Chapter 6: Distributions of Sample Statistics

Department of Mathematics
Izmir University of Economics

Week 8-9
2014-2015

Introduction

In this chapter we will focus on

- Chebyshev's Theorem and the empirical rule,
- the distributions of *the sample means*, *the sample proportions*, and *the sample variances*,
- how to use sample distributions to find probabilities,
- an important result in statistics, *The Central Limit Theorem*, which has important implications in applications of statistics.

Chebyshev's Theorem and Empirical Rule

A Russian mathematician, Pafnuty Lvovich Chebyshev, established data intervals for any data set, *regardless* of the shape of the distribution.

Chebyshev's Theorem

For any population with mean μ, standard deviation σ, and $k > 1$, the percent of observations that lie within the interval $[\mu \pm k\sigma]$ is

$$\textbf{at least } 100\left(1 - \frac{1}{k^2}\right)\%,$$

where k is the number of standard deviations.

For example, on an exam, if $\mu = 72$, $\sigma = 4$, and it is assumed that $k = 2$, then the scores in the interval $[72 \pm 2\,(4)] = [72 \pm 8] = [64, 80]$ is **at least** $100\left(1 - \frac{1}{2^2}\right) = 100\left(\frac{3}{4}\right) = 75\%$ **of the total**.

Chebyshev's Theorem and Empirical Rule

Empirical Rule

For many large populations (**bell-shaped**) the empirical provides an estimate of the approximate percentage of observations that are contained within one, two, or three standard deviations of the mean:

- **Approximately** 68% **of the observations** are in the interval $\mu \pm 1\sigma$.
- **Approximately** 95% **of the observations** are in the interval $\mu \pm 2\sigma$.
- **Almost all of the observations** are in the interval $\mu \pm 3\sigma$.

Chebyshev's Theorem and Empirical Rule

Example. A company produces lightbulbs with a mean lifetime of 1200 hours and a standard deviation of 50 hours.

a) Describe the distribution of lifetimes if the shape of the distribution is unknown.

b) Describe the distribution of lifetimes if the shape of the distribution is known to be bell-shaped.

Solution.
Using $\mu = 1200$ and $\sigma = 50$ we have the following intervals:

$$[\mu \pm 1\sigma] = [1500 \pm 1\,(50)] = [1150, 1250]$$
$$[\mu \pm 2\sigma] = [1500 \pm 2\,(50)] = [1100, 1300]$$
$$[\mu \pm 3\sigma] = [1500 \pm 3\,(50)] = [1050, 1350]$$

Chebyshev's Theorem and Empirical Rule

a) If the shape of the distribution is unknown, we can apply Chebyshev's theorem (being aware that $k > 1$). Therefore, we cannot make any conclusions about the percentage of bulbs that last between 1150 and 1250 hours. We can conclude that **at least** $100\left(1-\frac{1}{2^2}\right) = 75\%$ of the bulbs will last between 1200 and 1300 hours and that **at least** $100\left(1-\frac{1}{3^2}\right) = 88.89\%$ of the bulbs will last between 1050 and 1350 hours.

b) If the shape of the distribution is **bell-shaped**, then we can conclude that **approximately** 68% of the bulbs will last between 1150 and 1250 hours, that **approximately** 95% of the bulbs will last between 1200 and 1300 hours, and that **almost all** the bulbs will last between 1050 and 1350 hours.

Chebyshev's Theorem and Empirical Rule

Example. A random sample of data has a mean of 75 and a variance of 25.

a) Use Chebyshev's theorem to determine the percent of observations between 65 and 85.

b) If the data are mounded, use the empirical rule to find the approximate percent of observations between 65 and 85.

Chebyshev's Theorem and Empirical Rule

Example. An auditor finds that the values of a corporation's account receivable have a mean of \$1645 and a standard deviation of \$92.

a) It can be guaranteed that 96% of the values will be in which interval?

b) It can be guaranteed that 99% of the values will be in which interval?

Chebyshev's Theorem and Empirical Rule

Example. It is known that the final grades in a math course have a mean of 60 and a variance of 100. The instructor has stated that he will give A to students in the top 16%.

a) A student in this course wants to guarantee an A in the final. What grade will be enough to guarantee an A in the final?

b) What can you say about the proportion of the grades that are greater than 40?

Sampling from a Population

A *simple random sample* of n subjects is chosen in such a way that

- each member of the population has the same chance to be selected,
- the selection of one member is independent of the selection of any other member, and
- every possible sample of size n has the same probability of selection.

Sampling Distribution of Sample Means

Definition:

Let the random variables $X_1, X_2, \ldots, X_n$ denote a random sample from a population. The **sample mean** of these random variables is defined as:

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

- The sample mean, $\bar{X}$, itself is a random variable, which means that it has a distribution.
- In many situations this distribution can be assumed to be normal.

Some Important Properties of the Distribution of the Sample Means

1. The sampling distribution of $\bar{X}$
   - has mean $E\left[\bar{X}\right] = \mu$
   - has standard deviation $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$ (*standard error* of $\bar{X}$)
2. If the sample size, n, is not small compared to the population size, N, then the standard error of $\bar{X}$ is

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}}$$

3. If the parent population distribution is *normal*, then the r.v.

$$Z = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}}$$

has a *standard normal distribution* with a mean of 0 and a variance of 1.

Example. It is believed that the monthly salary of executive officers in a city is normally distributed with a mean of \$6000 and a standard deviation of \$600. A random sample of nine observations is obtained from this population, and the sample mean is computed. What is the probability that the sample mean will be greater than \$6420?

Central Limit Theorem

Let $X_1, X_2, \ldots, X_n$ be a set of n independent random variables having identical distributions with mean μ, variance σ^2, and $\bar{X}$ as the mean of these random variables. As n becomes large, the *Central Limit Theorem* states that the distribution of

$$Z = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}}$$

approaches the standard normal distribution.

Example. Given a population with a mean of $\mu = 400$ and a variance of $\sigma^2 = 1,600$, the central limit theorem applies when the sample size is $n \geq 25$. A random sample of size $n = 35$ is obtained.

a) What are the mean and variance of the sampling distribution for the sample mean?

b) What is the probability that $\bar{x} > 412$?

c) What is the probability that $393 \leq \bar{x} \leq 407$?

d) What is the probability that $\bar{x} \leq 389$?

Sampling Distribution of Sample Proportions

Sample Proportion

Let X denote the number of successes in a binomial sample of n observations with probability of success p. The **sample proportion** is defined as:

$$\hat{p} = \frac{X}{n}$$

- Note that the random variable X is the sum of n independent Bernoulli random variables, each with probability of success p. As a result, $\hat{p}$ is the mean of a set of independent random variables, and the results for sample means apply.

Some Important Properties of the Distribution of the Sample Proportions

1. The sampling distribution of $\hat{p}$ has mean

$$E[\hat{p}] = p$$

2. The sampling distribution of $\hat{p}$ has standard deviation

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

3. If the sample size, n, is large, then the random variable

$$Z = \frac{\hat{p} - p}{\sigma_{\hat{p}}}$$

is approximately distributed as a standard normal distribution. This approximation is appropriate if $np(1-p) > 5$.

Example. In 1992, Canadians voted in a referendum on a new constitution. In the province of Quebec, 42.4% of those who voted were in favor of the new constitution. A random sample of 100 voters from the province was taken.

a) What is the mean of the distribution of the sample proportion in favor of a new constitution?

b) What is the variance of the sample proportion?

c) What is the standard error of the sample proportion?

d) What is the probability that the sample proportion is more than 0.5?

Example. A charity has found that 42% of all donors from last year will donate again this year. A random sample of 300 donors from last year was taken.

a) What is the standard error of the sample proportion who will donate again this year?

b) What is the probability that more than half of these sample members will donate again this year?

c) What is the probability that the sample proportion is between 0.40 and 0.45?

Sampling Distribution of Sample Variances

Definition

Let $x_1, x_2, \ldots, x_n$ be a random sample of observations from a population with variance σ^2. The **sample variance** is defined as

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

and its square root, s, is called the *sample standard deviation*.

Some Important Properties of the Distribution of the Sample Variances

1. The sampling distribution of s^2 has mean

$$E\left[s^2\right] = \sigma^2$$

2. The variance of the sampling distribution of s^2 depends on the underlying population distribution. If that distribution is normal, then

$$Var(s^2) = \frac{2\sigma^4}{n-1}$$

3. If the population distribution is normal, then $\frac{(n-1)s^2}{\sigma^2}$ has a **chi-square distribution** with $n-1$ degrees of freedom:

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

4. The density function of the chi-square distribution is *asymmetric* with a long positive tail.

Example. It is believed that first-year salaries for newly qualified accountants follow a normal distribution with a standard deviation of \$2500. A random sample of 16 observations was taken.

a) Find the probability that the sample standard deviation is more than \$3000.

b) Find the probability that the sample standard deviation is less than \$1500.

Example. In a large city it was found that summer electricity bills for single-family homes followed a normal distribution with a standard deviation of \$100. A random sample of 25 bills was taken.

a) Find the probability that the sample standard deviation is less than \$75.

b) Find the probability that the sample standard deviation is more than \$133.

Example. Assume that the standard deviation of monthly rents paid by students in a particular town is \$40. A random sample of 100 students was taken to estimate the mean monthly rent paid by the whole student population.

a. What is the standard error of the sample mean monthly rent?

b. What is the probability that the sample mean exceeds the population mean by more than \$5?

c. What is the probability that the sample mean is more than \$4 below the population mean?

d. What is the probability that the sample mean differs from the population mean by more than \$3?

Example. A record store owner finds that 20% of the customers entering her store make a purchase. One morning 180 people, who can be regarded as a random sample of all customers, enter the store.

a. What is the mean of the distribution of the sample proportion of customers making a purchase?

b. What is the variance of the sample proportion?

c. What is the standard error of the sample proportion?

d. What is the probability that the sample proportion is less than 0.15?

Example. A random sample of size $n = 16$ is obtained from a normally distributed population with a population mean of $\mu = 100$ and a variance of $\sigma^2 = 25$.

a. What is the probability that $\bar{x} > 101$?

b. What is the probability that the sample variance is greater than 45?

Example. The downtime per day for a certain computing facility averages 4 hours with a standard deviation of 0.8 hours. Find the probability that the total downtime for the 31 days is less than 115 hours.