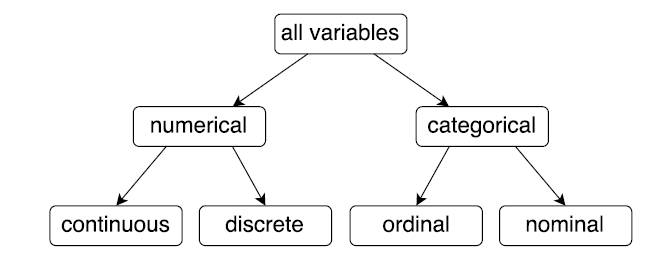
Numerical variables: can take a wide range of numerical values which are sensible to add, subtract, or average. a) Continuous variables: Can take on any value on the real number line. b) Discrete variables: Can only take numerical values with jumps.

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.

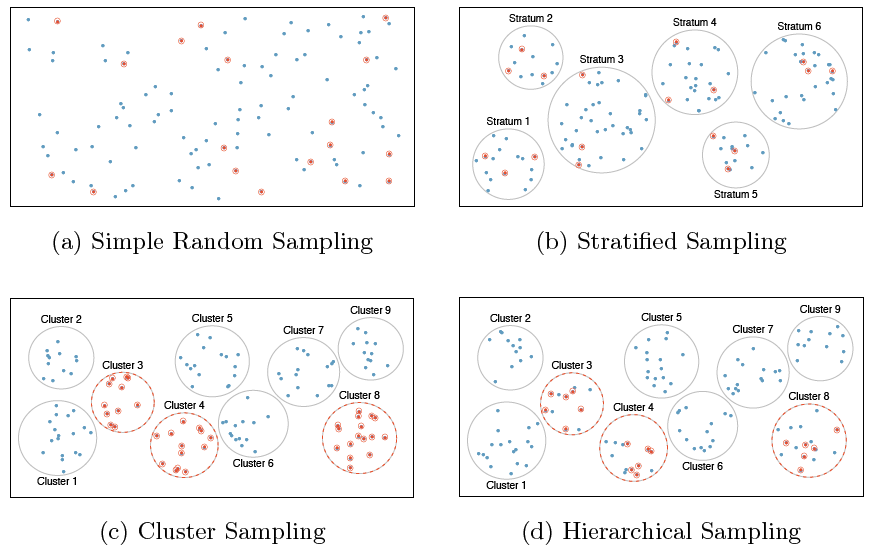
Stratified Sampling: Strata are made up of similar observations. We take a simple random sample from each stratum.

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.

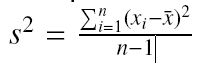
Statistic: A numerical summary about a sample. If the sample is good (representative of the population), sample statistics can serve as point estimates for the unknown population parameters. Represented by letters of the Latin alphabet.

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.

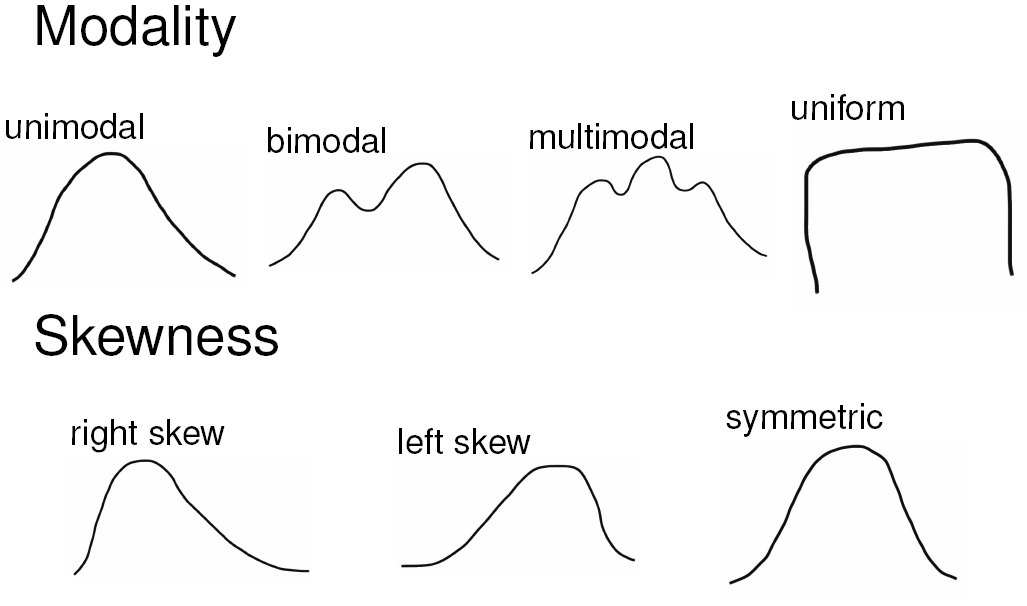
Histograms describe the distribution of a single numerical variable. Higher bars represent where the data are relatively more common. Histograms are especially convenient for describing the shape of the data distribution.

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.

Lecture 2

A random process is a situation in which we know what outcomes could happen, but we don’t know which particular outcome will happen.

Sample Space: The set of all possible outcomes of a random process. Denoted by S.

Event: Any subset of the sample space. Denoted by E.

Probability: The likelihood of an event occurring. Denoted by P(E). Alternative definition: The probability of an outcome is the proportion of times the outcome would occur if we observed the random process an infinite number of times. Computed by: number of elements in the event/number of elements in the sample

0 <= P(A) <=1 The probability of any event must be between 0 and 1.

The probabilities of all events in a sample space must sum to 1. P(A) + P(B) + …. = P(S) = 1

Disjoint outcomes: Two or more outcomes of a random process that cannot happen at the same time. Also

known as mutually exclusive outcomes. P(A and B) = 0

Non-disjoint outcomes, on the other hand, can happen at the same time. P(A and B) not = 0

Complementary events are disjoint events that are the only possible outcomes of a sample space. The probabilities of complementary events add up to 1. Head plus Tails

Union of Events: occurs if either A or B occurs (or both). A or B = A U B

Union Addition Rule: P(A U B) = P(A) + P(B) - P(A n B) disjoint P(A U B) = P(A) + P(B)

Intersection of Events: occurs if both A and B occur. A and B = A n B

Complement of an Event: occurs if an event, say A, does not occur. Not A = Ac = A bar

Conditional Probability: The probability of an event (A), given that another event (B) has already occurred.

P(A given B) = P(A | B) =

If A and B represent two events, then P(A ꓵ B) = P(A | B) X P(B)

If A and B represent two independent events, then P(A ꓵ B) = P(A) X P(B)

Two processes are independent if knowing the outcome of one provides no useful information about the outcome of the other. A and B are independent iff P(A | B) = P(A) A and B are independent iff P(A ꓵ B) = P(A) X P(B)

Marginal Probability: The probability of event A occurring. Also known as an unconditional probability. P(A)

Joint Probability: The probability of event A and event B occurring. This is equivalent to the intersection of two or more events. P(A and B), or P(A ꓵ B)

Conditional Probability: The probability of event A occurring, given that event B occurs. P(A | B)

Law of large numbers states that as more observations are collected, the proportion of occurrences with a particular outcome, ˆpn, converges to the probability of that outcome, p. Recall: ˆpn is a statistic, p is a parameter.

Lecture 3

A random variable (RV) is a numeric quantity whose value depends on the outcome of a random event. Then the probability can be written as: P(Y = y). Random variables (usually) have an associated mean, µ,and standard deviation Ϭ

Discrete random variables often take only finite or countably infinite values (ie, integers). Example: Number of credit hours, Difference in number of credit hours this term vs last.

. Continuous random variables take real (decimal) values . Example: Cost of books this term, Difference in cost of

books this term vs last.

Common discrete distributions Bernoulli, Binomial, Geometric

A Bernoulli random variable is a special type of discrete random variable which has exactly two levels, often denoted as 0/1 or success/failure. Each level has a fixed probability of occurring (recall: the probabilities must some to one). µX = p ϬX =

The Geometric distribution describes the waiting time until a success for independent and identically distributed (iid) Bernouilli random variables. Independence: outcomes of trials don’t affect each other

Identical: the probability of success is the same for each trial P(success on the nth trial) = (1 - p)n-1 p

µX = 1/p ϬX =

The Binomial distribution describes the probability of having exactly k successes in n independent Bernouilli trials with probability of success p. P(single scenario) = pk (1 - p)(n-k)

The choose function is useful for calculating the number of ways to choose k successes in n trials. , , µX = np ϬX =

Observations that are more than 2 standard deviations away from the mean are considered unusual

The Normal distribution Unimodal, symmetric, bell-shaped curve. Many variables are nearly normal, but none are exactly normal. A normal distribution with mean µ and standard deviation Ϭ is denoted by N(µ,Ϭ).

Standard Normal N(µ=0,Ϭ=1),

A percentile is the percentage of observations that fall below (to the left of) a given data point.

For nearly normally distributed data, about 68% falls within 1 SD of the mean, about 95% falls within 2 SD of the mean, about 99.7% falls within 3 SD of the mean.

A Normal probability plot, or Q-Q plot shows if data are normally distributed or if they deviate from normality Right skew - Points bend up and to the left of the line. Left skew- Points bend down and to the right of the line. Short tails (narrower than the normal distribution) - Points follow an S shaped-curve. Long tails (wider than the normal distribution) - Points start below the line, bend to follow it, and end above it.