

# 36-200: Reasoning with Data Final Exam Note Sheet:

Name: Sean Balbale Andrew Id: sbalbale August, 2022

## Formulas:

### Standard Deviation:

For a sample containing  $n$  data values, the **Sample Standard Deviation:**  $S = \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}}$

$X_i$  is the variable,  $\bar{X}$  is the sample mean.

For a population containing  $N$  data values, the **Population Standard Deviation:**  $\sigma = \sqrt{\frac{\sum_{i=1}^N (x_i - \mu)^2}{N}}$

$X_i$  is the variable,  $\mu$  is the population mean.

$S$  is a statistic,  $\sigma$  is a parameter.

**Binomial Formula:**  $P(X = x) = \binom{n}{x} p^x \cdot (1-p)^{n-x}$

$P$  is the probability that the outcome occurs,  $n$  is the total number of trials,

$X$  is the variable,  $x$  is the instance of the variable

**Binomial Coefficient:**  $nCx = \binom{n}{x} = \frac{n!}{x!(n-x)!}$

### Probability:

- $P(A)$ : What is the probability of  $A$  happening, Marginal Probability
- $P(A \cap B)$ :  $P(A \text{ and } B) =$   
**If statistically independent:**  $P(A) \cdot P(B)$ , Joint Probability  
**If statistically dependant:**  $P(A) \cdot P(B|A)$
- $P(A \cup B)$ :  $P(A \text{ or } B) =$   
**If mutually exclusive:**  $P(A) + P(B)$ , Disjoint Probability  
**If not mutually exclusive:**  $P(A) + P(B) - P(A \cap B)$
- $P(A|B)$ :  $P(A \text{ given } B) = \frac{P(A \cap B)}{P(B)}$ , Conditional Probability

## Definitions:

**Population:** The complete set of people or objects of interest or the infinite set of all possible values if the same person or item were repeatedly measured in the same way.

**Sample:** The subset of the population for which data can actually be obtained in a study.

**Parameter:** A fixed (but usually unknown) number summarizing some feature of a population.

**Statistic:** A computed number that summarizes sample data in some appropriate way, and which estimates a parameter.

**Inference:** Specifying the estimate of an unknown parameter.

**Mean:** The arithmetic average.

**Median:** The middle value after ordering.

**Standard Deviation:** The typical variation from the mean.

**Variance:** The square of the standard deviation.

**Quartiles:**  $Q_1$  is the 25th percentile,  $Q_3$  is the 75th percentile

**Five-Number Summary:** Minimum,  $Q_1$ , Median,  $Q_3$ , Maximum

**Inter Quartile Range (IQR):** The difference between quartiles.

**The Pearson Correlation Coefficient:** A unitless number that measures the direction and strength between two quantitative variables.

- Between -1 and 1
- The sign (+ or -) matches the direction of the relationship.
- Closer to Zero: *less* linear.
- Closer to 1 or -1: *more* linear.
- Correlation coefficient *exactly* 1 or *exactly* -1 only if all the points fall on a *perfect* line.

### Skewness and Modality:

- Skewed Right: Tail to the Right.
- Skewed Left: Tail to the Left.

- Unimodal: One peak.
- Multimodal: Multiple peaks.

**Statistical Independence:** A and B are statistically independent if and only if  $P(A|B) = P(A|B^C) = P(A)$ ,  $B^C$  is the opposite of B.

## Notation:

- **Uppercase vs. lowercase:** Uppercase denotes a random variable, Lowercase denotes a number.
- **P's:**
  - $\hat{P}$ : Denotes sample proportion, estimates  $p$ .
  - $p$ : Denotes the 'true' population proportion, is a population parameter.
  - $P( )$ : Denotes the "probability of..."
- **X vs.  $\bar{X}$ :**
  - $X$ : Denotes sample *count*, in the case of a *categorical* measurement.
  - $\bar{X}$ : Denotes sample *mean*, in the case of a *quantitative* measurement.