Comparing Groups: Statistical Tests

Luis Francisco Gomez Lopez

FAEDIS

2023-08-26

Contents

Please Read Me

- This presentation is based on (Chapman and Feit 2019, chap. 6)

Purpose

-

Consumer segmentation survey

- **Import data**

```r
segmentation <- read_csv(file = "http://goo.gl/qw303p")
segmentation |> head(n = 5)
```

```
# A tibble: 5 x 7
    age gender income  kids ownHome subscribe Segment
  <dbl> <chr>   <dbl> <dbl> <chr>   <chr>     <chr>
1  47.3 Male   49483.     2 ownNo   subNo     Suburb mix
2  31.4 Male   35546.     1 ownYes  subNo     Suburb mix
3  43.2 Male   44169.     0 ownYes  subNo     Suburb mix
4  37.3 Female 81042.     1 ownNo   subNo     Suburb mix
5  41.0 Female 79353.     3 ownYes  subNo     Suburb mix
```

Consumer segmentation survey

- Chi-squared test

```
segmentation |> count(Segment)
```

```
# A tibble: 4 x 2
  Segment       n
  <chr>     <int>
1 Moving up    70
2 Suburb mix  100
3 Travelers    80
4 Urban hip    50
```

```
segmentation |>
  count(subscribe, ownHome) |>
  pivot_wider(id_cols = subscribe,
              names_from = ownHome,
              values_from = n)
```

```
# A tibble: 2 x 3
  subscribe ownNo ownYes
  <chr>     <int>  <int>
1 subNo       137    123
2 subYes       22     18
```

Consumer segmentation survey

- Chi-squared test for given probabilities

$$H_0 : p_1 = \frac{1}{4} \wedge p_2 = \frac{1}{4} \wedge p_3 = \frac{1}{4} \wedge p_4 = \frac{1}{4}$$

$$H_1 : p_1 \neq \frac{1}{4} \vee p_2 \neq \frac{1}{4} \vee p_3 = \frac{1}{4} \vee p_4 \neq \frac{1}{4}$$

$$\chi^2 = \sum_{i=1}^{n} \frac{(Observed_i - Expected_i)^2}{Expected_i} =$$
$$\frac{70-300\frac{1}{4}}{300\frac{1}{4}} + \frac{100-300\frac{1}{4}}{300\frac{1}{4}} + \frac{80-300\frac{1}{4}}{300\frac{1}{4}} + \frac{50-300\frac{1}{4}}{300\frac{1}{4}}$$

- **Base R way**

```
chi_statistic <- table(segmentation$Segment) |>
  chisq.test(p = c(1/4, 1/4, 1/4, 1/4))
chi_statistic
```

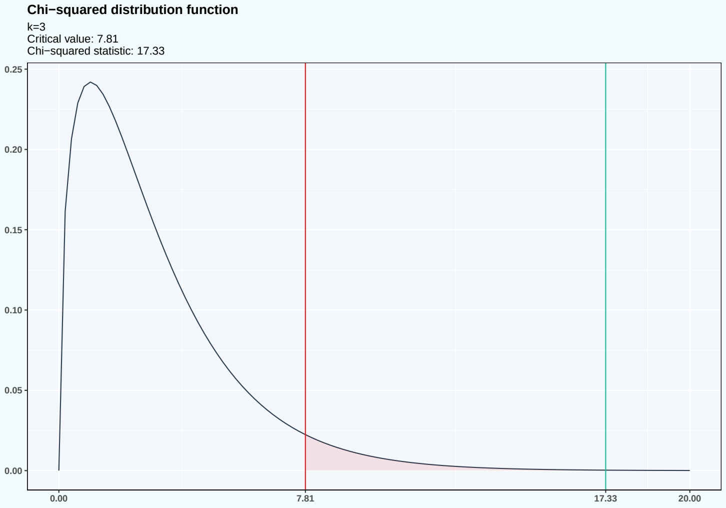
```

	Chi-squared test for given probabilities

data:  table(segmentation$Segment)
X-squared = 17.333, df = 3, p-value = 0.0006035
```

Consumer segmentation survey

- Chi-squared test for given probabilities



Consumer segmentation survey

- Chi-squared test for given probabilities

$$H_0 : p_1 = \frac{1}{4} \wedge p_2 = \frac{1}{4} \wedge p_3 = \frac{1}{4} \wedge p_4 = \frac{1}{4}$$

$$H_1 : p_1 \neq \frac{1}{4} \vee p_2 \neq \frac{1}{4} \vee p_3 = \frac{1}{4} \vee p_4 \neq \frac{1}{4}$$

$$\chi^2 = \sum_{i=1}^{n} \frac{(Observed_i - Expected_i)^2}{Expected_i} =$$
$$\frac{70-300\frac{1}{4}}{300\frac{1}{4}} + \frac{100-300\frac{1}{4}}{300\frac{1}{4}} + \frac{80-300\frac{1}{4}}{300\frac{1}{4}} + \frac{50-300\frac{1}{4}}{300\frac{1}{4}}$$

- **tidymodels way**

```
library(tidymodels)
segmentation |>
  chisq_test(response = Segment,
             p = c(1/4, 1/4, 1/4, 1/4))
```

```
# A tibble: 1 x 3
  statistic chisq_df  p_value
      <dbl>    <dbl>    <dbl>
1      17.3        3 0.000603
```

Consumer segmentation survey

- Pearson's Chi-squared test

$$H_0: p_{11} = \frac{260}{300}\frac{159}{300} \wedge p_{12} = \frac{260}{300}\frac{141}{300} \wedge p_{21} = \frac{40}{300}\frac{159}{300} \wedge p_{22} = \frac{40}{300}\frac{141}{300}$$

$$H_1: p_{11} \neq \frac{260}{300}\frac{159}{300} \vee p_{12} \neq \frac{260}{300}\frac{141}{300} \vee p_{21} \neq \frac{40}{300}\frac{159}{300} \vee p_{22} \neq \frac{40}{300}\frac{141}{300}$$

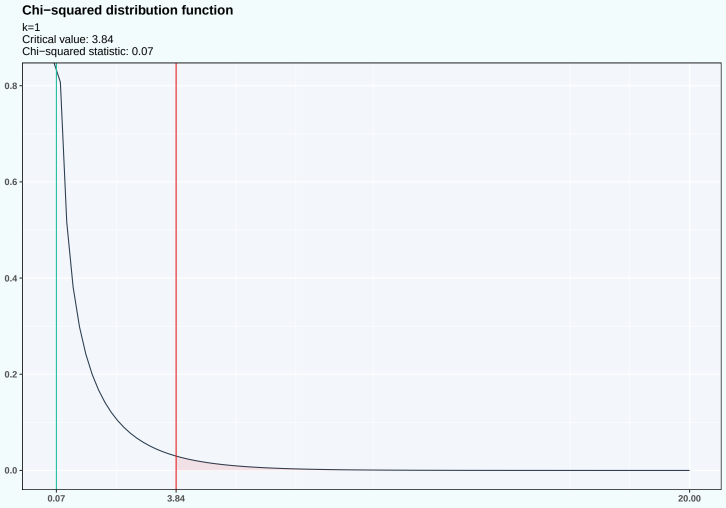
$$\chi^2 = \sum_{i=1}^{n} \frac{(Observed_i - Expected_i)^2}{Expected_i} =$$
$$\frac{(137-300\frac{260}{300}\frac{159}{300})^2}{300\frac{260}{300}\frac{159}{300}} + \frac{(123-300\frac{260}{300}\frac{141}{300})^2}{300\frac{260}{300}\frac{141}{300}} + \frac{(22-300\frac{40}{300}\frac{159}{300})^2}{300\frac{40}{300}\frac{159}{300}} + \frac{(18-300\frac{40}{300}\frac{141}{300})^2}{300\frac{40}{300}\frac{141}{300}}$$

- **Base R way**

```
chi_statistic <- chisq.test(table(segmentation$subscribe,
                segmentation$ownHome),
          correct = FALSE)
```

Consumer segmentation survey

- Pearson's Chi-squared test



Consumer segmentation survey

- Pearson's Chi-squared test

$$H_0: p_{11} = \frac{260}{300}\frac{159}{300} \wedge p_{12} = \frac{260}{300}\frac{141}{300} \wedge p_{21} = \frac{40}{300}\frac{159}{300} \wedge p_{22} = \frac{40}{300}\frac{141}{300}$$

$$H_1: p_{11} \neq \frac{260}{300}\frac{159}{300} \vee p_{12} \neq \frac{260}{300}\frac{141}{300} \vee p_{21} \neq \frac{40}{300}\frac{159}{300} \vee p_{22} \neq \frac{40}{300}\frac{141}{300}$$

$$\chi^2 = \sum_{i=1}^{n} \frac{(Observed_i - Expected_i)^2}{Expected_i} =$$

$$\frac{(137-300\frac{260}{300}\frac{159}{300})^2}{300\frac{260}{300}\frac{159}{300}} + \frac{(123-300\frac{260}{300}\frac{141}{300})^2}{300\frac{260}{300}\frac{141}{300}} + \frac{(22-300\frac{40}{300}\frac{159}{300})^2}{300\frac{40}{300}\frac{159}{300}} + \frac{(18-300\frac{40}{300}\frac{141}{300})^2}{300\frac{40}{300}\frac{141}{300}}$$

- **tidymodels way**

```
segmentation |>
  chisq_test(formula = subscribe ~ ownHome,
             correct = FALSE)
```

```
# A tibble: 1 x 3
  statistic chisq_df p_value
      <dbl>    <int>   <dbl>
1    0.0741        1   0.785
```

Consumer segmentation survey

- Exact binomial test

$H_0 : p = 0.5 \; H_1 : p \neq 0.5$

$B = \sum_{i=1}^{n} x_i = 157$ where $x_i \in 0, 1$

- **R base way**

```r
binom_test <- binom.test(x = 157, n = 300, p = 0.5,
           alternative = 'two.sided',
           conf.level = 0.95)
binom_test
```

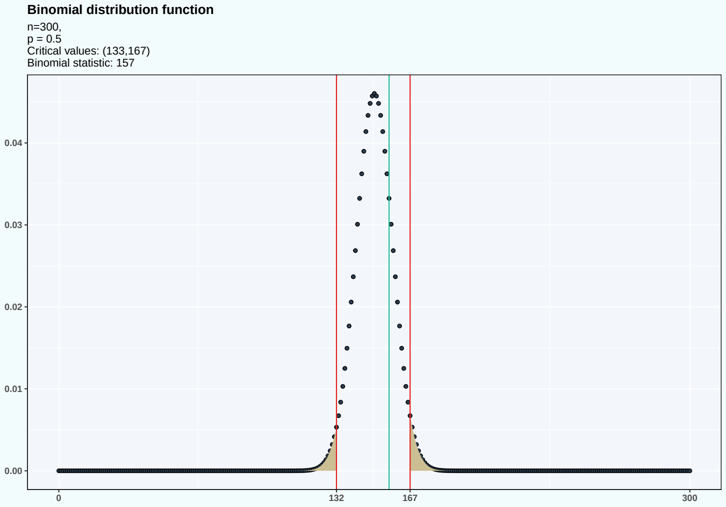
```

	Exact binomial test

data:  157 and 300
number of successes = 157, number of trials = 300, p-value = 0.453
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.4651595 0.5810418
sample estimates:
probability of success
             0.5233333
```

Consumer segmentation survey

- Exact binomial test



Consumer segmentation survey

- Exact binomial test
  - Confidence interval:

$$p_L < p < p_U$$

- p_L and p_U are random variables but p is not a random variable. Therefore $[p_L, p_U]$ is a random interval where we have that:

$$P(0.4651595 \approx p_L < p < p_U \approx 0.5810418) = 0.95$$

Consumer segmentation survey

- Exact binomial test

$H_0 : p = 0.5 \ H_1 : p \neq 0.5$

$B = \sum_{i=1}^{n} x_i = 157$ where $x_i \in 0, 1$

- **tidymodels way**

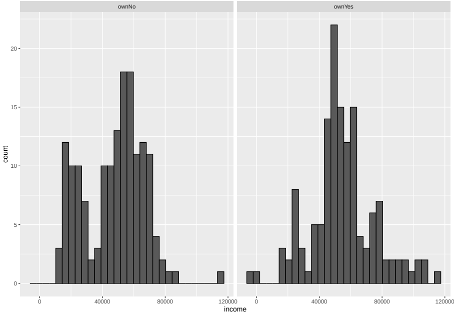
```
binom.test(x = 157, n = 300, p = 0.5,
           alternative = 'two.sided',
           conf.level = 0.95) |>
  tidy()
```

```
# A tibble: 1 x 8
  estimate statistic p.value parameter conf.low conf.high method      alternative
     <dbl>     <dbl>   <dbl>     <dbl>    <dbl>     <dbl> <chr>       <chr>
1    0.523       157   0.453       300    0.465     0.581 Exact bin~  two.sided
```

Consumer segmentation survey

- 2 sample t-test: independent samples

```
segmentation |> ggplot() +
  geom_histogram(aes(x = income), color='black') +
  facet_wrap(facets = vars(ownHome))
```



Consumer segmentation survey

- 2 sample t-test: independent samples

```r
segmentation |>
  group_by(ownHome) |>
  summarise(mean_income = mean(income),
            var_income = var(income),
            n = n())
```

```
# A tibble: 2 x 4
  ownHome mean_income var_income     n
  <chr>         <dbl>      <dbl> <int>
1 ownNo        47391. 358692875.   159
2 ownYes       54935. 430890091.   141
```

Consumer segmentation survey

- 2 sample t-test: independent samples

$$H_0 : \mu_{ownNo} - \mu_{ownYes} = 0 \; H_1 : \mu_{ownNo} - \mu_{ownYes} \neq 0$$

$$t = \frac{\overline{ownNo} - \overline{ownYes}}{\sqrt{\frac{s^2_{ownNo}}{n_{ownNo}} - \frac{s^2_{ownYes}}{n_{ownYes}}}} = \frac{47391.01 - 54934.68}{\sqrt{\frac{358692875}{159} - \frac{430890091}{141}}} \approx -3.273094$$

- **R base way**

```
t_test <- t.test(income ~ ownHome, data = segmentation,
                 alternative='two.sided', mu = 0,
                 conf.level = 0.95)
t_test
```

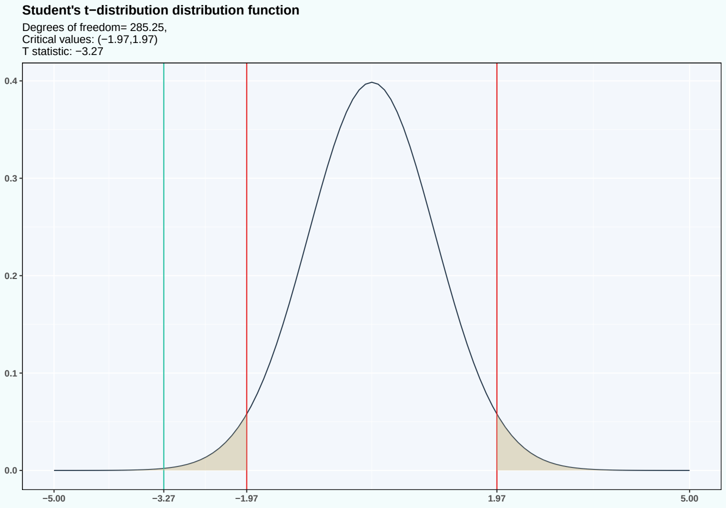
```

    Welch Two Sample t-test

data:  income by ownHome
t = -3.2731, df = 285.25, p-value = 0.001195
alternative hypothesis: true difference in means between group ownNo and group ownYes is not equal to 0
95 percent confidence interval:
 -12080.155  -3007.193
sample estimates:
 mean in group ownNo mean in group ownYes
            47391.01             54934.68
```

Consumer segmentation survey

- 2 sample t-test: independent samples



Consumer segmentation survey

- 2 sample t-test: independent samples
  - Confidence interval:

$$c_L < \mu_{ownNo} - \mu_{ownYes} < c_U$$

- $\mu_{ownNo} - \mu_{ownYes}$ is not a random variable so we need to use a random variable

$$P\left(t_L < \frac{\overline{x}_{ownNo} - \overline{x}_{ownYes} - (\mu_{ownNo} - \mu_{ownYes})}{\sqrt{\frac{s^2_{ownNo}}{n_{ownNo}} + \frac{s^2_{ownYes}}{n_{ownYes}}}} < t_U\right) = 0.95$$

- $\overline{x}_{ownNo} - \overline{x}_{ownYes}$ is a random variable

Consumer segmentation survey

- 2 sample t-test: independent samples
  - Confidence interval:
    - $\frac{\overline{x}_{ownNo}-\overline{x}_{ownYes}-(\mu_{ownNo}-\mu_{ownYes})}{\sqrt{\frac{s_{ownNo}^2}{n_{ownNo}}+\frac{s_{ownYes}^2}{n_{ownYes}}}}$ is also a random variable with student's t-distribution and $\nu \approx \frac{(\frac{s_{ownNo}^2}{n_{ownNo}}+\frac{s_2^2}{n_{ownYes}})^2}{\frac{(\frac{s_{ownNo}^2}{n_{ownNo}})^2}{n_{ownNo}-1}+\frac{(\frac{s_2^2}{n_{ownYes}})^2}{n_{ownYes}-1}} = 285.2521$ degrees of freedom
    - Also we need to specify t_L and t_U

```r
t_L <- qt(p = 0.025, df = 285.25, lower.tail = TRUE)
t_L
```

```
[1] -1.968315
```

```r
t_U <- qt(p = 0.975, df = 285.25, lower.tail = TRUE)
t_U
```

```
[1] 1.968315
```

Consumer segmentation survey

- 2 sample t-test: independent samples
  - Confidence interval:

$$P(-7543.674-1.968315\times2304.753 < \mu_{ownNo}-\mu_{ownYes} < -7543.674-1.968315\times2304.753) = 0.95$$

$$P(-12080.16 < \mu_{ownNo} - \mu_{ownYes} < -3007.193) = 0.95$$

- In the long run 95% of confidence intervals constructed in this manner will contain the true parameter

Consumer segmentation survey

- 2 sample t-test: independent samples

$$H_0 : \mu_{ownNo} - \mu_{ownYes} = 0 \; H_1 : \mu_{ownNo} - \mu_{ownYes} \neq 0$$

$$t = \frac{\overline{ownNo} - \overline{ownYes}}{\sqrt{\frac{s^2_{ownNo}}{n_{ownNo}} - \frac{s^2_{ownYes}}{n_{ownYes}}}} = \frac{47391.01 - 54934.68}{\sqrt{\frac{358692875}{159} - \frac{430890091}{141}}} \approx -3.273094$$

- **tidymodels way**

```
segmentation |>
  t_test(formula = income ~ ownHome,
         alternative = "two-sided",
         order = c("ownNo", "ownYes"),
         mu = 0,
         conf_level = 0.95)
```

```
# A tibble: 1 x 7
  statistic  t_df p_value alternative estimate lower_ci upper_ci
      <dbl> <dbl>   <dbl> <chr>          <dbl>    <dbl>    <dbl>
1     -3.27  285. 0.00119 two.sided      -7544.  -12080.   -3007.
```

Consumer segmentation survey

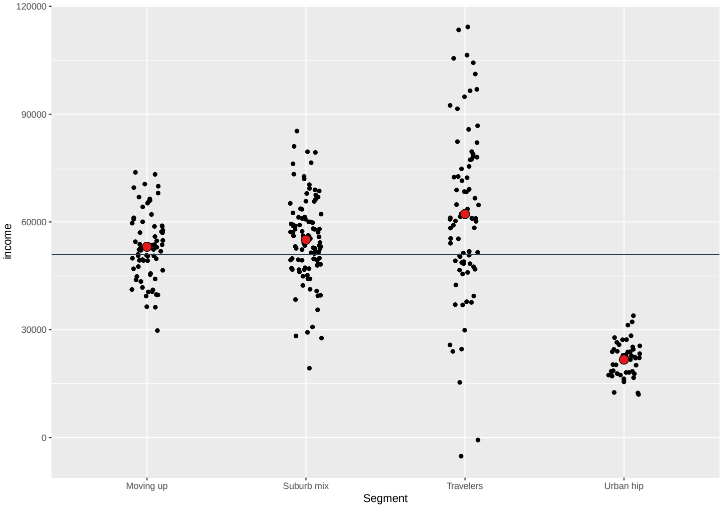
- Testing Multiple Group Means: Analysis of Variance (ANOVA)

```
segmentation |>
  group_by(Segment) |>
  summarise(mean = mean(income),
            variance = var(income),
            n = n())
```

```
# A tibble: 4 x 4
  Segment      mean   variance     n
  <chr>       <dbl>      <dbl> <int>
1 Moving up  53091.  92862689.    70
2 Suburb mix 55034. 142761527.   100
3 Travelers  62214. 564173979.    80
4 Urban hip  21682.  23885953.    50
```

Consumer segmentation survey

- Testing Multiple Group Means: Analysis of Variance (ANOVA)



Consumer segmentation survey

- Testing Multiple Group Means: Analysis of Variance (ANOVA)

$H_0 : \mu_{Moving\ up} = \mu_{Suburb\ mix} = \mu_{Travelers} = \mu_{Urban\ hip}$

H_1 : At least one group mean is different from the rest

$n = \sum_{j=1}^{4} n_j = n_1 + \cdots + n_4 = 70 + 100 + 80 + 50 = 300$

$\overline{income} = \frac{1}{n} \sum_{j=1}^{4} \sum_{i=1}^{n_j} income_{ij}$

$\overline{income}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} income_{ij}$

$$F = \frac{\frac{\sum_{j=1}^{4} \sum_{i=1}^{n_j} (\overline{income}_j - \overline{income})^2}{4-1}}{\frac{\sum_{j=1}^{4} \sum_{i=1}^{n_j} (income_{ij} - \overline{income}_j)^2}{300-4}} = \frac{\frac{54969675428}{3}}{\frac{66281072794}{296}} = \frac{18323225143}{223922543} = 81.82841$$

Consumer segmentation survey

- Testing Multiple Group Means: Analysis of Variance (ANOVA)
  - **R base way**

```
anova_table <- aov(data = segmentation, formula = income ~ Segment) |>
  anova()
anova_table
```

```
Analysis of Variance Table

Response: income
           Df     Sum Sq    Mean Sq F value    Pr(>F)
Segment     3 5.4970e+10 1.8323e+10  81.828 < 2.2e-16 ***
Residuals 296 6.6281e+10 2.2392e+08
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Consumer segmentation survey

- Testing Multiple Group Means: Analysis of Variance (ANOVA)

Consumer segmentation survey

- Testing Multiple Group Means: Analysis of Variance (ANOVA)
    - **tidymodels way**

```
anova_table <- aov(data = segmentation, formula = income ~ Segment) |>
  anova() |>
  tidy()
anova_table
```

```
# A tibble: 2 x 6
  term          df       sumsq      meansq statistic   p.value
  <chr>      <int>       <dbl>       <dbl>     <dbl>     <dbl>
1 Segment        3 54969675428. 18323225143.      81.8  1.41e-38
2 Residuals    296 66281072794.   223922543.      NA    NA
```

Consumer segmentation survey

- Testing Multiple Group Means: Analysis of Variance (ANOVA)

```
segmentation |>
  distinct(Segment) |>
  arrange(Segment) |>
  rowid_to_column(var = 'i')
```

```
# A tibble: 4 x 2
      i Segment
  <int> <chr>
1     1 Moving up
2     2 Suburb mix
3     3 Travelers
4     4 Urban hip
```

```
segmentation |>
  distinct(ownHome) |>
  rowid_to_column(var = 'j')
```

```
# A tibble: 2 x 2
      j ownHome
  <int> <chr>
1     1 ownNo
2     2 ownYes
```

Consumer segmentation survey

- Testing Multiple Group Means: Analysis of Variance (ANOVA)

```
segmentation |>
  count(Segment, ownHome, name = "n_ij")
```

```
# A tibble: 8 x 3
  Segment    ownHome  n_ij
  <chr>      <chr>   <int>
1 Moving up  ownNo      47
2 Moving up  ownYes     23
3 Suburb mix ownNo      52
4 Suburb mix ownYes     48
5 Travelers  ownNo      20
6 Travelers  ownYes     60
7 Urban hip  ownNo      40
8 Urban hip  ownYes     10
```

Consumer segmentation survey

- Testing Multiple Group Means: Analysis of Variance (ANOVA)

```
mu_ij <- segmentation |>
  group_by(Segment, ownHome) |>
  summarise(mean = mean(income)) |>
  ungroup()
mu_11 <- mu_ij$mean[1]
mu_11
```

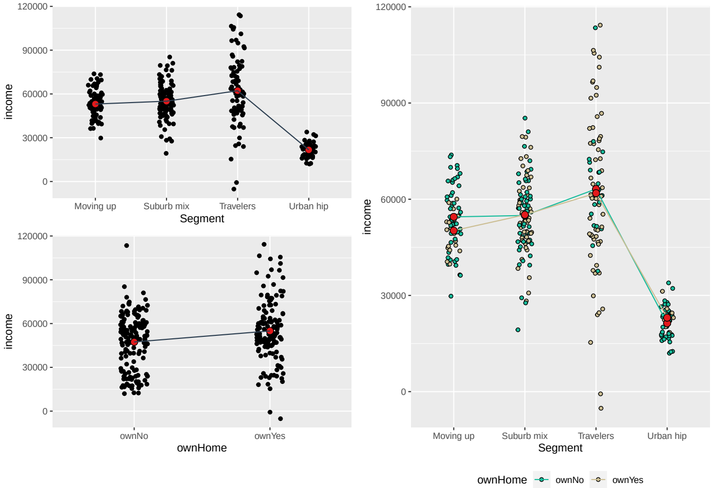
```
[1] 54497.68
```

```
segmentation |>
  select(income, Segment, ownHome) |>
  head(n=5)
```

```
# A tibble: 5 x 3
  income Segment    ownHome
   <dbl> <chr>      <chr>
1 49483. Suburb mix ownNo
2 35546. Suburb mix ownYes
3 44169. Suburb mix ownYes
4 81042. Suburb mix ownNo
5 79353. Suburb mix ownYes
```

Consumer segmentation survey

- Testing Multiple Group Means: Analysis of Variance (ANOVA)



Consumer segmentation survey

- Testing Multiple Group Means: Analysis of Variance (ANOVA)

$$
\begin{aligned}
income_{ijk} = & \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk} \\
& \text{where } + \epsilon_i \sim \mathcal{N}(0, \sigma^2) \\
& \text{and } i = 1, 2, 3, 4 \\
& j = 1, 2 \\
& k = 1, \ldots n_{ij} \\
& \mu = \mu_{11} \\
& \alpha_1 = \beta_1 = 0 \\
& (\alpha\beta)_{11} = (\alpha\beta)_{12} = 0 \\
& (\alpha\beta)_{21} = (\alpha\beta)_{31} = (\alpha\beta)_{41} = 0
\end{aligned}
$$

Consumer segmentation survey

- Testing Multiple Group Means: Analysis of Variance (ANOVA)

$$
\begin{aligned}
\widehat{income}_{ijk} =& \widehat{\mu} + \widehat{\alpha}_i + \widehat{\beta}_j + (\widehat{\alpha\beta})_{ij} + \hat{\epsilon}_{ijk} \\
& \text{and } i = 1, 2, 3, 4 \\
& j = 1, 2 \\
& k = 1, \dots n_{ij} \\
& \widehat{\mu} = \widehat{\mu}_{11} \\
& \widehat{\alpha}_1 = \widehat{\beta}_1 = 0 \\
& (\widehat{\alpha\beta})_{11} = (\widehat{\alpha\beta})_{12} = 0 \\
& (\widehat{\alpha\beta})_{21} = (\widehat{\alpha\beta})_{31} = (\widehat{\alpha\beta})_{41} = 0
\end{aligned}
$$

$$
income_{ijk} - \widehat{income}_{ijk} = \hat{\epsilon}_{ijk}
$$

Consumer segmentation survey

- Testing Multiple Group Means: Analysis of Variance (ANOVA)

```
segmentation |>
  select(income, Segment, ownHome) |>
  head(n=2) |>
  glimpse()
```

```
Rows: 2
Columns: 3
$ income  <dbl> 49482.81, 35546.29
$ Segment <chr> "Suburb mix", "Suburb mix"
$ ownHome <chr> "ownNo", "ownYes"
```

```
framed <- model_frame(formula = income ~
                                Segment +
                                ownHome +
                                Segment:ownHome,
            data = segmentation)

model_matrix(terms = framed$terms,
             data = framed$data) |>
  head(n = 2) |>
  glimpse()
```

```
Rows: 2
Columns: 8
$ `(Intercept)`                      <dbl> 1, 1
$ `SegmentSuburb mix`                <dbl> 1, 1
$ SegmentTravelers                   <dbl> 0, 0
$ `SegmentUrban hip`                 <dbl> 0, 0
$ ownHomeownYes                      <dbl> 0, 1
$ `SegmentSuburb mix:ownHomeownYes`  <dbl> 0, 1
$ `SegmentTravelers:ownHomeownYes`   <dbl> 0, 0
$ `SegmentUrban hip:ownHomeownYes`   <dbl> 0, 0
```

Consumer segmentation survey

- Testing Multiple Group Means: Analysis of Variance (ANOVA)
  - Model

$$\begin{bmatrix} 49482.81 \\ 35546.29 \\ \vdots \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_2 \\ \beta_2 \\ (\alpha\beta)_{13} \\ (\alpha\beta)_{14} \\ (\alpha\beta)_{22} \\ (\alpha\beta)_{32} \\ (\alpha\beta)_{42} \end{bmatrix}$$

- Coefficients to estimate using `aov`

$$\widehat{\mu} = \widehat{\mu}_{11}, \widehat{\alpha}_2, \widehat{\beta}_2, (\widehat{\alpha\beta})_{13}, (\widehat{\alpha\beta})_{14}, (\widehat{\alpha\beta})_{22}, (\widehat{\alpha\beta})_{32}, (\widehat{\alpha\beta})_{42}$$

References

Chapman, Chris, and Elea McDonnell Feit. 2019. *R For Marketing Research and Analytics*. 2nd ed. 2019. Use R! Cham: Springer International Publishing : Imprint: Springer. https://doi.org/10.1007/978-3-030-14316-9.