Types of Data

Qualitative or **categorical**:

- *Nominal*: blood type (A/B/AB/O), sex (M/F), color, etc.
- *Ordinal*: response to therapy (none/partial/complete), etc.

Quantitative or **numerical**:

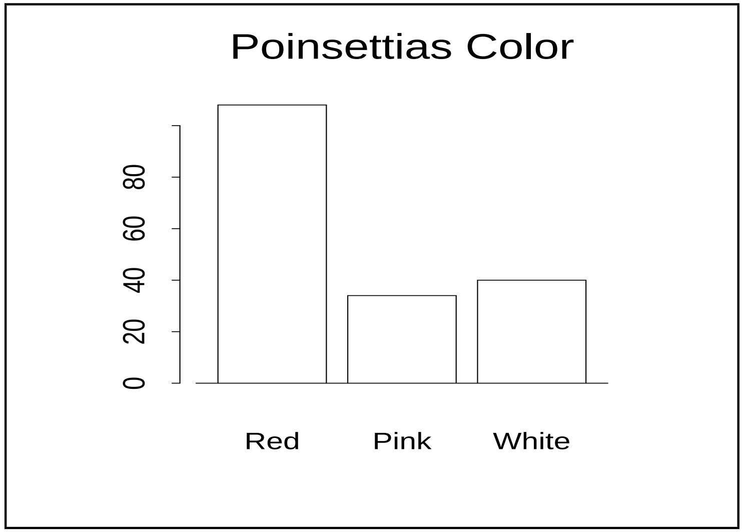
- *Continuous*: weight, concentration, length, etc.
- *Discrete*: number of eggs in nest, etc.

A data set is often called a **sample**. The "readings" are of the **observed variable** taken from the **observational units**. The number of readings in a sample is called the **sample size**.

Bar Plot for Categorical Data

Poinsettias can be red, pink, or white. The color of 182 poinsettias is summarized as follows.

Color	Freq.	Rel. Freq.
Red	108	0.593
Pink	34	0.187
White	40	0.220
Total	182	1.000



- Categories should be **mutually exclusive** and **exhaustive**.
- May use relative frequency on vertical axis. (alt.: pie chart)

Freq. Dist. of Numerical Data

Preening times (sec) of 20 fruitflies during a six-minute observation period are listed below.

SORTED DATA:

10	16	18	19
22	24	24	25
26	29	31	32
33	34	46	48
48	52	57	76

Range: $76 - 10 = 66$

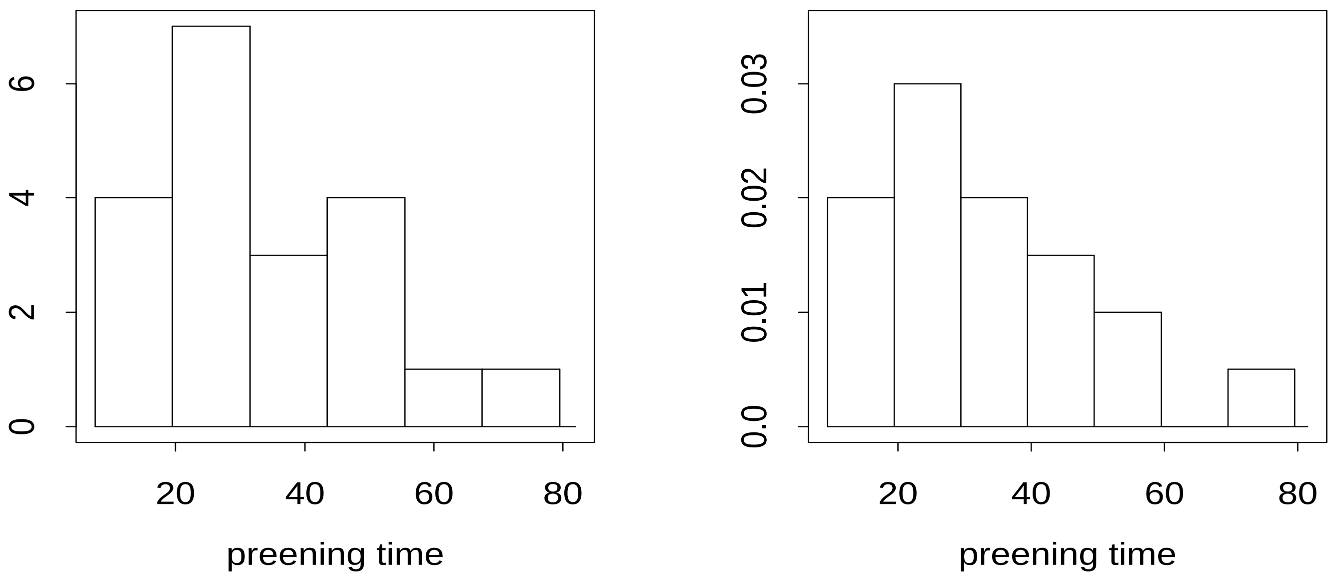
Class	Freq.
8-19	4
20-31	7
32-43	3
44-55	4
56-67	1
68-79	1

Class	Freq.
10-19	4
20-29	6
30-39	4
40-49	3
50-59	2
60-69	0
70-79	1

- The solution is not unique.

Histogram for Numerical Data

A **histogram** is simply a bar plot of frequency distribution.



For **class limits** 10-19, 20-29, etc., one has **class boundaries** 9.5-19.5, 19.5-29.5, etc., and **class width** 10.

More on Frequency Distribution and Histogram

- Classes in a frequency distribution should be *nonoverlapping* and of *equal width*. The latter is for the histogram to convey the correct visual perception of data density.
- There is a class number versus class width tradeoff. More classes (tighter class width) gets more details at the expense of "unstable" global picture.
- To be effective as data summarizing tools, transformations are sometimes needed, as the following example shows.

0.02	0.11	0.18	0.19	0.20	0.28	0.58	0.85	1.18	2.00	7.30
-1.68	-0.97	-0.75	-0.72	-0.71	-0.55	-0.24	-0.07	0.07	0.30	0.86

Stem-and-Leaf Display for Numerical Data

Stem-and-leaf display is a rotated histogram that keeps the "original" data. We use the fruitfly preening time to illustrate.

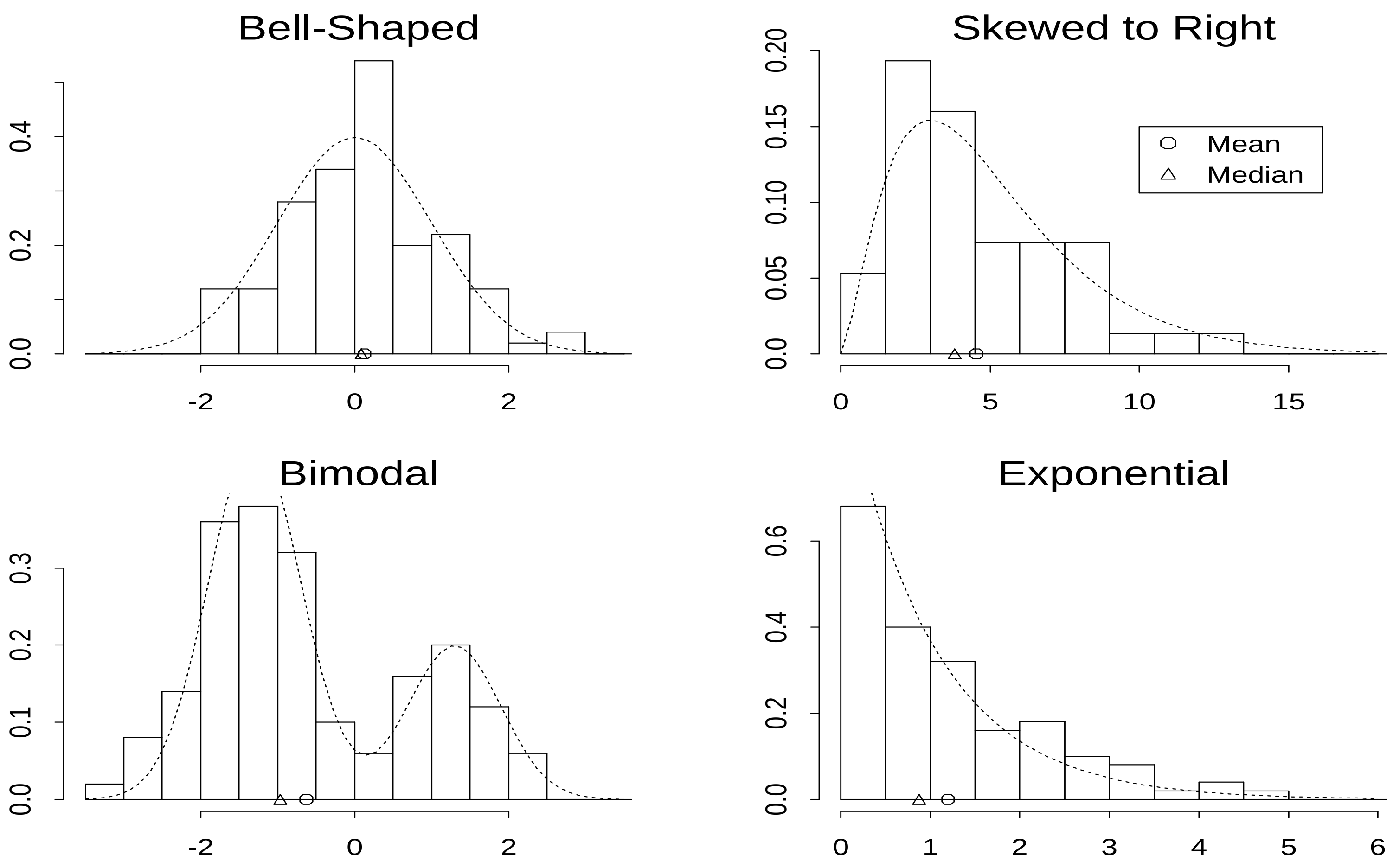
1	0 6 8 9
2	2 4 4 5 6 9
3	1 2 3 4
4	6 8 8
5	2 7
6	
7	6

- Need to specify the decimal place.
- Possible class limits: 2-, 5-, 10-leaf.

	...
1	6 8 9
2	2 4 4
2	5 6 9
3	1 2 3 4
	...

	...
2	
2	2
2	4 4 5
2	6
2	9
	...

Shapes of Frequency Distributions



Some R Commands

Colors of Poinsettias.

```
barplot(c(108,34,40),col=c("red","pink","white"))
pie(c(108,34,40),col=c("red","pink","white"))
```

Preening times of fruitflies.

```
x <- c(10,16,18,19,22,24,24,25,26,29,
       31,32,33,34,46,48,48,52,57,76) ## enter data
## x <- c(scan("file")) ## read data from file
table(cut(x,7.5+(0:6)*12)) ## 6 class freq. dist.
table(cut(x,9.5+(0:7)*10)) ## 7 class freq. dist.
hist(x); hist(x,bre=7.99+(0:6)*12,prob=T) ## histograms
stem(x); stem(x,scale=2); stem(x,s=4) ## stem-and-leaf
```

Measures of Location: Mean and Median

Data are often denoted by $x_1, x_2, \ldots, x_n$, with n the sample size.

Mean: $\bar{x} = \dfrac{x_1 + x_2 + \cdots + x_n}{n} = \dfrac{\sum_{i=1}^n x_i}{n}$.

Median: The number in the middle, that splits x_i's to half-half.

Toy example 1:

Data: 1 2 4 8 6 12

$$\bar{x} = \frac{1+2+4+8+6+12}{6} = 5.5$$

$$\text{Median} = \frac{4+6}{2} = 5$$

Toy example 2:

Data: 4 5 7 6 6

$$\bar{x} = \frac{4+5+7+6+6}{5} = 5.6$$

$$\text{Median} = 6$$

- The mean $\bar{x}$ is most commonly used, but can be misleading for highly skewed data. Consider $\{1, 1, 1, 1, 1, 10\}$: $\bar{x} = 2.5$ is in the middle of nowhere.

Measure of Variability: Standard Deviation

Variance: $s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1} = \frac{S_{xx}}{n-1} = \frac{\sum_{i=1}^n x_i^2 - n\bar{x}^2}{n-1}$.

Standard Deviation: $s = \sqrt{s^2}$.

Toy example 1:

Data: 1 2 4 8 6 12 with $\bar{x} = 5.5$.

$$s^2 = \frac{(1-5.5)^2 + \cdots}{5} = 16.7$$
$$s = \sqrt{16.7} = 4.08$$

Toy example 2:

Data: 4 5 7 6 6 with $\bar{x} = 5.6$

$$s^2 = \frac{(4-5.6)^2 + \cdots}{4} = 1.3$$
$$s = \sqrt{1.3} = 1.14$$

- s^2 is the average squared deviation from $\bar{x}$.
- s has the same unit as x_i's.

Percentiles and Quartiles

Percentile: The $100p$th percentile has $100p\%$ of data at or below it and $100(1-p)\%$ at or above.

Quartile: The 25th, 50th, and 75th percentiles are quartiles.

1	0 6 8 9
2	2 4 4 5 6 9
3	1 2 3 4
4	6 8 8
5	2 7
6	
7	6

$$Q_1 = (22+24)/2 = 23 \quad (np=5)$$
$$Q_2 = (29+31)/2 = 30 \quad (np=10)$$
$$Q_3 = (46+48)/2 = 47 \quad (np=15)$$
$$17\text{th} = 19 \quad (np=3.4)$$
$$93\text{rd} = 57 \quad (np=18.6)$$

Calculation: For $k=np$ an integer, average kth and $(k+1)$st ordered data; o.w. round k up and find the ordered datum.

Alternative Variability Measure

Interquartile Range: $\mathrm{IQR} = Q_3 - Q_1$.

Coefficient of Variation: $\mathrm{CV} = s/\bar{x}$.

1	0 6 8 9
2	2 4 4 5 6 9
3	1 2 3 4
4	6 8 8
5	2 7
6	
7	6

$$\bar{x} = 33.5$$

$$s = 16.31$$

$$Q_2 = 30$$

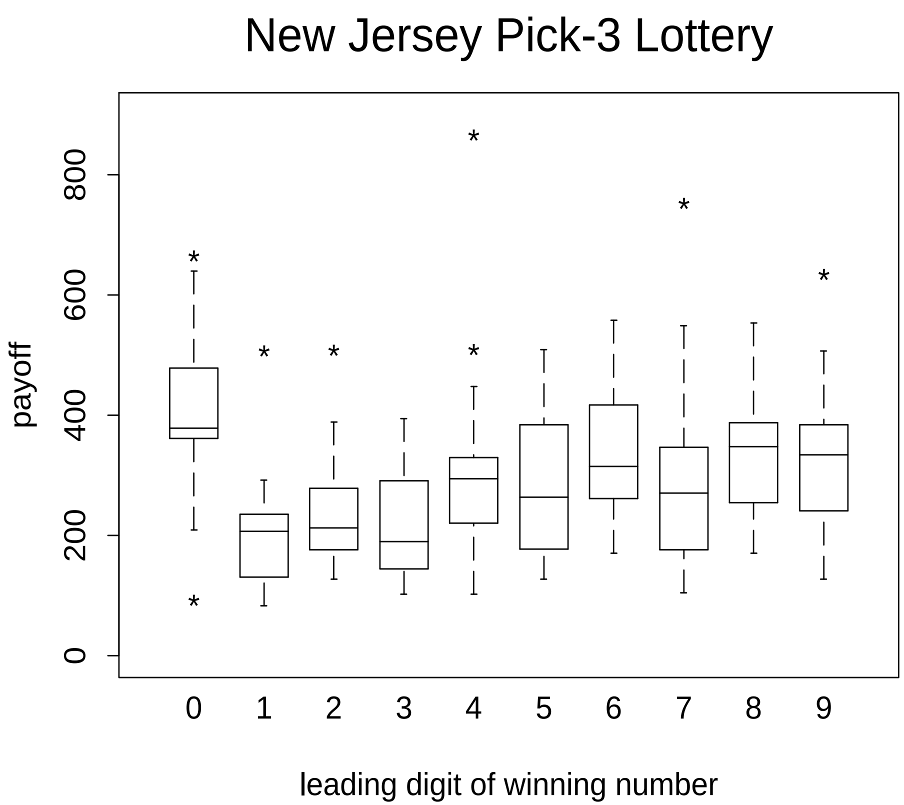
$$\mathrm{IQR} = Q_3 - Q_1 = 47 - 23 = 24$$

$$\mathrm{CV} = s/\bar{x} = 0.487 = 48.7\%$$

- For bell-shaped distribution, $Q_3 - Q_1 \approx 1.35s$.
- CV is unitless and is only meaningful for positive data.

Box Plots

Box plot sketches a distribution in a compact form, and is especially appropriate for comparative purposes.



- The box contains the center half of the data, with Q_1 and Q_3 on the edges and Q_2 inside.
- The lines extend to data within 1.5 IQR from the box.
- Outliers are marked individually.

Linear Transformation

Linear Transformation: $y = ax + b$, where a and b are constants. It shifts and scales but preserves the shape.

- $\bar{y} = a\bar{x} + b$. Similar results hold for other *location* measures.
- $s_y = |a|s_x$. Similar for other *dispersion* measures.
- With $b = 0$ and $a > 0$, $\mathrm{CV}_x = \mathrm{CV}_y$.

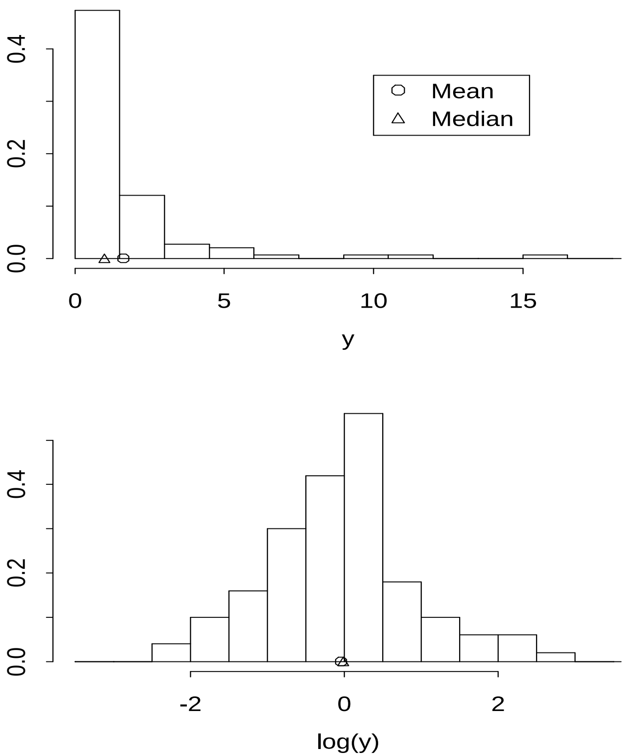
Example: Consider temperature measured in $y^o C$ or $x^o F$.

$$y = \frac{5}{9}(x - 32) = \frac{5}{9}x - \frac{160}{9}.$$

If $\bar{x} = 86$ and $s_x = 9$, then $\bar{y} = 30$ and $s_y = 5$. Note that it does not make sense to compute CV for temperature.

Nonlinear Transformation

Nonlinear Transformation: $x = f(y)$, where $f(y)$ is anything but $ay + b$. Examples of $f(y)$ include $\log(y)$, $\sqrt{y}$, etc.



- Shape of the distribution changes.
- No simple formula for mean and SD.
- Percentiles are "transparent" for *monotone* $f(y)$: for $f(y)$ increasing,

$$Q_3(x) = f(Q_3(y)).$$

Some R Commands

Data summaries and transformations.

```
mean(x); mean(x,trim=.1); median(x) ## location
sd(x); IQR(x) ## variability
mean(2*x+3); sd(2*x+3) ## linear transform
mean(exp(x)); exp(mean(x)) ## nonlinear transform
quantile(x); quantile(x,c(.05,.95)) ## percentiles
quantile(exp(x)); exp(quantile(x))
```

Boxplots.

```
## dump(c("lot.pay","lot.num"),"lottery.R")
source("lottery.R") ## restore dumped data
boxplot(split(lot.pay,lot.num%/%100))
boxplot(x,x+10,x-20)
```

Samples and Population

One usually collects **samples** to learn about **population**.

Poinsettias color: Observing 108 reds out of 182, can we conclude that about 60% of all poinsettias are red?

Fruitfly preening time: Seeing 10 of 20 fruitflies preen less than 30 sec, can we say half of all fruitflies preen less than 30 sec?

	Popu	Smpl
Mean	μ	$\bar{x}$
SD	σ	s
Prop	p	$\hat{p}$
Dist	dsty	hist

Sampling draws samples from population.

Inference infers population from sample.

- Samples should represent population.
- Inference is always with error.

Description of Data: Summary

Bar plot, histogram, stem-and-leaf display, and box plot plot **frequency distributions** which summarize data.

Location measures: mean, median, quartiles, etc.

Variability measures: SD, IQR, CV, etc.

Linear transformations shift and scale but do not reshape distributions, **nonlinear** ones change everything.

Samples serve as windows for us to look into **population**.