Course Notes

Introduction to Statistics

- Statistics is a discipline defining a set of procedures used to collect and interpret numerical data.
- The discipline of statistics serves two purposes:
    - Statistical procedures can be used to describe the relevant characteristics (dispersion, central tendency) of a body of data. This is called **Descriptive Statistics** .
    - Statistical procedures can be utilized to help us make inferences or predictions about a whole population based on information from a sample of the population. This is called **Inferential Statistics**.

Measurement and Sampling

- A **population** consists of all the observations with a given set of characteristics. It is all the possible observations within the group the researcher is studying.
- A **sample** is a portion of the population.
- In inferential statistics, a sample is selected to represent the population studied.
- A sample will, on average, be representative of the population if the procedure used to select the sample is unbiased.
- An unbiased sampling procedure is one in which each observation in the population has an equal chance of being chosen for the sample.
- The possible biasedness of a sampling procedure depends partially on the exact definition of the population.
    - For example if a researcher were studying CSULA male students, it would not be biased to select a sample from among only males.
    - If the researcher were studying CSULA students in general, then it would be biased to select only males.
- Random sampling is an unbiased procedure but there are other unbiased sampling procedures.
- Stratified sampling, in which the sample is purposely selected so that certain characteristics of the sample matches that of the population, is not completely random but nevertheless can be unbiased.

Distribution and the Visual Display of Data

- A frequency distribution illustrates the number of observations in a data set that fall into various classes of the data.
- A **category (or bin)** is an interval of the data.
    - Categories must be mutually exclusive. No observation in the data should fall within more than one category.
    - Categories must also be exhaustive. Each observation in the data must fall within a class.

- The number of observations falling into the various classes in the distribution is called **frequency** (denoted f_i).
- A relative frequency distribution reveals the percentage of observations that fall into the various classes of data.
- A histogram is a graphical representation of a frequency or relative frequency distribution.

Summary Description of Data

- The summary information we usually want to know about a data set is the center of the data (measure of centrality) and how disperse the data set is (measure of dispersion).
- A **parameter** is a numerical characteristic of the population.
- A **sample statistic** is a numerical characteristic of the sample.
- There are three measures of centrality: **mean, median** and **mode**.
    - The mean (arithmetic average) is the most common measure of centrality.
    - The mean of the population is denoted μ The mean of the sample is denoted $\overline{X}$.
    - The median is the middle value of the data ordered from lowest to highest.
    - The value in a data set that occurs with the greatest frequency is defined as the mode.
    - The value of the mean is sensitive to the outliers in a data set whereas the median is not.
    - The median will equal the mean if the distribution of the data is symmetric.
- A symmetric distribution is one in which the side of the distribution to the right of the mean is a mirror image of the left portion.
- If a distribution is skewed to the right, outlier values in the data much larger than the mean are pulling the value of the mean above the median.
- If a distribution is skewed to the left, outlier values in the data much smaller than the mean are pulling the value of the mean below the median.
- There are three measures of dispersion: **range, variance** and **standard deviation**.
    - Variance and standard deviation are both measures of how a data set varies with respect to its mean.
    - The variance of population data is: $\sigma^2 = \dfrac{\sum_i (x_i - \mu)^2}{N}$
    - The variance of sample data is: $S^2 = \dfrac{\sum_i (x_i - \bar{x})^2}{n-1}$
    - Standard deviation for either the population or the sample equals the square root of the variance.
    - The larger the variance or standard deviation, the greater the variation in the data around its mean.

- The **coefficient of variation** expresses standard deviation as a percentage of the mean: $CV = \frac{S}{\bar{x}} * 100\%$
- The coefficient of variation is useful in comparing the variation of data sets that have different means.

The Normal Distribution

- A continuous distribution is represented by a smooth curve.
- Probability is represented by the area under the curve.
- The **normal distribution** is the most common continuous distribution in statistics. Many variables in the social and natural world are normally distributed.
- The normal distribution is a formula that draws a family of symmetric curves.
- Each distinct normal curve has its own mean and variance.
- Characteristics of the normal distribution:
    - The normal curve is symmetric around the mean, μ.
    - The normal curve extends from negative to positive infinity.
    - The total area under the normal curve sums to one.
    - The mean, median and mode of the distribution equal one another.
    - Empirical Rule: If a variable follows a normal distribution,
        - 68% of its observations will be within one standard deviation of its mean.
        - 95% of its observations will be within two standard deviations.
        - 99% of its observations will be within three standard deviations.
- The standard normal, or Z-distribution is a specific normal curve with a mean $\mu=0$ and variance $\sigma^2=1$.
- The value of the variable Z represents standard deviations from the mean.
- If the variable X represents individual observations in a population, the formula $Z = \frac{x - \mu}{\sigma}$ transforms the variable into Z.

The Concept of Probability

- A **random experiment** is any activity whose outcome cannot be predicted with certainty (for example a coin toss).
- Each possible outcome of an experiment is called a **basic outcome**.
- An **event** is a collection of basic outcomes that share some characteristic.
    - For example, if the experiment consisted of randomly selecting a student in class, each student would represent a basic outcome whereas left-handed students would exemplify an event.
- An event (A) composed of three basic outcomes is denoted $A=\{O_1,O_2,O_3\}$.
- The probability of event A occurring, P(A), can be assigned using different approaches.

- **Relative Frequency Approach.** The experiment may be repeated n number of times and f_A, the frequency of event A could be observed. The relative frequency, $\frac{f_A}{n}$ could be used to approximate probability.
- **Equally Likely (or Theoretical) Approach.** If each basic outcome is equally likely to occur, the probability of event A can be calculated as the sum of the chances of its basic outcomes.

- The **Law of Large Numbers** states that if an experiment is repeated through many trials, the proportion of trials in which event A occurs will be close to the probability *P(A)*. The larger the number of trials the closer the proportion should be to *P(A)*.
- A **union** of two events is composed of all those basic outcomes that belong to at least one of the events.
- An **intersection** of events is composed of those basic outcomes that fall in both events simultaneously.
- **Conditional probability** is the probability of an event occurring conditional on another event having already arisen.
  - Suppose A={females} and B={right-handed people}.
  - $A \cup B$, the union of the events, consists of all those who are right-handed, female or both.
  - $A \cap B$, the intersection of the two events, consists of right-handed females.
  - $P(A|B)$, the probability of event A conditional on event B, is the probability of being female conditional on being right-handed.
  - In an experiment that selects students, $P(A|B)$, would be the probability of selecting a female if we chose only from right- handed students.
- The formula to calculate the probability of the union of the events A and B: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.
- The formula for conditional probability: $P(A|B) = \frac{P(A \cap B)}{P(B)}$.
- If events A and B are independent then the likelihood of one event occurring is not a function of the other event.
- If event A is independent of B, then $P(A|B) = P(A)$.
  - In the example, under independence, our chance of selecting a female is not altered if we condition our selection to those who are right handed.
- If the two events are independent, the probability of the intersection of events A and B is calculated as: $P(A \cap B) = P(A) \cdot P(B)$

Discrete Probability Distribution

- A **Discrete Probability Distribution** assigns probability to the possible values of the discrete variable X.
- The probabilities that make up any probability distribution must sum to 1 (100%).

- The notation expressing the probability that the variable X equals its specific value x_i is $P(X=x_i)$.
- The variable X is considered discrete if we are able to observe and count each of the different values of the variable.
    - For example, student attendance in a statistics class over the course of a quarter would be represent a discrete variable. The different values of the variable would be observable and countable.
- The **expected value**, or mean, of the variable X can be calculated using the information provided by its distribution.
- The formula for the expected value of the discrete variable X is $E(X) = \mu_x = \sum_i x_i \cdot P(X = x_i)$.
- The calculation of the mean for the discrete variable X differs from the simple formula for the average because we are using the information on probability given by the distribution, we are not utilizing individual values of X directly in the calculation.
- The formula for the variance of the discrete variable X is $\sigma^2 = \sum_i (x_i - \mu)^2 P(X = x_i)$.
- The standard deviation σ is the square root of the variance.
- Variance and standard deviation are indexes measuring the degree of variation in X.

Binomial Distribution

- The **binomial distribution** generates probabilities arising in a repeated experiment with only two possible outcomes for each repeated trial.
- The distribution derives from a formula that generates probability for the following type of experiment:
    - Suppose a fair coin is flipped five times:
        - The variable X would represent the number of "successes" which we can define as the number of times the coin lands heads. The variable takes on six different discrete values: {0, 1, 2, 3, 4, 5}.
        - The number of trials of the experiment would be n=5.
        - Given that we know there is a 50% probability that any one coin toss will result in heads we can generate a distribution assigning probability to each value of the variable X.
- The binomial formula: $P(X = x) = {}_nC_x p^x q^{n-x}$ where x=0, 1, 2...n; p is the probability of "success" and q equals 1-p.
- ${}_nC_x$ counts the number of ways the variable X can equal a given value.
    - In the above example, ${}_5C_2$ would, for instance, tell us how many possible ways we could get two successes (heads) in five coin flips {hhttt, hthtt, httht, etc.}.

- The formula for combinations is ${}_nC_x = \frac{n!}{x!(n-x)!}$.
- If X follows a binomial distribution :
  - The expected value or mean of X: E(X)=*np*
  - The variance of X: Var(X)=npq
  - The standard deviation of X: $\sigma = \sqrt{npq}$.
- As the number of trials increases, the distribution of X becomes more symmetric.
- If *np* and *nq* equal at least five, the binomial distribution of X may be approximated by the normal distribution.

From Samples to Population

- **Parameters** are numerical characteristics of the population.
- The numerical characteristics of a sample are called sample **statistics**.
  - The mean, variance and standard deviation of the population are denoted, respectively: μ, σ^2, σ.
  - The mean, variance and standard deviation of the sample are denoted, respectively: $\bar{X}$, S^2, S.
- Sample statistics serve as estimates of the population parameters.
- If the sample is drawn from the population in an unbiased manner, the sample mean, $\bar{X}$, , is an unbiased estimate of the population mean, μ.
- Unbiasedness means that, **on average**, the statistic $\bar{X}$ will equal the population parameter μ.
- In most cases there will be some difference between the sample statistic and population parameter. This difference is called **Sampling Error**.
- Characteristics of $\bar{X}$.
  - $\bar{X}$ is a variable. It is calculated from a sample and different samples taken from a population will typically generate different values of $\bar{X}$.
  - $\bar{X}$ follows a sampling distribution which assigns probability to the different possible values of the variable.
  - The expected value of the sample mean is the population mean: $E(\bar{X}) = \mu$.
  - The variance of $\bar{X}$ is $\frac{\sigma^2}{n}$ and the standard deviation is $\sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$ where *n* is the size of the sample taken to calculate $\bar{X}$.
- The variance and standard deviation of $\bar{X}$ can be estimated from sample data, in which case the calculated variance and standard deviation would be $\frac{S^2}{n}$ and $\frac{S}{\sqrt{n}}$.

- The variance and standard deviation of $\bar{X}$ decrease as sample size increases. This is seen from the above formulas for the variance and standard deviation in which *n* is in the denominator.
- The distribution of $\bar{X}$ can be approximated by the normal curve (with a specific mean and variance) if the sample size is at least thirty observations. This holds regardless of the distribution of the population that is sampled.
- A population proportion, denoted **P**, is the percentage of observations within a population that has a specific characteristic.
- A sample proportion, denoted $\hat{p}$, is the percentage of observations within a sample that has a specific characteristic.
- $\hat{p}$, calculated from a sample, is a variable with the following characteristics:
    - The expected value of the sample proportion is the population proportion, $E(\hat{p}) = P$.
    - The variance of $\hat{p}$ is $\frac{Pq}{n}$ and the standard deviation is $\sqrt{\frac{Pq}{n}}$ where *n* is sample size and *q=1-P*.
    - If *nP≥5 or nq≥5* the distribution of $\hat{p}$ can be approximated by the normal distribution with the relevant expected value and variance.

Interval Estimation of Population Mean and Proportion

- $\bar{X}$ is a point estimate for the population parameter μ.
- The point estimate for μ does not utilize information provided by the standard deviation of $\bar{X}$ on the possible magnitude of sampling error.
- A confidence interval for μ utilizes this information by generating an interval around $\bar{X}$ in which there is a 100(1-α)% probability that μ is within the interval.
    - *1-α* is the level of significance of the confidence interval, which is set by the researcher. It gives the probability that the population parameter will fall within the constructed interval.
- The formula that generates the confidence interval for μ around our sample $\bar{X}$ point estimate: $\bar{X} \pm Z_{\frac{\alpha}{2}} \sigma / \sqrt{n}$ where $Z_{\frac{\alpha}{2}}$ is a value in the standard normal distribution in which $P(Z > Z_{\frac{\alpha}{2}}) = \frac{\alpha}{2}$. $\frac{\sigma}{\sqrt{n}}$ is the standard deviation of $\bar{X}$.
    - The use of σ in the formula implies that we know the value of the population variance.
    - We are also to assume that the population that's sampled is normally distributed.

    - If these requirements do not hold, we can still use the Z-distribution to estimate the confidence interval at a given level of significance if our sample size is at least thirty observations.
    - In this case we would be estimating σ with the sample standard deviation, ***S***.
- There is a tradeoff made between interval size and level of confidence.
    - The greater the level of confidence (1-α) the larger the size of the interval calculated around $\overline{X}$.
    - The researcher wants the precision of a **small** interval with a **large** level of confidence.
- $\hat{p}$ is a point estimate for the population parameter ***P***, the population proportion.
- The formula that generates the confidence interval for ***P***, is $\hat{p} \pm Z_{\frac{\alpha}{2}}(\hat{p}\hat{q})$ where $\hat{q}$ equals $1 - \hat{p}$.

t-Distribution

- If the population is normal but the variance is unknown and the sample size is less than thirty, confidence intervals may be constructed using the t-distribution.
    - Characteristics of the t-distribution:
        - As in the case of the normal distribution, the t-distribution is a formula that draws a family of symmetric curves.
        - The mean of *t* is 0.
        - The variance of the t-distribution is always greater than one.
        - The variance of *t* is calculated as $\frac{\nu}{\nu - 1}$ where ν equals *n-1* and is called *degrees of freedom.*
- The formula to calculate a confidence interval for μ at the *1-α* level of significance is, $\overline{X} \pm t_{\frac{\alpha}{2},\nu} \cdot S/\sqrt{n}$ where $t_{\frac{\alpha}{2},\nu}$ is a t-value in which $P(t > t_{\frac{\alpha}{2},\nu}) = \frac{\alpha}{2}$. ***S*** is an estimate for σ.

Hypothesis Testing

- A hypothesis test involves testing an idea the researcher has about the value of a population parameter.
- A hypothesis test always involves two competing ideas as to the value of the population parameter. One hypothesis is placed within the null, H_0, the other is placed within the alternative, H_1.
- Going into the hypothesis test, the assumed value of the population parameter is that which is placed within H_0.
- The hypothesis placed within H_0 is considered in an advantaged position since, normally, only if sample evidence strongly suggests H_0 is incorrect will the researcher reject H_0.

- In the formal test, if the test statistic falls within a predetermined range of values (the critical or rejection region) the null hypothesis is rejected.
- The alternative hypothesis determines whether the test will be a one tailed or two tailed test. It will determine on which side the rejection region will be for the one tailed test.
- A type I error occurs if the researcher incorrectly rejects the null hypothesis. A type II error occurs when the researcher incorrectly fails to reject H_0.
- The researcher controls the probability of making a type I error by setting the level of significance of the test.