LECTURE 1 AND 2

ANISH SUNDARAM

Contents

1 Populations Samples and Descriptive Statistics

1.1 What are Statistics.

Definition 1. Uncertainty: Whenever data are involved, there is almost always uncertainty (aka randomness, stochasticity, error, etc). Data sets are invariably measured with some error.

Definition 2. Statistics: Statistics is the practice collecting and analyzing numerical data in large quantities, especially for the purpose of inferring proportions in a whole from those in a representative sample.

1.2 Population and Sampling.

Definition 3. Population(N): The entire group of objects $W_1....W_N$

Remark. Typically the size of Population N is very large, or its result isnt very meaningful

Date: September 5, 2021.

Definition 4. Sample(n): Samples are sub-sections of the populations where typically we look at different cross-sections of the population

Definition 5. Population Mean/Average(μ):

$$\mu = \frac{1}{N}\sum_{i=1}^{N} w_i = \frac{w_1 + w_2 + w_N}{N}$$

Definition 6. Sample Mean($\bar{x}$): Sample mean is the arithmetic average of the measurements in the sample, found mathematically by:

$$\bar{X_n} = \frac{1}{n}\sum_{i=1}^{n} = \frac{X_1 + X_2 + X_n}{n}$$

Remark. Naturally,$\bar{X_n}$ a "very good" approximation of the μ especially forlarge sample sizes n.

1.3 Sampling Error.

Definition 7. Statistic: A quanitity based on a smaple and meant to estiamate a population parameter

Remark. The sample meanX_n is a statistic used to estmate the pipulation parameter μ

1.4 Random Sampling.

Definition 8. Simple Random Sampling(SRS): Random sampling is a part of the sampling technique in which each sample has an equal probability of being chosen. A sample chosen randomly is meant to be an unbiased representation of the total population.

Remark. Simple Random Sampling is the hallmark of collecting representative samples from the population.

Definition 9. Double-blinded Randomized Trials: A randomized trial in which the subjects are divided into two groups, one with the treatment and one with a placebo, and neither the doctor handing treatment nor the patient know which pill is which, only the directors.

1.5 Other types of Sampling.

Definition 10. Convenience Sampling: Random samples may be impossible to come by and sometimes researchers settle at convenience sampling, sampling groups that may not be representative and pigeonholed yet easy to test. This is prone to **Selection Bias**

Definition 11. Voluntary Sampling: Samples where people voluteer themselves to be included into the study because only those with a strong opinion will respond which will polarize results. This is also prone to **Selection Bias**

Definition 12. Stratified Sampling: A method of sampling where we divide the population into homogeneous subgroups before doing an SRS.

1.6 Types of Data.

Definition 13. Types of Data:

(1) **Categorical Data**: Non-numerical values such as gender, blood type, ethnicity. Essentially char-based types
(2) **Numerical Data**: Real number values on a continuous interval, for example temperature, weight, insurance losses, concentration, etc. Ints or Floats
(3) **Ordinal/Count Data**: Typically, non-negative integer-valued data, e.g.,number of accidents on I-95 during the period of a week; number of foxes in a given area; number of gamma-ray bursts.

1.7 Measures of Location.

Definition 14. Outliers: Numerical Values much smaller or much smaller the average selection of values. Outliers can skew measurements which are sensitive to outliers like the Mean and can render them inappropriate.

Definition 15. Linearity: The ability of a set of data to be fit by a linear regression line. The mean is a linear function of the data.

Definition 16. Median: The Median is the literal middle value of a dataset, and is calculated the same way. The Median is robust to outliers such that outliers are not impactful.

1.8 Measures of Variability.

Definition 17. Population Standard Deviation(σ): Standard Deviation is a quantity calculated to indicate the extent of deviation for a group as a whole. It can be mathematically found by the equation:

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(w_1 - \mu)^2}$$

for a population

Definition 18. Population Variance(σ^2): Variance qualifies how dispersed the data is from the mean μ and is calculated as the square of the Standard Deviation σ in order to get a positive distance value and to amplify the effect.

Definition 19. Sample Standard Deviation and Variance(s and s^2): For samples the formula for finding the standard deviation and resulting variance are linked but use sample values as opposed to population parameters:

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_1 - \bar{X_n})^2}$$

and the variance(s^2) is simply the square.

1.9 Chebyshev's Inequality.

Theorem 1. *Chebyshev's Inequality*: The proportion of the population values farther than k¿1 standard deviations(σ) from the mean μ is no greater than $1 - \frac{1}{k^2}$.

2 More on descriptive and Graphical Statistcs

2.1 Stem-and-leaf and Dot plots.

Definition 20. Stem-and-leaf Plot: Graph consisting of a stem a which contains a shared point, and various leafs which all "branch from the stem". For example:

(1) 4 — 259
(2) 5 — 0111133556678
(3) 6 — 067789
(4) 7 — 0123344456666699
(5) 8 — 000012223344456668
(6) 9 — 013

Definition 21. Dot-plots: Useful for small to moderate data sets the dot plot involves placing stacked dots to represent number of times a value has been reached per value apparent. Very similar to Histograms

2.2 Histograms.

Definition 22. Histogram: Similar to dot-plots but using buckets for value ranges as opposed to including every single point value, essentially an abstraction of a dot-plot. Bins are arranged [x,y) aside from the last bin which is [x,y]

Definition 23. Relative Frequency Graph: A relative frequency histogram is a type of graph that shows how often something happens, in percentages. Essentially we relate how much of the total sample rests within each bin using the equation

$$R_i = \frac{F_i}{N}$$

Theorem 2. *Friedman-Diaconis rule*: A method of determining how many bins k or the bin-size.

$$k \approx \frac{range}{2 \cdot IQR} \cdot N^{1/3}$$

2.3 Box-plots.

Definition 24. Inter-Quartile Range(IQR): The difference between the 3rd and 1st quartiles in a dataset, in other words the middle 50 percent. The IQR is a measure of variability and can be found by:

$$IQR = q(0.75) - q(0.25)$$

Definition 25. Box-plots: A easy-to-grasp visual summary of location, variability, and outliers using quartiles and outlier points. The box itself is constructed from the 1st and 3rd quartiles with a line at the median, and whiskers indicating the most extreme values within 1.5 * IQR from the nearest quartile.

Remark. The outliers are all observations farther than 1.5×IRQ from thenearest quartile – they are all displayed as circles.