BME 3005
BIOSTATISTICS

Lecture 11: Confidence Interval, Regression Line, Correlation Coefficient

Burcu Tunç Çamlıbel

Chapter 7 Confidence Intervals

BAU
BAHÇEŞEHİR ÜNİVERSİTESİ

Confidence Intervals

- This approach yields exactly the same conclusions as the procedures we discussed earlier because it simply represents a different perspective on how to use concepts like the standard error, *t*, and normal distributions.

- Confidence intervals are also used to estimate the range of values that include a specified proportion of all members of a population, such as the "normal range" of values for a laboratory test.

	A	B
$\bar{X}$	10.3	15.3
SD	3.1	2.1

Confidence Intervals: THE SIZE OF THE TREATMENT EFFECT MEASURED AS THE DIFFERENCE OF TWO MEANS

$$(\bar{X}_1 - \bar{X}_2) - t_\alpha s_{\bar{X}_1-\bar{X}_2} < \mu_1 - \mu_2 < (\bar{X}_1 - \bar{X}_2) + t_\alpha s_{\bar{X}_1-\bar{X}_2}$$

$1 - t_\alpha \cdot 0.169 < \mu_1 - \mu_2 <$

$1 + t_\alpha \cdot 0.169$

- the 95% confidence interval for the true difference of the population means is $(\bar{X}_1 - \bar{X}_2) - t_{.05} s_{\bar{X}_1-\bar{X}_2} < \mu_1 - \mu_2 < (\bar{X}_1 - \bar{X}_2) + t_{.05} s_{\bar{X}_1-\bar{X}_2}$

- This equation defines the range that will include the true difference in the means for 95% of all possible experiments that involve drawing samples from the two populations under study.
- Since this procedure to compute the confidence interval for the difference of two means uses the t distribution, it is subject to the same limitations as the t test.
- In particular, the samples must be drawn from populations that follow a normal distribution at least approximately.*

Confidence Intervals:CONFIDENCE INTERVAL FOR THE POPULATION MEAN

- The procedure we developed above can be used to compute a confidence interval for the mean of the population from which a sample was drawn.
- We can compute the 100 (1 − α) percent confidence interval for the population mean by obtaining the value of tα corresponding to v = n − 1 degrees of freedom, in which is the sample size.
- Substitute this value for t in the equation and solve for μ (just as we did for μ1 − μ2 earlier).

$$\overline{X} - t_{\alpha} s_{\overline{X}} < \mu < \overline{X} + t_{\alpha} s_{\overline{X}}$$

Example

OX 7-1 • The Effect on Temperature of Using Polyethylene Bags to Keep Extreme ow Birth Weight Infants Warm

The skin temperature for the 70 infants wrapped in polyethylene bags was 36°C with a standard deviation of 1°C and 35°C with a standard deviation of 1°C for the 70 infants kept warm using traditional methods. To compute the 95% confidence interval for the difference in temperature, we first compute the observed mean difference in temperature

1

$$\bar{X}_{bag} - \bar{X}_{trad} = 36 - 35 = 1°C$$

P.E Bags 36°

Traditional 35°C

−1°C

and the standard error of the difference

3

$$s_{\bar{X}_{bag} - \bar{X}_{trad}} = \sqrt{\frac{s^2}{n_{bag}} + \frac{s^2}{n_{trad}}} = \sqrt{\frac{1^2}{70} + \frac{1^2}{70}} = .169°C$$

Text

−1°C

because

2

$$s^2 = \frac{(n_{bag} - 1)s^2_{bag} + (n_{trad} - 1)s^2_{trad}}{n_{bag} + n_{trad} - 2} = \frac{(70-1)1^2 + (70-1)1^2}{70 + 70 - 2} = 1°C$$

pooled variance

s^2

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s^2}{n} + \frac{s^2}{n}}}$$

There are $\nu = n_{bag} + n_{trad} - 2 = 70 + 70 - 2 = 138$ degrees of freedom associated with this estimate. From Table 4-1 the critical value of t that defines the 5% most extreme values of the t distribution for 138 degrees of freedom is 1.977, so the 95% confidence interval for the difference in temperature is

table 4.1

$$1 - 1.977 \cdot .169 < \mu_{bag} - \mu_{trad} < 1 + 1.977 \cdot .169$$

$$.67°C < \mu_{bag} - \mu_{trad} < 1.33°C$$

99% P=0.01

Because the 95% confidence interval does not include 0, we can reject the null hypothesis that the wrapping technique did not affect the infants' temperature ($P < .05$).

From Table 4-1, the critical value of t that defines the 1% most extreme values of the t distribution is 2.611, so the 99% confidence interval for the difference in temperature is

$$1 - 2.611 \cdot .169 < \mu_{bag} - \mu_{trad} < 1 + 2.611 \cdot .169$$

$$.54°C < \mu_{bag} - \mu_{trad} < 1.44°C$$

Because the 99% confidence interval also excludes 0, we can also reject the null hypothesis with $P < .01$. (Compare this result with Prob. 4-2.)

Confidence Intervals:THE SIZE OF THE TREATMENT EFFECT MEASURED AS THE DIFFERENCE OF TWO RATES OR PROPORTIONS

- If p1 and p2 are the actual proportions of members of each of the two populations with the attribute, and if the corresponding estimates computed from the samples are ˆp1 and ˆp2 , respectively,

$$(\hat{p}_1 - \hat{p}_2) - z_\alpha s_{\hat{p}_1 - \hat{p}_2} < p_1 - p_2 < (\hat{p}_1 - \hat{p}_2) + z_\alpha s_{\hat{p}_1 - \hat{p}_2}$$

$z\alpha$ is the value that defines the most extreme α proportion of the values in the normal distribution;* $z\alpha = z.05 = 1.960$ is commonly used, since is used to define the 95% confidence interval.

Page 133 (7th edition) (try to solve the example)

Confidence Intervals:FOR RELATIVE RISK AND ODDS RATIO

$$e^{\ln RR - z_{\alpha} s_{\ln RR}} < RR_{true} < e^{\ln RR + z_{\alpha} s_{\ln RR}}$$

$$e^{\ln OR - z_{\alpha} s_{\ln OR}} < OR_{true} < e^{\ln OR + z_{\alpha} s_{\ln OR}}$$

- Try to solve examples (pages 139-140)

Chapter 8 How to Test for Trends

The Best Straight Line through the Data: Regression Line

- Example (page 149-150 and 159)
- Variability about the regregssion line **excluded**.
- You can find the related formulas from the example.

The resulting line is called the *regression line* of y on x (in this case, the regression line of weight on height). Its equation is

$$\hat{y} = a + bx$$

$\hat{y}$ denotes the value of y on the regression for a given value of x. This notation distinguishes it from the observed value of the dependent variable Y. The intercept a is given by

$$a = \frac{(\sum Y)(\sum X^2) - (\sum X)(\sum XY)}{n(\sum X^2) - (\sum X)^2}$$

and the slope is given by

$$b = \frac{n(\sum XY) - (\sum X)(\sum Y)}{n(\sum X^2) - (\sum X)^2}$$

in which X and Y are the coordinates of the n points in the sample.[†]

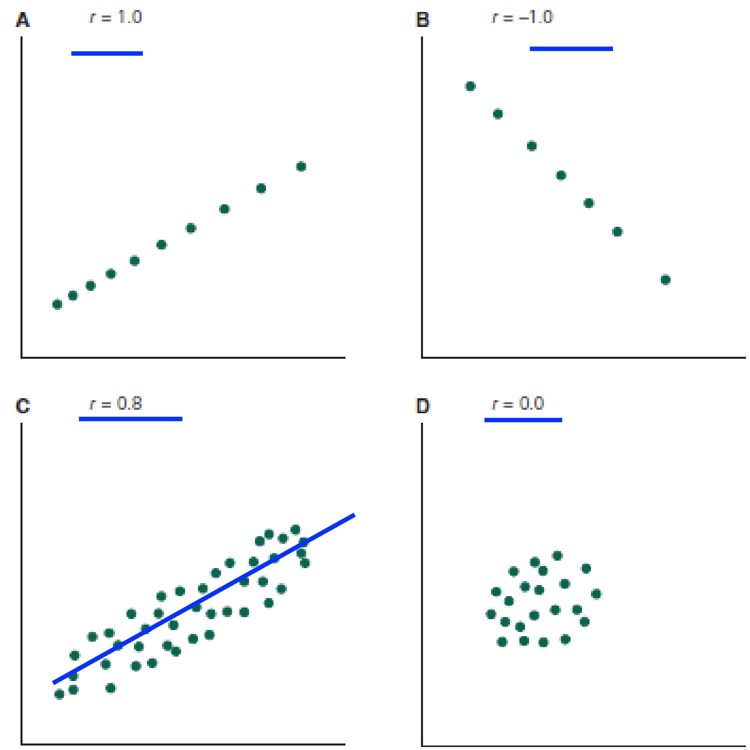
$$s_{y\cdot x} = \cancel{\sqrt{\frac{\sum(Y-\hat{y})^2}{n-2}}} = \sqrt{\frac{\sum[Y-(a+bX)]^2}{n-2}}$$

$$\cancel{s_{y\cdot x} = \sqrt{\frac{n-1}{n-2}(s_Y^2 - b^2 s_X^2)}}$$

Correlation and Correlation Coefficients

- The *correlation coefficient*, a number between −1 and +1, is often used to quantify the strength of this association.

- Figure 8-12 shows that the tighter the relationship between the two variables, the closer the magnitude of *r* is to 1; the weaker the relationship between the two variables, the closer *r* is to 0.



GURE 8-12. The closer the magnitude f the correlation coefficient is to 1, the ss scatter there is in the relationship etween the two variables. The closer e correlation coefficient is to 0, the eaker the relationship between the two riables.

Correlation and Correlation Coefficients

- We will examine two different correlation coefficients.
- The first, called the **Pearson product-moment correlation coefficient**, quantifies the strength of association between two variables that are **normally distributed.**
- When people refer to the correlation coefficient, they almost always mean the Pearson product moment correlation coefficient. (Read Page 165)

The Pearson product-moment correlation coefficient defined by

$$r = \frac{\Sigma(X - \overline{X})(Y - \overline{Y})}{\sqrt{\Sigma(X - \overline{X})^2 \Sigma(Y - \overline{Y})^2}}$$

- See example at page 166.

Correlation and Correlation Coefficients

- It is often desirable to test the hypothesis that there is a trend in a clinical state, measured on an **ordinal scale**, as another variable changes.
- The Pearson product-moment correlation coefficient is a parametric statistic designed to be used on data distributed normally along interval scales, so it cannot be used.
- It also requires that the trend relating the two variables be linear.
- <u>When the sample suggests that the population from which both variables were drawn from does not meet these criteria, it is possible to compute a measure of association based on the ranks rather than the values of the observations.</u>
- This new correlation coefficient, called the **Spearman rank correlation coefficient**, rs , is based on ranks and can be used for data quantified with an ordinal scale.*
- The Spearman rank correlation coefficient is a nonparametric statistic because it does not require that the observations be drawn from a normally distributed population.

- See page 170 ana 172 (Example for Spearman Rank correlation coef.) $r_s = 1 - \frac{6\sum d^2}{n^3 - n}$