15.075 Exam 4

Instructor: Cynthia Rudin
TA: Dimitrios Bisias

December 21, 2011

Grading is based on demonstration of conceptual understanding, so you need to show all of your work.

Problem 1

Choose Y or N for each (you do not need to justify):

Y N You can perform a t-test to determine whether the population mean is different from 10 using a sample of size 400 without the normality of data assumption.

Y N If $X \sim U[-1, 1]$ and $Y = X^2$, then cov(X,Y) = 0.

Y N The hypothesis

$$H_0 : \mu = 0$$
$$H_1 : \mu > 0$$

is rejected at $\alpha = 0.05$ if the 2-sided 90% CI does not contain 0

Y N In multiple regression if r^2 is less than 0.05, then the F statistic of the regression is not significant at the 0.1-level.

Problem 2

For the following scenarios, determine

a. the smallest set of assumptions you need to make in order to do a hypothesis test.

b. which test you might perform. If multiple tests apply, please choose a test that has the fewest assumptions, and among those, choose the strongest test.

c. the hypotheses for the test

d. how you might check the assumptions you made (if you made any)

2.1

You want to check if hedge fund B is riskier i.e. its returns have higher variance, than hedge fund A. For this you collect their yearly returns from 1995-2010.

2.2

You want to check if city A has taller people on average than city B. For this you measure the heights of 200 people from city A and 400 people from city B.

2.3

You want to check if there is a linear relationship between the salary of a CEO, his company's return-on-equity and his company's sales. For this you collect data from 200 large corporations and you regress the CEO salary on the return-on-equity and the company's sales.

2.4

You want to see if income of people is independent of their religion. For this you randomly collect 500 people and you classify them by their religion and whether or not their income is less than 50K per year.

2.5

Of 64 offspring of a certain cross between guinea pigs, 34 were red, 10 were black, and 20 were white. According to the genetic model, these numbers should be in ratio 9/3/4. You want to check if the data are consistent with the genetic model.

2.6

You would like to know whether your waiting time for the bus is, on average, greater than 5 minutes. You measure your waiting time on 20 different days. (Hint: do you know anything about waiting times for buses?)

2.7

You would like to compare LeBron James and Kobe Bryant regarding their performance in free throws. For this for each player you count how many free throws he scored in a season. Check if LeBron has higher success rate than Bryant. (Hint: the usual number of free throws for a good basketball player in a season is in the hundreds.)

Problem 3

We want to find the relationship between the college GPA and college entrance verbal and mathematics test scores. For this reason we collect data from $n = 40$ students as in Table 1. Let y be GPA, x_1 be verbal score, and x_2 be math score.

Letting the GPA be the dependent variable y, a regression is fit on the independent variables verbal score (x_1) and math score (x_2) yielding $\beta_0 = -1.570537$, $\beta_1 = 0.025732$, and $\beta_2 = 0.033615$.

GPA	3.49	2.89	$\cdots$	3.84
Math Score	87	99	$\cdots$	93
Verbal Score	81	68	$\cdots$	87
Fitted GPA	3.4383	3.5071	$\cdots$	3.7943

Table 1: GPA, math and verbal score.

The output below shows the estimates of the coefficients and their standard errors.

```
Coefficients:
             Estimate   Std. Error
(Intercept)-1.570537   0.493749
verbal      0.025732   0.004024
math        0.033615   0.004928
```

You are also told that $\bar{y} = 2.98$, and also that

$$(3.49 - 2.98)^2 + (2.89 - 2.98)^2 + \cdots + (3.84 - 2.98)^2 = 18.7736$$

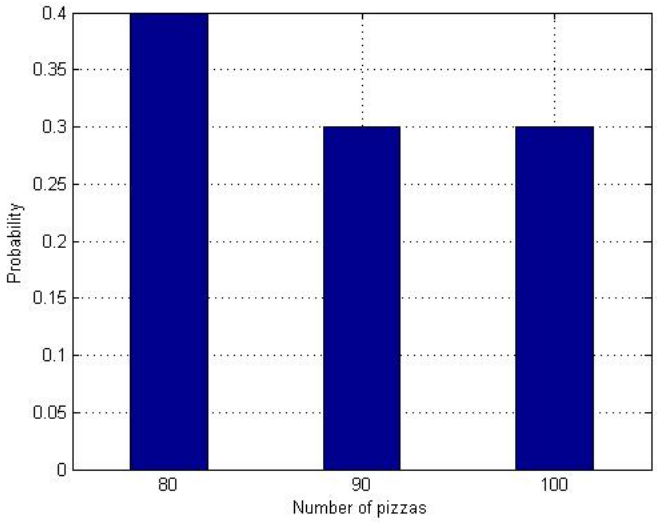
and further you are given that

$$(3.49 - 3.4383)^2 + (2.89 - 3.5071)^2 + \cdots + (3.84 - 3.7943)^2 = 5.9876$$

1. What is the coefficient of multiple determination r^2?
2. Fill the ANOVA table.
3. Test the hypothesis that all of the coefficients are zero at $\alpha = 0.05$.
4. Test the hypothesis that the coefficient for math score is zero at $\alpha = 0.05$.

Problem 4

You are the owner of an italian restaurant. Some days you sell a lot of pizzas and some other days you don't, independently of other days. In fact, the distribution of the number of pizzas you sell each day is shown below.



If we measure pizza sales for 100 days, what is the approximate probability that the average pizza sales (over the 100 days) will be more than 90 pizzas?

Problem 5

We want to explore the association between political party affiliation and age among the US population. For this reason, we take a random sample of 521 US citizens based on their SSN and then we record whether they are older than 40 or not and their political affiliation. The table below shows the results.

Gender	Party A	Party B	Row total
Age ≥ 40	173	125	298
Age < 40	150	73	223
Column Total	323	198	521

1. Which type of sampling is used here?
2. Set up the hypotheses to test whether there is an association between the age and the political party affiliation
3. Perform the test at the $\alpha = 0.05$ level.

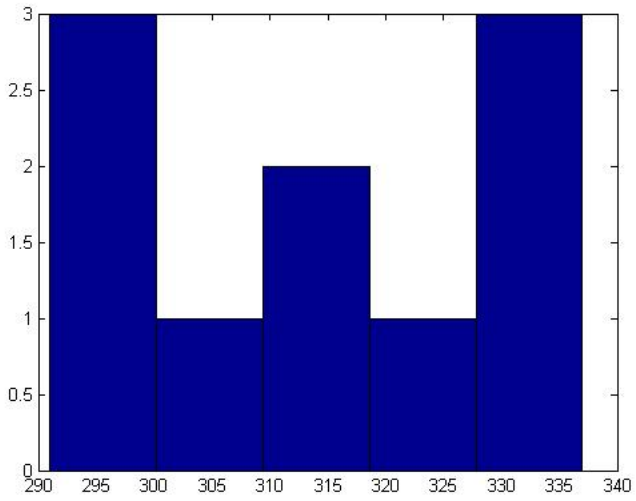
Problem 6

We would like to find some way to claim that the new strawberry cheesecakes at our bakery tend to have more than 300 grams of jam. Here's the measurements of jam for 10 cheesecakes (measured in grams):

Grams of jam	293	311	331	295	337	328	291	306	323	316

Table 2: Measurements of jam.

A histogram of these data is here:



Perform two different hypothesis tests using $\alpha = 0.05$ to determine whether the cheesecakes have significantly more jam than 300 grams. What is your overall conclusion from the tests?

15.075J / ESD.07J Statistical Thinking and Data Analysis

Fall 2011