Chapter 1

Continuous Distributions

1.1 Normal Distribution

Definition 1.1 (Normal distribution)**.** The random variable X subjects to the *normal distribution*, with two parameters μ and σ, if its density function is given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left[\frac{-1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right], \quad -\infty < x < \infty, \qquad (-\infty < \mu < \infty,\ \sigma > 0),\ \pi = 3.14. \tag{1.1}$$

If the random variable X is normally distributed with mean μ and variance σ^2 (later on we prove that the mean is μ and the variance is σ^2), we will write $X \sim N(\mu, \sigma^2)$. We will also use the notation $\Phi_{\mu,\sigma^2}(x)$ for the cumulative distribution function.

If in (1.1) $z = \dfrac{x-\mu}{\sigma}$, then

$$f_Z(z) = \frac{1}{\sqrt{2\pi}}\exp\left[\frac{-z^2}{2}\right], \quad -\infty < z < \infty,$$

is the density function of the random variable Z with two parameter values $\mu = 0$ and $\sigma^2 = 1$, which is called **standard** normal random variable, i.e. $Z \sim N(0, 1)$.

Properties of the normal curve

1. The curve is symmetric about a vertical axes through the mean μ and it has the bell-shape.

2. The mode, which is the point on the horizontal x-axes where the curve is a maximum, occurs at $x = \mu$.

3. The normal curve approaches the horizontal x-axes as $x \to \pm\infty$

4. The total area under the curve and above the x-axes is equal to 1.

Fortunately, to avoid the use of integral calculus, we are able to transform all of the observations of any normal random variable X to a new set of observations of a standard normal random variable Z with mean zero and variance one, using the transformation

$$Z = \frac{(X-\mu)}{\sigma},$$

where $E[Z] = \dfrac{(\mu-\mu)}{\sigma} = 0$ and $V[Z] = \dfrac{V(X-\mu)}{\sigma^2} = \dfrac{\sigma^2}{\sigma^2} = 1$.

Example 1.1. *Given the normally distributed random variable X with mean 18 and variance 6.25, find*

(i) $P(X < 15)$,

(ii) the value of k such that $P(X < k) = 0.2578$,

(iii) $P(17 < X < 21)$,

(iv) the value of k such that $P(X > k) = 0.1539$.

(i)

$$\begin{aligned} P(X < 15) &= P\left(\frac{X-18}{2.5} < \frac{15-18}{2.5}\right) \\ &= P(Z < -1.2) \\ &= \Phi(-1.2) = 0.1151, \end{aligned}$$

(ii)

$$\begin{aligned} P(X < k) = 0.2578 &\Rightarrow P\left(Z < \frac{k-18}{2.5}\right) = 0.2578 \\ &\Rightarrow \frac{k-18}{2.5} = -.65 \Rightarrow k = 16.375, \end{aligned}$$

(iii)

$$\begin{aligned} P(17 < X < 21) &= P\left(\frac{17-18}{2.5} < Z < \frac{21-18}{2.5}\right) \\ &= P(-.4 < Z < 1.2) \\ &= \Phi(1.2) - \Phi(-.4) \\ &= .8849 - .3446 = 0.5403, \end{aligned}$$

(iv)

$$\begin{aligned} P(X > k) = .1539 &\Rightarrow P\left(Z > \frac{K-18}{2.5}\right) = .1539 \\ &\Leftrightarrow P\left(Z < \frac{18-k}{2.5}\right) = .1539 \\ &\Rightarrow \frac{18-k}{2.5} = -1.02 \\ &\Rightarrow k = 20.55, \end{aligned}$$

Example 1.2. *If a set of grades on a statistic examination are approximately normally distributed with a mean of* 17 *and a standard deviation of* 7.9*, find (a) the lowest passing grade if the lowest* 10% *of the students are given* Fs*, (b) the highest* B *if the top* 5% *of the students are given* As*.*

Let $X \sim N(74, (7.9)^2)$ and shows the grades

(a) $P(X < k) = 0.1 \Rightarrow P\left(Z < \dfrac{k-74}{7.9}\right) = .1$

$\Rightarrow \dfrac{k-74}{7.9} = -1.28 \Rightarrow k \cong 64.$

(b) $P(X > B) = .05 \Rightarrow P\left(Z > \dfrac{B-74}{7.9}\right) = .05$

$\Leftrightarrow P\left(Z < \dfrac{74-B}{7.9}\right) = .05$

$\Rightarrow \dfrac{74-B}{7.9} = -1.65 \Rightarrow B \cong 87.$

Example 1.3. *In a mathematics examination the average grade was* 82 *and the standard deviation was* 5*. All students with grades from* 88 *to* 94 *received a grade of* B*. If the grades are approximately normally distributed and* 8 *students received a* B *grade, how many students took the examination?*

Let $X \sim N(82, 25)$ and shows the grade,
$P(88 < X < 94) = P(1.2 < Z < 2.4) = \Phi(2.4) - \Phi(1.2) = .9918 - .8849 = .1096 \Rightarrow n \times .1096 = 8 \Rightarrow n \cong 75.$

1.1.1 Normal approximation to the binomial

We shall now state (without proof) a theorem that allows us to use areas under the normal curve to approximate binomial probabilities when n is sufficiently large.

Theorem 1.1. *If X is a binomial random variable with mean $\mu = np$ and variance $\sigma^2 = npq$, then the limiting form of the distribution of*

$$Z = \frac{X - np}{\sqrt{npq}},$$

as $n \to \infty$, is the standardized normal distribution $N(0,1)$.

The probabilities of the binomial are approximated according to the following:
If a, b and c are positive integers, $0 \leq a, b, c \leq n$, then

1.
$$\begin{aligned} P(X = c) &= P(c - 0.5 \leq X \leq c + 0.5) \\ &= P\left(\frac{c - 0.5 - np}{\sqrt{npq}} \leq Z \leq \frac{c + 0.5 - np}{\sqrt{npq}}\right) \end{aligned}$$

2. $P(a \leq X \leq b) = P\left(\frac{a - 0.5 - np}{\sqrt{npq}} \leq Z \leq \frac{b + 0.5 - np}{\sqrt{npq}}\right)$,

3. $P(a < X < b)$, $P(a < X \leq b)$ and $P(a \leq X < b)$ should be transformed to closed interval probability and then apply (2).

Example 1.4. *A drug manufacturer claims that a certain drug cures a blood disease on the average 85% of the time. To check the claim, government testers used the drug on a sample of 100 individuals and decide to accept the claim if 75% or more are cured, what is the probability that the claim will be accepted when the cure probability is in fact 85%*

Let $X \sim b(100, 0.85)$ and shows the number of cured people, then

$$P(X \geq 75) = \sum_{x=75}^{100} C_x^{100} (0.85)^x (0.15)^{100-x},$$

but we notice that the number of individuals is large, so it is prefer to use the normal approximation to the binomial. Therefore

$$\begin{aligned} P(X \geq 75) &= P\left(Z \geq \frac{75 - 0.5 - (100)(0.85)}{\sqrt{(100)(0.85)(0.15)}}\right) \\ &= P(Z \geq -2.94) = 1 - \Phi(-2.94) = 1 - 0.0016 \\ &= 0.9984. \end{aligned}$$

Example 1.5. *A certain pharmaceutical company knows that, on the average, 5% of a certain type of pill has ingredient that is below the minimum strength and thus unacceptable. What is the probability that at least 2 in a sample of 200 pills will be unaccepted. Find also the mean and standard deviation of the accepted pills.*

Let $X \sim b(200, 0.05)$ and shows the number of unaccepted pills, then

$$\begin{aligned} P(X \geq 2) &= 1 - [P(X = 1) + P(X = 0)] = \\ &= 1 - [C_1^{200}(0.05)^1(0.95)^{199} + C_0^{200}(0.95)^{200}], \end{aligned}$$

but we notice that $n = 200$ is large, so it is prefer to use Poisson distribution with parameter $\lambda = (200)(0.05) = 10$ or use the normal approximation to the binomial. Therefore

$$\begin{aligned} P(X \geq 2) &= P\left(Z \geq \frac{2 - 0.5 - (200)(0.05)}{\sqrt{(200)(0.05)(0.95)}}\right) \\ &= P(Z \geq -2.75) = 1 - \Phi(-2.75) = 1 - 0.003 = 0.997. \end{aligned}$$

The mean and standard deviation of the accepted pills are equal, respectively, $n(1-p) = 200(0.95) = 190$ and $\sqrt{n(1-p)q} = \sqrt{200(0.95)(0.05)} = 3.08$.

Chapter 2

SAMPLING THEORY

Studying the relationships existing between a population and samples drawn from the population is called "sampling theory".

Sampling theory is useful in estimating the unknown population parameters and also in determining whether observed differences between two samples are really due to chance variation or whether they are actually.

The purpose of this chapter is to introduce the concept of sampling and to present some distribution results that are related by sampling.

2.1 Population and Samples

Definition 2.1 (Population)**.** The totality of all observations which are under discussion will be called the *population.*

Definition 2.2 (Simple random sample)**.** If a sample of size n, say $X_1, X_2, \ldots, X_n$, drawn from a population of size N in such a way that every possible sample of size n has the same probability of being selected, then it is called a *simple random sample.*

Definition 2.3 (Statistic)**.** A *statistic* is a random variable depends only on the observed sample.

Example 2.1. *If $X_1, X_2, \ldots, X_n$ is a random sample of size n, then each of the following represents a statistic.*

1. $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ *[sample mean].*

2. $\mu_r' = \frac{1}{n}\sum_{i=1}^{n} X_i^r$ *[r^{th} sample moment about 0].*

3. $\mu_r = \dfrac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^r$ *[r^{th} sample moment about $\bar{X}$].*

Definition 2.4 (Sample variance)**.** If $X_1, X_2, \ldots, X_n$ represent a random sample of size n, with mean $\bar{X}$, then

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}[X_i - \bar{X}]^2; \quad n > 1$$

is defined to be the *sample variance.*

Definition 2.5 (Sampling distribution)**.** The probability distribution of a statistic is called a *sampling distribution.*

To construct a sampling distribution, we proceed as follows:

1. From a finite population of size N, randomly draw all possible samples of size n.

2. Compute the statistic of interest, such as the mean, for each sample.

3. List in one column the different distinct observed values of the statistic, and in another column list the corresponding frequency of occurrence of each distinct observed value of the statistic.

Definition 2.6 (Standard error)**.** The standard deviation of the sampling distribution of a statistic is called the *standard error* of the statistic.

2.2 Sampling Distribution of the Mean

1. When σ is known.

   Let $X_1, X_2, \ldots, X_n$ be a random sample of size n drawn from a population of size N having mean μ and variance σ^2, then

$$\sigma^2_{\bar{X}} = \begin{cases} \dfrac{\sigma^2}{n}\left(\dfrac{N-n}{N-1}\right); & \text{if the population is finite \textbf{and} the sampling is without replacement,} \\[2ex] \dfrac{\sigma^2}{n}; & \text{if the population is infinite \textbf{or} the sampling is with replacement,} \end{cases}$$

   where $\sigma_{\bar{X}}$ is called the standard error of $\bar{X}$.

Remark 2.1. The factor $\left(\frac{N-n}{N-1}\right)$ is called "finite population correction" and can be ignored if N is very large (infinite population) or if n represents at most 5 percent from the population, i.e. $\frac{n}{N} \leq 0.05$, in this case $\sigma^2_{\bar{X}} = \frac{\sigma^2}{n}$.

Theorem 2.1. *If all possible random samples of size n are drawn with replacement from a finite population of size N with mean μ and variance σ^2, then the sampling distribution of the mean $\bar{X}$ will be approximately normally distributed with mean $\mu_{\bar{X}} = \mu$ and variance σ^2/n. Hence*

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

2. When σ is unknown.

   In this case we replace σ by S (standard deviation of the sample) and then we have the following two cases:

   (a) If $n \geq 30$, then
   $$Z = \frac{\bar{X}-\mu}{S/\sqrt{n}} \sim N(0,1).$$

   (b) If $n < 30$, then
   $$T = \frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t_\nu$$
   where $\nu = n-1$ is the degrees of freedom of t-distribution.

2.3 Sampling Distribution of the Difference of Means

If we are given two populations, the first with mean μ_1 and variance σ_1^2, and the second with mean μ_2 and variance σ_2^2. Let the values of the variable $\bar{X}_1$ represent the means of random samples of size n_1 drawn from the first population and similarly the values of $\bar{X}_2$ represent the means of random samples of size n_2 drawn from the second population such that the values of $\bar{X}_1$ are independent of the values of $\bar{X}_2$, then

$$\mu_{\bar{X}_1 \pm \bar{X}_2} = \mu_1 \pm \mu_2 \quad \text{and} \quad \sigma^2_{\bar{X}_1 \pm \bar{X}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}.$$

Theorem 2.2. *Suppose that two independent samples of sizes* n_1 *and* n_2 *are drawn from two large populations with means* μ_1 *and* μ_2 *and variances* σ_1^2 *and* σ_2^2*, respectively. Then the sampling distribution of* $\bar{X}_1 - \bar{X}_2$ *is approximately normally distributed with mean and standard error, given by*

$$\mu_{\bar{X}_1-\bar{X}_2} = \mu_1 - \mu_2 \quad \textit{and} \quad \sigma_{\bar{X}_1-\bar{X}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}.$$

Hence,

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1).$$

Example 2.2. *If the uric acid values in normal adult males are approximately normally distributed with a mean and standard deviation of* 5.7 *and* 1 *mg percent, respectively. Find the probability that a sample of size* 9 *will yield a mean:*
(a) Greater than 6*, (b) between* 5 *and* 6.

$$\mu = 5.7, \quad \sigma = 1, \quad n = 9$$

(a)

$$\begin{aligned} P(\bar{X} > 6) &= P\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} > \frac{6 - \mu}{\sigma/\sqrt{n}}\right) \\ &= P(Z > 0.9) = 1 - P(Z \leq 0.9) \\ &= 1 - 0.8159 = 0.1841. \end{aligned}$$

(b)

$$\begin{aligned} P(5 < \bar{X} < 6) &= P(-2.1 < Z < 0.9) \\ &= 0.8159 - 0.0143 = 0.8016. \end{aligned}$$

Example 2.3. *Suppose that a population consists of the following values: 1, 3, 5, 7. Construct the sampling distribution of* $\bar{X}$ *based on samples of size two selected without replacement from the above population. Find the mean and variance of the sampling distribution?*

The number of drawn samples is equal to $C_2^4 = \dfrac{4!}{2! \cdot 2!} = 6.$

Samples	$\bar{x}$
(1,3)	2
(1,5)	3
(1,7)	4
(3,5)	4
(3,7)	5
(5,7)	6

Then we have the following frequency distribution

i	$\bar{x}_i$	f_i	$\bar{x}_i f_i$	${\bar{x}_i}^2 f_i$
1	2	1	2	4
2	3	1	3	9
3	4	2	8	32
4	5	1	5	25
5	6	1	6	36
		6	24	106

$$E[\bar{X}] = \frac{\sum_i \bar{x}_i f_i}{\sum_i f_i} = \frac{24}{6} = 4.$$

$$\sigma^2_{\bar{X}} = \frac{\sum_i \bar{x}_i^2 f_i}{\sum_i f_i} - \left(\frac{\sum_i \bar{x}_i f_i}{\sum_i f_i}\right)^2 = \frac{106}{6} - 16 = \frac{5}{3}.$$

We can note that

$$\mu = \frac{1+3+5+7}{4} = 4 = \mu_{\bar{X}},$$

$$\frac{\sigma^2}{n}\frac{N-n}{N-1} = \frac{5}{2} \times \frac{2}{3} = \frac{5}{3} = \sigma^2_{\bar{X}}.$$

Example 2.4. *Suppose it has been established that for a certain type of client the average length of a home visit by a public health nurse is 45 minutes with a standard deviation of 15 minutes, and that for a second type of client the average home visit is 30 minutes long with a standard deviation of 20 minutes. If a nurse randomly visits 35 clients from the first and 40 for the second group, what is the probability that the average length of home visit will differ between the two groups by 20 or more minutes?*

$$\begin{array}{r|l} \mu_1 = 45 & \mu_2 = 30 \\ \sigma_1^2 = 15 & \sigma_2^2 = 20 \\ \mathrm{n}_1 = 35 & \mathrm{n}_2 = 40. \end{array}$$

We don't know here whether the two populations are normal or not. But, since $n_1 > 30$ and $n_2 > 30$, then the difference between two sample means is approximately normally distributed with the following mean and variance:

$$\mu_{\bar{X}_1 - \bar{X}_2} = \mu_1 - \mu_2 = 45 - 30 = 15,$$

$$\sigma^2_{\bar{X}_1-\bar{X}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} = \frac{(15)^2}{35} + \frac{(20)^2}{40} = 16.4286,$$

and hence

$$\begin{aligned} P((\bar{X}_1 - \bar{X}_2) \geq 20) &= P\left(Z \geq \frac{20-15}{4.05}\right) \\ &= P(Z \geq 1.23) \\ &= 1 - P(Z < 1.23) \\ &= 1 - 0.8907 = 0.1093. \end{aligned}$$

2.4 Sampling Distribution of the Sample Variance (S^2)

When we draw a sample of size n from a normal population with variance σ^2, and the sample variance s^2 is computed for each sample, then we have obtained the values of a statistic S^2. In practice, the sampling distribution of S^2 has little application in statistics. Instead, we shall consider the distribution of a random variable X^2, called chi-square, whose values are calculated from each sample by the formula

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}.$$

The distribution of $X^2 = \frac{(n-1)S^2}{\sigma^2}$ is referred to as the chi-square distribution with $\nu = n - 1$ degrees of freedom.

Example 2.5. *Find the probability that a random sample of size 25, from a normal population with $\sigma^2 = 6$, will have a variance*
(a) greater than 9.1. (b) between 3.462 and 10.745.

(a)

$$\begin{aligned} P(S^2 > 9.1) &= P\left(\frac{(n-1)S^2}{\sigma^2} > \frac{(n-1)9.1}{\sigma^2}\right) \\ &= P\left(X^2 > \frac{24 \times 9.1}{6}\right) \\ &= P(X^2 > 36.4) = \chi^2_{24}(36.4) = 0.05. \end{aligned}$$

(b)

$$\begin{aligned} P(3.462 < S^2 < 10.745) &= P\left(\frac{24 \times 3.462}{6} < X^2 < \frac{24 \times 10.745}{6}\right) \\ &= P(13.848 < X^2 < 42.98) \\ &= \chi^2_{24}(13.848) - \chi^2_{24}(42.98) \\ &= .95 - .01 = 0.94. \end{aligned}$$

Exercises V

(i) Calculate the variance of the sample $3, 5, 8, 7, 5$, and

A finite population consists of the numbers $2, 4$, and 7.

(i) Construct a frequency histogram for the sampling distribution of $\bar{X}$ when samples of size 4 are drawn with replacement.

(ii) Verify that $\mu_{\bar{X}} = \mu$ and $\sigma^2_{\bar{X}} = \sigma^2/n$

(iii) Between what two values would you expect the middle 68% of the sample means to fall?

The heights of 1000 students are approximately normally distributed with a mean of 68.5 inches and a standard deviation of 2.7 inches. If 200 random samples of size 25 are drawn from this population, determine

(i) The expected mean and standard deviation of the sampling distribution of the mean.

(ii) The number of sample means that fall between 66 and 69 inclusive.

(iii) The number of sample means falling below 65.

Chapter 3

POINT AND INTERVAL ESTIMATION

Estimation is the first of the two general areas of statistical inference. The second general area, hypothesis testing, is examined in the next chapter.

In this chapter we shall consider inferences about unknown population parameters such as the mean, variance and proportion.

Definition 3.1 (Statistical inference)**.** The procedure whereby inferences about a population are made on the basis of the results obtained from a sample drawn from that population is called *statistical inference.*

3.1 Methods of Estimation

A population parameter can be estimated by a *point* or an *interval.* A point estimate of some population parameter θ is a single numerical value $\hat{\theta}$ of the statistic $\hat{\Theta}$. For example, the value $\bar{x}$ of the statistic $\bar{X}$, computed from a sample of size n, is a point estimate of the population parameter μ. Similarly, s^2 is a point estimate of the population variance σ^2.

An interval estimate of a population parameter, θ, is given by two values which θ lies within them.

Definition 3.2 (Unbiased estimator)**.** A statistic $\hat{\Theta}$ is said to be an unbiased estimator of the parameter θ if $E(\hat{\Theta}) = \theta$.

The sample mean, the difference between two sample means, the sample proportion, the difference between two sample proportions are unbiased estimates of their corresponding parameters.

Example 3.1. *Prove that S^2 is an unbiased estimator of σ^2.*

Let $X_1, X_2, \ldots, X_n$ is a random sample of size n, that is $X_1, X_2, \ldots, X_n$ are independent and identically distributed, each with mean μ and variance σ^2. Then

$$\begin{aligned}
\because \bar{X} &= \frac{X_1 + X_2 + \cdots + X_n}{n} \\
\therefore E[\bar{X}] &= E\left(\frac{X_1}{n}\right) + E\left(\frac{X_2}{n}\right) + \cdots + E\left(\frac{X_n}{n}\right) \\
&= \frac{\mu}{n} + \frac{\mu}{n} + \cdots + \frac{\mu}{n} = \mu. \\
\text{Also, } V[\bar{X}] = E[(\bar{X} - \mu)^2] &= \frac{\sigma^2}{n^2} + \frac{\sigma^2}{n^2} + \cdots + \frac{\sigma^2}{n^2} = n\frac{\sigma^2}{n^2} = \frac{\sigma^2}{n}.
\end{aligned}$$

Now,

$$\begin{aligned}
\sum_{i=1}^{n}(X_i - \mu)^2 &= \frac{1}{n}\sum_{i=1}^{n}[(X_i - \bar{X}) + (\bar{X} - \mu)]^2 \\
&= \sum_{i=1}^{n}(X_i - \bar{X})^2 + 2(\bar{X} - \mu)\sum_{i=1}^{n}(X_i - \bar{X}) \\
&\quad + n(\bar{X} - \mu)^2 \\
&= (n-1)\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{(n-1)} + 0 + n(\bar{X} - \mu)^2 \\
&= (n-1)S^2 + n(\bar{X} - \mu)^2
\end{aligned}$$

Taking the expectation for both sides, then

$$\begin{aligned}
\sum_{i=1}^{n} E[(X_i - \mu)^2] &= \frac{n-1}{n}E[S^2] + nE[(\bar{X} - \mu)^2] \\
\therefore \sum_{i=1}^{n}\sigma^2 &= (n-1)E[S^2] + n\frac{\sigma^2}{n} \\
(n-1)\sigma^2 &= (n-1)E[S^2] \\
\therefore E[S^2] &= \sigma^2,
\end{aligned}$$

and hence S^2 is an unbiased estimator for σ^2.

3.2 Confidence Intervals

The interval I can be considered a confidence interval for the population parameter, θ, if we can compute the probability that I contains θ. This probability is called the confidence coefficient of the interval.

The procedure of obtaining a confidence interval is to obtain $Q(\hat{\Theta}, \theta)$, which is a function of the estimator $\hat{\Theta}$ and the parameter θ such that the distribution of this quantity does not depend on θ. For fixed α (usually 1% or 5%) we obtain the values Q_1 and Q_2 such that

$$P(Q_1 \leq Q(\hat{\Theta}, \theta) \leq Q_2) = 1 - \alpha.$$

By solving the inequality $Q_1 \leq Q(\hat{\Theta}, \theta) \leq Q_2$ with respect to θ, then

$$Q_1 \leq Q(\hat{\Theta}, \theta) \leq Q_2 \Longleftrightarrow T_1 \leq \theta \leq T_2.$$

Then we can write

$$P(Q_1 \leq Q(\hat{\Theta}, \theta) \leq Q_2) = P(T_1 \leq \theta \leq T_2) = 1 - \alpha,$$

where T_1 and T_2 are called the lower and upper limits, respectively, $1 - \alpha$ is called confidence coefficient.

3.2.1 Confidence interval for the population mean (μ) [σ known]

It is easy now to find a $(1 - \alpha)100\%$ confidence interval for μ of a normal distribution with known variance, σ^2. We know that $Z \sim N(0, 1)$, then by taking

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \equiv Q(\bar{X}, \mu)$$

$$\begin{aligned} P(-z_{\alpha/2} < Z < z_{\alpha/2}) = P(-z_{\alpha/2} < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < z_{\alpha/2}) &= 1 - \alpha \\ P\left(-z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \bar{X} - \mu < z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) &= 1 - \alpha \\ P\left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) &= 1 - \alpha \end{aligned}$$

Theorem 3.1. *A* $(1-\alpha)100\%$ *confidence interval for* μ*, based on a random sample of size* n *with mean* $\bar{X}$*, is*

$$\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \tag{3.1}$$

where $z_{\alpha/2}$ *is the value of standard normal random variable leaving an area* $\alpha/2$ *to the right, i.e.* $P(Z > z_{\alpha/2}) = \alpha/2$.

3.2.2 Confidence interval for the population mean (μ) [σ unknown]

In this case we replace σ^2 by S^2 (sample variance) and then the confidence interval becomes

$$\begin{array}{l}\bar{X} - z_{\alpha/2}\frac{S}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha/2}\frac{S}{\sqrt{n}}; \quad n \geq 30\\ \bar{X} - t_{\alpha/2}\frac{S}{\sqrt{n}} < \mu < \bar{X} + t_{\alpha/2}\frac{S}{\sqrt{n}}; \quad n < 30\end{array}$$

where $t_{\alpha/2}$ is the value of the random variable T having t-distribution, with $\nu = n-1$ degrees of freedom, leaving an area $\alpha/2$ to the right, i.e. $P(T > t_{\alpha/2}) = \alpha/2$.

3.2.3 Determination of sample size for estimating means

We present now a method for determining the sample size requires for estimating a population mean.

Let e denote the error in estimating the population mean represented for example by Inequality (6.1). So

$$e = z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \Longrightarrow n = \left[z_{\alpha/2}\frac{\sigma}{e}\right]^2.$$

Example 3.2. *The average number of heartbeats per minute for a sample of 49 subjects was found to be 90. If the sample is taken from a normal population with variance 100, find* 90%, 95% *and* 99% *confidence interval for the population mean.*

$\bar{X} = 90, \qquad \sigma^2 = 100, \qquad n = 49.$
$1-\alpha = 0.90 \longrightarrow \alpha = 0.1 \longrightarrow \alpha/2 = 0.05 \longrightarrow 1-\alpha/2 = 0.95. \Longrightarrow z_{\alpha/2} = 1.645.$
Since the confidence interval is given by

$$\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

Then 90% confidence interval for μ is given by

$$90 - 1.645 \times \frac{10}{7} < \mu < 90 + 1.645 \times \frac{10}{7}$$

Now,
$1 - \alpha = 0.95 \longrightarrow 1 - \alpha/2 = 0.975 \Longrightarrow z_{\alpha/2} = 1.96$,
$1 - \alpha = 0.99 \longrightarrow 1 - \alpha/2 = 0.995 \Longrightarrow z_{\alpha/2} = 2.6$.
So, 95% and 99% confidence intervals for μ are given, respectively, by

$$90 - 1.96 \times \frac{10}{7} < \mu < 90 + 1.96 \times \frac{10}{7},$$

$$90 - 2.6 \times \frac{10}{7} < \mu < 90 + 2.6 \times \frac{10}{7}.$$

Example 3.3. *Let $\bar{X}$ be the mean of a random sample of size n from a distribution which is $N(\mu, 9)$. Find n such that*

$$P(\bar{X} - 1 < \mu < \bar{X} + 1) = 0.9.$$

$\sigma^2 = 9, \qquad e = 1$
$1 - \alpha = 0.9 \longrightarrow 1 - \alpha/2 = 0.95 \Longrightarrow z_{\alpha/2} = 1.645$, then

$$n = \left[z_{\alpha/2} \frac{\sigma}{e} \right]^2 = \left[1.645 \times \frac{3}{1} \right]^2 \cong 24.$$

Chapter 4

TESTS OF HYPOTHESES

The purpose of hypothesis testing is to aid the clinician, researcher, or administrator in reaching a decision concerning a population by examining a sample from that population.

Definition 4.1 (Statistical hypothesis)**.** A *statistical hypothesis* is an assumption or statement, which may or may not be true, concerning one or more populations.

Hypothesis that we formulate with the hope of rejecting are called *null hypotheses*, denoted by H_0. The null hypothesis is sometimes referred to as a hypothesis of no difference, since it is a statement of agreement with (or no difference form) conditions presumed to be true in the population of interest. The rejection of H_0 leads to the acceptance of an *alternative hypothesis*, denoted by H_1.

Definition 4.2. A *type I* error has been committed if we reject the null hypothesis when it is true.

Definition 4.3. A *type II* error has been committed if we accept the null hypothesis when it is false.

Definition 4.4. The probability of committing a type I error is called the *level of significance* of the test and is denoted by α
i.e. $\alpha = P(\text{type I error})$.

If the alternative hypothesis is one-sided such as $H_1 : \theta > \theta_0$ or $\theta < \theta_0$, the test is called a *one-tailed* test. The critical region for the alternative hypothesis $\theta > \theta_0$ lies entirely in the right tail of the distribution, while the critical region $H_1 : \theta < \theta_0$ lies entirely in the left tail. If the alternative hypothesis $H_1 : \theta \neq \theta_0$, then it is called *two-tailed* test. The critical region here consists of two tails, one in left corresponds to $\theta < \theta_0$ and the other one in the right corresponds to $\theta > \theta_0$, see Fig. ( ).

A test is said to be *significant* if the null hypothesis is rejected at the 0.05 level of significance, and is considered highly significant if the null hypothesis is rejected at the 0.01 level of significance.

4.1 Tests Concerning Means

The steps for testing a hypothesis concerning a population parameter θ against some alternative hypothesis may be summarized as follows:

1. Formulate the null hypothesis, $H_0 : \theta = \theta_0$.
2. Formulate the alternative hypothesis, $H_1 : \theta > \theta_0$, $\theta < \theta_0 \, or \, \theta \neq \theta_0$.
3. Choose a level of significance equal to α which may be 0.05 or 0.01.
4. Select the appropriate test statistic and establish the critical region.
5. Compute the value of the statistic from a random sample of size n.
6. Conclusion: Reject H_0 if the statistic has a value in the critical region, otherwise accept H_0.

Example 4.1. *A doctor developed a new drug claims its efficiency with mean $\mu = 20$ and with standard deviation of 0.5. Test the hypothesis that $\mu = 20$ against the alternative that $\mu \neq 20$. If a random sample of 50 patients is tested and found a mean $\bar{x} = 19.8$. Use 0.01 level of significance.*

1. $H_0 : \mu = 20$.
2. $H_1 : \mu \neq 20$.
3. $\alpha = 0.01$.
4. Suppose that $z = \dfrac{\bar{x} - \mu}{\sigma/\sqrt{n}}$, so the critical region is
   $z < -z_{\alpha/2}$ and $z > z_{\alpha/2}$
   $z_t < -2.58$ $z_t > 2.58$.
5. Computation: $\bar{x} = 19.8$, $n = 50$
   $z_c = \dfrac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \dfrac{19.8 - 20}{0.5/\sqrt{50}} = -2.828.$

6. Conclusion: Reject H_0 since $z_c < z_t$, and conclude that the drug is highly significant.

Example 4.2. *Teat the hypothesis that the average weight of containers of a particular lubricant is* 10 *ounces if the weights of a random sample of* 10 *containers are* $10.2, 9.7, 10.1, 10.3, 10.1,$ $9.8, 9.9, 10.4, 10.3$ *and* 9.8 *ounces? Use a* 0.01 *level of significance and assume that the distribution of weights is normal.*

$$n = 10, \quad \mu = 10$$

$$\bar{x} = \frac{1}{10}\sum_{i=1}^{10} x_i = \frac{100.6}{10} = 10.06,$$

$$s^2 = \frac{1}{10 \times 9}\left[\sum_{i=1}^{10} x_i^2 - \left(\sum_{i=1}^{10} x_i\right)^2\right] \Longrightarrow s = 0.245.$$

1. $H_0 : \mu = 10.$
2. $H_1 : \mu \neq 10.$
3. $\alpha = 0.01.$
4. Suppose that $t = \dfrac{\bar{x} - \mu}{s/\sqrt{n}}$, so the critical region is
   $t < -t_{\alpha/2}$ and $t > t_{\alpha/2}$
   $t_t < -3.25$ $t_t > 3.25.$
5. Computation: $t_c = \dfrac{\bar{x} - \mu}{s/\sqrt{n}} = \dfrac{10.06 - 10}{0.245/\sqrt{10}} = 0.769.$
6. Conclusion: Accept H_0, since $-t_t < t_c < t_t$.

4.1.1 Contingency tables

The *contingency table* is used for the purpose of studying the relationship between two variables, each variable has different levels. Consider, for example, the factor A classified into n levels $(A_1, \ldots, A_n)$ and factor B classified into m levels $(B_1, \ldots, B_m)$ and it is desired to test the hypothesis that there is no relationship between the two factors A and B. If o_{ij} denote to the observed frequency in the i^{th} classification A_i for A and j^{th} classification B_j for B, and suppose that

$$R_i = \sum_{j=1}^{m} o_{ij}\,; \qquad i = 1, \ldots, n,$$

$$C_j = \sum_{i=1}^{n} o_{ij}\,; \qquad j = 1, \ldots, m, \quad \text{and}$$

$$N = \sum_{i=1}^{n} R_i = \sum_{j=1}^{m} C_j\,; \quad i = 1, \ldots, n \quad j = 1, \ldots, m,$$

A$\diagdown B$	B_1	B_2	$\ldots$	B_j	$\ldots$	B_m	Total
A_1	o_{11}	o_{12}	$\ldots$	o_{ij}	$\ldots$	o_{1m}	R_1
A_2	o_{21}	o_{22}	$\ldots$	o_{2j}	$\ldots$	o_{2m}	R_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
A_i	o_{i1}	o_{i2}	$\ldots$	o_{ij}	$\ldots$	o_{im}	R_i
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
A_n	o_{n1}	o_{n2}	$\ldots$	o_{nj}	$\ldots$	o_{nm}	R_n
Total	C_1	C_2	$\ldots$	C_j	$\ldots$	C_m	$N = \sum C_j = \sum R_i$

If the null hypothesis is satisfied (the two factors are independent), then we have, for $i = 1, \ldots, n, \quad j = 1, \ldots, m$,

$$e_{ij} = \frac{R_i \times C_j}{N} \quad \text{and then } \chi^2 = \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}.$$

The number of degrees of freedom $\nu = (n-1)(m-1)$, and the null hypothesis will be rejected if $\chi^2 > \chi^2_{\alpha,\nu}$.

Example 4.3. *A random sample of* 30 *adults are classified according to sex and the number of hours they watch television during a week.*

	Male	*Female*
Over 25 *hours*	5	9
Under 25 *hours*	9	7

Using a 0.01 *level of significance, test the hypothesis that a person's sex and time watching television are independent.*

1. H_0 : A person's sex and time watching television are independent.
2. H_1 : A person's sex and time watching television are dependent.
3. $\alpha = 0.01$.
4. We use $\chi_{(\alpha,\nu)^2} = \sum_i \sum_j \dfrac{(o_{ij} - e_{ij})^2}{e_{ij}}$,
   where the critical region is $\chi^2 > \chi^2_{\alpha,\nu} = \chi^2_{0.01,1} = 6.635$.

5. Computations:

$$e_{11} = \frac{14 \times 14}{30} = 6.5 \quad e_{12} = \frac{14 \times 16}{30} = 7.5$$
$$e_{21} = \frac{14 \times 16}{30} = 7.5 \quad e_{22} = \frac{16 \times 16}{30} = 8.5$$

	Male	Female	Total
Over 25 hours	5 (6.5)	9 (7.5)	14
Under 25 hours	9 (7.5)	7 (8.5)	16
Total	14	16	30

$$\begin{aligned}\chi_c^2 &= \sum_{i=1}^{2}\sum_{j=1}^{2}\frac{(o_{ij}-e_{ij})^2}{e_{ij}}\\ &= \frac{(5-6.5)^2}{6.5}+\frac{(9-7.5)^2}{7.5}+\frac{(9-7.5)^2}{7.5}+\frac{(7-8.5)^2}{8.5}\\ &= 0.538\end{aligned}$$

6. Conclusion: Accept H_0, since $\chi_c^2 < 6.635$.

REFERENCES

Guttman, I; Wilks, S and Hunter, J. (1982). *Introductory Engineering Statistics* . 3$^{\text{rd}}$ Ed., John Wiley & Sons, Inc.

Hogg, R.; McKean, J. and Craig, A. (2005). *Introduction to Mathematical Statistics.* 6$^{\text{th}}$ Ed. Pearson Prentice Hall.

Mood, A.; Graybill, F. and Boes, D. (1982). *Introduction to the Theory of Statistics.* 3$^{\text{rd}}$ Ed. 12$^{\text{th}}$ printing, McGraw-Hill.

Rosner, B. (1982). *Fundamentals of Biostatistics.* PWS Publishers, Duxbury Press, Boston, Massachusetts.