Hypothesis Testing

Parthiban Rajendran

October 12, 2018

Contents

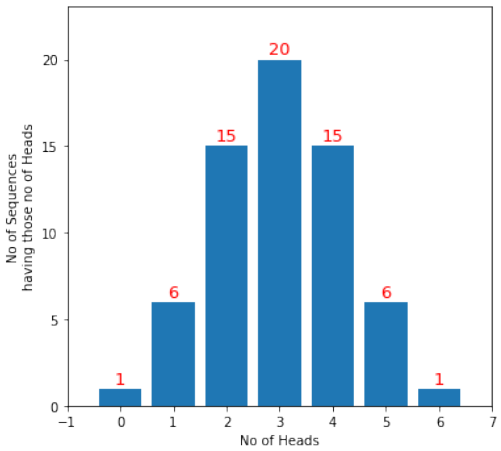
Chapter 1

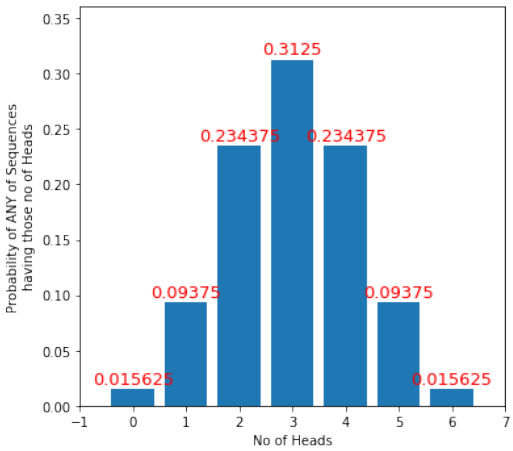
Testing hypotheses about a mean

1.1 Discrete Distribution

Example 1: Flip a Coin

Suppose we flip a fair coin, 6 times. We already have seen, flipping a coin once is a Bernoulli trial, and a number of times (including once also), gives us a binomial distribution for frequency and probability of no of heads (or tails depending on our interest) in the final combination.





Above distributions are the theoretical frequency and probability distribution of all outcomes possible for number of flips 6.

- From the distributions, you are already clear, if you conduct the experiment once (flipping 6 times), getting 3 heads in final outcome has the highest probability $P(X = 3) = 0.3125$.
- However, if you get 4 heads in final combination $X = 4$, that has about 23% probability, it is not bad, its just next to mean. So you still have ground to believe the mean was still $X = 3$.
- And, if you get 6 heads, then it is a rare case, that is, $P(X = 6) = 0.015625$ or 1.5625% only. We have reason to believe that, there is something else at play. Perhaps, coin was loaded (distribution skewed to right), that getting $X = 6$ was not a rarity at all.

This is kind of basis for hypothesis testing. We could define an uneventful hypothesis and then depending on probability of outcome we had from our experiment,we either believe that Hypothesis or reject it. If you had gotten $X = 3$, that has maximum probability of 31.25% of all outcomes, so we could very well accept our hypothesis that, indeed the mean is $X = 3$.

Alternate hypothesis: Mean has increased

Suppose we get $X = 4$:

1. We would first define a **null hypothesis** $H_0 : \mu = 3$ and alternate hypothesis $H_a : \mu > 3$
2. We would look at our experiment. Our outcome was $X = 4$. This has only 23% chance out of all possibilities if null hypothesis was true.
3. So we **cannot reject null hypothesis** that $H_0 : \mu = 3$. There is **lesser evidence** that the **mean has increased**, suggesting there is not enough data to believe alternate hypothesis $H_a : \mu > 3$

Suppose we get $X = 6$:

1. We would first define a **null hypothesis** $H_0 : \mu = 3$ and alternate hypothesis $H_a : \mu > 3$
2. We would look at our experiment. Our outcome was $X = 6$. This has only 1.5% chance out of all possibilities if null hypothesis was true.
3. So we **reject null hypothesis** that $H_0 : \mu = 3$ and say, there is **stronger evidence** that the **mean has increased**, suggesting alternate hypothesis $H_a : \mu > 3$

Alternate hypothesis: Mean has decreased

Suppose we get $X = 2$

1. We would first define a **null hypothesis** $H_0 : \mu = 3$ and alternate hypothesis $H_a : \mu < 3$
2. We would look at our experiment. Our outcome was $X = 2$. This has only 23% chance out of all possibilities if null hypothesis was true.
3. So we **cannot reject null hypothesis** that $H_0 : \mu = 3$. There is **lesser evidence** that the **mean has decreased**, suggesting there is not enough data to believe alternate hypothesis $H_a : \mu < 3$

Suppose we get $X = 0$

1. We would first define a **null hypothesis** $H_0 : \mu = 3$ and alternate hypothesis $H_a : \mu < 3$
2. We would look at our experiment. Our outcome was $X = 0$. This has again only 1.5% chance out of all possibilities if null hypothesis was true.
3. So we **reject null hypothesis** that $H_0 : \mu = 3$ and say, there is stronger evidence that the **mean has decreased**, suggesting alternate hypothesis $H_a : \mu < 3$

Significance level α

Who decides 1.5% was low enough to reject null hypothesis or 23% was high enough to avoid rejecting null hypothesis? Well, that is a standard taken by statisticians called significance level α. Suppose we take our $\alpha = 0.05$ or 5%, then we would say, if the probability of the outcome was below α, we reject the null hypothesis else we we will not.

For eg, in above cases,

- For $X = 4$, $P(X = 4) = 0.234$ then, $P(X = k) > \alpha$, so **cannot reject** null hypothesis $H_0 : \mu = 3$
- For $X = 6$, $P(X = 6) = 0.015$ then, $P(X = k) < \alpha$, so **reject** null hypothesis $H_0 : \mu = 3$
- For $X = 2$, $P(X = 2) = 0.234$ then, $P(X = k) > \alpha$, so **cannot reject** null hypothesis $H_0 : \mu = 3$
- For $X = 0$, $P(X = 0) = 0.015$ then, $P(X = k) < \alpha$, so **reject** null hypothesis $H_0 : \mu = 3$

Typically, α values are 0.05 (5%), 0.01 (1%), etc. and specified in the question. We had used α earlier in confidence intervals, for confidence level $1 - \alpha$. Also note it is better to say, we cannot reject null hypothesis than accepting it.

A basic hypothesis test

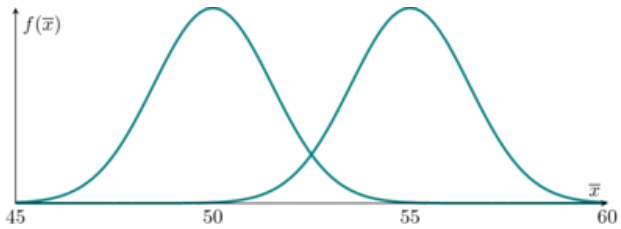
- If $P(X = k) < \alpha$, we will reject null hypothesis H_0
- If $P(X = k) > \alpha$, we will not reject null hypothesis H_0

But why did we reject or accept the hypothesis about the mean, when it was clearly $\mu = 3$? The μ was the population mean and most often in reality we would not know it. The binomial distribution shown above, was a theoretical distribution for equal probability of heads and tails. In reality, it may be the case that the coin was loaded (unequal probabilities) or population distribution was skewed. And we may not take enough trials to form a normal sampling distribution which then would give us hint about population mean due to CLT. Recall, $\left(\overline{X} \to \mu, S \to \sigma/\sqrt{n}\right)$. We would have one sample set output, and have to take best decision out of that. This is why we **assumed** null hypothesis, even though our theoretical mean was obvious. We will assume that above binomial distribution is indeed the case, and observe probability of our outcome, given that was case. Let us consider another example, this time continous.

1.2 Continuous Distribution

Let X be the breaking strength of a steel bar. If the bar is manufactured by process I, X has *population* distribution is $N(50, 36)$ and if process II, has *population* distribution $N(55, 36)$, a 5 units improvement. If hypothetically, we take extensive sample sets of size n and plot the means, we get another set of normal sampling distributions as below, with process I having sampling distribution $N(50, 36/\sqrt{n})$, and process II having sampling distribution $N(55, 36/\sqrt{n})$

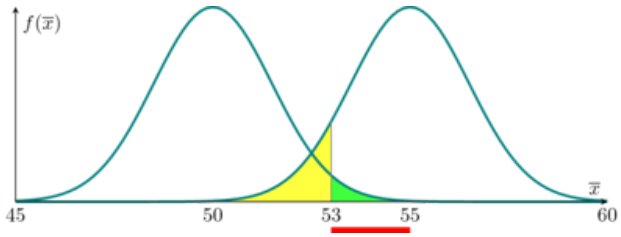
```
The tikzmagic extension is already loaded. To reload it, use:
  %reload_ext tikzmagic
```



Let us assume null hypothesis as $H_0 : \overline{X} = 50$ and alternate hypothesis as $H_a : \overline{X} = 55$. If now we take a sample set, $x = x_1, x_2, x_3..x_n$, and calculate $\overline{x}$.

- If $\overline{x} < 50$ we clearly **cannot reject** null hypothesis obviously $H_0 : \overline{X} = 50$, because the probability for mean of sampling distribution to be 55 is almost 0.
- If $\overline{x} > 55$, we clearly **reject** null hypothesis, as probability for mean of sampling distribution to be 50 is almost 0.

Of course, if $\overline{x}$ nears 45 or 60, we have similar hypothesis story waiting(?!). The interesting part is to wonder, what if $50 < \overline{x} < 55$. Note that, for $\overline{x} \geq 53$, the probability for $N(55, 36)$ is higher than that for $N(50, 36)$. Similarly, for $\overline{x} < 53$, the probability for $N(50, 36)$ is higher. This is highlighted with respective probability area below.



The rejection range for null hypothesis, which is $\overline{x} \geq 53$ is called the **critical region C** shown in red line above. Assuming null hypothesis is true, the probability for sampling set falling in critical region is called, again the significance level α. This is highlighted in green in above diagram.

Type I error:

Think about it. We decided if $\overline{x} \geq 53$, we would reject null hypothesis $H_0 : \overline{X} = 50$. However, there is still this slight probability α, that it could still be that the H_0 is true. Thus, due to our decision that we reject H_0 when $\overline{x} \geq 53$, we have α chance that, the reality is still H_0, and thus we would be committing an error. Rejecting null hypothesis H_0, when in reality, its true, is called Type I error. As I said, we have α chance for that, thus the probability of Type I error is α

Type II error:

It could happen the other way also as noted in yellow shade above. We decided if $\overline{x} < 53$ we would reject the alternate hypothesis $H_a : \overline{X} = 55$. But though less, there is still a probability shown in yellow color above, that H_a could be true, and we still choose to reject it. Rejecting the alternate hypothesis H_a, when in reality, it is true, is called Type II error. The associated probability of doing that error, as shown in yellow, is denoted by β

Critical region, Type I and II errors

- The values on $\overline{x}$ axis, where H_0 is rejected is called **Critical region C**.
- Rejecting null hypothesis H_0 when in reality it is true is **Type I error**. Its probability is called significance level of the test α
- Rejecting alternate hypothesis H_a when in reality it is true is **Type II error**. Its probability is β

Calculating α and β for given n

This is when we get in to problem of calculating the associated probabilities. Let sample set size $n = 16$.

Note, α is given H_0 is true, the probability of sample mean falling in critical region, shortly noted as $P(\overline{X} > 53; H_0)$

By transforming the sampling distribution of process I to Z, we could calculate the probability α. In other words, by calculating the Z score for $\overline{x} = 53$, we could calculate the probability $P(\overline{X} > 53; H_0)$. Note $\sigma^2 = 36 \rightarrow \sigma = 6$

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} = \frac{53 - 50}{6/\sqrt{16}}$$

```
In[6]: def get_zscore(x, mu, sigma, n):
           num = x - mu
           from math import sqrt
           den = sigma/sqrt(n)
           return round(num/den, 3)

       z = get_zscore(53, 50, 6, 16)
       print(z)
```

```
2.0
```

The Z score is 2. Now its easier to calculate the probability area.

$P(\overline{X} > 53; H_0) = P(Z > 2; H_0)$

```
In[7]: def get_zarea(z, tail='right'):
           from scipy import stats
           if tail == 'right':
               alpha = round(1 - stats.norm.cdf(z),4) # right tailed area
           else: # assume left tail
               alpha = round(stats.norm.cdf(z),4) # left tailed area
           return alpha

       za = get_zarea(z, 'right')
       print(za)
```

```
0.0228
```

Thus the significance level of the test with sample size 16, the probability of making Type I error, α is 0.0228 or 2.28%. Similarly one could calculate β as below. Note, β is from alternate hypothesis, so alternate sampling distribution $N(55, 36/16)$

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} = \frac{53 - 55}{6/\sqrt{16}}$$

```
In[8]: z = get_zscore(53, 55, 6, 16)
       za = get_zarea(z,'left')
       print(z, za)
```

```
-1.333 0.0913
```

Thus the probability of making Type II error, β is 0.0913 or 9.13%.

Adjusting α and β

Note that,

- If we decrease critical region C, then α reduces, however β increases
- If we increase critical region C, then β reduces, however α increases
- If we increase sample size n, that decreases $\alpha/sqrtn$ thus higher Z value implying reduced α and β

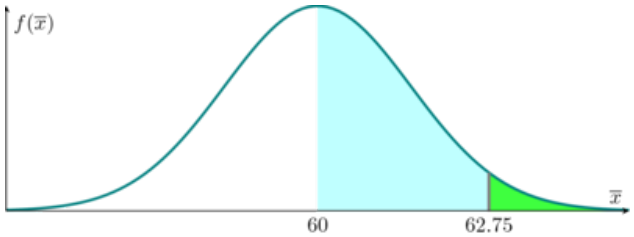
1.3 Composite Hypothesis:

What we saw so far was simple hypothesis test, because the alternate hypothesis was simple - $H_a : \mu = 55$. We had another process II and assumed, if not $\mu = 50$, it only could be $\mu = 55$. Often, we would have situations where we are not aware of process II or there could be more such possibilities. We could only say, if $\mu = 50$ or not (increased or decreased). In other words, instead of one alternate normal distribution N(55,36), we might have many, all with mean $\mu > 50$. So our alternate hypothesis should be $H_a : \mu > 50$. This could happen in other direction also, that the mean reduced leading to $H_a : \mu < 50$. Or we may only be interested if μ changed (suggesting $H_a : \mu \neq 50$). If we have any such alternate hypothesis, then we would call the test as **Composite Hypothesis** because it is composed of all possible alternate normal distributions.

Let us take one case $H_a : \mu > K$, where K is any value and analyze in detail.

Example

Assume that we have a population distribution which is normal with unknown mean μ but known variance $\sigma^2 = 100$. Say we are testing the simple null hypothesis $H_0 : \mu = 60$ against the composite alternative hypothesis $H_1 : \mu > 60$ with a sample mean X based on $n = 52$ observations. Suppose that we obtain the observed sample mean of $\overline{x} = 62.75$. Our situation is depicted below.



$$H_0 : \mu = 60$$
$$H_a : \mu > 60$$

We assume our sampling distribution is $N(60, 100)$ (that is, assuming null hypothesis is true), and then wondering what is the probability of getting $\overline{x} > 62.75$. Note few things carefully.

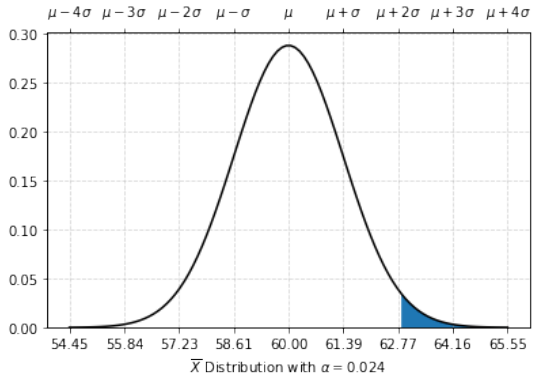
- We did not ask, what is $P(\overline{X} = 62.75)$ like we did in discrete distribution earlier. It is 0 for continuous anyway.

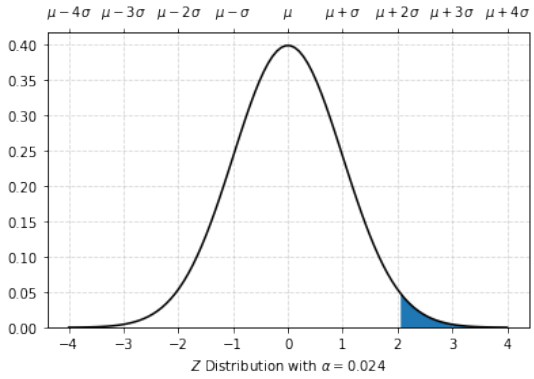
- We did not ask, what is $P(60 < \overline{X} < 62.75)$, our sample mean is greater than assumed mean, so if we assume null hypothesis, then this is definitely a higher probability as shown in blue above, reinforcing null hypothesis again.
- We could have done a continuity correction around $\overline{X} = 62.75$, but we need to derive more. If you recall earlier example, we said if $\overline{x} > 53$ we assume alternate hypothesis $\mu = 55$ to be true. Similarly, here, we need to define critical region, and if we get sample mean $\overline{x}$ within that, we assume alternate hypothesis to be true (thus establishing our chances to commit Type I error)

If we assume critical region to be C $\{\overline{x} : \overline{x} \geq 62.75\}$, then our probability of making Type I error is $\alpha = P(\overline{X} \geq 62.75)$. We could find that using Z score.

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} = \frac{62.75 - 60}{10/\sqrt{52}} = 1.983$$

$$\therefore P(\overline{X} \geq 62.75) = (Z \geq 1.983) = 0.024 \tag{1.1}$$





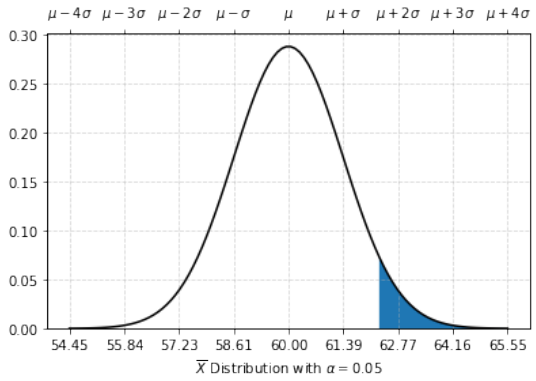
So if we decide critical region, C $= \{\overline{x} : \overline{x} \geq 62.75\}$, then our α would be 0.024. That is, there is about 2.4% chance of making Type I error.

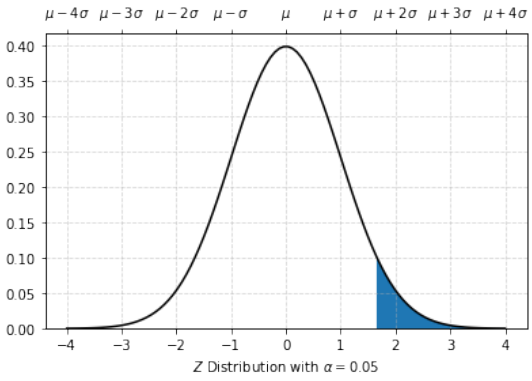
Typically, the α is decided up front. For eg, we would say, 5% probability allowed to make Type I error for the problem at hand. This means, $\alpha = 0.05$. We could then go in reverse, to calculate the value beyond which we could declare critical region.

$\alpha = 0.05 \implies Z_\alpha = 1.645$ because , $P(Z \geq 1.645) = P(Z \geq Z_\alpha) = 0.05$

$$\overline{X} = Z\frac{\sigma}{\sqrt{n}} + \mu = 1.645\left(\frac{10}{\sqrt{52}}\right) + 60 = 62.281$$

$$P(Z \geq Z_\alpha) = P(\overline{X} \geq 62.281) = 0.05 \tag{1.2}$$





Since our **given permitted** α is 0.05, our permitted critical region C is $\overline{x} : \overline{x} \geq 62.281$. That is, if we get a sample mean $\overline{x} \geq 62.281$ we would **reject** null hypothesis and take alternate hypothesis, even when there is 5% chance of committing Type I error. The sample mean we got was $\overline{x} = 62.75$, which had $\alpha = 0.024$. There is only a 2.4% probability that, the sample mean could be ≥ 62.75. Since 2.4% is within the permissble range of 5%, we reject the null hypothesis $H_0 : \mu = 60$ and support alternate composite hypothesis $H_a : \mu > 60$.

Now how we have traversed from the notion of quoting extreme low probability for sample set for rejection to a pre determined level for rejection α, and simply accept or reject based on if sample set probability fell within that region or not. Understandably, the α should be typically set by problem domain experts who have enough expertise to trade off between Type I and II errors. (Decreasing Type I probability may increase Type II probability, etc).

By the way, the earlier α we got from sample set is called **p-value** to differentiate it from preset α

Equations 1.1 and 1.2 could be further condensed as, if Z denotes the Z score of sample mean, $Z = \left(\frac{\overline{X} - \mu}{\sigma/\sqrt{n}}\right)$, then if, $Z \geq Z_\alpha$, **reject** null hypothesis.

From 1.2 we could also write,

$$P(Z \geq Z_\alpha) = P\left(\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \geq Z_\alpha\right) = 0.05 \tag{1.3}$$

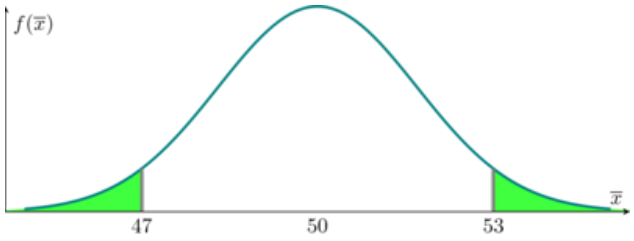
Example

A researcher is testing the hypothesis that consuming a sports drink during exercise improves endurance. A sample of $n = 50$ male college students is obtained and each student is given a series of three endurance tasks and asked to consume 4 ounces of the drink during each break between tasks. The overall endurance score for this sample is $M = 53$. For the general population of male college students, without any sports drink, the scores for this task average $\mu = 50$ with a standard deviation of $\sigma = 12$. Can the researcher conclude that endurance scores with the sports drink are significantly different than scores without the drink? Assume $\alpha = 0.05$

Solution:

Given population has $\mu = 50, \sigma = 12$. It is not known if its normal, but sample size $n = 50$ is > 30, so good enough to consider the resultant sampling distribution of sample means from this population to form a normal distribution $N(\mu = 50, S^2 = \sigma^2/n = 12^2/50)$. Given sample set has sample mean $\overline{x} = 53$. Our null hypothesis would be $H_0 : \mu = 50$. Alternate is $H_a : \mu \neq 50$.

As first step, we will try to define a temporary critical region. Since we are interested not in μ increase, but change, it could be both μ increasing or decreasing. That is, we have to define critical region for both directions. Taking the delta $\delta = 53 - 50 = 3$ on left side also, we could now establish a temporary critical region **C**: $\{\overline{x} : \overline{x} \leq 47 \quad \text{or} \quad \overline{x} \geq 53\}$. If a new hypothetical sample mean falls within this C, we would reject null hypothesis and take alternate hypothesis. Our situation is depicted below.



$$H_0 : \mu = 50$$
$$H_a : \mu \neq 50$$

Let us transform the above assumed sample distribution (provided null hypothesis is true) to Z distribution to get the respective probabilities.

For $\overline{x} = 53$, $Z_{53} = \dfrac{\overline{x} - \mu}{\sigma/\sqrt{n}} = \dfrac{53 - 50}{12/\sqrt{50}}$

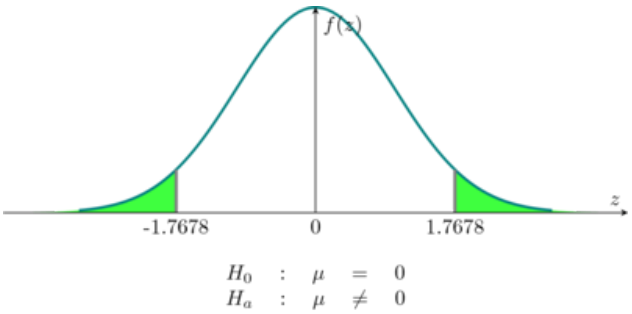
For $\overline{x} = 47$, $Z_{47} = \dfrac{\overline{x} - \mu}{\sigma/\sqrt{n}} = \dfrac{47 - 50}{12/\sqrt{50}}$

```
In[13]: def get_Zscore_1(x_bar, mu, sig, n):
            num = x_bar - mu
            from math import sqrt
            den = sig/sqrt(n)
            return round(num/den, 4)

        def get_Z_1(zs, tail='right'):
            from scipy.stats import norm
            if tail == 'left':
                return round(norm.cdf(zs),4)
            else:
                return round(1- norm.cdf(zs),4)

        mu, sig, n = 50, 12, 50
        zs = get_Zscore_1(53, mu, sig, n)
        a1 = get_Z_1(zs,'right')
        print('z_53:{}, area:{}'.format(zs,a1))
        zs = get_Zscore_1(47, mu, sig, n)
        a2 = get_Z_1(zs,'left')
        print('z_47:{}, area:{}'.format(zs,a2))
        print('Total area:{}'.format(a1+a2))
```

```
z_53:1.7678, area:0.0385
z_47:-1.7678, area:0.0385
Total area:0.077
```

Thus $(Z_{47}, Z_{53}) = (-1.7678, 1.7678)$. The total probability area would be $P(Z \geq 1.7678) + P(Z \leq -1.7678)$. The probability area of each tail would be 0.0385, thus total area, which is *p-value* would be 0.077

$$\begin{aligned}\therefore P(\overline{X} \geq 53 \quad \cup \quad \overline{X} \leq 47) &= P(Z \geq 1.7678 \quad \cup \quad Z \leq -1.7678) \\ &= P(Z \geq 1.7678) + P(Z \leq -1.7678) \\ &= 0.0385 + 0.0385 \\ &= 0.077\end{aligned}$$



We could straight away conclude from above finding. We are given $\alpha = 0.05$ which is the total allowed probability for making Type I error. If we select C as **C**: $\{\overline{x} : \overline{x} \leq 47 \text{ or } \overline{x} \geq 53\}$, our probability of making Type I error would be 0.077 which is greater than allowed 0.05 limit. Thus right away, we could **fail to reject null hypothesis** H_0, which is same as concluding there is no significant evidence to believe there is change in the mean.

We could also have concluded right away from Z value. Note total area allowed $\alpha = 0.05$. This means, tail end probabilities on both ends should be 0.025, so they add up to 0.05. We could easily find the respective Z value as $Z_{0.025} = Z_{\alpha/2} = 1.96$.

Since $Z_{53} < Z_{\alpha/2}$, it is already evident, Z_{53} occupies more probability area. Similarly, $Z_{47} > -Z_{\alpha/2}$, Z_{47} is occupying more area. In simpler terms,

$|\pm 1.7578| < |\pm 1.96| \implies |Z| < |Z_{\alpha/2}| = \left|\dfrac{\overline{X} - \mu}{\sigma/\sqrt{n}}\right| < |Z_{\alpha/2}|$ and we **fail to reject null hypothesis** H_0

Generalizing, we could write as, for two tailed situation,

$$P(|Z| \geq |Z_{\alpha/2}|) = P\left(\left|\frac{\overline{X} - \mu}{\sigma/\sqrt{n}}\right| \geq |Z_{\alpha/2}|\right) = 0.05$$

$$\therefore \text{If} \quad \left|\frac{\overline{X} - \mu}{\sigma/\sqrt{n}}\right| \geq |Z_{\alpha/2}|, \quad \text{reject null hypothesis } H_0$$

Table 1.1. When σ known, and $n \geq 30$

H_0	H_a	Tail	Reject H_0 when..
$\mu = \mu_0$	$\mu > \mu_0$	Right	$Z \geq Z_\alpha$
$\mu = \mu_0$	$\mu < \mu_0$	Left	$Z \leq -Z_\alpha$
$\mu = \mu_0$	$\mu \neq \mu_0$	Both	$\lvert Z\rvert \geq Z_{\alpha/2}$

1.4 When σ is unknown or small sample size

In reality, σ is also often unknown, thus like we did in confidence intervals, we could use the student's t distribution to evaluate the hypothesis test. As before in confidence intervals, this goes without proof for now. Also when the sample size is small $n \leq 30$, we use t distribution.

Thus, if S represents sample standard deviation, the right tail example from 1.3, becomes,

$$P(t \geq t_{(\alpha,n-1)}) = P\left(\frac{\overline{X} - \mu}{S/\sqrt{n}} \geq t_{(\alpha,n-1)}\right) = 0.05 \tag{1.4}$$

Similary, for left and double tailed examples, we would have,

$$P(t \leq -t_{(\alpha,n-1)}) = P\left(\frac{\overline{X} - \mu}{S/\sqrt{n}} \leq -t_{(\alpha,n-1)}\right) = 0.05$$

$$P(|t| \geq |t_{(\alpha,n-1)}|) = P\left(\left|\frac{\overline{X} - \mu}{S/\sqrt{n}}\right| \geq t_{(\alpha,n-1)}\right) = 0.05$$

It is not needed to remember the above formula, one could always just reason them out. Let us try a left tail example since we have not yet done that.

Example

Bags of a certain brand of tortilla chips claim to have a net weight of 14 ounces. The net weights actually vary slightly from bag to bag and are normally distributed with mean μ. A representative of a consumer advocacy group wishes to see if there is any evidence that the mean net weight is less than advertised. For this, the representative randomly selects 16 bags of this brand and determines the net weight of each. He finds the sample mean to be $X = 13.82$ and the sample standard deviation to be $S = 0.24$. Use these data to perform an appropriate test of hypothesis at 5% significance level.

Solution

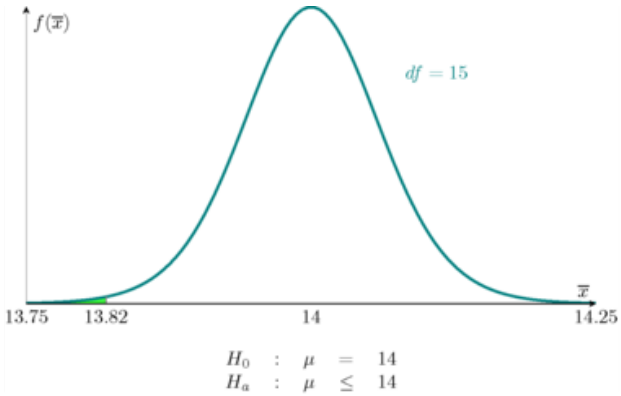
Given population $N(\mu = 14, \sigma^2)$. Sample set $n = 16$. Thus Sampling distribution will be $N(\mu = 14, \sigma^2/16)$

Given Sample set has $\overline{x} = 13.82, s = 0.24$

Given $\alpha = 0.05$

Forget about the formula. Only thing we need to remember is, we need to use *t distribution* because, σ is unknown and $n < 30$. Note even if either of the case, we would still have to use *t distribution*.

Let us start with defining critical region and thus arriving at our probability of making type I error, if we choose the critical region C to be $\{\overline{x} : \overline{x} \leq 13.82\}$. Our situation is depicted below. The area is so small its barely visible.



Note: The above *t distribution* was just a shifted and scaled one from standardized distribution with single sample standard deviation value $t = \dfrac{\overline{x} - 14}{0.24/\sqrt{16}}$ for illustrative purposes. However in reality, for each sample set calculated, teh sample standard deviation obviously varies, but the resulting histogram of 't' values calculated as $t = \dfrac{\overline{x} - 14}{s/\sqrt{16}}$ would resemble a standardized t distribution [1] As per our temporary critical region, if $\overline{x} \leq 13.82$ we would say, the μ has decreased. Let us calculate the probability of committing Type I error, if we do so, which is $P(\overline{X} \leq 13.82)$. To do that, we will proceed to calculating the *t score*.

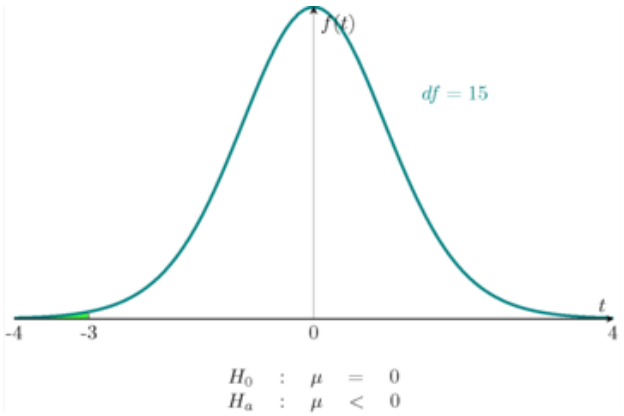
```
In[45]: def get_tscore_1(x_bar, mu, s, n):
            num = x_bar - mu
            from math import sqrt
            den = s/sqrt(n)
            return round(num/den, 4)

        def get_t_1(zs, df, tail='right'):
            from scipy.stats import t
            if tail == 'left':
                return round(t.cdf(zs, df),4)
            else:
                return round(1- t.cdf(zs, df),4)

        mu, s, n = 14, 0.24, 16
        ts = get_tscore_1(13.82, mu, s, n)
        a1 = get_t_1(ts,n-1,'left')
        print('t_p:{}, area:{}'.format(ts,a1))
```

```
t_p:-3.0, area:0.0045
```

$$\therefore \quad P(\overline{X} \leq 13.82) = P(t \leq -3) = 0.0013 \tag{1.5}$$

The standard t distribution with this area is depicted below.

[1] https://youtu.be/rePsvdAxwX8



Our probability of making a Type I error is only 0.0045 from the sample set, which is allowable under the limits of 0.05. That is, p *value* $< \alpha$. So we could reject the null hypothesis H_0 and say there is significant evidence for H_a, or that μ has decreased. You see, we did not even find $t_{\alpha,15}$, but for the sake of sticking to formula, we could find that and conclude as well. Finding the area helps better, because area is always positive so easy to compare.

```
In[47]: def get_tscore_2(sl, df, tail='left'):
            from scipy.stats import t
            if tail == 'left':
                return round(t.ppf(sl, df),4)
            else:
                return round(1- t.ppf(sl, df),4)

        print(get_tscore_2(0.05,15,'left'))
```

```
-1.7531
```

Thus, $P(-3 < -1.7531)$, that is, $P(t < t_{(\alpha,15)})$, we could **reject null hypothesis**.
Summaring for all cases,

Table 1.2. When σ unknown, and/or $n < 30$

H_0	H_a	Tail	Reject H_0 when..
$\mu = \mu_0$	$\mu > \mu_0$	Right	$t \geq t_{(\alpha,n-1)}$
$\mu = \mu_0$	$\mu < \mu_0$	Left	$t \leq -t_{(\alpha,n-1)}$
$\mu = \mu_0$	$\mu \neq \mu_0$	Both	$\lvert t \rvert \geq t_{(\alpha/2,n-1)}$

Tips to remember

It is always better to stick to calculating area to compare **p-value** with **significance level**. The signs in the formula could be confusing because often it is not obvious if right tail or left tail convention is used. In above example, $t_{(\alpha,15)}$ was -1.7531, but in above table, we wrote $t \leq -t_{(\alpha,n-1)}$, where, $t_{(\alpha,n-1)}$ meant 1.7531

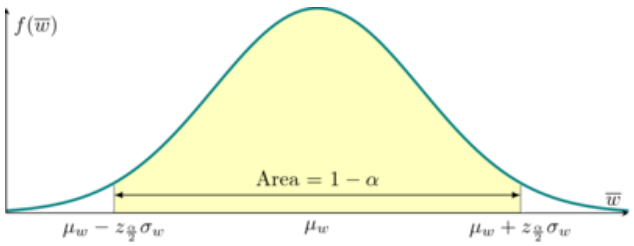
Chapter 2

Testing the difference between two means

Note the pre requisite to understand below material is to know confidence intervals for difference between two means as we straight away use the definitions from there. In fact, entire hypothesis testing concept is always understood better after learning confidence intervals and is the typical order in many textbooks.

2.1 σ known, sample sizes are high

Suppose that we are interested in comparing two approximately normal sampling distributions described by random variables $\overline{X} = N(\mu_{\overline{x}}, \sigma^2_{\overline{x}})$ and $\overline{Y} = N(\mu_{\overline{y}}, \sigma^2_{\overline{y}})$, created from population distributions described by random variables $X(\mu_x, \sigma^2_x)$ and $Y(\mu_y, \sigma^2_y)$. Note that $\overline{X}$ represents collection of sample means from sampled sets sampled from X and similarly for $\overline{Y}$. Since both $\overline{X}$ and $\overline{Y}$ are normally distributed, and assuming both are independent to each other, the distribution $W = \overline{X} - \overline{Y}$ would be again a normal distribution $W(\mu_w, \sigma^2_w)$, where $\mu_w = \mu_{\overline{x}} - \mu_{\overline{y}}$ and $\sigma^2_w = \sigma^2_{\overline{x}} + \sigma^2_{\overline{y}}$.



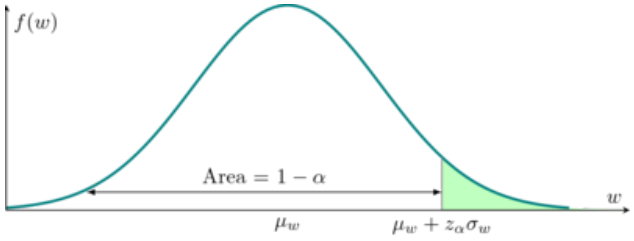
Then, we already know the confidence intervals could be calculated as below.

$$P\left(-z_{\frac{\alpha}{2}} \leq \frac{(\overline{X} - \overline{Y}) - (\mu_{\overline{x}} - \mu_{\overline{y}})}{\sqrt{\frac{\sigma^2_{\overline{x}}}{n} + \frac{\sigma^2_{\overline{y}}}{m}}} \leq z_{\frac{\alpha}{2}} \right) \approx 1 - \alpha$$

For Hypothesis testing, let the problem at hand is to wonder, if one mean is greater than the other. For eg, if $\mu_{\overline{x}} > \mu_{\overline{y}}$. This is another way of saying if $\mu_w > 0$. Then we could formulate our hypothesis as follows.

Null hypothesis: $H_0 : \mu_w = 0$
Alternate hypothesis: $H_a : \mu_w > 0$

Then the probability of making Type I error α, would be right hand tail area as follows. Note $z_{\frac{\alpha}{2}}$ becoming z_α as now its one side area we are interested in.

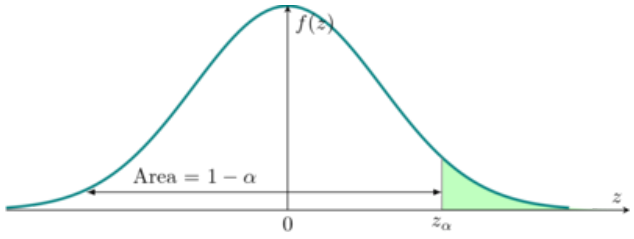


$$
\begin{aligned}
P(w \geq \mu_w + z_\alpha \sigma_w) &= \alpha \\
P(w - \mu_w \geq z_\alpha \sigma_w) &= \alpha \\
P\Big(\frac{w - \mu_w}{\sigma_w} \geq z_\alpha\Big) &= \alpha \\
P\left(\frac{(\overline{X} - \overline{Y}) - (\mu_{\overline{x}} - \mu_{\overline{y}})}{\sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}}} \geq z_\alpha\right) &= \alpha
\end{aligned}
$$

Since our null hypothesis is $\mu_w = 0$ or $\mu_{\overline{x}} = \mu_{\overline{y}}$, we could reduce the equation further as,

$$
P\left(\frac{\overline{X} - \overline{Y}}{\sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}}} \geq z_\alpha\right) = \alpha \tag{2.1}
$$

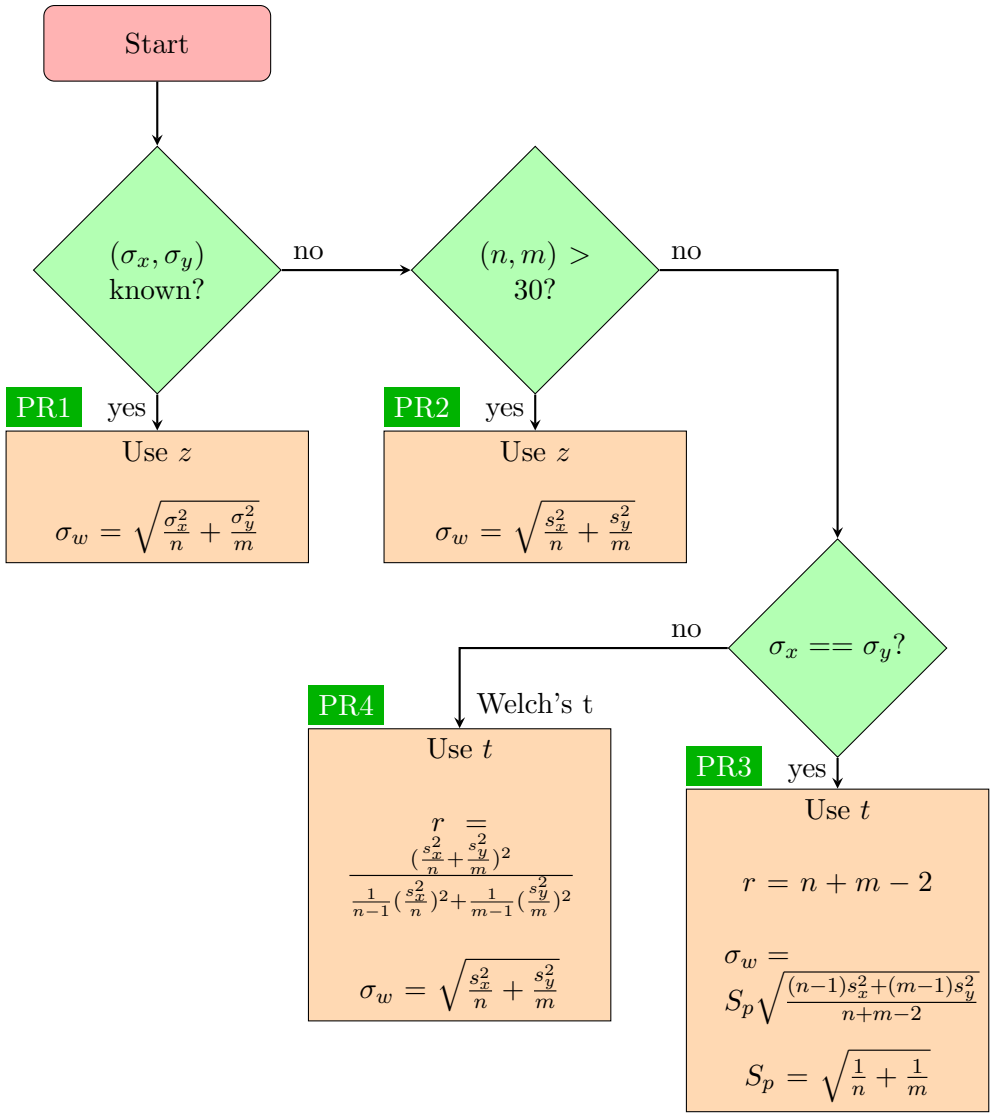
So if Z score of difference between sample means $Z = \dfrac{\overline{X} - \overline{Y}}{\sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}}}$, then the probability of making Type I error α is $P(Z \geq z_\alpha)$. This is depicted below.



So if the calculate Z score from the sample set values $(\overline{x}, \overline{y})$ exceeds z_α we could straight away **reject null hypothesis** because there is a stronger evidence that the alternate could be true. And we could derive similar Z score for μ decreasing or unequal, but it is much easier to directly tackling the problem than complicating the formula.

2.2 Visual Summary

Since we use the same components of confidence intervals in hypothesis testing, it helps to recall once the visual summary we have seen there.



2.3 Examples

2.3.1 σ unknown, sample sizes are high

As seen in visual summary (**PR2**), in this case, we still could use Z distribution, while we use sample set's unbiased standard deviations (s_x, s_y) in the place of (σ_x, σ_y) as best estimators. Since sample sizes are high, due to CLT, the sampling distribution would still be approximately normal, and our hypothesis testing approximately valid.

Lets assume we have two different ways to lose wieght, and we have to figure out which one is the most effective. We have 10000 people who received treatment A and their average loss is 10 pounds. The standard deviation of their loss is also 10 pounds. Lets consider a second treatment, Treatment B. We also applied it to 10000 people. The average loss in this case is 20 pounds and we also have a standard deviation of 20 pounds. Allowed false positive rate is 5%

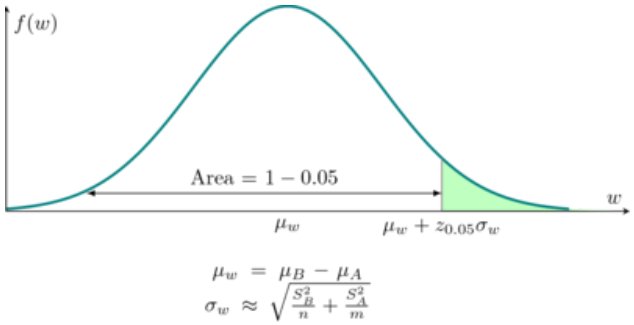
Also given is, null hypothesis $H_0 : \mu_A = \mu_B$

Alternate hypotheis is $H_a : \mu_B > \mu_A$

Solution

Whoa! Sample sizes are so high >>> 30. Also $W = B - A$ as we take the hint from alternate hypothesis. So we could rewrite equation 2.1 as per **PR2** in context as below

$$P\left(\frac{\overline{B} - \overline{A}}{\sqrt{\frac{S_B^2}{n} + \frac{S_A^2}{m}}} \geq z_\alpha\right) = \alpha \tag{2.2}$$



Given:

B: $n = 10000, \overline{b} = 20, s_B = 20$

A: $m = 10000, \overline{a} = 10, s_A = 10$

5% False positive rate would mean, we could be false 5% of the time while reality is true. This is type I error (rejecting null hypothesis, when null hypothesis is true in reality). Thus, $\alpha = 0.05$. So what would be $z_\alpha = z_{0.05}$?

```
In[7]: import scipy.stats as st
       z_a = st.norm.ppf(1-.05)  # as scipy is left tailed by default
       print(z_a)
```

```
1.6448536269514722
```

Therefore, $z_\alpha = z_{0.05} = 1.645$.

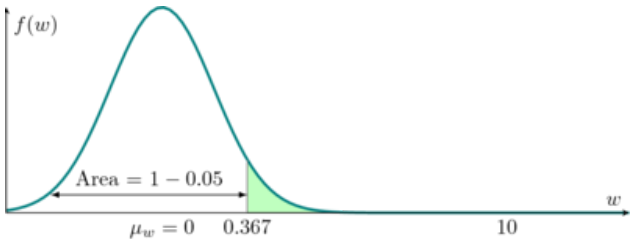
Let us try to create temporary critical region for W. Our given sample value $\overline{w} = 20 - 10 = 10$. We could say, if our hypothetical next sample means are if or above 10, we would reject the null hypothesis, and then wonder if that is the case, what would be our probability of making Type I error? Will we be in allowed limit of 0.05?

Note that, the critical region for permissible Type I probability of α starts at $(\mu_w + z_{0.05}\sigma_w)$. Since this is when null hypothesis is assumed, it is $z_{0.05}\sigma_w$ which is about 0.367. So any difference beyond 0.367, we could simply reject null hypothesis, that $\mu_A = \mu_B$.

```
In[8]: s_a, s_b, n, m = 10, 20, 10000, 10000
       from math import sqrt
       s_w = sqrt( (s_a**2)/m + (s_b**2)/n )
       print(s_w)
       print(s_w*z_a)
```

```
0.22360679774997896
0.3678004522900572
```

This situation is depicted below (not drawn at scale on x axis)



Now it would be evident beyond doubt that, we are well within permissible limits of 0.05 for making Type I error, which in fact is almost 0, to choose to reject null hypothesis, and suggest $\mu_B > \mu_A$. If we deployed equation 2.2 also we would have arrived at same conclusion. We could verify that as well. Rewriting 2.2,

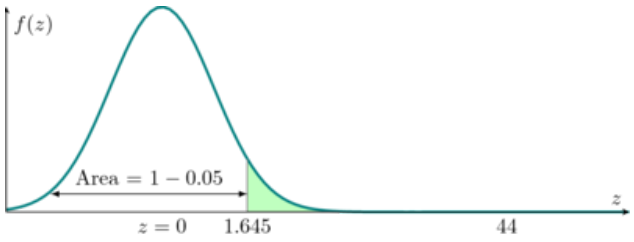
$$P\left(\frac{\bar{b} - \bar{a}}{\sqrt{\frac{s_B^2}{n} + \frac{s_A^2}{m}}} \geq 1.645\right) = 0.05$$

Recall, once the sample set is observed, there is no more probability about it. The calculated Z value is either above Z_α or not.

```
In[10]: b_bar, a_bar = 20, 10
        zs = (b_bar - a_bar)/s_w
        print(zs)
```

```
44.721359549995796
```

Our Z score 44 >> 1.645, so this again means, while the probability of Z score to be ≥ 1.645 was just 5%, provided null hypothesis was true. Looking at the rarity of this outcome if null hypothesis is true, it would be sane to conclude that this is a strong evidence that alternate hypothesis might be true. This strongly supports alternative hypothesis. This is depicted below (x axis not drawn at scale)



We are thus in a very good position to reject null hypothesis and support alternate hypothesis $H_a : \mu_B > \mu_A$

2.3.2 σ unknown, unequal and sample sizes are low

As seen in visual summary, we need to use **PR4**, that is Welch's t interval. Note the cumbersom calculation for calculating degrees of freedom. Some textbooks or platforms like Khan, [1] take

[1] https://www.khanacademy.org/math/ap-statistics/two-sample-inference/two-sample-t-test-means/v/two-sample-t-test-for-difference-of-means

conservative approach, that is, taking degrees of freedom $r = min(n, m)$. Nevertheless we will try to use Welch's and see what we get.

Independent random samples of 17 sophomores and 13 juniors attending a large university yield the following data on grade point averages.At the 5% significance level, do the data provide sufficient evidence to conclude that the mean GPAs of sophomores and juniors at the university differ?

Sample Data:

sophomor: $n = 17, \overline{x} = 2.84, s_x = 0.520$

juniors: $m = 13, \overline{y} = 2.9808, s_x = 0.3093$

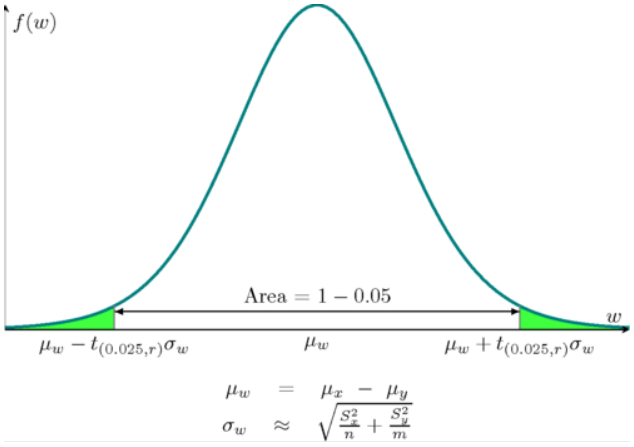
Solution:

The problem wonders if both the means **differ** so we would need to consider both tails.

Null hypothesis: $\mu_w = 0$ or $\mu_x = \mu_y$

Alternate hypothesis: $\mu_w \neq 0$ or $\mu_x \neq \mu_y$

$\alpha = 0.05$.



In welch's method, the degrees of freedom, r is the complicated one to calculate. It is given by integer part of below equation.

$$r = \frac{(\frac{s_x^2}{n} + \frac{s_y^2}{m})^2}{\frac{1}{n-1}(\frac{s_x^2}{n})^2 + \frac{1}{m-1}(\frac{s_y^2}{m})^2}$$

```
In[13]: s_x, s_y, n, m = 0.52, 0.3093, 17, 13

        num = ( s_x**2/n + s_y**2/m )**2
        den1 = (1/(n-1))*( s_x**2/n )**2
        den2 = (1/(m-1))*( s_y**2/m )**2
        den = den1+den2
        r = num/den
        print(r)
```

```
26.629678365237567
```

The degrees of freedom is the integer part of our result 26.629 which is $r = 26$. Let us then calculate the 't' score for our significance level α,

```
In[14]: from scipy import stats
        ts = stats.t.ppf(0.025, 26) # return value is left tailed by default..
        print(ts)
```

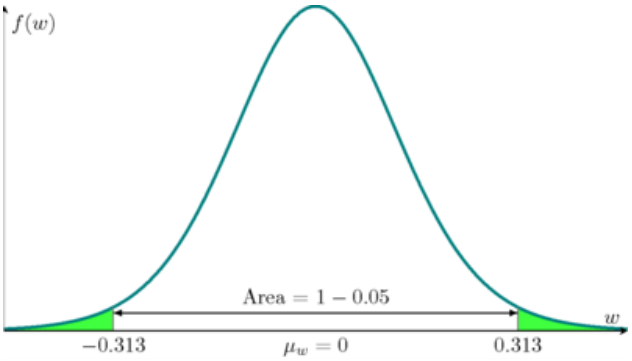
```
-2.0555294386428713
```

Therefore, $t_{(\alpha/2,r)} = t_{(0.025,26)} = 2.055$. We could now calculate the limits above or below which Type I error is allowed. Assuming $\mu_w = 0$ due to null hypothesis, $t_{(0.025,26)}\sigma_w$ should give us the limits.

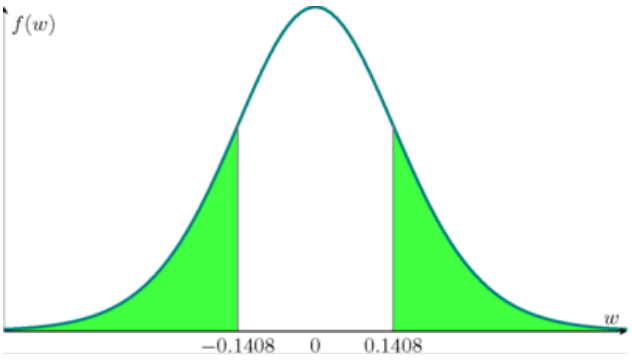
```
In[15]: from math import sqrt
        s_w = sqrt( (s_x)**2/n + (s_y)**2/m  )
        print(s_w, s_w*ts)
```

```
0.15252817156896606 -0.31352614688238034
```

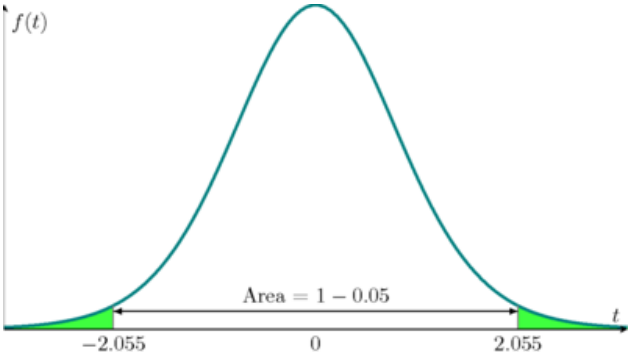
Thus our critical region for given α would be ± 0.313. Our situation could be depicted as below. We are allowed to reject null hypothesis, if our sample set mean difference is above 0.313 or below -0.313, with $\alpha = 0.05$ probability of making Type I error.



The difference of sample means we got is $\overline{x} - \overline{y} = 2.84 - 2.9808 = -0.1408$. This is far above -0.313, so we **cannot reject null hypothesis**. Taking ± 0.1408 as critical region would increase of probability of Type I error α enormously as shown below.



We could have also taken the difference the other way $\overline{y} - \overline{x} = 2.9808 - 2.84 = 0.1408$, and we still would have arrived at same conclusion because we are interested in only if the sample means of two sampling distributions differ or not (that is why two tails taken in above diagram).Also we could arrive at the same conclusion via 't' values only if we already know $t_{\alpha/2,r}$. We indeed calculated that earlier as 2.055. This means, in units of 't', critical region allowed is ± 2.055 beyond which we are allowed to make Type I error, whose total probability in critical region would be 0.05. This is depicted below.



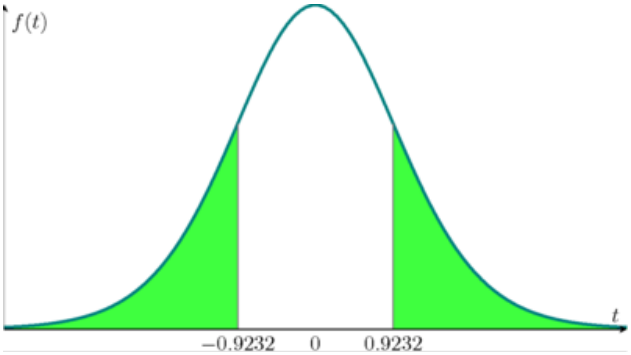
$$
\begin{aligned}
P\Big(-t_{(\alpha/2,r)} \leq \frac{W - \mu_w}{\sigma_w} \leq t_{(\alpha/2,r)} \Big) &= 1 - \alpha \\
2P\Big(\Big| \frac{W - \mu_w}{\sigma_w} \Big| \geq |t_{(\alpha/2,r)}| \Big) &= \alpha \\
P\Big(\Big| \frac{W - \mu_w}{\sigma_w} \Big| \geq |t_{(\alpha/2,r)}| \Big) &= \frac{\alpha}{2} \\
P\Big(\Big| \frac{W - \mu_w}{\sigma_w} \Big| \geq |t_{(0.025,r)}| \Big) &= 0.025 \\
P\Big(\Big| \frac{W - \mu_w}{\sigma_w} \Big| \geq |t_{(0.025,26)}| \Big) &= 0.025 \\
P\Big(\Big| \frac{W - \mu_w}{\sigma_w} \Big| \geq 2.055 \Big) &= 0.025 \\
P\left(\left| \frac{(\overline{X} - \overline{Y}) - (\mu_{\overline{x}} - \mu_{\overline{y}})}{\sqrt{\frac{s_x^2}{n} + \frac{s_y^2}{m}}} \right| \geq 2.055 \right) &= 0.025 \\
P\left(\left| \frac{(\overline{X} - \overline{Y})}{\sqrt{\frac{s_x^2}{n} + \frac{s_y^2}{m}}} \right| \geq 2.055 \right) &= 0.025
\end{aligned}
$$

When the sample set is observed, we could check if we are in critical region or not by calculating its t score.

$$t = \left| \frac{\overline{x} - \overline{y}}{\sqrt{\frac{s_x^2}{n} + \frac{s_y^2}{m}}} \right| = \left| \frac{2.84 - 2.9808}{0.1525} \right| = 0.9232$$

Our t score is 0.9232. This is well outside the critical region towards the null hypothesized zero mean difference, so if we assume this as critical region, we would be making high Type I error,

beyond 0.05 as depicted below.



So our conclusion is similar like earlier. We **cannot reject the null hypothesis**.

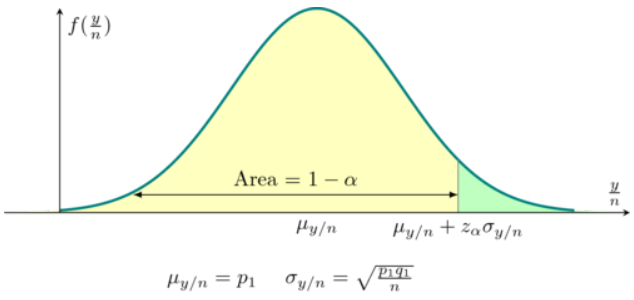
Chapter 3

Testing hypotheses about a proportion

As seen earlier in confidence intervals, using Wald's method for the sample proportions do not yield promising results as widely believed. So we will only stick to case when conditions are met to make the sampling distribution normalcy good enough.

3.1 When sample sizes are high

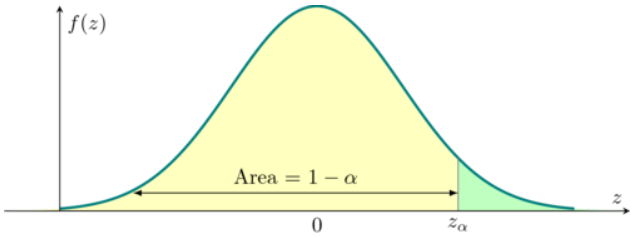
Suppose that we have a normal **sampling distribution** described by random variable $\frac{Y}{n} = N\left(p_1, \frac{p_1 q_1}{n}\right)$ created from a population distribution which is a Bernoulli distribution with mean p_1 and standard deviation $p_1 q_1$. Note that Y represents the sum of *successes* in a sample set, and thus $\frac{Y}{n}$ represents sample proportions. For example, for any *kth* sample set of $\frac{Y}{n}$, we calculate sample proportion statistic, $\frac{Y_k}{n} = \frac{1}{n}\sum_{i=1}^{n} Y_{ki}$, where Y_{ki} is ith sample in kth sample set of sampling distribution described by $\frac{Y}{n}$. If α is the significance level, then we could derive the conditions for hypothesis testing as follows. Below is our sampling distribution as null hypothesis, with α as significance level. This is for alternate hypothesis being $H_a : \mu > \mu_{y/n}$ so we consider the right tail area. One could try similar approach for left or both tails depending on if H_a is $H_a : \mu < \mu_{y/n}$ or $H_a : \mu \neq \mu_{y/n}$ respectively.



The significance level α, corresponds to the rest of $1 - \alpha$ area, that is green area as shown above.

$$P\left(\frac{Y}{n} \geq \mu_{y/n} + z_\alpha \sigma_{y/n}\right) = \alpha$$
$$\therefore P\left(\frac{\frac{Y}{n} - \mu_{y/n}}{\sigma_{y/n}} - \geq z_\alpha\right) = \alpha$$
$$P\left(\frac{\frac{Y}{n} - p_1}{\sqrt{\frac{p_1 q_1}{n}}} - \geq z_\alpha\right) = \alpha$$

Let the z score be, $z = \dfrac{\frac{Y}{n} - \mu_{y/n}}{\sigma_{y/n}}$, then $P(z \geq z_\alpha) = \alpha$



Our allowed critical region in sampling distribution is $(\mu_{y/n} + z_\alpha \sigma y/n, \infty)$, where the probability of making Type I error is α. Our allowed critical region in *standardized* sampling distribution would be (z_α, ∞). So if our z score falls within (z_α, ∞), we could reject the null hypothesis. This is also equivalent to saying, if our sample set proportion y/n falls within $(\mu_{y/n} + z_\alpha \sigma y/n, \infty)$, we could reject the null hypothesis.

Conditions

- One of the main condition to apply hypothesis testing to sample proportions is to ensure the sampling distribution is normal. This is usually ensured when $(np, nq) > 10$ if not population is already normal.
- You see, unlike sample means, there was no σ not known case in proportions, because we are testing against hypothesized mean p_1, so the associated σ would be simply $\sqrt{p_1 q_1}$. So p_1 is a pre requisite against which we need to test, so that is usually given or implicit in case of one proportion, so no σ unknown case arises here.

Example

It was claimed that many commercially manufactured dice are not fair because the "spots" are really indentations, so that, for example, the 6-side is lighter than the 1-side. To test, in an experiment, several such dice were rolled, to yield a total of $n = 8000$ *observations, out of which 6 resulted,* 1389 *times. Is there a significant evidence that dice favor a 6 far more than a fair die would? Assume* $\alpha = 0.05$

Solution:

Let us assume null hypothesis as a fair die, nothing to doubt about. The probability of getting a 6 in fair die is $p = 1/6$. So

$H_0 : \mu_{y/n} = p = 1/6$

$H_a : \mu_{y/n} = p \neq 1/6$

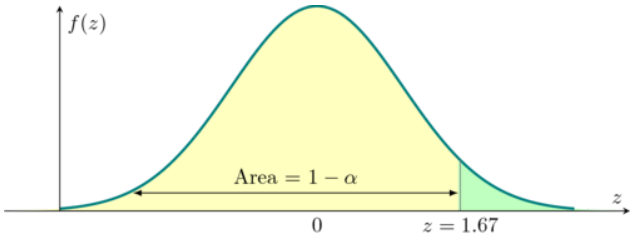
We have a sample size of $n = 8000$, so $np = (8000)(1/6) = 1333 >> 10, nq = (8000)(5/6) = 6666 >> 10$, so our normal condition is met. If we continue with sample sets of this size, we would get a good normal sampling distribution $\dfrac{Y}{n}$

Our z score is $z = \dfrac{\frac{Y}{n} - p_1}{\sqrt{\frac{p_1 q_1}{n}}} = \dfrac{(1389/5000) - (1/6)}{\sqrt{\frac{(1/6)(5/6)}{8000}}}$

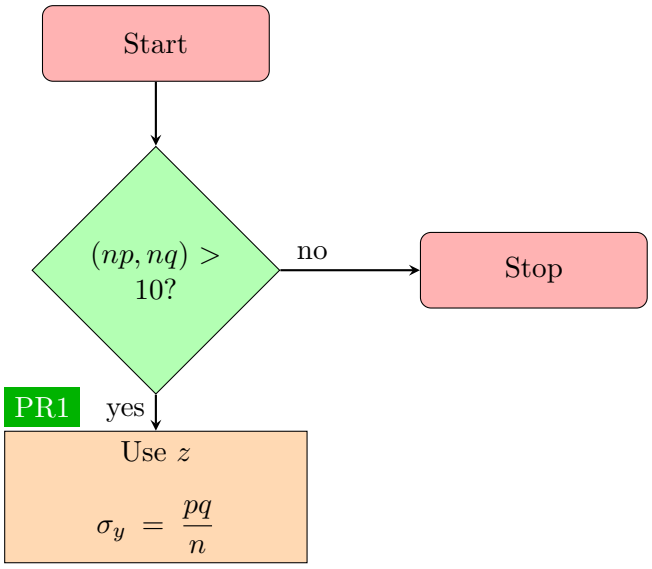
```
In[11]: Y,n,p_1,q_1 = 1389, 8000, 1/6,5/6
        num = (Y/n) - (p_1)
        from math import sqrt
        den = sqrt(p_1*q_1/n)
        zs = round(num/den, 4)
        print(zs)
```

```
1.67
```

Our **allowed** critical region starts from $z_{0.05} = 1.645$. The z score $z = 1.67$ is greater that that, which means, if we select this sample set as critical region's starting point, our probability of making Type I error is smaller than allowed $\alpha = 0.05$. So we **reject the null hypothesis** , thus suggesting there is stronger evidence for alternate H_a.



3.2 Conditions Summary



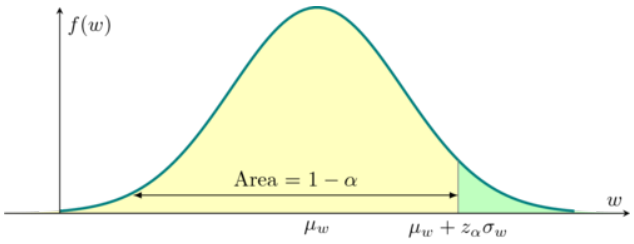
Chapter 4

Testing the difference between two proportions

4.1 When Successes and Failures are high enough

4.1.1 (p_1, p_2) known

Suppose that we are interested in comparing two approximately normal sampling distributions described by random variables $\frac{Y_1}{n_1} = N\Big(p_1, \frac{p_1q_1}{n_1}\Big)$ and $\frac{Y_2}{n_2} = N\Big(p_2, \frac{p_2q_2}{n_2}\Big)$, created from population distributions which are Bernoulli distributions. Note that Y_1 represents the sum of *successes* in a sample set, and thus $\frac{Y_1}{n_1}$ represents sample proportions. For example, for any *kth* sample set of $\frac{Y_1}{n_1}$, we calculate sample proportion statistic, $\frac{Y_{1k}}{n_1} = \frac{1}{n}\sum_{i=1}^{n} Y_{1ki}$, where Y_{1ki} is ith sample in kth sample set of sampling distribution described by $\frac{Y_1}{n_1}$. Similarly for $\frac{Y_2}{n_2}$. Then, if no of success and failures are high enough [1], that is at least > 10, as a general rule, we could assume that the random variable $W = \frac{Y_1}{n_1} - \frac{Y_2}{n_2}$ has approximately normal distribution $W = N(p_w, \sigma_w^2)$ where $p_w = p_1 - p_2$ and $\sigma_w = \sqrt{\frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2}}$ and has shown below, before standardization to Z. We *destandradize* from Z, because, each α could be linked to corresponding z score, which further could be linked to actual w or x axis in question.



$\mu_w = p_1 - p_2 \quad \sigma_w = \sqrt{\frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2}}$

[1] https://www.khanacademy.org/math/ap-statistics/two-sample-inference/two-sample-z-test-proportions/v/hypothesis-test-for-difference-in-proportions

The significance level α, corresponds to the rest of $1-\alpha$ area, that is green area as shown above.

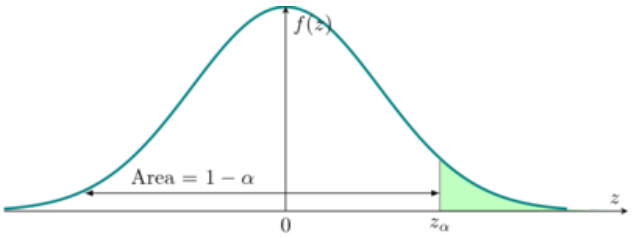
$$P(W \geq \mu_w + z_\alpha \sigma_w) = \alpha$$
$$\therefore P\Big(\frac{W - \mu_w}{\sigma_w} \geq z_\alpha\Big) = \alpha$$
$$P\left(\frac{(\frac{Y_1}{n_1} - \frac{Y_2}{n_2}) - (p_1 - p_2)}{\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}} \geq z_\alpha\right) = \alpha$$

Typically, null hypothesis is $p_1 = p_2$, so, assigning it to a common p, i.e $p1 = p2 = p$,..

$$P\left(\frac{\frac{Y_1}{n_1} - \frac{Y_2}{n_2}}{\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}} \geq z_\alpha\right) = \alpha$$
$$P\left(\frac{\frac{Y_1}{n_1} - \frac{Y_2}{n_2}}{\sqrt{pq(\frac{1}{n_1} + \frac{1}{n_2})}} \geq z_\alpha\right) = \alpha$$

Thus the z score for given sample data would be $z = \dfrac{\frac{Y_1}{n_1} - \frac{Y_2}{n_2}}{\sqrt{pq(\frac{1}{n_1} + \frac{1}{n_2})}}$

So if our alternate hypothesis is that $H_a : p_1 > p_2$, then we could calculate Z score as above and if that is beyond z_α we could reject null hypothesis.



We could simlarly derive for $H_a : p_1 < p_2$, and $H_a : p_1 \neq p_2$.

4.1.2 (p_1, p_2) unknown

Of course, the above section was for pedagogical purposes, to illustrate the concept. In reality, the individual p_1 and p_2 are not hypothesized typically, and usually compared only to see if there is significant evidence that if one is greater/smaller/different from the other. In which case we simply could use our best estimator $\hat{p}$ for calculating standard deviation in place of p. There are usually two ways, here.

Way 1: Calculate weighted p

This is usually given as $\hat{p} = \dfrac{Y_1 + Y_2}{n_1 + n_2}$. And then, 4.1 beocmes

$$P\left(\frac{\frac{Y_1}{n_1}-\frac{Y_2}{n_2}}{\sqrt{\hat{p}\hat{q}(\frac{1}{n_1}+\frac{1}{n_2})}}\geq z_\alpha\right)=\alpha \tag{4.1}$$

At the time of this writing, I could not find a derivation for the same, so over to next one.

Way 2: Use sample $\hat{p_1}, \hat{p_2}$

This is straight forward approach directly from 4.1, with $p_1 = p_2$

$$P\left(\frac{\hat{p_1}-\hat{p_2}}{\sqrt{\frac{\hat{p_1}\hat{q_1}}{n_1}+\frac{\hat{p_2}\hat{q_2}}{n_2}}}\geq z_\alpha\right)=\alpha \tag{4.2}$$

Tips

- Equation 4.2 would be the one mostly used for almost any of difference of proportions problems (of course adapted to right or left or both tails as needed)

Example

A machine shop that manufactures toggle levers has both a day and a night shift. A toggle lever is defective if a standard nut cannot be screwed onto the threads. Let p1 and p2 be the proportion of defective levers among those manufactured by the day and night shifts, respectively. We shall test the null hypothesis, $H_0 : p_1 = p_2$, against a two-sided alternative hypothesis based on two random samples, each of 1000 levers taken from the production of the respective shifts.

(a) Define the test statistic and a critical region that has an $\alpha = 0.05$ significance level. Sketch a standard normal pdf illustrating this critical region.

(b) If $y_1 = 37$ and $y_2 = 53$ defectives were observed for the day and night shifts, respectively, calculate the value of the test statistic. Locate the calculated test statistic on your figure in part (a) and state your conclusion.

This example was taken from exercise *8.3-11* in Robert et al. [1]

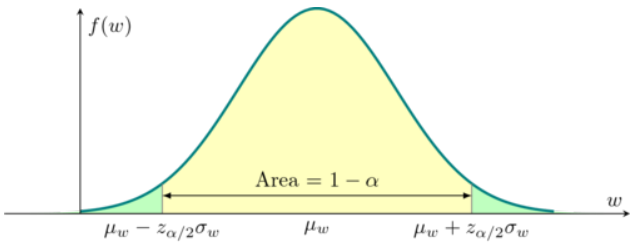
Solution:

Day: $y_1 = 37, n_1 = 1000, \hat{p_1} = \frac{Y_1}{n_1} = \frac{37}{1000} = 0.037$

Night: $y_2 = 53, n_2 = 1000, \hat{p_2} = \frac{Y_2}{n_2} = \frac{53}{1000} = 0.053$

(a)

It is said, "two sided alternative hypothesis", so below is our required test statistic. note we have used our best estimators $(\hat{p_1}, \hat{p_2})$ so result is only approximate.



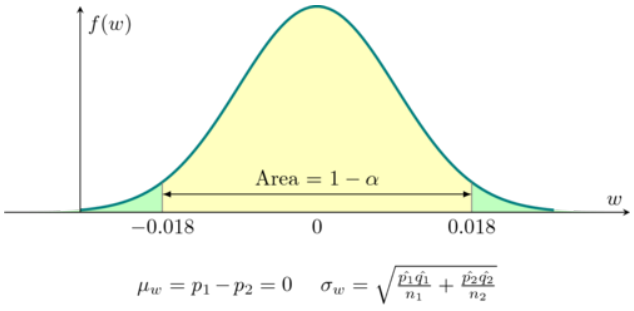
$$\mu_w = p_1 - p_2 = 0 \quad \sigma_w = \sqrt{\frac{\hat{p_1}\hat{q_1}}{n_1}+\frac{\hat{p_2}\hat{q_2}}{n_2}}$$

Calculating the values σ_w, we could arrive at $w = \mu_w \pm z_{\alpha/2}\sigma_w = \pm z_{\alpha/2}\sigma_w$ value beyond which we could define critical region α. Since it is double tailed, we already know $z_{0.025} = 1.96$.

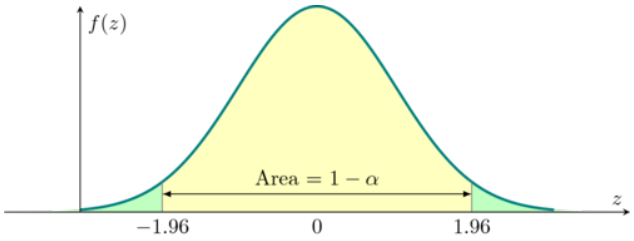
$\sigma_w = \sqrt{\frac{\hat{p_1}\hat{q_1}}{n_1} + \frac{\hat{p_2}\hat{q_2}}{n_2}} = \sqrt{\frac{(0.037)(1-0.037)}{1000} + \frac{(0.053)(1-0.053)}{1000}}$

```
In[14]: p_1_hat, q_1_hat, p_2_hat, q_2_hat, n_1, n_2 = 0.037, 1-0.037, 0.053, 1-0.053, 1000,
        1000
        z_0025 = 1.96

        from math import sqrt
        s_w = sqrt( (p_1_hat*q_1_hat/n_1) + (p_2_hat*q_2_hat/n_2) )
        print(s_w*z_0025)
```

```
0.018157472158866164
```



We could alreay take a call on our null hypothesis, Our $\hat{p_1} - \hat{p_2} = 0.037 - 0.053 = 0.016 < 0.0018$, so we cannot reject H_0. Our standardized test statistic would be simply z distribution as below.



(b)

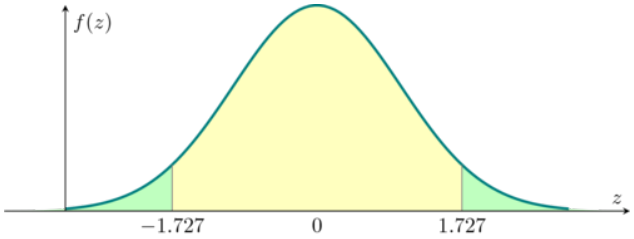
We have already kinda finished the solution, but for question's sake we could complete it fully by calculating the Z score.

Using 4.2, $z = \dfrac{\hat{p_1} - \hat{p_2}}{\sqrt{\frac{\hat{p_1}\hat{q_1}}{n_1} + \frac{\hat{p_2}\hat{q_2}}{n_2}}}$

```
In[19]: num = p_1_hat - p_2_hat
        den = s_w
        num/den
```

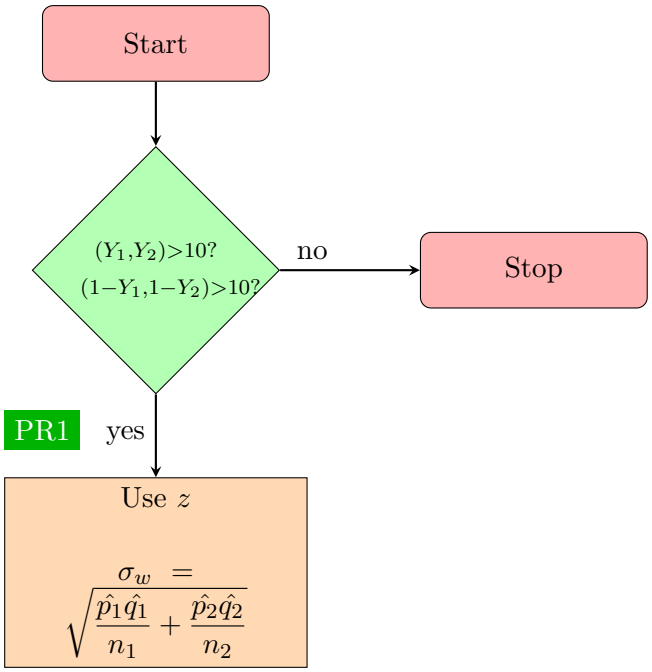
```
Out[19]: -1.7271126578424703
```

Being double tailed operation, our z score is thus ± 1.727. And since $\pm 1.727 < \pm 1.96$, we again **cannot reject null hypothesis** because then our probability of making Type I error would be more than allowed limit of $\alpha = 0.05$. Our standardized test statistic, with $\pm z_{\alpha/2} = \pm 1.727$ is shown below.



Though visibly not clear, one could use z table to find that $z_{1.727}$ takes more area than 0.05 which corresponds to $z_{1.96}$. Thus we conclude our answer.

4.2 Conditions Summary



Bibliography

[1] Robert, Elliot, and Dale. *Probability and Statistical Inference*. Pearson, 9th edition, 2015. URL http://www.nylxs.com/docs/thesis/sources/Probability%20and%20Statistical%20Inference%209ed%20%5B2015%5D.pdf.