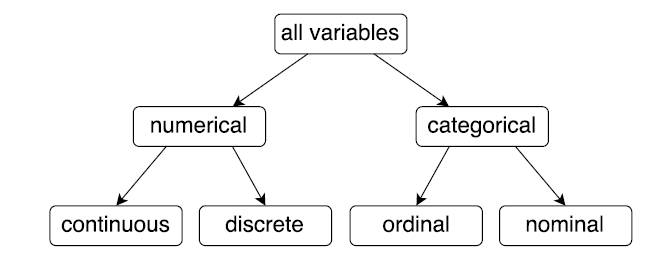
Numerical variables: can take a wide range of numerical values which are sensible to add, subtract, or average. a) Continuous variables: Can take on any value on the real number line. b) Discrete variables: Can only take numerical values with jumps.

Categorical variables: responses are categories; possible values are called levels. a) Ordinal variables: Levels have a natural ordering. b) Nominal variables: Levels do not have a natural ordering.



When two variables show some connection with one another, they are called associated, or dependent, variables.

If two variables are not associated, i.e. there is no evident connection between the two, then they are said to be independent.

Association does not imply causation!

Population: All members of a defined group that we are studying

Sample: Any subset of the population. If the sample is random, we can analyze it and use the results to make inference on the population as a whole.



Including the entire population is called a census

When you taste a spoonful of soup and decide the spoonful you tasted isn’t salty enough, that’s exploratory analysis.

If you generalize and conclude that your entire soup needs salt, that’s an inference.

For your inference to be valid, the spoonful you tasted (the sample) needs to be representative of the entire pot (the population).

Non-response: If only a small fraction of the randomly sampled people respond to a survey, the sample may no longer be representative of the population.

Voluntary response: Occurs when the sample consists of people who volunteer to respond because they have

strong opinions on the issue. Such a sample will also not be representative of the population.

Convenience sample: Individuals who are easily accessible are more likely to be included in the sample.

Simple Random Sampling: Randomly select cases from the population, where there is no implied

connection between the points that are selected.

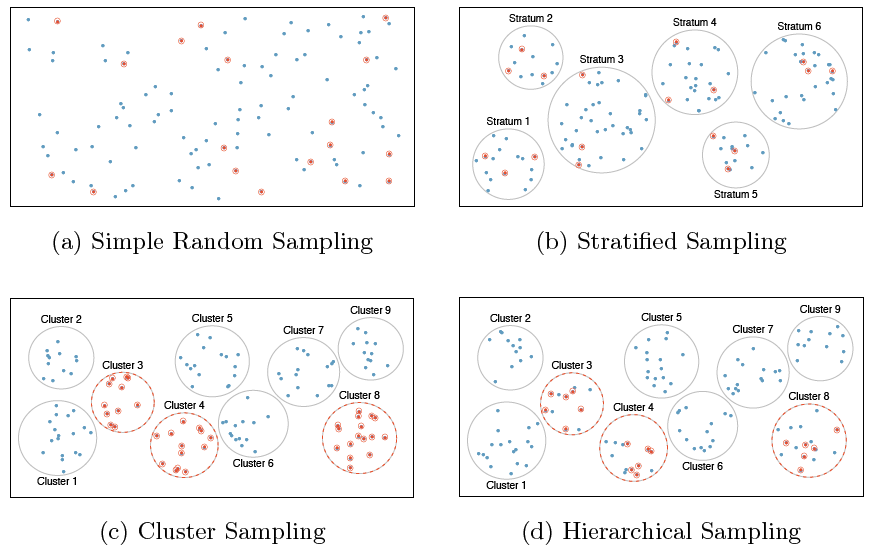
Stratified Sampling: Strata are made up of similar observations. We take a simple random sample from each stratum.

Cluster Sampling: Clusters are usually not made up of homogeneous observations. We take a simple

random sample of clusters, and then sample all observations in that cluster. Usually preferred for

economical reasons.

Multistage Sampling: Clusters are usually not made up of homogeneous observations. We take a simple random sample of clusters, and then take a simple random sample of observations from the sampled clusters. Also known as Hierarchical Sampling.



Observational study: Researchers collect data in a way that does not directly interfere with how the data arise, i.e. they merely “observe”, and can only establish an association between the explanatory and response variables. Observational studies always have a chance of confounding from unknown or unmeasurable confounders. We can only infer an association/correlation between variables, not causation. Explanatory variable might affect response variable

Experiment: Researchers randomly assign subjects to various treatments in order to establish causal connections between the explanatory and response variables. Experiments eliminate all possible confounding factors by randomly assigning treatments so confounding variables have an equally likely chance of being in control/experimental group. We can infer a causal relationship between variables. Explanatory variable affects Response variable.

Placebo: fake treatment, often used as the control group for medical studies.

Placebo effect: experimental units showing improvement simply because they believe they are receiving a special treatment.

Blinding: when experimental units do not know whether they are in the control or treatment group.

Double-blind: when both the experimental units and the researchers who interact with the patients do not know who is in the control and who is in the treatment group.

We consider gender a blocking variable. It is neither an explanatory nor a response variable. Blocking variables are characteristics the

experimental units come with that we would like to control for.

Parameter: A numerical summary about a population. Represented by letters of the Greek alphabet.

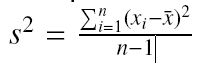
Statistic: A numerical summary about a sample. If the sample is good (representative of the population), sample statistics can serve as point estimates for the unknown population parameters. Represented by letters of the Latin alphabet.

Mean – an average of all the observations (x1+x2+x3)/n represented by xbar (sample) µ (population)

Median: The “middle” of a distribution; the value that splits the data in half when ordered in ascending order. In case of even number of data points, take the mean of the middle two numbers.

Mode: A prominent peak in the distribution.

Range: Maximum value - minimum value.

Variance: The average squared deviance from the mean. Population σ2 sample s2 

Standard deviation: The square root of the variance. 

Inter-quartile range (IQR): The 75th percentile (Q3) minus the 25th percentile (Q1), which gives the middle 50% of the data. IQR = Q3 - Q1

Boxplots are a useful for visualizing the distribution of a numerical variable, based on its median and IQG.

Draw a dark line denoting the median, which splits the data in half.

Draw a rectangle to capture the middle 50% of the data. The two boundaries of the box are called the first quartile

(the 25th percentile, i.e. 25% of the data fall below this value) and the third quartile (the 25th percentile).

The length of the box is the interquartile range (IQR) = Q3 - Q1.

Compute 1.5 X IQR. Draw the upper and lower whiskers to extend to the nearest data point that is not farther out than 1.5 X IQR from the median. Mark any point that extends beyond 1:5 X IQR from the

median as an outlier. An outlier is an observation that appears extreme relative to the rest of the data.

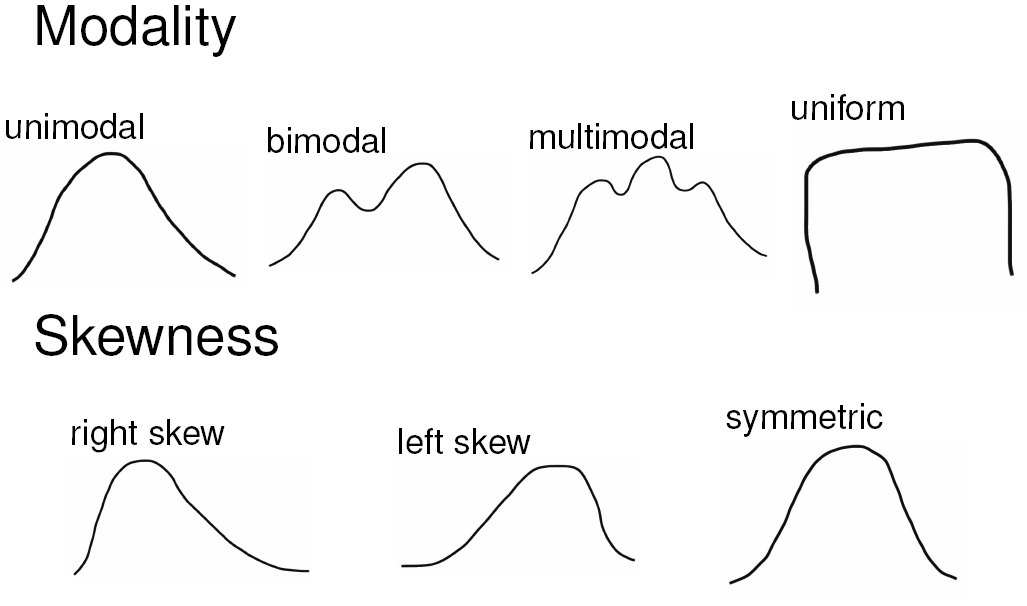
Histograms describe the distribution of a single numerical variable. Higher bars represent where the data are relatively more common. Histograms are especially convenient for describing the shape of the data distribution.

Unimodal: The histogram has a single prominent peak. Bimodal / Multimodal: The histogram has two or more

prominent peaks. Uniform: The histogram has no apparent peaks.

A histogram can be right skewed, left skewed, or symmetric.

Outliers: any unusual observations. 1.5 times the IQR above the third quartile or below the first quartile.



Scatterplots are useful for visualizing the relationship between two numerical variables. **Direction:** Positive/Negative **Shape:** Linear / Curved / None **Strength:** Strong / Weak **Outliers:** Note if any

Dot Plots are sometimes used for visualizing one numerical variable. Darker colors or stacked observations represent areas where there are more observations.

Median and IQR are more robust to skewness and outliers than mean and SD. Therefore, . For skewed distributions, it is often more helpful to use median and IQR to describe the center and spread. For symmetric distributions, it is often more helpful to use the mean and SD to describe the center and spread.

Transformation is a rescaling of the data using a function. When data are very strongly skewed, we sometimes transform them so they are easier to model. A common transformation is the log transformation.

Xnew = log(Xold)

A contingency table summarizes data for two categorical variables.

A bar chart is a common way to display a single categorical variable. A bar chart where proportions instead of frequencies are shown is called a relative frequency bar chart.

Segmented (or stacked) bar charts are made of different segments that are represented visually through colored sections. They are useful in comparing

different groups.

Mosaic plots display relative frequencies across both the horizontal and vertical axis, so they give more information than a bar chart, but they might be more difficult to visually make conclusions.

Pie charts, not good.