MAT 2379, Introduction to Biostatistics

Section 9.2 Sampling Distributions and Point Estimation

So far in this course, we learned:
-how to calculate probabilities associated to some events which arise from a random experiment;
-how to describe a measurement X which arises from a random experiment (we looked at its associated probabilities, its mean $\mu = E(X)$ and its variance $\sigma^2 = \text{Var}(X)$);
-how to an analyze a data set of measurements $x_1, \ldots, x_n$.

From now on, we will use these techniques to draw some conclusions about an unknown *parameter*, which is associated with a measurement X. The measurement X describes a certain characteristic of the *population*. Our conclusions will be based on a *random sample* selected from the population.

A **random sample** is a sample in which the subjects are selected randomly and each subject in the population has the same chance of of being selected. (There are various methods that can be used for selecting such a sample. These methods will not be investigated in this course.)

Typically, we are interested in the average $\mu = E(X)$ (μ is the unknown parameter).

Example 1. X is the growth of a randomly chosen seedling in a lab. μ is the mean growth of all seedlings in a lab (the population).

The measurement X itself is random (i.e. it cannot be predicted). But, if we select a sample of size n from the population, we observe some values $x_1, \ldots, x_n$ of X. These values are interpreted as follows:
-x_1 is the observed value of a (theoretical) random measurement X_1,
-x_2 is the observed value of a (theoretical) random measurement X_2, etc.
. . . . . .
-x_n is the observed value of a (theoretical) random measurement X_n, where

$X_1, X_2, \ldots, X_n$ are *independent* random variables with the *same distribution* as X.

By abuse of terminology, we will say that $X_1, \ldots, X_n$ is also a random sample. The **sample mean** of $X_1, \ldots, X_n$ is a new random variable $\bar{X}$ defined by:

$$\bar{X} := \frac{X_1 + \ldots + X_n}{n}$$

In general, $\bar{X}$ does *not* have the same distribution as the original measurement X.

Important Remark: There is a clear distinction between:

Theoretical Values:	$X_1, \ldots, X_n$	$\bar{X}$
Observed Values:	$x_1, \ldots, x_n$	$\bar{x}$

Example 1. (continued) As part of a study on plant growth, a researcher grew 6 soybean seedlings under identical environmental conditions and measured the total stem length (in cm) for each plant after 16 days. He obtained the following data:

$$x_1 = 20, \quad x_2 = 22, \quad x_3 = 19, \quad x_4 = 21, \quad x_5 = 23, \quad x_6 = 19$$

These can be regarded as the observed values for some random variables $X_1, \ldots, X_6$. The observed value of $\bar{X}$ (for this particular sample) is:

$$\bar{x} = \frac{20 + 22 + 19 + 21 + 23 + 19}{6} = 20.67$$

Since $E(X) = \mu$, and $X_1, \ldots, X_n$ have the same distribution as X, we have

$$E(X_1) = E(X_2) = \ldots = E(X_n) = \mu.$$

Hence,

$$\mu_{\bar{X}} = E(\bar{X}) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n}\sum_{i=1}^{n} \mu = \frac{1}{n}(n\mu) = \mu.$$

This shows that the average of $\bar{X}$ is equal to μ. Hence, the observed (numeric) value $\bar{x}$ is a good approximation for the unknown value μ. In statistics, we say that:

$$\bar{X} \text{ is an } \textit{estimator} \text{ of } \mu$$
$$\bar{x} \text{ is an } \textit{estimate} \text{ of } \mu.$$

The variance of $\bar{X}$ is also very important, since it shows how the theoretical value $\bar{X}$ fluctuates around μ. One can show that: (see Example 9.10 in the book)

$$\sigma^2_{\bar{X}} = \mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}, \quad \text{where} \quad \sigma^2 = \mathrm{Var}(X).$$

The standard deviation of $\bar{X}$ is:

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

Example 2. Consider the variable X which denotes the weight (in kg) of a female lamb at birth. The lambs were all born in April. Assume that X is normally distributed with mean $\mu = 6.2$ and standard deviation $\sigma = 0.6$. Let $\bar{X}$ be the mean of a sample of 49 female lambs. Then

$$E(\bar{X}) = 6.2, \quad \mathrm{Var}(\bar{X}) = \frac{0.6^2}{49} = 0.007, \quad \sigma_{\bar{X}} = \frac{0.6}{7} = 0.085$$

The most important thing is the distribution of $\bar{X}$. This is given by the following result: (see Theorem 9.5 in the book)

Central Limit Theorem (CLT): If the sample size n is large enough, then the distribution of $\bar{X}$ has *approximately a normal distribution* with mean $\mu_{\bar{X}} = \mu$ and standard deviation $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$. That is,

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \quad \text{has approximately a N(0,1) distribution.}$$

Remark: The CLT can be illustrated with a statistical software as follows: we ask the computer to generate a large number k of samples of size n each. Then we calculate the respective sample means $\bar{x}_1, \ldots, \bar{x}_k$ for the k samples. If n and k are large enough, then the density histogram of $\bar{x}_1, \ldots, \bar{x}_k$ should resemble the density of the normal distribution with mean μ and variance $\sigma/\sqrt{n}$. See the R instructions (Part 2), the item entitled "Illustration of the CLT."

Example 3. A botanist has planted tomato seedlings and is measuring their growth (in cm) after 30 days. Let X be the growth of a randomly chosen seedling. Assume that X is a random variable with mean $\mu = 10$ mm and standard deviation σ=3.5 mm. The botanist is selecting a random sample of 64 seedlings. Let $\bar{X}$ be the average of this sample. We would like to give an approximation for the probability that $\bar{X}$ is larger than 11.2mm.

Using the central limit theorem, $\bar{X}$ has approximately a normal distribution with the following mean and standard deviation:

$$\begin{aligned} \mu_{\bar{X}} &= \mu = 10 \\ \sigma_{\bar{X}} &= \sigma/\sqrt{n} = 3.5/\sqrt{64} = 0.4375 \end{aligned}$$

Note that the mean of $\bar{X}$ is the same as the mean of X, but the standard deviation of $\bar{X}$ is *different* than the standard deviation of X.

Therefore,

$$\frac{\bar{X} - 10}{0.4375} \quad \text{has approximately a N(0,1) distribution.}$$

We obtain:

$$P(\bar{X} > 11.2) = P\left(\frac{\bar{X} - 10}{0.05} > \frac{11.2 - 10}{0.4375}\right) \approx P(Z > 2.74)$$

$$= 1 - P(Z < 2.74) = 1 - 0.9969 = 0.0031$$

Remark: If we know that the original measurement X is normally distributed with mean μ and variance σ^2 (as it was the case in Example 2), then $\bar{X}$ **has a normal distribution** with mean $\mu_{\bar{X}} = \mu$ and standard deviation $\sigma_{\bar{X}} = \sigma/\sqrt{n}$. There is no need for a large sample size n. (This is Theorem 9.3 in the book.)