Confidence Intervals for a Population Mean

Let $X_1, \ldots, X_n$ be a simple random sample from a population with mean μ and variance σ^2. Let $\bar{X}$ be the sample mean, and S_n be the sum of sample observation. If n **is sufficiently large**,

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

and

$$S_n \sim N\left(n\mu, n\sigma^2\right)$$

Let $X_1, \ldots, X_n$ be a **large** $(n > 30)$ random sample from a population with mean μ and standard deviation σ, so that $\bar{X}$ is approximately normal. Then a level $100(1-\alpha)\%$ confidence interval for μ is

$$\bar{X} \pm z_{\frac{\alpha}{2}} \sigma_{\bar{X}}$$

where $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$. When the value of σ is unknown, it can be replaced with the sample standard deviation s.

Small Sample Confidence Intervals for a Population Mean

Let $X_1, \ldots, X_n$ be a small $(n < 30)$ sample from a *normal* population with mean μ. Then the quantity

$$\frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}}$$

has a Student's t distribution with $n-1$ degrees of freedom, denoted t_{n-1}. When n is large, the distribution of quantity $\frac{\bar{X}-\mu}{\frac{s}{\sqrt{n}}}$ is very close to normal, so the normal curve can be used, rather than the Student's t.

Hypothesis Testing

Large-Sample Tests for a Population Mean

Let $X_1, \ldots, X_n$ be a **large** $(n > 30)$ sample form a population with mean μ and standard deviation σ.

To test a null hypothesis of the form $H_0 : u \leq u_0, H_0 : u \geq \mu_0, H_0 := \mu_0$:

- Compute the z-score:

$$z = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

If σ is unknown it may be approximated with s.

- Computer the P-value. The P-value is an area under the normal curve, which depends on the alternate hypothesis as follows:

Alternate Hypothesis	P-Value
$H_1 : \mu > \mu_0$	Area to the right of z
$H_1 : \mu < \mu_0$	Area to the left of z
$H_1 : \mu = \mu_0$	Sum of the areas cut off by z and $-z$

Drawing Conclusions from the Results of Hypothesis Tests

Let α be any value between 0 and 1. Then, if $P \leq \alpha$,

- The result of the test is said to be statistically significant at the $100\alpha\%$ level.
- The null hypothesis is rejected at the $100\alpha\%$ level.
- When reporting the result of the hypothesis test, report the P-value, rather than just comparing it to the 5% or 1%.

Small-Sample Tests for a Population Mean

Let $X_1, \ldots, X_n$ be a **small** $(n \leq 30)$ random sample from a *normal* population with mean μ (unknown) and a standard deviation σ.
To test a null hypothesis of the form $H_0 : \mu \leq \mu_0$, $H_0 : \mu \geq \mu_0$, or $H_0 : \mu = \mu_0$:

- Compute the test statistic

$$t^* = \frac{\bar{X} - \mu_0}{\frac{s}{\sqrt{n}}}$$

- Compute the P-value. The P-value is an area under the Student's t curve with $n-1$ degrees of freedom, which depends on the alternate hypothesis as follows

Alternate Hypothesis	P-Value
$H_1 : \mu > \mu_0$	Area to the right of z
$H_1 : \mu < \mu_0$	Area to the left of z
$H_1 : \mu = \mu_0$	Sum of the areas cut off by z and $-z$

- If σ is known, the test statistic is $z = \frac{\bar{X}-\mu_0}{\frac{\sigma}{\sqrt{n}}}$

Large-Sample Tests for the Difference Between Two Means

Let $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$ be **large** ($n_x > 30$ and $n_y > 30$) independent random samples from populations with mean u_x and u_y and standard deviation σ_x and σ_y, respectively.
The test statistic is as follows:

$$z^* = \frac{(\bar{X} - \bar{Y}) - \delta_0}{\sqrt{\frac{\sigma_X^2}{n_x} + \frac{\sigma_Y^2}{n_y}}}$$

If σ_X and σ_Y are unknown they may be replaced by s_X and s_Y, respectively

Null Hypothesis	Alternative Hypothesis	p-value
$H_0 : \mu_x - \mu_y \leq \delta_0$	$H_1 : \mu_x - \mu_y > \delta_0$	$P(Z \geq z^*)$
$H_0 : \mu_x - \mu_y \geq \delta_0$	$H_1 : \mu_x - \mu_y < \delta_0$	$P(Z \leq z^*)$
$H_0 : \mu_x - \mu_y = \delta_0$	$H_1 : \mu_x - \mu_y \neq \delta_0$	$2 \times P(Z \geq \lvert z^* \rvert)$

Small-Sample Tests for the Difference Between Two Means

Population Variances Are Not Equal

Let $X_1, \ldots, X_{n,x}$ and $Y_1, \ldots, Y_{n,y}$ be samples from *normal* populations with means μ_X and μ_Y and standard deviations σ_X and σ_Y, respectively.
Assume the samples are drawn independently of each other.
If σ_x and σ_Y are not known to be equal, then, to test a null hypothesis of the form $H_0 : \mu_X - \mu_Y \leq \Delta_0, H_0 : \mu_X - \mu_Y \geq \Delta_0$, or $H_0 : \mu_X - \mu_Y = \Delta_0$.

- Rounding down to the nearest integer, calculate

$$\nu = \frac{\left[\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}\right]^2}{\frac{\left(\frac{s_X^2}{n_X}\right)^2}{n_X - 1} + \frac{\left(\frac{s_Y^2}{n_Y}\right)^2}{n_Y - 1}}$$

- Compute the test statistic

$$t = \frac{(\bar{X} - \bar{Y}) - \Delta_0}{\sqrt{S_X^2/n_x + S_Y^2/n_Y}}$$

- Compute the P-value. The P-value is an area under the Student's t curve with v degrees of freedom, which depends on the alternate hypothesis as follows:

Alternate Hypothesis	P-value
$H_1 : \mu_X - \mu_Y > \Delta_0$	Area to the right of t
$H_1 : \mu_X - \mu_Y < \Delta_0$	Area to the left of t
$H_1 : \mu_X - \mu_Y \neq \Delta_0$	Sum of the areas in the tails cut off

Population Variances Are Equal

Let $X_1, \ldots, X_{n,x}$ and $Y_1, \ldots, Y_{n,y}$ be samples from *normal* populations with means μ_X and μ_Y and standard deviations σ_X and σ_Y, respectively.
Assume the samples are drawn independently of each other.
If σ_X and σ_Y are known to be qual, then, to test a null hypothesis of the fortm $H_0 : \mu_X - \mu_Y \leq \Delta_0$, $H_0 : \mu_X - \mu_Y \geq \Delta_0$, or $H_0 : \mu_x - \mu_y = \Delta_0$:

- Compute

$$s_p = \sqrt{\frac{(n_X - 1)s_X^2 + (n_Y - 1)s_Y^2}{n_X + n_Y - 2}}$$

- Compute the test statistic

$$t = \frac{(\bar{X} - \bar{Y}) - \Delta_0}{s_p \sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}}$$

- Compue the P-value. The P-value is an area under the Student's t curve with $n_X + n_Y - 2$ degrees of freedom, which depends on the alternate hypothesis as follows.

Alternate Hypothesis	P-value
$H_1 : \mu_X - \mu_Y > \Delta_0$	Area to the right of t
$H_1 : \mu_X - \mu_Y < \Delta_0$	Area to the left of t
$H_1 : \mu_X - \mu_Y \neq \Delta_0$	Sum of the areas in the tails cut off

Correlation vs. Causation

Correlation

A correlation coefficient (denoted r) deasures the strength and direction of a linear relationship between two variables. Let (x_1, y_1), …, (x_n, y_n) represent bivariate data, then the correlation coefficient is

$$\begin{aligned} r &= \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{x_i - \bar{x}}{s_x}\right) \\ &= \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sqrt{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}\sqrt{\sum_{i=1}^{n} y_i^2 - n\bar{y}^2}} \qquad = \frac{SSR}{SST} \end{aligned}$$

The Least-Squares Line

For an equation of the form

$$y_1 = \beta_0 + \beta_1 x_i + \epsilon_i$$

y_i is called the **dependent variable**, x_i is called the **indepedent variable**, β is called the **regression coefficients** (the least squares coefficients), and ϵ_i is called the **error**.
Also, r^2 is the **proportion of variance in y explained by regression**.

$$\begin{aligned} e_1 &= y_i - \widehat{y}_i = y_i - \widehat{\beta}_0 - \widehat{B}_1 x_i \\ \widehat{\beta}_1 &= \widehat{\beta}_1 = \frac{\sum_i^n y_i - n\bar{x}\bar{y}}{\sum_i^n x_i^2 - n\bar{x}^2} \\ \widehat{\beta}_0 &= \bar{y} - \widehat{\beta}_1 \bar{x} \end{aligned}$$

Uncertainties in the Least-Squares Coefficients

Using some assumptions,

- The quantity $\widehat{\beta}$ is *normally distributed* random variables.
- The means of $\widehat{\beta}$ is the true values of $\widehat{\beta}$.
- The *standard deviations* of β is estimated with

$$s_{\beta_0} = s\sqrt{\frac{1}{n} + \frac{\bar{x}}{2}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$s_{\beta_1} = \frac{s}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

where

$$s = \sqrt{\frac{(1-r^2)\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-2}}$$

is an estimate of the error standard deviation σ.

Confidence Intervals for Coefficients

Under assumptions, the quantities $\frac{\widehat{\beta_1} - \beta_1}{s_{\beta_1}}$ and $\frac{\widehat{\beta_1} - \beta_1}{s_{\beta_1}}$ have Student's t distributions with $n-2$ degrees of freedom.
Level $100(1-\alpha)\%$ confidence intervals for β_0 and β_1 are given by

$$\bar{\beta_0} \pm t_{n-2} \times s_{\bar{\beta_0}} \qquad \bar{\beta_1} \pm t_{n-2} \times s_{\bar{\beta_1}}$$

Level $100(1-\alpha)\%$ confidence intervals for the quantity $\beta_0 + \beta_1 x$ is given by

$$\widehat{\beta_0} + \widehat{\beta_1} x \pm t_{n-2,\alpha/2} \times s_{\widehat{y}}$$

where

$$s_{\widehat{y}=s\sqrt{\frac{1}{n}+\frac{(x-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}}$$

Checking Assumptions

If the plot of residuals versus fitted values

- Shjows no substantial trend or curve, and
- Is **homoscedastic**, that is, the vertical spread does not varu too much along the horizontal length of the plot, except perhaps near the edges,

then it is *likely*, but not certain, that the assumptions of the linear model hold.

However, if the residual plot *does* show a substantial trend or curve, or is **heteroscedastic**, it is certain that the assumptions of the linear plot *do not* hold.

Miscellaneous Notes

- A **test statistic** is a function of the sample data whose value is used to test a hypothesis
- A **p-value** is a measure of the disagreement between a sample and H_0.
- The smaller the P-value, the more certain we can be that H_0 is false and vice versa.
- For **large samples**, we approximate the population standad deviation σ using the sample standard deviation s.
- The correlation coefficient is called the **sample correlation (r)**, and the it is an estimate of the population correleation (ρ).
- Some properties of the correlation coeffciient (r):

1. $-1 \leq r \leq 1$, r is unitless.
2. If the points lie exactly on a horizontal or vertical line, the correlation coefficient is undefined, because one of the standard deviations is equal to zero.
3. Whenever $r \neq 0$, x and y are said to be correlated. If $r = 0$, x and y are said to be uncorrelated.
4. Correlation coefficient is unaffected by the units in whicht he measurements are made.

- For **small samples**, s may be far from σ, which invalidates this large-sample method. However, when the population is approximately normal, the Student's t distribution can be used.
- The **pooled** sample variance is

$$s_p^2 = \frac{(n_X - 1)s_X^2 + (n_Y - 1)S_Y^2}{n_X + n_Y - 2}$$

- The correleation coefficient remains unchanged under each of the following operations

  - Multiplying each value of a variable by a positive constant.
  - Adding a constant to each of a variable.
  - Interchanging the values of x and y.

- A **goodness-of-fit statistic** measures how well a model explains a given set of data.
- SST = total sum of squares $= \sum_{i=1}^{n}(y_i - \bar{y})^2$
- SSE = error sum of squares $= \sum_{i=1}^{n}(y_i - \widehat{y})^2$
- SSR = regression sum of squares $= SST - SSE$
- The following assumptions are satisfied.

  - The errors $\epsilon_1, \ldots, \epsilon_0$ are random and independent. In particular, the magnitude of any error ϵ_i does not influence the value of the next error ϵ_{i+1}.
  - The errors $\epsilon_1, \ldots, \epsilon_0$ all have mean 0.
  - The errors $\epsilon_1, \ldots, \epsilon_0$ all have the same variance, which we denote by σ^2.
  - The errors $\epsilon_1, \ldots, \epsilon_0$ are normally distributed.

- Margin of error $= t_{\frac{\alpha}{2},\sqrt{n}} \times \frac{\sigma}{\sqrt{n}}$
- The sample variance is calculated

$$\frac{1}{N-1}\sum_{i=0}^{n}(x - \bar{x})^2$$