Lecture 7. Confidence Interval Estimation (Chapters 7 and 8)

Ping Yu

HKU Business School
The University of Hong Kong

Plan of This Lecture

- Confidence Interval Estimation: One Population
  - One Normal Mean, Known Population Variance
  - One Normal Mean, Unknown Population Variance
  - One Proportion, Large Samples
  - One Normal Variance
  - Confidence Intervals in Finite Populations
- (*) Sample-Size Determination
  - Large Populations
  - Finite Populations
- Confidence Interval Estimation: Two Populations
  - Matched Pair: Two Means
  - Independent Samples: Two Normal Means, Known Population Variances
  - Independent Samples: Two Normal Means, Unknown Equal Population Variances
  - Independent Samples: Two Normal Means, Unknown Unequal Population Variances
  - Independent Samples: Two Proportions, Large Samples
  - Independent Samples: Two Normal Variances [exercise]
- The discussion of this lecture is parallel to that in the last lecture.

Confidence Interval Estimation: One Population

Confidence Interval (CI)

- A confidence interval estimator of a population parameter is a rule for determining (based on the sample) an interval that is likely to include the parameter. The corresponding estimate is called a confidence interval estimate.
  - This concept was introduced by Jerzy Neyman in 1937, "Outline of a Theory of Statistical Estimation Based on the Classical Theory of Probability", *Philosophical Transactions of the Royal Society A*, 236 (767): 333-380.
  - The variability of a point estimator is not reflected in its estimate, but can be reflected in a CI estimate – when the variability is smaller, the CI is typically shorter.
  - The textbook calls a confidence interval estimate as a confidence interval, but we will use "confidence interval" to refer to both "confidence interval estimator" and "confidence interval estimate", depending on the context.
- A CI is an interval giving a range of values that:
  - Takes into consideration variation in sample statistics from sample to sample.
  - Based on observation from 1 sample (i.e., map one sample to an interval).
  - Gives information about closeness to unknown population parameters.
  - Stated in terms of level of confidence, so can never be 100% confident.

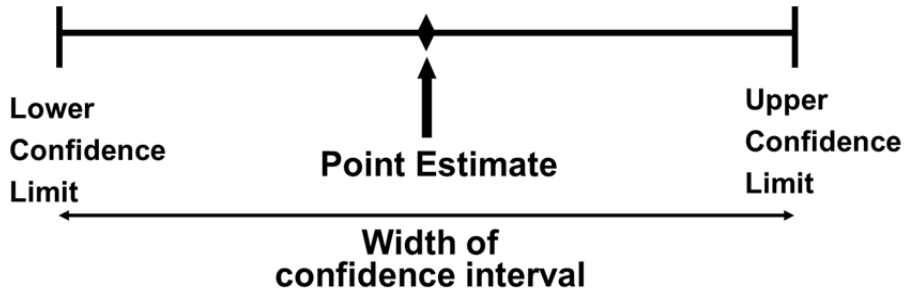
Confidence Level

- Suppose the CI of θ takes the form $[A,B]$, where A and B are random variables, i.e., $[A,B]$ is a random interval. If

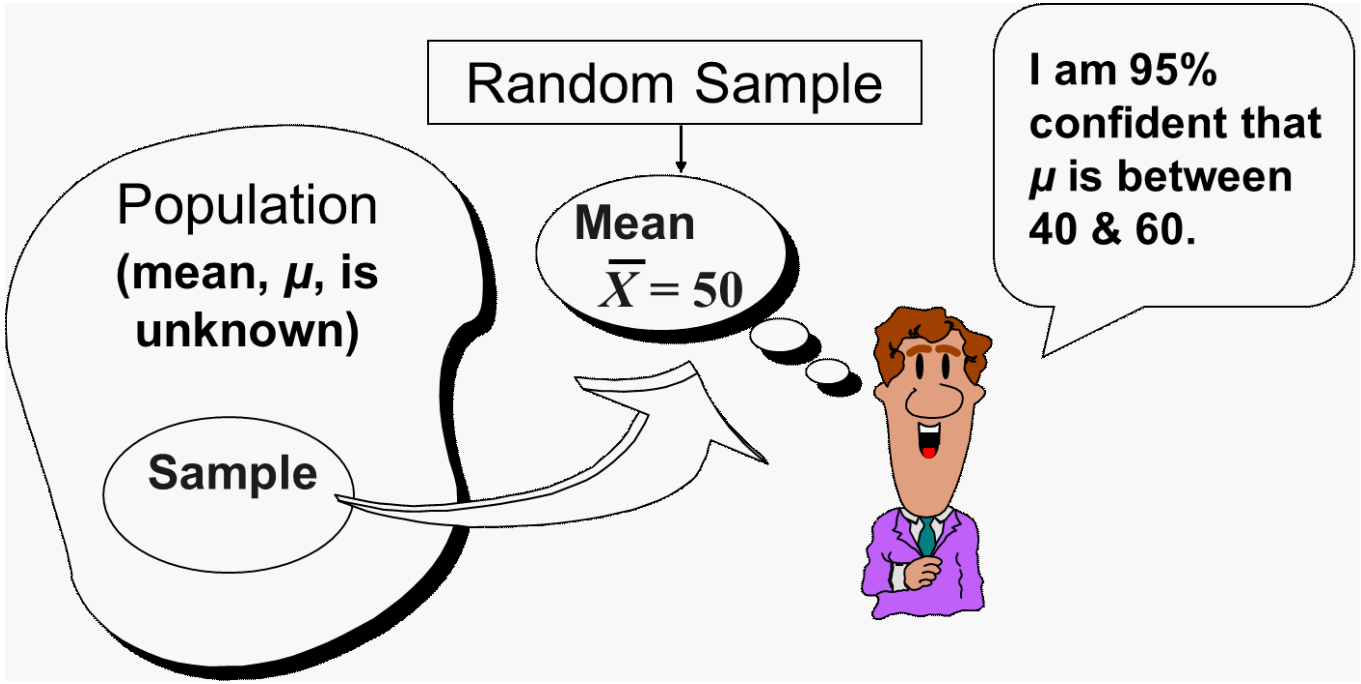
$$P(A \leq \theta \leq B) = 1-\alpha,$$

then $100(1-\alpha)\%$ is called the confidence level of the CI.
- We cannot say "θ falls in the CI with $(1-\alpha)$ probability" but only say "the CI covers θ with $(1-\alpha)$ probability", i.e., in repeated samples, $100(1-\alpha)\%$ (realized) intervals will cover θ, where note that given a realization of $[A,B]$, say $[a,b]$, either $\theta \in [a,b]$ or $\theta \notin [a,b]$, but we do not know which happens since θ is unknown.

- Analog: catch a butterfly using a net.
- A is the lower confidence limit (LCL) of the CI, B is the upper confidence limit (UCL) of the CI, and $B-A$ is the width of the CI:



Estimation Process



Margin of Error

- Because θ can be larger or smaller than a point estimator $\hat{\theta}$ of θ, the CI typically takes the form
$$\hat{\theta} \pm ME,$$
where the error factor ME is called the margin of error (or sampling error).
- ME should be increasing in $(1-\alpha)$.
- The UCL of the CI is given by
$$UCL = \hat{\theta} + ME.$$
- The LCL of the CI is given by
$$LCL = \hat{\theta} - ME.$$
- The width of the CI is equal to twice the ME:
$$w = 2(ME),$$
which is typically also random.
- Either an open interval $(\hat{\theta} - ME, \hat{\theta} + ME)$ or a closed interval $[\hat{\theta} - ME, \hat{\theta} + ME]$ is fine.

One Normal Mean, Known Population Variance

- From the last lecture, for a random sample $\{x_i\}_{i=1}^n$, where $x_i \sim N\left(\mu, \sigma^2\right)$ with unknown μ and known σ^2, if μ_0 is the true value of μ, then

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1),$$

which implies

$$\begin{aligned} 1-\alpha &= P\left(-z_{\alpha/2} \le z \le z_{\alpha/2}\right) = P\left(-z_{\alpha/2} \le \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}} \le z_{\alpha/2}\right) \\ &= P\left(-z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le \bar{x} - \mu_0 \le z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) \\ &= P\left(\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le \mu_0 \le \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right), \end{aligned}$$

i.e., the interval $\left[\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$ will cover μ_0 with probability $1-\alpha$, so it is a CI with confidence level $100(1-\alpha)\%$ or a $100(1-\alpha)\%$ CI.

- In this example, $\theta = \mu$, $\hat{\theta} = \bar{x}$ and $ME = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$.

A General Principle to Construct the CI: Inverting the Test Statistic

- Re-examining the procedure of constructing the CI above, we are actually inverting the test statistic in testing

$$H_0 : \mu = \mu_0 \text{ vs. } H_1 : \mu \neq \mu_0.$$

- Specifically, we try different μ_0's, and for each μ_0 value, we conduct the two-sided test with significance level α; if a μ_0 value is not rejected, then this μ_0 value is put in our CI. The interval collecting all μ_0 values that are not rejected is the CI with confidence level $(1-\alpha)$. [figure here]
- Conversely, if a value $\mu_0 \notin$CI, then we will reject $H_0 : \mu = \mu_0$ in favor of $H_1 : \mu \neq \mu_0$ at the level of $(1-$confidence level$)$.
- In summary, the hypothesis testing and CI construction are somewhat equivalent.

Table 7.2 Selected Confidence Levels and Corresponding Values of $z_{\alpha/2}$

CONFIDENCE LEVEL	90%	95%	98%	99%
α	0.100	0.05	0.02	0.01
$z_{\alpha/2}$	1.645	1.96	2.33	2.58

- The reliability factor $z_{\alpha/2}$ is the critical value for z at the significance level α.

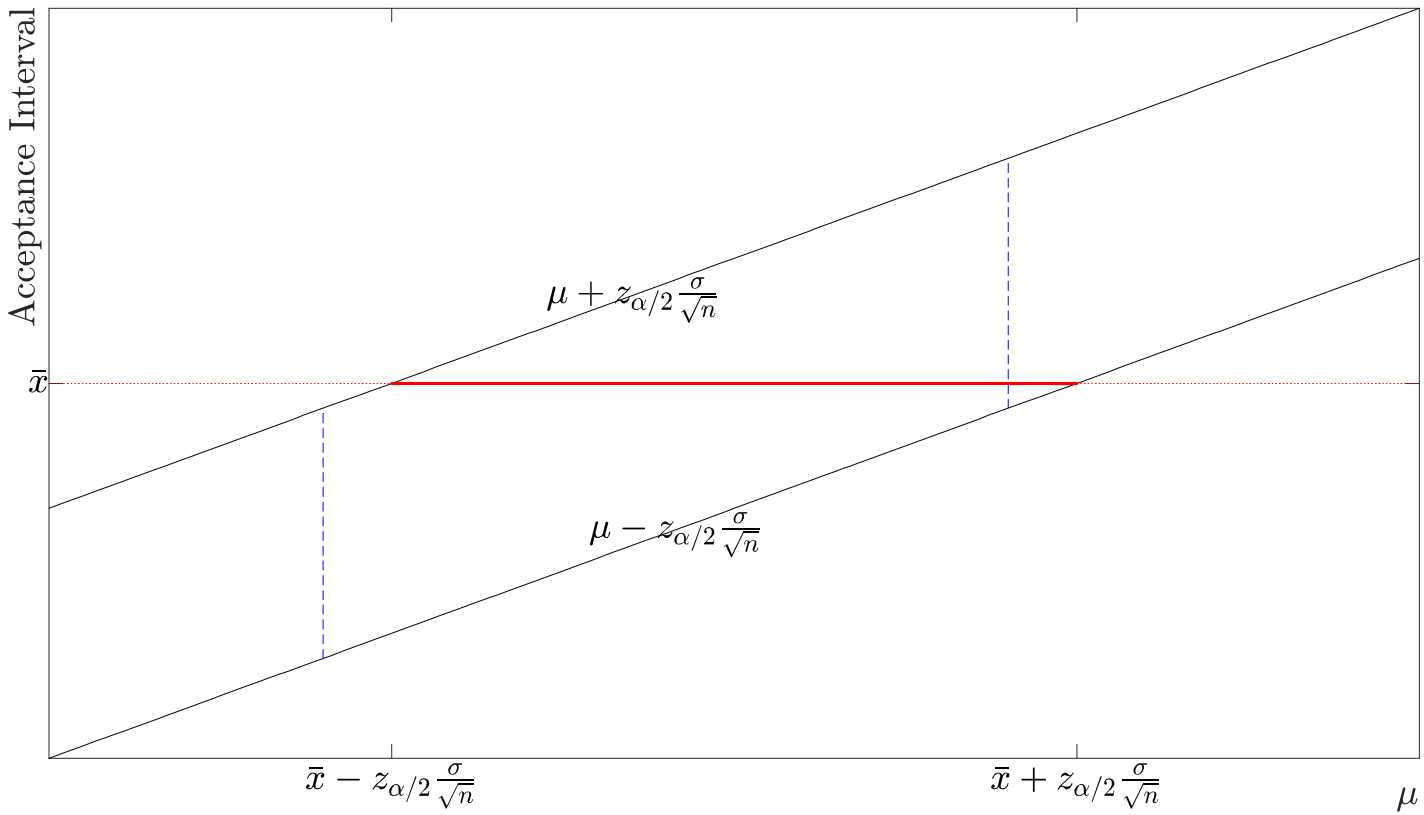


Figure: Test Statistic Inversion: recall that the acceptance interval for $\widehat{\mu}$ at μ is $\left[\mu - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \mu + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$

Example 7.3: Time at the Grocery Store

- Suppose the shopping times for customers are normally distributed with population standard deviation 20 minutes. A random sample of 64 shoppers had a mean time of 75 minutes. Construct the 95% CI for the population mean shopping time.
- Solution: Since

$$\bar{x} = 75 \text{ and } \sigma_{\bar{x}} = \sigma/\sqrt{n} = 20/\sqrt{64} = 2.5,$$

we have

$$\begin{aligned} ME &= z_{\alpha/2}\sigma_{\bar{x}} = 1.96 \times 2.5 = 4.9, \\ UCL &= \bar{x} + z_{\alpha/2}\sigma_{\bar{x}} = 75 + 4.9 = 79.9, \\ LCL &= \bar{x} - z_{\alpha/2}\sigma_{\bar{x}} = 75 - 4.9 = 70.1. \end{aligned}$$

So the 95% CI for the population mean shopping time is $[70.1, 79.9]$.

- Although the true mean may or may not be in this interval, 95% of intervals formed in this manner will contain the true mean.

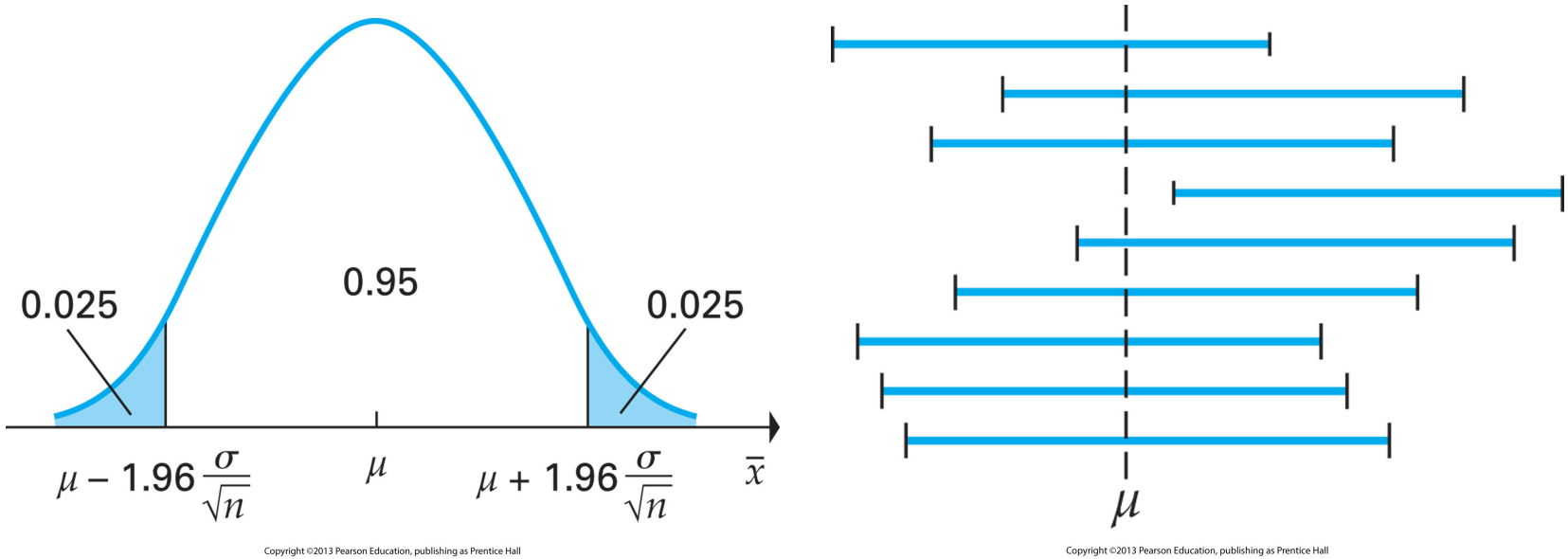


Figure: Sample Distribution of $\bar{x}$ and Schematic Description of 95% CI

- Intuition: $\bar{x}$ appears in $\left[\mu - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \mu + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$ with $(1-\alpha)$ probability, so extending $\bar{x}$ to the left and right by $z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$, i.e., $\left[\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$, will cover μ with $(1-\alpha)$ probability.

Reducing Margin of Error

- $ME = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ is decreasing in n and increasing in σ and $(1-\alpha)$.
  - Decreasing in n: if we get more information about the location of the butterfly, then we can use a smaller net.
  - Increasing in σ: if the information about the location of the butterfly is more vague, we must use a larger net.
  - Increasing in $(1-\alpha)$: to catch with a higher probability, we must use a larger net.
- To reduce the width of the CI ($=2\cdot ME$) while maintain the confidence level, we can either increase n or decrease σ (more information or better information).

continue

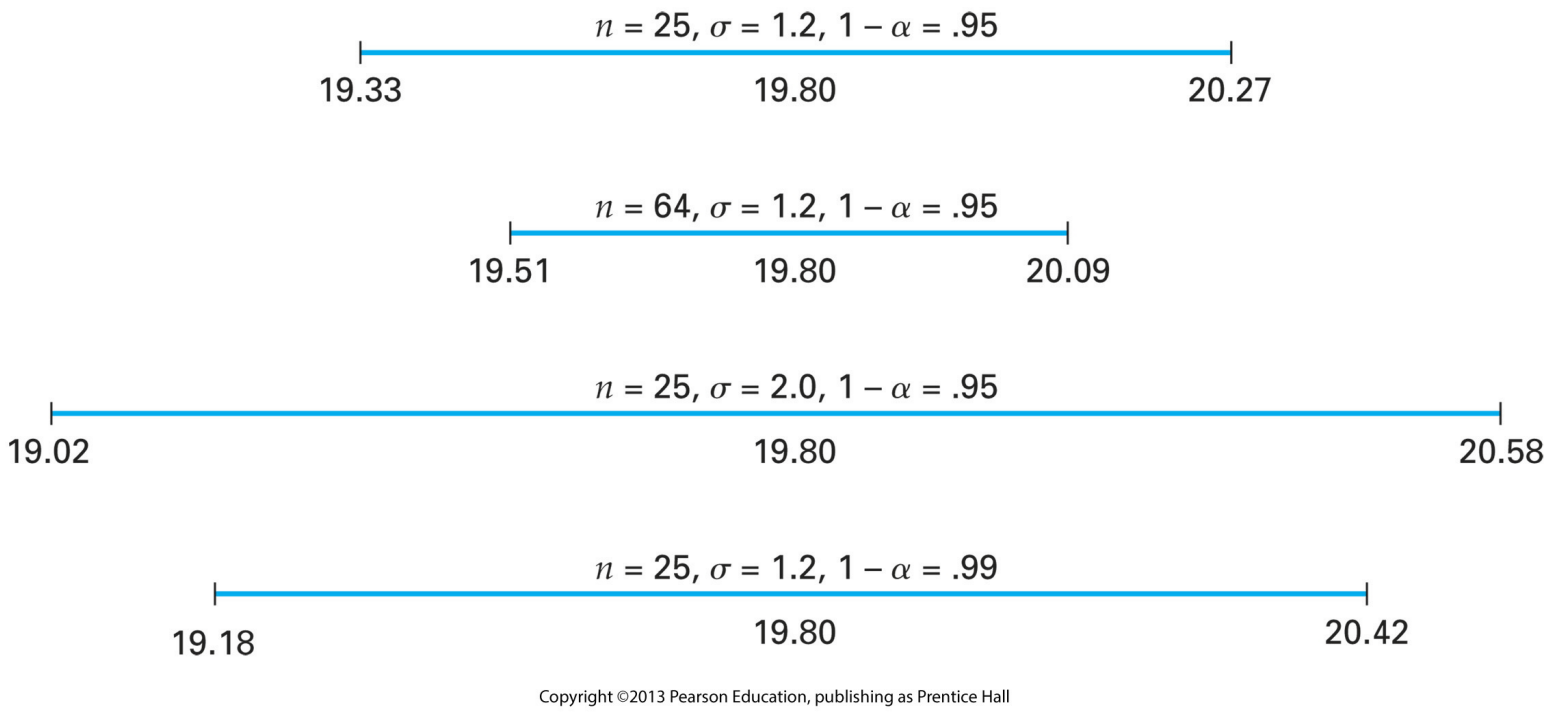


Figure: Effects of n, σ and $(1-\alpha)$ on CIs

One Normal Mean, Unknown Population Variance

- Inverting the test statistic

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

in testing $H_0 : \mu = \mu_0$ vs. $H_1 : \mu \neq \mu_0$, we have the $(1-\alpha)$ CI for μ is

$$\left\{ \mu_0 \middle| \left| \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \right| \leq t_{n-1,\alpha/2} \right\} = \left[\bar{x} - t_{n-1,\alpha/2} \frac{s}{\sqrt{n}}, \bar{x} + t_{n-1,\alpha/2} \frac{s}{\sqrt{n}} \right].$$

- $ME = t_{n-1,\alpha/2} \frac{s}{\sqrt{n}}$ is random, which is different from the known σ case where $ME = z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$ is fixed.

- Compared with $\left[\bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right]$, this CI should be wider because $t_{n-1,\alpha/2} > z_{\alpha/2}$ [table here]. This is the cost associated with replacing the unknown σ^2 by s^2.
  - When n gets large, $t_{n-1,\alpha/2} \approx z_{\alpha/2}$ and $s \approx \sigma$, so these two CIs are close.

Confidence Level	*t* (10 d.f.)	*t* (20 d.f.)	*t* (30 d.f.)	*z*
.80	1.372	1.325	1.310	1.282
.90	1.812	1.725	1.697	1.645
.95	2.228	2.086	2.042	1.960
.99	3.169	2.845	2.750	2.576

Figure: Comparison Between $t_{n,\alpha/2}$ and $z_{\alpha/2}$

A Numerical Example

- A random sample of $n = 25$ has $\bar{x} = 50$, and $s = 8$. Form a 95% CI for μ.
- Solution: The df $= n - 1 = 24$, so $t_{n-1,\alpha/2} = t_{24,.025} = 2.0639$.
- The CI is

$$\begin{aligned} & \left[\bar{x} - t_{n-1,\alpha/2}\frac{s}{\sqrt{n}}, \bar{x} + t_{n-1,\alpha/2}\frac{s}{\sqrt{n}}\right] \\ = & \left[50 - 2.0639 \times \frac{8}{\sqrt{25}}, 50 + 2.0639 \times \frac{8}{\sqrt{25}}\right] \\ = & [46.698, 53.302]. \end{aligned}$$

One Proportion, Large Samples

- We can invert the test statistic

$$z = \frac{\hat{p} - p_0}{\sqrt{p_0 (1 - p_0) / n}}$$

in testing $H_0 : p = p_0$ vs. $H_1 : p \neq p_0$, but z is a nonlinear function of p_0, so instead we replace p_0 in the denominator by $\hat{p}$ (which is consistent to p_0 as $n \to \infty$) to have the test statistic

$$\frac{\hat{p} - p_0}{\sqrt{\hat{p} (1 - \hat{p}) / n}}.$$

- The $(1 - \alpha)$ CI for p is

$$\left\{ p_0 \left| \left| \frac{\hat{p} - p_0}{\sqrt{\hat{p} (1 - \hat{p}) / n}} \right| \leq z_{\alpha/2} \right. \right\} = \left[\hat{p} - z_{\alpha/2} \sqrt{\frac{\hat{p} (1 - \hat{p})}{n}}, \hat{p} + z_{\alpha/2} \sqrt{\frac{\hat{p} (1 - \hat{p})}{n}} \right].$$

- $ME = z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ is random.

- The width of the CI can be reduced by either increasing n or decreasing $(1 - \alpha)$. [σ is $\sqrt{p (1 - p)}$ here, so out of control]

[Example] Proportion of Left-Handers

- A random sample of 100 people shows that 25 are left-handed. Form a 95% confidence interval for the true proportion of left-handers.

- Solution: First, $\hat{p} = 25/100 = 0.25$. Second, $\alpha = 0.05$, so $z_{\alpha/2} = 1.96$. Therefore, the 95% CI for p is

$$0.25 \pm 1.96\sqrt{\frac{0.25 \times (1-0.25)}{100}} = 0.25 \pm 0.085.$$

For the 99% CI for p, the ME increases from 0.085 to

$$2.58\sqrt{\frac{0.25 \times (1-0.25)}{100}} = 0.1117.$$

One Normal Variance

- Inverting the test statistic

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2},$$

in testing $H_0 : \sigma^2 = \sigma_0^2$ vs. $H_1 : \sigma^2 \neq \sigma_0^2$, we have the $(1-\alpha)$ CI for σ^2 is

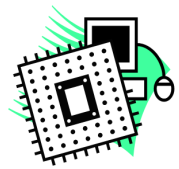
$$\left\{ \sigma_0^2 \left| \chi^2_{n-1,1-\alpha/2} \leq \frac{(n-1)s^2}{\sigma_0^2} \leq \chi^2_{n-1,\alpha/2} \right. \right\} = \left[\frac{(n-1)s^2}{\chi^2_{n-1,\alpha/2}}, \frac{(n-1)s^2}{\chi^2_{n-1,1-\alpha/2}} \right].$$

- This CI is not symmetric about s^2 (which is the unbiased estimator of σ^2) although it indeed includes s^2 because $\chi^2_{n-1,\alpha/2} > n-1$ and $\chi^2_{n-1,1-\alpha/2} < n-1$ for usual α's, so ME is not well defined.
- (*) In general, we can construct the CI for μ and σ^2 based on other tests (i.e., one-sided H_1) in the last lecture. However, such a CI may have an infinite length.
  - For example, inverting the test statistic for $H_0 : \mu \leq \mu_0$ vs. $H_0 : \mu > \mu_0$ with unknown normal variance, we have $\left[\bar{x} - t_{n-1,\alpha}\frac{s}{\sqrt{n}}, \infty\right)$, where compared with $\left[\bar{x} - t_{n-1,\alpha/2}\frac{s}{\sqrt{n}}, \bar{x} + t_{n-1,\alpha/2}\frac{s}{\sqrt{n}}\right]$, $\bar{x} - t_{n-1,\alpha}\frac{s}{\sqrt{n}} > \bar{x} - t_{n-1,\alpha/2}\frac{s}{\sqrt{n}}$, but $\infty > \bar{x} + t_{n-1,\alpha/2}\frac{s}{\sqrt{n}}$.

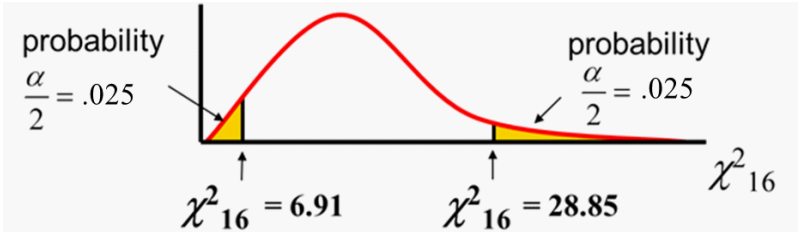
[Example] Speed of Computer Processors

- You are testing the speed of a batch of computer processors. You collect the following data (in Mhz):

Sample size	**17**
Sample mean	**3004**
Sample std dev	**74**



- Assume the population is normal. Determine the 95% confidence interval for σ^2.
- Solution: $n = 17$, so the chi-square distribution has $(n-1) = 16$ d.f..
- $\alpha = .05$, so use the the chi-square values with area .025 in each tail: $\chi^2_{n-1,\alpha/2} = \chi^2_{16,.025} = 28.85$, and $\chi^2_{n-1,1-\alpha/2} = \chi^2_{16,.975} = 6.91$.



- The 95% CI is $\left[\frac{(n-1)s^2}{\chi^2_{n-1,\alpha/2}}, \frac{(n-1)s^2}{\chi^2_{n-1,1-\alpha/2}}\right] = \left[\frac{(17-1)\times 74^2}{28.85}, \frac{(17-1)\times 74^2}{6.91}\right] = [3037, 12680]$.
- Converting to standard deviation, we are 95% confident that the population standard deviation of CPU speed is between 55.1 and 112.6 Mhz.

(*) Confidence Intervals in Finite Populations

- n is not much smaller than N in random sampling without replacement, e.g., $n > 0.05N$.
- n itself is large enough so that the CLT can be applied.
- One Mean, Unknown Population Variance, Large Samples: the $(1-\alpha)$ CI for μ is

$$\left[\bar{x} - z_{\alpha/2}\hat{\sigma}_{\bar{x}}, \bar{x} + z_{\alpha/2}\hat{\sigma}_{\bar{x}}\right],$$

where

$$\hat{\sigma}_{\bar{x}}^2 = \frac{s^2}{n}\frac{N-n}{N} = s^2\left(\frac{1}{n} - \frac{1}{N}\right)$$

rather than s^2/n is an unbiased estimator of $Var(\bar{x})$.
- $ME = z_{\alpha/2}\hat{\sigma}_{\bar{x}} < z_{\alpha/2}\frac{s}{\sqrt{n}}$.

- The $(1-\alpha)$ CI for the population total $N\mu$ (e.g., the total enrollment in business statistics when μ is the mean enrollment) is

$$\left[N\bar{x} - z_{\alpha/2}N\hat{\sigma}_{\bar{x}}, N\bar{x} + z_{\alpha/2}N\hat{\sigma}_{\bar{x}}\right].$$

- $ME = z_{\alpha/2}N\hat{\sigma}_{\bar{x}}$ is N times the ME of the CI for μ. [example here]

(**) Discussion

- As shown in Lecture 5,
$$t := \frac{\bar{x}-\mu}{\sqrt{\frac{s^2}{n}\frac{N-n}{N}}} \xrightarrow{d} N(0,1),$$
as $N \to \infty$, so the $(1-\alpha)$ CI for μ is
$$[\bar{x} - z_{\alpha/2}\hat{\sigma}_{\bar{x}}, \bar{x} + z_{\alpha/2}\hat{\sigma}_{\bar{x}}]$$
with $\hat{\sigma}^2_{\bar{x}}$ defined above.
- In the textbook, the authors replace $z_{\alpha/2}$ by $t_{n-1,\alpha/2}$ and $\frac{s^2}{n}\frac{N-n}{N}$ by $\frac{s^2}{n}\left(\frac{N-n}{N-1}\right)$. The second replacement is innocent given that $N \to \infty$, but the first one is not.
- The t-distribution is applied for a finite n. When n is finite, it is appropriate to assume N is also finite because $n > 0.05N$. In this case, it is not appropriate to think x_i follows a normal distribution (which requires an infinite population because a normal distribution can take infinite values). As a result, when n is finite, t follows a discrete distribution, not the t-distribution!
- The logic inconsistency of the textbook is obvious in assuming large samples (i.e., $n \to \infty$) and at the same time using t_{n-1} distribution.

[Example] CI for Population Total

- A firm has a population of 1000 accounts and wishes to estimate the value of the total population balance. A sample of 80 accounts is selected with average balance of \$87.60 and standard deviation of \$22.30. Find the 95% CI of the total balance.
- Solution: $N = 1000$, $n = 80$, $\bar{x} = 87.6$, and $s = 22.3$.
- In the textbook,

$$N^2\hat{\sigma}_{\bar{x}}^2 = N^2\frac{s^2}{n}\left(\frac{N-n}{N-1}\right) = 1000^2\frac{22.3^2}{80}\left(\frac{1000-80}{1000-1}\right) = 5724559.6,$$

so $N\hat{\sigma}_{\bar{x}} = \sqrt{5724559.6} = 2392.6$, and the resulting 95% CI for $N\mu$ is

$$\begin{aligned} N\bar{x} \pm t_{79,.025}N\hat{\sigma}_{\bar{x}} &= 1000 \times 87.6 \pm 1.9905 \times 2392.6 \\ &= [82837.53, 92362.47]. \end{aligned}$$

- Actually,

$$N^2\hat{\sigma}_{\bar{x}}^2 = N^2\frac{s^2}{n}\frac{N-n}{N} = 1000^2\frac{22.3^2}{80}\frac{1000-80}{1000} = 5718835,$$

so $N\hat{\sigma}_{\bar{x}} = \sqrt{5718835} = 2391.41$, and the resulting 95% CI for $N\mu$ is

$$N\bar{x} \pm z_{.025}N\hat{\sigma}_{\bar{x}} = 1000 \times 87.6 \pm 1.96 \times 2391.41 = [82912.84, 92287.16].$$

continue

- One Proportion, Large Samples: the $(1-\alpha)$ CI for p is

$$\left[\hat{p}-z_{\alpha/2}\hat{\sigma}_{\hat{p}},\hat{p}+z_{\alpha/2}\hat{\sigma}_{\hat{p}}\right],$$

where

$$\hat{\sigma}_{\hat{p}}^2=\frac{\hat{p}\left(1-\hat{p}\right)}{n-1}\frac{N-n}{N}$$

is an unbiased estimator of $Var\left(\hat{p}\right)$. [see Problem 5(iii) of Assignment III, $\frac{s^2}{n}=\frac{\hat{p}(1-\hat{p})}{n-1}$]

- $ME=z_{\alpha/2}\hat{\sigma}_{\hat{p}}=z_{\alpha/2}\sqrt{\frac{(N-n)n}{N(n-1)}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}<z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ if $n^2>N$ (or $n>N^{1/2}$).

(*) Sample-Size Determination

Large Populations

- If we think the CI is too wide, we can narrow it by increasing n.
- Fix the width of the CI, determine how large an n can achieve it.
- Consider only two cases below.
- One Normal Mean, Known Population Variance: Solving

$$ME = z_{\alpha/2}\frac{\sigma}{\sqrt{n}},$$

we have

$$n = \frac{z_{\alpha/2}^2\sigma^2}{ME^2}, \tag{1}$$

which increases in $1-\alpha$, σ^2 and decreases in ME, i.e., to make a $(1-\alpha)$ CI for μ extend a distance ME on each side of $\bar{x}$, we need $\frac{z_{\alpha/2}^2\sigma^2}{ME^2}$ (or the rounding-up $\left\lceil \frac{z_{\alpha/2}^2\sigma^2}{ME^2} \right\rceil$ if $\frac{z_{\alpha/2}^2\sigma^2}{ME^2}$ is not an integer) samples.

- Example: If $\sigma = 45$, what sample size is needed to estimate the mean within ± 5 with 90% confidence?
- Solution: $n = \frac{z_{\alpha/2}^2\sigma^2}{ME^2} = \frac{1.645^2\times 45^2}{5^2} = 219.19$, so the required sample size is $n = 220$.

continue

- One Proportion, Large Samples: We cannot solve

$$ME = z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},$$

to have the required n, since $\hat{p}$ is unobserved beforehand.

- Anyway, $\hat{p}(1-\hat{p}) \leq 0.25$, so solving

$$ME = z_{\alpha/2}\sqrt{\frac{0.25}{n}},$$

we conclude that

$$n = \frac{0.25 z_{\alpha/2}^2}{ME^2}$$

can **guarantee** that the CI extends no more than ME on each side of the $\hat{p}$, where n increases in $1-\alpha$ and decreases in ME.
- For one normal mean with unknown population variance, $ME = t_{n-1,\alpha/2}\frac{s}{\sqrt{n}}$ is not easy to solve since both $t_{n-1,\alpha/2}$ and s depend on n.

- Example 7.14: Electoral College: If an opinion survey on changing the Electoral College process reported that the poll has a 3% margin of error (with 95% confidence), how many citizens of voting age need to be sampled?
- Solution: $n = \frac{0.25 z_{\alpha/2}^2}{ME^2} = \frac{0.25 \times 1.96^2}{0.03^2} = 1067.11 \Longrightarrow n = 1068$.

Finite Populations

- One Mean, Known Population Variance: If $\sigma_{\bar{x}}^2$ is the target, then solving

$$\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right),$$

we have

$$n = \frac{N\sigma^2}{(N-1)\,\sigma_{\bar{x}}^2 + \sigma^2}.$$

- If ME is the target, then solving

$$ME = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$$

to have

$$n = \frac{n_0 N}{n_0 + (N-1)} \leq n_0,$$

where n_0 is given in (1). [Implicitly assume normality or the implied n is large]
- If the population total is of interest, then solve $ME = z_{\alpha/2} N \sigma_{\bar{x}}$.

- With nonresponse or missing data, practitioners may add a certain percent (like 10%) to the implied size n.

continue

- One Proportion, Large Samples: Solving

$$\sigma_{\hat{p}}^2 = \frac{p(1-p)}{n}\left(\frac{N-n}{N-1}\right)$$

to have

$$n = \frac{Np(1-p)}{(N-1)\sigma_{\hat{p}}^2 + p(1-p)}.$$

- Since $p(1-p)$ is unknown, the largest possible value of n (note that n is an increasing function of $p(1-p)$) is

$$n_{\max} = \frac{0.25N}{(N-1)\sigma_{\hat{p}}^2 + 0.25}.$$

- A 95% CI for p will extend approximately $1.96\sigma_{\hat{p}}$ on each side of $\hat{p}$.

- Example 7.16: Campus Survey: Suppose a random sample of the 1,395 U.S. colleges is taken to estimate the proportion for which the business statistics course is two semesters long. For a 95% CI to extend no further than 0.04 on each side of the sample proportion, how many samples should be taken?
- Solution: $1.96\sigma_{\hat{p}} = 0.04 \Longrightarrow \sigma_{\hat{p}} = 0.020408$. So

$$n_{\max} = \frac{0.25N}{(N-1)\sigma_{\hat{p}}^2+0.25} = \frac{0.25\times 1{,}395}{1{,}394\times 0.020408^2+0.25} = 419.88 \Longrightarrow n = 420.$$

Confidence Interval Estimation: Two Populations

Matched Pair: Two Means

- Inverting the test statistic

$$t = \frac{\bar{d} - \mu_0}{s_d / \sqrt{n}}$$

in testing $H_0 : \mu_d := \mu_x - \mu_y = \mu_0$ vs. $H_1 : \mu_d \neq \mu_0$, we have the $(1-\alpha)$ CI for μ_d is

$$\left\{ \mu_0 \middle| \left| \frac{\bar{d} - \mu_0}{s_d / \sqrt{n}} \right| \leq t_{n-1,\alpha/2} \right\} = \bar{d} \pm t_{n-1,\alpha/2} \frac{s_d}{\sqrt{n}},$$

where $\bar{d} = \bar{x} - \bar{y}$, and s_d is the sample standard deviation of $\{d_i\}_{i=1}^n$ with $d_i = x_i - y_i$.
- $ME = t_{n-1,\alpha/2} \frac{s_d}{\sqrt{n}}$.
- If the CI contains 0, then we cannot reject $\mu_x = \mu_y$ at the significance level α.

[Example] Weight Loss Program

- Six people sign up for a weight loss program. You collect the following data:

	Weight:		
Person	Before (x)	After (y)	Difference, d_i
1	136	125	11
2	205	195	10
3	157	150	7
4	138	140	−2
5	175	165	10
6	166	160	6
			42

  Form the 95% CI for $\mu_x - \mu_y$.
- Solution: $\bar{d} = 42/6 = 7$, and $s_d = 4.82$. $t_{n-1,\alpha/2} = t_{5,.025} = 2.571$.
- So the 95% CI for $\mu_x - \mu_y$ is $\bar{d} \pm t_{n-1,\alpha/2} \frac{s_d}{\sqrt{n}} = 7 \pm 2.571 \frac{4.82}{\sqrt{6}} = [-1.94, 12.06]$.
- Since this interval contains zero, we cannot be 95% confident, given this limited data, that the weight loss program helps people lose weight.

Independent Samples: Two Normal Means, Known Population Variances

- Inverting the test statistic

$$z = \frac{(\bar{x} - \bar{y}) - \mu_0}{\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}}$$

in testing $H_0 : \mu_d = \mu_0$ vs. $H_1 : \mu_d \neq \mu_0$, we have the $(1-\alpha)$ CI for μ_d is

$$\left\{ \mu_0 \left| \frac{(\bar{x} - \bar{y}) - \mu_0}{\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}} \right| \leq z_{\alpha/2} \right\} = (\bar{x} - \bar{y}) \pm z_{\alpha/2} \sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}.$$

- $ME = z_{\alpha/2} \sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}$.

Independent Samples: Two Normal Means, Unknown Equal Population Variances

- Inverting the test statistic

$$t=\frac{(\bar{x}-\bar{y})-\mu_0}{\sqrt{\frac{s_p^2}{n_x}+\frac{s_p^2}{n_y}}}$$

in testing $H_0:\mu_d=\mu_0$ vs. $H_1:\mu_d\neq\mu_0$, we have the $(1-\alpha)$ CI for μ_d is

$$\left\{\mu_0\left|\frac{(\bar{x}-\bar{y})-\mu_0}{\sqrt{\frac{s_p^2}{n_x}+\frac{s_p^2}{n_y}}}\right|\leq t_{n-2,\alpha/2}\right\}=(\bar{x}-\bar{y})\pm t_{n-2,\alpha/2}\sqrt{\frac{s_p^2}{n_x}+\frac{s_p^2}{n_y}},$$

where $n=n_x+n_y$, and the pooled sample variance

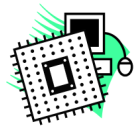
$$s_p^2=\frac{(n_x-1)s_x^2+(n_y-1)s_y^2}{n_x+n_y-2}.$$

- $ME=t_{n-2,\alpha/2}\sqrt{\frac{s_p^2}{n_x}+\frac{s_p^2}{n_y}}$.

[Example] Speed Difference Between Two CPUs

- You are testing two computer processors for speed. Form a CI for the difference in CPU speed. You collect the following speed data (in Mhz):

	CPU_x	CPU_y
Number Tested	17	14
Sample mean	3004	2538
Sample std dev	74	56



- Assume both populations are normal with equal variances, and use 95% confidence.
- Solution: The pooled variance is

$$s_p^2 = \frac{(n_x-1)s_x^2+(n_y-1)s_y^2}{n_x+n_y-2} = \frac{(17-1)\times 74^2+(14-1)\times 56^2}{17+14-2} = 4427.03.$$

- The t value for the 95% CI is $t_{n-2,\alpha/2} = t_{29,.025} = 2.045$.
- The 95% CI is

$$\begin{aligned}(\bar{x}-\bar{y}) \pm t_{n-2,\alpha/2}\sqrt{\frac{s_p^2}{n_x}+\frac{s_p^2}{n_y}} &= (3004-2538)\pm 2.045\sqrt{\frac{4427.03}{17}+\frac{4427.03}{14}} \\ &= [416.69, 515.31].\end{aligned}$$

Independent Samples: Two Normal Means, Unknown Unequal Population Variances

- Inverting the test statistic

$$t=\frac{(\bar{x}-\bar{y})-\mu_0}{\sqrt{\frac{s_x^2}{n_x}+\frac{s_y^2}{n_y}}}$$

in testing $H_0:\mu_d=\mu_0$ vs. $H_1:\mu_d\neq\mu_0$, we have the $(1-\alpha)$ CI for μ_d is

$$\left\{\mu_0\left|\frac{(\bar{x}-\bar{y})-\mu_0}{\sqrt{\frac{s_x^2}{n_x}+\frac{s_y^2}{n_y}}}\right|\le t_{v,\alpha/2}\right\}=(\bar{x}-\bar{y})\pm t_{v,\alpha/2}\sqrt{\frac{s_x^2}{n_x}+\frac{s_y^2}{n_y}},$$

where $v=\frac{\left[\left(\frac{s_x^2}{n_x}\right)+\left(\frac{s_y^2}{n_y}\right)\right]^2}{\left(\frac{s_x^2}{n_x}\right)^2/(n_x-1)+\left(\frac{s_y^2}{n_y}\right)^2/(n_y-1)}$ is defined in the last lecture.

- $ME=t_{v,\alpha/2}\sqrt{\frac{s_x^2}{n_x}+\frac{s_y^2}{n_y}}$.

- Whether $\sigma_x^2=\sigma_y^2$ or not can be tested using the test in the last lecture.

Independent Samples: Two Proportions, Large Samples

- Recall that the test statistic is

$$z=\frac{\hat{p}_x-\hat{p}_y}{\sqrt{\frac{\hat{p}_0(1-\hat{p}_0)}{n_x}+\frac{\hat{p}_0(1-\hat{p}_0)}{n_y}}},$$

  in testing $H_0: p_x-p_y=0$ vs. $H_1: p_x-p_y\neq 0$, which employs the null information $p_x=p_y$ in estimating the variance of $\hat{p}_x-\hat{p}_y$.
- In testing $H_0: p_x-p_y=p_0$ vs. $H_1: p_x-p_y\neq p_0$ for a general p_0 value, the proper test statistic is

$$z=\frac{(\hat{p}_x-\hat{p}_y)-p_0}{\sqrt{\frac{\hat{p}_x(1-\hat{p}_x)}{n_x}+\frac{\hat{p}_y(1-\hat{p}_y)}{n_y}}},$$

  inverting which to have the $(1-\alpha)$ CI for p_x-p_y as

$$\left\{p_0\left|\frac{(\hat{p}_x-\hat{p}_y)-p_0}{\sqrt{\frac{\hat{p}_x(1-\hat{p}_x)}{n_x}+\frac{\hat{p}_y(1-\hat{p}_y)}{n_y}}}\right|\le z_{\alpha/2}\right\}=(\hat{p}_x-\hat{p}_y)\pm z_{\alpha/2}\sqrt{\frac{\hat{p}_x(1-\hat{p}_x)}{n_x}+\frac{\hat{p}_y(1-\hat{p}_y)}{n_y}}.$$

  - $ME=z_{\alpha/2}\sqrt{\frac{\hat{p}_x(1-\hat{p}_x)}{n_x}+\frac{\hat{p}_y(1-\hat{p}_y)}{n_y}}$.

[Example] Difference Between the College Degree Rates

- Form a 90% confidence interval for the difference between the proportion of men and the proportion of women who have college degrees. In a random sample, 26 of 50 men and 28 of 40 women had an earned college degree.

- Solution: Men: $\hat{p}_x = \frac{26}{50} = 0.52$; Women: $\hat{p}_y = \frac{28}{40} = 0.70$.
- $\sqrt{\frac{\hat{p}_x(1-\hat{p}_x)}{n_x} + \frac{\hat{p}_y(1-\hat{p}_y)}{n_y}} = \sqrt{\frac{0.52\times 0.48}{50} + \frac{0.7\times 0.3}{40}} = 0.1012.$
- For the 90% CI, $z_{\alpha/2} = 1.645$.
- So the 90% CI for $p_x - p_y$ is

$$(\hat{p}_x - \hat{p}_y) \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_x(1-\hat{p}_x)}{n_x} + \frac{\hat{p}_y(1-\hat{p}_y)}{n_y}} = (.52-.70) \pm 1.645 \times 0.1012$$
$$= [-0.3465, -0.0135].$$

- Since this interval does not contain zero we are 90% confident that the two proportions are not equal – the college degree rate of men is lower than that of women.

Comparison of the Neyman-Pearson Approach, Fisher's p-Value Approach and Neyman's CI

- Consider testing $H_0 : \mu = \mu_0$ vs. $H_1 : \mu \neq \mu_0$ at the significance level α.
- The Neyman-Pearson approach can only make a decison for a fixed μ_0 and a fixed α each time.
- Fisher's p-value approach can make a decision for any α but a fixed μ_0 each time.
- Neyman's CI can make a decision for any μ_0 but a fixed α each time.