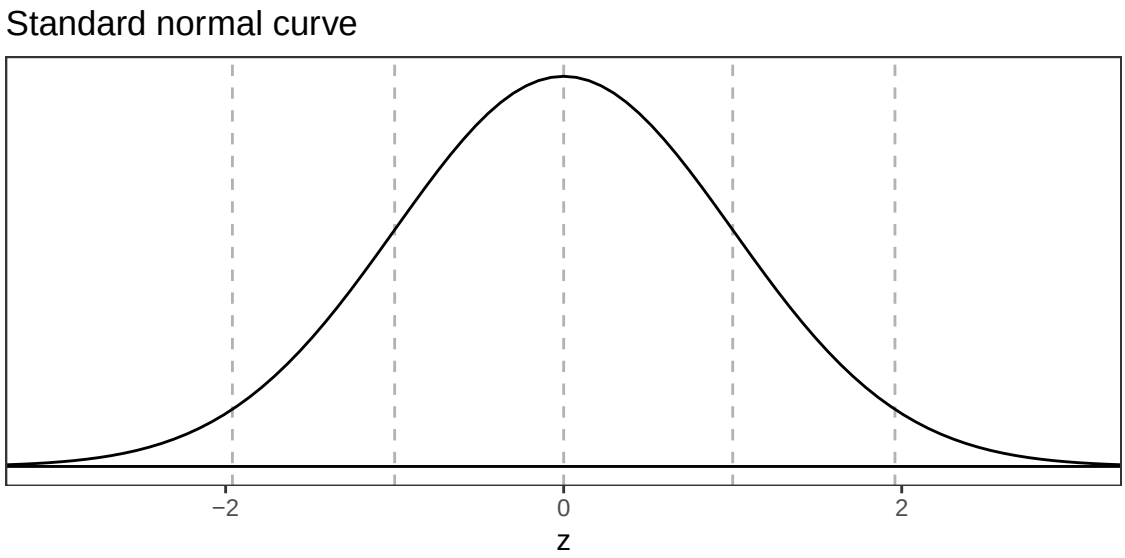
1 Short Answer

a) Below we have a standard Normal Z-curve along with 5 vertical dashed lines at z = -1.96, -1, 0, 1, and 1.96 cutting the x-axis into 6 segments. In the plot below, write down the 6 proportion of values under the Z-curve in each of the 6 segments. Hint: Your 6 proportions should sum to 100%.



a) Analysis of Variance (ANOVA) compares k group means for the following hypothesis test:

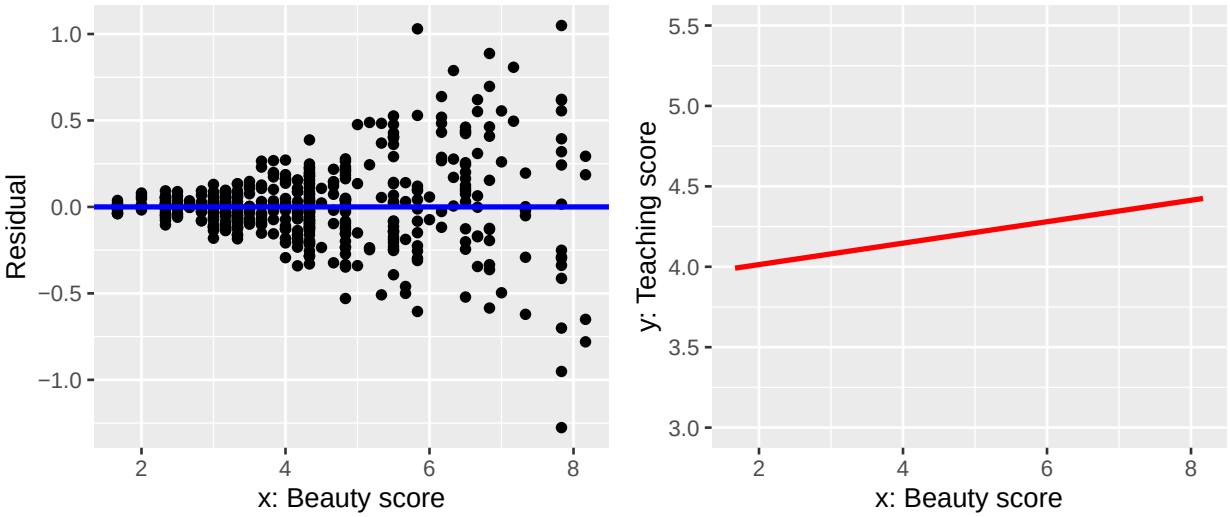
$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$$
$$\text{vs.} \quad H_A : \text{At least one of the } k \text{ means is different}$$

For example, in class we compared the mean life expectancy of countries in $k = 5$ continents. What other statistical technique covered in this course would allow us to similarly compare group means?

d) A *test statistic* is a X of the unknown population parameter of interest used for hypothesis testing. What is X?

f) The *null distribution* used in hypothesis testing for computing p-values is the X distribution of the test statistic assuming Y. What are X and Y?

d) Say we perform a regression to model an instructors' teaching score as a function of their beauty score, and obtain the following residual plot on the left which exhibits heteroskedasticity. Draw a *rough* sketch of what the scatterplot of x and y would look like given that the red line is the fitted regression line.



e) Say we perform a residual analysis of a regression model and find that the residuals exhibit very strong heteroskedasticity as above. What implications does this have for the results of our analysis?

2 Putting it all together: "Formatting is off"

The office of the president of a small liberal arts college in New England wants to promote the public launch of a fundraising campaign to all alumni. In particular, they would like the email to include a quote by American novelist and essayist Marilynne Robinson, followed by a link to donate money. However, the president is concerned that the formatting of the email will affect the "click-through rate" of the link: the proportion of those receiving the email that follow through and click on the link. In particular, they are **very** concerned about any possible differences in click-through rates arising due to the formatting of the quote attribution. So the office creates the two versions of the same email where the only difference is the quote attribution:

Version 1:

President's Office <president@[redacted].edu>
Reply-To: President's Office <president@[redacted].edu>
To: [redacted]

In the U.S., education, especially at the higher levels, is based around powerful models of community. We choose our colleges in order to be formed by them and supported by them in the identities we have or aspire to. If the graft takes, we consider ourselves ever after to be members of that community. As one consequence, graduates tend to treat the students who come after them as kin and also as heirs. They take pride in the successes of people in classes forty years ahead of or behind their own. They have a familial desire to enhance the experience of generations of students who are, in fact, strangers to them, except in the degree that the ethos and curriculum of the places does indeed form its students over generations.
Marilynne Robinson

Version 2:

President's Office <president@[redacted].edu>
Reply-To: President's Office <president@[redacted].edu>
To: [redacted]

In the U.S., education, especially at the higher levels, is based around powerful models of community. We choose our colleges in order to be formed by them and supported by them in the identities we have or aspire to. If the graft takes, we consider ourselves ever after to be members of that community. As one consequence, graduates tend to treat the students who come after them as kin and also as heirs. They take pride in the successes of people in classes forty years ahead of or behind their own. They have a familial desire to enhance the experience of generations of students who are, in fact, strangers to them, except in the degree that the ethos and curriculum of the places does indeed form its students over generations.

--Marilynne Robinson

Here is the sequence of events:

- They randomly select 25,138 alumni from the alumni database to send emails to.
- From these 25,138 alumni, they randomly choose 12,460 alumni and send them email Version 1. They send Version 2 to the remaining 12,678 alumni.
- Of those alumni who received Version 1 10,578 followed through and clicked the link for a rate of 84.9%. Of those alumni who received Version 2 11,169 followed through and clicked the link for a rate of 88.1%.

a) What kind of study are we considering: an observational study or a randomized experiment? Why?

b) In this scenario, can we establish the *causal* effect (and not just the *associated* effect) of the formatting on click-through rate? Why or why not?

c) Who is the study population in this scenario?

d) What is the statistical name of the population parameter of interest in this scenario?

e) What is the mathematical notation for the population parameter of interest in this scenario?

f) What is the statistical name for the point estimate (AKA sample statistic) of the population parameter of interest in this scenario?

g) What is the mathematical notation for the point estimate of interest in this scenario?

h) What is the numerical value of the point estimate of interest in this scenario?

i) The standard error of the point estimate in this question can roughly be approximated by the mathematical formula when constructing confidence intervals:

$$\text{SE}_{\widehat{p}_1-\widehat{p}_2} \approx \sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n_1} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{n_2}}$$

Construct a 95% confidence interval appropriate to answer the president's concerns.

j) We say that we are "95% confident" that this confidence interval captures the true value of the unknown population parameter. However, we say this as shorthand for the more involved statistical interpretation of a confidence interval. What is this precise statistical interpretation?

k) Write down the relevant hypothesis test using non-statistical language.

k) Write down the relevant hypothesis test using mathematical notation.

l) What is the statistical name of the relevant test statistic in this scenario?

m) What is the numerical value of the *observed* test statistic in this scenario?

m) What is being assumed throughout this hypothesis testing scenario?

m) The standard error of the point estimate in this question can roughly be approximated by the mathematical formula when conducting hypothesis tests:

$$\text{SE}_{\widehat{p}_1 - \widehat{p}_2} \approx \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n_1} + \frac{\widehat{p}(1-\widehat{p})}{n_2}}$$

where $\widehat{p}$ is the *pooled sample proportion* where you pool all observations in both groups into a single group and compute a single proportion:

$$\widehat{p} \quad = \quad \frac{\text{\# of Version 1 link clicks} + \text{\# of Version 2 link clicks}}{n_1 + n_2}$$

For hypothesis testing, why is this pooling appropriate?

n) Draw the null distribution. Hint: while not always the case, the sampling distribution of the point estimate of interest is normally shaped.

j) Recall that the president is **very** concerned about any possible differences in click-through rates arising due to the formatting of the quote attribution. If conducting a hypothesis test, would you use a "liberal" α value or a "conservative" α value?

k) Based on your response above, conduct the hypothesis test.

l) Based on your analysis, what do you tell the President? Keep in mind the President is very busy (monitoring the formatting of emails for example), so they would prefer a shorter response.

3 Evals continued

Recall the evals data of teaching evaluations of professors. Let say instead that these 463 professors are a randomly chosen set of instructors from all of the University of Texas system and not just UT Austin. Consider the following *simple linear regression* using only one numerical explanatory variable:

```
score_model <- lm(score ~ age, data = evals)
get_regression_table(score_model)
```

term	estimate	std_error	lower_ci	upper_ci
intercept	4.46	0.13	4.21	4.7
age	-0.01	0.00	-0.01	0.0

a) Interpret the slope coefficient for age.

b) Using statistical language, interpret the standard error for the slope for age.

c) Using non-technical language, interpret the standard error for the slope for age.

4 Confidence Intervals

Recall we saw an example of an NPR poll of $n = 2089$ young Americans' approval of Obama back in 2013. Of these respondents, 856 said they approved of Obama's job performance.

a) What is the numerical value of $\widehat{p}$, the point estimate of the population proportion p of all young Amercians who approve of Obama's job performance?

b) Say CBS conducted a similar poll with $n = 2089$ and finds that 860 young Americans approve of Obama, leading to one point estimate $\widehat{p}$ of p. Say NBC conducted a similar poll with $n = 2089$ and finds that 844 young Americans approve of Obama, leading to another point estimate $\widehat{p}$ of p. Say Buzzfeed News conducted a similar poll with $n = 2089$ and finds that 871 young Americans approve of Obama, leading to yet another point estimate $\widehat{p}$ of p. What is the name of the value that quantifies this variability?

c) Construct a 95% confidence interval for the population proportion p of all young Americans who approved of Obama's job performance. Note the following mathematical formula approximating the standard error:

$$\text{SE}_{\widehat{p}} \approx \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}$$

d) Marc-Edouard Vlasic states "I read on NPR that back in 2013, as little as 43% of **all** young Americans approved of Obama." What assumption must be met for Marc-Edouard's statement to be valid?

e) What assumption about the sampling distribution of $\widehat{p}$ must be met for the confidence interval in part c) to be valid?

5 Inference for Regression

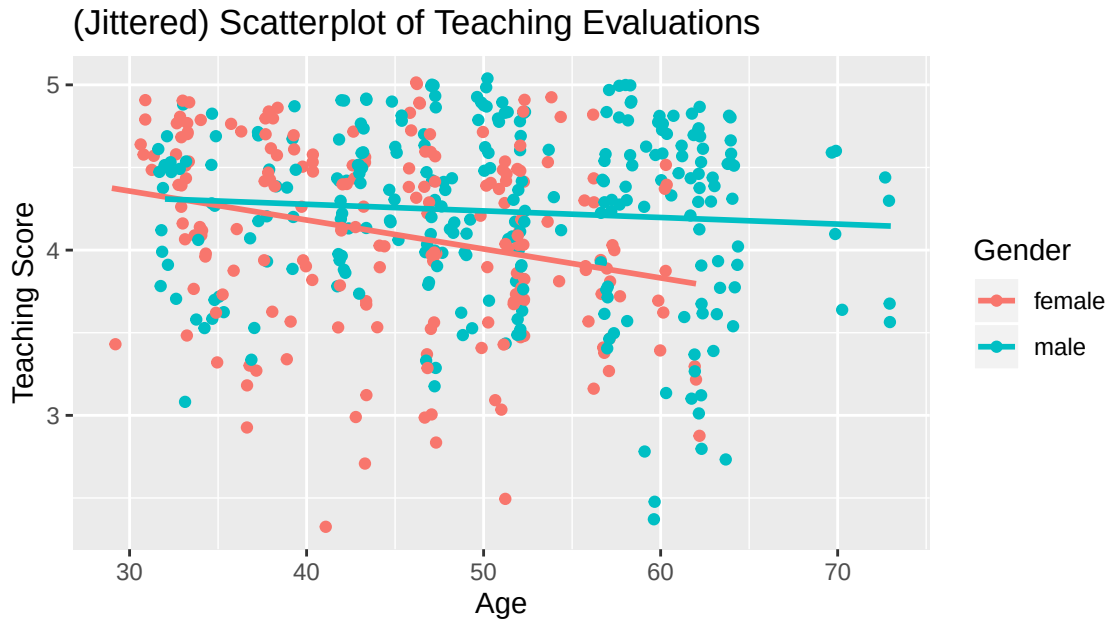
Recall our professor evaluations dataset based on the study from the University of Texas in Austin. In particular, we were interested in explaining a professor's teaching evaluation score using their gender and age as explanatory variables. Here is a random sample of 5 rows out of the $n = 463$ professors in dataset:

```
## # A tibble: 5 x 3
##   score gender   age
##   <dbl> <fct>  <int>
## 1   4.3 female    56
## 2   4.2 male      48
## 3   3.9 male      42
## 4   3.3 female    37
## 5   3.6 female    38
```

Recall we fit the following regression model *with an interaction term*:

$$\begin{aligned} \widehat{y} &= b_0 + b_1 x_1 + b_2 x_2 + b_3 x_1 x_2 \\ \widehat{\text{score}} &= b_0 + b_{\text{age}}\text{age} + b_{\text{male}}\mathbb{1}[\text{is male}] + b_{\text{age,male}}\text{age}\mathbb{1}[\text{is male}] \end{aligned}$$

Recall the visual representation of the our model. Hint: look at this closely.



Finally, recall the results of the regression with confidence intervals

```
evals_model <- lm(score ~ age * gender, data=evals)
get_regression_table(evals_model)
```

term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
intercept	4.88	0.20	23.8	0.00	4.48	5.29
age	-0.02	0.00	-3.9	0.00	-0.03	-0.01
gendermale	-0.45	0.26	-1.7	0.09	-0.97	0.08
age:gendermale	0.01	0.01	2.5	0.02	0.00	0.02

a) The table reports a p-value of 0 in the age row. Write down the corresponding hypothesis H_0 vs H_A in terms of the β_{age}, the true population associated effect of age on teaching score.

b) The p-value mentioned in part a) is 0. Report what this means for the hypothesis test corresponding to the two hypotheses above. Report this both in 1) statistical terms and 2) language that non-statisticians can understand.

c) Based on these results, among male professors at the University of Austin for every year increase in age, there is an associated X of on average Y units in teaching score. What are X and Y?

d) What conclusion is suggested by the 95% confidence interval for $\beta_{\text{age:gendermale}}$ of $(0.003, 0.024)$?

e) Say we relaxed the gender categorical variable to allow for the following three levels: female, male, and non-binary, and furthermore say some professors selected the new "non-binary" option. Describe precisely how the above plot would change.

f) BONUS 1 Describe precisely how the shape of the above regression table would change.

g) BONUS 2 The 95% confidence interval for $\beta_{\text{gendermale}}$ is $(-0.968, 0.076)$. Based on values in the table, write down your best guess of the formula that R uses to compute the left end point of -0.968. Your formula and the reported left endpoint of -0.968 should match up to 2 decimal places.

6 Regression

You run the code below to analyze departure delays from the 3 New York City airports, but for some weird reason, you only get the incomplete output below. Note AS corresponds to Alaska, F9 corresponds to Frontier, and AA corresponds to American.

```
library(dplyr)
library(nycflights13)
library(moderndive)

flights_subset <- flights %>%
  filter(carrier == "AS" | carrier == "F9" | carrier == "AA")

dep_delay_model <- lm(dep_delay ~ carrier, data = flights_subset)
get_regression_table(dep_delay_model, digits = 3)
```

term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
intercept	8.6	0.21	40.7	0.000	8.2	8.999
carrierAS	-2.8	1.43	-1.9	0.052	-5.6	0.025
carrierF9	11.6	1.46	8.0	0.000	8.8	NA

a) Interpret the 11.6 estimate value in the `carrierF9` row (third row, second column). Is its relationship of with the outcome variable meaningful?

b) Compute the missing right endpoint of the 95% confidence interval in the `carrierF9` row.

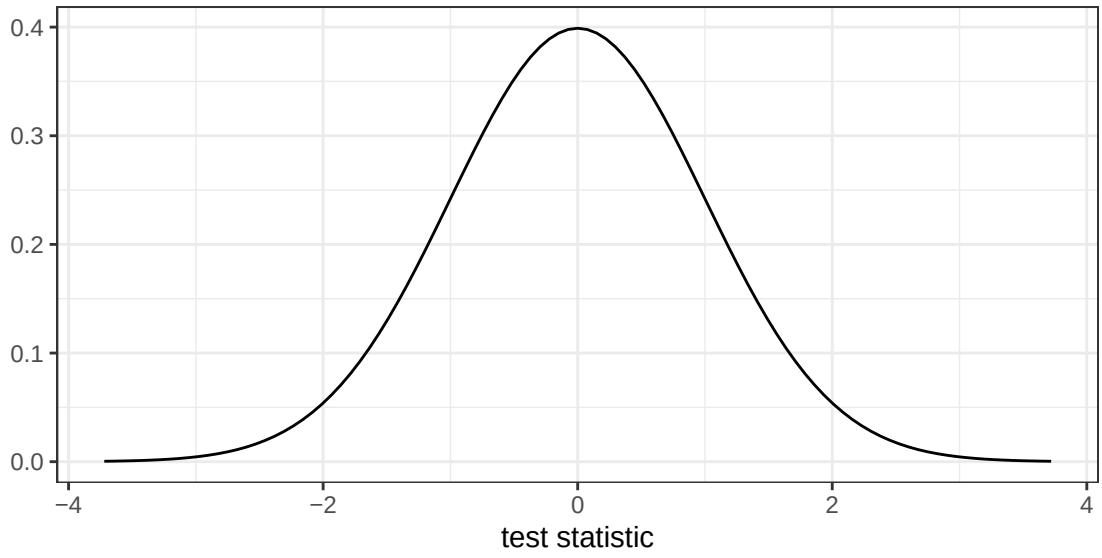
c) State the scientific conclusion reached based on the now complete 95% confidence interval.

e) Write down the hypothesis test corresponding to the `carrierAS` row using mathematical notation. Do not carry out the hypothesis test, simply state the two competing hypotheses.

f) Say you were given an α cutoff value of 0.01 for the hypothesis test above. Write down the conclusion of this hypothesis test both in statistical terms and using non-statistical language that an airline executive can understand.

c) In the second row, fifth column there is a p-value missing. What is the hypothesis test corresponding to this missing p-value?

d) Sketch on the follow plot of the corresponding *null* distribution what the missing p-value in the second row, fifth column is:


0.4
0.3
0.2
0.1
0.0
−4
−2
0
2
4
test statistic