R can do all of the common analyses that are available in SPSS, including:
R can do all of the common analyses that are available in SPSS, including:

- $\chi^2$ tests
R can do all of the common analyses that are available in SPSS, including:

- $\chi^2$ tests
- $t$-tests
Classical Hypothesis Testing in R

R can do all of the common analyses that are available in SPSS, including:

- $\chi^2$ tests
- $t$-tests
- ANOVA
R can do all of the common analyses that are available in SPSS, including:

- $\chi^2$ tests
- $t$-tests
- ANOVA
- Regression
Outline

Inference for a Proportion

Chi-Squared Tests of Independence

One-Sample t-test

Two-Sample t-test

Test for Homogeneity of Variance

Analysis of Variance

Tukey’s HSD Tests

Multifactor Analysis of Variance

Multiple Regression

Testing Nested Models
Inference for a Proportion

Let's read in gssData.csv. If you already have it read in, we don't have to do this again.
Inference for a Proportion

> gss = read.csv(file.choose())
Inference for a Proportion

The variable VOTE92 indicates whether the respondent voted in the 1992 presidential election.
The variable VOTE92 indicates whether the respondent voted in the 1992 presidential election.

Let’s examine a contingency table of this variable.
Inference for a Proportion

```r
> foo <- table(gss$VOTE92)
> foo

DID NOT VOTE    DK    VOTED
  383    117    999
```
In this example, there were 999 respondents who voted out of 1500 total respondents.
Inference for a Proportion

- In this example, there were 999 respondents who voted out of 1500 total respondents.
- We can carry out a hypothesis test using the prop.test() function.
Inference for a Proportion

> prop.test(999, 1500)

    1-sample proportions test with continuity correction

data:  999 out of 1500, null probability 0.5
X-squared = 164.6727, df = 1, p-value < 2.2e-16
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
  0.6413965 0.6897436
sample estimates:
    p
  0.666
Inference for a Proportion

Note that by default,

- The null hypothesis $\pi = 0.5$ is tested against the alternative hypothesis $\pi \neq 0.5$. 

- A 95% confidence interval for $\pi$ is computed.

- Both the test and CI employ a correction for continuity.

- Like any other function in R, these defaults can be changed.
Inference for a Proportion

Note that by default,

▶ The null hypothesis $\pi = 0.5$ is tested against the alternative hypothesis $\pi \neq 0.5$.
▶ A 95% confidence interval for $\pi$ is computed.
Inference for a Proportion

Note that by default,

- The null hypothesis $\pi = 0.5$ is tested against the alternative hypothesis $\pi \neq 0.5$.
- A 95% confidence interval for $\pi$ is computed.
- Both the test and CI employ a correction for continuity.
Inference for a Proportion

Note that by default,

- The null hypothesis $\pi = 0.5$ is tested against the alternative hypothesis $\pi \neq 0.5$.
- A 95% confidence interval for $\pi$ is computed.
- Both the test and CI employ a correction for continuity.
- Like any other function in R, these defaults can be changed.
Inference for a Proportion

For example, to test the null hypothesis that $\pi = 0.6$ against the alternative hypothesis $\pi > 0.6$ and a 99% CI (one-sided) for $\pi$, all without the correction for continuity.
> prop.test(999, 1500, p = 0.6, alternative = "greater",
  conf.level = 0.99, correct = FALSE)

1-sample proportions test without continuity correction

data: 999 out of 1500, null probability 0.6
X-squared = 27.225, df = 1, p-value = 9.055e-08
alternative hypothesis: true p is greater than 0.6
99 percent confidence interval:
  0.6371184 1.0000000
sample estimates:
  p
  0.666
Outline

Inference for a Proportion

Chi-Squared Tests of Independence

One-Sample t-test

Two-Sample t-test

Test for Homogeneity of Variance

Analysis of Variance

Tukey’s HSD Tests

Multifactor Analysis of Variance

Multiple Regression

Testing Nested Models
The `chisq.test()` function will compute Pearson’s chi-squared test statistics and the corresponding $p$-value.
The `chisq.test()` function will compute Pearson’s chi-squared test statistics and the corresponding $p$-value.

We will apply it to examine whether respondents from different age groups differ in the frequency with which they report their income level.
Chi-Squared Tests of Independence

> my.test <- chisq.test(gss$AGE, gss$INCOME)
> my.test

Pearson's Chi-squared test

data:  gss$AGE and gss$INCOME
X-squared = 117.8385, df = 8, p-value < 2.2e-16
Chi-Squared Tests of Independence

What happens when we issue the following commands?

- `mytest$observed`
Chi-Squared Tests of Independence

What happens when we issue the following commands?

- `mytest$observed`
- `mytest$expected`
Chi-Squared Tests of Independence

What happens when we issue the following commands?

- mytest$observed
- mytest$expected
- mytest$residuals
Chi-Squared Tests of Independence

In the case that we may be worried about the chi-squared approximation to the sampling distribution of the statistic, we can instead, use simulation to compute an approximate p-value.
Chi-Squared Tests of Independence

> chisq.test(gss$AGE, gss$INCOME, simulate.p.value = TRUE, 
  B = 1000)

Pearson's Chi-squared test with simulated p-value (based on 1000 replicates)

data:  gss$AGE and gss$INCOME
X-squared = 117.8385, df = NA, p-value = 0.000999
Outline

Inference for a Proportion
Chi-Squared Tests of Independence

**One-Sample t-test**

Two-Sample t-test

Test for Homogeneity of Variance

Analysis of Variance

Tukey’s HSD Tests

Multifactor Analysis of Variance

Multiple Regression

Testing Nested Models
One-Sample t-test

- The function to perform a one-sample t-test is `t.test()`.
The function to perform a one-sample t-test is `t.test()`.

Using the `wages` data, suppose we wanted to perform a one-sample t-test to examine whether the mean salary for all employees was different from $6000.
One-Sample t-test

> library(foreign)
> wages <- read.spss("Wages.sav", to.data.frame = T)
> t.test(wages$salary, mu = 6000, alternative = "two.sided",
       conf.level = 0.95)

    One Sample t-test

data:  wages$salary
t = -7.8781, df = 92, p-value = 6.45e-12
alternative hypothesis: true mean is not equal to 6000
95 percent confidence interval:
  5274.185 5566.460
sample estimates:
mean of x
  5420.323
Outline

Inference for a Proportion

Chi-Squared Tests of Independence

One-Sample t-test

Two-Sample t-test

Test for Homogeneity of Variance

Analysis of Variance

Tukey’s HSD Tests

Multifactor Analysis of Variance

Multiple Regression

Testing Nested Models
Two-Sample t-test

- The function to perform a two-sample t-test is also `t.test()`.
Two-Sample t-test

- The function to perform a two-sample t-test is also `t.test()`.
- Suppose we wanted to perform a two-sample t-test to examine whether the mean salary for all male employees was different than the mean salary for all female employees.
Two-Sample t-test

> t.test(salary ~ sex, data = wages, var.equal = TRUE)

Two Sample t-test

data:  salary by sex
t = -6.2926, df = 91, p-value = 1.076e-08
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  -1076.2465  -559.7985
sample estimates:
mean in group FEMALE 5138.852
mean in group MALE  5956.875
Two-Sample t-test

- R can also perform a paired test by setting the argument `paired = TRUE`
Two-Sample t-test

- R can also perform a paired test by setting the argument `paired = TRUE`
- The data can also be formatted so that the data for males is in one column and the data for females is in another.
Outline

Inference for a Proportion
Chi-Squared Tests of Independence
One-Sample t-test
Two-Sample t-test
Test for Homogeneity of Variance
Analysis of Variance
Tukey’s HSD Tests
Multifactor Analysis of Variance
Multiple Regression
Testing Nested Models
Test for Homogeneity of Variance

▶ We are not given a homogeneity of variance test unless we specifically ask for it.

levene.test() function is available in the car library.
Test for Homogeneity of Variance

- We are not given a homogeneity of variance test unless we specifically ask for it.
- The `levene.test()` function is available in the `car` library.
Test for Homogeneity of Variance

> library(car)
> levene.test(salary ~ sex, data = wages)

Levene's Test for Homogeneity of Variance

<table>
<thead>
<tr>
<th>Df</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>group</td>
<td>1</td>
<td>0.1876</td>
</tr>
</tbody>
</table>

91
Test for Homogeneity of Variance

- The results for this test will differ from those given in SPSS.
Test for Homogeneity of Variance

- The results for this test will differ from those given in SPSS.
- R uses the modified Levene’s test (which uses deviations from the median rather than from the mean).
Outline

Inference for a Proportion
Chi-Squared Tests of Independence
One-Sample t-test
Two-Sample t-test
Test for Homogeneity of Variance
Analysis of Variance
Tukey’s HSD Tests
Multifactor Analysis of Variance
Multiple Regression
Testing Nested Models
Analysis of Variance

- To perform an analysis of variance in R, we will use the `aov()` function.
To perform an analysis of variance in R, we will use the `aov()` function.

It is typical to assign this function to an object.
Analysis of Variance

> my.model <- aov(salary ~ sex, data = wages)
Analysis of Variance

We can then obtain the results of our ANOVA by applying the `anova()` function to our fitted model.
Analysis of Variance

> anova(my.model)

Analysis of Variance Table

Response: salary

<table>
<thead>
<tr>
<th></th>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>sex</td>
<td>1</td>
<td>14045183</td>
<td>14045183</td>
<td>39.597</td>
<td>1.076e-08</td>
</tr>
<tr>
<td>Residuals</td>
<td>91</td>
<td>32278107</td>
<td>354704</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

sex ***  
Residuals  
---

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
What does the command `plot(my.model)` do?
If the design is unbalanced the default results from `anova()` will not match those from SPSS.
Analysis of Variance

- If the design is unbalanced the default results from `anova()` will not match those from SPSS.
- The default sum of squares in R is type I.
Analysis of Variance

- If the design is unbalanced the default results from `anova()` will not match those from SPSS.
- The default sum of squares in R is type I.
- The `Anova()` function from the `car` library can be used to obtain type II or III SS.
Outline

Inference for a Proportion
Chi-Squared Tests of Independence
One-Sample t-test
Two-Sample t-test
Test for Homogeneity of Variance
Analysis of Variance
Tukey’s HSD Tests
Multifactor Analysis of Variance
Multiple Regression
Testing Nested Models
Tukey’s HSD Tests

- If we had more than two groups and found significant differences, we could perform Tukey HSD post hoc tests.
If we had more than two groups and found significant differences, we could perform Tukey HSD post hoc tests.

The `multcomp` library contains many other multiple comparison tests.
Tukey’s HSD Tests

> TukeyHSD(my.model)

Tukey multiple comparisons of means
95% family-wise confidence level

Fit: aov(formula = salary ~ sex, data = wages)

$sex

<table>
<thead>
<tr>
<th></th>
<th>diff</th>
<th>lwr</th>
<th>upr</th>
<th>p</th>
<th>adj</th>
</tr>
</thead>
<tbody>
<tr>
<td>MALE - FEMALE</td>
<td>818.0225</td>
<td>559.7985</td>
<td>1076.247</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>
Tukey’s HSD Tests

- We can also plot the simultaneous confidence intervals.
Tukey’s HSD Tests

- We can also plot the simultaneous confidence intervals.
- We apply the `plot()` function to our Tukey results.
Tukey's HSD Tests

> plot(TukeyHSD(my.model))

95% family-wise confidence level
Outline

Inference for a Proportion
Chi-Squared Tests of Independence
One-Sample t-test
Two-Sample t-test
Test for Homogeneity of Variance
Analysis of Variance
Tukey’s HSD Tests
Multifactor Analysis of Variance
Multiple Regression
Testing Nested Models
Multifactor Analysis of Variance

- We can add additional factors into the model using the + operator in our formula.

```r
my.model <- aov(salary ~ sex + race + sex:race, data=wages)
```
Multifactor Analysis of Variance

- We can add additional factors into the model using the + operator in our formula.
- We can add interaction terms into the model using the : operator in our formula.
Multifactor Analysis of Variance

- We can add additional factors into the model using the + operator in our formula.
- We can add interaction terms into the model using the : operator in our formula.
- my.model <- aov(salary sex + race + sex:race, data=wages)
Outline

Inference for a Proportion
Chi-Squared Tests of Independence
One-Sample t-test
Two-Sample t-test
Test for Homogeneity of Variance
Analysis of Variance
Tukey’s HSD Tests
Multifactor Analysis of Variance
Multiple Regression
Testing Nested Models
The function we used for the simple linear regression model can be expanded to include multiple independent variables.
Multiple Regression

- The function we used for the simple linear regression model can be expanded to include multiple independent variables.
- We can use the same operators in the formula as multifactor ANOVA in the `lm()` function.
Multiple Regression

- The function we used for the simple linear regression model can be expanded to include multiple independent variables.
- We can use the same operators in the formula as multifactor ANOVA in the `lm()` function.
- Suppose we wanted to model salary using sex, experience, and the sex by experience interaction terms.
Multiple Regression

> int.model <- lm(salary ~ sex + exper + sex:exper, data = wages)
Multiple Regression

> summary(int.model)

Call:
`lm(formula = salary ~ sex + exper + sex:exper, data = wages)`

Residuals:

<table>
<thead>
<tr>
<th></th>
<th>Min</th>
<th>1Q</th>
<th>Median</th>
<th>3Q</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-1342.94</td>
<td>-534.40</td>
<td>39.31</td>
<td>369.75</td>
<td>2139.91</td>
</tr>
</tbody>
</table>

Coefficients:

|                      | Estimate | Std. Error | t value | Pr(>|t|)   |
|----------------------|----------|------------|---------|------------|
| (Intercept)          | 4920.8166 | 115.2529  | 42.696  | < 2e-16 *** |
| sexMALE              | 1042.8877 | 187.1620  | 5.572   | 2.65e-07 *** |
| exper                | 2.1844    | 0.8804    | 2.481   | 0.0150 *    |
| sexMALE :exper       | -2.2507   | 1.3509    | -1.666  | 0.0992 .    |

---

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> anova(int.model)

Analysis of Variance Table

Response: salary

<table>
<thead>
<tr>
<th></th>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>sex</td>
<td>1</td>
<td>14045183</td>
<td>14045183</td>
<td>41.4070</td>
<td>6.059e-09</td>
</tr>
<tr>
<td>exper</td>
<td>1</td>
<td>1147920</td>
<td>1147920</td>
<td>3.3842</td>
<td>0.06916</td>
</tr>
<tr>
<td>sex:exper</td>
<td>1</td>
<td>941571</td>
<td>941571</td>
<td>2.7759</td>
<td>0.09921</td>
</tr>
<tr>
<td>Residuals</td>
<td>89</td>
<td>30188616</td>
<td>339198</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

sex       ***
exper     .
sex:exper .
Residuals
---

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Outline

Inference for a Proportion
Chi-Squared Tests of Independence
One-Sample t-test
Two-Sample t-test
Test for Homogeneity of Variance
Analysis of Variance
Tukey’s HSD Tests
Multifactor Analysis of Variance
Multiple Regression
Testing Nested Models
Testing Nested Models

- Suppose we wanted to fit a saturated model using the sex, education and experience variables (including all interactions) to model salary.
Testing Nested Models

- Suppose we wanted to fit a saturated model using the sex, education and experience variables (including all interactions) to model salary.
- What is the R command to do so?
Testing Nested Models

> full.model <- lm(salary ~ sex + exper +
                   educ + sex:exper + sex:educ + educ:exper +
                   sex:educ:exper, data = wages)
Testing Nested Models

How would we fit a main effects only model?
Testing Nested Models

```r
> me.model <- lm(salary ~ sex + exper + educ, data = wages)
```
We can test whether the additional terms in the full model are necessary by using `anova()` function.
We can test whether the additional terms in the full model are necessary by using \texttt{anova()} function.

We simply provide this function with two models — one of which is nested in the other.
> anova(me.model, full.model)

Analysis of Variance Table

Model 1: salary ~ sex + exper + educ
Model 2: salary ~ sex + exper + educ + sex:exper + sex:educ + educ:exper + sex:educ:exper

<table>
<thead>
<tr>
<th>Res.Df</th>
<th>RSS</th>
<th>Df</th>
<th>Sum of Sq</th>
<th>F</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>27883465</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>25727667</td>
<td>4</td>
<td>2155798</td>
<td>1.7806</td>
<td>0.1402</td>
</tr>
</tbody>
</table>