STA035B Midterm 2, Winter 2024

Name ____

Student ID ____

Section ____

problem	points
1a	
1b	
2a	
2b	
3a	
3b	
3c	
4.1	
4.2	
5a	
5b	
total	

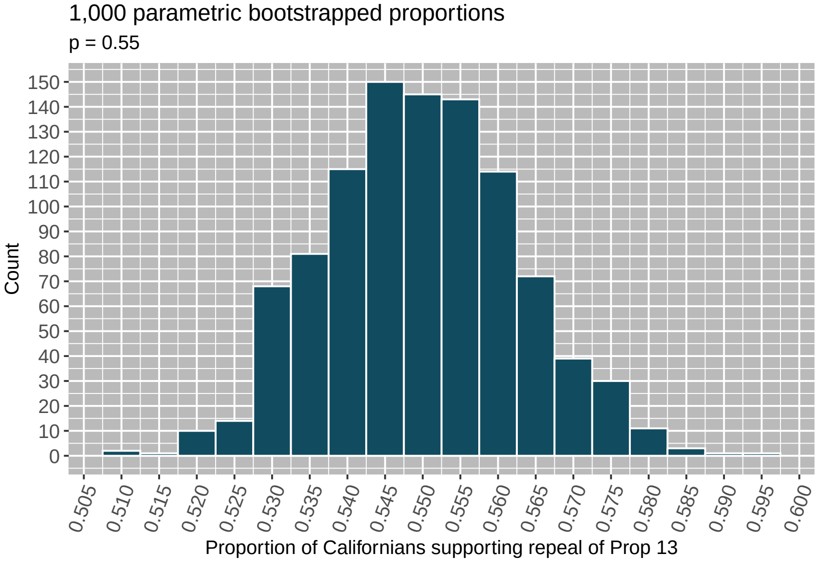
Problem 1

Suppose we are running a poll on whether or not a majority of Californians would like to repeal Proposition 13. The pollster takes a random sample of 1,563 people and reports that 58% of the people supported repeal. Suppose that in order for the repeal to go forward, 55% or more of voters must approve of the repeal.

(a) What are the null and alternative hypotheses for evaluating whether these data provide convincing evidence that, if voted on, Proposition 13 would be repealed in the US? (5 points)

- Null hypothesis H_0: Proportion of people supporting repeal $p \leq 0.55$
- Alternative hypothesis H_A: Proportion of people supporting repeal $p > 0.55$

(b) A parametric bootstrap simulation with 1,000 bootstrap samples was run and the resulting null distribution is displayed in the histogram below. Estimate the p-value using this distribution and conclude the hypothesis test in the context of the problem. (10 points)



The p-value corresponds to the **proportion** of observations to the right of 0.58 (not to the left, since the alternative is only that $p > 0.55$). We therefore want to estimate the number of bootstraps with proportions >= 0.58, then normalize by the total number of bootstraps (1000).

We can visually see there are about 10-15 bootstrap samples with ≥ 0.58. Thus, we estimate p-value of between 10/1000 and 15/1000, or 0.01 to 0.015.

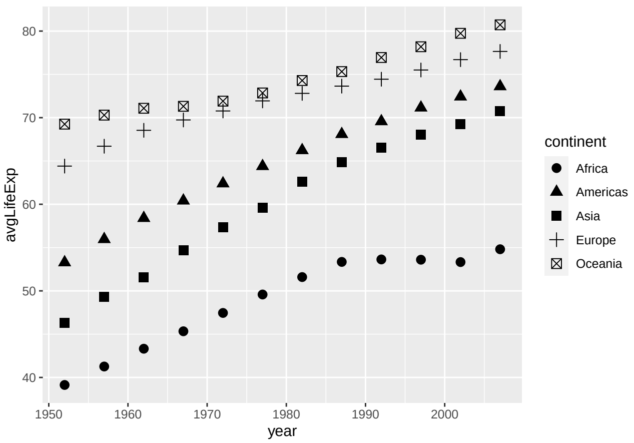
Using a default significance level of 0.05, we can therefore reject the null hypothesis at level 0.05. We therefore have significant evidence that if voted on, Prop 13 would be repealed.

Problem 2

Consider the dataset `gapminder`, which has observations of the life expectancy, population, and gdp per capita (in dollars, $) for years between 1952 and 2007 for different countries. A sample of 5 rows from this table is given as follows.

```
# A tibble: 5 x 6
  country      continent  year lifeExp       pop gdpPercap
  <fct>        <fct>     <int>   <dbl>     <int>     <dbl>
1 Denmark      Europe     1982    74.6   5117810    21688.
2 Egypt        Africa     1982    56.0  45681811     3504.
3 Brazil       Americas   2002    71.0 179914212     8131.
4 Finland      Europe     1997    77.1   5134406    23724.
5 Burkina Faso Africa     1962    37.8   4919632      723.
```

For each of the following, determine whether or not the code correctly computes the average life expectancy per year for each continent from 1952-2007 and makes the plot appearing below. If the code does not correctly do the computation and plot, explain what is wrong with the code.

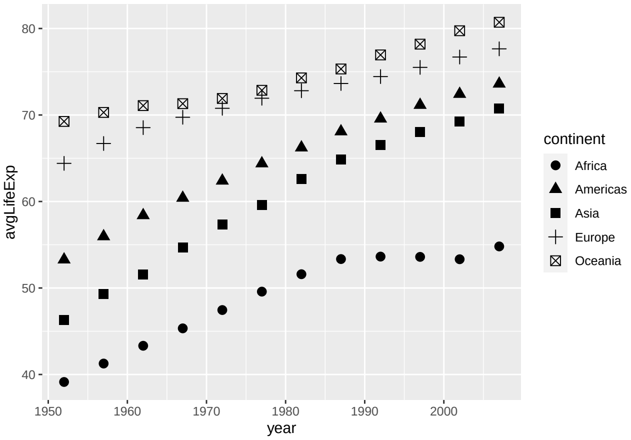


(a) 3 points

```
gapminder %>%
  group_by(year) %>%
  summarize(avgLifeExp = mean(lifeExp, na.rm=TRUE)) %>%
  ggplot(aes(x = year, y = avgLifeExp)) +
  geom_point(aes(shape = continent))
```

The code does not correctly do the computation. The issue is the `group_by()` is only grouping by year, but we need to group by both year and continent, since we want to compute average per year AND continent. The rest of the code is correct.

(The text and plot from the previous page are copied below for your convenience.) Determine whether or not the code correctly computes the average life expectancy per year for each continent from 1952-2007 and makes the plot appearing below. If the code does not correctly do the computation and plot, explain what is wrong with the code.



(b) 3 points

```r
gapminder %>%
  group_by(continent, year) %>%
  summarize(avgLifeExp = mean(lifeExp, na.rm=TRUE)) %>%
  ggplot(aes(x = year, y = avgLifeExp)) +
  geom_point()
```

This code does not produce the graph because it does not produce shapes according to the continent. We need to introduce this as an aesthetic into geom_point, or into the ggplot itself

Problem 3

Consider again the gapminder dataset from the previous problem.

Part (a), 10 points Suppose we wanted to produce a linear model which uses the gapminder data to try to predict the life expectancy of a country by its GDP per capita in 2007,

```r
lm(lifeExp ~ gdpPercap,
   data = gapminder %>% filter(year == 2007))
```

```

Call:
lm(formula = lifeExp ~ gdpPercap, data = gapminder %>% filter(year ==
    2007))

Coefficients:
(Intercept)    gdpPercap
  5.957e+01    6.371e-04
```

Describe the linear model provided by the above code. Write the resulting formula for predicting life expectancy using the GDP per capita, and interpret each of the quantities in the formula.

The linear model provides an estimate of the life expectancy as a function of GDP per capita. The formula described is

$$\text{LifeExp} = 59.57 + 0.0006371 * \text{gdpPercap}$$

The intercept represents the expected life expetancy for a country where gdp per capita is zero, which is 59.57 years. The slope of 0.0006371 = 6.371e-04 represents the amount of increase in life expectancy for every single unit (dollar) increase in gdp per capita.

Part (b), 3 points Suppose we consider a country whose GDP per capita is $10,000. According to the linear model above, what is the predicted life expectancy?

Substituting into the formula, we get

$$\text{LifeExp} = 59.57 + 6.37 \cdot 10^{-4} \cdot (10^4) = 65.941 \approx 66$$

We expect 66 years.

Part (c), 5 points The variable “continent” in `gapminder` is a factor with 5 levels: Africa, Americas, Asia, Europe, and Oceania. Suppose you fit the following linear model to predict life expectancy by continent in the year 2007,

```r
lm(lifeExp ~ continent,
   data = gapminder %>% filter(year == 2007))
```

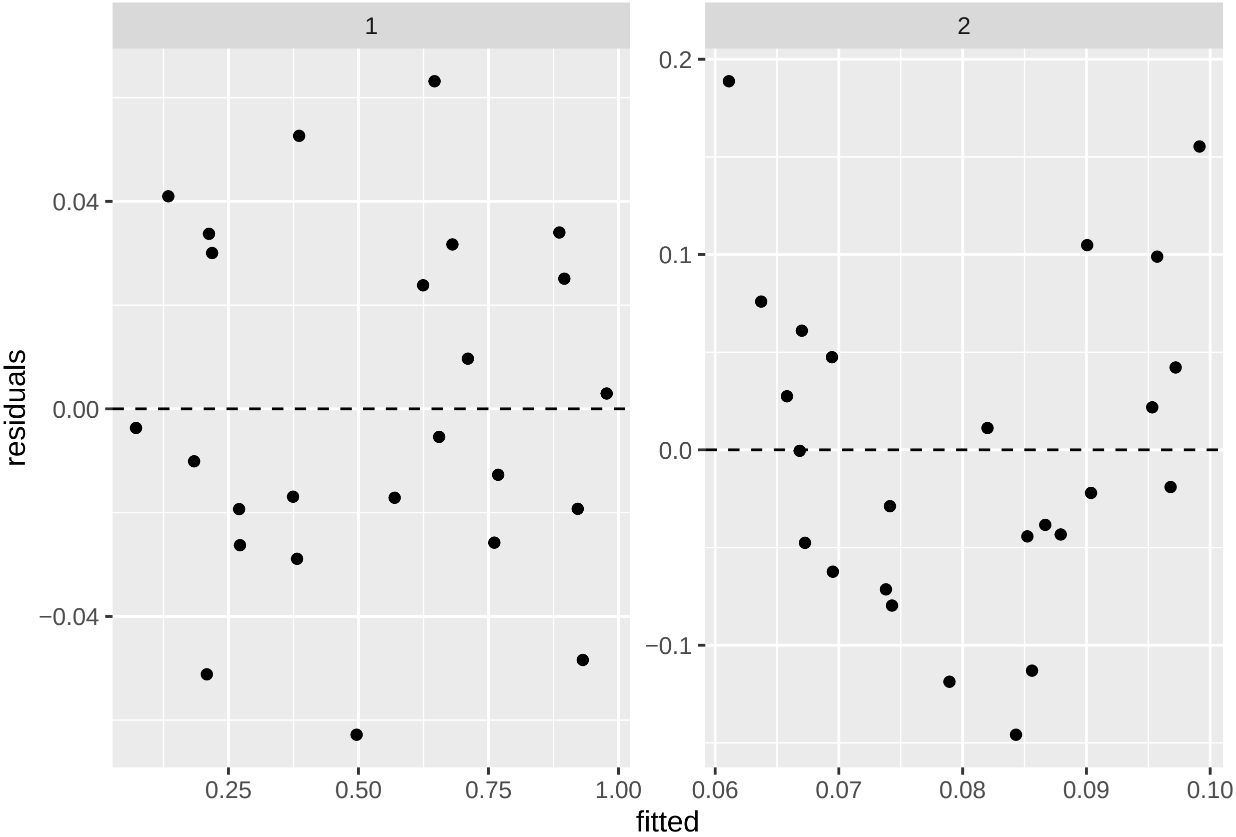
How many independent variables are in the resulting linear model? Circle your answer and provide an explanation.

(A) 1
(B) 2
(C) 3 **(D) 4**
(D) 5

The answer is D, 4 independent variables. This is because when we have a factor/categorical variable with N levels, we introduce $N - 1$ independent random variables. When all of these random variables are 0, we get the value that results from when the category is in the N-th level.

Problem 4

Consider the following residual plots.



For each of the residual plots above, describe which aspects of the plot (if any) indicates that a linear model is appropriate for modeling the data, and which aspects (if any) seem concerning for using a linear model.

Plot 1, 3 points The residual plot appears to show that a linear model is appropriate for the data. This is because there is no structure in the residuals, with relatively random positive and negative values, independent of the fitted values. There are no concerning things in this residual plot.

Plot 2, 3 points The residual plot has a clear problem: there is a significant structure to the residuals. When the fitted values are small or large, they are positive, while when the fitted values are in the middle, they are negative. There are no extreme outliers which is positive.

Problem 5

Suppose we know that the scores on the midterm exam approximately followed a normal distribution, with an average score of 83 and a standard deviation of 5.

Part (a), 3 points Which of the following correctly computes the score corresponding to the 99th percentile?

(A) `pnorm(0.99, mean = 83, sd = 5)`

(B) `qnorm(0.99, mean = 83, sd = 5)`

(C) `qnorm(0.01, mean = 83, sd = 5)`
(D) `pnorm(0.01, mean = 83, sd = 5)`

Part (b), 3 points Suppose Sal gets a 70 on the exam. Approximately what percent of students scored above Sal on the exam?

(A) Between 50% and 84%
(B) Between 84% and 95%

(C) Between 97% and 99.8%

(D) Between 99.8% and 100%