Chapter 5: Inference

Read: Chapter 5 of the text, omit binomial formulas, then read the notes and try the WEBCT assignment questions. If you need more practice, try the practice questions with answers available on the web.

Introduction

Statistical inference is the process of using our sample to draw conclusions about the population of interest. Some characteristic of the population is of primary importance to the researcher.

In this chapter you will see how this process of inference works using two examples; the sampling distribution for a proportion and the sampling distribution for the mean. The language of probability is used to help us describe what happens to the sample proportion and the sample mean when we repeat the process of sampling from the population many times. Remember we use repeated sampling to help us understand the process of

inference, in real life we usually only have one sample.

Counts and Proportions

Consider an opinion poll. The poll size is n = 1785 people. Each person is asked, "Have you driven a car in the last week"?

Record the number that say yes and call this variable X.

X is a count in a fixed number of trials, where the number of trials is n.The sample proportion is:

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

Example: 1000 people say "yes" to the above question, then:

$$\hat{p} = \frac{1000}{1785} = .56$$

Below is notation that will help us understand how the sample proportion can be used to make inferences about the population proportion.

Suppose we have a population π and a categorical variable X defined on π taking 2 distinct values.

For convenience we let these values be 0 and 1. Sometimes researchers label 0=failure and 1=success (X=0 or X=1).
Suppose we **randomly** select a sample. Then:

P(X=1) = proportion of individuals having X=1 (success)
= p

P(X=0) = 1 - P(X=1)
= 1 - p (failure)

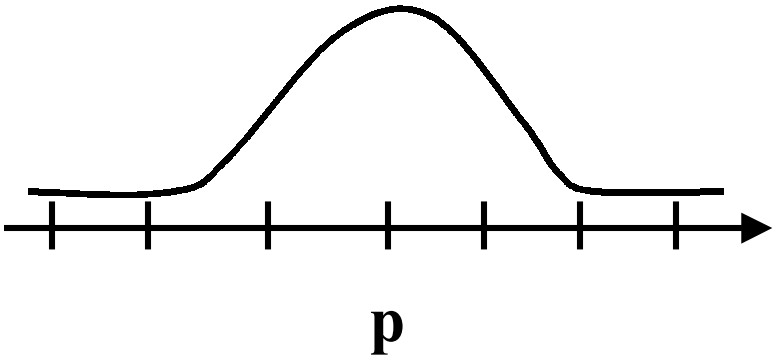
Without carrying out a census we will not know "p". Hence, we generate a random sample from the population. If the population is very large relative to the sample, we treat the values as independent. We then estimate "p" in the population by the sample proportion which using the above coding(0,1) is the mean.

$$\frac{1}{n} \sum x_i = \bar{x}$$

Question: How accurate is $\bar{x}$ an estimate of p?

Remember: We could substitute $\hat{p}$ for $\bar{x}$ in the above since $\hat{p}$ and $\bar{x}$ in this context are identical.

Answer:
We could run simulations for various values of p for fixed n and see how well the sampling distribution of $\bar{x}$ concentrates about p.



SPSS EXAMPLE ONE

In this SPSS example one, we know that p = .6. Consider a sample size of 100 (n) and repeatedly sample 1000 (N) times. The mean of the sampling

distribution is very close to p = .6 as you will see when you view the example. Note the range of values, is the range smaller than the example below?

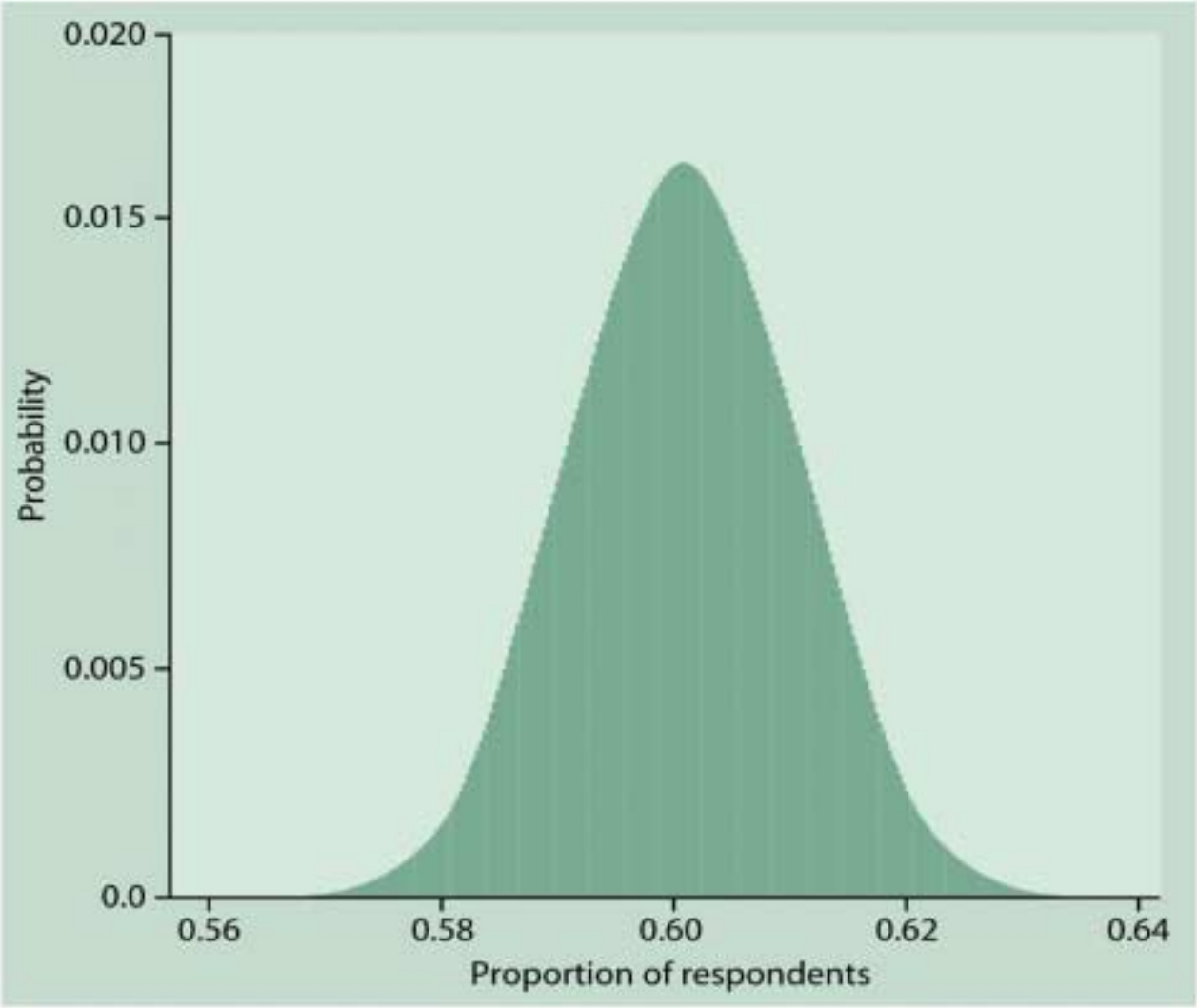


Figure 5.2 Probability histogram of the sample proportion p(hat) based on the binomial count with *n* = 2500 and p = 0.6. The distribution is very close to normal. The range of p(hat) is from about .57 to.63.

Fact: Let $\hat{p}$ be the sample proportion of successes in a SRS of size "n". If the population is large then the **sampling distribution** of $\hat{p}$ is normal with a mean equal to "p" and standard deviation equal to σ below:

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

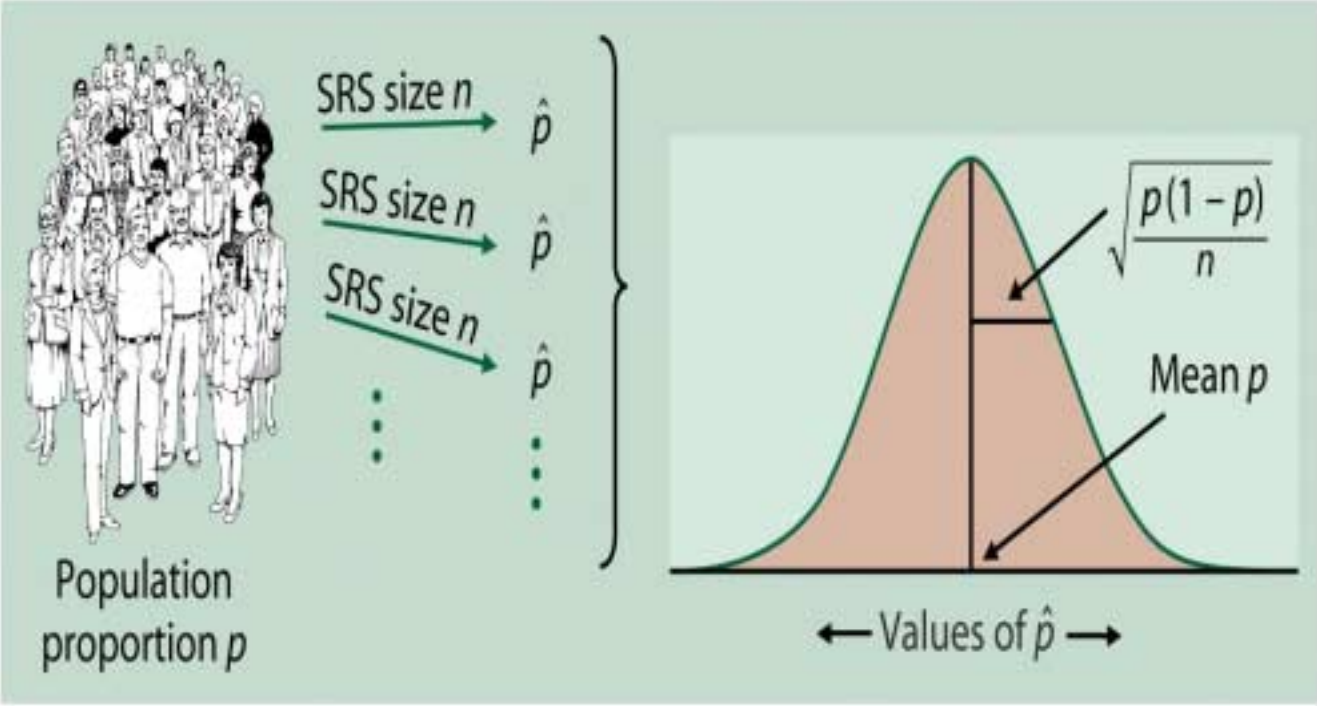


Figure 5.3 The sampling distribution of a sample proportion p(hat) is approximately normal with mean p and standard deviation √p(1-p)/n.

However, if x is a count, it is normally distributed with mean=np and corresponding standard deviation

$$\sigma = \sqrt{np\,(1 - p)}$$

Example: (Proportions)

A psychologist samples 50 people from the population of coffee drinkers at the university in order to see what proportion prefer instant coffee.

$$n = 50 \text{ people}$$

She finds the preference test results indicate that 19 of the 50 people sampled prefer instant coffee.

$$\hat{p} = \frac{19}{50} = .38$$

Suppose an executive at the coffee company believes the proportion of coffee drinkers that prefer instant is .5. If we assume the population proportion is .5 then:

$$p = .5$$

Subsequently,

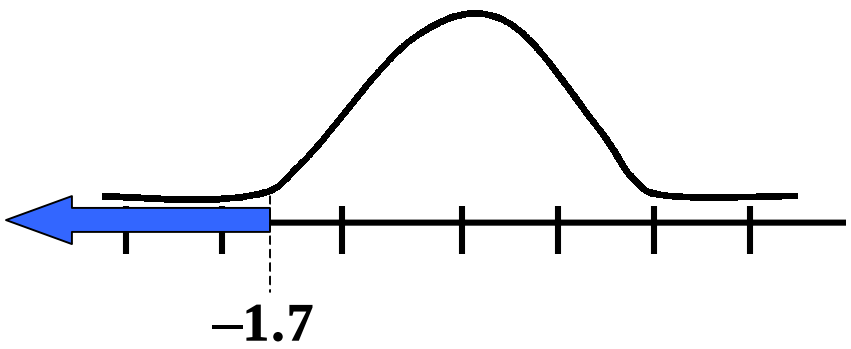
$$\sigma = \sqrt{\frac{p(1-p)}{50}} = .07$$

We now have the mean and standard deviation of the sampling distribution for sample sizes of 50. What is the probability of obtaining a sample value of .38 or something smaller assuming p=.5?

In other words, what are the chances of obtaining a sample proportion of .38 or something smaller if the executive is correct. Construct the standardized score (Z) from the information above, which will allow us to calculate the above probability.

$$Z = \frac{\bar{x} - \mu}{\sigma} = \frac{\hat{p} - p}{\sigma} = \frac{.38 - .5}{.07}$$

Therefore, $Z = -1.7$



P (Z < –1.7) is .045

(from our Z-tables)

If the population proportion is .5, then the probability of finding a sample of 50 with proportion = .38 or smaller is low (.045). In other words, we have evidence against the assumption that p=.5 (50% of coffee drinkers prefer instant).

You can recalculate this probability by changing the assumed value of p, i.e., one could assume a value of

.4. Note, however, the observed proportion would remain unchanged.

Note, we were able to infer the above since we knew what the sampling distribution of the proportion would look like if the executive was correct. Since the sample statistic had low probability under this assumption we conclude the executive was incorrect.

Quantitative Variables

Suppose you have a pop. π and a quantitative variable $\mathbf{X}$ defined on π. The distribution of $\mathbf{X}$ over π has a mean μ_X and standard deviation σ_X. We can't know μ_X, σ_X without carrying out a census. Hence, we generate a sample. Again if the population is large relative to n we can assume the observations are independent. It is natural to estimate μ_X by the sample mean.

$$\bar{x} = \frac{1}{n} \sum x_i$$

and, estimate σ_X by:

$$s_x = \sqrt{\frac{1}{n-1}\sum (x_i - \bar{x})^2}$$

How much error is there in these estimates?
To assess this we need a further assumption (for the moment) namely assume that:

$$X \sim N(\mu, \sigma)$$

Note: How do we check whether or not this assumption of normality makes sense? As in Chapter 1 we construct a **NORMAL PROBABILITY PLOT**.

Then if this assumption holds approximately we can simulate samples of n from the N(μ,σ) distribution for various values of μ and σ and look at the distributions of $\bar{x}$ and s_x.

Fact: **When $X_1, X_2,\ldots,X_n$ is independent and distributed normally $N(\mu,\sigma)$ then the exact distribution of $\bar{x}$ and s^2_x is**

$$\overline{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

and

$$\frac{(n-1)\, s^2_x}{\sigma^2} \sim \text{chi-square } (n-1)$$

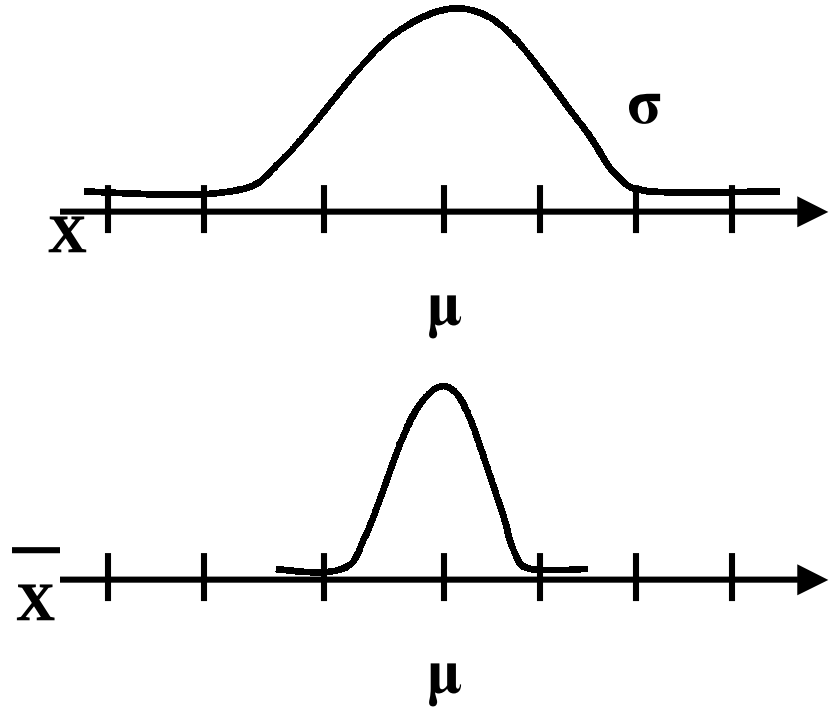
the variance is distributed as a chi-square distribution with n – 1 degrees of freedom. The **degrees of freedom** is a way statisticians correct for sample size. Small sample sizes have more variability than larger samples and the degrees of freedom is a correction for this. In general, the mean of the sampling distribution for $\bar{x}$ is the same as the mean of the population. Hence,

$$\mu_{\bar{x}} = \mu_x$$

The variance of the sampling distribution however is reduced.

$$\sigma^2_{\bar{x}} = \frac{\sigma^2_x}{n}$$

The first figure below gives the population distribution of X. The second figure gives the sampling distribution for the mean.



Question: Is the variance larger for the population or the sampling distribution? As you increase the size of the sample (the n of the sampling distribution) what do you think happens to the variance of the sampling distribution?

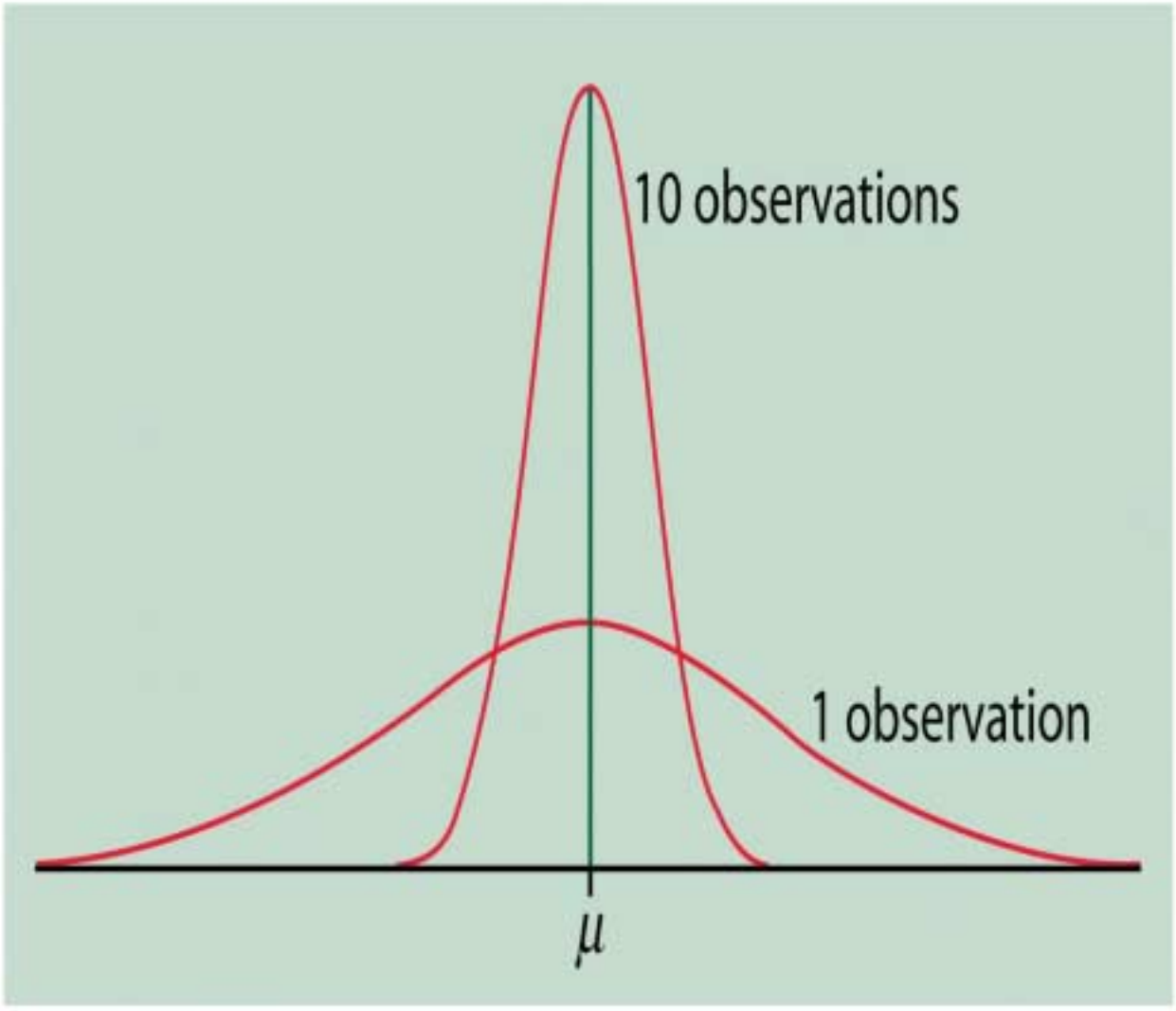


Figure 5.8 The sampling distribution of x(bar) for samples of size 10 compared with the distribution of a single observation.

The standard deviation of the sampling distribution is given below.

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

Note, the variance is reduced as the sample size increases. The mean of the sampling distribution is still centered at the population mean. What happens if the population is not normal? Does the sampling distribution for the mean remain normal? The answers are given in the following theorem.

View the Video: Sample Mean &Control Charts

The Central Limit Theorem (CLT)

When n is large $\bar{x}$ is approximately distributed as:

$$N\left(\mu_x, \frac{\sigma}{\sqrt{n}}\right)$$

no matter what the distribution of X is. In other words, if the population is normal the sampling distribution is normal. If the population is not normal the sampling distribution is still normal for large n (say > 50).

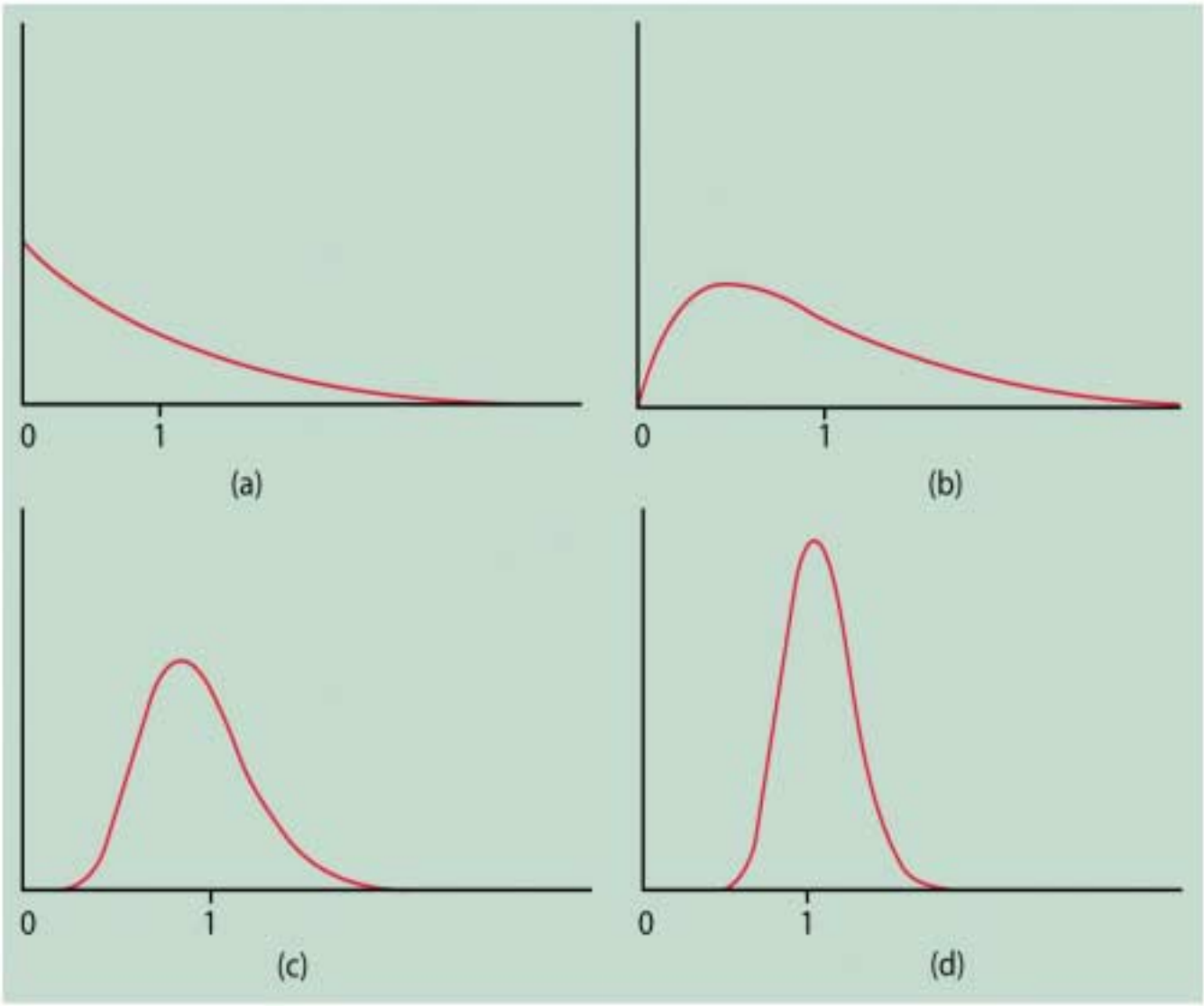


Figure 5.10 The central limit theorem in action: the distribution of sample means from a strongly nonnormal population becomes more normal as the sample size increases. (a) The distribution of 1 observation. (b) The distribution of x(bar) for two observations. (c) The distribution of x(bar) for 10 observations. (d) The distribution of x(bar) for 25 observations.

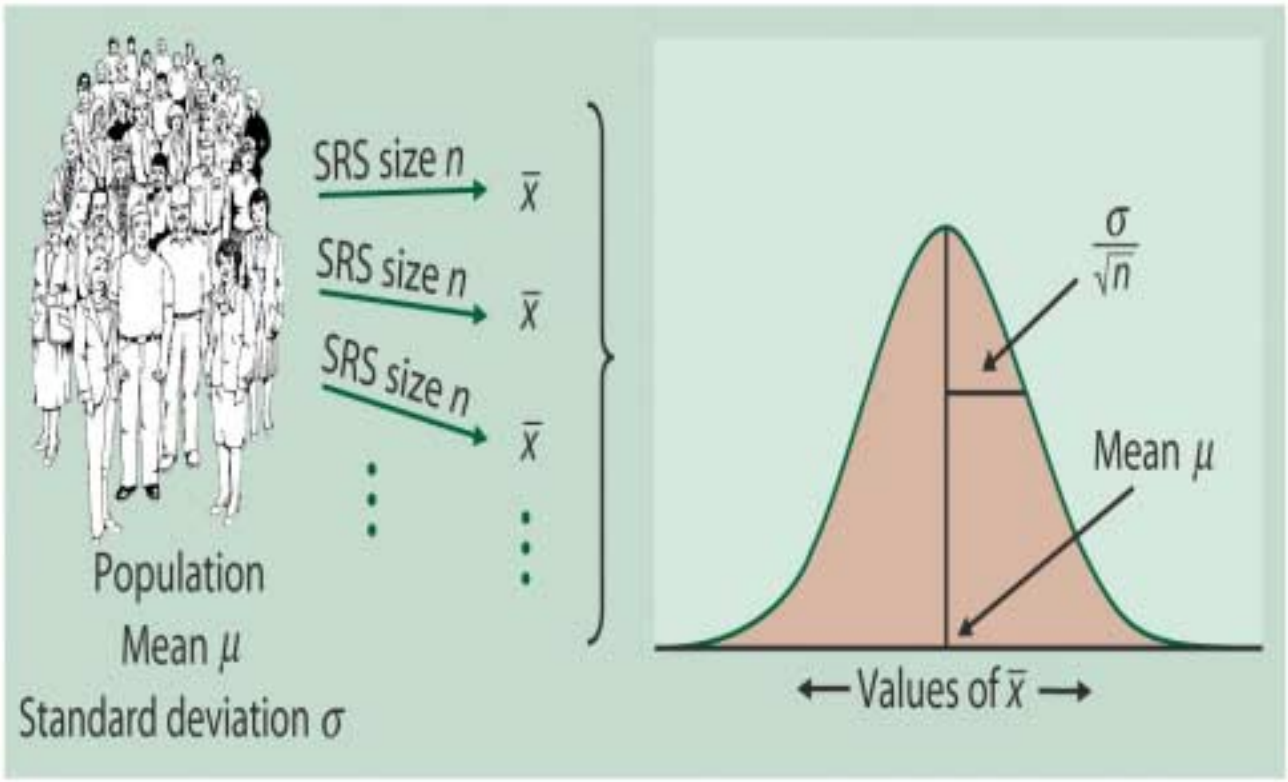
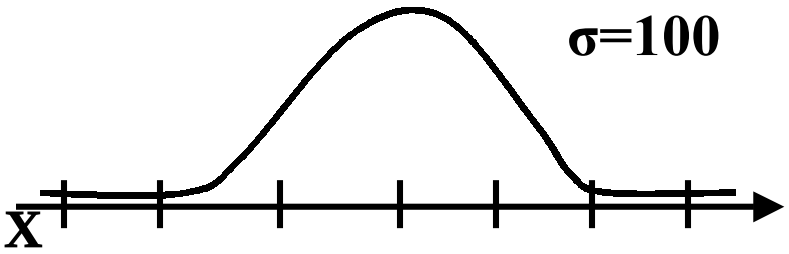


Figure 5.12 The sampling distribution of a sample mean x(bar) has mean μ and standard deviation σ/√n. The distribution is normal if the population distribution is normal; it is approximately normal for large samples in any case.

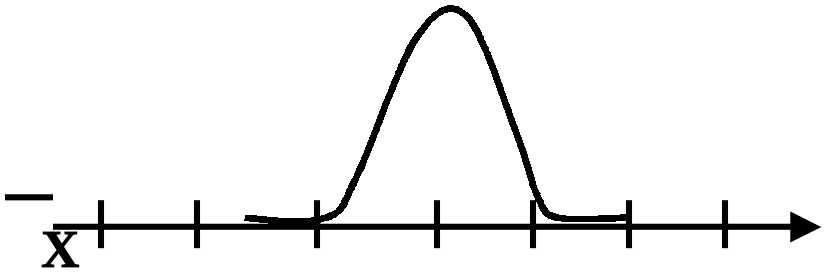
Example: The population of scores of the WISC IQ test has a normal distribution with μ = 500 and σ = 100. Take a random sample n = 25 students. What is the probability that the sample mean would be greater than 540?

The first figure gives the population and the second the sampling distribution for the mean.



$\mu = 500$

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



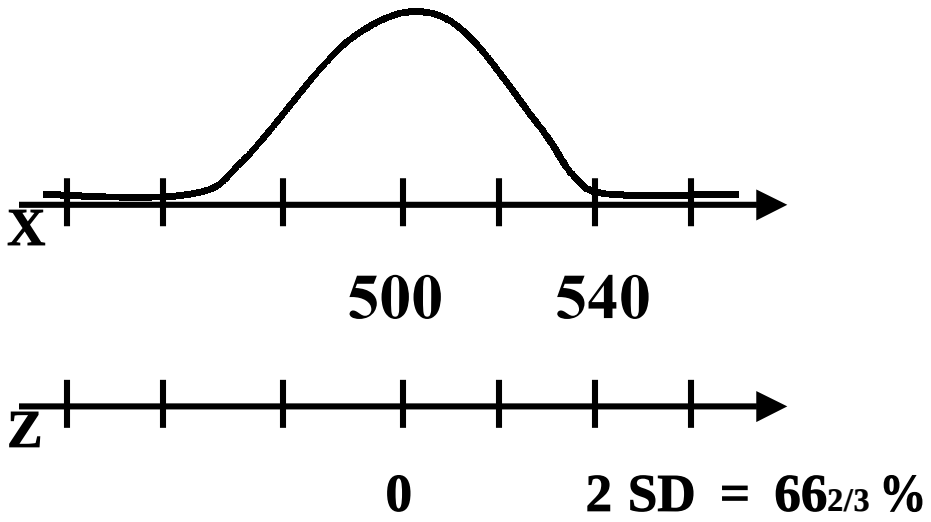
$\mu = 500$

First calculate the standard deviation of the sampling distribution.

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{100}{\sqrt{25}} = \frac{100}{5} = 20$$

Next we convert to a standard score (Z):

$$Z = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{540 - 500}{20} = \frac{40}{20} = 2$$



The mean of 540 is 2 standard deviations above the mean of 0. Using the Z tables in the text we have p = .0228 or 2.28% that the sample mean is greater than 540.

We sometimes use term standard error for the standard deviation of the sampling distribution.

The standard error is given below:

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

If n = 1 then the standard error = σ = standard deviation of the population. **Note:** In general the test statistic is constructed using the following formula:

$$\frac{\text{estimate} - \text{population value}}{\text{standard error of the estimate}}$$

The estimate is obtained from the data or sample. The population value is obtained from the literature, the researchers knowledge, a guess, etc.

SPSS EXAMPLE TWO

We use SPSS to simulate the sampling distribution for a normal population with mean 50 and standard deviation 15. The sample size is 25 and we draw 40 samples from the above population. Note the mean of the sampling distribution is very close to the mean of the population. This will be the case for any population according to the CLT.

Chapter 5 Summary

A **binary variable** is examined and the **sample proportion** is recorded. Formulas are given for the **mean** and **standard deviation of the sampling distribution** for the "sample proportion".

The sampling distribution for the proportion is normal. The sampling distribution for the **sample mean** is normally distributed as well. This is proven by the **central limit theorm** (CLT).

Remember: Tips for Success

1) Read the text.
2) Read the notes.
3) Try the assignment.
4) If needed, try the exercise questions.
5) Try the simulations and view the videos if you need more help with a concept.
6) Try the self tests for practice on each chapter of the text at www.whfreeman.com/ips
7) Steady Work = Success