Review of Probability and Statistics

Instructor: Aneesh Arora

Notes compiled from course materials/online resources/statistics books

---

Primary Textbook: James H. Stock and Mark W. Watson, *Introduction to Econometrics*, 4th Edition, Pearson.

Contents

Part I
Fundamentals of Probability

1 Chapter 1: Basic Concepts of Probability

This section introduces the foundational concepts of probability theory, which are essential for understanding random phenomena.

1.1 Experiments, Outcomes, and Events

- **Random Experiment:** A repeatable procedure that has a well-defined set of possible results. The outcome of the experiment is not known in advance.
- **Outcomes:** The mutually exclusive potential results of a random experiment.
- **Sample Space ($\mathcal{S}$):** The set of all possible outcomes of a random experiment.
- **Event ($\mathcal{E}$):** A subset of the sample space ($\mathcal{E} \subseteq \mathcal{S}$). An event is a collection of one or more outcomes.

1.2 Set Notation

- **Set:** A collection of objects, which are called elements of the set.
- **Union ($A \cup B$):** The set of all elements that are in either set A or set B (or both).
- **Intersection ($A \cap B$):** The set of all elements that belong to both set A and set B.
- **Empty Set ($\emptyset$):** A set with no elements.
- **Disjoint Sets:** Two sets are disjoint if their intersection is the empty set.
- **Subset ($A \subseteq B$):** A is a subset of B if every element of A is also an element of B.

1.3 Probability

Probability is a way of quantifying the likelihood that an event will occur.

- **Definition:** A probability is a mapping from all subsets of the sample space $\mathcal{S}$ to the interval $[0, 1]$. It can be interpreted as the long-run frequency of an outcome occurring in repeated experiments.
- **Axioms of Probability:** A probability measure $Pr(\cdot)$ must satisfy these properties:
  1. The probability of any event $\mathcal{E}$ is non-negative and no greater than 1: $0 \le Pr(\mathcal{E}) \le 1$.
  2. The probability of the entire sample space is 1: $Pr(\mathcal{S}) = 1$.
  3. For any set of mutually exclusive (disjoint) events $\mathcal{E}_1, \mathcal{E}_2, \dots$, the probability of their union is the sum of their individual probabilities: $Pr(\mathcal{E}_1 \cup \mathcal{E}_2 \cup \dots) = Pr(\mathcal{E}_1) + Pr(\mathcal{E}_2) + \dots$.

2 Chapter 2: Random Variables

A **random variable** is a function that assigns a real number to each outcome in the sample space $\mathcal{S}$.

2.1 Discrete Random Variables

A random variable is **discrete** if it can take on a countable number of distinct values, such as integers.

- **Example:** Rolling a die, where the sample space is $\mathcal{S} = \{1, 2, 3, 4, 5, 6\}$.
- **Probability Mass Function (PMF):** For a discrete random variable X, the PMF, denoted $f(x)$ or $p_X(x)$, gives the probability that X is exactly equal to some value x.

$$p = f(x) = Pr(X = x)$$

- **Cumulative Distribution Function (CDF):** The CDF, denoted $F(x)$, gives the probability that the random variable X is less than or equal to a particular value x.

$$F(x) = Pr(X \leq x) = \sum_{x_i \leq x} f(x_i)$$

2.2 Continuous Random Variables

A random variable is **continuous** if it can take on an uncountably infinite number of possible values within a given range.

- **Examples:** Measurements of length, weight, or temperature.
- **Probability Density Function (PDF):** For a continuous random variable, the distribution is described by a PDF, $f(x)$. The probability of X falling within an interval $[a, b]$ is the area under the PDF curve over that interval.

$$Pr(a \leq X \leq b) = \int_a^b f(x)dx$$

- **Cumulative Distribution Function (CDF):** The CDF for a continuous variable is the integral of the PDF from $-\infty$ to x. It represents the total area under the curve to the left of x.

$$F(x) = Pr(X \leq x) = \int_{-\infty}^{x} f(t)dt$$

  The PDF is the derivative of the CDF: $f(x) = \frac{dF(x)}{dx}$.

3 Chapter 3: Moments of a Distribution

Moments are a set of statistical parameters used to measure the shape of a probability distribution.

3.1 Expected Value (Mean)

The **expected value**, or mean, of a random variable is its probability-weighted average value, often denoted as $E[X]$ or μ_X.

- **For a discrete random variable:**

$$E[X] = \mu_X = \sum_{i=1}^{n} x_i f(x_i)$$

- **For a continuous random variable:**

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

The **expectation of a function** of a random variable, $g(X)$, is calculated as:

- Discrete: $E[g(X)] = \sum_{i=1}^{n} g(x_i)f(x_i)$
- Continuous: $E[g(X)] = \int_{-\infty}^{\infty} g(x)f(x)dx$

A key property of expectations is **linearity**: For constants a and b, and random variables X and Y:

$$E[aX + bY] = aE[X] + bE[Y]$$

3.2 Variance and Standard Deviation

The **variance** measures the spread or dispersion of a distribution around its mean. It is the expected value of the squared deviation from the mean.

$$var(X) = \sigma_X^2 = E[(X - \mu_X)^2]$$

A useful computational formula is:

$$var(X) = E[X^2] - (E[X])^2$$

The **standard deviation**, σ_X, is the square root of the variance. It is measured in the same units as the random variable, making it easier to interpret than the variance.

$$\sigma(X) = \sqrt{var(X)}$$

Properties of Variance: For constants a and b:

- $var(a) = 0$
- $var(a + bX) = b^2 var(X)$

3.3 Higher-Order Moments

- **k-th moment:** The expected value of X^k, which is $E[X^k]$. The first moment is the mean.
- **k-th central moment:** The expected value of the k-th power of the deviation from the mean, $E[(X - \mu_X)^k]$. The second central moment is the variance.

4 Chapter 4: Bivariate Distributions

This section explores the relationships between two random variables.

4.1 Joint, Marginal, and Conditional Distributions

Let X and Y be two random variables.

- **Joint Distribution:** Describes the probability of X and Y occurring simultaneously.
  - **Discrete (Joint PMF):** $f_{X,Y}(x, y) = Pr(X = x, Y = y)$.
  - **Continuous (Joint PDF):** The probability over a region is given by a double integral of the joint PDF, $f_{X,Y}(x, y)$.

- **Joint CDF:** $F_{X,Y}(x,y) = Pr(X \leq x, Y \leq y)$.

- **Marginal Distribution:** The probability distribution of a single variable, irrespective of the other.
  - **Discrete (Marginal PMF):** $f_X(x) = \sum_y Pr(X = x, Y = y)$.
  - **Continuous (Marginal PDF):** $f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dy$.

- **Conditional Distribution:** The distribution of one variable given that the other variable has taken a specific value.
$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)}$$

4.2 Independence

Two random variables X and Y are **independent** if knowing the value of one provides no information about the value of the other. This holds if and only if their joint distribution is the product of their marginal distributions.

$$f_{X,Y}(x,y) = f_X(x)f_Y(y)$$

If X and Y are independent, then:

- $Pr(A \cap B) = Pr(A) \cdot Pr(B)$
- $Pr(A|B) = Pr(A)$
- $f_{Y|X}(y|x) = f_Y(y)$

4.3 Covariance and Correlation

These measures describe the strength and direction of the linear relationship between two variables.

- **Covariance:** Measures how two variables move together.
$$cov(X,Y) = \sigma_{XY} = E[(X - \mu_X)(Y - \mu_Y)]$$
  A useful computational formula is:
$$cov(X,Y) = E[XY] - E[X]E[Y]$$

- **Correlation:** A standardized measure of covariance that is unit-free and ranges from -1 to 1.
$$corr(X,Y) = \rho_{XY} = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$$
  - $\rho = +1$: Perfect positive linear relationship.
  - $\rho = -1$: Perfect negative linear relationship.
  - $\rho = 0$: No linear relationship.

4.4 Variance of Sums

The variance of a sum of two random variables is:

$$var(aX + bY) = a^2 var(X) + b^2 var(Y) + 2ab \cdot cov(X,Y)$$

If X and Y are independent, $cov(X,Y) = 0$, and the formula simplifies.

5 Chapter 5: Conditional Distributions and Expectations

5.1 Conditional Probability and Bayes' Rule

- **Conditional Probability:** The probability of an event A occurring, given that event B has already occurred.

$$Pr(A|B) = \frac{Pr(A \cap B)}{Pr(B)}$$

- **Bayes' Rule:** Relates the conditional probability of two events.

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

5.2 Conditional Expectation

The **conditional expectation** $E[Y|X = x]$ is the expected value of Y given that X has taken the specific value x. It represents the best prediction of Y given information about X.

5.3 Law of Iterated Expectations

This law states that the unconditional expectation of Y is the weighted average of the conditional expectations of Y given X.

$$E[Y] = E[E[Y|X]]$$

For a discrete variable X:

$$E[Y] = \sum_i E[Y|X = x_i]Pr(X = x_i)$$

5.4 Conditional Variance

- **Conditional Variance:** The variance of a random variable Y given that another variable X has taken a specific value x.

$$var(Y|X) = E[(Y - E[Y|X])^2|X]$$

- **Law of Total Variance:** The unconditional variance of Y can be decomposed into two parts.

$$var(Y) = E[var(Y|X)] + var(E[Y|X])$$

Part II
Common Probability Distributions and Data

6 Chapter 6: Key Distributions in Statistics

Several probability distributions are fundamental to statistical analysis.

6.1 The Normal Distribution

A continuous distribution characterized by its symmetric, bell-shaped PDF. It is completely defined by its mean (μ) and variance (σ^2), denoted $Y \sim N(\mu, \sigma^2)$.

- Approximately 95% of the probability lies within $\mu \pm 1.96\sigma$.
- **Standard Normal Distribution:** $Z \sim N(0, 1)$. Any normal random variable Y can be standardized:
$$Z = \frac{Y - \mu}{\sigma} \sim N(0, 1)$$

6.2 The Chi-Square (χ^2) Distribution

The distribution of the sum of m squared independent standard normal random variables. It is defined by its **degrees of freedom**, m.

$$\text{If } Z_1, \ldots, Z_m \sim iid\ N(0, 1), \text{ then } \sum_{i=1}^{m} Z_i^2 \sim \chi_m^2$$

6.3 The Student-t Distribution

The distribution of the ratio of a standard normal random variable to the square root of an independently distributed chi-squared variable (divided by its degrees of freedom, m).

$$\text{If } Z \sim N(0, 1) \text{ and } W \sim \chi_m^2, \text{ then } \frac{Z}{\sqrt{W/m}} \sim t_m$$

It is bell-shaped with "fatter" tails than the normal distribution. As degrees of freedom increase, it converges to the standard normal distribution.

6.4 The F Distribution

The distribution of the ratio of two independently distributed chi-squared variables, each divided by its degrees of freedom (m and n).

$$\text{If } W \sim \chi_m^2 \text{ and } V \sim \chi_n^2, \text{ then } F = \frac{W/m}{V/n} \sim F_{m,n}$$

7 Chapter 7: Outliers and Data Types

7.1 Outliers

An **outlier** is a data point that differs significantly from other observations. Distributions with "thin tails," like the Normal distribution, have a very low probability of producing extreme outliers.

7.2 Data Types

- **Experimental Data:** Obtained from controlled experiments (e.g., randomized controlled trials).
- **Observational Data:** Collected without controlling subject assignment (e.g., surveys).
- **Cross-Sectional Data:** Data on different subjects at a single point in time.
- **Time-Series Data:** Data for a single subject collected at multiple points in time.
- **Longitudinal (Panel) Data:** Data on multiple subjects, each observed at multiple points in time.

Part III

Fundamentals of Statistics

8 Chapter 8: Estimation

Statistical inference involves using data from a sample to learn about a population.

8.1 Estimators and Estimates

- **Estimator:** A function of a sample of data used to infer the value of a population parameter. It is a random variable.
- **Estimate:** The specific numerical value of an estimator computed from a particular sample. It is a fixed number.

8.2 Estimating the Population Mean and Variance

- **Sample Mean ($\overline{Y}$):** The estimator for the population mean (μ_Y).

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

- **Sample Variance (s_Y^2):** The estimator for the population variance (σ_Y^2). The division by $n-1$ is a degrees of freedom adjustment.

$$s_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2$$

- **Standard Error of the Mean ($SE(\overline{Y})$):** An estimator of the standard deviation of the sample mean's sampling distribution.

$$SE(\overline{Y}) = \frac{s_Y}{\sqrt{n}}$$

8.3 Properties of Estimators

Desirable characteristics of an estimator include:

1. **Unbiasedness:** An estimator is unbiased if its expected value equals the true population parameter ($E[\overline{Y}] = \mu_Y$).
2. **Consistency:** An estimator is consistent if it converges in probability to the true parameter value as the sample size $n \to \infty$.
3. **Efficiency:** Among all unbiased estimators, the one with the smallest variance is the most efficient. The sample mean is the **Best Linear Unbiased Estimator (BLUE)** of the population mean.

9 Chapter 9: The Distribution of Sample Means

9.1 The Sampling Distribution

The probability distribution of the sample mean $\overline{Y}$ is called its **sampling distribution**. If observations are i.i.d., this distribution has:

- Mean: $E[\overline{Y}] = \mu_Y$
- Variance: $var(\overline{Y}) = \sigma^2_{\overline{Y}} = \frac{\sigma^2_Y}{n}$

9.2 The Law of Large Numbers (LLN)

The LLN states that the sample mean $\overline{Y}$ converges in probability to the true population mean μ_Y as the sample size n gets very large.

$$\overline{Y} \xrightarrow{p} \mu_Y$$

9.3 The Central Limit Theorem (CLT)

The CLT states that the sampling distribution of $\overline{Y}$ is approximately normal for large sample sizes, regardless of the population distribution.

$$\overline{Y} \sim N\left(\mu_Y, \frac{\sigma^2_Y}{n}\right) \quad \text{for large } n$$

10 Chapter 10: Hypothesis Testing

A formal procedure for deciding between two claims about a population parameter.

10.1 The Null and Alternative Hypotheses

- **Null Hypothesis (H_0):** The statement being tested, often a claim of "no effect." Example: $H_0 : \mu_Y = 20$.
- **Alternative Hypothesis (H_1 or H_A):** The statement accepted if H_0 is rejected.
  - **Two-Sided:** $H_1 : \mu_Y \neq 20$.
  - **One-Sided:** $H_1 : \mu_Y > 20$ or $H_1 : \mu_Y < 20$.

10.2 The p-Value

The **p-value** is the probability of observing a test statistic as extreme as, or more extreme than, the one calculated from the sample, assuming H_0 is true. A small p-value provides evidence against H_0.

10.3 The t-Statistic

When σ^2_Y is unknown, we use the **t-statistic**:

$$t = \frac{\overline{Y} - \mu_{Y,0}}{SE(\overline{Y})} = \frac{\overline{Y} - \mu_{Y,0}}{s_Y/\sqrt{n}}$$

For large samples, this t-statistic is approximately distributed as a standard normal, $N(0, 1)$.

10.4 Errors and Significance

- **Significance Level (α):** A threshold for rejecting H_0 (e.g., 0.05). We reject H_0 if p-value $< \alpha$.
- **Type I Error:** Rejecting a true null hypothesis ($P(\text{Type I Error}) = \alpha$).
- **Type II Error:** Failing to reject a false null hypothesis.
- **Power of the Test:** The probability of correctly rejecting a false null hypothesis.

11 Chapter 11: Confidence Intervals

A **confidence interval** is a range of values likely to contain the true population parameter. A 95% confidence interval for the population mean μ_Y is:

$$\left[\overline{Y} - 1.96 \cdot SE(\overline{Y}), \quad \overline{Y} + 1.96 \cdot SE(\overline{Y})\right]$$

12 Chapter 12: Comparing Two Populations

To compare the means of two independent populations (e.g., groups 'm' and 'w'):

- **Null Hypothesis:** $H_0 : \mu_m - \mu_w = d_0$ (where often $d_0 = 0$).
- **Standard Error of the Difference:**

$$SE(\overline{Y}_m - \overline{Y}_w) = \sqrt{\frac{s_m^2}{n_m} + \frac{s_w^2}{n_w}}$$

- **t-Statistic for the Difference in Means:**

$$t = \frac{(\overline{Y}_m - \overline{Y}_w) - d_0}{SE(\overline{Y}_m - \overline{Y}_w)}$$

Part IV
Problems and Applications

13 Exercise Set 1: Based on Stock & Watson, Ch. 2

1. Let Y be the number of "heads" that occur when two fair coins are tossed.
   - 1.a. Derive the probability distribution (PMF) of Y.
   - 1.b. Derive the cumulative probability distribution (CDF) of Y.
   - 1.c. Calculate the mean and variance of Y.
2. Using probability tables or software, compute the following probabilities:
   - 2.a. If $Y \sim N(1,4)$, find $Pr(Y \leq 3)$.
   - 2.b. If $Y \sim N(50,25)$, find $Pr(40 \leq Y \leq 52)$.
   - 2.c. If $Y \sim \chi^2_{10}$, find $Pr(Y > 18.31)$.
   - 2.d. If $Y \sim t_{15}$, find $Pr(Y > 1.75)$.
3. Given $\mu_Y = 100$ and $\sigma^2_Y = 43$, use the central limit theorem to compute $Pr(101 \leq \overline{Y} \leq 103)$ for a random sample of size $n = 64$.
4. Use the joint probability distribution table for employment status (Y) and college graduation (X) to answer the following:

Table 1: Joint Distribution of Employment and College Graduation (Sept 2017)

	Unemployed (Y=0)	**Employed (Y=1)**	**Total**
Non-College (X=0)	0.026	0.576	0.602
College (X=1)	0.009	0.389	0.398
Total	**0.035**	**0.965**	**1.000**

   - 4.a. Compute $E[Y]$.
   - 4.b. The unemployment rate is the fraction of the labor force that is unemployed. Show that this is equal to $1 - E[Y]$.
   - 4.c. Calculate $E[Y|X = 1]$ and $E[Y|X = 0]$.
   - 4.d. Calculate the unemployment rate for college graduates and non-college graduates.
   - 4.e. If a randomly selected person is unemployed, what is the probability they are a college graduate?
   - 4.f. Are educational achievement and employment status independent? Explain.

14 Exercise Set 2: Based on Stock & Watson, Ch. 3

1. In a population, $\mu_Y = 100$ and $\sigma^2_Y = 43$. Use the CLT to find:
   - 1.a. For $n = 100$, find $Pr(\overline{Y} < 101)$.
   - 1.b. For $n = 165$, find $Pr(\overline{Y} > 98)$.
2. In a survey of 400 likely voters, 215 plan to vote for the incumbent. Let p be the true fraction of voters who prefer the incumbent.

2.a. Estimate p.

2.b. Calculate the standard error of your estimate.

2.c. What is the p-value for the test $H_0 : p = 0.5$ vs. $H_1 : p \neq 0.5$?

2.d. Did the survey contain statistically significant evidence that the incumbent was ahead at the time of the survey?

3. Data on 5th-grade test scores for 420 California school districts yielded $\overline{Y} = 654.2$ and $s_Y = 19.1$.

3.a. Construct a 95% confidence interval for the mean test score.

3.b. The data was divided by class size, with the following results:

Table 2: Test Score Data by Class Size

Class Size	**Avg. Score** ($\overline{Y}$)	**Std. Dev.** (s_Y)	**n**
Small (< 20)	657.4	19.4	238
Large (≥ 20)	650.0	17.9	182

Is there statistically significant evidence that districts with smaller classes have higher average test scores? Explain.