Statistics 371, lecture 3

Cécile Ané

Spring 2012

Outline

Basic information

Objectives:

Introduction to modern statistical practice.

Understanding of the concepts, along with applications.

How should I collect my data?
Which method should I apply to my data?
How do I interpret the results?

Basic information

Read the syllabus: it's all in there.

`http://www.stat.wisc.edu/courses/st371-ane`
Refresh your browser!

`https://learnuw.wisc.edu/` for grades and blog.

Instructor & TAs:

Cecile Ane, `ane@stat.wisc.edu`

Dongguy Kim (331, 332, 333), `kimd@stat.wisc.edu`

Yi Liu (334), `liuyi@stat.wisc.edu`

Text: *The Analysis of Biological Data*, by Whitlock & Schluter.

Section switching: Come see me at end of lecture.

Homework assignments

Weekly.

Posted on Thursdays, due following Thursday in lecture.

If late: penalized except under extenuating circumstances.

Solution handouts will be posted.

should be well organized and **neat**.

Academic honesty: Talk to each other, discuss homework problems. Great! very stimulating! THEN, when writing up your assignment, do it all by yourself.

Exams and grading

Homework	weekly	15%
First midterm	February 28	15%
Second midterm	April 10	15%
Group project	report due May 1	15%
Final exam	May 18	40%

Second midterm + final exam will be **open book**.

If any **religious conflict** with exam dates: let me know within 1 week.

Grades: on **Learn@UW** 2-3 days after exams.

Group project

Goal: reduce weight of in-class exams; provide authentic experience.

Analyze real data,

Draw scientific conclusions, write up a report,

Evaluate reports from 2 other groups.

For **honors**: oral presentation of report.

All details in separate handout.

Computing

Modern statistics: need statistical software. We will use R.

- Free, easy to install,
- R tutorial on course webpage,
- first discussion (this week) will focus on R: installation, reading in data, plotting data. **Bring your laptop**.
- Use of R required in assignments, basic knowledge and interpretation of R outputs required in exams.

Get involved

Ask your questions! Feedback always welcome.

Sometimes I will ask you questions. Never trick questions.

Use the **discussion forum** / blog on Learn@UW: ask and answer questions.

Outline

Anthrax and vaccine experiment

1881, Pasteur. Reponse of sheep to Anthrax (bacterial disease, affects skin and lungs). All 48 animals were inoculated with a virulent culture.

	vaccinated	control
survived	24	0
died	0	24
% survival	100%	0%

No variation. Scientific conclusion: the vaccine works! categorical data: 2 categories. No numerical value.

What if...

Even without variation, need for statistical analysis to conclude that 24 sheep/group is enough evidence.

Comparing 2 drugs on mice

Success = survival for more than one week.

	drug A	drug B
success	71	45
failure	34	42
total	105	87
% success	67.6%	51.7%

Is this evidence real, or can it be explained by chance?
Answer in Chapter 12.

categorical data (success/failure)

2 samples (2 drugs)

Comparing 2 drugs on mice

Think design:

Which mice got drug A, and which got drug B?
The ones caught first got B.

The disease is contagious. Consider:

1. Mice are housed 3 in a cage,
2. Mice are in separate cages with separate food.

Comparing lizards from different locations

Compare tail length of male adults Southwestern earless lizards from 2 different locations: Big Bend (TX) and Box Canyon (NM). Same species, different populations.

Has tail length evolved differently in the 2 populations, perhaps in relation to habitat (rocky vs. thick vegetation)?

Big Bend lizard population

Data: (in cm)

8.8	10.4
9.7	11.9
10.8	7.6
7.1	8.0
6.6	8.5
9.9	9.4
10.2	9.4
8.6	...

$n = 24$ lizards.

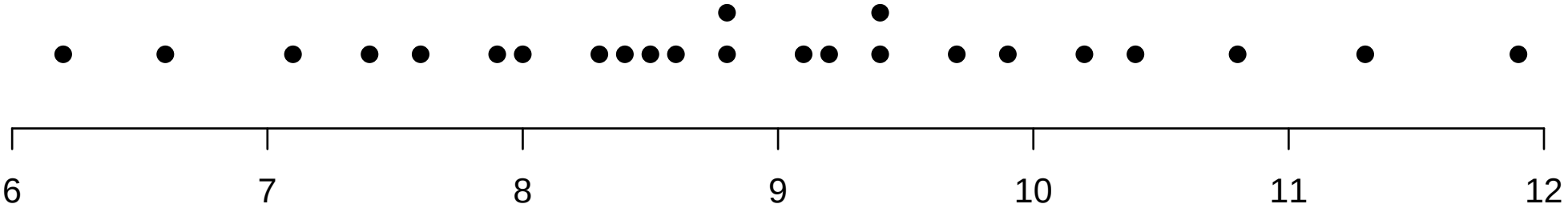
What is the average tail length of all male adults in the entire population of Big Bend lizards?

Answer: interval (8.3cm, 9.5cm), takes variation & uncertainty into account. Chapter 11.

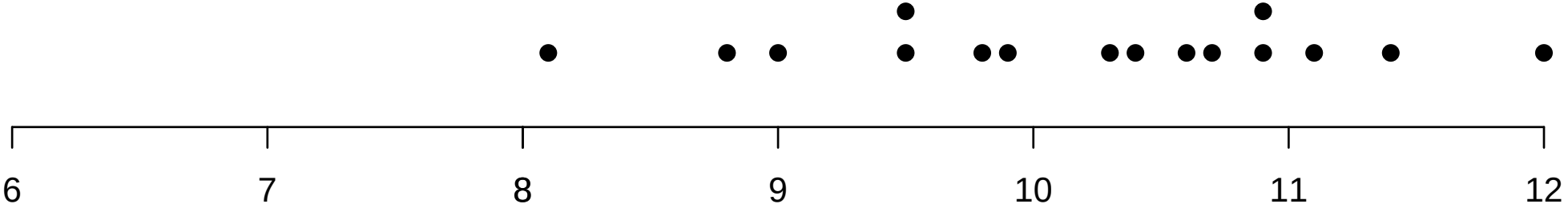
Sources of variability in the data:

- individual lizards themselves
- operator making measurement, measurement error
- time of measurement/season

Comparing the 2 lizard populations



Big Bend



Box Canyon

We wish to compare the means (averages) with respect to variability

Comparing the 2 lizard populations

Think design:

- how are lizards selected? representative of the target population, or just slowest moving?
- same operator for both locations?
- same habitat? same season?

Chapter 12:

- Continuous data
- 2 samples

Chapter 15 (ANOVA): extend to 3 or more locations.

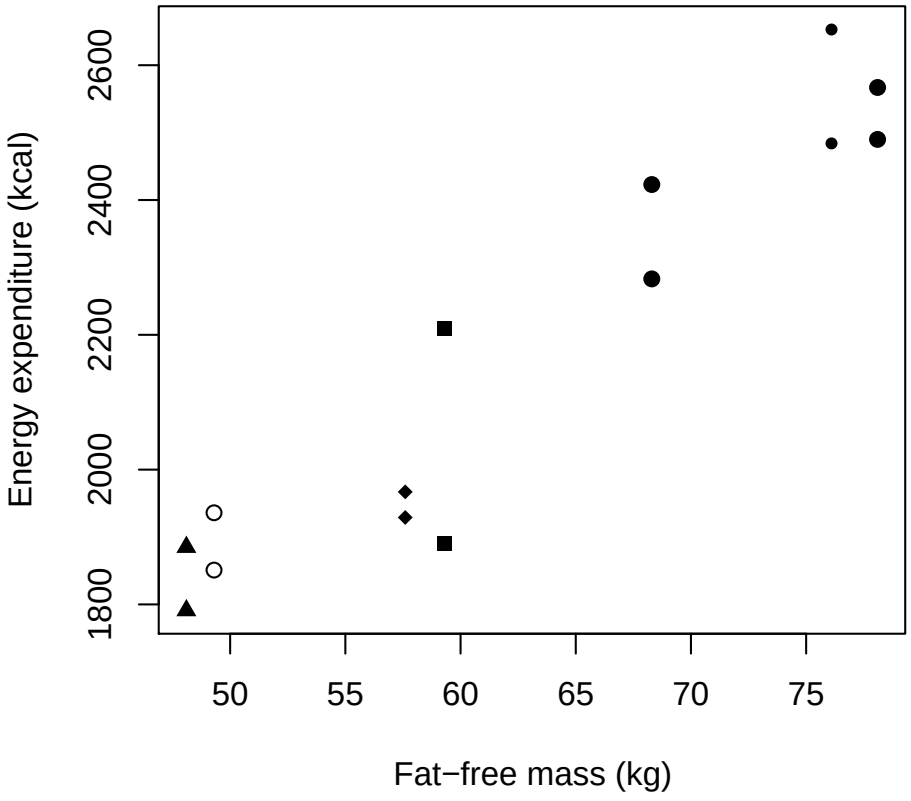
Nutritional requirement and body size

Does nutritional requirement depend on body size? How?

Expt: 7 men, 24-hour energy expenditure (kcal) was measured, in conditions of quiet sedentary activity, repeated twice.

Subject #	fat-free mass (kg)	energy expenditure (kcal)	
1	49.3	1851	1936
2	59.3	2209	1891
3	68.3	2283	2423
4	48.1	1885	1791
5	57.6	1929	1967
6	78.1	2490	2567
7	76.1	2484	2653

Nutritional requirement and body size



Find formula to predict energy expenditure then nutritional requirement as a function of body size.

Regression: Chap. 16-17

2 continous variables

Think design:

Were activity conditions really the same for all men?

7 subjects, 2 measurements on each, or
14 subjects and 1 single measurement on each?

Outline

Cow example: effect of diet on milk quality

A study assigned **50 cows** to **various diets** (based on the amount of an additive in the diet) and examined a number of **outcomes** associated with characteristics of the produced **milk**, amount of dry matter consumed, and weight gain of the cow. **Pre-treatment variables** include initial weight of the cow, number of lactations, and age of the cow. The primary purpose of the study was to examine the effect of the different diets on the outcome variables, controlling for effects of other covariates.

Cow variables

treatment	diet: CONTROL, LOW, MEDIUM, or HIGH
level	mg of additive per kg of feed
lactation	the number of lactations (pregnancies)
age	age of cow at beginning of study (months)
initial.weight	initial weight (pounds)
dry	mean daily weight of dry matter consumed (kg)
milk	mean daily amount of milk produced (pounds)
fat	percentage milk fat (grams of fat per 100g milk)
solids	% solids in milk (grams of solids per 100g milk)
final.weight	final weight of cow (pounds)
protein	% protein in milk (grams of protein per 100g milk)

Subset of cows' data

treatment	level	lactation	age	initial.weight	dry	milk	fat	solids	final.weight	protein
control	0	3	49	1360	15.429	45.552	3.88	8.96	1442	3.67
control	0	3	47	1498	18.799	66.221	3.40	8.44	1565	3.03
control	0	2	36	1265	17.948	63.032	3.44	8.70	1315	3.40
control	0	2	33	1190	18.267	68.421	3.42	8.30	1285	3.37
control	0	2	31	1145	17.253	59.671	3.01	9.04	1182	3.61
control	0	1	22	1035	13.046	44.045	2.97	8.60	1043	3.03
low	0.1	6	89	1369	14.754	57.053	4.60	8.60	1268	3.62
low	0.1	4	74	1656	17.359	69.699	2.91	8.94	1593	3.12
low	0.1	3	45	1466	16.422	71.337	3.55	8.93	1390	3.30
low	0.1	2	34	1316	17.149	68.276	3.08	8.84	1315	3.40
low	0.1	2	36	1164	16.217	74.573	3.45	8.66	1168	3.31
low	0.1	2	41	1272	17.986	66.672	3.43	9.19	1188	3.59
medium	0.2	3	45	1362	19.998	76.604	4.29	8.44	1273	3.41
medium	0.2	3	49	1305	19.713	64.536	3.94	8.82	1305	3.21
medium	0.2	3	48	1268	16.813	71.771	2.89	8.41	1248	3.06
medium	0.2	3	44	1315	15.127	59.323	3.13	8.72	1270	3.26
medium	0.2	2	40	1180	19.549	62.484	3.36	8.51	1285	3.21
medium	0.2	2	35	1190	19.142	70.178	3.92	8.94	1168	3.28
high	0.3	5	81	1458	20.458	71.558	3.69	8.48	1432	3.17
high	0.3	3	49	1515	19.861	56.226	4.96	9.17	1413	3.72
high	0.3	3	48	1310	18.379	49.543	3.78	8.41	1390	3.67
high	0.3	3	46	1215	18.000	55.351	4.22	8.94	1212	3.80
high	0.3	3	49	1346	19.636	64.509	4.16	8.74	1318	3.31
high	0.3	3	46	1428	19.586	74.430	3.92	8.75	1333	3.37

Types of data

Categorical (qualitative)

nominal: Blood type, Gender

ordered:

Numerical (quantitative)

continuous:

discrete:

Categorization of variables

Categorical or **Numerical** variables.

Experimental or **Observational** variables:

experimental: values under control of the researcher.
observational: values observed, not set by the researcher.

Response or **Explanatory** variables:

response: are considered as outcomes;
explanatory: are thought potentially to affect outcomes.

Example: Recombination

In the fruit fly *Drosophila melanogaster*, the gene *white* with alleles w^+ and w determines eye color (red or white) and the gene *miniature* with alleles m^+ and m determines wing size (normal or miniature). Both genes are located on the X chromosome, so female flies have two alleles for each gene while male flies have only one.

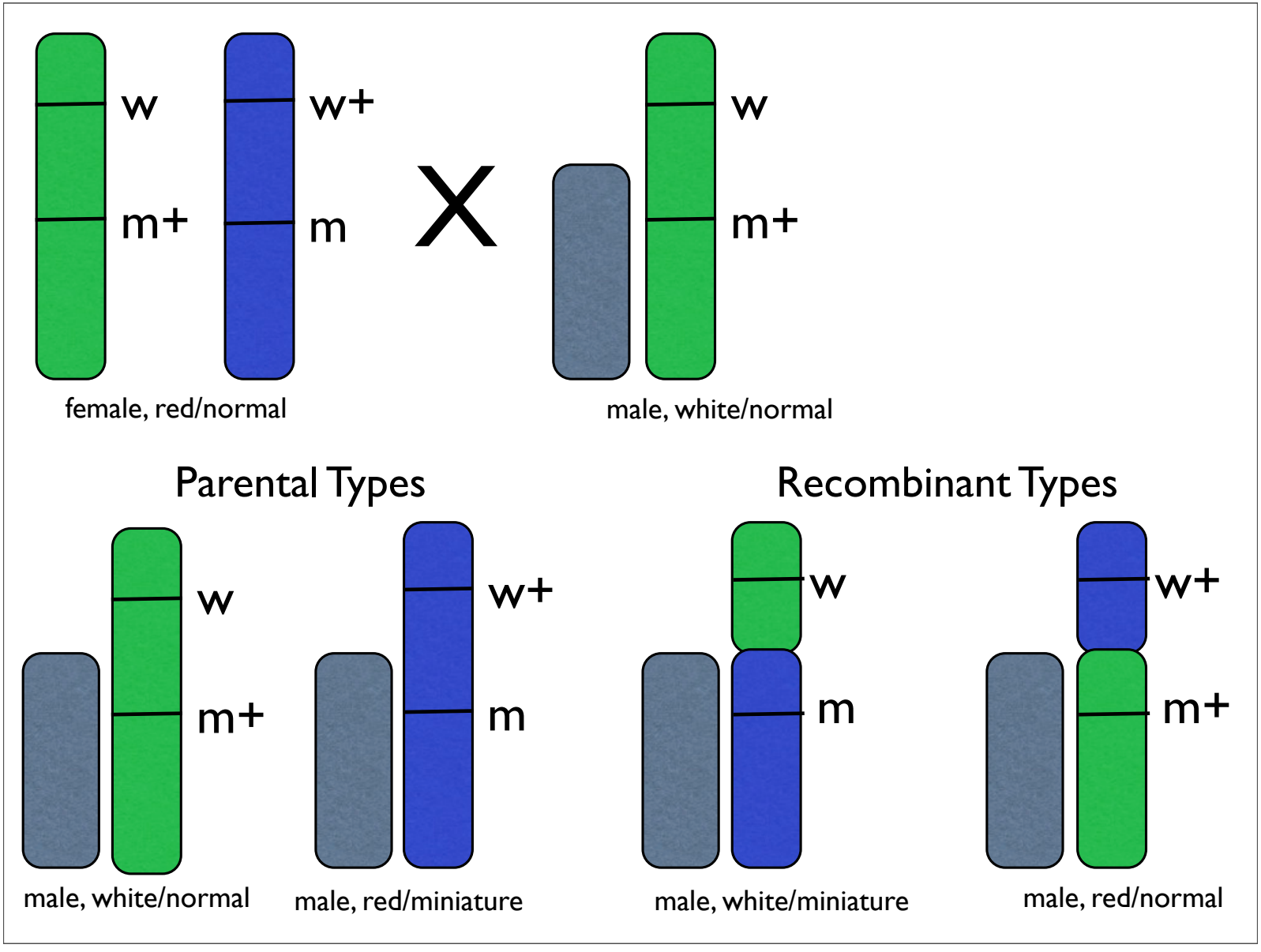
During meiosis (formation of gametes) in the female fly, if the X chromosome pair do not exchange segments, the resulting eggs will contain two alleles, each from the same X chromosome. However, if the strands of DNA cross-over during meiosis then some progeny may inherit alleles from different X chromosomes. This process is known as recombination. There is biological interest in determining the proportion of recombinants. Genes that have a positive probability of recombination are said to be genetically linked.

Recombination (cont.)

In a pioneering 1922 experiment to examine genetic linkage between the *white* and *miniature* genes, a researcher crossed wm^+/w^+m females with wm^+/Y male flies and looked at the traits of the male offspring. (Males inherit the Y chromosome from the father and the X from the mother.)

In the absence of recombination, we would expect half the male progeny to have the wm^+ haplotype and have white eyes and normal-sized wings while the other half would have the w^+m haplotype and have red eyes and miniature wings. This is not what happened.

Cross


w
m+
w+
m
X
w
m+
female, red/normal
male, white/normal
Parental Types
Recombinant Types
w
m+
w+
m
w
m
w+
m+
male, white/normal
male, red/miniature
male, white/miniature
male, red/normal

Recombination (cont.)

The phenotypes of the male offspring were as follows:

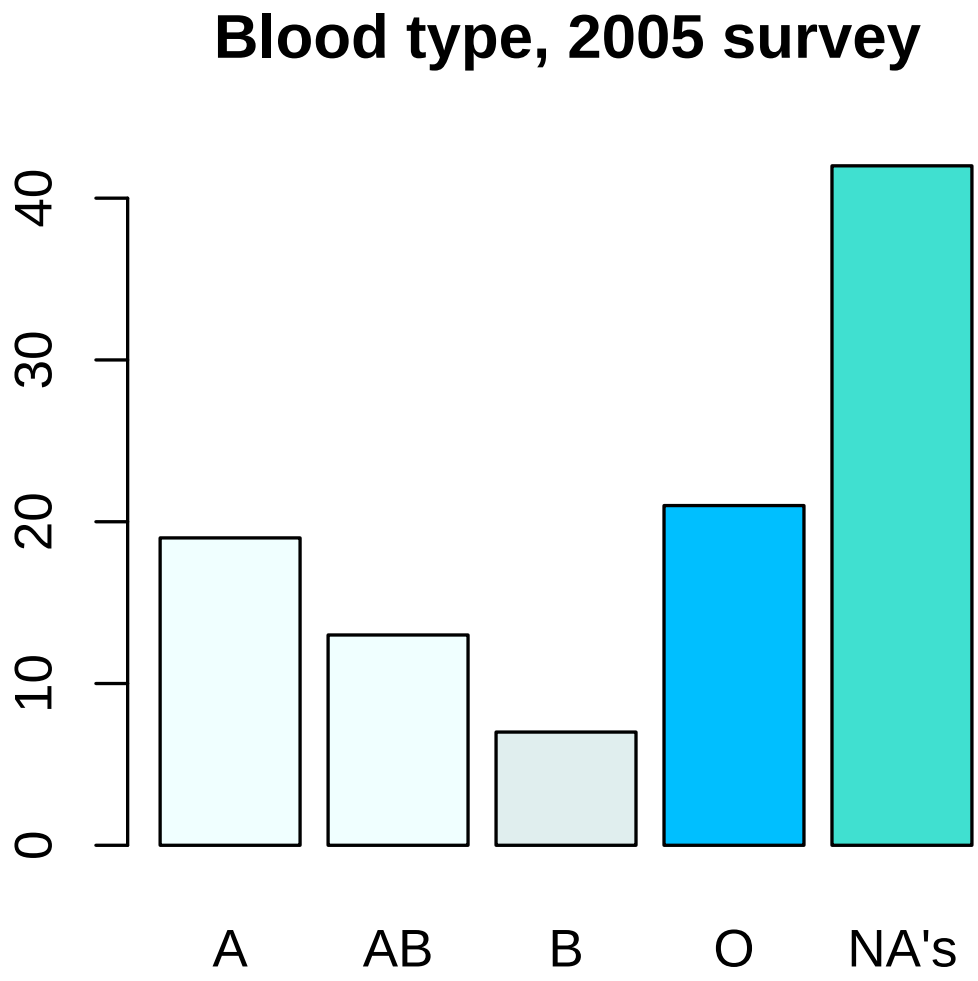
	Wing Size	
Eye color	normal	miniature
red	114	202
white	226	102

There were $114 + 102 = 216$ recombinants out of 644 total male offspring, a proportion of $216/644 \doteq 0.335$ or 33.5%.

Completely linked genes have a recombination probability of 0, unlinked genes have a recombination probability of 0.5. The *white* and *miniature* genes in fruit flies are incompletely linked. Measuring recombination probabilities is an important tool in constructing genetic maps, diagrams of chromosomes that show the positions of genes.

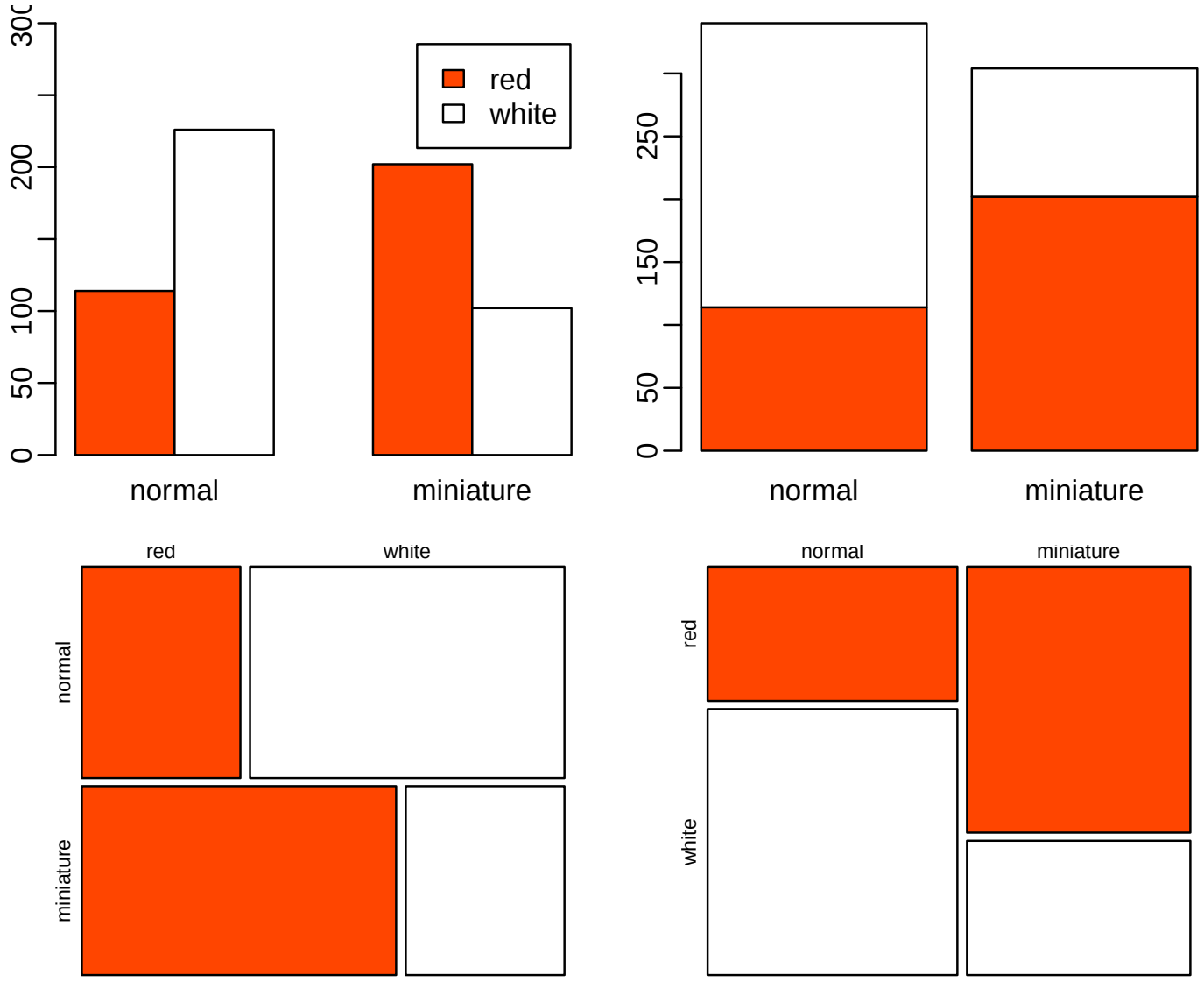
Outline

Bar plots: for categorical data



Space betweem bars: separate, discrete nature of categories.

Mosaic plots: 2 categorical variables



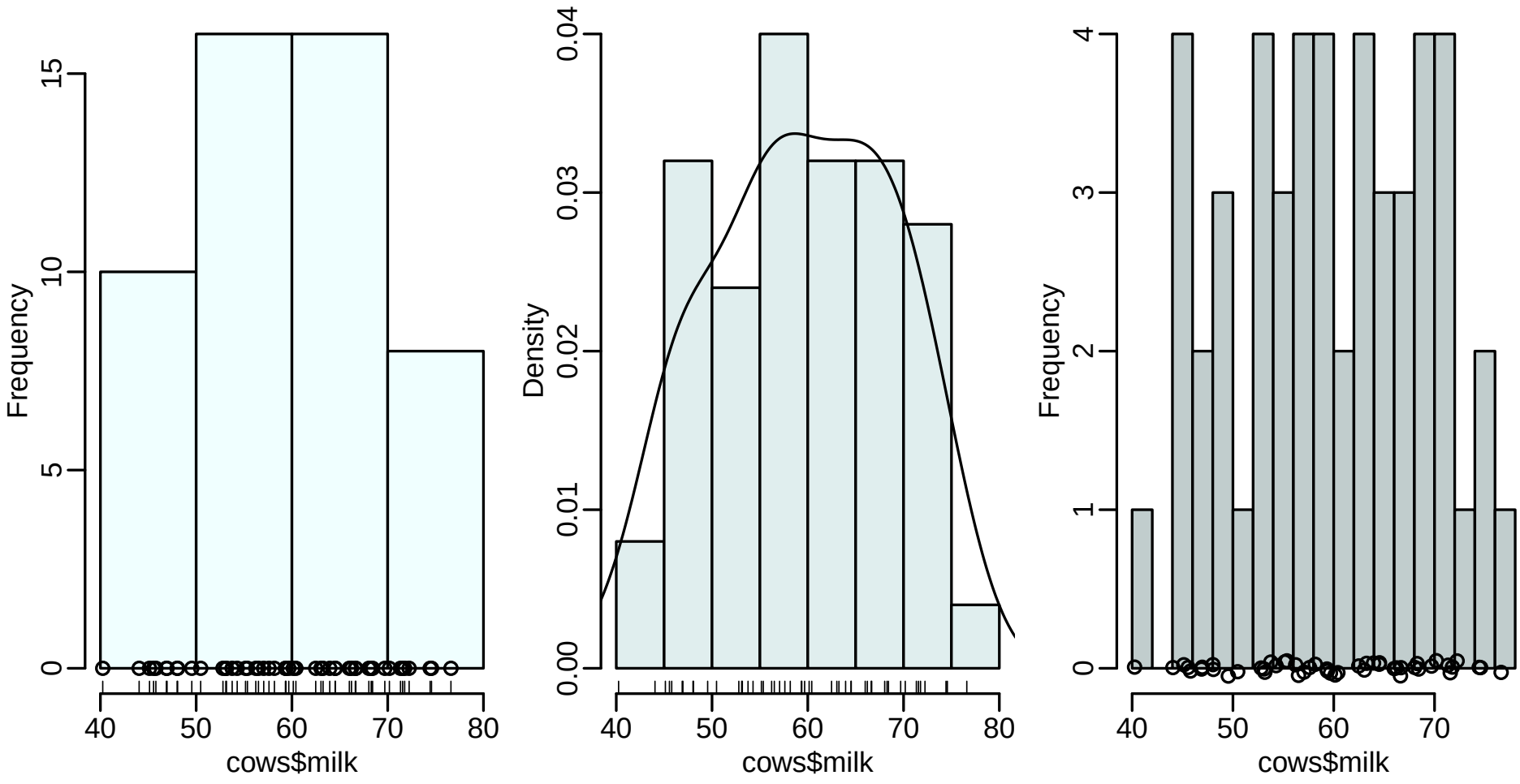
Bar and Mosaic plots: 2 categorical variables

```
> recomb = matrix( c(114,226,202,102),  2, 2)
> colnames(recomb) = c("normal","miniature")
> rownames(recomb) = c("red","white")
> recomb
      normal miniature
red      114       202
white    226       102
> t(recomb)
          red white
normal    114   226
miniature 202   102

> barplot(recomb, beside=TRUE, legend.text = rownames(recomb), col=c("red", "white"))
> barplot(recomb, beside=FALSE, col=c("red","white"))
> mosaicplot(t(recomb), col=c("red", "white"), dir=c("h","v"))
> mosaicplot(t(recomb), col=c("red", "white"))
```

Mosaic plots:

area $\leftrightarrow$ frequency (# flies), coordinates (height) $\leftrightarrow$ proportions.

Rug plots and Histograms: for numerical data



Rug/dot plots: display exact values. Points can be 'jittered' to avoid overlap.

Histogram: not unique. shows **frequency** (height=#)
or **density** (area=proportion)

Measure of location: Sample mean

```
> cows$milk
 [1] 45.552 66.221 63.032 68.421 59.671 44.045 55.153 46.957 63.948 65
[11] 57.603 63.254 57.053 69.699 71.337 68.276 74.573 66.672 72.237 58
[21] 48.063 60.412 45.128 53.759 52.799 76.604 64.536 71.771 59.323 62
[31] 70.178 48.013 60.140 56.506 40.245 45.791 59.373 54.281 71.558 56
[41] 49.543 55.351 64.509 74.430 68.030 46.888 53.164 53.096 50.471 66
> mean(cows$milk)
[1] 59.54314
```

Milk yield data: 45.552, 66.221, ... , 66.619. $y_1 = 45.552$, $y_2 = 66.221, \ldots, y_{50} = 66.619$.

$$\bar{y} = (45.552 + 66.221 + \cdots + 66.619)/50$$
$$= (y_1 + y_2 + \cdots + y_{50})/50 = 59.5 \text{ lbs/day}$$

Sample mean

$$\bar{y} = \frac{1}{n}(y_1 + y_2 + \cdots + y_n) = \frac{1}{n}\sum_{i=1}^{n} y_i$$

Measures of location: Sample median

Median = typical value. Half of observations are below, half are above.

Sort the data: 40.245 44.045 ... 76.604.

Find the middle value. If sample size *n* is odd, no problem. If *n* is even, there are 2 middle values (here 25^{th} and 26^{th}). The median is their average.

```
> sort(cows$milk)
 [1] 40.245 44.045 45.128 45.552 45.791 46.888 46.957 48.013 48.063 49.543
[11] 50.471 52.799 53.096 53.164 53.759 54.281 55.153 55.351 56.226 56.506
[21] 57.053 57.603 58.168 59.323 59.373 59.671 60.140 60.412 62.484 63.032
[31] 63.254 63.948 64.509 64.536 65.994 66.221 66.619 66.672 68.030 68.276
[41] 68.421 69.699 70.178 71.337 71.558 71.771 72.237 74.430 74.573 76.604
> median(cows$milk)
[1] 59.522
```

Mean and Median

Examples: data	median	mean $\bar{y}$
3, 7, 9, 11, 22		10.4
2, 6, 7, 12, 13, 16, 17, 20		11.625
2, 6, 7, 12, 13, 16, 17, 200		34.125

Mean and the median are usually close, unless the data are not symmetrical.

Mean $\bar{y} =$ balance point.

Measures of location: Quantiles and percentiles

25% **percentile** = 0.25 **quantile** = value such that 1/4 observations are below and 3/4 are above.

Example: 6 9 11 17 19 23 26 26

p quantile: value such that (about) a proportion *p* of observations are below and about $1 - p$ are above.

Median is a special case ($p = 0.5$)

First quartile Q_1: median of those values below the median
Third quartile Q_3: median of those values above the median

Example:

19 23 26 30 32 34 37 | 37 39 41 44 44 46 55

Measures of spread

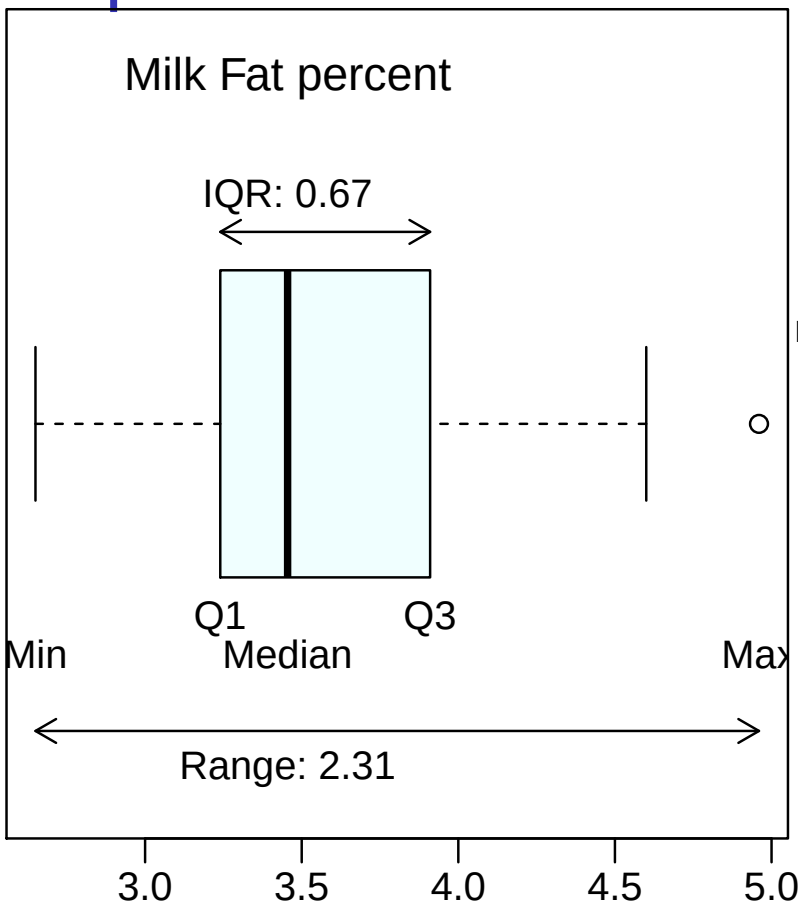
Range: maximum - minimum,
IQR: Inter Quartile Range $= Q_3 - Q_1$

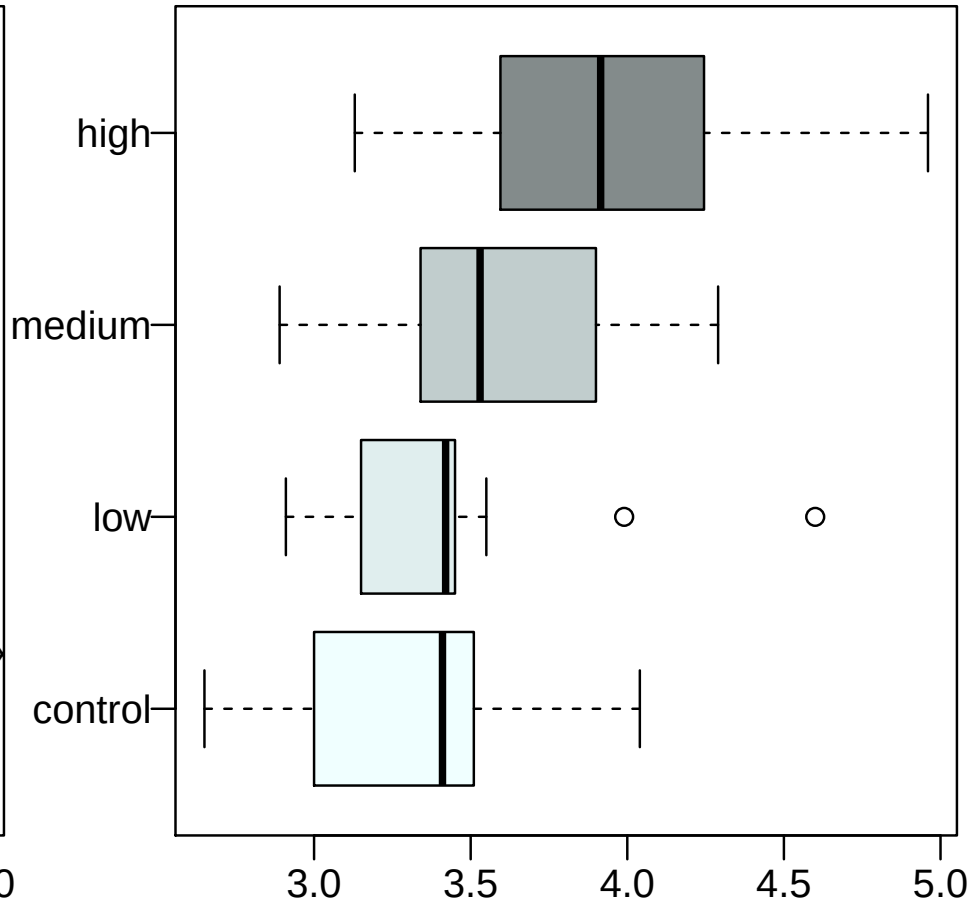
```
> min(cows$milk)
[1] 40.245
> max(cows$milk)          # range = 76.604 - 40.245
[1] 76.604                #       = 36.359 lbs/day
> IQR(cows$milk)
[1] 13.54575
> summary(cows$milk)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  40.24   53.11   59.52   59.54   66.66   76.60
> quantile(cows$milk)
      0%      25%      50%      75%     100%
40.24500 53.11300 59.52200 66.65875 76.60400
> quantile(cows$milk, p=0.10)
    10%
46.7783
```

Numerical summaries

```
> summary(cows)

  treatment     level        lactation          age        initial.weight
control:12  Min.   :0.00  Min.   :1.00   Min.   :21.00   Min.   : 900
high   :12  1st Qu.:0.10  1st Qu.:1.00   1st Qu.:26.25   1st Qu.:1119
low    :13  Median :0.15  Median :2.00   Median :37.00   Median :1266
medium :13  Mean   :0.15  Mean   :2.38   Mean   :42.16   Mean   :1258
            3rd Qu.:0.20  3rd Qu.:3.00   3rd Qu.:49.00   3rd Qu.:1369
            Max.   :0.30  Max.   :6.00   Max.   :95.00   Max.   :1656

      dry              milk            fat             solids
Min.   :11.42   Min.   :40.24   Min.   :2.650   Min.   :7.810
1st Qu.:14.56   1st Qu.:53.11   1st Qu.:3.265   1st Qu.:8.465
Median :16.69   Median :59.52   Median :3.455   Median :8.740
Mean   :16.43   Mean   :59.54   Mean   :3.577   Mean   :8.691
3rd Qu.:18.22   3rd Qu.:66.66   3rd Qu.:3.908   3rd Qu.:8.928
Max.   :20.46   Max.   :76.60   Max.   :4.960   Max.   :9.190

 final.weight      protein
Min.   : 968    Min.   :2.860
1st Qu.:1126    1st Qu.:3.172
Median :1234    Median :3.310
Mean   :1244    Mean   :3.328
3rd Qu.:1348    3rd Qu.:3.458
Max.   :1593    Max.   :3.800
```

Boxplot: for numerical data





Boxplots often better than histograms for comparing samples: easier to align.

```
> summary(cows$fat)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.650   3.265   3.455   3.577   3.908   4.960
> boxplot(cows$fat, horizontal=T)
> boxplot(fat ~ treatment, data=cows, horizontal=T)
```

Outlier display in boxplots

Fences: Observations outside fences are drawn individually.

Whiskers do not go beyond fences.

Typically, fence = 1.5 IQR

Milk fat example: IQR $= 0.67$. Fences are $1.5 * 0.67 = 1.005$ below Q_1 and above Q_3. Upper fence: $3.908 + 1.005 = 4.913$. Here largest values were:

```
> sort(cows$fat)
 [1] 2.65 2.89 2.91 2.97 2.99 2.99 3.01 3.08 3.13 3.13
...
[46] 4.27 4.29 4.52 4.60 4.96
```

so the whisker extends through 4.60, but 4.96 is displayed as a separate point.

Measures of spread: Variance

Deviation from the mean: $y_i - \bar{y}$.
Milk fat: cow in first row has deviation
$3.88 - 3.577 = +0.303$,
cow in second row has deviation $3.40 - 3.577 = -0.177$.

Variance: $s^2 \geq 0$ always!

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{1}{n-1}\left((y_1 - \bar{y})^2 + \cdots + (y_n - \bar{y})^2\right)$$

Equivalent formula:

$$s^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2\right)$$

For milk fat percent, we get

$$s^2 = \frac{1}{49}(0.303^2 + (-0.177)^2 + \dots) = 0.2347 \quad (\%^2)$$

Measures of spread: Standard deviation

Standard deviation: $s = \sqrt{\text{variance}} = \sqrt{s^2}$ is now in original units. s is the typical deviation.
Here, $s = 0.484$ (in % of milk weight)

```
> m=mean(cows$fat)
> m
[1] 3.5772
> var(cows$fat)
[1] 0.2346736
> s=sd(cows$fat)
> s
[1] 0.4844312
```

```
> (cows$fat < m+s) & (cows$fat > m-s)
 [1]  TRUE  TRUE  TRUE  TRUE FALSE FALSE FALSE
[13] FALSE FALSE  TRUE FALSE  TRUE  TRUE  TRUE F
[25]  TRUE FALSE  TRUE FALSE  TRUE  TRUE  TRUE
[37]  TRUE  TRUE  TRUE FALSE  TRUE FALSE FALSE
[49]  TRUE FALSE
> sum(cows$fat < m+s & cows$fat > m-s)
[1] 35
```

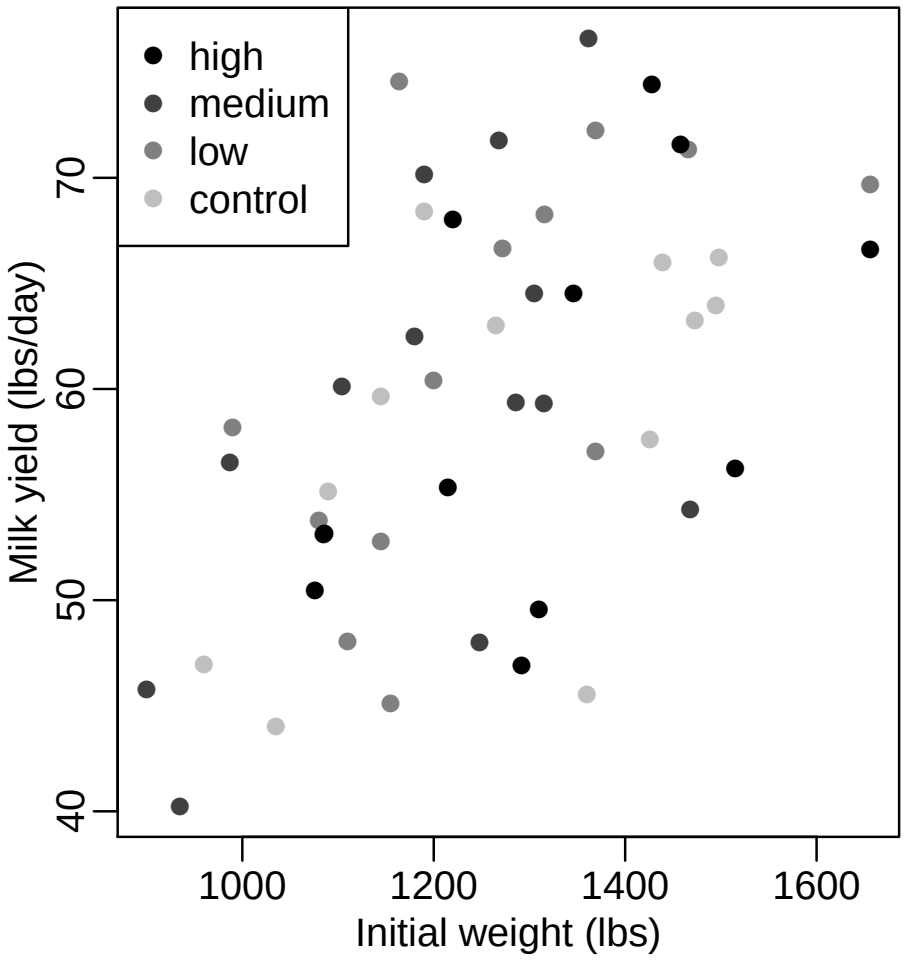
The empirical rule: for most "mound-shaped" distributions,

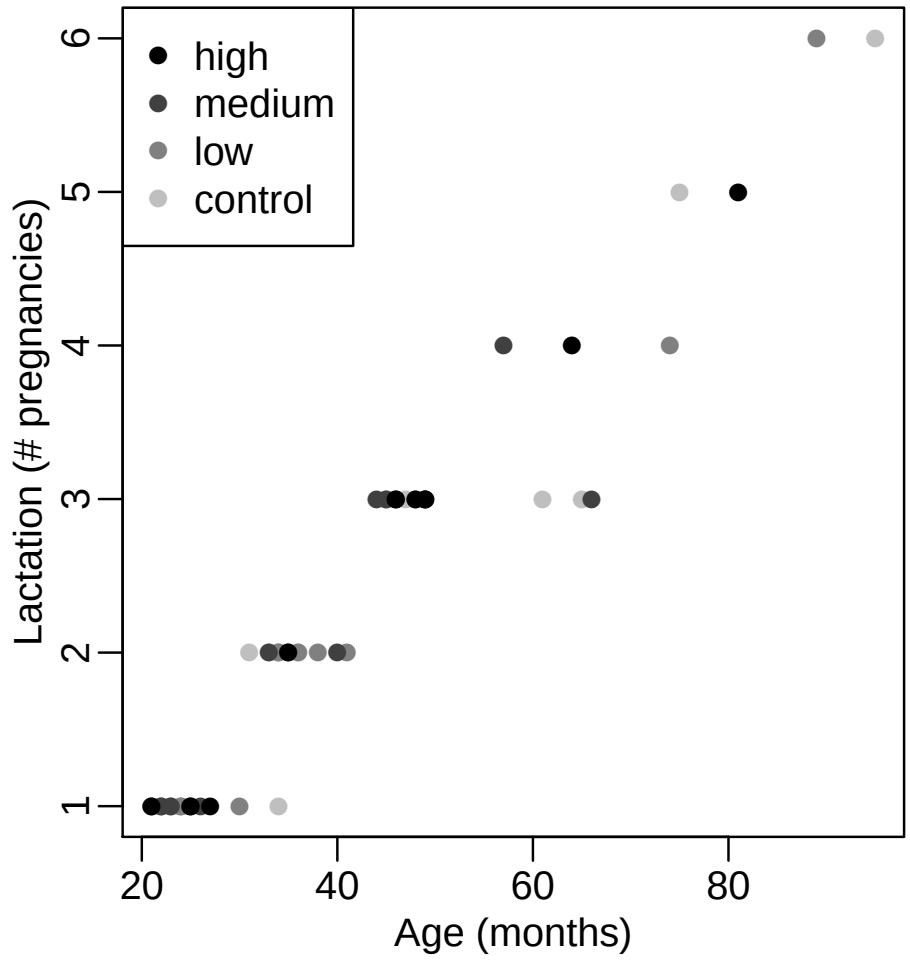
about 68% of observations lie within 1 standard deviation of the mean, (fat%: 35/50=70%)

about 95% lie within 2 s.d. of the mean (fat%: 48/50=96%)

about 99% lie within 3 s.d. of the mean (fat %: 100%)

Scatter-plot: 2 numerical variables





Here 3 variables displayed on each plot.

Scatter-plot: 2 numerical variables

```
layout(matrix(1:2,1,2))
par(mar=c(2.5,3.1,.1,.5), mgp=c(1.5,.5,0))

plot(milk~initial.weight, data=cows, pch=16,
     xlab="Initial weight (lbs)", ylab="Milk yield (lbs/day)",
     col=gray(1-as.numeric(treatment)/4) )

legend("topleft",pch=16,col=gray(1-(4:1)/4),
       legend=levels(cows$treatment)[4:1])

plot(lactation~age, data=cows, pch=16,
     xlab="Age (months)", ylab="Lactation (# pregnancies)",
     col=gray(1-as.numeric(treatment)/4) )

legend("topleft",pch=16,col=gray(1-(4:1)/4),
       legend=levels(cows$treatment)[4:1])
```