

**Economics 30330: Statistics for Economics**  
**Problem Set 2**  
*University of Notre Dame*  
**Instructor: Julio Garín**  
**Spring 2012**

## Descriptive Statistics (60 points)

1. Following a recent government shutdown, Minnesota Governor Mark Dayton proposes to give all state employees a \$500 raise.

(a) What would this do to the average monthly salary of state employees? To the SD? Explain.

*Suppose there are  $n$  state employees in Minnesota. Let  $x_i$  denote the wage for employee  $i$  before the raise, and let  $y_i$  denote the wage after the raise. Then we have*

$$y_i = x_i + 500$$

*Therefore,*

$$\begin{aligned}\mu_y &= \frac{\sum_i y_i}{N} \\ &= \frac{\sum_i (x_i + 500)}{N} \\ &= \frac{\sum_i x_i + 500N}{N} \\ &= \frac{\sum_i x_i}{N} + 500 \\ &= \mu_x + 500\end{aligned}$$

*Therefore, the mean increases by \$500. In other words, if employees across the entire salary distribution receive the same raise, this is simply a linear transformation of the mean, and the entire distribution will shift up by \$500.*

*Similarly, for the variance:*

$$\begin{aligned}\sigma_y &= \sqrt{\frac{\sum_i (y_i - \mu_y)^2}{N}} \\ &= \sqrt{\frac{\sum_i [(x_i + 500) - (\mu_x + 500)]^2}{N}} \\ &= \sqrt{\frac{\sum_i [(x_i + 500) - (\mu_x + 500)]^2}{N}} \\ &= \sqrt{\frac{\sum_i (x_i - \mu_x)^2}{N}} \\ &= \sigma_x\end{aligned}$$

*Hence, the standard deviation does not change. If employees across the entire salary distribution receive the same raise, the distribution itself will shift up but will not change shape.*

- (b) What would a 5 percent increase in salaries of all state employees do to the average monthly salary? To the SD? Explain.

$$y_i = 1.05x_i$$

Hence,

$$\begin{aligned}\mu_y &= \frac{\sum y_i}{N} \\ &= \frac{\sum (1.05x_i)}{N} \\ &= 1.05 \frac{\sum x_i}{N} \\ &= 1.05\mu_x\end{aligned}$$

Similarly,

$$\begin{aligned}&= \sqrt{\frac{\sum_i (y_i - \mu_y)^2}{N}} \\ &= \sqrt{\frac{\sum_i (1.05x_i - 1.05\mu_x)^2}{N}} \\ &= \sqrt{\frac{\sum_i [1.05(x_i - \mu_x)]^2}{N}} \\ &= \sqrt{1.05^2 \frac{\sum_i (x_i - \mu_x)^2}{N}} \\ &= 1.05 \sqrt{\frac{\sum_i (x_i - \mu_x)^2}{N}} \\ &= 1.05\sigma_x\end{aligned}$$

Therefore, both the mean and standard deviation increase by 5%. A 5% increase in salaries of all employees with both change the shape of and shift the salary distribution. This is because even though the percentage increase in salaries is the same for all employees, the level salary change is not (e.g., an employee with a \$30,000 salary will receive a \$1500 raise while an employee with a \$40,000 salary will receive a \$2000 raise).

2. The Associated Press Team Marketing Report listed the Dallas Cowboys at the team with the highest ticket prices in the National Football League (*USA Today*, October 20, 2009). Data showing the average ticket price for a sample of 14 teams in the NFL are as follows.

Team	Ticket Price (\$)	z-score	Team	Ticket Price (\$)	z-score
Atlanta Falcons	72	-0.111	Green Bay Packers	63	-0.444
Buffalo Bills	51	-0.888	Indianapolis Colts	83	0.296
California Panthers	63	-0.444	New Orleans Saints	62	-0.481
Chicago Bears	88	-0.481	New York Jets	87	0.444
Cleveland Browns	55	-0.740	Pittsburgh Steelers	67	-0.296
Dallas Cowboys	160	3.143	Seattle Seahawks	61	-0.518
Denver Broncos	77	0.074	Tennessee Titans	61	-0.518

- (a) Compute the range, interquartile range, and the median ticket price.

*Max: \$160 (Dallas Cowboys); Min: \$51 (Buffalo Bills):*

$$\mathbf{Range} = \$160 - \$51 = \$109$$

*Q<sub>1</sub>: \$61 (Tennessee Titans); Q<sub>3</sub>: \$83 (Indianapolis Colts):*

$$\mathbf{IQR} = \$83 - \$61 = \$22$$

*7th position: Green Bay Packers (\$63); 8th position Pittsburgh Steelers (\$67):*

$$\mathbf{Median} = \frac{63 + 67}{2} = 65$$

- (b) Compute the mean ticket price. The previous year, the mean ticket price was \$72.20. What was the percentage increase in the mean ticket price for the one-year period?

$$\bar{x} = \frac{\sum_i x_i}{n} = \frac{1050}{14} = 75$$

*Percentage increase:  $\frac{75.00 - 72.20}{72.20} = \frac{2.80}{0.0388}$ . Hence, ticket prices increased by 3.88%.*

- (c) Compute the sample variance and sample standard deviation. Interpret.

$$s^2 = \frac{\sum_i (x_i - \bar{x})^2}{n - 1} = \frac{9504}{13} = 731.08$$

$$s = \sqrt{s^2} = \sqrt{731.0769} = 27.04$$

*The ticket price of each team is, in average, \$27.04 away from the mean of the sample.*

- (d) Compute the standardized values (z-scores) for the listed ticket prices in the table above. Should the Dallas Cowboys' ticket price be considered an outlier? Explain.

*The z-score for the Cowboys' ticket price is 3.14. Since this is greater than 3 standard deviations from the mean, it would be considered an outlier.*

- (e) What are the mean and standard deviation of the standardized ticket prices? Explain.

*The mean of the standardized ticket prices is zero and the standard deviation is one. This is because the mean of any standardized values (i.e., z-scores) is always zero and the standard deviation is always one.*

3. The flashlight batteries produced by a Uruguayan manufacturer are known to have an average life of 60 hours with a standard deviation of 4 hours. Use Chebyshev's theorem to answer parts (a) through (c).

- (a) At least what percentage of batteries will have a life of 54 to 66 hours?

$$\bar{x} = 60 \text{ and } s = 4$$

$$[54, 66] = 60 \pm 4 = \bar{x} \pm 1.5s$$

$$1 - \frac{1}{1.5^2} = 0.5555$$

*By Chebyshev's theorem, at least 55.55% of batteries are within 1.5 standard deviations of the mean.*

- (b) At least what percentage of the batteries will have a life of 52 to 68 hours?

$$[52, 68] = 60 \pm 8 = \bar{x} \pm 2s$$

$$1 - \frac{1}{2^2} = 0.75$$

By Chebyshev's theorem, at least 75% of batteries are within 2 standard deviations of the mean.

- (c) Determine an interval for the battery lives that will be true for at least 80% of the batteries.

$$1 - \frac{1}{z^2} = 0.80$$

$$z = \sqrt{5}$$

$$\bar{x} \pm \sqrt{5} = 60 \pm \sqrt{5} = [51.06, 68.94]$$

- (d) Suppose we know that the battery lives have a normal distribution. Approximately what percentage of batteries will have a life of 50 to 70 hours? Why?

$$[50, 70] = 60 \pm 2.5s$$

Since  $2.5s > 2s$ , by the Empirical Rule, more than 95% of values will have a life of 50 to 70 hours.

- (e) Explain Chebyshev's Theorem in your own words.

One example: Chebyshev's Theorem tells us that in any distribution of data, almost all values will be close to the mean.

4. The following is the frequency distribution for the speeds of a sample of Notre Dame students driving from South Bend to Chicago.

Speed (MPH)	$f_i$	$M_i$	$f_i M_i$	$(M_i - \bar{x})^2$	$f_i (M_i - \bar{x})^2$
50-54	2	52	104	225	450
55-59	4	57	228	100	400
60-64	5	62	310	25	125
65-69	10	67	670	0	0
70-74	9	72	648	25	225
75-79	5	77	385	100	500

- (a) Complete the table.  
 (b) Compute the sample mean, variance, and standard deviation of the grouped sample.

$$\bar{x} = \frac{\sum_i f_i M_i}{n} = \frac{2345}{35} = 67$$

$$s^2 = \frac{\sum_i f_i (M_i - \bar{x})^2}{n - 1} = \frac{1700}{34} = 50$$

$$s = \sqrt{s^2} = \sqrt{50} = 7.07$$

- (c) What does  $M_i$  represent and why do we use it?

$M_i$  denotes the midpoint of each class  $i$ . We use it because we do not know the individual values for all observations and therefore cannot calculate the mean values.

5. The following table lists the study time and exam scores for a sample of 5 students in a college statistics class.

Score	Minutes Spent Studying	Hours Spent Studying
60	60	1
80	180	3
75	90	1.5
95	240	4
85	225	3.75

- (a) Calculate the sample covariance between test score and minutes spent studying.

Let  $x$  denote score and  $y$  denote minutes spent studying.

$$s_{xy} = \frac{\sum_i (x_i - \bar{x}) \sum_i (y_i - \bar{y})}{n - 1} = \frac{3870}{4} = 967.5$$

- (b) Calculate the sample covariance between test score and hours spent studying.

Let  $x$  denote score and  $z$  denote hours spent studying.

$$s_{xy} = \frac{\sum_i (x_i - \bar{x}) \sum_i (y_i - \bar{y})}{n - 1} = \frac{64.5}{4} = 16.125$$

- (c) Calculate the correlation coefficient between test score and minutes spent studying.

$$s_x = \sqrt{\frac{\sum_i (x_i - \bar{x})^2}{n - 1}} = \sqrt{\frac{670}{4}} = 12.94$$

$$s_y = \sqrt{\frac{\sum_i (y_i - \bar{y})^2}{n - 1}} = \sqrt{\frac{6480}{4}} = 80.5$$

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{967.5}{(12.94)(80.5)} = 0.9287$$

- (d) Calculate the correlation coefficient between test score and hours spent studying.

$$s_x = \sqrt{\frac{\sum_i (x_i - \bar{x})^2}{n - 1}} = \sqrt{\frac{670}{4}} = 12.94$$

$$s_z = \sqrt{\frac{\sum_i (y_i - \bar{z})^2}{n - 1}} = \sqrt{\frac{1.8}{4}} = 1.34$$

$$r_{xz} = \frac{s_{xy}}{s_x s_y} = \frac{16.125}{(12.94)(1.34)} = 0.9287$$

- (e) Why did we calculate two different measures to describe the relationship between time spent studying and test scores? Is one measure more useful than the other? Explain.

*Covariance depends on the unit of measurement, so our values for the covariance between time spent studying and test scores differed depending on whether we used hours or minutes to measure time. We calculated the correlation coefficient because this measure of the relationship between two variables does not depend on units of measurement. Since our interest is in the relationship between time spent studying and test scores, regardless of how we measure time, the correlation coefficient is a better measure of the relationship between these two variables.*

6. You will need *Excel*, *Stata*, or another statistical programming software to complete this part of the assignment. Please calculate the answers in the software of your choice and report them here. You do not need to turn in your actual *Excel* spreadsheet, code, etc.

NBA teams are increasingly using statistics to inform their decisions on the court. One example is the analysis of “bad” players. David Berri defines a “bad” player as one that has half the Wins Produced per 48 minutes (WP48) value of a good player, or a WP48 of 0.050 or less; an average player has a WP48 of 0.100. The dataset named ‘NBA’ contains data on “bad” players for NBA teams in 2011.

- (a) Are we working with a sample or a population here? Justify your answer.  
*Since our data include all teams in the NBA, we are working with the population of NBA teams.*

- (b) Calculate and interpret the following measures for *Bad Players*: mode, median, mean, and standard deviation.

*Mode: 8*

*Mean: 8.63*

*Median: 8*

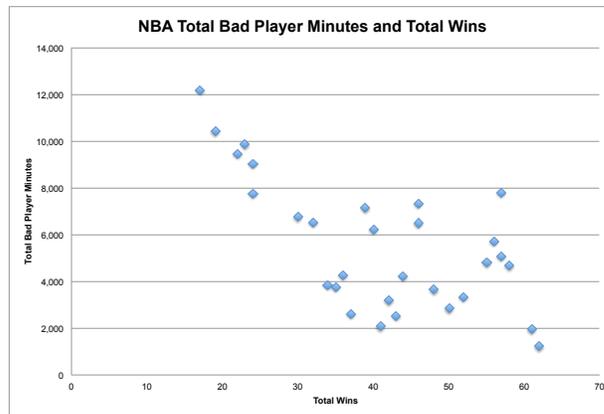
*Standard Deviation: 2.63*

- (c) Comment on the skewness of the data for *Bad Players* based on the measures you calculated in (a) (note: you do not actually need to calculate the skewness coefficient). What does the skewness tell us about the distribution of “bad” players?

*The mean is greater than the median, so the data are slightly skewed right.*

- (d) Create a scatterplot of *Total Bad Player Minutes* and *Total Wins* (print and attach a copy of the scatterplot to your assignment; be sure to label the axes). Based on the plot, what do you expect the direction and approximate magnitude of the correlation between these two variables to be? Explain.

*The plot shows a negative relationship between Total Bad Player Minutes and Total Wins, but the values are more dispersed as the number of Total Wins increases. We should therefore expect that the correlation will be about in the middle of 0 and -1.*



- (e) Calculate the covariance and correlation between *Bad Player Minutes* and *Total Wins*. What do you expect would happen to the covariance if you instead calculated *Bad Player Hours* instead of *Minutes*?

*Covariance: -23,911*

*Correlation: -0.67*

*If we calculated the covariance of Bad Player Hours instead of Minutes, the covariance should decrease.*