Introduction to Statistics Exam (0560.1823)- MOED A

Adi Sarid

Feb 2020

General Instructions

The test has five questions, each with 20 points.

The test is with open materials, e.g. books, formula pages, or whatever you want. You can use a calculator. You cannot use a laptop.

The test's duration is 3 hours.

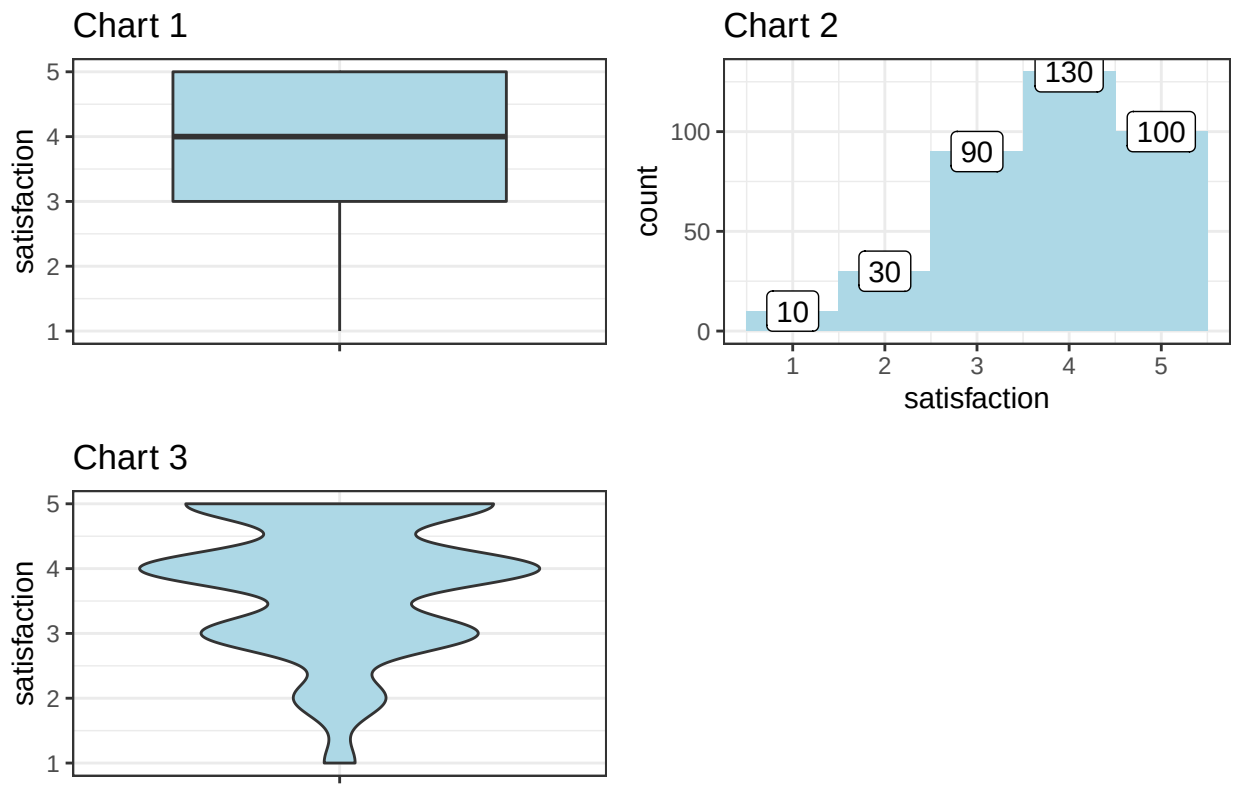
I will be circulating around in case you have any questions during the test.

Good luck!

Question 1 (20 pts)

A researcher conducted a survey with undergraduate students, about their satisfaction from their studies. The researcher included a question about the age of the respondent and a satisfaction question on a 5-point scale (i.e., 1 = extremely dissatisfied, 2 = dissatisfied, 3 = neither satisfied nor dissatisfied, 4 = satisfied, 5 = extremely satisfied).

The following charts show three different ways in which the distribution of satisfaction can be expressed.



Questions:

a. Which of the charts is better suited to show the distribution of **satisfaction**? explain. *The best answer here is the bar chart (the second chart), because the distribution of satisfaction is discrete and this is lost in the other two charts. The only benefit we get from the boxplot (the first chart) is the median, but it can also be deduced very easily from the bar chart. The third chart (violin plot) is not useful here, an in fact even confusing.*

b. What is the median satisfaction? *4, as can be seen easily from the boxplot or from computation of the bar chart.*

c. What chart contains the most information? *The bar chart has the labels so we can recreate the data of satisfaction entirely. It is the chart containing the most information. The boxplot contains the median which is not included in the bar chart but you can compute the median from the bar chart and cannot compute the sample size from the boxplot.*

d. Use the chart (from previous question) to compute: (1) the proportion of satisfied+extremely satisfied students; (2) the average satisfaction level (i.e., a number on the 1-5 scale). 0.6388889.

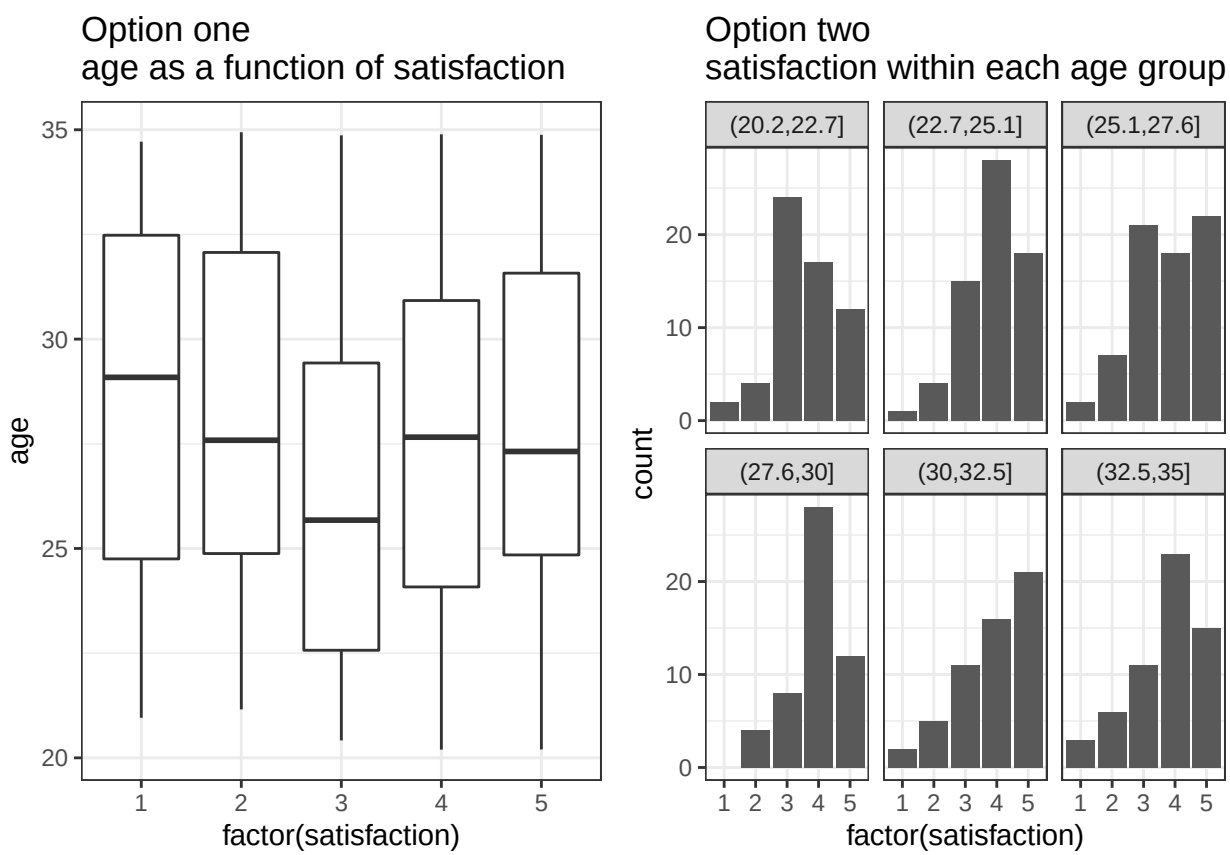
e. The researcher wants to visualize the relationship between age and satisfaction. What `geom_*` and

aesthetic mapping would you use for that? explain and draw an illustration (it doesn't have to be accurate). *A lot possible answers here. Some require pre-processing and some can be plotted directly. For example:*

```
option1 <- ggplot(students_survey, aes(x = factor(satisfaction), y = age)) +
   geom_boxplot() +
   theme_bw() +
   ggtitle("Option one\nage as a function of satisfaction")

option2 <- students_survey %>%
   mutate(age_group = cut(age, breaks = 6)) %>%
   ggplot(aes(x = factor(satisfaction))) +
   geom_bar() +
   facet_wrap(~ age_group) +
   theme_bw() +
   ggtitle("Option two\nsatisfaction within each age group")

cowplot::plot_grid(option1, option2)
```



Question 2 (20 pts)

Suppose we have a random sample of size $2n$ from a population denoted by X, and $E[X] = \mu, \mathrm{Var}[X] = \sigma^2$. Let

$$\bar{X}_1 = \frac{1}{2n}\sum_{i=1}^{2n} X_i \quad \text{and} \quad \bar{X}_2 = \frac{1}{n}\sum_{i=1}^{n} X_i$$

are two estimates for μ. Which is a better estimator of μ? Explain your choice (prove it!).

Answer

The variable $\bar{X}_1$ has a much lower variance then $\bar{X}_2$ which is what makes it better. Both are unbiased estimates for μ. Proof:

Unbiased:

$$E[\bar{X}_1] = \frac{1}{2n}\sum_{i=1}^{2n} EX_i = \mu$$

The standard error: for this part we have to assume that the X_i's are independent.

$$\mathrm{Var}\,\bar{X}_1 = \frac{1}{4n^2}\sum_{i=1}^{2n} \mathrm{Var}\,X_i = \frac{2n\sigma^2}{4n^2} = \frac{\sigma^2}{2n}$$

The estimate of $\bar{X}_2$ has a standard error of σ^2/n

Question 3 (20 pts)

The ministry for health is planning the purchase of flu vaccines for next year. For that purpose, data of vaccine consumption of the last few years has been concentrated:

- 2014: 1.52 million doses
- 2015: 1.56 million doses
- 2016: 1.63 million doses
- 2017: 1.91 million doses
- 2018: 1.53 million doses
- 2019: 1.83 million doses

Questions:

(a) Provide a 95% one-sided confidence interval for the number of doses used (an upper bound).

(b) Provide a 95% one-sided prediction interval for the number of doses used (an upper bound).

(c) Explain the difference between a confidence interval and a prediction interval.

(d) The minister wants to cover 95% probability that the vaccines do not run out for next year. What estimate would you use for planning? explain.

(e) Bonus: What kind of factors might intervene with the number of vaccine doses consumed during a specific year? How would you consider such factors (with what statistical model)?

Answers

In this question you just had to compute a one-sided confidence interval and one-sided prediction interval using the student's t-test table. The value of interest of the distribution is 2.0150484

```r
vaccine <- tibble::tibble(doses = c(1.52, 1.56, 1.63, 1.91, 1.53, 1.83))
t.test(vaccine$doses, alternative = "less")
```

```
##
##  One Sample t-test
##
## data:  vaccine$doses
## t = 24.46, df = 5, p-value = 1
## alternative hypothesis: true mean is less than 0
## 95 percent confidence interval:
##      -Inf 1.800363
## sample estimates:
## mean of x
##  1.663333
```

```r
# Alternatively here is the manual computation without using t.test:
mean(vaccine$doses) + sd(vaccine$doses)*sqrt(1/NROW(vaccine))*qt(0.95, df = NROW(vaccine) - 1)
```

```
## [1] 1.800363
```

```r
# prediction interval manually:
mean(vaccine$doses) + sd(vaccine$doses)*sqrt(1+1/NROW(vaccine))*qt(0.95, df = NROW(vaccine) - 1)
```

```
## [1] 2.02588
```

Confidence intervals are used to provide an interval of $1 - \alpha\%$ around the mean, where as prediction intervals provide a $1 - \alpha\%$ interval for a new observation.

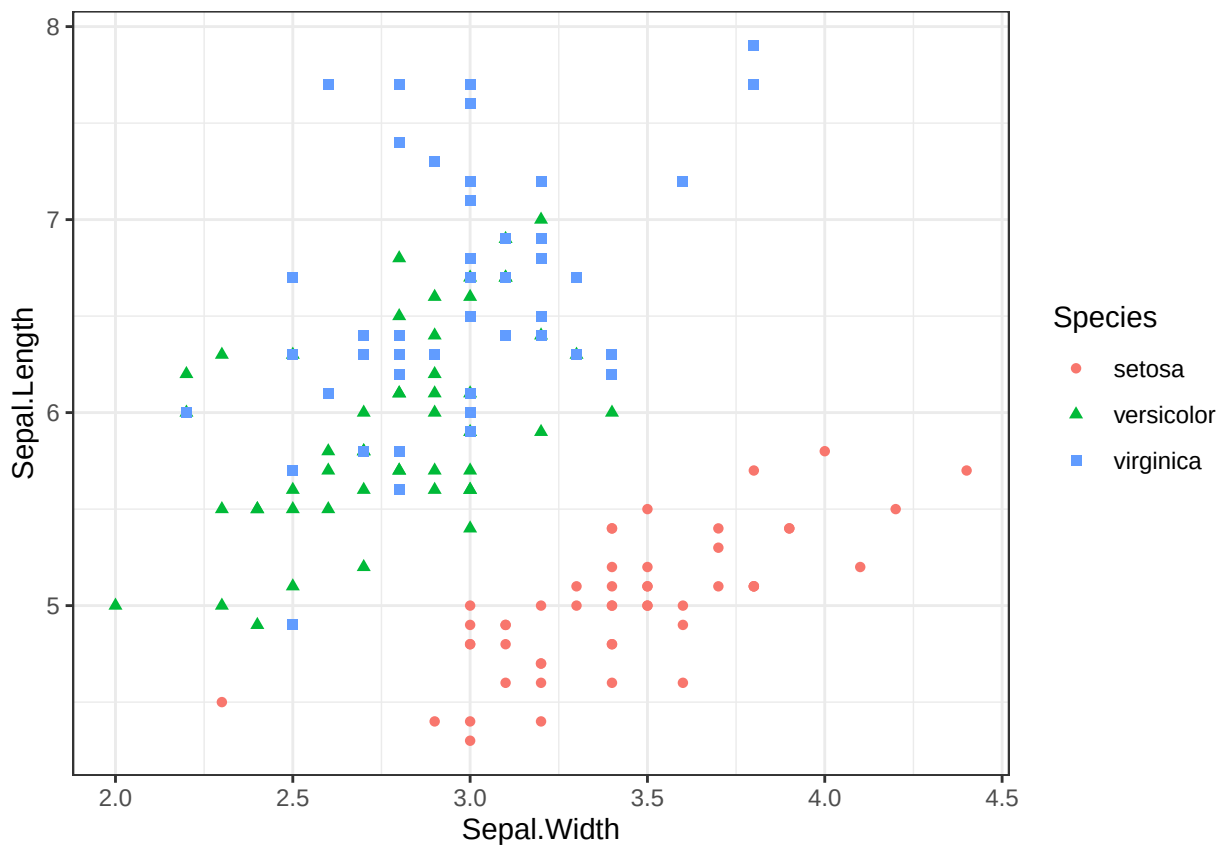
The minister should use a prediction interval, obviously, because this is a prediction on a new observation and for decision making we are interested in the value of next year, not the annual average.

For example, if there is a media covarage of a really lethal strand of the flu (e.g., swine influenze in 2009, or even the last flu season here in Israel). A statistical model (e.g., linear regression) with dummy variables indicating such factors might help express them in a model.

Question 4 (20 pts)

The `iris` dataset contains three types of flowers: setosa, virginica, and versicolor. The following chart shows the scatter plot of `Sepal.Width` and `Sepal.Length` variables for the species (the width and length of the parts which hold the base of the flower).

```
ggplot(iris, aes(Sepal.Width, Sepal.Length, color = Species, shape = Species)) +
  geom_point() +
  theme_bw()
```



The following linear models (simple linear regression models) show the relationship of the `Sepal.Width` and `Sepal.Length` variables. The first model includes all species and the second model only includes the setosa species.

Questions:

(a) What kind of measure can you use to compare the two models? *For example, R^2, but also the significance of the model and the significance of coefficients.*

(b) Explain which model is better and why. *The filtered model is better, as seen by either measure. The reason for this is that the relationship between Sepal.Width and Sepal.Length changes depending on the flower species, i.e., the slope (β coefficient) is different in each species.*

(c) In the second model (which includes just the setosa) what is the coefficient β of Sepal.Length? What is its meaning? (i.e., what is the interpretation of the coefficient in the realtionship between the Sepal.Width and Sepal.Length) $\beta = 0.7985$ a flower which has a Sepal.Length longer by 1cm will ave a Sepal.Width longer by 0.7985cm

(d) Can you think of a way to extend the first model so that it will still include all the species but have a

much better fit? explain!

```r
# The first model:
lm(formula = Sepal.Width ~ Sepal.Length, data = iris) %>% summary()
```

```
##
## Call:
## lm(formula = Sepal.Width ~ Sepal.Length, data = iris)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.1095 -0.2454 -0.0167  0.2763  1.3338
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.41895    0.25356   13.48   <2e-16 ***
## Sepal.Length -0.06188    0.04297   -1.44    0.152
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4343 on 148 degrees of freedom
## Multiple R-squared:  0.01382,    Adjusted R-squared:  0.007159
## F-statistic: 2.074 on 1 and 148 DF,  p-value: 0.1519
```

```r
# The second model:
lm(formula = Sepal.Width ~ Sepal.Length, data = iris %>% filter(Species == "setosa")) %>%
  summary()
```

```
##
## Call:
## lm(formula = Sepal.Width ~ Sepal.Length, data = iris %>% filter(Species ==
##     "setosa"))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.72394 -0.18273 -0.00306  0.15738  0.51709
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -0.5694     0.5217  -1.091    0.281
## Sepal.Length   0.7985     0.1040   7.681 6.71e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2565 on 48 degrees of freedom
## Multiple R-squared:  0.5514, Adjusted R-squared:  0.542
## F-statistic: 58.99 on 1 and 48 DF,  p-value: 6.71e-10
```

Answer for (d)

Make a regression model with the interaction of species

```r
lm(formula = Sepal.Width ~ Sepal.Length*Species, data = iris) %>%
  summary()
```

```
##
## Call:
```

```
## lm(formula = Sepal.Width ~ Sepal.Length * Species, data = iris)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.72394 -0.16327 -0.00289  0.16457  0.60954
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    -0.5694     0.5539  -1.028 0.305622
## Sepal.Length                    0.7985     0.1104   7.235 2.55e-11 ***
## Speciesversicolor               1.4416     0.7130   2.022 0.045056 *
## Speciesvirginica                2.0157     0.6861   2.938 0.003848 **
## Sepal.Length:Speciesversicolor  -0.4788     0.1337  -3.582 0.000465 ***
## Sepal.Length:Speciesvirginica   -0.5666     0.1262  -4.490 1.45e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2723 on 144 degrees of freedom
## Multiple R-squared:  0.6227, Adjusted R-squared:  0.6096
## F-statistic: 47.53 on 5 and 144 DF,  p-value: < 2.2e-16
```

Question 5 (20 pts)

Before you are four charts, for each chart write which distribution it depicts out of the following list:

- Student's t
- Chi square
- F

(Please note that there are additional parts of this question after the charts!)

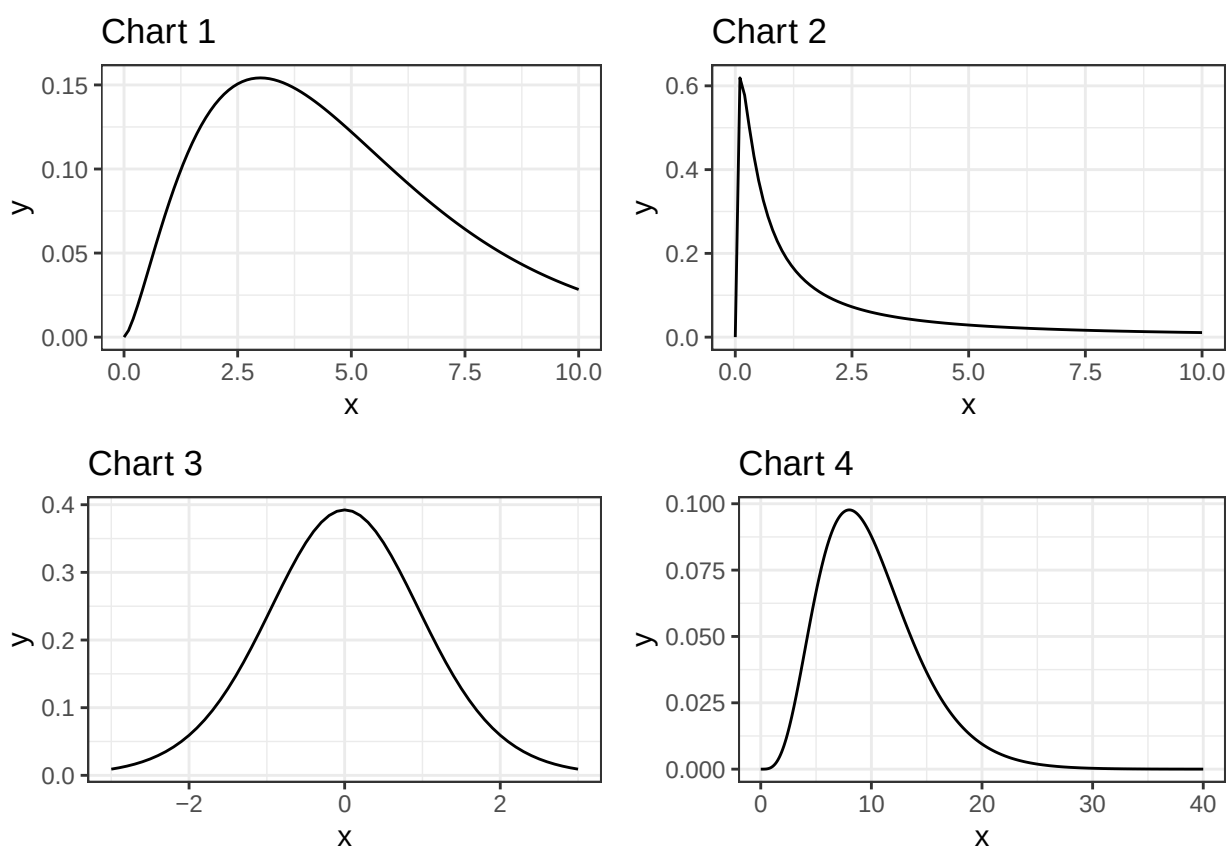
```
chisq <- tibble(x = seq(0,10, by = 0.1)) %>%
   mutate(y = dchisq(x, 5)) %>%
   ggplot(aes(x, y)) + geom_line() + theme_bw() + ggtitle("Chart 1")

fdist <- tibble(x = seq(0,10, by = 0.1)) %>%
   mutate(y = df(x, 3, 1)) %>%
   ggplot(aes(x, y)) + geom_line() + theme_bw() + ggtitle("Chart 2")

students_t <- tibble(x = seq(-3, 3, by = 0.1)) %>%
   mutate(y = dt(x, df = 15)) %>%
   ggplot(aes(x, y)) + geom_line() + theme_bw() + ggtitle("Chart 3")

chisq2 <- tibble(x = seq(0,40, by = 0.1)) %>%
   mutate(y = dchisq(x, 10)) %>%
   ggplot(aes(x, y)) + geom_line() + theme_bw() + ggtitle("Chart 4")

cowplot::plot_grid(chisq, fdist, students_t, chisq2)
```



For each of the following statistical tests, which distribution is used for the test statistic:

- Linear regression: a single β_i test, i.e., $H_0 : \beta_i = 0$ and $H_1 : \beta_i \neq 0$. *t-test*
- Linear regression: there exists a non-zero coefficient. I.e., $H_0 : \forall i \beta_i = 0$ and $H_1 : \exists i : \beta_i \neq 0$. *F-test*
- ANOVA test. *F-test*
- Goodness of fit. *Chi square*