Fundamental concepts

Fundamental concepts regarding Statistics and probability

0.1 Basics

- A **summary statistic** is a single number summarizing a large amount of data. Some important measures are - mean, standard deviation, inter quartile range.
- **Numerical data** refers to variables that not only take on numerical values but for which it is sensible to add, subtract and perform mathematical operations on. Among numerical variables we can have **discrete** and **continuous** variables taking on integer and floating point values respectively.
- Certain variables that contain string type values or even a few number of discrete values might qualify as **categorical** variables and they possible values they take on are called **levels**. When a categorical variable's levels have some sort of natural orderning (low, medium, high) then we say that the variable exhibits **ordinal** behaviour. Whereas a categorical variable showing no such natural ordering is said to be **nominal** in nature.
- When two variables show some sort of connection with each other they are called **associated** variables If two or more variables are not associated then are termed as **independent**.
- When we study the relationship among variables we often want to determine if a change one variable causes a change in the other variable or not. In this hypothesized relationship we want to know if the **explanatory variable** might affect the **response variable**. This is typically what is known as a **causal relationship**

0.2 Sampling

- We are typically interested in finding out key measures about a **population** which is the entire set of all possible observations. We do this by **sampling** and a **sample** refers to a subset of the population. It is advisable to have

random sampling instead of having **bias** in our sampling since that might skew our results.

- A **confounding** variable is a variable that is correlated with both the explanatory variable and the response variable.
- The **mean** of a dataset is a measure of the center of the distribution of data values.
- The **weighted mean** is basically the same thing as a mean except that in this case we tend to give more or less importance to certain values by giving them a weight in the computation of the mean. For example to get the average income across the country, if we are given state population and average income per state, we can weight each state's average income by their fraction of total population and then add up these values to obtain the weighted average.
- Suppose we want to plot the distribution of a single variable. Now if we are given a large number of values it is impractical to plot each and every observation. Rather we rely on first **binning** the data and then plot the frequency of occurrence of values in respective bins - the resulting figure is what is termed as a **histogram**.
- **Mode** refers to the value or observation with most occurrences in our dataset.
- The distance of an observation from its mean is called its **deviation**. The average squared deviation of observations from their mean is called the **variance** and gives us a measure of the variability in the data.
- A **boxplot** summarizes a dataset using 5 statistics. First we have the line denoting the **Median** which splits the data in half. Then we build a rectangle around this median line to represent the middle 50% of the data. The length of this box is called the **IQR**. The two box boundaries are the **first quartile** which is the value below which 25% of the data lie and the **third quartile** below which 75% of the values lie. $IQR = Q_3 - Q_1$. Then finally there are the **whiskers** that extend outward in either direction of the box which are computed as: $1.5 \times IQR$.
- Observations lying outisde these boxplot whiskers are called **outliers**. It refers to an observation that seems to be quite extreme relative to rest of the data.
- Median and IQR are called **robust statistics** because extreme values have little effect on their values.

0.3 Probability

- Two outcomes are said to be **disjoint** or **mutually exclusive** if they both cannot happen simultaneously.

- A **probability distribution** is a list or a plot of all possible outcomes along with their associated probabilities of occurrence while ensuring that - the outcomes are disjoint, all probabilities must add up to 1 and the probability lies between 0 and 1.
- Set of all possible outcomes is called the **sample space**.
- Probability is nothing but a rmeasure associated with andom process that results in an outcome. It is characterized by random variables and just as regular variables can be independent, random variables can also be independent if **knowing the outcome of one provides no useful information about the outcome of the other**.
- If the probability measure is based on a single outcome it is called **marginal probability** and if the probability of outcomes is based on two or more variables it is called as a **joint probability**. Conditional probability of outcome A given outcome B is shown as:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{1}$$

- The **Bayes theorem** is given by:

$$P(A_1|B) = \frac{P(B|A_1)P(A_1)}{P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + \cdots + P(B|A_k)P(A_k)} \tag{2}$$

- The **expected value** of a discrete random variable is given by:

$$E(X) = \sum_{i=1}^{n} x_i P(X = x_i) \tag{3}$$

- **Variance** of a discrete random variable is given by:

$$\sigma^2 = \sum_{j=1}^{n} (x_j - \mu)^2 P(X = x_j) \tag{4}$$

- When dealing with **continuous random variables** we refer to probability distributions as **probability density functions**.
- The **Z-score** of an observation is the number of standard deviation it falls above or below the mean.

$$Z = \frac{x - \mu}{\sigma} \tag{5}$$

0.4 Statistics

- Statistical inference is concerned with quantifying the uncertainty of parameter estimates.
- A population measure like the fraction of certain category of individuals or the mean value is called the **parameter**. Lets call it p. We attempt to estimate this parameter by finding a **point estimate** using a sample. Call it $\hat{p}$. Difference between the point estimate and the true parameter value is called the **error**.
- **Bias** refers to the systematic tendency for a sample estimate to over or under estimate the true value of a parameter.
- The distribution of sample means or sample proportions (with each sample mean obtained from repeated trials of an experiment) is called a **sampling distribution**.
- We know that the sample mean or sample proportion provides a point estimate of the true proportion. but we must note that this measure is not perfect and has some standard error associated with it. So when we are stating an estimate for a population parameter it is best to state a range of values. This range is called the **confidence interval**. Note that the standard error is called the standard deviation of the sample point estimate.
- We know that under the **Central limit theorem** the sample point estimate follows a normal distribution so we can say that 95% of the data values actually lie within 1.96 standard deviations from the mean on either side. Hence we are essentially constructing a 95% confidence inteval which means that we are 95% confident that our interval would capture the true parameter value.

  $$\text{point estimate} \pm 1.96 \times S.E \tag{6}$$

  This means that if we took many samples, each time time constructing a 95% confidence interval, then 95% of such intervals would contain the true value of the parameter.
- The **Null hypothesis** represents the skeptical perspective or the claim to be tested and the **alternate hypothesis** is the alternative claim under consideration. The null usually represents a perspective of 'no relation'. Under this framework we ask ourselves the question: does the data provide significant evidence that the true value of the mean or proportion is something other than the null value. Is the sample estimate deviation simply due to chance or is the true value of the parameter is actually different than the null specification.
- If the null value lies in the 95% confidence interval then we can say that the null value is not implausible.

- **Type 1 error** means rejecting the null hypothesis when it is actually true. **Type 2 error** is failing to reject the null hypothesis when the converse is true.
- Usually we design our test such that in cases when the null hypothesis is actually true, we do not want to **incorrectly reject** H_0 about 5% of the time. This is called the **significance level**. If the null hypothesis is actually true, the significance level tells us how often the data might lead us to incorrectly reject it.
- **P-value** is the probability of observing data atleast as favorable to the alternative hypothesis if the null hypothesis were true. If as per the null distribution, our sample estimate falls on a extremely far off tail point, then the p-value tells us the probability of observing such an extreme value by chance if the null value were true.
- If the p-value is less than α then we reject the null hypothesis.