Lecture Notes: Null Hypothesis Significance Testing

October 24, 2022

1 Introduction

In this chapter we will discuss a technique called Null Hypothesis Significance Testing (NHST) that can be used to assess the validity of a claim or hypothesis. The hypothesis under consideration can be about (i) a parameter of a statistical model, (ii) the probability distribution associated with a random variable or (iii) properties like mean and variance of a population. In this chapter we will only focus on hypotheses made about the mean of a population.

2 Null Hypothesis Significance Testing (NHST)

Here is a scenario for which NHST is typically applied. Engineers at the Aston Martin Formula 1 team have installed a new aerodynamics package on their car. The engineers believe the upgrade will improve the top speed of the car, which is currently measured to be 260 mph. Here is a brief description of the steps involved in testing this claim using the NHST framework.

The first step is to define the null and the alternative hypothesis of the problem. The null hypothesis is a statement that implies that the claim being assessed is incorrect. In the scenario being considered the null hypothesis would imply that there is no (null) improvement to the top speed caused by the changes made to the car, H_0; speed $s \leq 260$ mph. The alternative hypothesis denoted by H_A is the complement to the null hypothesis and states that the changes made do have an effect; H_A: speed $s > 260$ mph.

The next step is collect and analyze data. This might involve several independent measurements of the straight line speed of the car. Once the data is collected we quantify the disagreement between the observed data and the null hypothesis. If the disagreement is sufficiently strong we reject the null hypothesis. If not, we accept the null hypothesis in favor of the alternative hypothesis.

This testing procedure is similar to how people accused of crimes are treated in the criminal justice system. The default assumption or the null hypothesis is that the person accused of the crime is innocent. The prosecutors have the responsibility to collect and present strong enough evidence to overturn the claim of innocence. In the absence of strong evidence the defendant is acquitted, i.e the null is retained.

2.1 Example : One-sided hypothesis testing

An automobile engine manufactured by Honda emits on average 120 mg of oxides of nitrogen (NOx) per second when operating at 100 HP. New regulations demand that the NOx emission has to be lower than 100 mg per second when operating at 100 HP. Engineers have updated the design of the engine and believe that the emission rate has been cut down. 50 modified engines were built and tested. The sample mean of NOx emissions from the new engines was 92 mg/s and the sample standard deviation was 21 mg/s. The company wants to make a claim that its engines obey new emission regulations. Using NHST assess the validity of this claim. [1]

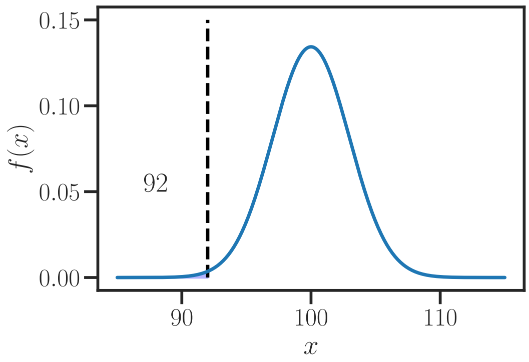

Figure 1: The graph shows the p.d.f. of the sample mean $\overline{X}_n \sim \mathcal{N}(100, 21/\sqrt{50})$. The area in shaded region ($x < 92$) gives the P-value.

The company's objective here is to state that there is statistically significant proof that the modified engines comply with the new emissions regulations. The null hypothesis in this case is that the modified engines do not comply with the new regulations and that there is no discernible effect due to the upgrades made. We can mathematically express this statement as $H_0; \mu \geq 100$ mg/s. Here μ is the mean emission rate of all engines in the population. The alternative hypothesis can be expressed as $H_A; \mu < 100$.

Let us assume that the null hypothesis is true, i.e. the engines have a mean emission rate which is greater than 100 mg/s. We now analyze the collected sample data to assess if it disagrees with this assumption. The sample mean emission rate $\overline{X}_n$ is 92 mg/s. This certainly seems to disagree with the null hypothesis. However, the sample mean $\overline{X}_n$ is a random variable and the engineers could have just gotten lucky with their sample. To ensure that this is not the case, we estimate a quantity called the P-value, which gives an estimate of how strongly the observation disagrees with the null hypothesis. The lower the P-value, the stronger is the disagreement.

2.1.1 P-value

The P-value is defined as the probability that sample data collected from this population will disagree at-least as strongly as the observed data $\overline{X}_n = 92$ mg/s, *under the assumption that the null hypothesis is true.* Since we need to compute probabilities associated with $\overline{X}_n$,

[1] This example has been taken from Statistics for Engineers and Scientists by William Navidi.

we need its probability distribution or p.d.f. Thanks to the central limit theorem, we know that $\overline{X}_n \sim \mathcal{N}(\mu, \sigma/\sqrt{n})$. Here μ is the mean emission rate of the population of modified engines. Since we have assumed that the null hypothesis is true, we choose $\mu = 100$.

Note that we could have chosen any value of μ greater than 100. We choose $\mu = 100$ since we want to test the data against the strongest candidate amongst all possible null hypothesis. $\mu \in [100, \infty)$. Typically the value of the null hypothesis which lies closest to the alternative hypothesis is the strongest candidate. If the population mean is 100 mg (null) it is conceivable that the sample mean could be 92 mg which is not that far off from 100 mg. However if we picked the null hypothesis to be say 200 mg, a sample mean observation of 92 mg seems extremely unlikely and implies that our null hypothesis is incorrect. Therefore the strongest candidate to test the data against would be $\mu = 100$ mg.

Let us assume that the sample standard deviation s is a good approximation for σ, the standard deviation of the population. Using C.L.T and the assumption that the null hypothesis is true ($\mu = 100$) we can approximate the p.d.f. of the sample mean as $\overline{X}_n \sim \mathcal{N}(100, 21/\sqrt{50})$. Data that is smaller than 92 mg/s disagrees more strongly than our observation $\overline{X}_n = 92$ mg/s, and therefore the P-value is given by

$$P = P(\overline{X}_n \leq 92) = \int_{-\infty}^{92} \frac{\sqrt{50}}{21\sqrt{2\pi}} e^{-\frac{(x-100)^2}{2(21/\sqrt{50})^2}} dx \tag{1}$$

One way to compute the P-value is to first map $\overline{X}_n$ to the standard normal $Z = (\overline{X}_n - 100)/(21/\sqrt{50})$. $\overline{X}_n = 92$ is equivalent to $Z = -2.69$. The P-value can then be estimated from the standard normal distribution

$$P = P(Z \leq -2.69) \int_{-\infty}^{-2.69} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz = 0.0035 \tag{2}$$

Statistical textbooks and papers which report NHST results typically estimate P-values using the above approach. Since P-values are estimated from a value of Z (standard normal) this testing procedure is commonly referred to as the z-test in literature. You can also use the cumulative distribution function of the standard normal to compute the above integral, or you could compute it using the following Python code.

```python
from scipy.stats import norm
import math

x_n    = 92

# Find p value from normal distribution
p_x = norm.cdf(x_n, loc = 100, scale = 21/math.sqrt(50))

# Compute z statistic to compute p value
z_val = (x_n - 100)/(21/math.sqrt(50))
```

```

# By default the parameters loc=0 and scale = 1
# loc corresponds to mean and scale to the standard deviation
p_z = norm.cdf(z_val, loc = 0, scale = 1)

print("P value from X: ",round(p_x,4), " P value from std. normal: ", round(p_z,4))
```

```
P value from X:  0.0035  P value from std. normal:  0.0035
```

The estimated P-value of 0.0035 is small. This means that the probability of observing data $\overline{X}_n < 92$ is very low *under the assumption that* $\mu = 100$. Therefore our assumption that the null hypothesis $\mu = 100$ is true is not a plausible explanation for the data and therefore we must reject it.

2.1.2 Significance Level

This raises the question of how small the P-value should be in order to reject the null hypothesis. Usually before the collection of data begins, a significance level α for the test is chosen. Typically it is set at $\alpha = 0.05$. The choice of α resides with the person conducting the test and has no scientific basis. If the computed P-value is less than α, we must reject the null hypothesis in favor of the alternative hypothesis. For the problem considered above we must reject $H_o; \mu \geq 100$ at the 5% ($\alpha = 0.05$) significance level. In other words we state that there is statistically significant evidence to conclude that the modified engines comply with the new regulations.

Let us perform a thought experiment. Assume that the modifications made by the engineers did not improve the mean emission rate μ. Let the ground truth (which is typically what are trying to find out) be $\mu = 100$ mg. If α is set to 0.05 even in situations where truth lies on the side of the null, there is a 5% chance that we reject it. In other words the engineers could have gotten lucky in their measurements and the company could end up making a false claim. Such an error is termed as a type I error in statistics literature. It is equivalent of pronouncing an innocent person as guilty and we want to avoid such a situation as much as possible. To reduce the possibility of such an error we can reduce the significance level to say 0.01 or 0.005. As the significance level is lowered the strength of the evidence required to reject the null becomes higher.

2.2 Example: Two-sided hypothesis testing

In the example we discussed above only data which is smaller than $\mu = 100$, is considered a disagreement. The computation of P-value therefore involves computing the area of one-side of the distribution of the sample mean $\overline{X}_n$ (see figure 1). Such tests are called one-sided or one-tailed hypothesis tests. In the next example, we will perform two-sided or two-tailed hypothesis testing.

A scientist has purchased a micro-balance for his lab. He measures the weight of 5 standardized samples using his new micro-balance. The weight of these samples is accurately known from other independent measurements. He documents the error of each sample measurement. The errors are 3, 19, 12, 21 and 5 μg. *The measurement errors are typically normally distributed. A perfectly calibrated micro-balance should have a mean error of zero. Perform a NHST at significance level 0.01 to assess if the device needs re-calibration.*

The claim that needs verification is that the mean error $\mu \neq 0$. Therefore the null hypothesis can be expressed as $H_0; \mu = 0$. It implies that the device functions perfectly and therefore does not need re-calibration. The next step is to compute the P-value or assess if the observed data is likely to occur given that the null hypothesis is true.

To do this we need to compute the probability of observing data that disagrees at least as strongly as the sample. Since the sample mean $\overline{X}_n$ is a reasonable statistic to estimate the population μ, we need to compute the probability associated with obtaining sample means which disagree at least as strongly as the observed sample mean. And for this we need the probability distribution of $\overline{X}_n$. We relied on the central limit theorem for this in the previous example. But here CLT cannot save us as the sample size is small.

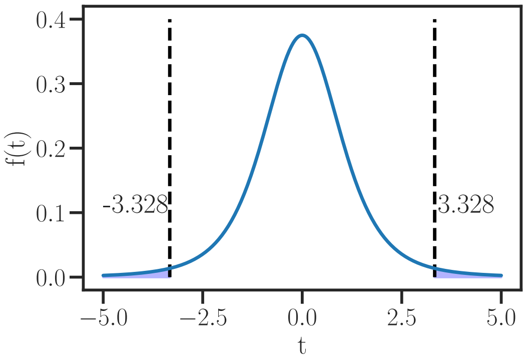

Figure 2: The graph shows the p.d.f. of the Student t-distribution with $\nu = 4$. The shaded region corresponds to the $|t| > 3.328$.

However, the measurement error is typically normally distributed and the observed data contains no outliers. Therefore we can do what William Gossett did many years ago and rely on the t distribution. The sample mean and standard deviation as computed from the data are $\overline{X}_n = 12$ and $s = 8.062$ and the sample size is $n = 5$. The t-statistic can be computed using the following expression.

$$t = \frac{\overline{X}_n - \mu}{s/\sqrt{n}} \tag{3}$$

Since we assume that the null hypothesis is true we set $\mu = 0$. The probability distribution of the t statistic written above is given by the Student's t distribution with the degrees of freedom parameter $\nu = n - 1 = 4$. As mentioned above the P-value is the probability (under the assumption that the null hypothesis is true), of observing data which disagrees at least as strongly with the null hypothesis as the sample data. In this example any measurement above or below zero, disagrees with the null hypothesis. The t statistic that corresponds to $\overline{X}_n = 12$ and $s = 8.062$ is $t = (12 - 0)/(8.061/\sqrt{5}) = 3.328$. Therefore the P-value is the area under the t-distribution with $\nu = 4$ in the shaded regions $t > 3.328$ and $t < -3.328$ (see shaded region in figure 2). Both these regions disagree with the null hypothesis at-least as strongly as the observed data. We can compute the area of the shaded regions using the cumulative distribution function of the t distribution.

$$\begin{aligned} P &= \int_{-\infty}^{-3.328} \frac{\Gamma(5/2)}{\sqrt{4\pi}\Gamma(2)}\left(1+\frac{t^2}{4}\right)^{-5/2} dt + \int_{3.328}^{\infty} \frac{\Gamma(5/2)}{\sqrt{4\pi}\Gamma(2)}\left(1+\frac{t^2}{4}\right)^{-5/2} dt \\ P &= F(-3.328) + 1 - F(3.328) = 0.0292 \end{aligned} \tag{4}$$

In the above equation F is the cumulative distribution function of the Student t-distribution with $\nu = 4$. Since the P-value is estimated from the t statistic, this testing procedure is commonly referred to as the t-test. Here is a small Python code snippet that computes the P-value from the Student-t distribution.

```python
from scipy.stats import t
nu    = 4
ts    = 3.328
p_val = t.cdf(-ts,nu) + 1 - t.cdf(ts,nu)

print(round(p_val,4))
```

```
0.0292
```

The P-value is estimated to be 0.0292. Since the P-value is greater than the significance level $\alpha = 0.01$, we cannot reject the null hypothesis. Therefore we conclude that the instrument does not need re-calibration. If however the significance level was set at $\alpha = 0.05$, then we would have rejected the null and concluded that the instrument needs re-calibration.

2.3 Conclusions of a hypothesis test

There can be only two outcomes of a hypothesis test.

1. Rejection of the null hypothesis. This happens when the P-value is less than the significance level. Rejection of the null hypothesis implies that it is an unlikely explanation for the observed data. It indirectly favors the alternative hypothesis as a possible explanation for the observed data.

2. Non rejection of the null hypothesis. This occurs when the P-value is greater than the significance level. The null hypothesis seems like a plausible explanation for the observed data and therefore we cannot rule it out.

2.4 Designing a hypothesis test

A light bulb manufacturer sells bulbs that have a lifetime of at least 25,000 hours. They claim to have come up with a new model with a much higher lifetime. Let us design a hypothesis test to verify the claims of the company.

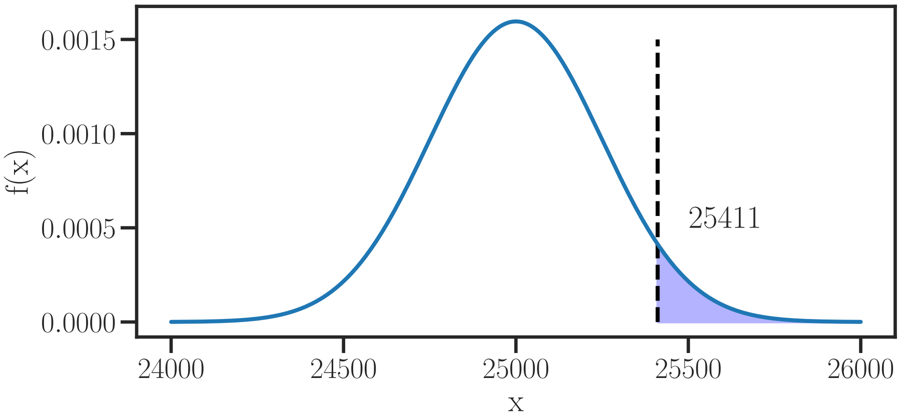

Figure 3: The blue curve is the probability distribution of the sample $\overline{X}_n$ under the assumption that the null hypothesis is true. $\overline{X}_n \sim \mathcal{N}(25000, 250)$. The area of the blue shaded region equals the significance level $\alpha = 0.05$. If the observed sample mean is greater than $\overline{X}_n > 25411$, it falls in the rejection region and results in the rejection of the null hypothesis.

Let us revisit the steps involved in a hypothesis test, this time focussing on how a test should be designed to find the ground truth. These ideas were first put forward by Neyman and Pearson and were an extension of the hypothesis testing as originally conceived by Fisher.. In particular let us focus on the decisions behind setting the significance level and the sample size. The null hypothesis must be contrary to the claim of the manufacturer and we can express it as $H_0; \mu \leq 25000$. Here μ is the mean lifetime of the light bulbs in the population. The alternative hypothesis is $H_A : \mu > 25000$. For now let us fix the sample size to be $n = 100$. Assume that the standard deviation of lifetimes is measured to be 2500 hours.

2.4.1 Significance level

Typically the significance level α is chosen in an ad-hoc manner. Let us pick $\alpha = 0.05$. To understand what this choice implies let us perform a thought experiment. Pretend like you are a God like entity and that you happen to know for a fact that the null hypothesis $\mu = 25000$ hours is indeed true.

Let $\overline{X}_n$ be the sample mean of the data collected from 100 light bulbs. In the scenario where $\mu = 25000$ is the ground truth, the central limit theorem tells us that the probability distribution of $\overline{X}_n$ can be approximated as $\mathcal{N}(25000, 2500/\sqrt{100})$. Figure 3 shows a plot of this distribution. The sample mean will be centered around 25000, and mostly will take a value that lies within two standard deviations of the mean $[24500, 25500]$. On rare occasions the sample mean, will be present in the tails of the distribution.

Values of $\overline{X}_n$ towards the right extreme of the plot in figure 3, strongly disagree with the null hypothesis. Let us find the value of $\overline{X}_n^*$, such that the area under the curve in the interval $[\overline{X}_n^*, \infty) = \alpha = 0.05$. One way to find $\overline{X}_n^*$, is to first map $\overline{X}_n$ to the standard normal via $Z = (\overline{X}_n - 25000)/250$. Finding $\overline{X}_n^*$, is equivalent to finding z^* such that the area under the standard normal between $[z^*, \infty) = 0.05$. Here is a Python code snippet to find z^*.

```python
from scipy.stats import norm
from scipy.optimize import fsolve
import numpy as np

def z_eqn_to_solve(xa, c = 0.05):

    # Return value of eqn 1 - F(x) - alpha = 0
    return 1 - norm.cdf(xa) - c

def get_z_limits ( alpha = 0.05 ):

    # Complete this function
    sol = fsolve(z_eqn_to_solve, x0 = 0.1, args=(alpha))
    return sol

print(np.around(get_z_limits(0.05),3))
```

```
[1.645]
```

The value of $z^* = 1.645$ and therefore $\overline{X}_n^* = 25000 + 1.645 * 250 \simeq 25411$. There is a 5% chance that a sample mean obtained from the population will be greater than 25411 hours (assuming the null hypothesis is true). And sample means with $\overline{X}_n > 25411$ will have a P-value less than $\alpha = 0.05$, leading to rejection of the null hypothesis. For example $\overline{X}_n = 25500$ has a P-value of 0.0227. Therefore the interval $[25411, \infty)$ is called the rejection region.

Even when the null hypothesis is the ground truth, in other words even if the light bulbs had a life time of 25000 hours or less, there is atleast a 5% chance of observing a sample mean which falls in the rejection region, which leads us to *erroneously* conclude that the null hypothesis should be rejected or that the bulbs have a lifetime greater than 25000 hours. If we choose to be conservative and reduce the chances of making such an error, we could lower the significance level to say 0.01 or 0.005.

2.4.2 Power of a test

The other error than can occur during hypothesis testing, is failing to reject the null hypothesis when the alternative hypothesis is the ground truth. Consider a scenario where the true mean lifetime of the bulbs is $\mu = 25500$ hours (ground truth). Therefore samples

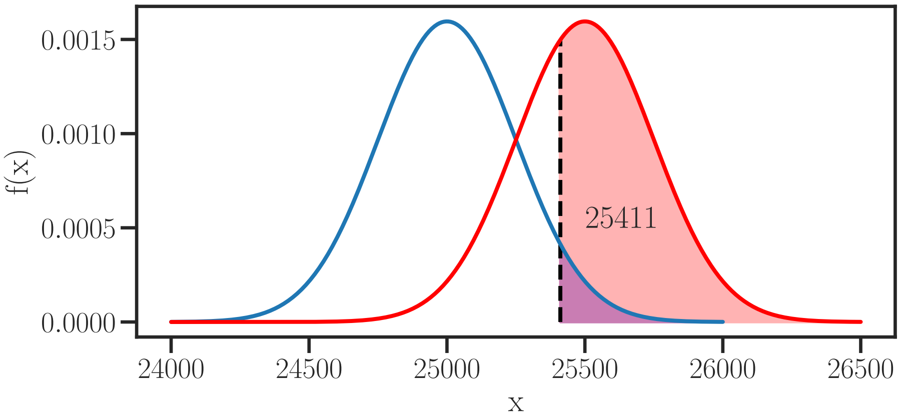

Figure 4: The blue curve is the probability distribution of the sample $\overline{X}_n$ under the assumption that the null hypothesis is true. $\overline{X}_n \sim \mathcal{N}(25000, 250)$. The area of the shaded region under the blue curve equals the significance level $\alpha = 0.05$. If the observed sample mean is greater than $\overline{X}_n > 25411$, it falls in the rejection region and results in the rejection of the null hypothesis. The red curve corresponds to the alternative hypothesis $\overline{X}_n \sim \mathcal{N}(25500, 250)$. The red-shaded region gives the area under the red curve that lies in the rejection region and it corresponds to the power of the test and it equals 0.6387.

collected from the population will have a sample mean that is normally distributed with mean 25500 hours and standard deviation $2500/\sqrt{(}100) = 250$ hours; $\overline{X}_n \sim \mathcal{N}(25500, 250)$. In figure 4, the red curve shows the p.d.f of $\mathcal{N}(25500, 250)$. Samples of bulbs collected will have a sample mean $\overline{X}_n$ which is drawn from this distribution (red curve), mostly likely lies in the interval $[25000, 26000]$ since the ground truth is $\mu = 25500$ *hours*.

As discussed above for significance level $\alpha = 0.05$, we reject the null hypothesis if $\overline{X}_n >$ 25411 hours. Now let us compute the probability that the sample mean will be greater than 25411 hours, or the probability that the sample mean falls in the rejection region. We will do this by first mapping $\overline{X}_n$ to the standard normal Z. Since the ground truth is $\overline{X}_n \sim \mathcal{N}(25500, 250)$, $Z = (\overline{X}_n - 25500)/250$. Computing the probability that $\overline{X}_n > 25411$ hours is equivalent to computing the area under the standard normal for $z = (25411 - 25500)/250 = -0.355$.

$$P(z > -0.355) = \int_{-0.355}^{\infty} \frac{\sqrt{1}}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz = 0.6387 \tag{5}$$

The above result implies that even when the alternative hypothesis is true ($\mu = 25500$), in other words even if the bulbs truly have a lifetime of 25,500 hours there is only a 64% chance that the observed sample mean falls in the rejection region at the 5% significance level. In other words the probability that the hypothesis test *correctly* rejects the null thereby favoring the truth is just 64%. This value gives us the power of the test. Power is related to what is termed as type II error in statistics literature and gives us the probability of correctly convicting a guilty person. Ideally we would want the power of a test to be as

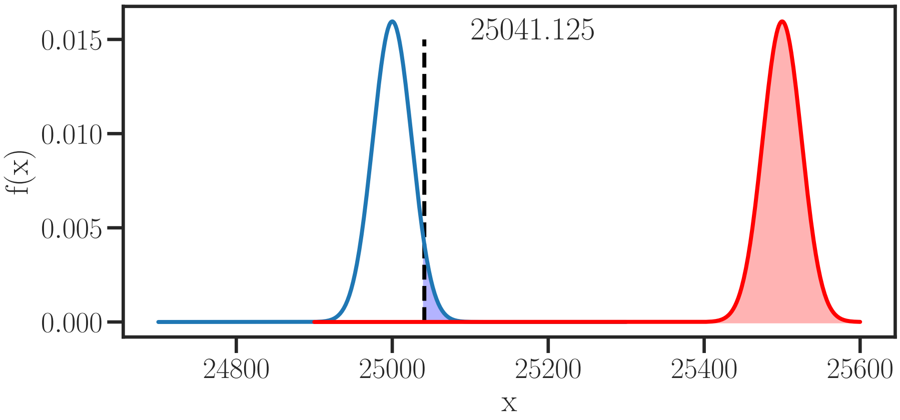

Figure 5: The blue curve is the probability distribution of the sample $\overline{X}_n$ under the assumption that the null hypothesis is true. $\overline{X}_n \sim \mathcal{N}(25000, 25)$. The area of the blue shaded region equals the significance level $\alpha = 0.05$. If the observed sample mean is greater than $\overline{X}_n > 25041.13$, it falls in the rejection region and results in the rejection of the null hypothesis. The red curve corresponds to the alternative hypothesis $\overline{X}_n \sim \mathcal{N}(25500, 25)$. Increasing the sample size has lead to both p.d.fs reducing their width or spread. Note the difference in the scale of the x-axis between figures 4 and 5. The shaded region under the red curve corresponds to the power of the test and it almost equals 1.0.

high (1) as possible. A rule of thumb is to design a test such that power is about 0.8 or 0.9.

In figure 4 we show a plot of the probability distribution corresponding to the null hypothesis (in blue). The blue shaded region corresponds to the rejection region of the test. Also shown in the figure is the probability distribution corresponding to the alternative hypothesis (in red). The red shaded region which overlaps with the blue shaded region corresponds to the power of the test.

Note that in order to compute the power, one needs to assume a certain alternative hypothesis to be the ground truth. In the above example we assumed the truth to be $\mu = 25,500$ hours in order to estimate the power. To further under why computing the power is important let us first look at how it changes with sample size.

2.4.3 Influence of sample size

Let us increase the sample size to $n = 10000$. Now let us estimate the power of the test for the alternative hypothesis (ground truth) $\mu = 25500$ at significance level $\alpha = 0.05$. The probability distribution that corresponds to the null hypothesis is $\mathcal{N}(25000, 2500/100)$. Following the procedure described above we can estimate the rejection rejection at confidence level $\alpha = 0.05$ to be $\overline{X}_n^* = 25041.125$. Let us compute the power of the test for alternative hypothesis $\mu = 25500$. This is given by the area shaded in red in figure 5 and it includes most of the area under the red-curve and is very close to 1. And this is an ideal scenario. Meaning when the sample size is as high as 10000, if the ground truth is

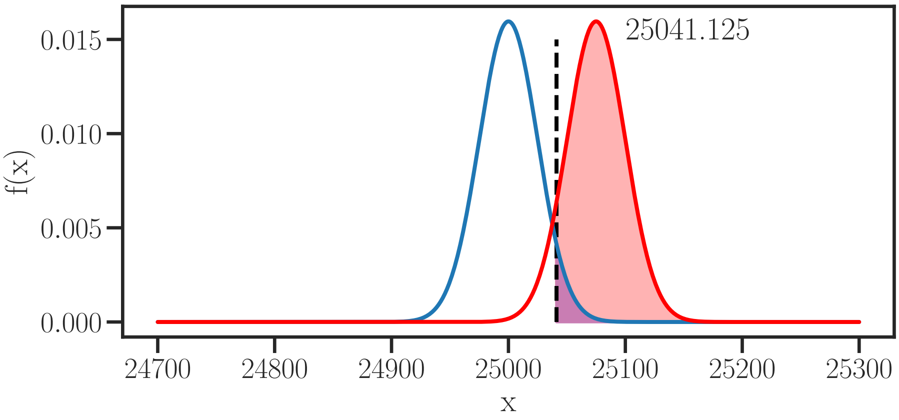

Figure 6: The blue curve is the probability distribution of the sample $\overline{X}_n$ under the assumption that the null hypothesis is true. $\overline{X}_n \sim \mathcal{N}(25000, 25)$. The area of the blue shaded region equals the significance level $\alpha = 0.05$. If the observed sample mean is greater than $\overline{X}_n > 25041.13$, it falls in the rejection region and results in the rejection of the null hypothesis. The red curve corresponds to the alternative hypothesis $\overline{X}_n \sim \mathcal{N}(25075, 25)$. Increasing the sample size has lead to both p.d.fs reducing their width or spread. Note the difference in the scale of the x-axis between figures 4 and 6. The shaded region under the red curve corresponds to the power of the test and it equals 0.912.

$\mu = 25500$, we will reject the null with almost 100% probability.

After looking at the case presented above you may wonder why one does not crank up the sample size for any test as it pushes the power close to 1. This is not done because obtaining samples consume time and effort and often there are resource based constraints. There is also another downside to increasing the sample size. Consider the scenario where the ground truth is $\mu = 25075$. The true lifetime is marginally higher than the null hypothesis $\mu = 25000$. The power for this alternative hypothesis $\mu = 25075$ can be estimated using the procedure detailed above and the result is 91.2% (see figure 6). The null hypothesis will be correctly rejected in favor of the truth ($\mu = 25075$) with a probability is 0.91. It appears that increasing the sample size has lead us to a favorable outcome of a high power value. The reason such an outcome is not desirable is because, the result (rejection of null) is not of much practical significance. In other words a 75 hr difference in lifetime is not a strong deviation from the null. It is a misconception to assume that only a substantial difference in effect from the null hypothesis will lead to its rejection. As n increases even marginal differences in effect will result in rejection of the null.

You can compare figures 4 and 6 to get a graphical explanation of the above idea. Power of a test is the overlap region between the probability distributions that correspond to the null hypothesis ($\mu = 25000$) and the alternative hypothesis ($\mu = 25075$). Increasing the sample size n, results in a reduction of the standard deviation $2500/\sqrt{n} = 25$, causing both null and the alternative probability distributions to tighten around its respective mean (see figure 6). The overlap region therefore increases leading to a high power value.

The rejection of the null merely implies statistical significance. In other words it implies that a sample mean has been observed which is several standard deviations away from the null hypothesis. As we crank up n, the standard deviation will reduce. And even if the truth is just marginally different in effect from the null hypothesis, it will produce a sample mean which is several standard deviations away from the null hypothesis, resulting in the rejection of the null. Therefore the result that the null hypothesis should be rejected is of no practical significance.

Typically during the test design the statistician makes a decision on what can be deemed as a significant difference in effect from the null. For the problem above it could be an improvement in lifetime by 500 hours or $\mu = 25500$ or $\mu = 26000$. The sample size is then chosen such that power at the desired effect is around 80 or 90%.

3 Things to remember while performing NHST

3.1 Report p - values

It is good practice to always state the P-value when reporting the findings of a hypothesis test. Both, $P = 0.049$ and $P = 0.00049$ lead to the same conclusion of rejecting the null hypothesis at significance level $\alpha = 0.05$. However the latter case offers a much stronger evidence against the null. Consider the case where $P = 0.051$. This evidence is not very far from the case where $P = 0.049$, yet our qualitative conclusions are dramatically different (rejection vs non rejection of the null). Therefore reporting the P-value of a hypothesis helps the reader make their own judgments regarding the validity of a hypothesis. The P-value is often misconstrued as the probability that the null-hypothesis is true. This is incorrect. It is a measure of disagreement of the observed data with the null hypothesis. Lower the P-value, stronger the disagreement.

3.2 Multiple testing

Suppose a researcher is testing a certain claim (e.g. drinking coffee improves your chess performance by 20 Elo points) at significance level 0.05. He analyzes the data he collected and finds a P-value above the 0.05 threshold and therefore his results cannot be published. The researcher proceeds to collect a new dataset and this time the P-value falls below 0.05. One might wonder if the researcher can fail to report the fact that he performed two sets of experiments and claim success based on one of the two data-sets. Let's say in this experiment the null hypothesis is the ground truth, i.e. coffee makes no difference to a chess player's performance. Since the significance value is = 0.05 there is a 5% chance of incorrectly rejecting the null. Now if two independent experiments are performed the probability of incorrectly rejecting the null at least once is $1 - P(0 \quad \text{rejections}) = 1 - (0.95 * 0.95) = 0.0975 = 9.75\%$. The probability of getting at least one success goes upto 22% if the same experiment is repeated 5 times. Therefore if the researcher collects two sets of data and claims success based on one of the two data-sets he is not performing

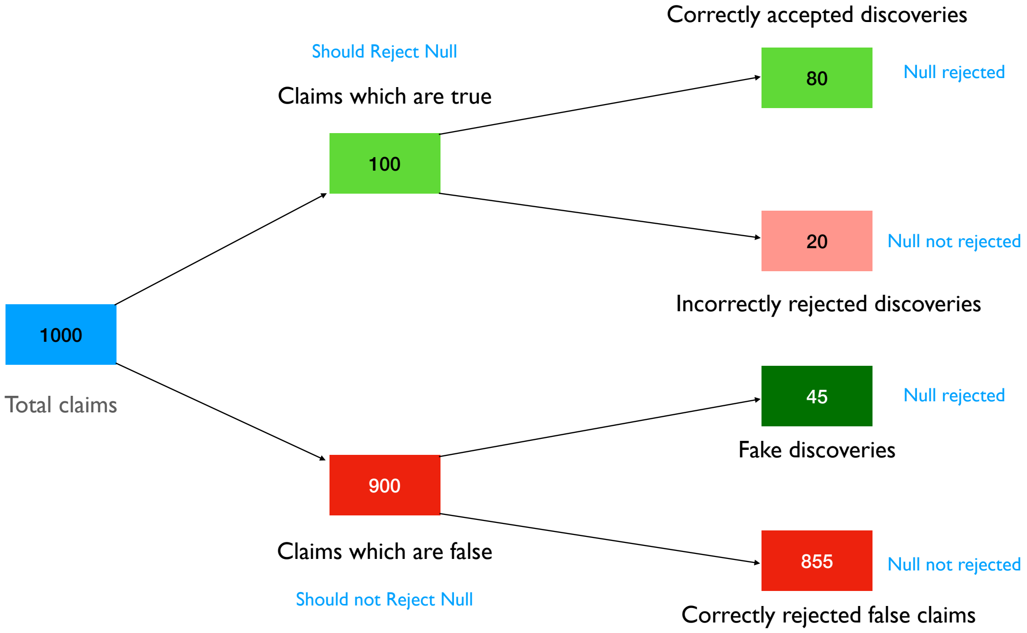

Figure 7: Schematic that illustrates the reproducibility crisis caused by NHST.

the tests at the 5% significance level anymore but rather at a much higher value.

3.3 Reproducibility crisis

David Spieghalter in his book The Art of Statistics uses the following example to illustrate the problems caused by NHST on the scientific community. Assume that a certain scientific field is ripe for new discovery and has an army of researchers working day and night to solve problems. Let's assume that in one year 1000 hypothesis are put forward and that 10% of them are indeed true scientific discoveries. Note that this is a very high number and arriving at new discoveries is much harder. Let's say all of these discoveries were based on results of hypothesis tests conducted at 5% significance level and 80% power.

Of the 1000 tests 900 are false claims and 100 are true discoveries. Of the 100 possible discoveries we will only correctly reject the null and accept the discovery 80 times as the tests are performed at power 80%. Also of the 900 false claims we will incorrectly claim a discovery $0.05 \times 900 = 45$ times. Therefore of all $(80 + 45 = 125)$ discoveries made in the year, roughly $45/120 \times 100 = 36\%$ turn out to be fake. This is a staggeringly high number and is something to keep in mind when reading scientific articles or trying to reproduce results from articles which claim discoveries based on hypothesis tests.

3.4 $P(H|D)$ vs $P(D|H)$

The NHST framework is an indirect method of testing the statistical significance of a hypothesis. The key quantity which decides if the null hypothesis should be rejected is the P-value, which is the probability of observing the sample data, under the assumption that a certain hypothesis is true $P(D|H)$. Here D denotes data and H denotes hypothesis. A more logical way to decide if a hypothesis is valid is by estimating $P(H|D)$, which is the probability that a certain hypothesis is true, given the observed data. Towards the end of next chapter we will learn how to perform hypothesis testing from $P(H|D)$.