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.

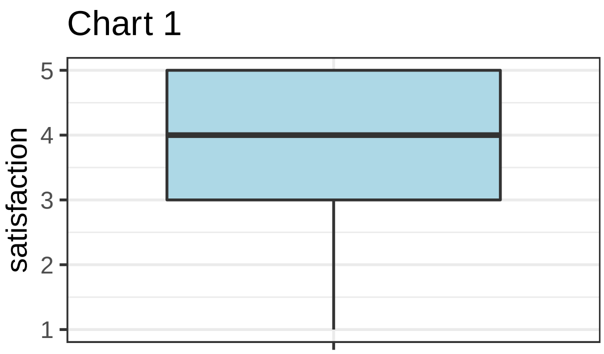
I will be circulating around in case you have any questions during the test.

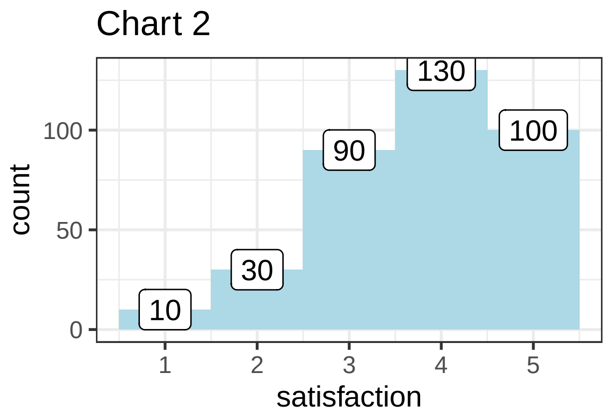
Good luck!

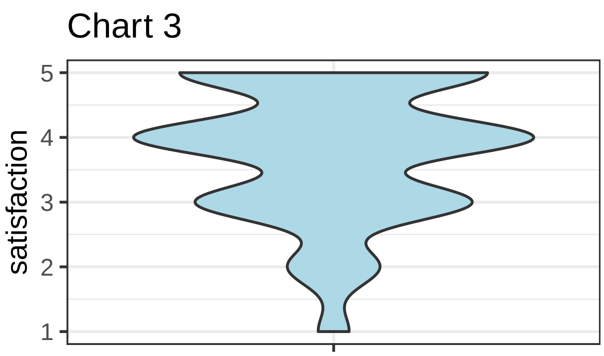
Question 1 (20 pts)

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.

b. What is the median satisfaction?

c. What chart contains the most information?

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).

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).

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!).

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.

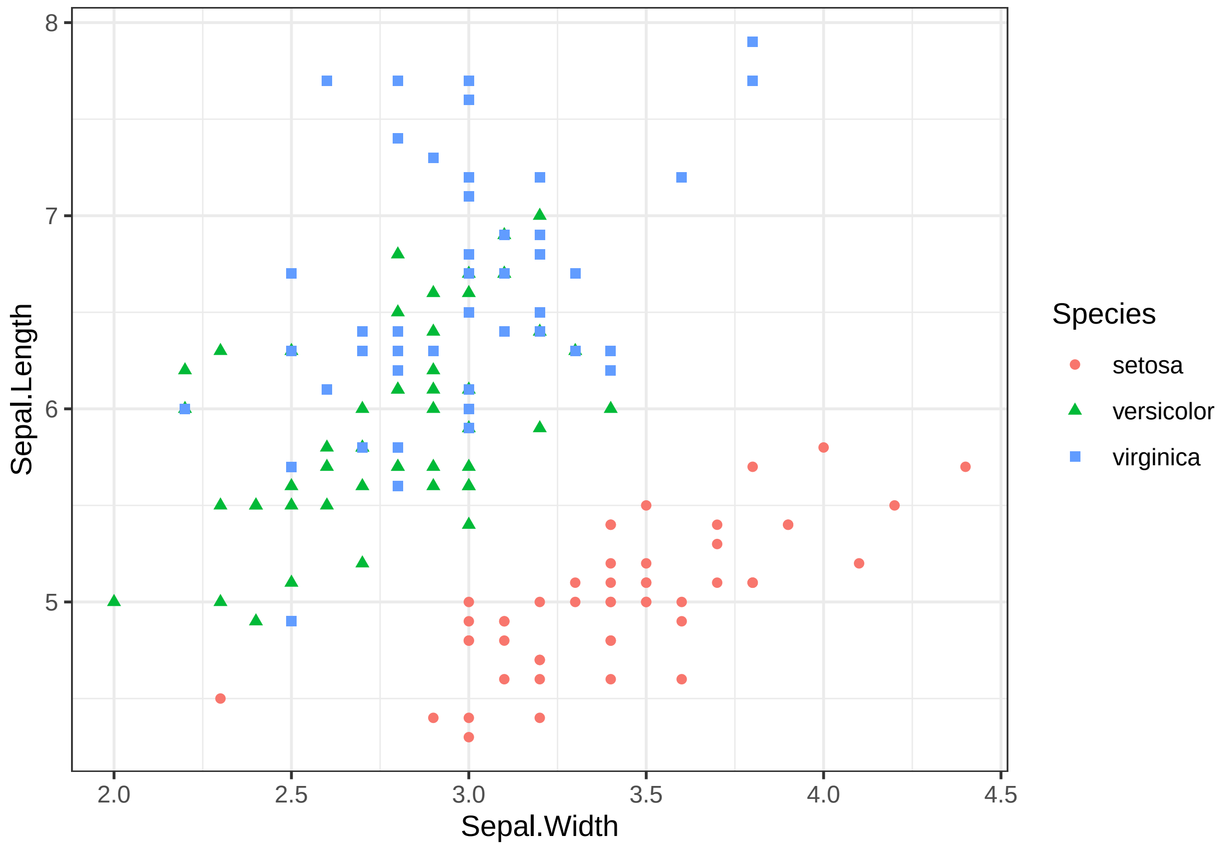
(d) The minister wants to cover 95% probability that the vaccines do not run out for next year. What estimate would you use for planning? explain.

(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)?

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).

```r
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?

(b) Explain which model is better and why.

(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)

(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
```

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!)

Chart 1

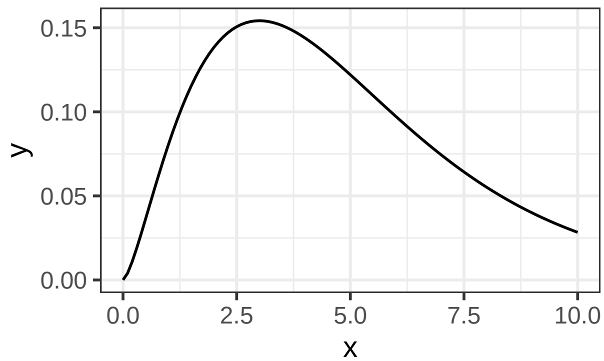


Chart 2

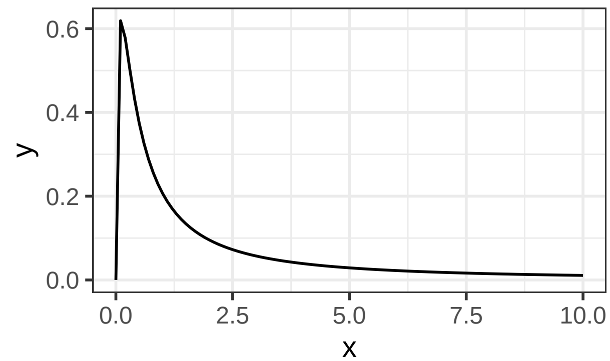


Chart 3

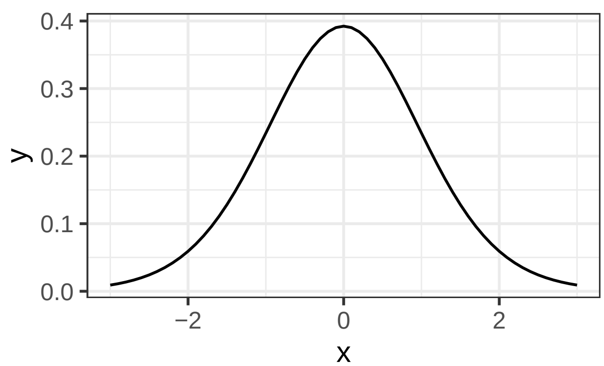
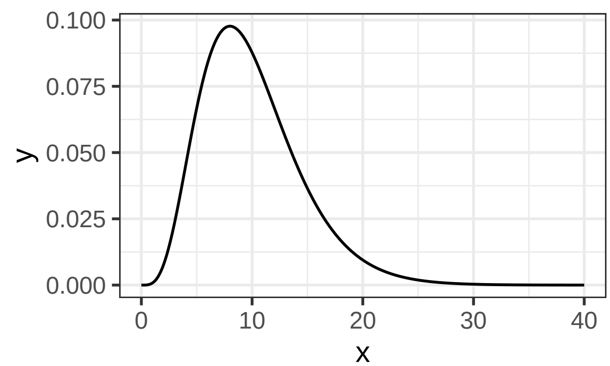


Chart 4



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