1 Exploratory data analysis via data wrangling

Recall the Google Forms survey you completed in Lecture 2 where:

- Students with an odd birthday (Ex: Nov 15th) were first asked if there are more or less than **14 countries** in Africa and then asked to guess how many countries there are in Africa.
- Students with an even birthday (Ex: Nov 14th) were first asked if there are more or less than **94 countries** in Africa and then asked to guess how many countries there are in Africa.

Let's refer to the numbers 14 and 94 as "priming" numbers since survey participants were "primed" with them in order to influence the number of countries they guessed. Furthermore all students were also asked their height (in inches), their graduation year (2019, 2020, 2021, or 2022), and whether or not they had previously been to Africa. A total of 41 students responded and the results are saved in a data frame `africa` with 41 rows:

```
## # A tibble: 41 x 5
##     year height been_to_africa priming      how_many_countries
##    <int>  <int> <chr>          <chr>                     <int>
##  1  2021     70 No             14 countries                 36
##  2  2020     67 No             94 countries                120
##  3  2021     69 No             14 countries                 30
##  4  2021     60 Yes            14 countries                 64
##  5  2021     66 No             14 countries                  1
##  6  2021     66 No             14 countries                 22
##  7  2022     65 No             14 countries                 16
##  8  2021     64 No             94 countries                100
##  9  2022     68 No             94 countries                 29
## 10  2021     62 No             94 countries                110
## # ... with 31 more rows
```

a) Write the pseudocode that will allow you to wrangle `africa` to obtain the median number of countries guessed for each of the two priming groups:

```
## # A tibble: 2 x 2
##   priming      median_guess
##   <chr>               <dbl>
## 1 14 countries           28
## 2 94 countries           57
```

b) Write the pseudocode that will allow you to wrangle `africa` to obtain only the year, priming group, and number of countries guessed for only the first-year students (class of 2022):

```
## # A tibble: 4 x 3
##    year priming      how_many_countries
##   <int> <chr>                     <int>
## 1  2022 14 countries                 16
## 2  2022 94 countries                 29
## 3  2022 14 countries                 30
## 4  2022 14 countries                 27
```

c) Write the pseudocode that will allow you to wrangle `africa` so that the rows are reordered from the largest number of countries guessed to the smallest (note we only show the first 5 out of 41 rows in the output below):

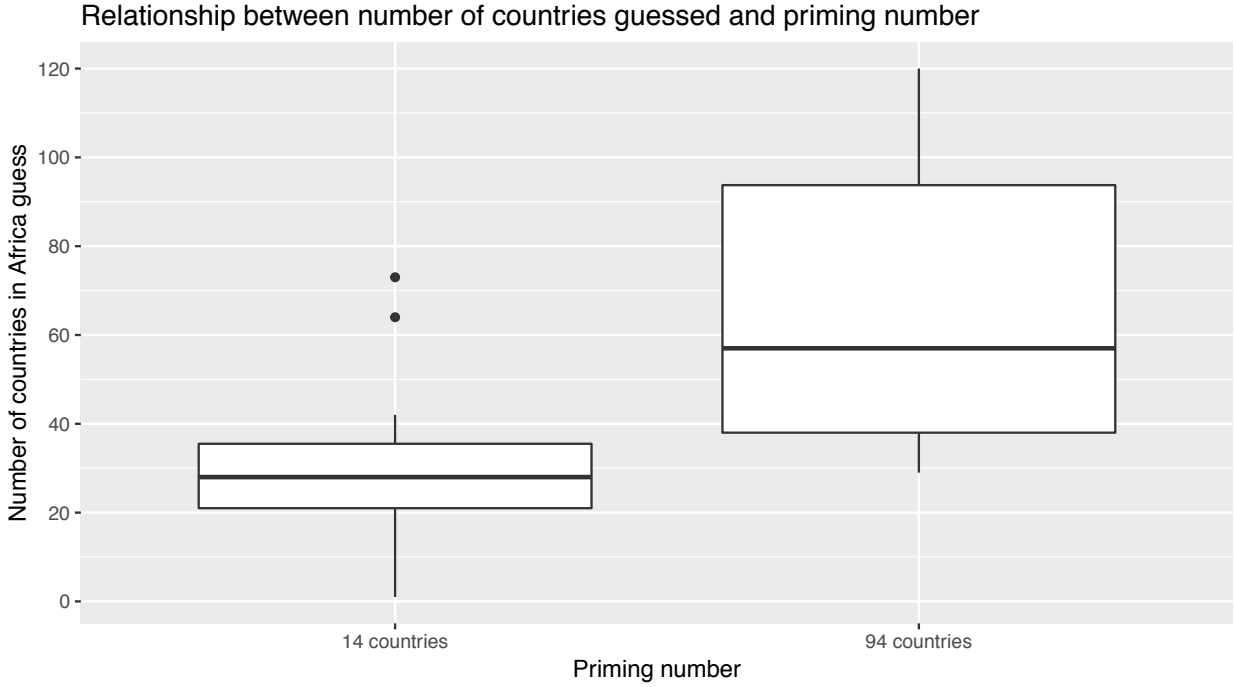
```
## # A tibble: 5 x 5
##    year height been_to_africa priming      how_many_countries
##   <int>  <int> <chr>          <chr>                     <int>
## 1  2020     67 No             94 countries                120
## 2  2021     62 No             94 countries                110
## 3  2021     64 No             94 countries                100
## 4  2021     60 No             94 countries                100
## 5  2019     68 No             94 countries                 75
```

2 Exploratory data analysis via visualizations

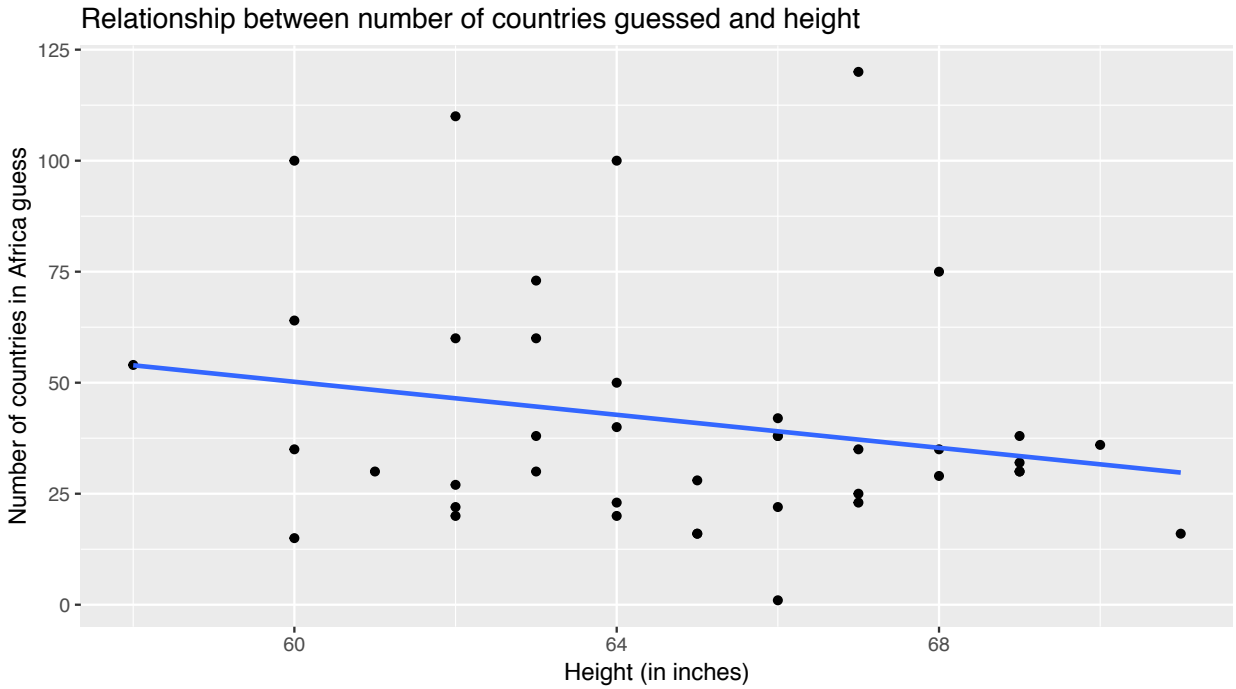
Continuing the previous `africa` question, for the remainder of this midterm let the outcome variable y be the number of countries a student guesses.

a) Name an ideal exploratory data visualization for the relationship between y and `height`.

b) We present an exploratory boxplot of the relationship between y and `priming`. It is a fact that there is more variation in responses amongst the students primed with the number 94. How is this apparent in the visualization? Compute the approximate values of a *summary statistic* we've seen in class to justify your answer.



c) The following graphic is created by the (incomplete) code snippet below.




```
ggplot(africa, aes(AAA, BBB)) +
  geom_CCC() +
  geom_DDD(method = "lm", se = FALSE) +
  labs(x = "Height (in inches)", y = "Number of countries in Africa guessed")
```

What precise code should be in place of `AAA`, `BBB`, `CCC`, and `DDD` in order to create this plot?

d) While an exploratory scatterplot of the relationship between y and `year` would be valid since `year` is numerical, why would a (vertical) boxplot with `year` on the x-axis also be acceptable *for this particular dataset*? Answer in one sentence.

3 Regression model using priming number

Continuing the previous `africa` question, we fit a regression where y is the number of countries guessed and x indicates which "priming group" a student was a part of:

```
model_countries_priming <- lm(how_many_countries ~ priming, data = africa)
get_regression_table(model_countries_priming)

## # A tibble: 2 x 7
##   term              estimate std_error statistic p_value lower_ci upper_ci
##   <chr>                <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept             29.5      4.18      7.05       0     21.1     38.0
## 2 priming94 countr~     34.7      7.16      4.84       0     20.2     49.2
```

a) What does the `intercept` term in the `estimate` column of the regression table tell us? Answer in one sentence.

b) What does the `priming94 countr~` term in the `estimate` column of the regression table tell us? Answer in one sentence.

c) Say instead of using only two priming numbers, we used three: 0, 14, and 94 countries. In other words, we assigned students to one of three priming groups. Write down what the three terms in the left-most `term` column of the above regression table would now be.

d) Say you perform data wrangling to compute the mean number of countries guessed for each of the two priming groups. What are `XXX` and `YYY` in the table below? Your answers should be numerical values. Show your work.

```
## # A tibble: 2 x 2
##   priming      mean_guess
##   <chr>        <chr>
## 1 14 countries XXX
## 2 94 countries YYY
```

e) Say we run the following code and focus only on the first two rows out of the output (out of 41), corresponding to the first two students in the `africa` dataset. What are `XXX`, `YYY`, `AAA`, and `BBB` below? Your answers should be numerical values. Show your work.

```
get_regression_points(model_countries_priming)
```

```
## # A tibble: 2 x 5
##      ID how_many_countries priming      how_many_countries_hat residual
##   <int>              <dbl> <chr>        <chr>                  <chr>
## 1     1                 36 14 countries XXX                    YYY
## 2     2                120 94 countries AAA                    BBB
```

f) Do you think the number of countries guessed by those primed by "14" differs *significantly* from the number of countries guessed by those primed with "94"? Why? You will receive full credit for merely making a good faith attempt at answering. A "right answer" is not expected as you don't have the tools to answer this question . . . yet.

4 Regression model using height

Continuing the previous `africa` question, say you run the following regression instead, using `height` instead of `priming` as the explanatory/predictor variable:

```
model_countries_height <- lm(how_many_countries ~ height, data = africa)
get_regression_table(model_countries_height)

## # A tibble: 2 x 7
##   term      estimate std_error statistic p_value lower_ci upper_ci
##   <chr>        <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept   162.       86.9       1.86   0.07    -14.1   338.
## 2 height       -1.86      1.34     -1.39   0.174    -4.57    0.854
```

a) Interpret the `intercept` term in the `estimate` column of the regression table, both mathematically and practically speaking ("practically" meaning in context of the data).

b) Give the precise interpretation of the slope for `height` in the `estimate` column of the regression table.

c) Say we run the following code and present only the first row of the output (out of 41 rows), corresponding to the first student in `africa`. What are `XXX` and `YYY`? Your answers should be numerical values. Show your work.

```
get_regression_points(model_countries_height)
```

```
## # A tibble: 1 x 5
##      ID how_many_countries height how_many_countries_hat residual
##   <int>              <dbl>  <dbl> <chr>                  <chr>
## 1     1                 36     70 XXX                    YYY
```

d) Based on the regression model above, someone predicts that someone of height 54 inches will guess 62 countries. Why might this prediction inappropriate? Base your answer only on the various output of the analysis/model so far, and not prior knowledge or hypotheses you may have about the relationship between height and knowledge of the number of countries in Africa.

e) What would it mean for the relationship between height and the number of countries guessed if the slope for `height` in the table above were 0? Answer in practical and not mathematical terms ("practical" meaning in context of the data).

f) Do you think the observed slope for `height` of `-1.86` is *significantly* different from 0? Why? You will receive full credit for merely making a good faith attempt at answering. A "right answer" is not expected as you don't have the tools to answer this question . . . yet.

1 Seattle House Prices

Recall the Seattle House Prices dataset you saw in the DataCamp course "Modeling with Data in the Tidyverse." Before we begin this question, let's perform a little data wrangling.

```
library(moderndive)
house_prices <- house_prices %>%
  mutate(
    log10_price = log10(price),
    log10_size = log10(sqft_living)
  ) %>%
  select(log10_price, log10_size, condition)
```

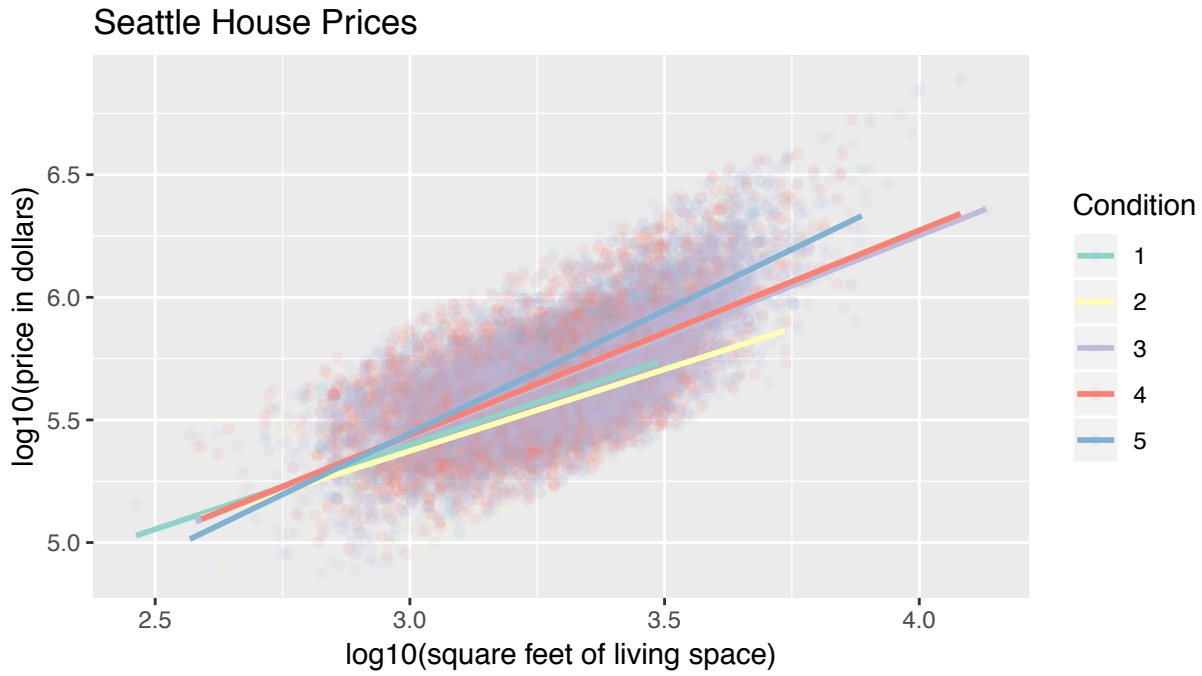
Now let's look at a random sample of 5 out of the 21,613 rows:

log10_price	log10_size	condition
6.0	3.6	4
5.8	3.3	3
5.8	3.3	4
5.2	3.2	4
5.5	3.4	3

We are interested in modeling the outcome variable y = log10 of house price in dollars as a function of two explanatory variables:

1. x_1: numerical explanatory/predictor variable log10 of the square footage of the house
2. x_2: categorical explanatory/predictor variable condition

You fit an interaction model, both graphically and using a regression model. Note the last 4 rows of the regression table got truncated; they should read `log10_size:condition2` through `log10_size:condition5`.



```
house_price_model <- lm(log10_price ~ log10_size * condition, data = house_prices)
get_regression_table(house_price_model)
```

```
## # A tibble: 10 x 7
##    term             estimate std_error statistic p_value lower_ci upper_ci
##    <chr>               <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
##  1 intercept            3.33     0.451     7.38    0        2.45     4.22
##  2 log10_size           0.69     0.148     4.65    0        0.399    0.98
##  3 condition2           0.047    0.498     0.094   0.925   -0.93     1.02
##  4 condition3          -0.367    0.452    -0.812   0.417   -1.25     0.519
##  5 condition4          -0.398    0.453    -0.879   0.38    -1.29     0.49
##  6 condition5          -0.883    0.457    -1.93    0.053   -1.78     0.013
##  7 log10_size:cond~    -0.024    0.163    -0.148   0.882   -0.344    0.295
##  8 log10_size:cond~     0.133    0.148     0.893   0.372   -0.158    0.424
##  9 log10_size:cond~     0.146    0.149     0.979   0.328   -0.146    0.437
## 10 log10_size:cond~     0.31     0.15      2.07    0.039    0.016    0.604
```

a) Why did we `log10()` transform the house price and house size in square feet variables first?

b) Using the numerical values in the above regression table, write the equation for the line for houses of condition 1.

c) Using the numerical values in the above regression table, write the equation for the line for houses of condition 5.

d) Say a house get puts on the market in Seattle and you know nothing other than its size is 1000 square feet and it is of condition 5. What is the above model's prediction of this house's sale price in dollars?

e) Say instead you ran the following regression model below. Write down what all the elements of the `term` variable would be in the resulting regression table.

```
house_price_model <- lm(log10_price ~ log10_size + condition, data = house_prices)
get_regression_table(house_price_model)
```

2 Sampling Scenarios

Consider the three scenarios below

- **Scenario 1**: You want to know the proportion of the balls in a sampling bowl of 2400 balls that are red. To this end, you mix the bowl first and use a shovel with 50 slots to pull out 50 balls. We observe that 20 of them are red.
- **Scenario 2**: We want to know the average year of minting of **all** pennies currently being used in the US. To this end, you go to Florence Bank in Downtown Northampton and ask the cashier to exchange a ten dollar bill for 1000 pennies. We observe that the average year of minting of these pennies is 2013.56
- **Scenario 3**: The instructor of SDS/MTH 220 wants to know what the effects are of priming with the numbers 14 and 94 on the number of countries Smith students guess are in Africa. To this end he conducts a priming experiment with all 38 of his students as done in class. He obtains the following fitted regression line based on the regression table output below:

$$\begin{aligned} \widehat{y} &= b_0 + b_1 \times x \\ \widehat{\text{countries}} &= b_0 + b_1 \times \mathbb{1}\,(\text{primed with 94}) \\ \widehat{\text{countries}} &= 29.5 + 34.7 \times \mathbb{1}\,(\text{primed with 94}) \end{aligned}$$

```
model_countries_priming <- lm(countries ~ priming, data = africa)
get_regression_table(model_countries_priming)

## # A tibble: 2 x 7
##   term      estimate std_error statistic p_value lower_ci upper_ci
##   <chr>        <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept     29.5      4.18      7.05       0     21.1     38.0
## 2 priming94     34.7      7.16      4.84       0     20.2     49.2
```

a) On the next page there is a table. For all cells with a question mark, fill in what those values should be.

Scenario	1	2	3
Population	$N = ?$	$N = ?$	$N = ?$
Population parameter name	?	?	Population slope
Population parameter mathematical notation	?	?	β_1
Sample size	$n = ?$	$n = ?$	$n = ?$
Point estimate name	?	?	Fitted slope
Point estimate mathematical notation	?	?	b_1
Point estimate numerical value	?	?	?

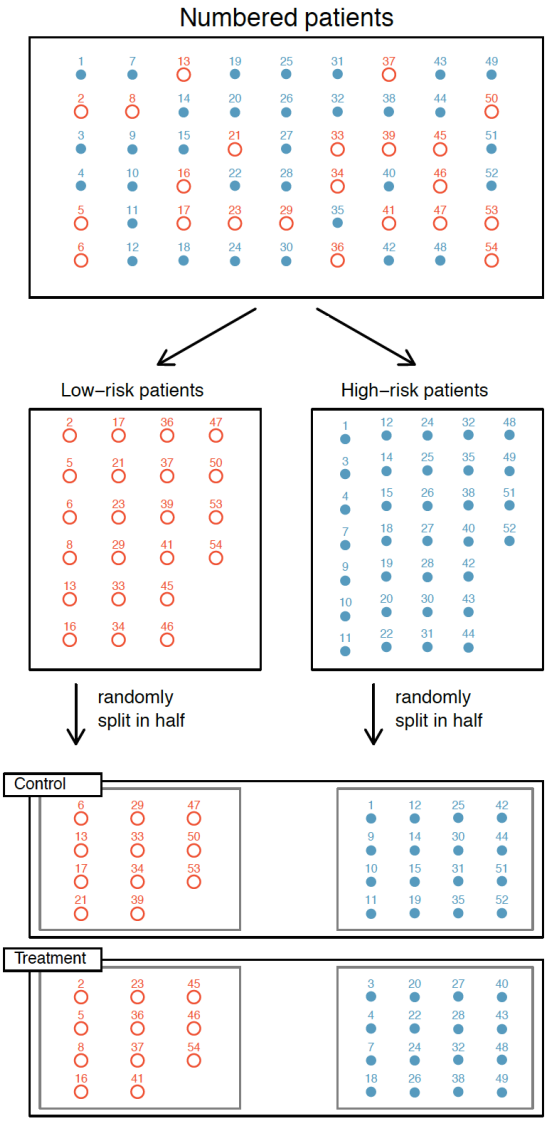
b) Is the point estimate for the population parameter in Scenario 1 a good one? Why or why not? Answer in three sentences or less.

c) Is the point estimate for the population parameter in Scenario 2 a good one? Why or why not? Answer in three sentences or less.

d) Is the point estimate for the population parameter in Scenario 3 a good one? Why or why not? Answer in three sentences or less.

3 Short Answer

a) Researchers are looking at the effect of a drug on heart attacks. They first split patients in the study into low-risk and high-risk groups, then randomly assign half the patients from each group to the control and the other half to the treatment, as shown in the figure below.

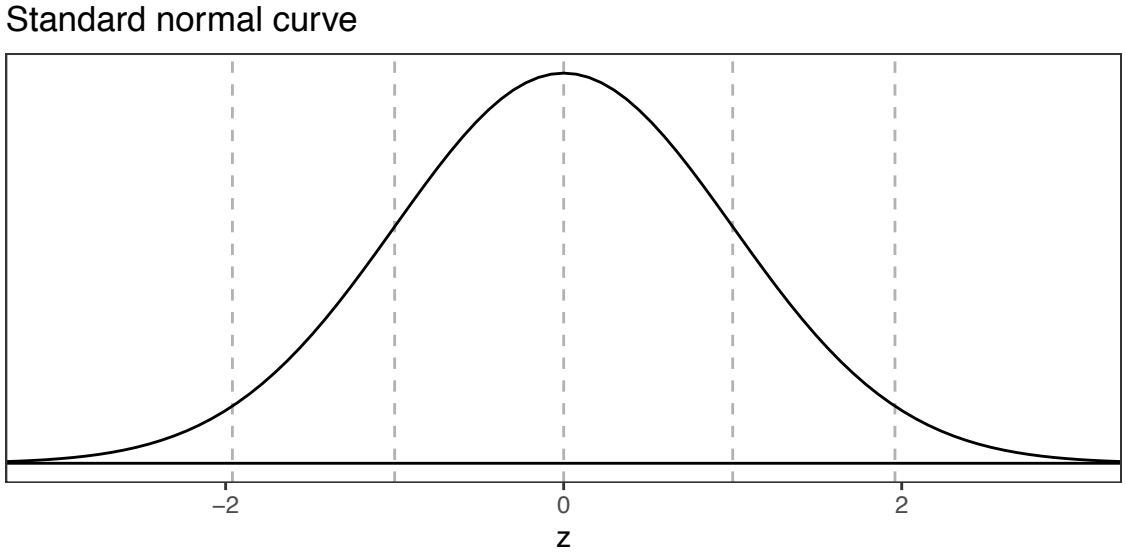


This is an example of what type of study: _____.

b) Alfred Kinsey, a sexologist in the 1940's, wanted to study the sexual behavior of American males. After interviewing a number of males, he asked them to refer other men they knew and conducted interviews with them. He repeated this process until 5500 men were interviewed. Based on an analysis of this data, he declared that "10% of all American males are exclusively homosexual." Comment on his research design and hence the validity of his conclusion using language from this course. **Answer in three sentences or less!**

c) The example from class where we studied the causal effect of shoes on the likelihood of waking up with a headache is an observational study. Why is it an observational study? **Answer in one sentence**.

d) 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%.



4 Sampling Distribution

a) Recall the virtual `bowl` consisting of 2400 balls from the `moderndive` package. Let's show the first 10 rows of the data set:

```
bowl
```

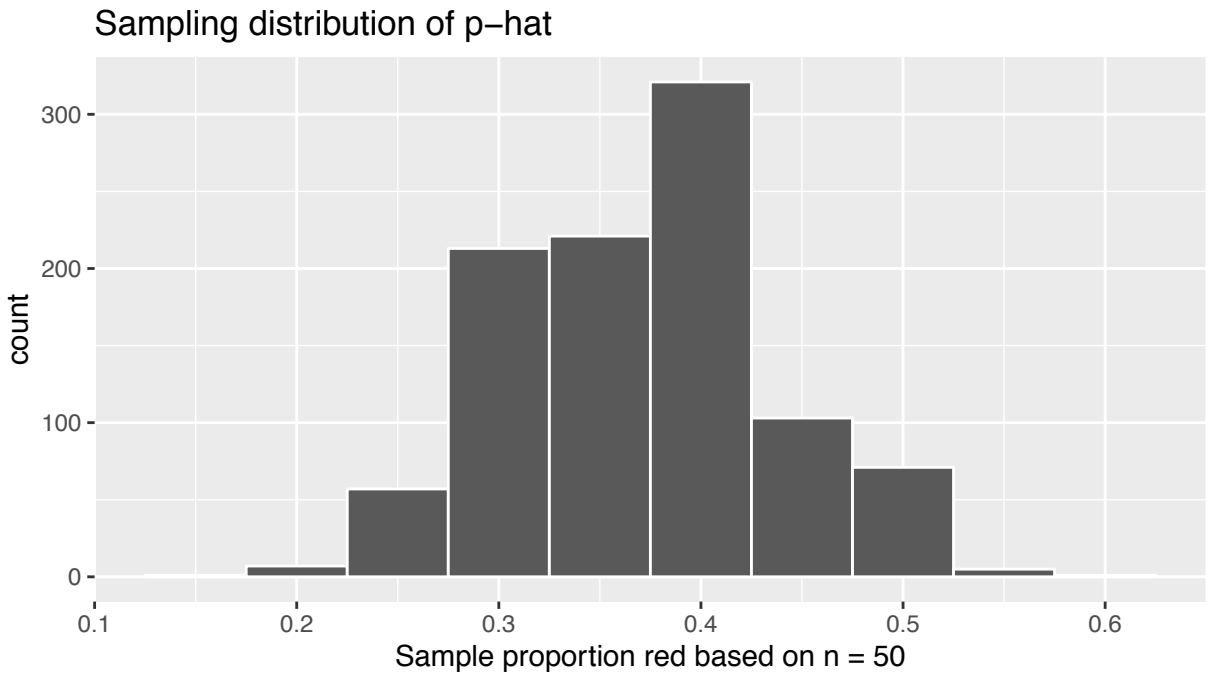
```
## # A tibble: 2,400 x 2
##    ball_ID color
##      <int> <chr>
##  1       1 white
##  2       2 white
##  3       3 white
##  4       4 red
##  5       5 white
##  6       6 white
##  7       7 red
##  8       8 white
##  9       9 red
## 10      10 white
## # ... with 2,390 more rows
```

From this virtual bowl we can draw virtual samples using the `rep_sample_n()` function. For example:

```
bowl %>%
  rep_sample_n(size = 3, reps = 2)
```

```
## # A tibble: 6 x 3
## # Groups:   replicate [2]
##   replicate ball_ID color
##       <int>   <int> <chr>
## 1         1    1144 red
## 2         1    1515 red
## 3         1     123 white
## 4         2    1874 white
## 5         2    1287 white
## 6         2     833 red
```

Recall that we ran a simulation creating the sampling distribution of $\widehat{p}$ based on 1000 samples of size $n = 50$ drawn using the virtual shovel:



Write out the pseudocode that will produce the visualization of the above sampling distribution. Feel free to write in actual code if you like. Hint: your pseudocode should start with `bowl` and use the `rep_sample_n()` function from earlier.

b) What will happen to the above sampling distribution if we used a virtual shovel with $n = 100$ slots?

c) What is the standard deviation of the above sampling distribution called?