Data Analytics and Mathematical Statistics
Day 3

Source: Towards Data Science, Spring Board, The Software Report

Agenda

Lecture 3 and Tutorial 3

- Using our sample and assessing the findings
  - Hypothesis testing
  - Investigating questions involving groups
  - Hands-on assessing sample findings

Using our sample and assessing the findings

What is probability and what does it have to do with reasoning with data?

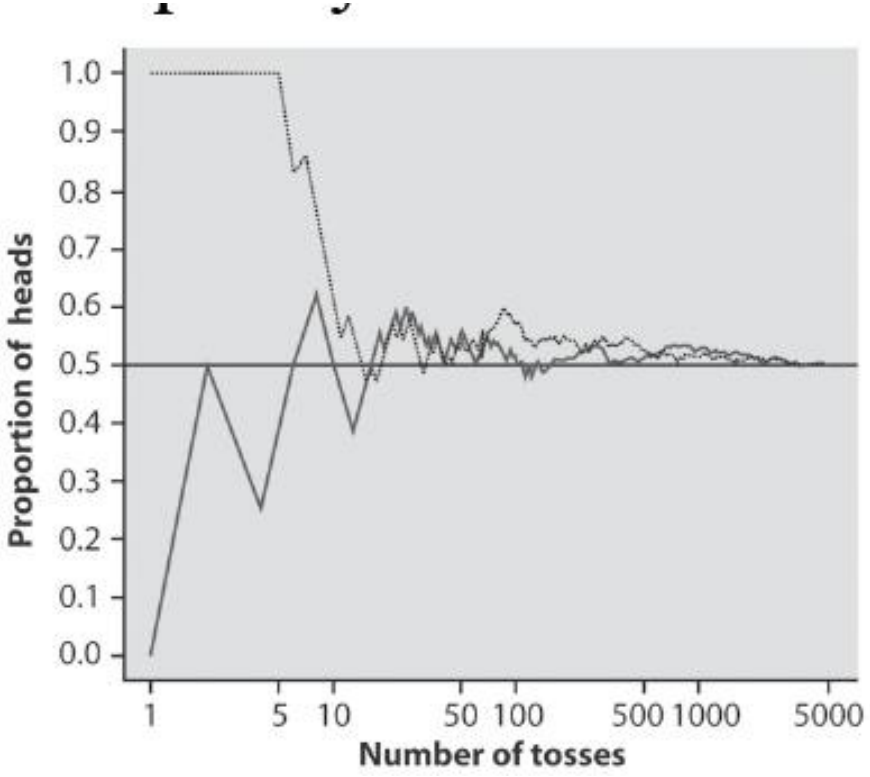
Source: Nintendo

Basic Probability

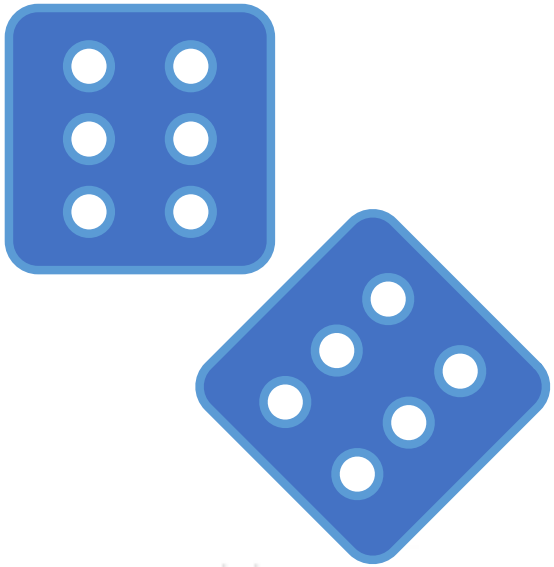
- Relative frequency (proportion of occurrences) of an outcome settles down to one value over the long run
- That one value is then defined as to be the probability of that outcome

Relative-Frequency Probabilities

Coin flipping:



Relevance of probability?

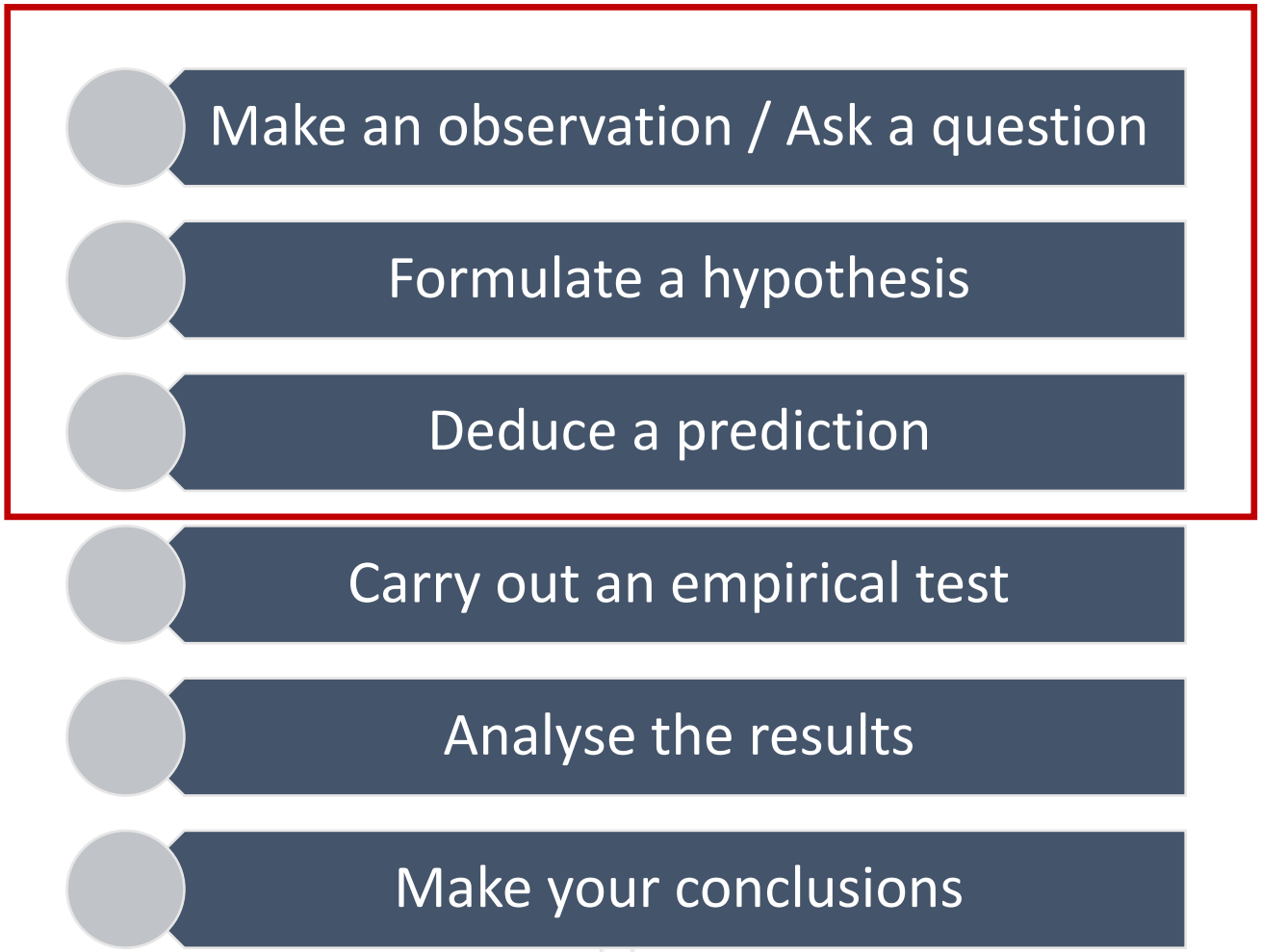


Using our sample

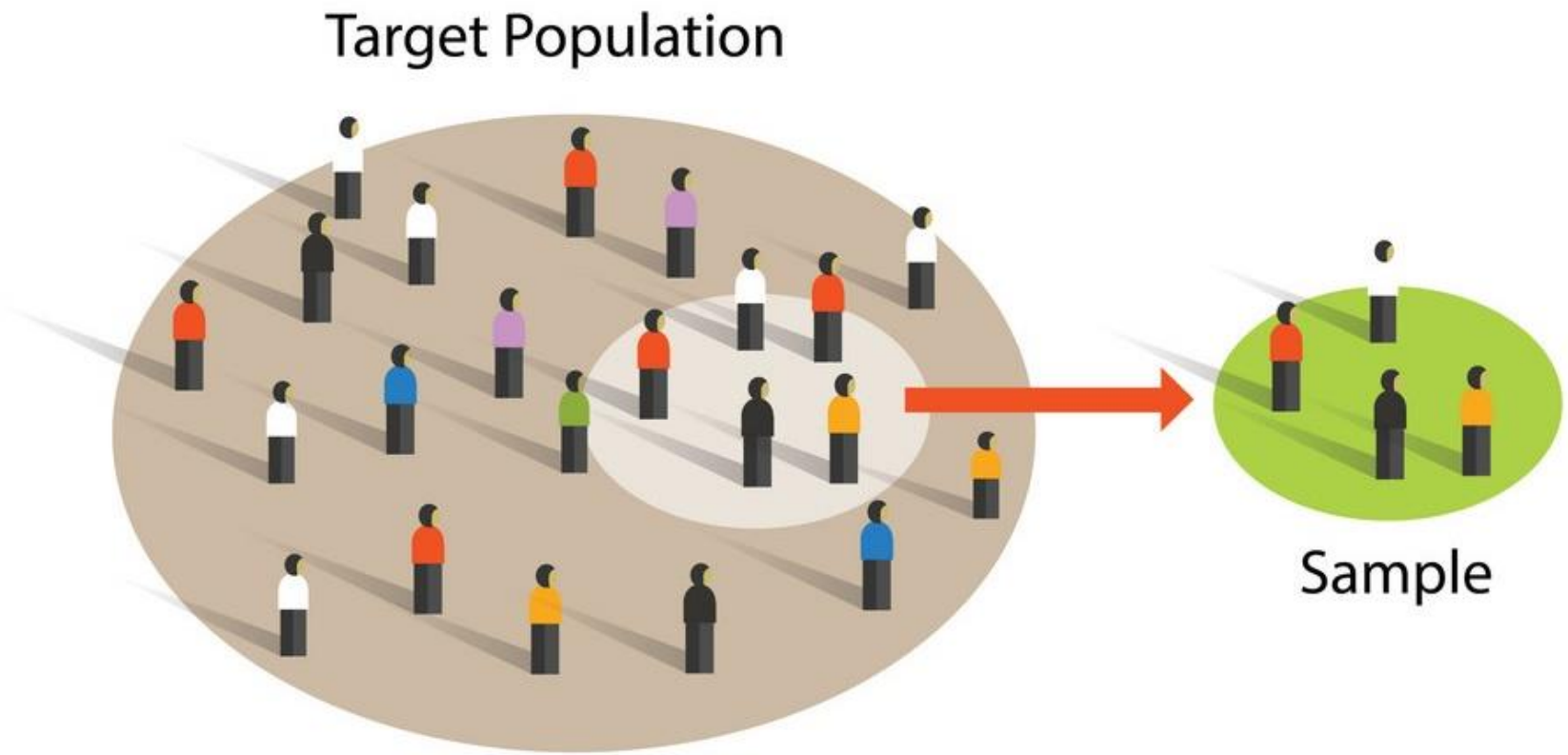
- Say my population of interest is the 800 students in my home faculty, and the dept is concerned about their well-being

- So let's use answer this question

Using our sample



Population and sampling



Source: VectorStock

Using our sample

- Since I am interested in the population mean μ, it seems logical to use the sample mean $\bar{X}$ as a way to "estimate" it
- My sample mean is known as a point estimator; different samples of 30 students will give me different values of my estimator
- A single sample will give me one value, say 67 kg – a point estimate

Using our sample

- If we obtain a point estimate for $\bar{X}$ of 67 kg:
  - Is this saying if μ > 60 kg or not?
- We do a hypothesis test

Using our sample

- We set up our null and alternative hypotheses:

- H_0: $\mu = 60$
- H_1: $\mu > 60$

- And we review the p-value

Using our sample

- If $\bar{X}$ = 67 kg, p-value is P($\bar{X}$ ≥ 67 kg | μ = 60 kg)
- H_0 is always specific, while H_1 is non-specific
- Hypothesis testing allows us to determine if our prediction is accurate, and therefore if our general hypothesis is supported

Using our sample

- What if we have a small p-value? It indicates it's not likely for us to observe such a value in $\bar{X}$ if H_0 is true, i.e. the value of $\bar{X}$ is too big

- So the fact that we are seeing this "extreme" value means it's not likely for H_0 to be true → reject H_0

Using our sample

- Conversely, a high p-value indicates that it's likely for us to observe such a value of $\bar{X}$, given that H_0 is true, i.e. $\bar{X}$ is close enough to 60 kg

- So we don't reject H_0

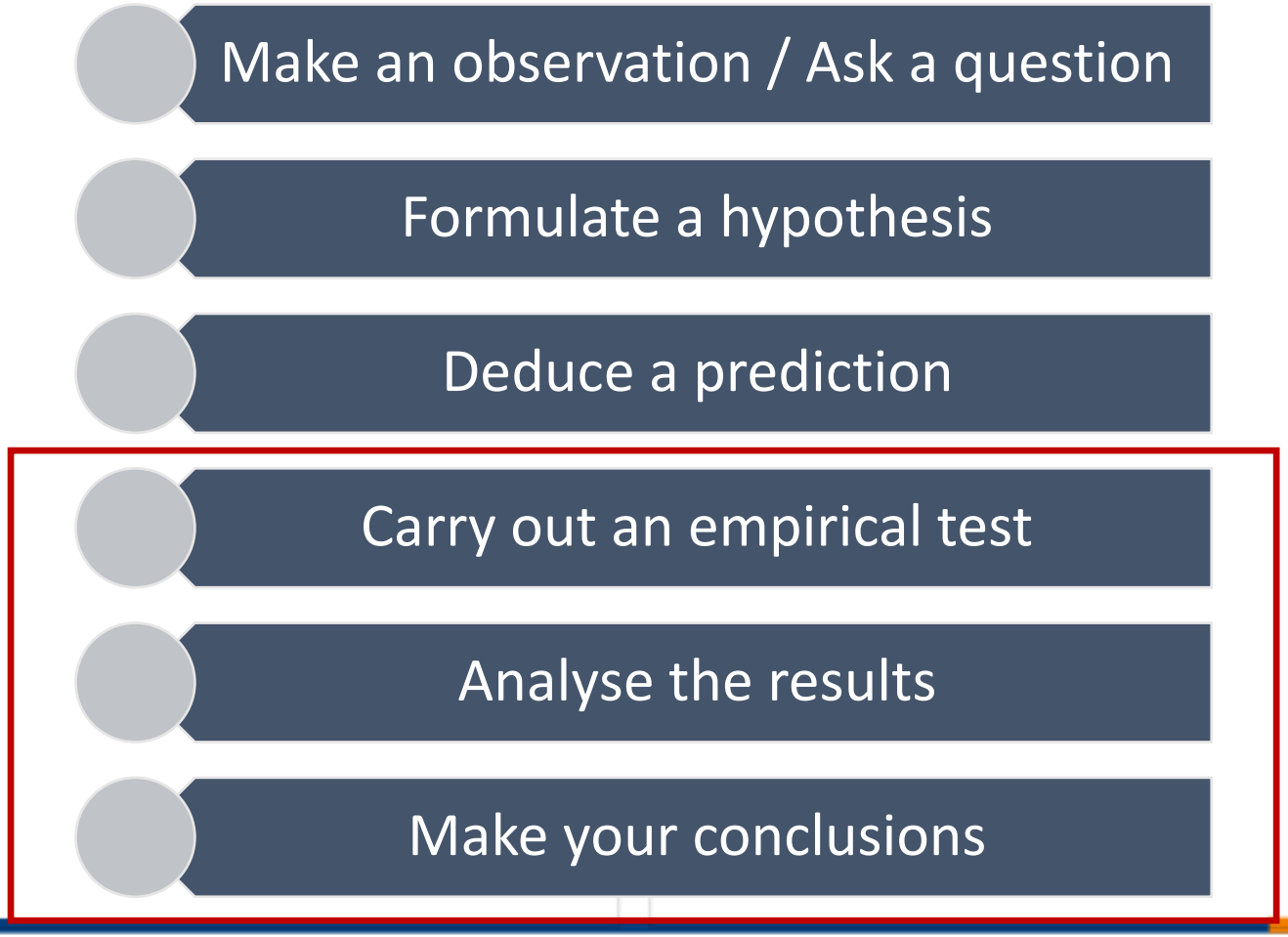
- How do we decide whether a p-value is high or low: compare against the level of significance, α

Using our sample

- The level of significance typically takes on values of 0.05, 0.01 and 0.001, and is up to the user's discretion to decide
  - we reject the null hypothesis as long as p-values fall below, say, 0.05

- In a way, the α represents our risk appetite: what is error rate that is acceptable to us?

Using our sample



Tutorial 3

- Let us carry out our empirical test, collect data from 30 students randomly chosen from our population of interest

- Let's analyse the results, and make our conclusions appropriately from there

Concluding

Tutorial 3 (continued)

- Did we come to the right conclusion? Do we know in this case?
- BUT, what if we chose the 30 students with the highest weight? Is it possible that we do this? Yes, not likely, but possible
- Re-do the hypothesis test and see if there is a difference in the conclusion we would draw

Takeaway

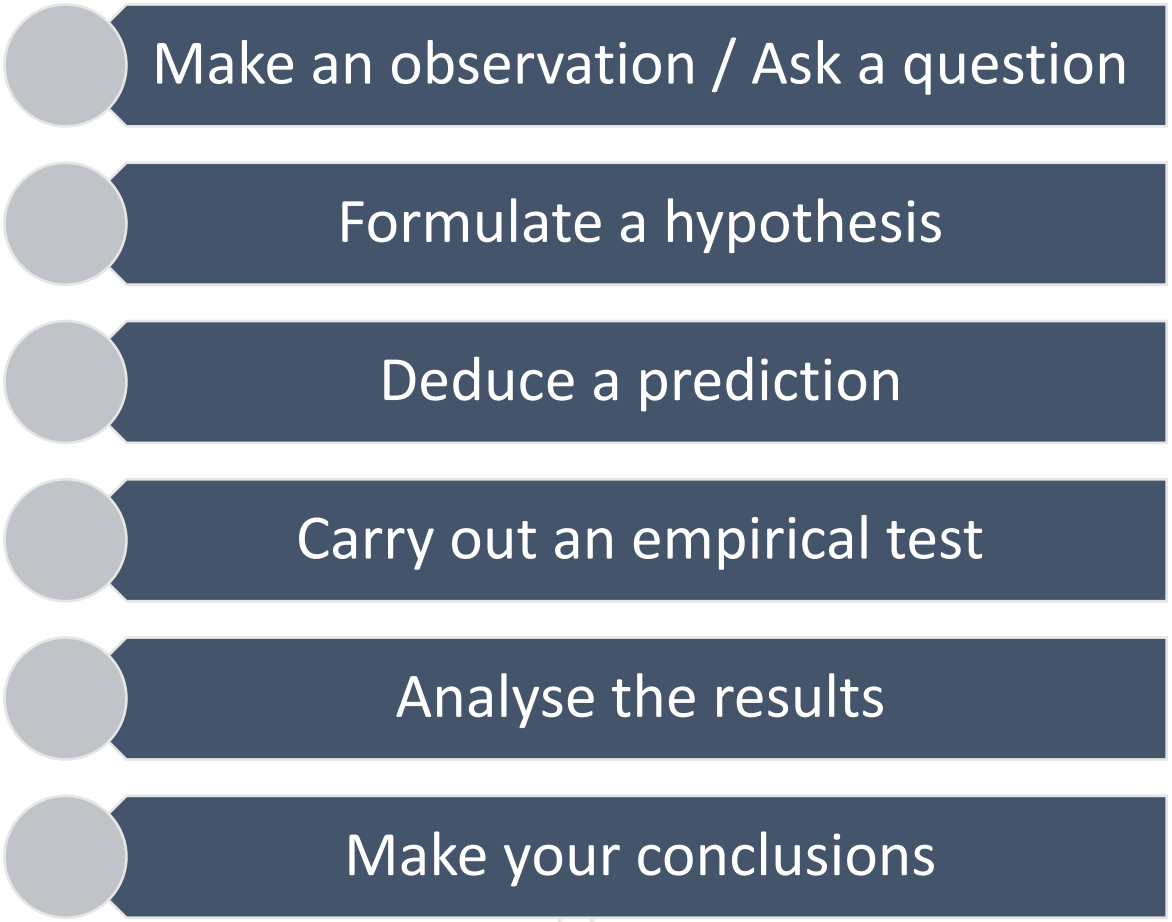
- Because we are reasoning with just a limited sample, and making conclusions about the wider population
- It is not possible for us to arrive at the right conclusions all the time
- What we can hope for is to be correct, most of the time

Answering questions involving groups

- We might also be interested in investigating questions on comparing the means between two groups

- e.g. is the air quality in the West of Singapore different than in the North?

Answering questions involving groups



Carry out empirical test

PSI in the North, μ_n

PSI in the West, μ_w,

$H_0: \mu_w = \mu_n$

$H_0: \mu_w \neq \mu_n$

Which statement is correct? Let's conduct our hypothesis test!

How should we conclude?

End of Day 3