STOR 155 - Comprehensive Review

Panagiotis Andreou

April 21, 2025

1 Learning Objectives

- Understand the basics of numeric data and recall measures of centrality and dispersion.
- Recall the basic axioms and rules of probability.
- Apply Bayes' Theorem to conditional probability problems.
- Work with random variables: build probability models, compute expectations and variances.
- Apply the Normal distribution: calculate z-scores, percentiles, and probabilities.
- Understand and apply the Geometric and Binomial distributions.
- Understand the concept and purpose of point estimates in statistical inference.
- Construct and interpret confidence intervals for proportions and means.
- Perform hypothesis tests for proportions and means, including formulating null and alternative hypotheses.
- Differentiate between one-tailed and two-tailed tests and compute the appropriate p-values.
- Apply the correct test statistic (z or t) depending on the type of hypothesis test we have.
- Conduct tests comparing two population parameters (e.g., difference of proportions or means).
- Understand the assumptions behind each statistical test and verify conditions for valid inference.
- Interpret the results of hypothesis tests and make conclusions in context.
- Understand the concept of linear regression and the relationship between two quantitative variables.
- Interpret the slope, intercept, and R^2 value in a linear model, and use the model for prediction.
- Use software output to conduct hypothesis test for the validity of adopting a linear model for our data.

2 Numerical Data

Setting: we have n data points, $x_1, \ldots, x_n$.

Centrality Measures

1. *Mean (Average)*:
$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

2. *Median*: The middle value when the data is sorted.
   - If n is odd: the median is the middle number.
   - If n is even: the median is the average of the two middle numbers.

3. *Mode*: The value that appears most frequently in the dataset.

Dispersion Measures

1. *Range*:
$$R = \max_{1 \le i \le n} x_i - \min_{1 \le i \le n} x_i$$

2. *Variance (Sample)*:
$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

3. *Standard Deviation*:
$$s = \sqrt{s^2} = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2}$$

3 Probability

Axioms of Probability

Let S be the sample space. A probability function P satisfies:

1. $P(A) \geq 0$ for any event A
2. $P(S) = 1$
3. If A and B are disjoint, then $P(A \cup B) = P(A) + P(B)$

Basic Probability Rules

- *Complement Rule:* $P(A^c) = 1 - P(A)$
- *Union Law:* $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- *Conditional Probability:* $P(A|B) = \frac{P(A \cap B)}{P(B)}$ (if $P(B) > 0$)

Independence

Events A and B are *independent* if

$$P(A \cap B) = P(A)P(B)$$

or, equivalently,

$$P(A|B) = P(A)$$

Bayes' Theorem

Used to reverse conditional probabilities when direct computation is hard.

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

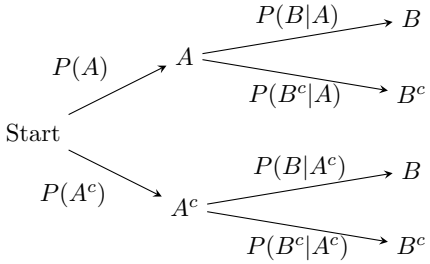


Diagram for Bayes' Theorem. Follow paths and multiply.

Discrete Distributions

The three main discrete distributions we studied are:

- Bernoulli

- Binomial
- Geometric

Bernoulli Distribution:

- Models a single trial with outcome success (1) or failure (0)
- $X \sim \text{Bernoulli}(p)$
- PMF: $P(X = k) = p^k(1-p)^{1-k}$ for $k \in \{0, 1\}$
- $\mathbb{E}[X] = p \quad , \quad \text{Var}(X) = p(1-p)$

Binomial Distribution:

- Models number of successes in n independent Bernoulli(p) trials
- $X \sim \text{Bin}(n, p)$
- PMF: $P(X = k) = \binom{n}{k} p^k(1-p)^{n-k}$ for $k = 0, 1, ..., n$
- $\mathbb{E}[X] = np \quad , \quad \text{Var}(X) = np(1-p)$

Geometric Distribution:

- Models number of trials until the first success (memoryless)
- $X \sim \text{Geom}(p)$
- PMF: $P(X = k) = (1-p)^{k-1}p$ for $k = 1, 2, 3, \ldots$
- $\mathbb{E}[X] = \frac{1}{p} \quad , \quad \text{Var}(X) = \frac{1-p}{p^2}$

4 Statistics

The three main components of statistical inference we studied are:

- Point Estimates
- Confidence Intervals
- Hypothesis Tests

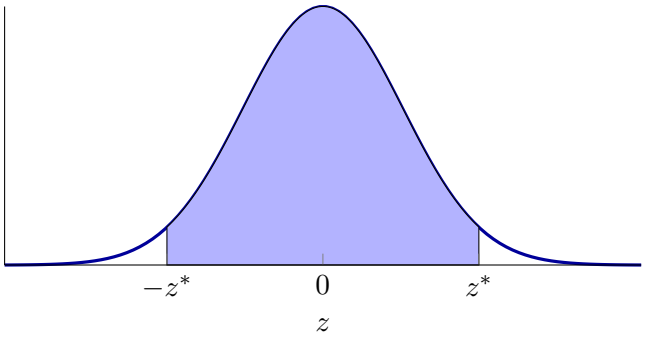
Point estimates are sample-based approximations of population parameters:

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

Confidence Intervals

The general form of a confidence interval is:

$$\text{point estimate} \pm \text{critical value} \cdot \text{standard error}$$



The percentile z^* is what we call the 'critical value'

1. *Proportion:*

$$\hat{p} \pm z_{1-\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

2. *Mean (unknown population variance):*

$$\bar{x} \pm t_{n-1;1-\alpha/2} \cdot \frac{s}{\sqrt{n}}$$

3. *Difference of means (unpaired):*

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\min(n_1-1,n_2-1);1-\alpha/2} \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Hypothesis Testing

The general form of the test statistic is:

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

- *Proportion:*

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

- *Mean (unknown σ):*

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

- *Difference of proportions:*

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

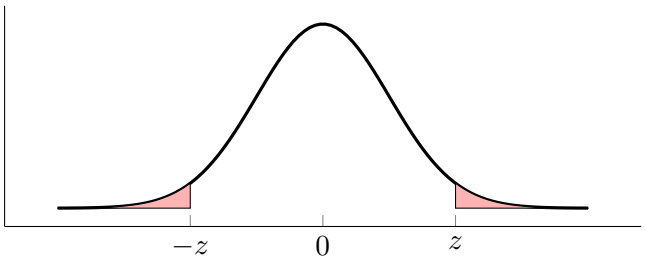
$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}_{\text{pooled}}(1-\hat{p}_{\text{pooled}})}{n_1} + \frac{\hat{p}_{\text{pooled}}(1-\hat{p}_{\text{pooled}})}{n_2}}}$$

- *Difference of means (unpaired):*

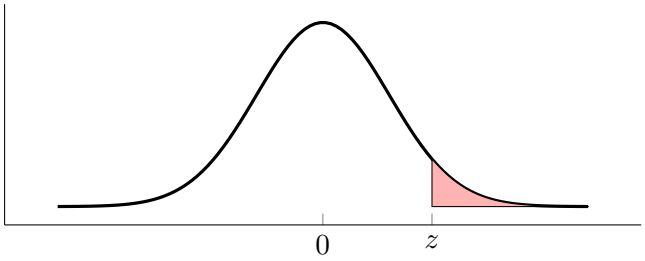
$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

p-values: based on the test statistic and the direction of H_A:

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



Two-tailed hypothesis test rejection regions.

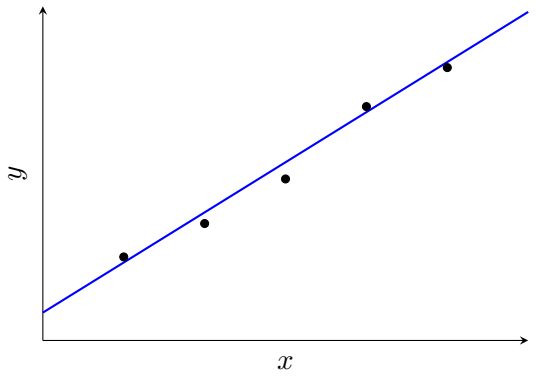


Right-tailed hypothesis test rejection region.

5 Linear Regression

We use linear regression to study the relationship between two numerical variables and to make predictions.

- Given data: $(x_1, y_1), \ldots, (x_n, y_n)$
- Model:
$$y_i = b_0 + b_1 x_i + e_i, \quad i = 1, \ldots, n$$
- Goal: Minimize the residual sum of squares: $\sum e_i^2$
- **Correlation coefficient (Pearson's r):**
$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$
- **Estimated slope and intercept:**
$$\hat{b}_1 = r \cdot \frac{s_y}{s_x}, \quad \hat{b}_0 = \bar{y} - \hat{b}_1 \bar{x}$$
- **Prediction:**
$$\hat{y} = \hat{b}_0 + \hat{b}_1 x$$
- **Coefficient of determination (R^2):**
$$R^2 = r^2 = \text{Proportion of variance explained by the model}$$



Fitted line $\hat{y} = b_0 + b_1 x$ through data points.