**Hypothesis Testing**

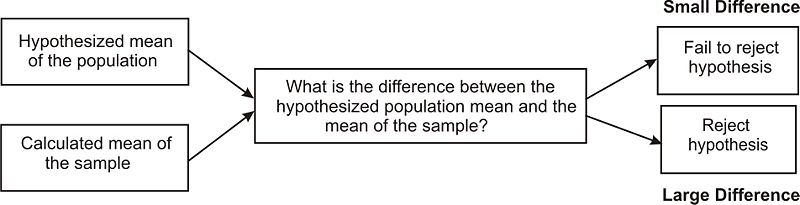
In everyday life, we often have to make decisions based on incomplete information. These may be decisions that are important to us such as, "Will I improve my prediction score If I spend more time fine tuning the model?” This section is about the use of hypothesis testing to help us with these decisions.

**Hypothesis testing** is a kind of statistical inference that involves asking a question, collecting data, and then examining what the data tells us about how to proceed.

In a formal hypothesis test, hypotheses are always statements about the population. The hypothesis tests we will examine in this chapter involve statements about the average values(means) of some variable in the population.

**Developing Null and Alternative Hypotheses**

Hypothesis testing involves testing the difference between a hypothesized value of a population parameter and the estimate of that parameter which is calculated from a sample. If the parameter of interest is the mean of the populations in hypothesis testing, we are essentially determining the magnitude of the difference between the mean of the sample and the hypothesized mean of the population. If the difference is very large, we reject our hypothesis about the population. If the difference is very small, we do not. Below is an overview of this process.



In statistics, the hypothesis to be tested is called the null hypothesis and given the symbol H0. The alternative hypothesis is given the symbol Ha.

The null hypothesis defines a specific value of the population parameter that is of interest. Therefore, the null hypothesis always includes the possibility of equality. Consider

H0: μ=3.2

Ha: μ≠3.2

In this situation if our sample mean, x̄, is very different from 3.2 we would reject H0. That is, we would reject H0 if x̄ is much larger than 3.2 or much smaller than 3.2. This is called a **2-tailed test**. An x̄ that is very unlikely if H0 is true is considered to be good evidence that the claim H0 is not true. Consider

H0: μ≤3.2

Ha: μ>3.2

In this situation we would reject H0 for very large values of x̄. This is called a **one tail test**. If, for this test, our data gives x̄=15, it would be highly unlikely that finding x̄ this different from 3.2 would occur by chance and so we would probably reject the null hypothesis in favor of the alternative hypothesis.

We test the null hypothesis against an alternative hypothesis, which is given the symbol Ha and includes the outcomes not covered by the null hypothesis. Basically, the alternative hypothesis states that there is a difference between the hypothesized population mean and the sample mean. The alternative hypothesis can be supported only by rejecting the null hypothesis.

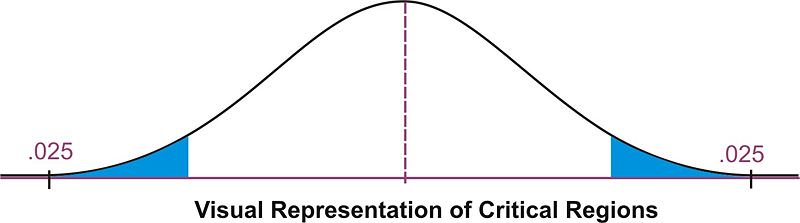
**Deciding Whether to Reject the Null Hypothesis: One-Tailed and Two-Tailed Hypothesis Tests**

When a hypothesis is tested, a statistician must decide on how much evidence is necessary in order to reject the null hypothesis. For example, if the null hypothesis is that the average height of a population is 64 inches a statistician wouldn't measure one person who is 66 inches and reject the hypothesis based on that one trial. It is too likely that the discrepancy was merely due to chance.

We use statistical tests to determine if the sample data give good evidence against the claim (H0). The numerical measure that we use to determine the strength of the sample evidence we are willing to consider strong enough to reject H0 is called the **level of significance** and it is denoted by **α**. If we choose, for example, α=0.01 we are saying that we would get data at least as unusual as the data we have collected no more than 1% of the time when H0 is true. Also, confidence level = 1- **α,** which shows the values lying under the **confidence interval**.

The most frequently used levels of significance are 0.05 and 0.01. If our data results in a statistic that falls within the region determined by the level of significance, then we reject H0. The region is therefore called the **critical region**. When choosing the level of significance, we need to consider the consequences of rejecting or failing to reject the null hypothesis. If there is the potential for occurring an event, we should use a more ‘conservative’ critical region with levels of significance such as .005 or .001.

When determining the critical regions for a two-tailed hypothesis test, the level of significance represents the extreme areas under the normal density curve. We call this a two-tailed hypothesis test because the critical region is located in both ends of the distribution. For example, if there was a significance level of 0.95 the critical region would be the most extreme 5 percent under the curve with 2.5 percent on each tail of the distribution.



Therefore, if the mean from the sample taken from the population falls within one of these critical regions, we would conclude that there was too much of a difference between our sample mean and the hypothesized population mean and we would reject the null hypothesis. However, if the mean from the sample falls in the middle of the distribution (in between the critical regions) we would fail to reject the null hypothesis.

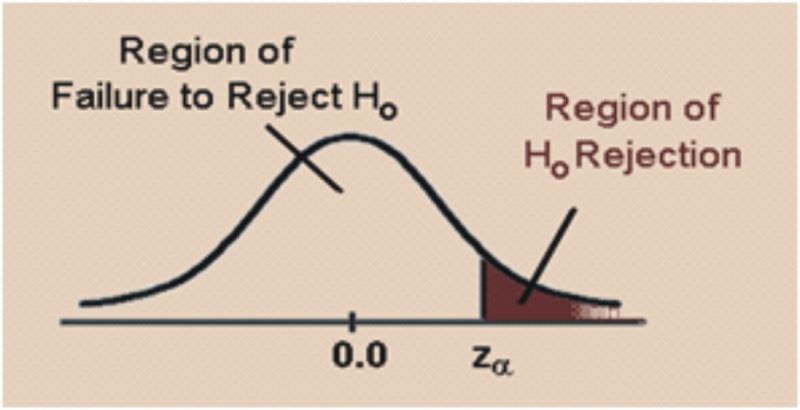
We calculate the critical region for the single-tail hypothesis test a bit differently.

When developing the alternative hypothesis in a single-tail hypothesis test we would use the symbols of greater than or less than. Using our example about average height of a population is 64 inch, our null and alternative hypothesis could look something like:

H0: μ=64

Ha: μ>64

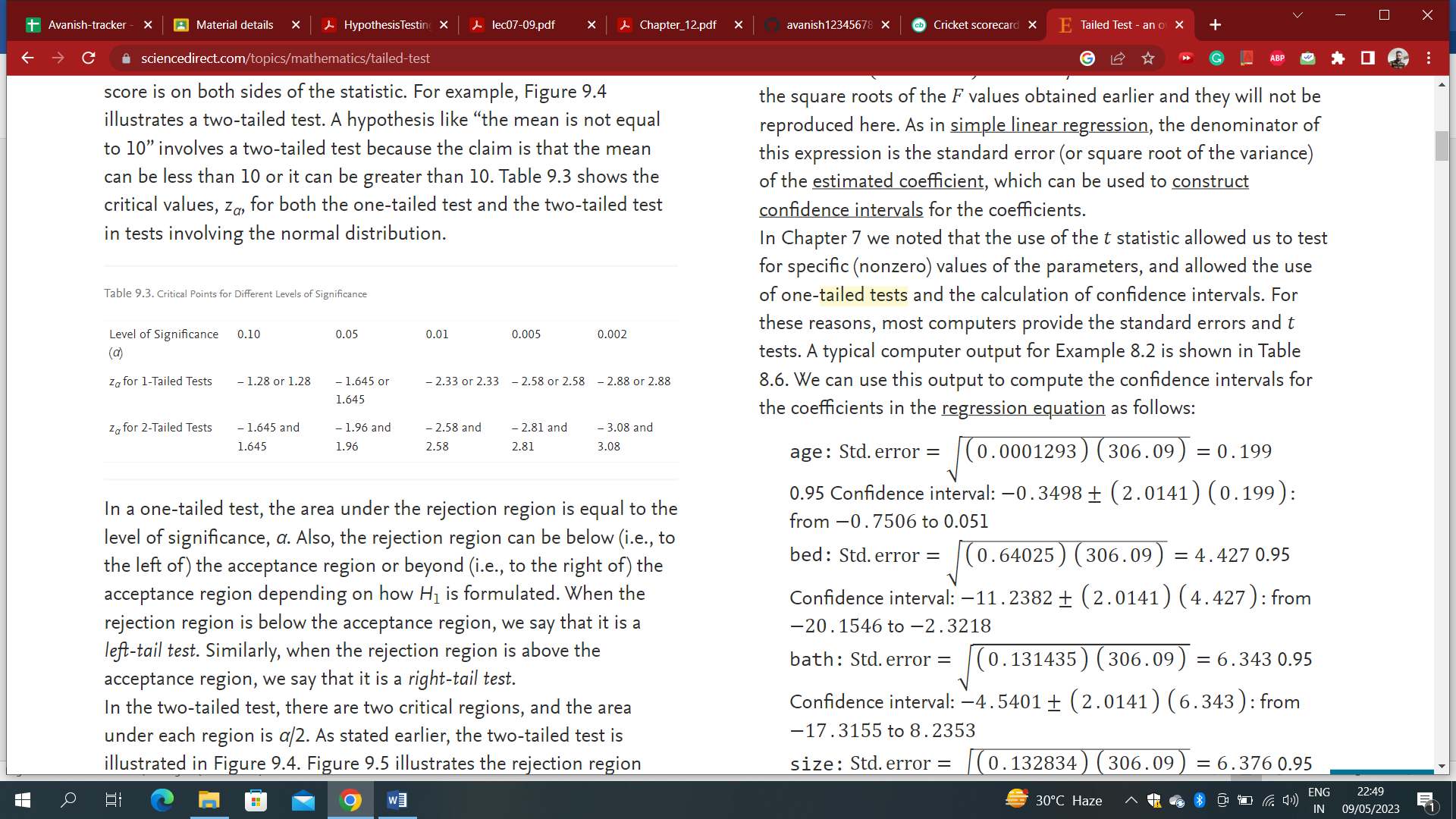
In this scenario, our null hypothesis states that the mean height would be equal to 64 while the alternate hypothesis states that the mean height would be greater than 64. A single-tail hypothesis test also means that we have only one critical region because we put the entire region of rejection into just one side of the distribution. When the alternative hypothesis is that the sample mean is greater, the critical region is on the right side of the distribution. When the alternative hypothesis is that the sample is smaller, the critical region is on the left side of the distribution (see below).



To calculate the critical regions, we must first find the critical values or the cut-offs where the critical regions start. To find these values, we use the critical values found specified by the z−distribution. These values can be found in a [z table](https://www.z-table.com/) that lists the areas of each of the tails under a normal distribution. Using this table, we find that for a 0.05 significance level, our critical values would fall at 1.96 standard errors above and below the mean. For a 0.01 significance level, our critical values would fall at 2.57 standard errors above and below the mean. Using the z−distribution we can find critical values (as specified by standard z scores) for any level of significance for either single-or two-tailed hypothesis tests.

Example: Determine the critical value for a single-tailed hypothesis test with a 0.05 significance level.

Using the z distribution table, we find that a significance level of 0.05 corresponds with a critical value of 1.645. If alternative hypothesis is the mean is greater than a specified value, the critical value would be 1.645. Due to the symmetry of the normal distribution, if the alternative hypothesis is the mean is less than a specified value the critical value would be -1.645.



**Calculating the Test Statistic**

Before evaluating our hypotheses by determining the critical region and calculating the test statistic, we need to confirm that the distribution is normal and determine the hypothesized mean μ of the distribution.

To evaluate the sample mean against the hypothesized population mean, we use the concept of z-scores to determine how different the two means are from each other. Based on the Central Limit theorem the distribution of X̄ is normal with mean, μ and standard deviation, σ/√n. The z score is calculated by using the formula:

z=(x̄−μ)/( σ/√n)

where:

z= standardized score

x̄= sample mean

μ= the population mean under the null hypothesis

σ= population standard deviation.

**Note:** If we do not have the population standard deviation and if n≥30, we can use the sample standard deviation, s. If n<30 and we do not have the population sample standard deviation we use a different distribution which will be discussed in a future lesson.

Once we calculate the z score, we can make a decision about whether to reject or to fail to reject the null hypothesis based on the critical values.

Following are the steps you must take when doing an hypothesis test:

1. Determine the null and alternative hypotheses.
2. Verify that necessary conditions are satisfied and summarize the data into a test statistic.
3. Determine the α level.
4. Determine the critical region(s).
5. Make a decision (Reject or fail to reject the null hypothesis)
6. Interpret the decision in the context of the problem.

Example: College A has an average SAT score of 1500. From a random sample of 125 freshman psychology students we find the average SAT score to be 1450 with a standard deviation of 100. We want to know if these freshman psychology students are representative of the overall population. What are our hypotheses and the test statistic?

1. Let’s first develop our null and alternative hypotheses:

H0: μ=1500

Ha: μ≠1500

2. The test statistic is z = (x̄−μ)/( σ/√n) = (1450−1500)/(100√125) ≈−5.59

3. Choose α = 0.05

4. This is a two sided test. If we choose α=.05, the critical values will be -1.96 and 1.96. That is we will reject the null hypothesis if the value of our test statistic is less than -1.96 or greater than 1.96.

5. The value of the test statistic is -5.59. This is less than -1.96 and so our decision is to reject H0.

6. Based on this sample we believe that the mean is not equal to 1500.

Example: A farmer is trying out a planting technique that he hopes will increase the yield on his pea plants. Over the last 5 years the average number of pods on one of his pea plants was 145 pods with a standard deviation of 100 pods. This year, after trying his new planting technique, he takes a random sample of 144 of his plants and finds the average number of pods to be 147. He wonders whether or not this is a statistically significant increase. What are his hypotheses and the test statistic?

1. First, we develop our null and alternative hypotheses:

H0: μ=145

Ha: μ>145

This alternative hypothesis is > since he believes that there might be a gain in the number of pods.

2. Next, we calculate the test statistic for the sample of pea plants.

z = (x̄−μ)/( σ/√n) = (147−145)100√144 ≈0.24

3. If we choose α=.05

4. The critical value will be 1.645. We will reject the null hypothesis if the test statistic is greater than 1.645. The value of the test statistic is 0.24.

5. This is less than 1.645 and so our decision is to accept H0.

6. Based on our sample we believe the mean is equal to 145.

**Finding the P-Value of an Event**

We can also evaluate a hypothesis by asking “what is the probability of obtaining the value of the test statistic we did if the null hypothesis is true?” This is called the **p−value**.

Example: Let’s use the example about the pea farmer. As we mentioned, the farmer is wondering if the number of pea pods per plant has gone up with his new planting technique and finds that out of a sample of 144 peas there is an average number of 147 pods per plant (compared to a previous average of 145 pods, the null hypothesis). To determine the p−value we ask what is P (z > .24)? That is, what is the probability of obtaining a z value greater than .24 if the null hypothesis is true? Using the z-score table, we find this probability to be .405. This indicates that there is a 40.5% chance that under the null hypothesis the peas will produce 147 or more pods.

•Reject H0 when P ≤ α

•Retain H0 when P > α

**Type I and Type II Errors**

When we decide to reject or not to reject the null hypothesis, we have four possible scenarios:

* The null hypothesis is true and we reject it.
* The null hypothesis is true and we do not reject it.
* The null hypothesis is false and we do not reject it.
* The null hypothesis is false and we reject it.

Two of these four possible scenarios lead to correct decisions: accepting the null hypothesis when it is true and rejections the null hypothesis when it is false.

Two of these four possible scenarios lead to errors: rejecting the null hypothesis when it is true and accepting the null hypothesis when it is false.

|  |  |  |
| --- | --- | --- |
|  | H0**is true** | H0**is false** |
| Accept H0 | Good Decision | Error (type II) |
| Reject H0 | Error (type I) | Good Decision |

Which type of error is more serious depends on the specific research situation, but ideally both types of errors should be minimized during the analysis.

The general approach to hypothesis testing focuses on the Type I error: rejecting the null hypothesis when it may be true. The level of significance, also known as the alpha level, is defined as the probability of making a Type I error when testing a null hypothesis. For example, at the 0.05 level, we know that the decision to reject the hypothesis may be incorrect 5 percent of the time.

α = P(rejecting H0|H0 is true) = P(making a type I error)

Calculating the probability of making a Type II error is not as straightforward as calculating the probability of making a Type I error. The probability of making a Type II error can only be determined when values have been specified for the alternative hypothesis. The probability of making a type II error is denoted by β.

β = P(accepting H0|H0 is false) = P(making a type II error)

Once the value for the alternative hypothesis has been specified, it is possible to determine the probability of making a correct decision (1−β). This quantity, **1−β**, is called the **power of the test**.

The goal in hypothesis testing is to minimize the potential of both Type I and Type II errors. However, there is a relationship between these two types of errors. As the level of significance or alpha level increases, the probability of making a Type II error (β) decreases and vice versa.

**Calculating the Power of a Test**

The power of a test is defined as the probability of rejecting the null hypothesis when it is false (that is, making the correct decision). Obviously, we want to maximize this power if we are concerned about making Type II errors. To determine the power of the test, there must be a specified value for the alternative hypothesis.

Example: Suppose that a doctor is concerned about making a Type II error only if the active ingredient in the new medication is greater than 3 milligrams higher than what was specified in the null hypothesis (say, 250 milligrams with a sample of 200 and a standard deviation of 50). Now we have values for both the null and the alternative hypotheses.

H0: μ=250

Ha: μ=253

By specifying a value for the alternative hypothesis, we have selected one of the many values for Ha. In determining the power of the test, we must assume that Ha is true and determine whether we would correctly reject the null hypothesis.

Calculating the exact value for the power of the test requires determining the area above the critical value set up to test the null hypothesis when it is re-centered around the alternative hypothesis. If we have an alpha level of .05 our critical value would be 1.645 for the one tailed test. Therefore,

z = (x̄−μ)/( σ/√n)

1.645=(x̄−250)/(50√200)

Solving for x̄ we find: x̄=1.645(50√200) + 250 ≈ 255.8

Now, with a new mean set at the alternative hypothesis Ha: μ=253 we want to find the value of the critical score when centered around this score when we center this x̄ around the population mean of the alternative hypothesis, μ=253. Therefore, we can figure that:

z = (x̄−μ)/( σ/√n) =(255.8−253)/ (50√200) ≈ 0.79

Recall that we reject the null hypothesis if the critical value is to the right of .79. The question now is what is the probability of rejecting the null hypothesis when, in fact, the alternative hypothesis is true? We need to find the area to the right of 0.79. You can find this area using a z table. The probability is .2148. This means that since we assumed the alternative hypothesis to be true, there is only a 21.5% chance of rejecting the null hypothesis. Thus, the power of the test is .2148. In other words, this test of the null hypothesis is not very powerful and has only a 0.2148 probability of detecting the real difference between the two hypothesized means.

There are several things that affect the power of a test including:

* Whether the alternative hypothesis is a single-tailed or two-tailed test.
* The level of significance α
* The sample size.

**T Test**

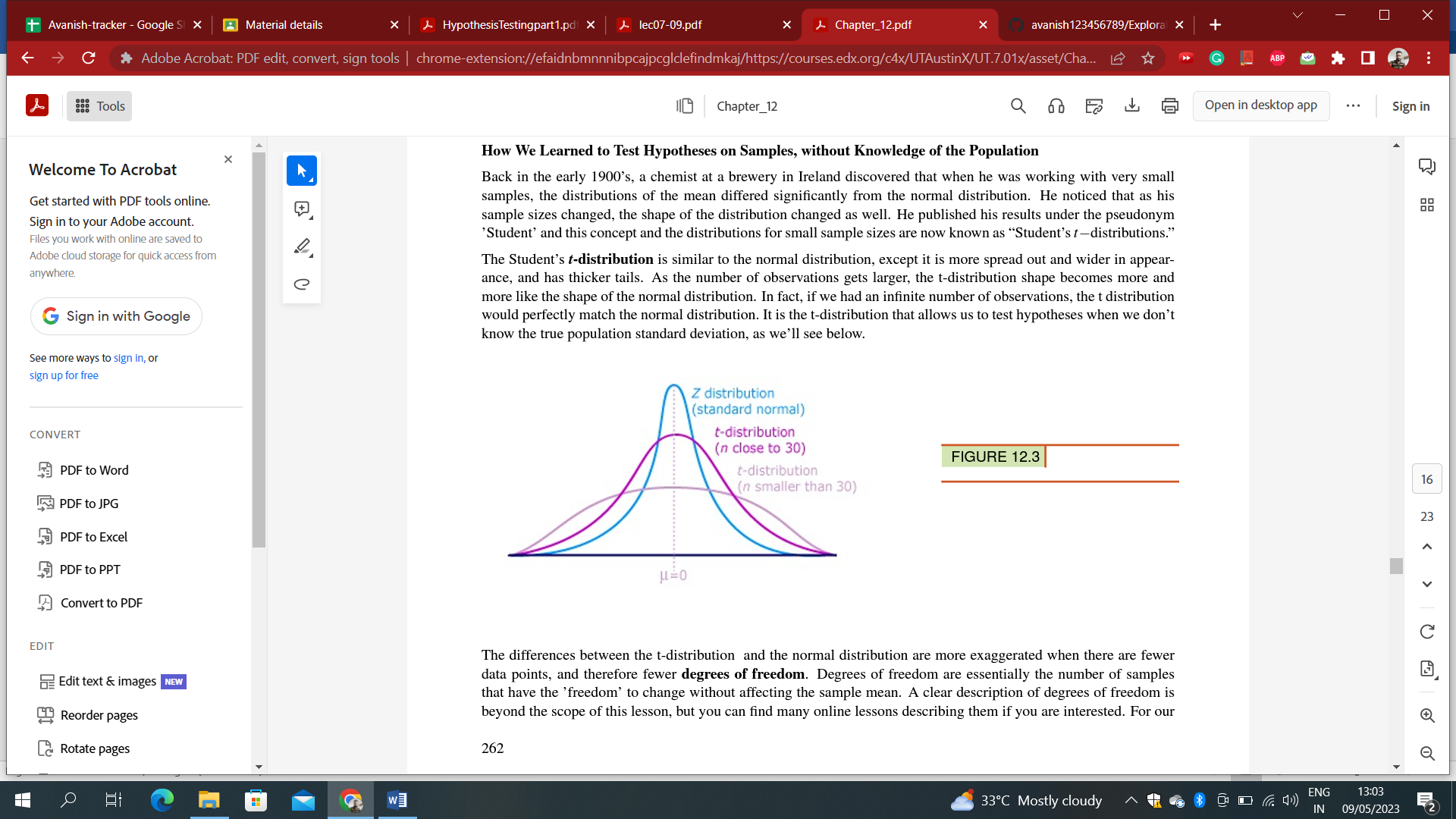
The z test conditions seldom hold in practice

–We don’t often know the population variance

–The sample size can be small

We use a T test instead –Assumes the sampling distribution is a t distribution.

The distributions for small sample sizes are known as t−distributions.



When we know the population standard deviation we use the normal distribution. When we don’t know the population standard deviation, we use the t-distribution. We use the t−distribution in hypothesis testing the same way that we use the normal distribution.

**Degrees of Freedom**

The differences between the t-distribution and the normal distribution are more exaggerated when there are fewer data points, and therefore fewer degrees of freedom. Degrees of freedom are essentially the number of samples that have the ’freedom’ to change without affecting the sample mean. For our purposes, all you really need to know about degrees of freedom is that there is always one less degree of freedom than the number of data points:

df = n −1

The reason you need to know how to find the number of degrees of freedom is quite simple: when you use a

t – distribution for a hypothesis test, there is a different critical value for each number of degrees of freedom. The larger your sample, the closer the critical value gets to the z-score for your alpha level. Below is the t-table. It shows you the critical value for the selected alpha level(across the top)for the degrees of freedom of your test. If you were conducting a two-tailed hypothesis test on a sample of 25students, your df = 24 and your critical value at α = 0.05 is ±2.064 because there is 0.025 in each tail.

**t−test statistic calculation**

t = (x̄−μ0)/( s/√n)

where:

t is the test statistic

x̄ is the sample mean

µ0 is the population mean under the null hypothesis

s is the sample standard deviation

n is the sample size

s/√n is the estimated standard error

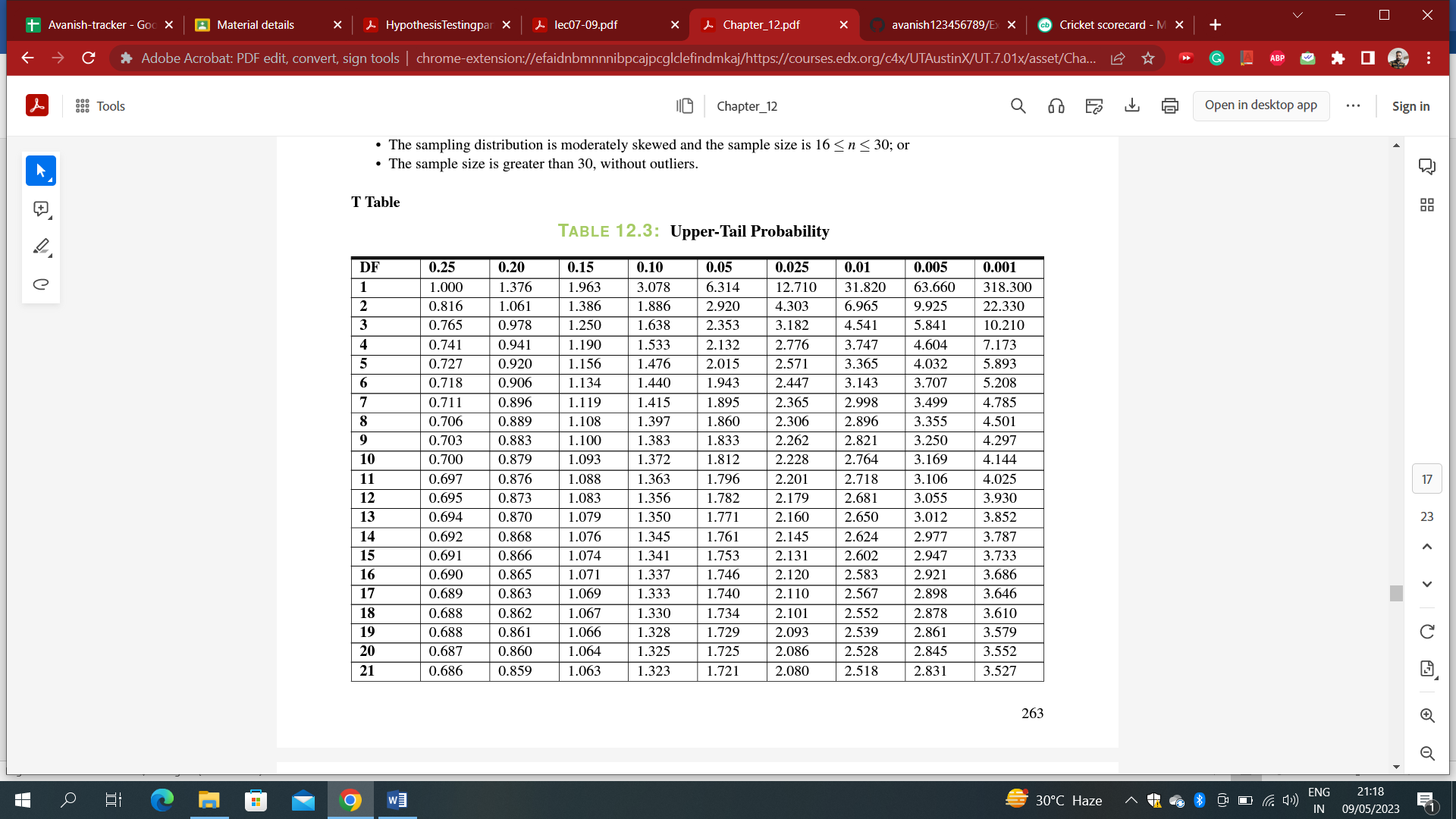
Assumptions of the single sample t-test:

• A random sample is used.

• The random sample is made up of independent observations

• The population distribution must be nearly normal, or the size of the sample is large.

T Table



Example

You have just taken ownership of a pizza shop. The previous owner told you that you would save money if you bought the mozzarella cheese in a 4.5pound slab. Each time you purchase a slab of cheese, you weigh it to ensure that you are receiving 72 ounces of cheese. The results of 7 random measurements are 70, 69, 73, 68, 71, 69 and 71 ounces. Are these differences due to chance or is the distributor giving you less cheese than you deserve?

a. State the hypotheses.

b. Calculate the test statistic.

c. Would the null hypothesis be rejected at the10% level? The 5% level? The 1% level?

Solution

a. For H0 the mean weight of cheese µ=72; and for Ha: µ=72.

b. Begin by determining the mean of the sample and the sample standard deviation.

x̄ = 70.143 and s = 1.676.

t = (x̄−μ0)/( s/√n) = (70.143−72)/( 1.676√7) t ≈ −2.9315

c. The test statistic computed in part b) was -2.9315.

d. The null hypothesis would be rejected at the .10 and the .05 levels, but not at the .01level.

Example

Duracell manufactures batteries that the CEO claims will last an average of 300 hours under normal use. A researcher randomly selected 20 batteries from the production line and tested these batteries. The tested batteries had a mean lifespan of 270 hours with a standard deviation of 50 hours. Do we have enough evidence to suggest that the claim of an average lifetime of 300 hours is false?

• Hypothesis Step 1: Clearly state the Null and Alternative Hypothesis

H0 : µ = 300

Ha : µ ≠ 300

• Hypothesis Step 2: Identify the appropriate significance level and confirm the test assumptions.

We’ll use the standard significance level of 0.05, and we assume a normal population distribution.

• Hypothesis Step 3: Analyze the data and compute the test statistic.

First, we need to calculate the Standard Error:

SE(x̄) = s√n

SE(x̄) = 50√20 = 11.18

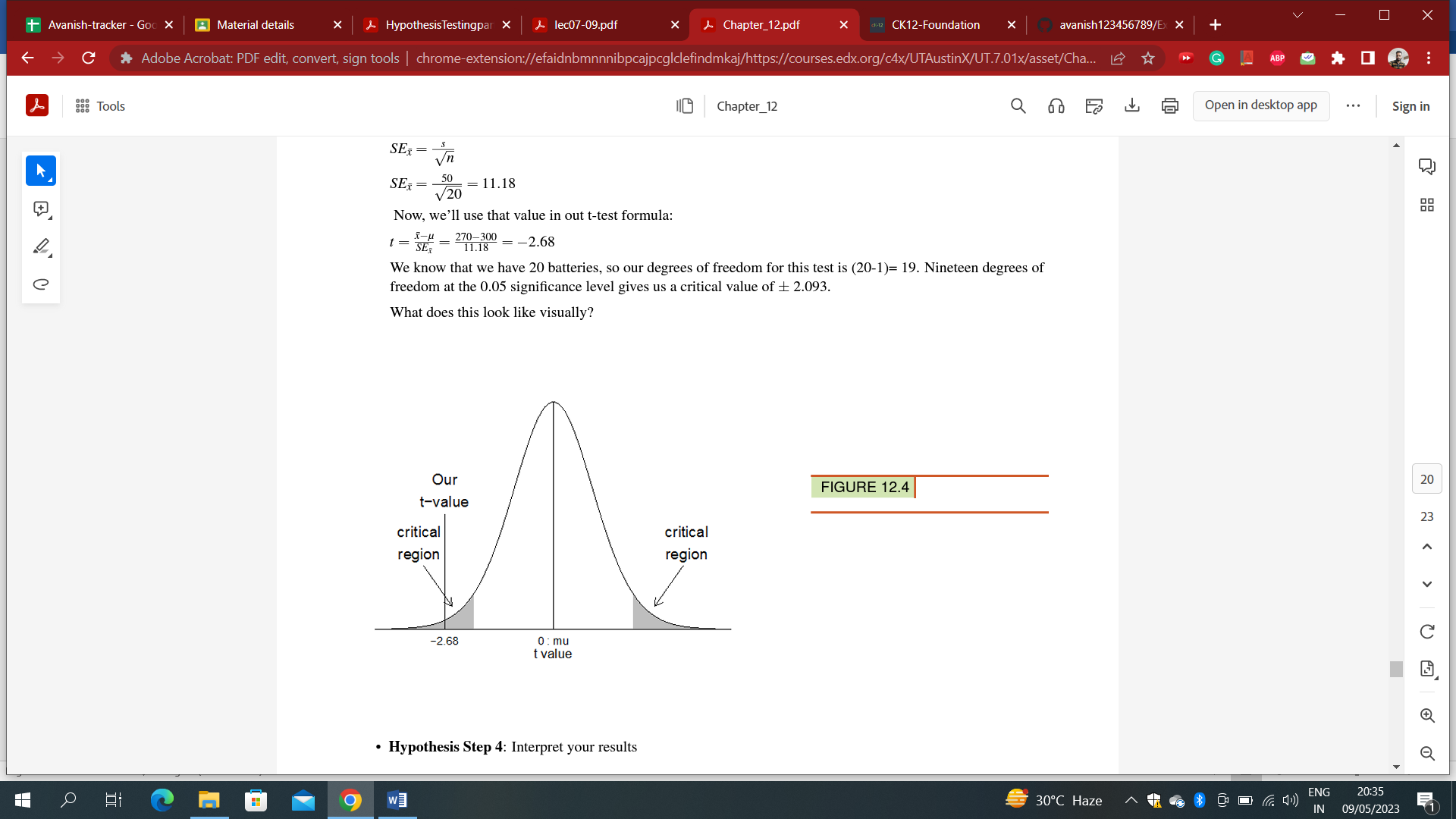
Now, we’ll use that value in out t-test formula:

t = (x̄−μ0)/( s/√n) = (270−300)/11.18 = −2.68

We know that we have 20 batteries, so our degrees of freedom for this test is (20-1) = 19.

Nineteen degrees of freedom at the 0.05 significance level gives us a critical value of ±2.093.

What does this look like visually?



• Hypothesis Step 4: Interpret your results Since our calculated t-test value is outside of our t-critical value–it lies in the critical region – we reject the Null Hypothesis. The average battery life of the sample is significantly different from the average battery life claim by the CEO.

**Margin of Error**

When a hypothesis is rejected, it is often useful to turn to estimation to try to capture the true value of the population mean. Now that we have rejected the claim that the average life time of a battery is 300 hours, we want to know how long these batteries do last. Let’s take a look at how we can use estimation to answer this question. We will choose a 95% confidence interval. This is essentially the same degree of confidence that we had in our t-test, since confidence =1-α.

To calculate the confidence interval, we need to know three things: the mean of our sample, the standard error, and the critical value.

confidence interval = x̄ ± Margin of Error

Where the Margin of Error is defined as:

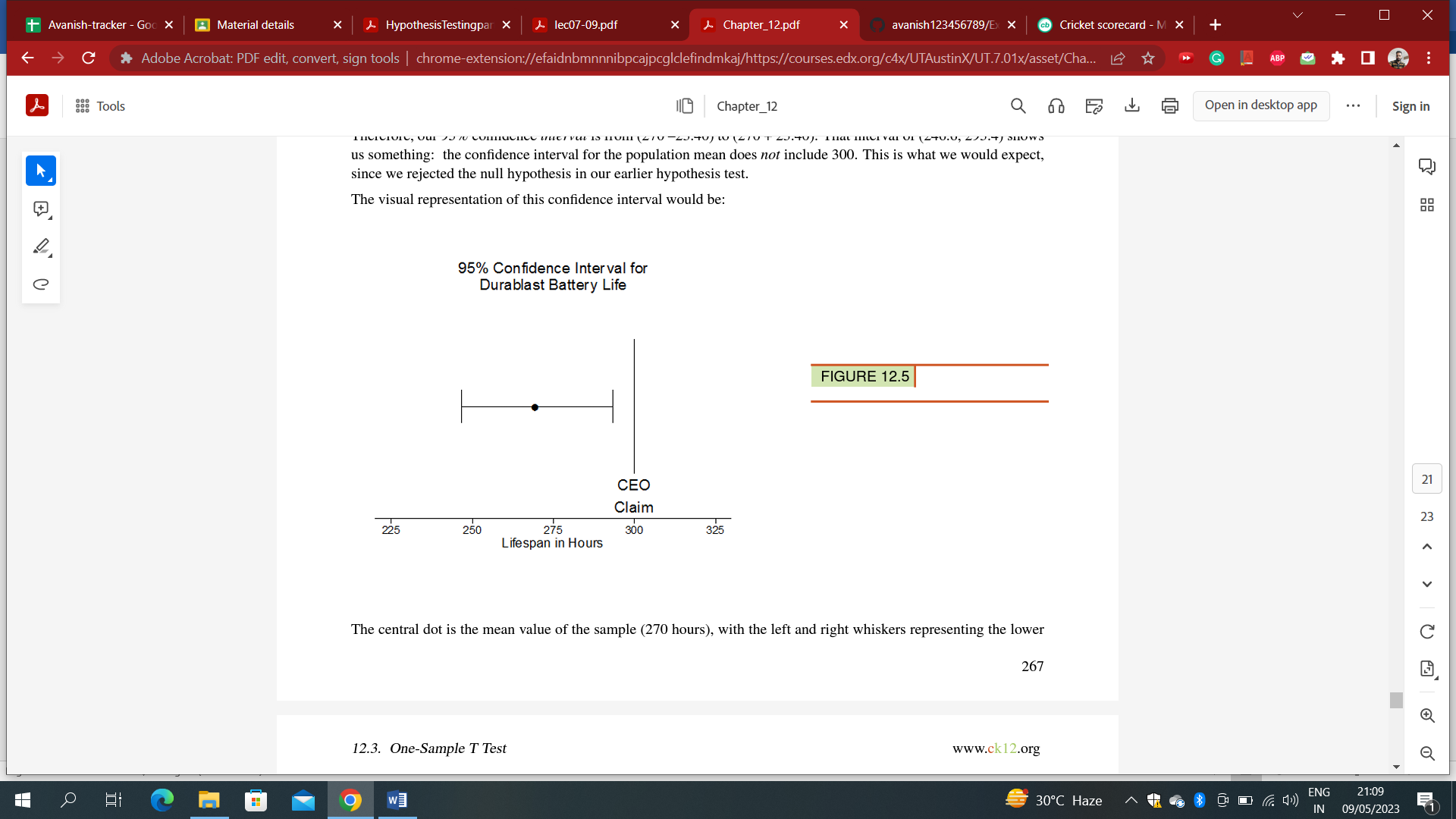
ME = t-critical\*SE(x̄)

Note that we are using the t-critical value (rather than z) because we are working only with sample data.

We do not know the true population mean or standard deviation. We have to use both our sample mean and our sample standard error to create the confidence interval.

So, for our above Durablast example, the Margin of Error would be (remember we have 19 degrees of freedom): ME = 2.093\*11.18 = 23.40

Therefore, our 95% confidence interval is from (270–23.40) to (270+23.40). That interval of (246.6,293.4) shows us something: the confidence interval for the population mean does not include 300.



This is what we would expect, since we rejected the null hypothesis in our earlier hypothesis test. The visual representation of this confidence interval would be: The central dot is the mean value of the sample(270hours), with the left and right whiskers representing the lower and upper values of the 95% confidence interval. If we were to interpret the 95% confidence interval we would say: “I am 95% confident that the true population mean of Durablast battery lifespan is between 246.6 and 293.4 hours.”

Also, ***Margin of Error = t\* √((p \* (1 – p)) / n)***

where:

p = sample proportion.

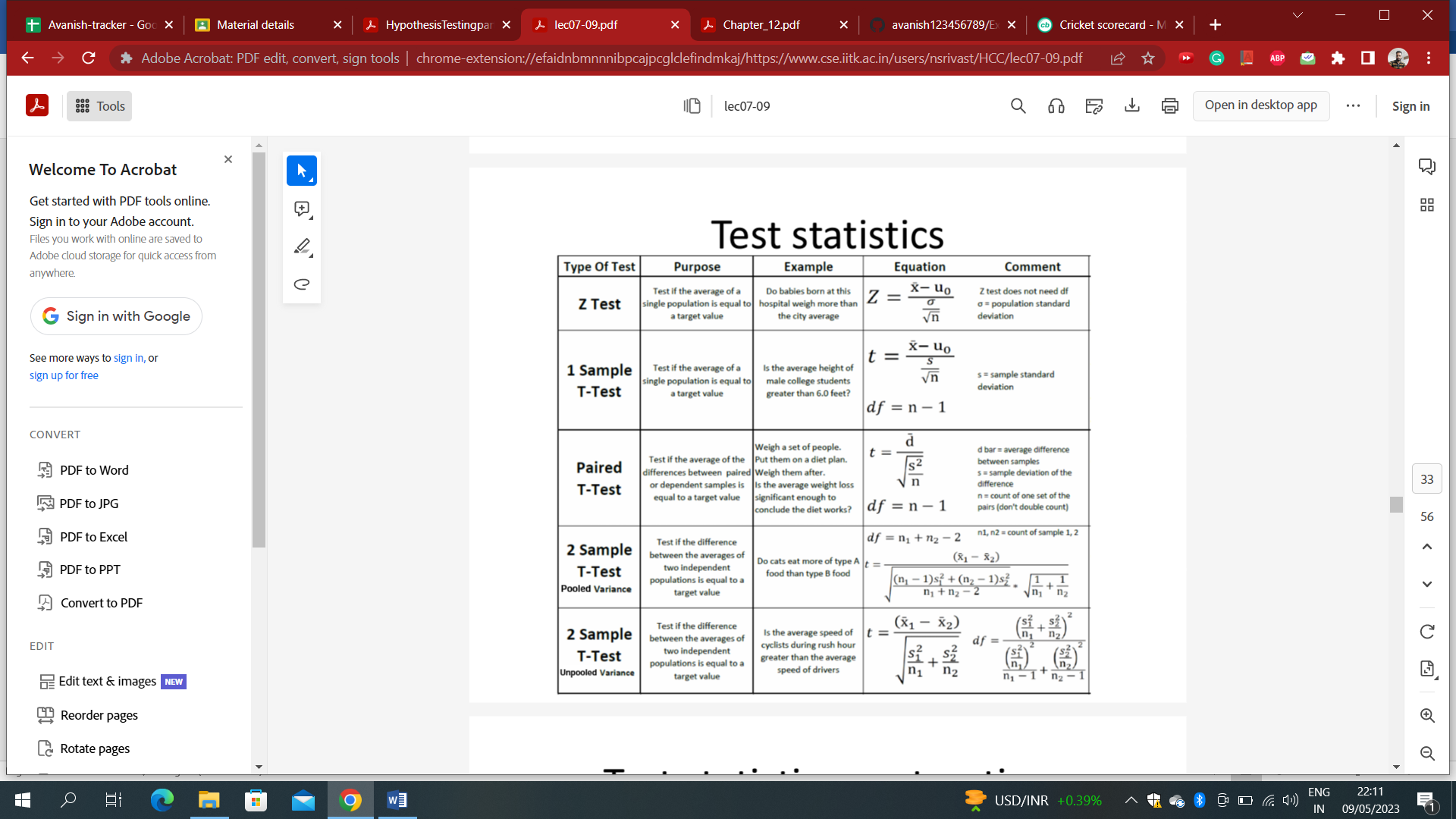
#### Example

**Let say you are launching a new health product in the market but you are confused which flavor will people like. You are confused between banana flavor and vanilla flavor and have decided to conduct a survey. Your population for that is 500,000 which is your target market and out of that you decided to ask the opinion of 1000 people and that will the sample. Assume that a confidence interval is 90%. Calculate the margin of error.**

**Solution:**

Once the survey is done, you came to know that 470 people liked the banana flavor and 530 has asked for vanilla flavor.

Margin of Error = 1.65 \* √((0.47 \* (1 – 0.47)) / 1000) = 2.60%



**Miscellaneous**

Example

Piston rings are mass-produced. The target internal diameter is 45 mm but records showthat the diameters are normally distributed with mean 45mm and standard deviation 0.05mm. An acceptable diameter is one within the range 44.95 mm to 45.05 mm. What proportion of the output is acceptable?

Upper limit: z = (45.05-45.00)/0.05 = 1

Lower limit: z = (44.95-45)/0.05 = -1

P(-1< z <1) = P(z = 1) – P(z = -1) = 0.8413-0.1587 = 0.68 = 68%(acceptable)

Example

A sample poll of 300 voters from district A and 200 voters from District B showed that 56% and 48% respectively were in favor of a given candidate. Using α = .05, test whether (a)the two districts significantly differ in their support of the candidate, (b)the candidate is preferred in district A.

P1 = 0.56, P2 =0.48

n1 = 300, n2 = 200

Hypothesis: H0: P1 = P2 , Ha : P1 ≠ P2

pooled P = (n1\*P1 + n2\*P2)/(n1 +n2) = 0.528

test statistic; Z = (P1-P2)/ √(P\*(1-P)\*(1/n1 + 1/n2)) = -1.7556

since, it is two tailed, |Z| = 1.7556

critical value(Zα/2) = Z0.025 = 1.96

So, fail to reject null hypothesis and thus P1 = P2

(b)If we wish to determine whether the candidate is preferred in district A, we must decide between the hypotheses (H0: p1=p2) and (H1: p1> p2), which involves a one-tailed test. On the basis of a one-tailed test at a 0.05 level of significance, we would reject H0 if Z were greater than 1.645. Since this is the case, we can reject H0at this level and conclude that the candidate is preferred in district A.

Example

A botanist has produced a new variety of hybrid soy plant that is better able to withstand drought than other varieties. The botanist knows the seed germination for the parent plants is 75%, but does not know the seed germination for the new hybrid. He tests the claim that it is different from the parent plants. To test this claim, 450 seeds from the hybrid plant are tested and 321 have germinated. Use a 5% level of significance to test this claim that the germination rate is different from 75%.

H0 : P = 0.75, Ha : P≠ 0.75

Z = (p̂ - P)/√(P(1-P)/n) = (0.713-0.75)/ √((0.75\*0.25)/450) = -1.81 { p̂ = 321/450 = 0.713

critical value(Zα/2) = Z0.025 = -1.96

For null hypothesis rejection, Z< Zα/2

So, fail to reject H0

**Factors affecting P- values**:

The p-value is a statistical measure used in hypothesis testing that quantifies the strength of evidence against a null hypothesis. It represents the probability of obtaining results as extreme as, or more extreme than, the observed data, assuming the null hypothesis is true. Let's examine how the p-value is affected by various factors:

1. Distribution is non-normal: The p-value is robust to violations of the normality assumption, particularly when the sample size is large. However, if the sample size is small and the distribution significantly deviates from normality, the p-value may be less reliable. In such cases, alternative non-parametric tests or transformations of the data may be considered.
2. Increase the sample size: As the sample size increases, the p-value tends to become more accurate and reliable. With a larger sample, the estimate of the underlying distribution improves, leading to more precise results. This increased precision may result in a smaller p-value if the observed effect is genuine and statistically significant.
3. Increase the number of tests: When conducting multiple tests, such as in multiple comparisons or hypothesis testing on multiple variables, it is important to consider multiple testing adjustments (e.g., Bonferroni correction) to control the overall false positive rate. Increasing the number of tests without adjusting for multiple testing increases the probability of obtaining at least one significant result by chance, leading to an inflated false positive rate.
4. Dealing with data outliers: Outliers can significantly affect the p-value, especially in small sample sizes. Outliers can either increase or decrease the p-value, depending on the specific circumstances. If the outliers are genuine extreme observations caused by the underlying phenomenon, they can strengthen the evidence against the null hypothesis, leading to a smaller p-value. Conversely, if the outliers are spurious or measurement errors, they can distort the results and potentially invalidate the conclusions. It is crucial to carefully assess the nature and impact of outliers and consider robust statistical methods that are less sensitive to extreme values.

In summary, the p-value can be influenced by various factors, including the distributional assumptions, sample size, number of tests conducted, and the presence of outliers. Understanding these influences and choosing appropriate statistical techniques are essential for accurate hypothesis testing and interpretation of results.

Top of Form