Homework Assignment 2

Carlos M. Carvalho
Statistics

Problem 1

I am interested in building a portfolio of stocks and bonds... a very convenient way is to invest in two ETFs (Exchange Traded Funds). Let's we choose VTI and VGLT as the ETFs. Build the efficient frontier combining these two ETFs. What allocation gives you the best Sharpe Ratio? If you decide on a 50-50 allocation, what is the probability you will get a return larger than 1% next month? (To make things a little easier, use monthly returns in the last 5 years.)

Problem 2

A credit card company collects data on 10,000 users. The data contained two variables: an indicator of the costumer status, i.e., current ($def = 0$) or in default ($def = 1$) and a measure of their loan balance relative to income, i.e., low ($bal = 1$), medium ($bal = 2$) and high ($bal = 3$). The data is in the following table:

```
     def
--------------------
bal |     0      1  |
--------------------
  1 | 8,940     64  |
  2 |   651    136 |
  3 |    76    133 |
--------------------
```

1. Compute the estimated marginal distribution of costumer status

   $\hat{p}(def = 0) = \frac{8{,}940+651+76}{10{,}000} = 0.9667$
   $\hat{p}(def = 1) = \frac{64+136+133}{10{,}000} = 0.0333$

2. What is the conditional distribution of bal, given $def = 1$?

   $\hat{p}(bal = 1|def = 1) = \frac{64}{64+136+133} = 0.1922$
   $\hat{p}(bal = 2|def = 1) = \frac{136}{64+136+133} = 0.4084$
   $\hat{p}(bal = 3|def = 1) = \frac{133}{64+136+133} = 0.3994$

3. Make a prediction for the status of a costumer with a high balance.

   $\hat{p}(def = 0|bal = 3) = \frac{76}{76+133} = 0.36$
   $\hat{p}(def = 1|bal = 3) = \frac{133}{76+133} = 0.64$

   Therefore, the prediction is $def = 1$, as it has the highest probability.

Problem 3

In a recent episode of Mythbusters, Jamie and Adam (the show's hosts) wanted to determine whether women are better multitaskers than men. To test this theory, they had 10 men and 10 women perform a set of tasks that required multitasking in order to have sufficient time to complete all of the tasks. They use a scoring system that produces scores between 0 and 100.

The women ended up with an average of 72 with a standard deviation of 5, while the men averaged 64 with a standard deviation of 9. In the show, The Mythbusters concluded that this 8 point difference confirms the myth that women are better multitaskers. Based on the results from the experiment, do you agree with their conclusion? Why?

We are looking to make a general statement about the how men and women differ ON AVERAGE. So far, the information in the problem tells us that women are better than men as the scored on average 72 versus 64 of men.

However, before we get to a final conclusion we need to acknowledge that these results are based on a SAMPLE of 10 men and 10 women and could be different if we have seen a different set of men and women. So, we need to figure out the variability in these estimates (ie, average for men and average for women) and think about how could they change if we were to see a different dataset!

The standard error for $\bar{X}$ is what allow us to evaluate this variability and with that we can build confidence intervals... so, we have: $\bar{X}_{Men} = 65$, $\bar{X}_{Women} = 72$, $s_{Men} = 9$, $s_{Women} = 5$.

The standard error for $\bar{X}$ for men is:

$$s_{\bar{X}_M} = \sqrt{\frac{s^2_{Men}}{n}} = \sqrt{\frac{9^2}{10}} = 2.85$$

so the 95% confidence interval is

$$\bar{X}_{Men} \pm 2 \times s_{\bar{X}_M} = 64 \pm 2 \times 2.85 = [58.30; 69.7]$$

The standard error for $\bar{X}$ for women is:

$$s_{\bar{X}_W} = \sqrt{\frac{s^2_{Women}}{n}} = \sqrt{\frac{5^2}{10}} = 1.58$$

so the 95% confidence interval is

$$\bar{X}_{Women} \pm 2 \times s_{\bar{X}_W} = 72 \pm 2 \times 1.58 = [68.84; 75.16]$$

Given the overlap in the confidence intervals we CANNOT conclude at the 95% level that men and women are different in general!

An alternative, more precise way to look into this problem is to build the confidence interval for the difference of means (difference between the mean for females and males in this case)... to that end, we need to compute the standard error for the difference of means, i.e.,

$$s_{(\bar{X}_F - \bar{X}_F)} = \sqrt{\frac{s_F^2}{n_F} + \frac{s_M^2}{n_M}} = \sqrt{\frac{5^2}{10} + \frac{9^2}{10}} = 3.256$$

The confidence interval is then:

$(\bar{X}_F - \bar{X}_M) \pm 2 \times s_{(\bar{X}_F - \bar{X}_F)}$
$= (72 - 64) \pm 2 \times 3.256$
$= (1.48; 14.5)$

Now, we are sure there is a difference between females and males on average... Females are indeed better in multitasking!

Problem 4

During a recent breakout of the flu, 850 out 6,224 people diagnosed with the virus presented severe symptoms. During the same flu season, a experimental anti-virus drug was being tested. The drug was given to 238 people with the flu and only 6 of them developed severe symptoms. Based only on this information, can you conclude, for sure, that the drug is a success?

The general estimate for the rate of people with severe symptoms is

$$\hat{p} = \frac{850}{6,224} = 0.136$$

with standard error

$$s_{\hat{p}} = \sqrt{\frac{(0.136) \times (1 - 0.136)}{6,224}} = 0.0044$$

leading us to the following 95% confidence interval: $[0.136 \pm 2 \times 0.0044] = [0.128; 0.145]$

The estimate for the rate for people that took the drug is:

$$\hat{p} = \frac{6}{238} = 0.025$$

with standard error

$$s_{\hat{p}} = \sqrt{\frac{(0.025) \times (1 - 0.025)}{238}} = 0.01$$

leading us to the following 95% confidence interval: $[0.025 \pm 2 \times 0.01] = [0.005; 0.045]$

Yes, given this information we can conclude that the drug is working... We could revisit this example using the confidence interval for the difference in proportions... however, given the large difference between the two intervals, I am pretty sure the results are going to be the same. You should check anyway!

Now, it turns out that the people who received this drug were all MBA students. Can you infer any causal connection between the the drug and the lack of severe symptoms? What are some potential confounding variables that may influence whether someone develop severe symptoms or not?

Is the drug working or are the people that took the drug generally more healthy and therefore more resistant to the virus' complications... my guess is that MBA students are on average healthier than the general population as they are wealthier, more educated, etc... so, by having data only on the drug for MBA students we may need to be a little skeptical of the real effectiveness of this drug!

Problem 5

Time to revisit the slides... rework the following examples:

- Oracle vs. SAP
- Gender gap in the Chicago banking industry
- Google's new search algorithm

Make sure you understand the computation of the standard errors and that you can answer the questions in the slides, including slide 53!

Problem 6

In 1960, census results indicated that the age at which American men first married had a mean of 23.3 years. It is widely suspected that young people today are waiting longer to get married. We want to find out if the mean age at first marriage has increased during the past 50 years. We plan to test our hypothesis by selecting a random sample of 40 men who married for the first time last year. The men in our sample married at an average age of 24.2 years, with a standard deviation of 5.3 years.

1. Based on a 99% confidence interval, what do you conclude?

   First we need to compute the standard error for the average age...

   $$s_{\bar{X}} = \sqrt{\frac{5.3^2}{40}} = 0.838$$

   so that our 99% confidence interval becomes:

   $$24.2 \pm 3 \times 0.838 = (21.68; 26.71)$$

   So NO, we cannot conclude for SURE that men are getting married later in life as the 23.3 is included in the confidence interval.

2. Now, use a t-stat... explain your conclusion The t-stat is computed as:

   $$t = \frac{24.2 - 23.3}{0.838} = 1.074$$

   so that the current average is only (about) one standard deviation away from the hypothesis (value in 1960) of 23.3!! Therefore, 24.2 and 23.3 are not that different statistically.... so we say that 24.2 is not statistically different from the results from the census of 23.3.

Problem 7

A SurveyUSA poll conducted on March 1, 2011 asked randomly sampled Los Angeles residents about their views on American vs. foreign-made products. One of the questions on the survey was "If an American-made product cost slightly more than a foreign-made product, which would you be more likely to buy?"

81 out of the 166 respondents between the ages of 18 and 34, and 248 out of the 334 respondents 35 years and older said they would prefer the American-made product. We are interested to see if younger people are less likely to choose American-made products. Test this hypothesis at the 5% level.

Based on the data, our current estimates for the proportions are $\hat{p}_Y = 81/166 = 0.488$ and $\hat{p}_O = 248/334 = 0.743$ (Y for "young" and O for "old").

We can now compute the standard error for the difference in proportions:

$$s_{(\hat{p}_Y - \hat{p}_O)} = \sqrt{\frac{\hat{p}_Y \times (1 - \hat{p}_Y)}{n_Y} + \frac{\hat{p}_O \times (1 - \hat{p}_O)}{n_O}} = 0.0455$$

Now, let's compute the 95% confidence interval for the difference in proportions:

$$(\hat{p}_Y - \hat{p}_O) \pm 2 \times s_{(\hat{p}_Y - \hat{p}_O)} = (0.488 - 0.742) \pm 2 \times 0.0455 = (-0.345; -0.163)$$

So, yes, younger people are definitely less likely to choose American-made products based on this dataset as the difference between the proportions is negative for sure.

What if I had decided to answer this question based on the t-stat?

$$t = \frac{(0.488 - 0.742) - 0}{0.0455} = -5.58$$

What do we conclude? Well, "zero" ("no difference" hypothesis) is more than 5 standard deviation away form the difference we see in the data... so yes, there is a difference!

Problem 8

Your dad claims that he shoots better than 50% from the three point line. You bet him $100 that doesn't, based on a challenge where he attempts 10 shots. He makes 7 of his shots and says that you owe him $100. You say that you do not, because you cannot reject the null hypothesis that he shoots exactly 50%. Who is right, here?

Your dad claims that he shoots better than 50% from the three point line. You bet him $100 that doesn't, based on a challenge where he attempts 10 shots. He makes 7 of his shots and says that you owe him $100. You say that you do not, because you cannot reject the null hypothesis that he shoots exactly 50%. Who is right, here?

Let's compute the confidence interval for your dad's 3-point shooting ability (proportion)... first we need the estimated proportion $\hat{p} = 7/10 = 0.7$... then, we need the standard error for the proportion

$$s_{\hat{p}} = \sqrt{\frac{\hat{p} \times (1 - \hat{p})}{n}} = \sqrt{\frac{0.7 \times (0.3)}{10}} = 0.145$$

so, the 95% confidence interval becomes:

$$0.7 \pm 2 \times 0.145 = (0.41; 0.99)$$

so NO, we can't rule out that your dad's shooting ability is actually 50% (or even less than that)!! You win the bet.