8: Independent Samples

Introduction

In the previous chapter we looked at paired samples. In this chapter, we look at independent samples. The distinction between paired samples and independent samples was made in Chapter 7 and will not be reiterated here.

Illustrative data. This illustration considers cholesterol levels in two groups of men. Group 1 are hard-driving, time-urgent Type A personalities. Group 2 are more laid-back Type B personalities. Data are simple random samples from populations of Type A and Type B men. Fasting cholesterol levels (mg/dl) are:

Group 1 ($n_1 = 20$): 233, 291, 312, 250, 246, 197, 268, 224, 239, 239, 254, 276, 234, 181, 248, 252, 202, 218, 212, 325

Group 2 ($n_2 = 20$): 344, 185, 263, 246, 224, 212, 188, 250, 148, 169, 226, 175, 242, 252, 153, 183, 137, 202, 194, 213

For data to be analyzed in SPSS, they must be reconfigured into two columns. One column is for the **response (dependent) variable** and the other is for the **explanatory (group) variable.** Data are stored online in the file wcgs.sav. Here is a screen shot of the first 22 observations in the data file:
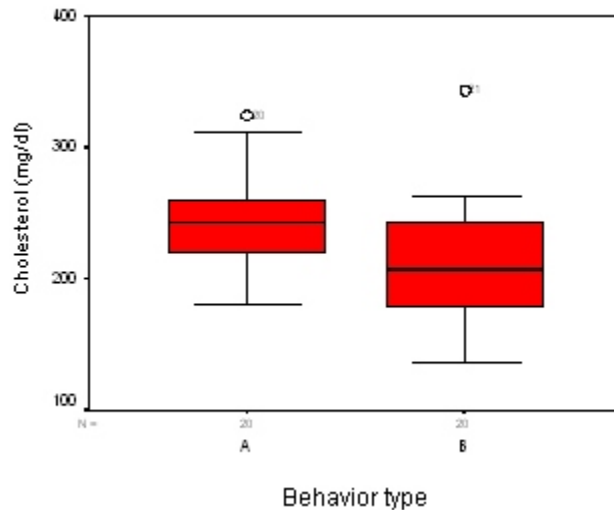
Exploratory and Descriptive Statistics

We can start be calculating **group means and standard deviations**. For the illustrative data

Type A men: $\bar{x}_1 = 245.05$, $s_1 = 36.64$, $n_1 = 20$.

Type B men: $\bar{x}_2 = 210.30$, $s_2 = 48.34$, $n_2 = 20$.

Side-by-side boxplots are often helpful for seeing differences in locations, spreads, and shapes:

The descriptive statistics and plot reveals that the Type A men had a higher cholesterol values on average. They also had less variability. There are outside values (potential outliers) in both groups.

SPSS. Descriptive statistics and side-by-side boxplots are computed by clicking `Analyze > Descriptive Statistics > Explore`. The variable `CHOL` is placed in the `Dependent List` and the variable `GROUP` is placed in the `Factor List`.

95% Confidence Interval

There are a couple of ways to calculate confidence intervals and *P*-values for independent means. We use the "equal variance assumed" method.

The sample mean difference $\bar{x}_1 - \bar{x}_2$ is the point estimator of population mean difference $\mu_1 - \mu_2$. For the illustrative data, $\bar{x}_1 - \bar{x}_2 = 245.05 - 210.30 = 34.75$.

The 95% confidence interval for μ_1 μ_2 is:

$$(\bar{x}_1 - \bar{x}_2) \pm (t_{df,.975})(se_{\bar{x}_1 - \bar{x}_2}) \tag{1}$$

where $\bar{x}_1 - \bar{x}_2$ represents the mean difference in the sample, $t_{df,.975}$ is the 97.5th percentile of a *t* random variable with *df* degrees of freedom, and $se_{\bar{x}_1 - \bar{x}_2}$ is the **standard error of the independent mean difference** given by:

$$se_{\bar{x}_1 - \bar{x}_2} = \sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \tag{2}$$

The symbol s_p^2 represents the **pooled estimate of variance**:

$$s_p^2 = \frac{(df_1)(s_1^2) + (df_2)(s_2^2)}{df} \tag{3}$$

with $df = df_1 + df_2 = (n_1 - 1) + (n_2 - 1) = n_1 + n_2 - 2$.

Illustrative example (`wcgs.sav`). For the illustrative data, the pooled estimate of variance $s_p^2 = \frac{(19)(36.64^2) + (19)(48.34^2)}{38} = 1839.557$ with $df = 20 + 20 - 2 = 38$. The standard error of the mean

difference: $se_{\bar{x}_1 - \bar{x}_2} = \sqrt{1839.557\left(\frac{1}{20} + \frac{1}{20}\right)} = 13.56$. For 95% confidence, $t_{38,.975} = 2.02$ (from the *t*

table). The 95% confidence interval for $\mu_1 - \mu_2 = (245.05 - 210.30) \pm (2.02)(13.56) = 34.75 \pm 27.4 = (7.4, 61.8)$. This permits us to say with 95% confidence that population mean difference $\mu_1 - \mu_2$ is between 7.4 and 61.8 (mg/dl).

Hypothesis Test

Hypotheses: Under the null hypothesis, there is no difference in population means: H_0: $\mu_1 = \mu_2$. The alternative hypothesis is either H_1: $\mu_1 \neq \mu_2$ (two-sided), H_1: $\mu_1 > \mu_2$ (one-sided to the right), or H_1: $\mu_1 < \mu_2$ (one-sided to the left). Two-sided alternative are more common in practice.

Test statistic: The test statistic is:

$$t_{stat} = \frac{(\bar{x}_1 - \bar{x}_2)}{se_{\bar{x}_1 - \bar{x}_2}} \tag{4}$$

where *se* represent the standard error of the mean difference (previous page). The test statistic has $df = df_1 + df_2 = (n_1 - 1) + (n_2 - 1) = n_1 + n_2 - 2$.

P-value: The t_{stat} is converted to a *P*-value as discussed in prior chapters. When a computer program is unavailable, use the *t* table to determine the approximate *P*-value by wedging the t_{stat} between *t* quantile landmarks Small *P*-values (less than 0.05 or 0.01) provide good evidence against the null hypothesis.

Significance statement (optional): The null hypothesis is rejected at the α level of significance when $P \leq \alpha$.

Illustrative example. We want to test the illustrative data to see if the population means are equivalence. On the prior page we established $\bar{x}_1 - \bar{x}_2 = 34.75$, $se_{\bar{x}_1 - \bar{x}_2} = 13.56$, and $df = 38$.

- The statistical hypotheses are H_0: $\mu_1 = \mu_2$ versus H_1: $\mu_1 \neq \mu_2$.

- The test statistic $t_{stat} = \dfrac{\bar{x}_1 - \bar{x}_2}{se_{\bar{x}_1 - \bar{x}_2}} = \dfrac{34.75}{13.56} = 2.56$ with $df = 38$.

- A $t_{stat} = 2.56$ with 38 df is shown on the t_{38} distribution in the figure to the right. Critical value landmarks from the *t* table are shown in relation to the test statistic. The critical values are $t_{38,.99} = 2.43$ (right tail of 0.01) and $t_{38,.995} = 2.71$ (right tail of 0.005). The one-sided *P*-value is more than 0.005 and less than 0.01. The two-sided *P*-value is twice this: $0.01 < P < 0.02$. The two-sided *P*-value $= 0.014$ (by computer). This provides significant evidence against H_0 (reject H_0).
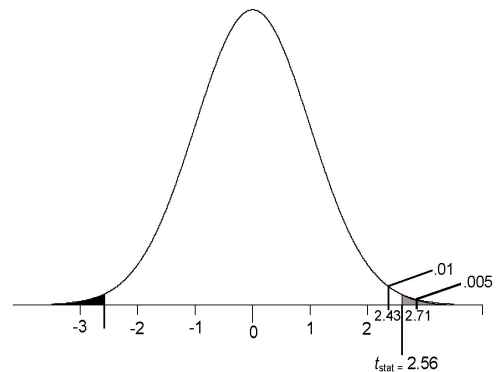- The test is significant at α 0.05 but is not at α 0.01.

Fig:t_stat=2.56.ai

SPSS. The confidence interval and test results are derived with `Analyze > Compare Means > Independent Samples T Test`. Output are shown below. The top table contains descriptive statistics. The bottom table contains inferential statistics. The techniques presented in the chapter correspond to the "Equal variances assumed" line in the second table. "Sig. (2-tailed)" refers to the *p* value for the two-sided tests. We have not covered Levene's test for equality of variance.

Group Statistics

	GROUP	N	Mean	Std. Deviation	Std. Error Mean
Cholesterol (mg/dl)	1	20	245.05	36.638	8.193
	2	20	210.30	48.340	10.809

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means					95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Cholesterol (mg/dl)	Equal variances assumed	1.293	.263	2.562	38	.014	34.75	13.563	7.293	62.207
	Equal variances not assumed			2.562	35.413	.015	34.75	13.563	7.227	62.273

Assumptions for Inference

All inferential methods require assumptions. We consider validity and distributional assumptions separately.

The **validity assumptions** include:

- "No information bias." This means data are accurate and measure what the purport to measure. Do you remember the GIGO (garbage in, garbage out) principle?
- "No selection bias." This means data are a random reflection of their respective populations.
- "No confounding." This means groups are comparable in all ways other that the factor that defines the groups.

The **distributional assumptions** are:

- "Independence." This means that groups are simple random samples from their respective populations. (Same as the "no selection" validity assumption).
- "Normality." This means either the underlying populations are Normal or the sample is large enough to compensate for underlying non-Normality by means of The Central Limit Theorem. (In practice it is often enough for data to be mound shaped without extreme skewness.)
- "Equal variance." This means the variability within the two populations is similar. There are several formal techniques to check this assumption, but it is often adequate to simply look at the hinge-spreads in the side-by-side boxplots. If one box is three times the height of the other, you may be dealing with unequal population variances. Another rule of thumb is to look at group standard deviations. If one of the sample standard deviations is three times larger than the other, you are probably dealing with unequal variance.

Notes:

- Distributional assumptions form the mnemonic INE ("line minus the L) for Independence, Normal, and Equal variance.
- Methods to assess distributional assumptions are covered in more advanced statistics course.
- Validity assumptions are generally much more important than distributional assumptions.

Sample Size Requirements

A quick way to determine an appropriate sample size to test H_0: $\mu_1 = \mu_2$ so that the test has 80% power at = .05 (two-sided) is to use the formula:

$$n = \frac{(16)(s^2)}{\Delta^2} + 1$$

where n represents the *per group* sample size requirements,
          s^2 represents the variance within groups (use s^2_p, if available), and
            represents "a difference worth detecting" (i.e., the size of the difference you are looking for).

Illustrative example.

*   Suppose you want to detect a mean difference of 50 for this same variable. Then,
    $n = \frac{(16)(45^2)}{50^2} + 1$ 14 per group. (Total sample size = 14 + 14 = 28).

*   Suppose you want to detect a mean difference of 25 for a variable with a standard deviation of
    45. We calculate $n = \frac{(16)(45^2)}{25^2} + 1 = 53$. Keep in mind that this is the sample size for each
    group. (The total sample size for both groups combined is 53 + 53 = 106).

*   Notice that you need a much larger sample size to detect a smaller difference.