MAT 2377 3X (Spring 2011)

Sampling Distributions and Point Estimation of Parameters

Refer to Sections 7-1, 7-2, 7-3

Random Sampling Terminology : In practice we often repeat an experiment n times.

Examples :

- Randomly selecting n items from a production line and verifying the conformity of each of the selected items.
- Measuring the current in a thin copper wire n times.

We will model each of the observations as a random variable, i.e. X_i will represent the ith observation.

Assumptions :

1. We will assume **independent** trials.
2. Furthermore, we are repeating the same experiment n times, hence the random variables $X_1, \ldots, X_n$ are **identically distributed**. This common distribution is called the **population**.

Random Sample : We call the following collection of independent and identically distributed random variables $X_1, \ldots, X_n$ a **random sample**.

Population Parameters : Here are some population parameters that often need to estimate in practice :

- The mean of a population μ, called the **population mean**.
- The variance of a population σ^2, called the **population variance**.
- The proportion p of items in a population that satisfy an attribute, called the **population proportion**.

Estimation : A function of a random sample is called a **statistic**. In practice we use a statistic

$$\widehat{\Theta} = h(X_1, X_2, \ldots, X_n)$$

to estimate an unknown population parameter θ, then we say that $\widehat{\Theta}$ is a point estimator of θ. Note that $\widehat{\Theta}$ is a random variable. The observed value of $\widehat{\Theta}$ which is

$$\widehat{\theta} = h(x_1, x_2, \ldots, x_n)$$

is called a point estimate of θ.

Here are some common statistics.

1. The **sample mean** $\overline{X}$ is defined as that average of $X_1, \ldots, X_n$, that is

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}.$$

Remarks :
- We will use $\overline{X}$ as a point estimator of the population mean μ.
- The observed value of $\overline{X}$ which is denoted $\overline{x}$ is a point estimate for μ.

2. The **sample variance** S^2 is defined as follows :

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1} = \frac{\left(\sum_{i=1}^{n} X_i^2\right) - n\,\overline{X}^{\,2}}{n-1}$$

Remarks :
- We will use S^2 as a point estimator of the population variance σ^2.
- The observed value of S^2 which is denoted s^2 is a point estimate for σ^2.

3. The **sample standard deviation** S is defined as $S = \sqrt{S^2}$.

Remarks :
- We will use S as a point estimator of the population standard deviation σ.
- The observed value of S which is denoted s is a point estimate for σ.

4. The **sample proportion** $\widehat{P}$ is defined as follows :

$$\widehat{P} = \frac{X}{n},$$

where X is the number of selected items that satisfy the attribute of interest among n items.

Remarks :
- We will use $\widehat{P}$ as a point estimator of the population proportion p.
- The observed value of $\widehat{P}$ which is denoted $\widehat{p} = x/n$ is a point estimate of p.

Example 1 : Consider the following data that represent the life of packaged magnetic disks exposed to corrosive gases (in hours) :

$$4, \quad 86, \quad 335, \quad 746, \quad 195$$

(a) Give point estimates for the mean and the standard deviation of the lifetime of a packaged magnetic disk exposed to corrosive gases.

(b) Let p denote the probability that a packaged magnetic disk exposed to corrosive gases will have a life of more than 100 hours. Find a point estimate for p.

Expectations and Variances of a Linear Function of the Random Sample

We will give a result from Section 5-5. These results will allow us to study the properties of some estimators.

Expectation : Consider the following random variables $X_1, \ldots, X_n$ and the following constants $c_0, c_1, \ldots, c_n$. If Y is defined as

$$Y = c_0 + c_1 X_1 + \ldots + c_n X_n,$$

then it can be shown that the expected value of Y is

$$E[Y] = c_0 + c_1 E[X_1] + \ldots + c_n E[X_n].$$

Variance : Consider the following random variables $X_1, \ldots, X_n$ and the following constants $c_0, c_1, \ldots, c_n$. If Y is defined as

$$Y = c_0 + c_1 X_1 + \ldots + c_n X_n,$$

and we assume that $X_1, \ldots, X_n$ are **independent**, then it can be shown that the variance of Y is

$$\sigma_Y^2 = c_1^2 \sigma_{X_1}^2 + \ldots + c_n^2 \sigma_{X_n}^2.$$

Corollary : Let X be a random varible and a and b constants, then

$$E[a X + b] = a E[X] + b \quad \text{and} \quad V[a X + b] = a^2 \sigma_X^2.$$

Example 2 :
(a) Consider a random sample $X_1, \ldots, X_n$ from a population with mean μ and variance σ^2. Show that the sample mean has the following mean and variance :

$$\mu_{\overline{X}} = \mu \quad \text{and} \quad \sigma^2_{\overline{X}} = \frac{\sigma^2}{n}.$$

(b) Let $\widehat{P} = X/n$ be the sample proportion from a population with probability of success p. Show that

$$E[\widehat{P}] = p \quad \text{and} \quad \sigma^2_{\widehat{P}} = \frac{p\,(1-p)}{n}.$$

Terminology : The standard deviation of a point estimator is called the standard error of the estimate. So the standard error of the mean and of the sample proportion are respectively :

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}} \quad \text{and} \quad \sigma_{\widehat{P}} = \sqrt{\frac{p\,(1-p)}{n}}.$$

Remark : The standard error usually involves an unknown parameter, so in practice when giving data we replace this unknown parameter by its point estimate. So the estimated standard error of the mean and of the sample proportion become in this case :

$$\widehat{\sigma}_{\overline{X}} = \frac{s}{\sqrt{n}} \quad \text{and} \quad \widehat{\sigma}_{\widehat{P}} = \sqrt{\frac{\widehat{p}\,(1-\widehat{p})}{n}}$$

Example 3 : Refer to Example 1. Give the estimated standard errors for the mean from part (a) and the sample proportion from part (b).

Sampling Distribution

The distribution of a statistic $h(X_1, \ldots, X_n)$ is known as a sampling distribution. The following two theorems concerns the sampling distribution of the sample mean $\overline{X}$.

Theorem : Let $X_1, \ldots, X_n$ be **independent normal** random variables such that X_i follows a $N(\mu_i, \sigma_i^2)$ distribution for $i = 1, \ldots, n$. If we define Y as

$$Y = c_0 + c_1 X_1 + \ldots + c_n X_n,$$

then it can be shown that Y follows a $N(\mu_Y, \sigma_Y^2)$ distribution.

Corollary : Let $X_1, \ldots, X_n$ be a random sample from a $N(\mu, \sigma^2)$ population, then the sample mean $\overline{X}$ follows a $N(\mu, \sigma^2/n)$ distribution.

Remark : So now we know that if the population is normally distributed, then the sample mean has a normal sampling distribution. Now in general a population is not necessarily normally distributed. The following theorem tells us that as long a we collect a large number of observations, then the sample mean is approximately normally distributed regardless of the underlying distribution.

Central Limit Theorem : Let $X_1, \ldots, X_n$ be a random sample from a population with mean μ and variance σ^2. Define Z as

$$Z = \frac{\overline{X} - \mu_{\overline{X}}}{\sigma_{\overline{X}}} = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}.$$

Then, as $n \to \infty$, it can be shown that Z follows a **standard normal distribution**.

Rule of thumb : We will take $n \geq 30$, as large enough. That is, if $n \geq 30$, then $\overline{X}$ follows approximately a $N(\mu, \sigma^2/n)$ distribution.

Example 4 : The compressive strength of concrete has a mean of 2500 psi and a standard deviation of 50 psi. What is the probability that the mean of 40 specimens will be less than 2490 psi ?

Comparing means from independent populations : Let $\overline{X}_i$ be the sample mean from a $N(\mu_i, \sigma_i^2)$ population, for $i = 1, 2$. Let n_i be the size of the ith random sample.

We know that

$$\overline{X}_i \text{ follows a } N(\mu_i, \sigma_i^2/n_i) \text{ distribution.}$$

Since $\overline{X}_1 - \overline{X}_2$ is a linear function of independent normal random variables, then it is also normally distributed. Hence,

$$Z = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$$

is a standard normal random variable. **Remark :** By the central limit theorem, if $n_1 \geq 30$ and $n_2 \geq 30$, then Z is approximately a standard normal random variable.

Example 5 : The effective life of a component in a jet-turbine aircraft is normally distributed with mean 5000 hours and standard deviation 40 hours. Suppose that when using an improved manufacturing process the effective is normally distributed with mean 5050 and standard deviation 30. Let $\overline{X}_1$ be the average life of $n_1 = 20$ components produced under the old manufacturing process and let $\overline{X}_2$ be the average life of $n_1 = 15$ components produced with the new manufacturing process. Find the following probability :

$$P\left(\overline{X}_1 - \overline{X}_2 > 10\right).$$

Properties of Point Estimators - Refer to Section 7-3

Definition : Let $\widehat{\Theta}$ be a point estimator for the population parameter θ. We say that it is an **unbiased** estimator if

$$E[\widehat{\Theta}] = \theta.$$

It the estimator is not unbiased, then we define the the difference

$$E[\widehat{\Theta}] - \theta$$

is called the bias.

Remark : We have already shown that

$$E[\overline{X}] = \mu \quad \text{and} \quad E[\widehat{P}] = p.$$

Hence the sample mean is an unbiased estimator for μ and the sample proportion is an unbiased estimator for the population proportion p.

Example 6 : Show that the sample variance

$$S^2 = \frac{(\sum_{i=1}^{n} X_i^2) - n\,\overline{X}^{\,2}}{n-1}$$

is an unbiased estimator for the population variance σ^2, that is show that $E[S^2] = \sigma^2$.

Comparing Unbiased Estimators

Suppose that $\widehat{\Theta}_1$ and $\widehat{\Theta}_2$ are unbiased estimators for the parameter θ. We will say that $\widehat{\Theta}_1$ is best for estimating θ, if

$$V[\widehat{\Theta}_1] < V[\widehat{\Theta}_2].$$

That is, we will prefer the estimator with the smaller variance.

Example 7 : Let X_1, X_2, X_3 be a random sample from a population with mean μ and variance σ^2.

$$\widehat{\mu}_1 = \frac{X_1 + X_3}{2} \quad \text{and} \quad \widehat{\mu}_1 = X_2.$$

(a) Show that both estimators are unbiased estimators for μ.

(b) Which is best for estimating μ?

Definition : The **mean squared error** of an estimator $\widehat{\Theta}$ for the parameter θ is defined as

$$\mathsf{MSE}(\widehat{\Theta}) = E[(\widehat{\Theta} - \theta)^2] = V(\widehat{\Theta}) + (\text{ bias })^2.$$

Remarks :

- If the estimator is unbiased (i.e. bias=0) then the mean squared error of the estimator is equal to the variance of the estimator.
- We can use the mean squared error to compare estimators. The relative efficiency of $\widehat{\Theta}_2$ compared to $\widehat{\Theta}_1$ is defined as

$$\frac{\mathsf{MSE}(\widehat{\Theta}_1)}{\mathsf{MSE}(\widehat{\Theta}_2)}.$$

  If the relative efficiency is less than 1, then we conclude that $\mathsf{MSE}(\widehat{\Theta}_1)$ is a more efficient estimator of θ than $\mathsf{MSE}(\widehat{\Theta}_2)$.

Example 8 : Let $X_1, \ldots, X_{20}$ be a random sample from a population with mean $\mu = 10$ and variance $\sigma^2 = 5$. Consider the following two estimators of μ :

$$\widehat{\mu}_1 = \frac{X_1 + \ldots + X_{19}}{20} \quad \text{and} \quad \widehat{\mu}_2 = \frac{X_1 + X_{20}}{2}.$$

Which of the two estimators is more efficient estimator of μ?