HW03 - Normal Curve and Summary Statistics

Stat 20 & 131A, Spring 2017, Prof. Sanchez

Due Feb-09

1) For the women age 18-24 in HANES2, the average height was about 64.3 inches; the SD was about 2.6 inches. Using the normal curve, estimate the percentage of women with the following heights: $_{0.6pts}$

a. below 66 inches.

b. between 60 inches and 66 inches.

c. above 72 inches.

2) Consider the following table 1. Fill in the blanks, using the options below. $_{0.6pts}$

Table 1. Selected percentiles for family income in the U.S. in 2004.

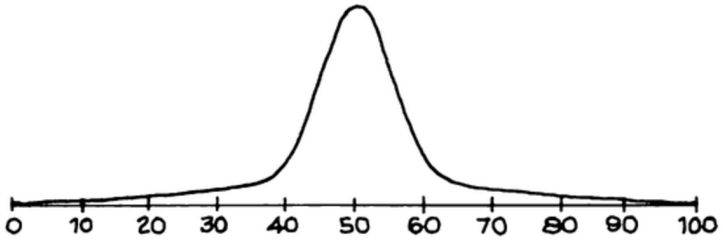
Percentile	Income
1	$0
10	$15,000
25	$29,000
50	$54,000
75	$90,000
90	$135,000
99	$430,000

a. The percentage of families with incomes below $90,000 was about _____.

b. About 25% of the families had incomes below _____.

c. The percentage of families with incomes between $15,000 and $125,000 was about _____.

3) A histogram is sketched below. $_{0.2pts}$



a. How is it different from the normal curve

b. Is the interquartile range around 15, 25, or 50?

4) For Berkeley freshmen, the average GPA (grade point average) is around 3.0; the SD is about 0.5. The histogram follows the normal curve. Estimate the 30th percentile of the GPA distribution. $_{0.2pts}$

5) Among applicants to one law school, the average LSAT score was about 169, the SD was about 9, and the highest score was 178. Did the LSAT scores follow the normal curve? $_{0.5pts}$

6) Among freshmen at a certain university, scores on the Math SAT followed the normal curve, with an average of 550 and an SD of 100. Fill in the blanks; explain briefly. $_{1pt}$

a. A student who scored 400 on the Math SAT was at the _____th percentile of the score distribution.

b. To be at the 75th percentile of the distribution, a student needed a score of about _____ points on the Math SAT.

7) True or False, and explain briefly. $_{1.2pts}$

a. If you add 7 to each entry on a list, that adds 7 to the average.

b. If you add 7 to each entry on a list, that adds 7 to the SD.

c. If you double each entry on a list, that doubles the average.

d. If you double each entry on a list, that doubles the SD.

e. If you change the sign of each entry on a list, that changes the sign of the average.

f. If you change the sign of each entry on a list, that changes the sign of the SD.

8) For women age 25-34 with full time jobs, the average income in 2004 was $32,000. The SD was $26,000, and 1/4 of 1% had incomes above $150,000. Was the percentage with incomes in the range from $32,000 to $150,000 about 40%, 50%, or 60%? Choose one option and explain briefly. $_{0.4pts}$

9) Consider the following options for a correlation coefficient: $_{0.4pts}$

- exactly -1
- close to -1
- close to 0
- close to 1
- exactly 1

a. If women always married men who were five years older, the correlation coefficient between the ages of husbands and wives would be _____. Choose one of the options above, and explain.

b. The correlation between the ages of husbands and wives in the U.S. is _____. Choose one of the options above, and explain.

10) True or False, and explain: if the correlation coefficient is 0.90, then 90% of the points are highly correlated. $_{0.2pts}$

The next questions involve working with R, and it is assumed that you have read the tutorial scripts `06-normal-curve.pdf`, `07-scatter-diagrams.pdf`, and `08-correlation.pdf`.

11) The heights of the NBA players averaged 78.96 inches; the SD was 3.48 inches. Assume that the height follows a symmetric bell-shaped distribution so that it is safe to use the normal approximation for data. Use the function `pnorm()`—showing R code—to find the percentage of players with heights: $_{1pt}$

a. below 75 inches.

b. above 80 inches.

c. between 77.5 and 82 inches.

d. below 73.5 inches or above 83.5 inches.

12) Use the function `qnorm()`—showing R code—to answer the following questions: $_{0.7pt}$

a. How tall a player should be to be at the 75th percentile of the distribution?

b. At least how tall a player should be to be in the top 10th percentile?

c. What are the height values for players to be between the 15th and 85th percentiles?

13) The following lines of code define three pairs of x, y values: $_{3pts}$

```
x1 = c(1, 1, 1, 1, 2, 2, 2, 3, 3, 4)
y1 = c(5, 3, 5, 7, 3, 3, 1, 1, 1, 1)

x2 = c(1, 1, 1, 1, 2, 2, 2, 3, 3, 4)
y2 = c(1, 2, 1, 3, 1, 4, 1, 2, 2, 3)

x3 = c(1, 1, 1, 1, 2, 2, 2, 3, 3, 4)
y3 = c(2, 2, 2, 2, 4, 4, 4, 6, 6, 8)
```

For each pair of variables, use R, showing your code, to:

1. plot a scatter diagram, and
2. find the correlation coefficient using the procedure described in the textbook (page 132):

   Convert each variable to standard units. The average of the products gives the correlation coefficients.

   r = average of (x in standard units) $\times$ (y in standard units)

Note: to convert to standard units, you need to obtain SD (not SD^+). This means that you have to adjust the value of `sd()` by the correction factor. You can use the function `cor()` only to verify that your computations are correct.