

STA260 Notes

By sytez

Preface

These notes were made for the course STA260: Probability and Statistics II at the University of Toronto Mississauga. They are based on the lectures and lecture slides of the 2024 summer offering. This course was instructed by Professor Luai Al Labadi (The Goat). These notes primarily consist of a summary of the definitions and theorems from the course. The notes are not exhaustive and may contain errors. The proofs provided are not complete and instead provide a brief outline of the proof containing the main ideas. Theorems that do not contain a proof are usually trivial to prove with exceptions of the ones whose proof is beyond the scope of this course.

Contents

7	Chapter 7 — Sampling Distributions and the Central Limit Theorem	3
7.1	Introduction	3
7.2	Sampling Distributions	4
7.3	The t-Distribution	6
7.4	The F-Distribution	7
7.5	The Central Limit Theorem	8
7.6	Normal Approximation to the Binomial	10
8	Chapter 8 — Estimation	11
8.1	Point Estimation	11
8.2	Evaluating the Goodness of an Estimator	13
8.3	Confidence Intervals	14
8.4	Large-Sample Confidence Intervals	16
9	Chapter 9 — Properties of Point Estimators and Methods of Estimation	17
9.1	Relative Efficiency	17
9.2	Consistency	18
9.3	Sufficiency	19
9.4	Completeness and Uniqueness	20
9.5	Exponential Family	21
9.6	The Rao-Blackwell Theorem	22
9.7	Unique Minimum Variance Unbiased Estimator (UMVUE)	23
9.8	Cramer-Rao Theorem	24
9.9	How to find Estimators	25
10	Chapter 10 — Hypothesis Testing	26
10.1	Elements of a Statistical Test	26
10.2	Neyman-Pearson Lemma	27
10.3	Likelihood Ratio Tests	28

7 Chapter 7 — Sampling Distributions and the Central Limit Theorem

7.1 Introduction

Definition 7.1.1. A **population** consists of the entire collection of the observations with which we are concerned.

Definition 7.1.2. A **sample** is a subset of a population.

Definition 7.1.3. A **parameter** is a numerical summary of a population. For example, the mean, the variance, etc. In practice, it is unknown.

Definition 7.1.4. A **statistic** is a numerical summary of a sample. For example, the sample mean, the sample variance, etc.

Definition 7.1.5. The statistic varies from sample to sample and hence it is a random variable and has a probability distribution called the **sampling distribution**.

The knowledge of the sampling distribution of a statistic helps to make an inference about the corresponding population (true) parameter.

Definition 7.1.6. We say that the random variables Y_1, Y_2, \dots, Y_n are **independent and identically distributed** (i.i.d.) if they are independent random variables and have the same probability distribution (same pdf/cdf).

For a random sample, we write

$$Y_1, \dots, Y_n \stackrel{\text{i.i.d.}}{\sim} f(y) \quad (\text{continuous})$$

$$Y_1, \dots, Y_n \stackrel{\text{i.i.d.}}{\sim} p(y) \quad (\text{discrete})$$

7.2 Sampling Distributions

Definition 7.2.1. The **sample mean** is defined as

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

Definition 7.2.2. The **sample variance** is defined as

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

Theorem 7.2.1. Let Y_1, \dots, Y_n be an iid random sample from a population with any distribution with finite mean μ and finite variance σ^2 . Then $E(\bar{Y}) = \mu$ and $V(\bar{Y}) = \frac{\sigma^2}{n}$.

Theorem 7.2.2. If $Y_1, \dots, Y_n \stackrel{\text{i.i.d.}}{\sim} N(\mu, \sigma^2)$, then

$$\bar{Y} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

Proof. Show MGF of \bar{Y} is the MGF of $N\left(\mu, \frac{\sigma^2}{n}\right)$ then conclude by uniqueness of MGF. \square

Definition 7.2.3. The **standard normal distribution** is defined as

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

Theorem 7.2.3. Let $Y_1, \dots, Y_n \stackrel{\text{i.i.d.}}{\sim} N(\mu, \sigma^2)$. If $Z_i = \frac{Y_i - \mu}{\sigma}$, then

$$\sum_{i=1}^n Z_i^2 = \sum_{i=1}^n \left(\frac{Y_i - \mu}{\sigma}\right)^2$$

has a χ^2 distribution with n degrees of freedom (df)

Proof. Note 3 facts:

1. $\chi_{(v)}^2 = \text{Gamma}(v/2, 2)$
2. $\chi_{(v_1)}^2 + \chi_{(v_2)}^2 \sim \chi_{(v_1+v_2)}^2$ assuming independence
3. $[N(0, 1)]^2 = Z^2 \sim \chi_{(1)}^2$

The proof becomes trivial from here. \square

Note that similarly we have the sum of many distributions can be studied easily.

$$Y_1, \dots, Y_n \stackrel{\text{iid}}{\sim} \text{Ber}(p) \implies \sum_{i=1}^n Y_i \sim \text{Bin}(n, p)$$

$$Y_1, \dots, Y_n \stackrel{\text{iid}}{\sim} \text{Poisson}(\lambda) \implies \sum_{i=1}^n Y_i \sim \text{Poisson}(n\lambda)$$

$$Y_1, \dots, Y_n \stackrel{\text{iid}}{\sim} \chi^2(v_i) \implies \sum_{i=1}^n Y_i \sim \chi^2\left(\sum_{i=1}^n v_i\right)$$

$$Y_1, \dots, Y_n \stackrel{\text{iid}}{\sim} N(\mu_i, \sigma_i^2) \implies \sum_{i=1}^n Y_i \sim N\left(\sum_{i=1}^n \mu_i, \sum_{i=1}^n \sigma_i^2\right)$$

Proof. Note the following fact:

Let $U = Y_1 + \dots + Y_n$ then we have $M_U(t) = M_{Y_1}(t) \cdot \dots \cdot M_{Y_n}(t)$.

That is that the MGF of the sum of random variables is the product of the MGFs of the random variables.

The proof follows trivially using the uniqueness of MGFs. \square

Theorem 7.2.4. As a corollary of theorem 7.2.2

If $Y_1, \dots, Y_n \stackrel{\text{i.i.d.}}{\sim} N(\mu, \sigma^2)$, then

$$U = \sum_{i=1}^n \left(\frac{Y_i - \mu}{\sigma} \right)^2 \sim \chi^2_{(n)}$$

similarly one sees that if $Y_i \sim N(\mu_i, \sigma_i^2)$, $i = 1, \dots, n$ (and independent), then

$$U = \sum_{i=1}^n \left(\frac{Y_i - \mu_i}{\sigma_i} \right)^2 \sim \chi^2_{(n)}.$$

Theorem 7.2.5. Let Y_1, Y_2, \dots, Y_n be a random sample from a normal distribution with mean μ and variance σ^2 . Then

$$\frac{(n-1)S^2}{\sigma^2} = \sum_{i=1}^n \frac{(Y_i - \bar{Y})^2}{\sigma^2} \sim \chi^2_{n-1}.$$

Additionally, \bar{Y} and S^2 are independent.

7.3 The t-Distribution

Definition 7.3.1. Let $Z \sim N(0, 1)$ and $W \sim \chi^2_{(v)}$. Then if Z and W are independent then we say

$$T = \frac{Z}{\sqrt{W/v}} \sim t_{(v)}$$

has a **t-Distribution** with v degrees of freedom.

So far it was assumed that the population standard deviation σ is known. However, this assumption may be unreasonable. We want to estimate both μ and σ . A natural statistic to deal with inferences on μ is

$$T = \frac{\bar{Y} - \mu}{S/\sqrt{n}} \sim t_{(n-1)}$$

Properties of the t-Distribution

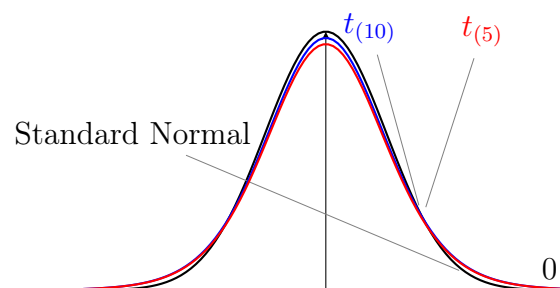


Figure 1: Comparison of Standard Normal and t -distribution curves with 5 and 10 degrees of freedom

- Each t_ν curve is bell-shaped and centered at 0.
- Each t_ν curve is spread out more than the standard normal (Z) curve.
- As ν increases, the spread of the corresponding t_ν curve decreases.
- As $\nu \rightarrow \infty$, the sequence of t_ν curves approaches the standard normal curve (the Z curve is called a t curve with $\text{df} = \infty$).

7.4 The F-Distribution

Definition 7.4.1. Let $W_1 \sim \chi^2_{(\nu_1)}$ and $W_2 \sim \chi^2_{(\nu_2)}$ be independent. Then

$$F = \frac{W_1/\nu_1}{W_2/\nu_2}$$

has a **F-Distribution** with ν_1 and ν_2 degrees of freedom.

Theorem 7.4.1. If S_1^2 and S_2^2 are the variances of independent random samples of size n_1 and n_2 taken from normal populations with variances σ_1^2 and σ_2^2 , respectively, then

$$\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F(n_1 - 1, n_2 - 1).$$

Proof. Recall that

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{(n-1)}$$

the proof trivially follows. □

Theorem 7.4.2. Let Y be a random variable such that $Y \sim F_{(v_1, v_2)}$. Then

$$\frac{1}{Y} \sim F_{(v_2, v_1)}$$

Theorem 7.4.3. Let $Y_1 \sim F_{(v_1, v_2)}$ then

$$\left(1 + \frac{v_1}{v_2} Y_1\right)^{-1} \sim \text{Beta}\left(\frac{v_2}{2}, \frac{v_1}{2}\right)$$

Proof. Recall that

$$\chi^2_{(v)} = \text{Gamma}\left(\frac{v}{2}, 2\right) \quad \text{and} \quad \frac{\text{Gamma}(\alpha, \gamma)}{\text{Gamma}(\alpha, \gamma) + \text{Gamma}(\beta, \gamma)} = \text{Beta}(\alpha, \beta)$$

the proof follows trivially using the definition of the F -distribution. □

Theorem 7.4.4. Let T be a random variable with a t -distribution with v degrees of freedom. Then $U = T^2$ has an F -distribution with 1 and v degrees of freedom.

7.5 The Central Limit Theorem

We already showed that if Y_1, Y_2, \dots, Y_n represents a random sample from any distribution with mean μ and variance σ^2 , then $E(\bar{Y}) = \mu$ and $V(\bar{Y}) = \frac{\sigma^2}{n}$.

In what follows, we will develop an approximation for the sampling distribution of \bar{Y} that can be used regardless of the distribution of the population from which the sample is taken.

Theorem 7.5.1. Central Limit Theorem

Let Y_1, Y_2, \dots, Y_n be a random sample from a population with finite mean μ and finite variance σ^2 , but unknown distribution. Then if n is sufficiently large, \bar{Y} is **approximately normally distributed** with mean μ and variance $\frac{\sigma^2}{n}$, i.e. $\bar{Y} \approx N(\mu, \frac{\sigma^2}{n})$.

The CLT can be written more formally as:

If

$$U_n = \frac{\sum_{i=1}^n Y_i - n\mu}{\sigma\sqrt{n}} = \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}}$$

then

$$U_n \xrightarrow{d} N(0, 1)$$

This implies that

$$\lim_{n \rightarrow \infty} P(U_n \leq u) = \int_{-\infty}^u \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt$$

Proof. Let $U_n = \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}}$, now we can rewrite $U_n = \frac{1}{\sqrt{n}} \sum_{i=1}^n Z_i$.

Now note that

$$M_{U_n}(t) = M_{\frac{1}{\sqrt{n}} \sum_{i=1}^n Z_i}(t) = M_{\sum_{i=1}^n Z_i}\left(\frac{t}{\sqrt{n}}\right) = \prod_{i=1}^n M_{Z_i}\left(\frac{t}{\sqrt{n}}\right) = \left[M_{Z_1}\left(\frac{t}{\sqrt{n}}\right)\right]^n$$

Now through the use of the McLaurin expansion of $M_{Z_1}\left(\frac{t}{\sqrt{n}}\right)$ it can be shown that

$$\lim_{n \rightarrow \infty} n \cdot M_{Z_1}\left(\frac{t}{\sqrt{n}}\right) - n = \frac{t^2}{2}$$

Now note that

$$\lim_{n \rightarrow \infty} b_n = b \implies \lim_{n \rightarrow \infty} \left(1 + \frac{b_n}{n}\right)^n = e^b$$

Thus we see that

$$\lim_{n \rightarrow \infty} M_{U_n}(t) = \lim_{n \rightarrow \infty} \left[M_{Z_1}\left(\frac{t}{\sqrt{n}}\right)\right]^n = e^{\frac{t^2}{2}}$$

Thus we see that $\lim_{n \rightarrow \infty} M_{U_n}(t) = M_{N(0,1)}(t)$ and thus by the uniqueness of MGFs we see that $U_n \xrightarrow{d} N(0, 1)$. □

The **Central Limit Theorem** states that the sample mean from any probability distribution (as long as mean and variance are finite) will have an approximate normal distribution, if the sample is sufficiently large.

“Large n ” means $n \geq 30$ in general, but in some cases may even be much less.

The larger the sample size, the more nearly normally distributed is the population of all possible sample means.

For fairly symmetric distributions, $n > 15$ will be sufficient.

7.6 Normal Approximation to the Binomial

The normal distribution (continuous) can be used to approximate binomial (discrete) probabilities when there is a very large number of trials and when np and $n(1 - p)$ are both large ($np \geq 5$ and $n(1 - p) \geq 5$).

Theorem 7.6.1. Normal Approximation to the Binomial

If $Y \sim \text{Bin}(n, p)$, then $Y \approx N(\mu = np, \sigma^2 = npq)$

Proof. The justification for the normal approximation to the binomial is based on the Central Limit Theorem. \square

To improve the accuracy of the approximation, we usually use a correction factor, called continuity correction, to take into account that the binomial random variable is discrete while the normal is continuous.

The basic idea is to treat the discrete value b as the continuous interval from $b - 0.5$ to $b + 0.5$ giving the following adjustments:

- $P(Y = b) = P(b - 0.5 \leq Y \leq b + 0.5)$
- $P(Y \leq b) = P(Y \leq b + 0.5)$
- $P(b \leq Y) = P(b - 0.5 \leq Y)$

8 Chapter 8 — Estimation

If $Y_1, Y_2, \dots, Y_n \stackrel{i.i.d.}{\sim} N(\mu, 1)$, where μ is unknown, then how do we estimate μ ?

An **estimator** is a rule, often expressed as a formula, that tells how to calculate the value of an estimate based on the measurements contained in a sample.

Let Y_1, Y_2, \dots, Y_n be i.i.d. with pdf/pmf $f = f(\cdot|\theta) = f_\theta$, where the parameter θ is unknown (f depends on θ). We want to find an estimate for the unknown parameter θ .

8.1 Point Estimation

Definition 8.1.1. A **point estimator** $\hat{\theta}$ of the parameter θ is a function of the underlying random variables and so it is a random variable with a distribution function.

A point estimate of θ is a function of the sample Y_1, Y_2, \dots, Y_n only. That is if y_1, \dots, y_n is the observed sample then $\hat{\theta} = U(y_1, \dots, y_n)$ is a number.

For example a distribution with mean μ and variance σ^2 we have that \bar{Y} is a point estimator for μ and S^2 is a point estimator for σ^2 .

We say that a point estimator $\hat{\theta}$ is "good" if it has the following desirable properties:

1. Unbiased
2. Consistent
3. Minimum Variance
4. Has a known probability distribution

Definition 8.1.2. Let $\hat{\theta}$ be a point estimator for a parameter θ . Then $\hat{\theta}$ is an **unbiased estimator** if $E(\hat{\theta}) = \theta$. If $E(\hat{\theta}) \neq \theta$, then $\hat{\theta}$ is said to be **biased**.

Note that $\hat{\theta}$ is a random variable whereas θ is a constant.

Definition 8.1.3. The **bias** of a point estimator $\hat{\theta}$ is given by $B(\hat{\theta}) = E(\hat{\theta}) - \theta$.

We see that a point estimator $\hat{\theta}$ is unbiased w.r.t. θ if $B(\hat{\theta}) = 0$.

Definition 8.1.4. The **mean square error** of a point estimator $\hat{\theta}$ is $MSE(\hat{\theta}) = E[(\hat{\theta} - \theta)^2]$.

Theorem 8.1.1. $MSE(\hat{\theta}) = V(\hat{\theta}) + [B(\hat{\theta})]^2$

It can be seen that if the estimator is unbiased then $MSE(\hat{\theta}) = V(\hat{\theta})$.

Examples of Well-Known Unbiased Estimators

Let Y_1, Y_2, \dots, Y_n be a random sample with $E(Y_i) = \mu_1$ and X_1, X_2, \dots, X_n be a random sample with $E(X_i) = \mu_2$. Then:

$$\begin{aligned} E(\bar{Y}) &= \mu_1 \quad \text{and} \quad E(\bar{X}) = \mu_2 \\ E(\bar{Y} - \bar{X}) &= \mu_1 - \mu_2 \end{aligned}$$

Let $Y \sim \text{Bin}(n, p_1)$ and $X \sim \text{Bin}(n, p_2)$. Then $\hat{p}_1 = \frac{Y}{n}$ is the proportion of successes in the sample.

$$E(\hat{p}_1) = E\left(\frac{Y}{n}\right) = \frac{1}{n}E(Y) = \frac{np_1}{n} = p_1$$

$$E(\hat{p}_2) = E\left(\frac{X}{n}\right) = \frac{1}{n}E(X) = \frac{np_2}{n} = p_2$$

$$E(\hat{p}_1 - \hat{p}_2) = p_1 - p_2$$

8.2 Evaluating the Goodness of an Estimator

Definition 8.2.1. Let $\sigma_{\hat{\theta}}^2$ be the variance of the sampling distribution of the estimator $\hat{\theta}$ (i.e. $V(\hat{\theta}) = \sigma_{\hat{\theta}}^2$), then $\sqrt{V(\hat{\theta})} = \sqrt{\sigma_{\hat{\theta}}^2} = \sigma_{\hat{\theta}}$ is called the **standard error** of the estimator.

That is, the **standard error** of $\hat{\theta}$ = the standard deviation of $\hat{\theta}$.

Note that we call it the standard error because it comes from the sampling process.

Definition 8.2.2. The **error of estimation** ε is the distance between an estimator and its target parameter. That is, $\varepsilon = |\hat{\theta} - \theta|$.

Definition 8.2.3. The **2-standard-error bound** on the error of estimation is given by $2SE(\hat{\theta}) = 2\sigma_{\hat{\theta}}$.

To place a 2-standard-error bound on the error of estimation is to find the probability that the error of estimation is within 2 standard errors of the estimator. That is to find $P(|\varepsilon| < 2SE(\hat{\theta})) = P(|\varepsilon| < 2\sigma_{\hat{\theta}})$.

8.3 Confidence Intervals

An alternative to reporting a single value for the parameter being estimated is to calculate and report an entire interval of plausible values; i.e., a **confidence interval** (CI).

Properties of the CI:

- It contains true parameter θ with high probability
- It is relatively narrow

The **lower** and **upper** endpoints of a CI are called the **lower** and **upper confidence limits**.

The probability that a CI will enclose θ is called the **confidence coefficient** or **confidence level**, denoted by $1 - \alpha$.

Definition 8.3.1. A $100(1 - \alpha)\%$ **confidence interval** for a parameter θ is a random interval $[\hat{\theta}_L, \hat{\theta}_U]$ such that

$$P(\hat{\theta}_L \leq \theta \leq \hat{\theta}_U) = 1 - \alpha,$$

regardless of the value of θ .

Properties of the CI:

- Here, $\hat{\theta}_L$ and $\hat{\theta}_U$ are functions of Y_i 's only.
- **Two-sided CI:** $[\hat{\theta}_L, \hat{\theta}_U]$
- **Lower one-sided CI:** $[\hat{\theta}_L, \infty)$
- **Upper one-sided CI:** $(-\infty, \hat{\theta}_U]$
- The confidence level is denoted by $100(1 - \alpha)\%$. The most common confidence levels are 90%, 95%, and 99%.
- The higher the confidence level, the more strongly we believe that the true value of the parameter being estimated lies within the interval.

One way of getting CI is through the **Pivotal Method**:

- We need to find a random variable, called a **pivot**, that is typically a function of the estimator of θ satisfying the following:
 1. It depends on and only on Y_1, \dots, Y_n and θ : $Q(y_1, \dots, y_n | \theta)$.
 2. Its probability distribution (pdf/cdf) does not depend on θ or any other unknown parameter.
- Use the distribution of the pivot to find the cutting points a and b :

$$P(a \leq Q \leq b) = 1 - \alpha.$$

- Write down $P(a \leq Q \leq b)$ to get $P(L(y_1, \dots, y_n) \leq \theta \leq U(y_1, \dots, y_n))$, if you can.
- Thus, $(1 - \alpha)100\%$ CI for θ is: $[\hat{\theta}_L, \hat{\theta}_U] = [L(y_1, \dots, y_n), U(y_1, \dots, y_n)]$.

In general we use 3 methods to show a pivot's distribution does not depend on θ :

1. Use the CDF of the pivot and show that it does not depend on θ .
2. Use the MGF of the pivot and the uniqueness of MGFs to show that its corresponding pdf does not depend on θ .
3. Directly show that the pivot has a similar distribution as a known distribution.

An example of method III:

Let Y be an observation from an exponential distribution. Then $\frac{Y}{\theta}$ is a pivot.

Proof. Note that $Y \sim \text{Exp}(\theta)$ and thus $Y \sim \text{Gamma}(1, \theta)$. Now note that as

$$\phi \text{Gamma}(\alpha, \beta) = \text{Gamma}(\alpha, \phi\beta)$$

we see that

$$\frac{Y}{\theta} \sim \frac{1}{\theta} \text{Gamma}(1, \theta) = \text{Gamma}(1, \frac{1}{\theta}\theta) = \text{Gamma}(1, 1) \sim \text{Exp}(1)$$

and thus does not depend on θ . □

How to find the cutting points:

Suppose we have a pivot U of Y_1, \dots, Y_n and θ and we want to find cutting points a and b such that $P(a \leq U \leq b) = 1 - \alpha$. We then see that $P(U \leq a) + P(b \leq U) = \alpha$. Thus there may be many solutions for cutting points. Thus as a convention we find a and b such that $P(U \leq a) = \frac{\alpha}{2}$ and $P(b \leq U) = \frac{\alpha}{2}$.

8.4 Large-Sample Confidence Intervals

If the target parameter θ is $\mu, p, \mu_1 - \mu_2$ or $p_1 - p_2$, then for large samples by CLT

$$Z = \frac{\hat{\theta} - \theta}{\sigma_{\hat{\theta}}} \sim N(0, 1)$$

The distribution of Z doesn't depend on θ . So it is **pivotal**.

The pivotal method can be employed to develop confidence intervals for the target parameter θ .

9 Chapter 9 — Properties of Point Estimators and Methods of Estimation

9.1 Relative Efficiency

9.2 Consistency

9.3 Sufficiency

9.4 Completeness and Uniqueness

9.5 Exponential Family

9.6 The Rao-Blackwell Theorem

9.7 Unique Minimum Variance Unbiased Estimator (UMVUE)

9.8 Cramer-Rao Theorem

9.9 How to find Estimators

10 Chapter 10 — Hypothesis Testing

10.1 Elements of a Statistical Test

10.2 Neyman-Pearson Lemma

10.3 Likelihood Ratio Tests