The history of statistics has its roots in biology

Sir Francis Galton

Inventor of fingerprints,
study of heredity of quantitative traits

Regression & correlation

Karl Pearson

Polymath

Studied genetics

Correlation coefficient
χ^2 test
Standard deviation

Sir Ronald Fisher

The Genetical Theory of Natural Selection
Founder of population genetics

Analysis of variance
likelihood
P-value
randomized experiments
multiple regression
etc., etc., etc.

Graphics

Chapter 2

Type of Data

- Categorical Variables

  (also known as Class variables, or Nominal variables)

- Numerical Variables (or Quantitative Variables)
  - Numerical variables are either continuous or discrete.

Categorical variables

- Genotype (e.g., AA, AG, GG)
- Drug treatment (e.g. aspirin vs. ibuprofen)
- Province
- Survival (i.e., live or die)

Numerical variables

- Height
- Weight
- Tail length
- Dose (e.g., in micrograms/gram)
- Longevity (i.e., number of years)

Discrete vs. Continuous

Can be counted

- Number of limbs
- Number of offspring
- Number of petals

Can be measured

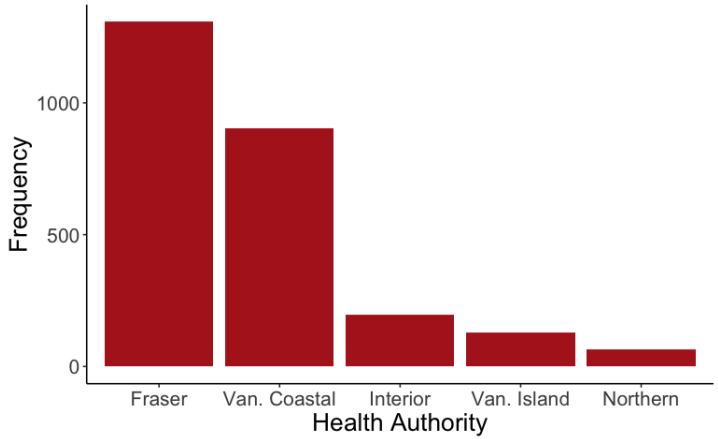
- Arm length
- Height
- Weight
- Age

Graphing categorical variables

Frequency table showing cases of COVID by BC Health Authority on June 2, 2020

Health Authority	Frequency
Fraser	1307
Vancouver Coastal	904
Interior	195
Vancouver Island	127
Northern	64

Bar graph



Graphing numerical variables

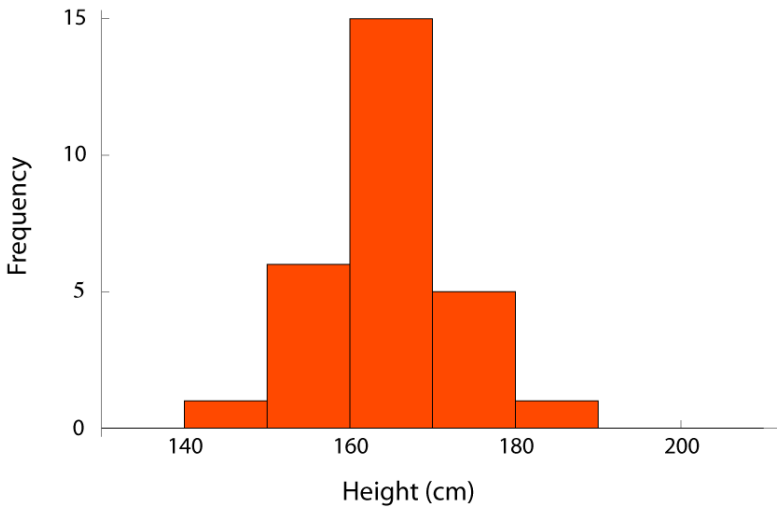
Heights of biostats students (cm)

165	168	163	173	170
163	170	155	152	190
170	168	142	160	154
165	156	177	173	165
165	175	155	166	168
165	180	165		

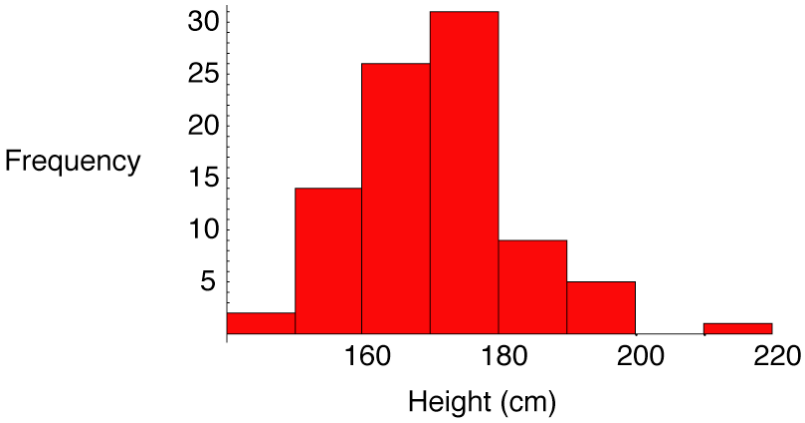
Frequency table

Height Group	Frequency
141-150	1
151-160	6
161-170	15
171-180	5
181-190	1

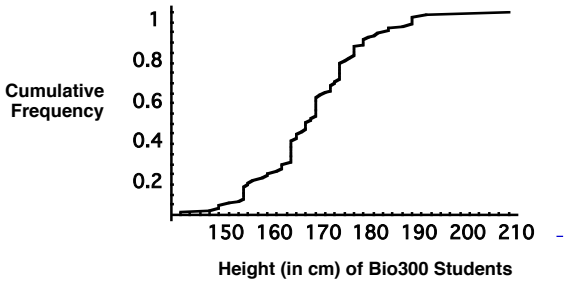
Histogram



Height histogram with more data

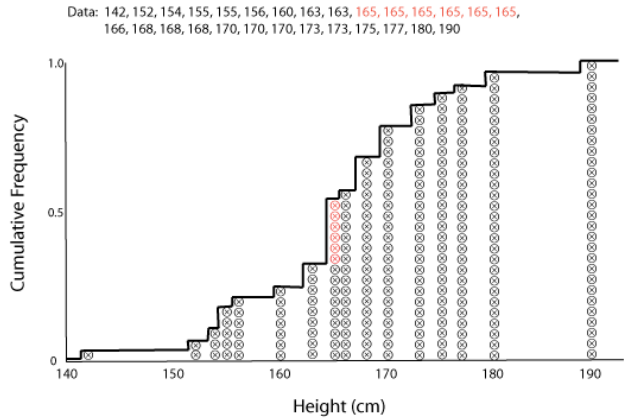


Cumulative Frequency Distribution



The cumulative frequency of a value is the proportion of individuals equal to or less than that value.

Making a CDF



15/28 (= 0.536) of the data are equal to or less than 165cm.

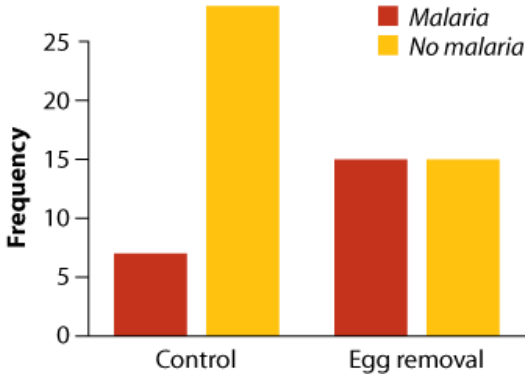
Associations between two categorical variables

Association between reproductive effort and avian malaria

Table 2.3A. Contingency table showing incidence of malaria in female great tits subjected to experimental egg removal.

	control group	egg removal group	row total
malaria	7	15	22
no malaria	28	15	43
column total	35	30	65

Grouped Bar Graph

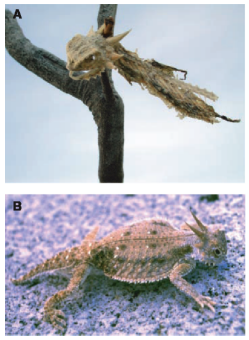


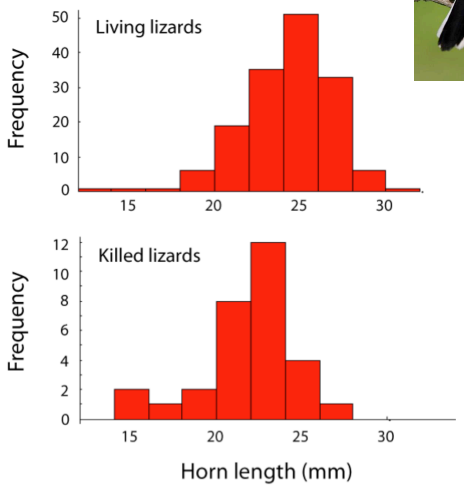
Mosaic plot

Associations between categorical and numerical variables

Multiple histograms

Shrike



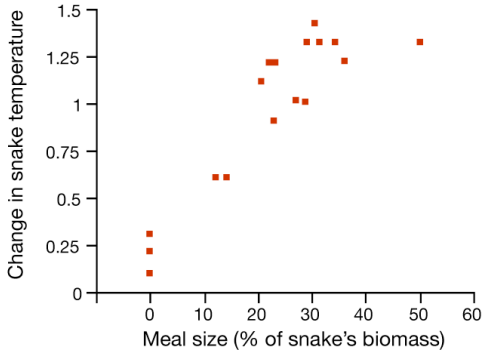


Young, K. V., E. D. Brodie Jr., and E. D. Brodie III. 2004. How the horned lizard got its horns. Science 304:65.

www.netcore.ca/~peleetom

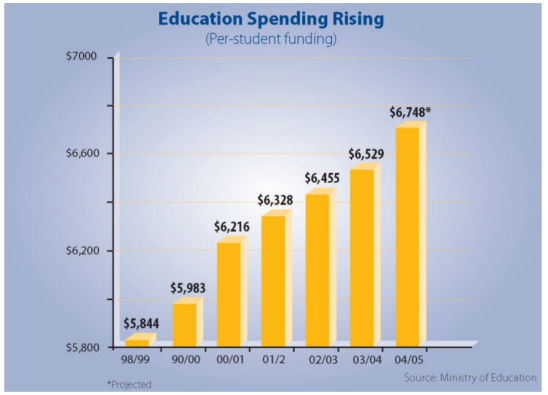
Associations between two numerical variables

Scatter plot

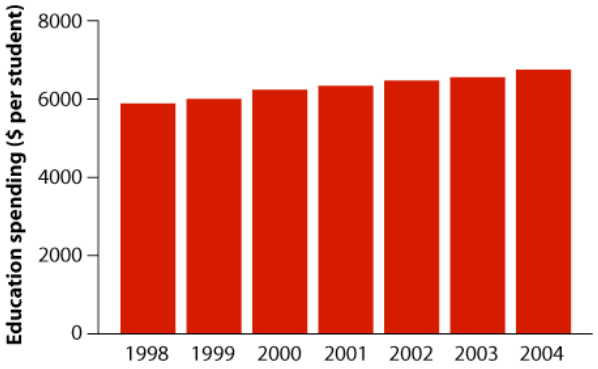


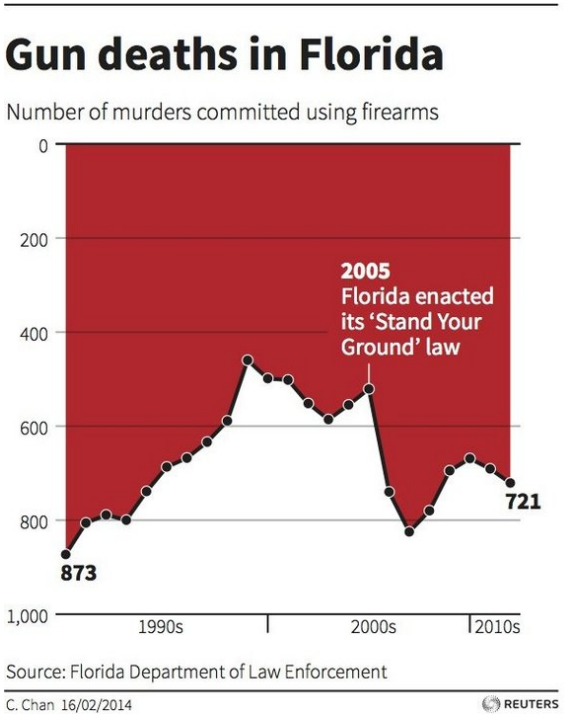
Tattersall et al. (2004) *Journal of Experimental Biology* 207:579-585

Don't mislead with graphics



Better representation of truth





Source: http://www.livescience.com/45083-misleading-gun-death-chart.html (Thanks, Phoebe!)

Summary: Graphical methods for frequency distributions

Type of Data	Method
Categorical data	Bar graph
Numerical data	Histogram Cumulative frequency distribution

Summary: Associations between variables

	Explanatory variable	
Response variable	**Categorical**	**Numerical**
Categorical	Contingency table Grouped bar graph Mosaic plot	
Numerical	Multiple histograms Cumulative frequency distributions	Scatter plot

Graphics in R

Bar graphs:

```
ggplot(data = BCCOVIDData, aes(x = HealthAuthority)) +
    geom_bar(stat = "count")
```

Histograms:

```
ggplot(classHeightDataFull, aes(x=height)) +
      geom_histogram(binwidth=5)
```

Scatterplots:

```
ggplot(data = snakeTempData,
      aes(x = mealSize, y = tempChange)) +
      geom_point()
```