EXAM CHEAT SHEET (PART I)

I. DESCRIPTIVE STATISTICS

Categorical Data: Nominal or Ordinal

Numerical Data: Continuous or Discrete

Quartile: Q1 (25%), Q2 (50%), Q3 (75%)

Percentile: Location of p-th percentile:

$$L_p = (n+1)\frac{p}{100}$$

Inter-Quartile Range (IQR):

$$IQR = Q_3 - Q_1$$

Population Mean: $\mu = \frac{1}{N}\sum_{i=1}^{N} X_i$

Sample Mean: $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

Population Variance:

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

Sample Variance:

$$\begin{aligned} s^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2 \\ &= \frac{1}{n-1}\left(\left(\sum_{i=1}^{n} X_i^2\right) - \frac{\left(\sum_{i=1}^{n} X_i\right)^2}{n}\right) \end{aligned}$$

Coefficient of Variance: $CV = \frac{\sigma}{\mu}$, or $cv = \frac{s}{\bar{X}}$

Population Covariance:

$$\begin{aligned} \sigma_{XY} &= \frac{1}{N}\sum_{i=1}^{N}(X_i-\mu_X)(Y_i-\mu_Y) \\ &= E(XY) - E(X)E(Y) \end{aligned}$$

Sample Covariance:

$$\begin{aligned} s_{XY} &= \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y}) \\ &= \frac{1}{n-1}\left(\left(\sum_{i=1}^{n} X_iY_i\right) - \frac{\left(\sum_{i=1}^{n} X_i\right)\left(\sum_{i=1}^{n} Y_i\right)}{n}\right) \end{aligned}$$

Population Correlation Coefficient: $\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X\sigma_Y}$

Sample Correlation Coefficient: $r_{XY} = \frac{s_{XY}}{s_X s_Y}$

Note: $-1 \le \rho_{XY},\ r_{XY} \le 1$

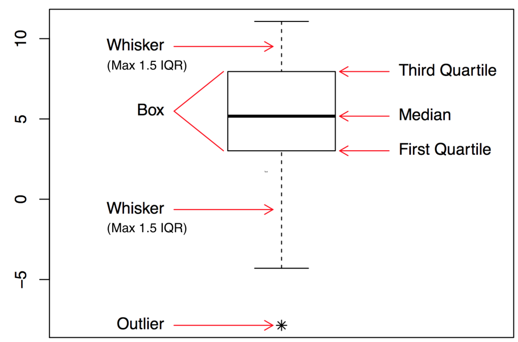
Skewness:

Zero Skewness: Symmetric (mean = median)

Positively Skewed: Long tail to the *right*

Negatively Skewed: Long tail to the *left*

Boxplots:



II. PROBABILITY

Mutually Exclusive $\Leftrightarrow P(A \cap B) = 0$

Independent $\Leftrightarrow P(A \cap B) = P(A) \cdot P(B)$

Law of Total Probability:

$$P(A) = \sum_{i=1}^{n} P(A \cap B_i)$$

where $B_1, B_2, \ldots, B_n$ are mutually exclusive and $B_1 \cup B_2 \cup \cdots \cup B_n = S$ (exhaustive).

Multiplication Rule:

$$P(A \cap B) = P(A|B) \cdot P(B) = P(B|A) \cdot P(A)$$
$$P(A \cap B) = P(A) \cdot P(B), \text{ if A,B independent}$$

Addition Rule:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$
$$P(A \cup B) = P(A) + P(B), \text{if mutually exclusive}$$

Complement Rule: $P(A^C) = 1 - P(A)$

III. DISCRETE PROBABILITY DISTRIBUTION

Random Variable: X, Y, Z

Realised Variable: x, y, z

Expected Value: $\mu = E(X) = \sum_{all\ x}(x \cdot p(x))$

Variance: $\sigma^2 = V(X) = E(X^2) - E^2(X)$

Joint Probability: $p(x, y) = P(\{X = x\} \cap \{Y = y\})$

Marginal Probability: $p_X(x) = \sum_{all\ y} p(x, y)$

If X, Y independent:

$$p(x, y) = p_X(x)p_Y(y), \text{ for all } x, y$$

$E(c) = c$, $E(cX) = cE(X)$

If X, Y independent: $E(XY) = E(X)E(Y)$

$V(c) = 0$, $V(X + c) = V(X)$, $V(cX) = c^2V(X)$

$E(aX + bY) = aE(X) + bE(Y)$

$V(aX + bY) = a^2V(X) + b^2V(Y) + 2abCov(X, Y)$

Covariance: $Cov(X, Y) = E(XY) - E(X)E(Y)$

Independent $\Rightarrow Cov(X, Y) = 0$

Binomial Distribution: $X \sim Bin(n, p)$

$P(X = x) = \binom{n}{x} p^x (1 - p)^{n-x}$

$E(X) = np$, $V(X) = np(1 - p)$

Binomial Table: Value is $P(X \leq k)$

IV. CONTINUOUS PROBABILITY DISTRIBUTION

Probability Density Function (PDF): $f(x)$

$P(a < X < b) = \int_a^b f(x)dx$

Expected Value:

$$\mu = E(X) = \int_{-\infty}^{\infty} xf(x)dx$$

Variance:

$$\sigma^2 = V(X) = \int_{-\infty}^{\infty} (x - \mu)^2 f(x)dx = \left(\int_{-\infty}^{\infty} x^2 f(x)dx\right) - \mu^2$$

Uniform Distribution: $X \sim U(a, b)$

$f(x) = \frac{1}{b - a}$, $a \leq x \leq b$

$E(X) = \frac{a + b}{2}$, $V(X) = \frac{(b - a)^2}{12}$

Normal Distribution: $X \sim N(\mu, \sigma^2)$

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \ -\infty < x < \infty$$

$E(X) = \mu$, $V(X) = \sigma^2$

Standard Normal Distribution:

$$Z = \frac{X - \mu}{\sigma} \sim N(0, 1)$$

z-table: Value is $P(Z < z)$

z-values frequently used:
$|z_{0.1}| = 1.282$, $|z_{0.05}| = 1.645$, $|z_{0.025}| = 1.96$, $|z_{0.01}| = 2.327$, $|z_{0.005}| = 2.576$, $|z_{0.001}| = 3.091$

V. SAMPLING DISTRIBUTION

Mean of $\bar{X}$: $\mu_{\bar{X}} = \mu$

Variance of $\bar{X}$ (Standard Error): $\sigma^2_{\bar{X}} = \frac{\sigma^2}{n}$

Central Limit Theorem:

$$\bar{X} \sim N(\mu_{\bar{X}} = \mu, \sigma^2_{\bar{X}} = \frac{\sigma^2}{n}), \text{ as } x \to \infty$$

Sample Proportion (of Bernoulli trials): $\hat{p} = \frac{X}{n}$

$$\hat{p} \sim N\left(\mu_{\hat{p}} = p, \sigma^2_{\hat{p}} = \frac{p(1 - p)}{n}\right)$$

VI. ESTIMATION

Point Estimator & Interval Estimator

Bias: $B(\hat{\theta}) = E(\hat{\theta}) - \theta$. Unbiased if $B(\hat{\theta}) = 0$

Mean Squared Error(MSE):

$$MSE(\hat{\theta}) = E((\hat{\theta} - \theta)^2) = V(\hat{\theta}) + B^2(\hat{\theta})$$

$\hat{\theta}$ is consistent if $MSE(\hat{\theta}) \to 0$ as $n \to \infty$

Relative Efficiency: $\text{eff}(\hat{\theta}_1, \hat{\theta}_2) = \frac{V(\hat{\theta}_2)}{V(\hat{\theta}_1)}$

$\hat{\theta}_1$ is better if eff > 1; $\hat{\theta}_2$ is better if eff < 1

Confidence Interval of $100(1 - \alpha)\%$: $\bar{X} \pm z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$

Lower confidence limit; Upper confidence limit; Cofidence level

Interpretation: In repeated sampling, $100(1 - \alpha)\%$ of such intervals created would contain the true population mean.