STOR 155 - Formula sheet

Panagiotis Andreou

April 21, 2025

Numerical Data

Centrality & Dispersion Measures

Centrality Measures:

- **Mean:** $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
- **Median:** Middle value when data is ordered
- **Mode:** Most frequent value

Dispersion Measures:

- **Range:** $R = \max - \min$
- **Variance (sample):** $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$
- **Standard deviation (sample):** $s = \sqrt{s^2}$

Probability

Axioms of Probability

Let P be a probability function on sample space S:

- $P(A) \geq 0$
- $P(S) = 1$
- If $A \cap B = \emptyset$, then $P(A \cup B) = P(A) + P(B)$

Basic Rules

$$\begin{aligned} P(A^c) &= 1 - P(A) \\ P(A \cup B) &= P(A) + P(B) - P(A \cap B) \\ P(A \cap B) &= P(A)P(B|A) \\ P(A|B) &= \frac{P(A \cap B)}{P(B)} \end{aligned}$$

Independence

Events A and B are independent if:

$$P(A \cap B) = P(A)P(B) \quad \text{or} \quad P(A|B) = P(A)$$

Bayes' Theorem

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^c)P(A^c)}$$

Bernoulli Distribution

$X \sim \text{Bernoulli}(p)$

$$\begin{aligned} P(X = k) &= p^k(1-p)^{1-k}, \quad k = 0, 1 \\ \mathbb{E}[X] &= p \\ \text{Var}(X) &= p(1-p) \end{aligned}$$

Binomial Distribution

$X \sim \text{Bin}(n, p)$

$$\begin{aligned} P(X = k) &= \binom{n}{k} p^k(1-p)^{n-k} \\ \mathbb{E}[X] &= np \\ \text{Var}(X) &= np(1-p) \end{aligned}$$

Geometric Distribution

$X \sim \text{Geom}(p)$

$$\begin{aligned} P(X = k) &= (1-p)^{k-1}p, \quad k = 1, 2, \dots \\ \mathbb{E}[X] &= \frac{1}{p} \\ \text{Var}(X) &= \frac{1-p}{p^2} \end{aligned}$$

Statistics

Point Estimates

- $\hat{p}$ estimates p (proportion)
- $\bar{x}$ estimates μ (mean)

Confidence Intervals

General form:

$$\text{point estimate} \pm \text{critical value} \cdot \text{SE}$$

- Proportion:
  $\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
- Mean (unknown σ):
  $\bar{x} \pm t^* \frac{s}{\sqrt{n}}$
- Difference of Means (unpaired):
  $(\bar{x}_1 - \bar{x}_2) \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

Hypothesis Testing

Test statistic:

$$z \text{ or } t = \frac{\text{point estimate} - \text{null value}}{\text{SE}}$$

- Proportion:

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

- Mean:

$$t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$$

- Difference of proportions:

$$\hat{p}_{\text{pool}} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$$

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}_{\text{pool}}(1-\hat{p}_{\text{pool}})}{n_1} + \frac{\hat{p}_{\text{pool}}(1-\hat{p}_{\text{pool}})}{n_2}}}$$

- Difference of means:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

p-values

$$\begin{cases} 2P(\text{stat} \geq |obs|) & \text{(two-tailed)} \\ P(\text{stat} \leq obs) & \text{(left-tailed)} \\ P(\text{stat} \geq obs) & \text{(right-tailed)} \end{cases}$$

Linear Regression

Simple Linear Regression

$$y_i = b_0 + b_1 x_i + e_i$$

- b_0: intercept
- b_1: slope
- e_i: residual $= y_i - \hat{y}_i$
- Goal: Minimize $\sum e_i^2$

Slope and Intercept Estimates

$$\hat{b}_1 = r \cdot \frac{s_y}{s_x} \qquad \hat{b}_0 = \bar{y} - \hat{b}_1 \bar{x}$$

Prediction

For new x:

$$\hat{y} = \hat{b}_0 + \hat{b}_1 x$$

Correlation Coefficient

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Coefficient of Determination

$$R^2 = r^2$$

Interpretation: Proportion of variance in y explained by the model.