1 Some Facts About Random Variables

Proposition 1 *Let X_1 and X_2 be any two random variables with finite expected values, then*

$$\mathrm{E}[X_1 + X_2] = \mathrm{E}[X_1] + \mathrm{E}[X_2] \ .$$

Proposition 2 *Let X be any random variable with a finite expected value and a, b real numbers, then*

$$\mathrm{E}[aX + b] = a\mathrm{E}[X] + b$$

and

$$\mathrm{Var}[aX + b] = a^2\mathrm{Var}[X] \ .$$

Proposition 3 *Let X_1 and X_2 be two independent random variables with finite variances, then*

$$\mathrm{Var}[X_1 + X_2] = \mathrm{Var}[X_1] + \mathrm{Var}[X_2] \ .$$

Proposition 4 *Let $X \sim N(\mu, \sigma^2)$ and a, b be two real number with $a \neq 0$. Then*

$$aX + b \sim N(a\mu + b, a^2\sigma^2) \ .$$

Theorem 1 (Standardization Theorem) *$X \sim N(\mu, \sigma^2)$ if and only if $Z = (X - \mu)/\sigma \sim N(0, 1)$.*

Proposition 5 *Let X_1 and X_2 be two independent random variables such that $X_1 \sim N(\mu_1, \sigma_1^2)$ and $X_2 \sim N(\mu_2, \sigma_2^2)$, then*

$$X_1 + X_2 \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$$

and

$$X_1 - X_2 \sim N(\mu_1 - \mu_2, \sigma_1^2 + \sigma_2^2) \ .$$

2 Central Limit Theorem

Theorem 2 (Central Limit Theorem) *Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed (iid) random variables with expected value μ and variance σ^2. If n is "large", then the random variable $\bar{X}$ defined by*

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

is approximately normally distributed with an expected value μ and variance σ^2/n.

Corollaries to the Central Limit Theorem are:

Corollary 1 *If $\bar{X}$ is the sample mean of a random sample (with or without replacement), of size n, take from a infinite population with a mean μ and variance σ, then for large n ($n \geq 30$) we have*

$$\bar{X} \sim N(\mu, \sigma^2/n) \ .$$

Corollary 2 *If $\bar{X}$ is the sample mean of a random sample with replacement, of size n, take from a finite population with a mean μ and variance σ, then for large n ($n \geq 30$) we have*

$$\bar{X} \sim N(\mu, \sigma^2/n) \ .$$

Corollary 3 *If $\hat{p}$ is the sample proportion of a random sample (with replacement if the population is of finite size) taken from a population with a population proportion p, then for large n ($np \geq 5$ and $n(1-p) \geq 5$) we have*

$$\hat{p} \sim N(p, p(1-p)/n) \ .$$

Corollary 4 *For large n ($np \geq 5$ and $n(1-p) \geq 5$) we have* $\text{Bin}(n,p) \sim N(np, np(1-p))$.

3 Estimation

A descriptive measure used to describe a population is called a *parameter*. Therefore a *parameter* is a characteristic of a population (like population mean, median, variance, etc.). A descriptive measure used to describe a sample is called a *statistic*. Therefore a *statistic* is a characteristic of a sample (like sample mean, median, variance, etc.).

An *estimator* is any method to guess an unknown parameter using sample data. For example, the sample mean is an estimator of the population mean. Note that an estimator is a random variable. An *estimate* is an observed value of an estimator. For example, if we use the sample median as an estimator for the population median, and the median of the sample we take is 5, then our estimate for the population median is 5.

In general, if Ω is the parameter to be estimated then we denote an estimator of Ω with $\hat{\Omega}$. Given a sample data $x_1, x_2, \ldots, x_n$ the value $\hat{\Omega}(x_1, x_2, \ldots, x_n)$ is an estimate for Ω.

Below we give several estimators for the population mean μ.

$$\begin{aligned}
\hat{\mu}^1(x_1, \ldots, x_n) &= \frac{1}{n}(x_1 + \ldots + x_n) \\
\hat{\mu}^2(x_1, \ldots, x_n) &= x_1 \\
\hat{\mu}^3(x_1, \ldots, x_n) &= \frac{1}{2}(\min\{x_1, \ldots, x_n\} + \max\{x_1, \ldots, x_n\} \\
\hat{\mu}^4(x_1, \ldots, x_n) &= \text{Median of } \{x_1, x_2, \ldots, x_n\}
\end{aligned}$$

Consider a sample consisting of $1, 4, 2, 5, 6$ take from a population with unknown mean. The following table gives, for each of the estimators given above, the estimate based on the given sample.

Estimator	Estimate
$\hat{\mu}^1$	4.2
$\hat{\mu}^2$	1.0
$\hat{\mu}^3$	3.5
$\hat{\mu}^4$	4.0

Now we are faced with the problem of choosing an estimator. We would like our estimator to satisfy the following properties:

Unbiasedness An estimator for a given parameter is *unbiased* if its expected value is equal to the parameter, i.e., $E(\hat{\Omega}) = \Omega$. The estimators μ^1 and μ^2 are unbiased estimators for the population mean. The other estimators are not unbiased (biased) estimators for the population mean.

Consistency An estimator is *consistent* if it is unbiased and its variance approaches zero as the sample size increases. The estimator μ^1 is a consistent estimator for the population mean but μ^2 is not.

Efficient An estimator $\hat{\Omega}^1$ is more efficient than an estimator $\hat{\Omega}^2$ is $\text{Var}[\hat{\Omega}^1] < \text{Var}[\hat{\Omega}^2]$.

The sample mean, which we will denote with $\bar{x}$ rather than $\hat{\mu}$, is the "best" estimator for the population mean. For the population variance we will use the estimator, which we will denote with s^2, define by

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \ ,$$

where $x_1, \ldots, x_n$ is the sample data. Note that this is not the variance of the data (sample) consisting of $x_1, \ldots, x_n$, but it is an estimate for the population variance based on the sample.

4 Confidence Interval

A *confidence interval* is an interval believed to include an unknown parameter. With the confidence interval we associated number representing a measure of the confidence we have that the interval does indeed contain the parameter of interest. Formally:

Definition 1 (Confidence Interval) *A $(1-\alpha)100\%$ confidence interval for a population parameter Ω is an interval $[l, u]$ (obtained from the sample information) such that*

$$P(l \leq \Omega \leq u) = 1 - \alpha \ .$$

Thus a $(1-\alpha)100\%$ confidence interval for the population mean is an interval $[l, u]$ such that

$$P(l \leq \mu \leq u) = 1 - \alpha \ .$$

Note that if $[l, u]$ is a $(1-\alpha)100\%$ confidence interval for Ω then we have

$$P(\Omega < l) + P(\Omega > u) = \alpha \ .$$

Thus to find a confidence interval for Ω we have to find any two real numbers l and u such that the above equation holds. We can simply the task of finding l and u which satisfies the above condition by requiring l and u to be such that

$$P(\Omega < l) = \frac{\alpha}{2} \quad \text{and} \quad P(\Omega > u) = \frac{\alpha}{2} \ .$$

Note that Ω in the above statement is not a random variable. The number l and u are calculated by using the sample information. Hence both l and u are random variables and to find their values we need the distributions of an estimator for Ω.

5 Distribution of some Estimators

Theorem 3 *Standardization Theorem for sample mean when population variance is not know: If* $\bar{x} \sim N(\mu, \sigma^2/n)$ *then*

$$t_{n-1} = \frac{\bar{x} - \mu}{s/\sqrt{n}} \sim T_{n-1} \ .$$

Sample mean If the population has a normal distribution or if the samples size is large ($n \geq 30$) then the distribution of sample means, $\bar{x}$, is normal (approximately normal if the distribution of the population is not normal) with a expected value of μ (the population mean) and variance

- $\frac{\sigma^2}{n}\frac{N-n}{N-1}$ if the population is finite and sampling is done without replacement.
- $\frac{\sigma^2}{n}$ otherwise, i.e., if the population is infinite or sampling is done with replacement.

Sample proportion If np and $n(1-p)$ are large (greater than 5) then the distribution of sample proportion is approximately normal with expected value np and variance $p(1-p)/n$.

Sample variance If the population is normally distributed then the distribution of $(n-1)s^2/\sigma^2$ is χ^2_{n-1} (chi-square) with $n-1$ degrees of freedom.

Ratio of variances If two populations are normally distributed then the distribution of $\frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2}$ is a F_{n_1-1,n_2-1} distribution with numerator degrees of freedom $n_1 - 1$ and denominator degrees of freedom $n_2 - 1$

5.1 Confidence Intervals for some Parameters

5.1.1 Confidence interval for population mean μ

If the population standard deviation, σ, is know, then the following interval gives a $100(1-\alpha)\%$ confidence interval for μ:

$$\left(\bar{x} - z_{1-\frac{\alpha}{2}}\sigma_{\bar{x}},\ \bar{x} + z_{1-\frac{\alpha}{2}}\sigma_{\bar{x}}\right)$$

where $\bar{x}$ is a sample mean, $z_{1-\frac{\alpha}{2}}$ is a number such that $P(z \leq z_{1-\frac{\alpha}{2}}) = 1 - (\alpha/2)$, and $\sigma_{\bar{x}}$ is equal to

- $\frac{\sigma}{\sqrt{n}}$, if the population is infinite, or sampling is done with replacement (n is the sample size).

- $\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$, if the population is finite, and sampling is done without replacement (N is the population size, and n is the sample size).

If the population standard deviation, σ, is unknown then we can find a confidence interval for μ by using the sample standard deviation, $s = \sqrt{(\sum_{i=1}^{n}(x_1 - \bar{x})^2)/(n-1)}$, i.e., replace σ in the above expressions with s. But, then we have to replace $z_{1-\frac{\alpha}{2}}$ in the expressions with $t_{n-1,1-\frac{\alpha}{2}}$. Thus a $100(1-\alpha)\%$ confidence interval for μ when population standard deviation is unknown is:

$$\left(\bar{x} - t_{n-1,1-\frac{\alpha}{2}}\sigma_{\bar{x}},\ \bar{x} + t_{n-1,1-\frac{\alpha}{2}}\sigma_{\bar{x}}\right)$$

where $\sigma_{\bar{x}}$ is equal to

- $\frac{s}{\sqrt{n}}$, if the population is infinite, or sampling is done with replacement.
- $\frac{s}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$, if the population is finite, and sampling is done without replacement.

Note that when n is large $t_{n-1,1-\frac{\alpha}{2}} \cong z_{1-\frac{\alpha}{2}}$.

5.1.2 Confidence interval for population proportion p

When np_s and $n(1-p_s)$ are large (greater than 5) then a $100(1-\alpha)\%$ confidence interval for population proportion is

$$\left(p_s - z_{1-\frac{\alpha}{2}}\sqrt{\frac{p_s(1-p_s)}{n}},\ p_s + z_{1-\frac{\alpha}{2}}\sqrt{\frac{p_s(1-p_s)}{n}}\right)$$

where p_s is the sample proportion.

5.1.3 Confidence interval for population variance σ^2

If the population is normally distributed, then a $100(1-\alpha)\%$ confidence interval for the population variance σ^2 is

$$\left(\frac{(n-1)s^2}{\chi^2_{n-1,1-\frac{\alpha}{2}}},\ \frac{(n-1)s^2}{\chi^2_{n-1,\frac{\alpha}{2}}}\right) \quad .$$

5.1.4 Confidence interval for the difference between two population means $\mu_1 - \mu_2$

If the population variances σ_1^2 and σ_2^2 are known the following interval gives a $100(1-\alpha)\%$ confidence interval for μ:

$$\left((\bar{x}_1 - \bar{x}_2) - z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}},\ (\bar{x}_1 - \bar{x}_2) + z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$$

where $\bar{x}_1$ and $\bar{x}_2$ are means of samples take from the two populations respectively, n_1 and n_2 are the samples sizes of the respective samples.

If the population variances are not know, then we have the following cases:

Unknown but equal variances: a $100(1-\alpha)\%$ confidence interval for $\mu_1 - \mu_2$ for this case is

$$\left((\bar{x}_1 - \bar{x}_2) - t_{n_1+n_2-2,1-\frac{\alpha}{2}} \sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{2}},\ (\bar{x}_1 - \bar{x}_2) + t_{n_1+n_2-2,1-\frac{\alpha}{2}} \sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}} \right)$$

where

$$s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2} \ .$$

Unknown but unequal variances a $100(1-\alpha)\%$ confidence interval for $\mu_1 - \mu_2$ for this case is

$$\left((\bar{x}_1 - \bar{x}_2) - t' \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}},\ (\bar{x}_1 - \bar{x}_2) + t' \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \right)$$

where

$$t' = \frac{\frac{s_1^2}{n_1} t_{n_1-1,1-\frac{\alpha}{2}} + \frac{s_2^2}{n_2} t_{n_2-1,1-\frac{\alpha}{2}}}{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \ .$$

Note that, if the sample sizes are large the intervals give above will approximately be equal to the interval

$$\left((\bar{x}_1 - \bar{x}_2) - z_{1-\frac{\alpha}{2}} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}},\ (\bar{x}_1 - \bar{x}_2) + z_{1-\frac{\alpha}{2}} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \right)$$

5.1.5 Confidence interval for the ratio of two population variances σ_1^2/σ_2^2

A confidence interval for the ratio population variances, σ_1^2/σ_2^2, of two normally distributed populations is

$$\left(\frac{F_{n_1-1,n_2-1,\frac{\alpha}{2}}}{s_1^2/s_2^2},\ \frac{F_{n_1-1,n_2-1,1-\frac{\alpha}{2}}}{s_1^2/s_2^2} \right) \ .$$

Note that $F_{n_1-1,n_2-1,\frac{\alpha}{2}} = 1/F_{n_2-1,n_1-1,1-\frac{\alpha}{2}}$.