004 - Exploring Data - Part II

EPIB 607 - FALL 2020

Sahir Rai Bhatnagar
Department of Epidemiology, Biostatistics, and Occupational Health
McGill University

sahir.bhatnagar@mcgill.ca

slides compiled on September 9, 2020

Summarizing relationships between two variables

Approaches for summarizing relationships between two variables vary depending on variable types:

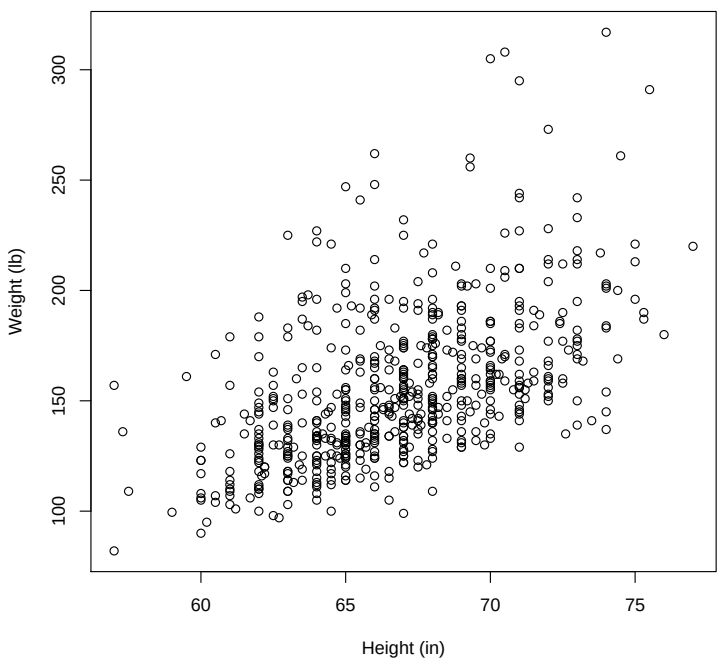
- Two numerical variables
- Two categorical variables
- One numerical variable and one categorical variable

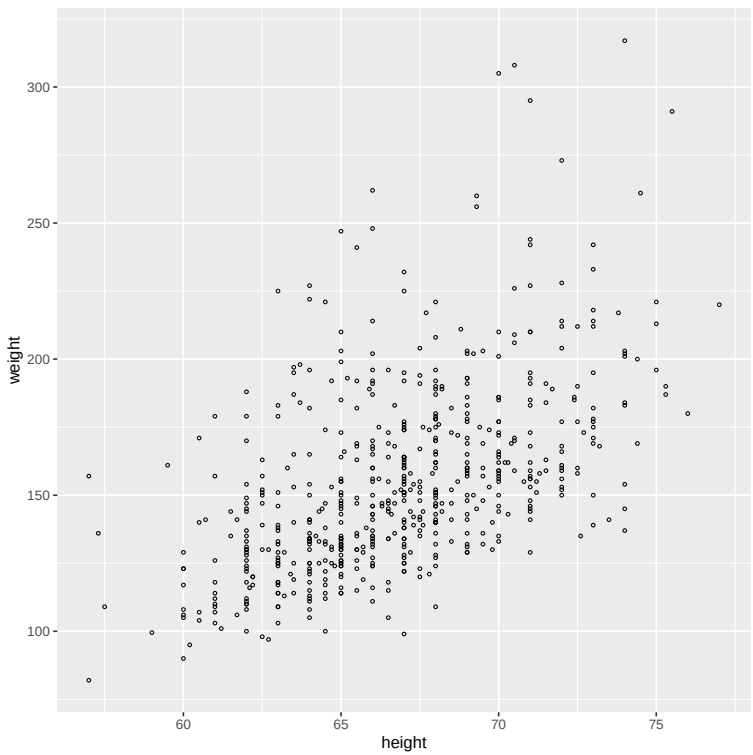
Scatterplots

```
library(ggplot2); library(oibiostat);
data(famuss)

plot(famuss$height, famuss$weight, xlab = "Height (in)", ylab = "Weight (lb)")

ggplot(data = famuss, mapping = aes(x = height, y = weight)) +
  geom_point(size = 0.8, pch = 21)
```





Pearson's correlation coefficient

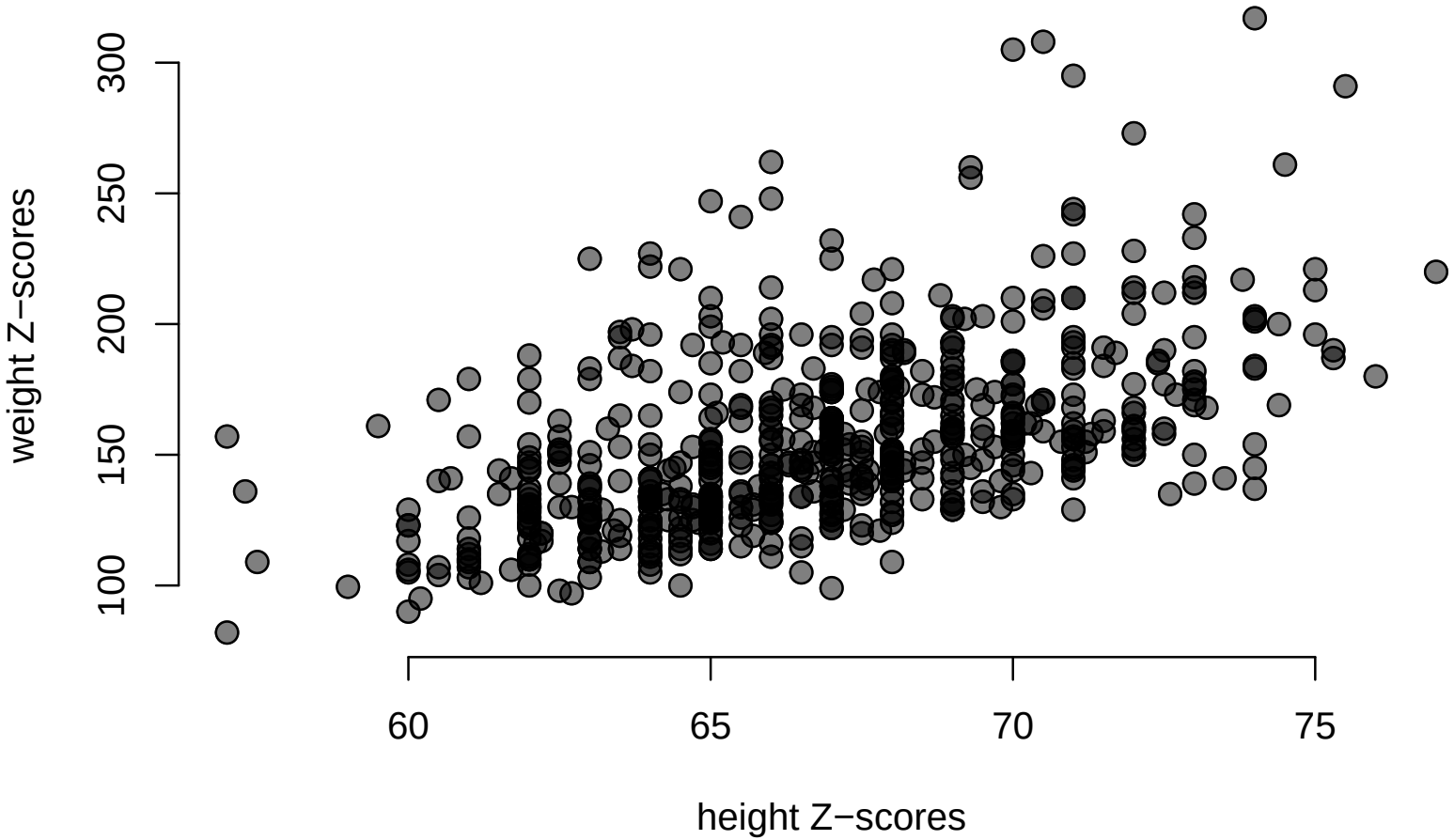
- The **sample** correlation (r) between two variables X and Y is given by:

$$r = \frac{1}{n-1} \sum_{i=1}^{n} z_X \cdot z_Y \tag{1}$$

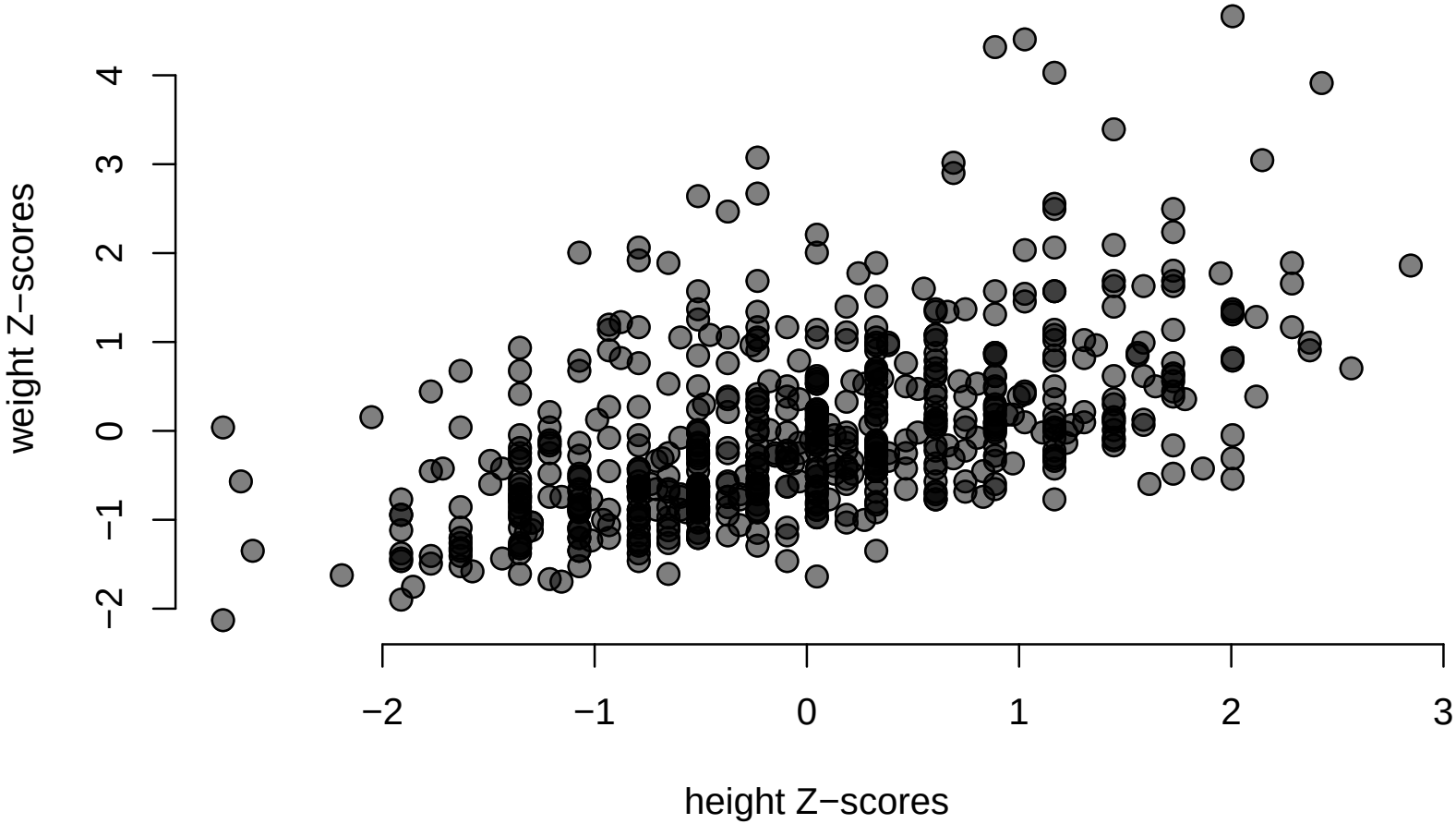
$$= \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_X} \right) \left(\frac{y_i - \bar{y}}{s_Y} \right) \tag{2}$$

- $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ the n paired sample values of X and Y
- z_X and z_Y are the sample Z-scores of the X and Y variables, respectively
- s_X and s_Y are the sample standard deviations of the X and Y variables, respectively
- $\bar{x}$ and $\bar{y}$ are the sample means of the X and Y variables, respectively
- The correlation coefficient quantifies the strength of a **linear** trend.

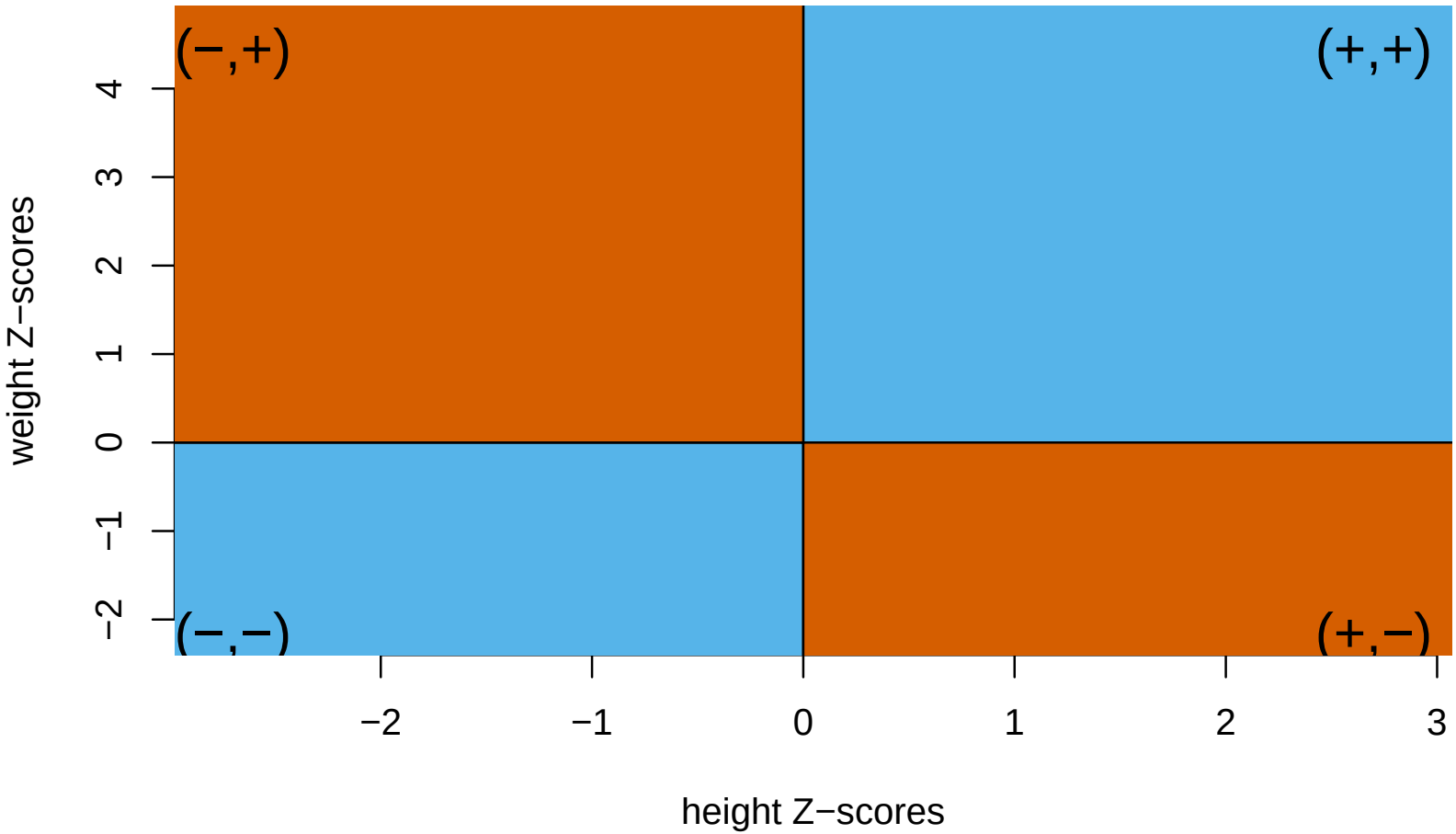
Plot of weight vs. height in `famuss` dataset



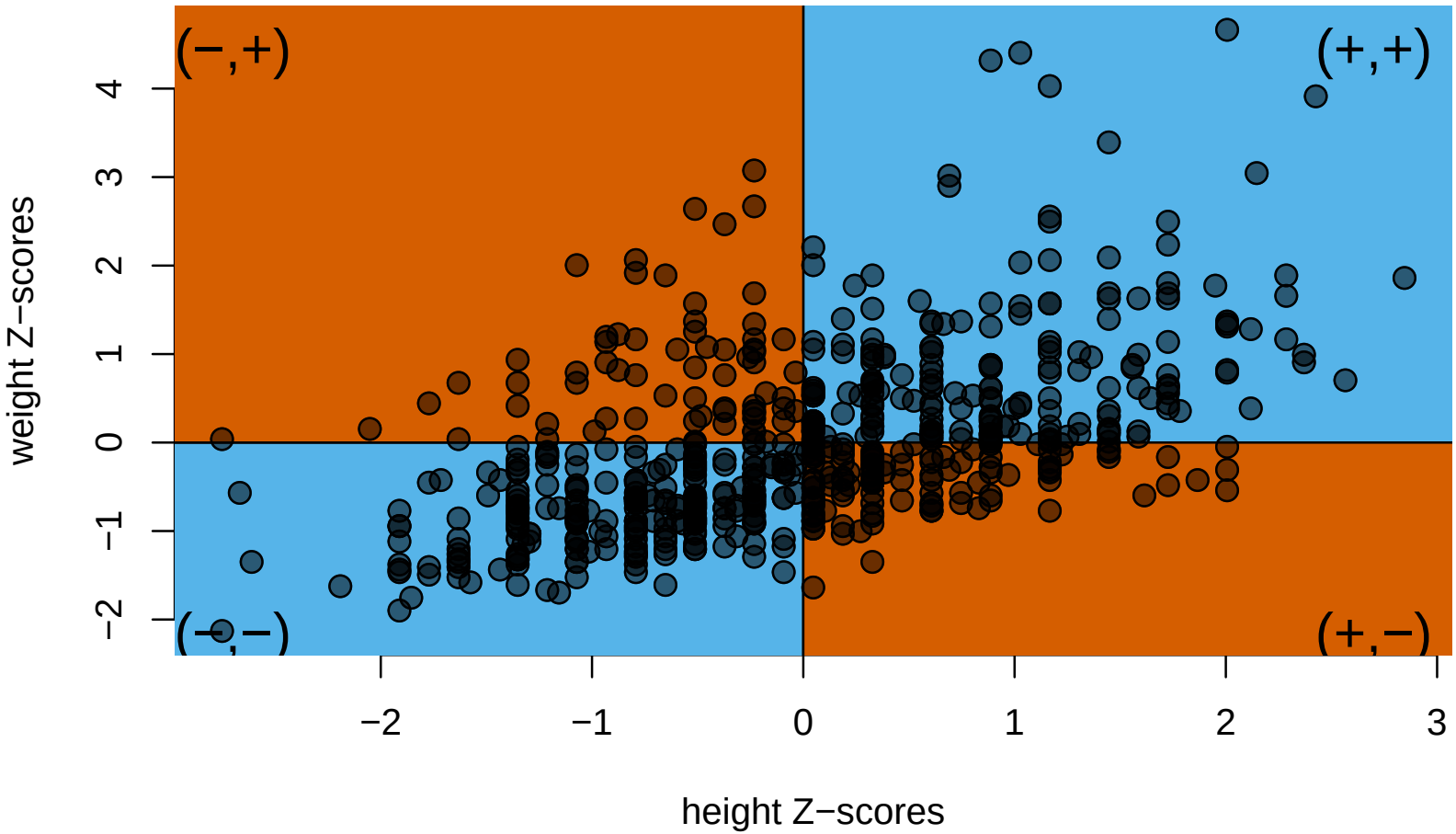
Plot of Z-scores weight vs. Z-scores height in `famuss` dataset



Partition the graph into four quadrants (x, y)



Correlation depends on which quadrants the points are on



Pearson's correlation coefficient

- The correlation coefficient r takes on values between -1 and 1.
- The closer r is to ± 1, the stronger the linear association.
- Two variables X and Y are
  - *positively associated* if Y increases as X increases ($r > 0$)
  - *negatively associated* if Y decreases as X increases ($r < 0$)
- Since the formula for calculating the correlation coefficient standardizes the variables, changes in scale or units of measurement will not affect its value

Exercise: Show mathematically that the correlation (r) is bounded by -1 and 1

Consider that we can't have higher correlation than when we compare a list to itself (perfect correlation).

Correlation and Simple linear Regression

- If we are predicting a random variable Y knowing the value of another variable $X = x$ using a regression line, then the formula for the regression can be given by:

$$\left(\frac{Y - \bar{y}}{s_Y}\right) = r\left(\frac{x - \bar{x}}{s_X}\right) \tag{3}$$

- This can be rewritten as:

$$Y = \bar{y} + r\left(\frac{x - \bar{x}}{s_X}\right) s_Y \tag{4}$$

Correlation in R

- Correlation between weight and height in the `famuss` dataset:

```
cor(famuss$height, famuss$weight)
## [1] 0.53
```

- We can also obtain the correlation between `weight` and `height` from a simple linear regression:

```
summary(lm(height ~ weight, data = famuss))
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  58.2952     0.5732   101.7   <2e-16 ***
## weight        0.0548     0.0036    15.2   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3 on 593 degrees of freedom
## Multiple R-squared: 0.282,^^IAdjusted R-squared: 0.281
## F-statistic:  233 on 1 and 593 DF,  p-value: <2e-16
```

- Exercise: calculate the correlation coefficient from the regression coefficient for weight.

Let's remind ourselves about random variability

- In many cases, we do not observe data for the entire population of interest but rather for a random sample.
- As with the mean and standard deviation, the sample correlation is the most commonly used estimator of the population correlation.
- This implies that the correlation we compute and use as a summary is a random variable.

Let's remind ourselves about random variability

Lets create a pseudo population from the 595 observations by sampling **with replacement**, and calculate the correlation. Lets repeat this process 1000 times:

```
B <- 1000; N <- 595
R <- replicate(B, {
        dplyr::sample_n(famuss, size = N, replace = TRUE) %>%
        dplyr::summarize(r = cor(height, weight)) %>%
        dplyr::pull(r)
})
```

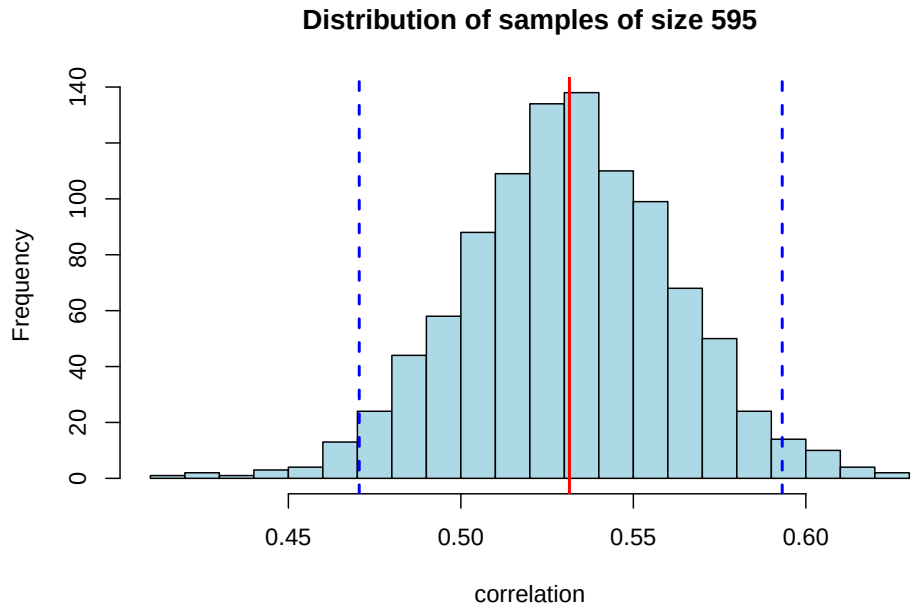
```
hist(R, breaks = 20, col = "lightblue", xlab = "correlation",
     main = "Distribution of samples of size 595")
abline(v = mean(R), col = "red", lwd = 2)
abline(v = quantile(R, probs = c(0.025, 0.975)), col = "blue",
       lty = 2, lwd = 2)
```

```
mean(R)

## [1] 0.53

quantile(R, probs = c(0.025, 0.975))

## 2.5%  98%
## 0.47 0.59
```

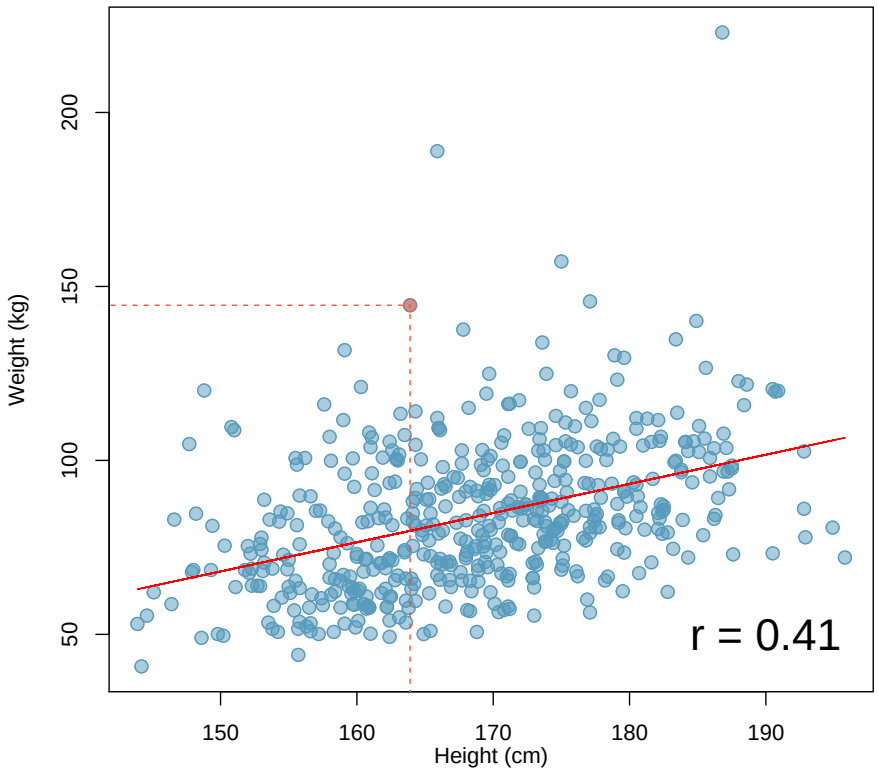


Another example: NHANES[2]

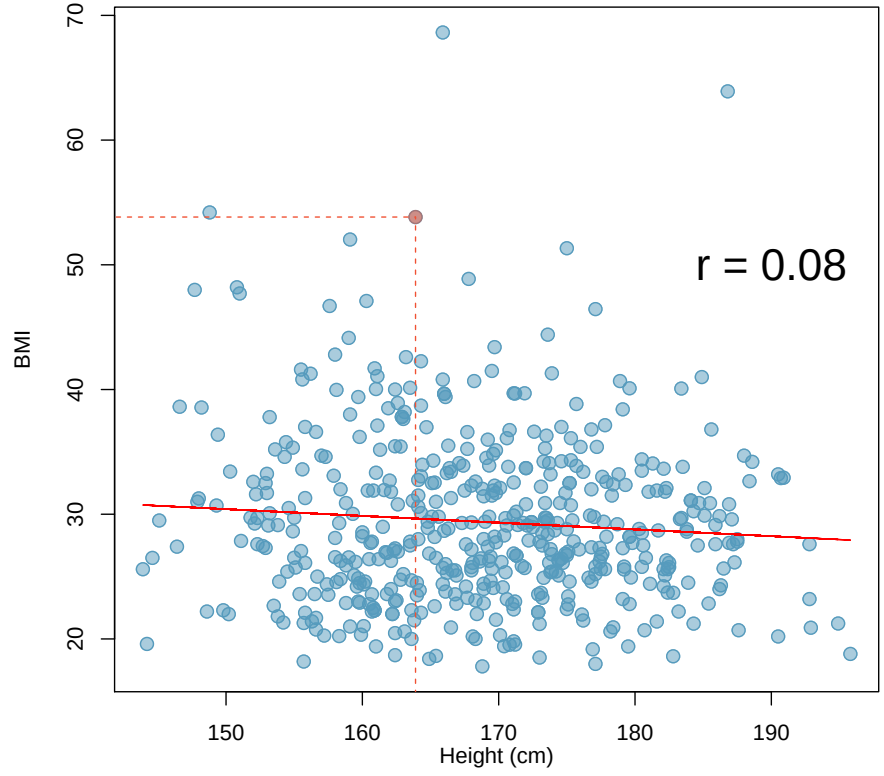
- The National Health and Nutrition Examination Survey (NHANES) consists of a set of surveys and measurements conducted by the US CDC to assess the health and nutritional status of adults and children in the United States.
- The following example uses data from a sample of 500 adults (individuals ages 21 and older) from the NHANES dataset[1].

[1] The sample is available as `nhanes.samp.adult.500` in the `R oibiostat` package

[2] http://www.cdc.gov/nchs/nhanes.htm



(a)



(b)

Figure: (a) A scatterplot showing height versus weight from the 500 individuals in the sample from NHANES. One participant 163.9 cm tall (about 5 ft, 4 in) and weighing 144.6 kg (about 319 lb) is highlighted. (b) A scatterplot showing height versus BMI from the 500 individuals in the sample from NHANES. The same individual highlighted in (a) is marked here, with BMI 53.83. Fitted regression lines are shown in red with correlation coefficient r. BMI = weight/height^2 $\times 703$.

Cautionary notes

- The formulas above are for a particular sample, hence the lower case letters r, x, y. In statistical terms, r is the **estimator** for the population-level correlation ρ (the **estimand**) of the random variables X and Y. The actual value of the sample correlation is denoted by $\widehat{r}$ and is called the **estimate**
- This implies that we are not 100% confident in our estimate and therefore should provide a confidence interval as well.
- A strong linear relationship is not necessarily a **causal** relationship, that is, just because $r \approx 1$ (or $r \approx -1$) does not mean that x **causes** changes in y (we may have a *spurious* correlation).
- Just because $r \approx 0$ does not mean that that x and y are unrelated, merely that they are **uncorrelated**. That is, it is possible to construct examples where x and y have a strong functional relationship, but where $r = 0$.
- X, Y independent $\Rightarrow r_{XY} = 0$
- $r_{XY} = 0 \not\Rightarrow X$, Y are independent

Anscombe's quartet[3]

```
library(datasets);data("anscombe")
```

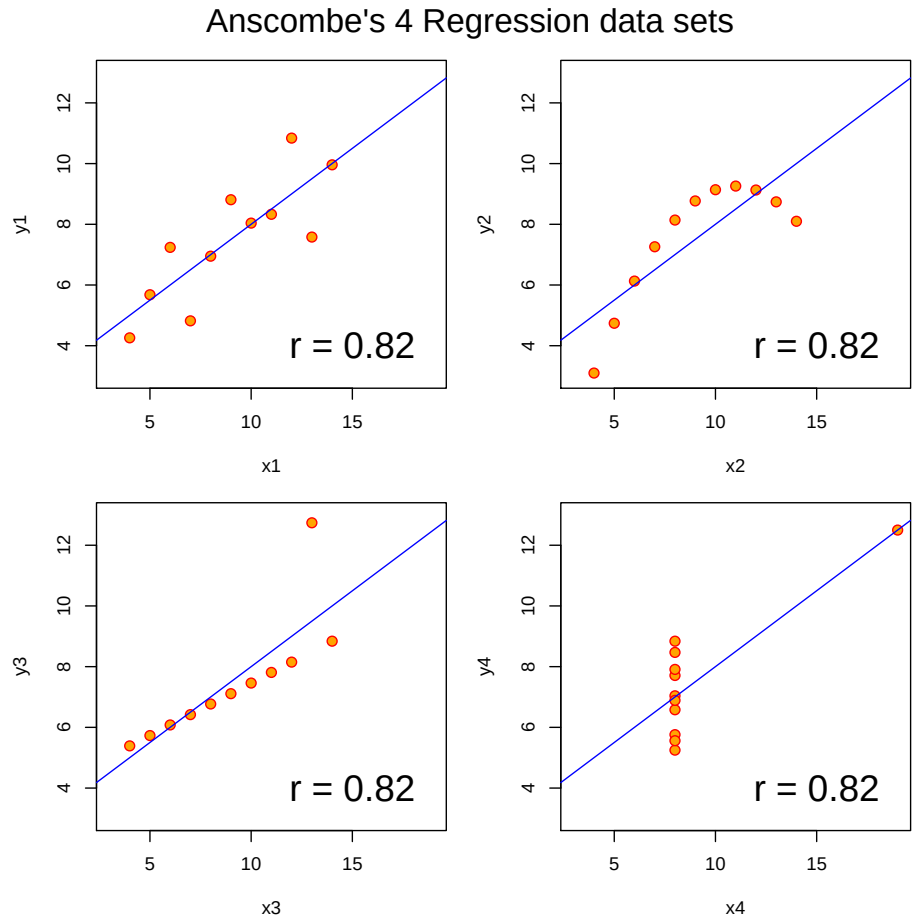
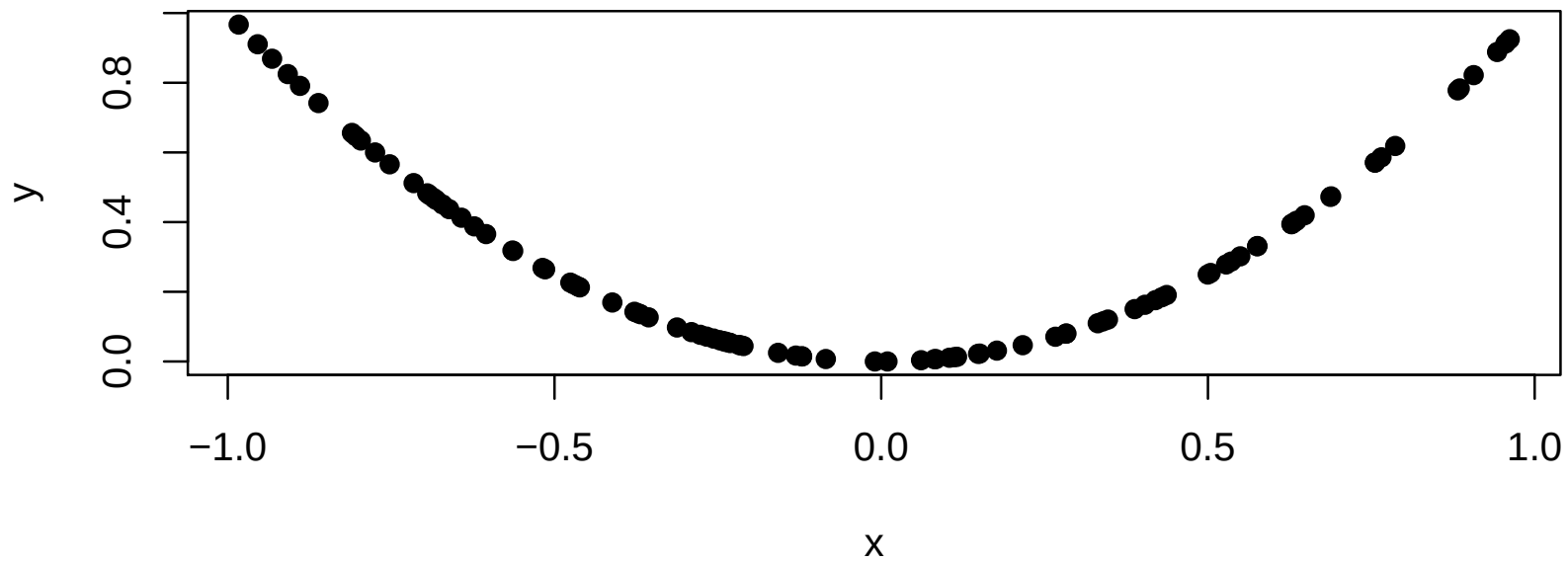


Figure: All four panels have the exact same linear correlation coefficient

[3] Anscombe, Francis J. (1973). Graphs in statistical analysis. The American Statistician, 27, 17–21. doi: 10.2307/2682899.

Zero linear correlation does not imply independence

```
set.seed(12)
x <- runif(100,-1,1)
y <- x^2
plot(x,y, pch = 19)
```



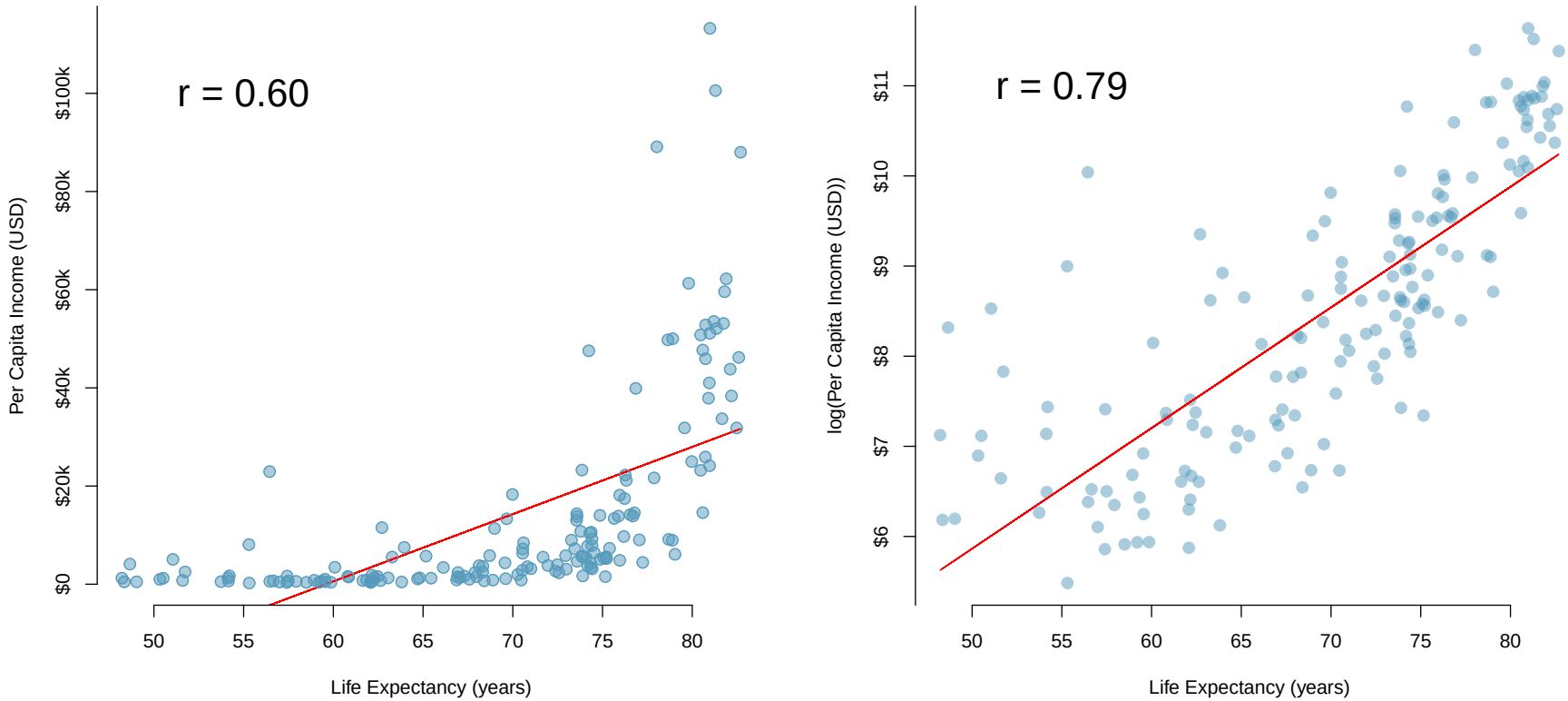

```
cor(x,y)

## [1] -0.023
```

Another example of same summary statistics but very different relationships

https://www.autodeskresearch.com/publications/samestats

Transformations to improve linear fit



(a) (b)

Figure: (a) per capita income vs. life expectancy (b) log per capita income vs. life expectancy. Fitted regression line in red with correlation coefficient r.[4]

[4] The World Development Indicators (WDI) is a database of country-level variables (i.e., indicators) recording outcomes for a variety of topics, including economics, health, mortality, fertility, and education

Rank correlation

The Pearson correlation, recall, is a measure of *linear* association. This may be undesirable for a number of reasons:

- y and x may be related, but not linearly (e.g., shape may be quadratic)
- one or both of y and x may be an ordered categorical variable (e.g., highest level of education attained, income category, age group, etc.) and the investigator may not wish to impose a particular numerical scale
- A nonparametric approach may be preferred if y or x are thought not to be Normally distributed

We can overcome these concerns using a correlation that is based on the ranks of the data, called **Spearman's rank correlation**.

Spearman's rank correlation

This is most easily understood through the use of an example.

Suppose we want to examine the correlation between gestational age (GA) and birthweight (BW).

Infant	1	2	3	4	5	6	7	8
BW (g)	2621	2863	3322	3508	3518	3770	3784	3801
GA (days)	270	271	267	268	276	282	288	278
BW rank	1	2	3	4	5	6	7	8
GA rank	3	4	1	2	5	7	8	6

Spearman's rank correlation

Spearman's rank correlation is based on the squared differences in rank for each individual:

Infant	1	2	3	4	5	6	7	8
Rank by BW	1	2	3	4	5	6	7	8
Rank by GA	3	4	1	2	5	7	8	6
Difference	-2	-2	2	2	0	-1	-1	2
Difference^2	4	4	4	4	0	1	1	4

Then Spearman's rank correlation coefficient is computed to be

$$r_s = 1 - \frac{6 \sum d^2}{n^3 - n}$$

In our example, this gives $r_s = 0.738$.

Spearman's rank correlation

Spearman's rank correlation is equivalent to calculating a Pearson's correlation on the ranks:

$$r = \text{Corr}(\text{Rank}_{\text{GA}}, \text{Rank}_{\text{BW}}) = 0.738$$

Rank correlation: Kendall's τ

There is another rank correlation, Kendall's τ, which we will again learn by example.

We study the correlation between gestational age and birthweight.

Infant	1	2	3	4	5	6	7	8
BW (g)	2621	2863	3322	3508	3518	3770	3784	3801
GA (days)	270	271	267	268	276	282	288	278
BW rank	1	2	3	4	5	6	7	8
GA rank	3	4	1	2	5	7	8	6

Rank correlation: Kendall's τ

Infant	1	2	3	4	5	6	7	8
Rank by b.weight	1	2	3	4	5	6	7	8
Rank by gest.age	3	4	1	2	5	7	8	6

First, we order the data according to one of the rankings (we chose to do so with birthweight).

Next, we sum the number of infants to the right of each cell with a **higher** ranking for the *other* variable (gestational age), and call this P:

$$P = 5 + 4 + 5 + 4 + 3 + 1 + 0 = 22$$

Rank correlation: Kendall's τ

Then Kendall's rank correlation coefficient is computed to be

$$\tau = \frac{2 \times P}{\frac{1}{2}n(n-1)} - 1$$

In our example, this gives $\tau = 0.57$.

We can perform hypothesis testing on Kendall's τ; the approximately Normal test statistic is

$$z = \frac{2 \times P}{\sqrt{n(n-1)(2n+5)/18}}$$

Rank correlation: Kendall's τ

In our example, if we wish to test $H_0 : \tau = 0$ vs. $H_A : \tau \neq 0$, this gives

$$\begin{aligned} z &= \frac{2 \times P}{\sqrt{n(n-1)(2n+5)/18}} \\ &= \frac{2 \times 22}{\sqrt{8 \times 7 \times 21/18}} \\ &= 5.444 \end{aligned}$$

which yields a p-value of $P(|Z| > 5.444) < 0.001$, indicating that there is a statistically significant association as measured by Kendall's rank correlation between gestational age and birthweight.

Rank correlation

Notes:

- Both Spearman's and Kendall's correlations lie between -1 and 1; positive values correspond to a positive association, negative values to a negative association.
- Both Spearman's and Kendall's correlations are nonparametric statistics.
- Corrections for ties are required (beyond the scope of this course). `R` handles it for you.

Two numerical variables and the correlation coefficient

Two categorical variables and contingency tables

A numerical variable and a categorical variable

Summary

Two categorical variables

A contingency table summarizes data for two categorical variables:

```
tab1 <- table(famuss$race,
              famuss$actn3.r577x)
tab1

##
##                CC  CT  TT
##   African Am   16   6   5
##   Asian        21  18  16
##   Caucasian   125 216 126
##   Hispanic      4  10   9
##   Other         7  11   5

addmargins(tab1)

##
##                CC  CT  TT Sum
##   African Am   16   6   5  27
##   Asian        21  18  16  55
##   Caucasian   125 216 126 467
##   Hispanic      4  10   9  23
##   Other         7  11   5  23
##   Sum         173 261 161 595
```

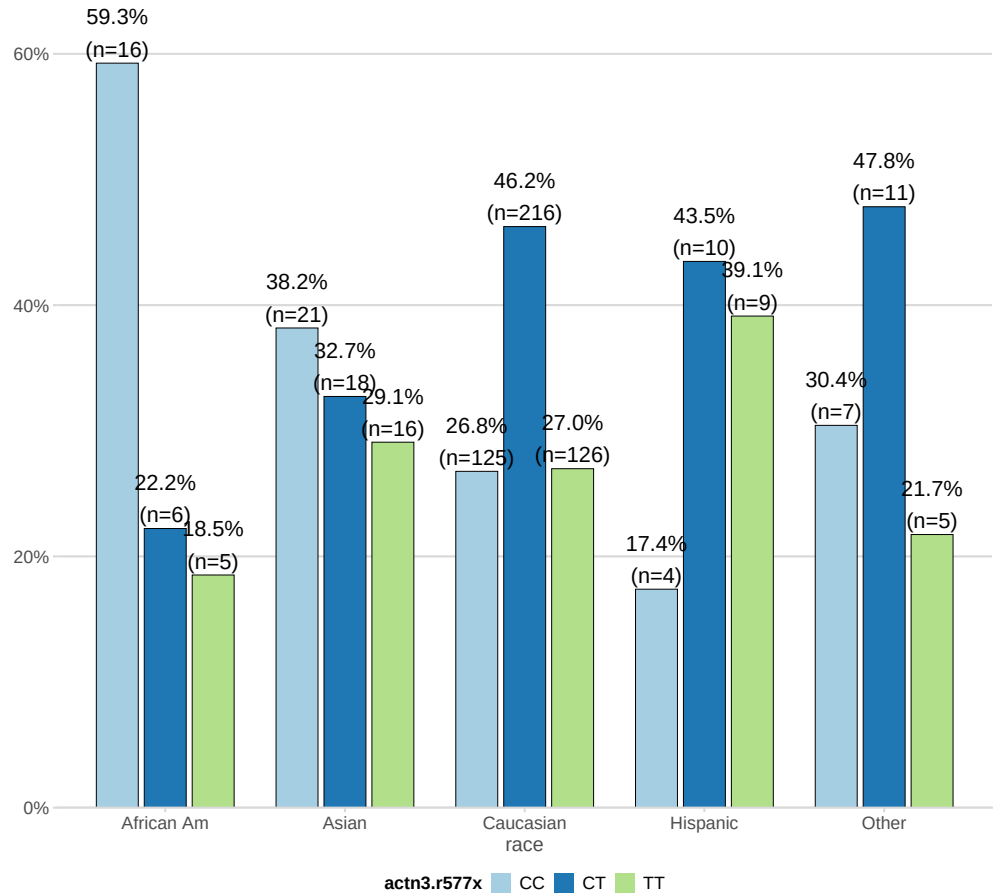
Conditional distribution of genotype *given* race

The distributions we create this way are called **conditional distributions**, because they show the distribution of one variable for just those cases that satisfy a condition on another variable

```
addmargins(
  prop.table(tab1, margin = 1)
)
```

```
##
##                 CC   CT   TT  Sum
##   African Am 0.59 0.22 0.19 1.00
##   Asian      0.38 0.33 0.29 1.00
##   Caucasian  0.27 0.46 0.27 1.00
##   Hispanic   0.17 0.43 0.39 1.00
##   Other      0.30 0.48 0.22 1.00
##   Sum        1.72 1.93 1.35 5.00
```

```
sjPlot::plot_xtab(famuss$race,
                  famuss$actn3.r577x,
                  margin = "row")
```

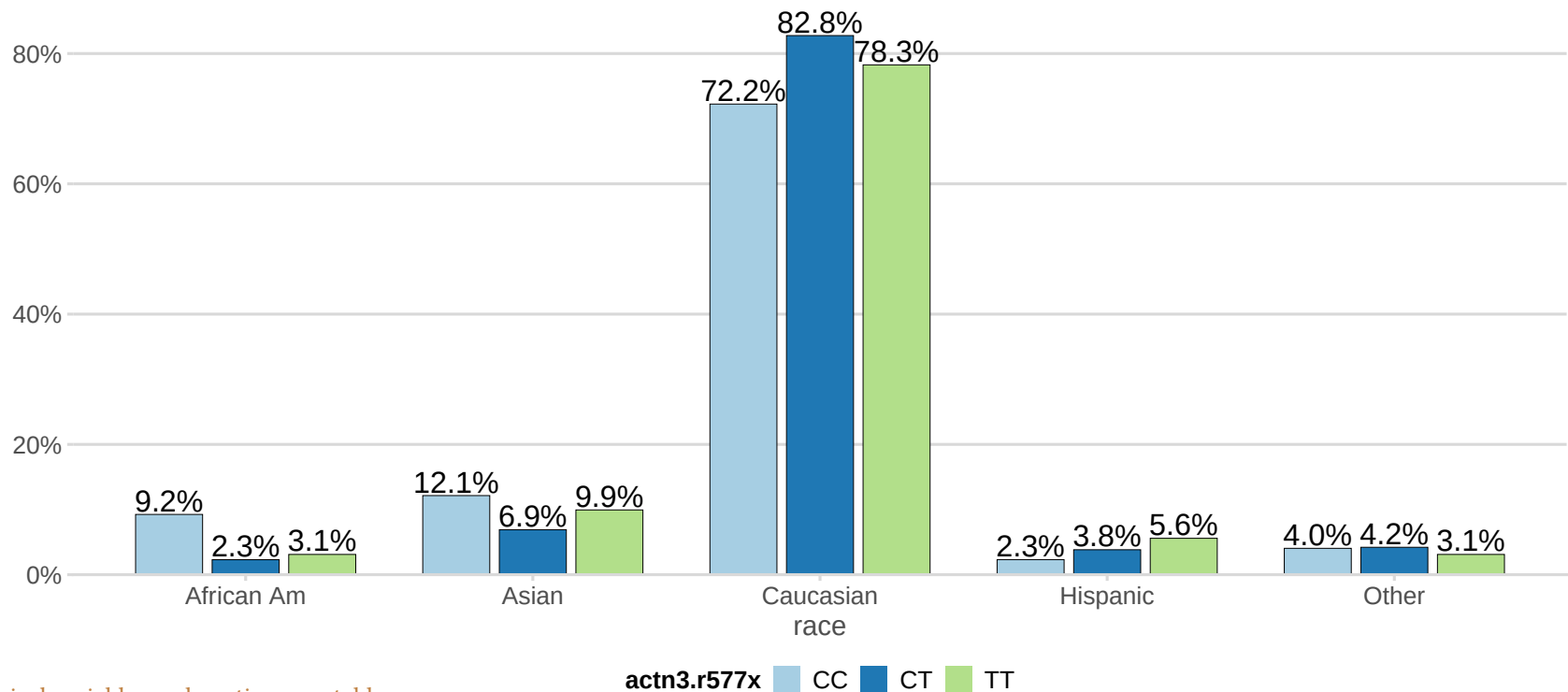


Conditional distribution of race *given* genotype

```
addmargins(prop.table(tab1, margin = 2))
```

```
##
##                  CC    CT    TT   Sum
##   African Am 0.092 0.023 0.031 0.147
##   Asian      0.121 0.069 0.099 0.290
##   Caucasian  0.723 0.828 0.783 2.333
##   Hispanic   0.023 0.038 0.056 0.117
##   Other      0.040 0.042 0.031 0.114
##   Sum        1.000 1.000 1.000 3.000
```

```
sjPlot::plot_xtab(famuss$race, famuss$actn3.r577x, margin = "col", show.total = F, show.n = F)
```



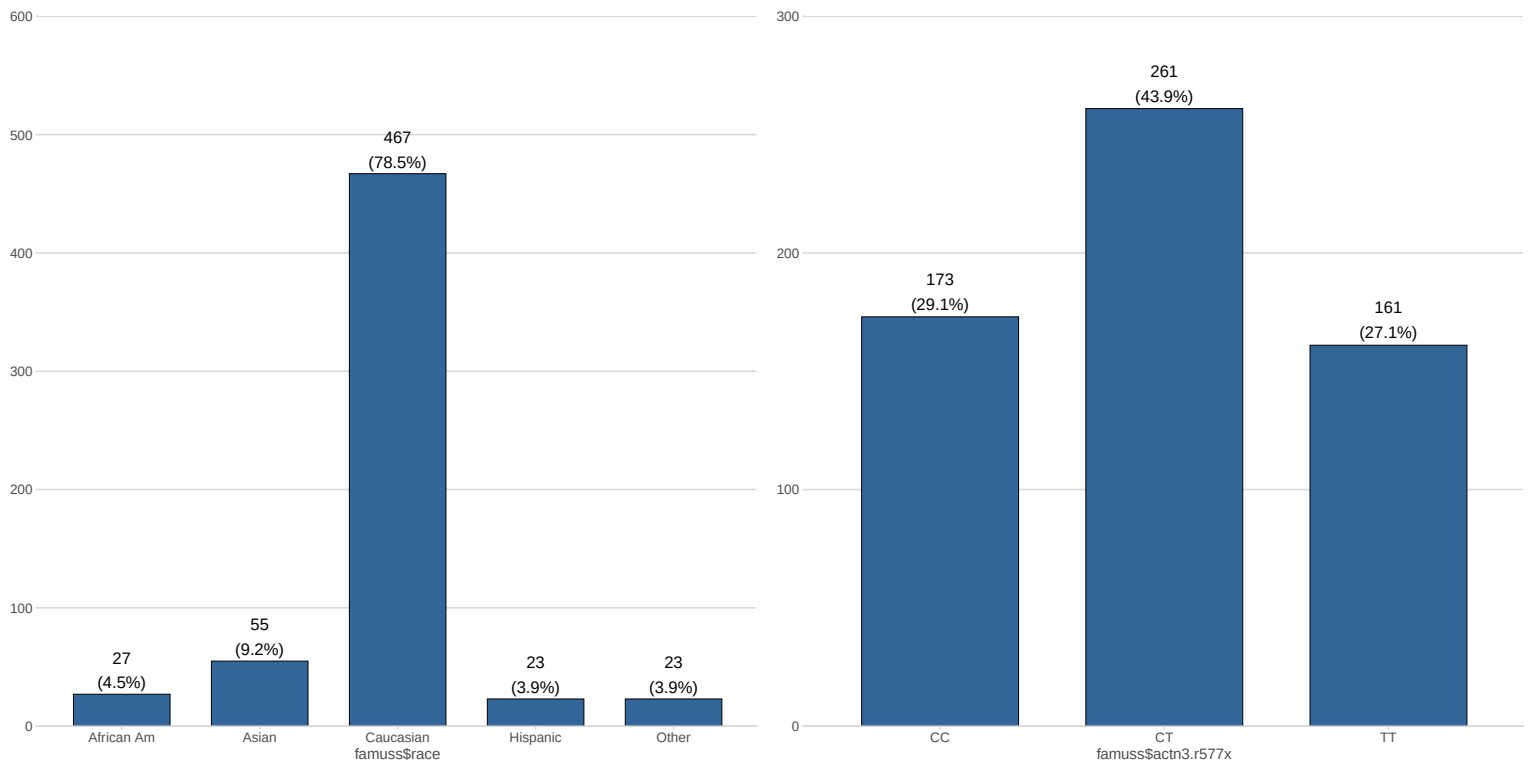
Marginal distributions of race and genotype

Given a contingency table, the frequency distribution of one of the variables is called its **marginal distribution**.

```
table(famuss$race) / nrow(famuss)

##
## African Am      Asian  Caucasian   Hispanic      Other
##      0.045      0.092      0.785      0.039      0.039

sjPlot::plot_frq(famuss$race)
sjPlot::plot_frq(famuss$actn3.r577x)
```

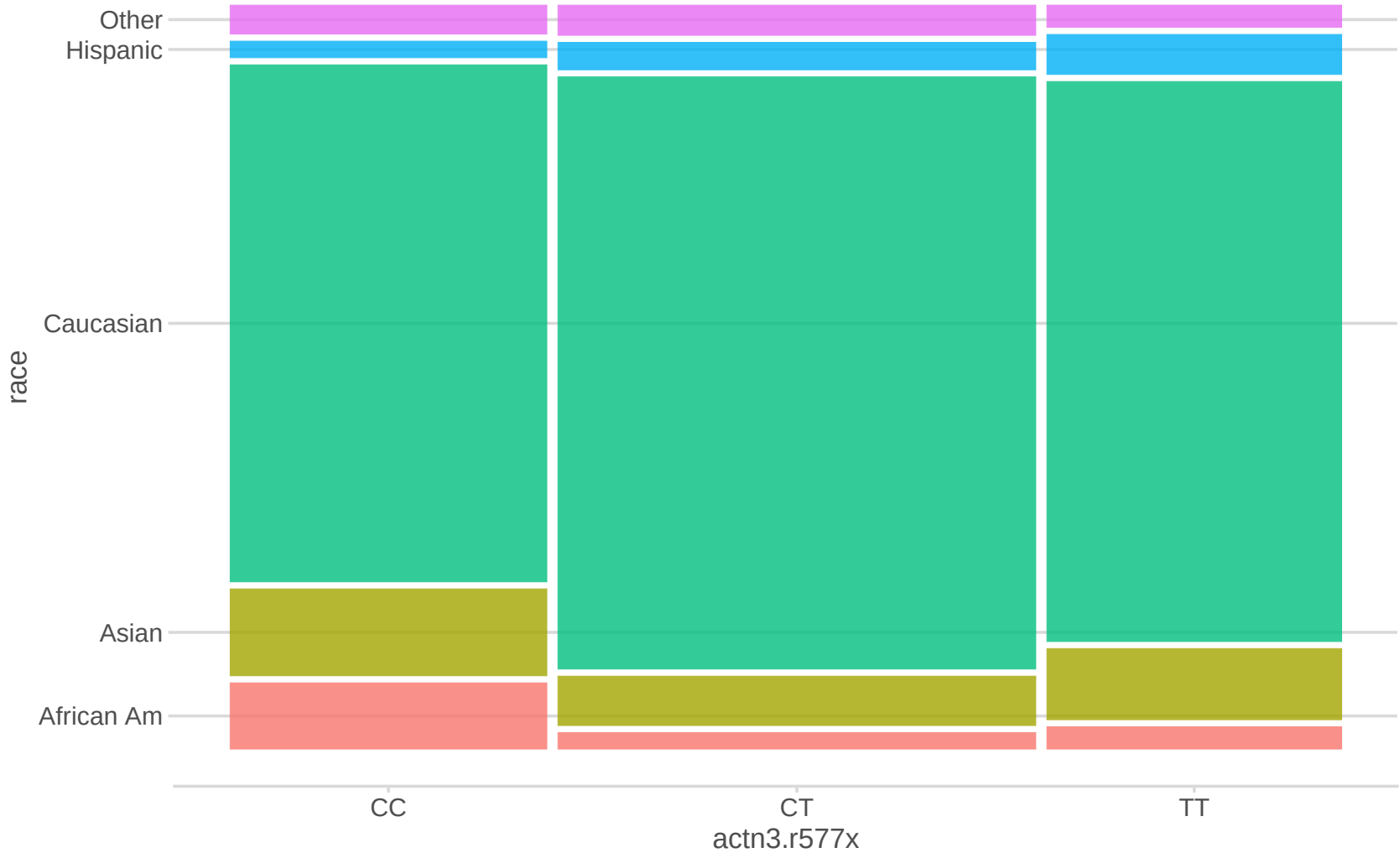


Mosaic plots

- A mosaic plot is a graphical display that allows you to examine the relationship among two or more categorical variables.
- The mosaic plot starts as a square with length one. The square is divided first into horizontal bars whose widths are proportional to the probabilities associated with the first categorical variable.
- Then each bar is split vertically into bars that are proportional to the conditional probabilities of the second categorical variable. Additional splits can be made if wanted using a third, fourth variable, etc.

Mosaic plots - race and genotype

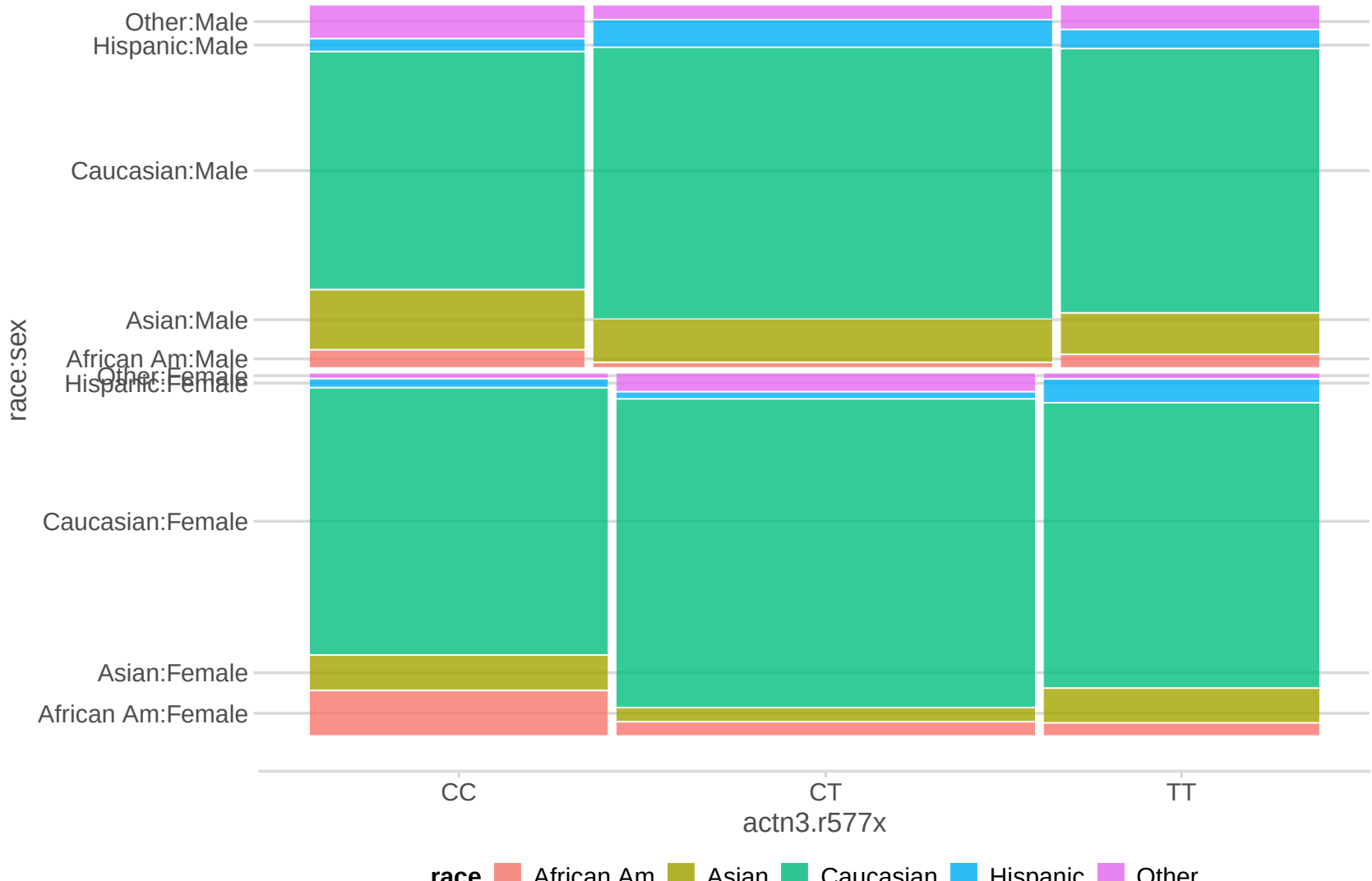
```
# devtools::install_github("haleyjeppson/ggmosaic")
pacman::p_load(ggmosaic)
ggplot(data = famuss) +
  geom_mosaic(aes(x = product(race, actn3.r577x),
                  fill = race))
```



race African Am Asian Caucasian Hispanic Other

Mosaic plots - race, genotype and sex

```
ggplot(data = famuss) +
  geom_mosaic(aes(x = product(race, actn3.r577x),
                  fill = race, conds = product(sex)),
                  divider = mosaic("v"))
```



Two numerical variables and the correlation coefficient

Two categorical variables and contingency tables

A numerical variable and a categorical variable

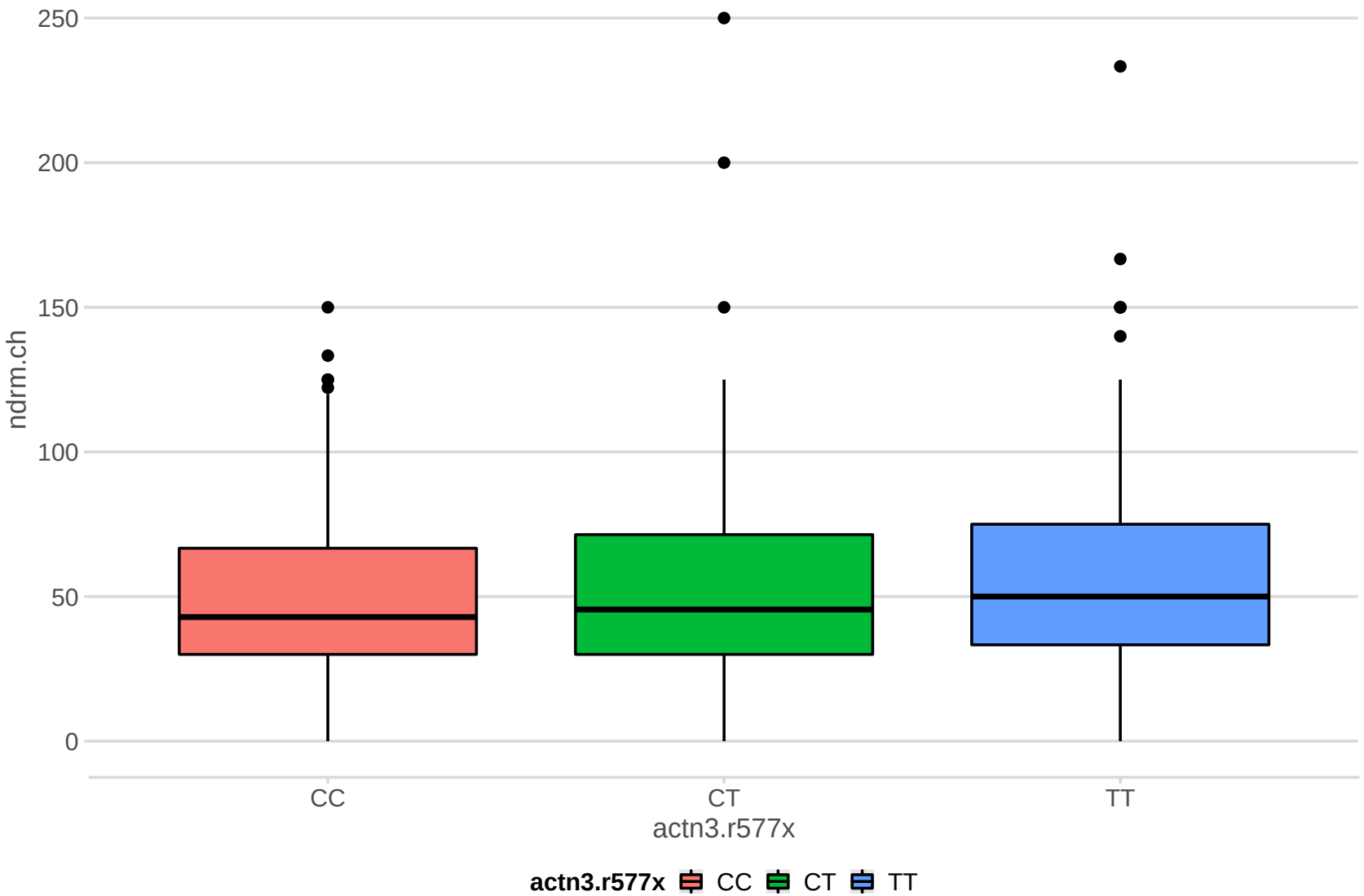
Summary

A numerical variable and a categorical variable

- *FAMuSS* was designed to study the relationship between genotype at the location *r577x* in the gene *ACTN3* and muscle strength.
- Muscle strength was assessed by the percent change in non-dominant arm strength after resistance training (`ndrm.ch`).
- What visualization would be a good choice to make this comparison?

A numerical variable and a categorical variable

```
ggplot(data = famuss, mapping = aes(x = actn3.r577x, y = ndrm.ch, fill = actn3.r577x)) +
  geom_boxplot()
```



Correlations

```
cor(famuss$actn3.r577x, famuss$ndrm.ch)

## Error in cor(famuss$actn3.r577x, famuss$ndrm.ch): 'x' must be numeric

cor(as.numeric(famuss$actn3.r577x), famuss$ndrm.ch, method = "pearson")

## [1] 0.1

cor(as.numeric(famuss$actn3.r577x), famuss$ndrm.ch, method = "kendall")

## [1] 0.077

cor(as.numeric(famuss$actn3.r577x), famuss$ndrm.ch, method = "spearman")

## [1] 0.098
```

Two numerical variables and the correlation coefficient

Two categorical variables and contingency tables

A numerical variable and a categorical variable

Summary

Summary of exploring data slides

- Two types of variables:
  - **Numeric**: Discrete, Continuous
  - **Categorical**: Ordinal, Nominal
- The collection of values for a numerical or categorical is called the distribution of that variable
- Measures of center include mean and median.
- Measures of spread include standard deviation, interquartile range
- Median and IQR are robust to outliers
- Histograms, boxplots, violin plots, and scatterplots are useful graphical summaries of numerical data, which can also be grouped by a categorical variable
- Bar plots, contingency tables, mosaic plots are useful summaries of categorical data

Summary of exploring data slides *continued*

- Correlation coefficient (r) quantifies the strength of a linear trend.
- The `multiple R-squared` in a simple linear regression output is equal to r^2.
- Transformation (e.g. log) can produce better linear associations for highly skewed data. But be careful about the interpretation!
- Given a contingency table, the frequency distribution of one of the variables is called its marginal distribution
- Conditional distributions show the distribution of one variable for just those cases that satisfy a condition on another variable
- See https://www.r-graph-gallery.com/ and https://www.data-to-viz.com/ for a collection of graphical displays

Session Info

```
R version 3.6.2 (2019-12-12)
Platform: x86_64-pc-linux-gnu (64-bit)
Running under: Pop!_OS 19.10

Matrix products: default
BLAS:   /usr/lib/x86_64-linux-gnu/openblas/libblas.so.3
LAPACK: /usr/lib/x86_64-linux-gnu/libopenblasp-r0.3.7.so

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

other attached packages:
 [1] ggmosaic_0.3.0      cowplot_1.0.0       openintro_2.0.0
 [4] usdata_0.1.0        cherryblossom_0.1.0 airports_0.1.0
 [7] oibiostat_0.2.0     NCStats_0.4.7       FSA_0.8.30
[10] forcats_0.5.0       stringr_1.4.0       dplyr_1.0.2
[13] purrr_0.3.4         readr_1.3.1         tidyr_1.1.2
[16] tibble_3.0.3        ggplot2_3.3.2.9000  tidyverse_1.3.0
[19] knitr_1.29

loaded via a namespace (and not attached):
 [1] nlme_3.1-143       fs_1.3.2           lubridate_1.7.4    RColorBrewer_1.1-2
 [5] insight_0.8.1      httr_1.4.1         backports_1.1.9    R6_2.4.1
 [9] sjlabelled_1.1.3   lazyeval_0.2.2     DBI_1.1.0          colorspace_1.4-1
[13] withr_2.2.0        tidyselect_1.1.0   emmeans_1.4.5      compiler_3.6.2
[17] performance_0.4.4  cli_2.0.2          rvest_0.3.5        pacman_0.5.1
[21] xml2_1.3.0         plotly_4.9.2       sandwich_2.5-1     labeling_0.3
[25] bayestestR_0.5.2   scales_1.1.1       mvtnorm_1.0-12     digest_0.6.25
[29] minqa_1.2.4        htmltools_0.5.0    pkgconfig_2.0.3    lme4_1.1-21
[33] dbplyr_1.4.2       highr_0.8          htmlwidgets_1.5.1  rlang_0.4.7
[37] readxl_1.3.1       rstudioapi_0.11    farver_2.0.3       generics_0.0.2
[41] zoo_1.8-7          jsonlite_1.7.0     sjPlot_2.8.3       magrittr_1.5
[45] parameters_0.5.0   Matrix_1.2-18      Rcpp_1.0.4.6       munsell_0.5.0
[49] fansi_0.4.1        lifecycle_0.2.0    stringi_1.4.6      multcomp_1.4-12
[53] snakecase_0.11.0   MASS_7.3-51.5      plyr_1.8.6         grid_3.6.2
[57] sjmisc_2.8.3       crayon_1.3.4       lattice_0.20-38    ggeffects_0.14.1
[61] haven_2.3.1        splines_3.6.2      sjstats_0.17.9     hms_0.5.3
[65] pillar_1.4.6       boot_1.3-24        estimability_1.3   effectsize_0.2.0
[69] codetools_0.2-16   reprex_0.3.0       glue_1.4.2         evaluate_0.14
[73] data.table_1.12.8  modelr_0.1.5       vctrs_0.3.4        nloptr_1.2.2.1
```