Lecture 20: Confidence Interval for One Population Proportion

- **Confidence Interval for a Population Proportion:**
  - Formula: $\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
  - Requirements:
    * Random sample
    * Large enough sample size (np and n(1-p) both ≥ 10)
  - Agresti-Coull Adjustment: Add 2 successes and 2 failures to the sample size.
  - Interpretation: We are [confidence level] confident that the true population proportion is between [lower bound] and [upper bound].
- **Standard Error vs. Standard Deviation:**
  - Standard error (SE) measures the variability of the sample proportion.
  - Standard deviation (SD) measures the variability of the population.
  - SE is used when the population proportion is unknown.
- **Steps for Constructing a Confidence Interval:**
  - Verify assumptions (random sample, large enough sample size).
  - Calculate the sample proportion $\hat{p}$.
  - Find the critical value $z_{\alpha/2}$ based on the confidence level.
  - Calculate the margin of error.
  - Construct the confidence interval.
  - Interpret the confidence interval in the context of the problem.
- **Sample Size Calculation:**
  - Formula: $n = \frac{z^2_{\alpha/2} \cdot \hat{p}(1-\hat{p})}{ME^2}$
  - Use this formula to determine the minimum sample size needed to achieve a desired margin of error.

Lecture 21: Inference about $\mu_1 - \mu_2$

- **Assumptions for Two-Sample t-Tests:**
  - Independent random samples from two populations.
  - Both populations are normally distributed.
  - Population standard deviations are unknown.
- **Central Limit Theorem (CLT):**
  - If sample sizes are large enough (n $\geq$ 30), the sampling distribution of the sample mean will be approximately normal, regardless of the population distribution.
- **Hypothesis Testing for the Difference in Means:**
  - Null Hypothesis (H_0): $\mu_1 - \mu_2 = 0$ (no difference in means)
  - Alternative Hypothesis (H_a):
    * $\mu_1 - \mu_2 \neq 0$ (two-sided)
    * $\mu_1 - \mu_2 > 0$ (right-tailed)
    * $\mu_1 - \mu_2 < 0$ (left-tailed)
  - Test Statistic: $t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$
  - P-value: Probability of observing a test statistic as extreme or more extreme than the one calculated, assuming the null hypothesis is true.
  - Conclusion:
    * If p-value ¡ α (significance level), reject H_0.
    * If p-value $\geq \alpha$, fail to reject H_0.
- **Confidence Interval for the Difference in Means:**
  - Formula: $(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2} \cdot s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$
  - Interpretation: We are [confidence level] confident that the true difference in population means is between [lower bound] and [upper bound].

Lecture 22: Confidence Interval for $\mu_1 - \mu_2$

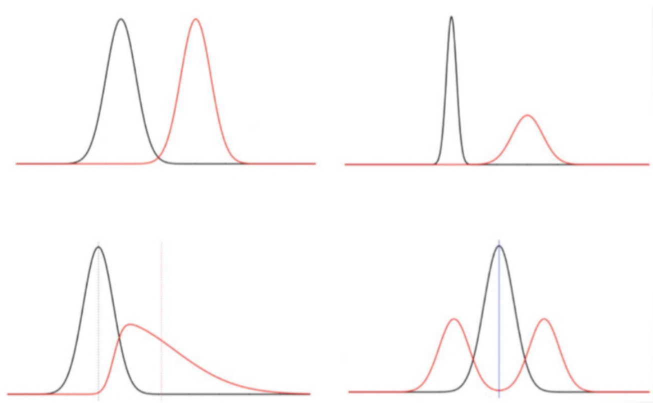
- **Pooled t-Test:**
  - Used when we assume that the population variances are equal.
  - Pooled variance: $s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$
- **Assumptions for Pooled t-Test:**
  - Independent random samples from two populations.
  - Both populations are normally distributed.
  - Population variances are equal.
- **Hypothesis Testing and Confidence Interval for Pooled t-Test:**
  - Similar to the regular two-sample t-test, but using the pooled variance s_p^2 in the calculations.
- **Paired t-Test:**
  - Used when the samples are dependent, meaning each observation in one sample is paired with a corresponding observation in the other sample.
  - Example: Before-and-after measurements on the same individuals.
- **Assumptions for Paired t-Test:**
  - The differences between paired observations are normally distributed.
- **Hypothesis Testing and Confidence Interval for Paired t-Test:**
  - Calculate the differences between paired observations.
  - Perform a one-sample t-test on the differences.

Lecture 23: Inference of $p_1 - p_2$

- **Inference for Two Population Proportions:**
  - Used to compare the proportions of successes in two independent populations.
- **Assumptions:**
  - Two independent random samples from two Bernoulli populations.
  - Expected successes and failures in each sample are at least 10.
- **Hypothesis Testing for the Difference in Proportions:**
  - Null Hypothesis (H_0): $p_1 - p_2 = 0$ (no difference in proportions)
  - Alternative Hypothesis (H_a):
    - $p_1 - p_2 \neq 0$ (two-sided)
    - $p_1 - p_2 > 0$ (right-tailed)
    - $p_1 - p_2 < 0$ (left-tailed)
  - Test Statistic: $z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}}$
  - P-value: Probability of observing a test statistic as extreme or more extreme than the one calculated, assuming the null hypothesis is true.
  - Conclusion:
    - If p-value ¡ α (significance level), reject H_0.
    - If p-value $\geq \alpha$, fail to reject H_0.
- **Confidence Interval for the Difference in Proportions:**
  - Formula: $(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$
  - Interpretation: We are [confidence level] confident that the true difference in population proportions is between [lower bound] and [upper bound].
- **Agresti-Coull Adjustment:**
  - Add 2 successes and 2 failures to each sample size to improve the accuracy of the confidence interval when sample sizes are small.

Other Considerations

- **Type I and Type II Errors:**
  - Type I Error: Rejecting the null hypothesis when it is actually true.
  - Type II Error: Failing to reject the null hypothesis when it is actually false.
  - The choice of significance level (α) affects the probability of making these errors.
- **Controlling Type I Errors:**
  - Use a smaller significance level (α).
  - Increase the sample size.
- **Practical Significance vs. Statistical Significance:**
  - Statistical significance: The result is unlikely to have occurred by chance.
  - Practical significance: The result has a meaningful impact in the real world.
  - A result can be statistically significant but not practically significant, and vice versa.

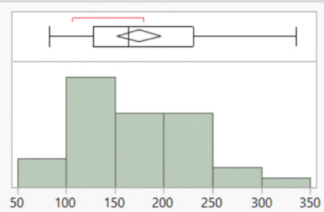


In all of these cases, if the sample sizes are large enough, then

$$\bar{X} \sim N\left(\mu_X, \frac{\sigma_X^2}{n_X}\right) \text{ (approximately)}$$
$$\bar{Y} \sim N\left(\mu_Y, \frac{\sigma_Y^2}{n_Y}\right) \text{ (approximately)}$$

and our methods for inference about $\mu_X - \mu_L$ will be *valid*, but the results might not be *meaningful*.

Sales Price



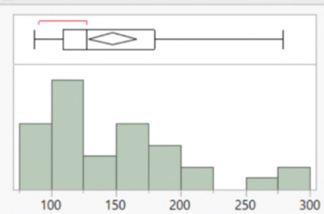
Quantiles

100.0%	maximum	335
99.5%		335
97.5%		335
90.0%		270
75.0%	quartile	230.125
50.0%	median	163
25.0%	quartile	127.175
10.0%		95.75
2.5%		82.5
0.5%		82.5
0.0%	minimum	82.5

Summary Statistics

Mean	174.82647
Std Dev	64.528529
Std Err Mean	11.066551
Upper 95% Mean	197.34154
Lower 95% Mean	152.3114
N	34

Assessed Value

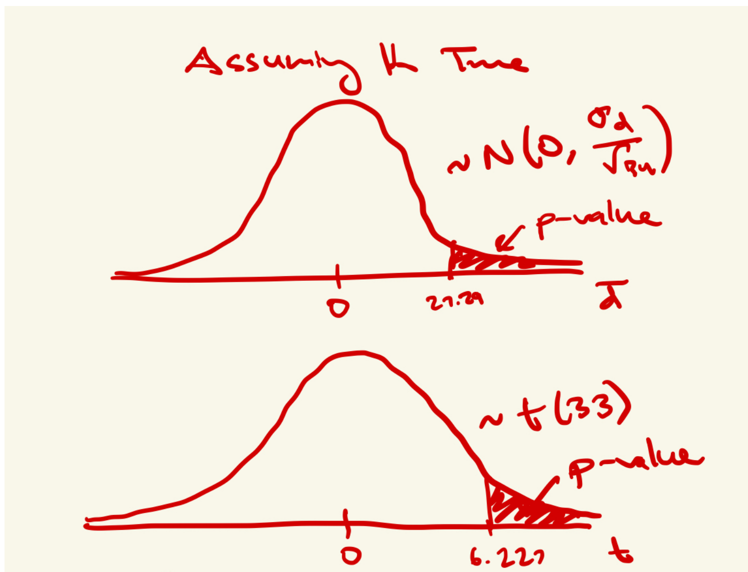


Quantiles

100.0%	maximum	279
99.5%		279
97.5%		279
90.0%		231.85
75.0%	quartile	179.9
50.0%	median	127.6
25.0%	quartile	108.6
10.0%		91.55
2.5%		87
0.5%		87
0.0%	minimum	87

Summary Statistics

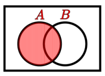
Mean	147.54118
Std Dev	53.287921
Std Err Mean	9.138803
Upper 95% Mean	166.13421
Lower 95% Mean	128.94814
N	34



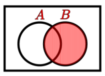
CI For	Sample Statistic	Margin of Error	Use When
Population mean (μ)	$\bar{x}$	$\pm z^* \dfrac{\sigma}{\sqrt{n}}$	X is normal, or $n \geq 30$; σ known
Population mean (μ)	$\bar{x}$	$\pm t^*_{n-1} \dfrac{s}{\sqrt{n}}$	$n < 30$, and/or σ unknown
Population proportion (p)	$\hat{p}$	$\pm z^* \sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}$	$n\hat{p},\ n(1-\hat{p}) \geq 10$
Difference of two population means ($\mu_1 - \mu_2$)	$\bar{x}_1 - \bar{x}_2$	$\pm z^* \sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}$	Both normal distributions or $n_1, n_2 \geq 30$; σ_1, σ_2 known
Difference of two population means $\mu_1 - \mu_2$	$\bar{x}_1 - \bar{x}_2$	$\pm t^*_{n_1+n_2-2} \sqrt{\dfrac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}}$	$n_1, n_2 < 30$; and/or $\sigma_1 = \sigma_2$ unknown
Difference of two proportions ($p_1 - p_2$)	$\hat{p}_1 - \hat{p}_2$	$\pm z^* \sqrt{\dfrac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \dfrac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$	$n\hat{p}, n(1-\hat{p}) \geq 10$ for each group

Probability Formulas

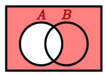
Ace Tutors



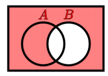
$P(A)$



$P(B)$

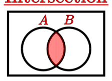


$P(A)^c$



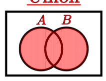
$P(B)^c$

Intersection



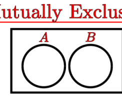
$P(A \cap B)$

Union



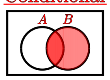
$P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Mutually Exclusive



$P(A \cap B) = 0$

Conditional



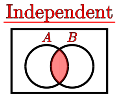
$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Bayes' Theorem

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

$$P(B|A) = \frac{P(A|B) \cdot P(B)}{P(A)}$$

Independent



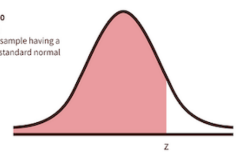
$$P(A \cap B) = P(A) \cdot P(B)$$
$$P(A|B) = P(A)$$

CI for	Sample Statistic	Margin of Error	Used when
Population mean (μ)	$\bar{x}$	$z^* \dfrac{\sigma}{\sqrt{n}}$	σ known; X is normal or $n \geq 30$
Population mean (μ)	$\bar{x}$	$t^*_{n-1} \dfrac{s}{\sqrt{n}}$	σ unknown; X is normal-looking and $n < 30$
Population proportion (p)	$\hat{p}$	$z^* \sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}$	$n\hat{p} \geq 10$ and $n(1-\hat{p}) \geq 10$
Difference of two population means ($\mu_1 - \mu_2$)	$\bar{x}_1 - \bar{x}_2$	$z^* \sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}$	σ_1, σ_2 known; X_1, X_2 both normal or $n_1, n_2 \geq 30$
Difference of two population means ($\mu_1 - \mu_2$)	$\bar{x}_1 - \bar{x}_2$	$t^*_{n_1+n_2-2} \sqrt{\dfrac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}}$	σ_1, σ_2 unknown; $\sigma_1 \neq \sigma_2$; X_1, X_2 normal-looking and $n_1, n_2 < 30$
Difference of two population means ($\mu_d = \mu_1 - \mu_2$)	$\bar{x}_1 - \bar{x}_2$	$t^*_{n-1} \dfrac{s_d}{\sqrt{n_d}}$	X_1, X_2 contain matched pairs; differences are normal-looking and $n_d < 30$
Difference of two proportions ($p_1 - p_2$)	$\hat{p}_1 - \hat{p}_2$	$z^* \sqrt{\dfrac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \dfrac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$	$n_1\hat{p}_1$, $n_1(1-\hat{p}_1)$, $n_2\hat{p}_2$, $n_2(1-\hat{p}_2)$ all greater than 10

NORM.S.DIST and NORMSDIST

Calculate the probability of having a value below x (percentile)

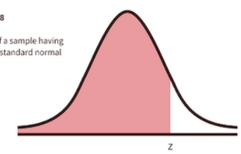
NORM.S.DIST (1.28) = 0.90
NORMSDIST (1.28) = 0.90
There's a 90% chance of a sample having a Z-score of 1.28 or less in a standard normal distribution



NORM.S.INV and NORMSINV

Calculate a Z-score based on the probability of having a value below x

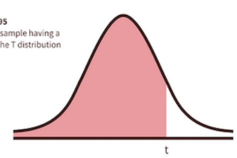
NORM.S.INV (0.90) = 1.28
NORMSINV (0.90) = 1.28
There's an 90% chance of a sample having a Z-score below 1.28 in a standard normal distribution



T.DIST(cumulative = true)

Calculates the probability of having a value below a certain T-value (percentile); not equivalent to TDIST

T.DIST (2.02, 5, true) = 0.95
There's a 95% chance of a sample having a T-score lower than 2.02 if the T distribution has 5 degrees of freedom



T.INV

Calculates a T-score based on the probability of having a value below that T-score (percentile); not equivalent to TINV

T.INV (0.95, 5) = 2.02
The T-score on the 95th percentile (95% of values are lower than it) with 5 degrees of freedom is 2.02