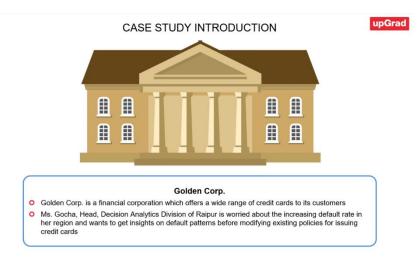


Transcription

Case Study: Analysis of Credit Defaulters



In this video, we will discuss application of interval estimation and hypothesis testing through a practical example. We will look at the case of Golden Corp, which is a financial institution offering wide range of credit cards.

Within Golden Corp, we have Ms. Gocha who heads a decision analytics team in Raipur. She is worried about the default rate in her region and thinks that the existing policies of the company need to be modified. Before doing that, she wants to analyse the historical data and get some insight on default and spending patterns of credit card holders.

	CASE STUDY : OBJECTIVE	upGra
1	What is the average credit score of defaulters?)
2	What should be the cut-off score for second level of scrutiny?)
3	Does the population default rate of those who have more than 10 accounts exceed 20%?)
4	Which group is a better target for Golden Corp: customers with home mortgages or those who rent?)

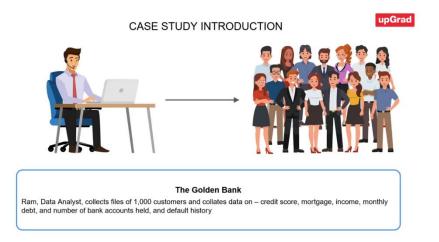
She goes on to discuss the project with her colleague Ram, and they decide on some of the more important objectives. The objectives are:

- One, what is the average credit score of defaulters?
- Two, what should be the cut-off score for a second scrutiny?
- Three, if the population default rate of those who have more than 10 bank accounts exceed 20%.

© Copyright 2019. UpGrad Education Pvt. Ltd. All rights reserved



• And four, which group is a better target for Golden Corp, customers with home mortgages or customers who rent their houses?



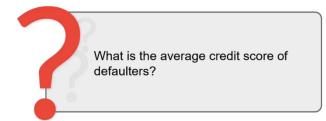
We look at each of these objectives in greater detail. After discussing the objectives, Ram collects random files of 1000 credit card holders from the data team. He collates information on credit score, mortgages, income, monthly spending, number of bank accounts held and customers default history.

	- ຍ 🖬 🖬 ທ - ປ 🔻					DataSet_C	edit Card Defaulter										- ur	oGra
Home	Insert Page Layout Form																	
	X Cut Calibri (Body) + 11 Copy + B I U +				p Text General ge & Center + 🛃 + % 3 🐾	•	Re	Normal Ball	wy T.	tood to	atral Calculation		• 	Delete F		∑ AutoSum •	ZY.	
	‡ × √ fr Default				-		ormatting as Table									Otar 1	Filter	
4							_											
4	A B	с	D	E	F	G	H N	1 1		ĸ	L	м	N		0	Р	Q	R
1 Cu	tomerID HomeOwnership					Default	_		_									
2	6600 Rent	744		8,257	10			Credit Scor		Income	Monthly Debt							
3	1791 Rent	695	84,135	21,118	5		Me		5 4	1,06,614								
4	4120 Home Mortgage		51,810	12,952	12	1	Mi	59	4 र 1 र	1,065								
5	3420 Home Mortgage		26,039	6,510	12		Ma	< 75	are	6,24,657	< 97,150							
0	492 Home Mortgage	719	1,26,168	39,491 12,726	7													
-	6720 Home Mortgage 3059 Rent	688	1,02,629 44,527	12,726	6													
0	4430 Home Mortgage	689	44,527	22,371	4													
0	3318 Home Mortgage	738	1,11,853	22,3/1 23.260	7													
11	6564 Home Mortgage	/38	33,867	8,467	3													
12	8940 Home Mortgage	705	2,28,296	26,482	3													
13	6315 Rent	709	1,23,698	28,574	10	- 1												
4	7729 Home Mortgage	713	1,95,548	15,214	25	1												
IS	3161 Home Mortgage	713	2,01,881	31.695	16													
6	1539 Rent		14.114	7.057	24													
7	5551 Rent	694	66.454	16.946	10													
18	7495 Home Mortgage	700	1.01.501	29.841	3													
19	150 Rent	745	1.12.330	9,952	5													
10	9072 Rent		51,848	25,924	8													
1	9029 Own Home	644	59,499	19,516	7													
12	8622 Rent		5,094	2,547	7	1												
13	4421 Home Mortgage	643		72,444	18													
14	5671 Rent		11,728	5,864	5													
25	1886 Rent		35,486	17,743	9													
16	6412 Rent	717	55,744	3,155	6													
17	6827 Rent	716	1,09,077	20,288	9													
18	494 Rent	693	86,952	1,374	4													
19	7996 Home Mortgage	678	1,39,394	29,970	9	1												
10	5135 Rent	741	1,10,401	22,953	9													
11	9936 Rent	746	2,11,299	31,484	8													
2	6131 Rent	746	1,59,521	13,368	10													
3	2165 Rent		35,352	17,676	3													
4	7810 Home Mortgage	729	1,49,620	15,860	10													
15	7771 Home Mortgage	717	1,97,600	41,496	16													
16	4852 Rent CreditCards +	739	70.609	13.345	5													

Let us take a look at the data. We have customer ID. We have home ownership, whether the customer owns a house, rents it or has a home loan. Then, we have credit score of customers, his or her monthly income, monthly debt or average credit card bill of the customer.

Then, there is number of open accounts, which is the total number of bank accounts held by the customer. Finally, we have default, which is one if the customer has defaulted on three or more consecutive bill payments and zero otherwise.





Coming back to the objectives, here is the first one, what is the average credit score of defaulters? How do we go about this? We have a sample data set, and based on the sample, we can estimate the 95% confidence interval for population mean of defaulters' credit score.

Untitled1* x		Environment Histo	ry Connections Tutorial	-
Source on Save 🔍 🖉 - 🚺	-+Run -+ Source - 2	Car Internet Car		E List - 3
<pre>1 data<- read.csv("CreditCards.csv")</pre>		Clobal Environmen		9
2		Data		
<pre>3 defaulters <- data[data\$Default == 1,] 4 bs <- defaulters\$Credit\$core</pre>		O data	1000 obs. of 7 variables	
<pre>4 ps <- defaultersscreditscore 5 cs<-na.omit(cs)</pre>		O defaulters	142 obs. of 7 variables	
5 Carrieronico(Ca)		Values		
		cs	int [1:116] 714 713 678 729 714 3	716 725 716 797 65
4:1 (Top Level) :	R Script ©		ages Help Viewer	-0
ionsole Terminal x Jobs x			Export • 9	
defailters data[dataSbefault 1,] head(defailters) Customer1D HomeCharening CreditScore Income MonthlyDebt Number 7619 Home Mortgage MA 51849, 66 12952,49 2 3318 Rent 714 121407,92 28574,10 3 6564 Home Mortgage 7,31 395548,60 15213.68	.of.Open.Accounts 12 16 25 7			

Let's look at that in R. So, we start with reading the data, we again use read.csv function here. The data is called credit card dot csv. Once we run it, we can do head data. We can also see the data here in this window. It has 1000 observation, which is on sample size.

Now, we want to estimate the confidence interval for credit score of defaulters. So, first we would need to subset the data. We only need the data of defaulters and defaulters would be all the data points or all the customers for whom default is equal to 1.

So, net subset R data, we have formed defaulters, which is data such that data dollar default equals to 1, and we can again do head defaulters. So, we have all the customers for whom default is 1.

Next, we want to estimate confidence interval for credit score, and you can see there are NAs here. So, before estimating the confidence interval, we would need to get rid of these NAs.



Let us default cs as the credit score of defaulters or the series of these credit scores, and remember how we get rid of the NAs. We will use na.omit function. So, we will write na.omit cs. So, once we run it, we have a sample of defaulters of credit score.

Envir	nvironment History Connections Tutorial	-
	🚰 🕞 🔐 Import Dataset 🖌 🔏	
	Clobal Environment -	
Date		
a[dataSDefault 1,]	data 1000 obs. of 7 variables	
	defaulters 142 abs. of 7 variables	
Value		
cs		65
n		
5		
n-1) se		
	t 1.98080/54110391	
R Script # Files	iles Plots Packages Help Viewer	-
sx	🗈 🕪 🏓 Zoom 📲 Export + O 🍯	
poge 713 195548.00 15213.68 25		
Rent NA 5093.52 2546.76 7		
gage 678 139393.50 29969.65 9		
goge NA 61012.80 15253.20 16		
itScore		

Now, let us estimate the 95% confidence interval for mean credit score amongst defaulters. So, for sample size would be length of cs and the mean can be calculated using mean cs, which is defined as n here. We have s or standard deviation equal to sdcs, and then we compute the standard error, which is s divided by square root of n.

Now, as population standard deviation is not known. We will use T distribution and for 95% confidence interval, alpha would be 0.05, and alpha by 2 would be 0.025. So, the T value at 95% confidence interval can be calculated using QT function. As we want to estimate two-side confidence interval, we will use alpha by 2 and the degree of freedom for this T distribution would be n minus 1, which is the sample size minus one.

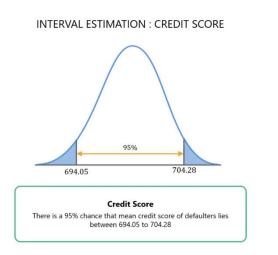
Untitled1* x	-	Emire	ament	History	Connect	tions Tutorial	upG	
				Import Datas			E List -	
Source on Save Q 2 - []					et • 🖉			
<pre>4 cs <- defaulters\$CreditScore</pre>		Glo Clo	obal Env	vironment +			9,	
5 cs<-na.omit(cs)		O dat	a		1000	obs. of 7 variable	is	
6		O defi	aulter	rs	142 o	bs. of 7 variables		
7 n<- length(cs)		Value	5					
8 m<-mean(cs)		ci			nue C	1:2] 694 704		
9 s<-sd(cs)		cs					729 714 716 725 716 707	65
10 se<-s/sqrt(n)						72413793103	123 124 120 123 120 101 1	1.0-11
11 12 t<-at(1- 0.05/2, n-1)					116L	/2413/93103		
12 t<-qt(1- 0.05/2, n-1) 13 ci<-m+c(-se*t, se*t)		n						
14 (14)		5				03052413862		
15 library("BSDA")		se				19354161224		
16 tsum.test(mean.x = m, s.x = s, $n.x = n$)		t			1.980	80754110391		
6:40 (Top Level) :	R Script 0	Files	Plots	Packages	Help	Viewer		-
onsole Terminal × Jobs ×				Zoom -				
/Documents/ 🔅								
5199 Home Mortgage NA 61812.88 15 1 1 1								

Now, we have T, we have standard error, and we have mean, so we can compute the confidence interval by adding and subtracting se into t from the mean. So, here is the confidence interval, which is 694.05 to 704.28.

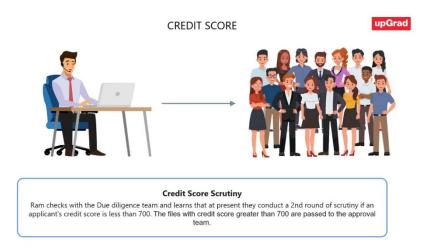
Remember, we can also use t sum dot s function here. For that, we first need to call the library BSDA, and then we can call ts dot test function, mean dot x would be n, s dot x or standard deviation of sample would be s, and n dot x as the sample size.



If we run it, we have the confidence interval 694.05 to 704.28. So, based on this sample, we are 95% confident that the mean credit score of defaulters lie between 694 and 704.3.



We also have the plot here, there's a 95% chance that mean credit score of defaulters lie between 694 and 704.3.



Ram and Gocha also discuss about proposing a second scrutiny or inspection for those who have a very low credit score. Further examination of such applicants can help code and called identify potential defaulters, and it can then refuse such applications.

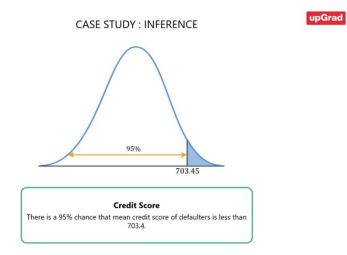
How do we define such a score or what should be the cut-off credit score for a second scrutiny? By cut-off score, we want to estimate the upper bound of credit score amongst defaulters, such that anyone below that score can be scrutinized again. In interval estimating term, we want to find one-sided confidence interval or the upper bound of credit score amongst defaulters.

🕨 • 🤫 🕼 😭 🚍 🔺 Go to file/function 🔢 💀 • Addins •			upGra
Untitled1*	Environment Hi	story Connections Tutorial	-
(a) D □ Source on Save Q / · □ → Run → □ Source · ≥	🔐 🔒 🖙 Impo	rt Dataset 🔹 🥑	🗏 List + 💽
10 se<-s/sqrt(n)	Global Environn	nent +	Q,
11	O data	1000 obs. of 7 variables	
12 t<-qt(1- 0.05/2, n-1)	O defaulters	142 obs. of 7 variables	
13 ci<-m+c(-se*t, se*t)	Values		
14 15 library("BSDA")	ci	num [1:2] 694 704	
16 tsum.test(mean.x = m, s.x = s, n.x = n)	cs	int [1:116] 714 713 678 729 71	4 716 725 716 707 65
17		699.172413793103	
18 t<-qt(1-0.05, df = n-1)	n	116L	
19 m+t*se	5	27.8003052413862	
28	se	2.58119354161224	
<pre>21 tsum.test(mean.x = m, s.x = s, n.x = n, alternative = "less") 22 </pre>	t	1.65821183003114	
22:1 (Top Level) C R Script	Files Plots Pa	ickages Help Viewer	-0
Console Terminal × Jobs ×	a a P. Zoon	2 Export • 0	
~/Documents/ @			
> m+t*se			
[1] 703.4526			
<pre>> tsum.test(mean.x = m, s.x = s, n.x = n, alternative = "less")</pre>			
One-sample t-Test			
data: Summarized x			
t = 270.87, df = 115, p-value = 1			
alternative hypothesis: true mean is less than 0			
95 percent confidence interval: NA 703.4526 1			
sample estimates:			
mean of x			
699.1724			
Worning message:			
<pre>In tsum.test(mean.x = m, s.x = s, n.x = n, alternative = "less") :</pre>			
argument 'var.equal' ignored for one-sample test.			
>			

Earlier, we build a two-sided confidence interval. Now, we just want to get the critical value at the right side of the distribution. So, going back to R, first, we need the right T value at 95% confidence interval. So, we use QT function, and we pass 1 minus 0.05. Now, the 5% region is only at the right side of the distribution, and therefore, we use alpha 0.05 and not alpha by 2.

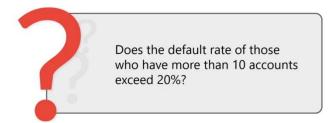
Next, we can calculate the upper bound as m+t x se. So, this is the upward bound, 703.45. With 95% confidence, population mean of credit defaulters is less than 703.4. We can also estimate confidence interval using t sum dot test function.

As we want to estimate the upper bound in the t sum dot test function, we will pass alternative as less. Once we run it, we have one-sided confidence interval, which is the upward bound of 703.4.



We can also see it graphically. Based on the given sample, we found that on average, there is a 95% chance that the credit score of defaulters is less than 703.4. Therefore, anyone with a score less than 703.4 should be inspected further.





Moving on, we also want to check if the population default rate of those who have more than 10 bank accounts exceed 20%. We can perform a hypothesis test for this.

DEFINING H	IYPOTHESIS	upGrad
Null Hypothesis $ \begin{array}{c} \hline $	Alternate Hypothesis $\overline{\mathbf{k}}$ 1. Default rate of customers with more than 10 accounts is less than 20% $H_0 < 0.2$	

So, a null would be that default rate of customers with more than 10 accounts is more than 20% or pi is greater than an equal to 0.2. And our alternative hypothesis would be default rate of customers with more than 10 accounts is less than 20% or pi is less than 0.2. This is a population proportion test. Let's do it in R.

Untitled1* x	-0	Environment	History I	Connections	Tutorial	upGr
Source on Save Q Z - D	Run 🐤 Source + 🖻		Import Dataset		1 DOM NOT	List -
		Global Em				0.
21 tsum.test(mean.x = m, s.x = s, n.x = n, alternati 22	ve = "less")					104
<pre>23 df <- data[data\$Number.of.Open.Accounts > 10,]</pre>				699.1724137	93103	
24 #H_0: pi>=0.2		n		116L		
25 #H_1: pi< 0.2		p		0.241286863	2/0///	
26		pi				
27 pi <- 0.2		s		27.80030524		
<pre>28 p <- nrow(df[dfSDefault = 1,])/nrow(df)</pre>		se		2.581193541		
29 n<- nrow(df) 30 sigma <- sqrt(pi*(1-pi)/n)		signa		0.037139067		
30 sigma <- sqrt(pi*(1-pi)/n) 31 Z_test <- (p-pi)/n		t		1.658211830		
32 Z<-gnorm(0.95)		Z		1.644853626		
33		Z_test		0.000355921	235092909	
33:1 (Top Level) :	R Script ©	Files Plots	Packages	Help View	br	-
onsole Terminal × Jobs ×	-0	0.017	Zoom J J D	port - O		
/Documents/ 🔗						
Default						
1						
0						
1						
1						
0						
0 0						
0 0 p <- nrow(df[df\$Default == 1,])/nrow(df)						
0 0 0 0 0 0 0 0 0 0 0 0 0 0						
0 p < - nrow(df[df\$Default == 1,])/nrow(df) pi < 0.2						
i 0 p <- nrom(df[df3Default 1,])/nrom(df) pi <- 0.2 p 0.2412869						
0 0 p <- nrow(df[dfSDefault -= 1,])/nrow(df) p < 0.2 p] 0.2412869 sigma <- sqrt(pi*(1-pi)/n)						
0 0 0 p <- nrcw(df[dflDefault == 1,])/nrom(df) pi<<0.2 0 0.412869 signe <- sart(pi*(1-pi)/n) Z.test <- (p-pi)/n						
0 0						
$\label{eq:constraint} \begin{array}{c} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 $						
0 0						

© Copyright 2019. UpGrad Education Pvt. Ltd. All rights reserved



For population proportion hypothesis test, we would want to identify the sample proportion of default amongst those who have more than 10 accounts. For that, we would subset the data and we'll define DF as data such that number of open accounts is greater than 10.

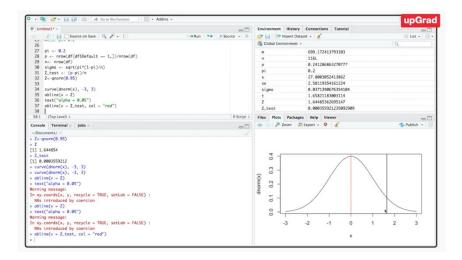
So, in DF, we have all the customers who hold more than 10 bank accounts in our sample. Based on the sample, we want to test the null or default rate of customers who have more than 10 accounts is 20% or 0.2 and the null hypothesis would be that by pi is less than 0.2. Pi is 0.2, so the pi or population proportion in the hypothesis is 0.2.

Next, we need to estimate the sample proportion or the default rate in DF. So, we can calculate sample proportion or P using this command. In the denominator, we have all the customers who are more than 10 accounts, which is the number of rows in DF. And in the numerator, we have those customers in DF who have defaulted, that is where default is equal to 1.

So, P or sample proportion of default rate for customers who hold more than 10 bank accounts is 0.24. Now to test this hypothesis, we would want to calculate sigma. We can define sample size as nrow DF, and this is how we would calculate sigma in proportion test. So, it is square root of pi x 1 minus pi divided by N.

The test statistic that can be used is Z test. So, we'll define the test statistic as P minus pi divided by N, and we can calculate the critical value using qnorm function. So, we will define C as to qnorm 0.95.

This is a one-tailed test, as our alternative hypothesis is pi less than 0.2, and therefore, we use 5% as rejection region. The critical value is 1.64, while the test statistic is 0.0003, and since the Z test statistic is less than critical value, we do not reject the null that pi is greater than equal to 0.2.



We can also plot the graph here, just remember how we will plot the graph using curve function. We will first plot the normal distribution and we will define the critical region or the rejection region.

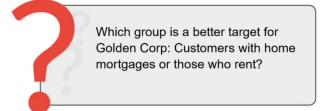
So, this is the rejection region, and our test statistic is almost 0, it lies here. The test statistic does not line the rejection region. Hence, we don't reject the null.



Based on the hypothesis test, we fail to reject the null that for customers with more than 10 accounts, the default rate exceeds 20%.

So, we can conclude that based on the given sample, we fail to reject the null that customers who have more than 10 bank accounts, the default rate exceeds 20%. Golden Corp should thus be more cautious in approving credit cards to such applicants.

upGrad



Finally, the last objective is to identify which group should Golden Corp target in its new marketing strategy. They want to identify which group can generate more revenue for Golden Corp. Is it the customers with home mortgage or is it the customers who are living on rent?

We do not include the customers who own houses here, as they are secured borrowers, and that any way given appropriate offers by Golden Corp, they want to make customized policy for the other group. The other two groups are also fundamentally comparable based on home ownership as they both pay a considerable proportion of income on houses, either in terms of EMI or in terms of rents.



— • —	
Default Rate Credit	Spending
rate of customers with home spending of mortgages and customers who home mortg	erence in credit customers with lages and those o rent?

So, in order to identify the better target or the higher revenue generating group amongst the two, Ram decides to check for two factors, default rate and credit spending. So, he wants to check if there is a difference between the default rate of two groups. For this, we use hypothesis testing for difference in proportion.

• 👒 🐨 • 🔒 🔂 💩 🔺 Co to file/function			upGra
Untitled1* =	Environment History	Connections Tutorial	
🚛 🔐 🔂 Source on Save 🔍 🖉 - 📋 👘 🛶 Run 😁 🔅 Source - ≷	👉 🔒 🛛 🐨 Import Datas	iet - 🔏	List + 💽 -
30 stynu <- sqrt(pr (1-pt)/n) 31 Z_test <- (p-pi)/n	Global Environment -	9	
32 Z<-qnorm(0.95)	cs	int [1:116] 714 713 678 729 714 716 725 71	6 707 65
33		699.172413793103	
34 hm<-data[data\$HomeOwnership "Home Mortgage",]	=1	0.220918043707171	
<pre>35 rent<-data[data\$HomeOwnership "Rent",]</pre>	m2	0.188569558117841	
36 37	1	116L	
<pre>37 38 p1<- nrow(rent[rentSDefault = 1.])/nrow(rent)</pre>	n1	409L	
39 p2 <- nrow(hm[hmSDefault == 1,])/nrow(hm)	n2	5031	
40	9	0.241286863270777	
41 #H0: p1 = p2	p1	0.136919315403423	
42 #H1: p1!=p2]	p2	0.137176938369781	
43 2.12 (Top Level) : R Script :	Files Plots Packages		
			-
onsole Terminal × Jobs ×	🦛 🐞 🏓 Zoom 🛛 🚬	Export - O 🥑	Publish - 1 🔅
/Documents/ 🕫 🚽			
hm=-dstiddstdBendbership==="Nom Wortgage"] Fend-dstiddstdBendbership==="Ref"] Ji==read(refledBefsail==1,j)/read(ref) Ji==read(refledBefsail==1,j)/read(ref)	dnorm(x) 0.0 0.1 0.2 0.3 0.4		

So again, we will first need to subset the data to compute the proportion or default rate of the two groups. So, we'll define HM as the customers who have home mortgage, and we define rent as customers who live on rent. So, we subset on the home ownership column, and we create two subsets HM and rent here.

Next, we compute the default rate of each group. So, the default rate would simply be the number of defaulters in that group divided by the total number of customers and then 2. So, for example, P1 or default rate of those who rent houses would be equal to nrow, number of defaulters in the rent group divided by and nrow of rent group. And similarly, we will compute the default rate for those who have home mortgage.

Untitled1* x		Environment History	v Connections Tutorial	upGra
Source on Save Q Z .	-+ Run + Source - 2	Carl Internet Da		= Ust • 6
<pre>30 pic- mow(renc)rencipiencipiencipiencipienc) 39 p2 <- nrow(hm[hmSDefault = 1,])/nrow(hm)</pre>		Global Environment	•	9
40		5	27.8003052413862	
41 #H0: p1 = p2		51	0.139526142182626	
42 #H1: p11=p2		\$2	0.076213929856029	
43 44 nl<-nrow(rent)		se	2.58119354161224	
44 n1<-nrow(rent) 45 n2<-nrow(hm)		sigma	0.022898092582112	
46 p<-(n1*p1+n2*p2)/(n1+n2)		spending_hm	num [1:503] 0.25 0.25 0.313	0.124 0.2
<pre>47 sigma<-sqrt(p*(1-p)*(1/n1 + 1/n2))</pre>		spending_rent	num [1:409] 0.0991 0.251 0.	201 0.231 0.5
48		t	1.65821183003114	
49 Z_test <- (p1-p2)/sigma		Z	1.95996398454005	
50 Z<-qnorm(0.975) 51		Z_test	-0.0112508483156189	
i1:1 (Top Level) 0	R Script 2	Files Plots Packag	ges Help Viewer	-
onsole Terminal × Jobs ×		🚓 🐘 🏓 Zoom 🖙	Export - O 🥑	S- Publish -
/Documents/ 🤿				
hmc-data[data\$HomeOwnership == "Home Wortgage",] rent-data[data\$HomeOwnership == "Rent",] plc-nrow(rent[rentSb@foult == 1,])/nrow(rent) p2 <- nrow(rent] n1c-nrow(rent) n2c-ornow(rent)		3 0.4		

Now, we want to test the null hypothesis that P1 is equal to P2, and the alternative would be P1 not equal to P2. We can define the sample size here, using again nrow command, and further for hypothesis testing of difference in proportion, we would need to compute P which is the pool proportion.

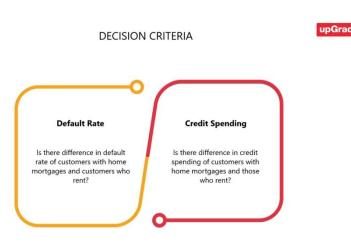
And if you remember, this is the formula for computing pool proportion, n1 into into P1 plus n2n2 P2 divided by n1 plus n2, and the sigma would be square root of P x 1 - P multiplied by 1 upon n1 plus 1 upon n2.

We have seen this formula in lectures. We have the sigma. So, we can calculate the test statistic. It would be P1 minus P2 divided by sigma. So, the test statistic is minus 0.01, and we can compute the critical value using qnorm function. Because it is a 2-tailed test, we'll use alpha by two, which is 2.5 and 1 minus 2.5 is 97.5. So, we will use qnorm of 0.975.

			upGra
= -	Environment Hist	tory Connections Tutorial	-
arce + 🗧	🖙 🔒 📪 Import	Dataset - 🖌	📃 List + 🕥
	Global Environme	ent •	9,
	5	27.8003052413862	
	s1	0.139526142182626	
	\$2	0.076213929856029	
	se	2.58119354161224	
	signa	0.022898092582112	
	spending_hm	num [1:503] 0.25 0.25 0.313	0.124 0.2
	spending_rent	num [1:409] 0.0991 0.251 0.3	201 0.231 0.5
	t	1.65821183003114	
	Z	1,95996398454005	
	Z_test	-0.0112508483156189	
R Script 2	Files Plots Pac	kages Help Viewer	-
			Se Publish •
	dnorm(x) 0.0 0.1 0.2 0.3 0.4		1 2 3
	R Sonpt :	ures *** @*** [**** @**** @***** @****** @******* @************************************	wree + 10 Image: Data - 1 Image: Data - 1 Image: Data - 1

Critical value is 1.9. The test value is less than the critical value. We can also see it in the graph here. We are again using the curve function. This is a 2-tailed test. So, the rejection region would be on both side of the curves, and our test statistics lies here. It does not lie in the rejection region.

So based on the sample, we do not reject the null that default rate of customers with home mortgage is equal to default rate of customers who live on rents.



So, we have checked for the first factor default rate, we could not reject the null that the two groups have same default. Had they been different, and we could identify a group with a lower default rate, Golden Corp would have focused on that group to increase its revenue.

Next, we will check if there is difference in credit cards spending of customers with home loans and customers will live on rental properties. Generally speaking, customers or group who have a higher credit card bill are generally more valued by a bank or by a credit card company. It is a group which is generating more revenue for them.

	\$ @ • B B = ♦ ∞	to file/function	- Addins -					upGra	
0 Ur	titled1* x			- Env	ironment	History	Connections Tutorial	-0	
	🔊 🔒 🗍 Source on Save 🛛	1.10	-+ Run -+ Source	. 2 @	B 1 1791	mport Datase	n = 🧭	📃 List + 🗌 🙄 •	
49 Z_test <- (p1-p2)/sigma					Clobal Envi	Q.			
50 Zqnorm(0.975)			Or	rent 409 obs. of 8 variables					
51				Val	Values				
52 curve(dnore(x), -3, 3)				c	ci		num [1:2] 694 704		
53 abline(v = Z)				c	s		int [1:116] 714 713 678 729 714 716 725 716 707 65.		
<pre>54 abline(v = -Z) 55 abline(v = Z_test, col = "red")</pre>							699.172413793103		
55 dollne(v = 2_test, col = "red") 56					1		0.220918043707171		
57					2		0.188569558117841		
58 rentSCreditSpending <- rentSMonthlyDebt/rentSIncome				n	n		116L		
59				n	1		409L		
60 61	60 #Null: CreditSpending of renters = CS of those who have a house loon 61 #Alternate: Credit Spending of renters != CS of those with a house loan			n	2		503L		
61:54		g of renters I= CS		cript : File	s Plots	Bertreese	Help Viewer		
							xport - 0	😁 🗆	
Conse	le Terminal × Jobs ×			-0		coom A	xport - O 2	Publish • 1	
	cuments/ 🕫								
	<pre>ine(v = Z_test, col = "red"</pre>								
	CreditSpending <- hmSMonthl tSCreditSpending <- rentSMo								
	d(rent)	ith Lypebt/rentsince	me		7.0				
	stomerID HomeOwnership Cred	itScore Income I	fonthlyDebt Number.of.Open.Accounts		0				
1	6600 Rent	744 83321.33	8257.02 10		~				
z	1791 Rent	695 84135.17	21117.74 5	-	0.3				
7	492 Rent	731 44526.50	8949.95 4	×)	2				
12	3318 Rent 1140 Rent	714 123697.92 NA 14113.96	28574.10 16 7056.98 24	dnorm(x)	0 -				
15	1140 Rent 6315 Rent	694 66454.08	7656.98 24 16945.72 10	h	100				
	fault CreditSpending	034 00434.00	10343.72 10		- 0.1				
De	0 0.09909851					/			
De 1					8-	-			
De 1 2	0 0.25099778				-	1	1 1		
1 2 7	0 0.20100277								
1 2 7 12	0 0.20100277 1 0.23099904					-3	-2 -1 0	1 2 3	
De 1 2 7 12 15 16	0 0.20100277					-3	-2 -1 0 x	1 2 3	

So, let's now check the hypothesis if we can see a difference in credit card spending of customers who have home loans versus customers who live on rent. First, we will approximate the credit spending using monthly debt or monthly credit card bill divided by the monthly income.

So, we will define a new column in the data set called credit spending, which is the monthly credit card bill as proportion of their income. So, we define this variable in both the groups. So, we have a new column called credit spending.

We want to conduct a difference in mean hypothesis. We want to test if credit spending of renters is equal to credit spending of those who have a house loan. So, we can define the null as credit spending of renters is equal to credit



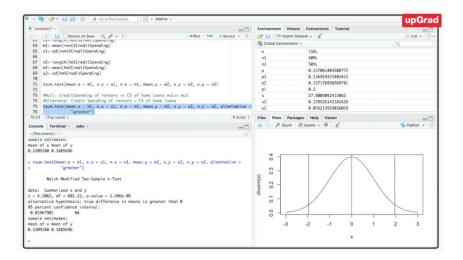
spending of those who have a house loan, and the alternative would be credit spending of renters is not equal to credit spending of those with a house loan.

Untitled1* x	Environment	History C		Tutorial	upGra	
					Tutonal	= Ust - (8)
Car I al I H C Source on Save C Y • I C	Source • C					
60 #Null: CreditSpending of renters = CS of those who have a house loan	Global Envi	ironment -		Q.		
61 #Alternate: Credit Spending of renters I= CS of those with a house l	oan	n		116L		
62		n1		409L		
63 n1<-length(rentSCreditSpending)		nZ		503L		
64 m1<-mean(rentSCreditSpending)		p		0.1370614	03508772	
65 s1<-sd(rent\$CreditSpending)		p1		0.1369193		
66		p2		0.1371769		
<pre>67 n2<-length(hm%CreditSpending)</pre>				0.13/1/05	39303/91	
<pre>68 m2<-mean(hm%CreditSpending)</pre>		pi				
<pre>69 s2<-sd(hmSCreditSpending) 70</pre>		5		27.800305		
70 71 tsum.test(mean.x = m1, s.x = s1, n.x = n1, mean.y = m2, s.y = s2, n.	73	s1		0.1395261		
71 tsum.test(mean.x = m1, s.x = s1, n.x = n1, mean.y = m2, s.y = s2, n. 72	y = hz	s2		0.0762139	29856029	
72.1 (Top Level) :	R Script c	Files Plots	Packages	Help Vie	wer	-0
Console Terminal × Jobs ×	-0					Se Publish •
~/Documents/ @						
s1<-sd(rent\$CreditSpending)						
<pre>> n2<-length(hm\$CreditSpending) > m2<-mean(hm\$CreditSpending)</pre>				<u></u>		
 m2<-mean(nmsCreditSpending) s2<-sd(hm\$CreditSpending) 		7 - 6				
					/	
<pre>stsum.test(mean.x = m1, s.x = s1, n.x = n1, mean.y = m2, s.y = s2, n.y = s</pre>	n2)	0.3			/	
Welch Modified Two-Sample t-Test		dnorm(x) 0.2 0			/	
Weich Modified Iwo-Sample t-Test		0.2		/	<u> </u>	
data: Summarized x and y		5 0 7				
t = 4.2062, df = 601.22, p-value = 2.993e-05		þ		/		
alternative hypothesis: true difference in means is not equal to 0				/		
PS percent confidence interval:		-	1			
0.01724473 0.04745224		8-	_			
sample estimates:		0 1	1 1		r 1	1 1 1
mean of x mean of y			-3 -3	2 -	1 0	1 2 3
					. 0	, 2 3
0.2209180 0.1885696					×	

Next, we can define the variables. So, we have n1 has lend of rent dollar credit spending, m1 as mean of rent dollar credit spending and standard deviation is s1, which is SD rent dollar credit spend.

Similarly, we have n2, m2, and s2 for those who have house loan. So, we can conduct the difference in mean test using t sum dot test function, we define mean dot x s dot x and dot x as n1, s1 and n1. And similarly for the second sample, we define m2, s2, and n2.

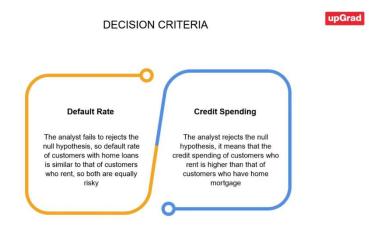
Once you run this, you can see that the 95% confidence interval for difference in means is 0.01 to 0.04. A null is the mean of first sample is equal to mean of second sample that is mu1 minus mu2 is equal to zero. Since zero does not lie in this interval, we reject the null that credit spending of the two groups is same.



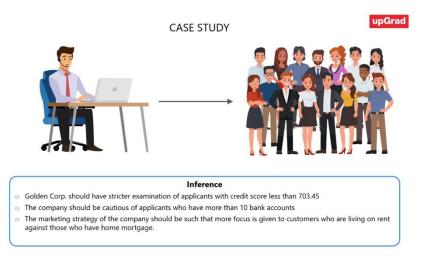
We can further test the hypothesis that spending of renters is less than spending of home loans. So, we will define the null as credit spending of renters less than equal to credit spending of home loans or mu1 is less than equal to mu2, and the alternative will be credit spending of renters is greater than credit spending of home loans.



So, as our alternative is greater, in the t sum dot test function, we'll add another argument alternative equal to greater. So, the one-sided confidence interval or lower bound is 0.01, that is, based on this sample, we are 95% confident that the difference between credit spending of renters and credit spending of home loans is greater than or equal to 0.01. Since 0 does not lie in this interval, we reject the null that credit spending of those who rent is less than or equal to credit spending of those who have home loans.



So, coming back to choosing a better target, Ram concludes that based on the above sample, he cannot reject the hypothesis that default rate between the two groups is same. Further, he is able to reject the hypothesis that means spending of those who rent is less than means spending of those who have a house loan.



Finally, based on the sample study, Ram concludes three points and reports them to Ms. Gocha.

- 1. Golden Corp should have stricter examination of applicants with credit score less than 703.4. We found that on average, there is a 95% chance that the credit score of defaulters is less than 703.4.
- 2. Secondly, Golden Corp should be cautious of applicants who have more than 10 bank accounts. Based on the sample, we fail to reject the null that default rate of customers with more than 10 accounts exceed 20%.



3. Finally, as we saw in the last test, it should target its marketing strategies to attract customers who are living on rent versus those who have a home mortgage.

No part of this publication may be reproduced, transmitted, or stored in a retrieval system, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without the prior permission of the publisher.