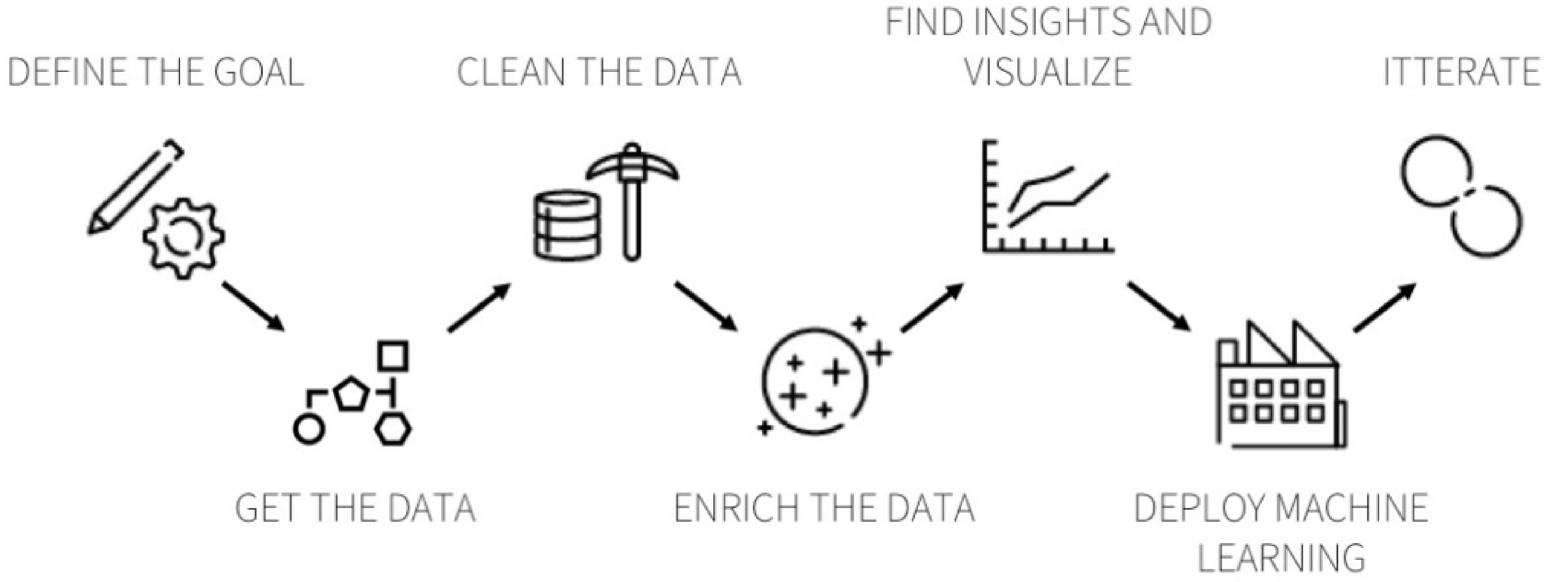
Exploring Data with `Statistics`

The Data Science Process



Agenda

- Classical (Frequentist) Hypothesis Testing
- Type I and Type II Errors
- p-values and Confidence Intervals
- Hypothesis Testing using Python

Hypothesis Testing

Whenever we observe an apparent effect (such as a difference in means between groups) in a sample drawn from a population, we must ask the question -

> *Could this have happened by chance?*

Hypothesis Testing presents us with a framework for quantifying this probability. The approach is similar to the mathematical notion of **Proof by Contradiction** where in order to prove a statement S1, we *start by assuming that S1 is False*, then we compute some metric and if these calculations lead to a *contradiction*, we conclude that S1 must be True.

On these lines, to prove a hypothesis like *The Mean score for Group-A is different from the Mean score for Group-B* we

- assume temporarily that it's the same and call this the **Null Hypothesis,** denoted by H_0.
- compute the probability of observing a such an extreme difference in Means and call it the **p-value.**
- If this probability is too low, we conclude that the data does not support our initial assumption and the H_0 is unlikely to be True.

Conventionally, we set the **threshold** for the p-value *before* we conduct the analysis, calling it the **Level of Significance** and denoting it by the letter α. This is an arbitrary choice and the magnitude of the p-value is dependent upon the test statistic chosen and the model of the Null Hypothesis. Traditionally, we set $\alpha = 0.05$

Example

In a medical study we observe that 10% of the women and 12.5% of the men suffer from heart disease. If there are 20 people in the study, we would probably be hesitant to declare that women are less prone to suffer from heart disease than men; it is very possible that the results occurred by chance. However, if there are 20,000 people in the study, then it seems more likely that we are observing a real phenomenon.

> Hypothesis testing makes this intuition precise; it is a framework that allows us to decide whether patterns that we observe in our data are likely to be the result of random fluctuations or not.

The Hypothesis

In the example above, we might specify our H_0 as *"heart disease is at least as prevalent in men as in women"*

If the null hypothesis holds, then whatever pattern we are detecting in our data that seems to support our conjecture is just a fluke. There just happen to be a lot of men with heart disease (or women without) in the study.

The Test

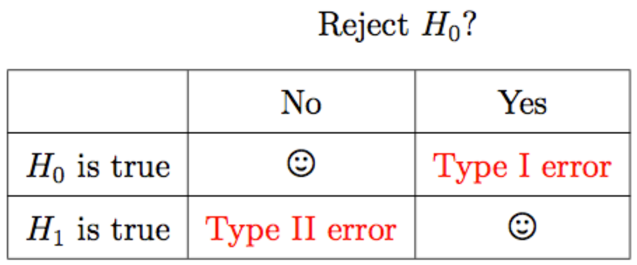
We use the data to calculate a metric known as the **Test Statistic**, and then we compare it against a pre-computed **threshold** at the decided significance level. The test then rejects the H_0 if the test statistic falls under a rejection region. Note that if we fail to reject the null hypothesis, this does not mean that we consider it likely, we just don't have enough information to discard it.

> For example, in a $t - test$, we compute the $t - statistic$ from the data, and compare it against the t_{crit} or the critical value of t set against $\alpha = 0.05$

The Errors

- A Type I error is a **False Positive**, when we reject the null hypothesis even though it is True.
- A Type II error is a **False Negative**, we fail to reject the null hypothesis even though it is False.

> In hypothesis testing, our priority is to control Type I errors.
> A statistically significant result at $\alpha = 0.05$ means that the probability of having committed a Type I error is bounded by 5%.



Reject H_0?

	No	Yes
H_0 is true	☺	Type I error
H_1 is true	Type II error	☺

Steps in Hypothesis Testing

1. Choose an appropriate **test statistic** to quantify the **effect**
2. Define a **Null Hypothesis** (typically based on the assumption that the apparent effect is not real.)
3. Compute the **probability** of observing an effect equal to or larger than the effect observed under the assumption that the Null Hypothesis is True. This is the **p-value**
4. If the p-value is lower than a pre-meditated threshold, we conclude that our assumption is invalid, and the effect is in fact statistically significant and is unlikely to have occurred by chance alone.

p-values

If you take two samples from the same population there will always be a difference between them. Statistical significance is the likelihood that the observed difference between the two groups could just be **an accident of sampling**.

> The statistical significance is usually calculated as a 'p-value', the probability that a difference of at least the same size would have arisen by chance, even if there really were no difference between the two populations.

For differences between the means of two groups, this p-value would normally be calculated from a 't-test'. By convention, if $p < 0.05$ (i.e. below 5%), the difference is taken to be large enough to be 'significant'; if not, then it is 'not significant'.

Problems with p-values

The p-value depends essentially on two things (1) the size of the effect and (2) the size of the sample.

One would get a 'significant' result either

- if the effect were very big (despite having only a small sample) or
- if the sample were very big (even if the actual effect size were tiny).

Confidence Intervals

It is important to know the statistical significance of a result, since without it there is a danger of drawing firm conclusions from studies where the sample is too small to justify such confidence.

However, statistical significance alone does not tell you the most important thing: the size of the effect. Therefore, it is recommended to report the effect size, together with an estimate of its likely 'margin for error' or 'confidence interval'.

> If an effect size is calculated from a very large sample it is likely to be more accurate than one calculated from a small sample.

To calculate a 95% confidence interval,

- assume that the observed effect it true
- calculate the amount of variation in this estimate by repeatedly taking new samples of the same size

If this confidence interval includes zero, then that is the same as saying that the result is not statistically significant. If, on the other hand, zero is outside the range, then it is 'statistically significant at the 5% level'. Using a confidence interval is a better way of conveying this information since it keeps the emphasis on the effect size - which is the important information - rather than the p-value.

Hypothesis Testing in Python

Student's t-tests

1-sample t-test: testing the value of a population mean

`scipy.stats.ttest_1samp()` tests if the mean of observed data is likely to be equal to a given value. It returns the T statistic, and the p-value. If the p-value is less than 0.05, we conclude that the mean is different from 0. For example, to test if the population mean of `my_data` is equal to 0, we would write

```
stats.ttest_1samp(my_data, 0)
```

2-sample t-test: testing for difference across populations

To test if a found effect size is statistically significant, we run a two-sample t-test on the two arrays

```
# for independent samples
stats.ttest_ind(female_height, male_height)

# if the data come from measurements made on the same entities at different points in time (before/after)
stats.ttest_rel(y_before, y_after)
```

Quiz

1. Which hypothesis is typically assumed to be true in hypothesis testing?

    - The null.
    - The alternative.

2. The type I error rate controls what?

3. A confidence interval for the mean contains:

```
- All of the values of the hypothesized mean for which we would fail to reject with α =1 − Conf.Level.
- All of the values of the hypothesized mean for which we would fail to reject with 2α = 1 − Conf.Level.
- All of the values of the hypothesized mean for which we would reject with α = 1 − Conf.Level.
- All of the values of the hypothesized mean for which we would reject with 2α = 1 − Conf.Level
```

Assignments

- Load the data set `mtcars`. Assume that it is a random sample. Compute the mean MPG, $\bar{x}$, of this sample. Test whether the true MPG is μ_0 or smaller using a one-sided 5% leveltest. $(H_0 : \mu = \mu_0 - versus - H_a : \mu < \mu_0)$. Using that data set and a Z test: Based on the mean MPG of the sample $\bar{x}$, and by using a Z test: what is the smallest value of μ_0 that you would reject for (to two decimal places)?
- Consider again the `mtcars` dataset. Use a two-group $t - test$ to test the hypothesis that the 4 and 6 cyl cars have the same mpg. Do you reject or fail to reject the H_0? And why?
- You believe the coin that you're flipping is biased towards heads. You get 55 heads out of 100 flips. Do you reject at the 5% level that the coin is fair?