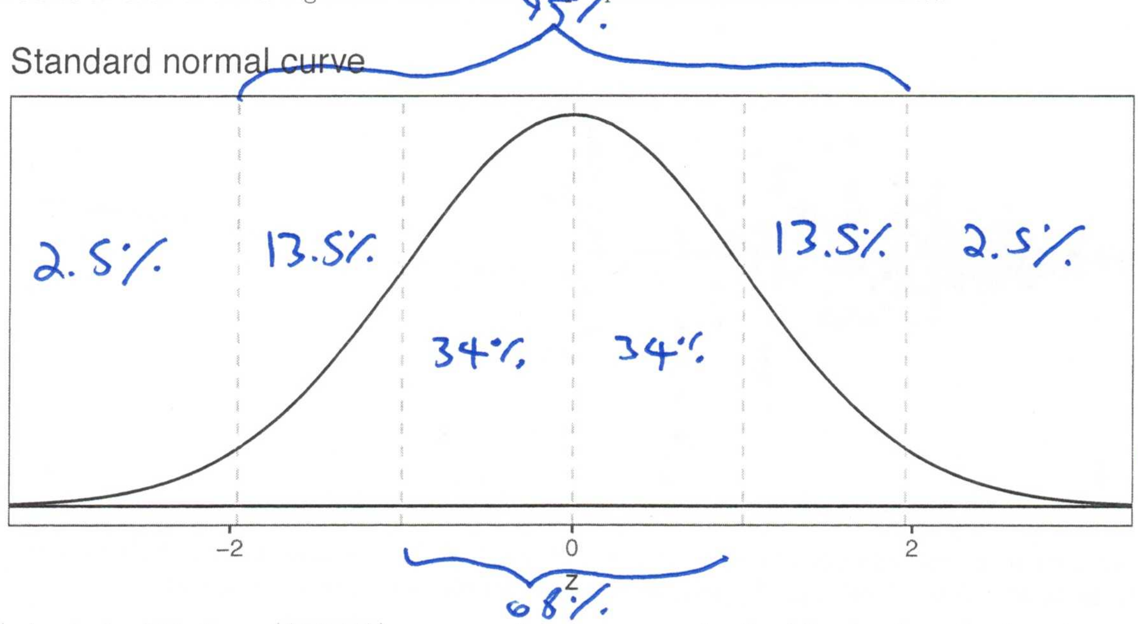
1 Short Answer

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%.



why 13.5%
95%
- 68%
= 27%
then
divided
by 2

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?

Testing for equality of group means is achieved by regression with single categorical predictor

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

point estimate

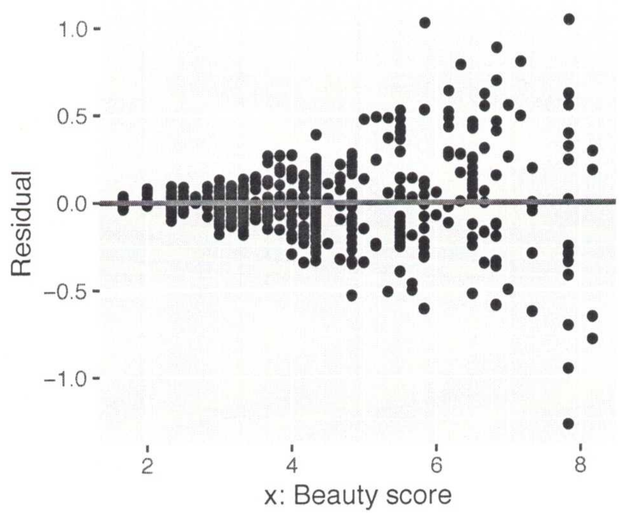
f) The *null distribution* used in hypothesis testing for computing p-values is the X distribution of the test statistic assuming Y. What are X and Y?

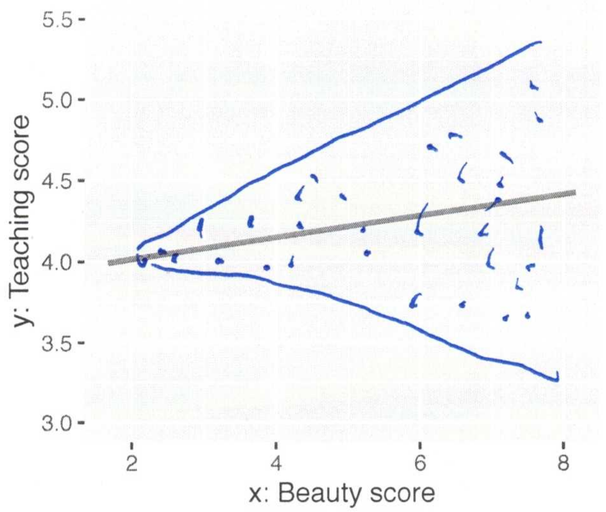
X = "sampling"

Y = the null hypothesis is true

AKA under Ho

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?

For p-values & confidence intervals for regression to have valid interpretation, all 4 conditions for inference must be met. If violated, we'll need to improve the model somehow

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.

For every ↑ of 1 year in age, there is an associated decrease of on avg 0.01 teaching units

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

If we were to take more samples of professors from the same or similar population & repeated this analysis, the SE quantifies the variation in the slope for age due to sampling variation

~~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?

$$\frac{856}{2089} = 0.41 = 41\%$$

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?

Standard error (which is the standard deviation of the point estimate $\hat{p}$)

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:

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

* If sampling distribution is normal, 95% CI is

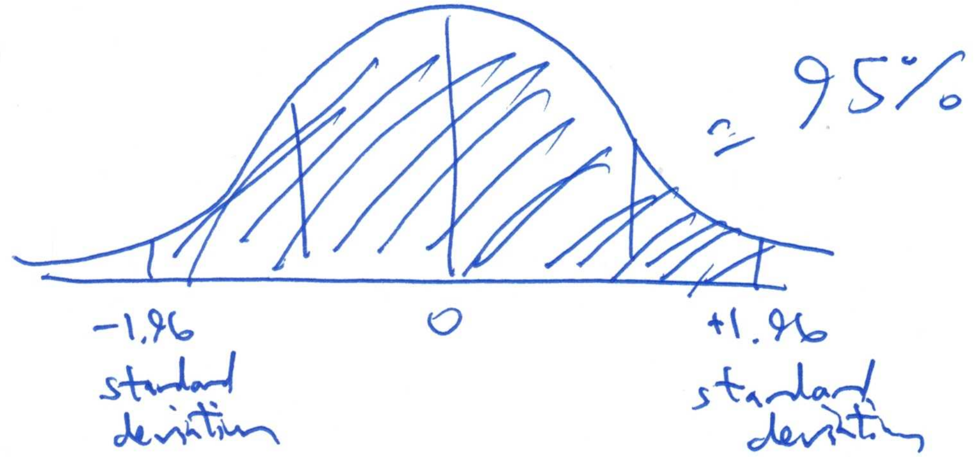
$$\begin{aligned} & PE \pm 1.96 \times SE \\ &= \hat{p} \pm 1.96 \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \\ &= 0.41 \pm 1.96 \times 0.011 \\ &= 0.41 \pm 0.02 \\ &= [0.39, 0.43] \end{aligned}$$

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?

the sampling had to be representative. We can ensure this by sampling @ random

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

has to be normally shaped to use



rule

5 Inference for Regression

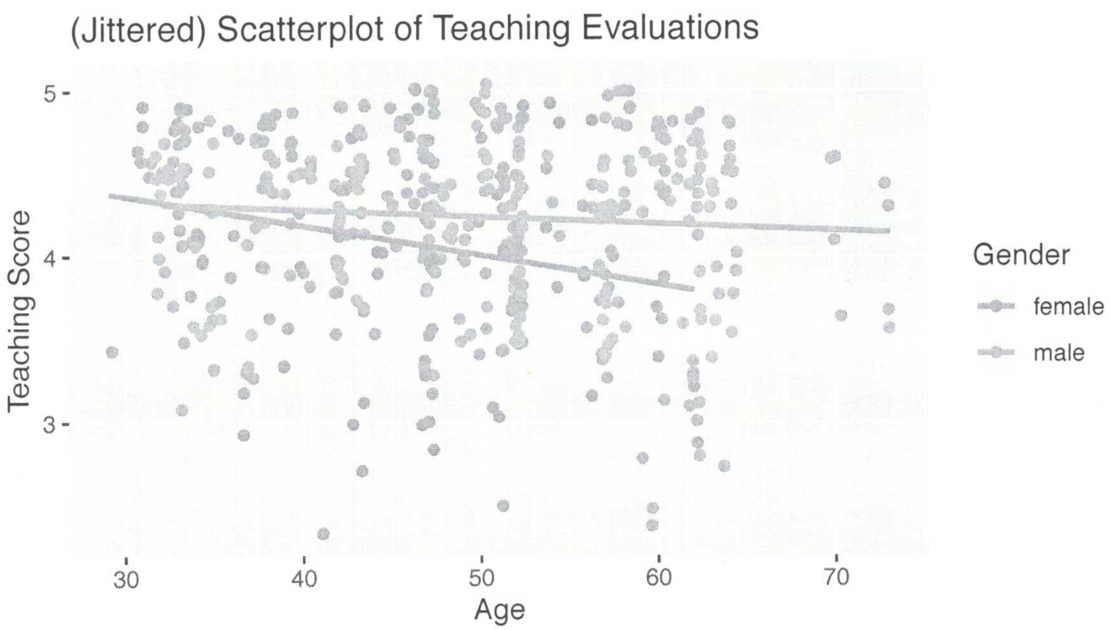
Recall our professor evaluations dataset based on the study from the University of Texas in Austin. In particular, we were interested in explaining a professor's teaching evaluation score using their gender and age as explanatory variables. Here is a random sample of 5 rows out of the $n = 463$ professors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

$H_0: \beta_{age} = 0$

vs $H_A: \beta_{age} \neq 0$

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.

1) Reject H_0 in favor of H_A

2) Reject hypothesis that there is no relationship between age & teaching score in favor of hypothesis that there is!

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?

X = decrease

$\boxed{Y = 0.02}$

Should be
y = - 0.02 + 0.01 = - 0.01

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

Since 0 is NOT in CI, there <u>IS</u> a diff in slope for age for males vs baseline group females. Age affects female faculty different. Consistent with visualization

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.

- third regression line
- third color of points

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

two extra rows

gender non binary = offset in intercept

age:gender non binary = difference in slope for age

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.

sampling distribution of PE $b_{\text{gendermale}}$ is roughly normal, so 95% CI is PE $\pm 1.96 \times$ SE

$= -0.45 \pm 1.96 \times 0.26$

$= -0.45 \pm 0.51$

$= [-0.96, 0.06]$

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)
```

table values are rounded

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?

Diff in departure delay for Frontier vs AA on avg 11.6.

Since 95% CI is [8.8, NA]
part b)

does not contain 0,
~~So~~ yes, it is meaningful.

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

$$PE \pm 1.96 \times SE$$
$$= 11.6 \pm 1.96 \times 1.46$$
$$= 11.6 \pm 2.86$$
$$= [8.74, 14.46]$$

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

Since CI does not contain 0, it suggests there is a significant difference in avg departure delay between AA & Frontier

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.

$H_0: \beta_{carrierAS} = 0$

vs $H_A: \beta_{carrierAS} \neq 0$

We don't have evidence to reject hypothesis that there is no difference in delays between AA and AS

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.

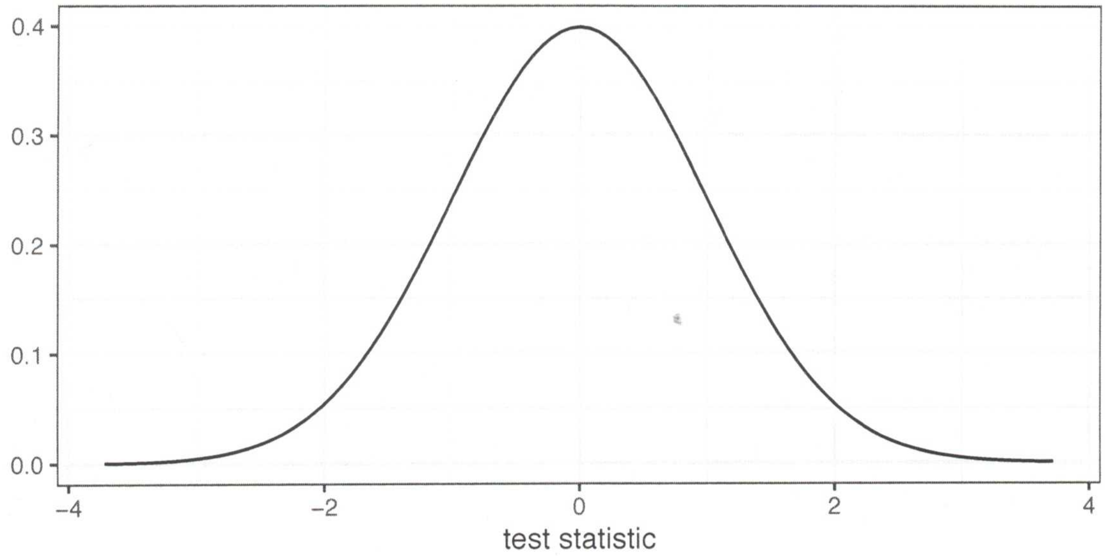
Since p-value $0.052 > \alpha = 0.01$ we do not reject H_0.
i.e. we do not have evidence to suggest the avg difference in delays for AS vs AA is not 0.

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

mistake question!

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

ignore.


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