

AP Statistics
Review for Test 6: **Data Analysis**

The 6-weeks test will be on Tuesday, November 11.

Step 1: Make sure you can answer all of the following questions

Step 2: Make sure you know how to work all the problems in Reviews 1 – 5.

1. What is a sample?

It is a part of the whole population. Hopefully it's selected randomly.

2. What is the difference between a parameter and a statistic?

Parameters describe a population, and statistics describe a sample. Statistics are used to estimate parameters. The mean of an entire population is a parameter. The mean of a sample is a statistic.

For example, when we want to estimate the percentage of all voters that voted for Obama in the presidential election, we take a sample and we look at the percentage of that sample that voted for Obama. The percentage of the sample is a statistic, which is an estimate of the percentage of the population.

3. What is the difference between categorical and quantitative data?

Categorical data is data that is separated into categories. Quantitative data is measured by numbers and some sort of units (feet, degrees, liters, etc.).

4. What are three ways of displaying categorical data? What are their strengths and weaknesses?

Pie charts are good for showing proportions. They are not particularly good for showing actual counts.

Bar charts are better at showing actual counts, but not as good as pie charts at showing proportions. Bar charts are also good for comparing counts. It's easy to tell if a bar is taller than another bar, whereas it's not always quite as easy to tell if a pie slice is bigger or smaller than another pie slice.

Segmented bar charts (aka ribbon charts) serve a similar purpose as pie charts: they show proportions.

For all of the above ways of illustrating categorical data, the **area principle** must be obeyed: this says that the area that represents a certain category must correspond to the actual percentage of that category. This is the main reason that 3-D charts are discouraged—they violate this principle.

Categorical data can also be displayed in a table, whether it is a simple one-variable frequency table, or a two variable contingency table. Page 24 in your book shows a good example of a contingency table with two variables: Class and Survival.

5. What is the difference between a marginal distribution and a conditional distribution?

This is where contingency tables come in handy. Look at the table on p. 24. The marginal distribution is the counts or percentages at the margins. Marginal distributions just measure one variable—they don't care about the other variable. The marginal distribution of survival in the table on p. 24 is 711 (32.3%) alive and 1490 (67.7%) dead. The marginal distribution of class is 325 (14.8%) first class, 285 (12.9%) second class, 706 (32.1%) third class and 885 (40.2%) crew. The conditional distribution of survival for those in third class is 178 (25.2%) alive and 528 (74.8%) dead.

6. What does it mean for categorical variables to be independent?

It means that the conditional and the marginal distributions will be virtually identical when viewed as a percentage. Using the example from p. 24, since the conditional distribution of survival for those in third class is 25.2% alive and 74.8% dead, and the marginal distribution of survival is 32.3% alive and 67.7% dead, the variables of Class and Survival are NOT independent. If the variables were independent, the percentage of survivors would not change from class to class. One can see from the table that the percentage of survivors in first class was much higher than those in other classes.

7. What are four ways to illustrate univariate (one variable) quantitative data? What are strengths and weaknesses of each?

Histogram: good for showing shape of data; good for larger data sets; not that great for small data sets. Not good for seeing individual data values.

Stemplot: good for illustrating data without a calculator or computer; good for seeing the shaped of data; not good for large data sets (too time consuming). Good for seeing individual data points.

Dotplot: good for illustrating data without a calculator or computer; good first step on the way to making a histogram.

Boxplot: good for seeing exact values of 5 pt summary; good for identification of outliers; good for comparing data sets. Not as good as histogram for seeing shape.

8. When describing univariate quantitative data, what four things should you look for?

Center (mean or median)

Unusual features (multiple modes, outliers)

Shape (symmetry, skew)

Spread (standard deviation, variance, IQR, range)

9. Compare and contrast the two ways of measuring the center of a set of data.

The mean is used for statistical inference, so we will concentrate on the mean much more than the median for the rest of the year. That being said, the median is the more appropriate measure (along with IQR) when the data is skewed either left or right.

The median is also better if what you're looking for is a "typical" value. When we want to focus on the typical price of a car, or the typical salary, or the height of a typical student, it makes more sense to use the median.

10. What is the five-number summary?

Min, Q1, median, Q3, max.

11. What are four ways of measuring spread?

Range = max - min.

IQR = range of the middle 50% of data = Q3 - Q1.

Variance = Average of squared distances from each data point to the mean =

$$\sigma^2 = \frac{1}{n} \sum_{j=1}^n (x_j - \mu)^2 \text{ if you're measuring the whole population, or}$$

$$s^2 = \frac{1}{n-1} \sum_{j=1}^n (x_j - \bar{x})^2 \text{ if you're using just a sample of the population.}$$

Standard deviation = square root of the variance.

12. How does one determine if a data point is an outlier or not?

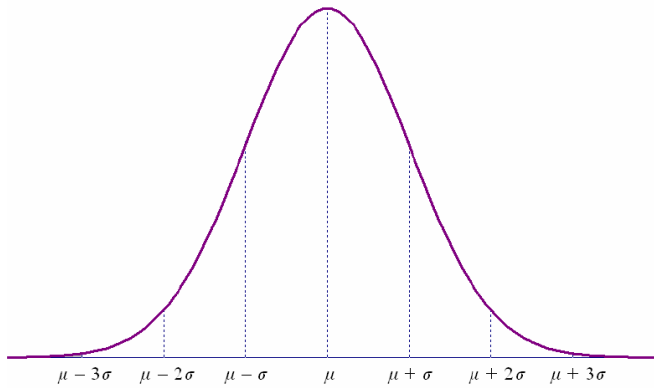
Anything greater than $Q3 + 1.5(IQR)$ or less than $Q1 - 1.5(IQR)$ is considered an outlier.

13. What does “resistant to outliers” mean? Which statistics are resistant to outliers and which aren’t?

Take a maximum or minimum value of a data set and change its value. The summary statistics that do not change are “resistant” to outliers (e.g. median, IQR, Q1, Q3). The summary statistics that change are “sensitive” to outliers (e.g. mean, standard deviation, variance, min, max, range).

14. What is the Normal model? What is the Standard Normal model?

The normal model looks like this:



A Standard Normal model has a mean of $\mu = 0$ and a standard deviation of $\sigma = 1$.

15. What is a z-score? Why do we care?

A z-score is like a percentile, in that it tells you where an individual data point lies in relation to the rest of the data. A data point that is one standard deviation greater than the mean has a z-score of 1. A data point that is two standard deviations less than the mean has a z-score of -2.

16. What is the empirical rule?

In a Normal Model, about 68% of the data lie within one standard deviation of the mean (z-scores between -1 and 1). About 95% of the data lie within 2 standard deviations of the mean (z-scores between -2 and 2). And about 99.7% of the data (virtually all of it) lie within 3 standard deviations of the mean (z-scores between -3 and 3). It is therefore VERY unusual to have a z-score that is more than 3 or less than -3.

17. How does one decide if data is “approximately normally distributed?”

A dataset whose histogram looks somewhat like the normal model is approximately normally distributed. In particular, the data needs to have one mode, and it needs to be fairly symmetrical.

18. What is the most popular way to illustrate bivariate (two variable) data?

A scatterplot.

19. When describing bivariate data, what four things should you look for?

Direction (positive or negative association).

Unusual things (outliers, influential points).

Form (linear, something else).

Strength (how well the points adhere to the model).

20. What does “association does not imply causation” mean anyway?

If two variables relate to one another in some way, it doesn't necessarily mean that one of them is causing the other one.

21. What is the Least Squares Regression Line? What point does the LSR line go through?

It is the unique line that minimizes the sum of the squared residuals. It is considered to be the line that "fits" a set of data the best. It always goes through the point (\bar{x}, \bar{y}) .

22. What is association? What is correlation?

If two variables are associated, it just means that there is some relationship between them. A correlation is a linear association.

23. What is the correlation coefficient r ?

The correlation coefficient r is one measure of the strength of the linear relationship. If $r = 1$ or $r = -1$, then the data points all lie on a line, and the response variable y can be predicted exactly from the explanatory variable x . If $r = 0$, then the explanatory variable is no use at all in predicting the response variable.

It is also the slope of the Least Squares Regression line of the z-scores. It can also be used to calculate the slope of the least squares line using the formula $b_1 = r \frac{s_y}{s_x}$.

That formula tells us that if the explanatory variable is k standard deviations above the mean, then we would predict the response variable to be kr standard deviations above the mean.

24. What is the coefficient of determination r^2 ?

It's the square of the correlation coefficient. It's also the percentage of the variance of the response variable that can be predicted from the linear model.

25. What are residuals? Why are they important?

Residuals are the difference between the actual data and the predicted data (using a least squares line). In particular, it is $y - \hat{y}$. If you predict that a value will be 4, but the actual value is 4.2, then your residual is 0.2.

The sum of all residuals of a data set is 0. Therefore, the mean of all residuals is also 0.

When we look at a scatterplot of residuals, we hope to see no pattern at all. The point of looking at residuals is that sometimes we can see a pattern in the residuals that we didn't see in the original scatterplot. The pattern might make us believe that our linear model is not as appropriate as we had hoped.