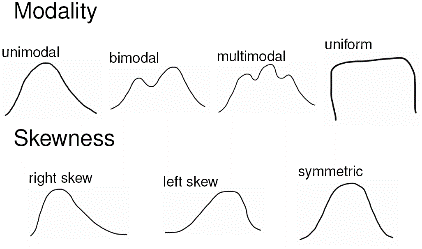
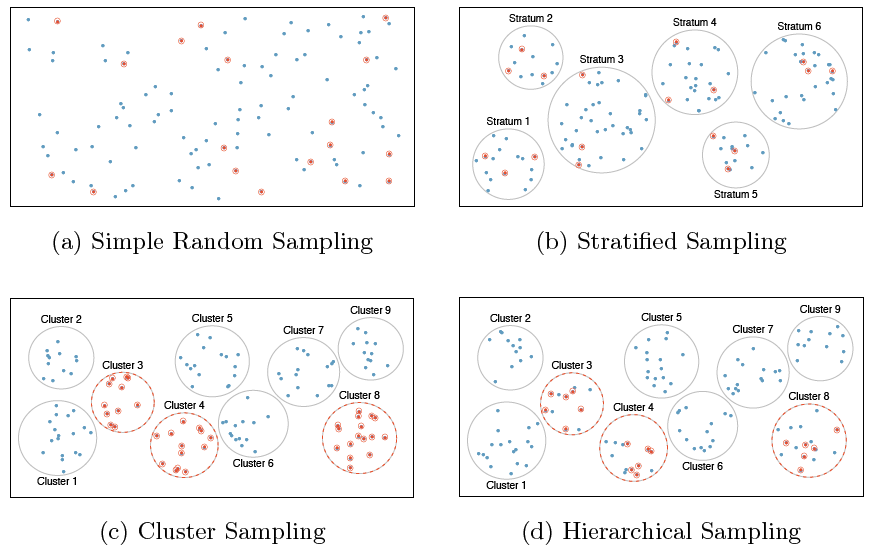
Numerical variables: numbers can add, subtract, or average. a) Continuous variables: any value on the real number line. b) Discrete variables: Can only take numerical values with jumps.

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 vars

If two variables not associated, there is no evident connection between the two, then they are independent.

Association does not imply causation!

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 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 (Hierarchical): Clusters are usually not made of homogeneous observations. We take a simple random sample of clusters, and then take a simple random sample of observations from the sampled clusters.

Sample aka cases: 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.

Bias: Non-response: only small fraction of randsampled people respond, the sample may not be representative of population.

Bias: Voluntary response: Sample consists of people who volunteer because they have strong opinions on the issue. Such a sample will also not be representative of the population.

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

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.

Blocking variables are characteristics the experimental units come with that we would like to control for (gender). It is neither an explanatory nor a response variable.

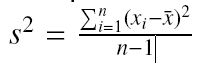
Parameter: A numerical summary about a population. Represented by letters of the Greek alphabet.

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 % (Q3) minus the 25th % (Q1), which gives the middle 50% of the data. IQR = Q3 - Q1

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

Boxplots are a useful for visualizing the distribution of a numerical variable, based on its median and IQG. Steps 1) Draw a dark line denoting the median 2) Draw a rectangle to capture the middle 50% of the data. The boundaries are Q1 and Q3 (IQR) 3)

Compute 1.5 X IQR. Draw the upper and lower whiskers to the nearest data point that is not more than 1.5 X IQR from the median. 4) Mark any point that extends beyond 1:5 X IQR from the median as an outlier.

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.

Transformation is rescaling the data via a function. When data are very strongly skewed, we can 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 on both the horizontal and vertical axis but they might be more difficult to visually make conclusions.

Pie charts, not good.

Lecture 2

Sample Space: The set of all possible outcomes of a random process S. Event: Any subset of the sample space. Denoted by E.

Probability: The likelihood of an event occurring. Denoted by P(E). 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. P(A and B) = 0

Non-disjoint outcomes: 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). P(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 event (A), given that 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 information about the outcome of the other. TEST FOR INDEPENDENCE 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

Common discrete distributions Bernoulli, Binomial, Geometric

Bernoulli random variable is a 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. µX = p ϬX =

Geometric distribution describes the waiting time until a success for independent & 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. Requires 1) independent trials, 2) fixed number of trials 3) only 2 outcomes 4) probability of success remains constant 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.

Lecture 4

Central Limit Theorem: The distribution of the sample mean is well approximated by a normal model:

¯x ~ N(mean = µ, SE = ) where SE is represents standard error, which is defined as the standard deviation of the sampling distribution. If σ is unknown, use s. Sampling distributions are symmetric and centered at the true population mean. Note that as n increases, SE decreases.

Criteria for CLT (conditions for inference): Independence and Sample size/Skew distribution is normal or if skewed sample is large > 30

confidence interval (1-α)% is defined as: point estimate (xbar) ± Zα/2 x SE In a confidence interval, Zα/2 x SE is called the margin of error, and for a given sample, the margin of error changes as the confidence level changes.

Wider Interval (less precise) = smaller α & N, Larger σ Narrower Interval (more precise) = larger α & N, smaller α

SE = , Z = , P(xbar>var| µ=val) = P(Z>) if P(Z) is lower than 5% we reject H0

Type 1 Error: reject the null hypothesis when H0 is true. Type 2 Error: fail to reject the null hypothesis when HA is true.

Increasing α increases the Type 1 error rate. Decreasing α increases type 2 error rate

Hypothesis Testing set hypotheses, check assumption conditions, calculate a test statistic and p value, make a decision and interpret it in context of the research question If p-value < α, reject H0, data provide evidence for HA, If p-value > α, do not reject H0, data do not provide evidence for HA

The point estimate x(bar) = or p(hat) =

Clinical Significance: Real differences between the point estimate and null value are easier to detect with larger samples.

However, very large samples will result in statistical significance even for tiny differences between the sample mean and the null value (effect size), even when the difference is not clinically significant.