Chapter 7

7.1 SUMS OF RANDOM VARIABLES

Let $X_1, X_2, \ldots, X_n$ be a sequence of random variables, and let S_n be their sum:

$$S_n = X_1 + X_2 + \cdots + X_n. \tag{7.1}$$

In this section, we find the mean and variance of S_n, as well as the pdf of S_n in the important special case where the X_j's are independent random variables.

7.1.1 Mean and Variance of Sums of Random Variables

In Section 6.3, it was shown that *regardless of statistical dependence, the expected value of a sum of n random variables is equal to the sum of the expected values:*

$$E[X_1 + X_2 + \cdots + X_n] = E[X_1] + \cdots + E[X_n]. \tag{7.2}$$

Thus knowledge of the means of the X_j's suffices to find the mean of S_n.

The following example shows that in order to compute the variance of a sum of random variables, we need to know the variances and covariances of the X_j's.

Example 7.1

Find the variance of $Z = X + Y$.

From Eq. (7.2), $E[Z] = E[X + Y] = E[X] + E[Y]$. The variance of Z is therefore

$$\begin{aligned}
\text{VAR}(Z) &= E[(Z - E[Z])^2] = E[(X + Y - E[X] - E[Y])^2] \\
&= E[\{(X - E[X]) + (Y - E[Y])\}^2] \\
&= E[(X - E[X])^2 + (Y - E[Y])^2 + (X - E[X])(Y - E[Y]) \\
&\quad + (Y - E[Y])(X - E[X])] \\
&= \text{VAR}[X] + \text{VAR}[Y] + \text{COV}(X, Y) + \text{COV}(Y, X) \\
&= \text{VAR}[X] + \text{VAR}[Y] + 2\,\text{COV}(X, Y).
\end{aligned}$$

In general, the covariance $\text{COV}(X, Y)$ is not equal to zero, so the variance of a sum is not necessarily equal to the sum of the individual variances.

The result in Example 7.1 can be generalized to the case of n random variables:

$$\begin{aligned}
\text{VAR}(X_1 + X_2 + \cdots + X_n) &= E\left\{\sum_{j=1}^{n}(X_j - E[X_j])\sum_{k=1}^{n}(X_k - E[X_k])\right\} \\
&= \sum_{j=1}^{n}\sum_{k=1}^{n} E[(X_j - E[X_j])(X_k - E[X_k])] \\
&= \sum_{k=1}^{n}\text{VAR}(X_k) + \sum_{\substack{j=1 \\ j\neq k}}^{n}\sum_{k=1}^{n}\text{COV}(X_j, X_k).
\end{aligned} \tag{7.3}$$

Thus *in general, the variance of a sum of random variables is not equal to the sum of the individual variances.*

An important special case is when the X_j's are independent random variables. *If $X_1, X_2, \ldots, X_n$ are independent random variables*, then $\text{COV}(X_j, X_k) = 0$ for $j \neq k$ and

$$\text{VAR}(X_1 + X_2 + \cdots + X_n) = \text{VAR}(X_1) + \cdots + \text{VAR}(X_n). \tag{7.4}$$

Example 7.2 Sum of iid Random Variables

Find the mean and variance of the sum of n **independent, identically distributed** (iid) random variables, each with mean μ and variance σ^2.

The mean of S_n is obtained from Eq. (7.2):

$$E[S_n] = E[X_1] + \cdots + E[X_n] = n\mu.$$

The covariance of pairs of independent random variables is zero, so by Eq. (7.4),

$$\text{VAR}[S_n] = n\,\text{VAR}[X_j] = n\sigma^2,$$

since $\text{VAR}[X_j] = \sigma^2$ for $j = 1, \ldots, n$.

THE SAMPLE MEAN AND THE LAWS OF LARGE NUMBERS

Let X be a random variable for which the mean, $E[X] = \mu$, is unknown. Let $X_1, \ldots, X_n$ denote n independent, repeated measurements of X; that is, the X_j's are **independent, identically distributed** (iid) random variables with the same pdf as X. The **sample mean** of the sequence is used to estimate $E[X]$:

$$M_n = \frac{1}{n}\sum_{j=1}^{n} X_j. \tag{7.15}$$

In this section, we compute the expected value and variance of M_n in order to assess the effectiveness of M_n as an estimator for $E[X]$. We also investigate the behavior of M_n as n becomes large.

The following example shows that the relative frequency estimator for the probability of an event is a special case of a sample mean. Thus the results derived below for the sample mean are also applicable to the relative frequency estimator.

The sample mean is itself a random variable, so it will exhibit random variation. A good estimator should have the following two properties: (1) On the average, it should give the correct value of the parameter being estimated, that is, $E[M_n] = \mu$; and (2) It should not vary too much about the correct value of this parameter, that is, $E[(M_n - \mu)^2]$ is small.

The expected value of the sample mean is given by

$$E[M_n] = E\left[\frac{1}{n}\sum_{j=1}^{n} X_j\right] = \frac{1}{n}\sum_{j=1}^{n} E[X_j] = \mu, \tag{7.17}$$

since $E[X_j] = E[X] = \mu$ for all j. Thus the sample mean is equal to $E[X] = \mu$, on the average. For this reason, we say that the sample mean is an **unbiased estimator** for μ.

Equation (7.17) implies that the mean square error of the sample mean about μ is equal to the variance of M_n, that is,

$$E[(M_n - \mu)^2] = E[(M_n - E[M_n])^2].$$

Note that $M_n = S_n/n$, where $S_n = X_1 + X_2 + \cdots + X_n$. From Eq. (7.4), $\mathrm{VAR}[S_n] = n\,\mathrm{VAR}[\mathrm{X_j}] = n\sigma^2$, since the X_j's are iid random variables. Thus

$$\mathrm{VAR}[M_n] = \frac{1}{n^2}\mathrm{VAR}[S_n] = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}. \tag{7.18}$$

Equation (7.18) states that the variance of the sample mean approaches zero as the number of samples is increased. This implies that the probability that the sample mean is close to the true mean approaches one as n becomes very large. We can formalize this statement by using the Chebyshev inequality, Eq. (4.76):

3 THE CENTRAL LIMIT THEOREM

Let $X_1, X_2, \ldots$ be a sequence of iid random variables with finite mean μ and finite variance σ^2, and let S_n be the sum of the first n random variables in the sequence:

$$S_n = X_1 + X_2 + \cdots + X_n. \tag{7.25}$$

In Section 7.1, we developed methods for determining the exact pdf of S_n. We now present the central limit theorem, which states that, as n becomes large, the cdf of a properly normalized S_n approaches that of a Gaussian random variable. This enables us to approximate the cdf of S_n with that of a Gaussian random variable.

The central limit theorem explains why the Gaussian random variable appears in so many diverse applications. In nature, many macroscopic phenomena result from the addition of numerous independent, microscopic processes; this gives rise to the Gaussian random variable. In many man-made problems, we are interested in averages that often consist of the sum of independent random variables. This again gives rise to the Gaussian random variable.

From Example 7.2, we know that if the X_j's are iid, then S_n has mean $n\mu$ and variance $n\sigma^2$. The central limit theorem states that the cdf of a suitably normalized version of S_n approaches that of a Gaussian random variable.

Central Limit Theorem Let S_n be the sum of n iid random variables with finite mean $E[X] = \mu$ and finite variance σ^2, and let Z_n be the zero-mean, unit-variance random variable defined by

$$Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}}, \tag{7.26a}$$

then

$$\lim_{n\to\infty} P[Z_n \leq z] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-x^2/2}\, dx. \tag{7.26b}$$

Note that Z_n is sometimes written in terms of the sample mean:

$$Z_n = \sqrt{n}\frac{M_n - \mu}{\sigma}. \tag{7.27}$$

The amazing part about the central limit theorem is that the summands X_j can have *any* distribution as long as they have a finite mean and finite variance. This gives the result its wide applicability.

Figures 7.2 through 7.4 compare the exact cdf and the Gaussian approximation for the sums of Bernoulli, uniform, and exponential random variables, respectively. In all three cases, it can be seen that the approximation improves as the number of terms in the sum increases. The proof of the central limit theorem is discussed in the last part of this section.

Chapter 8

Statistical Inference and Estimation

Let $X_1, X_2, \ldots, X_n$ denote a set of n random sam
from a population, the objective is to formulate a stati
computed from the sample data to be used for estimatio
of population parameters. Common point estimation:

Parameters: μ Population mean

σ^2 population variance

statistics: $\bar{X} = \frac{1}{n}\sum_{j=1}^{n} X_j$ sample mean

S^2 sample variance

Note: earlier $M_n = \frac{1}{n} \cdot S_n$, where $S_n = X_1 + X_2 + \cdots + X_n$
and M_n is the sample mean, $M_n = \frac{S_n}{n}$

$$E[M_n] = E\left[\frac{1}{n}\sum_{j=1}^{n} X_j\right] = \frac{1}{n}\sum_{j=1}^{n} E[X_j] = \mu$$

$$Var[M_n] = E\left[(M_n - E[M_n])^2\right]$$

$$= VAR\left[\frac{S_n}{n}\right] = \frac{1}{n^2} Var[S_n] = \frac{1}{n^2}\left[n\sigma^2\right]$$

$$= \frac{\sigma^2}{n}$$

1 Properties of Estimators

Ideally, a good estimator should be equal to the parameter θ, on the average. We say that the estimator $\hat{\Theta}$ is an **unbiased estimator** for θ if

$$E[\hat{\Theta}] = \theta. \tag{8.11}$$

The **bias** of any estimator $\hat{\Theta}$ is defined by

$$B[\hat{\Theta}] = E[\hat{\Theta}] - \theta. \tag{8.12}$$

From Eq. (8.4) in Example 8.1, we see that *the sample mean is an unbiased estimator for the mean* μ. However, biased estimators are not unusual as illustrated by the following example.

Example 8.5 The Sample Variance

The sample mean gives us an estimate of the center of mass of observations of a random variable. We are also interested in the spread of these observations about this center of mass. An obvious estimator for the variance σ_X^2 of X is the arithmetic average of the square variation about the sample mean:

$$\hat{S}^2 = \frac{1}{n}\sum_{j=1}^{n}(X_j - \overline{X}_n)^2 \tag{8.13}$$

where the sample mean is given by:

$$\overline{X}_n = \frac{1}{n}\sum_{j=1}^{n}X_j. \tag{8.14}$$

Let's check whether $\hat{S}^2$ is an unbiased estimator. First, we rewrite Eq. (8.13):

$$\begin{aligned}
\hat{S}^2 &= \frac{1}{n}\sum_{j=1}^{n}(X_j - \overline{X}_n)^2 = \frac{1}{n}\sum_{j=1}^{n}(X_j - \mu + \mu - \overline{X}_n)^2 \\
&= \frac{1}{n}\sum_{j=1}^{n}\left\{(X_j - \mu)^2 + 2(X_j - \mu)(\mu - \overline{X}_n) + (\mu - \overline{X}_n)^2\right\} \\
&= \frac{1}{n}\sum_{j=1}^{n}(X_j - \mu)^2 + \frac{2}{n}(\mu - \overline{X}_n)\sum_{j=1}^{n}(X_j - \mu) + \frac{1}{n}\sum_{j=1}^{n}(\mu - \overline{X}_n)^2 \\
&= \frac{1}{n}\sum_{j=1}^{n}(X_j - \mu)^2 + \frac{2}{n}(\mu - \overline{X}_n)(n\overline{X}_n - n\mu) + \frac{n(\mu - \overline{X}_n)^2}{n} \\
&= \frac{1}{n}\sum_{j=1}^{n}(X_j - \mu)^2 - 2(\overline{X}_n - \mu)^2 + (\overline{X}_n - \mu)^2 \\
&= \frac{1}{n}\sum_{j=1}^{n}(X_j - \mu)^2 - (\overline{X}_n - \mu)^2.
\end{aligned} \tag{8.15}$$

The expected value of $\hat{S}^2$ is then:

$$E[\hat{S}^2] = E\left[\frac{1}{n}\sum_{j=1}^{n}(X_j - \mu)^2 - (\overline{X}_n - \mu)^2\right]$$

$$= \frac{1}{n}\sum_{j=1}^{n}[E[(X_j - \mu)^2] - E[(\overline{X}_n - \mu)^2]]$$

$$= \sigma_X^2 - \frac{\sigma_X^2}{n} = \frac{n-1}{n}\sigma_X^2 \tag{8.16}$$

where we used Eq. (8.2) for the variance of the sample mean. Equation (8.16) shows that the simple estimator given by Eq. (8.13) is a *biased* estimator for the variance. We can obtain an **unbiased estimator for** σ_X^2 by dividing the sum in Eq. (8.15) by $n - 1$ instead of by n:

$$\hat{\sigma}_n^2 = \frac{1}{n-1}\sum_{j=1}^{n}(X_j - \overline{X}_n)^2. \tag{8.17}$$

Equation (8.17) is used as the standard estimator for the variance of a random variable.

Basic Concepts

Definition

Independent random variables $X_1, X_2, \ldots, X_n$ with the same distribution are called a **random sample.**

Definition

A **statistic** is a function of a random sample.

Definition

The probability distribution of a statistic is called its **sampling distribution.**

The sample mean estimator $\overline{X}_n$ provides us with a single numerical value for the estimate of $E[X] = \mu$, namely,

$$\overline{X}_n = \frac{1}{n}\sum_{j=1}^{n} X_j. \tag{8.48}$$

This single number gives no indication of the accuracy of the estimate or the confidence that we can place on it. We can obtain an indication of accuracy by computing the sample variance, which is the average dispersion about $\overline{X}_n$:

$$\hat{\sigma}_n^2 = \frac{1}{n-1}\sum_{j=1}^{n} (X_j - \overline{X}_n)^2. \tag{8.49}$$

If $\hat{\sigma}_n^2$ is small, then the observations are tightly clustered about $\overline{X}_n$, and we can be confident that $\overline{X}_n$ is close to $E[X]$. On the other hand, if $\hat{\sigma}_n^2$ is large, the samples are widely dispersed about $\overline{X}_n$ and we cannot be confident that $\overline{X}_n$ is close to $E[X]$. In this section we introduce the notion of confidence intervals, which approach the question in a different way.

Instead of seeking a single value that we designate to be the "estimate" of the parameter of interest (i.e., $E[X] = \mu$), we attempt to specify an *interval or set of values* that is highly likely to contain the true value of the parameter. In particular, we can specify some high probability, say $1 - \alpha$, and pose the following problem: Find an interval $[l(\mathbf{X}), u(\mathbf{X})]$ such that

$$P[l(\mathbf{X}) \leq \mu \leq u(\mathbf{X})] = 1 - \alpha. \tag{8.50}$$

In other words, we use the observed data to determine an interval that by design contains the true value of the parameter μ with probability $1 - \alpha$. We say that such an interval is a $(1 - \alpha) \times 100\%$ **confidence interval**.

This approach simultaneously handles the question of the accuracy and confidence of an estimate. The probability $1 - \alpha$ is a measure of the consistency, and hence degree of confidence, with which the interval contains the desired parameter: If we were to compute confidence intervals a large number of times, we would find that approximately $(1 - \alpha) \times 100\%$ of the time, the computed intervals would contain the true value of the parameter. For this reason, $1 - \alpha$ is called the **confidence level**. The width of a confidence interval is a measure of the accuracy with which we can pinpoint the estimate of a parameter. The narrower the confidence interval, the more accurately we can specify the estimate for a parameter.

The probability in Eq. (8.50) clearly depends on the pdf of the X_j's. In the remainder of this section, we obtain confidence intervals in the cases where the X_j's are Gaussian random variables or can be approximated by Gaussian random variables. We will use the equivalence between the following events:

$$\begin{aligned}\left\{-a \leq \frac{\overline{X}_n - \mu}{\sigma_X/\sqrt{n}} \leq a\right\} &= \left\{\frac{-a\sigma_X}{\sqrt{n}} \leq \overline{X}_n - \mu \leq \frac{a\sigma_X}{\sqrt{n}}\right\} \\ &= \left\{-\overline{X}_n - \frac{a\sigma_X}{\sqrt{n}} \leq -\mu \leq -\overline{X}_n + \frac{a\sigma_X}{\sqrt{n}}\right\} \\ &= \left\{\overline{X}_n - \frac{a\sigma_X}{\sqrt{n}} \leq \mu \leq \overline{X}_n + \frac{a\sigma_X}{\sqrt{n}}\right\}.\end{aligned}$$

The last event describes a confidence interval in terms of the observed data, and the first event will allow us to calculate probabilities from the sampling distributions.

4.1 Case 1: X_j's Gaussian; Unknown Mean and Known Variance

Suppose that the X_j's are iid Gaussian random variables with unknown mean μ and known variance σ_X^2. From Example 7.3 and Eqs. (7.17) and (7.18), $\overline{X}_n$ is then a Gaussian random variable with mean μ and variance σ_X^2/n, thus

$$\begin{aligned}1 - 2Q(z) &= P\left[-z \leq \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \leq z\right] \\ &= P\left[\overline{X}_n - \frac{z\sigma}{\sqrt{n}} \leq \mu \leq \overline{X}_n + \frac{z\sigma}{\sqrt{n}}\right].\end{aligned} \tag{8.51}$$

Equation (8.51) states that the interval $\left[\overline{X}_n - z\sigma/\sqrt{n}, \overline{X}_n + z\sigma/\sqrt{n}\right]$ contains μ with probability $1 - 2Q(z)$. If we let $z_{\alpha/2}$ be the critical value such that $\alpha = 2Q(z_{\alpha/2})$, then *the* $(1 - \alpha)$ *confidence interval for the mean* μ is given by

$$\left[\overline{X}_n - z_{\alpha/2}\sigma/\sqrt{n}, \overline{X}_n + z_{\alpha/2}\sigma/\sqrt{n}\right]. \tag{8.52}$$

The confidence interval in Eq. (8.52) depends on the sample mean $\overline{X}_n$, the known variance σ_X^2 of the X_j's, the number of measurements n, and the confidence level $1 - \alpha$. Table 8.1 shows the values of z_α corresponding to typical values of α. We can use the Octave function `normal_inv`$(1 - \alpha/2, 0, 1)$ to find $z_{\alpha/2}$. This function was introduced in Example 4.51.

When X is not Gaussian but the number of samples n is large, the sample mean $\overline{X}_n$ will be approximately Gaussian if the central limit theorem applies. *Therefore if n is large, then Eq. (8.52) provides a good approximate confidence interval.*

Example 8.17 Estimating Signal in Noise

A voltage X is given by

$$X = v + N,$$

where v is an unknown constant voltage and N is a random noise voltage that has a Gaussian pdf with zero mean, and variance $1\mu V$. Find the 95% confidence interval for v if the voltage X is measured 100 independent times and the sample mean is found to be 5.25 μV.

From Example 4.17, we know that the voltage X is a Gaussian random variable with mean v and variance 1. Thus the 100 measurements $X_1, X_2, \ldots, X_{100}$ are iid Gaussian random variables with mean v and variance 1. The confidence interval is given by Eq. (8.52) with $z_{\alpha/2} = 1.96$:

$$\left[5.25 - \frac{1.96(1)}{10}, 5.25 + \frac{1.96(1)}{10}\right] = [5.05, 5.45].$$

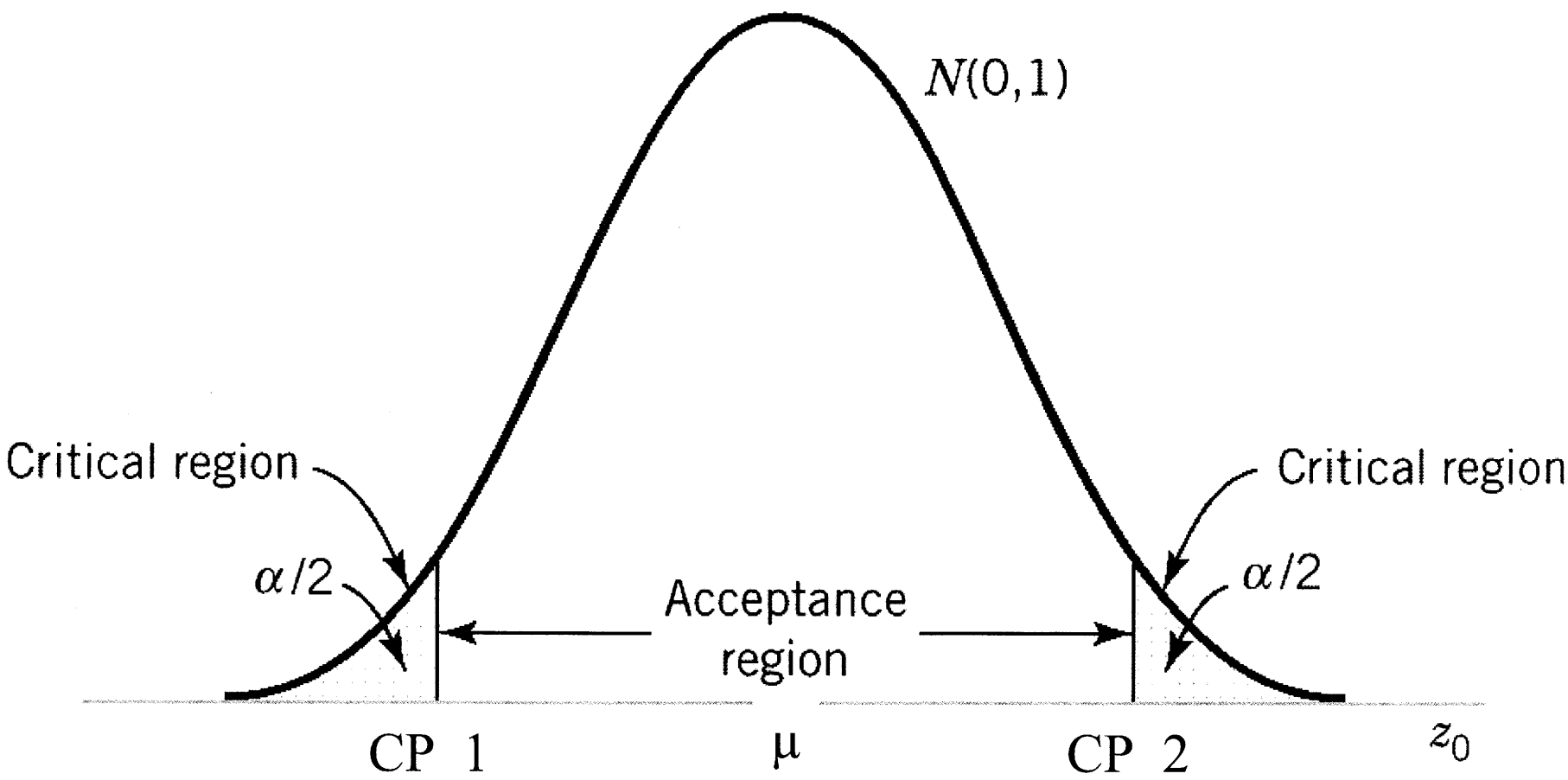


Figure 4-8 The distribution of Z_0 when H_0: $\mu = \mu_0$ is true, with critical region for H_1: $\mu \neq \mu_0$.

5 HYPOTHESIS TESTING

In some situations we are interested in testing an assertion about a population based on a random sample $\mathbf{X}_n$. This assertion is stated in the form of a hypothesis about the underlying distribution of X, and the objective of the test is to accept or reject the hypothesis based on the observed data $\mathbf{X}_n$. Examples of such assertions are:

- A given coin is fair.
- A new manufacturing process produces "new and improved" batteries that last longer.
- Two random noise signals have the same mean.

We first consider significance testing where the objective is to accept or reject a given "null" hypothesis H_0. Next we consider the testing of H_0 against an alternative hypothesis H_1. We develop decision rules for determining the outcome of each test and introduce metrics for assessing the goodness or quality of these rules.

In the general case we wish to test a hypothesis H_0 about a parameter θ of the random variable X. We call H_0 the **null hypothesis**. The objective of a **significance test** is to accept or reject the null hypothesis based on a random sample $\mathbf{X}_n = (X_1, X_2, \ldots, X_n)$. In particular we are interested in whether the observed data $\mathbf{X}_n$ is *significantly* different from what would be expected if the null hypothesis is true. To specify a decision rule we partition the observation space into a **rejection or critical region** $\tilde{R}$ where we reject the hypothesis and an *acceptance region* $\tilde{R}^c$ where we accept the hypothesis. The **decision rule** is then:

$$\begin{aligned} &\text{Accept } H_0 \quad \text{if} \quad \mathbf{X}_n \in \tilde{R}^c \\ &\text{Reject } H_0 \quad \text{if} \quad \mathbf{X}_n \in \tilde{R}. \end{aligned} \tag{8.65}$$

Two kinds of errors can occur when executing this decision rule:

$$\begin{aligned} &\textbf{Type I error:} \quad \text{Reject } H_0 \text{ when } H_0 \text{ is true.} \\ &\textbf{Type II error:} \quad \text{Accept } H_0 \text{ when } H_0 \text{ is false.} \end{aligned} \tag{8.66}$$

If the hypothesis is true, then we can evaluate the probability of a Type I error:

$$\alpha \triangleq P[\text{Type I error}] = \int_{\mathbf{x}_n \in \tilde{R}} f_{\mathbf{X}}(\mathbf{x}_n \mid H_0)\, d\mathbf{x}_n. \tag{8.67}$$

If the null hypothesis is false, we have no information about the true distribution of the observations $\mathbf{X}_n$ and hence we cannot evaluate the probability of Type II errors.

We call α the **significance level** of a test, and this value represents our tolerance for Type I errors, that is, of rejecting H_0 when in fact it is true. The level of significance of a test provides an important design criterion for testing. Specifically, the rejection region is chosen so that the probability of Type I error is no greater than a specified level α. Typical values of α are 1% and 5%.

Hypothesis testing steps

(1.) State the null & alternative Hypothesis

H_0 : what is believed to be true

H_1 : what is being claimed

e.g. the average height of a population is $\mu = 100$.

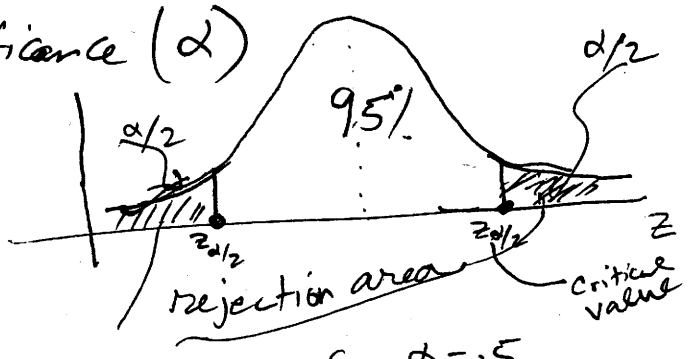
$H_0 : \mu = 100$

The alternative Hypothesis What is being claimed $H_1 : \mu \neq 100$

for $\neq$ we use 2 tailed test

for $>$ or $<$, we use 1 tailed test.

(2) Choose the level of significance (α)



for $\alpha = .5$

The two tail areas $= .025$

Therefore the acceptance area would be 95%.

(3) Find the Critical values

it could be z or t values

We use z values when σ (the population S.D.) is known

We use t values when σ unknown or sample size < 30

assume $\sigma = 15$ is given

(4) find test statistics

find the z value for our Sample

S : the sample S.D.

test statistics $= \dfrac{\bar{X} - \mu}{\sigma/\sqrt{n}}$ if σ known

$= \dfrac{\bar{X} - \mu}{S/\sqrt{n}}$ " " unknown

(5) If test statistic falls in rejection area, we reject the null hypothesis H_0.

MAXIMUM LIKELIHOOD ESTIMATION

We now consider the maximum likelihood method for finding a point estimator $\hat{\Theta}(\mathbf{X}_n)$ for an unknown parameter θ. In this section we first show how the method works. We then present several properties that make maximum likelihood estimators very useful in practice.

The maximum likelihood method selects as its estimate the parameter value that maximizes the probability of the observed data $\mathbf{X}_n = (X_1, X_2, \ldots, X_n)$. Before introducing the formal method we use an example to demonstrate the basic approach.

Example 8.9 Poisson Distributed Typos

Papers submitted by Bob have been found to have a Poisson distributed number of typos with mean 1 typo per page, whereas papers prepared by John have a Poisson distributed number of typos with mean 5 typos per page. Suppose that a page that was submitted by either Bob or John has 2 typos. Who is the likely author?

In the maximum likelihood approach we first calculate the probability of obtaining the given observation for each possible parameter value, thus:

$$P[X = 2 \mid \theta = 1] = \frac{1^2}{2!}e^{-1} = \frac{1}{2e} = 0.18394$$

$$P[X = 2 \mid \theta = 5] = \frac{5^2}{2!}e^{-5} = \frac{25}{2e^5} = 0.084224.$$

We then select the parameter value that gives the higher probability for the observation. In this case $\hat{\Theta}(2) = 1$ gives the higher probability, so the estimator selects Bob as the more likely author of the page.

Let $\mathbf{x}_n = (x_1, x_2, \ldots, x_n)$ be the observed values of a random sample for the random variable X and let θ be the parameter of interest. The **likelihood function** of the sample is a function of θ defined as follows:

$$l(\mathbf{x}_n; \theta) = l(x_1, x_2, \ldots, x_n; \theta) = \begin{cases} p_X(x_1, x_2, \ldots, x_n | \theta) & X \text{ discrete random variable} \\ f_X(x_1, x_2, \ldots, x_n | \theta) & X \text{ continuous random variable} \end{cases} \quad (8.21)$$

where $p_X(x_1, x_2, \ldots, x_n | \theta)$ and $f_X(x_1, x_2, \ldots, x_n | \theta)$ are the joint pmf and joint pdf evaluated at the observation values if the parameter value is θ. Since the samples $X_1, X_2, \ldots, X_n$ are iid, we have a simple expression for the likelihood function:

$$p_X(x_1, x_2, \ldots, x_n | \theta) = p_X(x_1 | \theta) p_X(x_2 | \theta) \ldots p_X(x_n | \theta) = \prod_{j=1}^{n} p_X(x_j | \theta) \quad (8.22)$$

and

$$f_X(x_1, x_2, \ldots, x_n | \theta) = f_X(x_1 | \theta) f_X(x_2 | \theta) \ldots f_X(x_n | \theta) = \prod_{j=1}^{n} f_X(x_j | \theta). \quad (8.23)$$

The **maximum likelihood method** selects the estimator value $\hat{\Theta} = \theta^*$ where θ^* is the parameter value that maximizes the likelihood function, that is,

$$l(x_1, x_2, \ldots, x_n; \theta^*) = \max_{\theta} l(x_1, x_2, \ldots, x_n; \theta) \quad (8.24)$$

where the maximum is taken over all allowable values of θ. Usually θ assumes a continuous set of values, so we find the maximum of the likelihood function over θ using standard methods from calculus.

It is usually more convenient to work with the **log likelihood function** because we then work with the sum of terms instead of the product of terms in Eqs. (8.22) and (8.23):

$$L(\mathbf{x}_n | \theta) = \ln l(\mathbf{x}_n; \theta) = \begin{cases} \sum_{j=1}^{n} \ln p_X(x_j | \theta) = \sum_{j=1}^{n} L(x_j | \theta) & X \text{ discrete random variable} \\ \sum_{j=1}^{n} \ln f_X(x_j | \theta) = \sum_{j=1}^{n} L(x_j | \theta) & X \text{ continuous random variable.} \end{cases} \quad (8.25)$$

Maximizing the log likelihood function is equivalent to maximizing the likelihood function since $\ln(x)$ is an increasing function of x. We obtain the maximum likelihood estimate by finding the value θ^* for which:

$$\frac{\partial}{\partial \theta} L(\mathbf{x}_n | \theta) = \frac{\partial}{\partial \theta} \ln l(\mathbf{x}_n | \theta) = 0. \quad (8.26)$$

Example 8.11 Estimation of α for Poisson random variable

Suppose we perform n independent observations of a Poisson random variable with mean α. Find the maximum likelihood estimate for α.

Let the counts in the n independent trials be given by $k_1, k_2, \ldots, k_n$. The probability of observing k_j events in the jth trial is:

$$p_X(k_j|\alpha) = \frac{\alpha^{k_j}}{k_j!}e^{-\alpha}.$$

The log likelihood function is then

$$\ln l(k_1, k_2, \ldots, k_n; \alpha) = \sum_{j=1}^{n} \ln p_X(x_j|\alpha) = \sum_{j=1}^{n} (k_j \ln \alpha - \alpha - \ln k_j!)$$

$$= \ln \alpha \sum_{j=1}^{n} k_j - n\alpha - \sum_{j=1}^{n} \ln k_j!.$$

To find the maximum, we take the first derivative with respect to α and set it equal to zero:

$$0 = \frac{d}{d\alpha} \ln l(k_1, k_2, \ldots, k_n; \alpha) = \frac{1}{\alpha} \sum_{j=1}^{n} k_j - n. \tag{8.29}$$

Solving for α, we obtain:

$$\alpha^* = \frac{1}{n} \sum_{j=1}^{n} k_j.$$

The maximum likelihood estimator for α is the sample mean of the event counts.

Example 8.12 Estimation of Mean and Variance for Gaussian Random Variable

Let $\mathbf{x}_n = (x_1, x_2, \ldots, x_n)$ be the observed values of a random sample for a Gaussian random variable X for which we wish to estimate two parameters: the mean $\theta_1 = \mu$ and variance $\theta_2 = \sigma_X^2$. The likelihood function is a function of two parameters θ_1 and θ_2, and we must simultaneously maximize the likelihood with respect to these two parameters.

The pdf for the jth observation is given by:

$$f_X(x_j \mid \theta_1, \theta_2) = \frac{1}{\sqrt{2\pi\theta_2}} e^{-(x_j-\theta_1)^2/2\theta_2}$$

where we have replaced the mean and variance by θ_1 and θ_2, respectively. The log likelihood function is given by:

$$\begin{aligned} \ln l(x_1, x_2, \ldots, x_n; \theta_1, \theta_2) &= \sum_{j=1}^{n} \ln f_X(x_j \mid \theta_1, \theta_2) \\ &= -\frac{n}{2} \ln 2\pi\theta_2 - \sum_{j=1}^{n} \frac{(x_j - \theta_1)^2}{2\theta_2}. \end{aligned}$$

We take derivatives with respect to θ_1 and θ_2 and set the results equal to zero:

$$\begin{aligned} 0 = \frac{\partial}{\partial\theta_1} \sum_{j=1}^{n} \ln f_X(x_j \mid \theta_1, \theta_2) &= -2 \sum_{j=1}^{n} \frac{(x_j - \theta_1)}{2\theta_2} \\ &= -\frac{1}{\theta_2} \left[\sum_{j=1}^{n} x_j - n\theta_1 \right] \end{aligned} \tag{8.30}$$

and

$$\begin{aligned} 0 = \frac{\partial}{\partial\theta_2} \sum_{j=1}^{n} \ln f_X(x_j \mid \theta_1, \theta_2) &= -\frac{n}{2\theta_2} + \frac{1}{2\theta_2^2} \sum_{j=1}^{n} (x_j - \theta_1)^2 \\ &= -\frac{1}{2\theta_2} \left[n - \frac{1}{\theta_2} \sum_{j=1}^{n} (x_j - \theta_1)^2 \right]. \end{aligned} \tag{8.31}$$

Equations (8.30) and (8.31) can be solved for θ_1^* and θ_2^*, respectively, to obtain:

$$\theta_1^* = \frac{1}{n} \sum_{j=1}^{n} x_j \tag{8.32}$$

$$\theta_2^* = \frac{1}{n} \sum_{j=1}^{n} (x_j - \theta_1^*)^2. \tag{8.33}$$

Thus, θ_1^* is given by the sample mean and θ_2^* is given by the biased sample variance discussed in Example 8.5. It is easy to show that as n becomes large, θ_2^* approaches the unbiased $\hat{\sigma}_n^2$.