Formula and Concept Sheet

STAT 107: Data Science Discovery
Karle Flanagan and Wade Fagen-Ulmschneider

Box Plots

Quartiles equally divides data into four segments. The boundaries between the quartiles are labeled as $\mathbf{Q_1}$, $\mathbf{Q_2}$, and $\mathbf{Q_3}$.

Inner-Quartile Range (IQR) = Q_3 - Q_1
IQR tells us the width of the middle 50%, which is exactly the size of the box in a box plot.

Whiskers extend out from the box plot to include all data points smaller than $\mathbf{Q_1}$ or larger than $\mathbf{Q_3}$ and are **not** outliers.

Outliers are extreme values far from the median, drawn as individual points outside of the "box" and "whiskers":
High Outliers > Q_3 + (1.5 × IQR)
Low Outliers < Q_1 - (1.5 × IQR)

Probability Rules

P(A) is the **probability** of event **A** occurring.
P(A | B) is the **probability** of event A occurring **given** event B occurred.

Multiplication ("AND") Rule:
P(A **and** B) = P(**A**) × P(**B** | **A**)

Addition ("OR") Rule:
P(A **or** B) = P(**A**) + P(**B**) - P(**A** and **B**)

Negation and DeMorgan's Law:
P(**A**) = 1 - P(**not A**)
P(**"at least one"**) = **1 -** P(**"none"**)
P("**not all**") = **1 -** P(**"all"**)

Conditional Probability:
P(**B** | **A**) = P(**A** and **B**) / P(**A**)

Bayes' Rule:
P(**A | B**) = P(**B | A**) × (P(**A**) / P(**B**))

Common Symbols/Notations:
"AND": intersection, ∩, or ∧
"OR": union, ∪, or ∨
"NOT": complement, A^C, $\overline{A}$, or $\neg A$

Random Variables

Z-Score for random variables: $Z = (Value - EV) / SE$

Discrete Random Variables:
Expected Value: $E(X) = X_1P_1 + \ldots + X_nP_n$

Standard Error: $\sigma_X = \sqrt{(X_1 - \mu_X)^2P_1 + \ldots + (X_n - \mu_X)^2P_n}$

Binomal Random Variables:
Expected Value= n*p

Standard Error= $\sqrt{np(1-p)}$

Means:

Expected Value: $EV_{avg} = E(\overline{X}) = \mu$

Standard Error: $SE_{avg} = SD / \sqrt{n}$

Percents:

Expected Value: EV%=population percent=p

Standard Error: $SE\% = (SD / \sqrt{n}) * 100\%$

SD for a population with 0s and 1s (yes and no): $\sqrt{p(1-p)}$

Regression

Equation of the regression line: $\hat{y} = b_0 + b_1 * x_1$

Slope of the regression line= r*SDy/SDx

Y Intercept of the regression line= average of y - slope * average of x

$RMSE = SD_{errors} = \sqrt{1 - r^2} \times SD_y$

residual= actual- predicted

Sampling

General z-score formula: z = (value-average)/SD

Quick formula for choosing how many people to poll:

$n=(100\%*z*SD/M_{of}E)^2$