Statistical Significance

WE'LL COVER THE FOLLOWING

What Is Statistical Significance?

Statistics is guesswork based on mathematics. It is not an exact science, so when dealing with statistical results, we need to know how close our *guess* is to reality. When someone claims that some data proves their point, we can't just accept it, as if all the juggling with complex statistics led to results that can't be questioned! We need to use statistical significance to reach conclusions instead. **Statistical significance is a measure of whether our findings are meaningful or just a result of random chance.**

As with most skills and concepts in life, breaking things down into sub-skills or basic components is a great way to approach learning. In this lesson we are going to chunk statistical significance into its base components and then put all the pieces together to understand this concept in an intuitive bottom-up approach.

Components of Statistical Significance

Statistical significance can be broken down into three base components:

- Hypothesis Testing

- Normal Distribution
- P-values

We will first understand these three components theoretically and then we will put it all together with the help of a practical example.

1. Hypothesis Testing

Hypothesis testing is a technique for evaluating a theory using data. The hypothesis is the researcher's initial belief about the situation before the study. The commonly accepted fact is known as the **null hypothesis** while the opposite is the **alternate hypothesis**. The researcher's task is to reject, nullify, or disprove the null hypothesis. In fact, the word "null" is meant to imply that it's a commonly accepted fact that researchers work to nullify (zero effect).

For example, if we consider a study about cell phones and cancer risk, we might have the following hypothesis:

- **Null hypothesis:** *"Cell phones have no effect on cancer risk."*
- **Alternative hypothesis** (the one under investigation): *"Cell phones affect the risk of cancer."*

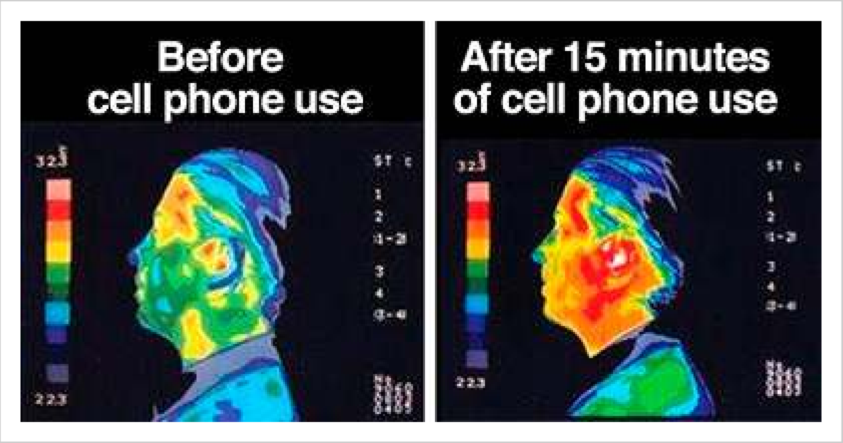


Image Credits: Penn State's SC200 course site, https://sites.psu.edu/siowfa15/2015/09/30/can-cell-phone-usage-cause-cancer/

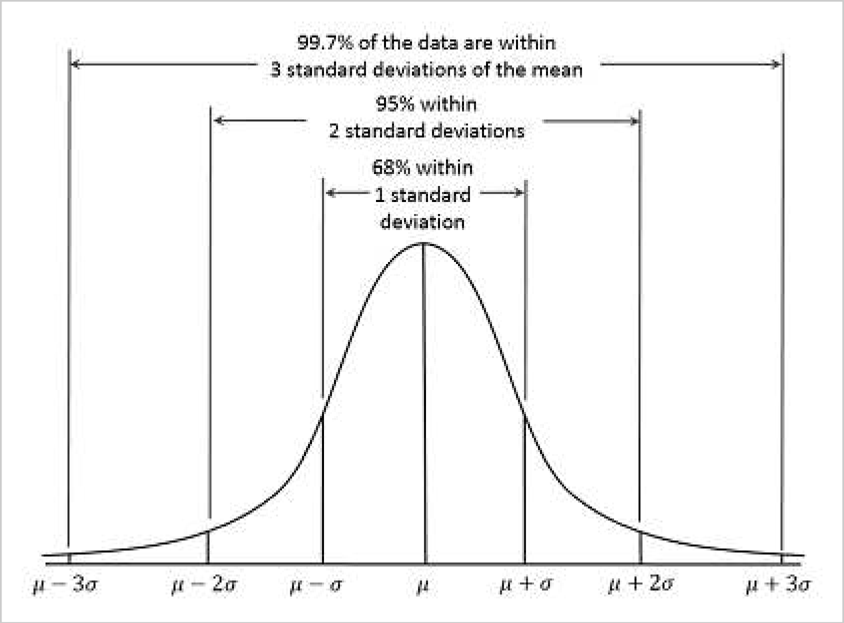
These studies can be anything from a medical trial to a study evaluating customer retention. The common goal among these studies is to determine which of the two hypotheses is better supported by the evidence found from our data. This means that we need to be able to test these hypotheses— the *testing* part of hypothesis testing. How do we do test?

There are many hypothesis tests that work by making comparisons either between two groups or between one group and the entire population. We are going to look at the most commonly used **z-test**.

Does *z-test* ring any bells? In the previous lessons, we came across the concept of z-scores while learning about normal distributions. Remember?? The second building block of statistical significance is built upon normal distributions and z-scores. *If you need a refresher, before continuing further, revisit the section on normal distributions.*

2. Normal Distribution

As we learned earlier, the normal distribution is used to represent the distribution of our data and it is defined by the mean, μ (center of the data), and the standard deviation, σ (spread in the data). These are two important measures because any point in the data can then be represented in terms of its standard deviation from the mean:


99.7% of the data are within 3 standard deviations of the mean
95% within 2 standard deviations
68% within 1 standard deviation
$\mu - 3\sigma$
$\mu - 2\sigma$
$\mu - \sigma$
μ
$\mu + \sigma$
$\mu + 2\sigma$
$\mu + 3\sigma$

For the normal distribution, the values less than one standard deviation away from the mean account for 68% of the set; while two standard deviations from the mean account for 95%; and three standard deviations account for 99.7%. Image Credits: Wikipedia

Standardizing the results by using z-scores where you subtract the mean from the data point and divide by the standard deviation gives us the standard normal distribution.

From z-score to z-test: A z-test is a statistical technique to test the Null

From z-score to z-test: A z-test is a statistical technique to test the Null Hypothesis against the Alternate Hypothesis. This technique is used when the

sample data is normally distributed and the population size is greater than 30. Why 30?

According to the **Central Limit Theorem** as the sample size grows and number of data points exceeds 30, the samples are considered to be normally distributed. So whenever sample size exceeds 30, we assume data is normally distributed and we can use the z-test.

As the name implies, z-tests are based on z-scores, which tell us where the sample mean lies compared to the population mean:

$$z = \frac{\overline{x}-\mu}{\frac{\sigma}{\sqrt{n}}}$$

where,
$\bar{x}$: mean of sample,
μ: mean of population,
σ: standard deviation of the population,
n: number of observations

z-scores on the extremes – higher end or lower end – indicate that our result is meaningful because it is less likely to have occurred just by chance.

But what determines *how high the high should be and how low the low should be* in order for us to accept the results as meaningful?

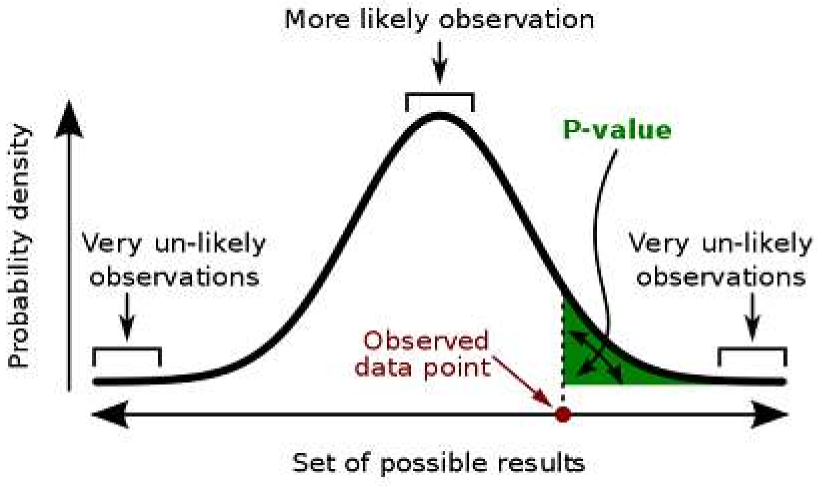
To quantify the *meaningfulness* of our results, we need to understand the third component of statistical significance, p-values.

3. P-value

The p-value quantifies the rareness in our results. It tells us how often we'd see the numerical results of an experiment (our z-scores) if the null hypothesis is true and there are no differences between the groups. This means that we can use p-values to reach conclusions in significance testing.

More specifically, we compare the p-value to a **significance level** α to make conclusions about our hypotheses:

If the p-value is very small or lower than the significance level we chose, it means the numbers would rarely occur by chance alone, and we can reject the null hypothesis in favor of the alternative hypothesis. On the other hand, if the **p-value is greater than or equal to the significance level**, then we fail to reject the null hypothesis. This doesn't mean we accept the null hypothesis though!



A p-value (shaded green area) is the probability of an observed (or more extreme) result assuming that the null hypothesis is true.

Image Credits: Wikipedia

But where does this a come from?

Although the choice of *α* depends on the situation, 0.05 is the most widely used value across all scientific disciplines. This means that ***p<.05*** is the threshold beyond which study results can be declared to be *statistically significant*, i.e., it's unlikely the results were a result of random chance. If we run the experiment 100 times, we'd see these same numbers, or more extreme results, 5 times, assuming the null hypothesis is true.

Again, a p-value of less than .05 means that there is **less than a 5% chance of seeing our results, or more extreme results, in the world where the null hypothesis is true**.

Note that p<.05 does not mean there's less than a 5% chance that our

Note that p<.05 does not mean there's less than a 5% chance that our experimental results are due to random chance. The false-positive rate for experiments can be much higher than 5%!

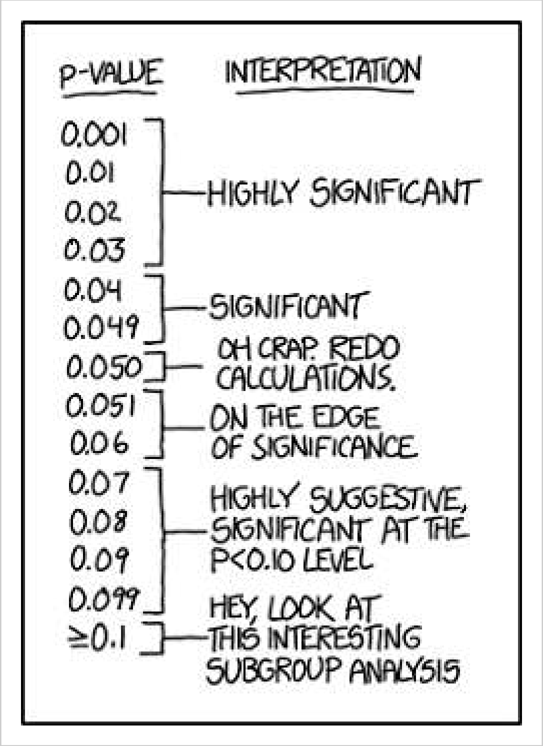

P-VALUE
INTERPRETATION
0.001
0.01
0.02
0.03
HIGHLY SIGNIFICANT
0.04
0.049
SIGNIFICANT
0.050
OH CRAP. REDO CALCULATIONS.
0.051
0.06
ON THE EDGE OF SIGNIFICANCE
0.07
0.08
0.09
0.099
HIGHLY SUGGESTIVE, SIGNIFICANT AT THE P<0.10 LEVEL
≥0.1
HEY, LOOK AT THIS INTERESTING SUBGROUP ANALYSIS

Image Credits: https://xkcd.com/

> **Note**: Since this is a tricky concept that most get wrong but is important to understand it well, again: p-value doesn't necessarily tell us if our experiment was a success or not, it doesn't prove anything! It just gives us the probability that a result at least as extreme as that observed would have occurred if the null hypothesis is true. The lower the p-value, the more significant the result because it is less likely to be caused by noise.

From z-score to p-value: The z-score is called our test-statistic. Once we have a test-statistic, we can either use the old-fashioned approach of looking at tables or use any programming language of our choice to convert z-scores into p-values. For example, in Python, we can use SciPy library's `scipy.stats` module that provides many handy statistical functions.

Now, putting it all together; if the observed p-value is lower than the chosen threshold α, then we conclude that the **result is statistically significant**.

As a final note, an important take away is that at the end of the day calculating p-values is not the hardest part here! **The real deal is to interpret the p-**

p-values is not the hardest part here! **The real deal is to interpret the p-values so that we can reach sensible conclusions**. *Does 0.05 work as the threshold for your study or should you use 0.01 to reach any conclusions instead? And what is our p-value really telling us?*

Example

Let's put all the pieces together by looking at an example from start to finish.

A company claims that it has a high hiring bar which is reflected in its employees having an IQ above the average. Say a random sample of their 40 employees has a mean IQ score of 115. Is this sufficient evidence to support the company's claim given the mean population IQ is 100 with a standard deviation of 15?

1. **State the Null hypothesis**: the accepted fact is that the population mean is 100 — $H_0 : \mu = 100$.

2. **State the Alternate Hypothesis**: the claim is that the employees have above average IQ scores — $H_1: \mu > 100$.

3. **State the threshold for the p-value – α level**: we will stick with the most widely used value of 0.05.

4. **Find the test statistic** using this formula for the z-test:

$$z = \frac{115 - 100}{15/\sqrt{40}} = 6.32$$

The company mean score is 115, which is 6.32 standard error units from the population mean of 100.

5. **Get the p-value from the z-score**: Using an online calculator for converting z-scores to p-values, we see that the probability of observing a standard normal value below 6.32 is < .00001.

6. **Interpret the p-value**: our result is significant at $p < 0.05$, so we can reject the null hypothesis — the 40 employees of interest have an unusually higher IQ score compared to random samples of similar size from the entire population.

Final Thoughts

Final Thoughts

There were quite a few concepts in this lesson. To make sure that our understanding is crystal clear, we will engrain these concepts with some **exercises**.

Also, this was the last lesson on Statistics, so well-done for having come this far 🙌 Let’s keep going, there are fun Machine Learning lessons awaiting us ahead!