# Sections 2.1 and 2.2

Timothy Hanson

Department of Statistics, University of South Carolina

Stat 205: Elementary Statistics for the Biological and Life Sciences

# Variables

- **Variables** are random phenomena measured on people, mice, bacteria, air, or any other **observational unit**.
- They are either **categorical** or **numeric**.
- Categorical: records which category.
- Numeric: records amount of something.

# Categorical variables

- **Nominal** categorical variables do not have an obvious order.
- Examples: blood type (A, B, AB, O), gender (male, female), eye color (blue, brown, other).
- **Ordinal** categorical variables have order.
- Examples: cancer stage (I, II, III, IV), salsa taste assessment (mild, moderate, hot).

# Numeric variables

- **Continuous** numeric variables are measured on a continuous scale.
- Examples: baby weight, cholesterol level, lifetime, miles per gallon.
- The outcome of a **discrete** numeric variable can be listed.
- Examples: kitten litter size $(1, 2, 3, \dots)$, earthquakes in an hour, lifetime in years.

# What kind of variable is each of these?

**Four types**: nominal categorical, ordinal categorical, continuous numeric, and discrete numeric.

- Newborn giraffe height in cm.
- Type of vehicle: sedan, SUV, truck, van, or other.
- Number of pizzas ordered.
- Amount of whipped cream on a sundae: none, light, medium, heavy.
- Movie rated on a scale from 1 to 10. (Two types!)

# Observational unit

- A collection of $n$ persons or things are collected and one or more variables measured.
- The **observational unit** is the person or thing being measured, also called **experimental units**.
- Example: The gender of 73 trapped Cecropia moths is recorded. What is the observational unit?
- Example: 150 babies born in a hospital are weighed in *kg*. What is the observational unit?

# Notation for variables and observations

- Variables are uppercase, e.g. $X$ or $Y$.
- The observations are lowercase, $x$ or $y$.
- Example: $Y =$ birthweight and $y = 7.9$ lb.
- Later on: $Y$ is random and has probability attached to it; $y$ is a fixed number.

# Frequency distributions

- A frequency distribution is a bar chart or table giving the frequency of variable occurrences.
- For categorical variables we can make a bar chart or table easily.
- For continuous numeric variables, we will have to group or bin observations (more later).

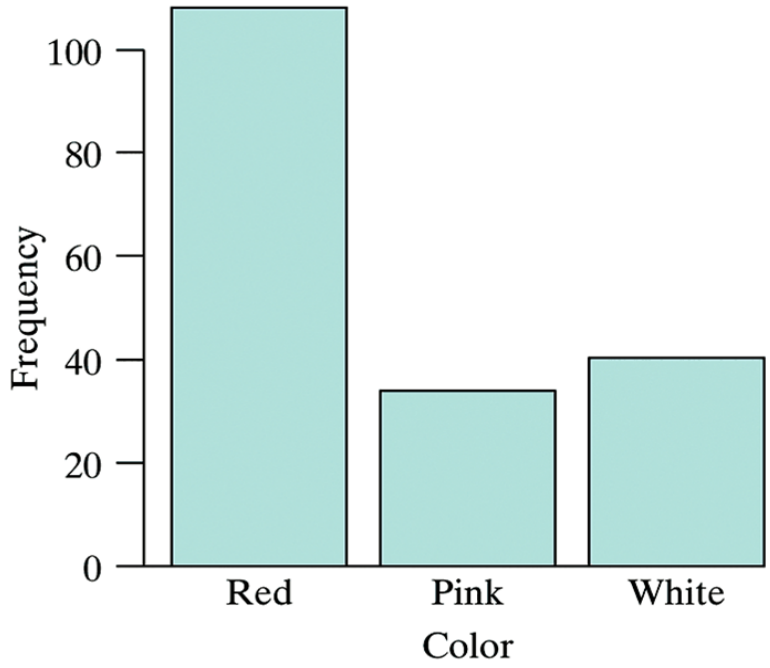
# Example 2.2.1 Poinsettias

- Poinsettias can be red, pink, or white.
- The color of 182 progeny of a certain parental cross were recorded.
- Questions of interest: what is the most common color? How do the frequencies compare?
- Lets look at a bar chart and table of frequencies

# Distribution of poinsettia colors

**Table 2.2.1** Color of one hundred eighty-two poinsettias

| Color | Frequency (number of plants) |
|---|---|
| Red | 108 |
| Pink | 34 |
| White | 40 |
| Total | 182 |

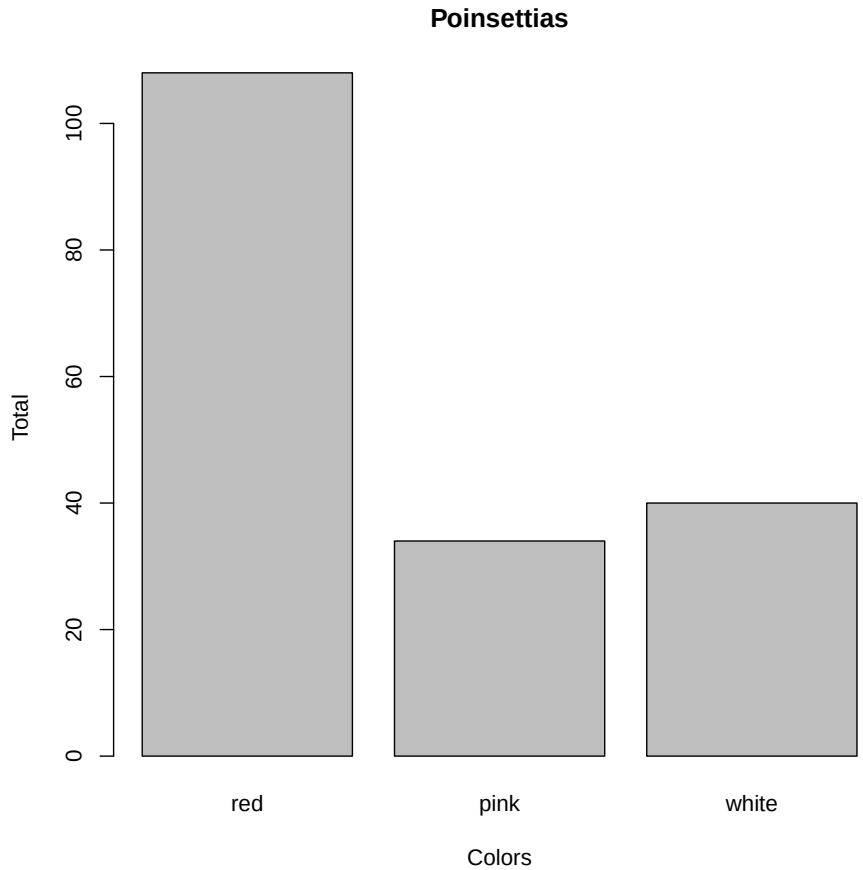



(a) Tabular frequency distribution (b) Graphical frequency distribution

Figure: Poinsettia colors

# R code

```
barplot(height=c(108,34,40),names.arg=c("red","pink","white"),
 main="Poinsettias",xlab="Colors",ylab="Total")
```



# Dotplots and histograms

- **Dotplot**: draw number line along $x$-axis and place a dot at each data value.
- If theres more than one of the same value, stack them.
- Side-by-side dotplots (for different **strata**) can illustrate differences in distribution.
- A **Histogram** a bar chart for numeric variables. The bars are in numeric order and their heights are the frequencies.

# Example 2.2.3 Infant mortality rates for 12 South American countries

**Table 2.2.3** Infant mortality in 12 South American countries

| Country | Infant mortality rate |
|---|---|
| Argentina | 11.4 |
| Bolivia | 44.7 |
| Brazil | 22.6 |
| Chile | 7.7 |
| Colombia | 18.9 |
| Ecuador | 20.9 |
| Guyana | 30.0 |
| Paraguay | 24.7 |
| Peru | 28.6 |
| Suriname | 18.8 |
| Uruguay | 11.3 |
| Venezuela | 26.5 |

(a) Tabular frequency distribution

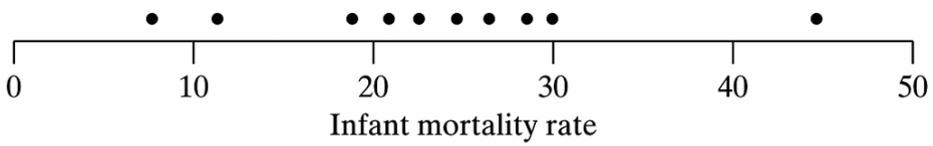



**Figure 2.2.3** Dotplot of infant mortality in 12 South American countries

(b) Dotplot

Figure: Infant mortality in South America

# Example 2.2.4 Numbers of surviving piglets from 36 sows

- $n = 36$ two-year old sows ($\frac{3}{4}$ Duroc & $\frac{1}{4}$ Yorkshire) bred with Yorkshire boars.
- Number of piglets surviving 3 weeks recorded for each sow.

**Table 2.2.4** Number of surviving piglets of 36 sows

| Number of piglets | Frequency (number of sows) |
|---|---|
| 5 | 1 |
| 6 | 0 |
| 7 | 2 |
| 8 | 3 |
| 9 | 3 |
| 10 | 9 |
| 11 | 8 |
| 12 | 5 |
| 13 | 3 |
| 14 | 2 |
| Total | 36 |

# Example 2.2.4 surviving piglets

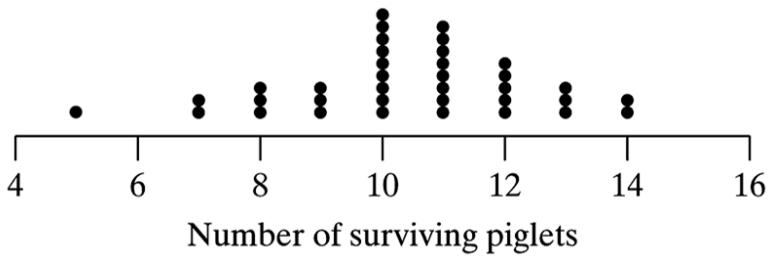



**Figure 2.2.4** Dotplot of number of surviving piglets of 36 sows

(a) Dotplot

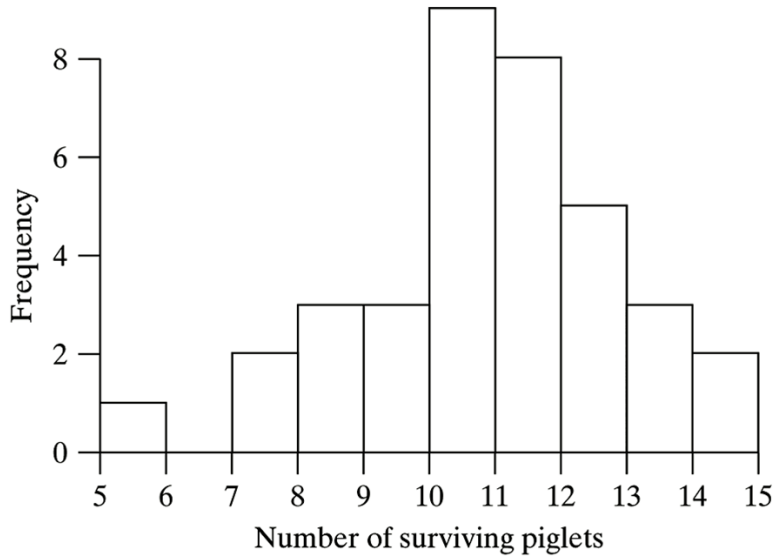



**Figure 2.2.5** Histogram of number of surviving piglets of 36 sows

(b) Histogram

Figure: Number of surviving piglets

# Relative frequency

$$\text{Relative frequency} = \frac{\text{Frequency}}{n}$$

- Divide frequency by total sample size n to get a proportion or percentage.
- Use the proportion rather than the frequency.
- All proportions then add up to one.

# Poinsettias continued...

**Table 2.2.5** Color of one hundred eighty-two poinsettias

| Color | Frequency | Relative frequency | Percent frequency |
|---|---|---|---|
| Red | 108 | .59 | 59 |
| Pink | 34 | .19 | 19 |
| White | 40 | .22 | 22 |
| Total | 182 | 1.00 | 100 |

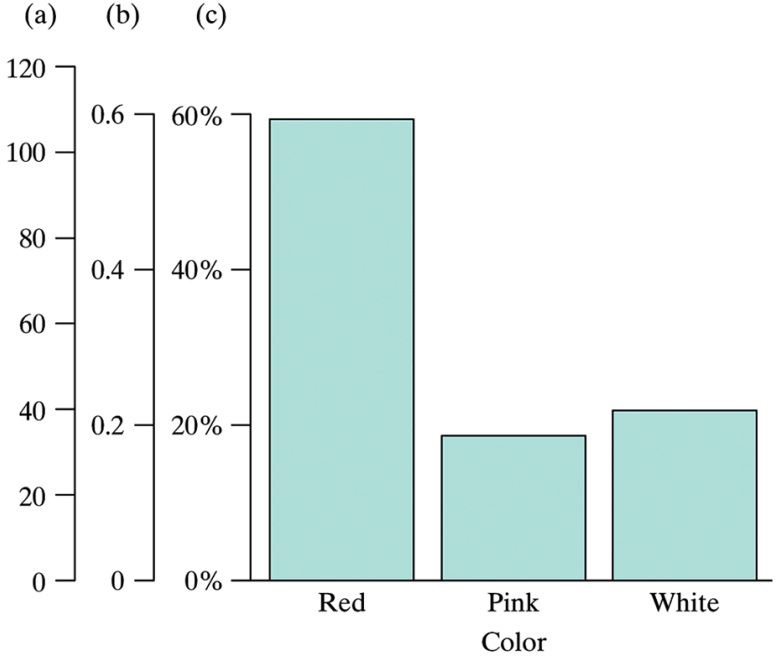



Figure: (a) frequency (b) relative frequency (c) percentage.

# Histogram for continuous data

- For continuous data, often no tied values.
- We can instead collect data into bins.
- Record relative frequency in each bin and make a bar chart: histogram.
- The bins you choose affect what the histograms looks like. Software can make the choice for us.

# Example 2.2.6 Creatine phosphokinase (CK) from 36 men (units U/l)

**Table 2.2.6** Serum CK values for 36 men

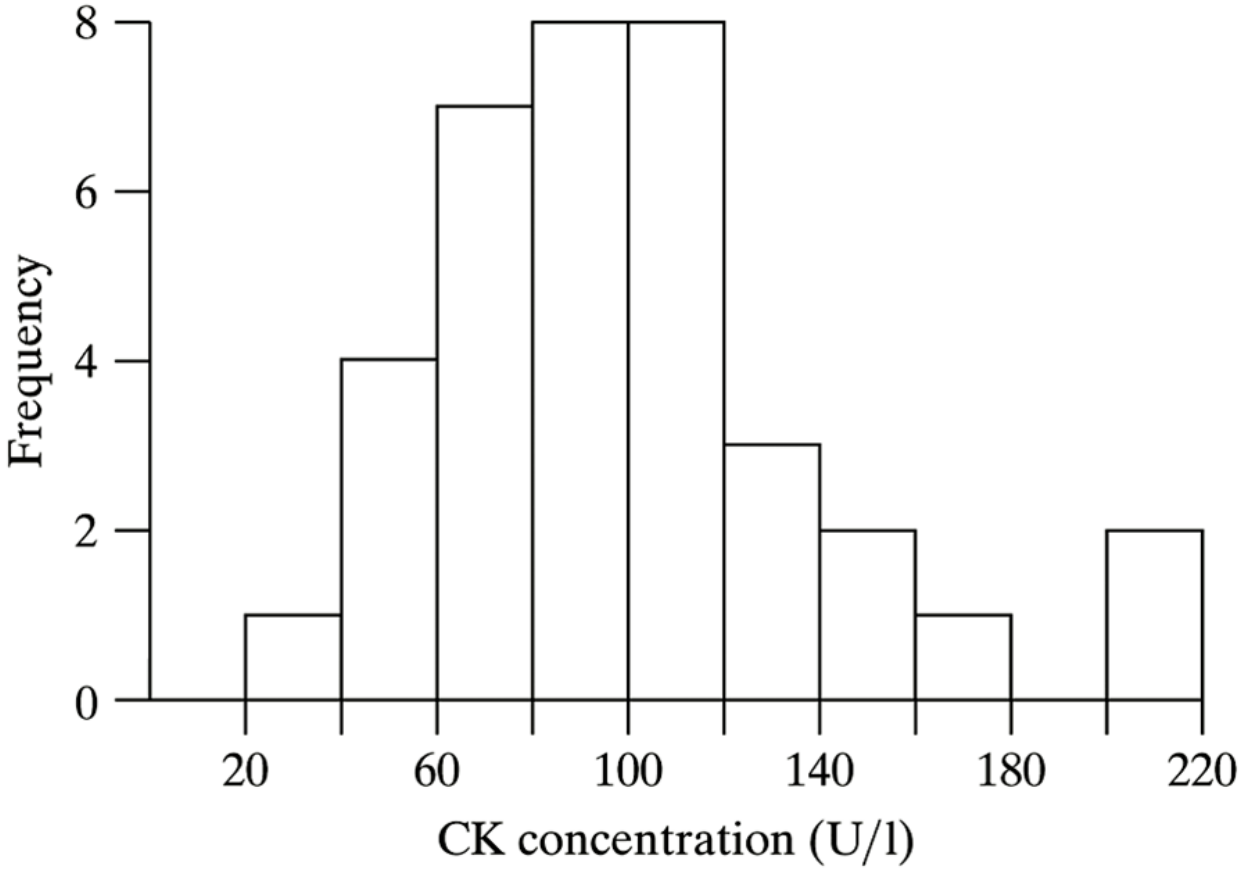
| | | | | | |
|---|---|---|---|---|---|
| 121 | 82 | 100 | 151 | 68 | 58 |
| 95 | 145 | 64 | 201 | 101 | 163 |
| 84 | 57 | 139 | 60 | 78 | 94 |
| 119 | 104 | 110 | 113 | 118 | 203 |
| 62 | 83 | 67 | 93 | 92 | 110 |
| 25 | 123 | 70 | 48 | 95 | 42 |

# Bins are all of equal length 20 units

**Table 2.2.7** Frequency distribution of serum CK values for 36 men

| Serum CK (U/l) | Frequency (number of men) |
|---|---|
| [20,40) | 1 |
| [40,60) | 4 |
| [60,80) | 7 |
| [80,100) | 8 |
| [100,120) | 8 |
| [120,140) | 3 |
| [140,160) | 2 |
| [160,180) | 1 |
| [180,200) | 0 |
| [200,220) | 2 |
| Total | 36 |

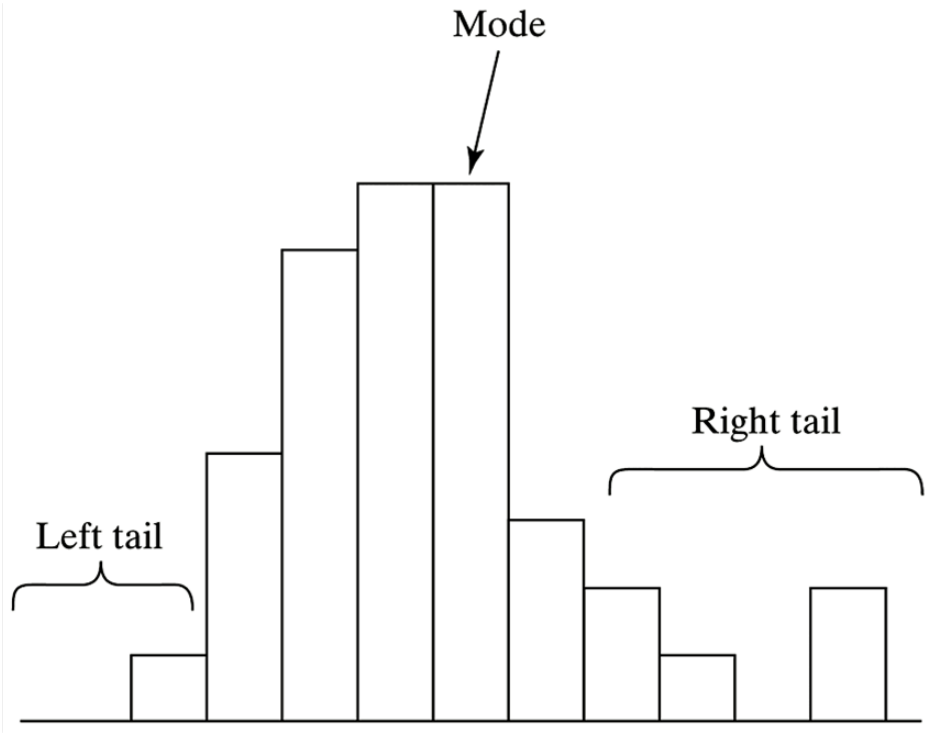
# One particular histogram for these data



# Describing histogram shapes

- Highest peak is called the **mode**.
- On the left and right, where the frequencies decline are called the **tails**.
- If the right tail is longer, or more stretched out, than the left tail the histogram is **skewed to the right**.
- If the left tail is more stretched out than the right tail the histogram is **skewed to the left**.

# Mode and tails for CK data



# How do the bins affect histogram shape?

- The *number* and *location* of histogram bins affect its shape.
- Each bin is also called a *class*.
- Good idea to look at different histograms for the same data before discussing skew and modality.
- Example 2.2.7 $n = 510$ college student reported their height.

# Height data: few classes

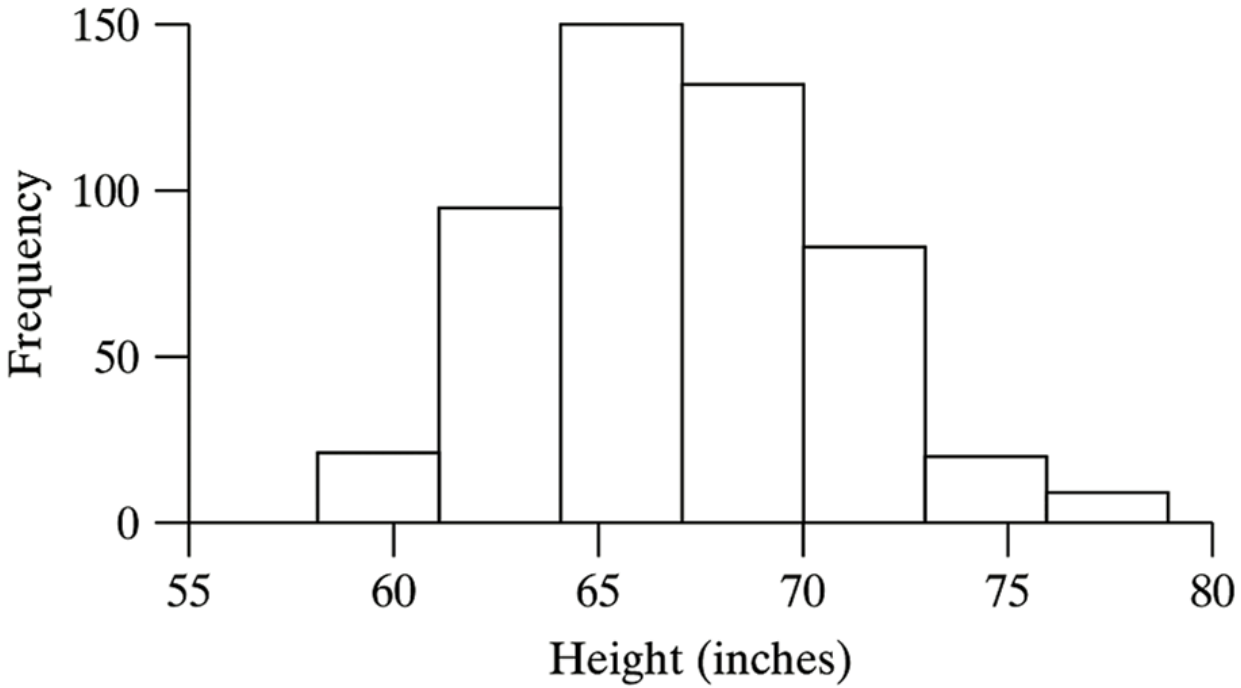



Figure: Heights of students using 7 classes (class width = 3)

# Height data: more classes

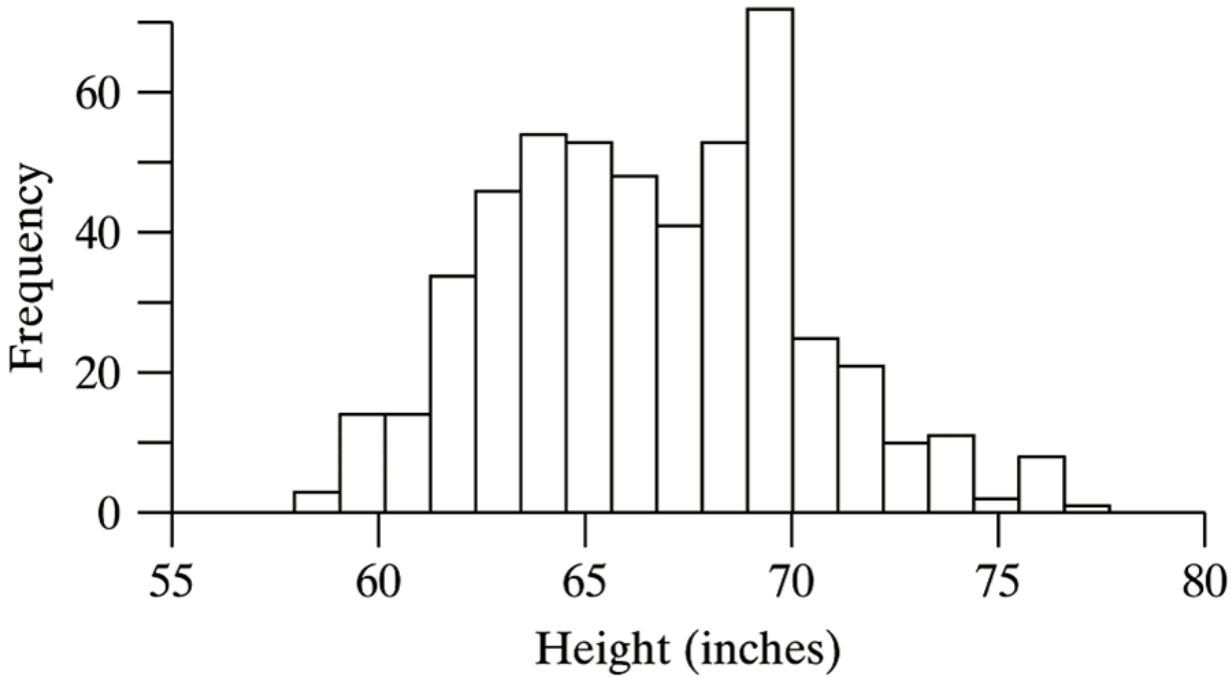



**Figure 2.2.10** Heights of students, using 18 classes (class width = 1.1)

# Height data: too many classes

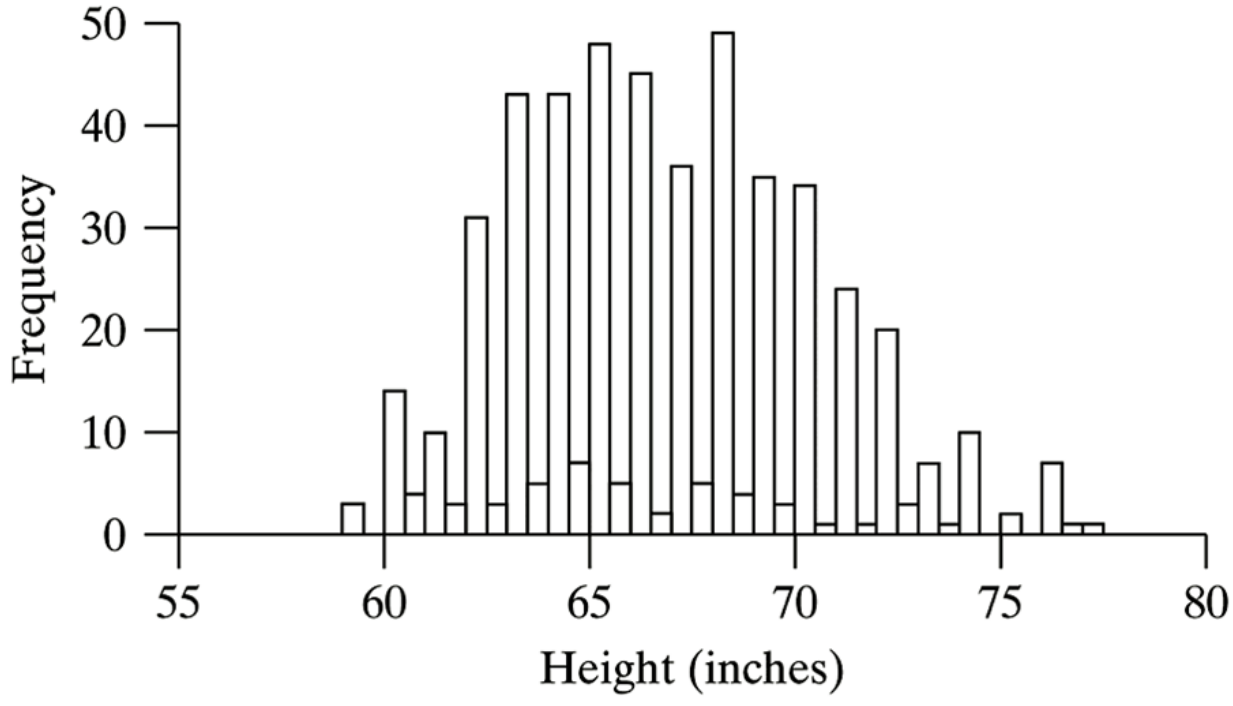
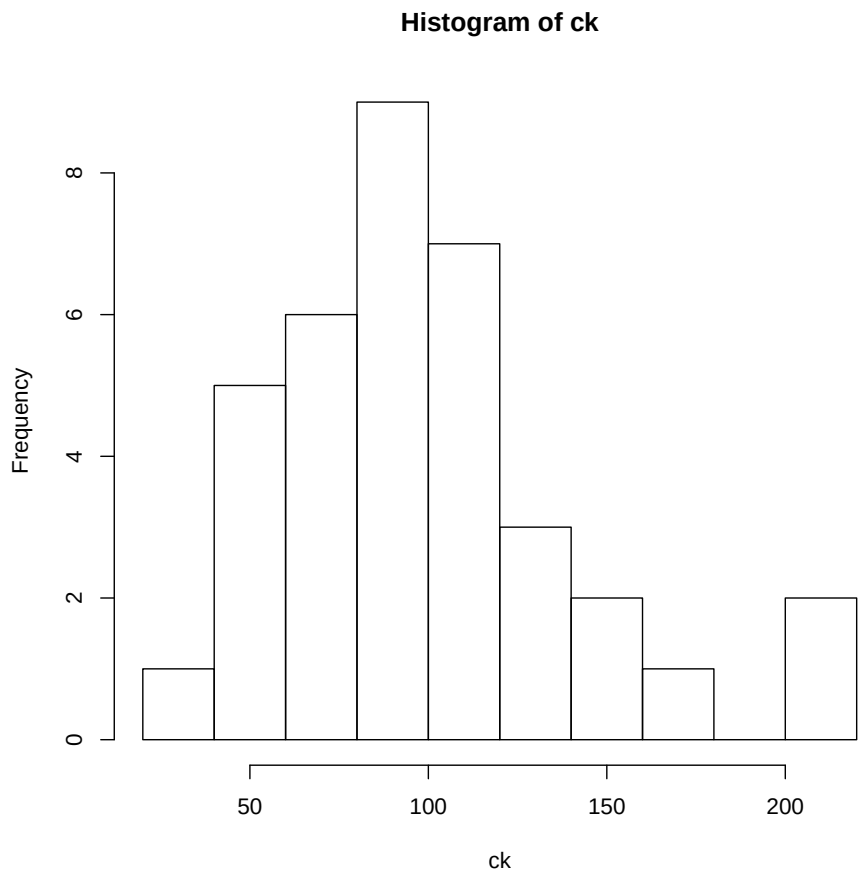



**Figure 2.2.11** Heights of students, using 37 classes (class width = 0.5)

# R code for creatine phosphokinase data

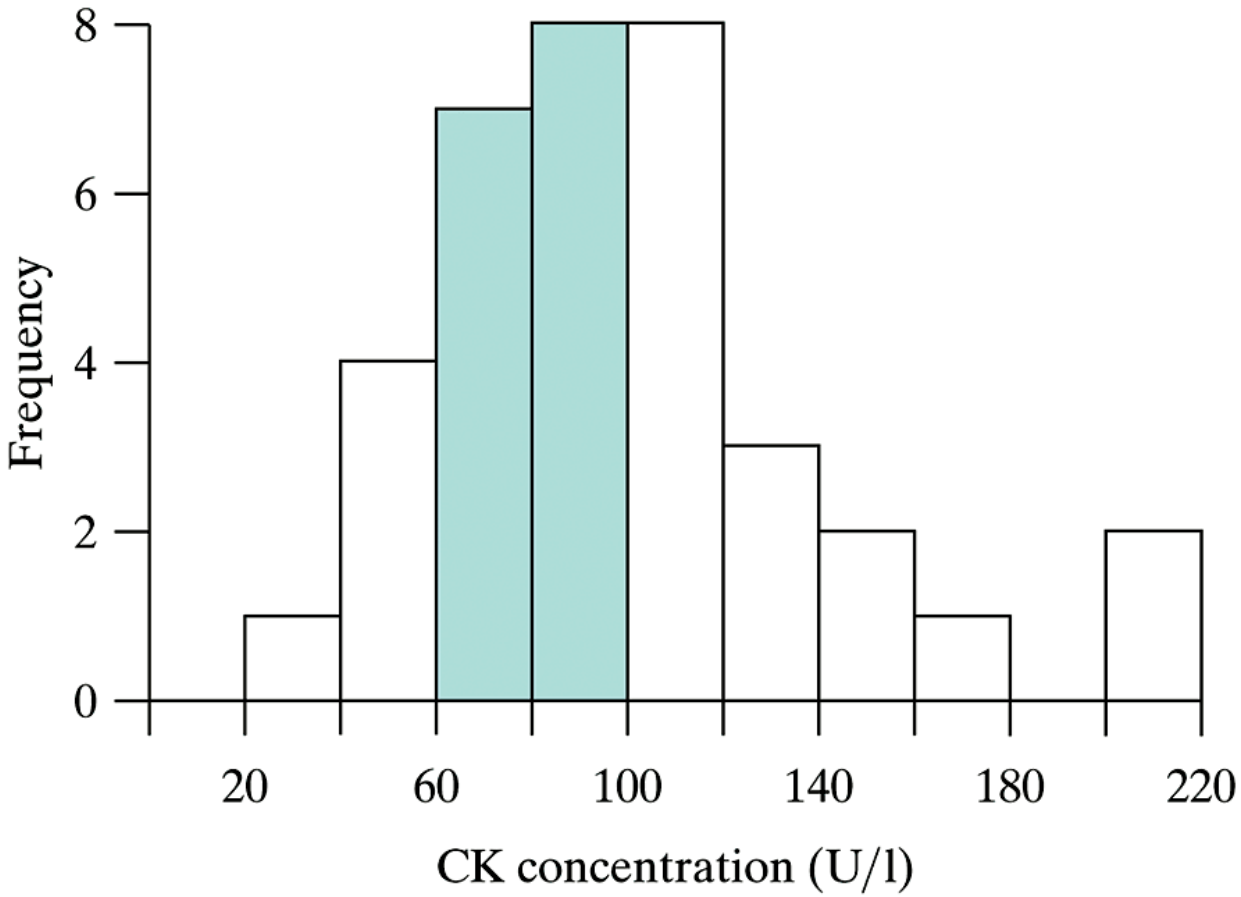
```
ck=c(121,82,100,151,68,58,95,145,64,201,101,163,84,57,139,60,78,94,
     119,104,110,113,118,203,62,83,67,93,92,110,25,123,70,48,95,42)

hist(ck)              # default R histogram
hist(ck,breaks=5)     # 5 classes
hist(ck,freq=FALSE)   # total area is one
```



# Area under histogram

- Area of a single bar proportional to frequency corresponding to class.
- So area of several bars proportional to frequency of several classes combined into one larger class.
- On next slide,
  7/36 CK values between 60 and 80 U/l,
  8/36 CK values between 80 and 100 U/l.
  So $15/36 = 0.42$ or 42% between 60 and 100 U/l.
- Area = *probability* (more later...)

# 42% of CK concentrations between 60 and 100 U/l



# Shapes of distributions

- Shortly we will define smooth curves that approximate histograms in large samples; such a curve is called a **density**.
- A **unimodal** density has one mode; a **bimodal** density has two modes; more than two modes is **multimodal**.
- If one half of the density is the mirror-image of the other, the density is **symmetric**.
- If one tail is longer than the other, the density is **skewed**. The direction of skew (**right** or **left**) indicates which tail is longer.

# Smooth density: unimodal, skewed right

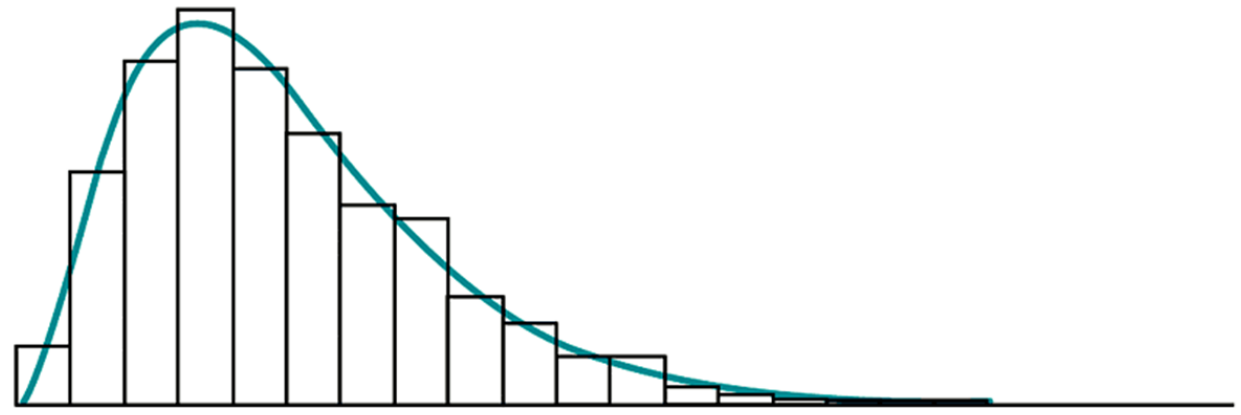
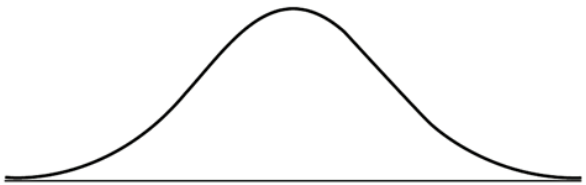
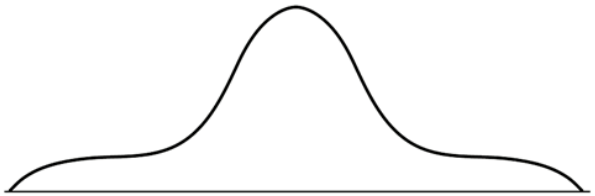


**Figure 2.2.13**
Approximation of a histogram by a smooth curve

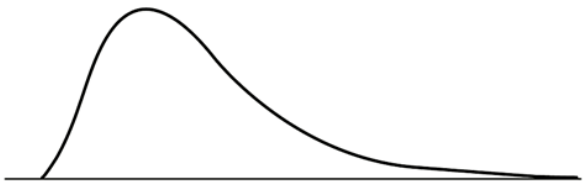
# Density shapes



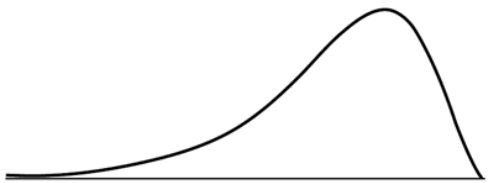
(a) Symmetric, bell-shaped



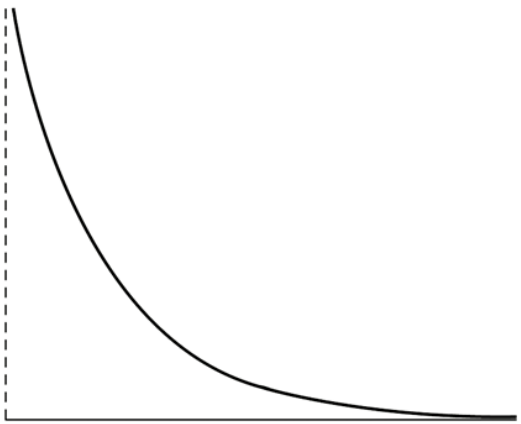
(b) Symmetric, not bell-shaped



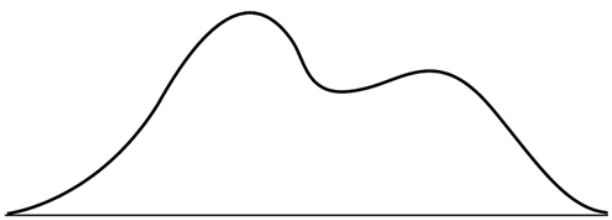
(c) Skewed to the right



(d) Skewed to the left

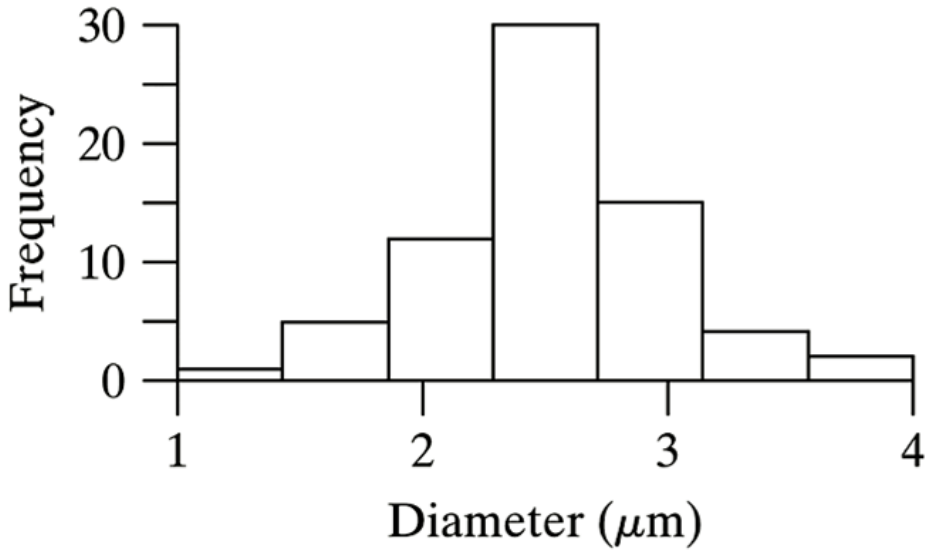


(e) Exponential



(f) Bimodal

# Example 2.2.8: microscopic fossils biological?




In 1977 paleontologists discovered microscopic fossil structures, resembling algae, in rocks 3.5 billion years old. Are they biological in origin? This unimodal and symmetric shaped distribution looks like known microbial populations, but not like nonbiological structures.