Chapter 2

Elements of Statistical Inference

2.1 Principles of Data Reduction

Example 2.1. Cholesterol levels continued.
Suppose we want to make inference on a mean cholesterol level on the second
day after a heart attach of such a population of people in a north eastern American
state. We have got data of 28 patients, which are a realization of the random
sample of size \( n = 28 \). One of the first things we can do is to calculate the
estimate of the population mean (\( \mu \)) and of the population variance (\( \sigma^2 \)). To do
this we can use some functions of the random sample, such as the sample mean
(\( \bar{X} \)) and the sample variance (\( S^2 \)), respectively, where
\[
\bar{X} = \frac{1}{28} \sum_{i=1}^{28} X_i, \quad S^2 = \frac{1}{27} \sum_{i=1}^{28} (X_i - \bar{X})^2.
\]
Here we have
\[
\bar{x} = 257 \quad \text{and} \quad s^2 = 32.
\]
Note that \( \bar{X} \) and \( S^2 \) are random variables, as they are functions of random vari-
able, while \( \bar{x} \) and \( s^2 \) are their values obtained for the particular values of the rvs;
in this case, for the 28 patients who took part in the study. A different group of 28
patients in the state who suffered a heart attack, would give different values of \( \bar{X} \)
and \( S^2 \).

\[ \square \]

A random variable which is a function of a random sample, \( T(X_1, \ldots, X_n) \), is
called an estimator of a population parameter \( \vartheta \), while its value is called an esti-
mate of the population parameter \( \vartheta \).
CHAPTER 2. ELEMENTS OF STATISTICAL INference

Notation

Special symbols such as $\bar{X}$ or $S^2$, are used to denote estimators of some common parameters, in these cases the population mean and variance. Then their “small letter” counterparts denote the respective estimates.

Hat over a symbol of the parameter of interest, for example $\hat{\vartheta}$, indicates an estimator of $\vartheta$. Here we have to be careful because $\hat{\vartheta}$ is often used to denote an estimate of $\vartheta$ as well. It is better to write $\hat{\vartheta}_{\text{obs}}$ to indicate that this is a value of the estimator $\hat{\vartheta}$ obtained for the observed values of a random sample.

A random sample will often be denoted by a bold capital letter, while its realization by a bold small case letter. For example, $X$ denotes a random sample $(X_1, \ldots, X_n)$ and $x$ denotes its realization $(x_1, \ldots, x_n)$.

We often know the family of distributions related to the variable of interest, which we denote by $P = \{ P_\vartheta : \vartheta \in \Theta \subset \mathbb{R}^s \}$ (see Definition 1.10). The family depends on a parameter vector $\vartheta$ which belongs to a parameter space $\Theta$. For example, for the family of normal distributions $P = \{ f(y; \mu, \sigma^2) : -\infty < \mu < \infty, \sigma^2 > 0 \}$, the parameter vector is two-dimensional, $\vartheta = (\mu, \sigma^2)$, and the parameter space is $\Theta = \{ \mathbb{R} \times \mathbb{R}_+ \setminus \{0\} \}$.

Properties of the estimators of the unknown parameters and the methods of finding them will be our topics for the next several lectures.

2.1.1 Sufficient statistics

Statistic $T(Y)$ represents a random sample $Y = (Y_1, \ldots, Y_n)$ and serves as an estimator of a population parameter $\vartheta$ or of a function of $\vartheta$. For the inference based on the random sample, it is often enough to consider $T(Y)$, or a function of it, rather than the whole random sample. Hence, the statistic should represent the data well and should retain all the information about the parameter which is contained in the original observations.

Such a statistic has the property of sufficiency. Any additional information in the sample, besides the value of the sufficient statistic, does not contain any more information about the parameter of interest. The principle of sufficiency says that if $T(Y)$ is a sufficient statistic for $\vartheta$ then any inference about $\vartheta$ should depend on the sample through the value of the statistic only. That is, if $y_1$ and $y_2$ are two realizations of the random sample $Y$ such that $T(y_1) = T(y_2)$, then the inference
about $\theta$ should be the same whichever of the realizations, $y_1$ or $y_2$, is used.

**Definition 2.1.** Let $X = (X_1, \ldots, X_n)$ denote a random sample from a probability distribution with unknown parameters $\theta = (\theta_1, \ldots, \theta_p)$. Then the statistics $T = (T_1, \ldots, T_q)$ are said to be jointly sufficient for $\theta$ if the conditional distribution of $X$ given $T$ does not depend on $\theta$.

**Example 2.2.** Let $X = (X_1, X_2, \ldots, X_n)$ be a random sample from a population of Bernoulli distribution with probability of success equal to $p$, i.e.,

$$P(X_i = x_i) = p^{x_i}(1 - p)^{1-x_i}, \quad i = 1, 2, \ldots, n,$$

where $x_i \in \{0, 1\}$, $p \in (0, 1)$. Then the joint mass function of $X$ is

$$P(X = x) = \prod_{i=1}^{n} p^{x_i}(1 - p)^{1-x_i} = p^{\sum_{i=1}^{n} x_i}(1 - p)^{n-\sum_{i=1}^{n} x_i}.$$

Let us define a statistic $T = T(X)$ as

$$T(X) = \sum_{i=1}^{n} X_i.$$

This is a random variable which denotes the number of success in $n$ trials and it has a Binomial distribution, that is,

$$P(T = t) = \binom{n}{t} p^t(1 - p)^{n-t}, \quad t = 0, 1, \ldots, n.$$

The joint probability of $X$ and $T$ is zero unless $\sum_{i=1}^{n} x_i = t$. That is, for $\sum_{i=1}^{n} x_i = t$ we can write

$$P(X = x, T = t) = P(X = x) = p^{\sum_{i=1}^{n} x_i}(1 - p)^{n-\sum_{i=1}^{n} x_i} = p^t(1 - p)^{n-t}.$$

Hence, the conditional distribution of $X|T$ is

$$P(X = x|T = t) = \frac{P(X = x, T = t)}{P(T = t)} = \frac{p^t(1 - p)^{n-t}}{\binom{n}{t} p^t(1 - p)^{n-t}} = \binom{n}{t}^{-1}.$$

The conditional distribution does not depend on the parameter $p$, so the function $T(X) = \sum_{i=1}^{n} X_i$ is a sufficient statistic for $p$. 
It can be interpreted as follows: if we know the number of successes in \( n \) trials, \( t = \sum_{i=1}^{n} x_i \), then the information on the order of the successes occurring in the Bernoulli experiments does not give any more information about the probability of success, which could be estimated as \( \sum_{i=1}^{n} x_i / n \).

We can identify sufficient statistics more easily using the following result.

**Theorem 2.1. Neyman’s factorization.**

The statistics \( T(Y) \) is sufficient for \( \vartheta \) if and only if the joint pdf (pmf) of \( Y \) can be factorized as

\[
f_Y(y; \vartheta) = g(T(y); \vartheta)h(y),
\]

where \( h(y) \) does not depend on the parameters and function \( g \), which depends on the parameters, is a function of the values of \( Y \) only through the statistic \( T \).

**Proof.** Discrete case for a single parameter.

First, assume that \( T(Y) \) is a sufficient statistic for \( \vartheta \). If \( T(Y) = t \), then

\[
P(Y = y) = P(Y = y, T(Y) = t) = \frac{P(T(Y) = t)P(Y = y|T(Y) = t)}{=g_0(T(Y)) \quad =h(y) \text{ if } T \text{ is a suff. stat.}}.
\]

So, we get the factorization.

Conversely. Let the factorization hold, that is

\[
P(Y = y) = g[T(y); \vartheta]h(y).
\]

Then,

\[
P(Y = y|T(Y) = t) = \frac{P(Y = y, T(Y) = t)}{P(T(Y) = t)}
\]

\[
= \left\{ \begin{array}{ll}
0, & \text{if } T(Y) \neq t; \\
\frac{P(Y = y)}{P(T(Y) = t)}, & \text{if } T(Y) = t \\
\end{array} \right.
\]

\[
= \left\{ \begin{array}{ll}
0, & \text{if } T(Y) \neq t; \\
\frac{P(Y = y)}{\sum_{y: T(y) = t} P(Y = y)}, & \text{if } T(Y) = t \\
\end{array} \right.
\]

\[
= \left\{ \begin{array}{ll}
0, & \text{if } T(Y) \neq t; \\
\frac{g[T(y); \vartheta]h(y)}{g[T(y); \vartheta]\sum_{y: T(y) = t} h(y)}, & \text{if } T(Y) = t \\
\end{array} \right.
\]

\[
= \left\{ \begin{array}{ll}
0, & \text{if } T(Y) \neq t; \\
\frac{h(y)}{\sum_{y: T(y) = t} h(y)}, & \text{if } T(Y) = t \\
\end{array} \right.
\]

This does not depend on the parameter \( \vartheta \), hence \( T \) is a sufficient statistic for \( \vartheta \). \( \square \)
2.1. **PRINCIPLES OF DATA REDUCTION**

As we will later see, sufficient statistics provide optimal estimators and pivots for confidence intervals and test statistics.

**Example 2.3.** Suppose that \( Y = (Y_1, \ldots, Y_n) \) is a random sample from a \( \text{Poisson}(\lambda) \) distribution. Then we may write

\[
P(Y = y; \lambda) = \prod_{i=1}^{n} \frac{\lambda^{y_i} e^{-\lambda}}{y_i!} = \frac{\lambda^{\sum_{i=1}^{n} y_i} e^{-n\lambda}}{\prod_{i=1}^{n} y_i!} = \lambda^{\sum_{i=1}^{n} y_i} e^{-n\lambda} \times \frac{1}{\prod_{i=1}^{n} y_i!}.
\]

It follows that \( \sum_{i=1}^{n} Y_i \) is a sufficient statistic for \( \lambda \).

**Exercise 2.1.** Suppose that \( Y = (Y_1, \ldots, Y_n) \) is a random sample from an \( \text{Exp}(\lambda) \) distribution. Obtain a sufficient statistic for \( \lambda \).

**Example 2.4.** Suppose that \( Y = (Y_1, \ldots, Y_n) \) is a random sample from a \( \mathcal{N}(\mu, \sigma^2) \) distribution. Then we may write

\[
f_Y(y; \mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left\{ -\frac{(y_i - \mu)^2}{2\sigma^2} \right\} = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu)^2 \right\} = (2\pi)^{-\frac{n}{2}} (\sigma^2)^{-\frac{n}{2}} \exp \left\{ -\frac{1}{2\sigma^2} \left( \sum_{i=1}^{n} y_i^2 - 2\mu \sum_{i=1}^{n} y_i + n\mu^2 \right) \right\} \times (2\pi)^{-\frac{n}{2}}.
\]

Hence, \( \sum_{i=1}^{n} Y_i \) and \( \sum_{i=1}^{n} Y_i^2 \) are jointly sufficient statistics for \( \mu \) and \( \sigma^2 \).

**Example 2.5.** Suppose that \( Y = (Y_1, \ldots, Y_n) \) is a random sample from a \( \text{U}[0, \vartheta] \) distribution. Then we may write

\[
f_Y(y; \vartheta) = \prod_{i=1}^{n} \frac{1}{\vartheta^n} = \frac{1}{\vartheta^n}, \quad \max_i y_i \leq \vartheta.
\]
Now, this may be rewritten as
\[ f_Y(y; \vartheta) = \frac{1}{\vartheta^n} I_{\{\max_i y_i \leq \vartheta\}} \times 1. \]

It follows that \( \max_i Y_i \) is a sufficient statistic for \( \vartheta \).

There can be many sufficient statistics for a parameter of interest. Hence, the question is which one to use. As the purpose of a statistic is to reduce data without loss of information about the parameter we would like to find such a function of the random sample that achieves the most data reduction while still retaining all the information about the parameter.

**Definition 2.2.** A sufficient statistic \( T(Y) \) is called a **minimal sufficient statistic** if, for any other sufficient statistic \( T'(Y) \), \( T(Y) \) is a function of \( T'(Y) \).

**Lemma 2.1.** Let \( f(y; \vartheta) \) be the pdf of a random sample \( Y \). If there exists a function \( T(Y) \) such that for two sample points \( x \) and \( y \), the ratio \( f(x; \vartheta)/f(y; \vartheta) \) is constant as a function of \( \vartheta \) if and only if \( T(x) = T(y) \) then \( T(Y) \) is a minimal sufficient statistic for \( \vartheta \).

Analogous Lemma holds for discrete r.v.s.

**Example 2.6.** Suppose that \( Y = (Y_1, \ldots, Y_n) \) is a random sample from a \( \text{Poisson}(\lambda) \) distribution. From the factorization theorem we know that \( T(Y) = \sum_{i=1}^n Y_i \) is a sufficient statistic for \( \lambda \). The ratio of the joint probability mass functions is
\[
\frac{P(Y = x; \lambda)}{P(Y = y; \lambda)} = \frac{\prod_{i=1}^n x_i! \lambda^{\sum_{i=1}^n x_i} e^{-n\lambda}}{\prod_{i=1}^n y_i! \lambda^{\sum_{i=1}^n y_i} e^{-n\lambda}} = \prod_{i=1}^n \frac{y_i!^{x_i}}{x_i!^{y_i}} \lambda^{\sum_{i=1}^n x_i - \sum_{i=1}^n y_i}
\]
This is constant in terms of \( \lambda \) iff \( \sum_{i=1}^n y_i = \sum_{i=1}^n x_i \) that is \( T(Y) = \sum_{i=1}^n Y_i \) is a minimal sufficient statistic.

Note that \( Y \) is also a minimal sufficient statistic for \( \lambda \). In general, a minimal sufficient statistic is not unique. Any one-to-one function of a minimal sufficient statistic is also a minimal sufficient statistic.
Exercise 2.2. Suppose that $Y = (Y_1, \ldots, Y_n)$ is a random sample from an $\text{Exp}(\lambda)$ distribution. Show that $T(Y) = \sum_{i=1}^n Y_i$ is a minimal sufficient statistic for $\lambda$.

Example 2.7. Suppose that $Y = (Y_1, \ldots, Y_n)$ is a random sample from a $\mathcal{N}(\mu, \sigma^2)$ distribution. Then $T_1 = \sum_{i=1}^n Y_i$ and $T_2 = \sum_{i=1}^n Y_i^2$ are jointly sufficient statistics for $\mu$ and $\sigma^2$. The ratio of the joint densities is

$$
\frac{f_Y(x; \mu, \sigma^2)}{f_Y(y; \mu, \sigma^2)} = \exp \left\{ -\frac{1}{2\sigma^2} \left[ \sum_{i=1}^n x_i^2 - \sum_{i=1}^n y_i^2 - 2\mu \left( \sum_{i=1}^n x_i - \sum_{i=1}^n y_i \right) \right] \right\}.
$$

The ratio does not depend on the parameters $\mu, \sigma^2$ iff $\sum_{i=1}^n x_i^2 = \sum_{i=1}^n y_i^2$ and $\sum_{i=1}^n x_i = \sum_{i=1}^n y_i$. Hence, $(\sum_{i=1}^n Y_i, \sum_{i=1}^n Y_i^2)$ are jointly minimal sufficient statistics for these parameters.

Note that $\overline{Y}$ and $S^2 = \frac{1}{n-1} \sum_{i=1}^n (Y_i - \overline{Y})^2 = \frac{1}{n-1} \left[ \sum_{i=1}^n Y_i^2 - n\overline{Y}^2 \right]$ are one-to-one functions of $(\sum_{i=1}^n Y_i, \sum_{i=1}^n Y_i^2)$. Hence, $(\overline{Y}, S^2)$ are also jointly minimum sufficient statistics for $\mu$ and $\sigma^2$. □
CHAPTER 2. ELEMENTS OF STATISTICAL INFERENCE

2.2 Point Estimation

Suppose that we have data $y_1, y_2, \ldots, y_n$ which are a realization of $Y_1, Y_2, \ldots, Y_n$. We are interested in the population from which the data are sampled. Suppose that the probability model is that $Y_i \sim N(\mu, \sigma^2)$ independently for $i = 1, 2, \ldots, n$. Then the probability model for the population is completely specified, except for the unknown parameters $\mu$ and $\sigma^2$.

In general, $Y_1, Y_2, \ldots, Y_n$ have a joint distribution which is specified, except for the unknown parameters $\vartheta = (\vartheta_1, \vartheta_2, \ldots, \vartheta_p)^T$. We are interested in estimating some function of the parameters $\phi = g(\vartheta_1, \ldots, \vartheta_p)$. Note that $\phi$ could be just a single parameter, such as $\mu$ or $\sigma^2$. We will use a sample statistic $T(Y_1, \ldots, Y_n)$ as a point estimator of $\phi$. For example, $T(Y_1, \ldots, Y_n) = \overline{Y}$ is an estimator of $\mu$. The observed value of $T$, $T(y_1, \ldots, y_n)$, is the point estimate of $\phi$.

2.2.1 Properties of Point Estimators

Estimator $T(Y)$ is a random variable (a function of rvs $Y_1, \ldots, Y_n$) and the estimate is a single value taken from the distribution of $T(Y)$. Since we want our estimate, $t$, to be close to $\phi$, the random variable $T$ should be centred close to $\phi$ and have small variance. Also, we would want our estimator to be such that, as $n \to \infty$, $T \to \phi$ with probability one.

Below are definitions of some measures of goodness of the estimators.

**Definition 2.3.** Let $Y = (Y_1, Y_2, \ldots, Y_n)^T$ be a random sample from a distribution $P_{\vartheta}$ and let $\phi = g(\vartheta)$ be a function of the parameters and $T(Y)$ its point estimator. Then,

1. The **error** in estimating $\phi$ by estimator $T(Y)$ is $T - \phi$;

2. The **bias** of estimator $T(Y)$ for $\phi$ is

   $$\text{bias}(T) = \mathbb{E}(T) - \phi,$$

   that is, the expected error;

3. Estimator $T(Y)$ is **unbiased** for $\phi$ if

   $$\mathbb{E}(T) = \phi,$$

   that is, it has zero bias.
A weaker property of unbiasedness is defined below.

**Definition 2.4.** The estimator $T(Y)$ is asymptotically unbiased for $\phi$ if $E(T) \to \phi$ as $n \to \infty$, that is, the bias tends to zero as $n \to \infty$. □

Another common measure of goodness of an estimator is given in the following definition.

**Definition 2.5.** The mean square error of $T(Y)$ as an estimator of $\phi$ is

$$\text{MSE}(T) = E \{ (T - \phi)^2 \}.$$ □

Note: The mean square error can be written as a sum of variance and squared bias of the estimator.

$$\text{MSE}(T) = E \left( \{ T - E(T) \}^2 + \{ E(T) - \phi \}^2 \right)$$

$$= E \left( \{ T - E(T) \}^2 + 2 \{ T - E(T) \} \{ E(T) - \phi \} + \{ E(T) - \phi \}^2 \right)$$

$$= E \left( \{ T - E(T) \}^2 \right) + 2 \{ E(T) - E(T) \} \{ E(T) - \phi \} + \{ E(T) - \phi \}^2$$

$$= \text{var}(T) + \{ \text{bias}(T) \}^2.$$  

It seems natural that if we increase the sample size $n$ we should get a more precise estimate of $\phi$. The estimators which have this property are called consistent. In the following definition we use the notion of convergence in probability:

A sequence of rvs $X_1, X_2, \ldots$ converges in probability to a random variable $X$ if for every $\varepsilon > 0$

$$\lim_{n \to \infty} P(\{|X_n - X| < \varepsilon\}) = 1.$$  

**Definition 2.6.** A sequence $T_n = T_n(Y_1, Y_2, \ldots, Y_n)$, $n = 1, 2, \ldots$, is called a consistent sequence of estimators for $\phi$ if $T_n$ converges in probability to $\phi$, that is for all $\varepsilon > 0$ and for all $\vartheta \in \Theta$

$$\lim_{n \to \infty} P(\{|T_n - \phi| < \varepsilon\}) = 1.$$ □
Note: Consistency means that the probability of our estimator being within some small \( \varepsilon \) of \( \phi \) can be made as close to one as we like by making the sample size \( n \) sufficiently large.

The following Lemma gives a useful tool for establishing consistency of estimators.

**Lemma 2.2.** A sufficient condition for consistency is that \( \text{MSE}(T) \to 0 \) as \( n \to \infty \), or, equivalently, \( \text{var}(T) \to 0 \) and \( \text{bias}(T) \to 0 \) as \( n \to \infty \).

**Example 2.8.** Suppose that \( Y_1, \ldots, Y_n \) are independent Poisson(\( \lambda \)) random variables and consider using \( \bar{Y} \) to estimate \( \lambda \). Then we have

\[
E(\bar{Y}) = \frac{1}{n} E(Y_1 + \ldots + Y_n)
= \frac{1}{n} \{ E(Y_1) + \ldots + E(Y_n) \}
= \frac{1}{n} n \lambda = \lambda,
\]

and so \( \bar{Y} \) is an unbiased estimator of \( \lambda \). Next, by independence, we have

\[
\text{var}(\bar{Y}) = \frac{1}{n^2} \{ \text{var}(Y_1) + \ldots + \text{var}(Y_n) \}
= \frac{1}{n^2} n \lambda = \frac{\lambda}{n},
\]

so that \( \text{MSE}(\bar{Y}) = \lambda/n \). Thus, \( \text{MSE}(\bar{Y}) \to 0 \) as \( n \to \infty \), and so \( \bar{Y} \) is a consistent estimator of \( \lambda \).

A more general result about the mean is given in the following theorem, which states that the sample mean approaches the population mean as \( n \to \infty \), and it holds for any distribution.

**Theorem 2.2. Weak Law of Large Numbers**

Let \( X_1, X_2, \ldots \) be IID random variables with \( E(X_i) = \mu \) and \( \text{var}(X_i) = \sigma^2 < \infty \). Define \( \overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i \). Then, for every \( \varepsilon > 0 \),

\[
\lim_{n \to \infty} P(|\overline{X}_n - \mu| < \varepsilon) = 1,
\]

that is, \( \overline{X}_n \) converges in probability to \( \mu \).
2.2. POINT ESTIMATION

Proof. The proof is a straightforward application of Markov’s Inequality, which says the following: Let $X$ be a random variable and let $g(\cdot)$ be a non-negative function. Then, for any $k > 0$,

$$P(g(X) \geq k) \leq \frac{E[g(X)]}{k}. \quad (2.1)$$

[Markov’s Inequality]

Proof of Markov’s inequality (continuous case):

For any rv $X$ and a nonnegative function $g(\cdot)$ we can write

$$E[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x) dx \geq \int_{x: g(x) \geq k} g(x) f_X(x) dx \geq \int_{x: g(x) \geq k} k f_X(x) dx = k P(g(X) \geq k).$$

Hence (2.1) holds. Then,

$$P(|X_n - \mu| < \varepsilon) = 1 - P(|X_n - \mu| \geq \varepsilon)$$

$$= 1 - P((X_n - \mu)^2 \geq \varepsilon^2)$$

$$\geq 1 - \frac{E(X_n - \mu)^2}{\varepsilon^2} = 1 - \frac{\sigma^2}{n \varepsilon^2} \to 1 \text{ as } n \to \infty.$$

Markov’s inequality is a generalization of Chebyshev’s inequality. If we take $g(X) = (X - \mu)^2$ and $k = \sigma^2 t^2$ for some positive $t$, then the Markov’s inequality states that

$$P((X - \mu)^2 \geq t^2) \leq \frac{E[(X - \mu)^2]}{\sigma^2 t^2} = \frac{\sigma^2}{\sigma^2 t^2} = \frac{1}{t^2}.$$

This is equivalent to

$$P(|X - \mu| \geq \sigma t) \leq \frac{1}{t^2}$$

which is known as Chebyshev’s inequality.

Example 2.9. The rule of three-sigma.

This rule basically says that a chance that an rv will have values outside the interval $(\mu - 3\sigma, \mu + 3\sigma)$ is close to zero. Chebyshev’s inequality, however, says that for any random variable the probability that it will deviate from its mean by more than three standard deviations is at most 1/9, i.e.,

$$P(|X - \mu| \geq 3\sigma) \leq \frac{1}{9}.$$
CHAPTER 2. ELEMENTS OF STATISTICAL INFEERENCE

This suggests that the rule should be considered with caution. The rule of three-sigma works well for normal rvs or for the variables whose distributions are close to normal. Let $X \sim \mathcal{N}(\mu, \sigma^2)$. Then

$$P(|X - \mu| \geq 3\sigma) = 1 - P(|X - \mu| \leq 3\sigma) = 1 - P(\mu - 3\sigma \leq X \leq \mu + 3\sigma) = 1 - P(-3\sigma \leq X - \mu \leq 3\sigma) = 1 - [\Phi(3) - \Phi(-3)] = 1 - [\Phi(3) - (1 - \Phi(3))] = 2(1 - \Phi(3)) = 2(1 - 0.99865) \approx 0.0027.$$

Another important result, which is widely used in statistic is the Central Limit Theorem (CLT) which uses the notion of convergence in distribution. We have already used this notion, but here we have its formal definition.

**Definition 2.7.** A sequence of random variables $X_1, X_2, \ldots$ converges in distribution to a random variable $X$ if

$$\lim_{n \to \infty} F_{X_n}(x) = F_X(x),$$

at all points $x$ where the cdf $F_X(x)$ is continuous.

The following theorem represents the large sample behaviour of the sample mean (compare the Weak Law of Large Numbers).

**Theorem 2.3. The Central Limit Theorem**

Let $X_1, X_2, \ldots$ be a sequence of independent random variables such that $\mathbb{E}(X_i) = \mu$ and $\text{var}(X_i) = \sigma^2 < \infty$. Define

$$U_n = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}.$$

Then the sequence of random variables $U_1, U_2, \ldots$ converges in distribution to a random variable $U \sim \mathcal{N}(0, 1)$, that is

$$\lim_{n \to \infty} F_{U_n}(u) = F_U(u),$$

where $F_U(u)$ is a cdf of a standard normal rv.

**Proof.** The proof is based on the mgf and so it is valid only for the cases where the mgf exits. Here we assume that $M_{X_i}(t)$ exists in some neighborhood of zero, say for $|t| < h$ for some positive $h$. 


2.2. POINT ESTIMATION

We have

\[ U_n = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} = \frac{1}{n} \sum_{i=1}^{n} \frac{X_i - \mu}{\sigma / \sqrt{n}} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{X_i - \mu}{\sigma / \sqrt{n}} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Y_i, \]

where \( Y_i = \frac{X_i - \mu}{\sigma} \). The variables \( X_i \) are IID, hence by Theorem 1.15 the variables \( Y_i \) are also IID. Hence,

\[ M_{U_n}(t) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Y_i(t) = E \left( e^{\frac{1}{\sqrt{n}} \sum_{i=1}^{n} Y_i t} \right) = E \left( \prod_{i=1}^{n} e^{\frac{1}{\sqrt{n}} Y_i t} \right) = \prod_{i=1}^{n} M_{Y_i} \left( \frac{1}{\sqrt{n}} t \right) = \left[ M_{Y_i} \left( \frac{1}{\sqrt{n}} t \right) \right]^n. \]

The common mgf of \( Y_i \) can be expanded into a Taylor series around zero, that is

\[ M_{Y_i} \left( \frac{1}{\sqrt{n}} t \right) = \sum_{k=0}^{\infty} M_{Y_i}^{(k)}(0) \frac{(t/\sqrt{n})^k}{k!}, \]

where \( M_{Y_i}^{(k)}(0) = \frac{d^k}{dt^k} M_{Y_i}(t) |_{t=0} \). Since the mgfs exist for \(|t| < h\), for some positive \( h \), the Taylor series expansion is valid if \( t < \sqrt{n}h \).

Now, \( M_{Y_i}^{(1)}(0) = E(Y_i) = 0 \), \( M_{Y_i}^{(2)}(0) = E(Y_i^2) = \text{var}(Y_i) = 1 \) and \( M_{Y_i}^{(0)}(0) = M_{Y_i}(0) = E(e^{tY_i}) |_{t=0} = E(1) = 1 \). Hence, the series can be written as

\[ M_{Y_i} \left( \frac{1}{\sqrt{n}} t \right) = 1 + 0 + \frac{(t/\sqrt{n})^2}{2!} + \sum_{k=3}^{\infty} M_{Y_i}^{(k)}(0) \frac{(t/\sqrt{n})^k}{k!} \]

\[ = 1 + \frac{(t/\sqrt{n})^2}{2!} + R_{Y_i}(t/\sqrt{n}), \]

where the remainder term \( R_{Y_i}(t/\sqrt{n}) \) tends to zero faster than the highest order explicit term, that is

\[ \lim_{n \to \infty} \frac{R_{Y_i}(t/\sqrt{n})}{(t/\sqrt{n})^2} = 0. \]

This means that also \( \lim_{n \to \infty} \frac{R_{Y_i}(t/\sqrt{n})}{1/n} = \lim_{n \to \infty} nR_{Y_i}(t/\sqrt{n}) = 0 \) (as \( t \) is fixed here and can be taken equal to 1).
Thus,
\[
\lim_{n \to \infty} M_{U_n}(t) = \left[ M_{Y_1} \left( \frac{1}{\sqrt{n}} t \right) \right]^n
\]
\[
= \lim_{n \to \infty} \left[ 1 + \frac{(t/\sqrt{n})^2}{2!} + R_{Y_1}(t/\sqrt{n}) \right]^n
\]
\[
= \lim_{n \to \infty} \left\{ 1 + \frac{1}{n} \left( \frac{t^2}{2} + nR_{Y_1}(t/\sqrt{n}) \right) \right\}^{n \to t^2/2}
\]
\[
= e^t^2/2 = M_U(t).
\]
which is the mgf of a standard normal rv. □

The CLT means that the sample mean is asymptotically normally distributed whatever the distribution of the original rvs is. However, we have no way of knowing how good the approximation is. This depends on the original distribution of $X_i$.

**Exercise 2.3.** Let $Y = (Y_1, \ldots, Y_n)^T$ be a random sample from the distribution with the pdf given by
\[
f(y; \theta) = \begin{cases} 
\frac{2}{\theta^2}(\theta - y), & y \in [0, \theta], \\
0, & \text{elsewhere.}
\end{cases}
\]
For an estimator $T(Y) = 3Y$ calculate its bias and variance and check if it is a consistent estimator for $\theta$.

**Exercise 2.4.** Let $X_i \sim \text{Bernoulli}(p)$. Denote by $\bar{X}$ the sample mean.

(a) Show that $\hat{p} = \bar{X}$ is a consistent estimator of $p$.

(b) Show that $\hat{pq} = \bar{X}(1 - \bar{X})$ is an asymptotically unbiased estimator of $pq$, where $q = 1 - p$.  