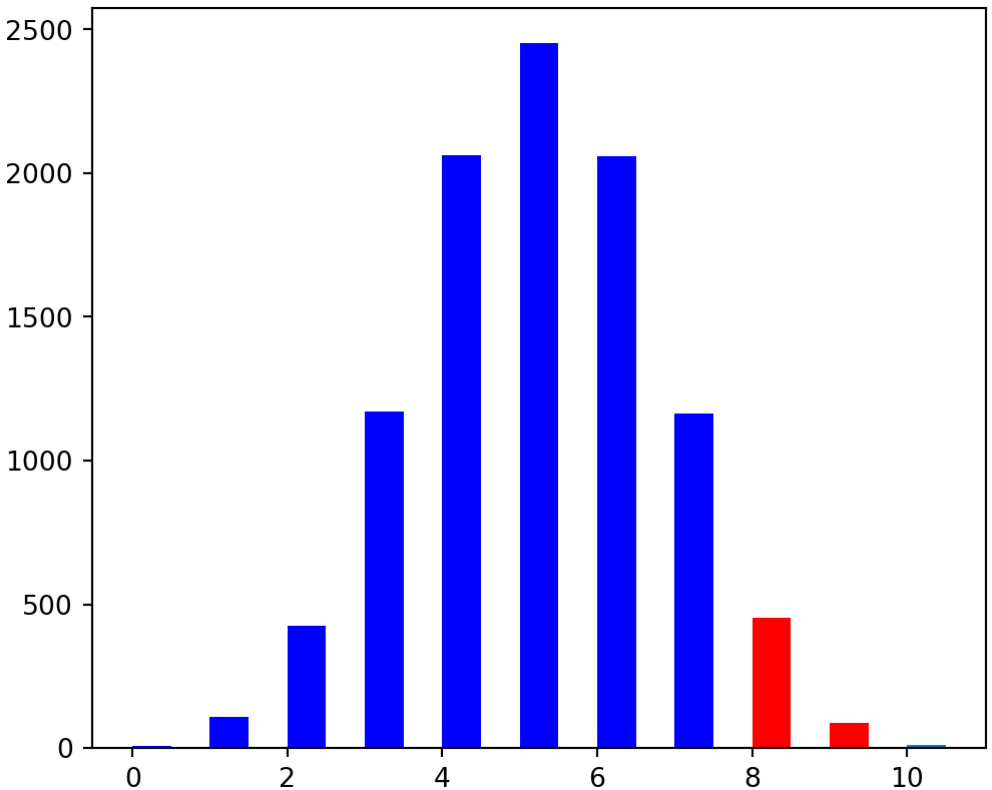
Method 1: Simulation (as before)

```
null_res = np.random.binomial(10, 0.5, size=10000)
```

```
pd.value_counts(null_res, normalize=True).sort_index()
```

```
0     0.0009
1     0.0108
2     0.0424
3     0.1171
4     0.2063
5     0.2452
6     0.2060
7     0.1164
8     0.0452
9     0.0087
10    0.0010
dtype: float64
```

▶ Code




```
one_sided_p_value = np.mean(null_res >= 8)
print(f'P(heads >= 8 | H0) = {one_sided_p_value: .3f}')

two_sided_p_value = np.mean(null_res >= 8) + np.mean(null_res<=2)
print(f'P(heads>=8 or heads<=2 | H0) = {two_sided_p_value: .3f}')
```

```
P(heads >= 8 | H0) =  0.055
P(heads>=8 or heads<=2 | H0) =  0.109
```

Method 2: Binomial Dist.

- Denote by X the number of heads in 10 tosses.
- Under $H_0 : P(\text{head}) = \frac{1}{2}$ what is the distribution of X?

$X \sim Bin(10, \frac{1}{2})$

- One sided: $P(X \geq 8) = \left(\binom{10}{8} + \binom{10}{9} + \binom{10}{10}\right) \cdot 2^{-10} = 0.055$
- Two sided: $2 \cdot 0.055 = 0.11$
- **Important lesson: proper simulation and proper mathematical analysis should give similar results**

Type-I and Type-II Errors

INTRODUCTION

TO DATA SCIENCE

The Alternative Hypothesis

- Let us introduce the simple alternative hypothesis
- In the courtroom example:

$$H_0 : \text{innocent}$$

$$H_1 : \text{guilty}$$

- In the coin example, $X \sim Bin(10, p)$, and:

$$H_0 : p = 0.5$$

$$H_1 : p = 0.8$$

- In the paintings example, μ is the diff in red between impressionist and realist, and:

$$H_0 : \mu = 0$$

$$H_1 : \mu = 20$$

Type I and Type II Errors

- What could go wrong?

- We **reject** H_0 when it is true, Type-I error: $\alpha = P(\text{reject } H_0 | H_0 \text{ true})$
- We **don't reject** H_0 when we should, Type-II error: $\beta = P(\text{not reject } H_0 | H_1 \text{ true})$
- (notice the jargon)

The meaning of these terms is often better understood in a table:

Reality\Decision	**Not Reject H_0**	**Reject H_0**
H_0	Confidence: $1 - \alpha$	Type I Error: α
H_1	Type II Error: β	Power: $1 - \beta$

Statistical power is often written as $1 - \beta$, or: $\pi = P(\text{reject } H_0 | H_1 \text{ true})$

Highly important, see later.

Two Common Approaches to Testing

1. Compute p-value and compare to some threshold α (1%, 5%)
   - If p-value $\leq \alpha \Rightarrow$ reject H_0
   - If p-value $> \alpha \Rightarrow$ don't reject H_0

2. Rejection area: looking at some statistic of the sample $T(X)$, by fixing α at some "significance level", extract a critical value C and compare to it:
   - If $T(X) \geq C \Rightarrow$ reject H_0
   - If $T(X) < C \Rightarrow$ don't reject H_0

(for a one-sided hypothesis)

Method 3: Rejection Area

- Denote by X the number of heads in 10 tosses.
- $X \sim Bin(10, p)$, under $H_0 : p = \frac{1}{2}$
- $T(X) = X$, the outcome itself
- Set $\alpha = 0.01$
- Extract C:

$$\alpha = 0.01 \approx P(X \geq C) \Rightarrow C = 9$$

- If $X \geq 9 \Rightarrow$ reject H_0
- If $X < 9 \Rightarrow$ don't reject H_0
- Got $X = 8$ in sample so not rejecting H_0
- **Crucial Question**: Did we need H_1?

Or by simulation

Back to paintings, set $\alpha = 0.01$ and extract C the 99th quantile of the null:

```
print(f'C (above which 1% of null distribution) = {np.quantile(null_mean_diffs, 0.99):
```

```
C (above which 1% of null distribution) =  23.21
```

We got 15 points difference, so.

Again: Did we need H_1?

Intro. to Power

- It is only when we want to calculate the power when we need a specific H_1.
- In the coin example, $X \sim Bin(10, p)$, and:

$$H_0 : p = 0.5$$

$$H_1 : p = 0.8$$

- The test was set at:
  - If $X \geq 9 \Rightarrow$ reject H_0 (rejection area)
  - If $X < 9 \Rightarrow$ don't reject H_0

Power =

$$P(\text{reject } H_0 | H_1 \text{ true}) = P_{H_1}(X \geq 9) = \binom{10}{9} \cdot 0.8^9 \cdot 0.2 + \binom{10}{10} \cdot 0.8^{10} = 0.376$$

So for this test we have less than 40% chance of rejecting H_0 when we should!

Or by simulation?

How would we calculate the power for the paintings test?

Intro to Data Science