Statistics

Chapter 7: Interval Estimation

Where We're Going

- Use the sampling distribution of the statistic to form a *confidence interval* for the parameter
- Select the proper sample size when estimating a parameter

7.1: Large-Sample Confidence Interval for a Population Mean

- Suppose a sample of 225 college students spend an average of 7 hours per week on social media, with a standard deviation of 3 hours.
  - What can we conclude about *all* college students' social media time?

7.1: Large-Sample Confidence Interval for a Population Mean

- Suppose a sample of 225 college students spend an average of 7 hours per week on social media, with a standard deviation of 3 hours.
- Since the sample size is large, it is not unreasonable to assume that the sample mean is approximately normally distributed (CLT).

7.1: Large-Sample Confidence Interval for a Population Mean

- We can be 95%* sure that

$$-1.96 < \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} < 1.96$$

Rewriting this, we get

$$\bar{x} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}$$

*In the standard normal distribution, exactly 95% of the area under the curve is in the interval

-1.96 … +1.96

7.1: Large-Sample Confidence Interval for a Population Mean

- An **interval estimator** or **confidence interval** is a formula that tell us how to use sample data to calculate an interval that estimates a population parameter.

$$\boxed{\mu = \bar{x} \pm z\sigma_{\bar{x}}}$$

In the college student social media example, the 95% confidence interval is (7-1.96*3/15,7+1.96*3/15)

(6.608,7.392)

7.1: Large-Sample Confidence Interval for a Population Mean

- The **confidence coefficient** is the probability that a randomly selected confidence interval encloses the population parameter.
- The **confidence level** is the **confidence coefficient** expressed as a percentage.

  (90%, 95% and 99% are very commonly used.)

95% sure $\mu = \bar{x} \pm z\sigma_{\bar{x}}$

7.1: Large-Sample Confidence Interval for a Population Mean

- The area outside the confidence interval is denoted by α

$$\mu = \bar{x} \pm z\sigma_{\bar{x}}$$

So we are left with (1 – 95)% = 5% = α uncertainty about μ

7.1: Large-Sample Confidence Interval for a Population Mean

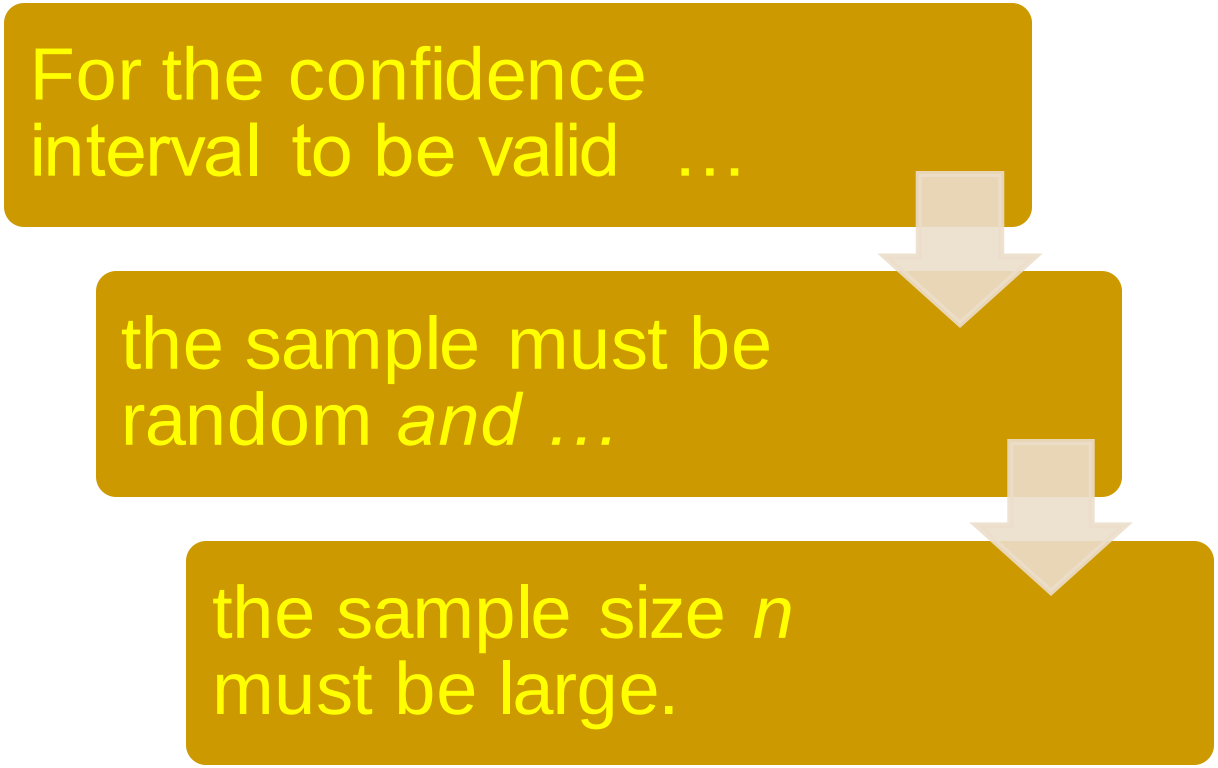
- Large-Sample (1- α)% Confidence Interval for μ

$$\mu = \overline{x} \pm z_{a/2}\sigma_{\overline{x}} = \overline{x} \pm z_{a/2}\frac{\sigma}{\sqrt{n}}$$

- If **σ** is unknown and *n* is large, the confidence interval becomes

$$\mu \cong \overline{x} \pm z_{a/2}s_{\overline{x}} = \overline{x} \pm z_{a/2}\frac{s}{\sqrt{n}}$$

7.1: Large-Sample Confidence Interval for a Population Mean



For the confidence interval to be valid …

the sample must be random *and* …

the sample size *n* must be large.

If *n* is large, the sampling distribution of the sample mean is normal, and *s* is a good estimate of **σ**

7.2: Small-Sample Confidence Interval for a Population Mean

Large Sample

- Distribution on X can be anything
- s is close to σ or σ is known.

Standard Normal (*z*) Distribution

$$\mu = \bar{x} \pm z_{a/2} \frac{\sigma}{\sqrt{n}}$$

Small Sample

- Distribution on X is normal
- Unknown **σ** and small *n*

Student's *t* Distribution (with *n-1* degrees of freedom)

$$\mu = \bar{x} \pm t_{a/2} \frac{s}{\sqrt{n}}$$

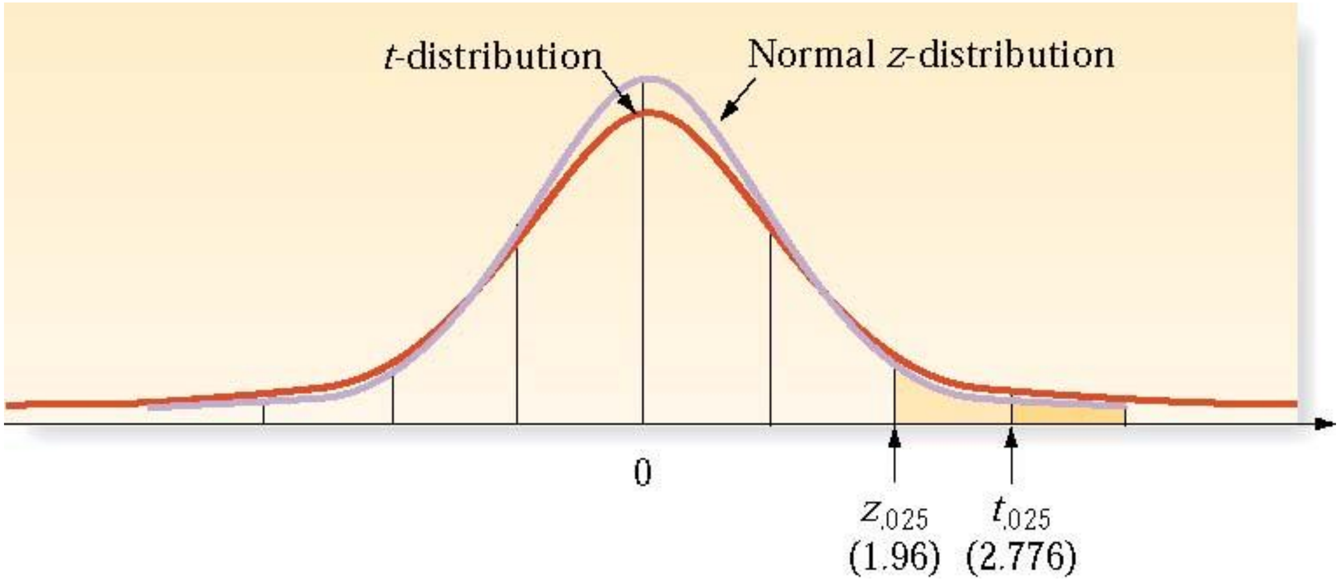
7.2: Small-Sample Confidence Interval for a Population Mean

Large Sample

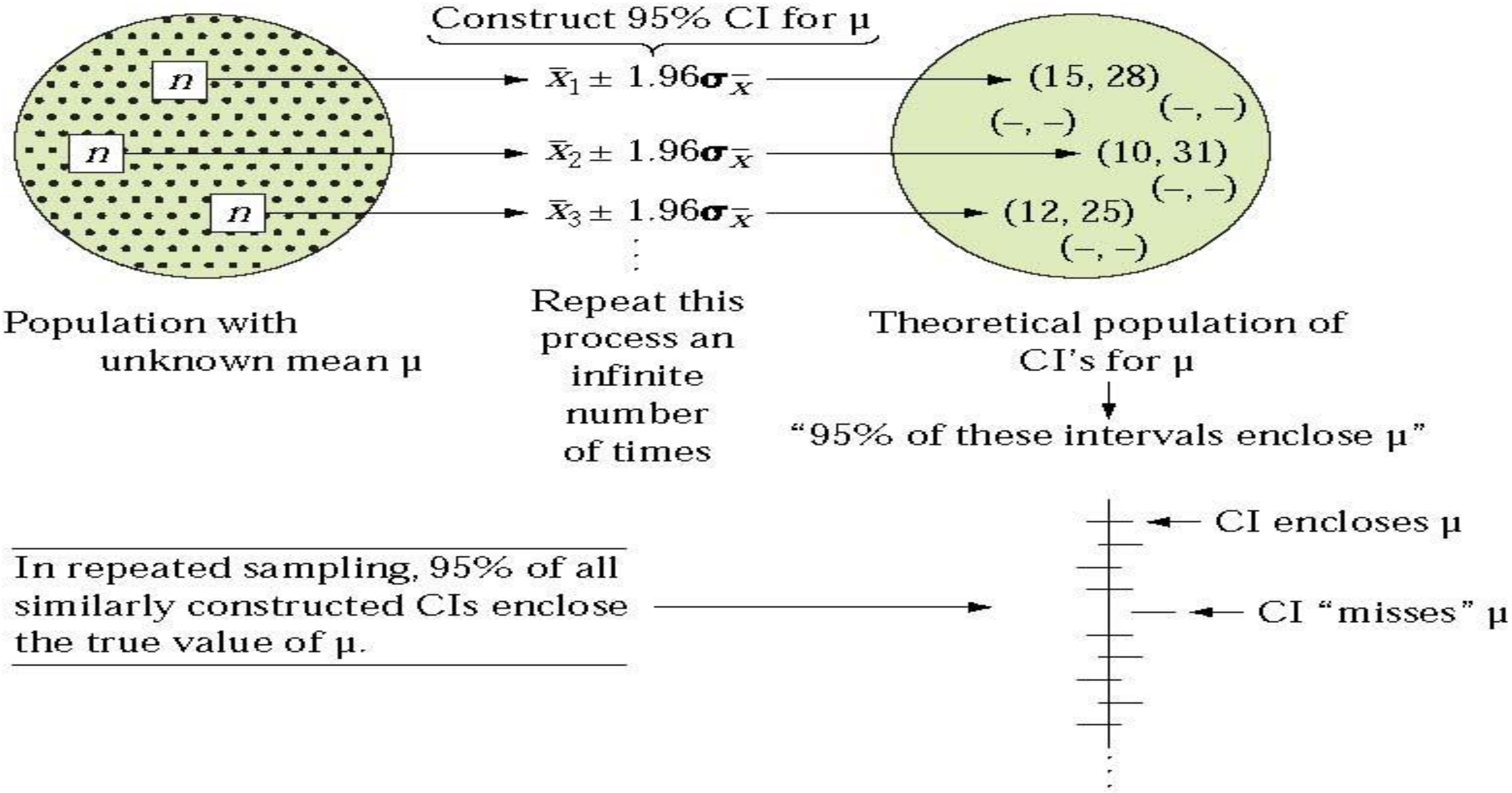
$$\mu = \bar{x} \pm z_{a/2} \frac{\sigma}{\sqrt{n}}$$

Small Sample

$$\mu = \bar{x} \pm t_{a/2} \frac{s}{\sqrt{n}}$$



7.2: Small-Sample Confidence Interval for a Population Mean


Construct 95% CI for μ
n
n
n
$\bar{x}_1 \pm 1.96\boldsymbol{\sigma}_{\bar{x}}$
$\bar{x}_2 \pm 1.96\boldsymbol{\sigma}_{\bar{x}}$
$\bar{x}_3 \pm 1.96\boldsymbol{\sigma}_{\bar{x}}$
(15, 28)
(–, –)
(–, –)
(10, 31)
(–, –)
(12, 25)
(–, –)
Population with unknown mean μ
Theoretical population of CI's for μ
CI encloses μ
CI “misses” μ

Repeat this process an infinite number of times

“95% of these intervals enclose μ”

In repeated sampling, 95% of all similarly constructed CIs enclose the true value of μ.

7.2: Small-Sample Confidence Interval for a Population Mean

- Suppose a sample of 25 college students spend an average of 7 hours per week on social media, with a standard deviation of 3 hours.
  - What can we conclude about *all* college students' social media time?

7.2: Small-Sample Confidence Interval for a Population Mean

- Assuming a normal distribution for social media engagement, we can be 95% sure that the average is in the interval

(7-2.064*3/5,7+2.064*3/5)

(5.76,8.24)

Note : qt(0.975,24)=2.064

7.3: Large-Sample Confidence Interval for a Population Proportion

- Sampling distribution of $\hat{p}$
  - The mean of the sampling distribution is *p*, the population proportion.
  - The standard deviation of the sampling distribution is

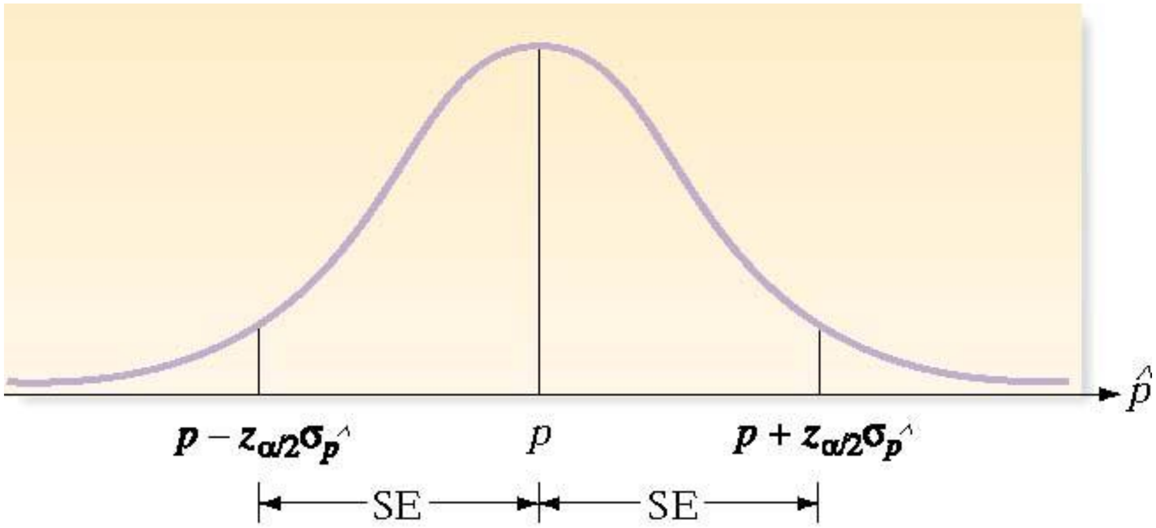
$$\sigma_{\hat{p}} = \sqrt{\frac{pq}{n}} \qquad \text{where} \quad q = 1 - p$$

  - For large samples the sampling distribution is approximately normal. Large is defined as

$$0 < \hat{p} \pm 3\sigma_{\hat{p}} < 1$$

7.3: Large-Sample Confidence Interval for a Population Proportion

- Sampling distribution of


$\hat{p}$
$p - z_{\alpha/2}\sigma_{\hat{p}}$
p
$p + z_{\alpha/2}\sigma_{\hat{p}}$
SE
SE

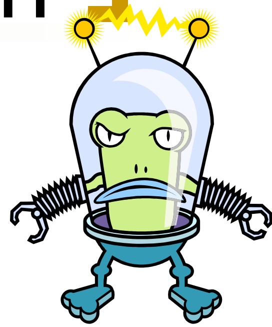
7.3: Large-Sample Confidence Interval for a Population Proportion

We can be 100(1-α)% confident that

$$p = \hat{p} \pm z_{\alpha/2}\sigma_{\hat{p}} = \hat{p} \pm z_{\alpha/2}\sqrt{\frac{pq}{n}} \cong \hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}$$

where $\hat{p} = \dfrac{x}{n}$ and $\hat{q} = 1 - \hat{p}$

7.3: Large-Sample Confidence Interval for a Population Proportion



A nationwide poll of nearly 1,500 people … conducted by the syndicated cable television show **Dateline: USA** found that more than 70 percent of those surveyed believe there is intelligent life in the universe, perhaps even in our own Milky Way Galaxy.

What proportion of the entire population agree, at the 95% confidence level?

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

$$p = .70 \pm 1.96\sqrt{\frac{(.70)(.30)}{1500}}$$

$$p = .70 \pm (1.96)(.012)$$

$$p = .70 \pm .023$$

7.3: Large-Sample Confidence Interval for a Population Proportion

- If *p* is close to 0 or 1, **Wilson's adjustment for estimating *p*** yields better results

$$\tilde{p} \pm z_{\alpha/2}\sqrt{\frac{\tilde{p}(1-\tilde{p})}{n+4}}$$

where $\tilde{p} = \dfrac{x+2}{n+4}$

7.3: Large-Sample Confidence Interval for a Population Proportion

Suppose in a particular year the percentage of firms declaring bankruptcy that had shown profits the previous year is .002. If 100 firms are sampled and one had declared bankruptcy, what is the 95% CI on the proportion of profitable firms that will tank the next year?

$$p = \tilde{p} \pm z_{\alpha/2}\sqrt{\frac{\tilde{p}(1-\tilde{p})}{n+4}}$$

$$\tilde{p} = \frac{x+2}{n+4} = \frac{1+2}{100+4} = .0289$$

$$p = .0289 \pm 1.96\sqrt{\frac{.0289(1-.0289)}{100+4}}$$

$$p = .0289 \pm .032$$

7.4: Determining the Sample Size

- To be within a certain sampling error (SE) of μ with a level of confidence equal to 100(1-α)%, we can solve

$$z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) = SE$$

for *n*:

$$n = \frac{(z_{\alpha/2})^2 \sigma^2}{SE^2}$$

7.4: Determining the Sample Size

- The value of **σ** will almost always be unknown, so we need an estimate:
  - *s* from a previous sample
  - approximate the range, *R*, and use *R*/4
- Round the calculated value of *n* upwards to be sure you don’t have too small a sample.

7.4: Determining the Sample Size

- Suppose we need to know the mean driving distance for a new composite golf ball within 3 yards, with 95% confidence. A previous study had a standard deviation of 25 yards. How many golf balls must we test?

7.4: Determining the Sample Size

Suppose we need to know the mean driving distance for a new composite golf ball within 3 yards, with 95% confidence. A previous study had a standard deviation of 25 yards. How many golf balls must we test?

$$n = \frac{(z_{\alpha/2})^2 \sigma^2}{SE^2}$$

$$n = \frac{1.96^2 25^2}{3^2}$$

$$n = 266.78 \cong 267$$

7.4: Determining the Sample Size

- For a confidence interval on the population proportion, *p*, we can solve

$$z_{\alpha/2}\sqrt{\frac{pq}{n}} = SE$$

for n: $$n = \frac{(z_{\alpha/2})^2(pq)}{SE^2}$$

To estimate *p*, use the sample proportion from a prior study, or use *p* = .7.

Round the value of *n* upward to ensure the sample size is large enough to produce the required level of confidence.

7.4: Determining the Sample Size

- How many cellular phones must a manufacturer test to estimate the fraction defective, *p*, to within .01 with 90% confidence, if an initial estimate of .10 is used for *p*?

7.4: Determining the Sample Size

How many cellular phones must a manufacturer test to estimate the fraction defective, *p*, to within .01 with 90% confidence, if an initial estimate of .10 is used for *p*?

$$SE = z_{\alpha/2}\sqrt{\frac{pq}{n}}$$

$$n = \frac{(z_{\alpha/2})^2(pq)}{(SE)^2}$$

$$n = \frac{(1.645)^2(.1)(.9)}{(.01)^2}$$

$$n = 2435.4 \cong 2436$$

7.5: Comparing Two Population Means: Independent Sampling

Point Estimators → $\bar{x} \rightarrow \mu$ $\quad \bar{x}_1 - \bar{x}_2 \rightarrow \mu_1 - \mu_2$

To construct a confidence interval or conduct a hypothesis test, we need the standard deviation:

Singe sample → $\hat{\sigma}_{\bar{x}} = s / \sqrt{n}$

Two samples → $$\hat{\sigma}_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

7.5: Comparing Two Population Means: Independent Sampling

The Sampling Distribution for $(\bar{x}_1 - \bar{x}_2)$

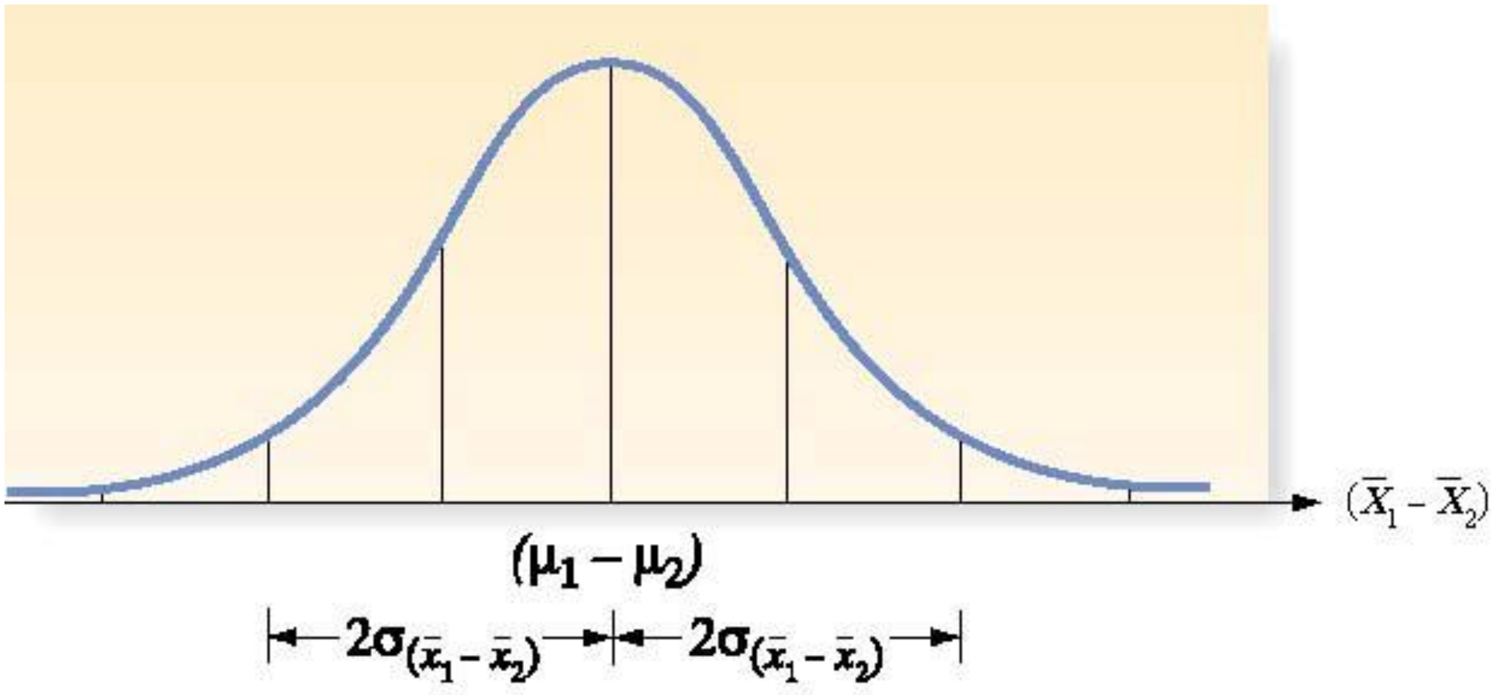
1. The mean of the sampling distribution is $(\mu_1 - \mu_2)$.
2. If the two samples are independent, the standard deviation of the sampling distribution (the *standard error*) is

$$\hat{\sigma}_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

3. The sampling distribution for $(\bar{x}_1 - \bar{x}_2)$ is approximately normal for large samples.

7.5: Comparing Two Population Means: Independent Sampling

The Sampling Distribution for $(\bar{x}_1 - \bar{x}_2)$



$(\bar{X}_1 - \bar{X}_2)$
$(\mu_1 - \mu_2)$
$2\sigma_{(\bar{x}_1 - \bar{x}_2)}$
$2\sigma_{(\bar{x}_1 - \bar{x}_2)}$

7.5: Comparing Two Population Means: Independent Sampling

Large Sample Confidence Interval for (μ_1 - μ_2)

$$(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2}\sigma_{(\bar{x}_1-\bar{x}_2)} = (\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$$\cong (\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

7.5: Comparing Two Population Means: Independent Sampling

Two samples concerning retention rates for first-year students at private and public institutions were obtained from the Department of Education's data base to see if there was a significant difference in the two types of colleges.

Private Colleges

- *n*: 71
- Mean: 78.17
- Standard Deviation: 7.35
- Variance: 91.17

Public Universities

- *n*: 32
- Mean: 84
- Standard Deviation: 7.88
- Variance: 97.64

What does a 95% confidence interval tell us about retention rates?

Source: National Center for Education Statistics

7.5: Comparing Two Population Means: Independent Sampling

Private Colleges

- *n*: 71
- Mean: 78.17
- Standard Deviation: 7.35
- Variance: 91.17

Public Universities

- *n*: 32
- Mean: 84
- Standard Deviation: 7.88
- Variance: 97.64

$$(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2}\sigma_{(\bar{x}_1 - \bar{x}_2)} \cong (\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$$78.17 - 84 \pm 1.96\sqrt{\frac{91.1}{71} + \frac{97.64}{32}}$$

$$-5.83 \pm 4.08$$

7.5: Comparing Two Population Means: Independent Sampling

Private Colleges

- *n*: 71
- Mean: 78.17
- Standard Deviation: 7.35
- Variance: 91.17

Public Universities

- *n*: 32
- Mean: 84
- Standard Deviation: 7.88
- iance: 97.64

$$(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2}$$

$$\pm z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$$78.17$$

$$\frac{97.64}{32}$$

$$-5.83 \pm 4.08$$

Since 0 is not in the confidence interval, the difference in the sample means appears to indicate a real difference in retention.

7.5: Comparing Two Population Means: Independent Sampling

For small samples, the *t*-distribution can be used with a **pooled sample estimator of σ^2**, $\boldsymbol{s_p^2}$

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

7.5: Comparing Two Population Means: Independent Sampling

Small Sample Confidence Interval for (μ_1 - μ_2)

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2}\sigma_{(\bar{x}_1-\bar{x}_2)} = (\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2}\sqrt{s_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$$

The value of *t* is based on (n_1 + n_2 -2) degrees of freedom.

7.6: Comparing Two Population Means: Paired Difference Experiments

$$\underline{\text{Paired Difference Confidence Interval for } \mu_{\mathrm{d}} = \mu_1 - \mu_2}$$

$$\text{Large Sample: } \bar{x}_d \pm z_{\alpha/2} \frac{\sigma_d}{\sqrt{n_d}} \cong \bar{x}_d \pm z_{\alpha/2} \frac{s_d}{\sqrt{n_d}}$$

$$\text{Small Sample: } \bar{x}_d \pm t_{\alpha/2} \frac{s_d}{\sqrt{n_d}} \text{ with } (n_d - 1) \text{ degrees of freedom,}$$

$$\text{where } \bar{x}_d = \text{sample mean difference}$$

$$\text{where } s_d = \text{sample standard deviation of differences}$$

$$\text{where } n_d = \text{number of pairs observed}$$

7.6: Comparing Two Population Means: Paired Difference Experiments

Suppose ten pairs of puppies were housetrained using two different methods: one puppy from each pair was paper-trained, with the paper gradually moved outside, and the other was taken out every three hours and twenty minutes after each meal. The number of days until the puppies were considered housetrained (three days straight without an accident) were compared. Nine of the ten paper-trained dogs took longer than the other paired dog to complete training, with the average difference equal to 4 days, with a standard deviation of 3 days.

What is a 90% confidence interval on the difference in successful training?

7.6: Comparing Two Population Means: Paired Difference Experiments

$$\bar{x}_d = 4$$

$$s_d = 3$$

$$90\% \text{ confident } \mu_1 - \mu_2 = \bar{x}_d \pm t_{\alpha/2, df = n_d - 1} \frac{s_d}{\sqrt{n_d}}$$

$$= 4 \pm 1.833 \frac{3}{\sqrt{10}} = 4 \pm 1.74$$

Since 0 is not in the interval, one program does seem to work more effectively.

7.7: Comparing Two Population Proportions: Independent Sampling

Two groups may or may not have similar proportions regarding particular characteristics.

We can make inferences about p_1 and p_2 by examining $\hat{p}_1$ and $\hat{p}_2$.

7.7: Comparing Two Population Proportions: Independent Sampling

Properties of the Sampling Distribution of $(\hat{p}_1 - \hat{p}_2)$

1. $E(\hat{p}_1 - \hat{p}_2) = (p_1 - p_2)$

2. $\sigma_{(\hat{p}_1 - \hat{p}_2)} = \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}} \cong \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}$

3. If the samples are large, the sampling distribution is approximately normal.

7.7: Comparing Two Population Proportions: Independent Sampling

$$\text{Large-Sample } 100(1\text{-}\alpha)\% \text{ Confidence Interval for } (p_1 - p_2)$$

$$(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2}\sigma_{(\hat{p}_1 - \hat{p}_2)} = (\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2}\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}$$

$$\cong (\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}$$

7.7: Comparing Two Population Proportions: Independent Sampling

- A group of men and women were asked their opinions on the following important issue:

  Are the Three Stooges funny?

  The results are as follow:

	Men	Women
Yes	290	200
No	50	50
n	340	250

7.7: Comparing Two Population Proportions: Independent Sampling

- Calculate a 95% confidence interval on the difference in the opinions of men and women.

	Men	Women
p	.85	.80
q	.15	.20
n	340	250

$$(\hat{p}_M - \hat{p}_W) \pm z_{\alpha/2}\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}} =$$

$$(.85 - .80) \pm 1.96\sqrt{\frac{.85 \times .15}{340} + \frac{.80 \times .20}{250}} =$$

$$.05 \pm .062$$

7.7: Comparing Two Population Proportions: Independent Sampling

- Calculate a 95% confidence interval on the difference in the f men and women.

	Men
p	.85
q	.15
n	340

Since 0 is in the confidence interval, we cannot rule out the possibility that both genders find the Stooges equally funny.
Nyuk nyuk nyuk.

$$q_1 + \frac{p_2 q_2}{n_2} =$$

$$\sqrt{\frac{5\times.15}{340}+\frac{.80\times.20}{250}} =$$

$$.05 \pm .062$$