AP Statistics Notes

Stasya

Fall 2024 & Spring 2025

The purpose of this course is to introduce students to the major concepts and tools for collecting, analyzing, and drawing conclusions from data. Students are exposed to four broad conceptual themes: Exploring Data: Observing patterns and departures from patterns; Planning a Study: Deciding what and how to measure; Anticipating Patterns: Producing models using probability and simulation; and Statistical Inference: Confirming models.

Contents

1 Exploring One-Variable Data **2**
- 1.1 Representing Categorical and Quantitative Variables with Graphs . . . 2
- 1.2 Representing Quantitative Variables with Graphs . . . 3
- 1.3 Describing Distributions of Quantitative Variables . . . 3
- 1.4 Comparing Distributions of Quantitative Variables . . . 4
- 1.5 Z-Scores and the Empirical Rule . . . 4
- 1.6 The Standard Normal Curve . . . 5

2 Exploring Two-Variable Data **6**
- 2.1 Two Categorical Variables . . . 6
- 2.2 Scatterplots and Correlation . . . 6
- 2.3 Linear Regression . . . 7
- 2.4 Influential Points and Departure from Linearity . . . 7

3 Collecting Data **8**
- 3.1 Planning a Study . . . 8
- 3.2 Selecting a Random Sample . . . 9
- 3.3 Experimental Design . . . 9

4 Probability, Random Variables, and Probability Distributions **10**

5 Sampling Distributions **11**

6 Inference for Categorical Data: Proportions **12**

7 Inference for Quantitative Data: Means **13**

8 Inference for Categorical Data: Chi-Square **14**

9 Inference for Quantitative Data: Slopes **15**

1 Exploring One-Variable Data

1.1 Representing Categorical and Quantitative Variables with Graphs

Data contains information about a group of individuals. The information is organized using variables.

Individuals are objects described by a set of data. Individuals may be people but may be animals or inanimate objects.

Variables are characteristics of individuals. A variable may take on different values of different variables. Variables can be split into two types: categorical or quantitative.

Categorical variables place individuals into specific groups.

Quantitative variables takes on numerical values for which it makes sense to do arithmetic operations like adding and averaging. Quantitative variables fall into two categories: discrete and continuous.

Be careful - just because it is a number doesn't make it quantitative.

Discrete variables are numerical values where counting makes sense; in other words, decimals would not be an appropriate way to record the data.

Continuous variables are numerical values where decimals are appropriate; it usually involves some form of measuring.

The difference between discrete and continuous isn't always clear. An example of this would be age.

One of the easiest ways to display categorical data is with a table.

Count is the amount of that category in a table and relative count is

$$\frac{\text{count}}{\text{total}}$$

If you wanted to display two categorical variables at a time, we could make a two way table.

To better visualize the data, there are graphs that we can make from the data. We want to visualize the graphs to get a better idea of the distribution.

Distribution of a variable tells us what values the variable takes and how often it takes these values.

Bar Graphs have the following characteristics:

- Label each axis clearly
- The x-axis will contain the categorical variable and the y-axis will display teh counts
- Each category has its own bar and the bars cannot touch
- Order is not important when creating the x-axis

To make a histogram, we need to put the data into even intervals that capture our data. We will do this first by hand by counting how many data scores are in each bin.

To find the interval width, we can use the formula

$$\frac{\text{max-min}}{\text{\# of wanted intervals}}$$

To make the histogram:

- Draw rectanges for each interval with height representing the count
- Bars must touch
- Label the x-axis with the lower bound values of each interval

1.2 Representing Quantitative Variables with Graphs

Stemplots (or stem and leaf plots) are an alternate way to illustrate data using a semi-graph. It is similar to a histogram, but the data isn't lost. If the data has two digits, the stem is the first digit and the leaf is the second. If the data has 3 digits, the stem is the first two digits. You must always add a key to the graph.

Back-to-Back Stemplots are created when you can separate the data into two categories. The stems are the same, but the data can be split into different categories. You still must have a key for both sides.

Split Stemplots are the last type of stem and leaf plots. You can split the stem; in a similar way to creating more bins on a histogram if the bin width resulted in a skyscraper.

Dot plots are a very simple type of graph that involves plotting data values with dots above the values on a number line.

To construct a dot plot:

- Label your axis and title your graph. Draw a horizontal line and label it with the variable.
- Scale the axis based on the values of the variable.
- Mark a dot above the number on the horizontal axis corresponding to each data value.

Cumulative relative frequency graphs display percentiles. A percentile will tell you what percent of data falls above a value.

In order to create one of these graphs, you must make a table of the cumulative relative frequencies in order to graph it. This can be done by first finding the relative frequencies and then you add them together to get the cumulative relative frequency. This graph is also called an ogive.

1.3 Describing Distributions of Quantitative Variables

We can describe the distribution of a quantitative variable by its shape, outliers, center, and spread.

After graphing the distribution, the first thing we identify is the shape. A unimodal shape has one peak, a bimodal shape has two peaks, a uniform shape has no peak. If a graph is skewed it will be skewed in the direction in which the tail is located.

There are three measures of center: mean, median, and mode. Mean is defined as

$$\bar{x} = \frac{\sum x_i}{n}$$

, where $\bar{x}$ is the mean of sample, x_i is each individual observation, and n is the number of observations. Mean is called non-resistant because the mean is strongly influenced by outliers.

Median is a different measure of center that is resistant. The median is found by ordering the data and finding the middle value in the list. The location of the median is

$$\frac{n+1}{2}$$

Mode is the most occurring value in a data set.

To summarize, a unimodal symmetric graph will have the median, mean, and mode similar to each other. A unimodal, left-skewed graph will have mean $<$ median $<$ mode, and a unimodal, right-skewed graph will have mode $<$ median $<$ mean.

There are three measures of spread: range, standard deviation, and IQR.

Range is the difference between the maximum number and the minimum number.

The standard deviation is the average deviation of an observation from the mean of the data set. It is calculated by

$$s_x = \sqrt{\frac{\sum (x_i - \bar{x})^2}{(n-1)}}$$

, where s_x is the standard deviation, and s_x^2 is the variance of sample.

The variance is the average squared deviation. The quantitative variable typically varied by the mean by "standard deviation units".

Standard deviation has a few properties:

- The standard deviation is always positive.
- The standard deviation is always 0 when all observations are equal.
- The standard deviation has the same units of measure as the original variable measured.
- The standard deviation is non-resistant.
- The greater the standard deviation, the greater the distribution.

IQR is inter-quartile range and it uses percentiles to describe the spread of distribution. The 0th percentile is the minimum, the 25th percentile is Quartile 1, the 50th percentile is the median, the 75th percentile is Quartile 3, and the 100th percentile is the maximum.

An outlier is an individual piece of data that falls outside the overall pattern of the distribution. We can determine if a point is an outlier:

- First, find the five-number summary (the percentiles mentioned above)
- Find the IQR by subtracting the value of Quartile 1 from the value of Quartile 3
- Compute Quartile 1 - (1.5 * IQR). Any data above this number is an outlier.
- Compute Quartile 3 - (1.5 * IQR). Any data above that number is an outlier.

Using this data, we can make box plots or modified box plots depending on if the data set as an outlier determined.

1.4 Comparing Distributions of Quantitative Variables

In the world of statistics, it is not enough to just report the graph or just report the summary statistics.

Given a set of data, first you must create a graph for the data. If the data is categorical, use a bar graph. If the data is quantitative, if it's discrete, use a dot plot, stem plot, or boxplot. If it is continuous, use a histogram or box plot.

You want to summarize your findings after this. First, describe the shape. Next, describe if there are any outliers. Then, use the mean or median for the center of data. Lastly, describe the spread using range or IQR if you described the center with the median, or standard deviation if you described the center with mean.

1.5 Z-Scores and the Empirical Rule

Normal distributions are appropriate for many distributions whoes shapes are unimodal and approximately symmetric.

In the normal distribution, Mean $= \mu$, Standard deviation $= \sigma$, and this is written as

$$N(\mu, \sigma)$$

A normal distribution curve has these properties:

- Symmetric around the mean
- mean = median = mode
- 50% of observations are greater than the mean and 50% are less than the mean
- The standard deviation tells us how measurements for a group of observations are spread out from the mean
- In a normal distribution, approximately all the data lies 3 standard deviations above and below the mean

In a normal distribution, 68% of the data is likely to be found within 1 standard deivation from the mean, 95% found 2 standard deviations away, and 99.7% 3 standard deviations away.

To calculate how many standard deviations away from the mean an observation is, we use a z-score.

$$z = \frac{x - \mu}{\sigma}$$

Observations larger than the mean have positive z-scores, and observations smaller than the mean have negative z-scores.

1.6 The Standard Normal Curve

Using a calculator, we can determine what percentage of data falls above, below, or between specific z-scores.

Using the calculator commands: 2ND $\rightarrow$ Vars $\rightarrow$ 2:normalcdf(), we can find the percentage of data in between two values.

We can also find the z-score that corresponds to a percentile.

Using the calculator commands: 2ND $\rightarrow$ Vars $\rightarrow$ 3:invNorm(), we can find this.

2 Exploring Two-Variable Data

2.1 Two Categorical Variables

A side by side bar graph merges two bar graphs into one, in an attempt to compare the distributions of the two categorical variables.

A segmented bar graph is another way to display data, where each group is split by its relative frequency.

A mosaic plot is similar to a segmented bar graph, but it draws attention to the sizes of each group.

Joint relative frequencies are the ratio of the frequency in a cell and the total number of data values.

Marginal relative frequencies is the ratio of the sum in a row or column and the total number of data values.

Conditional relative frequencies are the ratio of a joint relative frequency and related marginal relative frequency.

Basically - joint relative frequency is the cell count divided by the table total, marginal is the row/column total divided by the table total and the conditional relative frequency is the intersection divided by the row/column total.

2.2 Scatterplots and Correlation

We have worked with bar graphs, box plots, and histograms. This is called variate data.

When we compare two variables or bivariate data, we are exploring the relationship between the two.

You should start with a scatter plot. A scatterplot shows the representation between two quantitative variables measured on the same individuals. The values of one variable appear on the horizontal axis and the values of the other variable appear on the vertical axis. Each individual in the data appears as a point in the plot fixed by the values of both variables for that individual.

A response variable measures the outcome of a study or an observation.

An explanatory variable helps explain or influences change in a response variable.

Commonly, explanatory variables are called independent and respondent are called dependent.

You can describe the overall pattern of a scatterplot by the direction, form, and strength of the relationship. An important kind of deviation is an outlier.

Form is the overall pattern or deviations from the pattern. Direction is whether the graph has a positive or negative slope. Strenght is how close the points lie to a simple form (such as a line).

Here are the steps to creating a scatterplot on a calculator:

- Load the x-values into list 1 and y-values into list 2.
- Using StatPlot - highlight the mini-scatterplot: XList:L1 and YList:L2
- Zoom:9 to fit the scatterplot and then graph

In order to strengthen the analysis when comparing two variables, we can attach a number, called the correlation coefficient (r), to describe the linear relationship between two variables.

The correlation measures the strength and direction of the linear relationship between two quantitative variables.

The formula to find the correlation is:

$$r = \frac{1}{n-1} \sum \left(\frac{x_i - \overline{x}}{s_x} \right) \left(\frac{y_1 - \overline{y}}{s_y} \right)$$

Correlation is a number between -1 and 1. The strongest correlations are closer to 1 or -1.

Correlation describes only linear relationship between two variables.

Correlation does not have units and changing units on either axis will not affect correlation.

Switching the x and y axes will not change the correlation.

Correlation is very strongly affected by outliers.

2.3 Linear Regression

Linear regression or least squares regression allows you to fit a line to a scatterplot in order to be able to better interpret the relationship between two variables as well as to make predictions about the response variable.

The fitted line is called the line of best fit and has an equation:

$$\hat{y} = a + bx$$

The way the line is fitted to the data is through a process called the method of least squares. The main idea is that the square of the vertical distance between each data point and the line is minimized.

Using a calculator we can find the slope of the least squares regression line by doing:

$$\text{Slope} = r\left(\frac{S_y}{S_x}\right)$$

Using this, we can find the y-intercept as well:

$$\text{y-intercept} = \overline{y} - \text{slope} \cdot \overline{x}$$

We can get the line by using Stat $\rightarrow$ Calc $\rightarrow$ 8:LinReg(a+bx)

Correlation does not imply causation.

To describe the strength of a prediction, we use the coefficient of determination. Basically we just use r^2, and this gives the proportion of variation in the values of y that is explained by least-squares regression on y on x.

A residual is a vertical distance between an observed value of the response variable and the value predicted by the regression line.

Residual value is the Actual value minus the Predicted value.

A residual plot is plotting residuals against the explanatory variable. Essentially, it turns the regression line horizontal.

In order to draw a residual plot, you must first perform a LinReg. Next, create a StatPlot where XList is L1 and YList is RESID (From 2nd $\rightarrow$ Stat -¿ :(this one depends on the calculator))

2.4 Influential Points and Departure from Linearity

The standard deviation of the residuals gives the approximate size of a "typical" prediction error. Large values means our line is expected to give larger residuals.

An influential point in a data set, is a point that has leverage on the correlation and regression line.

If your scatterplot when graphing data does not show a linear pattern, or the residual plot pattern is not random, consider transforming the graph.

Some of the most common patterns involve transforming the x or y variables by the natural log or a square root.

3 Collecting Data

3.1 Planning a Study

In order to better understand the characteristics of a population, statisticians and researchers often use a sample from that population and make inferences based on the summary results from the sample.

A population is the entire group we want information from.

A sample is a part of the population we actually examine.

A census collects data from every individual in the population.

An observational study observes individuals and measures variables of interest but does not attempt to influence the responses.

An experiment deliberately imposes some treatment on individuals to measure their responses.

It is only appropriate to make generalizations about a population based on samples that are renadomly selected or otherwise representative of that population.

A convenience sample uses subjects that are readily avaliable.

A voluntary response sample is a sample obtained by allowing subjects to decide whether or not to respond.

A simple random sample consists of n individuals from the popultion chosen in such a way that every set of n individuals has an equal chance in the sample selected.

A stratified random sampling is when you divide the population into groups of similar individuals then select a SRS within each strata. Combine the SRSs from each strata to form your full sample.

Cluster sampling is dividing the population into sections (clusters) then randomly choose a few of these clusters. Every member of the cluster becomes your sample.

Systematic random samplling is one randomly selects an arbitrary starting point and then select every kth member of the population.

When an item from a population can only be selected once, this is called without replacement. When it can be selected more than once, it is called with replacement.

Samples are biased if they are systematically not representative of the desired population.

Voluntary response is when a sample is comprised entirely of volunteers or people who choose to participate, the sample will typically not be representative of the population.

Undercoverage occurs when some groups in the population are left out of the process of choosing a sample.

Non-response occurs when an individual chosen for a sample can't be contacted or refuses to respond.

Response bias is bias cuased by the behavior of the respondent or interviewer.

Untruthful answers occur when people give untruthful answers for several reasons.

Ignorance is when people will give silly answers just so that they appear to know something about the subject.

Lack of Memory is giving a wrong answer simply because the respondent doesn't remember the correct answer.

Timing is when a survey is taken can have an impact on answers.

Phrasing is subtle differences that can make a large difference in results.

When drawing a sample, two types of errors may occur:

Sampling Error: The difference between a sample result and the true population result. This error results from chance variation.

Non-sampling Error: Occurs when the sample data are incorrectly collected, recorded, or analyzed. Usually occurs when the sample is selected in a non-random fashion.

3.2 Selecting a Random Sample

3.3 Experimental Design

4 Probability, Random Variables, and Probability Distributions

5 Sampling Distributions

6 Inference for Categorical Data: Proportions

7 Inference for Quantitative Data: Means

8 Inference for Categorical Data: Chi-Square

9 Inference for Quantitative Data: Slopes