Sections 4.5-4.7: Two-Sample Problems

Paired t-test (Section 4.6)

Examples of Paired Differences studies: • Similar subjects are paired off and one of two treatments is given to each subject in the pair.

or

• We could have two observations on the same subject.

The key: With paired data, the pairings cannot be switched around without affecting the analysis.

We typically wish to perform inference about the mean difference, denoted δ .

Example 1: Twelve cars were equipped with radial tires and driven over a test course. Then the same 12 cars (with the same drivers) were equipped with regular belted tires and driven over the same course. After each run, the cars' gas economy (in km/l) was measured. Is there evidence that radial tires produce better fuel economy? (Assume normality of data, and use $\alpha = .05$.)

 Car

 Gas eco.
 | 1 2 3 4 5 6 7 8 9 10 11 12

 Y1 (radial)
 | 4.2 4.7 6.6 7.0 6.7 4.5 5.7 6.0 7.4 4.9 6.1 5.2

 Y2 (belted)
 | 4.1 4.9 6.2 6.9 6.8 4.4 5.7 5.8 6.9 4.7 6.0 4.9

Calculate differences: $d = Y_1 - Y_2$

d_i:

Example 1(a): Find a 95% CI for the mean difference in gas economy between radial tires and belted tires.

Interpretation: With 95% confidence:

Note: With paired data, the two-sample problem really reduces to a one-sample problem on <u>the sample of differences</u>.

Paired t-test in R:

> radial <- c(4.2, 4.7, 6.6, 7.0, 6.7, 4.5, 5.7, 6.0, 7.4, 4.9, 6.1, 5.2) > belted <- c(4.1, 4.9, 6.2, 6.9, 6.8, 4.4, 5.7, 5.8, 6.9, 4.7, 6.0, 4.9) > t.test(radial, belted, paired = T, alternative =

Paired t-CI in R:

"greater")

> t.test(radial, belted, paired = T, conf.level=0.95)\$conf.int

Two Independent Samples t-test (Section 4.5)

Sometimes there's no natural pairing between samples.

Example 1: Collect sample of males and sample of females and ask their opinions on whether capital punishment should be legal.

Example 2: Collect sample of iron pans and sample of copper pans and measure their resiliency at high temperatures.

No attempt made to pair subjects – we have two <u>independent</u> samples.

We could rearrange the order of the data and it wouldn't affect the analysis at all.

Comparing Two Means

Our goal is to compare the mean responses to two treatments, or to compare two population means (we have two separate samples).

We assume both populations are normally distributed (or "nearly" normal).

We're typically interested in the difference between the mean of population 1 (μ_1) and the mean of population 2 (μ_2).

We may construct a CI for $\mu_1 - \mu_2$ or perform one of three types of hypothesis test:

Note: H_0 could be written H_0 : $\mu_1 - \mu_2 = 0$.

The parameter of interest is

Notation: \overline{Y}_1 = mean of Sample 1 \overline{Y}_2 = mean of Sample 2 σ_1 = standard deviation of Population 1 σ_2 = standard deviation of Population 2 s_1 = standard deviation of Sample 1 s_2 = standard deviation of Sample 2 n_1 = size of Sample 1 n_2 = size of Sample 2

The point estimate of $\mu_1 - \mu_2$ is

This statistic has standard error

but we use

since σ_1 , σ_2 unknown.

Since the data are normal, we can use the t-procedures for inference.

<u>Case I</u>: Equal population variances $(\sigma_1^2 = \sigma_2^2)$

In the case where the two populations have equal variances, we can better estimate this population variance with the <u>pooled</u> sample variance:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

Formula for $(1 - \alpha)100\%$ CI for $\mu_1 - \mu_2$ is:

where the d.f. = $n_1 + n_2 - 2$.

To test H_0 : $\mu_1 = \mu_2$, the test statistic is:

<u>H_a</u>	Rejection region	P-value
$\mu_1 \neq \mu_2$	$t < -\mathbf{t}_{\alpha/2} \text{ or } t > \mathbf{t}_{\alpha/2}$	2*(tail area)
$\mu_1 < \mu_2$	$t < -t_{\alpha}$	left tail area
$\mu_1 > \mu_2$	$t > t_{\alpha}$	right tail area

where the d.f. = $n_1 + n_2 - 2$.

Example: What is the difference in mean water absorbency of cotton fiber and acetate fiber?
Let μ₁ = mean absorbency (in %) for cotton, let μ₂ = mean absorbency (in %) for acetate.

Find 99% CI for $\mu_1 - \mu_2$.

Does $\sigma_1^2 = \sigma_2^2$? Could test this formally using an F-test (Sec. 4.7) or (preferably) could simply compare spreads of box plots for samples 1 and 2.

Let's suppose we can safely assume $\sigma_1^2 = \sigma_2^2$ here.

99% CI for $\mu_1 - \mu_2$:

Interpretation:

Testing whether cotton has a greater mean absorbency: $H_0: \mu_1 = \mu_2 \text{ vs. } H_a: \mu_1 > \mu_2 \text{ (at } \alpha = .10)$

Test statistic:

Case II: Unequal population variances $(\sigma_1^2 \neq \sigma_2^2)$.

In this case, no exact t-procedure exists, but an approximate t-test is available.

An R example:

```
> y1 <- c(102,86,98,109,92,99)
> y2 <- c(81,165,97,134,92,87,114)
> boxplot(y1, y2)
> t.test(y1, y2, alternative = "two.sided",
paired = FALSE, var.equal = FALSE)
```

Inference about Two Proportions

We now consider inference about $p_1 - p_2$, the difference between two population proportions.

Point estimate for $p_1 - p_2$ is

For large samples, this statistic has an approximately normal distribution with mean $p_1 - p_2$ and standard

deviation
$$\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

So a $(1 - \alpha)100\%$ CI for $p_1 - p_2$ is

 \hat{P}_1 = sample proportion for Sample 1 \hat{P}_2 = sample proportion for Sample 2 n_1 = sample size of Sample 1 n_2 = sample size of Sample 2

Requires large samples:

- (1) Need $n_1 \ge 20$ and $n_2 \ge 20$.
- (2) Need number of "successes" <u>and</u> number of "failures" to be 5 or more in <u>both</u> samples.

Test of H₀: $p_1 = p_2$

Test statistic:

(Use pooled proportion because under H_0 , p_1 and p_2 are the same.)

Pooled sample proportion

 \hat{p} =

Example: Let p_1 = the proportion of urban residents who support the construction of a nuclear power plant, and let p_2 = the proportion of suburban residents who support the construction. Take a random sample of 100 urban residents; 61 support the construction. Take a random sample of 120 suburban residents; 65 support the construction.

Find a 95% CI for the difference in the true proportion of urban residents and the true proportion of suburban ones who support it. Hypothesis Test: Is the true proportion of urban residents who support construction significantly different from the proportion of suburban ones? Use $\alpha = 0.05$.

In R:
> prop.test(c(61,65), n = c(100,120), correct=F,
alternative="two.sided")