PRINCIPLES OF DATA SCIENCE

2021 - 2022

Assoc. Prof. Dr. Salha Alzahrani

Department of Computer Science

جامعة الطائف
TAIF UNIVERSITY

CHAPTER 8: ADVANCED STATISTICS

- POINT ESTIMATES
- SAMPLING DISTRIBUTIONS
- CONFIDENCE INTERVALS
- HYPOTHESIS TESTS
  - Conducting a hypothesis test
  - One sample t-tests
  - Type I and type II errors
  - Hypothesis test for categorical variables

POINT ESTIMATES

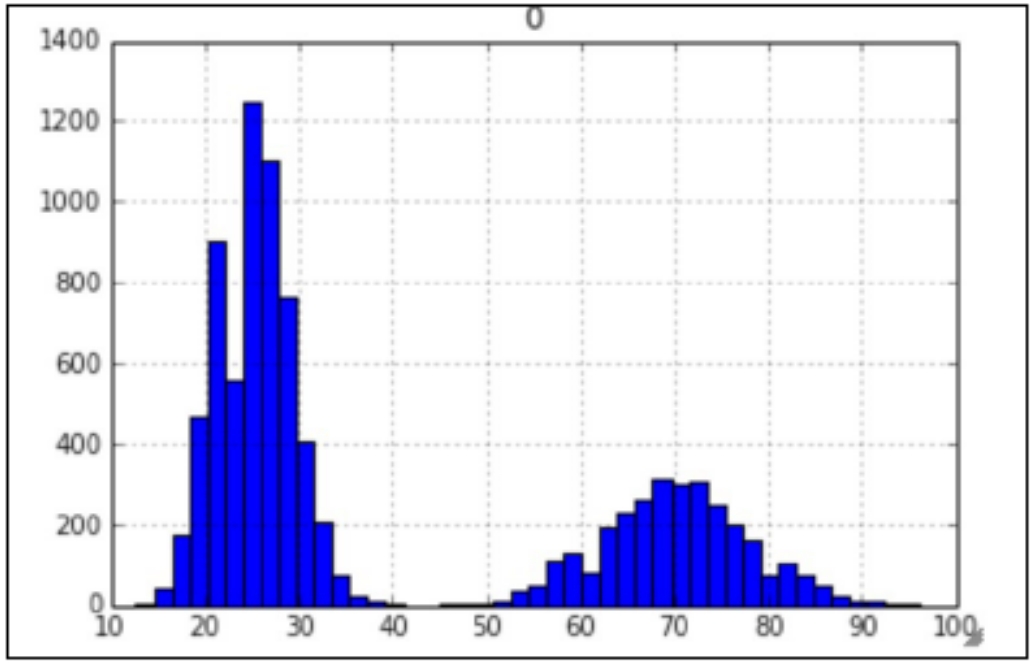
Point estimates

- In the previous chapter, we mentioned how difficult it was to obtain a population parameter; so, we had to use sample data to calculate a statistic that was an estimate of a parameter. When we make these estimates, we call them point estimates.
- A point estimate is an estimate of a population parameter based on sample data.
- We use point estimates to estimate population means, variances, and other statistics.
- To obtain these estimates, we simply apply the function that we wish to measure for our population to a sample of the data. For example, suppose there is a company of 9,000 employees and we are interested in ascertaining the average length of breaks taken by employees in a single day. As we probably cannot ask every single person, we will take a sample of the 9,000 people and take a mean of the sample. This sample mean will be our point estimate.

SAMPLING DISTRIBUTIONS

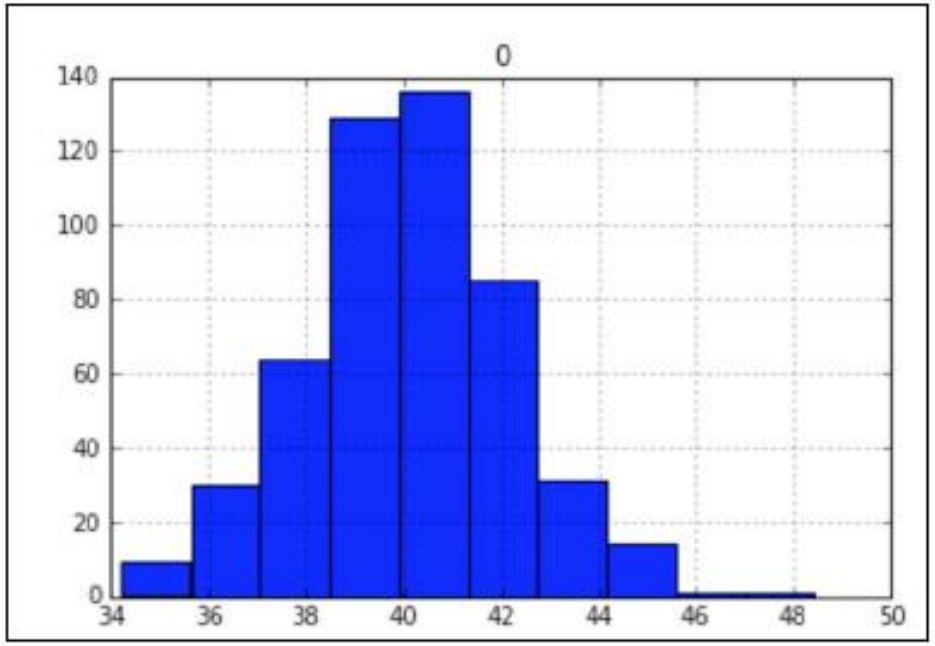
Sampling distributions

- In Chapter 7, Basic Statistics, we mentioned how much we love when data follows the normal distribution.
- One of the reasons for this is that many statistical tests rely on data that follows a normal pattern, and for the most part, a lot of real-world data is not normal (surprised?)
- Example, the data in the graph is definitely not following a normal distribution, it appears to be bi-modal, which means that there are two peaks.



Sampling distributions

- As our data is not normal, many of the most popular statistics tests may not apply, however, if we follow the given procedure, we can create normal data!
- We will need to utilize what is known as a sampling distribution, which is a distribution of point estimates of several samples of the same size.
- 1. Take 500 different samples of the x of size 100 each.
- 2. Take a histogram of these 500 different point estimates (revealing their distribution).
- The number of elements in the sample (100) was arbitrary, but large enough to be a representative sample of the population.
- The number of samples I took (500) was also arbitrary, but large enough to ensure that our data would converge to a normal distribution



CONFIDENCE INTERVALS

Confidence intervals

- While point estimates are okay estimates of a population parameter and sampling distributions are even better, there are the following two main issues with these approaches:
  - Single point estimates are very prone to error (due to sampling bias among other things)
  - Taking multiple samples of a certain size for sampling distributions might not be feasible, and may sometimes be even more infeasible than actually finding the population parameter
- For these reasons and more, we may turn to a concept, known as confidence interval, which is a range of values based on a point estimate that contains the true population parameter at some confidence level.
- Confidence is an important concept in advanced statistics. Its meaning is sometimes misconstrued. Informally, a confidence level does not represent a "probability of being correct"; instead, it represents the frequency that the obtained answer will be accurate. For example, if you want to have a 95% chance of capturing the true population parameter using only a single point estimate, we would have to set our confidence level to 95%.

Confidence intervals

- Calculating a confidence interval involves finding a point estimate, and then, incorporating a margin of error to create a range. The margin of error is a value that represents our certainty that our point estimate is accurate and is based on our desired confidence level, the variance of the data, and how big your sample is.
- There are many ways to calculate confidence intervals; for the purpose of brevity and simplicity, we will look at a single way of taking the confidence interval of a population mean. For this confidence interval, we need the following:
  - A point estimate. For this, we will take our sample mean of break lengths from our previous example.
  - An estimate of the population standard deviation, which represents the variance in the data.
  - This is calculated by taking the sample standard deviation (the standard deviation of the sample data) and dividing that number by the square root of the population size.
  - The degrees of freedom (which is the -1 sample size).

HYPOTHESIS TESTS

Hypothesis tests

- Hypothesis tests are one of the most widely used tests in statistics. They come in many forms; however, all of them have the same basic purpose.
- A hypothesis test is a statistical test that is used to ascertain whether we are allowed to assume that a certain condition is true for the entire population, given a data sample. Basically, a hypothesis test is a test for a certain hypothesis that we have about an entire population. The result of the test then tells us whether we should believe the hypothesis or reject it for an alternative one.
- You can think of the hypothesis tests' framework to determine whether the observed sample data deviates from what was to be expected from the population itself. Now this sounds like a difficult task but, luckily, Python comes to the rescue and includes built-in libraries to conduct these tests easily.
- A hypothesis test generally looks at two opposing hypotheses about a population.
- We call them the null hypothesis and the alternative hypothesis. The null hypothesis is the statement being tested and is the default correct answer; it is our starting point and our original hypothesis. The alternative hypothesis is the statement that opposes the null hypothesis. Our test will tell us which hypothesis we should trust and which we should reject..

Hypothesis tests

- Based on sample data from a population, a hypothesis test determines whether or not to reject the null hypothesis. We usually use a p-value (which is based on our significance level) to make this conclusion.
- A very common misconception is that statistical hypothesis tests are designed to select the more likely of the two hypotheses. This is incorrect. A hypothesis test will default to the null hypothesis until there is enough data to support the alternative hypothesis.

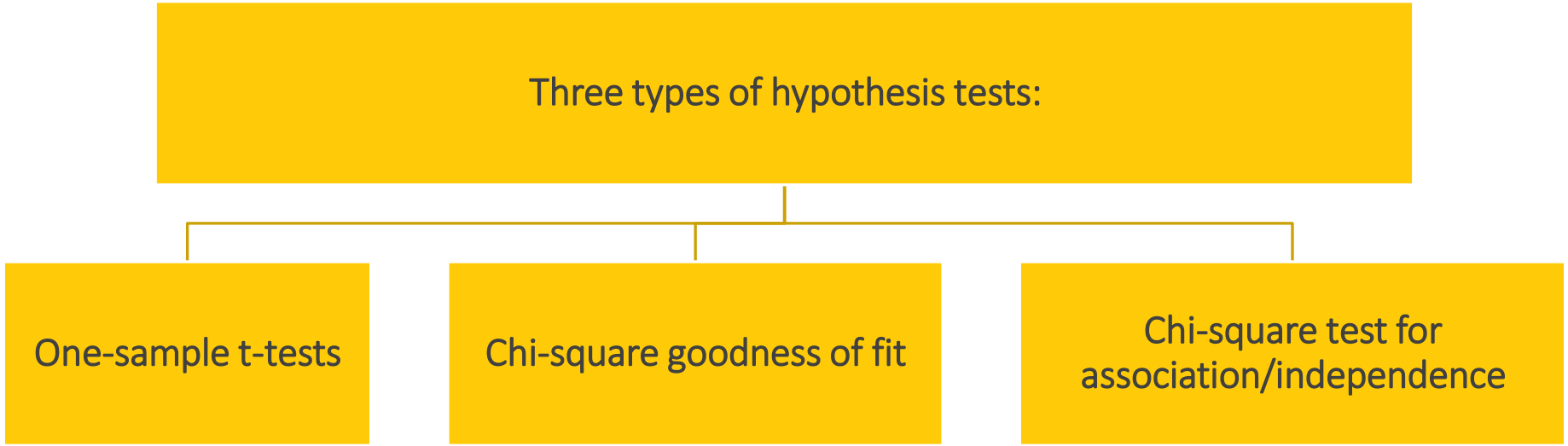
- The following are some examples of questions you can answer with a hypothesis test:
  - Does the mean break time of employees differ from 40 minutes?
  - Is there a difference between people who interacted with website A and people who interacted with website B (A/B testing)?
  - Does a sample of coffee beans vary significantly in taste from the entire population of beans?

Conducting a hypothesis test

There are five basic steps that most hypothesis tests follow, which are as follows:

1. Specify the hypotheses:
   - Here, we formulate our two hypotheses: the null and the alternative
   - We usually use the notation of H0 to represent the null hypothesis and Ha to represent our alternative hypothesis
2. Determine the sample size for the test sample:
   - This calculation depends on the chosen test. Usually, we have to determine a proper sample size in order to utilize theorems, such as the central limit theorem, and assume the normality of data.
3. Choose a significance level (usually called alpha or α):
   - A significance level of 0.05 is common
4. Collect the data:
   - They collect a sample of data to conduct the test
5. Decide whether to reject or fail to reject the null hypothesis:
   - This step changes slightly based on the type of test being used. The final result will either yield in rejecting the null hypothesis in favor of the alternative or failing to reject the null hypothesis.

Hypothesis tests



- There are many more tests. However, these three are a great combination of distinct, simple, and powerful tests. One of the biggest things to consider when choosing which test we should implement is the type of data we are working with, specifically, are we dealing with continuous or categorical data.
- In order to truly see the effects of a hypothesis, I suggest we dive right into an example. First, let's look at the use of a t-tests to deal with continuous data.

One sample t-tests

- The one sample t-test is a statistical test used to determine whether a quantitative (numerical) data sample differs significantly from another dataset (the population or another sample).

- Assumptions of the one sample t-tests:
  Before diving into the five steps, we must first acknowledge that t-tests must satisfy the following two conditions to work properly:
  - The population distribution should be normal, or the sample should be large ($n \geq 30$).
  - In order to make the assumption that the sample is independently randomly sampled, it is sufficient to enforce that the population size should be at least 10 times larger than the sample size ($10n < N$).
- Note that our test requires that either the underlying data be normal (which we know is not true for us), or that the sample size be at least 30 points large. For the t-test, this condition is sufficient to assume normality. This test also requires independence, which is satisfied by taking a sufficiently small sample
- The basic idea is that our sample must be large enough to assume normality (through conclusions similar to the central limit theorem) but small enough as to be independent from the population.

Type I and type II errors

- A type I error occurs if we reject the null hypothesis when it is actually true. This is also known as a false positive. The type I error rate is equal to the significance level α, which means that if we set a higher confidence level, for example, a significance level of 99%, our α is .01, and therefore our false positive rate is 1%.

- A type II error occurs if we fail to reject the null hypothesis when it is actually false. This is also known as a false negative. The higher we set our confidence level, the more likely we are to actually see a type II error.

HYPOTHESIS TEST FOR CATEGORICAL VARIABLES

Hypothesis test for categorical Variables

- T-tests (among other tests) are hypothesis tests that work to compare and contrast quantitative variables and underlying population distributions.
- We will explore two new tests, both of which serve to explore qualitative data. They both
- are a form of test called chi-square tests. These two tests will perform the following two tasks for us:
  - Determine whether a sample of categorical variables is taken from a specific population (similar to the t-test)
  - Determine whether two variables affect each other and are associated to each other.

- Chi-square goodness of fit test : The one-sample t-test was used to check whether a sample mean differed from the population mean. The chi-square goodness of fit test is very similar to the one sample t-test in that it tests whether the distribution of the sample data matches an expected distribution, while the big difference is that it is testing for categorical variables.

s.zahrani@tu.edu.sa