8: Independent Samples and Their Differences

Introduction | EDA | Estimation | Hypothesis Test | Sample Size for the Independent *t* Test

Introduction

In the previous chapter we studied paired samples. In this chapter we study independent samples. Unlike paired samples, independent samples represent measurements on separate, unrelated groups. There is *no* pairing, coupling, or matching of observations within the samples.

Comparison of paired and independent samples. Suppose we want to determine the effects of oral contraceptive use on blood pressure. We could employ paired samples or independent samples to address this question. *Using paired samples* we could measure say, 4 women's blood pressure before (`SBP1`) and then after (`SBP2`) taking the contraceptive. Data might look something like this:

Woman's ID number	SBP1	SBP2	DELTA (SBP2 - SBP1)
1	125	120	-5
2	115	115	0
3	125	125	0
4	130	135	+5

Using independent sample we could randomly assign the contraceptive or a placebo to 8 women and compare the change in blood pressure in the two groups. Data might look something like this:

Woman's ID number	Group (1 = OC, 2 = Placebo)	Change in Systolic Blood Pressure (mm Hg)
1	1	+5
2	1	+8
3	1	+2
4	1	0
5	2	+5
6	2	-1
7	2	0
8	2	+6

Notice that with paired samples, women serve as their own controls (so-called "self controls"). This is *not* a truly controlled situation. In contrast, use of independent samples provides a true control group. The independent design is generally preferable.

Illustrative data (`wcgs.sav`). Let us use data from the Western Collaborative Group Study (Selvin, 1991, p. 41) to illustrate methods this chapter. Data are serum cholesterol levels (mg/dl) in Type A and Type B men.

Group 1 (Type A men): 233, 291, 312, 250, 246, 197, 268, 224, 239, 239, 254, 276, 234, 181, 248, 252, 202, 218, 212, 325

Group 2 (Type B men): 344, 185, 263, 246, 224, 212, 188, 250, 148, 169, 226, 175, 242, 252, 153, 183, 137, 202, 194, 213

SPSS. Data are restructure with a *dependent variable* and *group variable* to as follows:

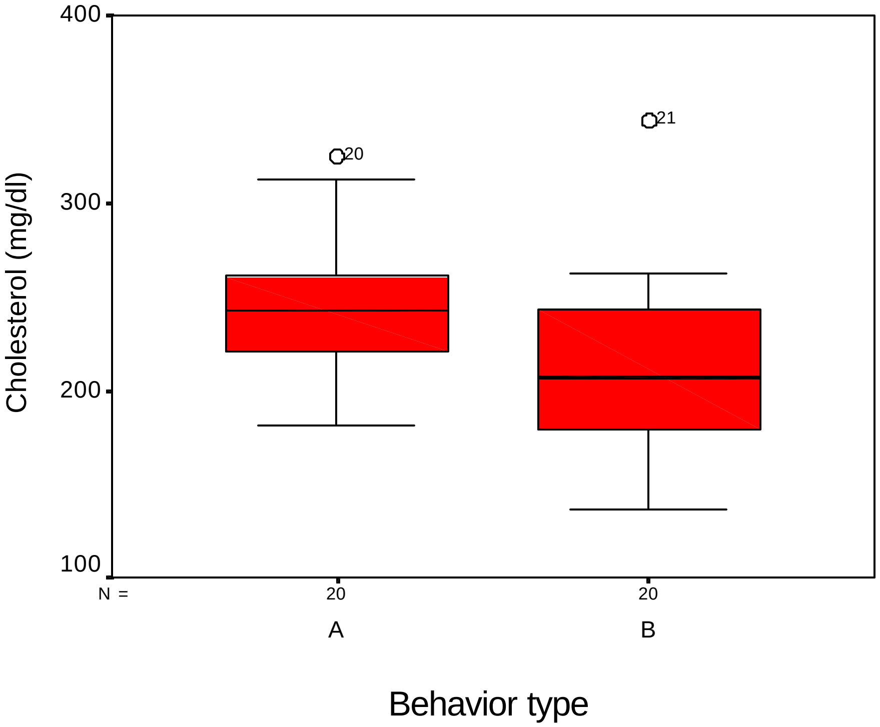
Observation	CHOL	GROUP
1	233	1
2	291	1
3	312	1
↓	↓	↓
38	202	2
39	194	2
40	213	2

Only the first three and last three observations are shown in order to save printing costs. The full data set may be downloaded from the data directory.

Exploratory Data Analysis

Insights are gained by plotting the data in the form of a side-by-side stem-and-leaf plot or boxplot. We take this opportunity to review the boxplot method. (See Chapter 3 for boxplot basics.)

This figure below shows side-by-side boxplots for the illustrative data.



Notice that group A has higher cholesterol values on average and has less variability. Each group has an outside value (on top).

SPSS. The graphs can be contructed with `Analyze> Descriptive Statistics > Explore`. The variable `CHOL` was placed in the `Dependent List` and the variable `GROUP` was placed in the `Factor List`.

Estimation

Let n_i represent the sample size for group i {i: 1, 2}, $\bar{x}_i$ represent the mean of group i, and s_i represent the standard deviation of group i. For the illustrative data we note:

Group 1:	$n_1 = 20$	$\bar{x}_1 = 245.05$	$s_1 = 36.64$
Group 2:	$n_2 = 20$	$\bar{x}_2 = 210.30$	$s_2 = 48.34$

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

We also want to derive an **interval estimate** for $\mu_1 - \mu_2$ in the form of a confidence interval.

Before calculating the confidence interval we determine the **pooled estimate of variance (s_p^2)** as:

$$s_p^2 = \frac{(\mathrm{df}_1)(s_1^2) + (\mathrm{df}_2)(s_2^2)}{\mathrm{df}} \qquad \textbf{(8.1)}$$

where $\mathrm{df}_1 = n_1 - 1$, $\mathrm{df}_2 = n_2 - 1$, and $\mathrm{df} = \mathrm{df}_1 + \mathrm{df}_2$. For the illustrative data, $\mathrm{df}_1 = 20 - 1 = 19$, $\mathrm{df}_2 = 20 - 1 = 19$, $\mathrm{df} = 19 + 19 = 38$. Therefore, $s_p^2 = \frac{(19)(36.64^2) + (19)(48.34^2)}{38}$ = 1839.557 (mg/dl)^2. This pooled estimate of variance is a weighted average of the two sample variances.

We then estimate the **standard error of the independent mean difference** as follows:

$$se_{\bar{x}_1 - \bar{x}_2} = \sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \qquad \textbf{(8.2)}$$

For the illustrative data, $se_{\bar{x}_1 - \bar{x}_2} = \sqrt{1839.557\left(\frac{1}{20} + \frac{1}{20}\right)}$ = 13.56.

The **95% confidence interval for μ_1 - μ_2** is given by:

$$(\bar{x}_1 - \bar{x}_2) \pm (t_{df,.975})(se_{\bar{x}_1 - \bar{x}_2}) \qquad \textbf{(8.3)}$$

where $t_{\mathrm{df},.975}$ is the 97.5th percentile on a t distribution with df degrees of freedom.. For the illustrative data, $t_{38,.975} \cong 2.02$. (The t table does not include a row for df = 38, so we use the closest available degrees of freedom: $t_{38} \approx t_{40}$.) The 95% confidence interval for $\mu_1 - \mu_2$ = (245.05−210.30) ± (2.02)(13.56) = 34.75 ± 27.4 = (7.4, 61.8). This permits us to say with 95% confidence that the mean difference in the population is between 7.4 mg/dl and 61.8 mg/dl.

Hypothesis Test

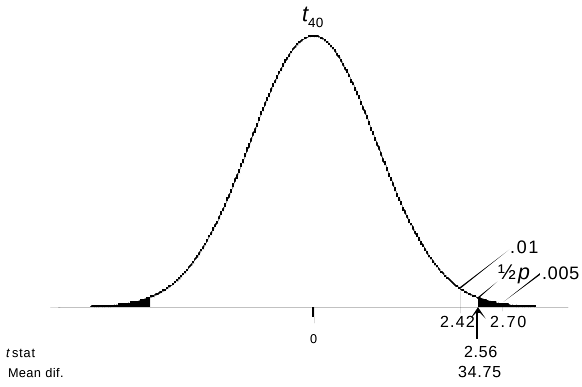
An *independent t statistic* is used to test H_0:$\mu_1 - \mu_2$ = 0,or equivalently H_0:$\mu_1 = \mu_2$. The two-sided alternative is H_1:$\mu_1 - \mu_2 \neq 0$, or equivalently, H_1:$\mu_1 \neq \mu_2$.

The test statistic is:

$$t_{\text{stat}} = \frac{(\bar{x}_1 - \bar{x}_2)}{se_{\bar{x}_1 - \bar{x}_2}} \qquad \textbf{(8.4)}$$

The standard error is calculated according to formula 8.2. The test statistic has $(n_1 - 1) + (n_2 - 1) = n_1 + n_2 - 2$ degrees of freedom.

Illustrative example. Recall that for the illustrative data $\bar{x}_1 - \bar{x}_2$ = 34.75, $se_{\bar{x}_1 - \bar{x}_2}$ = 13.56, and df = 38 (see prior page). In testing H_0: $\mu_1 = \mu_2$, $t_{\text{stat}} = \dfrac{34.75}{13.56}$ = 2.56. The two-tailed *p* value is twice the area under the curve beyond the test statistic in the t_{38} distribution. Since the *t* table does not include a row for *df* = 38, we note $t_{38} \cong t_{40}$. Using 40 degrees of freedom, *half* the *p* value is between .005 and.01.



Thus, the two-sided *p* is between .01 and .02 (.01 < *p* < .02). A precise *p* value can be computed with *StaTable* or any other probability calculator (*p* = .015).

This test assumes observations are *independent*, the sampling distribution is *normal*, and the groups are *equal*. These assumptions can be remembered with the mnemonic = "ine," or "line" without the "l.". Although we should be aware of assumptions, the test allows for considerable departures from the normality and equal variance while still providing accurate results. The test is particularly robust with large samples ($n_i \geq 30$) and when groups are of equal size ($n_1 = n_2$) (Zar, 1996, p. 128).

SPSS. The test is calculated with `Analyze > Compare Means > Independent Samples T Test`. Select the variable being analyzed as the `Test variable` and identify the `Group variable`. Click the `Define Groups` button to identify the codes used to identify groups.

Sample Size for the Independent *t* Test

To achieve 80% power at $\alpha = .05$ (two-sided), use this formula to determine the *n* of *each group*:

$$\frac{(16)(s^2)}{\Delta^2} + 1 \qquad \textbf{(8.5)}$$

where Δ = a difference worth detecting and s^2 = a good variance estimate (use the pooled estimate of variance, when available).

Illustrative example. Suppose we want to detect a mean difference of 25 for a variable with a variance of 45^2. Thus, $n = \frac{(16)(45^2)}{25^2} + 1 = 53$ for each group. In contrast, if we wanted to detect a mean difference of 50 for this same variable, $n = \frac{(16)(45^2)}{50^2} + 1 \cong 14$ per group.