Vero Beach Invitational
March 2005
Team

Statistics Round | Question #1 | Distributions/probability

HDS is a package delivery service that operates in the United States. Ace Ventura, the chief executive officer of HDS, is conducting an efficiency study for delivery. Two routes in particular are being studied: Richmond to Dallas and New York to Los Angeles.

On the Richmond to Dallas route the company is able to track the delivery times of 17 packages. The delivery times on this route were (in days) 1, 2, 3, 4, and 5.

The distribution of delivery days on this route is summarized below:

Days	1	2	3	4	5
Probability	0.10	0.25	0.30	0.20	0.15

On the significantly busier New York to Los Angeles route the company is able to track the delivery times of 712 packages. The delivery times on this route were (in days) 3, 4, 5, 6, 7.

The distribution of delivery days on this route is summarized below:

Days	3	4	5	6	7
Probability	0.15	0.20	0.30	0.25	0.10

Part (A) : What is the expected delivery time of a package from Richmond to Dallas?

Part (B) : What is the expected delivery time of a package from New York to Los Angles?

Part (C) : The distribution that best approximates the sample mean delivery time of the 17 packages from Richmond to Dallas is the normal distribution. True or False? Explain if false.

Part (D) : The distribution that best approximates the sample mean delivery time of the 712 packages from New York to Los Angles is the normal distribution. True or False? Explain if false.

Part (E) : Assume the distribution above is the underlying distribution, what is the *approximate* probability that a sample of 78 packages from New York to Los Angeles will have an average delivery time in excess of 6.4 days?

March 2005

Statistics Round | Question #2 | Interval Estimation

At a lake in Albemarle County a test is conducted on alligators. An alligator is categorized as a "threat" if it is 60 inches or longer. An alligator is categorized as "adorable" if it is less than 60 inches. 27 Alligators are temporarily caught, measured, and released with the result that 19 were classified as "adorable".

Part (A) : Find a 98% confidence interval for the proportion of alligators in the entire lake that are "threats".

Part (B) : Which of the following claims is valid based on this confidence interval?

Claim 1) This process is 98% effective in finding the population mean
Claim 2) If this statistical sampling process was repeated a large number of times, the resulting confidence intervals would actually contain the population mean 98% of the time.
Claim 3) 98% of the sample means would fall inside the interval.

A laboratory in Waco, Texas conducted a study on blood pressure medication. The lab conducted a study and gathered data on the drug "Cholester-lower". They found that in their sample of 1000 subjects the average cholesterol drop was 175 milligrams per liter with a sample standard deviation of 74.

Part (C) : Find a 80% confidence interval for the population mean drop in cholesterol.

Part (D) : Find a 90% confidence interval for the population mean drop in cholesterol *if it is known that the population standard deviation is 79.4.*

Part (E) : Give an intuitive answer for why your methods in (C) and (D) differ. You will need to reference the student t-distribution and the normal distribution properties. Be as thorough and as detailed as possible without going into excess.

Part (F) : Assume that the study was conducted again finding the average drop was 170 milligrams per liter with a sample standard deviation of 75 for a sample of 100 subjects. Using this data, find a 95% confidence interval for the population variance, σ^2.

March 2005

Statistics Round | Question #3 | Linear Modeling

The consulting firm Cheny, Rumsfeld, and Bush has been hired to conduct a study on the effects of Rove Brand Cigarettes on the lung capacity of its customers. Their report will be used to make policy about cigarette company regulation.

For the study, the two variables below were analyzed.

X = The average number of cigarettes smoked per day by the customer.
Y = Their lung capacity measurement (as a percentage of average healthy person their age).

To enable your calculations, the data are shown below. To aid in your calculations, some information is included to the right of the data to avoid mindless number crunching.

X	Y
13	70.13
6	86.45
4	86.87
10	75.14
11	72.11
7	82.75
7	81.45
8	80.11
10	74.35
15	69.83
4	90.11
5	84.48
0	98.22
1	96.75
8	82.8
10	74.31
15	66.92
13	71.43
3	94.37
12	69.66

$\sum_{i=1}^{20} X_i = 162$ $\quad\quad$ $\sum_{i=1}^{20} Y_i = 1608.24$

$\sum_{i=1}^{20} (X_i)^2 = 1682$ $\quad\quad$ $\sum_{i=1}^{20} (Y_i)^2 = 131068.4898$

$\sum_{i=1}^{20} X_i Y_i = 12240.22$ $\quad\quad$ *Number of subjects* = 20

Part (A) : Calculate a regression line using the least squares error criterion. Regress Y on X (lung capacity on average number of cigarettes smoked). Write your regression equation on the answer sheet for part (A). To receive credit, you must use the form $y = \beta_0 + \beta_1 X + e$.

Part (B) : Give the interpretation for your value of $\hat{\beta}_0$.

Part (C) : Give the interpretation for your value of $\hat{\beta}_1$.

Part (D) : Based on your value of $\hat{\beta}_1$, is this a positive or negative association between X and Y?

Part (E) : Calculate the value of r, the correlation coefficient. Make sure to denote what units your coefficient is measured in. Round your answer to the nearest thousandth.

Part (F) : What is the range of possible values for the correlation coefficient in general? Express your answer in interval notation, e.g. [2,4] or (-2,1), etc.

Part (G) : What percentage of variation in the lung capacity of subjects can be explained by the average number of cigarettes smoked per day?

Part (H) : How would the value of the correlation coefficient change if you measured the number of cigarettes smoked as a number of packs instead of individual cigarettes?

Part (I) : Based on the given data, is it reasonable to attempt to predict the lung capacity of someone who consumes an average of 22 cigarettes daily? Why or why not? If it is reasonable, perform the calculation and in addition to your explanation, provide the predicted value.

Part (J) : Based on the given data, is it reasonable to attempt to predict the lung capacity of someone who consumes an average of 9 cigarettes daily? Why or why not? If it is reasonable, perform the calculation and in addition to your explanation, provide the predicted value.

Part (K) : Based on the given data, is it reasonable to attempt to predict the lung capacity of someone who only smokes two days a month, consuming five cigarettes each of those days? Why or why not? If it is reasonable, perform the calculation and in addition to your explanation, provide the predicted value.

Part (L) : Test the null hypothesis that $\beta_1 = 0$ versus the two sided alternative. Provide a value for your test statistics as well as your conclusion written out with an explanation. You must *clearly* state your test statistic and result (reject or not) with an explanation to receive credit.

(Notes: Allow yourself a type I error probability of 0.05; It is not necessary to provide a P-Value, only the test statistic)

Part (M) : A government official with the Food and Drug Administration (FDA) is interested in regulating the cigarette industry to protect consumers from what he perceives as negative effects of cigarettes. He testifies in front of the United States Congress "This study clearly shows that cigarette consumption causes decreased lung capacity among consumers." Using your knowledge gained from Parts (A) thru (L), is his statement accurate or misleading to the Congress? Explain.

Part (N) : If you created a residual plot for this data and your regression line, what kind of pattern would you expect in the residual plot if the linear model was appropriate for this data?

March 2005

Statistics Round Question #4 Grab Bag

Below are the summary statistics for the number of inmates per year (measured in hundreds of inmates) in the Westchester County, New York Department of Corrections for 84 years (1921-2005).

N	84
Mean	470.75
Median	472.74
St Dev	59.40
SE Mean	6.48
Min	316.92
Max	581.81
Q1	431.32
Q3	504.02

Part (A) : Describe in detail a procedure to detect outliers in this data.

Part (B) : Are there any outliers in the data? If so are they low-outliers or high-outliers? Appeal to your method in Part (A) to justify.

Part (C) : Suppose in 2006 that Westchester County imprisons 54900 people and the state governor objects saying that the county is "jailing more people than you should reasonably expect." Evaluate his statement.

Part (D) : Now suppose (only for this part) that the prisoner data is normally distributed with mean 470 and variance 3600. Using the empirical rule for the normal distribution, how likely is it that the county will imprison between 29000 and 65000 people?

Part (E) : Now suppose that a new Attorney General is elected and vows to crackdown on crime with sweeping reforms of the police department in 2006. Does this affect the validity of your answer to Part (D)?

March 2005

Statistics Round | Question #5 | Goodness of Fit and χ^2

NOTE : For this question, do NOT round intermediate answers.

Several years ago the Supreme Court ruled that affirmative action was not an acceptable standard for admission to Bob James University. The Chancellor of the University stands accused of having unfair admissions practices and is depending on you to defend him.

You are presented with the following data depicting the racial compositions of the entire applicant pool from 2002 and the racial composition of those offered admission in 2002. The Chancellor claims that the school's admissions policy guarantees offers of admission that accurately reflect the racial composition of the entire applicant pool.

	Entire Applicant Pool	Offers of Admission
Caucasian/White	26320	275
African American	14708	168
Hispanic/Latino	12980	112
Asian/Pacific Islander	9502	92
American Indian	3876	41
Totals	67386	688

Part (A) : If the offers of admission WERE in fact reflective of the entire applicant pool, how many students from each ethnic background would you expect to be offered admission? Round your answer in this part to two decimal places, e.g. #.##

Part (B) : Using the answers from part (A) that are rounded to two decimal places, calculate the chi-squared statistic for the goodness-of-fit test. You should specifically show the value of the statistics and the associated degrees of freedom. Do not round interim calculations, but round your final value of the test statistic to two decimal places, e.g. #.##

Part (C) : Test the claim of proportional offers of admission at the $\alpha = 0.05$ level. Clearly state your conclusion and reasoning.

Part (D) : Assume that the US Court of Appeals has set the standard that proportional representation ±3% for each category is acceptable. However, if the University deviates from that 6% interval for each category, they are assessed a fine of $100,000 per category (ethnic group) violated. What would the fine amount have been for the University in 2002? (Note: Use the "expected value" answers from part A that are rounded to two decimal places.)