Section 3: DATA DISTRIBUTIONS

Mean, variance and standard deviation

Population: The entire group of subjects that we're interested in

Sample: A sub-section of the population

Population mean:

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

Sample mean:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Variance:

- for a population:

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

- for a sample (biased):

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}$$

- for a sample (unbiased):

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

Standard deviation:

- for a population:

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

- for a sample:

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

Frequency histograms and polygons, and density curves

Histograms, frequency histogram: Shows the frequency at which each category occurs

Relative frequency histogram: The same as a regular histogram, except that we display the frequency of each category as a percentage of the total of the data

Frequency polygon: The shape created by connecting the top of each bar in a frequency histogram

Density curve: The shape created by smoothing out the frequency polygon. The area under the density curve will always represent %100 of the data or 1.0

Symmetric and skewed distributions and outliers

Distribution: The distribution of data across a range of values.

Symmetric distributions: The distribution's mean and median are at the very center of the distribution, with an equal about of data to the left and right.

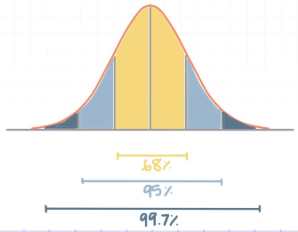
Normal distribution: A symmetric, bell-shaped distribution

Skewed distribution: A non-symmetric distribution that leans right and left. Negatively / left-skewed / left-tailed distributions have their tail on the left, and, from left-to-right, they have their mean, then median, then mode. Positively / right-skewed / right-tailed distributions have their tail on the right, and, from left-to-right, they have their mode, then median, then mean.

1.5-IQR rule: Low outliers are defined as values less than $Q_1 - 1.5(IQR)$, while high outliers are defined as values greater than $Q_3 + 1.5(IQR)$.

Normal distributions and z-scores

Empirical rule, 68-95-99.7 rule: For any normal distribution, there's a 68% chance, 95% chance, and 99.7% chance a data point falls within 1, 2, and 3 standard deviations of the mean, respectively



Percentile: n percent of the values in the data set lie below the n^{th} percentile of the data set

z-score: The z-score for a data point x is the score that tells us the number of standard deviations between x and the mean of μ. A data point is generally considered unusual if its z-score is $z = \pm 3$.

$$z = \frac{x - \mu}{\sigma}$$

Threshold: Some pre-defined cutoff point in the data set

Chebyshev's Theorem

Chebyshev's Theorem: At least $(1 - 1/k^2)\%$ of the data must fall within k standard deviations of the mean, for $k > 1$, regardless of the shape of the data's distribution. Because this theorem applies to distributions of all shapes, it's more conservative than the Empirical Rule.

- At least 75% of the data must be within $k = 2$ standard deviations of the mean.
- At least 89% of the data must be within $k = 3$ standard deviations of the mean.
- At least 94% of the data must be within $k = 4$ standard deviations of the mean.

Covariance

Covariance: How much two random variables vary together. Covariance reflects the directional relationship between two random variables, but not the magnitude of the relationship.

Population covariance

$$\sigma_{xy} = \frac{\sum_i (x_i - \mu_x)(y_i - \mu_y)}{N}$$

Sample covariance

$$s_{xy} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

Direction of covariance:

- **Positive covariance:** Positive linear relationship between the variables
- **Approximately 0 covariance:** No linear relationship between variables
- **Negative covariance:** Negative linear relationship between the variables

Correlation coefficient

Correlation: The degree of the relationship between variables. If two variables are correlated, then a change in one variable results in a change in the other variable.

- **Perfectly correlated:** A specific change in one variable results in an exact and predictable change in the other variable
- **Somewhat correlated:** A change in one variable results in a predictable change in some general direction in the other variable
- **Uncorrelated:** A change in one variable doesn't result in any kind of predictable change in the other variable

Pearson correlation coefficient:

Population

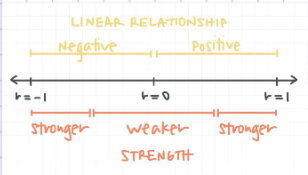
$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \frac{\left(\frac{\sum_i (x_i - \mu_x)(y_i - \mu_y)}{N}\right)}{\sigma_x \sigma_y}$$

Sample

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{\left(\frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{n-1}\right)}{s_x s_y},$$

where $s_x = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$ and $s_y = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n-1}}$

Value of the coefficient: The correlation coefficient indicates both the strength and direction of the relationship between two variables.



- Any value of the correlation coefficient between -0.7 and -1 (or 0.7 and 1) indicates a strong negative (or positive) correlation.
- Any value of the correlation coefficient between -0.3 and -0.7 (or 0.3 and 0.7) indicates a moderate negative (or positive) correlation.
- Any value of the correlation coefficient between 0 and -0.3 (or 0 and 0.3) indicates a weak negative (or positive) correlation.

Weighted means and grouped data

Weighted mean

- Weighted population mean:

$$\mu = \frac{\sum_{i=1}^{N} w_i x_i}{\sum_{i=1}^{N} w_i}$$

- Weighted sample mean:

$$\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

Grouped mean

- Grouped population mean:

$$\mu = \frac{\sum_{i=1}^{N} f_i M_i}{N}$$

- Grouped sample mean:

$$\bar{x} = \frac{\sum_{i=1}^{n} f_i M_i}{n}$$

Grouped variance

- Grouped data population variance:

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

- Grouped data sample variance:

$$s^2 = \frac{\sum_{i=1}^{n} f_i (M_i - \bar{x})^2}{n-1}$$