**The p-value**

The **p-value** (short for probability value) is the most critically viewed number in statistics. It is defined as the probability of receiving a result at least as extreme as the one observed, given that the null hypothesis is true. My favourite resource of describing the p-value in simple words is by [Cassie Kozyrkov](https://www.youtube.com/watch?v=9jW9G8MO4PQ):

*The p-value tells you, given the evidence that you have (data), if the null hypothesis looks ridiculous or not […] The lower the p-value, the more ridiculous the null hypothesis looks.*

The p-value is a value between 0% and 100% and can be retrieved from the null hypothesis, sampling distribution, and the data. Generally, it is calculated with the help of statistical software or reading off a distribution table with set parameters (degrees of freedom, alpha level etc.). Distribution tables with the most common parameters can be found online for most test statistics, like [t-score](https://www.statisticshowto.com/tables/t-distribution-table/), [chi-squared score](https://www.itl.nist.gov/div898/handbook/eda/section3/eda3674.htm), or [Wilcoxon-rank-sum](http://www.socr.ucla.edu/Applets.dir/WilcoxonRankSumTable.html).

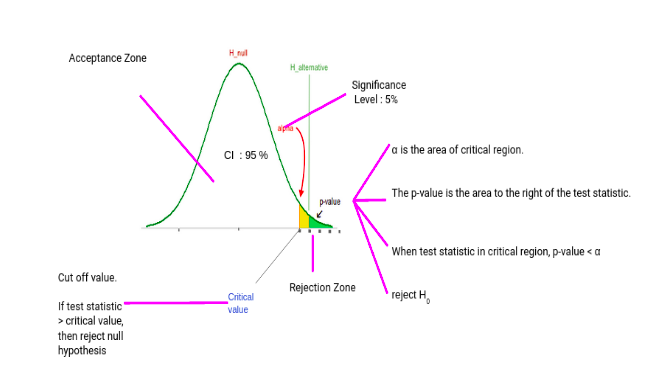
Critical Value (p-value)

We will understand the logic of Hypothesis Testing with the graphical representation for Normal Distribution.

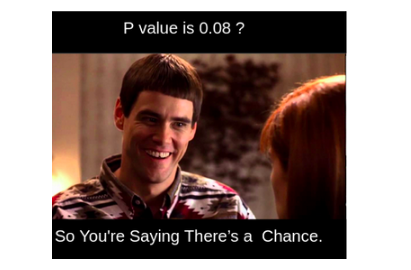
Typically, we set the Significance level at 10%, 5%, or 1%. If our test score lies in the Acceptance Zone we fail to reject the Null Hypothesis. If our test score lies in the critical zone, we reject the Null Hypothesis and accept the Alternate Hypothesis.

Critical Value is the cut off value between Acceptance Zone and Rejection Zone. We compare our test score to the critical value and if the test score is greater than the critical value, that means our test score lies in the Rejection Zone and we reject the Null Hypothesis. On the opposite side, if the test score is less than the Critical Value, that means the test score lies in the Acceptance Zone and we fail to reject the null Hypothesis.

But why do we need a p-value when we can reject/accept hypotheses based on test scores and critical values?



p-value has the benefit that we only need one value to make a decision about the hypothesis. We don’t need to compute two different values like critical values and test scores. Another benefit of using a p-value is that we can test at any desired level of significance by comparing this directly with the significance level.



This way we don’t need to compute test scores and critical values for each significance level. We can get the p-value and directly compare it with the significance level.

The p-value is the probability that, given the null hypothesis is true, we observe a result at least as extreme as the test statistic. If the p-value is low (generally below a predetermined significance level like 0.05), we reject the null hypothesis in favour of the alternative.

Let’s illustrate with an example:

# Importing necessary libraries  
import numpy as np  
from scipy.stats import ttest\_1samp  
  
# Let's assume that we have a sample data from an experiment  
data\_sample = np.random.normal(loc=15, scale=5, size=50)  
  
# We hypothesize that the population mean is 20  
pop\_mean = 20  
  
# Perform one-sample t-test  
t\_statistic, p\_value = ttest\_1samp(data\_sample, pop\_mean)  
  
print("T-statistic: ", t\_statistic)  
print("p-value: ", p\_value)  
  
# Result  
# T-statistic: -6.716038599410901  
# p-value: 1.8271420699335366e-08

Let’s break it down.

T-statistic: The T-statistic or T-score is a ratio of the departure of the estimated value of a parameter from its hypothesized value to its standard error. The T-statistic in our example is negative, indicating that the sample's observed mean is less than the hypothesized population mean. The magnitude (6.716) is the difference's size relative to the data's variance. In this case, we are using a **one-sample t-test**.

P-value: The p-value is a probability that measures the evidence against the null hypothesis. Smaller p-values provide stronger evidence against the null hypothesis. The extremely small p-value (1.8271420699335366e-08) in our example means there is a very small probability that we would have observed such an extreme test statistic if the null hypothesis were true.

Given a significance level of 0.05, we can reject the null hypothesis (that the sample mean equals the hypothesized population mean of 20), as our p-value is less than the significance level. This suggests that the population mean is likely different from 20.

**1. Non-Normal Distribution**

* **Effect on p-value:** In many statistical tests (e.g., t-test, ANOVA), non-normality can lead to unreliable p-values because these tests assume normality. If the distribution is skewed or has heavy tails, the p-value might not reflect the true significance, often resulting in incorrect conclusions.
* **Example:** Let's check how a t-test behaves with a non-normal distribution using Python.

import numpy as np

from scipy import stats

# Generate normal and non-normal distributions

np.random.seed(0)

normal\_data = np.random.normal(0, 1, 100)

non\_normal\_data = np.random.exponential(1, 100)

# Perform t-test

t\_stat, p\_value\_normal = stats.ttest\_1samp(normal\_data, 0)

t\_stat\_non, p\_value\_non\_normal = stats.ttest\_1samp(non\_normal\_data, 0)

print(f'P-value (Normal distribution): {p\_value\_normal}')

print(f'P-value (Non-normal distribution): {p\_value\_non\_normal}')

* **Conclusion:** The p-value might be more unreliable for non-normal data, as seen when comparing p-values from different distributions.

**2. Increasing the Sample Size**

* **Effect on p-value:** Increasing the sample size generally makes statistical tests more powerful. With larger samples, small differences are easier to detect, which can lead to smaller p-values, even if the effect size is small.

np.random.seed(0)

# Small sample size

small\_sample = np.random.normal(0, 1, 30)

t\_stat\_small, p\_value\_small = stats.ttest\_1samp(small\_sample, 0)

# Large sample size

large\_sample = np.random.normal(0, 1, 3000)

t\_stat\_large, p\_value\_large = stats.ttest\_1samp(large\_sample, 0)

print(f'P-value (Small sample size): {p\_value\_small}')

print(f'P-value (Large sample size): {p\_value\_large}')

**Conclusion:** As the sample size increases, the p-value tends to decrease, even when the mean difference is the same.

**3. Increasing the Number of Tests**

* **Effect on p-value:** When performing multiple tests, the probability of obtaining a significant p-value (false positive) by chance increases. This leads to the "multiple testing problem." A common approach to control this is the Bonferroni correction.
* **Example:** Simulating multiple t-tests.

p\_values = []

for i in range(50): # Performing 50 t-tests

sample = np.random.normal(0, 1, 30)

t\_stat, p\_val = stats.ttest\_1samp(sample, 0)

p\_values.append(p\_val)

# Applying Bonferroni correction

corrected\_p\_values = np.array(p\_values) \* 50 # Adjust p-values

print(f'Original p-values: {p\_values}')

print(f'Corrected p-values: {corrected\_p\_values}')

**Conclusion:** Increasing the number of tests inflates the chance of false positives, but techniques like Bonferroni correction can adjust the p-values.

**4. Dealing with Data with Outliers**

* **Effect on p-value:** Outliers can heavily influence the p-value by inflating the variability and skewing the results, potentially leading to inaccurate conclusions. Removing or adjusting for outliers (e.g., using robust statistical methods) is essential.
* **Example:** Compare p-values with and without outliers.

# Data with an outlier

data\_with\_outliers = np.append(np.random.normal(0, 1, 100), [10, -10])

data\_without\_outliers = np.random.normal(0, 1, 100)

# Perform t-tests

t\_stat\_outliers, p\_value\_outliers = stats.ttest\_1samp(data\_with\_outliers, 0)

t\_stat\_no\_outliers, p\_value\_no\_outliers = stats.ttest\_1samp(data\_without\_outliers, 0)

print(f'P-value (With outliers): {p\_value\_outliers}')

print(f'P-value (Without outliers): {p\_value\_no\_outliers}')

**Conclusion:** Outliers can significantly affect the p-value by increasing variability and potentially hiding true patterns.