Putting Things Together Part 1

These exercise blend ideas from various graphs (histograms and boxplots), differing shapes of distributions, and values summarizing the data. Data for 1, 5, and 6 are in the instructor’s shared folder on LakerApps: Putting Things Together Part 1.

1. School districts in Kentucky bid on contracts for milk. The boxplots below are for winning bids from districts in the northern (254 bids) and southern (100) regions of the state.

![Boxplots of milk prices for northern and southern regions](image)

Answer the following using only the boxplots above.

a) For the Northern region:
   - State the 5-number summary
   - Determine the Range and Interquartile Range (IQR).
   - Approximately what percent of contracts had a winning bid above $0.140 per pint?
   - Guess the standard deviation. Do it two ways: First, using the Range. Second, using a “reduced Range” computed by ignoring outliers.

b) For the Southern region:
   - State the 5-number summary
   - Determine the Range and IQR.
   - Approximately what percent of contracts had a winning bid below $0.145 per pint?
   - Guess the standard deviation – both ways again.

c) Are the mean bids for the two regions closer together or further apart than the median bids? Explain. (Hint: a little skew is evident, and there are outliers.)

2. A specialty food company sells gourmet hams by mail order. The hams vary in size from 4.00 to 7.25 pounds, with a mean weight of 6.00 pounds and a standard deviation of 0.65 pounds. The quartiles and median are 5.50, 6.20 and 6.55 pounds.

a) Find the Range and IQR of the weights.

b) Is the distribution of the weights symmetric or skewed? If skewed, which way? Why?
3. John has a radar gun, and collects data on the speed of cars passing his house. The mean is 32.5 mph with standard deviation of 2.5 mph.

a) Make a rough guess at the percentage of cars that go between 30.0 and 35.0 mph.

b) Suppose that cars traveling outside this interval are equally likely to be going faster than 35.0 mph or slower than 30.0 mph. (This is symmetry.) What percentage of cars go faster than 35.0 mph? Slower than 30.0 mph? What can you now say about the percentile ranks for 30.0 and 35.0 mph?

c) Make a rough guess at the percentage of cars that go between 27.5 and 37.5 mph. Again assuming symmetry: What can you say about the percentile ranks for 27.5 and 32.5 mph?

4. A popular band on tour played a series of concerts in large venues. They always drew a large crowd, averaging 21,359 fans. While the band did not announce (and probably never calculated) the standard deviation, which of these values do you think is most likely to be correct: 20, 200, 2000, or 20000 fans? Explain your choice.

5. The histogram shows the budgets of 110 major movie releases in a recent year. Try answering the questions below just using the histogram. (You can check answers using the data.)

![Histogram of Movie Budgets]

a) What shape is this distribution?

b) How many of the movies had budget less than $40 million?

c) What percent of the movies had budget less than $40 million? (This is the percentile rank of $40 million.)

d) The median is approximately $_______ million.

e) The mode is approximately $_______ million.

f) Guess the standard deviation of budgets.

g) Which of the following is closest to the mean budget?

- $25 million
- $40 million
- $52 million
- $90 million
6. Surveying members of her church, a young woman obtains the following data on the number of marriages for each adult male (of whom there are 24):

   0 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 3 3

Obtain the 5 number summary for this data. Notice that the first, second and third quartiles are identical. Our working definition of a percentile is “k% of the data is lower and (100-k)% is higher.” In this situation it makes no sense to say all three of the following simultaneously:

- 25% of the data is below 1; 75% above 1
- 50% of the data is below 1; 50% above 1
- 75% of the data is below 1; 25% above 1

This data is far too discrete for percentiles and percentile ranks to be useful.

Here’s the most effective summary for this type of highly discrete data

<table>
<thead>
<tr>
<th># of marriages</th>
<th>% of males</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>21%</td>
</tr>
<tr>
<td>1</td>
<td>58%</td>
</tr>
<tr>
<td>2</td>
<td>13%</td>
</tr>
<tr>
<td>3</td>
<td>8%</td>
</tr>
</tbody>
</table>

7. Would it make good sense to summarize data with the five number summary for the following variables? (While you’re at it: Circle only the variable; underline only the units.)

a) The number of goals scored in a professional soccer game. (If you don’t know anything about scoring in soccer, try an internet search of “scores of soccer games.” You’ll quickly learn something.)

b) The mean number of goals per game scored by professional soccer teams over the course of a lengthy season.

c) Running times of movies (as displayed on a DVD player).

d) The number of living grandparents of currently enrolled college students.

e) Favorite brand of beer of 25-year-old males.
Solutions

1. | a) Northern | b) Southern |
---|---|---|
5-#-Summary: { 0.1041, 0.1224, 0.1280, 0.1400, 0.1800 } | { 0.1064, 0.1374, 0.1447, 0.1525, 0.1690 } |
Range | 0.0759 | 0.0626 |
IQR | 0.0176 | 0.0152 |
The above values don’t have to be exact, but you should be within 0.001 on each. | % above 0.140? Since 0.140 is Q3, 25% of the data are above 0.140. | % below 0.145? Since 0.145 is awful close to the median, 50% of the data are above 0.145. |
Guess SD | 0.0759 / 4 = 0.019 | 0.0626 / 4 = 0.0157 |
| (0.166 – 0.104) / 4 = 0.0155 | (0.168 – 0.119) / 4 = 0.0125 |
Actual SD = 0.01579* | Actual SD = 0.01329* |
This somewhat demonstrates that, at least for fairly large data sets, ignoring outliers and using a “usual” range yields a better guess of SD. | |
c) Means | The outliers will pull on the mean. For the Northern region the mean will be above the median; for the Southern, below. Consequently, the means are closer than the medians. | |
Mean = 0.13309* | Mean = 0.14306* |

*Here’s what Minitab gives¹:

Descriptive Statistics: $ per Pint

<table>
<thead>
<tr>
<th>Variable</th>
<th>Market</th>
<th>Minimum</th>
<th>Q1</th>
<th>Median</th>
<th>Q3</th>
<th>Maximum</th>
<th>IQR</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>$ per Pint</td>
<td>Northern</td>
<td>0.10410</td>
<td>0.12230</td>
<td>0.12800</td>
<td>0.14000</td>
<td>0.18000</td>
<td>0.01770</td>
<td>0.07590</td>
</tr>
<tr>
<td></td>
<td>Southern</td>
<td>0.10640</td>
<td>0.13705</td>
<td>0.14470</td>
<td>0.15250</td>
<td>0.16900</td>
<td>0.01545</td>
<td>0.06260</td>
</tr>
</tbody>
</table>

Difference in medians: 0.0167

<table>
<thead>
<tr>
<th>Variable</th>
<th>Market</th>
<th>Mean</th>
<th>StDev</th>
</tr>
</thead>
<tbody>
<tr>
<td>$ per Pint</td>
<td>Northern</td>
<td>0.13309</td>
<td>0.01579</td>
</tr>
<tr>
<td></td>
<td>Southern</td>
<td>0.14306</td>
<td>0.01329</td>
</tr>
</tbody>
</table>

Difference in means: 0.0098

¹ Minitab uses a slightly different method for percentiles. Consequently, it reports quartiles a bit different from what you can get in a spreadsheet. However: They are very close – which is always the case for large data sets.
2.

a) Range = 3.25 pounds; IQR = 1.05 pounds.

b) This would be a left skewed distribution. The mean is below the median, which hints at left skew or outliers to the left. The simple boxplot is shown. Clearly data falls much further to the left of center than to the right.

![Boxplot](image)

3. a) About 68% of cars go between 30.0 and 35.0 mph, so about 32% do not.

b) If they are split symmetrically, then 16% go less than 30 mph. So 30 mph is the 16th percentile. Similarly, 16% go faster than 35.0 mph, so 35.0 mph is the 84th percentile. These are only rough guesses – which is the best one can do without more information.

c) 95% of cars go between 27.5 and 37.5 mph. 27.5 mph has percentile rank of 2.5; 37.5 has percentile rank of 97.5. (Again, these are only rough guesses.)

4. 2000 is the best choice. This would put about 2 in 3 concerts (2/3 is quite close to 68%) with attendance between 19359 and 23359 and most (95%) between 17359 and 25359. A standard deviation of 20000 would imply about 68% of shows having between 359 and 41,359 as attendance – and the problem states that concerts always drew a large crowd (values about 360 are ruled out). A standard deviation of 20 makes the attendance within 60 of 21,359 for almost all concerts, which is hard to believe. A standard deviation of 200 is not the worst answer (partial credit), but if that’s the value then attendance never falls outside of 20,759 to 21,959, which is still a fairly narrow range – especially given that the band would be playing in venues (arenas, probably) of a variety of capacities.

5. a) This is a right skewed distribution.

b) 55 of the movies had budget less than $40 million. (5 + 11 + 19 + 20 = 55).

c) 55 of 110 is 50%. The percentile rank of $40 million is about 50.

d) So: $40 million is approximately 50th percentile. $40 million is the median. (In fact, the median is $40 million.)

e) The mode is about $35 million ($30 million is an acceptable answer).

f) The range is about $180 million, yielding a guess of $45 million for the standard deviation. A better guess applies the principle from #1 above: Ignoring the outliers, the usual range is about $135 million, which leads to a guess of about $34 million for the standard deviation. (In fact, the standard deviation is $36.60 million.)

g) The mean is $52.46 million. So $52 million is the best choice.

---

2 The data are available. Just compute it and check.
3 See the previous footnote.
4 See the previous footnote.
6. The 5-number summary is \{ 0, 1, 1, 1, 3 \}. Not very informative – except that you get the feel that 1 is most common.

<table>
<thead>
<tr>
<th>Units</th>
<th>Variable</th>
<th>Percentiles?</th>
</tr>
</thead>
<tbody>
<tr>
<td>a) Soccer games</td>
<td>Number of goals scored</td>
<td>No</td>
</tr>
<tr>
<td>b) Soccer teams</td>
<td>Mean number of goals scored</td>
<td>Yes</td>
</tr>
<tr>
<td>c) Movies</td>
<td>Running time</td>
<td>Yes</td>
</tr>
<tr>
<td>d) College students</td>
<td>Number of living grandparents</td>
<td>No</td>
</tr>
<tr>
<td>e) 25-year-old males</td>
<td>Favorite brand of beer</td>
<td>Ridiculous</td>
</tr>
</tbody>
</table>

e is not even quantitative. It makes no sense to talk about “how much data is below” (which is what percentiles are about) when the data consists of brands of beer.

a and d are very discrete. These situations are very similar to what’s going on in #6 above – there’d only be a handful of values, replicated quite often. So a 5 number summary would not be that effective.

The variables described in b and c are both fairly continuous – especially movie running times which are easily measured to the nearest 1 second, and on directors’ computers are displayed to the nearest 0.001 second. You wouldn’t expect many soccer teams, after a long season, to have exactly the same average number of goals. While there might be a handful of “ties,” they would be relatively uncommon, and so percentiles / quartiles would provide a reasonable summary of the data.