

# 009 - Random Variables

EPIB 607

Sahir Rai Bhatnagar

Department of Epidemiology, Biostatistics, and Occupational Health  
McGill University

`sahir.bhatnagar@mcgill.ca`

<https://sahirbhatnagar.com/EPIB607/randomVariables.html>

slides compiled on September 23, 2021





# Introduction

- This central chapter addresses a fundamental concept, namely the laws governing the variance of a sum of 2, or (especially)  $n$  random variables – and even more importantly – the laws governing the variance of a difference of two random variables.
- The latter is central, not just to simple contrasts involving 2 sample means or proportions, but also in the much wider world of regression, since the variance (sampling variability) of any regression slope can be viewed as the variance of a linear combination of random errors, or random deviations, or random variables.
- So, if there is one master formula to pay attention to and to own, it is the one for the variance of a linear combination of random variables. All others are special cases of this.

# Objectives

- Understand the equations for expectation and variance of both a continuous and discrete random variable.
- Why it is that, when dealing with the sum of two or more independent random variables, it is not their standard deviations that sum (add), but rather their variances.
- Likewise, when dealing with the difference of two independent random variables, or some linear combination of  $n$  independent random variables involving positive and negative weights, why it is that the component variances add, and with what weights.

# Discrete Random Variable (RV)

## Definition 1 (Discrete Random Variable).

A random variable that assumes only a finite (or countably infinite) number of distinct values. Discrete random variables have a finite or countably infinite number of possible values, each with positive or zero probability.

## Definition 2 (Probability Mass Function (PMF)).

The probability mass function (PMF) of a discrete random variable  $Y$  provides the possible values  $y$  and their associated probabilities by  $P(Y = y)$ . The sum of all probabilities must sum to 1, i.e.  $\sum_y P(Y = y) = 1$ .

# Continuous Random Variable

## Definition 3 (Continuous Random Variable).

A random variable is continuous if both of the following apply:

1. Its set of possible values consists either of all numbers in a single interval on the number line (possibly infinite in extent, e.g., from  $-\infty$  to  $\infty$ ) or all numbers in a disjoint union of such intervals.
2. No possible value of the variable has positive probability, that is,  $P(X = c) = 0$  for any possible value  $c$ .

## Definition 4 (Probability Density Function (PDF)).

Let  $Y$  be a continuous random variable. Then a probability distribution or probability density function (pdf) of  $Y$  is a function  $f(y)$  such that for any two numbers  $a$  and  $b$  with  $a \leq b$ ,

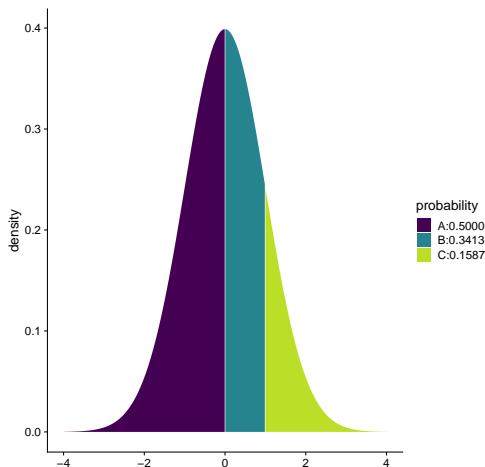
$$P(a \leq Y \leq b) = \int_a^b f(y) dy.$$

That is, the probability that  $Y$  takes on a value in the interval  $[a, b]$  is the area above this interval and under the graph of the density function

# How Can Every Value Have Probability 0?

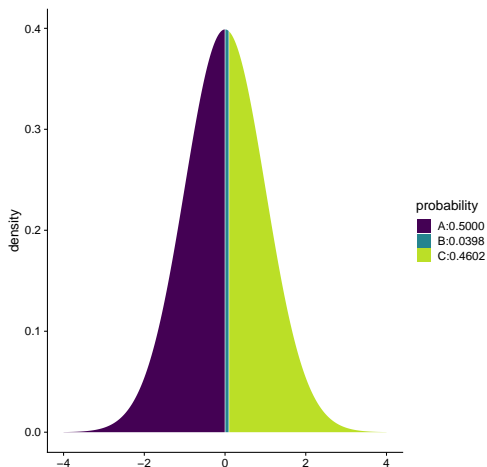
We can find a probability for any interval of z-scores. But the probability for a single z-score is zero. How can that be? Let's look at the standard Normal random variable,  $Z$ . We could find that the probability that  $Z$  lies between 0 and 1 is  $P(0 \leq Z \leq 1) = 0.3413$ .

```
mosaic::xpnorm(c(0,1))
```



$$P(0 \leq Z \leq 0.1)$$

```
mosaic::xpnorm(c(0,1/10))
```

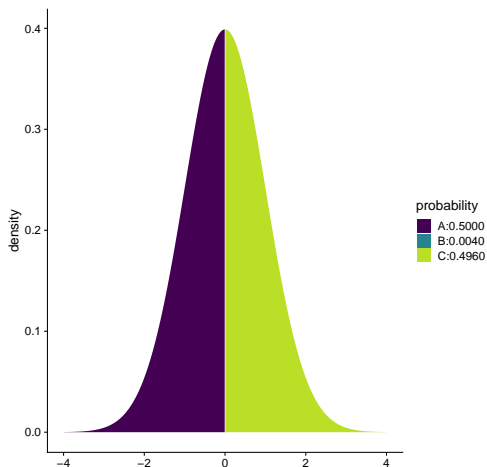


```
## [1] 0.5000000 0.5398278
```



$$P(0 \leq Z \leq 0.01)$$

```
mosaic::xpnorm(c(0,1/100))
```



```
## [1] 0.5000000 0.5039894
```

# How Can Every Value Have Probability 0?

- So, what's the probability that  $Z$  is exactly 0? Well, there's no area under the curve right at  $x = 0$ , so the probability is 0.
- It's only **intervals** that have positive probability, but that's OK.
- In real life we never mean exactly 0.0000000000 or any other value. If you say “exactly 164 pounds,” you might really mean between 163.5 and 164.5 pounds or even between 163.99 and 164.01 pounds, but realistically not 164.000000000 . . . pounds



# Expected value for a discrete RV

## Definition 5.

Let  $Y$  be a discrete random variable with set of possible values  $D = \{y_1, y_2, \dots, y_k\}$  and corresponding probabilities for each value, e.g.,  $y_1$  with probability  $P(y_1)$ ,  $y_2$  with probability  $P(y_2)$ ,  $y_3$  with probability  $P(y_3)$ ,  $\dots$ ,  $y_k$  with probability  $P(y_k)$ . Furthermore, let  $g(Y)$  be some real-valued function of  $Y$ . Then the expected value of  $g(Y)$  is:

$$E(g(Y)) = \sum_{y \in D} g(y) \times P(y) .$$

i.e. it is a weighted mean of the  $g(y)$ 's, with  $P(y)$ 's as weights.

Let  $c$  be a constant and  $Z$  another random variable

- $g(Y) = Y + c \rightarrow$
- $g(Y) = cY \rightarrow$
- $g(Y, Z) = Y + Z \rightarrow$

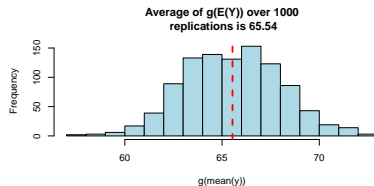
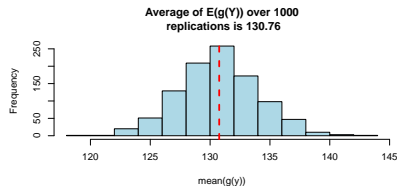
## Exercise: $E(g(Y)) = g(E(Y))$ ?

- $Y$  = Noon Temperature (C) in Montreal on a randomly selected day of the year;  $g(Y)$  = Temperature (F) =  $32 + (9/5) Y$
- $Y_1$  and  $Y_2$  are two random variables that might or might not be related;  $g(Y_1, Y_2) = Y_1 + Y_2$
- $g(Y_1, Y_2) = \frac{Y_1 + Y_2}{2}$
- $g(Y_i) = \frac{1}{n} \sum_{i=1}^n Y_i$
- $Y$  = diameter of a randomly chosen sphere;  $g(Y)$  = Volume of sphere =  $\frac{\pi}{6} Y^3$

Example:  $g(Y) = \text{Volume of sphere} = \frac{\pi}{6} Y^3$

# Example: Checking via simulation

```
g.y <- function(y) {  
  (pi / 6) * y^3  
}  
  
set.seed(12)  
B <- 1000; N <- 2000  
E_g.y <- replicate(B, {  
  diameter <- runif(N, min = 0, max = 10)  
  mean(g.y(diameter)) # E(g(y))  
})  
  
g_E.y <- replicate(B, {  
  diameter <- runif(N, min = 0, max = 10)  
  g.y(mean(diameter)) # g(E(y))  
})  
  
par(mfrow = c(1,2))  
hist(E_g.y, col = "lightblue", xlab = "mean(g(y))",  
     main = sprintf("Average of E(g(Y)) over 1000\nreplications  
is %.2f", mean(E_g.y)))  
abline(v = mean(E_g.y), col = "red", lty = 2, lwd = 3)  
  
hist(g_E.y, col = "lightblue", xlab = "g(mean(y))",  
     main = sprintf("Average of g(E(Y)) over 1000\nreplications  
is %.2f", mean(g_E.y)))  
abline(v = mean(g_E.y), col = "red", lty = 2, lwd = 3)
```



## Example: Checking the results theoretically

- We **know** for  $Y \sim \text{Uniform}(a, b)$ , the  $n^{\text{th}}$  moment  $E(Y^n)$  is given by:

$$E(Y^n) = \frac{b^{n+1} - a^{n+1}}{(n+1)(b-a)}$$

- Therefore, we know, theoretically, for  $Y \sim \text{Uniform}(0, 10)$ :

$$E(Y) = \quad (1)$$

$$E(Y^3) = \quad (2)$$

- It follows that, theoretically,

$$E(g(Y)) = 130.9 \quad (3)$$

$$g(E(Y)) = 65.45 \quad (4)$$





# Variance of an RV

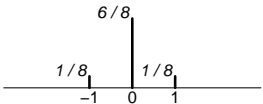
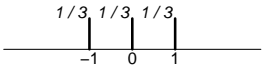
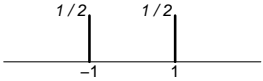
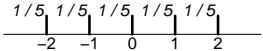


## Definition 6.

The variance of the random variable  $Y$  is given by

$$\text{Var}(Y) = E[(Y - \mu)^2].$$

- Discrete RV:  $\text{Var}(Y) = \sum_y (y - E(Y))^2 \times f(y)$
- Continuous RV:  $\text{Var}(Y) = \int_y (y - E(Y))^2 \times f(y) dy$

# Graphical representation of variance of a RV

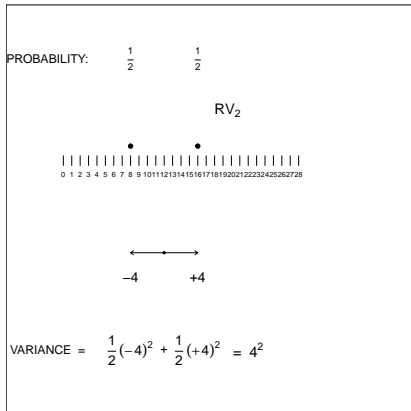
	Mean Absolute Deviation	Mean Squared Deviation (Variance)	Root Mean Squared Deviation (SD; écart type)
	0.25	0.25	0.5
	0.67	0.67	0.82
	1	1	1
	1.2	2	1.41
	2	4	2
	2.5	6.5	2.55

# Variance of a function of a RV

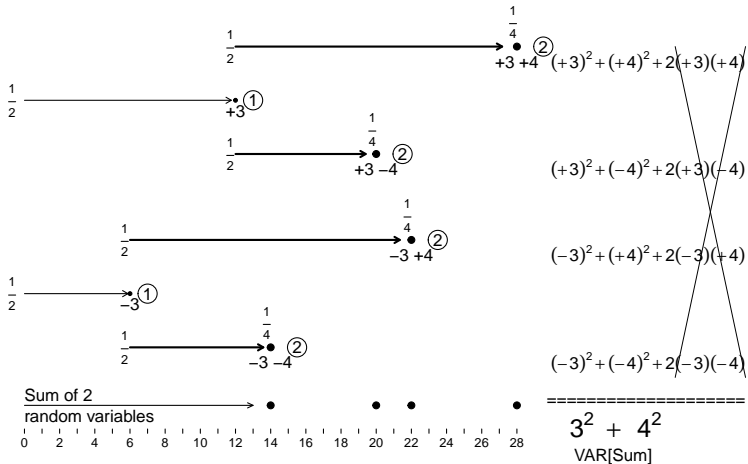
- $Y$  = Noon Temperature (C) in Mtl on a randomly selected day of the year;  $g(Y)$  = Temperature (F) =  $32 + (9/5) Y$
- $Y$  = Years of publication of all the books in the McGill Library, with Years measured from 1439 AD;  $W = Y - 1439$  vs.  $W = 2020 - Y$



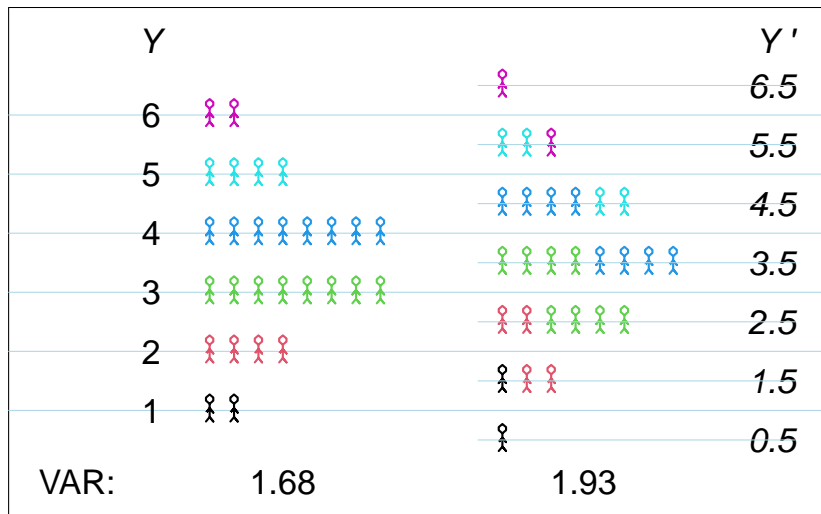
## Sums, means, differences of random variables



# The Squared Deviations



## Application: Measurement errors





## A sum of $n$ random variables

- Up to now, to keep things general, we used  $n$  non-identical – but independent – random variables. If we consider the Variance and the sum of  $n$  **identical** – and independent – random variables, so the  $n$  Variances (each abbreviated to Var) are all equal, the laws simplify:
- First, since the variances add, we have that

$$\text{Var}(RV_1 + RV_2 + \cdots + RV_n) = \text{Var}_1 + \text{Var}_2 + \cdots + \text{Var}_n = n \times \text{each Var}.$$

- Taking square roots,

$$SD(RV_1 + RV_2 + \cdots + RV_n) = \sqrt{n \times \text{each Var}} = \sqrt{n} \times \text{each SD}$$

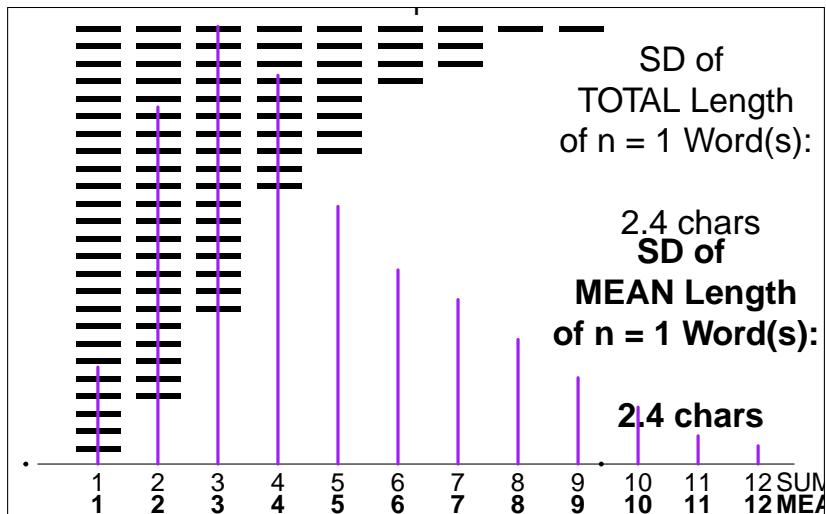
•

$$SD\left(\frac{RV_1 + RV_2 + \cdots + RV_n}{n}\right) = \frac{\sqrt{n} \times \text{each SD}}{n} = \frac{\text{common SD}}{\sqrt{n}}$$

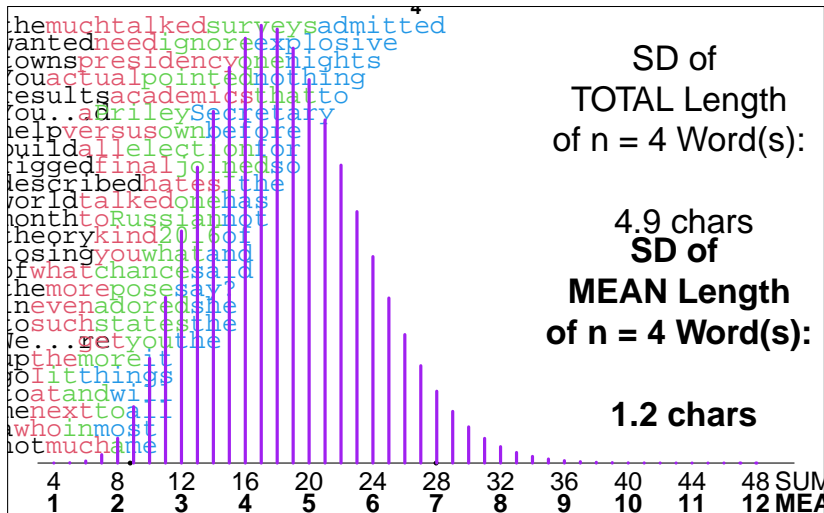
•

$$\text{Var}\left(\frac{RV_1 + RV_2 + \cdots + RV_n}{n}\right) = \frac{\text{common Var}}{n}$$

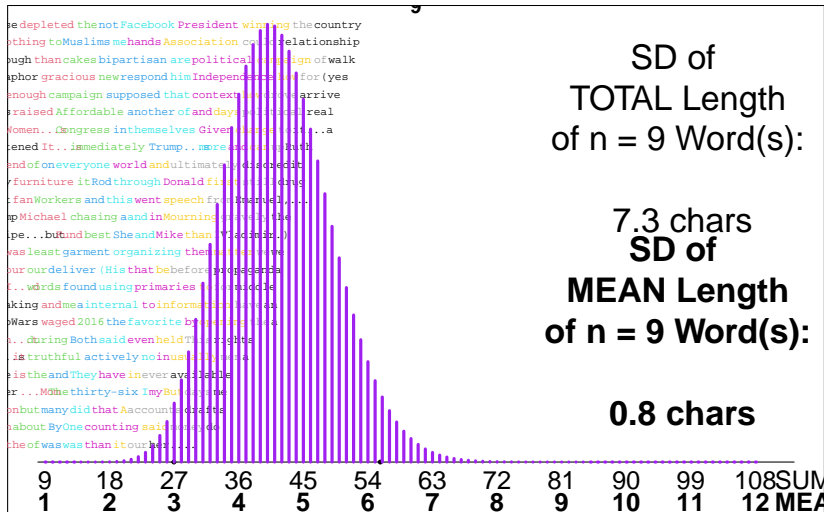
## Example: Length of Words in a Book - $n = 1$ word



$n = 4$  words



$n = 9$  words

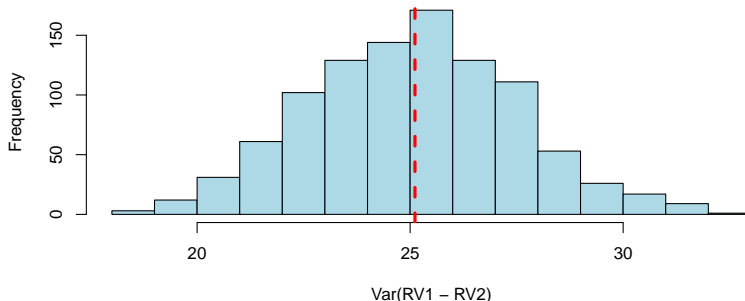


# Difference of 2 Random Variables via Simulation

```
set.seed(12)
B <- 999; N <- 200
var_diff <- replicate(B, {
  RV1 <- rnorm(N, mean = 2, sd = 3)
  RV2 <- rnorm(N, mean = 4, sd = 4)
  var(RV1 - RV2)
})

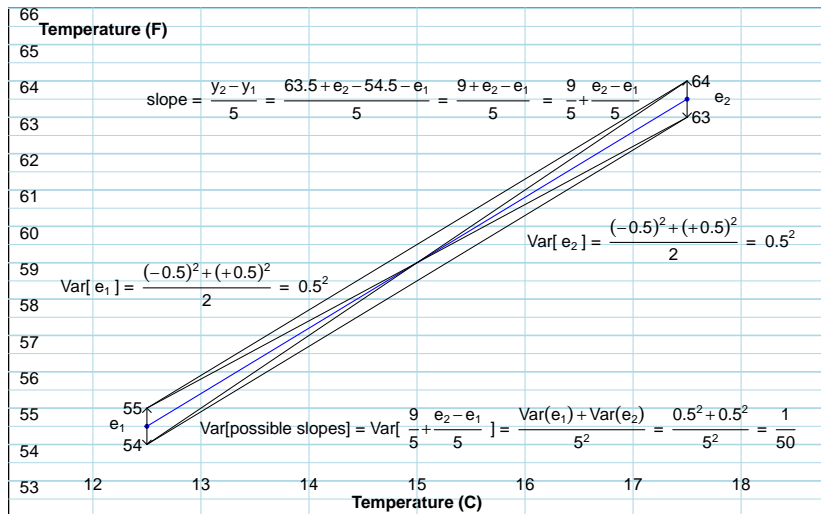
hist(var_diff, col = "lightblue", xlab = "Var(RV1 - RV2)",
     main = sprintf("Median of Var(RV1-RV2) over 999 replications is %0.2f", median(var_diff))),
abline(v = median(var_diff), col = "red", lty = 2, lwd = 3)
```

**Median of  $\text{Var}(\text{RV1}-\text{RV2})$  over 999 replications is 25.11**





# Linear Combinations of random variables



# Session Info

```
R version 4.0.4 (2021-02-15)
Platform: x86_64-pc-linux-gnu (64-bit)
Running under: Pop!_OS 21.04

Matrix products: default
BLAS:   /usr/lib/x86_64-linux-gnu/openblas-pthread/libblas.so.3
LAPACK: /usr/lib/x86_64-linux-gnu/openblas-pthread/libopenblas-p-r0.3.13.so

attached base packages:
[1] tools      stats      graphics  grDevices  utils      datasets  methods
[8] base

other attached packages:
[1] DT_0.16 mosaic_1.7.0 Matrix_1.3-2 mosaicData_0.20.1
[5] ggformula_0.9.4 ggstance_0.3.4 lattice_0.20-41 kableExtra_1.2.1
[9] socviz_1.2 gapminder_0.3.0 here_0.1 NCStats_0.4.7
[13] FSA_0.8.30 forcats_0.5.1 stringr_1.4.0 dplyr_1.0.7
[17] purrr_0.3.4 readr_1.4.0 tidyr_1.1.3 tibble_3.1.3
[21] ggplot2_3.3.5 tidyverse_1.3.0 knitr_1.33

loaded via a namespace (and not attached):
[1] fs_1.5.0 lubridate_1.7.9 webshot_0.5.2 httr_1.4.2
[5] rprojroot_2.0.2 backports_1.2.1 utf8_1.2.2 R6_2.5.1
[9] DBI_1.1.1 colorspace_2.0-2 withr_2.4.2 tidyselect_1.1.1
[13] gridExtra_2.3 leaflet_2.0.3 curl_4.3.2 compiler_4.0.4
[17] cli_3.0.1 rvest_1.0.0 pacman_0.5.1 xml2_1.3.2
[21] ggdendro_0.1.22 labeling_0.4.2 mosaicCore_0.8.0 scales_1.1.1
[25] digest_0.6.27 foreign_0.8-81 rmarkdown_2.9.7 rio_0.5.16
[29] pkgconfig_2.0.3 htmltools_0.5.2 highr_0.9 dbplyr_1.4.4
[33] fastmap_1.1.0 htmlwidgets_1.5.3 rlang_0.4.11 readxl_1.3.1
[37] rstudioapi_0.13 farver_2.1.0 generics_0.1.0 jsonlite_1.7.2
[41] crosstalk_1.1.1 zip_2.2.0 car_3.0-9 magrittr_2.0.1
[45] Rcpp_1.0.7 munsell_0.5.0 fansi_0.5.0 abind_1.4-5
[49] lifecycle_1.0.0 stringi_1.7.3 carData_3.0-4 MASS_7.3-53.1
[53] plyr_1.8.6 grid_4.0.4 blob_1.2.1 ggrepel_0.8.2
[57] crayon_1.4.1 cowplot_1.1.0 haven_2.3.1 splines_4.0.4
[61] hms_1.0.0 pillar_1.6.2 reprex_0.3.0 glue_1.4.2
[65] evaluate_0.14 data.table_1.14.0 modelr_0.1.8 vctrs_0.3.8
[69] tweenr_1.0.1 cellranger_1.1.0 gtable_0.3.0 polyclip_1.10-0
[73] assertthat_0.2.1 TeachingDemos_2.12 xfun_0.25 ggforce_0.3.2
[77] openssl_1.5.1 broom_0.7.2 viridisLite_0.4.0 ellipsis_0.3.2
```