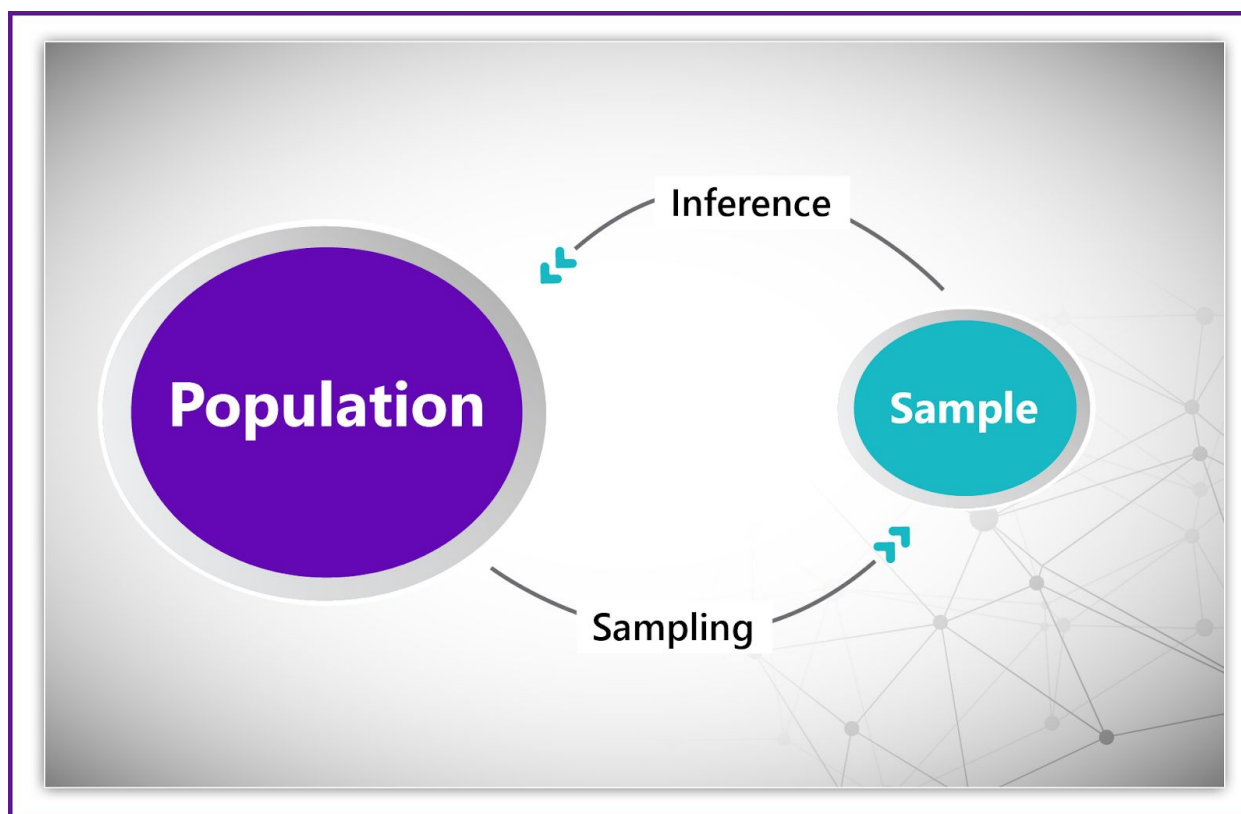


What is Inferential Statistics?

The role of a data scientist is to answer questions on data and give insights. Think of a data scientist trying to do research on fellow data scientists. If we want to research about the Data Scientists some of the examples of research questions we would like to ask are as follows

- What is the average salary of a Data Scientist?
- What percentage of Data Scientists hold a doctorate?
- Are data scientists paid more than data engineers?

To answer the above questions accurately we will have to conduct a census and ask each and every data scientist the questions, but as you may have realized we cannot ask every person in this `population` of data scientists. Instead what we can do is pick a random `sample` out of the population and ask these questions and `make inference` from the response received.

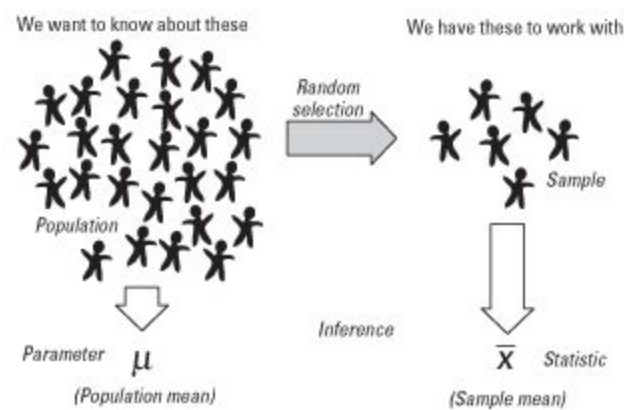


For the above-mentioned research questions, the population comprises of all the Data Scientists on planet earth. The key point here for the data scientist is deciding whether the observation on the small sample hold for the complete data or not. For example, if a data scientist's average salary based on the sample is found to be \$100,000 does it hold true for the entire population of data scientists? The inference is all about finding answers to such questions.

Statistical inference is the process of making a judgment about a population based on sampling properties. An important aspect of statistical inference is using estimates to approximate the value of an unknown population parameter.

Let's try and understand the words "Sample" & "Population" and break down the above statement using an example.

Leading up to U.S. presidential elections it could be very useful to know the political leanings of every single eligible voter, but surveying every voter is not feasible. Instead, we could poll some subset of the population, such as a thousand registered voters, and use that data to make inferences about the population as a whole.



This "subset" of the population is nothing but the sample data. We carry out various tests on the Sample to gain insight into the larger population out there!

Statistical inference helps us to answer various such questions based on the samples drawn from the entire population. There are two broad areas of statistical inference that we are going to discuss in this chapter viz. Statistical Estimation and Hypothesis Testing.

Statistical Estimation

Following are some of the questions that we can answer using Statistical Estimation

- What is the average salary of a Data Scientist in the bay area?
- What proportion of Data Scientists holds a doctorate degree?

Here as you can see we are trying to find the values of population parameters based on the sample.

A parameter is a descriptive measure of the population. - Example: Population mean, Population variance, etc. A statistic is a descriptive measure of the sample. - Example: Sample mean, Sample variance, etc.

Population Parameter	Sample Statistic
Population Mean - μ	Sample mean - \bar{x}

Population Std Dev - σ	Sample Std Dev - s
Population Variance - σ^2	Sample Variance - s^2

Hypothesis Testing

Following are some of the questions that we can answer using Hypothesis Testing

- Is the average salary of Data Scientists in the Bay Area and Montreal the same?
- Is the salary of the Data Scientist and his Education independent of each other?
- Is the average salary of Data Scientists in Montreal greater than \$100k

Here you can see that we are trying to answer certain questions regarding the population. Most of the times we have certain assumptions about population parameters, hypothesis testing is a way to decide whether these assumptions stand true based on the data from a sample.

Central Limit Theorem (CLT)

Statement of CLT

The central limit theorem (CLT) is a statistical theory that states that given a sufficiently large sample size from a population with a finite level of variance, the `mean` of all samples from the same population will be approximately equal to the `mean` of the population. Furthermore, on taking multiple samples from the same population, the mean of the individual samples will form a normal distribution pattern, with all variances being approximately equal to the variance of the population divided by each sample's size.

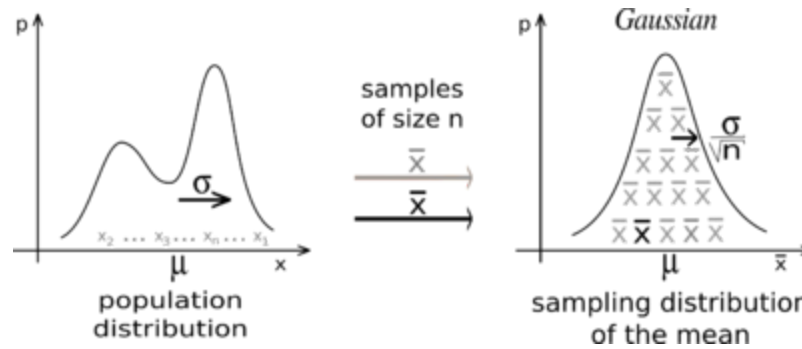
Seems esoteric? Don't worry. First, let's understand the central limit through the video [Central Limit Theorem](#)

Now, let's break this statement into a series of steps:

Simplifying CLT

- Take a random sample (S_1) of size (n) from your data/population
- Take the average of this sample (\bar{x}_1)
- Take another sample (S_2) of the same size and calculate its average (\bar{x}_2)
- In this way calculate $\bar{x}_3, \bar{x}_4, \dots$ etc
- Plot all the sample averages using a histogram i.e. plot $\bar{x}_1, \bar{x}_2, \bar{x}_3, \dots, \bar{x}_n$

The shape that you observe will look like a normal distribution bell curve. The image below is a pictorial representation of what the Central Limit Theorem does.



Pythonic implementation of CLT

Lets now see this theorem in action with Python with the underlying population distributions as Flat, Exponential, and Beta:

```
# provides the capability to define a function with partial arguments
from functools import partial
import numpy as np
import matplotlib.pyplot as plt
%matplotlib inline

N = 1000 # number of times n samples are taken. Try varying this number.
nobb = 101 # number of bin boundaries on plots
n = np.array([1,2,3,5,10,100,200]) # number of samples to average over

exp_mean = 3 # mean of exponential distribution
a,b = 0.7,0.5 # parameters of beta distribution

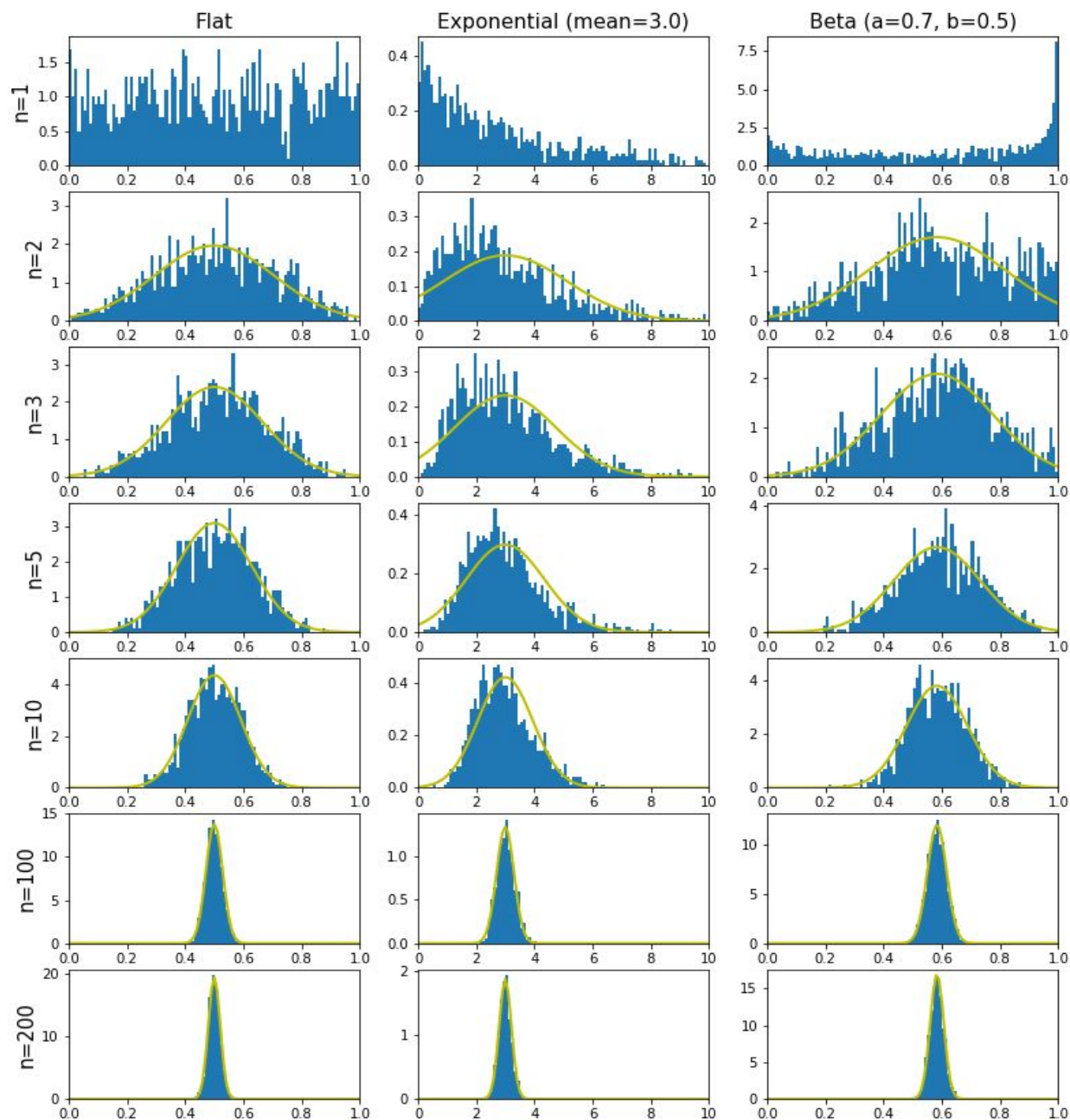
dist = [
    partial(np.random.random), partial(np.random.exponential, exp_mean), partial(np.random.beta, a, b)
]
title_names = ["Flat", "Exponential (mean=%.1f)" % exp_mean, "Beta (a=%.1f, b=%.1f)" % (a, b)]
drange = np.array([[0,1],[0,10],[0,1]]) # ranges of distributions
means = np.array([0.5, exp_mean, a/(a+b)]) # means of distributions
var = np.array([1/12, exp_mean**2, a*b/((a+b+1)*(a+b)**2)]) # variances of distributions

binrange = np.array([np.linspace(p,q,nobb) for p,q in drange])
ln,ld = len(n),len(dist)
plt.figure(figsize=((ld*4)+1, (ln*2)+1))

for i in range(ln): # loop over the number of n samples to average over
    for j in range(ld): # loop over the different distributions
        plt.subplot(ln,ld,i*ld+1+j)
        plt.hist(np.mean(dist[j]((N,n[i])),1),binrange[j],normed=True)
        plt.xlim(drangle[j])
        if j==0:
            plt.ylabel('n=%i' % n[i],fontsize=15)
        if i==0:
            plt.title(title_names[j], fontsize=15)
        else:
            clt=(1/(np.sqrt(2*np.pi*var[j]/n[i]))) * np.exp(-(((binrange[j]-means[j])**2)*n[i]/(2*var[j])))
```

```
plt.plot(binrange[j],clt,'y',linewidth=2)
plt.show()
```

The output of the above code snippet is the image below. You can clearly observe that as the sample size (n) increases, the distribution of the sample mean approaches a more normal curve, irrespective of the underlying distribution.



Important pointers

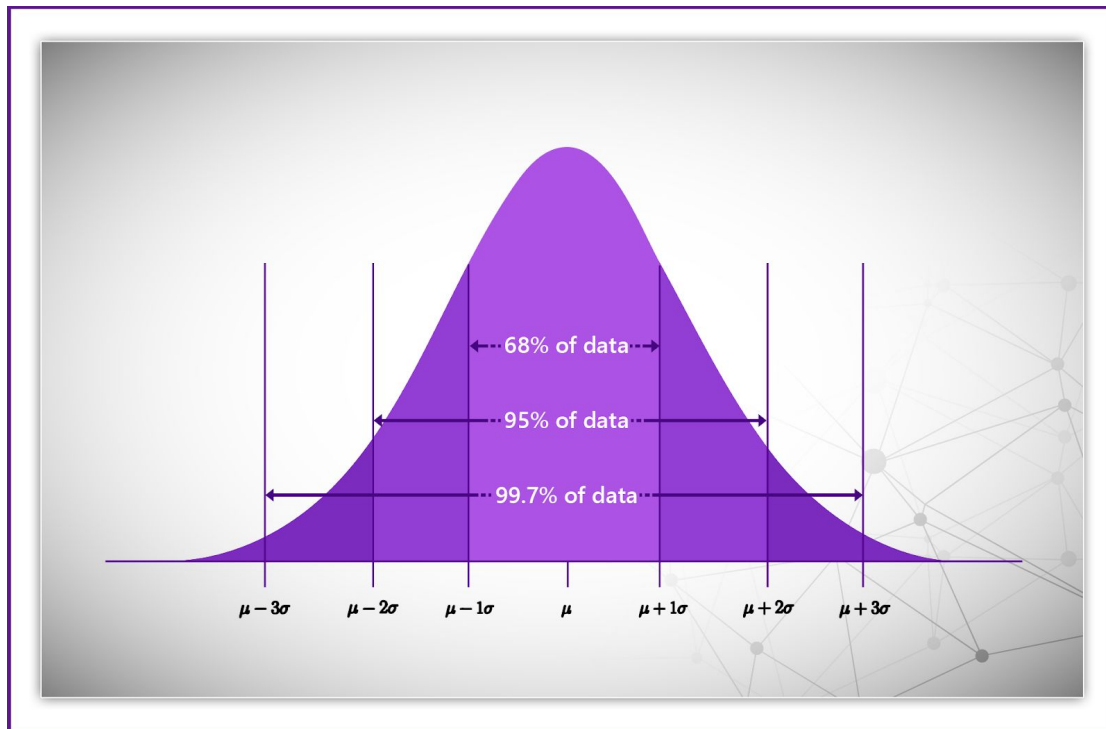
- The mean of a sample of data will be closer to the mean of the overall population in question as the sample size increases, notwithstanding the actual distribution of the data, and whether it is normal or non-normal.
- Sample sizes equal to or greater than 30 are considered sufficient for this theorem to hold, meaning the distribution of the sample means is fairly normally distributed.

Applications of Central Limit Theorem

Perhaps one of the most misunderstood things in statistics is the Central Limit Theorem. The central limit theorem can be used to help evaluate data from various distribution patterns. Using this theorem we can apply statistical methods that would otherwise only apply to normal distributions of data.

The first question that comes to the mind is: Okay, I have a bell curve now, who cares?

Once you have a normal bell curve, I now know something very powerful. Known as the 68,95,99 rule, I know that 68% of my sample is going to be within one standard deviation of the mean. 95% will be within 2 standard deviations and 99.7% within 3. It can then be used to calculate something called a p-value which will help us in making inferences and forms the final step in hypothesis testing.



Example:

The mean is 200 and standard deviation is 24 of a population. Sample size is 30. What is the mean of sampling distribution?

Ans: 200

Explanation:

The central limit theorem states that the mean of the sampling distribution of the mean is the mean of the population from which the scores were sampled.

Variance and Degrees of Freedom

Estimates first!

Recall that standard deviation for population is denoted by σ while the sample standard deviation is given by $s = \frac{\sigma}{\sqrt{n}}$. This sample standard deviation s is said to be an estimate of the population standard deviation σ and can be used as a substitute for the same in case we don't know σ . You will learn more about **estimates** in the next topic. Similarly, the sample mean (\bar{x}) can be considered as an estimate of the population mean (μ).

What is the degree of freedom and how is it related to estimates?

Some estimates are based on more information than others. For example, an estimate of the variance based on a sample size of 200 is based on more information than an estimate of the variance based on a sample size of 10. The concept of degrees of freedom is central to the principle of estimating statistics of populations from samples of them. Degrees of freedom is commonly abbreviated to **df**.

You can think of **df** as a mathematical restriction that needs to be put in place when estimating one statistic from an estimate of another. **The degrees of freedom (df) of an estimate is the number of independent pieces of information on which the estimate is based.**

Examples calculating degree of freedom

As an example, let's say that you know that the mean height of all the employees at your company is 6 feet and wish to estimate the variance of their heights. Let's consider separate instances and look at the outcomes:

- **Case I:** We randomly sample one employee out of all and find that the employee's height is 8 feet. Recall that the variance is defined as the mean squared deviation of the values from their population mean. We can compute the squared deviation of our value of 8 from the population mean of 6 to find a single squared deviation from the mean. This single squared deviation from the mean, $(8 - 6)^2 = 4$, is an estimate of the mean squared deviation for all the employees. Therefore, based on this sample of one, we would estimate that the population variance is 4.

Since this estimate is based on a single piece of information, therefore it has 1 df. If we sampled another employee and obtained a height of 5 feet, then we could compute a second estimate of the variance, $(5 - 6)^2 = 1$. We could then average our two estimates (4 and 1) to obtain an estimate of 2.5. **Since this estimate is based on two independent pieces of information, it has two degrees of freedom.**

NOTE: The two estimates are independent because they are based on two independently and randomly selected employees.

- **Case II:** Usually it is pretty rare that we know the population mean when we are estimating the variance. Instead, we have to first estimate the population mean (μ) with the sample mean (\bar{x}). The process of estimating the mean affects our degrees of freedom as described below.

Returning to your problem of estimating the variance in the employee heights, let's assume you do not know the population mean (μ) and therefore we have to estimate it from the sample mean (\bar{x}). You have sampled two employees and found that their heights are 8 and 5. Therefore \bar{x} , your estimate of the population mean, is $\bar{x} = (8 + 5)/2 = 6.5$.

You can now compute two estimates of variance:

$$\text{Estimate 1} = (8 - 6.5)^2 = 2.25$$

$$\text{Estimate 2} = (5 - 6.5)^2 = 2.25$$

Are these two estimates independent? The answer is **no** because each height contributed to the calculation of \bar{x} . Since the first employee's height of 8 influenced \bar{x} , it also influenced Estimate 2. If the first height had been, for example, 10, then \bar{x} would have been 7.5 and Estimate 2 would have been $(5 - 7.5)^2 = 6.25$ instead of 2.25.

The important point is that the two estimates are not independent and therefore we do not have two degrees of freedom. Another way to think about the non-independence is to consider that if you knew the mean and one of the scores, you would know the other score. For example, if one score is 5 and the mean is 6.5, you can compute that the total of the two scores is 13 and therefore that the other score must be $13 - 5 = 8$.

In general, the degrees of freedom for an estimate is equal to the number of values minus the number of parameters estimated en route to the estimate in question. In the employee's example, there are two values (8 and 5) and we had to estimate one parameter (μ) on the way to estimate the parameter of interest (σ). Therefore, the estimate of variance has $2 - 1 = 1$ degree of freedom. If we had sampled 12 employees instead, then our estimate of the variance would have had 11 degrees of freedom. Therefore, the degrees of freedom of an estimate of variance is equal to $n - 1$, where n is the number of observations.

And this is the reason why the denominator while calculating sample variance is $n - 1$ instead of n so that it provides an unbiased estimator.

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n - 1}$$

Point Estimates

In the previous topic, you came across the term estimates. They are approximations of a statistic as a parameter when the actual value of the parameter is unknown. Now let's learn about the first type of estimates i.e. point estimates.

What is a point estimate?

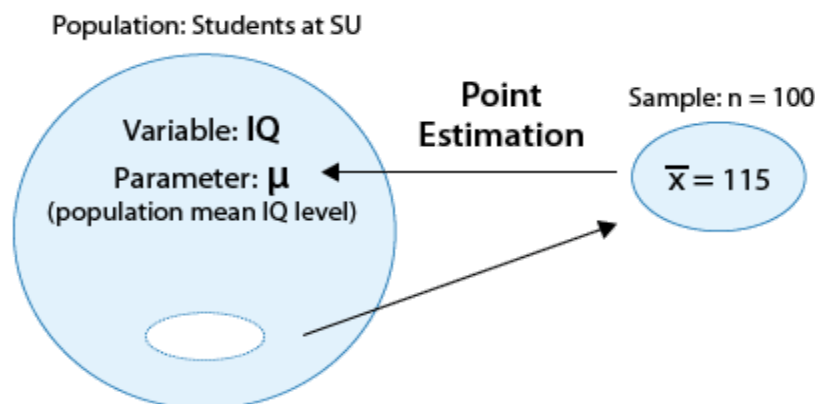
In point estimation, we estimate an unknown parameter using a single number that is calculated from the sample data. Here, a single value estimates the population parameter. The sample mean (\bar{x}) can be considered as a point estimate of the population mean (σ). Let's take a more concrete example to understand point estimation.

Example of point estimate

Consider the following situation:

Suppose that we are interested in studying the IQ levels of students at State University (SU). In particular (since IQ level is a quantitative variable), we are interested in estimating μ , the mean IQ level of all the students at SU.

A random sample of 100 SU students was chosen, and their (sample) mean IQ level was found to be 115. If we wanted to estimate μ , the population mean IQ level, by a single number based on the sample, it would make intuitive sense to use the corresponding quantity in the sample, the sample mean. We say that 115 is the point estimate for μ , and in general, we'll always use as the point estimator for μ . (Note that when we talk about the specific value (115), we use the term estimate, and when we talk in general about the statistic, we use the term estimator. The following figure summarizes this example:

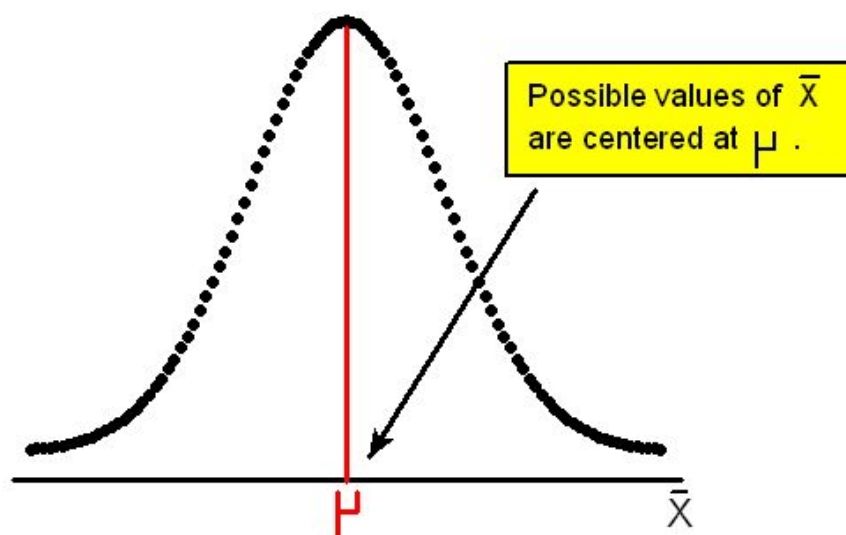


Intuitive explanation of point estimate

You may feel that since it is so intuitive, you could have figured out point estimation on your own as our intuition tells us that the best estimator for μ should be \bar{x} .

Probability theory does more than this; it actually gives an explanation of why \bar{x} is a good choice as a point estimator for μ . You already know from the section on Central Limit Theorem that as long as a sample is taken at random, the distribution of sample means is exactly centered at the value of the population mean. \bar{x}

is therefore said to be an unbiased estimator for μ . Any particular sample mean (\bar{x}_i) might turn out to be less than the actual population mean (μ), or it might turn out to be more. But in the long run, such sample means are *on target* in that they will not underestimate any more or less often than they overestimate. Hence \bar{x} can be considered as what we term as an unbiased estimator of the population mean μ under the condition that the samples are drawn at random.



Interval Estimates

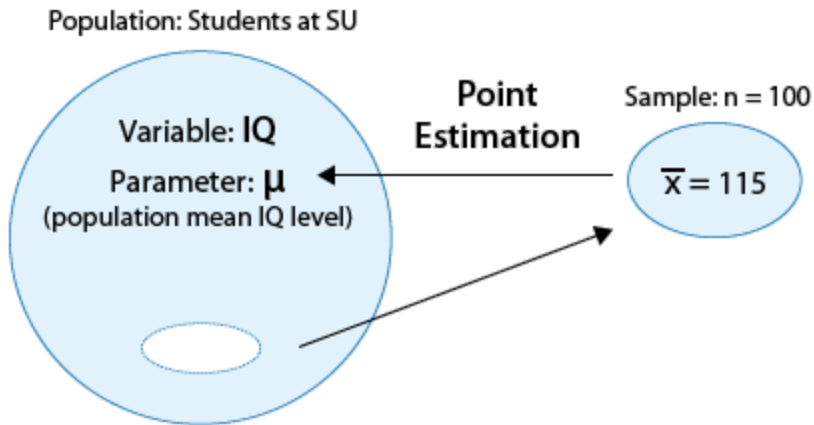
Problem with point estimates alone

In the previous topic, you learned about point estimation where you estimate a population parameter (μ) with the sample statistic (\bar{x}). But this is a bit problematic. Why? When we estimate, say, μ by the sample mean (\bar{x}), we are almost guaranteed to make some kind of error. Even though we know that the values of \bar{x} fall around μ , it is very unlikely that the value of \bar{x} will fall exactly at μ .

Interval estimation to the rescue

In point estimation, we used \bar{x} as the point estimate for μ . However, we had no idea of what the estimation error involved in such an estimation might be. Interval estimation takes point estimation a step further and says something like we are 95% sure that \bar{x} lies between X_1 and X_2 .

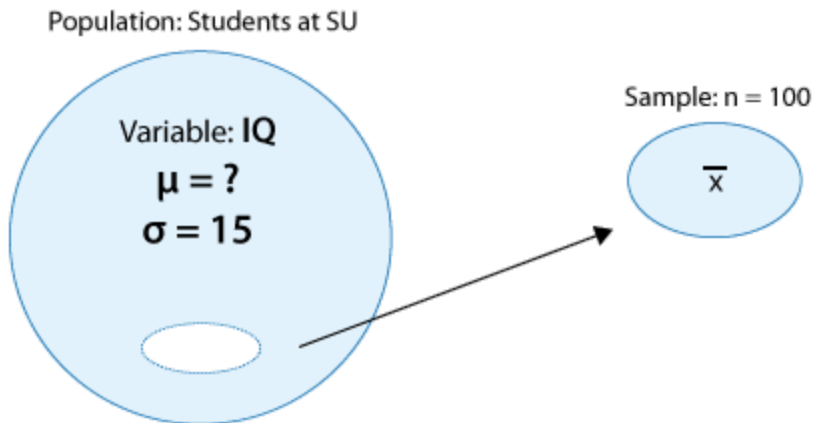
Now let's understand interval estimation in context of our previous example:



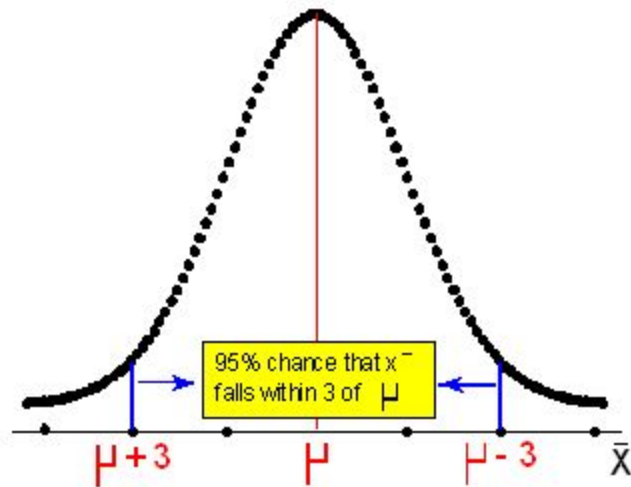
Let's say that you are 95% confident that by using the point estimate \bar{x} to estimate μ , you are off by no more than 3 IQ points. In other words, you are 95% confident that μ is within 3 of 115, or between 112 ($115 - 3$) and 118 ($115 + 3$).

Construction and Interpretation of the confidence interval

Let's also assume that you know the population standard deviation σ , which is 15. The problem now looks somewhat like this:



- You already know that according to the Central Limit Theorem, the sampling distribution of the sample mean is approximately normal with a mean of μ and a standard deviation of σ/n where n is the sample size. In our example, the possible values of \bar{x} , the sample mean IQ level of 100 randomly chosen students, is approximately normal, with mean μ and standard deviation $15/\sqrt{100} = 1.5$
- Next, apply the Standard Deviation Rule for the normal distribution. We are considering a 95% confidence level which means that we are 95% sure that the sample mean (\bar{x}) falls within this interval. The z -value corresponding to 95 is 2 (from the normal distribution curve). So, there is a 95% chance that the sample mean we get in our sample falls within $2 * 1.5 = 3$ units of μ .



- Finally, you can say that there is a 95% chance that the sample mean \bar{x} falls within 3 units of μ or You are 95% confident that the population mean μ falls within 3 units of \bar{x} .

Other Confidence Levels calculation

Now lets calculate the confidence levels with other confidence levels. Their calculations are shown below:

- 90% confidence interval for μ is

$$\bar{x} \pm 1.645 \frac{\sigma}{\sqrt{n}} = 115 \pm 1.645 \frac{15}{\sqrt{100}} = 115 \pm 2.5 = (112.5, 117.5)$$



- 95% confidence interval for μ is μ is

$$\bar{x} \pm 2 \frac{\sigma}{\sqrt{n}} = 115 \pm 2 \frac{15}{\sqrt{100}} = 115 \pm 3 = (112, 118)$$

- 99% confidence interval for μ is μ is

$$\bar{x} \pm 2.576 \frac{\sigma}{\sqrt{n}} = 115 \pm 2.576 \frac{15}{\sqrt{100}} = 115 \pm 4 = (111, 119)$$

General structure of confidence interval calculation

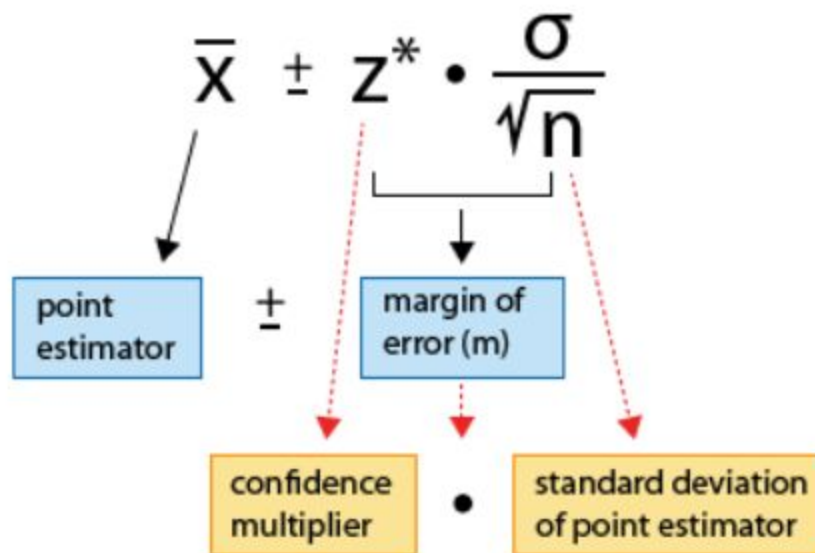
In general, the confidence interval has the form: $\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}$ where:

- z^* is a general notation for the multiplier that depends on the level of confidence.
As we discussed before:

For a 90% level of confidence, $z^* = 1.645$

For a 95% level of confidence, $z^* = 2$ (or 1.96 if you want to be really precise)

For a 99% level of confidence, $z^* = 2.576$



When is it safe to use the confidence interval you developed?

Now you have learned everything that you need in order to calculate the confidence intervals. Time to learn when you can actually use it to calculate confidence intervals. Let's look at them:

- Firstly, the sample must be random and not biased
- Assuming that the sample is random, recall that the Central Limit Theorem works when the sample size is large (a common rule of thumb for *large* is $n > 30$), or, for smaller sample sizes, if it is known that the quantitative variable of interest is distributed *normally* in the population.
- The only situation in which we cannot use the confidence interval is when the sample size is small and the variable of interest is not known to have a normal distribution.

In this task, you will calculate the confidence interval based on a random sample of size 100 with a 95% confidence level.

Instructions

- Dataset has been loaded for you in the variable `data`
- Initialize a variable `sample_size` with value 100. This is the size of the random sample that you will be taking from the population
- Calculate the z-score for the 95% confidence level using `stats.norm.ppf(q=0.95)`. Save it as `z_critical`
- Now time to take a sample. Do it using `data.sample(n=sample_size, random_state=0)` and save it as `data_sample`. You will calculate confidence intervals based on this sample
- Find the mean of `SalePrice` for this sampled dataset (`data_sample`) and save it as `sample_mean`
- Similarly, find the standard deviation of `SalePrice` for this population dataset (`data` and not `data_sample`) and save it as `population_std`
- Next find the margin of error using the formula $z^* \frac{\sigma}{\sqrt{n}}$ where z^* is `z_critical`, σ is the population standard deviation and n is the sample size which is 100 in this case. Save it as `margin_of_error`
- Calculate the confidence interval as `sample_mean ± margin_of_error` and save it as `confidence_interval`
- Calculate the mean of `Saleprice` for dataset `data` and store it in a variable `true_mean`

```
8 data = pd.read_csv(path)
9 # code starts here
10 # sample size
11 sample_size=100
12 # z-critical Score
13 z_critical = stats.norm.ppf(q = 0.95)
14 # sampling the dataframe
15 data_sample = data.sample(n=sample_size, random_state=0)
16 # finding the mean of the sample
17 sample_mean = data_sample['SalePrice'].mean()
18 # finding the standard deviation of the population
19 population_std = data['SalePrice'].std()
20 # finding the margin of error
21 margin_of_error = z_critical * (population_std/math.sqrt(sample_size))
22 # finding the confidence interval
23 confidence_interval = (sample_mean - margin_of_error, sample_mean + margin_of_error)
24 print("Sample Mean:", sample_mean)
25 print("Confidence interval:", confidence_interval)
26 # finding the true mean
27 true_mean=data['SalePrice'].mean()

(165168.45042165995, 184522.44957834008)
180796.0600682594
```

ISSUES

The value of lower confidence level is incorrect

The value of upper confidence interval is incorrect

Examples:

A random sample of 30 households was selected as part of a study on electricity usage, and the number of kilowatt-hours (kWh) was recorded for each household in the sample for the March quarter of 2006. The average usage was found to be 375kWh. In a very large study in the March quarter of the previous

- year it was found that the standard deviation of the usage was 81kWh. Assuming the standard deviation is unchanged and that the usage is normally distributed, provide an expression for calculating a 99% confidence interval for the mean usage in the March quarter of 2006. Exactly one option must be correct). Z score for 0.99 is 2.575.



$$375 \pm 2.575 \times \frac{81}{\sqrt{30}}$$

$$375 \pm 2.575 \times \frac{81}{\sqrt{30}}$$

Explanation:

The formula to calculate confidence interval is

$$\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}$$

we know $z = 2.575$ as well as the other values. Substitute the values.

- A sample of size $n = 100$ produced the sample mean of $\bar{x} = 16$. Assuming the population standard deviation $\sigma = 3$, compute a 95% confidence interval for the population mean μ . Click [here](#) for Z-table.



$$15.412 - 16.588$$

$$15.412 - 16.588$$

Explanation:

A 95% confidence interval for μ is

$$\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}$$

where $z = 1.96$ from the table of Normal distribution. Then, the 95% confidence interval for μ is $16 \pm ((1.96)3/\sqrt{100}) = 16 \pm 0.588 = [15.412, 16.588]$

Confidence Intervals

A confidence interval is an interval estimate for a distribution's mean (μ). It is parameterized by a confidence level, which determines how frequently the confidence interval will contain the true distribution mean.

What is a hypothesis?

In the previous chapter, we learned about statistical methods of estimation where we estimated the population parameters using the data from the sample. We were trying to answer the question like, 'What is the average salary of a Data Scientist in the Bay Area and Montreal?'. In this section, we will be answering slightly different questions like 'Is the average salary of Data Scientist in the Bay Area and Montreal different?'.

If you closely look at the two questions stated above there is a clear distinction, in the first question we are trying to find a value for the population mean, whereas in the second we have a certain assumption about the population average and we just want to answer whether this assumption is correct or not.

Let us now take a closer look into understanding what exactly a hypothesis is and a brief about the process of conducting a hypothesis test.

Null and Alternative Hypothesis

Hypothesis is a statement about the population that might be true. The beauty of these Hypotheses is that they can be TESTED!

Example: We have a hypothesis that the 'average salary of the Data Scientists in the Bay Area and Montreal is different'.

Statistical hypothesis tests are based on a statement called the null hypothesis that assumes nothing interesting is going on between whatever variables you are testing.



Therefore, in our case, the Null Hypothesis is that 'The average salary of Data Scientists in the Bay Area and Montreal is same'

- The purpose of a hypothesis test is to determine whether the null hypothesis is likely to be true given sample data.
- If there is little evidence against the null hypothesis given the data, you accept the null hypothesis.
- If the null hypothesis is unlikely given the data, you might reject the null in favor of the alternative hypothesis: that something interesting is going on.

Alternate Hypothesis This is nothing but the question you ask which kind of "opposes" the Null Hypothesis. Therefore, in our case the Alternative Hypothesis is that: "The average salary of Data Scientists in Bay Area and Montreal is different"

Only 1 Hypothesis out of Null and Alternate can be right. In hypothesis testing, we test a sample, with the goal of accepting or rejecting a null hypothesis which is our assumption or the default position. The test tells us whether or not our primary hypothesis is true.

Important: The null hypothesis is assumed true and statistical evidence is required to reject it in favor of an alternative hypothesis.

<p>A tyre manufacturer believes that an experiment with</p> <p>1. the tyres has increased the average life of tyres by 25%. What is the alternate hypothesis in this case?</p> <p></p>	<p>$p \neq 0.25$</p> <p>where p denotes the percentage increase</p> <p>i.e The increase in the average life is not equal to 25%</p>	<p>$p \neq 0.25$</p> <p>where p denotes the percentage increase</p> <p>i.e The increase in the average life is not equal to 25%</p>
<p><u>Explanation:</u></p> <p>The alternate hypothesis is the opposite of the null hypothesis.</p> <p>In this case, since the percentage of increase is given, the null hypothesis will be:</p> $H_o : p = 0.25$ <p>indicating there was a 25% increase in the average life of tyres.</p> <p>So the alternate hypothesis will be :</p> $H_a : p \neq 0.25$ <p>i.e There was not a 25% increase in the average life of tyres.</p>		
<p>A researcher believes that the average cost of college textbooks is 180 . She samples 30 textbooks and calculates the mean of the sample, X, to be 205. What is the null hypothesis in this situation?</p> <p>2. </p>	<p>The average cost of college textbooks is not 180?</p>	<p>The average cost of college textbooks is 180?</p>
<p><u>Explanation:</u></p> <p>The null hypothesis suggests nothing special is going on and refers to the existing belief that there is no difference from the traditional state of affairs. So the null hypothesis should be:</p>		

Process of Hypothesis Testing

First, let's watch a video that would introduce hypothesis testing in general - [Introduction to Hypothesis Testing](#).

Let us understand the process involved in Hypothesis testing with a simple example in this video - [An example walkthrough for Hypothesis Testing](#). Now let's try to work through another example to understand the process better.

A company uses a semi-automatic process to fill coffee powder in 200 gm jars and this fill is known to have a standard deviation of 4 gm. For long, the amount of coffee powder filled is observed to be normally distributed with a mean of 200 gm. The manager is concerned with ensuring that the process is working satisfactorily so that the average amount of coffee powder filled in jars is 200 gm. She has currently taken a sample of 25 jars, weighs the amount of coffee in each of them and finds the average amount equal to 202 gm. Now, her problem is as to how this difference of 2 gm is interpreted. Is it a

small difference and may be ignored or is it large enough to conclude that the process is not working properly and some action is warranted?

We will break down this example into four broad steps and delve with them in detail one by one.

1. State the Hypothesis

In this example, we are concerned about the average amount of coffee powder that is being filled by the machine in the jars. As discussed in the previous topic, the null hypothesis is that which assumes nothing interesting is going on i.e. in our case we can state our null hypothesis that 'the machine is filling 200gms of coffee on an average'. To put more details, what we are assuming is that for the entire population of the bottles filled by the machine the average amount is 200 gms. We represent it as follows

H_0 : The machine is filling 200gms of coffee on an average or $\mu=200$

The alternate or opposing hypothesis to this is that the machine is not filling 200gms of coffee on an average. This can be represented as

H_1 : The machine is not filling 200gms of coffee on an average or $\mu \neq 200$ (Not equal to)

Note that we are concerned about the average amount of coffee filled by the machine and not with each and every individual bottle.

This completes our first step where based on the problem statement we have stated our hypothesis.

2. Formulate the plan for analysis

Based on the hypothesis that we have stated in the previous step now we need to determine a plan to make our decision.

The first thing that we need to determine, is the test statistic that we will be calculating for our data. We can determine this test statistic based on what population parameter we are testing. In our example, we want to check the average amount of coffee filled by the machine. For such cases, we use a z-statistic calculated as follows

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

where

\bar{x} —sample mean

μ —population mean

σ —population standard deviation

n —sample size

We will take a look at different types of test statistics in later topics.

3. Analyze sample data

Now that we have decided upon which test statistic to use, we need to calculate the value for the same based on the information that we have regarding the sample and the population. From the example given above, we can clearly write the information as follows

population mean= μ =200gms

population standard deviation= σ =4gms

sample size= n =25

sample mean= \bar{x} =202gms

Here we have all the information that we need to calculate our test statistic.

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{202 - 200}{\frac{4}{\sqrt{25}}} = 2.5$$

4. Interpreting the result

This is one of the most critical steps in hypothesis testing. Based on the test statistic that we have calculated in the previous step we need to make a decision regarding our hypothesis. As stated in the earlier topic we are looking for evidence to reject the null hypothesis. In this step, we will be calculating p -value based on which we will be interpreting the result.

Now, p -value is quite misunderstood by a lot of people. Before proceeding further, let's go through a video explaining it through a simple example - [p-values made easy](#). With the intuitive understanding in the background, let's do a more technical deep dive - [p-values clearly explained](#).

The p -value, or calculated probability, is the probability of finding the observed, or more extreme, results when the null hypothesis (H_0) of a study question is true – the definition of 'extreme' depends on how the hypothesis is being tested.

In our example, p -value is the probability of finding the average coffee filled to be 202grams when the null hypothesis is correct.

So, if p -value is very small then we can safely say that there is a strong evidence against the null hypothesis whereas on the flipside if p -value is large we can say that there is weak evidence against the null hypothesis.

Now the question is how do we calculate this p -value. From the central limit theorem, we are aware of the Central Limit Theorem that the sampling distribution of sample means is normally distributed. As we are dealing with samples, in this case, we can use a normal distribution to calculate the p -value.

In our example, we can calculate the p -value as follows

$$p\text{-value} = P(z \geq 2.5) = 0.0062 \text{ (remember the calculation from probability)}$$

Now we know that the probability of finding an average of 202 grams is 0.0062 when the null hypothesis is true. As you can observe this is a very small probability, but we need some threshold to define what probability is small or big. This threshold is called the significance level.

The choice of significance level at which you reject H_0 is arbitrary. Conventionally the 5% (less than 1 in 20 chance of being wrong), 1% and 0.1% ($p < 0.05$, 0.01 and 0.001) levels have been used.

Now for our example, we select a level of significance of 5% i.e. 0.05. As our calculated p -value is less than the threshold we can say that there is strong evidence against the Null Hypothesis and hence we reject the null hypothesis in favor of the alternate hypothesis

Conclusion: From our test of the hypothesis we can say that the average amount of coffee filled by the machine is significantly different than 200 grams and hence some action is warranted. Now as you are well versed with the process of hypothesis testing we will apply the test to some of our hypotheses to the housing dataset in the next topic. We will use the python library of `statsmodels` to perform this testing.

The life span of light bulbs manufactured by a particular company follows a normal distribution with a standard deviation of 120 hours and its half-life is guaranteed under warranty for a minimum of 800

1. hours. At random, a sample of 50 bulbs from a lot is selected and it is revealed that the half-life is 750 hours. With a significance level of 0.01, should the lot be rejected by not honoring the warranty? Click [here](#) for Z-table.



Yes, there is enough evidence to support null hypothesis .

No, there is not enough evidence to support the null hypothesis.

Explanation:

1. State the null and alternative hypothesis:

$$H_0 : \mu \geq 800$$

$$H_1 : \mu < 800$$

2. Calculate the limit of acceptance:

$$\alpha = 0.01; z_{\alpha} = 2.33$$

Calculate the confidence interval:

Confidence Interval

3. Verify:

$$x = 750$$

4. Decide:

The null hypothesis, H_0 , cannot be accepted with a significance level of 1%.

A principal at a certain school claims that the student in his school are above average intelligence. A random sample of 36 students IQ score has a mean score of 112. Is there sufficient evidence to support the principal's claim? The mean population IQ is 100 with a standard deviation of 15. IQ scores are normally distributed. p-value is 0.05. Click [here](#) for Z-table.



Yes, there is enough evidence to support principal's claim.

Yes, there is enough evidence to support principal's claim.

Explanation:

Null hypothesis: Mean IQ score is 100

Alternate hypothesis : Mean IQ score is greater than 100

Here we will want to check the z-score at 95%, so if we look into the table, we find out that the z-score for 95% is 1.645.

To support our claim of mean IQ score being more than 100, we want the Z-score of the mean to be more than 1.645 so that we can reject the null hypothesis.

So, we will calculate the Z-score for mean IQ score of 112. We already have the formula, so we will substitute the values:

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{112 - 100}{\frac{15}{\sqrt{36}}} = 4.8$$

So, 4.8 is greater than 1.645, so we will reject the null hypothesis and accept alternate hypothesis.

Hypothesis testing using statsmodels library

We can perform the test of the hypothesis of means z-test with the help of statsmodels library in Python.

We are using the housing dataset from the previous chapter. Let us formulate and test a hypothesis regarding the `SalePrice` of the houses from this dataset.

You as a buyer of a house think that the average price of a house is \$175,000. Let us check whether this assumption is correct or not based on the sample data we have. Our null and alternate hypothesis over here can be stated as follows

H_0 : The average SalePrice of house is 175,000 i.e. $\mu=175,000$

H_1 : The average SalePrice of house is not 175,000 i.e. $\mu \neq 175,000$

Now we can now use the `statsmodel` library to perform the next steps of hypothesis testing i.e. calculating the z-statistic and p-value.

```
z_statistic, p_value = ztest(df.SalePrice,value=175000)
print("Z-statistics = ",z_statistic)
print("p-value = ",p_value)
```

Output

```
Z-statistics =  3.9272834547730024

p-value =  8.591071063153728e-05
```

We can see that the p-value is less than the significance level of 5% i.e. 0.05 hence we have enough evidence against the null hypothesis. In this case, we reject the null hypothesis in favor of the alternate hypothesis. So from the sample, we can statistically conclude that the average SalePrice of the houses is not \$175,000.

Two-sided vs. One-sided tests

The hypothesis for the tests that we have discussed earlier are of the following pattern

$H_0: \mu = \text{some value}$

$H_1: \mu \neq \text{some value}$

In these tests, we are testing whether the population mean is equal to some value or not. These are called a two-sided hypothesis test i.e. the null hypothesis is rejected if the value is greater or smaller.

Let us now understand the one-sided hypothesis test. For this, let us rephrase the previous hypothesis on SalePrice. You as a buyer think that the average SalePrice of the house is greater than \$175,000. In this case, our alternate hypothesis will change as follows

H_0 : The average SalePrice of house is 175,000 i.e. $\mu = 175,000$

H_1 : The average SalePrice of house is greater than 175,000 i.e. $\mu > 175,000$

In this scenario the calculation of p-value changes. Let us use statsmodel to conduct the test. We need to set the parameter of `alternative` to `larger` to get the result.

```
z_statistic, p_value = ztest(df.SalePrice, value=175000, alternative='larger')
print("Z-statistics = ", z_statistic)
print("p-value = ", p_value)
```

Output

```
Z-statistics = 3.9272834547730024
p-value = 4.295535531576864e-05
```

Over here you can observe that the value of `z-statistic` is similar to the one that we calculated in the previous example, but the p-value has changed to `4.295535531576864e-05`. But this p-value is still less than the significance level of 5%. Hence we can still reject the null hypothesis and accept the alternate hypothesis - the average SalePrice of the house is greater than \$175,000

A similar one-sided hypothesis test can also be conducted for the hypothesis .the average SalePrice of the house is less than \$175,000

Is the average Lot.Area less than 1200?

Let us try to test the hypothesis that the Lot.Area is less than 1200. Assume the level of significance to be 5%

Instructions

- We will use the dataframe `data` for this task
- Call the function `ztest` for the feature of `Lot.Area` with parameters `value=1200` and `alternative='smaller'`
- Print the values of z-statistic and p-value
- Use if statement to determine whether the null hypothesis is `Accept` or `Reject` based on the p-value and store the result in variable `inference`
- Print the `inference`

Skills Covered:

Probability and Statistics

Reference Solution

```
1 from statsmodels.stats.weightstats import ztest
```

```
2
```

```
1 from statsmodels.stats.weightstats import ztest
2
3 # apply ztest
4 z_statistic, p_value = ztest(data['Lot.Area'],value=1200,alternative='smaller')
5
6 # print z statistic and p value
7 print("Z-statistics = ",z_statistic)
8 print("p-value = ",p_value)
9
10 # check the p-value
11 if p_value > 0.05:
12     inference = "Accept"
13 else:
14     inference = "Reject"
15 print(inference)
16
```

OUTPUT

RESULT

```
Z-statistics = 61.46512878748129
p-value = 1.0
Accept
```

Errors in hypothesis testing

To understand what types of errors can be committed in hypothesis testing, first, let's look at a video explaining the various types of errors - [Type I and Type II Errors](#)

If we again think of the hypothesis test as a criminal trial then it makes sense to frame the verdict in terms of null and alternate hypothesis:

- Null Hypothesis: Defendant is innocent
- Alternate Hypothesis: Defendant is guilty

What type of error is being committed in the following circumstances?

- Declaring the defendant guilty when they are actually innocent? Essentially it implies that the evidence leads the jury to convict an innocent person. By analogy, we reject a true null hypothesis and accept a false alternative hypothesis.
- Declaring the defendant innocent when they are actually guilty? It implies that the evidence leads the jury to declare a defendant not guilty when he is in fact guilty. By analogy, we fail to reject a null hypothesis that is false. In other words, we do not accept an alternative hypothesis when it is really true.

Type I error: The first one is a `Type I` error also known as a `false positive (FP)` or `false hit`. `Type I` error describes a situation where you reject the null hypothesis when it is actually true. The type I error rate is equal to the significance level, so setting a lower significance level reduces the chances of getting a false positive. It is typically denoted by α .

Type II error: The second one is a `Type II` error also known as a `false negative (FN)` or `miss`. `Type II` error describes a situation where you fail to reject the null hypothesis when it is actually false. The higher your confidence level (1-significance level), the more likely you are to make a type II error. It is denoted by β .

		Fact (The Truth)	
Our Prediction (Model)		Ho is True	Ho is False
	Ho is True	Correct Decision (True Positive)	Type II Error (False Negative)
	Ho is False	Type I Error (False Positive)	Correct Decision (True Negative)

Since statistical decisions are based on evidence gathered through sampling, and due to randomness involved, sampling evidence will sometimes fool everyone. As long as we are making a decision, we will never be able to eliminate the potential for these two types of errors.

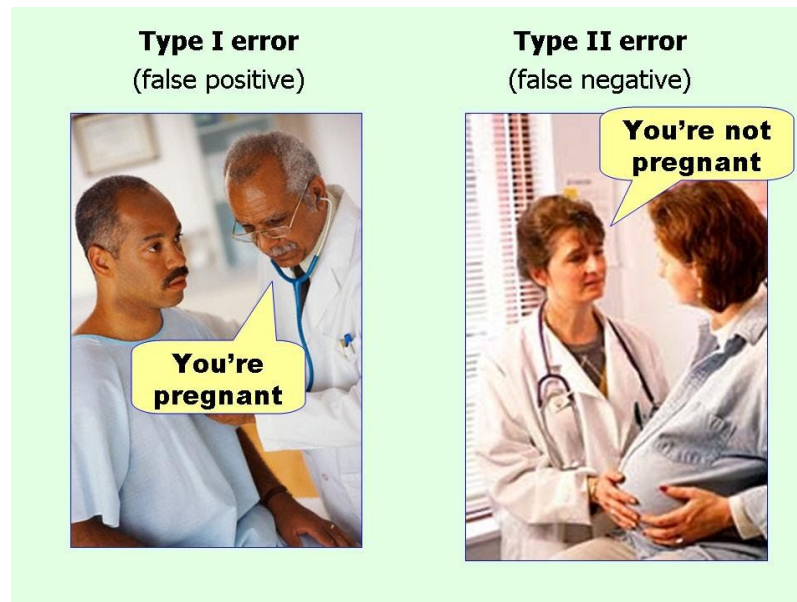
What is the probability that we will make a Type I error?

If the significance level is 5 percent ($\alpha=0.05$), then 5 percent of the time, we will reject the null hypothesis, even if it is true. Obviously we will not know whether the null hypothesis is true. But if it is, the natural variability that we expect in random samples will produce *rare* results 5 percent of the time. Similarly, if the significance level is 1 percent, then we can expect the sample results to lead us to reject the null hypothesis 1 percent of the time. Therefore, the probability of a Type I error is α or the significance level.

What is the probability that we will make a Type II error?

The probability of a Type II error is much more complicated to calculate but *it is inversely related to the probability of making a Type I error*. Thus, reducing the chance of making a Type II error causes an increase in the likelihood of a Type I error.

This image will make your understanding of Type I and Type II errors more clear.



Which type (Type I or Type II) to minimize?

You must have already realized that Type I and Type II errors are inversely related i.e. increasing one decreases the other and vice-versa. And since they are errors they must be reduced (if not removed) to prevent making erroneous decisions. In general, the choice of error type to reduce depends on the business problem at hand. Let's look at some examples where we will illustrate which error type to reduce depending on the business context.

- Cancer detection: In cancer detections, two types of errors can arise:
 - The patient is cancer-free but detected as a cancer patient (Type I)
 - The patient has cancer but detected as cancer-free (Type II)
- So, which error can be more disastrous? Obviously the second situation i.e. Type II error. In this case, our choice of the error to reduce must be the Type II error by reducing the number of patients having cancer but detected as cancer-free (False Negatives)

- Guilty conviction: Here also you can come across two situations:
 - A person is judged as guilty when the person actually did not commit the crime i.e. convicting an innocent person (Type I)
 - A person is judged not guilty when they actually did commit the crime i.e. letting a guilty person go free (Type II)
- In modern society, the social costs of sending an innocent person to prison and denying them their personal freedoms is considered an almost unbearable act and hence in this case Type I error should be the focus.

t - test

Let us relook at the formula used to calculate the z-statistic

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

Over here we assume that the population standard deviation (σ) is known to us, but in reality that is hardly the case. We generally don't have information about the population and that is when we use a t-statistic given by



$$z = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

where, s - sample standard deviation

In this case, we use a t-distribution to calculate the p-value instead of normal distribution in z-test.

The key differences between a t-test and z-test are as follows

The difference between t-test and z-test can be drawn clearly on the following grounds:

- The t-test is based on Student's t-distribution. On the contrary, z-test relies on the assumption that the distribution of sample means is normal. Both student's t-distribution and normal distribution appear alike, as both are symmetrical and bell-shaped. However, they differ in the sense that in a t-distribution, there is less space in the center and more in the tails.
- One of the important conditions for adopting a t-test is that population variance is unknown. Conversely, population variance should be known or assumed to be known in the case of a z-test.

- Z-test is used when the sample size is large, i.e. $n > 30$, and the t-test is appropriate when the size of the sample is small, in the sense that $n < 30$.

We can use the `scipy` library to conduct the test on the data in a similar fashion as we saw for the z-test.

Test of equality of two means

The examples that we discussed up till now involved testing the value of one sample. There may be cases where we need to test that the values of two means from the same population are equal. Such a test is called a two-sample t-test. Let us look at an example. From our data, we have a feature of `Sale.Condition` which specified the condition of sale of the house. The categories of this condition are Normal, Partial, Abnormal, Family, Alloca, AdjLand. Now we are interested to know whether the average `Sale.Price` for `Sale.Condition` Normal and Partial are the same or not.

The hypothesis can be formulated as follows

H_0 : The mean `Sale.Price` of Normal and Partial condition homes are same i.e.

$$\mu_N = \mu_P$$

H_1 : The mean `Sale.Price` of Normal and Partial condition homes are not same i.e.

$$\mu_N \neq \mu_P$$

In this scenario, we will use a two-sample t-test to validate our hypothesis.

```
# subset the dataframe
normal = df[df['Sale.Condition'] == 'Normal']['SalePrice']
partial = df[df['Sale.Condition'] == 'Partial']['SalePrice']

# conduct two sample t-test
t_stat, p_value = scipy.stats.mstats.ttest_ind(normal,partial)

# print the results
print('t-statistic = ',t_stat)
print('p-value = ',p_value)
```

Output

```
t-statistic = -19.6793088005836
p-value = 1.3616391052606246e-80
```

Over here we can observe that the p-value is less than the significance level of 5% hence we have evidence against the null hypothesis and conclude that the mean `Sale.Price` of Normal and Partial condition homes are not the same.

Example:

With Code:

Is the average sale price same for Family and Alloca condition homes

In this task, we will apply the two-sample t-test to find out whether the average sale price of the houses with condition Family and Alloca is the same.

Instructions

- We have the housing dataset stored in the dataframe `data`
- Subset the `SalePrice` for `Sale.Condition == 'Family'` and store the same in variable `family`
- Subset the `SalePrice` for `Sale.Condition == 'Alloca'` and store the same in variable `alloca`
- Call the function `ttest_ind` with appropriate parameters, store the values of t-statistic and p-value in variables `t_stat` and `p_value` respectively
- Print the values of z-statistic and p-value
- Use if statement to determine whether the null hypothesis is `Accept` or `Reject` based on the p-value and store the result in variable `inference`
- Print the `inference`

Skills Covered:

Probability and Statistics

```
1 import scipy
2 # subset the dataframe
3 family = data[data['Sale.Condition'] == 'Family']['SalePrice']
4 alloca = data[data['Sale.Condition'] == 'Alloca']['SalePrice']
5 # conduct two sample t-test
6 t_stat, p_value = scipy.stats.mstats.ttest_ind(family, alloca)
7 # print the results
8 print(t_stat)
9 print(p_value)
10 if p_value > 0.05:
11     inference = "Accept"
12 else:
13     inference = "Reject"
14 print(inference)
```

OUTPUT

RESULT

```
-0.2600957325150498
0.7955758684627952
Accept
```

Average heart rate for Americans is 72 beats/minute.

A group of 25 individuals participated in an aerobics fitness program to lower their heart rate. After six months the group was evaluated to identify is the

1. program had significantly slowed their heart. The mean heart rate for the group was 69 beats/minute with a standard deviation of 6.5. Was the aerobics program effective in lowering heart rate? $\alpha = 0.05$. Here you can find the [t-table](#)



There is an insignificant effect of the ind. var. of fitness

There is a significant effect of the ind. var. of fitness

Explanation:

The pop. mean is given as 72 beats per minutes.

The sample of 25 has an average of 69 with a standard dev. of 6.5.

Step One: You need to solve for st. error. St. error = 1.30

Step Two: Solve for t test for single samples $t = -2.31$

Step Three: Evaluate. The critical value is 2.064. The computed value exceeds this value so there is a significant effect of the ind. var. of fitness.

- In the population, the average IQ is 100. A team of scientist wants to test a new medication to see if it has a positive or negative or no effect at all. A random
2. sample of 25 participants who have taken the medication has a mean of 140 with a standard deviation of 20. Did the medication effect intelligence?
 $\alpha = 0.05$. Here you can find the [t-table](#)



medications did not affect the intelligence

medications deeply affected intelligence

Explanation:

Null hypothesis ;

$$H_0 : \mu = 100$$

Alternate hypothesis ;

$$H_1 : \mu \neq 100$$

here $\alpha = 0.05$ is given

Now, we will calculate the degrees of freedom, the formula for which is:

$$df = N - 1$$

Where N is the number of values in the data set (sample size). Take a look at the sample computation.

$$df = 25 - 1 = 24$$

From the t-table we will find the t- score for $\alpha(0.05)$ and $df(24)$ that is 2.064.

So, if t is less than -2.064 or more than 2.064 we reject the null hypothesis.

Now we will calculate the t-statistic, we have the formula as :

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

where s - sample standard deviation

so by substituting the values we get t-score as 10.96 which is clearly greater than 2.064. So, we will reject the null hypothesis.

So, the conclusion is that the medications deeply affected intelligence.

Chi-squared test of independence

The chi-squared test of independence is used when you have two categorical variables from the same population. The test of independence is used to determine whether the two categorical variables are inter-related to each other.

Many times it would occur that the features in a dataset are related to each other, such features do not add much information while building a machine learning model and it is best to remove one of these related features. The chi-squared test of independence is one of the ways to determine this.

The procedure for the test remains the same as discussed in the previous topic. The hypothesis for a chi-square test of independence is as follows.

H_0 : Variable A and Variable B are independent.

H_1 : Variable A and Variable B are not independent.

In this case we will be calculating a chi-square test statistic as follows

$$\chi^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

In the formula, observed is the actual observed count for each category and expected is the expected count based on the distribution of the population for the corresponding category.

Let us understand it with an example from the data. We will be using the `scipy` library to perform the test.

From our dataset, we have divided the sale price of the houses into three categories viz. High, Medium and Low. We have then mapped these to the features of Land.Contour. The observed values for these two variables now as per the following table

Land.Contour/SalePrice	High	Medium	Low
Bnk	61	34	22
HLS	21	23	76
Low	15	18	27
Lvl	884	905	844

Now we want to determine whether the Land.Contour of the property is dependent on the Sale.Price of the property. We will apply a simple chi-squared test of independence as follows

H_0 : Land.Contour and Sale.Price are independent.

H_1 : Land.Contour and Sale.Price are not independent.

```
import scipy.stats as stats

# categorize the SalePrice into three buckets
price = pd.qcut(df['SalePrice'], 3, labels = ['High', 'Medium', 'Low'])

# make a frequency table with Land.Conotur
observed = pd.crosstab(df['Land.Contour'], price)

print(observed)

# conduct the chi-square test with the above frequency table
chi2, p, dof, ex = stats.chi2_contingency(observed)

print("Chi-square statistic = ", chi2)
```

```
print("p-value = ",p)
```

Output

```
SalePrice      High  Medium  Low
Land.Contour
Bnk           61      34     22
HLS           21      23     76
Low           15      18     27
Lvl           884     905    844
Chi-square statistic = 5856.025316455696
p-value = 0.0
```

Here we can observe that the p-value is less than 5% significance level hence there is enough evidence against the null hypothesis and hence we can state that the variable of Land.Contour and Sale.Price are not independent of each other.

Example:

A sample of 100 voters are asked which of four candidates they would vote for in an election. The number supporting each candidate is given below:

Higgins	Reardon	White	Osborne	
41	19	24	16	
				Voters do not prefer the four candidates equally.
				Voters do not prefer the four candidates equally.

Does the data suggest that all candidates are equally popular? [Chi-Square = 14.96, with 3 d.f. and $p < 0.05$].

Explanation:

A Chi-Squared Goodness-of-Fit test is appropriate here. The null hypothesis is that there is no preference for any of the candidates: if this is so, we would expect roughly equal numbers of voters to support each candidate. Our expected frequencies are therefore $100/4 = 25$ per candidate.

O	E	(O-E)	$\frac{(O-E)^2}{E}$
41	25	16	10.24
19	25	-6	1.44
24	25	-1	0.04
16	25	-9	3.24

Adding the last column gives us a value of $10.24 + 1.44 + 0.04 + 3.24 = 14.96$, with $4 - 1 = 3$ degrees of freedom.

The critical value of Chi-Square for a 0.05 significance level and 3 d.f. is 7.82. Our obtained Chi-Square value is bigger than this, and so we conclude that our obtained value is unlikely to have occurred merely by chance. In fact, our obtained value is bigger than the critical Chi-Square value for the 0.01 significance level (13.28). In other words, it is possible that our obtained Chi-Square value is due merely to chance, but highly unlikely: a Chi-Square value as large as ours will occur by chance only about once in a hundred trials. It seems more reasonable to conclude that our results are not due to chance, and that the data do indeed suggest that voters do not prefer the four candidates equally.

Example:

Is gender independent of education level?

A random sample of 395 people were surveyed and each person was asked to report the highest education level they obtained. The data that resulted from the survey is summarized in the following table:

	High School	Bachelors	Masters	Ph.d.	Total
Female	60	54	46	41	201
Male	40	44	53	57	194
Total	100	98	99	98	395



There is a relationship between gender and education

There is a relationship between gender and education

Are gender and education level dependent at 5% level of significance? In other words, given the data collected above, is there a relationship between the gender of an individual and the level of education that they have obtained? (Hint : Use Chi-Square test of independence)

The expected count values are be given as

	High School	Bachelors	Masters	Ph.d.	Total
Female	50.886	49.868	50.377	49.868	201
Male	49.114	48.132	48.623	48.132	194
Total	100	98	99	98	395

Explanation:

The null hypothesis and alternate hypothesis as follows:

Null hypothesis - The two features (I.e gender and education level) are independent

Alternate hypothesis - The two features are dependent

The degrees of freedom can be given as $(r - 1) \times (c - 1)$ where **r** and **c** are the number of rows and columns in the table mentioned in the question.

The observation table given in the question has 2 rows and 4 columns excluding the **Total** row and **Total** column. So the degree of freedom is $(2 - 1) \times (4 - 1) = 3$

The Chi-Square test statistic is defined as

$$\chi^2 = \sum \frac{(Observed\ Count - Expected\ Count)^2}{Expected\ Count}$$

Using the above values, the chi-square test statistic can be calculated to be 8.006

The critical χ^2 value for a 5% significance level with degrees of freedom 3 is 7.81 (From the chi-square statistic table)

Since $8.006 > 7.81$, therefore we reject the null hypothesis and conclude that the education level depends on gender at a 5% level of significance.

- A company is engaged in the manufacture of car tyres. Their mean life is 42,000 km with the standard deviation of 3,000 km. A change in the production process is believed to improve the quality of tyres. A test sample of 30 tyres has a mean life of 43,500 km. Do you think that the new car tyres are significantly superior to the earlier one? Test the hypothesis at 5% level of significance.



The new tyres are not significantly better than the old one

The new tyres are significantly better than the old one

Explanation:

- First we will define null and alternate hypothesis:
 - Null hypothesis : The change has not improved the quality.
 - Alternate hypothesis : The change has improved the quality.
- 0.05 is the alpha level for this example. As this is a one-tailed test, the alpha will be 0.05.
- Find the z-score associated with your alpha level. You're looking for the area in one tail only. A z-score $(0.5-0.05 = 0.45)$ is 1.645.
- Find the test statistic using the z score formula:

$$Z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

if z score is less than or more than 1.645 we will reject the null hypothesis and accept the alternate hypothesis.

CLT

CLT

Let's now find out if Central Limit Theorem holds for `installment` column

- An array of sample sizes to take ('sample_size') is given
- Create subplot with (nrows = 3 , ncols = 1) and store it in variable's `fig ,axes`
- Create `i` loop with range (len(sample_size)).
- Inside the `i` loop, initialise a list 'm' and create `j` loop with range (1000) (for no. of samples='1000')
- Inside the `j` loop create a dataset sample of `data['installment']` with `n = sample_size[i]` using "`sample()`", find the mean of `installment` column of that sample and append it to list 'm'
- Outside the `j` loop (but still inside the `i` loop), convert 'm' into a series called 'mean_series'
- Then, using `axes[i]`, plot the corresponding histogram for `mean_series`

Things to ponder upon:

* Does the central limit theory hold for the 'installment' column?

```
1 import matplotlib.pyplot as plt
2 import numpy as np
3
4 #Different sample sizes to take
5 sample_size=np.array([20,50,100])
6
7 #Code starts here
8 fig, axes = plt.subplots(nrows = 3, ncols = 1, figsize=(10,20))
9 for i in range(len(sample_size)):
10     m = []
11     for j in range(1000):
12         mean = data['installment'].sample(sample_size[i]).mean()
13         m.append(mean)
14     mean_series = pd.Series(m)
15     axes[i].hist(mean_series,normed=True)
16 plt.show()
```

Congrats! You have successfully plotted the the mean sample for different sample sizes.

CONTINUE



Small Business Interests

The bank manager believes that people with `purpose` as 'small_business' have been given `int.rate` more due to the risk associated

Let's do a hypothesis testing(one-sided) on that belief

Null Hypothesis $H_0 : \mu = 12\%$

Meaning: There is no difference in interest rate being given to people with `purpose` as 'small_business'

Alternate Hypothesis $H_1 : \mu > 12\%$

Meaning: Interest rate being given to people with `purpose` as 'small_business' is higher than the average interest rate

- From the column `int.rate` of 'data', remove the % character and convert the column into float.
- After that divide the values of `int.rate` with 100 and store the result back to the column '`int.rate`'
- Apply "`ztest()`" with `x1` as `data[data['purpose']=='small_business']['int.rate']` and value as `data['int.rate'].mean()`, `alternative = 'larger'` (**WHY?**) and save the results in '`z_statistic`' and '`p_value`' respectively
- If '`p-value`' is less than 0.05, you can reject the null hypothesis, If '`p-value`' is greater than 0.05, you can't reject the null hypothesis,

```
1 #Importing header files
2 from statsmodels.stats.weightstats import ztest
3 #Code starts here
4 # Removing the last character from the values in column
5 data['int.rate'] = data['int.rate'].map(lambda x: str(x)[-1])
6 #Dividing the column values by 100
7 data['int.rate']=data['int.rate'].astype(float)/100
8 #Applying ztest for the hypothesis
9 z_statistic, p_value = ztest(x1=data[data['purpose']=='small_business']['int.rate'], value=data['int.rate'].mean(),
10                             alternative='larger')
11 print(('Z-statistic is :{}'.format(z_statistic)))
12 print(('P-value is :{}'.format(p_value)))
13 if p_value > 0.05:
14     print("Accepted the Null Hypothesis")
15 else:
16     print("Rejected the Null Hypothesis")
```

OUTPUT

RESULT

Z-statistic is :12.321276240200591
P-value is :3.4792131906806856e-35
Rejected the Null Hypothesis

Installment vs Loan Defaulting

The bank thinks that monthly installments (`installment`) customers have to pay might have some sort of effect on loan defaulters

Let's do hypothesis testing(two-sided) on that

Null Hypothesis $H_0 : \mu(D(yes)) = \mu(D(no))$

Meaning: There is no difference in installments being paid by loan defaulters and non defaulters

Alternate Hypothesis $H_1 : \mu(D(yes)) \neq \mu(D(no))$

Meaning: There is difference in installments being paid by loan defaulters and loan non defaulters

- Apply "`ztest()`" with `x1` as `data[data['paid.back.loan']=='No']['installment']` and `x2` as `data[data['paid.back.loan']=='Yes']['installment']` and save the results in '`z_statistic`' and '`p_value`' respectively
- If '`p-value`' is less than 0.05, you can reject the null hypothesis, If '`p-value`' is greater than 0.05, you can't reject the null hypothesis,

```
1 #Importing header files
2 from statsmodels.stats.weightstats import ztest
3
4 #Code starts here
5 z_statistic,p_value = ztest( x1 = data[data['paid.back.loan'] == 'No']['installment'],
6                             x2 = data[data['paid.back.loan'] == 'Yes']['installment'])
7 print(z_statistic)
8 print(p_value)
9 if p_value > 0.05:
10     print("Accepted the Null Hypothesis")
11 else:
12     print("Rejected the Null Hypothesis")
13
```

Congrats! You have successfully implemented two sided hypothesis test

CONTINUE

OUTPUT

RESULT

4.894575287952092
9.85182562491764e-07
Rejected the Null Hypothesis

Purpose vs Loan Defaulting

Purpose vs Loan Defaulting

Another thing bank suspects is that there is a strong association between purpose of the loan(`purpose` column) of a person and whether that person has paid back loan (`paid.back.loan` column)

Since both are categorical columns, we will do chi-square test to test the same.

Null Hypothesis : Distribution of purpose across all customers is same.

Alternative Hypothesis : Distribution of purpose for loan defaulters and non defaulters is different.

- Create a variable `'yes'` which is the value counts of `'purpose'` when `paid.back.loan` in `'data'` is `Yes`
- Create a variable `'no'` which is the value counts of `'purpose'` when `paid.back.loan` in `'data'` is `No`
- Concat `'yes.transpose()'` (transpose of `'yes'`) and `'no.transpose()'` (transpose of `'no'`) along `axis = 1` with keys = `['Yes','No']` and store it in a variable called `'observed'`
- Apply `"chi2.contingency()"` on `'observed'` and store the result in variables named `chi2`, `p`, `dof`, `ex` respectively.
- Compare `chi2` with `critical_value` (given)
- If chi-squared statistic exceeds the critical value, reject the null hypothesis that the two distributions are the same, else null hypothesis cannot be rejected.

```
1 #Importing header files
2 from scipy.stats import chi2_contingency
3 #Critical value
4 critical_value = stats.chi2.ppf(q = 0.95, # Find the critical value for 95% confidence
5 | | | | | df = 6) # Df = number of variable categories(in purpose) - 1
6 #Code starts here
7 #Subsetting the dataframe
8 yes=data[data['paid.back.loan']=='Yes']['purpose'].value_counts()
9 no=data[data['paid.back.loan']=='No']['purpose'].value_counts()
10 #Concatting yes and no into a single dataframe
11 observed=pd.concat([yes.transpose(),no.transpose()], 1,keys=['Yes','No'])
12 print(observed)
13 chi2, p, dof, ex = chi2_contingency(observed)
14 print("critical value")
15 print(critical_value)
16 print("Chi Statistic")
17 print(chi2)
18 #Code starts here
19
```

RESULT

```
all_other      Yes    No
credit_card    1944   387
debt_consolidation 3354  603
educational    274    69
home_improvement 522  107
major_purchase 388    49
small_business 447   172
Critical value
12.591587243743977
Chi Statistic
96.98469589063261
```

```
{
  "name": "stderr",
  "text": "/opt/greyatom/kernel-gateway/runtime-environments/lib/python3.6/site-packages/ipykernel_launcher.py:11: FutureWarning: Sorting because non-concatenation axis is not
```