#### Week 2 Practice Quiz

**TOTAL POINTS 5**

1.

Question 1

Which of the below data sets has the **lowest** standard deviation? You do not need to calculate the exact standard deviations to answer this question.

**1 / 1 point**



0,1,3,3,3,5,6



100, 100, 100, 100, 100, 100, 101



0, 25, 50, 100, 125, 150, 1000



0,1,2,3,4,5,6

**Correct**

This question refers to the following learning objective(s):

Note that there are three commonly used measures of center and spread:

* center: mean (the arithmetic average), median (the midpoint), mode (the most frequent observation)
* spread: standard deviation (variability around the mean), range (max-min), interquartile range (middle 50% of the distribution)

The dataset with the most repeated observations has the least variability, hence the lowest standard deviation.

2.

Question 2

**True or False:** The statistic mean/median (mean divided by median) can be used as a measure of skewness (either right or left). Suppose we are dealing with a distribution where the minimum is 0.5. If this statistic (mean/median) is less than 1, the distribution is most likely left skewed.

**1 / 1 point**



False



True

**Correct**

This question refers to the following learning objective(s):

Identify the shape of a distribution as symmetric, right skewed, or left skewed, and unimodal, bimodoal, multimodal, or uniform.

In a left skewed distribution the median tends to be greater than the mean, therefore we would expect mean/median to be less than 1.

3.

Question 3

**True or False:** You are going to collect income data from a right-skewed distribution of incomes of politicians. If you take a large enough sample from that distribution, the sample mean and the sample median will always have the same value.

**1 / 1 point**



True



False

**Correct**

This question refers to the following learning objective(s):

Define a robust statistic (e.g. median, IQR) as a statistics that is not heavily affected by skewness and extreme outliers, and determine when such statistics are more appropriate measures of center and spread compared to other similar statistics.

Think about the unrealistic case where instead of collecting a sample of income data from politicians you instead collect income data from all politicians. In that case, your sample mean and sample median will be close to or equal to the population’s values, which are not equal to each other in a skewed distribution.

4.

Question 4

**True or False:** A mosaic plot is useful for visualizing the relationship between a numerical and a categorical variable.

**1 / 1 point**



True



False

**Correct**

This question refers to the following learning objective(s):

Use contingency tables and segmented bar plots or mosaic plots to assess the relationship between two categorical variables.

5.

Question 5

Does meditation cure insomnia? Researchers randomly divided 400 people into two equal- sized groups. One group meditated daily for 30 minutes, the other group attended a 2-hour information session on insomnia. At the beginning of the study, the average difference between the number of minutes slept between the two groups was about 0. After the study, the average difference was about 32 minutes, and the meditation group had a higher average number of minutes slept. To test whether an average difference of 32 minutes could be attributed to chance, a statistics student decided to conduct a randomization test. She wrote the number of minutes slept by each subject in the study on an index card. She shuffled the cards together very well, and then dealt them into two equal-sized groups. Which of the following best describes the outcome?

**1 / 1 point**



The average difference between the two stacks of cards will be about 0 minutes.



If meditation is effective, the average difference between the two stacks of cards will be more than 32 minutes.



The average difference between the two stacks of cards will be about 32 minutes.

**Correct**

This question refers to the following learning objective:

Note that an observed difference in sample statistics suggesting dependence between variables may be due to random chance, and that we need to use hypothesis testing to determine if this difference is too large to be attributed to random chance. Set up null and alternative hypotheses for testing for independence between variables, and evaluate the data support for these hypotheses using a simulation technique.

Since we’re randomly splitting the cards into two groups, we would expect similar averages in the two groups, yielding a difference of 0 in the averages.

#### Week 2 Practice Quiz

**TOTAL POINTS 5**

1.

Question 1

Which of the below data sets has the **lowest** standard deviation? You do not need to calculate the exact standard deviations to answer this question.

**1 / 1 point**



0,1,3,3,3,5,6



100, 100, 100, 100, 100, 100, 101



0, 25, 50, 100, 125, 150, 1000



0,1,2,3,4,5,6

**Correct**

This question refers to the following learning objective(s):

Note that there are three commonly used measures of center and spread:

* center: mean (the arithmetic average), median (the midpoint), mode (the most frequent observation)
* spread: standard deviation (variability around the mean), range (max-min), interquartile range (middle 50% of the distribution)

The dataset with the most repeated observations has the least variability, hence the lowest standard deviation.

2.

Question 2

**True or False:** The statistic mean/median (mean divided by median) can be used as a measure of skewness (either right or left). Suppose we are dealing with a distribution where the minimum is 0.5. If this statistic (mean/median) is less than 1, the distribution is most likely left skewed.

**1 / 1 point**



False



True

**Correct**

This question refers to the following learning objective(s):

Identify the shape of a distribution as symmetric, right skewed, or left skewed, and unimodal, bimodoal, multimodal, or uniform.

In a left skewed distribution the median tends to be greater than the mean, therefore we would expect mean/median to be less than 1.

3.

Question 3

**True or False:** You are going to collect income data from a right-skewed distribution of incomes of politicians. If you take a large enough sample from that distribution, the sample mean and the sample median will always have the same value.

**1 / 1 point**



True



False

**Correct**

This question refers to the following learning objective(s):

Define a robust statistic (e.g. median, IQR) as a statistics that is not heavily affected by skewness and extreme outliers, and determine when such statistics are more appropriate measures of center and spread compared to other similar statistics.

Think about the unrealistic case where instead of collecting a sample of income data from politicians you instead collect income data from all politicians. In that case, your sample mean and sample median will be close to or equal to the population’s values, which are not equal to each other in a skewed distribution.

4.

Question 4

**True or False:** A mosaic plot is useful for visualizing the relationship between a numerical and a categorical variable.

**1 / 1 point**



True



False

**Correct**

This question refers to the following learning objective(s):

Use contingency tables and segmented bar plots or mosaic plots to assess the relationship between two categorical variables.

5.

Question 5

Does meditation cure insomnia? Researchers randomly divided 400 people into two equal- sized groups. One group meditated daily for 30 minutes, the other group attended a 2-hour information session on insomnia. At the beginning of the study, the average difference between the number of minutes slept between the two groups was about 0. After the study, the average difference was about 32 minutes, and the meditation group had a higher average number of minutes slept. To test whether an average difference of 32 minutes could be attributed to chance, a statistics student decided to conduct a randomization test. She wrote the number of minutes slept by each subject in the study on an index card. She shuffled the cards together very well, and then dealt them into two equal-sized groups. Which of the following best describes the outcome?

**1 / 1 point**



The average difference between the two stacks of cards will be about 0 minutes.



If meditation is effective, the average difference between the two stacks of cards will be more than 32 minutes.



The average difference between the two stacks of cards will be about 32 minutes.

**Correct**

This question refers to the following learning objective:

Note that an observed difference in sample statistics suggesting dependence between variables may be due to random chance, and that we need to use hypothesis testing to determine if this difference is too large to be attributed to random chance. Set up null and alternative hypotheses for testing for independence between variables, and evaluate the data support for these hypotheses using a simulation technique.

Since we’re randomly splitting the cards into two groups, we would expect similar averages in the two groups, yielding a difference of 0 in the averages.

#### Week 2 Quiz

**LATEST SUBMISSION GRADE**

1.

Question 1

Which of the below data sets has the **highest** standard deviation? You do not need to calculate the exact standard deviations to answer this question.

**1 / 1 point**



0,1,1,1,1,1,2



0,1,2,3,4,5,6



0, 100, 200, 300, 400, 500, 600



0, 25, 25, 25, 25, 25, 25

**Correct**

This question refers to the following learning objective(s):

Note that there are three commonly used measures of center and spread:

* center: mean (the arithmetic average), median (the midpoint), mode (the most frequent observation)
* spread: standard deviation (variability around the mean), range (max-min), interquartile range (middle 50% of the distribution)

The dataset with the least repeated observations that are farthest from the center has the most variability, hence the highest standard deviation.

2.

Question 2

The distribution of exam scores (ranging from 0 - 100%) where the mean score is 75%, the standard deviation is 12%, and the median is 78% is most likely

**1 / 1 point**



right skewed



symmetric



uniform



left skewed

**Correct**

This question refers to the following learning objective(s):

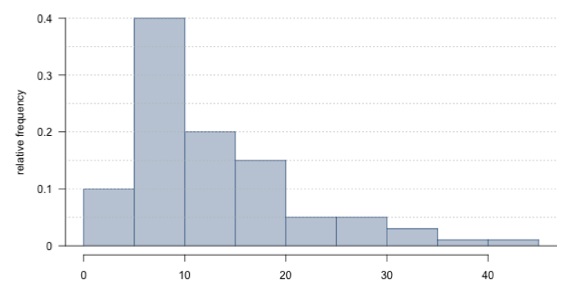
Identify the shape of a distribution as symmetric, right skewed, or left skewed, and unimodal, bimodoal, multimodal, or uniform.

The mean is lower than the median, and the mean is closer to the upper bound of the distribution with a large standard deviation, yielding a left skew.

3.

Question 3

Based on the relative frequency histogram below, which of the following statements is supported by the plot?



**1 / 1 point**



The mean of the distribution is smaller than its median.



It is not possible to estimate the median without knowing the sample size.



The distribution is multimodal.



The IQR of the distribution is roughly 10.



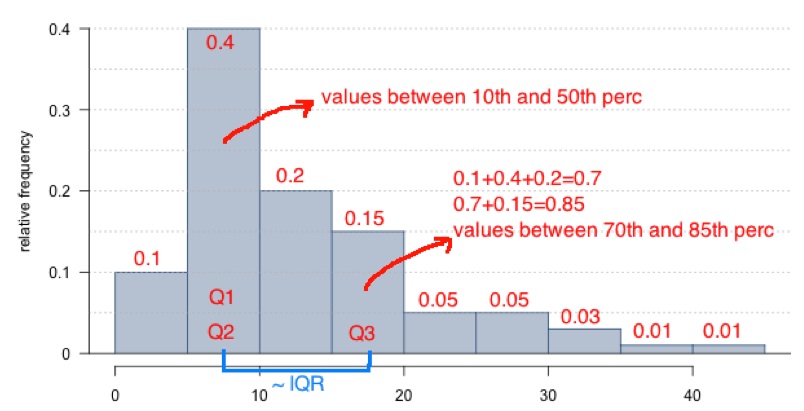
There are no outliers in the distribution.

**Correct**

This question refers to the following learning objective(s):

Use histograms and box plots to visualize the shape, center, and spread of numerical distributions, and intensity maps for visualizing the spatial distribution of the data.

Using the relative frequency histogram, we can tell that 10% of observations are below 5 (in the first bin), 40% are between 5 and 10, 20% are between 10 and 15, and 15% between 15 and 20. Therefore Q1 is in the second bin (between 5 and 10) and Q3 is in the fourth bin (between 15 and 20). This confirms that the IQR is roughly 10.



4.

Question 4

Which is more affected by extreme observations, the mean or median? And how about the standard deviation or IQR?

**1 / 1 point**



mean, IQR



mean, SD



median, IQR



median, SD

**Correct**

This question refers to the following learning objective(s):

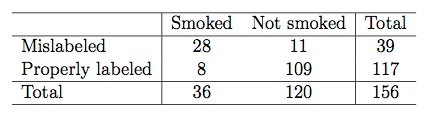
Define a robust statistic (e.g. median, IQR) as a statistics that is not heavily affected by skewness and extreme outliers, and determine when such statistics are more appropriate measures of center and spread compared to other similar statistics.

For a classic example illustrating this idea, suppose there are a few people (30 for example) in a bar and they all earn fairly normal incomes. The mean, SD, median, and IQR of their salaries are at fairly normal levels. Now suppose a very wealthy businessman who earned $700 million this year walks into the bar. Suddenly, the mean and SD of the incomes of people in the bar go up by a lot due to the wealthy businessman being an outlier and due to mean and SD not being robust to outliers. However, the median income and IQR would still be very close to what they were before the businessman walked in.

5.

Question 5

It is relatively common for fish to be mislabeled in supermarkets and even in restaurants. The table below shows the results of a study where a random sample of 156 fish for sale were collected and genetically tested. The researchers classified each sample as being labeled properly or being mislabeled. What fraction of smoked fish in the sample were mislabeled? Choose the closest answer.



**1 / 1 point**



9%



78%



18%



28%



72%

**Correct**

This question refers to the following learning objective(s):

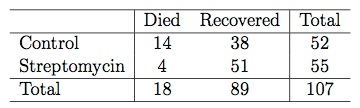
Use contingency tables and segmented bar plots or mosaic plots to assess the relationship between two categorical variables.

Of the 36 smoked fish, 28 are mislabeled: 28/36 ≈ 78%.

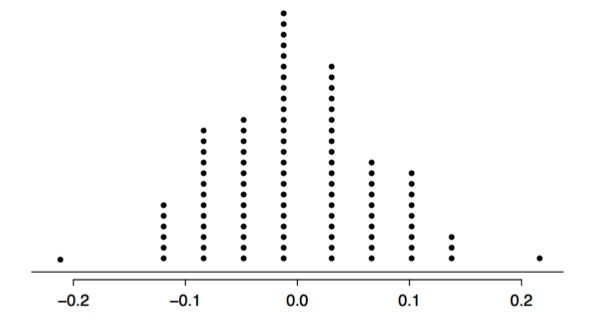
6.

Question 6

In 1948, Austin Bradford Hill, designed a study to test a new treatment for tuberculosis that at the beginning of the study there was no evidence whether it would be any better or worse than bed rest. He randomly assigned some patients who volunteered to be a part of this study to receive the treatment Streptomycin, an antibiotic. The other patients received only bed rest as the control group. Hill then observed the patients’ outcomes: which patients died and which recovered. The results of the study are shown below.



We use the following simulation test if there is a difference between the recovery rates under the two treatments: We write “died” on 18 index cards and “survived” on 89 index cards to indicate whether or not a patient died. Next, we shuffle the cards and deal them into two groups of 52 and 55, for control and treatment, respectively. We then calculate the simulated difference between the recovery rates in Streptomycin and control groups (p̂Streptomycin − p̂Control), and record this value. We repeat this simulation 100 times. The histogram below shows the distribution simulated difference between the recovery rates in these 100 simulations.



Which of the following is correct? Choose all that apply (there are multiple correct answers).

**0.778 / 1 point**



Streptomycin treatment appears to be effective in treating tuberculosis since the observed difference in recovery rates would be considered unusual based on the simulation results.



Based on this study we can conclude a causal relationship between Streptomycin and better tuberculosis recovery rate.

**Correct**

This question refers to the following learning objective:

Note that an observed difference in sample statistics suggesting dependence between variables may be due to random chance, and that we need to use hypothesis testing to determine if this difference is too large to be attributed to random chance. Set up null and alternative hypotheses for testing for independence between variables, and evaluate the data support for these hypotheses using a simulation technique.

Also, since this is an experiment we can deduce causation.



The alternative hypothesis is that the Streptomycin treatment is more effective than bed rest.



Hill’s study is observational.



The difference between the survival rates in the control and treatment groups appear to be simply due to chance.



The alternative hypothesis should be that there is a difference between the recovery rates under the two treatments.

**Correct**

This question refers to the following learning objective:

Note that an observed difference in sample statistics suggesting dependence between variables may be due to random chance, and that we need to use hypothesis testing to determine if this difference is too large to be attributed to random chance. Set up null and alternative hypotheses for testing for independence between variables, and evaluate the data support for these hypotheses using a simulation technique.

The evidence could go either way so we should consider any difference between the two treatments.



Streptomycin treatment does not appear to be effective in treating tuberculosis since the observed number of deaths in the treatment group would not be considered unusual based on the simulation results.



If Streptomycin and bed rest are equally effective in curing tuberculosis, the probability of observing a difference in the recovery rates at least as high as the one observed is 2%.



The conclusion of this study is generalizable to all tuberculosis patients.

Lesson Learning Objectives

**Suggested reading:** [OpenIntro Statistics, 3rd edition](https://www.openintro.org/stat/textbook.php?stat_book=os), Chapter 1, Section 1.6

**LO 1.** Use scatterplots for describing the relationship between two numerical variables making sure to note the direction (positive or negative), form (linear or non-linear) and the strength of the relationship as well as any unusual observations that stand out.

**LO 2.**When describing the distribution of a numerical variable, mention its shape, center, and spread, as well as any unusual observations.

**LO 3.** Note that there are three commonly used measures of center and spread:

* center: mean (the arithmetic average), median (the midpoint), mode (the most frequent observation).
* spread: standard deviation (variability around the mean), range (max-min), interquartile range (middle 50% of the distribution).

**LO 4.** Identify the shape of a distribution as symmetric, right skewed, or left skewed, and unimodal, bimodal, multimodal, or uniform.

**LO 5.**Use histograms and box plots to visualize the shape, center, and spread of numerical distributions, and intensity maps for visualizing the spatial distribution of the data.

**LO 6.** Define a robust statistic (e.g. median, IQR) as a statistic that is not heavily affected by skewness and extreme outliers, and determine when such statistics are more appropriate measures of center and spread compared to other similar statistics.

**LO 7.**Recognize when transformations (e.g. log) can make the distribution of data more symmetric, and hence easier to model.

***Test yourself:***

1. *Describe what is meant by robust statistics and when they are used.*
2. *Describe when and why we might want to apply a log transformation to a variable.*

Lesson Learning Objectives

**Suggested reading:** [OpenIntro Statistics, 3rd edition](https://www.openintro.org/stat/textbook.php?stat_book=os" \t "_blank), Chapter 1, Sections 1.7 - 1.8

**LO 1.** Use frequency tables and bar plots to describe the distribution of one categorical variable.

**LO 2.** Use contingency tables and segmented bar plots or mosaic plots to assess the relationship between two categorical variables.

**LO 3.**Use side-by-side box plots for assessing the relationship between a numerical and a categorical variable.

***Test yourself:***

1. *Interpret the plot in Figure 1.30 of OpenIntro Statistics (page 39).*
2. *You collect data on 100 classmates, 70 females and 30 males. 10% of the class are smokers, and smoking is independent of gender. Calculate how many males and females would be expected to be smokers. Sketch a mosaic plot of this scenario.*

## Suggested Readings and Practice

**Suggested reading for this week:** [OpenIntro Statistics, 3rd edition](https://www.openintro.org/stat/textbook.php?stat_book=os" \t "_blank), Chapter 1, Sections 1.6, 1.7, 1.8

**Practice exercises:** End of chapter exercises Chapter 1: 1.39, 1.41, 1.45, 1.49, 1.51, 1.55, 1.59, 1.63, 1.65, 1.67, 1.69

(Reminder: the solutions to the end of chapter exercises are at the end of the OpenIntro Statistics book)