Credit Risk Analysis: Testing For The Difference in Means

Elie Diwambuena,

June 20, 2023

Language: Python

Part 2. Modeliing

In Part 1, we did the prelimary univarite and bivariate analysis of variables that paved our way to Part 2. Here, we will take further steps into constructing our model. You may recall that the data visualization in Part 1 revealed some apperent differences between different groups of individuals such as gender, civil status, education and employment type. Thus, the focus in this Part 2 is to seeking evidence for these seemingly existing differences observed across different groups. That is, we test for the difference in means and proportions between groups using statistical tests such as z test, annova and chisquare.

This is how this article is structured. In section 1, we briefly introduce the notion of difference in means and the Z test. In secion 2, we perform the statistical tests to examine the difference and in section 3 we conclude this part 2 with some notes.

$$

1. Testing for the difference in mean

Almost every statistical study involve some forms of statistical test. And there exists several statistical tests that can be performed depending on the question being asked and the type of variables being involved. This is because every statistical test relies on some form of probability distribution to determine whether an event is likely to occur or whether it is simply an hazard. If you are not familiar with the notion of probability distribution and statistical testing, we have an article where we explain these concepts and several other fundamental statistical concepts in very easy-to-grasp terms. Thus, we would strongly advice you to consult Diwambuena (2023).

The starting point of every statistical test is that we are trying to challenge a common belief. Suppose your friend claims that every girl love the pink color. Would you consider this as truth or as a stereotype? This would definitely depends on how much you trust your friend or on how critical you are. If you are very critical, you will certainly not take it for a fact. You would go out and start asking every girl you meet her preferred color and colect a sample that is large enough to verify whether your friend claim is true. The point we are trying to make here is that if there is nothing to prove or disprove, there is no need for a statistical test. Hence, every statistical test involve an initial belief that we are trying to prove or disprove.

Depending on the question that we ask ourself, the statistical test needed will differ. For e.g., if the question is “do girl like pink?”, the statistical test will involve sampling girl color preference, getting an estimated statistic (i.e., the proportion of girls that liked pink in your sample) and then testing this estimated statistic to the general belief (i.e., your friend claim). However, if the question is "do girls prefer pink more than boy?", the statistical test will involve sampling boy's preferred colors and girl's preferred colors and then comparing them. Hence, asking the correct question is very important before starting a statistical analysis.

Once we have clarified the question, then we should be able to choose the appropriate statistical test. In this article, one of the questions we are going to examine is whether income earned by male application is different from income earned by female applicant. We are NOT going to examine whether income earned by male application is greater or less than income earned by female applicant. Even this small difference matters. It matters for our decission to reject or fail to reject the null hypothesis. Again, if this notion sounds less familiar, please consult Diwambuena (2023).

Hence, the statistical test that is used in this analysis is the Z test for the difference in means. What this Z test does is that it compares the means of two groups and tells use whether they are **statistically** different or not. It is very common that means of two different samples are **numerically** different. It is possible that the number of girls who like pink in your school will be different to the number of girls who like pink in your hometown. Or it is also possible that the average salary of people in your hometown is different to the average salary of people in your cousin hometown. However, just because two numbers are numerically different does not mean that they are statistically different. The idea with statistical difference is that we are not measuring the difference based on general numbers but based on standard errors. This concept is also explained in the aforementioned article.

$$

- **Z test for the different in means**

Every statisitical test will involve the following steps :

1. Define the Null and alternative hypothesis
2. Define a significance level
3. Compute the statistical test score
4. Compare the test score to the critical value
5. Reject or fail to reject the null hypothesis

The Z test makes no difference. In section 2, we demonstrate how this is done in practice. However, we think it will be necessary to introduce the logic behind the Z test for the difference in means. The Z test like other statistical tests gives us a score based on which we will be able to determine whether to reject or fail to reject the null hypothesis. We generaly distinguish two approaches for computing the Z test score or Z score for short:

- Z test when the variance of the two groups are not equal

$$Z = \frac{(\overline{x_A} - \overline{x_B}) - (\mu_A - \mu_B)}{\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}}$$

Where $\overline{x_A}$ and $\overline{x_B}$ are sample means for group A and group B, μ_A and μ_B are population mean for group A and group B, and s_A^2 and s_B^2 are sample variance for group A and group B respectively. n_A and n_B are the sample size for group A and group B respectively.

- Z test when the variance of the two groups are equal

$$Z = \frac{(\overline{x_A} - \overline{x_B}) - (\mu_A - \mu_B)}{s_{pooled}\sqrt{\frac{1}{n_A + n_B}}}$$

where

$$s_{pooled} = \sqrt{\frac{(n_A - 1)s_A^2 + (n_B - 1)s_B^2}{n_A + n_B - 2}}$$

s_{pooled} is known as the pooled sample standard deviation. It is simply the weighted average standard deviation.

The mathematical expressions may seem intimidating but they are easy to grasp. Before we explain the logic, we assume at this point that you are familiar with the notions of mean and variance and their respective mathematical expressions which are explained in the aforementioned article as well.

The idea is that we are comparing the difference between the sample mean of group A and group B $(\overline{x_A} - \overline{x_B})$ to the population mean of group A and group B $(\mu_A - \mu_B)$ in terms of standard deviations of group A and group B combined $(\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}})$ *or* $(s_{pooled}\sqrt{\frac{1}{n_A + n_B}})$. While the numerator stays constant, the denominator changes depend on whether the variance of both groups are equal or not.

Let start with the latter case, when the variances are not equal. The idea with the s_{pooled} is that each group variance is weighted according to its degree of freedom. Let the degree of freedom of group A (df_A) be $(n_A - 1)$, the degree of freedom of group B (df_B) be $(n_B - 1)$ and the total degree of freedom for both groups (df_{AB}) be $(n_A + n_B - 2 = (n_A - 1) + (n_B - 1))$. We can rewrite the pooled standard deviation as:

$$s_{pooled}^2 = \frac{df_A s_A^2 + df_B s_B^2}{df_{AB}}$$

$$s_{pooled} = \sqrt{\frac{df_A s_A^2 + df_B s_B^2}{df_{AB}}}$$

Because these two groups have the same variance, we want the difference in sample size to be accounted for. Haviing the same variance means in essence coming from the same population. If A and B are coming exactly from the same pie, then the group with the larger size should be accounted for more than the group with the smaller size so that at the end the pie 1:1 ratio is not distorted. If A is 70% of the pie and B is 30%, then taking them at their respective size will preserve the pie size at 100%.

For the second case, when the group variance of both groups are not equally. This means that they do not come fom the same population. Then, we simply weight them equally according to their respective sample sze.

2. Testing for the difference in means

From this section and forward, we return to our case of credit risk analysis for AB bank. We observed in Part 1 through data visualization some "apparent" differences between mean income or mean loan amount across group of sex, civil status, education level and employment type. Here, we are trying to find evidence to support or reject the fact that those "apparent" differences are statistically significant.

2.1. Is there any difference in income between gender?

To verify whether the mean income of male customers is different from the mean income of female customers, we use the two sample Z test also known as the Z test for the difference in means between two groups. As aforementioned, there are 5 big steps. Since this is our first difference test, we will give as much detailed as possible about every steps. However, from point 2.2, we will not be as detailled but keep in mind that even when we do not explicitly mention the step, we implicitly go through it.

1. Define the Null and Alternative hypothesis

The alternative hypothesis as its name suggests is the alternative belief that we hold against the general belief. Recall the example of your friend claiming that every girl likes the pink color. In this case, the alternative hypothesis is that we (or you) believe not every girl like pink. This is the belief that we (or you) are opposing to a general belief or claim. In the case at hand, for the question of whether the mean salary of men is different from the mean salary of women in our pool of loan applicants, the general belief is that there is no difference between the mean salary of men and the mean salary of women. Because why should there be a diference in salary between men and women at the first place? Put differently, there is no reason to believe that it is different. However, we may believe that a difference exists based on either our experience, culture and many other factors. That means that our alternative hypothesis is that we believe there is a difference and we are trying to disprove the null hypothesis that claims that there is no difference.

Beware that stating the hypothesis in the wrong way can lead to misleading results. In general, there are three types of hypotheses that can be tested with a Z test:

1. Two-sided or two-tailed Z test: when we want to test for the difference between two groups without specifying any direction (greater or lower). Like in our case, we simply want to test if income of men is different from that of women. We do **NOT** test if the income of men is higher than the income of women or vice versa. Thus, the null hypothesis is that mean income of male is equal to the mean income of female and the alternative hypothesis assumes that the mean income of male is **NOT** equal to the mean income of female.
2. One-sided or one-tail Z test: this is the opposite of the two tailed Z test. It specifies a direction such that one group's mean is greater or smaller than the other. In this case, the null hypothesis could be that mean income of male is **NOT** greater than the mean income of female and the alternative hypothesis could be that the mean income of male is greater than the mean income of female.
3. Paired Z test : this is a bit different frol the two discussed above. Here, we do **NOT** test for the difference between two different groups. Rather, we test for the difference within the same group **BUT** at two different points in time. For e.g., if we want to test whether the income of female today is different from the income of female 50 years ago.

2. Define a significance level

A significance level is the minimum level (probability) that is required for the null hypothesis to be “significant”. That is, if the probability value (**p-value**) of the null hypothesis is higher than or equal to the significance level, we cannot reject the null hypothesis. Put differently, we should “accept” the fact that the null hypothesis does not occur by pure hazard. We are using the quotes because we don’t usual accept the null hypothesis but let keep this discussion for later.

A common practice in social sciences is to take a significance level of 5% or 0.005. That is, we want to accept the fact that the null hypothesis is not a hazard if it has a probability (p-value) of at least 5%. This makes sense if we look at it from a different angle. Consider this example. You flip a coin 100 times every day for 100 days and every time you get exactly 5 heads. Can this really be a hazard? No it cannot. There should be something with your coin that makes it return extacly 5 heads every 100 times you flip it. Likewise, finding a p-value of at least 5% means that there is something in the population that makes the null hypothesis “not a hazard”. Here we say 5% because our significance level is 5%. If the significance level is 10%, then we would need a p-value of at least 10% or otherwise we should reject the null hypothesis because it does not meet the minimum. Thus, let summarize this as follows:

- If **p-value < sig. level: reject the null hypothesis**
- If **p-value sig. level: fail to reject (accept) the null hypothesis**

3. Compute the statistical test score

We will need the help of a statistical software such as Python to compute the Z score since doing it manually is tedious and can take hours. However, before using any statistical test, there are conditions that we should meet. In the case of a Z test there are generally two main conditions:

- Data for both groups should follow a normal distrubition
- We should know whether the variance of both groups are equal or different

For the first condition, all we need to do is to plot the data in a histogram or as a cumulative probability distribution. We also devote an interesting discussion about probability distribution in our article Diwambuena (2023). Therefore, we highly suggest you consulting it for a quick refresh on the notions of probability distribution. Moreover, we have already touched on the second point in the previous section where we discussed how the calculation of the Z score varies dependent on whether both sampleq have the same variances or not. Thus, we need to conduct a test that tells us whether they are equal or not and for that we use the **Levene test**. Although this is a completely different test, the idea remains the same. The levene test assumes the two groups variance are equal in general (null hypothesis). Hence, if the p-value is below our significance level of 5%, we should reject the null hypothesis and accept that the two variances are different.

Now let use **Python** to verify if we meet these conditions and if so, to calculate the Z score and compare it directly to the Z critical value. That is, Python will handle all the calculations side and will only return us the p-value with some additional information. This means that steps 4 & 5 will be handled together by Python.

* **Python**

- **Importing the libraries**

[61]:

```
import pandas as pd
import numpy as np
import seaborn as sn
import matplotlib.pyplot as plt
import scipy.stats as sps
from math import sqrt
```

- **Preparing the data set**

We have already demonstrate the necessary data cleasing process in Part 1. Thus, we do not spend too much time on it here. If you would like to read more, please consult Part 1.

[62]:

```
# loading the data set
data = pd.read_excel("loan_data.xlsx")

# get rid of missing observations
data.dropna(inplace=True)
any(data.isna().any())
```

[62]:

```
False
```

[63]:

```
# checking for data types and other details
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1555 entries, 0 to 1594
Data columns (total 14 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   CustrNo      1555 non-null   object
 1   Gender       1555 non-null   object
 2   Age          1555 non-null   int64
 3   Civil        1555 non-null   object
 4   Dependents   1555 non-null   object
 5   Educ         1555 non-null   object
 6   Self-emp     1555 non-null   object
 7   AppIncome    1555 non-null   int64
 8   CoAppIncome  1555 non-null   float64
 9   LoanAmt      1555 non-null   int64
 10  LoanTerm     1555 non-null   float64
 11  DefaultHist  1555 non-null   object
 12  Prop_Value   1555 non-null   int64
 13  OthDebts     1555 non-null   object
dtypes: float64(2), int64(4), object(8)
memory usage: 182.2+ KB
```

[64]:

```
# rearraging data to have male and female as columsns with income as rows
inc_gender = pd.pivot_table(data,␣
 ↪columns='Gender',values='AppIncome',index=data.CustrNo,aggfunc=np.sum)
```

[65]:

```
# spliting the two groups
m_inc = inc_gender.Female.dropna()
f_inc = inc_gender.Male.dropna()
```

- **Calculating the means, standard deviations and the number of observations**

[66]:

```
# Compute the mean of both groups
mean_female_inc = f_inc.mean()
mean_male_inc = m_inc.mean()
```

[67]:

```
# Compute std of both groups
std_female_inc = f_inc.std()
std_male_inc = m_inc.std()
```

[68]:

```
# Compute num of obs
num_obs_female= len(f_inc)
num_obs_male = len(m_inc)
```

[71]:

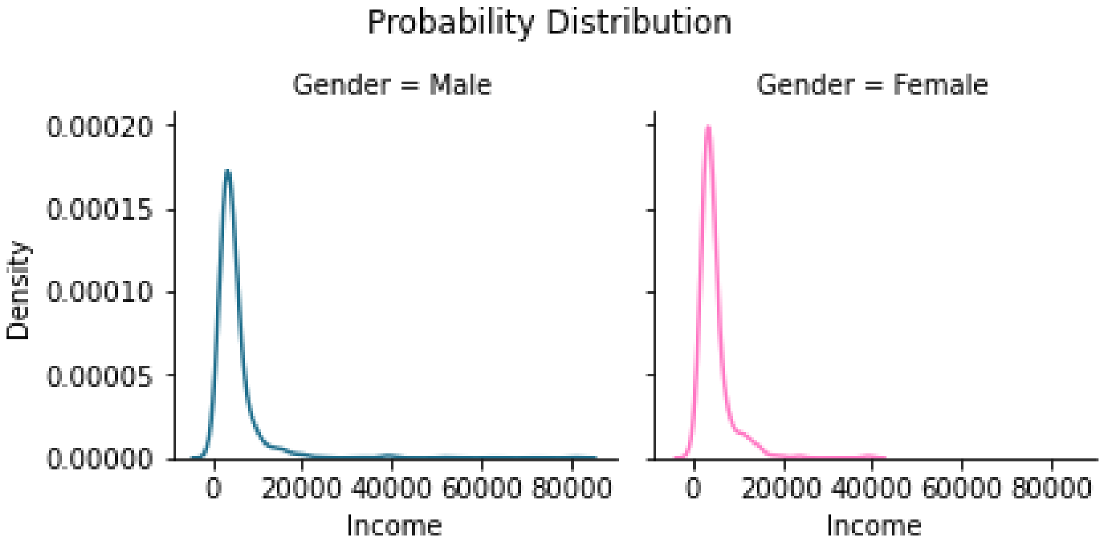
```
print(f"\nThe average incomes are {mean_male_inc:.3f} and {mean_female_inc:.3f}␣
 ↪for male and female with sample size of {num_obs_female} and {num_obs_male}␣
 ↪respectively.")
print(f"\nAnd the standard deviation of incomes are {std_male_inc:.3f} and␣
 ↪{std_female_inc:.3f} for male and female respectively.")
```

```

The average incomes are 4878.996 and 5200.722 for male and female with sample
size of 1277 and 278 respectively.

And the standard deviation of incomes are 3767.982 and 6044.315 for male and
female respectively.
```

- **Verifying the normal distribution condition?**

[72]:

```
# Normality
facet = sn.FacetGrid(data, col="Gender", hue = "Gender", palette =␣
 ↪["#1F6E8C","#FF78C4"])
facet.map_dataframe(sn.kdeplot,x="AppIncome")
facet.set_xlabels("Income")
facet.fig.subplots_adjust(top=0.8)
facet.fig.suptitle("Probability Distribution")
plt.show()
```



It is clear that both groups do not follow a normal distribution in the stricter sense. However, they do show some distributions that are a bit close to the normal. In practice, real-world data will rarely follow a perfect normal distribution. Thus, this is not unusual. The good news is even when the data do not follow a perfect normal distribution, we can still run the test. We are allow to do that thanks in part to **the central limit theoreom**. What this theorem says in a nutshell is that as long as the sample size is greater than 30, we can be sure that the **sampling distribution of sample means** follows a **normal distribution**. That is, we can be be sure that the population is normally distributed. All these concepts are also discussed in the aforementioned article.

- **Testing for the difference of variances: Levene test**

```
[73]: # Are the two groups variances equal?
m_inc.var()/f_inc.var() # rule of thumb if 4:1 then true, else false
p = sps.levene(m_inc,f_inc)[1]
print(f"Based on the Levene test, the p-value is {p:.3f} or {p*100:.2f}%.")
```

```
Based on the Levene test, the p-value is 0.195 or 19.45%.
```

Since we found a p-value of 19.45% which is higher than our significance level, we cannot reject the null hypothesis. This means that the two groups variances are equal.

- **Testing for the difference in means: Z test**

```
[12]: p=sps.ttest_ind_from_stats(mean1=mean_female_inc, std1=std_female_inc,␣
      ↪nobs1=num_obs_female,
                                                     ␣
      ↪mean2=mean_male_inc,std2=std_male_inc,nobs2=num_obs_male)[1]
print(f"Based on Z test, the p-value is {p:.3f}.")
```

```
Based on Z test, the p-value is 0.394.
```

Here too, we find a p-value greater than the confidence level. Therefore, we cannot reject the null hypothesis. This means the mean income for male is equal to the mean income for female customers. This shows that although the mean income of male customers differs from the mean income of female customers numerically, statistically they are the same.

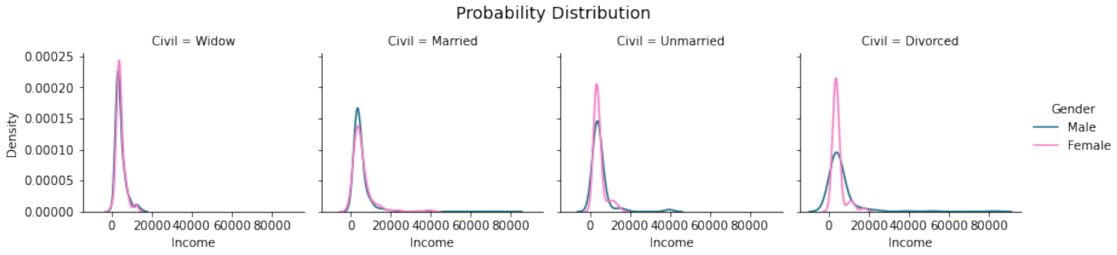
We repeat the same process for other variables while not being as detailled for the sake of time. Additionally, we should mention that the **Annova test** is used when we are comparing more than two groups. However, keep in mind that the steps and the logic are always the same. The Annova test's null hypothesis assumes no difference between all groups (null: A=B=C) whereas the alternative hypothesis assumes that there is a difference at least between two groups (A=B C). The other test that we are going to use is the **chi-square test**. The chi-square test is like a Z test or annova but the difference is that the variable is discrete or categorical like civil status but not numerical like income.

2.2 Is there any difference in income between civil status?

[13]:
```
inc_civ = pd.pivot_table(data, columns='Civil',values='AppIncome',index=data.
 ↪CustrNo,aggfunc=np.sum)
```

[14]:
```
m_inc = inc_civ.Married.dropna()
um_inc = inc_civ.Unmarried.dropna()
d_inc = inc_civ.Divorced.dropna()
w_inc = inc_civ.Widow.dropna()
```

[15]:
```
# Normality
facet = sn.FacetGrid(data, col="Civil", hue = "Gender", palette =␣
 ↪["#1F6E8C","#FF78C4"])
facet.map_dataframe(sn.kdeplot,x="AppIncome")
facet.set_xlabels("Income")
facet.fig.subplots_adjust(top=0.8)
facet.fig.suptitle("Probability Distribution", fontsize= 14)
facet.add_legend()
plt.show()
```



[16]:
```
# Homogeneity of variance: null: variances are the same
p = sps.levene(m_inc,um_inc,d_inc,w_inc)[1]
```

```
print(f"Based on the Levene test, the p-value is {p:.3f} or {p*100:.2f}%. This␣
 ↪suggest that the variances are equal.")
```

```
Based on the Levene test, the p-value is 0.089 or 8.85%. This suggest that the
variances are equal.
```

[17]:

```
# Annova test: null: there is no difference
p = sps.f_oneway(m_inc,um_inc,d_inc,w_inc)[1]
print(f"Based on Z test, the p-value is {p:.3f}.")
```

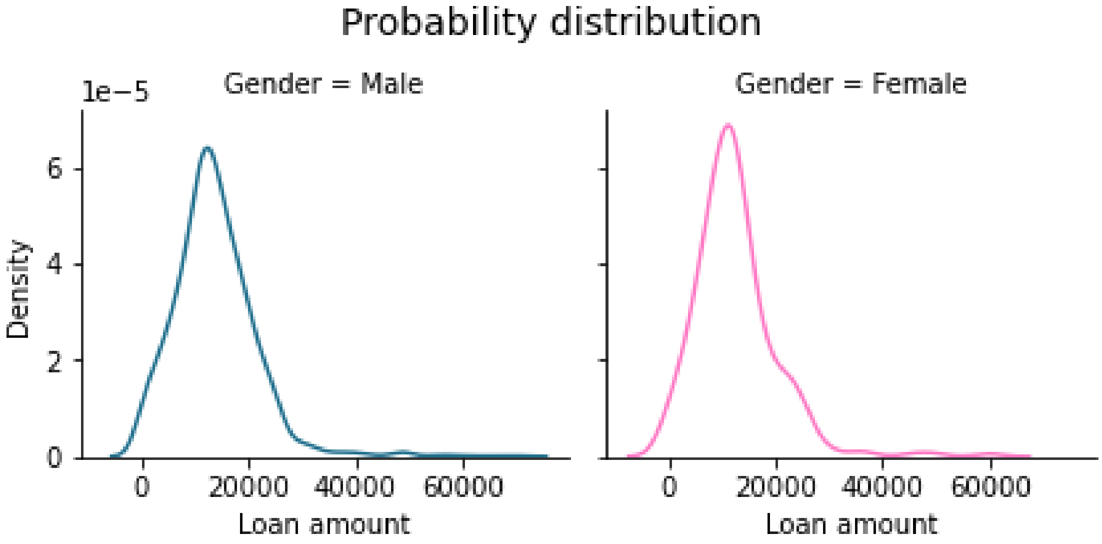
```
Based on Z test, the p-value is 0.116.
```

We find a p-value greater than the significance level. Therefore, we cannot reject the null hypothesis. This means the mean income for all groups widow, married, unmaried and divorced are equal. This shows that although the mean incomes between these groups seem to differ numerically, statistically they are the same. This is in line with our observation in Part 1 section 2.2.

2.3 Is there really a difference in loan amount between gender?

[18]:

```
# normality
facet = sn.FacetGrid(data, col="Gender", hue = "Gender", palette =␣
 ↪["#1F6E8C","#FF78C4"])
facet.map_dataframe(sn.kdeplot,x="LoanAmt")
facet.set_xlabels("Loan amount")
facet.fig.subplots_adjust(top=0.8)
facet.fig.suptitle("Probability distribution", fontsize=14)
plt.show()
```



[19]:

```
loan_gender = pd.pivot_table(data,columns="Gender",␣
  ↪values="LoanAmt",index="CustrNo", aggfunc=np.sum)
```

[20]:

```
f_loan = loan_gender.Female.dropna()
m_loan = loan_gender.Male.dropna()
```

[21]:

```
# homogeneity of variance
p = sps.levene(f_loan, m_loan)[1]
print(f"Based on the Levene test, the p-value is {p:.3f} or {p*100:.2f}%. This␣
  ↪suggest that the variances are equal.")
```

```
Based on the Levene test, the p-value is 0.324 or 32.45%. This suggest that the
variances are equal.
```

[22]:

```
# Annova test
p =sps.f_oneway(f_loan,m_loan)[1]
print(f"Based on Z test, the p-value is {p:.3f}.")
```

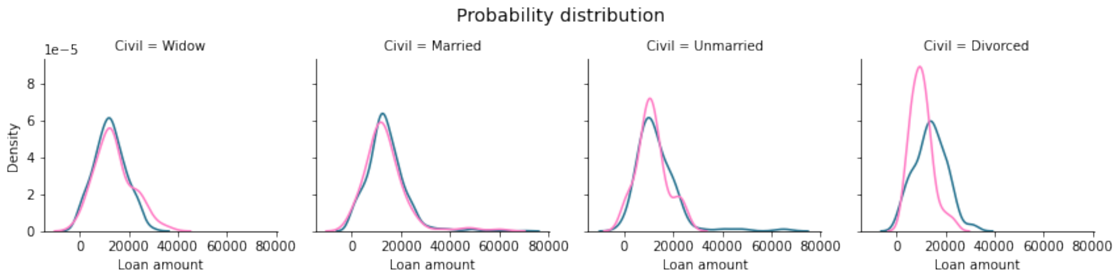
```
Based on Z test, the p-value is 0.028.
```

We find a p-value smaller than the confidence level. Therefore, we can reject the null hypothesis. This means the mean loan amount asked by male customers is indeed different from the mean loan amount asked by female customers. This insight can be very useful to our marketing department for instance. They might undertake actions to increase the number of female that apply for a loan and they might chose to target a specific audience based on gender when it comes to loan offers. Additionally, this is in line with our visualization in Part 1 section 2.3 where we observed that female widow ask for on average far higher loan amount than their counterparts and divorced male also ask for higher loan amount on average compared to other male categories. But we still have not verify whether the difference is significant across civil status.

2.4 Is there really a difference in loan amount between civil status?

[23]:

```
# normality
facet = sn.FacetGrid(data, col="Civil", hue = "Gender", palette =␣
  ↪["#1F6E8C","#FF78C4"])
facet.map_dataframe(sn.kdeplot,x="LoanAmt")
facet.set_xlabels("Loan amount")
facet.fig.subplots_adjust(top=0.8)
facet.fig.suptitle("Probability distribution", fontsize=14)
plt.show()
```



```
[24]: loan_civ = pd.pivot_table(data,columns="Civil",␣
        ↪values="LoanAmt",index="CustrNo", aggfunc=np.sum)
```

```
[25]: m_loan = loan_civ.Married.dropna()
      um_loan = loan_civ.Unmarried.dropna()
      d_loan = loan_civ.Divorced.dropna()
      w_loan = loan_civ.Widow.dropna()
```

```
[26]: # Homogeneity of variance: null: variances are the same
      p = sps.levene(m_loan,um_loan,d_loan,w_loan)[1]
      print(f"Based on the Levene test, the p-value is {p:.3f} or {p*100:.2f}%. This␣
        ↪suggest that the variances are equal.")
```

```
Based on the Levene test, the p-value is 0.287 or 28.65%. This suggest that the
variances are equal.
```

```
[27]: # Annova test: null: there is no difference
      p = sps.f_oneway(m_loan,um_loan,d_loan,w_loan)[1]
      print(f"Based on Z test, the p-value is {p:.3f}.")
```

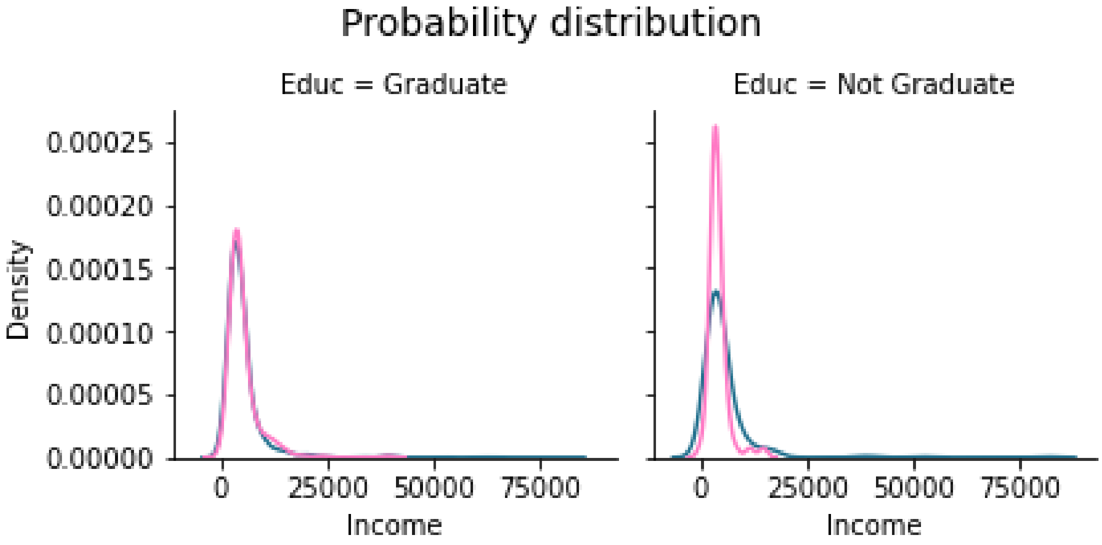
```
Based on Z test, the p-value is 0.059.
```

We find a p-value greater than the confidence level. Therefore, we cannot reject the null hypothesis. This means the mean loan amount for all groups widow, married, unmaried and divorced are equal. This shows that although the mean loan amounts between these groups seem to differ numerically, statistically they are the same. However, this is not in line with our visualization in Part 1 section 2.3 where we observed that female widow ask for on average far higher loan amount than their counterparts and divorced male also ask for higher loan amount on average compared to other male categories. But since we found evidence for the difference between sex, this means that what we observed is better explained by gender than by civil status.

2.5 Is there really a difference in income between education levels?

```
[28]: # Normality
```

```
facet = sn.FacetGrid(data, col="Educ", hue = "Gender", palette =␣
 ↪["#1F6E8C","#FF78C4"])
facet.map_dataframe(sn.kdeplot,x="AppIncome")
facet.set_xlabels("Income")
facet.fig.subplots_adjust(top=0.8)
facet.fig.suptitle("Probability distribution", fontsize=14)
plt.show()
```



[29]:
```
data = data.replace("Not Graduate","NotGraduate")
inc_educ = pd.pivot_table(data,columns="Educ",␣
 ↪values="AppIncome",index="CustrNo", aggfunc=np.sum)
```

[30]:
```
inc_grad = inc_educ.Graduate.dropna()
inc_notgrad = inc_educ.NotGraduate.dropna()
```

[31]:
```
# homogeneity of variance
p = sps.levene(inc_grad, inc_notgrad)[1]
print(f"Based on the Levene test, the p-value is {p:.3f} or {p*100:.2f}%. This␣
 ↪suggest that the variances are equal.")
```

```
Based on the Levene test, the p-value is 0.882 or 88.24%. This suggest that the
variances are equal.
```

[32]:
```
# Annova test
p = sps.f_oneway(inc_grad, inc_notgrad)[1]
print(f"Based on Z test, the p-value is {p:.3f}.")
```

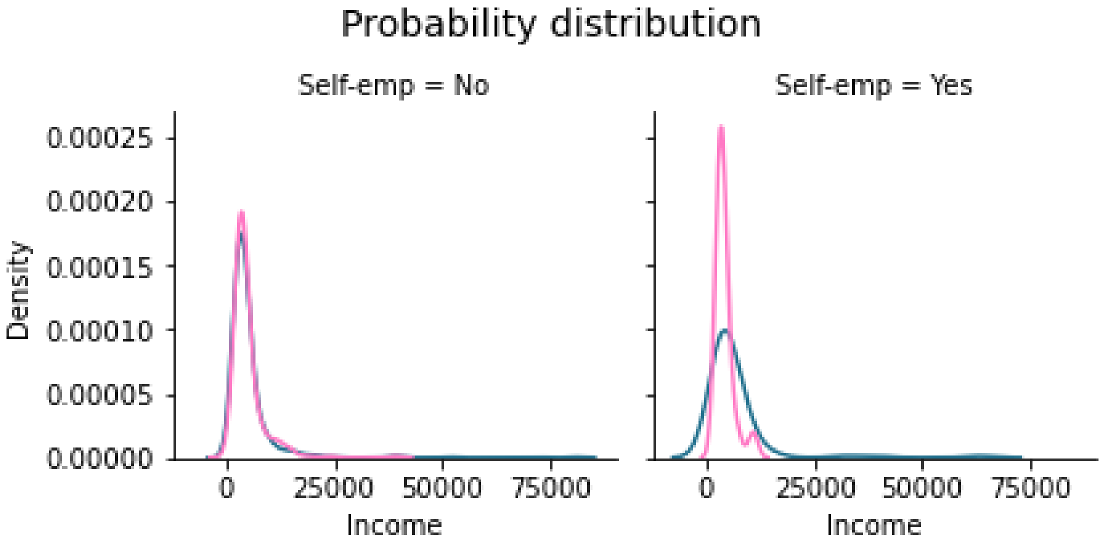
```
Based on Z test, the p-value is 0.883.
```

We find a p-value greater than the confidence level. Therefore, we cannot reject the null hypothesis. This means the mean income for graduate is equal to the mean income for not graduate applicants.Here, the population is the loan applicants not the entire world. Let that be clear. Hence, we do not mean that the mean salary of a typical graduate is not different from the mean salary of a typical non-graduate. We simply mean that for the average income of a typical person who have applied for a loan and has a degree is not different from the average income of a typical person who have applied for a loan and has no degree. In addition, this alligns with our observation in Part 1 section 2.4. There, we observe that the difference in income distribution seems to come from employment types rather than education level. Let examine the difference in employment type.

2.6 Is there really a difference in income between employment types?

[33]:

```
facet = sn.FacetGrid(data, col="Self-emp", hue = "Gender", palette =␣
  ↪["#1F6E8C","#FF78C4"])
facet.map_dataframe(sn.kdeplot,x="AppIncome")
facet.set_xlabels("Income")
facet.fig.subplots_adjust(top=0.8)
facet.fig.suptitle("Probability distribution", fontsize=14)
plt.show()
```



[34]:

```
inc_self = pd.pivot_table(data,columns="Self-emp",␣
  ↪values="AppIncome",index="CustrNo", aggfunc=np.sum)
```

[35]:

```
inc_selfYes = inc_self.Yes.dropna()
inc_selfNo = inc_self.No.dropna()
```

[36]:

```
# homogeneity of variance
p = sps.levene(inc_selfYes, inc_selfNo)[1]
print(f"Based on the Levene test, the p-value is {p:.3f} or {p*100:.2f}%. This␣
 ↪suggest that the variances are equal.")
```

```
Based on the Levene test, the p-value is 0.301 or 30.07%. This suggest that the
variances are equal.
```

[37]:

```
# Annova test
p=sps.f_oneway(inc_selfYes, inc_selfNo)[1]
print(f"Based on Z test, the p-value is {p:.3f}.")
```

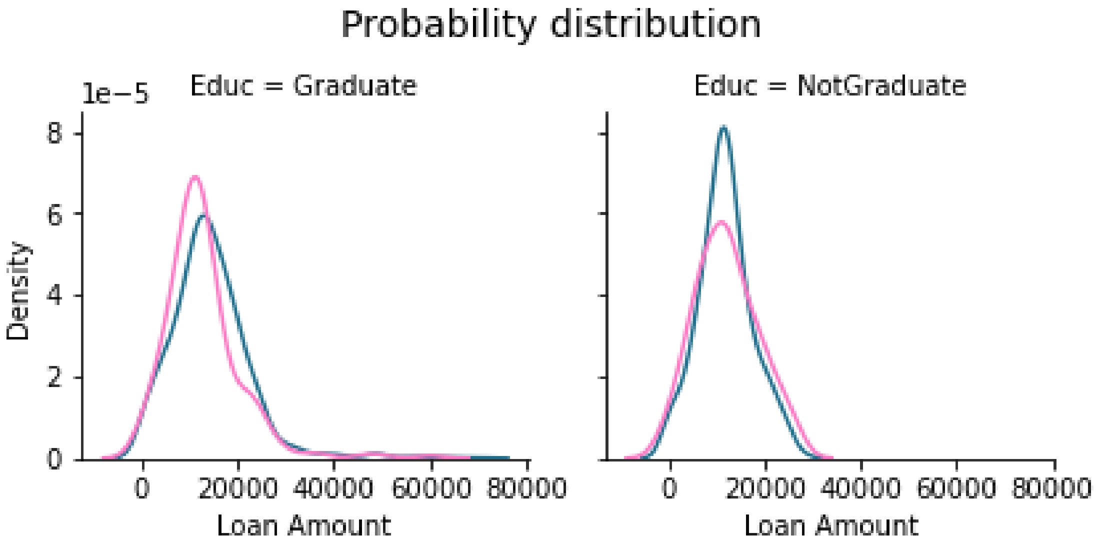
```
Based on Z test, the p-value is 0.110.
```

We find a p-value greater than the confidence level. Therefore, we cannot reject the null hypothesis. This means the mean income for a self-employed is equal to the mean income for not self-employed applicants. However, this doesn't allign with our observation in Part 1 section 2.4. There, we observe that the difference in income distribution seems to come from employment types rather than education level. But the test reveals no statistical significance for the difference.

2.7 Is there really a difference in loan amount between education levels?

[38]:

```
facet = sn.FacetGrid(data, col="Educ", hue = "Gender", palette =␣
 ↪["#1F6E8C","#FF78C4"])
facet.map_dataframe(sn.kdeplot,x="LoanAmt")
facet.set_xlabels("Loan Amount")
facet.fig.subplots_adjust(top=0.8)
facet.fig.suptitle("Probability distribution", fontsize=14)
plt.show()
```



[39]:

```
loan_educ = pd.pivot_table(data,columns="Educ",␣
  ↪values="LoanAmt",index="CustrNo", aggfunc=np.sum)
```

[40]:

```
loan_grad = loan_educ.Graduate.dropna()
loan_notgrad = loan_educ.NotGraduate.dropna()
```

[41]:

```
# homogeneity of variance
p = sps.levene(loan_grad, loan_notgrad)[1]
print(f"Based on the Levene test, the p-value is {p:.3f} or {p*100:.2f}%. This␣
  ↪suggest that the variances are different.")
```

```
Based on the Levene test, the p-value is 0.000 or 0.01%. This suggest that the
variances are different.
```

Therefore, we should adjust our formula to account for this difference.

[42]:

```
p = sps.kruskal(loan_grad, loan_notgrad)[1]
print(f"Based on Z test, the p-value is {p:.3f}.")
```

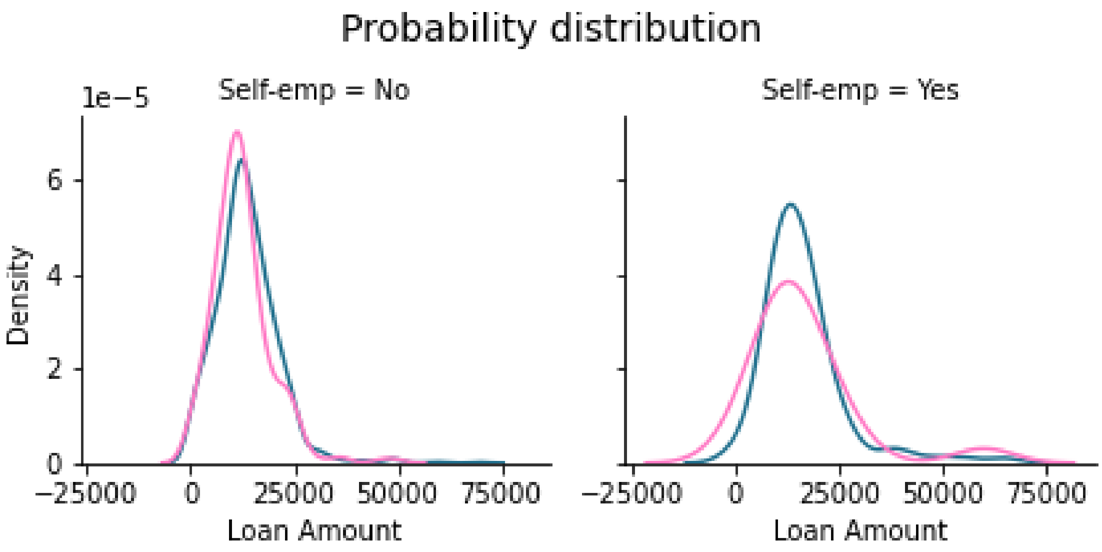
```
Based on Z test, the p-value is 0.000.
```

We find a p-value smaller than the confidence level. Therefore, we can reject the null hypothesis. This means the mean loan amount asked by graduate customers is indeed different from the mean loan amount asked by non-graduate customers. This insight can be very useful to our marketing department. For instance, they might chose to target a speific audience audience when it comes to loan offers. However, in Part 1 section 2.4, we did not observe a big difference in the distribution across education level.

2.8 Is there really a difference in loan amount between employment types?

[74]:

```
facet = sn.FacetGrid(data, col="Self-emp", hue = "Gender", palette =␣
  ↪["#1F6E8C","#FF78C4"])
facet.map_dataframe(sn.kdeplot,x="LoanAmt")
facet.set_xlabels("Loan Amount")
facet.fig.subplots_adjust(top=0.8)
facet.fig.suptitle("Probability distribution", fontsize=14)
plt.show()
```



[75]:
```
loan_self = pd.pivot_table(data,columns="Self-emp",␣
  ↪values="LoanAmt",index="CustrNo", aggfunc=np.sum)
```

[76]:
```
loan_selfYes = loan_self.Yes.dropna()
loan_selfNo = loan_self.No.dropna()
```

[77]:
```
# homogeneity of variance
p = sps.levene(loan_selfYes, loan_selfNo)[1]
print(f"Based on the Levene test, the p-value is {p:.3f} or {p*100:.2f}%. This␣
  ↪suggest that the variances are different.")
```

```
Based on the Levene test, the p-value is 0.036 or 3.63%. This suggest that the
variances are different.
```

[78]:
```
# annova test
p = sps.kruskal(loan_selfYes, loan_selfNo)[1]
print(f"Based on Z test, the p-value is {p:.3f}.")
```

```
Based on Z test, the p-value is 0.006.
```

We find a p-value smaller than the confidence level. Therefore, we can reject the null hypothesis. This means the mean loan amount asked by self-employed customers is indeed different from the mean loan amount asked by non self-employed customers. This is in line with our observation in Part 1 section 2.4.

2.9 Test for independancy among categorical variables

[79]:

```
gender = data[["CustrNo","Gender"]]
educ = pd.pivot_table(data,columns="Educ", values="Gender",␣
 ↪index="CustrNo",aggfunc=np.count_nonzero)
civil= pd.pivot_table(data,columns="Civil", values="Gender",␣
 ↪index="CustrNo",aggfunc=np.count_nonzero)
depend= pd.pivot_table(data,columns="Dependents", values="Gender",␣
 ↪index="CustrNo",aggfunc=np.count_nonzero)
selfemp= pd.pivot_table(data,columns="Self-emp", values="Gender",␣
 ↪index="CustrNo",aggfunc=np.count_nonzero)
default = pd.pivot_table(data,columns="DefaultHist", values="Gender",␣
 ↪index="CustrNo",aggfunc=np.count_nonzero)
othdebt= pd.pivot_table(data,columns="OthDebts", values="Gender",␣
 ↪index="CustrNo",aggfunc=np.count_nonzero)
```

[80]:

```
gender_educ = pd.merge(educ, gender, on="CustrNo")
gender_educ = gender_educ.groupby(by="Gender").count()
gender_educ.drop(columns="CustrNo", inplace=True)
```

- **Does education level depend on gender ?**

[81]:

```
f"p-value is {round(sps.chi2_contingency(gender_educ)[1],5)}"
```

[81]:

```
'p-value is 0.48856'
```

We find a p-value greater than the confidence level. Therefore, we cannot reject the null hypothesis. This means education level does not depend on gender. In other words, there is no evidence to suggest that men are more likely to be of a certain education level and that women are more likely to be of another education level.

- **Does civil status depend on gender ?**

[82]:

```
gender_civil = pd.merge(civil, gender, on="CustrNo")
gender_civil = gender_civil.groupby(by="Gender").count()
gender_civil.drop(columns="CustrNo", inplace=True)
```

[83]:

```
f"p-value is {round(sps.chi2_contingency(gender_civil)[1],5)}"
```

[83]:

```
'p-value is 0.0'
```

We find a p-value smaller than the confidence level. Therefore, we can reject the null hypothesis. This means that civil status choice depends on gender. It could be that women are more likely to be of a certain civil status and men are more likely to be of another civil status within our population of loan applicants.

- **Do employment type depends on gender ?**

[84]:

```
gender_selfemp = pd.merge(selfemp, gender, on="CustrNo")
gender_selfemp = gender_selfemp.groupby(by="Gender").count()
gender_selfemp.drop(columns="CustrNo", inplace=True)
```

```
[85]: f"p-value is {round(sps.chi2_contingency(gender_selfemp)[1],5)}"
```

```
[85]: 'p-value is 0.57289'
```

We find a p-value greater than the confidence level. Therefore, we cannot reject the null hypothesis. This means employment types does not depend on gender. In other words, there is no evidence to suggest that men are more likely to be of a certain employment types and that women are more likely to be of another employment type.

- **Does default history depend on gender ?**

```
[86]: gender_default = pd.merge(default, gender, on="CustrNo")
      gender_default = gender_default.groupby(by="Gender").count()
      gender_default.drop(columns="CustrNo", inplace=True)
```

```
[87]: f"p-value is {round(sps.chi2_contingency(gender_default)[1],5)}"
```

```
[87]: 'p-value is 0.13259'
```

We find a p-value greater than the confidence level. Therefore, we cannot reject the null hypothesis. This means defaulting does not depend on gender. In other words, there is no evidence to suggest that men are more likely to default or not and that women are more likely to default or not.

- **Does having other debts depend on gender ?**

```
[88]: gender_oth = pd.merge(othdebt, gender, on="CustrNo")
      gender_oth = gender_oth.groupby(by="Gender").count()
      gender_oth.drop(columns="CustrNo", inplace=True)
```

```
[89]: f"p-value is {round(sps.chi2_contingency(gender_oth)[1],5)}"
```

```
[89]: 'p-value is 0.39819'
```

We find a p-value greater than the confidence level. Therefore, we cannot reject the null hypothesis. This means having debt does not depend on gender. In other words, there is no evidence to suggest that men are more likely to have debt or not and that women are more likely to have debt or not.

- **Does having dependencies depends on gender ?**

```
[90]: gender_depend = pd.merge(depend, gender, on="CustrNo")
      gender_depend = gender_depend.groupby(by="Gender").count()
      gender_depend.drop(columns="CustrNo", inplace=True)
```

```
[91]: f"p-value is {round(sps.chi2_contingency(gender_depend)[1],5)}"
```

```
[91]: 'p-value is 4e-05'
```

We find a p-value smaller than the confidence level. Therefore, we can reject the null hypothesis. This means that having dependencies depends on gender. It could be that women are more likely to have people that depends on them or not and that men are more likely to bhave people that depends on them or not. We do not know which of the gender is more likely to have dependents but we find evidence that one of them is more likely to do. That is all we can say. This is not surprising however since civil status also differes by gender.

3. Conclusion

We conclude this analysis with some important notes. We found significant differences in loan across gender, education levels and employment types. However,we did not find significant differences in income and loan amount across other groups. Additionally, we found some relationships between cvili status and gender, and between dependencies and gender. This information can be important for other purposes such as marketing & sales for AB bank. They could use it to target specific groups to market their products. However, we still cannot predict which customer is likely to default or not. This is what we intend to do in Part 3.

If you enjoyed reading this analysis, don't hesitate to contact me for any inquiry at eliediwa9@gmail.com.