Distributions for Quantitative Data: Vocabulary

Individuals	Individuals are the objects described by a set of data. Individuals may be people, but they may also be animals or things.
Variable	A statistical variable is a characteristic that describes individuals in the data. A variable can take different values for different individuals.
Categorical variable	A categorical variable places an individual into one of several groups or categories. Example: a person's college major.
Quantitative variable	A quantitative variable takes a number value, which you can add or average. Example: number of people in household.
Distribution of a variable	A statistical distribution is an arrangement of the values of a variable (often a graph) showing their observed frequency of occurrence. Here is an other definition: "a representation that shows the possible values of a variable and how often the variable takes those values."
Describing the distribution of a quantitative variable	To describe a distribution, describe the shape, center, spread and outliers. **Shape:** describe the overall trends in the data. Typical ways to describe shape include symmetric, left or right skew, uniform. **Center:** give a single number that represents the data; a typical value or average (We will make this more precise in Topic 2.2) **Spread:** give a single number that measures how much the data varies. Range (maximum value minus minimum value) is one way to measure spread. (We will learn other ways to measure spread in Topic 2.3 and 2.4) **Outliers:** unusual data values.
Shape	**Symmetric:** The right and left sides of the graph are mirror images of each other **Skewed to the right:** The graph has more spread in the upper half than the lower half. There may be outliers to the right, so the graph has a longer tail to the right. **Skewed to the left:** The graph has more spread in the lower half than the upper half. There may be outliers to the left, so the graph has a longer tail to the left. **Uniform:** The graph is shaped like a rectangle. Each value occurs with the same frequency.

Experiments: Vocabulary

Population	The population is the entire group of individuals or objects that we want to study. Usually, it is not possible to study the whole population, so we collect data from a part of the population, called a *sample.*
Sample	A sample is a part of the population from which we actually collect information or data. In college-level Statistics we will use a sample to draw conclusions about the entire population.
Individual	The individuals are the members of the population, so a sample contains some of the individuals. Individuals may be people, but they may also be animals or things. Individuals can also be participants in an experiment.
Variable	A variable is the information we collect from individuals in the study. A variable can be a characteristic of an individual (such as gender) or a measurement (such as height).
Observational Study	An observational study collects information about individuals without intentionally manipulating variables. The main purpose of an observational study is to describe a group of individuals or to investigate an association between two variables. All of the research questions about a population in Activity 1.1 Exercise 1 could be investigated with an observational study.
Experiment	An experiment intentionally manipulates one variable in an attempt to cause an effect on another variable. The primary goal of an experiment is to provide evidence for a cause-and-effect relationship between two variables.
Response variable	The response variable measures the outcome of a study.
Explanatory variable	The explanatory variable may explain or even cause changes in the response variable.
Confounding variables	A confounding variable (also known as a lurking variable) is not defined as an explanatory variable in the study, but it may influence the response variable. Confounding variables can cloud or confuse the link between explanatory and response variables.

Experiments: Additional Vocabulary

Experiment	An experiment intentionally manipulates one variable in an attempt to cause an effect on another variable. The primary goal of an experiment is to provide evidence for a cause and effect relationship between two variables.
Explanatory variable	If we think a variable may explain or even cause changes, we call it an explanatory variable. In an experiment the explanatory variable is also called the treatment variable because the researcher manipulates the explanatory variable by assigning subjects to different treatments.
Response variable	The response variable is the outcome we measure for each individual.
Confounding variable	A confounding variable is a variable that is not measured in the study. It has an influence on the response variable, and its effects cannot be separated from the explanatory variable. It is related to both the explanatory and response variables and may explain the apparent association between the two. So it confounds our ability to draw a cause+ and+effect relationship between explanatory and response variables.
Direct Control	When confounding variables are directly addressed up front in the experiment design, this is called direct control.
Random Assignment	In an experiment when subjects are randomly assigned to treatments, the goal is to create treatment groups that are similar. This equalizes the effects of confounding variables across the treatment groups so that any differences in the response are not due to confounding variables.
Control Group	In an experiment a control group is the group that receives no treatment or a placebo. The control group provides a baseline for comparison.
Placebo	A placebo is a treatment with no active ingredients, such as a sugar pill.
Placebo effect	When placebo treatments produce positive improvements, we call this the placebo effect. Placebo effects arise from the power of the body to heal itself when the subject receives a placebo but believes that they are receiving treatment.
Blinding	When the researcher does not know which subjects received which treatment in an experiment, we say the researcher is "blind." Blinding the researcher controls the researcher's intentional or unintentional bias. The subjects can also be blind when they do not know which treatment they receive. Double blinding is when both researcher and subjects are blind.

Inference: Vocabulary

Parameter	A number that describes an entire **population**. Example of a parameter: At LMC 70% of all students are eligible for financial aid.
Statistic	A number that we calculate from a **sample**. Example of statistics: In a randomly selected Math 34 class, 62% of students are eligible for financial aid, but in a randomly selected English 100 class, 74% of students are eligible for financial aid.

Notation for parameters and statistics:

	(Population) Parameter	**(Sample) Statistic**
Proportion (e.g. probability) *Categorical data*	p	$\hat{p}$
Mean *Quantitative data*	μ	$\bar{x}$
Standard Deviation	σ	s

Sampling distribution	A mathematical model that describes the sample proportions from all possible random samples of size *n* from the population.
Variability in samples	When we make a calculation from a sample, the proportions, percents and means will differ from parameters.
Law of large numbers	When we have a large sample, the statistics are closer to the actual parameters. In general, the larger the sample, the closer to the actual parameters. A sampling distribution for large samples has less variability.
Patterns in large numbers of repetitions	There is less variability in a large number of repetitions. This means that in the long run, we will see a pattern, so we are more confident about estimating the probability of an event using empirical probability with a large number of repetitions.
Sample size n	We use the letter n to represent how many people, animals or objects are in the sample.
Center	We use the mean as the measure of center in a Normal distribution graph (bell curve). The mean of the sample proportions is the population proportion.

Standard Deviation SD of population proportion (also Standard Error of Sample Proportion)	The standard deviation tells us how much the graph is stretched or compressed. The formula for the standard deviation of the sample proportions is: $$\sqrt{\frac{p(1-p)}{n}}$$
Estimated Standard Error SE of sample proportion	When the population proportion is unknown, we estimate it using p. This is known as the standard error of the sample proportion. The formula is: $$\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$
Normal Approximation	If samples are repeatedly drawn from a population, the distribution of p will be approximately Normal if the number of "success" and "failure" responses is at least 10. We calculate: $np \geq 10$ and $n(1-p) \geq 10$
Success	Note that a *success* is the category of interest. It is what we are counting. For example, if we are counting blue M&Ms, a success is a blue M&M. A red M&M would be considered a failure.
Expected outcome of success	We multiply the number of the sample size times the probability of success. The formula is: $Expected\ success = np$
Expected outcome of failure	We multiple the number of the sample size times the probability of failure, 1-p. The formula is: $Expected\ failure = n(1-p)$
Z-score	The z-score measures how many standard deviations a particular statistic is away from the population proportion in a Normal distribution.
Z-score formula	$$z = \frac{statistic - parameter}{standard\ deviation} = \frac{\hat{p} - p}{SD}$$
Margin of Error	The error distance from the center of confidence interval. The z-score multiplied by the standard deviation (or standard error).
Confidence Interval	$$sample\ statistic\ \pm margin\ of\ error$$