

Solutions to Review Descriptive Statistics

STAT 401F Fall 2007

1. Salaries of professional athletes receive a good deal of attention in the press. The 1990 salaries of the non-pitchers on the Chicago Cubs baseball team are listed below, units are in thousands of dollars (e.g., the value of 100 units below means 100,000 dollars).

100	100	111	114	165	210	225	225
230	575	1200	1900	2100	2100	2650	3300

(a) Calculate the median and the mean for these data. *sample mean/average $\bar{y} = 956.5625$ & sample median $M = (225 + 230)/2 = 227.5$*

(b) Suppose the owner of the team says that a typical player earns close to a million dollars (1000 units in thousands of dollars). Which measure of center is the owner using? Would it be accurate to say that a typical player earns close to a million dollars? Explain.

The owner would be using the sample mean/average $\bar{y} = 956.5625$, which makes sense in that the owner also would like to argue against paying higher salaries (i.e., looks good for the owner to claim “a pitcher earns close to a million dollars on average”). On the other hand, a typical pitcher does not really earn close to a million dollars since the median is 227.5 (in thousands of dollars); that means 50% of the pitchers earn less than 227,500 dollars. Note the sample mean is much higher than the sample median because the distribution of salaries (data) is skewed right. However, overall, the owner is paying out in salaries close to a million dollars $\bar{y} = 956.5625$ per pitcher.

(c) Calculate a five number summary for these data. *max = 100, $Q_1 = (114+165)/2 = 139.5$, $M = 227.5$, $Q_3 = (1900 + 2100)/2 = 2000$, max = 3300*

(d) If each player received an extra 200,000 dollars (or 200 units in thousands of dollars), how would the mean and median change? **(You don't have to change the data and recalculate the mean and median. Instead, think about what each measures and how that is affected by adding a constant.)**

Both \bar{y} and M would increase by 200 units, since the center/typical value in the data is increasing by 200

(e) If each player received an extra 200,000 dollars (or 200 units in thousands of dollars), how would the standard deviation change? **(You don't have to change the data and calculate the standard deviation. Instead, think about what it measures (i.e., spread) and how that is affected by adding a constant.)**

The spread among in the individual observations in the sample, or sample standard deviation s , would NOT change by adding 200 units to each observation. Consider two samples for illustration

Sample A: 1, 2, 3 Sample B (Sample A + 200): 201, 202, 203

In Sample A, the sample mean is $\bar{y}_A = 2$ and, in Sample B, the sample mean is $\bar{y}_B = 202 = 200 + \bar{y}_A$. But the SPREADS among observations in each sample are the same (the standard deviation $s = 1$ for both data sets).

2. The sample average age of 5 persons in a room is 30 years. A 36-year-old person walks into the room. What is now the average age of the persons in the room? Suppose the median age is 30 years and a 36-year-old person enters the room. Can you find the new median age from this information?

Let y_i denote the age of person i in the room. With five people in the room, the sample average/mean age is

$$\bar{y} = \frac{\sum_{i=1}^5 y_i}{n} = \frac{\sum_{i=1}^5 y_i}{5} = 30$$

This implies the sum $\sum_{i=1}^5 y_i$ (in our notation) of the ages of the five people in the room must be $\sum_{i=1}^5 y_i = 30 \times 5 = 150$. If a sixth person walks in the room, the sum of all six ages is

$$\sum_{i=1}^6 y_i = \sum_{i=1}^5 y_i + y_6 = \sum_{i=1}^5 y_i + 36 = 150 + 36 = 186$$

So the sample average of six weights is $\bar{y} = \frac{\sum_{i=1}^6 y_i}{6} = \frac{186}{6} = 31$.

If a sixth person of age 36 walks into the room, you can't figure out the median. This is because you need to order the new observation "age 36" among the previous five ages to find how the median (as the average of the middle two ordered numbers).

3. This is a sample variance context. You must give a list of two numbers chosen from the whole numbers 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, with repeats allowed.
- Give a list of two numbers with the largest sample variance such a list can possibly have.
Pick 0,9. Then the sample mean/average is $\bar{y} = 4.5$ (the "center" of the data) and both "0" and "9" are equally far off from the center $\bar{y} = 4.5$, which increases the sample standard deviation s . Intuitively, the spread in the data is maximized by this choice since "0" and "9" are as far apart as you can get.
 - Give a list of two numbers with the smallest sample variance such a list can possibly have. 0,0 is one possibility; so is 1,1 and so on.
 - Does either part (a) or (b) have more than one correct answer? Part (a) has one answer; part (b) has ten possibilities.

4. An experiment to study the lifetime (in hours) for a certain type of component involved putting ten components into operation and observing them for 100 hours. Eight of the ten components failed during that period, and those lifetimes were recorded. Denote the lifetimes of the two components still functioning after 100 hours by 100+. The resulting sample observations were:

48, 79, 100+, 35, 92, 86, 57, 100+, 17, 29

Which of the measures of location/center can be calculated, and what are the values of those measurements?

Ordering the observations gives

17 29 35 48 57 79 86 92 100+ 100+

In terms of the median, it doesn't matter what the two "100+" values actually are. The median of the data set is $M = (57 + 79)/2 = 68$. In terms of the sample mean \bar{y} , it DOES matter what the two "100+" values actually are. If the two "100+" values turn out to be 101 and 102 or turn out to be 100001 and 57983457547, you could get different sample means \bar{y} . You need to know all the values in the data to compute \bar{y} here.

5. The following frequency distribution of storm duration (in minutes) for 74 storms appeared in the article "Lightning Phenomenology in the Tampa Bay Area" (*J. of Geophysical Research* (1984): 789-805).

Storm duration in minutes, x	frequency	Storm duration in minutes, x	frequency
$0 \leq x < 25$	1	$150 \leq x < 175$	5
$25 \leq x < 50$	17	$175 \leq x < 200$	4
$50 \leq x < 75$	14	$200 \leq x < 225$	3
$75 \leq x < 100$	11	$225 \leq x < 250$	2
$100 \leq x < 125$	8	$250 \leq x < 275$	0
$125 \leq x < 150$	8	$275 \leq x < 300$	1

- (a) Sketch a histogram from the table above.
- (b) As a measure of center for the distribution, can you identify the interval in the above table that contains the median?

There are 74 observations (storm durations in minutes) and the median M would be the average of the middle two observations if you could list and order all the observations from smallest to largest. The middle two observations would be the 37th ($74/2=37$) and 38th in the ordered list; you need to find the interval that contains the 37th and 38th ordered observations. Note the interval $(0,25]$ contains one observation (the smallest); the interval $(0,50]$ contains $1+17=18$ observations; the interval $(0,75]$ contains $1+17+14=32$ observations; and the interval $(0,100]$ contains $1+17+14+11=44$ observations. That implies that in the interval $(75,100]$ fall the 33rd, 34th, ..., 44th ordered observations (ordered smallest to largest). So M is in the interval $(75,100]$.

- (c) From the information provided in the table, can you compute the mean? Explain. No, you would need the sum of all the observations, or the individual observations themselves, to compute the sample mean/average \bar{y} .