Statistics Summary

Mathematical Notes

October 19, 2025

Contents

1 Descriptive Statistics

1.1 Measures of Central Tendency

Definition 1.1. For a sample $x_1, x_2, \ldots, x_n$:

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

1.2 Measures of Dispersion

Definition 1.2.
- **Variance**: $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$
- **Standard Deviation**: $s = \sqrt{s^2}$
- **Range**: $\max(x_i) - \min(x_i)$
- **Interquartile Range**: $Q_3 - Q_1$

2 Parameter Estimation

2.1 Point Estimation

Definition 2.1. A **point estimator** $\hat{\theta}$ is a statistic used to estimate a population parameter θ.

2.2 Properties of Estimators

Definition 2.2. An estimator $\hat{\theta}$ is:

- **Unbiased** if $E[\hat{\theta}] = \theta$
- **Consistent** if $\hat{\theta} \xrightarrow{p} \theta$ as $n \to \infty$
- **Efficient** if it has minimum variance among unbiased estimators

2.3 Maximum Likelihood Estimation

Definition 2.3. The **maximum likelihood estimator** (MLE) is the value of θ that maximizes the likelihood function $L(\theta) = \prod_{i=1}^{n} f(x_i|\theta)$.

2.4 Method of Moments

Definition 2.4. The **method of moments** estimator equates sample moments to population moments:

$$\frac{1}{n}\sum_{i=1}^{n} X_i^k = E[X^k]$$

3 Confidence Intervals

3.1 Definition

Definition 3.1. A **confidence interval** for parameter θ is an interval $[L, U]$ such that $P(L \leq \theta \leq U) = 1 - \alpha$.

3.2 Common Confidence Intervals

3.2.1 Normal Mean (Known Variance)

For $X \sim \mathcal{N}(\mu, \sigma^2)$ with known σ^2:

$$\bar{X} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

3.2.2 Normal Mean (Unknown Variance)

For $X \sim \mathcal{N}(\mu, \sigma^2)$ with unknown σ^2:

$$\bar{X} \pm t_{\alpha/2, n-1} \frac{s}{\sqrt{n}}$$

3.2.3 Proportion

For binomial proportion p:

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

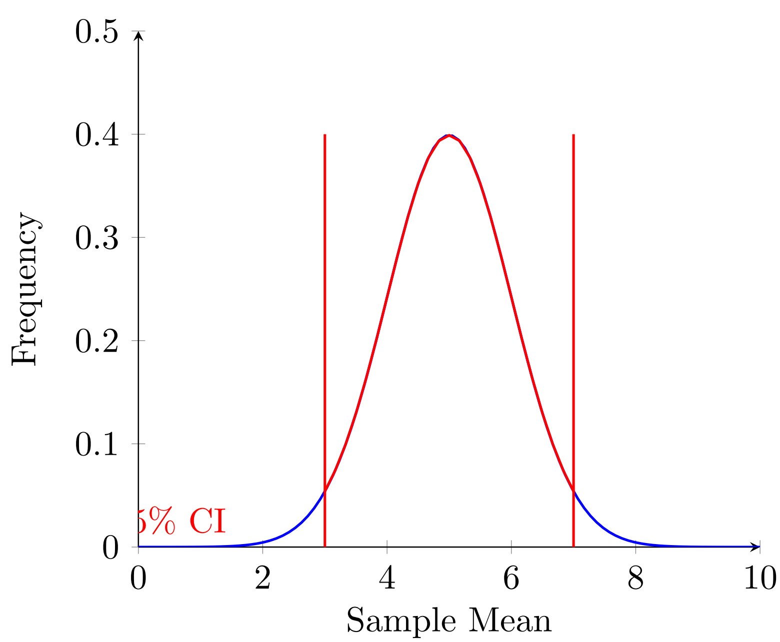

Figure 1: Confidence interval for normal distribution

4 Hypothesis Testing

4.1 Basic Concepts

Definition 4.1. A **hypothesis test** is a procedure for deciding between two competing hypotheses:

- H_0: Null hypothesis
- H_1: Alternative hypothesis

4.2 Types of Errors

Definition 4.2.

- **Type I Error**: Reject H_0 when it's true (probability α)
- **Type II Error**: Fail to reject H_0 when it's false (probability β)
- **Power**: $1 - \beta = P(\text{reject } H_0 | H_1 \text{ true})$

4.3 Test Statistics

4.3.1 Z-Test

For testing mean with known variance:

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1)$$

4.3.2 t-Test

For testing mean with unknown variance:

$$t = \frac{\bar{X} - \mu_0}{s/\sqrt{n}} \sim t_{n-1}$$

4.3.3 Chi-Square Test

For testing variance:

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2} \sim \chi^2_{n-1}$$

4.4 p-Values

Definition 4.3. The **p-value** is the probability of observing a test statistic as extreme or more extreme than the observed value, assuming H_0 is true.

5 Regression Analysis

5.1 Simple Linear Regression

Definition 5.1. The simple linear regression model is:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$.

5.2 Least Squares Estimation

The least squares estimators are:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

5.3 Coefficient of Determination

Definition 5.2. The **coefficient of determination** is:

$$R^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}}$$

where SSR is sum of squares due to regression, SSE is sum of squared errors, and SST is total sum of squares.

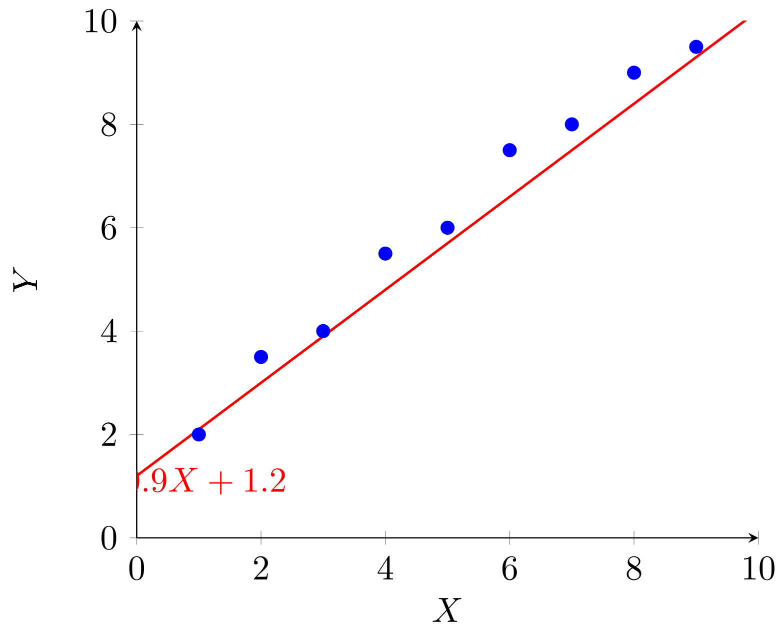

Figure 2: Simple linear regression

5.4 Multiple Linear Regression

Definition 5.3. The multiple linear regression model is:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + \epsilon_i$$

5.5 ANOVA

Definition 5.4. **Analysis of Variance** (ANOVA) tests whether the means of several groups are equal:

$$F = \frac{\text{MSB}}{\text{MSW}} \sim F_{k-1,n-k}$$

where MSB is mean square between groups and MSW is mean square within groups.

6 Bayesian Inference

6.1 Bayes' Theorem

Theorem 6.1.

$$P(\theta|data) = \frac{P(data|\theta)P(\theta)}{P(data)} = \frac{L(\theta)\pi(\theta)}{\int L(\theta)\pi(\theta)d\theta}$$

where $\pi(\theta)$ is the prior distribution and $P(\theta|data)$ is the posterior distribution.

6.2 Prior Distributions

Definition 6.1. Common conjugate priors:

- Normal-Normal: $X|\mu \sim \mathcal{N}(\mu, \sigma^2)$, $\mu \sim \mathcal{N}(\mu_0, \tau^2)$
- Beta-Binomial: $X|p \sim \text{Binomial}(n, p)$, $p \sim \text{Beta}(\alpha, \beta)$
- Gamma-Poisson: $X|\lambda \sim \text{Poisson}(\lambda)$, $\lambda \sim \text{Gamma}(\alpha, \beta)$

6.3 Bayesian Estimation

Definition 6.2. Bayesian point estimators:

- **Posterior Mean**: $E[\theta|data]$
- **Posterior Median**: Median of posterior distribution
- **Maximum A Posteriori (MAP)**: Mode of posterior distribution

7 Nonparametric Methods

7.1 Goodness of Fit Tests

7.1.1 Kolmogorov-Smirnov Test

Definition 7.1. Tests whether a sample comes from a specified distribution:

$$D_n = \sup_x |F_n(x) - F_0(x)|$$

where F_n is the empirical CDF and F_0 is the hypothesized CDF.

7.1.2 Chi-Square Goodness of Fit

Definition 7.2.

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_{k-1}$$

where O_i are observed frequencies and E_i are expected frequencies.

7.2 Rank Tests

7.2.1 Wilcoxon Rank-Sum Test

Tests whether two independent samples come from the same distribution.

7.2.2 Mann-Whitney U Test

Nonparametric alternative to the two-sample t-test.

8 Time Series Analysis

8.1 Stationarity

Definition 8.1. A time series is **stationary** if:

- $E[X_t] = \mu$ (constant mean)
- $\text{Var}(X_t) = \sigma^2$ (constant variance)
- $\text{Cov}(X_t, X_{t+k}) = \gamma(k)$ (covariance depends only on lag)

8.2 ARIMA Models

Definition 8.2. An **ARIMA**(p,d,q) model is:

$$\phi(B)(1-B)^d X_t = \theta(B)\epsilon_t$$

where B is the backshift operator, $\phi(B)$ is the AR polynomial, and $\theta(B)$ is the MA polynomial.

9 Design of Experiments

9.1 Randomized Controlled Trials

Definition 9.1. A **randomized controlled trial** randomly assigns subjects to treatment and control groups to minimize bias.

9.2 Factorial Designs

Definition 9.2. A **factorial design** studies the effect of multiple factors simultaneously:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}$$

9.3 Blocking

Definition 9.3. Blocking groups similar experimental units together to reduce variability and increase precision.

10 Multivariate Statistics

10.1 Multivariate Normal Distribution

Definition 10.1. A random vector $\mathbf{X} = (X_1, \ldots, X_p)^T$ has a multivariate normal distribution if:

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\mathbf{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

10.2 Principal Component Analysis

Definition 10.2. Principal Component Analysis (PCA) finds linear combinations of variables that explain maximum variance:

$$\mathbf{Y} = \mathbf{A}\mathbf{X}$$

where $\mathbf{A}$ is chosen to maximize variance of $\mathbf{Y}$.

10.3 Canonical Correlation

Definition 10.3. Canonical correlation finds linear combinations of two sets of variables that are maximally correlated.

11 Applications

11.1 Clinical Trials

Statistics is essential for:

- Sample size determination
- Randomization procedures
- Interim analyses
- Safety monitoring

11.2 Quality Control

Applications include:

- Control charts
- Process capability analysis
- Design of experiments
- Reliability analysis

11.3 Survey Sampling

Used in:

- Population estimation
- Stratified sampling
- Cluster sampling
- Nonresponse adjustment

12 Important Theorems

12.1 Central Limit Theorem

Theorem 12.1. If $X_1, X_2, \dots$ are i.i.d. with mean μ and variance σ^2, then:

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \xrightarrow{d} \mathcal{N}(0,1)$$

12.2 Slutsky's Theorem

Theorem 12.2. If $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{p} c$, then:

- $X_n + Y_n \xrightarrow{d} X + c$
- $X_n Y_n \xrightarrow{d} cX$
- $X_n / Y_n \xrightarrow{d} X/c$ (if $c \neq 0$)

12.3 Delta Method

Theorem 12.3. If $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$ and g is differentiable at θ, then:

$$\sqrt{n}(g(\hat{\theta}_n) - g(\theta)) \xrightarrow{d} \mathcal{N}(0, [g'(\theta)]^2 \sigma^2)$$