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 8, 2020

Summarizing relationships between two variables

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

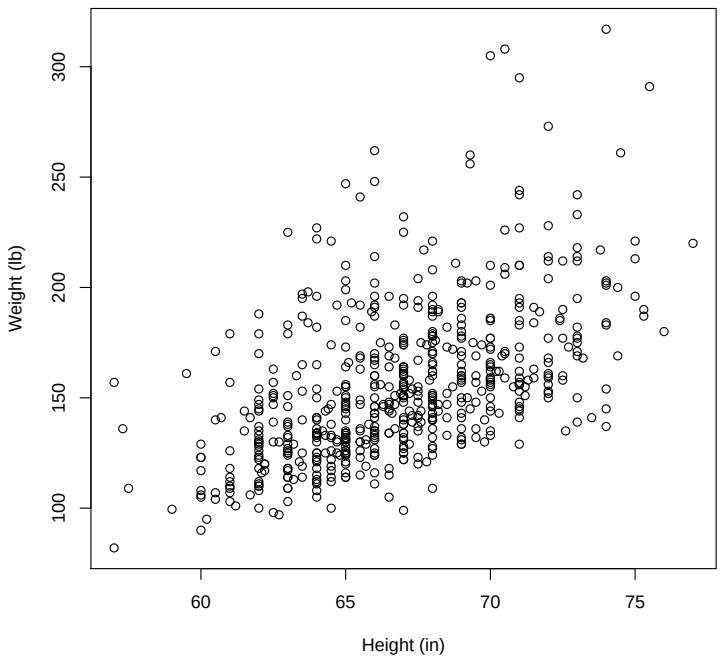
- Two numerical variables
- Two categorical variables
- One numerical variable and one categorical variable

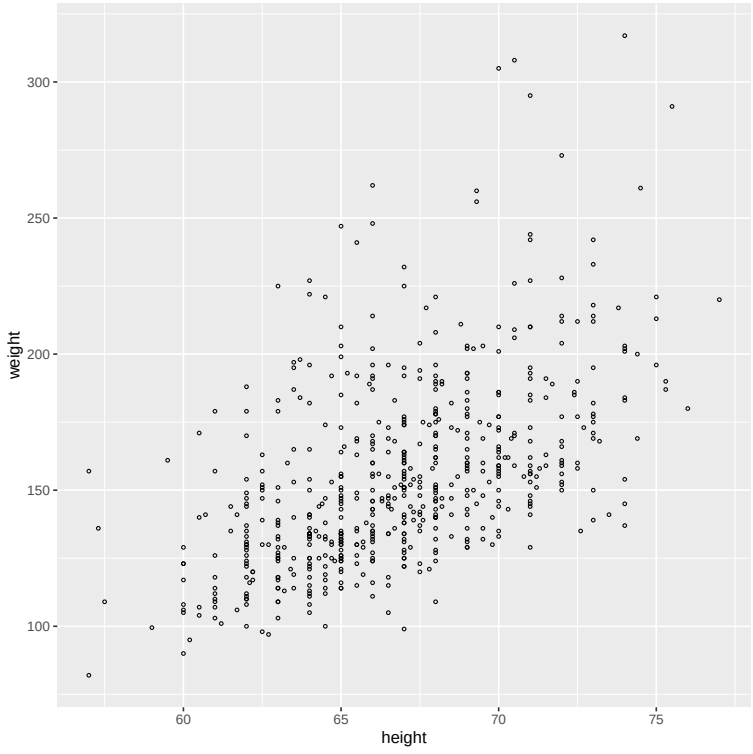
Scatterplots

```
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)
```





Correlation coefficient

- The correlation between two variables x and y is given by:

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

where $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ are the n paired values of x and y, and s_x and s_y are the sample standard deviations of the x and y variables, respectively.

Correlation coefficient

- The correlation between two variables x and y is given by:

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

  where $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ are the n paired values of x and y, and s_x and s_y are the sample standard deviations of the x and y variables, respectively.

- The correlation coefficient quantifies the strength of a **linear** trend.

Correlation coefficient

- The correlation between two variables x and y is given by:

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

  where $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ are the n paired values of x and y, and s_x and s_y are the sample standard deviations of the x and y variables, respectively.
- The correlation coefficient quantifies the strength of a **linear** trend.
- The correlation coefficient r takes on values between -1 and 1.

Correlation coefficient

- The correlation between two variables x and y is given by:

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

  where $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ are the n paired values of x and y, and s_x and s_y are the sample standard deviations of the x and y variables, respectively.
- The correlation coefficient quantifies the strength of a **linear** trend.
- The correlation coefficient r takes on values between -1 and 1.
- The closer r is to ± 1, the stronger the linear association.

Correlation coefficient

- The correlation between two variables x and y is given by:

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

  where $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ are the n paired values of x and y, and s_x and s_y are the sample standard deviations of the x and y variables, respectively.
- The correlation coefficient quantifies the strength of a **linear** trend.
- 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 ($r < 0$)

Correlation in R

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

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

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:

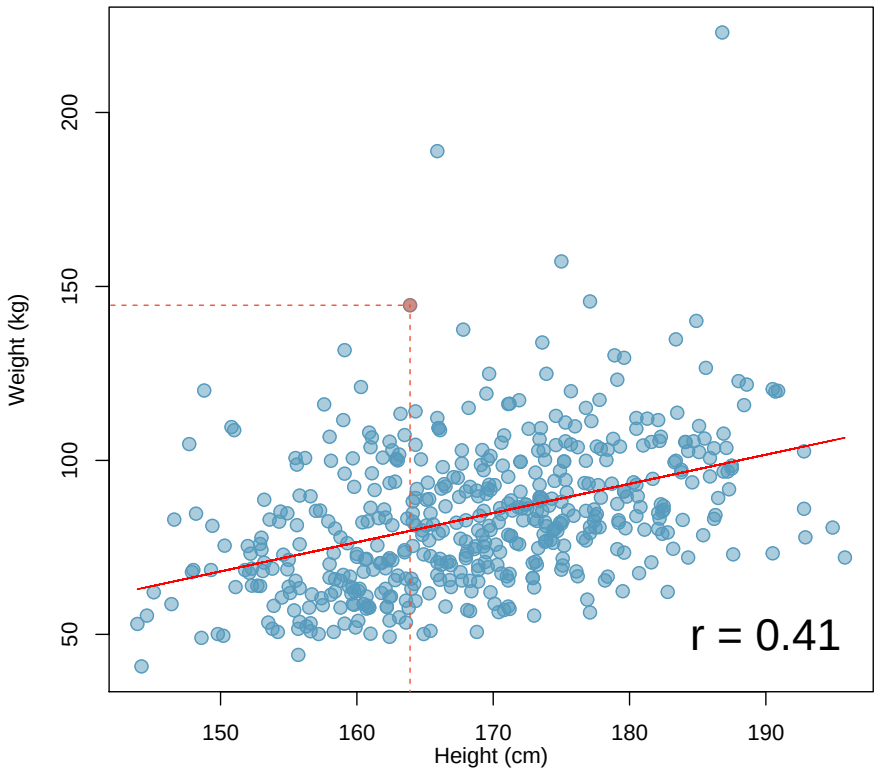
```
summary(lm(height ~ weight, data = famuss))
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  58.2952     0.5732   101.7   <2e-16 ***
## weight        0.0548     0.0036    15.2   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## 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
```

Another example: NHANES[2]

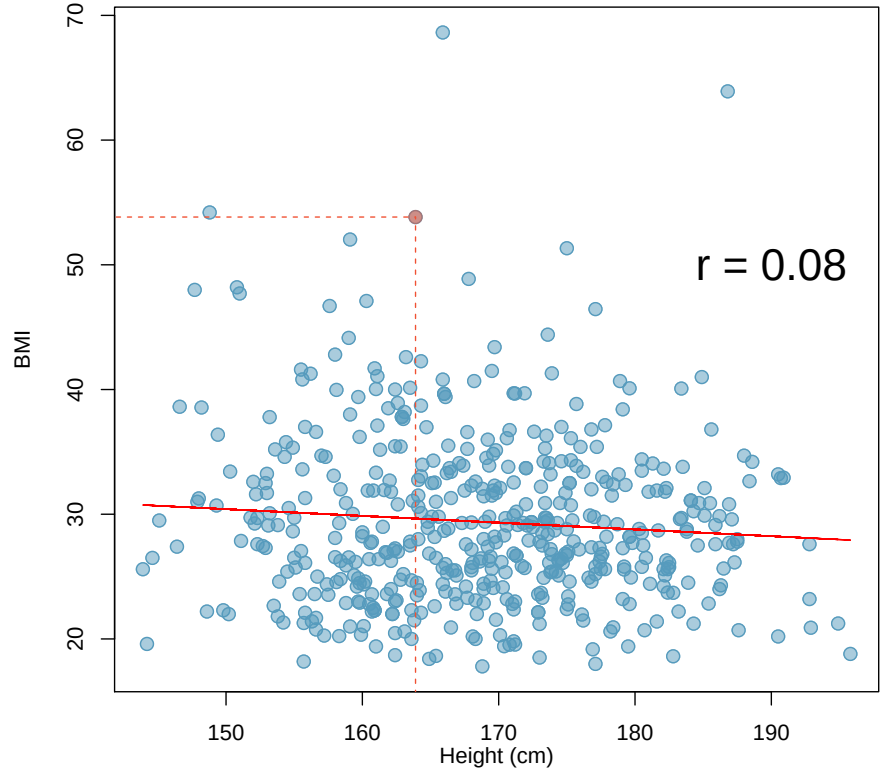
- The National Health and Nutrition Examination Survey (NHANES) consists of a set of surveys and measurements conducted by the US CDC to assess the health and nutritional status of adults and children in the United States.
- The following example uses data from a sample of 500 adults (individuals ages 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 $\times 703$.

Anscombe's quartet[3]

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

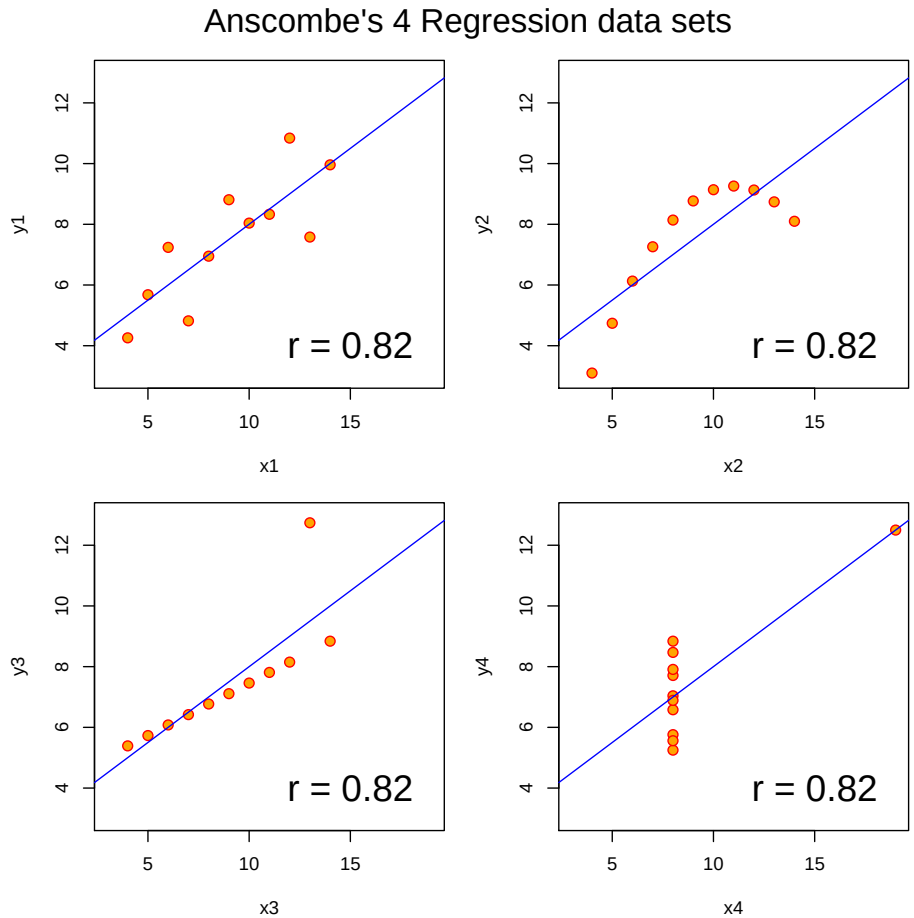
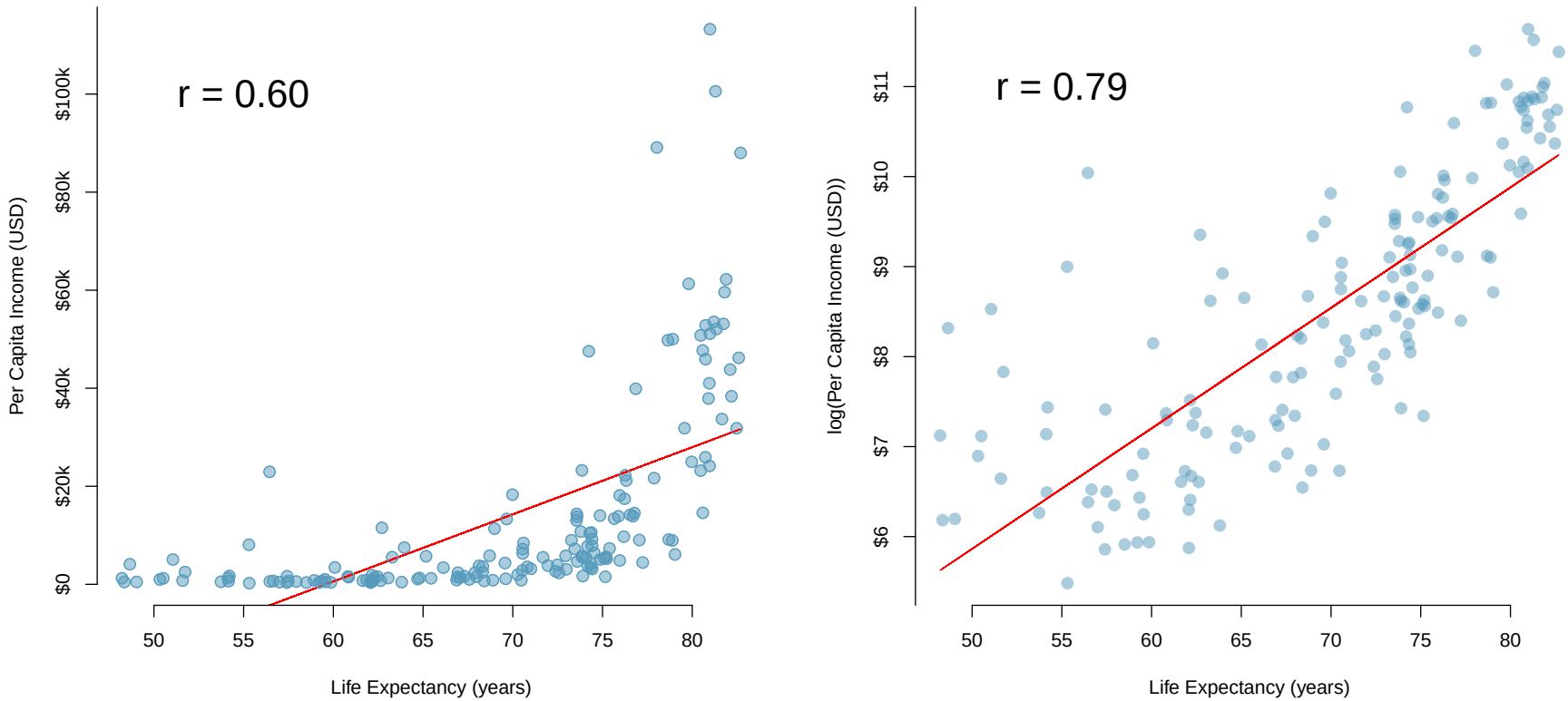


Figure: All four panels have the exact same linear correlation coefficient

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



(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

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
```

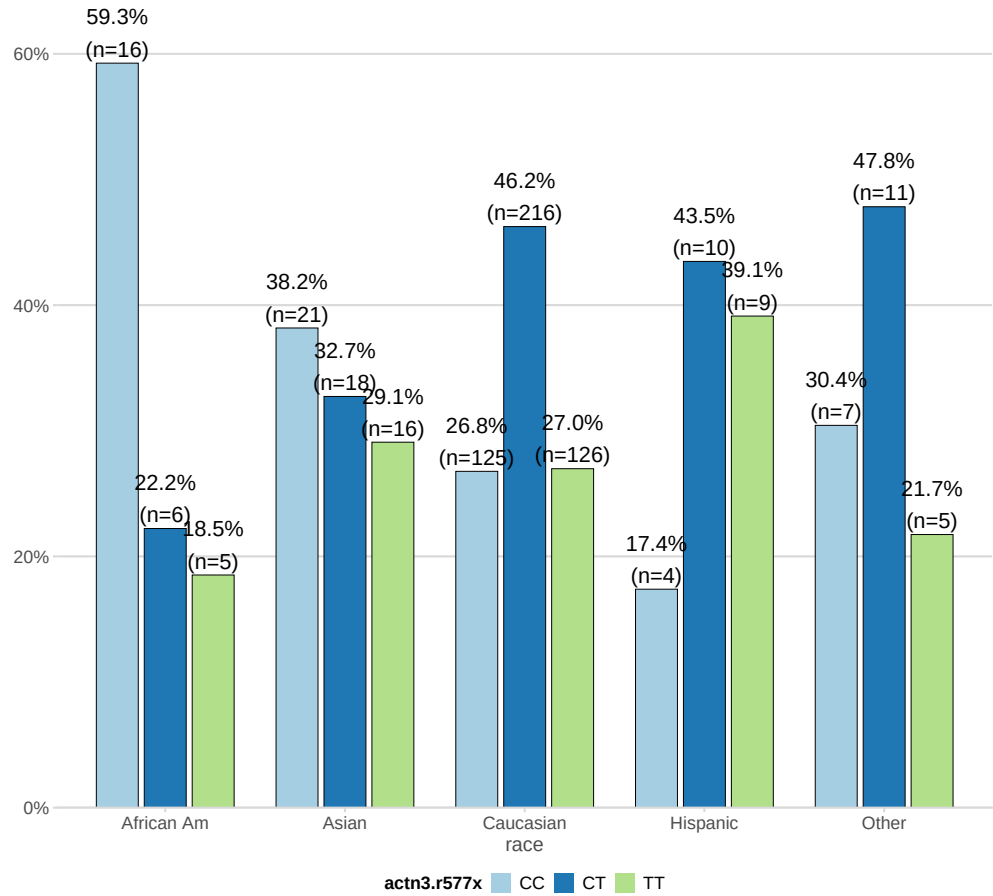
Conditional distribution of genotype *given* race

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")
```

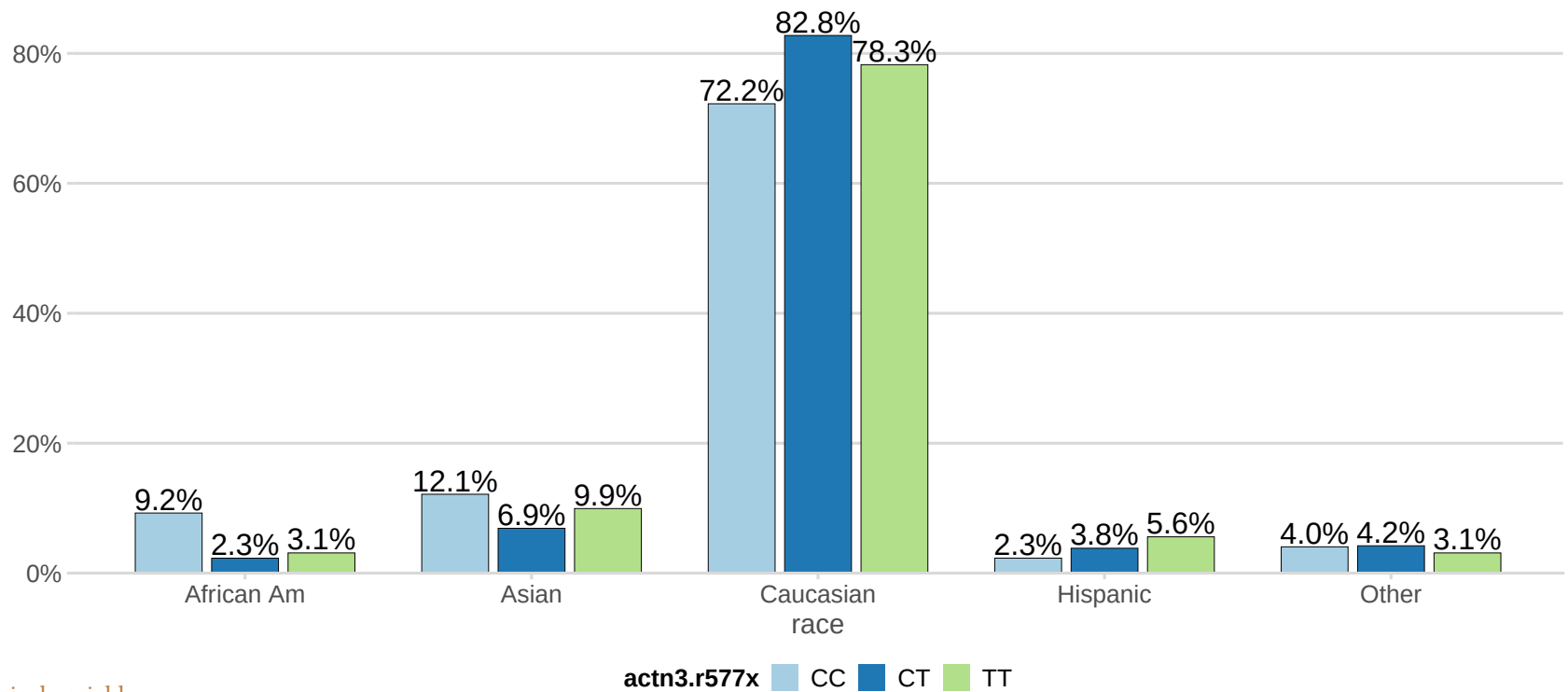


Conditional distribution of race *given* genotype

```
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)
```



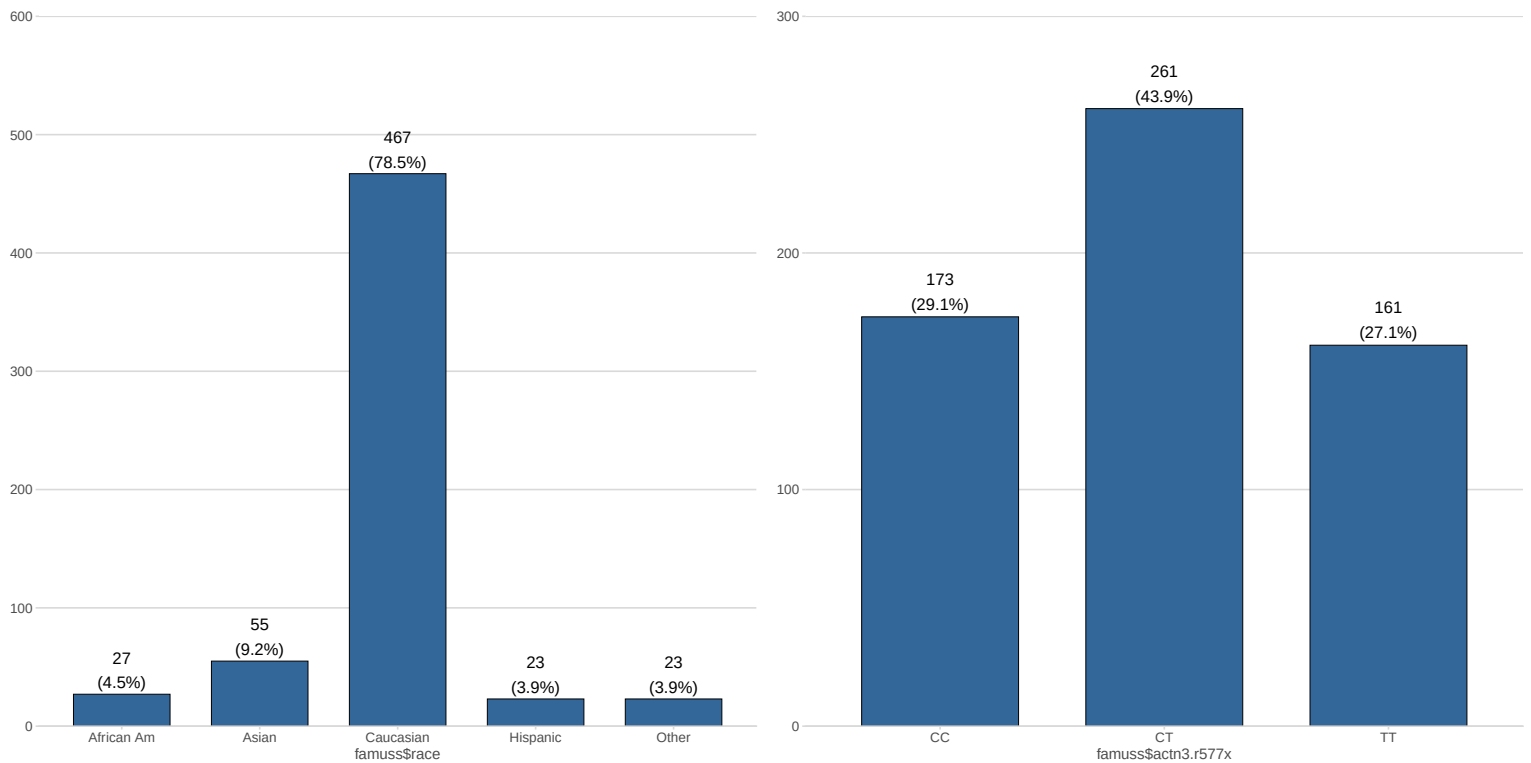
Marginal distributions of race and genotype

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)
```

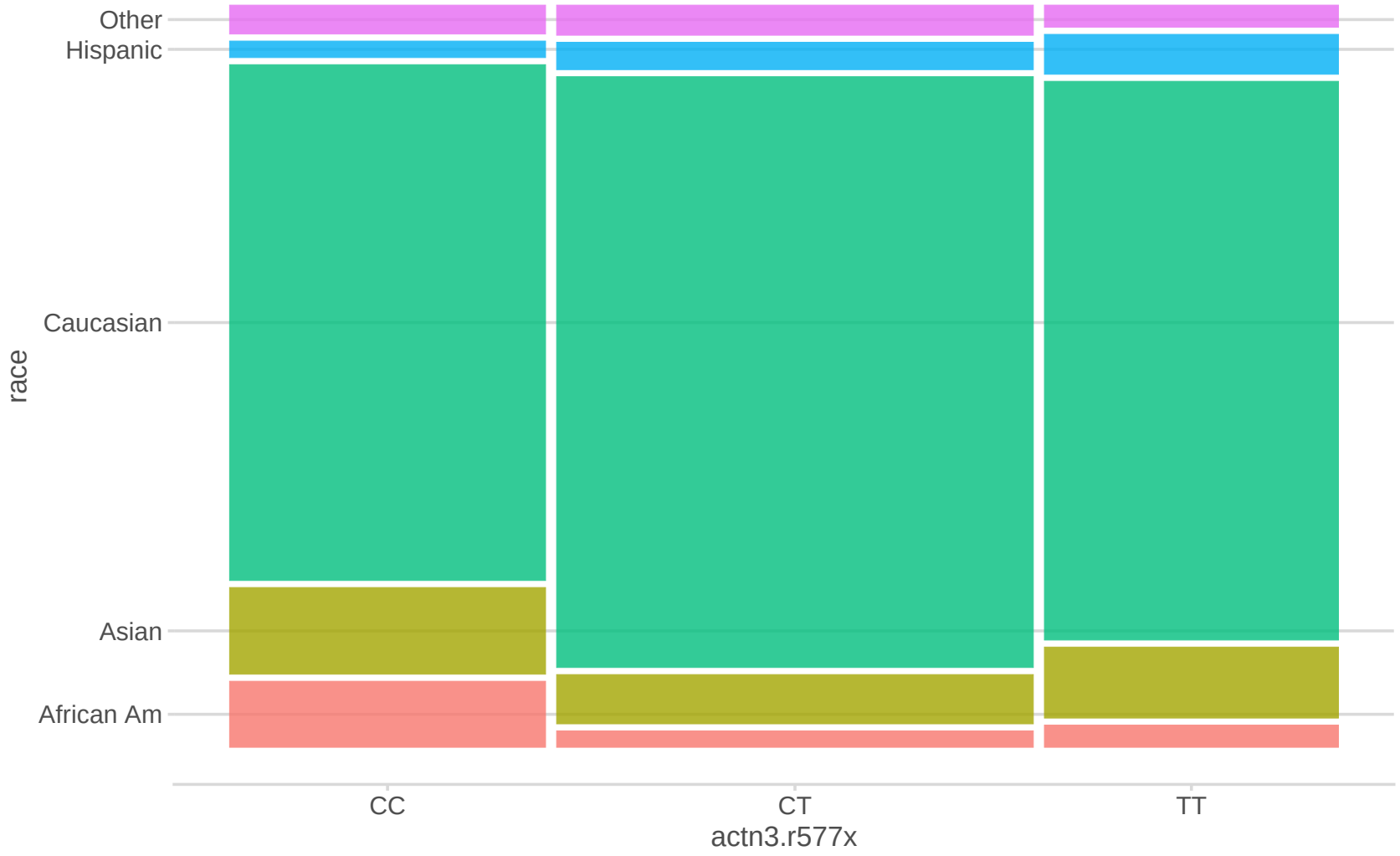


Mosaic plots

- 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.

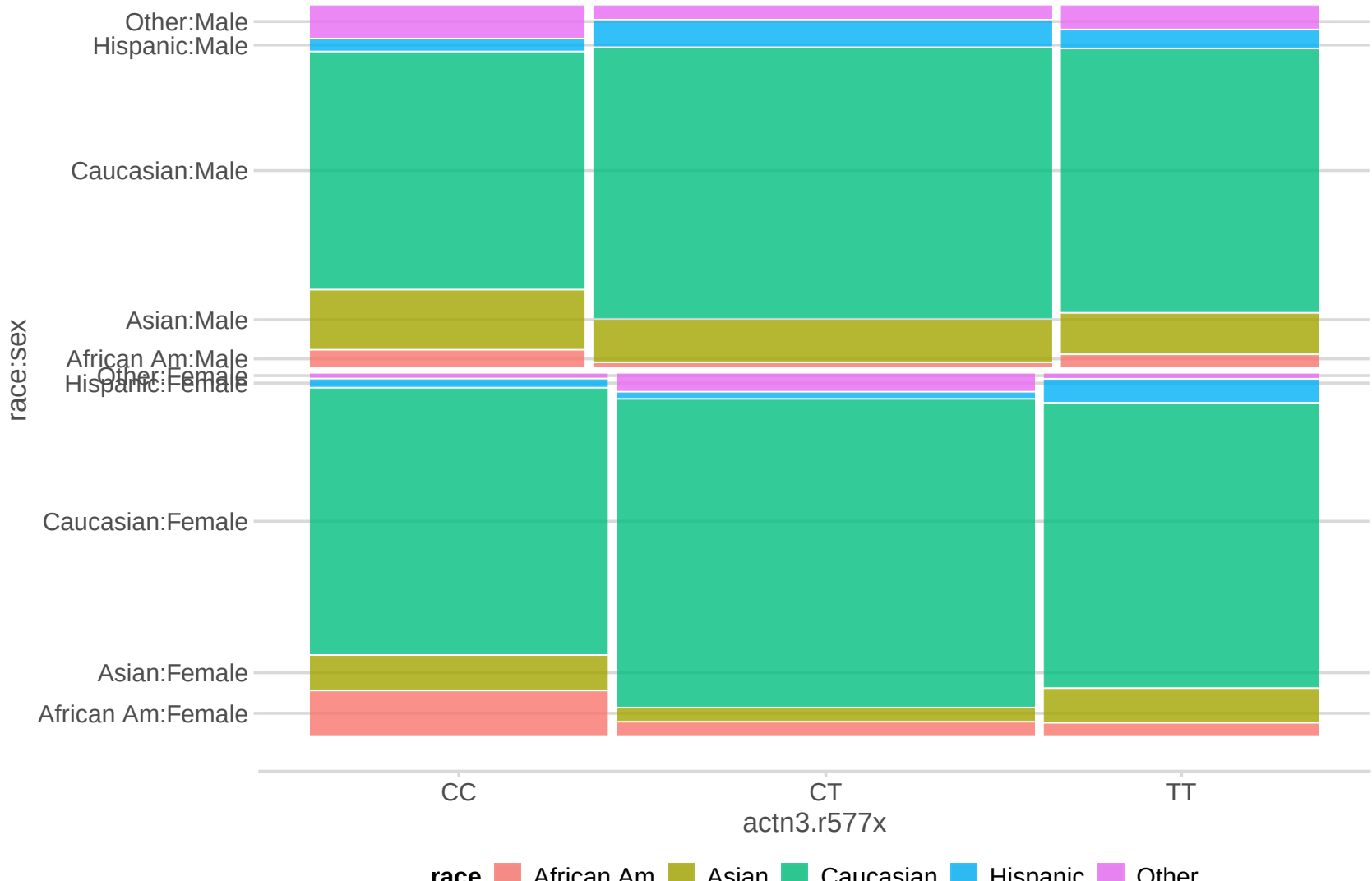
Mosaic plots - race and genotype

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



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"))
```

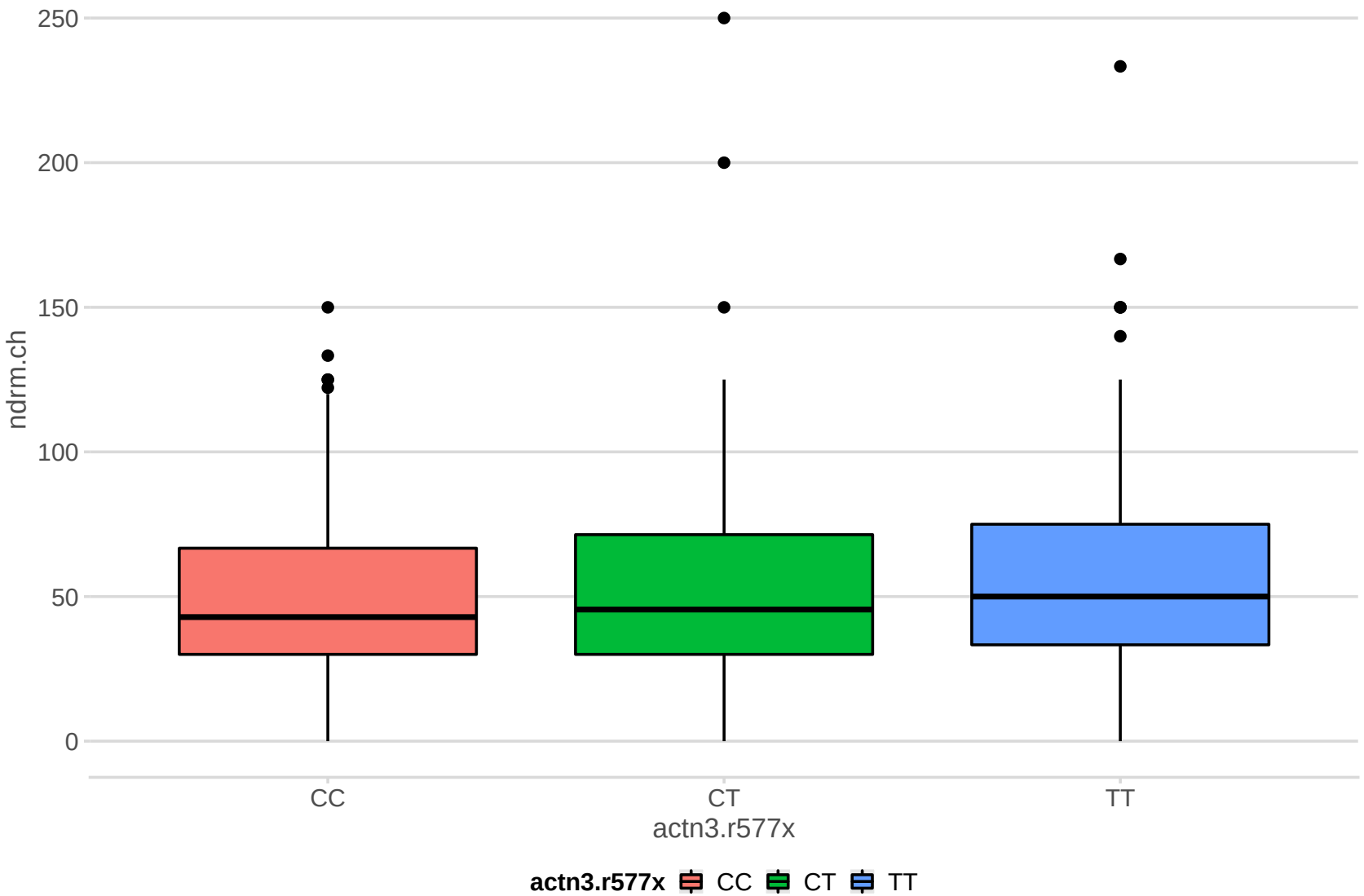


A numerical variable and a categorical variable

- *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()
```



Two numerical variables

Two categorical variables

A numerical variable and a categorical variable

Summary

Summary of exploring data slides

- Two types of variables:
  - **Numeric**: Discrete, Continuous
  - **Categorical**: Ordinal, Nominal

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

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

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!

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
```