7: Paired Samples

Data and exploration

This chapter considers the analysis of a quantitative outcome based on paired samples. **Paired samples** (also called **dependent samples**) are samples in which natural or matched couplings occur. In paired samples, each data point in one sample is matched to a unique data point in the second sample. An example of a paired sample is a pre-test/post-test study design in which a factor is measured before and after an intervention. Paired samples may also be achieved by matching individuals on personal characteristics such as age and sex.

In contrast to paired samples, **independent samples** consider two or unrelated separate groups. Two samples are independent when the data points in one sample are unrelated to the data points in the second sample.

It is important to differentiate between paired samples and independent samples. Consider two ways to study the effects of oral contraceptives on blood pressure. Using paired samples, we could take two measurements on individuals in the sample before and again after starting oral contractive. These data are paried because each data point associated with oral contraceptive use would be uniquely matched to the same individual when they were not taking oral contraceptives.

We could also study the effects of oral contraceptives on blood pressure in two separate groups of women, one group taking the oral contraceptive and the other taking a placebo. In this instance, data are obtained from unrelated groups and data points in one sample would not be matched to the data points in the second sample.

You might ask which type of sample is better, paired or independent? The paired sample offers the benefit of controlling for some of the extraneous factors that influence a woman's blood pressure. However, sequential samples such as this may still be confounded by independent factors that differ over time. Health outcome may show changes over time due to the influence of these extraneous factors. On the other hand, independent sample have the benefit of having a concurrent control group. Perhaps the best design is a third option in which change within individuals is compared in two independent groups.

Illustrative data (`oatbran.sav`). A study investigated whether oat bran lowers serum cholesterol levels. Fourteen individuals were randomly assigned a diet that included either oat bran or corn flakes. After two weeks on the initial diet, serum low-density lipoprotein levels (mg/dl) were measured. Each subject was then 'crossed-over" to the alternate diet. After two-weeks on the second diet, low-density lipoprotein levels were was once again recorded. Data are shown on the next page:

Illustrative data:

ID	CORNFLK	OATBRAN	DELTA
1	4.61	3.84	0.77
2	6.42	5.57	0.85
3	5.40	5.85	-0.45
4	4.54	4.80	-0.26
5	3.98	3.68	0.30
6	3.82	2.96	0.86
7	5.01	4.41	0.60
8	4.34	3.72	0.62
9	3.80	3.49	0.31
10	4.56	3.84	0.72
11	5.35	5.26	0.09
12	3.89	3.73	0.16
13	2.25	1.84	0.41
14	4.24	4.14	0.10

We will refer to the CORNFLK data as sample 1 and the OATBRAN data as sample two. Subscripts are attached to summary statistics, so $\bar{x}_1 = 4.444$ mg/dl and $\bar{x}_2 = 4.081$ mg/dl.

Further analysis requires the creation of a new variable to hold information about the difference within pairs. Let us call this variable DELTA. When creating DELTA values, it makes little difference whether you subtract sample 1 values from sample 2 values or vice versa. It is important, however, to keep track of the direction of the difference. Let DELTA = CORNFLK - OATBRAN. Positive values reflect higher LDL cholesterol on the corn flake diet. **All further analyses are directed toward the DELTA variable.** Statistics for DELTA are denoted with a subscript of $_d$:

$$n_d = 14 \qquad \bar{x}_d = 0.363 \qquad s_d = 0.4060$$

Exploratory graphs are useful in focusing our attention where it ought to be. Here's a stemplot of the DELTA values:

```
-0|24
 0|011334
 0|667788
×1 (mmol/L)
```

Notice that 12 of the 14 observations showed a decrease in LDL levels. The distribution is mound-shaped with no apparent outliers, with a median of 0.3 and range of −0.4 to 0.8.

Confidence Interval for μ_d

The parameter μ_d in a matched pair study is the mean difference in responses to the two treatments within matched pairs of individuals in the entire population. The sample mean of DELTA $\bar{x}_d$ estimates ths value. We want to get a better sense of how well $\bar{x}_d$ estimates the true value of μ_d by calculating a confidence interval for μ_d

The approach presented previously is used to calculate the confidence. The point estimate is surrounded with a margin of error. The sample mean of DELTA is the point estimator of μ_d The margin of error is based on a *t* critical value and the estimated standard error of the mean difference according to the following formula:

$$\bar{x}_d \pm (t_{n-1,1-\alpha/2})(se_{\bar{x}_d}) \quad \textbf{(7.1)}$$

where $t_{n-1,1-\alpha/2}$ represents the – percentile on a *t* distribution with *n*– degrees of freedom and $se_{\bar{x}_d} = \frac{s_d}{\sqrt{n}}$.

Illustrative example (`oatbran.sav`), 95% confidence interval. For the illustrative data, $n = 14$, $\bar{x}_d = 0.363$, $s_d = 0.4060$, and $se_d = \frac{0.4060}{\sqrt{14}} = 0.1085$. Therefore, 95% confidence interval for μ_d is

$= 0.3629 \pm (t_{13,\ 1-(.05/2)})(0.1085)$
$= 0.3629 \pm (t_{13,\ .975})(0.1085)$
$= 0.3629 \pm (2.16)(0.1085)$
$= 0.3629 \pm 0.2344$
$= (0.129, 0.597)$

We are 95% confident the true mean difference lies between 0.129 and 0.597.

Illustrative example (`oatbran.sav`), 90% confidence interval. For 90% confidence, use $\alpha = .10$. The 90% confidence interval

$= 0.3629 \pm (t_{13,\ 1-(.10/2)})(0.1085) = 0.3629 \pm (t_{13,\ .95})(0.1085)$
$= 0.3629 \pm (1.77)(0.1085)$
$= 0.3629 \pm 0.1920$
$= (0.171, 0.555)$

We are 90% confident the true mean difference lies in this interval.

Hypothesis Test

Hypotheses: Under the null, we expect no mean. The null hypothesis is H_0: $\mu_d = 0$. The alterative hypothesis is either H_1: $\mu_d < 0$ (one-sided to left), H_1: $\mu_d > 0$ (one-sided to right), or H_1: $\mu_d \neq 0$ (two-sided).

Test statistic: The paired *t* statistic is:

$$t_{stat} = \frac{\bar{x}_d}{se_{\bar{x}_d}} \qquad \textbf{(7.2)}$$

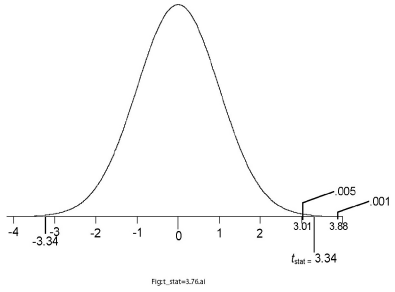
where $\bar{x}_d$ represents the mean difference in the sample and $se_{\bar{x}_d} = \frac{s_d}{\sqrt{n}}$. This test statistic has $df = n - 1$.

***P*-value:** The t_{stat} is converted to a *P*-value in the usual fashion. Use either the t table or a utility program (e.g., StaTable) for this purpose. Low P-value provide evidence against H_0.

Significance (optional): The test is said to be significant at the alpha level of significance when $P < \alpha$, in which case H_0 is rejected.

Illustrative example. We've established in mean decline in LDL cholesterol on oatbran is $\bar{x}_d = 0.363$ ($n_d = 14$, and $s_d = 0.4060$).

- **Hypotheses:** H_0: $\mu_d = 0$ versus H_1: $\mu_d \neq 0$
- **Test statistic:** The standard error of the mean difference $se_d = \frac{0.4060}{\sqrt{14}} = 0.1085$, the $t_{stat} = \frac{0.3629}{0.1085} = 3.34$, and $df = 14 - 1 = 13$.
- ***P*-value:** The critical landmarks surrounding the t_{stat} are $t_{13..995} = 3.01$ and $t_{13..999} = 3.88$. Therefore, the one-tailed *P*-value is less than 0.005 and more than 0.001. The two-sided $0.002 < P < 0.01$. The precise $P = 0.0053$ (by computer). This provides good evidence against H_0.



- **Significance (optional):** Data are significant at the 0.01; reject H_0.

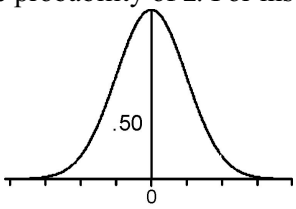
Power

If H_0 is false and we fail to reject it, we commit a type II error. The probability of a type II error is β, and the probability of avoiding a type II error is 1 − β or "power."

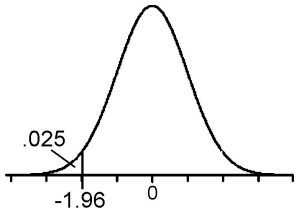
The approximate power of a *t* test at = .05 (two-sided) is equal to $\phi\left(-1.96+\frac{\Delta\sqrt{n}}{\sigma}\right)$, where Δ represents "a difference worth detecting," is the population standard deviation (or a reasonably good estimate of this value), and ϕ (phi) is the area under a standard normal curve to the left of the value in parentheses.

The difference worth detecting is the difference the investigator is looking for. It is a meaningful difference, and is up to the discretion of the investigator. In a study of an anti-hypertensives for instance, a drop of 10 mm Hg might be worth detecting, while a drop of 1 mm Hg might *not* be worth detecting. In a study on weight loss, a drop of 5 pounds might be meaningful in a population of runway models, but may be meaningless in a morbidly obese population. The amount of the difference worth detecting is a matter of the research.

The ϕ(z) function may require further explanation. This is the cumulative probability of z. For instance, (0) = .5000, since half the area under the Z curve is to the left of 0, i.e., .



As an additional example, (−1.96) = .025, which looks like this:



Illustrative example. We want to determine the power of a *t* test in which a mean difference of 2 is worth detecting. We study 30 observations, and prior studies suggests the variable being studied has a standard deviation of 6. What is the power of the study at α = .05 two-sided?

ANS: $\phi\left(-1.96+\frac{|2|\sqrt{30}}{6}\right)=\phi(-0.13)$. We use a *Z* table to determine that the area under the curve to the right of −0.13 is 0.4483 (about 45%). This means the test had less than a 50/50 chance of revealing a significant difference of 2 (or more). A test should have *at least* 80% power. Thus, this test has inadequate power. Increasing the sample size of the study will increase its power.

Sample Size Requirements (Optional)

When planning a hypothesis test, we need to have an idea of an appropriate size to ensure sufficient power of the test before the study begins. Without such a plan, we might find out that the study had inadequate power to detect a meaningful difference. When conducting statistical tests, an adequate sample size (n) is one that will reject H_0 at a given α level with adequate power ($1-\beta$) when a difference worth detecting (Δ) is present in the population.

To determine adequate n, we must make the assumption that sampling distributional of the observed mean difference is normal. We must also present some reasonable assumption regarding the variability (σ) of the underlying variable. The value of σ is generally unknown and must therefore be estimated from either (a) a preliminary (pilot) study, (b) previously published results from a similar population, (c) intuition, or (d) other past experience.

Given these assumptions, the sample size needed to detect a difference of Δ at $\alpha = .05$ (two-sided) for 80% power is equal to $\frac{(16)\sigma^2}{\Delta^2} + 1$ (http://www.tufts.edu/~gdallal/SIZE.HTM).

Illustrative example. To detect $\Delta = 2$ for a variable with $\sigma = 6$, you need $n = [(16)(6^2) / 2^2] + 1 = 145$.