Statistics Project – Math 141

Dashiell Ward, Gavin Rimmer, Bailee Brunsmann, Harrison Nicholls

CIA Factbook

Introduction

Hypothesis

There is a correlation between the proportion of the population using the internet and life expectancy at birth within a given nation.

$H_0 : B_1 = 0$

$H_A : B_1 \neq 0$

Approach, Declaration of Goals:

There is potentially a correlation between the variables. The average life expectancy in a country could influence internet usage and/or internet usage in a country could influence the average life expectancy.

Through our analysis, our goals are to observe how internet usage and life expectancy affect each other in different nations. Further exploration of this correlation, specifically in countries in Asia, is provided in the first background research article listed. Both background research articles provide insight about other factors that could be influencing or are influenced by the variables we chose to focus on. Economic development appears to have an important connection to technological advancements, which impacts internet accessibility and further, internet usage. Additional research related to this subject may be found below.

Lee, Cheng-Wen, The Relationship between Internet Environment and Life Expectancy in Asia

Article explores this question specifically in Asia, with relevant take-away in emphasising the disparity between countries with advanced telecommunication services and those without in regards to their general economic development in a globalized market.

Alzaid, Ahmed, Musleh Alsulami, Komal Komal, Adel-Maraghi, Examining the Relationship between the Internet and Life Expectancy

Article explores this question Globally, finding that the economic development of a country is greatly bolstered by internet development, and has both direct and indirect impacts on the average life expectancy of its citizens.

Data Exploration

Dataset:

Our dataset was Details on Countries, and it was sourced from the CIA factbook. A summary of variables may be found below.

Variables:

Country - Categorical - Nominal Countries recognized by the CIA.

Area - Numerical - Continuous Land in Square km

Infant Mortality Rate - Numerical - Continuous Infant mortality rate compares the number of deaths of infants under one year old in a given year per 1,000 live births in the same year. This rate is often used as an indicator of the level of health in a country.

Population - Numerical - Discrete Population compares estimates from the US Bureau of the Census based on statistics from population censuses, vital statistics registration systems, or sample surveys pertaining to the recent past and on assumptions about future trends.

Population growth rate - Numerical - Continuous Population growth rate compares the average annual percent change in populations, resulting from a surplus (or deficit) of births over deaths and the balance of migrants entering and leaving a country. The rate may be positive or negative.

Birth Rate - Numerical - Continuous Birth rate compares the average annual number of births during a year per 1,000 persons in the population at midyear; also known as crude birth rate.

Death rate - Numerical - Continuous Death rate compares the average annual number of deaths during a year per 1,000 population at midyear; also known as crude death rate.

Net migration rate - Numerical - Continuous Net Migration rate compares the difference between the number of persons entering and leaving a country during the year per 1,000 persons (based on midyear population).

Maternal mortality rate – Numerical - Continuous The Maternal mortality rate (MMR) is the annual number of female deaths per 100,000 live births from any cause related to or aggravated by pregnancy or its management (excluding accidental or incidental causes).

Life expectancy at birth – Numerical - Discrete Life expectancy at birth compares the average number of years to be lived by a group of people born in the same year, if mortality at each age remains constant in the future. Life expectancy at birth is also a measure of overall quality of life in a country and summarizes the mortality at all ages.

Internet users – Numerical - Discrete Internet users compares the number of users within a country that access the Internet. Statistics vary from country to country and may include users who access the Internet at least several times a week to those who access it only once within a period of several months.

> *Note: We also used the "countrycode" package in R to add the categorical variable of continent to our data.*

Methods

In performing an observational study, we are limited in the viability of performing randomization tests on our data. We can find correlations in our plots using R^2 and find the differences in means between different categorical groups, in our case Continents seems apt for a geopolitical analysis of the data. However, given that our sample population is itself the true population of all countries, there are no methods by which we can fabricate more samples via bootstrapping.

We noticed that when graphing life expectancy against internet usage while coloring the points by continent, almost the entire lower section of outliers was comprised of countries in Africa. We therefore theorized that internet usage only begins to correlate linearly with life expectancy at some threshold value (~65 years), which hardly any African countries meet. Based on this, we decided to see how the data would look with Africa excluded. Figures 7 and 8 reflect the results of this exploration.

Figure Six, Figure Seven, and Figure Eight: Here we have generated linear models based on Figure 1, and computed R^2 of each linear model within the graphs. The greatest R^2 value, and thereby the best-fit linear model, is calculated to be that of Figure Six, the graph which includes all countries, rather than Figure Seven or Figure Eight which exclude countries in the continent of Africa.

To further explore our findings, we plan to make plots of the residuals and include lines of best fit. The residuals will give us insight into the biases in our scatterplots, and allow us to determine how appropriate

linear models are for comparing internet usage and life expectancy. Although excluding Africa yields a worse R^2 value, it may be that the residuals form a distribution closer to normal, meaning that it is more suited to a linear model than the alternative.

Figure 1

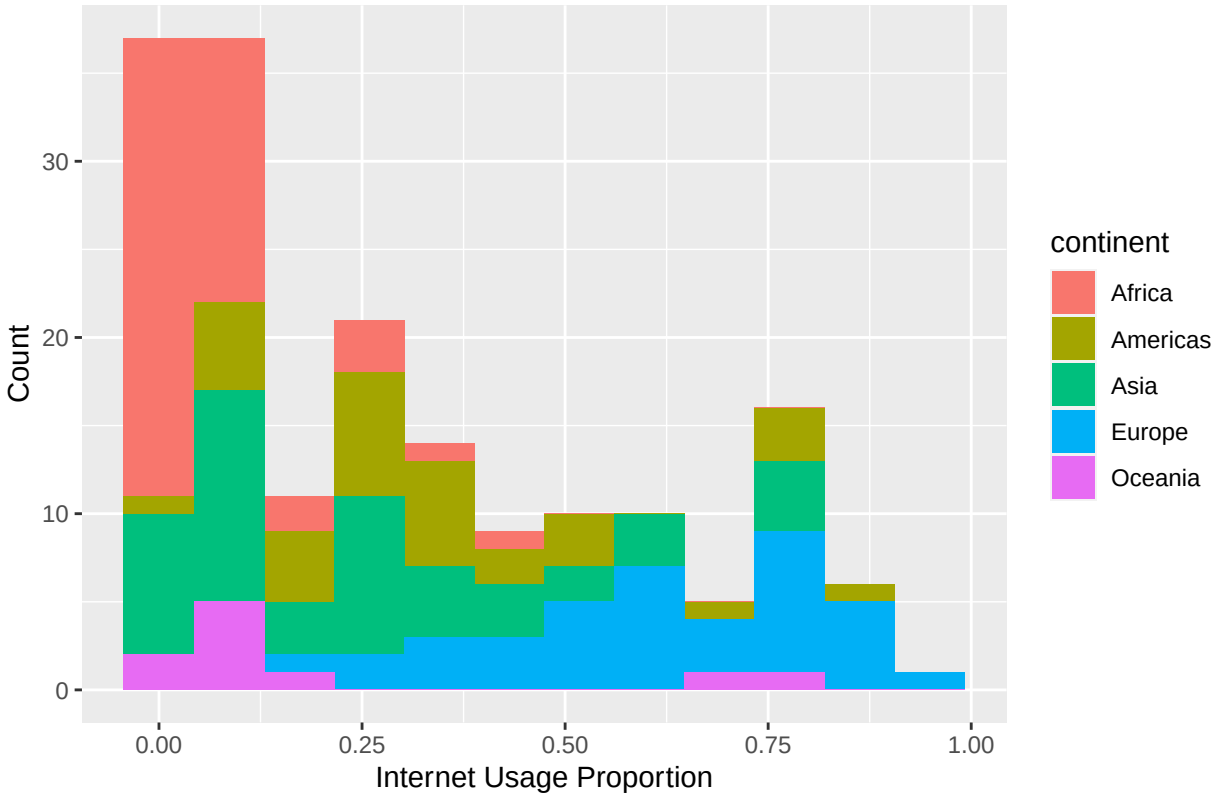

Figure 2

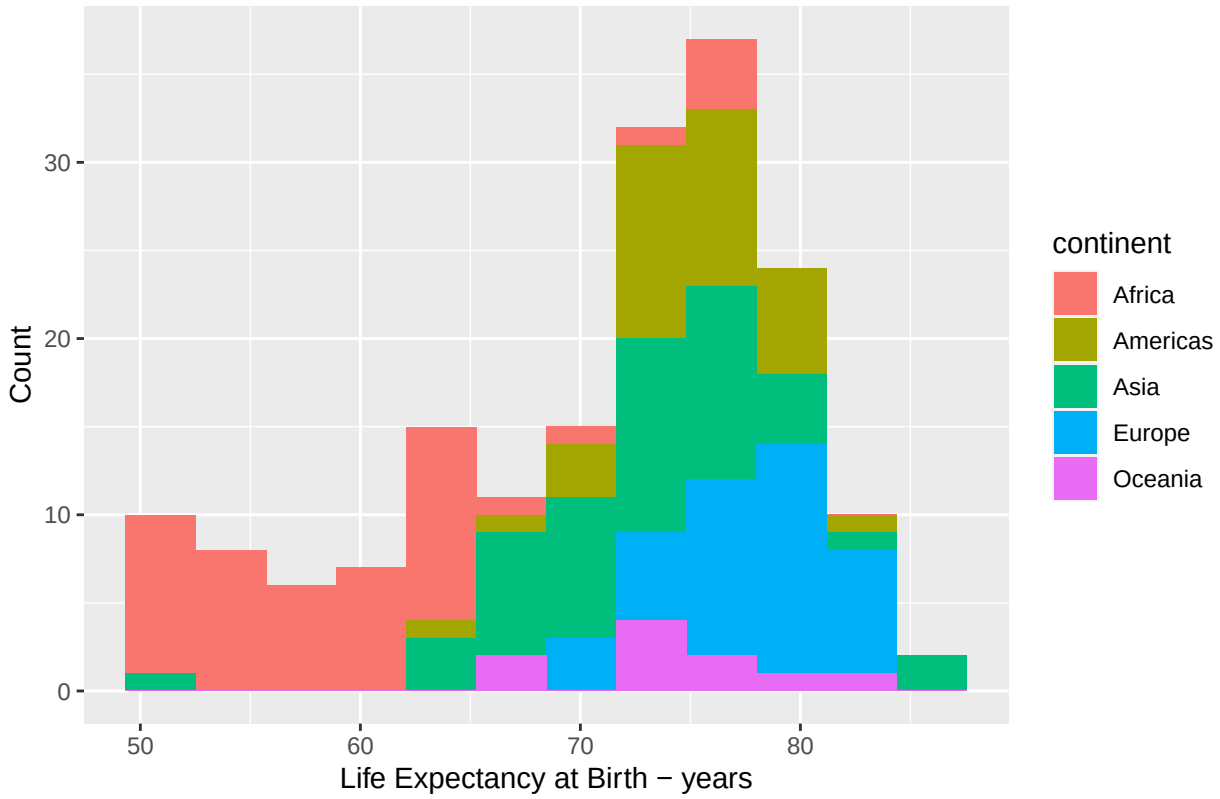
Count
30
20
10
0
50
60
70
80
Life Expectancy at Birth − years
continent
Africa
Americas
Asia
Europe
Oceania

Figure 3

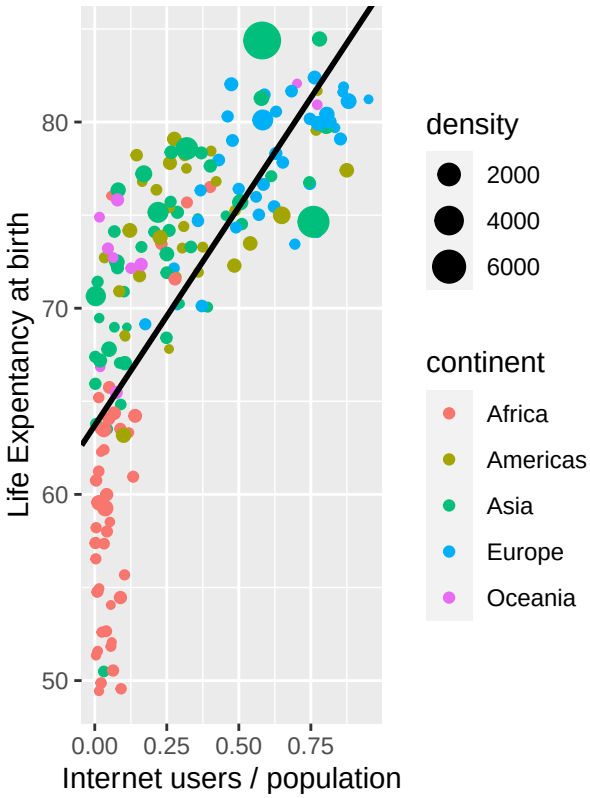

Figure 4

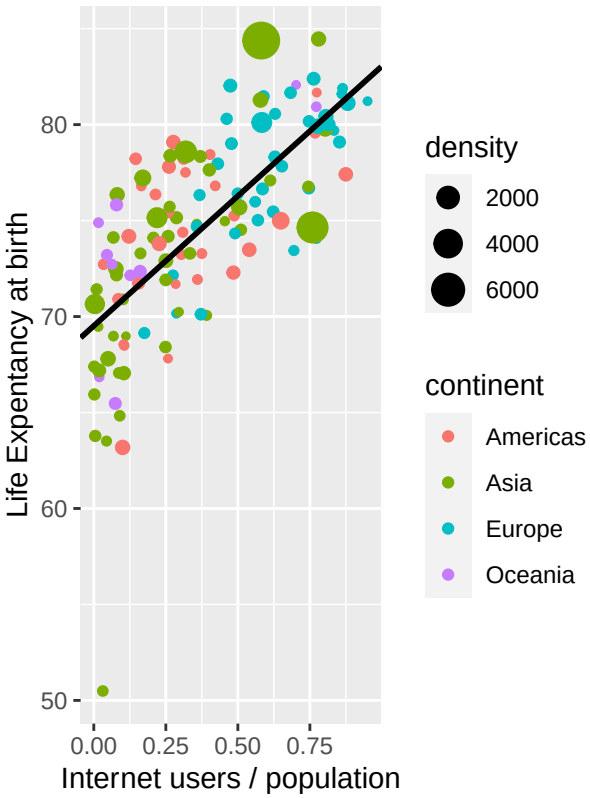

Figure 5

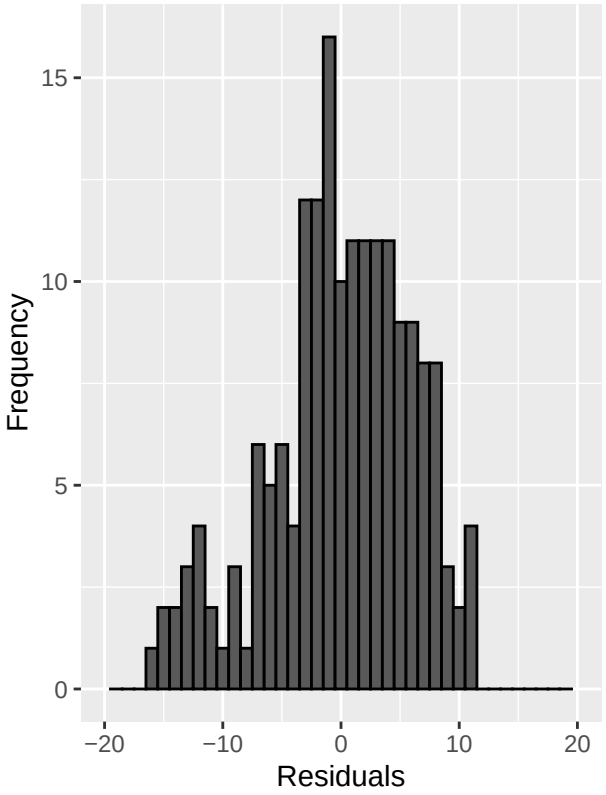

Figure 6

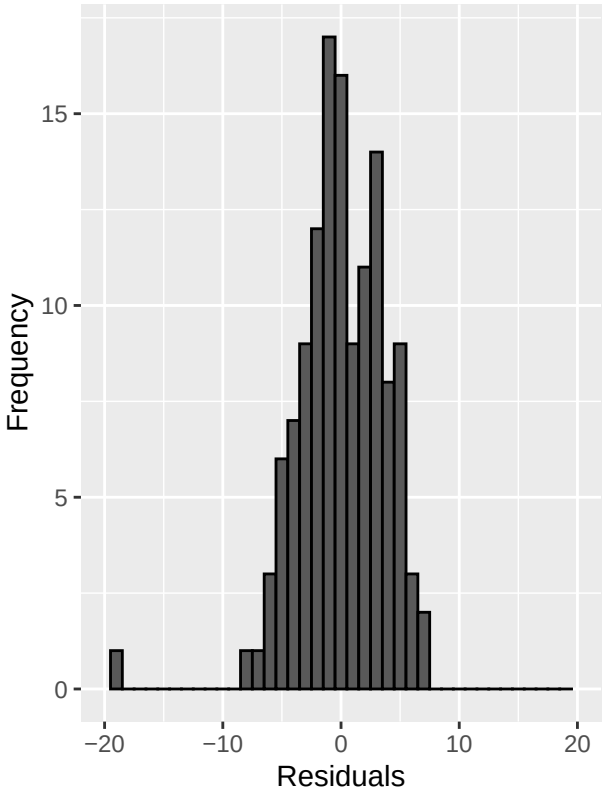

Figure 7

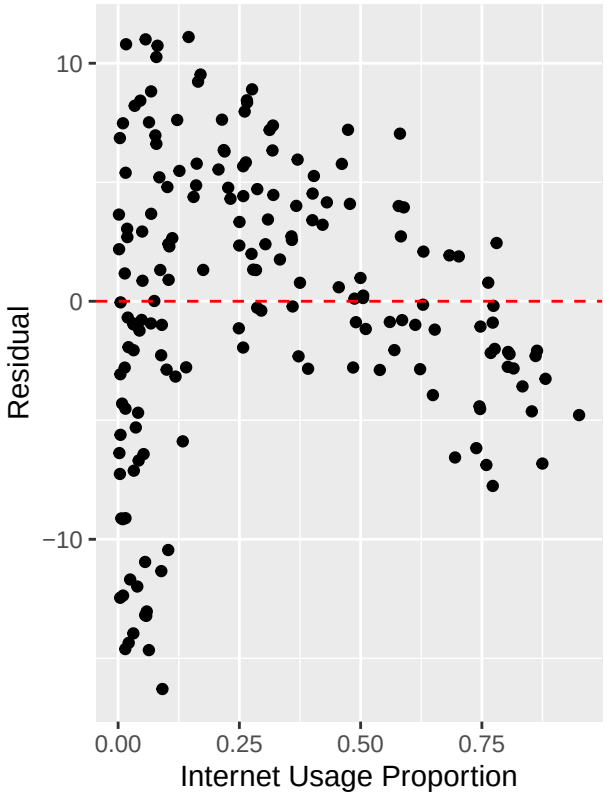

Figure 8

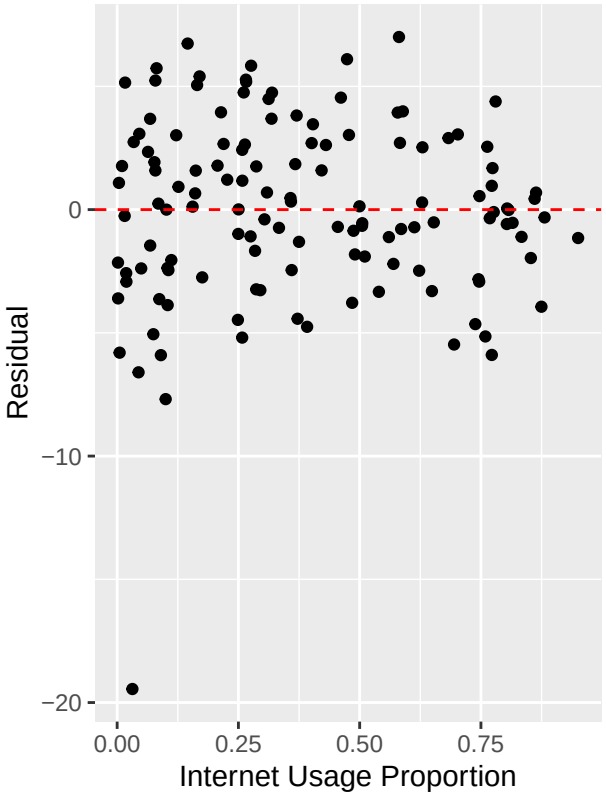

Figure 9

```
##
## Call:
## lm(formula = life_exp_at_birth ~ internet_usage_proportion, data = cia)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.2903  -2.8736   0.1377   4.4186  11.1027
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 63.7069     0.6732   94.63   <2e-16 ***
## internet_usage_proportion  23.4594     1.6731   14.02   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.066 on 175 degrees of freedom
## Multiple R-squared:  0.5291, Adjusted R-squared:  0.5264
## F-statistic: 196.6 on 1 and 175 DF,  p-value: < 2.2e-16
```

Figure 10

```
##
## Call:
## lm(formula = life_exp_at_birth ~ internet_usage_proportion, data = noafrica)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -19.4508  -2.3693   0.0082   2.6411   7.0030
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 69.5157     0.5602  124.08   <2e-16 ***
## internet_usage_proportion  13.5291     1.2007   11.27   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.678 on 127 degrees of freedom
## Multiple R-squared:  0.4999, Adjusted R-squared:  0.496
## F-statistic:   127 on 1 and 127 DF,  p-value: < 2.2e-16
```

Figure 11

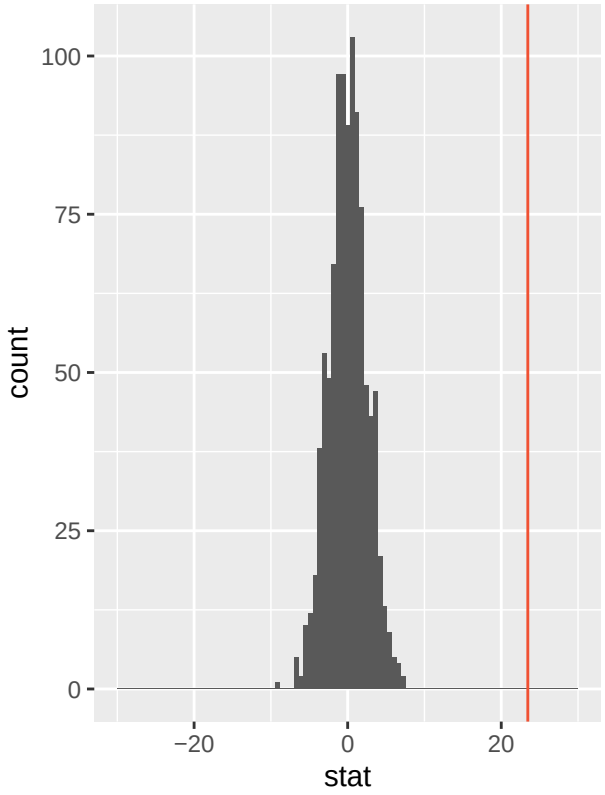

Figure 12

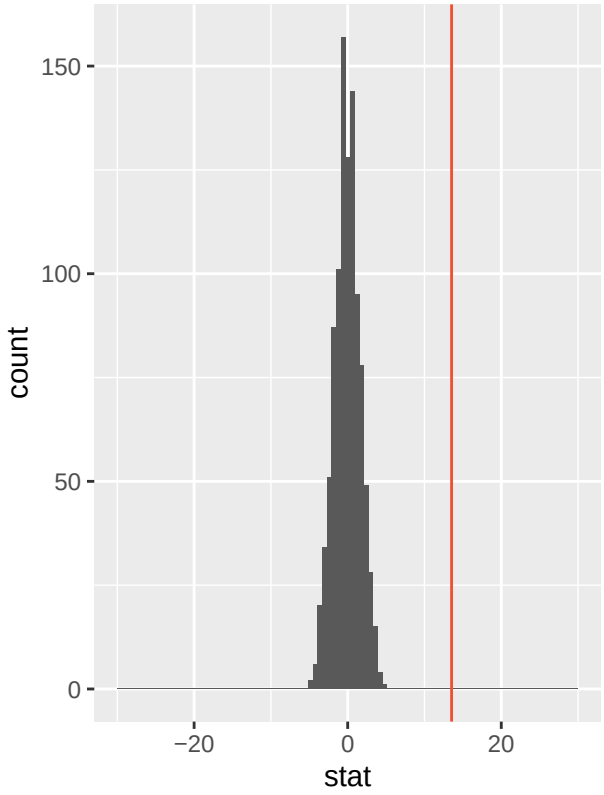

Figure 13

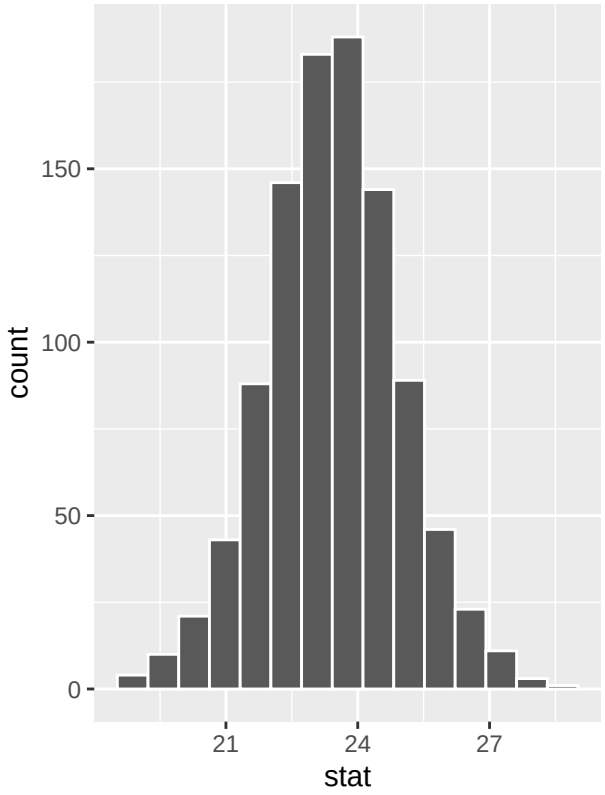

Figure 14

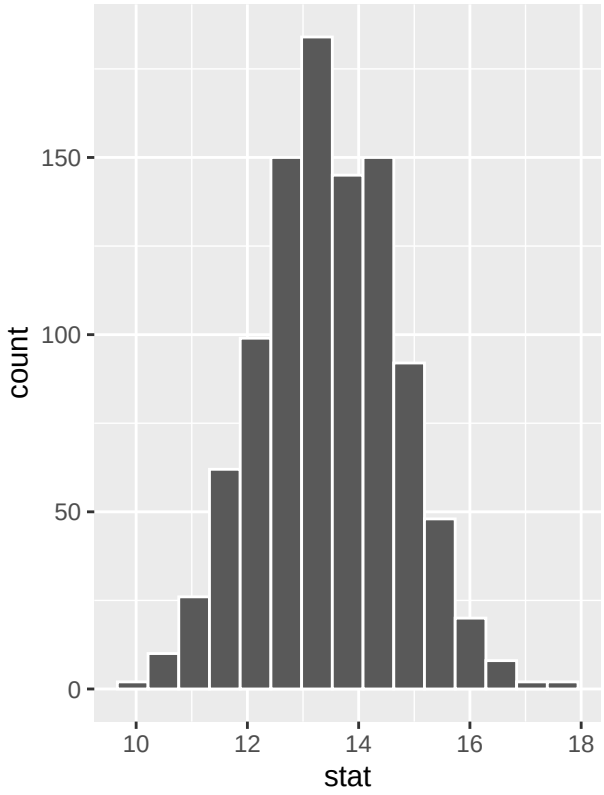

Figure 15

```
##     2.5%    97.5%
## 20.39227 26.50567
```

Figure 16

```
##     2.5%    97.5%
## 11.10236 15.85847
```