**3 Most Common Misinterpretations: Hypothesis Testing, Confidence Interval, P-Value**

A refresher on how to interpret statistical inference correctly



Photo by [Scott Graham](https://unsplash.com/@homajob?utm_source=medium&utm_medium=referral) on [Unsplash](https://unsplash.com/?utm_source=medium&utm_medium=referral)

Statistical inference, such as hypothesis testing and confidence interval, are well-known concepts in statistics. However, many people including myself would sometimes misinterpret the result from statistical inference. Hopefully, this article would serve as a refresher on how to interpret statistical inference correctly.

**What is Statistical Inference?**

If we are studying the average household income in Los Angeles, it is impossible to collect information from every household. Hence we randomly draw sample data from the population and infer the conclusion based on the sample data from the whole population.

*This process of analyzing sample data and making conclusions about the parameters of a population is called****Statistical Inference****.*

**Parameters**are **fixed** numbers that represent characteristics or measures of the population, such as,

* Mean, e.g., the average household income in Los Angeles.
* Variance, e.g., the variance of household income in Los Angeles.
* Standard Deviation, e.g., the standard deviation of household income in Los Angeles.
* Proportion, e.g., the share of households with income higher than $200,000 in Los Angeles.

There are three common forms of Statistical Inference. Each one has a different way of using sample data to make conclusions about the population. They are:

* **Point Estimation**
* **Interval Estimation**
* **Hypothesis Testing**

**Point Estimation**

For **Point Estimate**, we infer an unknown population parameter using a **single value** based on the sample data.

For example, we can use the average household income from the sample data (i.e., sample mean) to infer the average house income in Los Angeles (i.e., population mean).

***Point Estimate****is easy to understand and present. But it is not reliable because it will likely be off the mark.*

If we took another set of sample data (with the same sample size) and do a point estimate again, it is very likely we would end up making a different conclusion about the population. That means there is **uncertainty**when we draw conclusions about the population based on the sample data. Point estimation doesn’t give us any idea as to how good the estimation is.

*To account for this****uncertainty****, using Interval Estimation has advantages over Point Estimation.*

**Interval Estimation**

For **Interval Estimation**, we use an **interval of values** (aka **Confidence Interval**) to estimate an unknown population parameter, and state how confident we are that this interval would include the true population parameter.

To construct the confidence interval, we would need two metrics:

* the point estimate of the population parameter (e.g., sample mean)
* the standard error of the point estimate (e.g., the standard error of the sample mean), which indicates how different the population parameter is likely to be from one sample statistic.

We can easily compute the point estimate from one sample. But **how do we solve the standard error of the point estimate?**

Theoretically, we need to take **repeated samples** of the same size from the same population. Next compute the sample statistic (e.g., sample mean) for each set of samples.

The frequency distribution of sample statistics over repeated samples from the sample population is called the **Sampling Distribution**.

Then we can obtain the standard error by computing the standard deviation of these sample statistics.

Based on the “**Central Limit Theorem**”, we know the sampling distribution of statistics (e.g., sample means) follows a **normal distribution**. The 95% confidence interval can be computed as “sample mean +/- standard error \*2”

This sounds complicated, right? :(

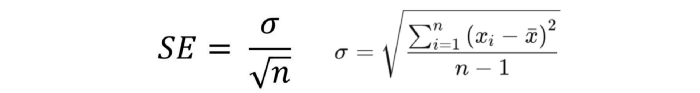
*Fortunately, in the real world, we****DON’T need to take repeated samples****to estimate the standard error for****Interval Estimation****. With a reasonable sample size, there is sufficient information within one set of samples.*

When we take repeated samples from a population, the sample statistics likely vary from one to other. The variation of samples depends on two factors:

* Variation of the population: The bigger the population variation, the bigger the variation between samples.
* Sample size: The bigger the sample size, the more likely they will resemble the population and each other. the smaller the variation between different sets of samples.

We rarely know the variation of the population, therefore we use sample standard deviation to estimate it. Then we divide it by the square root of the sample size to estimate the standard error of the sample mean.

The formula for Standard Error of Sample Mean:



* SE is the standard error of the sample mean
* σ is the sample standard deviation
* n is the sample size

In repeated samples, we would compute a confidence interval for each set of samples. We would expect about 95% of 95% confidence intervals from repeated independent sampling contain the true population parameter and about 5% don’t contain the true population parameter. Therefore,

**Misinterpretation #1:**

*It is****NOT correct****to say there is a 95%****probability****that the true population parameter lies between the 95% confidence interval.*

*It is****correct****to say we’re 95%****confident****that the true population parameter lies between the 95% confidence interval.*

Because the true population parameter is a fixed number, it is either in the confidence interval (p=100%) or not in the confidence interval (p=0%).

*Confidence interval indicates how****confident****we are about the conclusion, not the****probability****that the conclusion is correct.*

**Hypothesis Testing**

Unlike point estimate and interval estimate which are used to infer population parameters based on sample data, the purpose of hypothesis testing is to **evaluate the strength of evidence** from the sample data for making conclusions about the population.

In a hypothesis test, we evaluate **two mutually exclusive statements** about the population. They are

* **Null Hypothesis** (H0): H0 always uses =. For example, a null hypothesis could be “The average household income in Los Angeles is $65,000”.
* **Alternative Hypothesis** (Ha): Ha can use > or < or ≠. For example, an alternative hypothesis could be “The average household income in Los Angeles is above $65,000”, or “The average household income in Los Angeles is below $65,000”, or “The average household income in Los Angeles is NOT $65,000”.

The alternative Hypothesis using > or < indicates a **one-tailed**test because it only allows the alternative effect in **one direction**. The alternative Hypothesis using ≠ indicates a **two-tailed**test because it allows the alternative effect in **two directions**.

**Misinterpretation #2:**

*A hypothesis test is****NOT****designed to prove the null or alternative hypothesis. Instead, it evaluates****the strength of evidence******AGAINST the null hypothesis****using sample data.*

If we just want to find evidence against the null hypothesis, can we just compare the point estimate (e.g., sample mean) against the claim in the null hypothesis? If they turn out to be different (which will very likely happen), then does it mean we successfully falsify the null hypothesis?

The answer is obviously NO.

We can only observe the sample data rather than the whole population. There is almost always a difference (aka **Sampling Error)**between the sample data and the population.

*Therefore, there is a****likelihood****that the sample mean is different from the claimed number in the null hypothesis due to the****sampling error****, even if the claimed number in the null hypothesis is the true population parameter. The hypothesis test is a tool to****assess the likelihood of this possibility.***

Fortunately, we can create a probability distribution of the sample statistic (e.g., the sample mean) without repeated independent sampling thanks to two popular distributions,

* Z-distribution: When the sample size is greater than 30 and the population standard deviation is known, the Z-distribution is recommended.
* T-distribution: When the sample size is less than 30 or the population standard deviation is known, the T-distribution is recommended.
* population standard deviation is usually unknown and T-distribution is very similar to the normal distribution when the sample size is greater than 30, therefore, we would mostly apply T-distribution for a hypothesis test.

Let’s explain these concepts using an example.

It is claimed that the average household income in Los Angeles is $65,000. We randomly collect 10 samples from the population. It turns out that the sample mean is $70,000 with a sample standard deviation of $8,000. Assume normal distribution. Use a significance level of 0.05 to test the following hypothesis:Let's use µ to denote the population mean.  
H0: µ = $65,000  
Ha: µ > $65,000

*Let’s compute the****P-value****, which indicates the probability that the sample mean would be****as extreme or more extreme than the observed one****(e.g., the sample mean) assuming the null hypothesis is correct.*

**Misinterpretation #3:**

*It is****NOT correct****to say P-value is the probability that the null hypothesis is correct because the population parameter stated in the null hypothesis is a fixed number,****the probability that the null hypothesis is correct is either 0% or 100%.***

In other words, if the population mean is truly $65,000, then the observed sample mean of $70,000 would be considered to be extreme because it is not close to $65,000. But what does “more extreme” mean? It simply implies more extreme cases **in the same direction as the alternative hypothesis**.

Therefore, the P-value is computed as the probability that the sample mean is greater than $70,000, P(x̄ > 70,000).

Then we can rewrite the P-value by transforming x̄ into the **test statistic t**.

P(x̄ > 70,000) = P(t = (x̄ —µ) / (s/√n) > (70000–35000)/(8000/√10)) = P(t > 1.976)

From the t-distribution table with 9 (n-1 = 10 -1) degrees of freedom, we can compute that P(t > 1.976) = 0.03979.

The P-value of 0.03979 is smaller than the significance level of 0.05, therefore, we have **enough evidence to reject the null hypothesis** that the population mean is $65,000.

**What is the Significance Level (α)?**

The significant level is often denoted alpha or α. It indicates the **threshold**for statistical significance. In other words,

*The significant level is the****probability that we’re willing to accept extreme events****against the null hypothesis given that the null hypothesis is correct.*

For example, a significance level of 0.05 means we would expect a 5% probability to see evidence against the null hypothesis even if the null hypothesis is correct.

*1%, 5%, and 10% are all common thresholds of significant level. But they’re****NOT exhaustive****. We can choose any reasonable threshold of significant level based on how tolerant we’re willing to accept the false positive for a particular event given the null hypothesis is correct.*

**How to interpret the result of the P-Value?**

After we compute the P-value based on t or z statistic, how do we know if a P-value is small or big? Do we consider p-value = 0.02 small or p-value = 0.09 too big? This is where the Significance Level comes into play.

*We would need to use****the Significance Level****as the****benchmark****to determine whether a****p-value****is small or big.*

**If a p-value is less than the Significance Level**, then we can conclude the p-value is small. In other words, the probability that observing extreme events against the null hypothesis given that the null hypothesis is correct is very low. That means the evidence (against the null hypothesis) we found from the sample data would rarely occur by chance. Therefore, **We have enough evidence to reject the null hypothesis. It doesn’t mean we’ve proved the alternative hypothesis is correct. It only means we accept the alternative hypothesis or we’re more confident that the alternative hypothesis is correct.**

**If a p-value is greater than the Significance Level**, then we can conclude the p-value is big. In other words, the probability that observing extreme events against the null hypothesis given that the null hypothesis is correct is very high. That means the evidence (against the null hypothesis) we found from the sample data very likely occur by chance. Therefore, **We DON’T have enough evidence to reject the null hypothesis. It doesn’t mean we’ve proved the null hypothesis is correct nor do we accept the null hypothesis.**

**How do we interpret “extreme or more extreme” in a two-tailed test?**

Let’s continue with the above example.

Let's use µ to denote the population mean.  
H0: µ = $65,000  
Ha: µ ≠ $65,000

The P-value is computed as the probability of observing the extreme or more extreme cases **in the same direction as the alternative hypothesis**(e.g., µ ≠ $65,000).

We know the sample mean $70,000 is $5,000 higher than the claimed population mean $65,000. **In a two-tail test, “extreme or more extreme " means there is at least $5,000 difference between the sample mean and $65,000.**So that is

P(x̄ > 70,000 or x̄ < 60,000) = P(t > 1.976 | t < -1.976) = 2\*P(t > 1.976) = 0.07957

The P-value of 0.07957 is greater than the significance level of 0.05, therefore, we **DON’T have** **enough evidence to reject the null hypothesis** that the population mean is $65,000.

In this post, I’ve covered various misinterpretations of statistical inference and also the correct ways we should interpret these statistical concepts. The main takeaways are

* The point estimate is almost always different from the population parameter due to sampling error.
* The confidence interval tells us how confident we are about the conclusion, not the probability that the conclusion is correct.
* A hypothesis test doesn’t prove the null or alternative hypothesis. Instead, it only evaluates the strength of evidence AGAINST the null hypothesis using sample data.

Stay tuned! I will be posting more articles about statistics and data science.