

Inferences for Two Samples

Corresponds to Chapter 8 of
Tamhane and Dunlop

Slides prepared by Elizabeth Newton (MIT),
with some slides by Ramón V. León
(University of Tennessee)

Introductory Remarks

- A majority of statistical studies, whether experimental or observational, are comparative
- Simplest type of comparative study compares two populations
- Two principal designs for comparative studies
 - Using independent samples
 - Using matched pairs
- Graphical methods for informal comparisons
- Formal comparisons of means and variances of normal populations
 - Confidence intervals
 - Hypothesis tests

Independent Samples Design

Example: Compare Control Group to Treatment Group


- See page 270 in course textbook.

Independent samples design:

Sample 1: x_1, x_2, \dots, x_{n_1} ←

Sample 2: y_1, y_2, \dots, y_{n_2} ←

Different Numbers

A diagram illustrating independent samples design. It shows two samples: Sample 1 with values x_1, x_2, \dots, x_{n_1} and Sample 2 with values y_1, y_2, \dots, y_{n_2} . Arrows from the end of each sample list point to a rectangular box containing the text "Different Numbers".

- The two samples are independent
- Independent sample design relies on random assignment to make the two groups equal (on the average) on all attributes except for the treatment used (**treatment factor**).

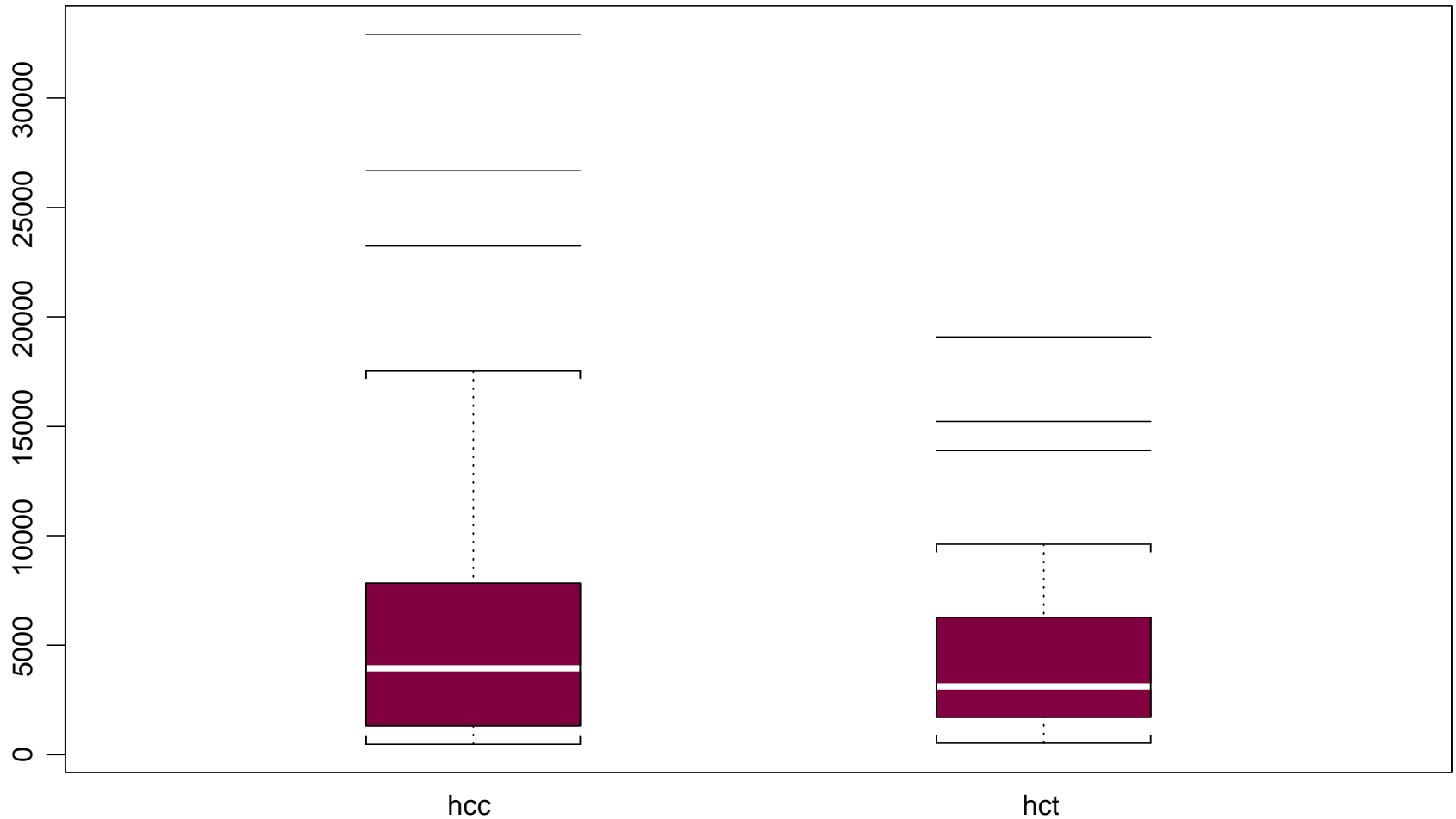
Graphical Methods for Comparing Two Independent Samples

Plot of the order statistics ordered pairs $(x_{(i)}, y_{(i)})$ which are the $\left(\frac{i}{n+1}\right)$ quantiles of the respective samples

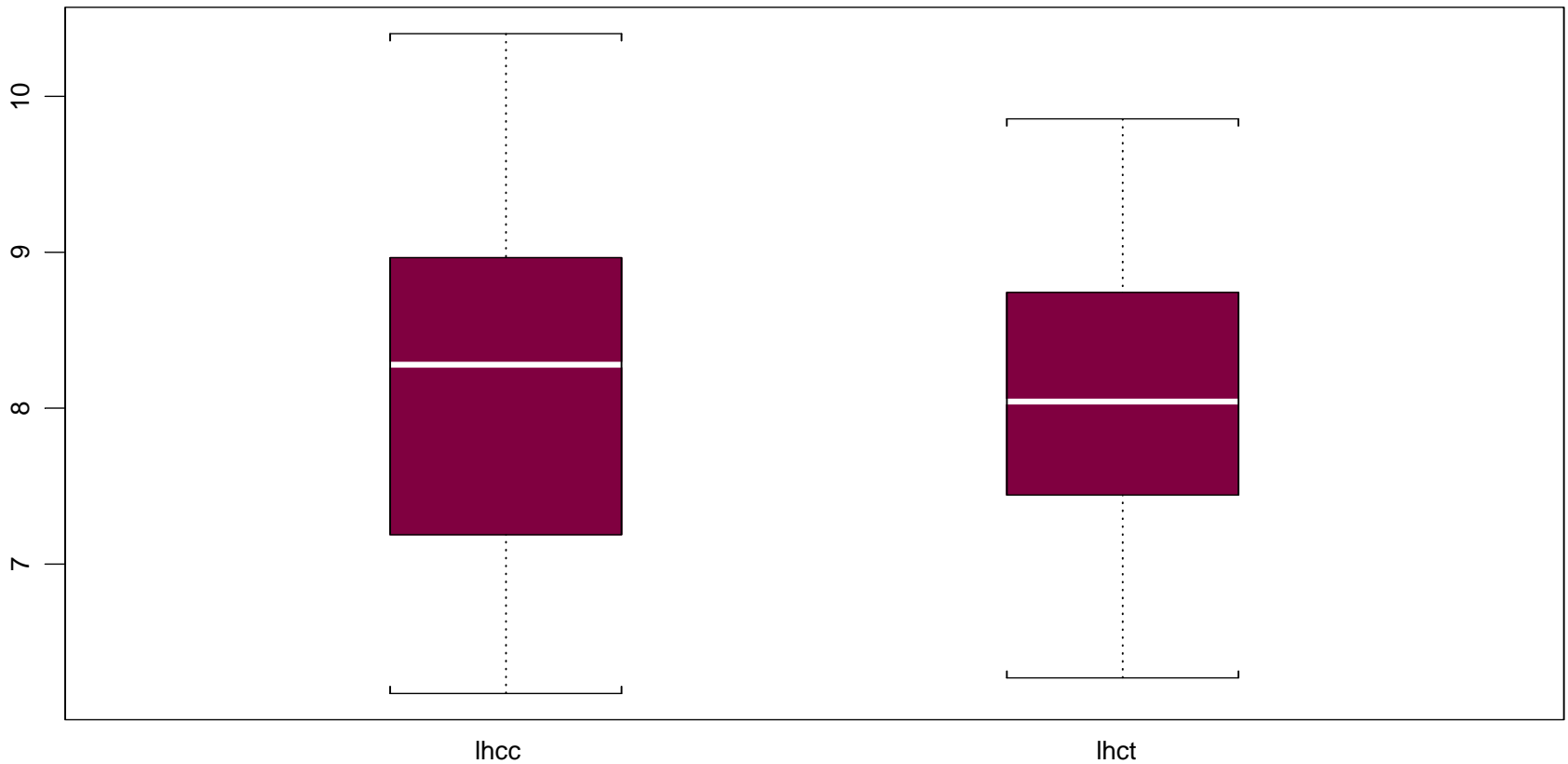
See Table 8.1 and Figure 8.1, which is a Q-Q Plot. Plot suggests that treatment group costs are less than control group costs. But is it true?

Book discusses how to prepare this graph when the two samples are not of the same size (interpolation).

Box plots of hospitalization cost data



Box plots of logs of hospitalization cost data



Graphical Displays of Data from Matched Pairs

- Plot the pairs (x_i, y_i) in a scatter plot. Using the 45° line as a reference, one can judge whether the two sets of values are similar or whether one set tends to be larger than the other
- Plots of the differences or the ratios of the pairs may prove to be useful
- A Q-Q plot is meaningless for paired data because the same quantiles based on the ordered observations do not, in general, come from the same pair.

Comparing Means of Two Populations: Independent Samples Design (Large Samples Case)

Suppose that the observations x_1, x_2, \dots, x_{n_1} and y_1, y_2, \dots, y_{n_2} are random samples from two populations with means μ_1 and μ_2 and variances σ_1^2 and σ_2^2 . Both means and variances are assumed to be unknown. The goal is to compare μ_1 and μ_2 in terms of their difference $\mu_1 - \mu_2$. We assume that n_1 and n_2 are large (say > 30).

Comparing Means of Two Populations: Independent Samples Design

$$E(\bar{X} - \bar{Y}) = E(\bar{X}) - E(\bar{Y}) = \mu_1 - \mu_2$$

$$Var(\bar{X} - \bar{Y}) = Var(\bar{X}) + Var(\bar{Y}) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

Therefore the standardized r.v.

$$Z = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} \quad \text{has mean} = 0 \text{ and variance} = 1$$

If n_1 and n_2 are large, then Z is approximately $N(0,1)$ by the Central Limit Theorem though we did not assume the samples came from normal populations. (We also use fact that the difference of independent normal r.v.'s is also normal.)

Large Sample (Approximate) $100(1-\alpha)\%$ CI for $\mu_1 - \mu_2$

$$(\bar{x} - \bar{y}) - z_{\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \leq \mu_1 - \mu_2 \leq (\bar{x} - \bar{y}) + z_{\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Note s_i^2 has been substituted for σ_i^2 because samples are large, i.e., bigger than 30.

Example 8.2: See Example 8.2 in course textbook.

Large Sample (Approximate) Test of Hypothesis

$H_0 : \mu_1 - \mu_2 = \delta_0$ vs. $H_1 : \mu_1 - \mu_2 \neq \delta_0$ (Typically $\delta_0 = 0$)

$$\text{Test statistics: } z = \frac{(\bar{x} - \bar{y}) - \delta_0}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$$

Inference for Small Samples

Case 1: Variances σ_1^2 and σ_2^2 assumed equal.

Assumption of normal populations is important since we cannot invoke the CLT

Pooled estimate of the common variance:

$$S^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{(n_1 - 1) + (n_2 - 1)} = \frac{\sum (X_i - \bar{X})^2 + \sum (Y_i - \bar{Y})^2}{n_1 + n_2 - 2}$$

Note: $S^2 = (S_1^2 + S_2^2) / 2$ if sample sizes are equal

$$T = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{S \sqrt{1/n_1 + 1/n_2}} \text{ has } t\text{-distribution with } n_1 + n_2 - 2 \text{ d.f.}$$

Inference for Small Sample: Confidence Intervals and Hypothesis Tests

Case 1: Variances σ_1^2 and σ_2^2 assumed equal.

Two-sided $100(1-\alpha)\%$ CI is given by:

$$\bar{x} - \bar{y} - t_{n_1+n_2-2, \alpha/2} s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \leq \mu_1 - \mu_2 \leq \bar{x} - \bar{y} + t_{n_1+n_2-2, \alpha/2} s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Test of Hypothesis: $H_0 : \mu_1 - \mu_2 = \delta_0$ vs. $H_1 : \mu_1 - \mu_2 \neq \delta_0$

Test statistics: $t = \frac{\bar{x} - \bar{y} - \delta_0}{s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$

Reject H_0 if $|t| > t_{n_1+n_2-2, \alpha/2}$

Hospitalization Cost Example

- See Example 8.2 on page 276 of course textbook.

Contrast this conclusion with apparent difference seen on the Q-Q plot in Figure 8.1

t.test in S-Plus to test difference in means of logs of hospitalization cost data

```
t.test(lhcc,lhct)
```

Standard Two-Sample t-Test

data: lhcc and lhct

$t = 0.6181$, $df = 58$, $p\text{-value} = 0.5389$

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-0.3731277 0.7064981

sample estimates:

mean of x mean of y

8.250925 8.08424

Interpretation of Difference in Means on the Log Scale

$$\text{Mean}(\log \text{ Cost}) = \text{Median}(\log \text{ Cost}) = \log(\text{Median Cost})$$

Because distribution of log cost is symmetric

Because the log preserves ordering

This Interpretation is not in your textbook

$$\begin{aligned} -0.373 &\leq \text{Mean}(\log \text{ Cost}_C) - \text{Mean}(\log \text{ Cost}_T) \leq 0.707 \\ -0.373 &\leq \log(\text{Median Cost}_C) - \log(\text{Median Cost}_T) \leq 0.707 \\ -0.373 &\leq \log\left(\frac{\text{Median Cost}_C}{\text{Median Cost}_T}\right) \leq 0.707 \\ .689 &= \exp(-0.373) \leq \frac{\text{Median Cost}_C}{\text{Median Cost}_T} \leq \exp(0.707) = 2.028 \end{aligned}$$

95% confidence interval for the ratio of median costs

Inference for Small Samples

Case 2: Variances σ_1^2 and σ_2^2 unequal.

$$T = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \text{ does not have a Student } t \text{ - distribution}$$

It can be shown that distribution of T depends on the ratio of unknown variances, hence T is not a pivotal quantity. However, when n_1 and n_2 are large T has an approximate $N(0,1)$ distribution

Inference for Small Samples

Case 2: Variances σ_1^2 and σ_2^2 unequal.

For small samples

$$T = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \text{ has an approximately } t\text{-distribution}$$

$$\text{with } \nu = \frac{(w_1 + w_2)^2}{w_1^2/(n_1 - 1) + w_2^2/(n_2 - 1)} \text{ degrees of freedom}$$

$$\text{where } w_1 = (\text{SEM}(\bar{x}))^2 = \frac{s_1^2}{n_1} \text{ and } w_2 = (\text{SEM}(\bar{y}))^2 = \frac{s_2^2}{n_2}$$

Note: d.f. are estimated from the data and are not a function of the samples sizes alone

Note: ν is not usually an integer but is rounded down to the nearest integer

Inference for Small Samples

Case 2: Variances σ_1^2 and σ_2^2 unequal.

Approximate $100(1-\alpha)\%$ two-sided CI for $\mu_1 - \mu_2$:

$$\bar{x} - \bar{y} - t_{v,\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \leq \mu_1 - \mu_2 \leq \bar{x} - \bar{y} + t_{v,\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Test statistics for $H_0 : \mu_1 - \mu_2 = \delta_0$ vs. $H_1 : \mu_1 - \mu_2 \neq \delta_0$

is $t = \frac{\bar{x} - \bar{y} - \delta_0}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$

Reject H_0 if $|t| > t_{v,\alpha/2}$.

Hospitalization Costs: Inference Using Separate Variances

See Example 8.4 on page 280 of course textbook.

t.test in S-Plus to test differences in means of hospitalization data, unequal variances

```
t.test(lhcc,lhct,var.equal=F)
```

Welch Modified Two-Sample t-Test

data: lhcc and lhct

$t = 0.6181$, $df = 54.61$, $p\text{-value} = 0.5391$

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-0.3738420 0.7072124

sample estimates:

mean of x mean of y

8.250925 8.08424

Testing for the Equality of Variances

Section 8.4 covers the classical F test for the equality of two variances and associated confidence intervals. However, this method is not robust against departures from normality. For example, p -values can be off by a factor of 10 if the distributions have shorter or longer tails than the normal.

A robust alternative is *Levene's Test*. His test applies the two-sample t-test to the absolute value of the difference of each observation and the group mean

$$|Y_{1i} - \bar{Y}_1|, i = 1, 2, \dots, n_1$$

$$|Y_{2i} - \bar{Y}_2|, i = 1, 2, \dots, n_2$$

This method works well even though these absolute deviations are not independent.

In the *Brown-Forsythe* test the response is the absolute value of the difference of each observation and the group median.

Independent Sample Design: Sample Size Determination Assuming Equal Variances

$$H_0 : \mu_1 - \mu_2 = 0 \text{ vs. } H_1 : \mu_1 - \mu_2 \neq 0$$

$$n = n_1 = n_2 = 2 \left[\frac{\sigma(z_{\alpha/2} + z_{\beta})}{\delta} \right]^2$$

Because we assume a known variance this n is a slight underestimate of sample size

Smallest difference of practical importance that we want to detect

Using S-Plus to compute sample size

```
normal.sample.size(mean2=.693,power=0.9)
```

```
mean1 sd1 mean2 sd2 delta alpha power n1 n2 prop.n2
```

```
0      1   0.693  1   0.693  0.05  0.9  44 44    1
```


Matched Pairs Design

Example:

See Section 8.3.2, page 283 in course textbook.

Statistical Justification of Matched Pairs Design

See Section 8.3.2, page 283 in course textbook.

Sample Size Determination

$$n = \left[\frac{(z_{\alpha} + z_{\beta})\sigma_D}{\delta} \right]^2 \quad (\text{One-Sided Test})$$

$$n = \left[\frac{(z_{\alpha/2} + z_{\beta})\sigma_D}{\delta} \right]^2 \quad (\text{Two-Sided Test})$$

- One needs a planning value for σ_D
- These formulas come from the one-sample formulas applied to the differences

Comparing Variances of Two Populations

- Application arises when comparing instrument precision or uniformities of products.
- The methods discussed in the book are applicable only under the assumption of normality of the data. They are highly sensitive to even modest departures from normality
- In case of nonnormal data there are nonparametric and other robust methods for comparing data dispersion.

Comparing Variances of Two Populations

Independent sample design:

Sample 1: x_1, x_2, \dots, x_{n_1} is a random sample from $N(\mu_1, \sigma_1^2)$

Sample 2: y_1, y_2, \dots, y_{n_2} is a random sample from $N(\mu_2, \sigma_2^2)$

$F = \frac{S_1^2 / \sigma_1^2}{S_2^2 / \sigma_2^2}$ has an F distribution $n_1 - 1$ and $n_2 - 1$ d.f. respectively

$$P \left\{ f_{n_1-1, n_2-1, 1-\alpha/2} \leq \frac{S_1^2 / \sigma_1^2}{S_2^2 / \sigma_2^2} \leq f_{n_1-1, n_2-1, \alpha/2} \right\} = 1 - \alpha$$

$$P \left\{ \frac{1}{f_{n_1-1, n_2-1, \alpha/2}} \frac{S_1^2}{S_2^2} \leq \frac{\sigma_1^2}{\sigma_2^2} \leq \frac{1}{f_{n_1-1, n_2-1, 1-\alpha/2}} \frac{S_1^2}{S_2^2} \right\} = 1 - \alpha$$

(1- α)-level CI (two-sided):

$$\frac{1}{f_{n_1-1, n_2-1, \alpha/2}} \frac{S_1^2}{S_2^2} \leq \frac{\sigma_1^2}{\sigma_2^2} \leq \frac{1}{f_{n_1-1, n_2-1, 1-\alpha/2}} \frac{S_1^2}{S_2^2}$$

An Important Industrial Application: Example 8.8

(See Table 8.8 in course textbook.)

Do the two labs have equal measurement precision?