I. INTRODUCTION

A. Course objectives

1. Statistical tools: two sample, several sample (ANOVA), multiple comparisons, simple regression

2. Strategies for data analysis

3. Dealing with assumptions

4. Interpretation and communication
B. Creative writing example

1. Data: Students in a creative writing class were randomly assigned to one of 2 groups. One received an “intrinsic” and the other an “extrinsic” questionnaire. Afterwards they wrote poems that were scored for creativity.

2. Question: Did the “intrinsic” questionnaire lead to higher scores?
3. Computer work

Two-sample t Test

Data
- Data Set: Case0101
- Variable 1: SCORE
- Variable 2: CODE

Hypotheses
- Mean Under Null Hypothesis: 0
- Alternative Hypothesis: two.sided

Confidence Interval
- Confidence Level: 0.95

Results
- Save As:
- Print Results

Standard Two-Sample t-Test

data:  x: SCORE with CODE = 1 , and y: SCORE with CODE = 2
t = -2.9259, df = 45, p-value = 0.0054
alternative hypothesis: true difference in means is not equal to
95 percent confidence interval:
-6.996973 -1.291432
sample estimates:
mean of x mean of y
15.73913 19.88333
4. Communication of results

As apparent in the boxplots below, there is strong evidence that receiving the intrinsic motivation questionnaire caused a study subject to receive a higher creativity score than if they would have received an extrinsic questionnaire (two-sided p-value = .005 from a two-sample t-test). The estimated improvement in score was 4.1 points on the 0 to 40 scale (95% confidence interval for improvement: 1.3 to 7 points).
C. Sex Discrimination Example

1. The data: Starting salaries were available for all male and all female entry-level clerical employees of the Harris Bank.

2. One question: Did male starting salaries tend to be larger than female starting salaries? (more so than could be explained by chance?)
3. Computer work
4. Summary of Findings

Boxplots of the male and female salaries appear in the display. The discrepancy in salary distributions is larger than can be explained by chance (p-value < .0001, from a two-sample t-test). It is estimated that the mean salary for males is $820 more than the mean salary for females (95% confidence interval: $560 to $1,080 more).
II. Drawing Conclusions

A. Inference

1. An *inference* is a conclusion from data about some broader context that the data represent.

2. A *statistical inference* is an inference justified by a probability model linking the data to the broader context.
3. Two “broader contexts” in statistics:

a. *Population inference*: an inference about population characteristics, like the difference between two population means.

b. *Causal inference*: an inference that a subject would have received a different numerical outcome had they belonged to a different group.
4. Two probability models linking data to the broader context

a. *Random sampling* means *selection* of population units for a sample, using a chance mechanism.

The chance involved in selection links the sample data characteristics to the population characteristics. (Ch. 2)
b. **Randomization** means assignment of available experimental units to treatment groups using a chance mechanism.

The chance involved in assignment links the data to the causal context. (Details coming soon. Note: an experiment in which randomization is used is called a *randomized experiment*.)
B. Causal Inference

1. Main lesson: statistical inferences of causation can be made from randomized experiments, but not otherwise.

2. Definition: an *observational study* is one in which group status (like male or female) is observed; i.e. beyond the control of the researcher.
3. We cannot make causal inferences from observational studies because of the possibility of confounding variables.
4. Since the sex discrimination problem is an observational study, we may find evidence of an association between sex and salary but we cannot be sure that it is sex that is the cause.
Suppose “intrinsic” creativity scores are higher. This must be due either to the treatment or to the luck of the randomization. ("p-value" is the prob. of the latter)
6. The creativity scores tended to be larger in the “intrinsic” than in the “extrinsic” group. Either the intrinsic questionnaire caused a higher score or else the more creative writers happened to be placed in the “intrinsic” group. The probability associated with this latter possibility is 0.011.
C. Randomization Test $p$-value

1. A hypothesis for the two group randomized experiment is that group status has no effect on the outcome.

2. A test statistic is a numerical data summary for testing a hypothesis. In the creativity study it could be $\bar{Y}_2 - \bar{Y}_1$.

3. The observed value of this test statistic can be different from 0 because
either (a) there is an effect of treatment or (b) the random assignment resulted in an uneven mix.

4. A *randomization test p-value* is the probability associated with explanation (b).

5. The smaller the p-value, the less believable (b) is as an explanation.
D. Exact calculation of the randomization test p-value

1. This is important for understanding the connection of the p-value to the chance mechanism, but...

2. in practice we often use an approximation (the two-sample t-test of Ch. 2)
3. Calculation for creativity study

a. Observed test statistic: \( \bar{Y}_2 - \bar{Y}_1 = 4.14 \)
b. The p-value is the probability that
\[ |\bar{Y}_2 - \bar{Y}_1| \geq 4.14 \] if, in fact, there is no treatment effect (and based on the chance model of randomization)

c. Important starting point: if there is no treatment effect then the creativity score for an individual would have been the same had they been assigned to the other group.
d. Exact calculation of p-value

i. Calculate $\bar{Y}_2 - \bar{Y}_1$ for every possible grouping of the 47 numbers into groups of size 23 and 24

ii. The p-value is the proportion of the regroupings with $|\bar{Y}_2 - \bar{Y}_1| \geq 4.14$

e. Using a computer program, the p-value is .011 (see Display 1.8)
f. Conclusion: either
i. there is no treatment effect and we happened to get an uneven randomization, or
ii. there is a treatment effect

The probability associated with (i) is .011. So, either there is a treatment effect or we obtained an unusual (one-in-a-hundred) randomization.
4. Def: The *randomization distribution* of a statistic describes its possible values over all the ways the randomization could have turned out.

5. The p-value of the randomization test is the proportion of the randomization distribution that is more extreme* than the observed outcome.

* greater than (1-sided), less than (1-sided), or farther from zero (2-sided)
E. Easier, Approximate P-value for 2-group Randomization Test

1. Fact: the randomization distribution of the “t-statistic” (Ch. 2) is approximated well by a t-distribution.

2. Upshot: we will use a “t-test” (justified in Ch. 2 in a random sampling context) but interpret p-value as tied to the chance involved in randomization.
F. Chance Mechanisms in the Real World

1. Types of “2-group” studies:
   a. Randomized experiment (2 groups)
   b. Random samples from 2 populations
   c. Haphazard samples from 2 pops.
   d. Two entire populations are available
2. A *bonafide* statistical inference follows when the probability model is induced by an actual chance mechanism (as in 1 a and b above).

3. The *statistical inference* may be used informally if the chance mechanism is fictitious (with appropriate caution!). E.g.: are the samples just as random as if random sampling was used?
III. Inference using t-distributions

A. Schizophrenia example

1. Observational “Paired data” (2.1.2)

2. Question: is the discrepancy in hippocampus volumes greater than can be explained by chance?

3. Chance model: random sampling (fictitious) from a single population
4. Computer work

1. Select the data set "case0202".
2. Choose the target column to transform and select "difference".
3. Enter the expression "unaffected - affected".
4. Click OK to apply the transformation.
One-sample t-Test

data: difference in case0202
t = 3.2289, df = 14, p-value = 0.0061
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval: 0.0667041 0.3306292
sample estimates: mean of x 0.1986667
B. Paired t-test (one-sample t-test)

1. Review definitions for one population

   a. *Population parameters* \((\mu, \sigma)\)

      Population variance = \(\sigma^2 = \text{ave. size of } (Y - \mu)^2 \text{ in the population}\)

   b. *Sample statistics* \((\bar{Y}, s)\)

      Sample variance, \(s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \bar{Y})^2\)
2. *Sampling distribution* of an average

a. See Displays 2.3 and 2.4

b. Standard deviation of $\bar{Y}$:

$$S.D.(\bar{Y}) = \sigma / \sqrt{n}$$

c. *Standard error* of $\bar{Y}$:  $S.E.(\bar{Y}) = s / \sqrt{n}$

(= estimated standard deviation of the sampling distribution of $\bar{Y}$)
3. **Z-ratio (in general):**

\[
Z\text{-ratio (1-group)} = \frac{\bar{Y} - \mu}{S.D.(\bar{Y})} = \frac{\bar{Y} - \mu}{\sigma / \sqrt{n}}
\]

4. **t-ratio (in general):**

\[
t\text{-ratio (1-group)} = \frac{\bar{Y} - \mu}{S.E.(\bar{Y})} = \frac{\bar{Y} - \mu}{s / \sqrt{n}}
\]
5. Distribution of the t-ratio

a. Fact from statistical theory:
   If* the population distribution of $Y$ is normal then the sampling distr. of
   i. the z-ratio is Standard Normal
   ii. the t-ratio is Student’s t on $n-1$ degrees of freedom

*(We will study the “If” part later)

b. See Display 2.5
6. Testing a hypothesis about $\mu$

a. Could the difference of $\bar{Y}$ from $\mu^*$ (a hypothesized value for $\mu$) be due to chance (in random sampling)?

b. *Null hypothesis*, $H_0$: $\mu = \mu^*$

c. *t-statistic* for $H_0$: \[
\frac{\bar{Y} - \mu^*}{S.E.(\bar{Y})}
\]

d. If $H_0$ is true then t-statistic = t-ratio
e. The (2-sided) p-value is the proportion of random samples with t-ratios as “far from” zero as the t-statistic.

Schizophrenia example:

t-statistic = 3.23
7. Schiz. ex.: $\mu = \text{mean diff.}$ \hspace{1cm} H$_0$: $\mu = 0$

a. p-value (2-sided, paired t-test) = .006

b. So, either

i. the null hypothesis is incorrect, OR

ii. the null hypothesis is correct and we happened to get one of those samples with a particularly unusual $\bar{Y}$. (Only 6 out of 1000 are as unusual)
c. Some S-PLUS “commands”

Do calculations on this data
“<-” means “is assigned”
Average of the variable “diff”
Sample variance

pt(3.2,14) gives Pr(t_{14} < 3.2)

Note: these calculations were done automatically on p. 31
8. **Confidence interval** for $\mu$

a. Fact: 95% of all possible samples (of size 15) have t-ratios between -2.145 and 2.145 (see Display 2.5)

b. *If* the observed sample is one of these “usual” ones then

$$-2.145 < \frac{0.199 - \mu}{0.0615} < 2.145$$

or $0.067 \text{ cm}^3 < \mu < 0.331 \text{ cm}^3$ (95% C.I.)
c. Formula for 95% confidence interval for a single mean:

\[ \bar{Y} - \text{halfwidth to } \bar{Y} + \text{halfwidth} \]

where \( \text{halfwidth} = t_{14} \times (.975) \times SE(\bar{Y}) \)

97.5th percentile of \( t_{14} \) distribution

d. Success rate of procedure = 95%
C. Bumpus example

1. Structure: 2 independent samples observational study (Section 2.1.1)

2. Question: Do humerus bone lengths tend to be different in the birds that survived than in those that perished?

3. Computer work:
St 411/511, Schafer  
Winter, 2003
Two-sample t Test

Data
Data Set: Case201
Variable 1: HUMERUS
Variable 2: CODE

Hypotheses
Mean Under Null Hypothesis:
Alternative Hypothesis:
two-sided

Confidence Interval
Confidence Level: 0.95

Results
SaveAs:

OK Cancel Apply

Standard Two-Sample t-Test

data: x: HUMERUS with CODE = 1 , and y: HUMERUS with CODE = 2
t = -1.777, df = 57, p-value = 0.0809
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-21.446053 1.279386
sample estimates:
mean of x mean of y
727.9167 738
D. Two-sample t-test

1. Population parameters $\mu_1, \sigma_1, \mu_2, \sigma_2$

2. Equal spread model: $\sigma_1 = \sigma_2$ (call it $\sigma$)

3. Statistics from samples of size $n_1$ and $n_2$ from pops. 1 and 2: $\bar{Y}_1, \bar{Y}_2, s_1^2$ (n$_1$-1 d.f.), $s_2^2$ (n$_2$ - 1 d.f.)
4. **Pooled est. of $\sigma^2$:**

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

degrees of freedom = $(n_1 - 1) + (n_2 - 1)$

(See Display 2.8 for an example of calculations.)
5. Parameter of interest: $\mu_2 - \mu_1$

6. Sampling distribution of $\overline{Y}_2 - \overline{Y}_1$
   (from stat. theory): see Display 2.7

7. $SD(\overline{Y}_2 - \overline{Y}_1) = \sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$ ({$\sigma_1 = \sigma_2$ model})

8. $SE(\overline{Y}_2 - \overline{Y}_1) = s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$ (df = $n_1 + n_2 - 2$)
9. If population distributions are normal with equal \( \sigma \), then the t-ratio for

\[ \bar{Y}_2 - \bar{Y}_1 \]  
has a \( t_{n_1+n_2-2} \) distribution

10. T-test for \( H_0: \mu_2 - \mu_1 = \theta^* \)

\[ t\text{-statistic} = \frac{(\bar{Y}_2 - \bar{Y}_1) - \theta^*}{S.E.(\bar{Y}_2 - \bar{Y}_1)} \]
11. $(1-\alpha)\times100\%$ confidence interval:

$$\bar{Y}_2 - \bar{Y}_1 \pm t_{n_1+n_2-2}[1-(\alpha/2)] \times SE(\bar{Y}_2 - \bar{Y}_1)$$

12. 95% confidence interval ($\alpha=.05$):

$$(\bar{Y}_2 - \bar{Y}_1) \pm t_{n_1+n_2-2}(.975) \times SE(\bar{Y}_2 - \bar{Y}_1)$$
E. Notes about tests, p-values

1. Interpretation of p-value:
   a. Formally: the probability of random sampling (or randomization) leading to a test statistic as extreme or more extreme than the observed one, if $H_0$ is true
   b. Informally: the degree of credibility in $H_0$
2. Conclusions from p-values

a. Small p-values mean:
   either $H_0$ is wrong or we obtained an unusual sample (or random assign.)

b. Large p-values mean:
   either $H_0$ is correct or the study isn’t strong enough to conclude otherwise (i.e. the data are consistent with $H_0$ being true; but do not prove it)
3. So what p-values are small and large?

a. For reference: chance of 3 heads in 3 coin tosses is 0.125

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>4</td>
<td>0.063</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>0.031</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>0.016</td>
</tr>
<tr>
<td>7</td>
<td>7</td>
<td>0.008</td>
</tr>
<tr>
<td>8</td>
<td>8</td>
<td>0.004</td>
</tr>
</tbody>
</table>

b. See Display 2.12 for subjective guide
4. Examples

a. Bumpus: p-value = .08 implies... suggestive but inconclusive evidence that mean humerus lengths differ

b. Sex discrimination: p-value < .0005... overwhelming evidence that males and females had different mean salaries

(Note: also report C.I. in both cases)
5. *Rejection region approach* to testing

Reject $H_0$ at level $\alpha$ if the t-statistic is in the rejection region, otherwise “fail to reject” (Usually $\alpha=.05, .01, \text{or } .001$)
6. Why the rejection region should not be used in practice

a. p-values of .049 and .051 are essentially equivalent measures of credibility in $H_0$ but lead to opposite “rejection” conclusions (at level .05)

b. p-values of .049 and .0000001 are very different measures of credibility in $H_0$ but lead to the same conclusion
7. **One-sided and two-sided tests**

What is the evidence against

\[ H_0: \mu_2 - \mu_1 = 0 \]

if the alternative possibility is

a. \( H_1: \mu_2 - \mu_1 > 0 \) (one-sided)

b. \( H_1: \mu_2 - \mu_1 \neq 0 \) (two-sided)
IV. A Closer look at assumptions

A. Introduction

1. The ideal *assumptions* upon which the t-tests and CIs are based are never met
2. The tools are still quite usable if we...
   a. understand their *robustness* and *resistance*
   b. consider transformation, e.g. log(Y)
c. have a strategy for outliers
d. be prepared to label inference as “approximate”

3. Additionally, we have alternative tools (Ch. 4)

4. Cloud seeding and rainfall example
   a. See Displays 3.1 and 3.2
   b. Computer work:
Standard Two-Sample t-Test
data: x: lograin with code = 1 , and y: lograin with code = 2
t = -2.5444, df = 50, p-value = 0.0141
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  -2.046697  -0.240865
sample estimates:
mean of x  mean of y
  3.990406  5.134187

1. transform
2. Change box color, style
3. Edit axis title: left click on title, then right click
B. Robustness of two-sample t-tools (p-value, CI)

1. A statistical procedure is *robust to departures from a particular assumption* if it is valid even when the assumption is not met exactly.

2. Statisticians know something about robustness from advanced theory and from computer simulation.
3. Assumptions for 2-sample t-tools
   
a. Population distributions are normal
   
b. \( \sigma_1 = \sigma_2 \)
   
c. Observations are independent
      
      i. within groups
      
      ii. between groups
      
      (*independent*: knowledge of one observation can’t help to predict another)
4. How important is normality?

a. Distributions skewed, same amount: p-value and CI are still ok if \( n_1 \approx n_2 \) or if both \( n_1 \) and \( n_2 \) are large

b. Populations have different skewnesses: Tools ok if \( n_1 \) and \( n_2 \) are both large

c. Distributions are long-tailed (so extreme outliers are present):
t-tools are not reliable in this case.

d. See Display 3.4 for clarification of “large” n in some situations (via simulation).

5. How important is equal-σ assumption?

a. t-tools are still pretty good when

$$\sigma_1 \neq \sigma_2 \text{ if } n_1 \approx n_2$$
b. otherwise, they may be misleading
c. See Display 3.5

6. How to deal with non-independence
   a. Common violations of independence assumption
      i. *Cluster effects* (Y’s from same
cluster are similar; Ch. 9-14)

ii. *Serial effects* (Y’s close together in time or space are similar; Ch. 15)

b. If we suspect non-independence, then we turn to Ch. 9-15 techniques

C. Outliers and resistance

1. *Outliers* are observations relatively far from their estimated means

2. Outliers may arise either
a. if the pop. distribution is long-tailed
b. or if they don’t belong to the populations of interest (they come from a contaminating population)

3. A statistical procedure is resistant if one or a few outliers can not have an undue influence on its result

4. Illustration for understanding resistance: the sample average is not resistant
tant; the sample median is

Sample: 9, 3, 5, 8, 100
Average with outlier: 25, without: 6.2
Median with outlier: 8, without: 6.5

5. t-tools (which are based on averages) are not resistant to outliers
6. Two approaches for two-sample comparison when there are suspected outliers

a. Use t-tools with outlier examination strategy (see box in Section 3.4)

b. Use resistant alternatives (like the rank-sum test in Ch. 4)

7. Comparison of the two approaches
a. Sometimes more information can be learned from the t-tools with outlier strategy

b. Rank-sum test is convenient for bypassing the outlier problem

c. Note: t-tools extend to more complicated data structures

D. Practical two-sample strategy
1. *Think* about independence; use Ch.9-15 tools if there’s a potential problem

2. Use *graphical displays* to assess: *normality, spread, outliers* (subjectively)

3. Consider transforming the data

4. Assess whether skewness, sample size, and relative spreads permit t-tools
a. validity not affected much by non-normality, except when population distributions are long-tailed

b. validity not affected much by unequal spreads if $n_1 \approx n_2$

5. If there are outliers, use the outlier strategy (or go to Ch. 4 alternatives)

6. Example: Dioxin levels in veterans; see Display 3.3, 3.7 & computer work:
1. Create columns without outlier
   - code1 gets code if dioxin < 40
   - dioxin1 gets dioxin if dioxin < 40

2. Do t-tools with and without outlier

---

Standard Two-Sample t-Test (with complete data)

data:  x: dioxin with code = 1 , and y: dioxin with code = 2
  t = 0.263, df = 741, p-value = 0.7926
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-0.4615229  0.6305128

Standard Two-Sample t-Test (with largest value removed)

data:  x: dioxin1 with code1 = 1 , and y: dioxin1 with code1 = 2
  t = 0.0489, df = 740, p-value = 0.961
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-0.4435973  0.4662617
E. Transformation

1. Transformation (like $\log(Y)$, $1/Y$) is a good idea if ...

a. the transformed data more nearly meet the ideal, two-sample model, and

b. the questions of interest can be answered
2. A note about robustness:
The robustness results indicate that the t-tools may be valid even if the ideal assumptions aren’t met. Even if they are valid, though, there might be some other tool that is better (e.g. narrower confidence interval).

So, if transformed scale works better, use it rather than count on robustness.
3. Illustrations

a. Log for moderately skewed data

b. Reciprocal for severely skewed data
4. The most important transformation: \[ \log(Y) \text{ [or } \log_{10}(Y)\text{]} \] (for \( Y > 0 \))

a. Notes about logs:
   i. \( \log_{10}(10) = 1, \log_{10}(100) = 2, \text{ etc.} \)
   \[ \log(2.718) = 1, \log(2.718^2) = 2, \text{ etc.} \]
   ii. \( \log(a \times b) = \log(a) + \log(b) \)

b. Why it’s important:
   i. Many real data more closely meet the ideal model after log transform
ii. it permits interpretation about \textit{multiplicative effects}

Example of \textit{additive effect}: the mean salary for males is $500 more than the mean for females $(\mu_{\text{male}} = \mu_{\text{fem}} + 500)$

Example of \textit{multiplicative effect}: the median salary for males is 15\% more than the median for females $(\text{Med}_{\text{male}} = \text{Med}_{\text{fem}} \times 1.15)$
5. What indicates that log might help?

a. Distributions are skewed

b. Spread is greater in the distribution with larger center

c. The data values differ by orders of magnitude; e.g., as a *rough* guide, the ratio of the largest to the smallest is > 10 (or perhaps > 4)

d. Multiplicative statement is desirable
6. Procedure for two samples
   a. Initial inspection and indicators in 5 above may cause one to try log(Y)
   b. Transform to get two new columns:
      \[ Z_1 = \log(Y_1), \quad Z_2 = \log(Y_2) \]
   c. Graphically examine \( Z_1, Z_2 \)
   d. If appropriate, use t-tools on \( Z_1, Z_2 \)
   e. Interpret results on original scale
7. Interpretation (population version)

a. Testing: if dist’s of $\log(Y_1)$ & $\log(Y_2)$ differ, dist’s of $Y_1$ & $Y_2$ also differ

b. Calculate $w = [\exp(\bar{Z}_2 - \bar{Z}_1) - 1] \times 100$

If $w > 0$: “It is estimated that the median for pop. 2 is $w\%$ greater than the median for pop. 1.”

If $w < 0$: “...median for pop. 2 is $-w\%$ less than the median for pop. 1”
8. Explanation of interpretation

a. First, note: \( \mu\{\log(Y)\} \neq \log\{\mu(Y)\} \)
   (the mean of the log-transformed data is not equal to the log of the mean of the un-transformed data)

   ex: \( Y: 1 \ 10 \ 100 \ \text{ave} = 37 \)
   \( \log_{10}(Y): 0 \ 1 \ 2 \ \text{ave} = 1 \neq \log(37) \)

b. But \( Med\{\log(Y)\} = \log\{Med(Y)\} \)
   \( Med(Y) = 10, \ Med\{\log_{10}(Y)\} = 1 \)
c. If a population distribution is symmetric, then the mean = the median

d. So, if dist. of log(Y) is symmetric

\[ \mu\{\log(Y)\} = Med\{\log(Y)\} = \log\{Med(Y)\} \]

because of c

because of b

e. If \( Z_1 = \log(Y_1), Z_2 = \log(Y_2) \), \( \bar{Z}_2 - \bar{Z}_1 \)

estimates \( \mu\{\log(Y_2)\} - \mu\{\log(Y_1)\} \)
f. If the distributions of $Z_1$ and $Z_2$ are symmetric, $\bar{Z}_2 - \bar{Z}_1$ also estimates

$$\text{Med}\{\log(Y_2)\} - \text{Med}\{\log(Y_1)\},$$

which is

$$\log\{\text{Med}(Y_2)\} - \log\{\text{Med}(Y_1)\},$$

which is

$$\log\{\text{Med}(Y_2) / \text{Med}(Y_1)\}$$

because $\log(a) - \log(b) = \log(a/b)$

g. So $\exp(\bar{Z}_2 - \bar{Z}_1)$ estimates

$$\text{Med}(Y_2) / \text{Med}(Y_1)$$
h. We say: “it is estimated that ... the median for pop. 2 is \( \exp(\bar{Z}_2 - \bar{Z}_1) \) times the median of pop. 1.”

i. Or, if \( \bar{Z}_2 - \bar{Z}_1 \) is positive: “... the median for pop. 2 is \( \left[ \exp(\bar{Z}_2 - \bar{Z}_1) - 1 \right] \times 100\% \) more than the med. of pop. 1.”

j. If, for example, \( \exp(\bar{Z}_2 - \bar{Z}_1) = 1.15 \), “... median for pop. 2 is 15\% more than the median for pop. 1”
F. Final Notes on Chap. 3

1. A difficulty: checking *population characteristics* using *samples*

9 two-sample data sets \( (n_1=30, n_2=30) \) from normal populations with equal spread
2. The problem of assessment is most difficult with small sample sizes

9 two-sample data sets (n_1=30, n_2=30) from populations that are normal with equal spread after log transformation
V. Alternatives to t-Tools
A. Wilcoxon Rank-Sum Test

1. Introduction
   a. Resistant alternative for 2 samples
   b. It’s distribution-free (nonparametric): not based on dist’n assumption
   c. Usually “better” that t-test for non-normal pops. and ...
d. ... nearly as good for normal ones

e. Useful for both randomized experiments and obs. studies

f. Drawbacks: i. C.I. is difficult to get
   ii. It doesn’t extend (for Ch. 7-15)

2. The Idea:

   a. Replace the data by their ranks in the combined sample
b. Compute a test statistic based on the ranks (like difference in ave. ranks or the sum of ranks in one group)
c. Find a p-value from the permutation distribution of the test statistic, or an approximation

3. Cognitive load example
   (See Displays 4.2-4.7 and computer-work that follows)
Wilcoxon rank-sum test

data: x: time with treatment = 1, and y: time with treatment = 2
rank-sum normal statistic with correction Z = -3.0183, p-value = 0.0025
alternative hypothesis: true mu is not equal to 0
4. Continuity correction to improve normal approximation

- If uncorrected Z-stat is $> 0$, subtract .5 from numerator.
- If Z-stat is $< 0$, add .5 to numerator.

Dist. of $T$ is “discrete” we want area of boxes above and to the right of $T$.

Approximate this by area under continuous normal curve.

If uncorrected Z-stat is $> 0$, subtract .5 from numerator. If Z-stat $< 0$, add .5 to numerator.
5. Small example

\textbf{samp1:} 3, 9, 12 \hspace{1cm} \textbf{ranks:} 2, 4, 5

\textbf{samp2:} 1, 4 \hspace{1cm} \textbf{ranks:} 1, 3

\textbf{rank-sum} = T = 2 + 4 + 5 = 11

\underline{Normal approx:}

\textbf{R} = 3.0, \textbf{s}_R = 1.58

\textbf{Mean}_0(T) = 3 \times 3.0 = 9

\textbf{SD}_0(T) = 1.58 \times (3 \times 2/5)^{0.5} = 1.73

\textbf{Z} = (11-9)/1.73 = 1.15; 1-sided p-value = .12

\textbf{Z}_{cc} = (11-9-.5)/1.73 = .86; p-value = .19
Possible Values for Sum of Ranks for Group 1

2-sided p-value = proportion as far or farther from mean(T) as observed T is

1-sided p-value = .2
B. Review: p-value can mean the...

1. chance that random sampling alone explains such an extreme test statistic*

2. chance that randomization alone explains such an extreme test statistic*

*test statistic may be: t-stat, $\bar{Y}_2 - \bar{Y}_1$, or rank-sum
3. For observational studies in which random sampling was not used, we can interpret the p-value as either:

a. same as #1, supposing samples are just as good as random samples

b. the chance that a random reshuffling of the available numbers leads to such an extreme test statistic (permutation test) (good if there are no pops.)
C. Permutation tests

1. p-value = proportion of groupings (into groups of $n_1$ and $n_2$) that lead to a test statistic as extreme as the observed one

2. *Permutation test* is the general name; *randomization test* is the name when applied to a randomized experiment (special name—stronger inference)
3. Calculation

a. Exact p-value can sometimes be calculated by doing the regroupings manually (or with computer) or by using combinatorics (p. 91-92; we’ll skip that)

b. If the t-stat is a suitable test statistic, then the t-dist’n is a good approximation for large n’s (use usual t-tools)
4. Importance

a. Conceptual: permits interpretation for randomized experiments and for internal significance of obs. studies with no populations (B3b above)

b. Idea used in rank-sum & other tests

c. Last resort (difficult calculations)

5. Space shuttle (Displays 4.1, 4.9, 4.10) (don’t worry about the calculations)
D. Welch’s t-test

1. Doesn’t assume equal variances

2. It’s the usual t-test and CI, but with

\[ SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_1}} \]; d.f. (approx.) = (p. 93)

3. Drawbacks: a. conceptual difficulty (Display 4.11); b. theory is approximate; c. doesn’t extend well
4. S-PLUS

5. Advice: prefer standard t-tools; consider transformation, robustness
E. Dist’n Free Tests for Paired Data

1. Sign test (p. 95): quick, easy, weak

2. Signed-rank test (p. 96):
   Good alternative to paired t-test

---

for paired data

not “exact”
F. Further Issues

1. Statistical significance ≠ practical significance

   a. p-value indicates credibility of $H_0$ in light of avail. data, but not the practical importance of the finding

   b. p-value depends on sample size: If $H_0$ is false, then as sample sizes increase, p-value decreases
c. So: report a C.I. along with a p-value (Is size of effect *practically* meaningful?)

d. Use the word “significance” cautiously. Suggestion: if you are talking about statistical significance then say “statistical significance”

2. Presentation of Findings (p. 98)
G.  Review of Chaps. 1-4

1. Main topics: causation, 2-sample t-test, paired t-test, p-value, C.I., robustness, resistance, outlier strategy, log transformation, rank-sum test

2. Need to recognize in a data problem:
   a. Paired or two independent samples?
   b. Randomized exper. or obs. study?
3. Factors affecting width of a C.I.

\[ (\bar{Y}_2 - \bar{Y}_1) \pm t_{df} \left( 1 - \frac{\alpha}{2} \right) \frac{s}{\sqrt{n_1} + \frac{1}{n_2}} \]

- Larger confidence level \( \Rightarrow \) wider interval
- Improved measurement (if possible) \( \Rightarrow \) smaller \( \sigma \), narrower
- Larger \( n_1, n_2 \) \( \Rightarrow \) narrower

\[ \text{SE}(Y_2 - Y_1) \]
H. Review of Ch. 1-4 case studies

1. Motivation and creativity
   a. Randomized exper; 2 treatments
   b. Use 2-sample t-test as an approximation to randomization test
   c. Causation implied

There is convincing evidence that for this group of volunteers an “intrinsic” questionnaire lead to a greater creativity score than an “extrinsic” questionnaire (two-sided p-value from a two-sample t-test = .005). The amount by which the score would be greater is estimated to be 4.1 points on the 0 to 40 point scale (95% confidence interval: 1.3 to 7.0 points).
2. Sex Discrimination

a. Two-sample observational study
b. No causation (confounding vars.?)
c. The groups were not random samples from populations.

There is convincing evidence that the disparity between the male and female salaries in these employees was larger than could be explained by chance alone (one-sided p-value < .00001 from a two-sample t-test). The estimated size of the difference is $820 (95% confidence interval: $560 to $1,080).
3. Bumpus’s sparrows
   a. Two sample, observational study
   b. No causation

There is suggestive but inconclusive evidence that humerus bones tended to be longer in the birds that survived than in those that perished (two-sided p-value = .08 from a two-sample t-test). A 95% confidence interval for the survivor mean minus the non-survivor mean is -.001 to .021 inches.
4. Twins schizophrenia study
   a. Single sample of pairs; obs. study
   b. Paired t-test (1-sample t-test on differences)
   c. Could have also used signed-rank test

There is convincing evidence that the left hippocampus volumes were larger in the unaffected twins than in the schizophrenic twins (two-sided p-value = .006 from a paired t-test). A 95% confidence interval for the mean difference is .07 to .33 cm$^3$. 
5. Cloud seeding and rainfall
  a. 2 treatment randomized experiment
  b. Use 2-sample t-test (as approx. to randomization test) on log rainfall
  c. Diff. in aves. of log rain = 1.14; back-transform: \( \exp(1.14) = 3.10 \) (310%)

There is convincing evidence that seeding the clouds caused more rainfall (one-sided p-value = .007 from a two-sample t-test on log rainfall). It is estimated that a cloud would produce 210% more rain if seeded than if unseeded (95% confidence interval: 30% more to 670% more).
6. Agent orange in Vietnam Vets
   a. Two-sample observational study
   b. Results do not change when unusually large dioxin values are set aside

The data are consistent with the theory that the mean dioxin levels are the same in Vietnam and other veterans (one-sided p-value = .4 from a two-sample t-test). A 95% confidence interval provides a set of all plausible values for the difference in means (Vietnam minus others): -.48 to .63 parts per trillion. These results may be biased if the samples are not truly representative of the target populations.
7. Space shuttle

a. Two-sample, observational study

b. Conditions inappropriate for t-test; too many ties for rank-sum test

The greater number of O-ring incidents in launches under 65 degrees than in launches over 65 degrees was more than can be explained by chance alone (one-sided p-value = .01 from a permutation test of the two-sample t-statistic).
VI. Comparing Several Groups

A. Introduction

1. Ch. 5-6: Tools and Issues for Comparisons among I (≥ 2) groups

2. Ch. 5 lessons:
   a. Pooled estimate of $\sigma$ from all groups
   b. One-way analysis of variance & “F-test” for $H_0: \mu_1 = \mu_2 = \ldots = \mu_I$
B. Example: diet restriction

1. Randomized experiment with 6 treatments (see Displays 5.1-5.3)

2. Equal spread ($\sigma_1=\sigma_2=...=\sigma_6$) model seems okay to use (from boxplots)

3. Pool estimates of $\sigma$ from all groups: $s_p$

4. Initial screening: $H_0$: $\mu_1=\mu_2=...=\mu_6$

5. T-tests for 5 planned comparisons
Test equality of all groups
<table>
<thead>
<tr>
<th></th>
<th>Estimate</th>
<th>Std.Error</th>
<th>Lower Bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>N/N85-N/R40</td>
<td>-12.400</td>
<td>1.24</td>
<td>-14.900</td>
</tr>
<tr>
<td>N/N85-N/R50</td>
<td>-9.610</td>
<td>1.19</td>
<td>-11.900</td>
</tr>
<tr>
<td>N/N85-R/R40</td>
<td>5.290</td>
<td>1.30</td>
<td>2.730</td>
</tr>
<tr>
<td>N/N85-R/R50</td>
<td>-10.200</td>
<td>1.26</td>
<td>-12.700</td>
</tr>
<tr>
<td>N/N85-lopro</td>
<td>-6.990</td>
<td>1.26</td>
<td>-9.470</td>
</tr>
<tr>
<td>N/R40-N/R50</td>
<td>2.820</td>
<td>1.17</td>
<td>0.516</td>
</tr>
<tr>
<td>N/R40-NP</td>
<td>17.700</td>
<td>1.29</td>
<td>15.200</td>
</tr>
<tr>
<td>N/R40-lopro</td>
<td>2.230</td>
<td>1.24</td>
<td>-0.210</td>
</tr>
<tr>
<td>N/R50-NP</td>
<td>5.430</td>
<td>1.24</td>
<td>2.990</td>
</tr>
<tr>
<td>N/R50-R/R50</td>
<td>14.900</td>
<td>1.24</td>
<td>12.500</td>
</tr>
<tr>
<td>N/R50-lopro</td>
<td>2.610</td>
<td>1.19</td>
<td>0.264</td>
</tr>
</tbody>
</table>
C. Populations and samples

Population 1 distribution

Population 2 distribution

Population 3 distribution

Random samples

Samples statistics

$\mu_1$

$\mu_2$

$\mu_3$

$Y_{11}, Y_{12}, ..., Y_{1n_1}$

$Y_{21}, Y_{22}, ..., Y_{2n_2}$

$Y_{31}, Y_{32}, ..., Y_{3n_3}$

$\bar{Y}_1 = 63.1$

$s_1 = 6.0$

$\bar{Y}_2 = 74.2$

$s_2 = 7.6$

$\bar{Y}_3 = 62.7$

$s_3 = 6.7$
D. Comparing any 2 means

1. Suppose interest is in $\mu_3 - \mu_2$

2. Sampling distribution of $Y_3 - Y_2$ has...

   a. mean = $\mu_3 - \mu_2$

   b. S.D. = $\sigma \sqrt{\frac{1}{n_3} + \frac{1}{n_2}}$, if $\sigma$'s are equal

3. Do t-test, C.I. as usual but estimate $\sigma$ from all groups, use df = $n-I$ (Sect. 5.2)
E. Extra-sum-of-squares F-test

1. Introduction

a. Primary interest in Ch. 5 is in 1-way ANOVA F-test for $H_0: \mu_1=\mu_2=...=\mu_I$

...which is an application of the more general extra-sum-of-squares F-test

b. We will learn the general principle and terminology; with particular attention to 1-way ANOVA
2. Example-Spock Trial

a. See Display 5.4

b. Steps in analysis
   i. Initial graphical exploration
   ii. Initial screening $H_0: \mu_1 = \mu_2 = \ldots = \mu_7$
   iii. Test equality of judges 2-7:
        $H_0: \mu_2 = \ldots = \mu_7 (= \mu, \text{ say})$
   iv. Compare Spock’s judge’s mean to other judges mean, $H_0: \mu_1 = \mu$
A factor is a code for groups. To use ANOVA, a numerical code must be converted to the data type “factor.”
1-way ANOVA F-Test p-value

Response variable
Grouping factor
3. Full and Reduced Models

a. A *full model* is one we believe adequately describes the data.

b. The *reduced model* is the special case of the full model if $H_0$ is true.

c. For $H_0$: $\mu_1 = \mu_2 = \ldots = \mu_7$

Group: 1 2 3 4 5 6 7

Full model: $\mu_1 \mu_2 \mu_3 \mu_4 \mu_5 \mu_6 \mu_7$

Red. model: $\mu \mu \mu \mu \mu \mu \mu \mu \mu \mu \mu$
4. A *residual* is the difference of an observation from its estimated mean (according to some model)

a. Residuals from full (separate means) model have the form: \( Y_{ij} - \bar{Y}_i \)

\( jth \ obs. \ in \ group \ i \) \quad \( Sample \ ave. \ from \ group \ i \)

b. Residuals from the reduced (single mean) model: \( Y_{ij} - \bar{Y} \) \quad \( Grand \ Average \)
5. A test statistic for $H_0: \mu_1 = \mu_2 = \ldots = \mu_7$

a. If $H_0$ true, then full=reduced model

b. If $H_0$ true, residuals from both models should be about the same size

c. A possible test statistic: *extra sum-of-squares* = sum of squared residuals from reduced model minus sum of squared residuals from full model (see Display 5.8)
d. This should be close to zero if $H_0$ is true, and large (positive) otherwise.

6. **Extra SS F-statistic** is a scaled version of the extra SS, whose sampling distribution is an **$F$-dist.**, if $H_0$ is true.

$$F-stat = \frac{(\text{Extra SS})/(\text{Extra df})}{\hat{\sigma}^2_{\text{full}}}$$

- $SS(\text{reduced}) - SS(\text{full})$ (Best est. of $\sigma^2$, full model)
- $\hat{\sigma}^2_{\text{full}}$
- $\text{df}(\text{reduced}) - \text{df}(\text{full})$
7. If $H_0$ is true, then the F-stat has an $F_{d_1,d_2}$ distribution
where $d_1 =$ numerator d.f. =
# of par. for mean in full model -
# of par. for mean in reduced model
d2 = denominator d.f. =
d.f. associated with $\sigma^2_{\text{full}}$
(see Display 5.8 and Appendix A.4)
(Sec. 5.4.1 details assumptions used)
8. The p-value for $H_0$ is the prob. of obtaining a sample with an F-statistic $\geq$ the observed one, if $H_0$ true

9. Note, If “full”= “separate means”

$$\hat{\sigma}^2_{full} = s_p^2 = \frac{(n_1 - 1)s_1^2 + \ldots + (n_7 - 1)s_7^2}{(n_1 - 1) + \ldots + (n_7 - 1)} =$$

$$= \frac{\sum_{j=1}^{n_1}(Y_{1j} - \bar{Y}_1)^2 + \ldots + \sum_{j=1}^{n_7}(Y_{7j} - \bar{Y}_7)^2}{n - 7} = \frac{(Y_{11} - \bar{Y}_1)^2 + (Y_{12} - \bar{Y}_1)^2 + \ldots + (Y_{7n_7} - \bar{Y}_7)^2}{n - 7}$$

= SS residuals (full)/[d.f.(full)]
10. Calculations for Spock data (n=46)
   a. Sum of squared residuals from full model with 7 separate means: 1864.45
   b. Sum of squared residuals from reduced model with 1 mean: 3791.53
   c. Extra SS = 3791.53 - 1864.45 = 1927.08
   d. Extra df = 7 - 1 = 6
   e. F-stat = \[ \frac{(1927.08)/6}{(1864.45/39)} \]
      = 6.72
f. p-value = 0.000061

(See output on p. 124 above)

Note: In S-PLUS commands window, get the right tail area with:
> 1 - pf(6.72, 6, 39)
F. Special Considerations for One-Way ANOVA F-test with I groups

1. $H_0$: $\mu_1 = \mu_2 = \ldots = \mu_I$

2. $H_a$: at least 1 mean is different

3. Used for initial screening (some further analysis usually follows)

4. Based on ideal normal, constant variance model (robustness: Sect. 5.5)
5. Notation:
   a. SS Full = Within-groups SS (df = n-1)
   b. SS Reduced = Total SS (df = n-1)
   c. ExtraSS = Between-groups SS (df = I-1)
   d. Mean Square = SS/df

6. ANOVA table (see Display 5.9)

7. Robustness and resistance: similar to t-test (see Sect. 5.5.1); check model by plotting residuals (full) vs. est. means.
G. Advanced Extra SS F-test ex.
a. \(H_0: \mu_2 = \ldots = \mu_7\)
   Full model means: \(\mu_1, \mu_2, \mu_3, \mu_4, \mu_5, \mu_6, \mu_7\)
   Reduced: \(\mu_1, \mu_0, \mu_0, \mu_0, \mu_0, \mu_0, \mu_0\)

b. Est. means (full): \(\bar{Y}_1, \bar{Y}_2, \bar{Y}_3, \bar{Y}_4, \bar{Y}_5, \bar{Y}_6, \bar{Y}_7\)
   Reduced: \(\bar{Y}_1, \bar{Y}_0, \bar{Y}_0, \bar{Y}_0, \bar{Y}_0, \bar{Y}_0, \bar{Y}_0\)

c. Get Sum of residuals\(^2\) (full and reduced) (fit ANOVA with 2 codes)
\[
F \text{-stat} = \frac{(2191 - 1864)/(44 - 39)}{1864/39} = 1.37
\]

\[Pr(F_{5,39} > 1.37) = .26\]
H. Spock summary, so far

1. The data are consistent with judges 2-7 having the same mean percentage of women on venires (p-value = .25 from an extra-sum-of-squares F-test). There is overwhelming evidence that the Spock judge mean is different from the mean of the others, when their venires are treated as being from one population (p-value = .000001 from a two-sample t-test) [Note: F-test for 2 groups is identical to a 2-sided 2-sample t-test]

2. What could still be done:
   a. C.I. for $\mu_{\text{Spock}} - \mu_{\text{others}}$ (using 2-sample t-tools),
   b. and/or test, CI for $\mu_I - (\mu_2 + \mu_3 + ... + \mu_7)/6$,
      i.e. not assuming 2-7 equal (Ch. 6)
VII. Linear combinations and multiple comparisons of means

A. Introduction

1. Linear combination example:

\[ \mu_1 - (\mu_2 + \mu_3 + \ldots + \mu_7)/6 \]

2. Multiple comparison example:

Which means differ from which? (For 7 groups, this requires 21 p-values)
B. *Linear combinations of means*

1. $\gamma = C_1 \mu_1 + C_2 \mu_2 + C_3 \mu_3 + \ldots + C_I \mu_I$

   - the linear combination (a parameter)
   - coefficients of the linear combination

2. *C’s are numbers selected to address a specific research question*

3. Spock: $C_1 = 1, C_2 = \ldots = C_7 = -(1/6)$
4. Estimated linear combination

\[ g = C_1 \bar{Y}_1 + C_2 \bar{Y}_2 + C_3 \bar{Y}_3 + \ldots + C_I \bar{Y}_I \]

5. Spock: (with averages from Disp. 5.8)

\[ g = 14.6 - \frac{34.1 + 33.6 + 29.1 + 27 + 27 + 26.8}{6} \]

\[ = (1)14.5 + \frac{-1}{6}34.1 + \frac{-1}{6}33.6 + \ldots + \frac{-1}{6}26.8 \]

6. \( SE(g) \), using equal-\( \sigma \) model:

\[ SE(g) = s_p \sqrt{\frac{C_1^2}{n_1} + \frac{C_2^2}{n_2} + \ldots + \frac{C_I^2}{n_I}} \]
7. Spock:

\[ SE(g) = s_p \sqrt{\frac{1^2}{n_1} + \frac{(-1)^2}{n_2} + \frac{(-1)^2}{n_3} + \ldots + \frac{(-1)^2}{n_7}} } \]

\[ = 6.9142 \sqrt{\frac{1}{9} + \frac{(-1)^2}{5} + \frac{(-1)^2}{6} + \ldots + \frac{(-1)^2}{9}} \]

8. Do t-test and confidence interval using this SE and using d.f. in \( s_p \).

Computer work follows:
*** Analysis of Variance Model ***

Short Output:
Call:
   aov(formula = percent ~ code, data = case0502, contrasts =
   list(code = contr.treatment), na.action =
   na.exclude)

Terms:  
   code Residuals

Sum of Squares: 1927.081  1864.445
Deg. of Freedom:  6   39

Residual standard error: 6.914209
Estimated effects may be unbalanced

Tables of means
Grand mean
   26.583

   code
   1  2  3  4  5  6  7
   14.622 34.12 33.617 29.1 27 26.967 26.8
rep  9.000 5.000 6.000 9.0 2 6.000 9.0
"Calculator work" with S-PLUS commands
9. Important linear combinations

a. Comparing average of means
   Ex: $\gamma = (\mu_1 + \mu_2)/2 - (\mu_5 + \mu_6 + \mu_7)/3$

b. Inference about a rate of change
   Ex. Diet restriction, increase in mean lifetime per kcal: $\gamma = (\mu_6 - \mu_3)/10$

c. Linear trends (note: Ch. 7 is better)

10. Def: A linear combination is a *contrast* if the sum of its coefficients is zero
C. Simultaneous Inferences

1. Compounded uncertainty

a. If the chance that a $|t\text{-ratio}| > 2$ is .05, the chance that at least 1 of several $|t\text{-ratios}| > 2$ is more than .05.

b. If the success rate of a 95% CI is 95%, the simultaneous success rate of several 95% CIs is < 95%.

c. See definitions, p. 159
2. Is there a problem?

a. For *planned comparisons*, in which each of several tests or CIs is related to a pre-defined research question, no; use pairwise or *comparison-wise* tests or CIs (as in Sect. 5.2)

b. For *unplanned comparisons*, in which the research answer depends on picking out the results from a fam-
ily of tests or CIs, as in “which means differ from which others?” or “which treatments differ from control?” yes; use family-wise tests or CIs (Sect. 6.4)

c. For *data snooping*, in which one or several comparisons are made on the basis of how the data turned out, yes; this is a form of unplanned comparison
D. Multiple comparison methods

1. Goal: widen CIs depending on number of statements in the family so that family-wise success rate is 95\%, say

2. Tukey-Kramer method

\[ 95\% \text{ CI} : (\bar{Y}_2 - \bar{Y}_1) \pm \left\{ q_{I,n-I}(.95) / \sqrt{2} \right\} \times SE \]
a. It’s appropriate for comparing every mean to every other mean
b. It’s based on theory for the distribution of the largest $|t\text{-ratio}|$
c. It’s called *Tukey HSD* if all $I$ sample sizes are equal (honest signif. dif.)
d. Example: handicap study (Disp. 6.1)
   $I = 5$, $n = 70$, $s_p = 1.63$ (on 65 d.f.)
   $SE(Y_j - Y_k) = .617$ (p. 157 of book)
95% CI: \((\bar{Y}_j - \bar{Y}_k) \pm \left\{ \frac{q_{I,n-I}(.95)}{\sqrt{2}} \right\} \times SE \) \\
For: \\
\(\mu_1 - \mu_2: \quad .417 \pm 2.81 \times .617 = -1.26 \text{ to } 2.20\) \\
\(\mu_1 - \mu_3: \quad -1.020 \pm 2.81 \times .617 = -2.75 \text{ to } 0.71\)
### Analysis of Variance Model

**Terms:**
- CODE Residuals

**Sum of Squares:**
- CODE: 30.5214, 173.3214
- Residuals: 173.3214, 2.861581

**Deg. of Freedom:**
- CODE: 4, 65
- Residuals: 65

**Residual standard error:** 1.632937

Estimated effects are balanced

<table>
<thead>
<tr>
<th>Df</th>
<th>Sum of Sq</th>
<th>Mean Sq</th>
<th>F Value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CODE</td>
<td>4</td>
<td>30.5214</td>
<td>7.630357</td>
<td>2.861581</td>
</tr>
<tr>
<td>Residuals</td>
<td>65</td>
<td>173.3214</td>
<td>2.666484</td>
<td></td>
</tr>
</tbody>
</table>

**Residuals:**
CODE: 0.03012686

Formulas:
- Formula: `SCORE^CODE`
Compare every mean to every other mean
3. *Protected LSD* method

a. If p-value from 1-way F-test is > .05 then STOP; do not compare means. If p-value < .05, then compare means with usual t-tests (F protection)

b. LSD: “least significant difference”

c. This is widely used, better than nothing, but not as good as other multiple comparison methods
4. *Bonferonni* method

a. For $K$ 95% CIs use as the multiplier $t_{df}(1 - .025/K)$ rather than $t_{df}(1 - .025)$

b. Then the probability of simultaneous coverage of all $K$ intervals is at least 95%

c. This is conservative (intervals tend to be too wide) but it’s a method that can be applied to many situations
d. Demonstration (Handicap study)

For $K = 10$ CIs with family-wise confidence level at least 95%, use $t_{65}(1 - .025/10) = t_{65}(.9975)$

(From S+: $qt(.9975,65) = 2.906$)

$$(\bar{Y}_j - \bar{Y}_k) \pm 2.906 \times SE$$

See Display 6.6
Comparison Type

- **mcc**: compare every group to a control
- **mca**: compare every group to every other group

Compare to Level

Specify control level (if mcc)

For comparing every group to a control

LSD: only choice if Error Type is comparison-wise (this does usual, unadjusted C.I.’s)
5. Clarification:

a. For **planned comparisons** a single CI or test indicates the uncertainty of an answer to one question. No adjustment is needed.

b. For **unplanned comparisons**, a single question is answered by many CIs or tests (Which means differ?). A family-wise set of comparisons is needed.
E. One-way Layout, Review

1. For equal $\sigma$ model, estimate $\sigma$ by $s_p$

2. 1-way ANOVA F-test, $H: \mu_1 = \ldots = \mu_I$

3. Extra SS F-test (Ex: $H: \mu_2 = \ldots = \mu_7$)

4. t-test and CI for linear combinations

5. Multiple comparison adjustments
   a. All pairwise comparisons: Tukey
   b. All treatments to control: Dunnet
Case0602. (a) Does proportion depend on particular pair? (b) Is proportion > .5?

(a) ANOVA

(b) H: \( \mu = .5 \)
(b) \( H: (\mu_1 + \mu_2 + \ldots + \mu_6)/6 = .5 \)

2-sided p-value

95% CI
VIII. Simple Linear Regression

A. Introduction

1. Data: \((Y_i, X_i)\) for \(i = 1,...,n\)

2. Interest is in the probability distribution of \(Y\) as a function of \(X\)

3. Simple regression model: mean of \(Y\) is a straight line function of \(X\)

4. Steer example (see Display 7.3)
5. Estimated regression line

Equation for estimated regression line: $6.98 - 0.73X$
Regression

*** Linear Model ***

Coefficients:

<table>
<thead>
<tr>
<th></th>
<th>Value</th>
<th>Std. Error</th>
<th>t value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>6.9836</td>
<td>0.0485</td>
<td>143.8974</td>
</tr>
<tr>
<td>ltime</td>
<td>-0.7257</td>
<td>0.0344</td>
<td>-21.0786</td>
</tr>
</tbody>
</table>

Pr(>|t|)

(Intercept) 0.0000
ltime 0.0000

Residual standard error: 0.08226 on 8 degrees of freedom
Multiple R-Squared: 0.9823
F-statistic: 444.3 on 1 and 8 degrees of freedom, the p-value is 2.695e-008
B. Regression Terminology

1. \textit{Regression}: the mean of a \textit{response variable} as a function of one or more \textit{explanatory variables}: \( \mu \{Y \mid X\} \)

2. \textit{Regression model}: an ideal formula to approximate the regression

3. \textit{Simple linear regression model}

\[ \mu \{Y \mid X\} = \beta_0 + \beta_1 X \]

"mean of \( Y \) given \( X \)" or "regression of \( Y \) on \( X \)"

\( \beta_0 \) \text{ Intercept} \quad \beta_1 \text{ Slope} \quad (\text{unknown parameters})
4. $Y = \text{response variable} = \text{variable whose probability distribution is to be explained} (= \text{dependent variable})$

5. $X = \text{explanatory variable} = \text{variable used to explain dist. of} \ Y \ \text{(also called independent variable)}$

6. $\beta_0 \ \text{and} \ \beta_1 \ \text{are regression coefficients}$

7. (See Display 7.5)

8. Note: $Y = \beta_0 + \beta_1X$ is NOT simple reg.
C. Estimated coefficients

Choose $\hat{\beta}_0$ and $\hat{\beta}_1$ to make the residuals “small”
1. **Fitted value** for obs. $i$ is its estimated mean: 

$$ \text{fit}_i = \hat{\mu}\{Y \mid X_i\} = \hat{\beta}_0 + \hat{\beta}_1 X_i $$

**Residual** for obs. $i$: 

$$ \text{res}_i = Y_i - \text{fit}_i $$

2. The **statistical estimation method** of **LEAST SQUARES** finds those estimates that minimize the sum of squared residuals.

3. Solution (from calculus) on p. 182 of **Sleuth**
D. Tests and CIs for $\beta_0$, $\beta_1$

1. $\hat{\sigma}^2 = \frac{\text{sum of squared residuals}}{n-2}$

2. *Degrees of freedom* = sample size - number of coefficients (n-2)

3. Standard errors (p. 182)

4. t-ratio = $\frac{(\hat{\beta}_1 - \beta_1)}{\text{SE}(\hat{\beta}_1)}$ has a t-distribution on $n-2$ d.f. (ideal normal model)

5. Do t-tests and CIs as usual (df=n-2)
<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>170.00</td>
<td>0.03</td>
</tr>
<tr>
<td>2</td>
<td>290.00</td>
<td>0.03</td>
</tr>
<tr>
<td>3</td>
<td>-130.00</td>
<td>0.21</td>
</tr>
<tr>
<td>4</td>
<td>-70.00</td>
<td>0.26</td>
</tr>
<tr>
<td>5</td>
<td>-185.00</td>
<td>0.28</td>
</tr>
<tr>
<td>6</td>
<td>-220.00</td>
<td>0.28</td>
</tr>
<tr>
<td>7</td>
<td>200.00</td>
<td>0.45</td>
</tr>
<tr>
<td>8</td>
<td>290.00</td>
<td>0.50</td>
</tr>
<tr>
<td>9</td>
<td>270.00</td>
<td>0.50</td>
</tr>
<tr>
<td>10</td>
<td>200.00</td>
<td>0.63</td>
</tr>
<tr>
<td>11</td>
<td>300.00</td>
<td>0.80</td>
</tr>
<tr>
<td>12</td>
<td>-30.00</td>
<td>0.90</td>
</tr>
<tr>
<td>13</td>
<td>650.00</td>
<td>0.90</td>
</tr>
<tr>
<td>14</td>
<td>150.00</td>
<td>0.90</td>
</tr>
<tr>
<td>15</td>
<td>500.00</td>
<td>0.90</td>
</tr>
<tr>
<td>16</td>
<td>920.00</td>
<td>1.00</td>
</tr>
<tr>
<td>17</td>
<td>450.00</td>
<td>1.10</td>
</tr>
<tr>
<td>18</td>
<td>500.00</td>
<td>1.10</td>
</tr>
</tbody>
</table>
Get CIs “manually”
> # Confidence interval for beta 0, using regression report output
> half.width <- qt(.975,22)*.1185
> lower <- .3991 - half.width
> upper <- .3991 + half.width
> lower
> [1] 0.153346
> upper
> x.berkson
> 0.00006391007
>
> # 95% confidence interval for beta l
> half.width <- qt(.975,22)*.0002
> lower <- .0014 - half.width
> upper <- .0014 + half.width
> lower
> [1] 0.0009852254
> upper
> [1] 0.001814775
>
> # A p-value for a non-zero hypothesis must be obtained manually
> # Suppose interest is in H: slope = .0010
> t.stat <- (.0014 - .0010)/.0002
> p.value <- pt(-t.stat,22) # 1-sided p-value
> p.value
> [1] 0.02899785
E. Review so far

1. \( \mu \{Y|X\} = \beta_0 + \beta_1 X; \ \text{var} \{Y|X\}=\sigma^2 \)

2. \( \hat{\beta}_1 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2}; \ \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} \)

3. \( res_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i \) (for \( i \) from 1 to \( n \))

4. \( \hat{\sigma}^2 = \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2 / (n-2) \); \( \hat{\sigma} = \sqrt{\hat{\sigma}^2} \)

5. \( SE(\hat{\beta}_1) = \hat{\sigma} / \sqrt{(n-1)s_x^2}; SE(\hat{\beta}_0) = \hat{\sigma} \sqrt{(1/n) + \bar{X}^2 / (n-1)s_x^2} \)
F. Other Inference Tools

1. Test and CI for mean of Y at some X

a. Estimate the mean of Y at $X = X_0$ by

$$\hat{\mu}\{Y \mid X_0\} = \hat{\beta}_0 + \hat{\beta}_1 X_0$$

b. 

$$SE[\hat{\mu}\{Y \mid X_0\}] = \hat{\sigma} \sqrt{\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{(n-1)s^2_X}}$$

c. Do t-test and confidence interval in the usual way (df = n-2)
Create new data set with \( X_0 \)'s at which \( \mu \{ Y \mid X_0 \} \) desired.

Fit regression line with primary data set; fill in “predict” tab (next page).
\[ X_0 \]

\[ \hat{\mu} \{ Y \mid X_0 \} \]

\[ SE[\hat{\mu} \{ Y \mid X_0 \}] \]
2. *Prediction* of a future $Y$ at $X = X_0$

a. $\text{Pred}(Y|X_0) = \hat{\mu}\{Y \mid X_0\}$

b. Standard error of prediction

$$SE\{\text{Pred}(Y \mid X_0)\} = \sqrt{\hat{\sigma}^2 + (SE[\hat{\mu}\{Y \mid X_0\}])^2}$$

There is variability of $Y$ about its mean and uncertainty in the estimated mean

c. 95\% prediction interval:

$$\text{Pred}(Y|X_0) \pm t_{df}(0.975) \times SE\{\text{Pred}(Y|X_0)\}$$
d. A *prediction band* is a collection of prediction intervals on a plot.

Many stat programs will draw this automatically. In S+ it is necessary, unfortunately, to make a new data set with fake $X_0$’s in the range of interest, get prediction intervals at each of these fake $X_0$’s, and then connect the lower and upper bounds, as follows for the steer data...
Create 100 numbers between 0 and 2.1.

Estimate regression with original data; estimate means and SEs for new "data".
95% prediction intervals for y, at each x
3. **Calibration interval** = values of $X$ for which $Y_0$ is in a prediction interval.

**Confidence band** = a set of confidence intervals for $\mu\{Y \mid X_0\}$ at many values $X_0$.

**95% confidence interval** for $\mu\{Y \mid 1.0\}$

**95% prediction interval** interval for $Y$ at $X = 1.5$
4. Notes about conf. and pred. bands

a. Both are narrowest at \( X = \bar{X} \)

b. Beware of \textit{extrapolation}

c. Width of CI is zero if \( n \) is large enough; this is not true of the PI
IX. A Closer Look at Assumptions

A. Introduction

1. Robustness
2. Model checking
3. Log transformation (of $Y$, $X$, or both)
4. Analysis of Variance F-test for

**Full:** $\mu \{Y|X\} = \beta_0 + \beta_1 X$

**Reduced:** $\mu \{Y|X\} = \beta_0$
B. Robustness

1. A model for simple linear regression
   a. $\mu \{Y | X\} = \beta_0 + \beta_1 X$ (straight line)
   b. var$\{Y | X\} = \sigma^2$ (constant variance)
   c. Dist. of $Y$’s at any $X$ is normal
   d. Given $X_i$’s, the $Y_i$’s are independent

2. Consequences of violations of model assumptions
a. If “linearity” (1a) is violated, misleading conclusions may occur (but the degree of the problem depends on the degree of non-linearity)
b. If “constant variance” (1b) is violated, LS estimates are still unbiased but SEs, tests, CIs, and PIs are incorrect (but the degree ...)
c. If “normality” (1c) is violated, LS estimates are still unbiased; tests and CIs are quite robust; PIs are not
d. If “independence” (1d) is violated, LS estimates are still unbiased, but everything else can be misleading (but degree...)

Plotting code is litter (5 mice from each of 5 litters)
C. Tools for model checking

1. Scatterplot of $Y$ vs. $X$ (see Disp. 8.6)*

2. Scatterplot of residuals vs. fits*
   *Look for curvature, non-constant variance, and outliers

3. Normal prob. plot (p.224) is sometimes useful—for checking if dist. is symmetric or normal (i.e. for PIs)
4. Lack of fit F-test when there are replicates (Section 8.5).

**Full:** 1-way ANOVA (I parameters)

**Reduced:** $\beta_0 + \beta_1 X_i$

Get sums of squared residuals by fitting each of these. Form extra-SS F-statistic; compare to $F_{I-2,n-I}$ distr.
D. Interpretation after log transf.

1. If response is logged:
   a. $\mu\{\log(Y) | X\} = \beta_0 + \beta_1 X$ is the same as: $\text{Median}\{Y | X\} = e^{\beta_0 + \beta_1 X}$ (if dist’n of $\log(Y)$ given $X$ is symmetric)
   b. As $X$ increases by 1, what happens?

   $$\frac{\text{Median}\{Y \mid X = x + 1\}}{\text{Median}\{Y \mid X = x\}} = \frac{e^{\beta_0 + \beta_1 (x+1)}}{e^{\beta_0 + \beta_1 x}} = e^{\beta_1}$$
c. So \( \text{Median}\{Y|x+1\}=e^{\beta_1}\text{Median}\{Y|x\} \)

d. “As \( X \) increases by 1, the median of \( Y \) changes by the multiplicative factor of \( e^{\beta_1} \).” Or, better:

e. If \( \beta_1 > 0 \): “As \( X \) increases by 1, the median of \( Y \) increases by \( (e^{\beta_1}-1)\times100\% \).”

f. If \( \beta_1 < 0 \): “As \( X \) increases by 1, the median of \( Y \) decreases by \( (1 - e^{\beta_1})\times100\% \).”
g. Ex: $\mu\{\log(Y) \mid X\} = 19 - .5X; \quad 1 - e^{-0.5} = 0.4$

“It is estimated that the median failure time decreases by 40% with each 1kV increase in voltage.”
2. **If explanatory is logged.**

   a. If $\mu \{ Y|\log(X) \} = \beta_0 + \beta_1 \log(X)$ then:
      
      “Associated with each two-fold increase (i.e. doubling) of $X$ is a $\beta_1 \log(2)$ change in the mean of $Y$."

   b. Example: $Y = \text{pH}, X = \text{time after slaughter (hrs.)}$, estimated model:
      
      $\mu \{ Y|\log(x) \} = 6.98 - .73\log(X)$.
      
      Since $-.73 \times \log(2) = -.5$: 
“It is estimated that for each doubling of time after slaughter (between 0 and 8 hours) the mean pH decreases by .5.”
3. If both $Y$ and $X$ are logged.

$$\mu\{\log(Y) | \log(X)\} = \beta_0 + \beta_1 \log(X)$$

a. If $\beta_1 > 0$: “Associated with each doubling of $X$ is a $(e^{\beta_1 \log(2)} - 1) \times 100\%$ increase in the median of $Y$.”

b. If $\beta_1 < 0$: “Associated with each doubling of $X$ is a $(1 - e^{\beta_1 \log(2)}) \times 100\%$ decrease in the median of $Y$. ”
c. Example: $Y =$ number of species on an island; $X =$ island area.

Estimate of model $\mu \{ \log(Y) | \log(X) \} = 1.94 + .25 \times \log(x)$.  

Since $e^{.25\log(2)} - 1 = .19$:

“Associated with each doubling of island area is a 19% increase in the median number of bird species”
One way to plot log species vs log area is to specify log log plotting scale; then draw a scatter plot of species vs. area.
Regress log species on log area
E. Some applications of the extra SS F-test

1. F-test for significance of regression
   a. $H_0: \beta_1 = 0$
      
      Full: $\beta_0 + \beta_1 X$
      
      Reduced: $\beta_0$
   
   b. The computations are laid out in the regression ANOVA table.
St 411/511, Schafer  

Winter, 2003
c. The F-statistic is identical to the square of the t-statistic for $H_0: \beta_1 = 0$. The F-test p-value = 2-sided t-test p-value.

d. Although the t-test and F-test are the same here, they extend to different things in multiple regression.
2. \( R^2 \)

a. If X is no help at all in predicting Y then SS due to regression is zero

b. If X can be used to predict Y exactly (so all residuals are 0) SS due to regression = SS Total

c. \( R^2 = \frac{\text{SS due to regression}}{\text{SS Total}} \)
d. It summarizes the degree to which $X$ can predict $Y$

e. $R^2 = 0$ if $X$ is no help, $R^2 = 1$ if all residuals are 0.

f. $R^2$ is also the square of the correlation between $Y$ and the fitted values.
g. In simple regression it is also the square of the correlation between Y and X

h. $R^2$ is useful as a unitless summary of the strength of linear association, but

i. it is NOT useful for assessing model adequacy of significance
3. F-test for lack of fit (if there are replicate values of Y at some of the X’s)

a. Full model: one-way ANOVA
   Reduced model: simple linear reg.

b. The one-way ANOVA specifies different means of Y for different X’s, but no pattern. The simple reg. model is a special case of this in which the means fall on a straight line.