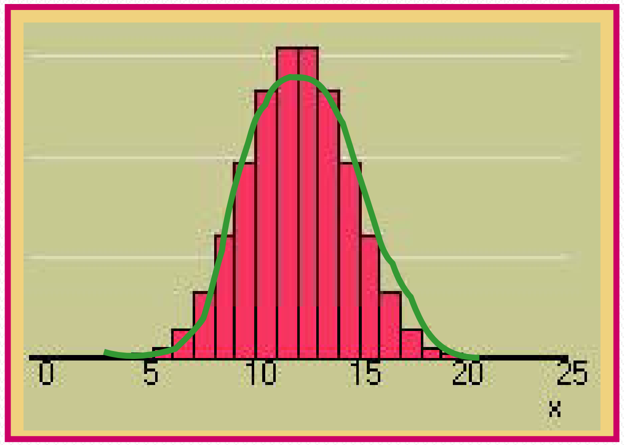
BUS 172
Descriptive Statistics

Lecture 16

The Normal Approximation to the Binomial

- We can calculate binomial probabilities using
  - The binomial formula
  - The cumulative binomial tables
- When n is large, and p is not too close to zero or one, areas under the normal curve with mean np and variance npq can be used to approximate binomial probabilities.



- **Normal approximation to a binomial probability distribution with n= 100 and p =0.10 showing the probability of 12 errors**

- With a continuous probability distribution, probabilities are computed as areas under the probability density function. As a result, the probability of any single value for the random variable is zero.
- Thus to approximate the binomial probability of 12 successes, we compute the area under the corresponding normal curve between 11.5 and 12.5. The 0.5 that we add and subtract from 12 is called a **continuity correction factor.** It is introduced because a continuous distribution is being used to approximate a discrete distribution.

Example

- Make sure that ***np* and *nq* are both greater than 5** to avoid inaccurate approximations!
- Standardize the values of *x* using $z = \dfrac{x - np}{\sqrt{npq}}$

$$P(11.5 \le x \le 12.5) \approx P(\frac{11.5 - 10}{3} \le z \le \frac{12.5 - 10}{3})$$

$$\approx P(0.5 \le z \le 0.83) = 0.7967 - 0.6915 = 0.1052$$

Example

Suppose *x* is a binomial random variable with *n* = 30 and *p* = .4. Using the normal approximation to find P($x \le 10$).

***n* = 30 *p* = .4 *q* = .6**

***np* = 12 *nq* = 18**

The normal approximation is ok!

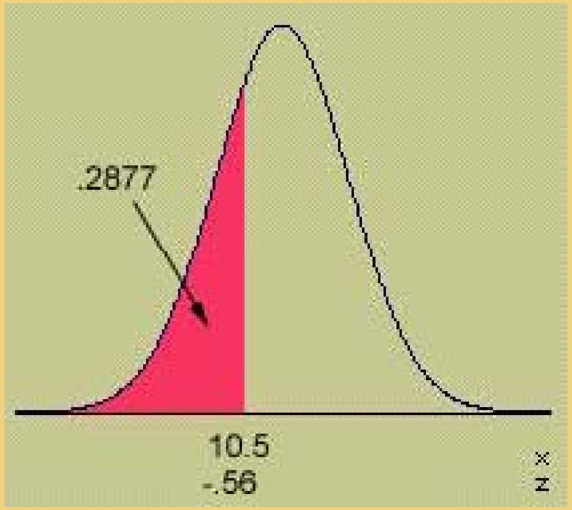
Calculate

$$\mu = np = 30(.4) = 12$$

$$\sigma = \sqrt{npq} = \sqrt{30(.4)(.6)} = 2.683$$

$$P(x \le 10) \approx P(z \le \frac{10.5 - 12}{2.683})$$

$$= P(z \le -.56) = .2877$$



Example

A production line produces AA batteries with a reliability rate of 95%. A sample of *n* = 200 batteries is selected. Find the probability that at least 195 of the batteries work.

Success = working battery ***n*** **= 200**

p **= .95** ***np*** **= 190** ***nq*** **= 10**

The normal approximation is ok!

$$P(x \geq 195) \approx P\left(z \geq \frac{194.5 - 190}{\sqrt{200(.95)(.05)}}\right)$$

$$= P(z \geq 1.46) = 1 - .9278 = .0722$$

Sampling Distributions

- Numerical descriptive measures calculated from the sample are called **statistics**. Statistics vary from sample to sample and hence are random variables.
- The probability distributions for statistics are called **sampling distributions**.
- To estimate the value of a population parameter, we compute a corresponding characteristic of the sample, referred to as a sample statistic.

Sampling Distributions

- For example, to estimate the population mean μ and the population standard deviation σ, we use the corresponding sample statistics the sample mean and sample standard deviation.
- Here the sample mean as the point estimator of the population mean μ.
- The sample standard deviation s as the point estimator of the population standard deviation σ,
- and the sample proportion as the point estimator of the population proportion p.

The Sampling Distribution of the Sample Mean

- ✓A random sample of size ***n*** is selected from a population with mean μ and standard deviation σ.
- ✓The sampling distribution of the sample mean $\overline{x}$ will have mean μ and standard deviation $\sigma / \sqrt{n}$ for an infinite population.
- ✓If the original population is **normal,** the sampling distribution will be normal for any sample size.
- ✓If the original population is **nonnormal,** the sampling distribution will be normal when *n* is large (central limit theorem).
- ✓the sample mean is a random variable and its probability distribution is called the sampling distribution

The standard deviation of *x*-bar is sometimes called the STANDARD ERROR (SE) of the mean. Standard error refers to the standard deviation of a point estimator.

STANDARD DEVIATION OF $\bar{x}$

Finite Population

$$\sigma_{\bar{x}} = \sqrt{\frac{N-n}{N-1}}\left(\frac{\sigma}{\sqrt{n}}\right)$$

Infinite Population

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

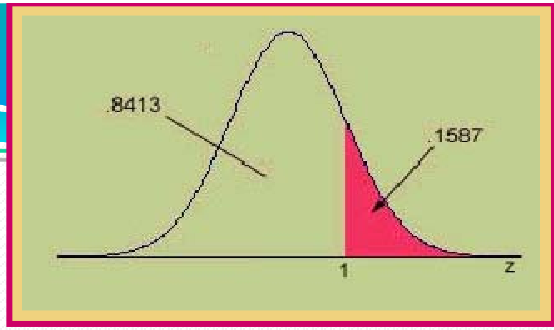
USE THE FOLLOWING EXPRESSION TO COMPUTE THE STANDARD DEVIATION OF $\bar{x}$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \qquad \textbf{(7.3)}$$

whenever

1. The population is infinite; or
2. The population is finite *and* the sample size is less than or equal to 5% of the population size; that is, $n/N \leq .05$.

Finding Probabilities for the Sample Mean



✓If the sampling distribution of $\bar{x}$ is normal or approximately normal, *standardize or rescale* the interval of interest in terms of

$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

✓Find the appropriate area using Table 3.

Example: A random sample of size $n = 16$ from a normal distribution with $\mu = 10$ and $\sigma = 8$.

$$P(\bar{x} > 12) = P(z > \frac{12 - 10}{8 / \sqrt{16}})$$

$$= P(z > 1) = 1 - .8413 = .1587$$

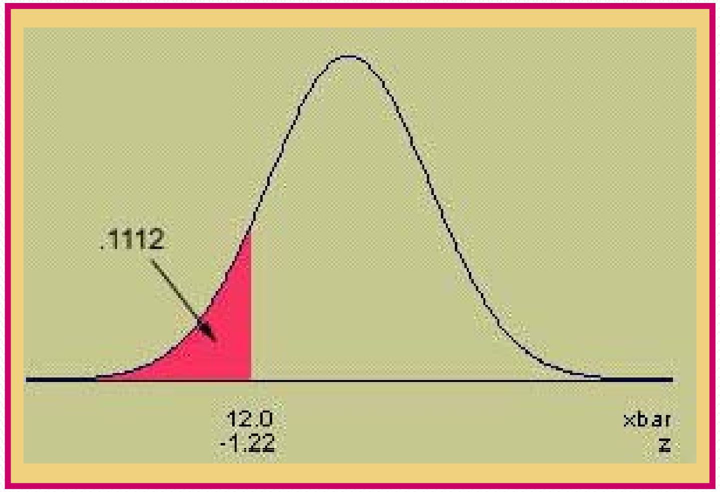
Example

A soda filling machine is supposed to fill cans of soda with 12 fluid ounces. Suppose that the fills are actually normally distributed with a mean of 12.1 oz and a standard deviation of .2 oz. What is the probability that the average fill for a 6-pack of soda is less than 12 oz?

$$P(\bar{x} < 12) =$$

$$P\left(\frac{\bar{x} - \mu}{\sigma / \sqrt{n}} < \frac{12 - 12.1}{.2 / \sqrt{6}}\right) =$$

$$P(z < -1.22) = .1112$$



Example

What is the probability that the sample mean of salaries computed using a simple random sample of 30 managers will be within \$500 of the population mean \$51,800 with population standard deviation of \$4,000.

$$P(51300 \le \bar{x} \le 52300) = P\left(\frac{51300 - 51800}{\frac{4000}{\sqrt{30}}} \le z \le \frac{52300 - 51800}{\frac{4000}{\sqrt{30}}}\right)$$

$$= P(-0.68 \le z \le 0.68) = P(z \le 0.68) - P(z \le -0.68)$$

$$= 0.7517 - 0.2483 = 0.5034$$

The Sampling Distribution of the Sample Proportion

✓The sample proportion is the point estimator of the population proportion P.

✓The sample proportion, $\hat{p} = \frac{x}{n}$ is simply a *rescaling* of the binomial random variable *x*, dividing it by *n*.

x = the number of elements in the sample that possess the characteristic of interest

n = sample size

✓From the Central Limit Theorem, the sampling distribution of will also be **approximately normal,** with a *rescaled* mean and standard deviation.

The Sampling Distribution of the Sample Proportion

✓A random sample of size ***n*** is selected from a binomial population with parameter ***p***.

✓The sampling distribution of the sample proportion,

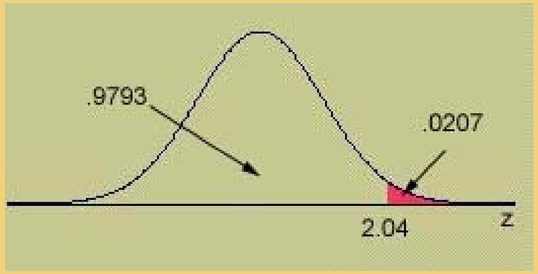
$$\hat{p} = \frac{x}{n}$$

✓will have mean ***p*** and standard deviation $\sqrt{\frac{pq}{n}}$

✓If n is large, and p is not too close to zero or one, the sampling distribution of $\hat{p}$ will be **approximately normal.**

The standard deviation of *p*-hat is sometimes called the STANDARD ERROR (SE) of *p*-hat.

Finding Probabilities for the Sample Proportion



- ✓If the sampling distribution of $\hat{p}$ is normal or approximately normal, *standardize or rescale* the interval of interest in terms of $$z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}}$$
- ✓Find the appropriate area using Table 3.

Example: A random sample of size $n = 100$ from a binomial population with $p = .4$.

$$P(\hat{p} > .5) = P\left(z > \frac{.5 - .4}{\sqrt{\frac{.4(.6)}{100}}}\right)$$

$$= P(z > 2.04) = 1 - .9793 = .0207$$

Example

The soda bottler in the previous example claims that only 5% of the soda cans are underfilled. A quality control technician randomly samples 200 cans of soda. What is the probability that more than 10% of the cans are underfilled?

***n* = 200**

S: underfilled can

***p* = P(S) = .05**

***q* = .95**

***np* = 10 *nq* = 190**

OK to use the normal approximation

$$P(\hat{p} > .10)$$

$$= P\left(z > \frac{.10 - .05}{\sqrt{\frac{.05(.95)}{200}}}\right) = P(z > 3.24)$$

$$= 1 - .9994 = .0006$$

This would be very unusual, if indeed *p* = .05!