1 CI single sample

1.1 Scenario 1: CI single small sample

Let $x_1, x_2, ..., x_n$ be iid (independent and identically distributed). $N(\mu, \sigma^2)$ where both μ and σ are unknown and $n < 30$. Then a $100(1-\alpha)\%$ CI is given by:

$$(L, R) = \overline{x} \pm t_{(n-1), \frac{\alpha}{2}} * \frac{s}{\sqrt{n}}$$

Conditions Required for a Valid Small-Sample Confidence Interval for μ 1. A random sample is selected from the target population. 2. The population has a relative frequency distribution that is approximately normal.

1.2 Scenario 2: CI single small sample, σ known

Let $x_1, x_2, ..., x_n$ be iid (independent and identically distributed). $N(\mu, \sigma^2)$ where μ is unknown and $n < 30$. σ is known. Then a $100(1-\alpha)\%$ CI is given by: (Conditions: same as 1.1)

$$(L, R) = \overline{x} \pm Z_{\frac{\alpha}{2}} * \frac{\sigma}{\sqrt{n}}$$

1.3 Scenario 3: CI single sample large

Let $x_1, x_2, ..., x_n$ be iid (independent and identically distributed) with μ and σ unknown. Given $n \geq 30$: don't need to assume population is normal since (CLT: central limit theorem). Then a CI for μ of $100(1-\alpha)$ is given by:

$$(L, R) = \overline{x} \pm Z_{\frac{\alpha}{2}} * \frac{s}{\sqrt{n}}$$

Conditions Required for a Valid Large-Sample Confidence Interval for μ 1. A random sample is selected from the target population. 2. The sample size n is large (i.e., $n \geq 30$). (Due to the Central Limit Theorem, this condition guarantees that the sampling distribution of $\overline{x}$ is approximately normal. Also, for large n, s will be a good estimator of σ.)

1.4 Scenario 4: CI proportion single sample large

Let $x_1, x_2, ..., x_n$ be iid (independent and identically distributed) Bernoulli r.v. (i.e. with $P(X_i = 1) = p$ and $P(X_i = 0) = 1 - p$). (P is what you try to estimate). Suppose if P is unknown, then if n is large enough a $100(1-\alpha)\%$ CI for P is given by:

$$\overline{p} = \frac{\#\text{success in sample}}{n}$$

$$(L, R) = \overline{p} \pm Z_{\frac{\alpha}{2}} * \sqrt{\frac{\overline{p}(1-\overline{p})}{n}}$$

Note: For n to be large enough, the following condition must be satisfied:

$$n\overline{p} \geq 15 \quad \text{and} \quad n(1-\overline{p}) \geq 15$$

Conditions: Conditions Required for a Valid Large-Sample Confidence Interval for p 1. A random sample is selected from the target population. 2. The sample size n is large. (This condition will be satisfied if both $npn \geq 15$ and $nqn \geq 15$. Note that npn and nqn are simply the number of successes and number of failures, respectively, in the sample.

1.5 CI interpretation

Practical: We are are $x\%$ confident that μ, the mean [specify context] in the population is between $(x_1; x_2)$
Theoretical: To be more precise, if we were to do this study infinitely many times and each time a $x\%$ confident interval is constructed using the same technique as above, $x\%$ of theses intervals would include the true mean duration [Specify context]

2 Hypothesis ERRORS

Type 1 Error: We reject the H_0 (the null hypothesis) when it is in fact true. "A Type I error occurs if the researcher rejects the null hypothesis in favor of the alternative hypothesis when, in fact, H0 is true. The probability of committing a Type I error is denoted by α."
Type 2 Error: We reject the the H_a when in fact it is true. (i.e. we do not reject H_0 (keep it) when when it is in fact false). " A Type II error occurs if the researcher accepts the null hypothesis when, in fact, H0 is false. The probability of committing a Type II error is denoted by β."

3 Hypothesis Decision And conclusion

Decision: Since $1,93 > 1,74$ we reject H_0 in favour of H_a, at the $\alpha = x$
Since $1,93 < 1,74$ we do not reject H_0 in favour of H_a, at the $\alpha = x$

Conclusion: we have evidence to conclude that the true mean in the population [context] is [bigger, smaller, not the same] compared to [context] at the $\alpha = x$ level
We do not have enough evidence to reject the null hypothesis that the [context] true mean is μ_0 at the $\alpha = x$ level

4 Hypothesis Testing single sample

4.1 Scenario 1: hypothesis single small

Suppose that $x_1, x_2, ..., x_n$ is a random sample from a normal distribution with unknown μ and σ and $n < 30$. Given α then:

$$H_0 : \mu = \mu_0$$
$$T = \frac{\overline{x} - \mu_0}{s/\sqrt{n}}$$

T will have a t-distribution with $(n-1)$ degrees of freedom. The rejection region (RR) depends of H_a.

$H_0 : \mu = \mu_0 \quad H_a : \mu > \mu_0$
$RR = \{T > t_{(n-1), \alpha}\}$

$H_0 : \mu = \mu_0 \quad H_a : \mu < \mu_0$
$RR = \{T < -t_{(n-1), \alpha}\}$

$H_0 : \mu = \mu_0 \quad H_a : \mu \neq \mu_0$
$RR = \{T > t_{(n-1), \frac{\alpha}{2}} \quad \text{OR} \quad T < -t_{(n-1), \frac{\alpha}{2}}\}$

$RR = \{|T| > t_{(n-1), \frac{\alpha}{2}}\}$

Conditions Required for a Valid Small-Sample Hypothesis Test for μ 1. A random sample is selected from the target population. 2. The population from which the sample is selected has a distribution that is approximately normal.

4.2 Scenario 2:hypothesis single large

Suppose that $x_1, x_2, ..., x_n$ in a random sample (iid) with unknown μ and σ and $n \geq 30$. Given α (conditions same as 1.3):

$$H_0 : \mu = \mu_0$$
$$T = \frac{\overline{x} - \mu_0}{s/\sqrt{n}}$$

Since $n \geq 30$, by central limit theorem. T is approximately normal:

$H_0 : \mu = \mu_0 \quad H_a : \mu > \mu_0$
$RR = \{T > Z_\alpha\}$

$H_0 : \mu = \mu_0 \quad H_a : \mu < \mu_0$
$RR = \{T < -Z_\alpha\}$

$H_0 : \mu = \mu_0 \quad H_a : \mu \neq \mu_0$
$RR = \{T > Z_{\frac{\alpha}{2}} \quad \text{OR} \quad T < -Z_{\frac{\alpha}{2}}\}$
$RR = \{|T| > Z_{\frac{\alpha}{2}}\}$

4.3 Scenario 3: hypothesis single large proportions

Let $x_1, x_2, ..., x_n$ be a random sample (iid) of Bernoulli r.v with unknown p (probability of success), where n is large enough [i.e. $n\overline{p} \geq 15$ and $n(1-\overline{p}) \geq 15$]. Given α:

$H_0 : P = P_0$

$$\overline{p} = \frac{\#\text{of success in sample}}{n}$$

$$T = \frac{\overline{p} - p_0}{\sqrt{\frac{P_0(1-P_0)}{n}}}$$

$H_0 : p = p_0 \quad H_a : p > p_0$
$RR = \{T > Z_\alpha\}$

$H_0 : p = p_0 \quad H_a : p < p_0$
$RR = \{T < -Z_\alpha\}$

$H_0 : p = p_0 \quad H_a : p \neq p_0$
$RR = \{T > Z_{\frac{\alpha}{2}} \quad \text{OR} \quad T < -Z_{\frac{\alpha}{2}}\}$
$RR = \{|T| > Z_{\frac{\alpha}{2}}\}$

Conditions Required for a Valid Large-Sample Hypothesis Test for p 1. A random sample is selected from a binomial population. 2. The sample size n is large. (This condition will be satisfied if both $np \geq 15$ and $nq \geq 15$.)

5 Two Sample Problems

5.1 Scenario 1: TWO INDP; SMALL

Let $x_1, x_2, ..., x_n$ be a random sample from a normal distribution with unknowns μ_1 and σ_1. Let $y_1, y_2, ..., y_n$ be a random sample from a normal distribution with unknowns μ_2 and σ_2. If $n < 30$ and $m < 30$. We assume that both samples are normally distributed and are independent of one another. Further suppose that $\sigma_1 = \sigma_2$. A $100(1-\alpha)\%$ is given by:

$$(L, R) = (\overline{x}_1 - \overline{x}_2) \pm t_{(m+n-2), \frac{\alpha}{2}} * S_p * \sqrt{\frac{1}{m} + \frac{1}{n}}$$

$$S_p = \sqrt{\frac{(m-1)s_1{}^2 + (n-1)s_2{}^2}{m+n-2}}$$

$H_0 : \mu_1 - \mu_2 = 0 \quad \text{OR} \quad H_0 : \mu_1 = \mu_2$

$$T = \frac{\overline{x}_1 - \overline{x}_2}{s_p * \sqrt{\frac{1}{m} + \frac{1}{n}}}$$

$H_a : \mu_1 - \mu_2 > 0$
$RR = \{T \geq t_{(m+n-2), \alpha}\}$

$H_a : \mu_1 - \mu_2 < 0$
$RR = \{T \leq -t_{(m+n-2), \alpha}\}$

$H_a : \mu_1 - \mu_2 > 0$
$RR = \{T \geq t_{(m+n-2), \frac{\alpha}{2}} \quad \text{OR} \quad T \leq -t_{(m+n-2), \frac{\alpha}{2}}\}$

$RR = \{|T| \geq t_{(m+n-2), \frac{\alpha}{2}}\}$

Conditions Required for Valid Small-Sample Inferences about $\mu_1 - \mu_2$ 1. The two samples are randomly selected in a independent manner from the two target populations. 2. Both sampled populations have distributions that are approximately normal. 3. The population variances are equal (i.e., $\sigma_1^2 = \sigma_2^2$).

5.2 Scenario 2: TWO INDP LARGE

Suppose $x_1, x_2, ..., x_n$ with unknown μ_1 and σ_1 and $y_1, y_2, ..., y_n$ with unknown μ_2 and σ_2. Furthermore, if $m \geq 30$ and $n \geq 30$ and the x's are independent of the y's. A $100(1-\alpha)\%$ CI for $(\mu_1 - \mu_2)$ is given by:

$$(L, R) = (\overline{x}_1 - \overline{x}_2) \pm Z_{\frac{\alpha}{2}} * \sqrt{\frac{s_1{}^2}{m} + \frac{s_2{}^2}{n}}$$

$H_0 : \mu_1 - \mu_2 = 0$

$$T = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{s_1{}^2}{m} + \frac{s_1{}^2}{n}}}$$

$H_0 : \mu_1 - \mu_2 = 0 \quad H_a : \mu_1 - \mu_2 > 0$
$RR = \{T > Z_\alpha\}$

$H_0 : \mu_1 - \mu_2 = 0 \quad H_a : \mu_1 - \mu_2 < 0$
$RR = \{T < -Z_\alpha\}$

$H_0 : \mu_1 - \mu_2 = 0 \quad H_a : \mu_1 - \mu_2 \neq 0$
$RR = \{T > Z_{\frac{\alpha}{2}} \quad \text{OR} \quad T < -Z_{\frac{\alpha}{2}}\}$
$RR = \{|T| > Z_{\frac{\alpha}{2}}\}$

Conditions Required for Valid Large-Sample Inferences about $\mu_1 - \mu_2$ 1. The two samples are randomly selected in an independent manner from the two target populations. 2. The sample sizes,n_1 and n_2 ,are both large(i.e.,$n_1 \geq 30$ and $n_2 \geq 30$). (By the central limit theorem, this condition guarantees that the sampling distribution of ($\overline{x}_1$ and $\overline{x}_2$) will be approximately normal, regardless of the shapes of the underlying probability distributions of the populations. Also s_1^2 and s_2^2 will provide good approximations to σ_1^2 and σ_2^2)

5.3 Scenario 3: TWO PAIR DEP SMALL

Let $x_1, x_2, ..., x_n$ and $y_1, y_2, ..., y_n$ be paired or matched observations from distributions with unknown μ_1 and μ_2 respectively, and $n < 30$. Let $D_i = x_i - y_i$ $(i = 1, 2, ...)$. So $D_1, D_2, ...D_n$ is a random sample from a normal distribution, with unknown $\mu_d = \mu_1 - \mu_2$ and variance $\sigma_d{}^2$ (variance of population of differences). A $100(1-\alpha)\%$ CI for μ_d is given by:

$$(L, R) = \overline{D} \pm t_{(n-1), \frac{\alpha}{2}} * \frac{s_d}{\sqrt{n}}$$

$H_0 : \mu_1 = \mu_2 \quad \text{OR} \quad H_0 : \mu_d = 0$

$$T = \frac{\overline{D}}{s_d/\sqrt{n}}$$

$H_0 : \mu_d = 0 \quad H_a : \mu_d > 0$
$RR = \{T > t_{(n-1), \alpha}\}$

$H_0 : \mu_d = 0 \quad H_a : \mu_d < 0$
$RR = \{T < -t_{(n-1), \alpha}\}$

$H_0 : \mu_d = 0 \quad H_a : \mu_d \neq 0$
$RR = \{T > t_{(n-1), \frac{\alpha}{2}} \quad \text{OR} \quad T < -t_{(n-1), \frac{\alpha}{2}}\}$

$RR = \{|T| > t_{(n-1), \frac{\alpha}{2}}\}$

Conditions Required for Valid Small-Sample Inferences about μ_d 1. A random sample of differences is selected from the target population of differences. 2. The population of differences has a distribution that is approximately normal.

5.4 Scenario 4: TWO PAIR DEP LARGE

Let $x_1, x_2, ..., x_n$ and $y_1, y_2, ..., y_n$ be paired or matched observations from distributions with unknown μ_1 and μ_2 respectively, and $n \geq 30$. Let $D_i = x_i - y_i$ $(i = 1, 2, ...)$. So $D_1, D_2, ...D_n$ is a random sample with unknown $\mu_d = \mu_1 - \mu_2$ and SD σ_d (Normality not required). A $100(1-\alpha)\%$ CI for $\mu_d = \mu_1 - \mu_2$ is given by:

$$(L, R) = \overline{D} \pm Z_{\frac{\alpha}{2}} * \frac{s_d}{\sqrt{n}}$$

$H_0 : \mu_1 = \mu_2 \quad \text{OR} \quad H_0 : \mu_d = 0$

$$T = \frac{\overline{D}}{s_d/\sqrt{n}}$$

$H_a : \mu_1 > \mu_2 \quad H_a : \mu_d > 0$
$RR = \{T > Z_\alpha\}$

$H_a : \mu_1 < \mu_2 \quad H_a : \mu_d < 0$
$RR = \{T < -Z_\alpha\}$

$H_a : \mu_1 \neq \mu_2 \quad H_a : \mu_d \neq 0$
$RR = \{T > Z_{\frac{\alpha}{2}} \quad \text{OR} \quad T < -Z_{\frac{\alpha}{2}}\}$
$RR = \{|T| > Z_{\frac{\alpha}{2}}\}$

Conditions Required for Valid Large-Sample Inferences about μ_d 1. A random sample of differences is selected from the target population of differences. 2. The sample size nd is large (i.e., $nd \geq 30$).(by the CLT...)

5.5 Scenario 5: TWO PROPORTION LARGE INDP

Let $x_1, x_2, ..., x_m$ be random sample of bernoulli random variable with unknown probability of success p_1 and let $y_1, y_2, ..., y_n$ be a random sample of bernoulli r.v. with unknown probability of success p_2. Further, suppose that the X_i's are independent of the y_i's and that both sample sizes are large enough:$n\overline{p}_1 \geq 15$ and $n(1-\overline{p}_1) \geq 15$ and $n\overline{p}_2 \geq 15$ and $n(1-\overline{p}_2) \geq 15$. A $100(1-\alpha)\%$ CI for $(p_1 - p_2)$ is given by:

$$(L, R) = (\overline{P}_1 - \overline{P}_2) \pm Z_{\frac{\alpha}{2}} \sqrt{\frac{\overline{P}_1(1-\overline{P}_1)}{m} + \frac{\overline{P}_2(1-\overline{P}_2)}{n}}$$

$H_0 : P_1 = P_2 \quad \text{OR} \quad H_0 : P_1 - P_2 = 0$

$$T = \frac{\overline{P}_1 - \overline{P}_2}{\sqrt{\overline{P}(1-\overline{P})[\frac{1}{m} + \frac{1}{n}]}}$$

$\overline{P} = \frac{X+Y}{m+n}$ X,Y success in populations

$H_0 : P_1 = P_2 \quad H_a : p_1 > p_2$
$RR = \{T > Z_\alpha\}$

$H_0 : P_1 = P_2 \quad H_a : p_1 < p_2$
$RR = \{T < -Z_\alpha\}$

$H_0 : P_1 = P_2 \quad H_a : p_1 \neq p_2$
$RR = \{T > Z_{\frac{\alpha}{2}} \quad \text{OR} \quad T < -Z_{\frac{\alpha}{2}}\}$
$RR = \{|T| > Z_{\frac{\alpha}{2}}\}$

Conditions Required for Valid Large-Sample Inferences about $p_1 - p_2$ 1. The two samples are randomly selected in an independent manner from the two target populations. 2. The sample sizes, n_1 and n_2, are both large, so the sampling distribution of $(\overline{p}_1 - \overline{p}_2)$ will be approximately normal. (cond. be satisfied if $\geq$)

6 P-Values

The observed significance level, or p-value, for a specific statistical test is the probability (assuming H0 is true) of observing a value of the test statistic that is at least as contradictory to the null hypothesis, and supportive of the alternative hypothesis, as the actual one computed from the sample data.

$\underline{H_a : \mu > \mu_0}$
$p = p(z \geq t_{obs})$
$p = p(t_\nu \geq t_{obs})$
$\underline{H_a : \mu < \mu_0}$
$p = p(z \leq -t_{obs})$
$p = p(t_\nu \leq -t_{obs})$
$\underline{H_a : \mu \neq \mu_0}$
$p = 2 * p(z \geq |t_{obs}|)$
$p = 2 * p(t_\nu \geq |t_{obs}|)$

if $p < \alpha$ we reject H_0. If $p > \alpha$ we do not reject H_0. Interpretation:

7 Discrete Distributions

7.1 Bernoulli Distribution

A random variable X is said to have a bernoulli distribution with paramater p ($0 \leq p \leq 1$) if ($P(x = 1) = p$ and $P(x = 0) = (1 - p)$).

$$E(x) = p$$
$$VAR(x) = p(1-p)$$
$$SD(x) = \sqrt{p(1-p)}$$

7.2 Binomial Setup

Characteristics of a Binomial Random Variable 1. The experiment consists of n identical trials. 2. There are only two possible outcomes on each trial. We will denote one outcome by S (for Success) and the other by F (for Failure). 3. The probability of S remains the same from trial to trial. This probability is denoted by p, and the probability of F is denoted by $q = 1 - p$. 4. The trials are independent. 5. The binomial random variable x is the number of S's in n trials.

$$p(X = x) = \binom{n}{x} * p^x * q^{n-x}$$

p is prob a success in one trial; q is $(1-p)$; n is number of trials; x is number of success in n trials.

$$E(x) = np$$
$$VAR(x) = np(1-p)$$
$$SD(x) = \sqrt{np(1-p)}$$

Interpretation of $E(x)$: we expect that on average $expectationValue$ [context]