Statistical Foundations

Prof. Eric A. Suess

January 25, 2021

Statistical Foundations

The authors of our book make many important observations about Data Science at the beginning of Chapter 7 (2e Chapter 9).

- "The objective of Data Science is to extract meaning from data."
- "Visualization is good for seeing patterns in noisy data."
- "It is important to be able to see when the patterns we see are strong enough that they are not mere accidents."
- **"Statistical methods quantify patterns and their strengths."**
- "Some people think that"big data" has made statistics obsolete. The argument is that with lots of data, the data can speak clearly for themselves. This is wrong, as we shall see."
- "This chapter will illustrate a Data Science *workflow*."

Samples and Populations

- Recall the Law of Large Numbers
- Recall the Central Limit Theorem

LLN

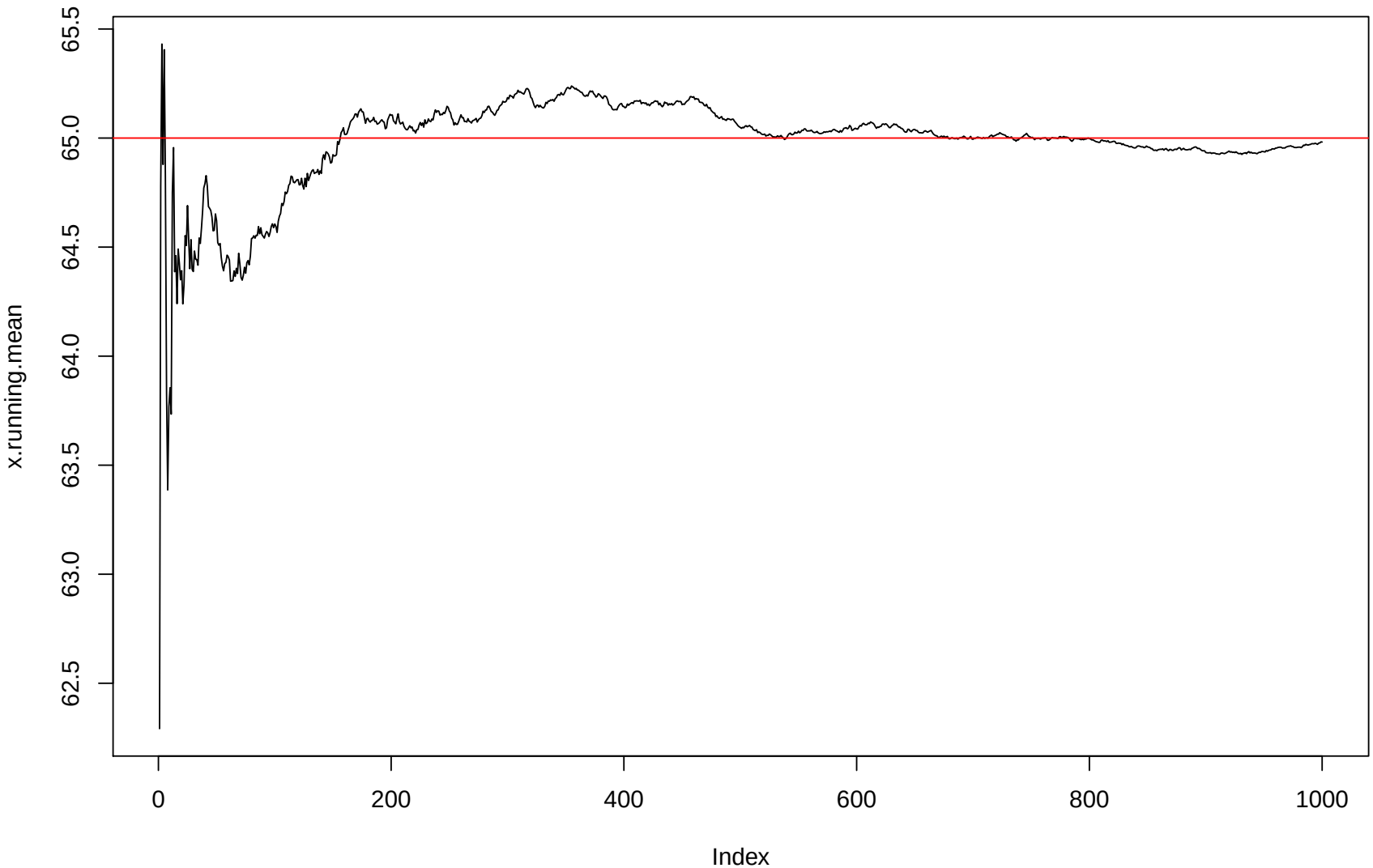
Sample mean $\bar{x}$ converges to the population mean μ.

Simulation assuming the populations parameters are known.

```
mu <- 65; sigma <- 3;  # population parameters assumed
B <- 1000
x.sample <- rnorm(B, mu, sigma)
x.running.mean <- numeric(B)
for(i in 1: B){
  x.running.mean[i] <- sum(x.sample[1:i])/i
}
```

LLN

```
plot(x.running.mean, type="l")
abline(h=mu, col="red")
```



CLT

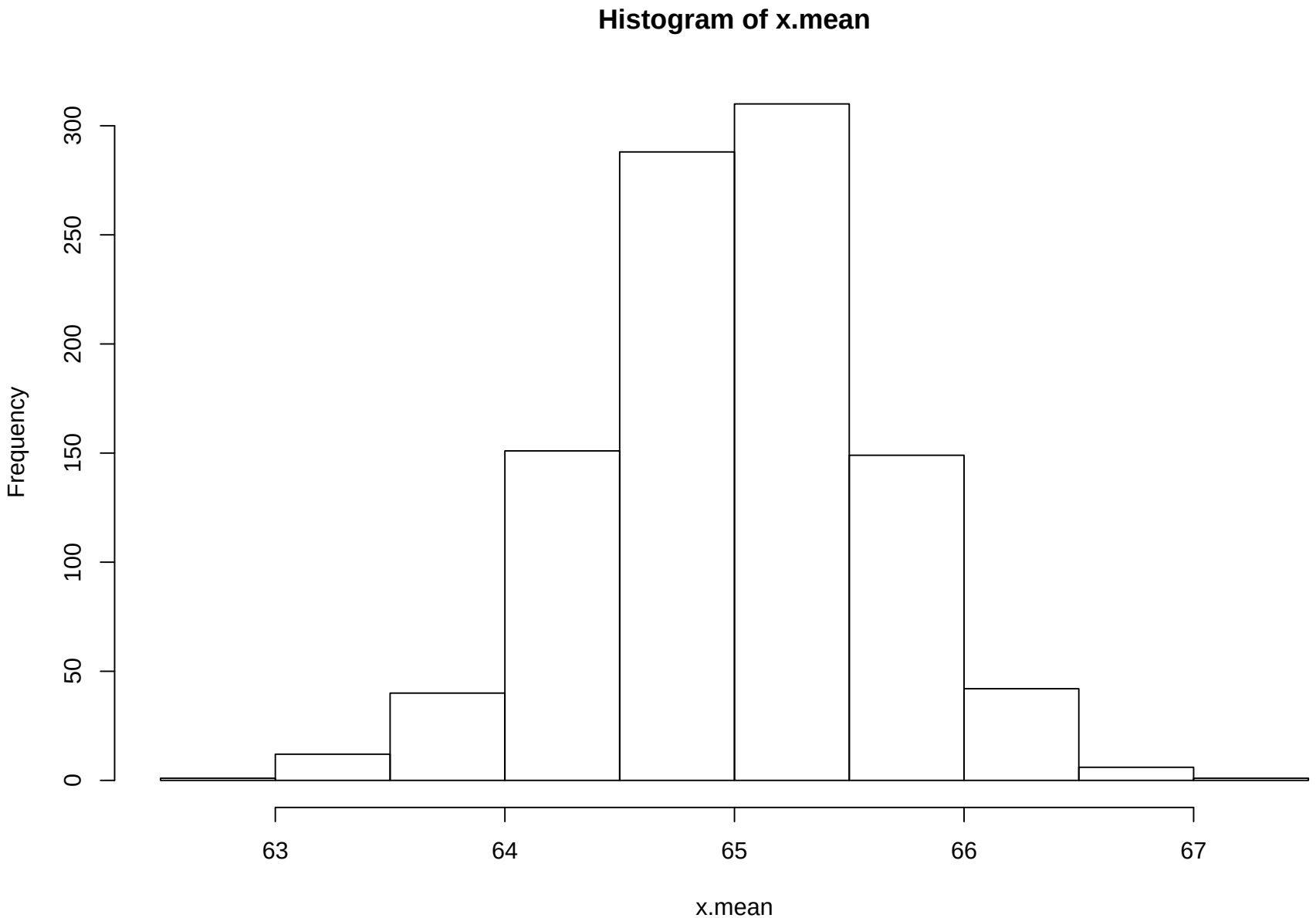
The *sampling distribution* of $\bar{x}$ is approximately $N(\mu, \frac{\sigma^2}{n})$.

Simulation assuming the populations parameters are known.

```
mu <- 65; sigma <- 3; n <- 25
B <- 1000
x.mean <- numeric(B)
for(i in 1: B){
  x.mean[i] <- mean(rnorm(n, mu, sigma))
}
```

CLT

```
hist(x.mean)
```



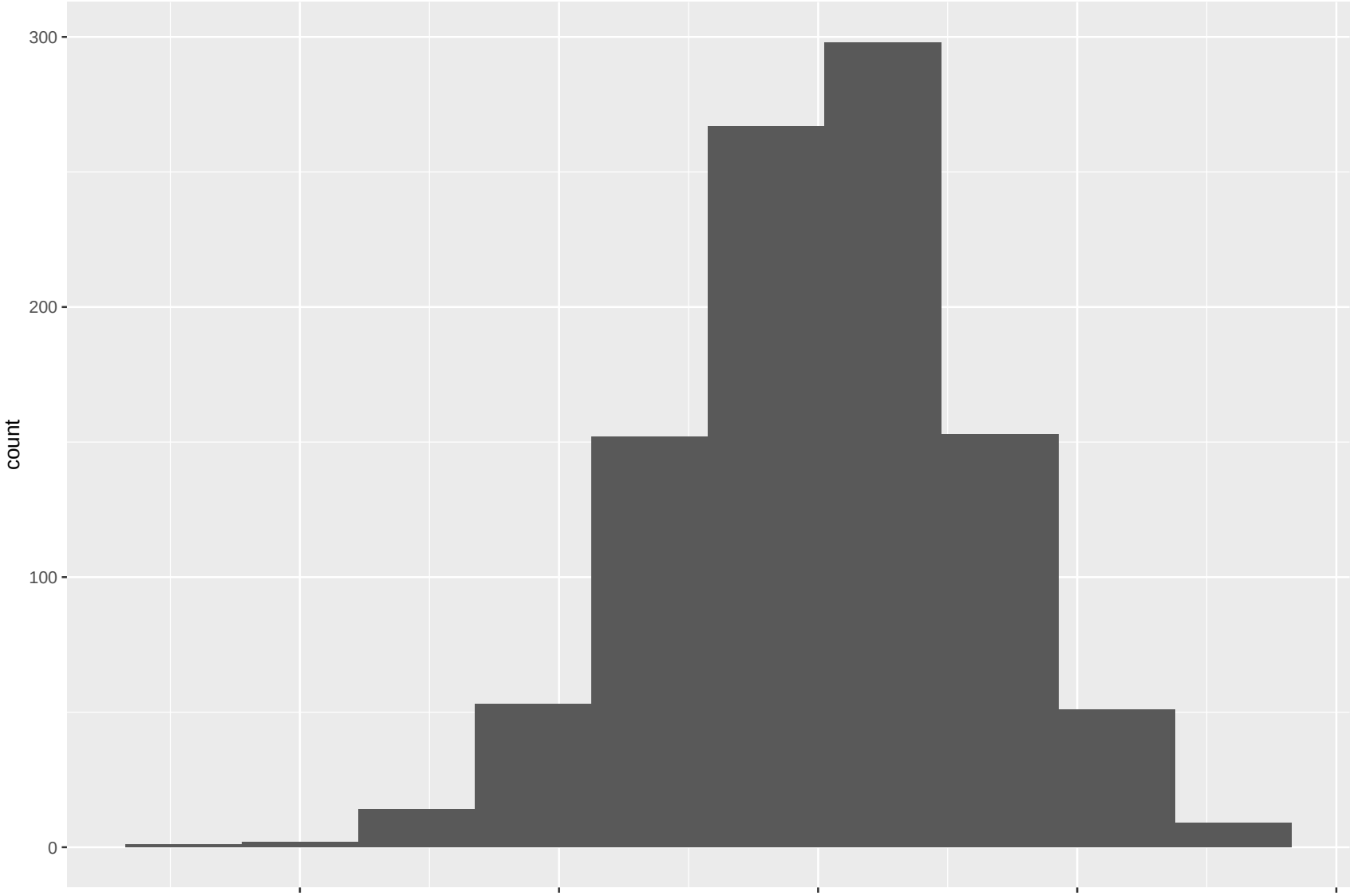
CLT in the Tidyverse

Simulation assuming the populations parameters are known.

```
library(pacman)
p_load(mosaic, tidyverse)
Trials_n <- do(1000) * mean( rnorm(n, mu, sigma) )
```

CLT in the Tidyverse

```
Trials_n %>% ggplot(aes(mean)) +
  geom_histogram(bins=10)
```



ModernDive Chapter 7 Sampling

For a ModernDive into sampling and the Central Limit Theorem try the code in Chapter 7 Sampling.

In this chapter the infer R package is ued. This package has a number of modern functions to easily simulate resampling from a datafram.

For further modern code, check out the tidymodels R package rsample.

Bootstrap

The *bootstrap* is a statistical method that allows us to approximate the *sampling distribution* even without access to the population.

Bootstrapping is a *resampling* method.

Bootstrapping uses *sampling with replacement.*

Sketch the difference

Note that the main difference between the CLT and the Bootstrap is that for the CLT the sample size n goes to infinity and with the Bootstrap the sample size remains fixed and the number of samples B goes to infinity.

ModernDive Chapter 8 Bootstrapping and Confidence Intervals.

For a ModernDive into Bootstrapping and Confidence Intervals try the code in Chapter 8 Bootstrapping and Confidence Intervals.

In this chapter the infer R package is ued. This package has a number of modern functions to easily simulate resampling from a dataframe.

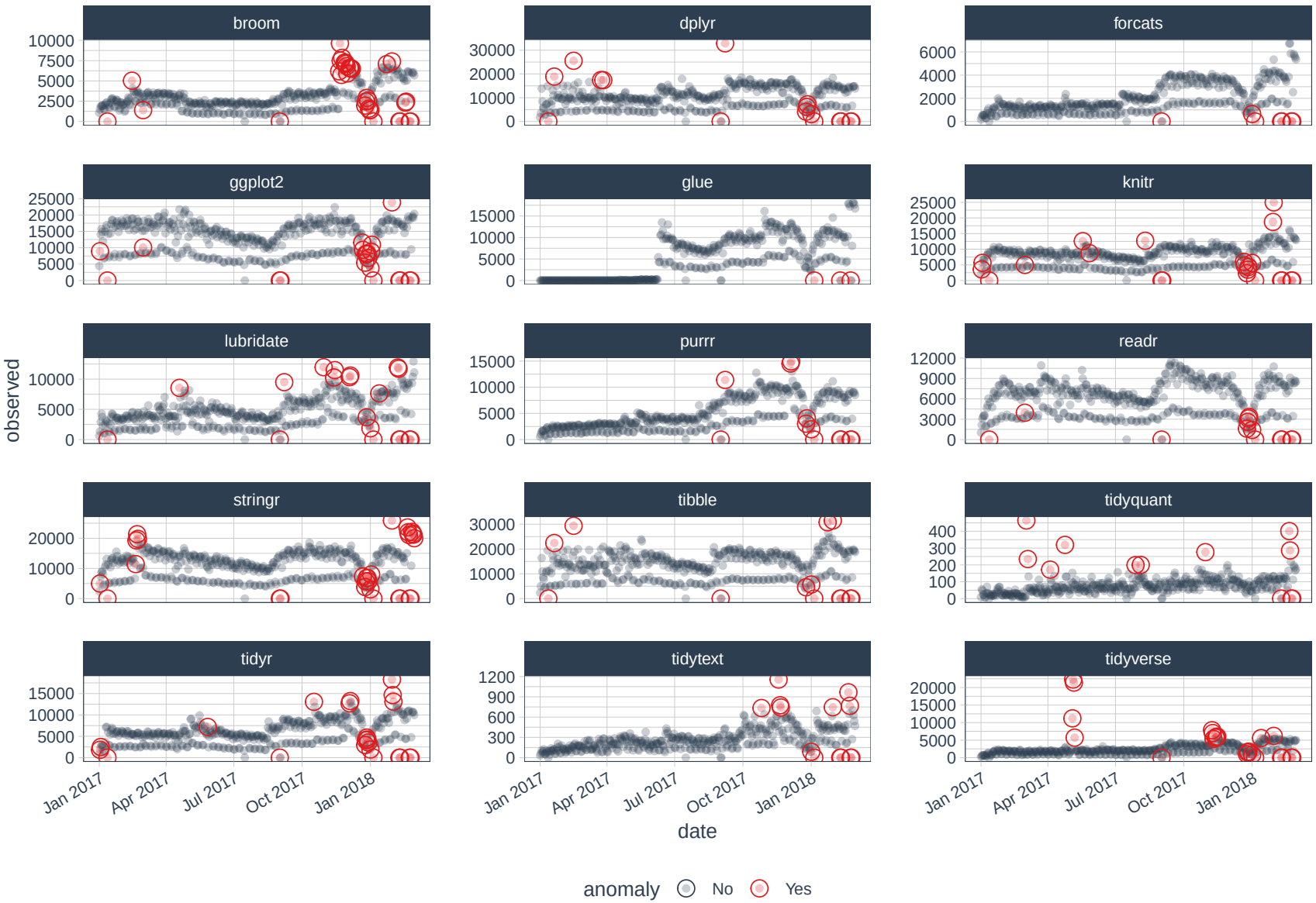
For further modern code, check out the tidymodels R package rsample.

Outliers

Outliers should never be dropped unless there is a clear rationale. If outliers are dropped this should be clearly reported.

```
p_load(anomalize, tibbletime)

tidyverse_cran_downloads %>%
    time_decompose(count, merge = TRUE) %>%
    anomalize(remainder) %>%
    time_recompose() %>%
    plot_anomalies(ncol = 3, alpha_dots = 0.25)
```

Outliers


broom
dplyr
forcats
ggplot2
glue
knitr
lubridate
purrr
readr
stringr
tibble
tidyquant
tidyr
tidytext
tidyverse
observed
date
Jan 2017
Apr 2017
Jul 2017
Oct 2017
Jan 2018
anomaly
No
Yes

Statistical Models

Statistical models are used to explain variation between *response variables* and *explanatory variables*.

Linear Regression models are commonly used to build models. They are fit using the least squares algorithm. This algorithm leads to unbiased estimators that have minimum variance.

- We know that the estimators of the parameters in the model are computed used **optimization**.
- We know that the estimators are **unbiased**.

Confounding Variables

What does the correlation coefficient measure? **Answer:** ???

Recall "Correlation does not imply causation."

The gold standard is a controlled experiment. The authors describe the idea of *A/B testing*.

Most data collected today is observational. So no designed experiment has been used.

Recall **Simpson's Paradox**.

Problems with **p-values**

Everyone is using *many many many* p-values all assuming $\alpha = 0.05$.

This causes much higher overall error rates.

When using *multiple comparisons* and overall error rate should be addressed.

Appendix E.

Be sure to read Appendix E at the end of the book.

It includes a very nice summary of fitting Multiple Linear Regression.

Confounding variables

A confounding variable is another variable that influences the other variables.

Simpson's Paradox

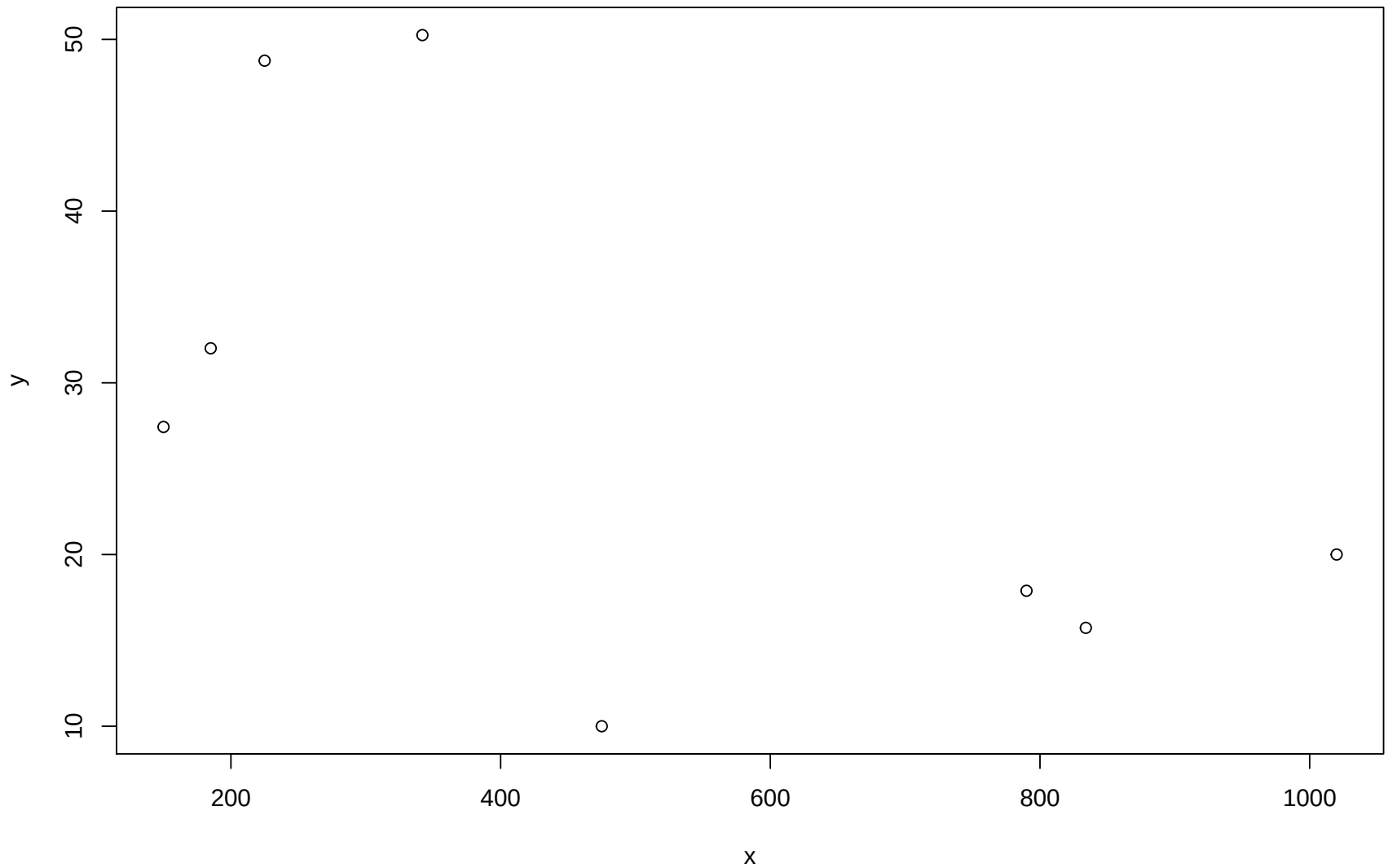
Edward Simpson: Bayes at Bletchley Park

Example of Simpson's Paradox

```
### synthetic data

# Consider book price (y) by number of pages (x)

z = c("hardcover","hardcover",
      "hardcover","hardcover",
      "paperback", "paperback","paperback",
      "paperback")

x1 = c( 150, 225, 342, 185)
y1 = c( 27.43, 48.76, 50.25, 32.01 )

x2 = c( 475, 834, 1020, 790)
y2 = c( 10.00, 15.73, 20.00, 17.89 )

x = c(x1, x2)
y = c(y1, y2)
```

Example of Simpson's Paradox

```
plot(x,y)
```



Example of Simpson's Paradox

```
# correlation

cor(y, x)
```

```
## [1] -0.5949366
```

```
cor(y1, x1)
```

```
## [1] 0.8481439
```

```
cor(y2, x2)
```

```
## [1] 0.9559518
```

Example of Simpson's Paradox

```
# linear regression

lm(y ~ x)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Coefficients:
## (Intercept)            x
##    41.15238     -0.02665
```

Example of Simpson's Paradox

```
# linear regression

lm(y1 ~ x1)
```

```
##
## Call:
## lm(formula = y1 ~ x1)
##
## Coefficients:
## (Intercept)           x1
##     13.0613       0.1177
```

Example of Simpson's Paradox

```
# linear regression

lm(y2 ~ x2)
```

```
##
## Call:
## lm(formula = y2 ~ x2)
##
## Coefficients:
## (Intercept)           x2
##     1.72389      0.01819
```

Example of Simpson's Paradox

Summary: Simpson's Paradox is the changing of the direction of a relationship with the introduction of another variable.

The relationship between Price and Number of pages in a book changes with the introduction of the variable Type of Book (Hardcover, Paperback).

See the R Markdown document SimpsonsParadox available on RPubs.com/esuess.