Northwestern University

STAT 202-0 Section 23

Assignments

Fall 2013

Instructor: Jiangtao Gou

Practical Question 1

The First Ladies of the United States of America

Much has been written about U.S. Presidents, but what about their wives? I guess you know much more about Presidents than First Ladies. In this practical question, we will investigate zodiac signs and ages of first ladies. We will use the data as provided by Wikipedia, via <http://en.wikipedia.org/wiki/List_of_First_Ladies_of_the_United_States>.

You may google "list of First Ladies wiki" to find this list.

There are 12 columns in this list.

1. Number
2. Image
3. Full name
4. DOB
5. Date of marriage
6. President
7. Date tenure began
8. Age at tenure start
9. Date tenure ended
10. Length of retirement
11. Date of death
12. Lifespan

In this practical question, we will use column 4, DOB, and column 12, lifespan.

You need to round the lifespan to the nearest integer year. For example, lifespan 61 years 182

days will be rounded down to 61 years, and lifespan 84 years 183 days will be rounded up to 85 years.

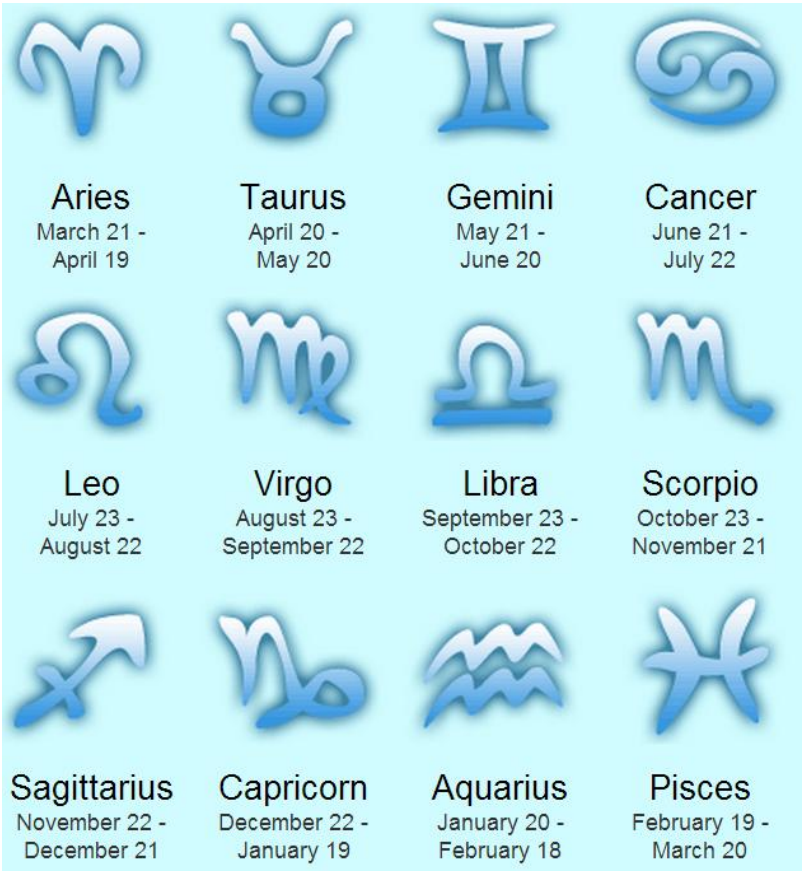
A snapshot of the beginning of this list is shown below.

#	Image	First Lady	Date of birth	Date of marriage	President (Husband)	Date tenure began	Age at tenure start	Date tenure ended	Length of retirement	Date of death	Lifespan
1		**Martha Washington** [12][13] (née Dandridge)	June 2, 1731	January 6, 1759	George Washington	April 30, 1789	57 years, 332 days	March 4, 1797	1,904 days	May 22, 1802	25,921 days (70 years, 354 days)
2		**Abigail Adams** [14][15] (née Smith)	November 11, 1744	October 25, 1764	John Adams	March 4, 1797	52 years, 113 days	March 4, 1801	6,447 days	October 28, 1818	27,013 days (73 years, 351 days)

At the very beginning you need to summarize your data, here is an example.

Name	Star Sign	Element	Age at Death
M. Washington	Capricorn	Earth	71

The relations between star signs and birthdates are listed in this table.



Each sign is associated with one of the four classical elements.

Element	Signs
Fire	Aries; Leo; Sagittarius
Air	Gemini; Libra; Aquarius
Earth	Taurus; Virgo; Capricorn
Water	Cancer; Scorpio; Pisces

(a) For all 51 First Ladies, **(a-1)** make a bar graph according to the categorical variable "Star Sign", and **(a-2)** make a pie chart according to the categorical variable "Element". **(a-3)** What is your observation?
(Hint: You may need to use a frequency table to summarize the data at first.)

(b) For 45 First Ladies who are not alive, **(b-1)** make a histogram or stemplot according to the quantitative variable "lifespan", **(b-2)** compute the five-number summary of First Ladies' lifespan, and make a boxplot (please use the second type of boxplot which I have introduced in class, which marks possible outliers, and is the most commonly used boxplot). **(b-3)** What is your observation?
(Hint: You need to round the lifespan to integer year at first.)

(c) For 45 First Ladies who are not alive, **(c-1)** compute the mean of lifespan, **(c-2)** compute the SD of lifespan, **(c-3)** what is the ratio of IQR to SD?
(Hint: SD = standard deviation, IQR = interquartile ragne.)

Practical Question 2

Height and Weight

We made an anonymous survey on October 9, 2013, and got 18 responses. We assigned each response a number. In the table below, we summarized gender, weight (pounds), and height (feet-inches).

No.	gender	weight	height
1	M	137	5'8"
2	M	155	5'10"
3	M	235	5'10"
4	M	162	5'10"
5	F	120	5'5.5"
6	M	140	5'8"
7	M	210	6'5"
8	F	145	5'4"
9	F	122	5'4"
10	M	170	6'2"
11	M	250	6'3"
12	M	140	5'11"
13	M	155	5'10"
14	F	135	5'6"
15	F	158	5'5"
16	F	110	5'5"
17	M	165	5'10"
18	M	145	5'6"

We also need to learn how to get information from a real statistical report. In this assignment, the real statistical report is ***National Health Statistics Reports, Number 10, October 22, 2008***. You may either find it via CDC (Centers for Disease Control and Prevention) website <http://www.cdc.gov/nchs/data/nhsr/nhsr010.pdf>, or from Blackboard under folder ***Week Three***.


National Health Statistics Reports

Number 10 ■ October 22, 2008

Anthropometric Reference Data for Children and Adults: United States, 2003–2006

by Margaret A. McDowell, Ph.D., M.P.H., R.D.; Cheryl D. Fryar, M.S.P.H.; Cynthia L. Ogden, Ph.D.; and Katherine M. Flegal, Ph.D.

(a) **(a-1)** Download National Health Statistics Report No. 10, and take a look at Table 12 on Page 16 about Male (20 years old or above) Standing height in the United States. We use one row in this table of all races and ethnicity groups, 20 years and over. What is the mean, median, lower quartile and upper quarter? **(a-2)** Try to estimate SD of the height of male, all races and ethnicity groups, 20 years and over, based on the information in this row. (Hint: Take a look at

the percentiles. Do you remember any relation between SD and percentiles?) **(a-3)** Based on the heights of 12 male students in our survey, what is the average height and SD? Compare the mean and SD in (a-2) and (a-3), and write down what you observe.
(Note: there is a column in tables in National Health Statistics Report No. 10 called standard error, which is different from standard deviation, and we will introduce this concept later in this course.)

(b) If you want to be an Air Force Pilot, the Height restrictions for male is 62" - 78". Assume that height of male, all races and ethnicity groups, 20 years and over is normally distributed and based on the mean and SD from part (a), what is the percentage of male adults in the United States who can pass this height measurement?

(c) **(c-1)** Make a scatter plot between height and weight, including all 18 students, and use two different symbols to indicate male students and female students in your plot. **(c-2)** Compute the correlations between height and weight for (i) male students, (ii) female students, and (iii) all students. Do you observe something unexpected? Try to explain it.

Practical Question 3

Regression: From shoe size to height

We made an anonymous online survey from October 16th to 17th, and got 12 responses. We assigned each response a number. In the table below, we summarized gender, shoe size (US system), and height (feet-inches).

	Gender	Shoe Size	Height
1	F	6.5	5'4"
2	M	8	5'10"
3	F	5.5	5'5.5"
4	M	14	6'3"
5	M	9	5'11"
6	M	11.5	5'10"
7	F	9	5'6"
8	M	13	6'5"
9	F	6.5	5'5"
10	M	12	5'10"
11	M	11	6'2"
12	M	10.5	5'10"

In the United States, the most common system for shoe size is
man shoe size = 3 × foot length in inches – 22
woman shoe size = 3 × foot length in inches – 20.5

You need to calculate each individual's foot length based on gender and his/her shoe size at first.

(a) Please find the least-squares regression line for predicting height (response variable) from foot length (explanatory variable). **(a-1)** Make a scatterplot and draw your line on the plot. **(a-2)** Report your slope, intercept, correlation and coefficient of determination. Do you think that the linear model is appropriate for these data? **(a-3)** Usually extrapolation causes much more troubles than interpolation. Shaquille "Shaq" O'Neal is a famous American retired basketball player. He is also famous for his physical stature, at 7 feet 1 inch, 325 lb, and US shoe size 23. Try to apply your linear model to predict Shaq's height from his shoe size. Does your model work reasonably well?

(b) Criminal investigation. **(b-1)** Compute residuals based on the values of real height and predicted height, and make a residual plot. Does the residual plot look good? **(b-2)** Compute SD of residuals, which is a rough (not exact) measure of the spread for predicting an individual response. **(b-3)** Criminal investigators often need to predict unobserved characteristics of

individuals from observed characteristics. For example, if a footprint is left at the scene of a crime, how accurately can we estimate that person's height based on the length of the footprint? Suppose that the footprint left is 10 inches long, by using the linear prediction model you get in part (a) to compute your predicted height (treat it as the average of a lot of possible heights), and using the SD of residuals which you get in (b-2) (treat it as the standard deviation of an individual's height), please give a height range that approximately 95% of the observations fall in. In other word, you may conclude that with 95% confidence that this criminal's height is in this height range.
(Hint: Normal distribution, 68-95-99.7 Rule)

(c) In class, we have learned how to compute slope and intercept of a simple linear regression model, which is

$$Y = a + bX.$$

There is another type of linear model called a simple linear regression through origin, which is

$$Y = \beta X,$$

where the intercept is 0 (fixed). The formula to compute β is

$$\beta = \frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2}.$$

(c-1) Try to compute the slope of a simple linear regression through origin based on our data set.
(c-2) Without calculation, could you know which model has a smaller sum of squared residuals, the simple linear regression model or the simple linear regression through origin model? Try to explain your conclusion.

Practical Question 4

Introduction to Inference: Characteristics of Same-Sex Couple Households
All United States Census Bureau demographic surveys collect information about same sex couples. The level of detail collected varies, as well as the availability of other characteristics of

the partners. In this assignment, we will take a look at Characteristics of Same-Sex Couple Households (2012), which can be downloaded from http://www.census.gov/hhes/samesex/ , Latest Releases, where you can find an Excel spreadsheet file. There are three spreadsheets in this document, name Tab 1, Tab 2, and Tab 3.

The table below shows the categorization method which US census bureau used.

Unmarried same-sex couples			
Unmarried male-male couples		Unmarried female-female couples	
Male-male couples reported as spouses	Unmarried male-male partners	Feale-female couples reported as spouses	Unmarried female-female partners

(a) Find the Total households (Number) and the Average Household Income (dollars) of Male-male couples reported as spouses, Unmarried male-male partners, Feale-female couples reported as spouses, and Unmarried female-female partners in Tab 2. Summarize these 8 numbers in a Table.

(b) (b-1) Based on the 8 numbers you find from Tab 2 in part (a), try to compute the average Household Income (dollars) of Unmarried male-male couples, Unmarried female-female couples, and Unmarried same-sex couples. (Actually you can find these three numbers form Tab 1, but you need to compute them based on the information in Tab 2.) (b-2) Based on the information in Tab 1, find the United States' average household incomes, including both opposite-sex couples and same-sex couples.

(c) (c-1) Find the average household income (dollars) and standard errors of Married opposite-sex couples and Unmarried same-sex couples. (You should get two pairs of numbers.) (c-2) Find the 95% confidence intervals for the true mean household income of Married opposite-sex couples and Unmarried same-sex couples. Do the two intervals overlap? Can you draw any conclusion from your computation? (c-3) We notice that the standard error of Married opposite-sex couples' average household income is less than the standard error of Unmarried same-sex couples' average household income. Can you explain it?