Recall

Binomial Probability Distribution

1. The procedure must have a **fixed number of trials, *n*.**
2. The trials must be **independent.**
3. Each trial must have all outcomes classified into **two categories** (commonly, **success** and **failure**).
4. The probability of success *p* remains the same in all trials (the probability of failure is $q=1-p$)

Approximation of a Binomial Distribution with a Normal Distribution

If $np \geq 5$ and $nq \geq 5$

Then $\mu = np$ and $\sigma = \sqrt{npq}$

and the random variable has

a  **distribution.**

(normal)

Procedure for Using a Normal Distribution to Approximate a Binomial Distribution

1. Verify that both $np \geq 5$ and $nq \geq 5$. If not, you cannot use normal approximation to binomial.

2. Find the values of the parameters μ and σ by calculating $\mu = np$ and $\sigma = \sqrt{npq}$.

3. Identify the discrete whole number x that is relevant to the binomial probability problem. Use the continuity correction (See continuity corrections discussion later) Draw a normal curve and enter the values of μ, σ, and either $x - 0.5$ or $x + 0.5$, as appropriate.

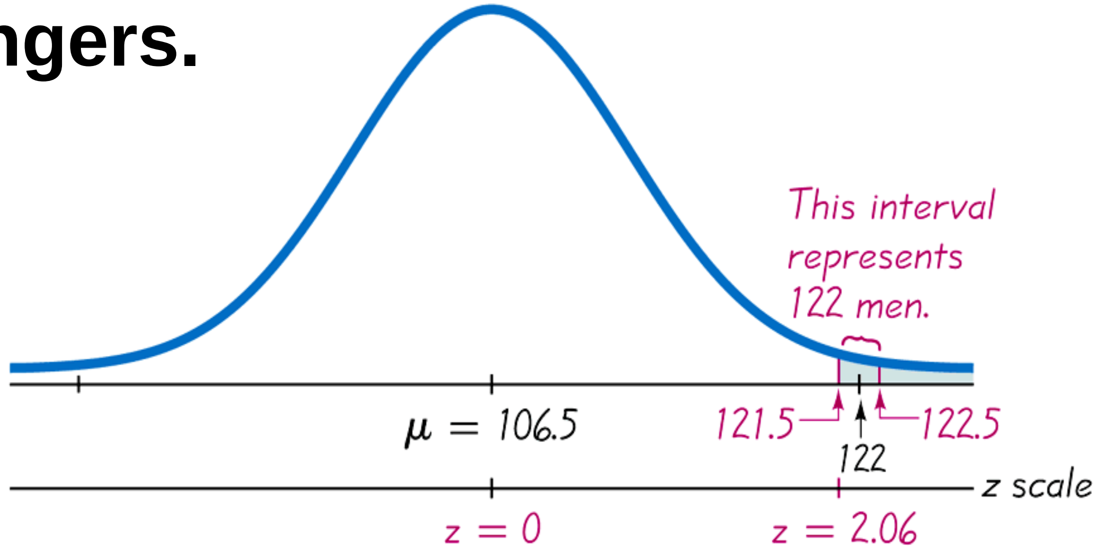
Procedure for Using a Normal Distribution to Approximate a Binomial Distribution

4. Change x by replacing it with x – 0.5 or x + 0.5, as appropriate.

5. Using x – 0.5 or x + 0.5 (as appropriate) in place of x, find the area corresponding to the desired probability by first finding the z score and finding the area to the left of the adjusted value of x.

Example – Number of Men Among Passengers

Finding the Probability of "At Least 122 Men" Among 213 Passengers.



$\mu = np = 213 * 0.5 = 106.5$

and σ = sqrt(npq) = *sqrt(213*0.5*0.5)=7.29*

Consider x-0.5 and x+0.5 for continuity correction

122 - 0.5=121.5 and 122 + 0.5 = 122.5

Z= (121.5-106.5)/7.29= 2.06

Definition

When we use the normal distribution (which is a continuous probability distribution) as an approximation to the binomial distribution (which is discrete), a continuity correction is made to a discrete whole number x in the binomial distribution by representing the single value x by the interval from

$$x - 0.5 \text{ to } x + 0.5$$

(that is, adding and subtracting 0.5).

Procedure for Continuity Corrections

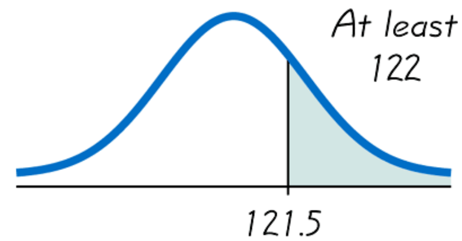
1. When using the normal distribution as an approximation to the binomial distribution, always use the continuity correction.

2. In using the continuity correction, first identify the discrete whole number x that is relevant to the binomial probability problem.

3. Draw a normal distribution centered about μ, then draw a vertical strip area centered over x . Mark the left side of the strip with the number x – 0.5, and mark the right side with x + 0.5. For x = 122, draw a strip from 121.5 to 122.5. Consider the area of the strip to represent the probability of discrete whole number x.

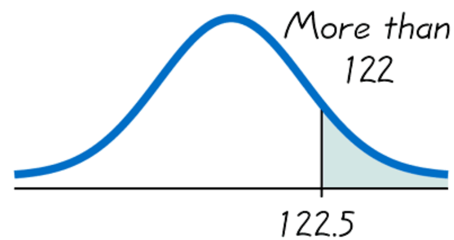
Procedure for Continuity Corrections

4. Now determine whether the value of *x* itself should be included in the probability you want. Next, determine whether you want the probability of at least *x*, at most *x*, more than *x*, fewer than *x*, or exactly *x*. Shade the area to the right or left of the strip, as appropriate; also shade the interior of the strip itself **if and only if *x* itself** is to be included. The total shaded region corresponds to the probability being sought.



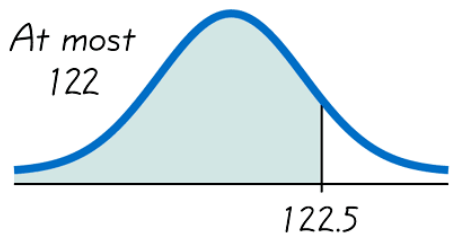
X = **<u>at least</u> 122**

(includes 122 and above)



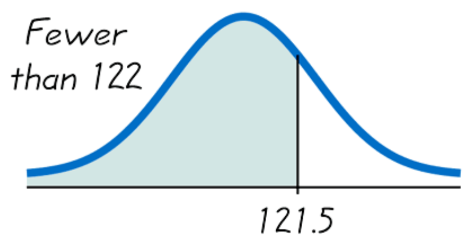
X = **<u>more than</u> 122**

(doesn’t include 122)



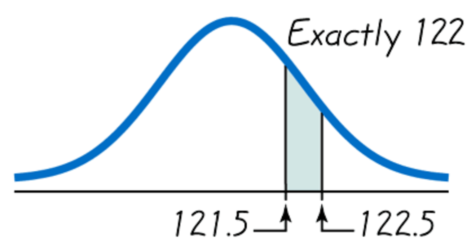
X = **<u>at most</u> 122**

(includes 122 and below)



X = **<u>fewer than</u> 122**

(doesn’t include 122)



X = **<u>exactly</u> 122**

Ex. Continuity correction

EXAMPLE 2

Internet Penetration Survey A recent Pew Research Center survey showed that among 2822 randomly selected adults, 2060 (or 73%) stated that they are Internet users. If the proportion of all adults using the Internet is actually 0.75, find the probability that a random sample of 2822 adults will result in *exactly* 2060 Internet users.

Step 1:

$$np = 2822 \cdot 0.75 = 2116.5 \qquad \text{(Therefore } np \geq 5.)$$

$$nq = 2822 \cdot 0.25 = 705.5 \qquad \text{(Therefore } nq \geq 5.)$$

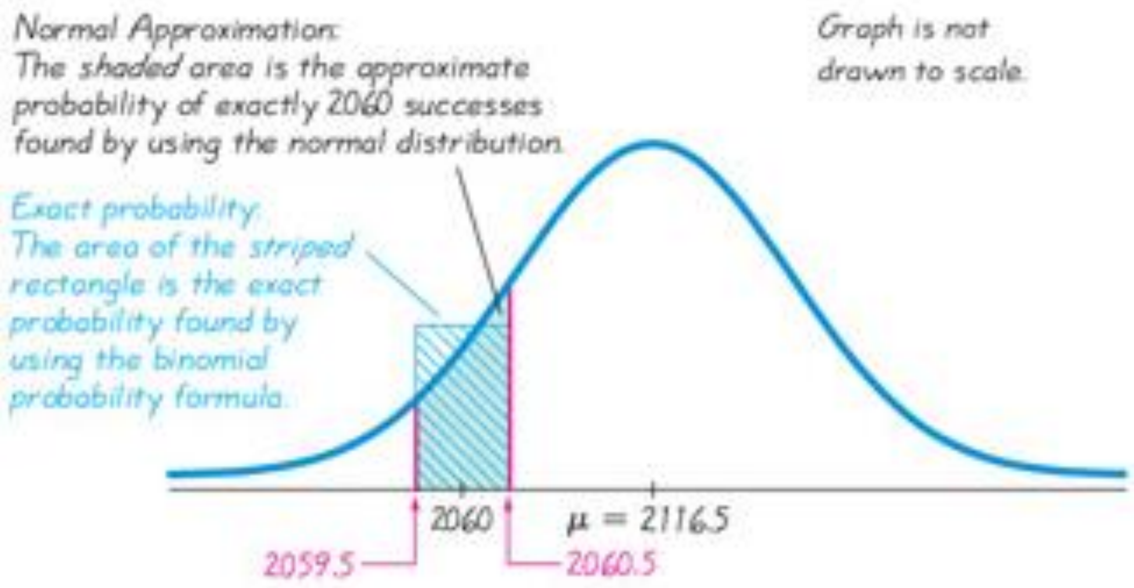
Step 2:

$$\mu = np = 2822 \cdot 0.75 = 2116.5$$

$$\sigma = \sqrt{npq} = \sqrt{2822 \cdot 0.75 \cdot 0.25} = 23.002717$$

Step 3: probability of *exactly* 2060

strip from 2059.5 to 2060.5.



$$z = \frac{2060.5 - 2116.5}{23.002717} = -2.43$$

$$z = \frac{2059.5 - 2116.5}{23.002717} = -2.48$$

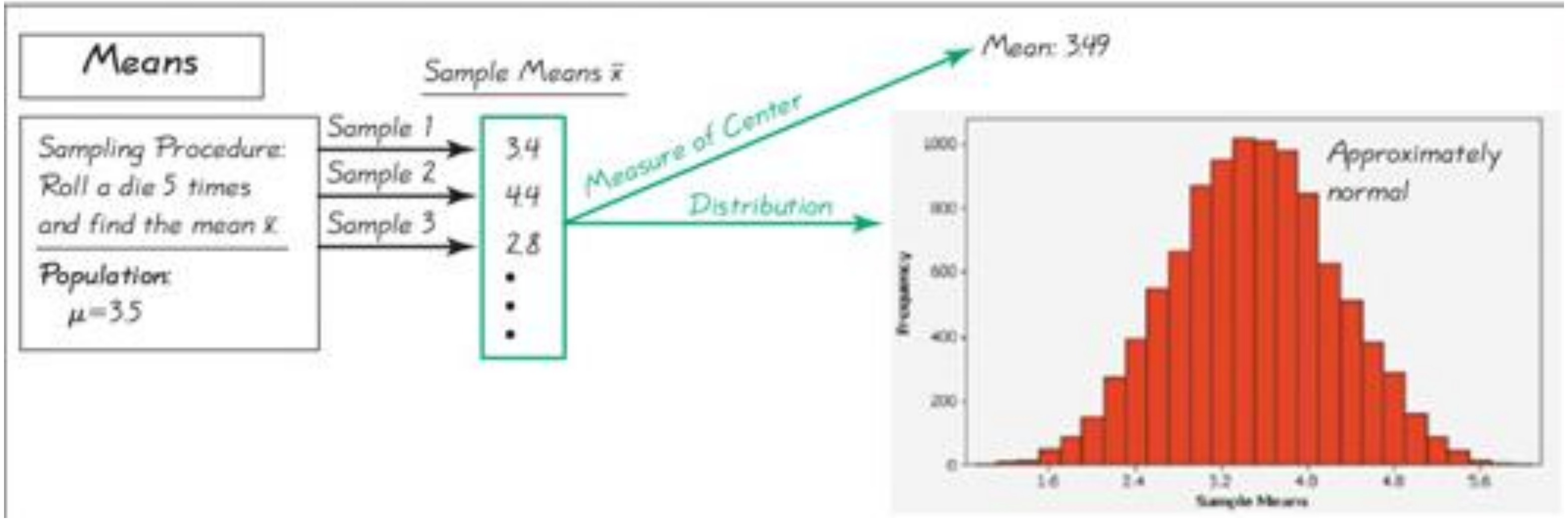
We use Table A-2 to find that $z = -2.48$ corresponds to a probability of 0.0066, which is the total area to the left of 2059.5. The shaded area is $0.0075 - 0.0066 = 0.0009$.

Recall: Sampling Distribution of the mean

- The sampling distribution of the mean is the distribution of the sample means, with all samples having the same sample size n taken from the same population

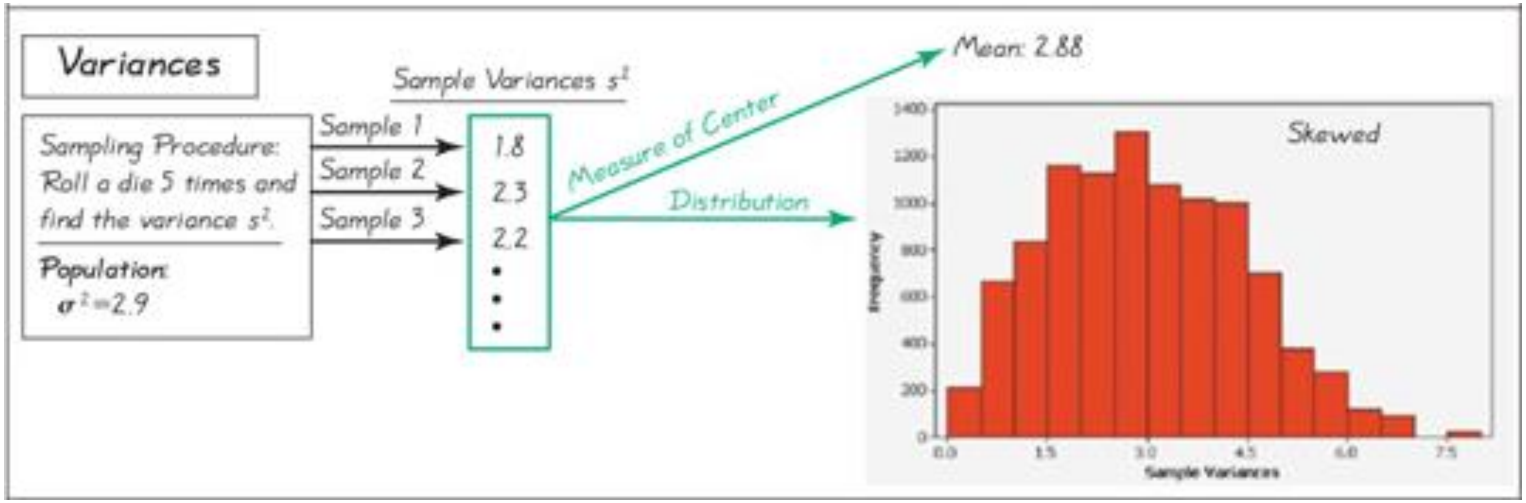
Recall: Example

- Consider rolling a die 5 times and find the mean $\bar{x}$ of the results.



Recall: Sampling distribution of the variance

- The sampling distribution of the variance is the distribution of the sample variances, with all samples having the same size n taken from the population



Estimators
Estimates and Sample Size

Some statistics work much better than others as estimators of the population. The example that follows shows this.

Review

- **Chapters 2 & 3 used "descriptive statistics" when we summarized data using tools such as graphs, and statistics such as the mean and standard deviation.**
- **Chapter 6 introduced <u>critical values</u>:**
  z_α denotes the z score with an area of α to its right.
  If $\alpha = 0.025$, the critical value is $z_{0.025} = 1.96$.
  That is, the critical value $z_{0.025} = 1.96$ has an area of 0.025 to its right.

Preview

This chapter presents the beginning of inferential statistics.

- **The two major activities of inferential statistics are (1) to use sample data to estimate values of a population parameters, and (2) to test hypotheses or claims made about population parameters.**
- **We introduce methods for estimating values of these important population parameters: proportions, means, and variances.**
- **We also present methods for determining sample sizes necessary to estimate those parameters.**

Section 7-2
Estimating a Population Proportion

Key Concept

In this section we present methods for using a sample proportion to estimate the value of a population proportion.

- **The sample proportion is the best point estimate of the population proportion.**
- **We can use a sample proportion to construct a confidence interval to estimate the true value of a population proportion, and we should know how to interpret such confidence intervals.**
- **We should know how to find the sample size necessary to estimate a population proportion.**

Definition

A point estimate is a single value (or point) used to approximate a population parameter.

Definition

The sample proportion $\hat{p}$ is the best point estimate of the population proportion p.

Example:

In a Research Center poll, 70% of 1501 randomly selected adults in the U. S. believe in global warming, so the sample proportion is = 0.70. Find the best point estimate of the proportion $\hat{p}$ of all adults in the United States who believe in global warming.

Because the sample proportion is the best point estimate of the population proportion, we conclude that the best point estimate of *p* is 0.70. When using the sample results to estimate the percentage of all adults in the U.S. who believe in global warming, the best estimate is 70%.

But how reliable (accurate) is this estimate?

We will see that its margin of error- coming after- is 2.3%. This means the true proportion of adults who believe in global warming is between 67.7% and 72.3%. This gives an interval (from 67.7% to 72.3%) containing the true (but unknown) value of the population proportion.

Example:

In a recent poll, 70% of 1501 randomly selected adults said they believed in global warming.

Q: What is the proportion of the adult population that believe in global warming?

Notation: *p* is the population proportion (an unknown parameter).

$\hat{p}$ **is the sample proportion (computed).**
From the poll data $\hat{p}$ **= 0.70.**

Apparently, 0.70 will be the best estimate of the proportion of all adults who believe in global warming.

Example

The sample proportion $\hat{p}$ = 0.70 is the best estimate of the population proportion *p*.

A 95% confidence interval for the unknown population parameter is

$$0.677 < p < 0.723$$

"will know later where these numbers come from"

What does it mean, exactly?

Interpreting a Confidence Interval

We are 95% confident that the interval from 0.677 to 0.723 actually does contain the true value of the population proportion *p*.

This means that if we were to select many different samples of size 1501 and construct the corresponding confidence intervals, then 95% of them would actually contain the value of the population proportion *p*.

Caution

Know the correct interpretation of a confidence interval.

It is wrong to say "the probability that the population parameter belongs to the confidence interval is 95%"

because the population parameter is not a random variable, it does not change its value.

Caution

Do not confuse two percentages: the proportion may be represented by percents (like 70% in the example), and the confidence level may be represented by percents (like 95% in the example).

Proportion may be any number from 0% to 100%.

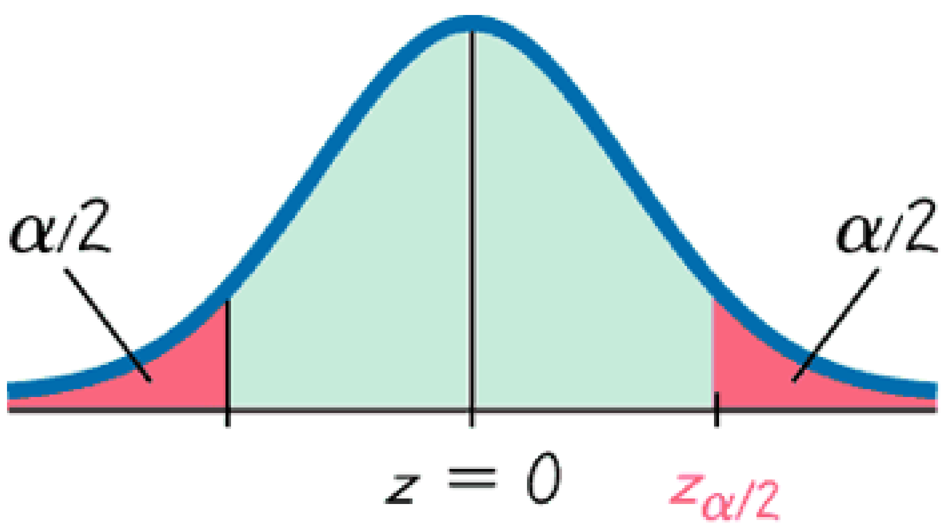
Confidence level is usually 90% or 95% or 99%.

Next we learn how to construct confidence intervals

Critical Values

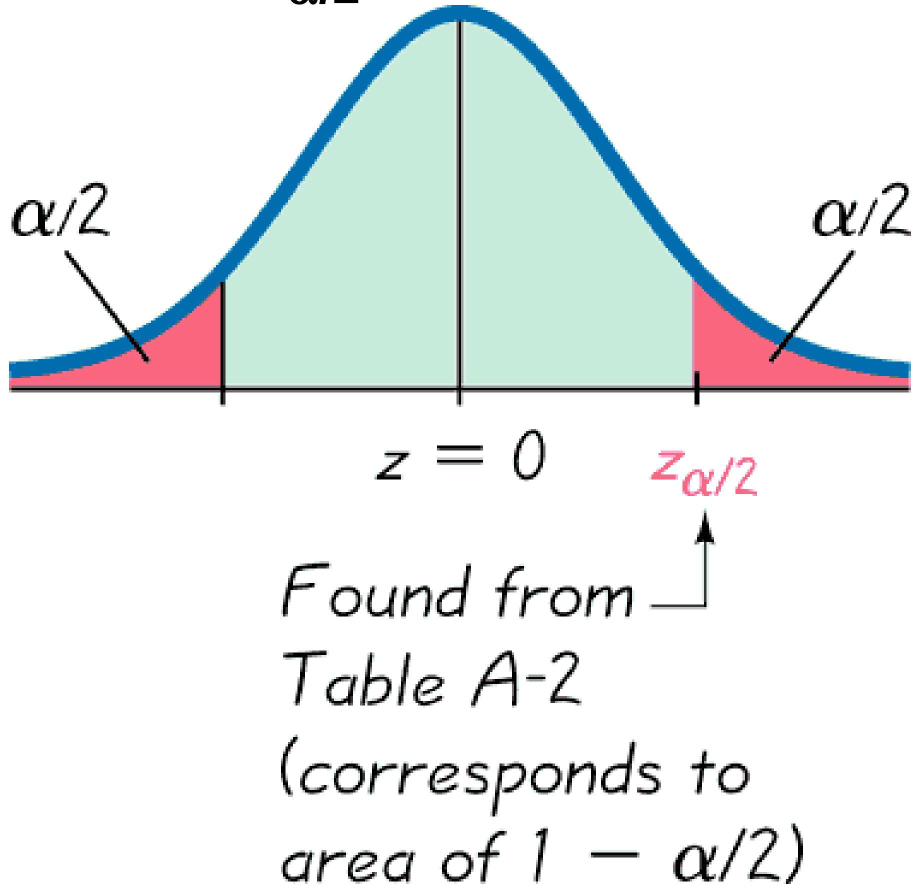
A *z* score can be used to distinguish between sample statistics that are likely to occur and those that are unlikely to occur. Such a *z* score is called a **critical value**.

The standard normal distribution is divided into three regions: middle part has area $1-\alpha$ and two tails (left and right) have area $\alpha/2$ each:



Critical Values

The *z* scores separate the middle interval (likely values) from the tails (unlikely values). They are $z_{\alpha/2}$ and $-z_{\alpha/2}$, found from Table A-2.



Definition

A critical value is the number on the borderline separating sample statistics that are likely to occur from those that are unlikely to occur.

Notation for Critical Value

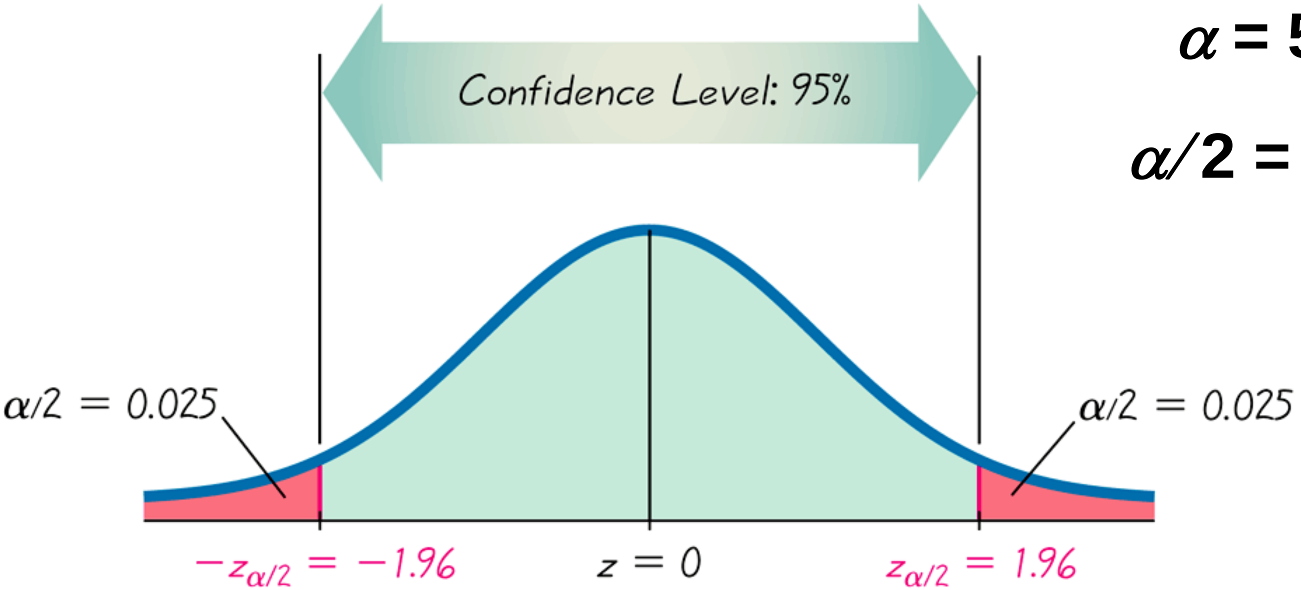
The critical value $z_{\alpha/2}$ separates an area of $\alpha/2$ in the right tail of the standard normal distribution. The value of $-z_{\alpha/2}$ separates an area of $\alpha/2$ in the left tail.

The subscript $\alpha/2$ is simply a reminder that the z score separates an area of $\alpha/2$ in the tail.

Finding $z_{\alpha/2}$ for a 95% Confidence Level

$\alpha = 5\%$

$\alpha/2 = 2.5\% = .025$



$-z_{\alpha/2}$ $z_{\alpha/2}$

Critical Values

Definition

Margin of error, denoted by E, is the maximum likely difference (with probability $1 - \alpha$, such as 0.95) between the observed proportion $\hat{p}$ and the true value of the population proportion p.

The margin of error E is also called the maximum error of the estimate.

Margin of Error for Proportions

$$E = z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}$$

Notation

E = margin of error

$\hat{p}$ = sample proportion

$\hat{q} = 1 - \hat{p}$

n = number of sample values

Example (continued)

- **We say that 0.70, or 70% is be the best point estimate of the proportion of all adults who believe in global warming.**
- **But how reliable (accurate) is this estimate?**
- **We will see that its margin of error is 2.3%. This means the true proportion of adults who believe in global warming is between 67.7% and 72.3%.** $0.677 < p < 0.723$ **.This gives an interval (from 67.7% to 72.3%) containing the true (but unknown) value of the population proportion.**

Confidence Interval for a Population Proportion *p*

$$\hat{p} - E < p < \hat{p} + E$$

where

$$E = z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}$$

Confidence Interval for a Population Proportion p

$$\hat{p} - E < p < \hat{p} + E$$

$$\hat{p} \pm E$$

$$(\hat{p} - E, \hat{p} + E)$$

Finding the Point Estimate and *E* from a Confidence Interval

Point estimate of $\hat{p}$:

$$\hat{p} = \frac{\text{(upper confidence limit)} + \text{(lower confidence limit)}}{2}$$

Margin of Error:

$$E = \frac{\text{(upper confidence limit)} - \text{(lower confidence limit)}}{2}$$

Round-Off Rule for Confidence Interval Estimates of *p*

Round the confidence interval limits for *p* to

three significant digits.

Sample Size

Suppose we want to collect sample data in order to estimate some population proportion. The question is how many sample items must be obtained?