STA260 Notes

By Sanchit Manchanda

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 summer 2024 offering. This course was instructed by Professor Luai Al Labadi. 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

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 $Y_1, Y_2, \ldots, Y_n$ 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

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

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

7.2 Sampling Distributions

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

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

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

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

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

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

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

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

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

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

Theorem 7.2.3. Let $Y_1, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} N(\mu, \sigma^2)$. 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. $\chi^2_{(v)} = \text{Gamma}(v/2, 2)$
2. $\chi^2_{(v_1)} + \chi^2_{(v_2)} \sim \chi^2_{(v_1+v_2)}$ assuming independence
3. $[N(0,1)]^2 = Z^2 \sim \chi^2_{(1)}$

The proof becomes trivial from here. □

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

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

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

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

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

Proof. Note the following fact:
Let $U = Y_1 + \ldots + Y_n$ then we have $M_U(t) = M_{Y_1}(t) \cdot \ldots \cdot M_{Y_n}(t)$.
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 unqiueness of MGFs. □

Theorem 7.2.4. As a corollary of theorem 7.2.2
If $Y_1, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} N(\mu, \sigma^2)$, then

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

similarly one sees that if $Y_i \sim N(\mu_i, \sigma_i^2)$, $i = 1, \ldots, n$ (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 $Y_1, Y_2, \ldots, Y_n$ be a random sample from a normal distribution with mean μ and variance σ^2. Then

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

Additionally, $\overline{Y}$ 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

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

Properties of the t-Distribution

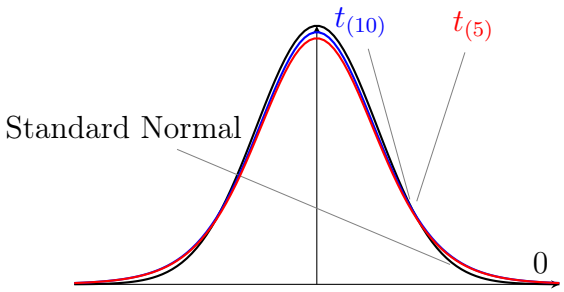

Figure 1: Comparison of Standard Normal and t-distribution curves with 5 and 10 degrees of freedom

- 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 $\nu \to \infty$, the sequence of t_ν curves approaches the standard normal curve (the Z curve is called a t curve with 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 defintion 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 $Y_1, Y_2, \ldots, Y_n$ represents a random sample from any distribution with mean μ and variance σ^2, then $E(\overline{Y}) = \mu$ and $V(\overline{Y}) = \frac{\sigma^2}{n}$.
In what follows, we will develop an approximation for the sampling distribution of $\overline{Y}$ 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 $Y_1, Y_2, \cdots, Y_n$ 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, $\overline{Y}$ is approximately normally distributed with mean μ and variance $\frac{\sigma^2}{n}$, i.e. $\overline{Y} \approx N(\mu, \frac{\sigma^2}{n})$.

The CLT can be written more formally as:
If

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

then

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

This implies that

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

Proof. Let $U_n = \frac{\overline{Y}-\mu}{\sigma/\sqrt{n}}$, 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

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

Now note that

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

Thus we see that

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

Thus we see that $\lim_{n\to\infty} M_{U_n}(t) = M_{N(0,1)}(t)$ 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. □

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 $Y_1, Y_2, \ldots, Y_n \overset{i.i.d.}{\sim} N(\mu, 1)$, 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 $Y_1, Y_2, \ldots, Y_n$ 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 $Y_1, Y_2, \ldots, Y_n$ only. That is if $y_1, \ldots, y_n$ is the observed sample then $\hat{\theta} = U(y_1, \ldots, y_n)$ is a number.

For example a distribution with mean μ and variance σ^2 we have that $\overline{Y}$ 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

$$MSE(\hat{\theta}) = E\left[(\hat{\theta} - \theta)^2\right]$$

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

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

Examples of Well-Known Unbiased Estimators
Let $Y_1, Y_2, \ldots, Y_n$ be a random sample with $E(Y_i) = \mu_1$ and $X_1, X_2, \ldots, X_n$ be a random sample with $E(X_i) = \mu_2$. Then:

$$E\left(\overline{Y}\right) = \mu_1 \quad \text{and} \quad E\left(\overline{X}\right) = \mu_2$$
$$E\left(\overline{Y} - \overline{X}\right) = \mu_1 - \mu_2$$

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.

$$E\left(\hat{p}_1\right) = E\left(\frac{Y}{n}\right) = \frac{1}{n}E(Y) = \frac{np_1}{n} = p_1$$
$$E\left(\hat{p}_2\right) = E\left(\frac{X}{n}\right) = \frac{1}{n}E(X) = \frac{np_2}{n} = p_2$$
$$E\left(\hat{p}_1 - \hat{p}_2\right) = p_1 - p_2$$

8.2 Evaluating the Goodness of an Estimator

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

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

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 $Y_1, \ldots, Y_n$ and θ: $Q(y_1, \ldots, y_n \mid \theta)$.
  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 $P(L(y_1, \ldots, y_n) \leq \theta \leq U(y_1, \ldots, y_n))$, if you can.
- Thus, $(1-\alpha)100\%$ CI for θ is: $[\hat{\theta}_L, \hat{\theta}_U] = [L(y_1, \ldots, y_n), U(y_1, \ldots, y_n)]$.

Note that the Pivot is not unique.

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 distritbution. 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 $Y_1, \ldots, Y_n$ 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 $P(U \leq a) = \frac{\alpha}{2}$ or $P(a \leq U) = 1 - \frac{\alpha}{2}$ and $P(b \leq U) = \frac{\alpha}{2}$. Thus it can be seen the cutting points are $U_{\frac{\alpha}{2}}$ and $U_{1-\frac{\alpha}{2}}$. This shortcut can be especially helpful when the pivot is distributed similar to a known distribution.

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.

Theorem 8.4.1. It can be seen that the confidence interval for any θ on a large sample is given by

$$\left[\hat{\theta} - Z_{\frac{\alpha}{2}}\sigma_{\hat{\theta}}, \hat{\theta} + Z_{\frac{\alpha}{2}}\sigma_{\hat{\theta}}\right] = \hat{\theta} \mp Z_{\frac{\alpha}{2}}\sigma_{\hat{\theta}}$$

Definition 8.4.1. We denote $Z_{\frac{\alpha}{2}}\sigma_{\hat{\theta}}$ as the **margin of error** of $\hat{\theta}$. It is a bound on the error of estimation.

Consider $\frac{\overline{Y}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$. Is this a pivotal quantity for μ?
Indeed the distribution does not depend on μ however it may depend σ. If σ is known then it is pivotal however if it is unknown then it is not pivotal. To this end we develop a pivotal quantity with the t-distribution.

Theorem 8.4.2. Let

$$T = \frac{\overline{Y} - \mu}{s/\sqrt{n}} \sim t_{n-1}$$

Then T is a pivotal quantity for μ. Moreover the $100(1-\alpha)\%$ CI for μ is given by

$$\left[\overline{Y} - t_{\alpha/2}\frac{S}{\sqrt{n}}, \overline{Y} + t_{\alpha/2}\frac{S}{\sqrt{n}}\right]$$

Suppose we want to compare between two datasets. Then we study the difference in their means or proportions. We can use the t-distribution to develop a pivotal quantity for the difference in means or proportions.
Let $Y_1, Y_2, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} N(\mu_1, \sigma_1^2)$ and $X_1, X_2, \ldots, X_n \overset{\text{i.i.d.}}{\sim} N(\mu_2, \sigma_2^2)$.
Assume the two popoulations are independent and that $\sigma_1^2 = \sigma_2^2 = \sigma^2$, where σ is unknown.
Then we have $\overline{Y} - \overline{X} \sim N\left(\mu_1 - \mu_2, \sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right)$.

Definition 8.4.2. We define the **pooled variance** as

$$S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{(n_1-1)+(n_2-1)}$$

where S_1^2 and S_2^2 are the sample variances.

Theorem 8.4.3. T defined as below is a pivotal quantity for $\mu_1 - \mu_2$:

$$T = \frac{Z}{\sqrt{W/(n_1+n_2-2)}} = \frac{\overline{Y} - \overline{X} - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim t(n_1+n_2-2)$$

The $100(1-\alpha)\%$ CI for $\mu_1 - \mu_2$ is given by

$$\left[\overline{Y} - \overline{X} - t_{\alpha/2}S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}},\ \overline{Y} - \overline{X} + t_{\alpha/2}S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\right] = \overline{Y} - \overline{X} \mp t_{\alpha/2}S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$$

Proof. We see that

$$Z = \frac{\overline{Y} - \overline{X} - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim N(0,1)$$

and

$$W = \frac{n_1+n_2-2}{\sigma^2}S_p^2 \sim \chi^2_{n_1+n_2-2}$$

Then we have

$$T = \frac{Z}{\sqrt{W/(n_1+n_2-2)}} = \frac{\frac{\overline{Y}-\overline{X}-(\mu_1-\mu_2)}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}}{\sqrt{\frac{\frac{n_1+n_2-2}{\sigma^2}S_p^2}{(n_1+n_2-2)}}} = \frac{\frac{\overline{Y}-\overline{X}-(\mu_1-\mu_2)}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}}{\sqrt{\frac{S_p^2}{\sigma^2}}} = \frac{\frac{\overline{Y}-\overline{X}-(\mu_1-\mu_2)}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}}{\frac{S_p}{\sigma}}$$

It follows then

$$T = \frac{\overline{Y} - \overline{X} - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim t(n_1+n_2-2)$$

□

Consider the case of finding a confidence interval for σ^2.

Theorem 8.4.4. Let $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$, where μ and σ^2 are unknown.
$(n-1)\frac{S^2}{\sigma^2} \sim \chi^2_{n-1}$ is a pivotal quantity for σ^2.
$100(1-\alpha)\%$ CI for σ^2 is:

$$\left[\frac{(n-1)S^2}{\chi^2_{\alpha/2}}, \frac{(n-1)S^2}{\chi^2_{1-\alpha/2}}\right]$$

Proof. As $(n-1)\frac{S^2}{\sigma^2} \sim \chi^2_{n-1}$ we have

$$P\left(\chi^2_{\alpha/2} \geq (n-1)\frac{S^2}{\sigma^2} \geq \chi^2_{1-\alpha/2}\right) = 1-\alpha$$

It can then be seen that

$$P\left(\frac{(n-1)S^2}{\chi^2_{\alpha/2}} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi^2_{1-\alpha/2}}\right) = 1-\alpha$$

□

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

9.1 Relative Efficiency

Definition 9.1.1. Given two unbiased estimators $\hat{\theta}_1$ and $\hat{\theta}_2$ of a parameter θ, with variances $V(\hat{\theta}_1)$ and $V(\hat{\theta}_2)$, respectively, then the **efficiency** of $\hat{\theta}_1$ relative to $\hat{\theta}_2$, denoted $\text{eff}(\hat{\theta}_1, \hat{\theta}_2)$, is defined to be the ratio

$$eff(\hat{\theta}_1, \hat{\theta}_2) = \frac{V(\hat{\theta}_2)}{V(\hat{\theta}_1)}$$

Definition 9.1.2. If $\hat{\theta}_1$ and $\hat{\theta}_2$ denote two unbiased estimators for the same parameter θ, then we say that $\hat{\theta}_1$ **is relatively more efficient** than $\hat{\theta}_2$ if $V(\hat{\theta}_2) > V(\hat{\theta}_1)$.

It can be seen that $\text{eff}(\hat{\theta}_1, \hat{\theta}_2) > 1 \implies \hat{\theta}_1$ is relatively more efficient than $\hat{\theta}_2$.
Likewise it can be seen that $\text{eff}(\hat{\theta}_1, \hat{\theta}_2) < 1 \implies \hat{\theta}_2$ is relatively more efficient than $\hat{\theta}_1$.
It can also be seen that $\text{eff}(\hat{\theta}_1, \hat{\theta}_2) = 1 \implies \hat{\theta}_1$ and $\hat{\theta}_2$ are equally efficient.

In practice when given two unbiased estimators we choose the one with the smaller variance.
If $\text{eff}(\hat{\theta}_1, \hat{\theta}_2) > 1 \implies$ choose $\hat{\theta}_1$.
If $\text{eff}(\hat{\theta}_1, \hat{\theta}_2) < 1 \implies$ choose $\hat{\theta}_2$.

9.2 Consistency

Definition 9.2.1. The estimator $\hat{\theta}_n$ is said to be a **consistent** estimator of θ if $\hat{\theta}_n$ converges in probability to θ. That is, for any positive number ϵ,

$$\lim_{n\to\infty} P\left(|\hat{\theta}_n - \theta| \leq \epsilon\right) = 1$$

or, equivalently,

$$\lim_{n\to\infty} P\left(|\hat{\theta}_n - \theta| > \epsilon\right) = 0$$

Theorem 9.2.1. Variance Rule
An unbiased estimator $\hat{\theta}_n$ for θ is consistent if

$$\lim_{n\to\infty} V\left(\hat{\theta}_n\right) = 0$$

Theorem 9.2.2. Weak Law of Large Numbers (WLLN)
Let $Y_1, \ldots, Y_n$ be independent random variables with expected value μ and variance σ^2. Then the sample mean

$$\overline{Y}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i \xrightarrow{p} \mu$$

- The WLLN states that as the number of observations in a sample increases, the sample mean approaches the population mean.
- In simpler terms, the more data you collect, the closer your average will be to the true average of the entire population.
- In fact, $\frac{1}{n}\sum_{i=1}^{n} g(Y_i) \xrightarrow{p} E[g(Y)]$.

Proof. Follows directly from using Chebyshev's Inequality □

Theorem 9.2.3. Continuous Mapping Theorem: Let g be a function that is continuous at θ. If $\hat{\theta}_n \xrightarrow{p} \theta$, then $g(\hat{\theta}_n) \xrightarrow{p} g(\theta)$.

9.3 Sufficiency

Definition 9.3.1. Let $Y_1, Y_2, \ldots, Y_n$ denote a random sample from a probability distribution with unknown parameter θ. $(Y_1, Y_2, \ldots, Y_n \sim f(y|\theta))$
Then the statistic $U = U(Y_1, Y_2, \ldots, Y_n)$ is said to be **sufficient** for θ if the conditional distribution of $Y_1, Y_2, \ldots, Y_n$, given U, does not depend on θ. That is $f(y_1, y_2, \ldots, y_n|U)$ does not require θ.

Interpretation of a Sufficient Statistic:

- All information about the value of the parameter θ is contained in the value of the sufficient statistic.
- A statistic is sufficient if no other statistic that can be calculated from the same sample provides any additional information as to the value of the parameter.
- A statistic is sufficient if the sample from which it is calculated gives no additional information than does the statistic.
- Sufficient estimators are important in constructing "good" estimators: MVUE
- **How to find a sufficient statistic for θ?**

Definition 9.3.2. Let $y_1, y_2, \ldots, y_n$ be sample observations taken on corresponding random variables $Y_1, Y_2, \ldots, Y_n$ whose distribution depends on a parameter θ. Then, if $Y_1, Y_2, \ldots, Y_n$ are discrete (continuous) random variables, the **likelihood of the sample**,
$L(\theta) = L(y_1, y_2, \ldots, y_n \mid \theta)$, is defined to be the joint probability (density) of $y_1, y_2, \ldots, y_n$. That is,

- **Discrete case:**

$$L(\theta) = L(y_1, \ldots, y_n \mid \theta) = p(y_1, \ldots, y_n) = p(y_1 \mid \theta) \times \cdots \times p(y_n \mid \theta) = \prod_{i=1}^{n} p(y_i \mid \theta)$$

- **Continuous case:**

$$L(\theta) = L(y_1, \ldots, y_n \mid \theta) = f(y_1, \ldots, y_n) = f(y_1 \mid \theta) \times \cdots \times f(y_n \mid \theta) = \prod_{i=1}^{n} f(y_i \mid \theta)$$

Theorem 9.3.1. Factorization Theorem
Let U be a statistic based on the random sample $Y_1, Y_2, \ldots, Y_n$. Then U is a sufficient statistic for the estimation of a parameter θ if and only if the likelihood $L(\theta) = L(y_1, y_2, \ldots, y_n \mid \theta)$ can be factored into two nonnegative functions,

$$L(y_1, y_2, \ldots, y_n \mid \theta) = g(U, \theta) \times h(y_1, y_2, \ldots, y_n),$$

where $g(u, \theta)$ is a function only of u and θ and $h(y_1, y_2, \ldots, y_n)$ is not a function of θ.

9.4 Completeness and Uniqueness

- Let $Y_1, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} f_\theta$
- **Notation:** $g(U) = 0$ a.e. means $P(\{U : g(U) \neq 0\}) = 0$

Definition 9.4.1. A statistic $U = U(Y_1, \ldots, Y_n)$ is said to be **complete** if and only if for every function $g(U)$ such that

$$E(g(U)) = 0 \text{ for all } \theta \implies g(U) = 0 \text{ a.e.}$$

Theorem 9.4.1. Under completeness, any unbiased estimator of is unique.

Proof. If $E(g_1(U)) = \theta$ and $E(g_2(U)) = \theta$, then $g_1(U)$ and $g_2(U)$ are unbiased estimators of θ. Then it follows that

$$E(g_1(U) - g_2(U)) = \theta - \theta = 0$$

By completeness, $g_1(U) - g_2(U) = 0$ a.e. $\Rightarrow g_1(U) = g_2(U)$ a.e. □

How to find a complete and a sufficient statistic?
Two methods:

- **Method I:** *(not required)*
  1. Obtain a sufficient statistic by the factorization theorem.
  2. Check completeness via definition: $E(g(U)) = 0$ for all θ implies that $g(U) = 0$ a.e.
  3. It is often challenging to directly verify completeness from the definition.
- **Method II:** Using the Exponential Family property, as described on the next page.

9.5 Exponential Family

Definition 9.5.1. Exponential Family
Consider the pdf/pmf $f(\cdot \mid \theta)$.

1. $f(y \mid \theta)$ is said to be a member of the **exponential class** if it can be written in the form:

$$f(y \mid \theta) = e^{p(\theta)K(y)+q(\theta)+S(y)}, \quad \text{if } a < y < b,$$

2. $f(y \mid \theta)$ is said to be a **regular** member of the exponential class (family) if, in addition to (1) above:
   (a) a and b are free of θ
   (b) $p(\theta)$ and $K(y)$ are non-trivial [$\iff$ $p'(\theta) \neq 0$ and $K'(y) \neq 0$]
   (c) $S(y)$ is continuous

Theorem 9.5.1. Consider the pdf/pmf $f(y \mid \theta)$, where $\theta \in \Omega$. If $f(y \mid \theta)$ is a regular member of the exponential class, then $U = \sum_{i=1}^{n} K(y_i)$ is sufficient and complete.

9.6 The Rao-Blackwell Theorem

Definition 9.6.1. Let $Y_1, \ldots, Y_n$ be a random sample of size n from $f(y \mid \theta)$. An estimator $T = T(Y_1, Y_2, \ldots, Y_n)$ is called a **minimum variance unbiased estimator (MVUE)** of the parameter θ if T is unbiased, that is, $E(T) = \theta$, and if the variance of T is less than or equal to the variance of every other unbiased estimator of θ.

Theorem 9.6.1. The Rao-Blackwell Theorem
Let $\hat{\theta}$ be an unbiased estimator for θ such that $V(\hat{\theta}) < \infty$. If U is a sufficient statistic for θ, define $\hat{\theta}^* = E(\hat{\theta} \mid U)$. Then, for all θ,

$$E(\hat{\theta}^*) = \theta \quad \text{and} \quad V(\hat{\theta}^*) \leq V(\hat{\theta}).$$

That is, $\theta^* = E(\text{unbiased}|\text{sufficient})$ is a MVUE.

Rao-Blackwell Theorem gives a procedure to improve two given estimators as follows:

1. Start with an unbiased estimator $\hat{\theta}$ for a parameter θ and the sufficient statistic U obtained through the factorization criterion.
2. Then $\hat{\theta}^* = E(\hat{\theta} \mid U)$ is **"better"** than $\hat{\theta}$ since it is unbiased as $\hat{\theta}$ and has variance less than that of $\hat{\theta}$ and is **"better"** than U since it is a function of the sufficient statistic and unbiased.

9.7 Unique Minimum Variance Unbiased Estimator (UMVUE)

Theorem 9.7.1. Lehmann-Scheffé Theorem
Suppose $Y_1, Y_2, \ldots, Y_n$ are i.i.d. with pdf $f(y \mid \theta)$. If $U_1 = U_1(Y_1, \ldots, Y_n)$ is an unbiased estimator of θ and $U_2 = U_2(Y_1, \ldots, Y_n)$ is sufficient and complete, then $E(U_1 \mid U_2)$ is the Unique Minimum Variance Unbiased Estimator (UMVUE) of θ.

That is $E(\text{unbiased}|\text{sufficient and complete}) = \text{UMVUE}$.

It is unique because of the completeness property.
It has minimum variance by Rao-Blackwell Theorem.

We have two procedures to find the UMVUE:

- **Procedure 1:** Two steps
    1. Get U_1 and U_2 such that U_1 is an unbiased estimator and U_2 is sufficient and complete.
    2. Then $E(U_1 \mid U_2)$ is the UMVUE.
- **Procedure 2:** Two steps
    1. Find a complete and sufficient statistic U.
    2. Find $\phi(U)$ such that $E(\phi(U)) = \theta$. Then $\phi(U)$ is the UMVUE.

9.8 Cramer-Rao Theorem

Theorem 9.8.1. Cramer-Rao Theorem
Let $Y_1, Y_2, \ldots, Y_n$ be i.i.d. $f_\theta(\cdot) = f(\cdot|\theta)$. Let $U = U(Y_1, \ldots, Y_n)$ be an **unbiased** estimator of θ then

$$\text{Var}(U) = E(U-\theta)^2 \geq \frac{1}{E\left(\frac{\partial}{\partial\theta}\ln(L(\theta))\right)^2}$$

where $L(\theta) = L(y_1, y_2, \ldots, y_n|\theta) = f(y_1, y_2, \ldots, y_n) = \prod_{i=1}^{n} f(y_i|\theta)$

- The above theorem gives the following **procedure**: if U is an unbiased estimator and if $\text{Var}(U) = \frac{1}{E\left(\frac{\partial}{\partial\theta}\ln(L(\theta))\right)^2}$, then U is a MVUE for θ.
- The quantity $\frac{1}{E\left(\frac{\partial}{\partial\theta}\ln(L(\theta))\right)^2}$ is called the **Cramer-Rao lower bound**.

Definition 9.8.1. The quantity

$$\frac{\partial}{\partial\theta}\ln(L(\theta))$$

is called the **score function**.

Theorem 9.8.2.

$$E(\text{score function}) = E\left[\frac{\partial}{\partial\theta}\ln(L(\theta))\right] = 0$$

Proof. First we illustrate the proof in a single variable case. Let $Y \sim f(y|\theta)$. Then we have

$$\int_{S_y} f(y|\theta)\,dy = 1$$

Taking the derivative with respect to θ

$$\frac{\partial}{\partial\theta}\int_{S_y} f(y|\theta)\,dy = 0$$

Now, if S_y does not depend on θ, we can exchange the derivative and the integral

$$\int_{S_y} \frac{\partial}{\partial\theta} f(y|\theta)\,dy = 0$$

Now note that

$$\frac{d}{dz}\ln(f(z)) = \frac{1}{f(z)}\cdot f'(z)$$
$$\frac{d}{dz}\ln(f(z)) = \frac{\frac{d}{dz}f(z)}{f(z)}$$
$$\frac{d}{dz}\ln(f(z))\cdot f(z) = \frac{d}{dz}f(z)$$

Thus this can be rewritten as

$$\int_{S_y} \frac{\partial}{\partial \theta} \ln(f(y|\theta)) \cdot f(y|\theta)\, dy = 0$$

Since $f(y|\theta)$ is the pdf, the expectation with respect to $f(y|\theta)$ gives

$$E\left[\frac{\partial}{\partial \theta} \ln(f(y|\theta))\right] = 0$$

The proof for the multivariate case follows similarly.
Given that

$$L(\theta) = \prod_{i=1}^{n} f(y_i|\theta) = f(y_1, \ldots, y_n|\theta) \quad \text{(assuming the independence of pdf's)}$$

The integral over the support $S_{y_1}, \ldots, S_{y_n}$ gives

$$\int_{S_{y_1}} \cdots \int_{S_{yn}} L(\theta)\, dy_1 \ldots dy_n = 1$$

Taking the derivative with respect to θ

$$\frac{\partial}{\partial \theta} \int_{S_{y_1}} \cdots \int_{S_{yn}} L(\theta)\, dy_1 \ldots dy_n = 0$$

As S_{y_i} does not depend on θ

$$\int_{S_{y_1}} \cdots \int_{S_{yn}} \frac{\partial}{\partial \theta} L(\theta)\, dy_1 \ldots dy_n = 0$$

This can be rewritten as

$$\int_{S_{y_1}} \cdots \int_{S_{yn}} \frac{\partial}{\partial \theta} \ln(L(\theta)) \cdot L(\theta)\, dy_1 \ldots dy_n = 0$$

Since $L(\theta)$ is the pdf

$$E\left[\frac{\partial}{\partial \theta} \ln(L(\theta))\right] = 0$$

□

Technique For Finding Expectation of Distributions
Let Y be a distribution that depends on a parameter θ then we have $E\left[\frac{\partial}{\partial \theta} \ln(f(y|\theta))\right] = 0$. Now we can compute $\frac{\partial}{\partial \theta} \ln(f(y|\theta))$ explicitly and this should result in a linear function of y and thus using the linearity of expectation we can compute the expectation.

Example: Poisson(λ)

$$E[Y|\lambda] = \frac{e^{-\lambda}\lambda^y}{y!}, \quad y = 0, 1, 2, \ldots$$

The log-likelihood function is:

$$\ln(f(y|\theta)) = -\lambda + y\ln\lambda - \ln y!$$

Taking the derivative with respect to λ:

$$\frac{\partial}{\partial\lambda}\ln(f(y|\theta)) = -1 + \frac{y}{\lambda} + 0$$

The expectation of the derivative is:

$$E\left[\frac{\partial}{\partial\lambda}\ln(f(y|\theta))\right] = E\left[\frac{y}{\lambda} - 1\right] = 0$$

This implies:

$$E\left[\frac{y}{\lambda}\right] = 1 \quad \Rightarrow \quad \frac{1}{\lambda}E[Y] = 1$$

Thus, $E[Y] = \lambda$.

Theorem 9.8.3.

$$E\left(\frac{\partial}{\partial\theta}\ln(L(\theta))\right)^2 = \mathrm{Var}\left(\frac{\partial}{\partial\theta}\ln(L(\theta))\right)$$

Proof. Use the fact that $E(\frac{\partial}{\partial\theta}\ln(L(\theta))) = 0$ □

Definition 9.8.2. The quantity $E\left(\frac{\partial}{\partial\theta}\ln(L(\theta))\right)^2 = \mathrm{Var}\left(\frac{\partial}{\partial\theta}\ln(L(\theta))\right)$ is called the **Fisher Information** and is denoted by $I_n(\theta)$.

In what follows we will develop some equivalent formulations of the Fisher Information.

Theorem 9.8.4.

$$I_n(\theta) = -E\left(\frac{\partial^2}{\partial\theta^2}\ln(L(\theta))\right)$$

Proof. First note that

$$E[\text{Score function}] = 0$$

$$\int_{\mathcal{S}_Y}\frac{\partial}{\partial\theta}\ln(L(\theta))L(\theta)\,dy = 0$$

Differentiating both sides with respect to θ, we obtain

$$\int_{\mathcal{S}_Y}\frac{\partial^2}{\partial\theta^2}\ln(L(\theta))L(\theta)\,dy + \int_{\mathcal{S}_Y}\frac{\partial}{\partial\theta}\ln(L(\theta))\cdot\frac{\partial L(\theta)}{\partial\theta}\,dy = 0$$

Notice that the second term can be rewritten as

$$\int_{\mathcal{S}_Y} \frac{\partial}{\partial\theta} \ln(L(\theta)) \cdot \frac{\frac{\partial L(\theta)}{\partial\theta}}{L(\theta)} \cdot L(\theta)\, dy$$

This simplifies to

$$E\left[\frac{\partial^2}{\partial\theta^2} \ln(L(\theta))\right] + E\left[\left(\frac{\partial}{\partial\theta} \ln(L(\theta))\right)^2\right] = 0$$

Therefore

$$I_n(\theta) = -E\left[\frac{\partial^2}{\partial\theta^2} \ln(L(\theta))\right]$$

□

Definition 9.8.3. We define

$$I(\theta) = -E\left(\frac{\partial^2}{\partial\theta^2} \ln f(y_i|\theta)\right) \text{ or } I(\theta) = E\left(\frac{\partial}{\partial\theta} \ln f(y_i|\theta)\right)^2$$

Note it is easily provable that the two defintions of $I(\theta)$ are equivalent by applying the proof of theorem 9.8.4 with $L(\theta)$ replaced by $f(y_i|\theta)$.

Theorem 9.8.5.

$$I_n(\theta) = nI(\theta)$$

Proof. We have

$$L(\theta) = \prod_{i=1}^{n} f(y_i \mid \theta)$$

Taking the natural logarithm we see

$$\ln L(\theta) = \sum_{i=1}^{n} \ln f(y_i \mid \theta)$$

Differentiating with respect to θ, we obtain

$$\frac{\partial}{\partial\theta} \ln L(\theta) = \sum_{i=1}^{n} \frac{\partial}{\partial\theta} \ln f(y_i \mid \theta)$$

Now, the Fisher information $I_n(\theta)$ is given by the variance of the score function

$$I_n(\theta) = \text{Var}\left[\frac{\partial}{\partial\theta} \ln L(\theta)\right]$$

Substituting the expression for the score function, we get

$$I_n(\theta) = \text{Var}\left[\sum_{i=1}^{n} \frac{\partial}{\partial\theta} \ln f(y_i \mid \theta)\right]$$

Since the observations are independent, the variance of the sum is the sum of the variances

$$I_n(\theta) = \sum_{i=1}^{n} \text{Var}\left[\frac{\partial}{\partial\theta} \ln f(y_i \mid \theta)\right]$$

By the independence assumption, each term has the same variance $I(\theta)$, so

$$I_n(\theta) = \sum_{i=1}^{n} I(\theta) = nI(\theta)$$

Thus, we have

$$I_n(\theta) = nI(\theta)$$

□

How to find estimators?

There are two main methods for finding estimators:

1. Method of Moments
2. Method of Maximum Likelihood

9.9 Method of Moments (MME)

The **method of moments** is a very simple procedure for finding an estimator for one or more population parameters.

Recall that: the kth moment of a random variable is $E(Y^k)$.

The corresponding kth sample moment is $\frac{1}{n}\sum_{i=1}^{n} Y_i^k$.

Method of Moments:

1. Equate population moments to sample moments
2. Solve for θ

Specifically we set

$$E[Y] = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

$$E[Y^2] = \frac{1}{n}\sum_{i=1}^{n} Y_i^2$$

$$E[Y^3] = \frac{1}{n}\sum_{i=1}^{n} Y_i^3$$

and so on.

Now in practice we usually have only one or two moments to work with. More specifically we set

$$\mu = E(Y) = \overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

and equivalent to the second moment we have

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

Note that sometimes the MME is not accurate. Thus this leads us to develop the following method.

9.10 Method of Maximum Likelihood (MLE)

Definition 9.10.1. Let $L(\theta)$ be the sample likelihood function. The **maximum likelihood estimator (MLE)** of θ is the value of θ that maximizes $L(\theta)$. That is, $\hat{\theta}$ is the MLE of θ if and only if

$$\hat{\theta} = \sup_{\theta} L(\theta)$$

The idea behind the MLE is to select the value within the parameter space that yields the highest "likelihood" for the observed data. In other words, we aim to find the value of θ that makes the observed data most probable.

Definition 9.10.2. We define $l(\theta) = \ln(L(\theta))$ as the **log-likelihood function**.

The maximum of the likelihood function is the same as the maximum of the log-likelihood function. This is because the logarithm is a monotonic function. It is often easier to maximize $l(\theta)$.

We find the MLE using one of the following cases:

- **CASE 1:** $l(\theta)$ is differentiable and support is free of θ.
  - Find θ such that $\frac{d}{d\theta} l(\theta) = 0$. Denote the obtained θ by $\hat{\theta}$.
  - Check that $\hat{\theta}$ is a maximum by the 2nd derivative test (only if asked):

$$\frac{d^2}{d\theta^2} l(\theta)\Big|_{\theta=\hat{\theta}} < 0$$

- **CASE 2:** $l(\theta)$ is NOT differentiable or support depends on θ.
  - Use inspection/monotonicity.

Theorem 9.10.1. The MLE has the invariance property. That is if $\hat{\theta}$ is the MLE of θ, then for any function $g(\theta)$, the MLE of $g(\theta)$ is $g(\hat{\theta})$.

It is important to note that the MLE is not unique and not always unbiased. Additionally the MLE is a function of a sufficient statistic.

10 Chapter 10 — Hypothesis Testing

10.1 Elements of a Statistical Test

Many times, the objective of a statistical test is to test a hypothesis concerning the values of one or more population parameters.

The structure of a problem on **hypothesis testing** is as follows:
Let Ω denote the total parameter space (i.e., the set of all possible values of the parameter θ). If $Y_1, Y_2, \ldots, Y_n$ are i.i.d with pdf $f_\theta(\cdot) = f(\cdot \mid \theta)$, where $\theta \in \Omega$. If $\{\Omega_0, \Omega_a\}$ is a partition for Ω [i.e., $\Omega = \Omega_0 \cup \Omega_a$ and $\Omega_0 \cap \Omega_a = \varnothing$], we need to check (test) whether $\theta \in \Omega_0$ or $\theta \in \Omega_a$.

Definition 10.1.1. A statistical hypothesis is an "assertion" about a specific distribution under consideration.

Definition 10.1.2. $\theta \in \Omega_0$ is called the **"null hypothesis"** and is denoted by H_0.

Definition 10.1.3. $\theta \in \Omega_a$ is called the **"alternative hypothesis"** and is denoted by H_a (or H_1).

Thus, we want to test

$$H_0 : \theta \in \Omega_0 \quad \text{vs} \quad H_a : \theta \in \Omega_a.$$

If Ω_0 (or Ω_a) is a **singleton** (i.e., contains one point), then H_0 (or H_a) is called **"simple hypothesis"**. Otherwise, it is called **"composite"**.

We will consider the following cases:

1. testing a simple hypothesis H_0 versus a simple alternative H_a.
2. testing a simple hypothesis H_0 versus a composite alternative H_a.
3. testing a composite hypothesis H_0 versus a composite alternative H_a.

Definition 10.1.4. A statistical test is a "rule" that determines when to accept or to reject H_0.

Definition 10.1.5. The test statistic is a function based on the data. It is used to reject or to accept H_0.

A statistical test determines a "critical region" or "rejection region"

- **Critical region** is denoted C
- **Rejection region** denoted by RR

Definition 10.1.6. The **Critical region** or **Rejection region** is defined as the set of sample points that lead to us rejecting the null hypothesis. That is

$$\text{Reject } H_0 \text{ if } (y_1, \ldots, y_n) \in C$$

There are 4 possibilities on the truth of the null hypothesis and the decision made by the test:

H_0	**True**	**False**
Reject	Type I Error	✓
Accept	✓	Type II Error

Definition 10.1.7. A **type I error** is made if H_0 is rejected when H_0 is true. The probability of a type I error is denoted by α. The value of α is called the level of the test. (α is also called the size of the test and the significance level).

In other words a **type I error** is a false negative.

Definition 10.1.8. A **type II error** is made if H_0 is accepted when H_a is true (H_0 is not true). The probability of a type II error is denoted by β.

In other words a **type II error** is a false positive.

Thus,

$$\begin{aligned}\alpha &= P(\text{false negative}) \\ &= P(\text{reject } H_0 \text{ when } H_0 \text{ is true}) \\ &= P(Y_1, \ldots, Y_n \in RR \mid H_0 \text{ is true})\end{aligned}$$

and

$$\begin{aligned}\beta &= P(\text{false positive}) \\ &= P(\text{accept } H_0 \text{ when } H_0 \text{ is false}) \\ &= P(Y_1, \ldots, Y_n \notin RR \mid H_a \text{ is true})\end{aligned}$$

Definition 10.1.9. The **power of a test** is defined as

$$\begin{aligned}\text{power}(\theta) = W(\theta) = 1 - \beta &= P(\text{reject } H_0 \text{ when } H_a \text{ is true}) \\ &= P(Y_1, \ldots, Y_n \in RR \mid H_a \text{ is true})\end{aligned}$$

10.2 Neyman-Pearson Lemma

Suppose RR is a rejection region of size α.

Now our objective is to find the "best" rejection region given a fixed value α of the level of significance, whenever it exists.

One may define the "best" rejection region as the one that maximizes the power of the test.

Case 1: Simple Hypothesis vs Simple Alternative
Consider the problem of testing a simple hypothesis $H_0 : \theta = \theta_0$ versus a simple alternative $H_a : \theta = \theta_a$.

That is, test

$$H_0 : \theta = \theta_0 \quad \text{vs} \quad H_a : \theta = \theta_a$$

at significance level α.

The solution to this problem is given by the following theorem

Theorem 10.2.1. Neyman-Pearson Lemma
Suppose that we wish to test the **simple** null hypothesis $H_0 : \theta = \theta_0$ versus the **simple** alternative hypothesis $H_a : \theta = \theta_a$, based on a random sample $Y_1, Y_2, \ldots, Y_n$ from a distribution with parameter θ. Let $L(\theta)$ denote the likelihood of the sample when the value of the parameter is θ. Then, for a given α, the test that maximizes the power at θ_a has a rejection region, RR, determined by

$$\frac{L(\theta_0)}{L(\theta_a)} < k.$$

The value of k is chosen so that the test has the desired value for α. Such a test is a most powerful α-level test for H_0 versus H_a.

What this means in practice is that we can take the set created by $\left\{\frac{L(\theta_0)}{L(\theta_a)} < k\right\} \subseteq S_{Y_1,\ldots,Y_n}$ as the rejection region. Note that this works as $\frac{L(\theta_0)}{L(\theta_a)}$ is a function of $y_1, \ldots, y_n$.
The only issue is that this rejection region depends on the unknown k. To find this k we set $\alpha = P(Y_1, \ldots, Y_n \in RR \mid H_0 \text{ is true})$ and now as α is fixed and RR depends on k we can solve for k.

Case 2: Simple Hypothesis vs Composite Alternative
When testing a simple H_0 vs a composite H_a, we need to test

$$H_0 : \theta \in \Omega_0 \quad \text{vs} \quad H_a : \theta \in \Omega_a$$

at the level of significance α, where Ω_0 is a singleton and Ω_a is not a singleton.

Consider

$$H_0 : \theta = \theta_0 \quad \text{vs} \quad H_a : \theta > \theta_a$$

Use Neyman-Pearson Lemma (which is the most powerful level α test for testing $H_0 : \theta = \theta_0$ vs $H_a : \theta = \theta_a$). If the resulting test is independent of θ_a (i.e., we get exactly the same test for all $\theta_a \in \Omega_a$), then the resulting test is said to be the "**Uniformly Most Powerful (UMP)**" level α test.

If the test obtained from the Neyman-Pearson Lemma "changes with" (or depends on) θ_a, then there does not exist a UMP test for testing the hypothesis

$$H_0 : \theta = \theta_0 \quad \text{vs} \quad H_a : \theta \in \Omega_a.$$