

# Unit 1: Exploring Data

Statistics 102 Teaching Team

January 27, 2020

Data basics

Numerical data

Categorical data

Relationships between two variables

Case study: molecular cancer classification

## Data basics

## EXAMPLE: THE *FAMuSS* STUDY

*The Functional SNPs Associated with Muscle Size and Strength* (FAMuSS) study and data is introduced in *OI Biostat*, Section 1.2.2.

One goal of the study—examine the association of demographic, physiological and genetic characteristics with muscle strength.

- In simpler terms, study the “sports gene” *ACTN3*.

## FOUR ROWS FROM *FAMuSS* DATA MATRIX

*Ol Biostat* Table 1.6

sex	age	race	height	weight	actn3.r577x	ndrm.ch
Female	27	Caucasian	65.0	199	CC	40.0
Male	36	Caucasian	71.7	189	CT	25.0
Female	24	Caucasian	65.0	134	CT	40.0
Female	30	Caucasian	64.0	134	CC	43.8

## *FAMuSS* VARIABLES AND THEIR DESCRIPTIONS

<b>Variable</b>	<b>Description</b>
sex	Sex of the participant
age	Age in years
race	Recorded as African Am (African American), Caucasian, Asian, Hispanic, Other
height	Height in inches
weight	Weight in pounds
actn3.r577x	Genotype at the location r577x in the ACTN3 gene.
ndrm.ch	Percent change in strength in the non-dominant arm, comparing strength after to before training

# TYPES OF VARIABLES

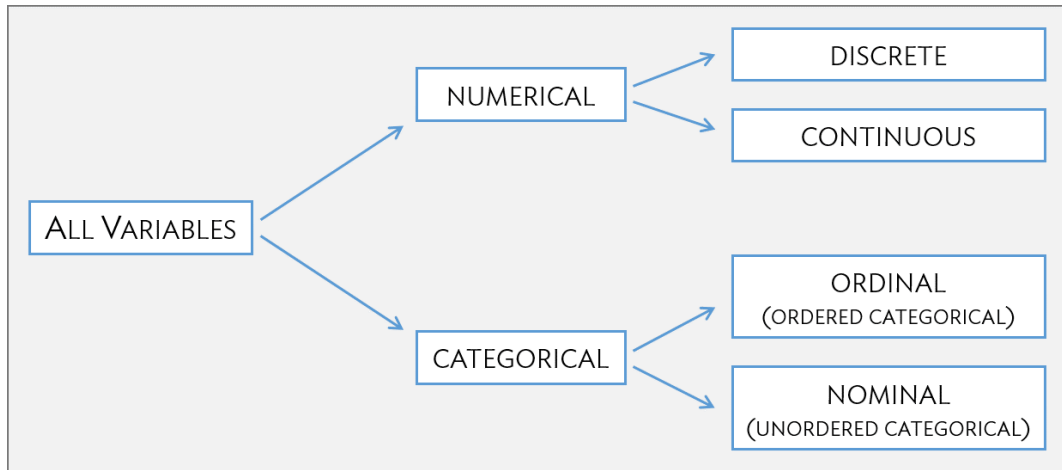
**Numerical variables** take on numerical values, such that numerical operations (sums, differences, etc.) are reasonable.

- Discrete: only take on integer values (e.g., # of family members)
- Continuous: can take on any value within a specified range (e.g., height)

**Categorical variables** take on values that are names or labels; the possible values are called the variable's *levels*.

- Ordinal: exists some natural ordering of levels (e.g., education)
- Nominal: no natural ordering of levels (e.g., gender)

# TYPES OF VARIABLES





# EXPLORING DATA WITH SIMPLE TOOLS

Techniques for exploring and summarizing data differ for numerical versus categorical variables.

Numerical and graphical summaries are useful for examining variables one at a time, but also for exploring the relationships between variables.

## Numerical data

# DISTRIBUTIONS AND SUMMARY MEASURES

The collection of values for a numerical, continuous variable (e.g., weight) is the *distribution* for that variable.

Numerical and graphical summaries convey characteristics of a distribution without listing all the values.

Important characteristics include. . .

- Center: where is the middle of the distribution?
  - Measures of center: mean, median
- Spread: how similar or varied are the values to each other?
  - Measures of spread: standard deviation, interquartile range

## MEASURES OF CENTER

The *sample mean* of a variable is the sum of all observations divided by the number of observations:

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

where  $x_1, x_2, \dots, x_n$  represent the  $n$  observed values in a sample.

The mean weight in famuss is 155.65 pounds:

```
mean(famuss$weight)
```

```
## [1] 155.6479
```

## MEASURES OF CENTER . . .

The *median* is the value of the middle observation in a sample.

If the number of observations is

- Odd, the median is the middle observation
- Even, the median is the average of the two middle observations

The median is the 50<sup>th</sup> percentile; 50% of observations lie below/above the median.

```
median(famuss$weight)
```

```
## [1] 150
```

## MEASURES OF SPREAD

The *standard deviation* measures (approximately) the distance between a typical observation and the mean.

- An observation's *deviation* is the distance between its value  $x$  and the sample mean  $\bar{x}$ :  $x - \bar{x}$ .
- The *sample variance*  $s^2$  is the sum of squared deviations divided by the number of observations minus 1.

$$s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n - 1},$$

where  $x_1, x_2, \dots, x_n$  represent the  $n$  observed values.

- The *standard deviation*  $s$  is the square root of the variance.

$$s = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n - 1}}$$

## MEASURES OF SPREAD: PERCENTILES/QUARTILES

The  $p^{th}$  percentile is the observation such that  $p\%$  of the remaining observations fall below this observation.

- The *first quartile* ( $Q_1$ ) is the 25<sup>th</sup> percentile.
- The *second quartile* ( $Q_2$ ), i.e., the median, is the 50<sup>th</sup> percentile.
- The *third quartile* ( $Q_3$ ) is the 75<sup>th</sup> percentile.

The *interquartile range* ( $IQR$ ) is the distance between the third and first quartiles.

$$IQR = Q_3 - Q_1$$

```
sd(famuss$weight)
```

```
## [1] 34.58999
```

```
IQR(famuss$weight)
```

```
## [1] 42
```

## ROBUST ESTIMATES

The median and IQR are called *robust estimates* because they are less likely to be affected by extreme values than the mean and standard deviation.

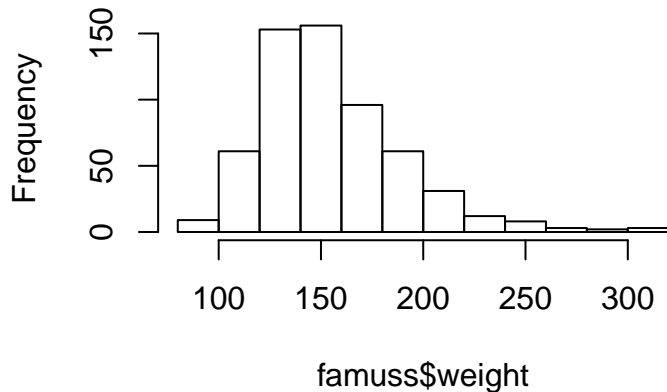
For distributions containing extreme observations, the median and IQR provide a more accurate sense of center and spread.



# HISTOGRAMS

```
hist(famuss$weight)
```

**Histogram of famuss\$weight**



## HISTOGRAMS . . .

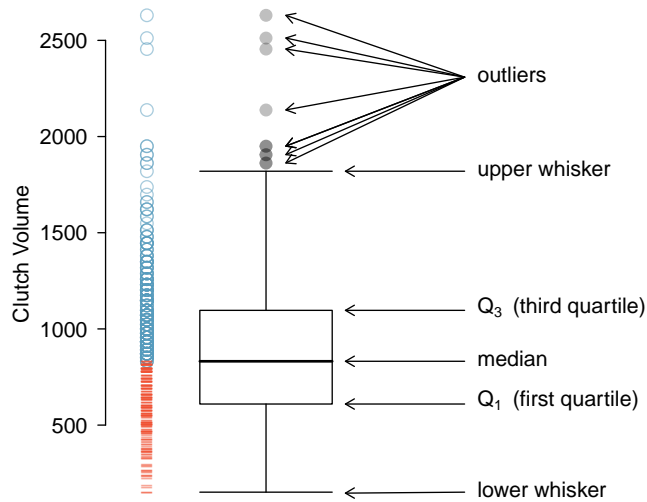
Histograms show important features of the shape of a distribution:

- Symmetry, or lack of it (skew)
- Minimum and maximum values
- Regions of high frequency (modes)

Histograms not so good for:

- Displaying median, quartiles
- Showing subtle skewing
- Identifying extreme values

## *Ol Biostat*, FIGURE 1.20, FROG DATA



# BOXPLOTS

A boxplot indicates the positions of the first, second, and third quartiles of a distribution in addition to potential **outliers**, observations that are far from the center of a distribution.

- Large outliers: values  $> Q_3 + (1.5 \times IQR)$
- Small outliers: values  $< Q_1 - (1.5 \times IQR)$

On a boxplot. . .

- The rectangle extends from the first quartile to the third quartile, with a line at the second quartile (median).
- Whiskers capture data between  $Q_1 - (1.5 \times IQR)$  and  $Q_3 + (1.5 \times IQR)$  ; whiskers must end at data points.
- Potential outliers shown with dots.

## Categorical data

# TABLES

A table for a single variable, a *frequency table* or *one-way table*, summarizes the distribution of observations among categories.

Based on the table, describe the distribution of genotype at the location *actn3.r577x* among the study participants.

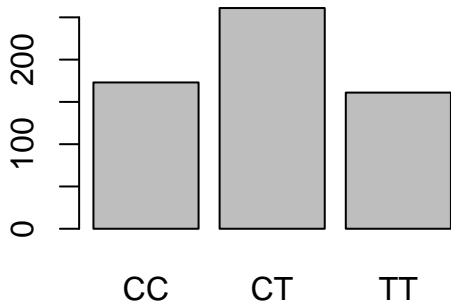
```
table(famuss$actn3.r577x)
```

```
##  
##   CC   CT   TT  
## 173 261 161
```

## BAR PLOTS FOR CATEGORICAL DATA

A bar plot is a common way to display a single categorical variable.

```
barplot(table(famuss$actn3.r577x))
```



## Relationships between two variables



# 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

## TWO NUMERICAL VARIABLES

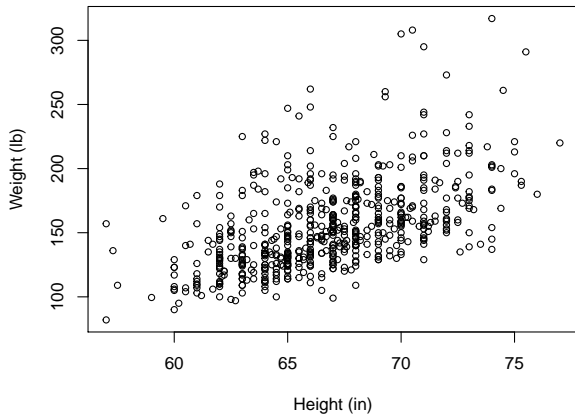
Two variables  $x$  and  $y$  are

- *positively associated* if  $y$  increases as  $x$  increases.
- *negatively associated* if  $y$  decreases as  $x$  increases.

Height and weight are positively associated.

## TWO NUMERICAL VARIABLES ...

```
plot(famuss$height, famuss$weight,  
     xlab = "Height (in)", ylab = "Weight (lb)", cex = 0.8)
```



## TWO NUMERICAL VARIABLES ...

Correlation is a numerical summary that measures the strength of a linear relationship between two variables.

- Introduced in *OI Biostat* Section 1.6.1; details in Ch. 6.
- The correlation coefficient  $r$  takes on values between -1 and 1.
- The closer  $r$  is to  $\pm 1$ , the stronger the linear association.

```
cor(famuss$height, famuss$weight)
```

```
## [1] 0.5308787
```

## TWO CATEGORICAL VARIABLES

A contingency table summarizes data for two categorical variables.

```
addmargins(table(famuss$race, famuss$actn3.r577x))
```

```
##  
##           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
```

## TWO CATEGORICAL VARIABLES ...

*#row proportions*

```
addmargins(prop.table(table(famuss$race, famuss$actn3.r577x), 1))
```

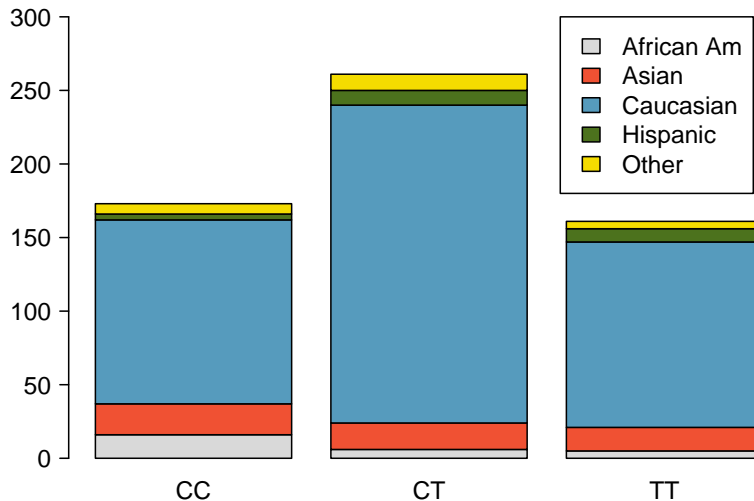
```
##
##           CC           CT           TT           Sum
## African Am 0.5925926 0.2222222 0.1851852 1.0000000
## Asian      0.3818182 0.3272727 0.2909091 1.0000000
## Caucasian  0.2676660 0.4625268 0.2698073 1.0000000
## Hispanic   0.1739130 0.4347826 0.3913043 1.0000000
## Other      0.3043478 0.4782609 0.2173913 1.0000000
## Sum        1.7203376 1.9250652 1.3545972 5.0000000
```

*#column proportions*

```
addmargins(prop.table(table(famuss$race, famuss$actn3.r577x), 2))
```

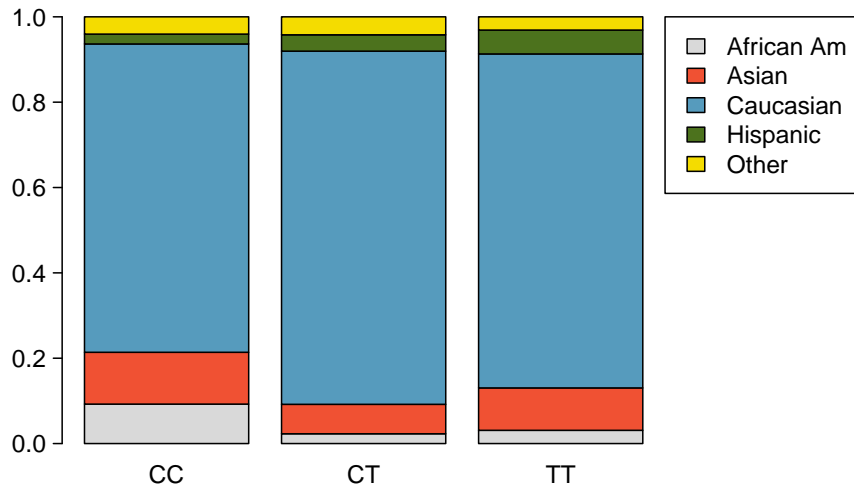
```
##
##           CC           CT           TT           Sum
## African Am 0.09248555 0.02298851 0.03105590 0.14652996
## Asian      0.12138728 0.06896552 0.09937888 0.28973168
## Caucasian  0.72254335 0.82758621 0.78260870 2.33273826
## Hispanic   0.02312139 0.03831418 0.05590062 0.11733618
## Other      0.04046243 0.04214559 0.03105590 0.11366392
## Sum        1.00000000 1.00000000 1.00000000 3.00000000
```

## TWO CATEGORICAL VARIABLES ...



*Ol Biostat* Figure 1.35a, segmented bar plot

## TWO CATEGORICAL VARIABLES ...



*Ol Biostat* Figure 1.35b, standardized segmented bar plot



## TWO CATEGORICAL VARIABLES ...

*Relative risk* (RR) is one way of summarizing data presented in a two-way table of study outcome by participant group.

More in Lab 1 ...

## A NUMERICAL VARIABLE AND A CATEGORICAL VARIABLE

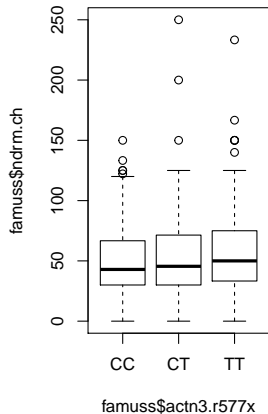
*FAMuSS* was designed to study the relationship between genotype at the location *r577x* in the gene *ACTN3* and muscle strength.

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 ...

```
boxplot(famuss$ndrm.ch ~ famuss$actn3.r577x)
```



## Case study: molecular cancer classification

## THE POTENTIAL VALUE OF GENOMIC DATA IN CANCER

The majority of cancers are diagnosed by an expert pathologist examining slides of malignant cells.

Can that be done more accurately by characterizing the genetic makeup of the malignancy?

- This is perhaps the major potential of genomic characterizations of tumors.

There are many forms of childhood leukemia.

- Acute myeloblastic leukemia (AML) and acute lymphoblastic leukemia (ALL) are the most common.
- AML is a cancer of the bone marrow, where white blood cells (lymphocytes) are produced.
- ALL is a cancer of the lymphocytes and is designated as B-cell (ALLB) or T-cell (ALLT).

## PROGNOSIS OF THE TWO CANCERS

The probability that a child diagnosed with ALL survives at least 5 years after the diagnosis is approximately 90%.

Approximately 65% of children diagnosed with AML survive at least 5 years.

The diagnosis of leukemia type determines the therapy that will be given to the child, and the successful treatments for ALL and AML are different.

In 1999, Todd Golub from the Dana-Farber and the Broad Institute examined the possibility of classifying leukemia through using a genetic analysis of a blood sample.

## ANALYZING THE GOLUB DATA

We can re-analyze the Golub data using tools from graphical and numerical summaries.

Our analysis will not be identical to the Golub analysis, but will be similar in spirit.

The tools are straightforward. . .

- Thinking through the problem and assembling the tools is the hard part.
- The process is more important than the final recipe.

## GENE EXPRESSION (DETAILS IN *Ol Biostat*)

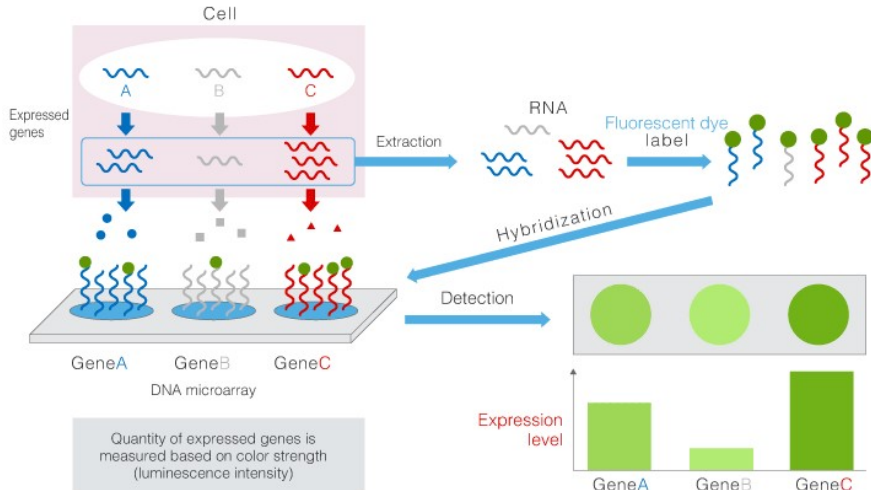
- The genetic code stored in DNA contains the information for producing the proteins that determine an organism's phenotype.
- Genes that are transcriptionally active (i.e. turned “on”) are transcribed into messenger RNA (mRNA) that gets translated into proteins.
- Genes can be switched on or off, and expressed at varying levels. Variations in gene expression produce the range of physical, biochemical, and developmental differences in cells and tissues.
- Quantifying the amount of RNA produced in a cell allows for a measure of gene expression.
- The transcriptome, or expression profile, is the complete set of RNA transcripts produced by the genome in a cell or set of cells.



## MICROARRAYS (DETAILS IN *Ol Biostat*)

- Microarray technology is based on hybridization between two DNA strands, in which complementary nucleotide sequences specifically pair together.
- The mRNA from a sample is converted into complementary-DNA (cDNA), labeled with a fluorescent dye, and added to the microarray.
- When cDNA from the sample encounters complementary DNA probes, the two strands will hybridize, allowing the cDNA to adhere to specific spots on the slide.
- When the chip is illuminated and scanned, the intensity of fluorescence detected at each spot corresponds to the amount of bound cDNA.
- DNA microarrays do not directly quantify gene expression levels or quantity of mRNA present in a sample.
- The fluorescence intensity data only provide a relative measure of gene expression, showing which genes on the chip seem to be more or less active in relation to each other.

# MICROARRAYS



# THE GOLUB CLINICAL DATA

Demographic variables described in *Ol Biostat* Table 1.54:

Variable	Description
Samples	Sample or chip number. The material from each patient was examined on a separate chip and experimental run.
BM.PB	Type of patient material. BM denotes bone marrow; PB denotes a peripheral blood sample.
Gender	F for female, M for male.
Source	Hospital where the patient was treated.
tissue.mf	A variable showing the combination of type of patient material and sex of the patient. BM:f denotes bone marrow from a female patient, etc.
cancer	The type of leukemia; aml is acute myeloblastic leukemia, allB is acute lymphoblastic leukemia which started in B-cells (cells that mature into plasma cells) origin, and allT is acute lymphoblastic leukemia with T-cell origin (T-cells are a type of white blood cell).

# THE GOLUB EXPRESSION DATA

The expression data is contained in the last 7,129 columns.

Each column is a variable with a name corresponding to the name of the probe on the microarray.

The expression levels record fluorescence intensity for each gene.

- The intensity levels have no inherent biological meaning.
- Data have been normalized to adjust for variability between the separate arrays used for each patient.

## SELECTED VARIABLES AND COLUMNS FROM GOLUB DATA

*Ol Biostat* Table 1.40

Samples	Gender	cancer	AFFX-BioB-5_at	AFFX-BioB-M_at	AFFX-BioB-3_at
39	F	allB	-1363.28	-1058.59	-541.47
40	F	allB	-796.29	-1167.10	7.54
42	F	allB	-679.14	-1069.83	-690.30
47	M	allB	-1164.40	-1109.94	-990.13
48	F	allB	-1299.65	-1402.00	-1077.54

# ANALYZING THE GOLUB LEUKEMIA DATA

We will do an analysis in class using some of the simple but surprisingly powerful ideas behind numerical and graphical summaries.

The goal of the Golub study was to develop a procedure for distinguishing between AML and ALL based only on the gene expression levels of a patient. There are two major issues to be addressed:

1. Which genes are the most informative for making a prediction?
2. What is a workable strategy for predicting leukemia type from expression data for a specific set of genes?

## STARTING SMALL...

##	cancer	A	B	C	D
## 69	allB	39307.96	35232.401	41170.76	35792.79
## 67	allT	32281.88	41432.024	59328.51	49608.14
## 55	allB	47429.94	35568.928	56074.96	42857.78
## 56	allB	25533.87	16983.749	28056.75	32693.92
## 59	allB	35960.55	24191.746	27637.90	22240.75
## 52	aml	46177.95	6189.465	12557.24	34485.41
## 53	aml	43790.70	33661.825	38380.30	29758.25
## 51	aml	53420.05	26109.245	31427.20	23809.70
## 50	aml	41241.59	37589.773	47325.77	30099.36
## 54	aml	41300.57	49198.412	66026.10	56248.62