CH 3 & 4 – Selected Parts

1. Types of data: Categorical, Numerical
2. Describing data for a single (numerical) variable
   - Summary Statistics
   - Graphs
3. Random variables and probability distributions
4. The Normal (Gaussian) distribution
5. The "Empirical Rule" and Chebyshev's Rule
6. Sampling distribution of the sample mean

<u>Examples:</u>

1. Normal probabilities in R

1. Types of Data

- **Categorical/Qualitative/Factor Variables:** can be placed into categories.

  Examples: Eye Color, Gender

  Note: Can be coded as numbers (Ex: M=0, F=1)

- **Numerical/Quantitative/Measurement Variables:** those for which we can record a numerical value and then order respondents according to those values.

  Examples: Age, Time

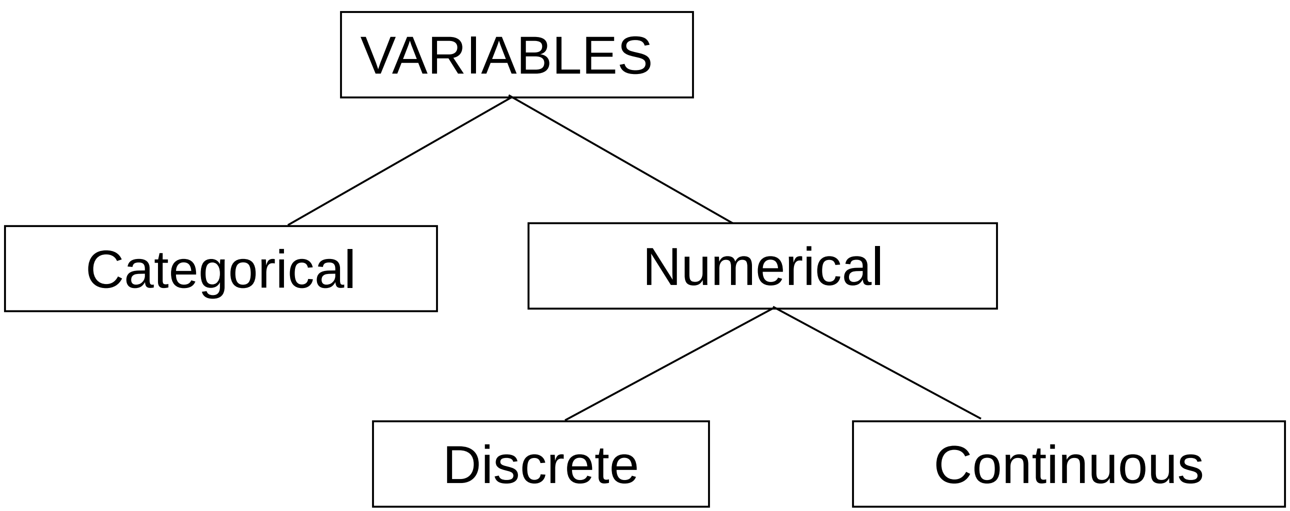
  Note: Can be further categorized as discrete or continuous, but distinction not always clear.

- **<u>Discrete</u> Variables** can only take some values; often obtained by counting.

  Examples: Number of Children, Age (in years)

- **<u>Continuous</u> Variables** can take any value within a given interval.

  Examples: Height, Weight



2. Describing Data for a Single (Numerical) Variable

Measures of Central Tendency

- The **mode** is the value that occurs most often (with the highest frequency).
- The **median** is the middle value in the ordered data set.
- The **mean** (denoted $\bar{y}$) is the sum of the values divided by the number of observations.

$$n = \text{sample size}$$

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

- The **pth percentile** of a set of n measurements arranged in order is the value that has p% of the measurements below it.
- Hence the median is the 50^{th} percentile.
- Q1 is the 25th percentile and Q3 is the 75^{th} percentile.
- The "**five number summary**" includes min, Q1, median, Q3 and max values for a data set.
- We will see that a boxplot is graphical display of the five number summary.

Measures of Variability

- The **range** is the difference between the largest and smallest values.
- The **interquartile range (IQR)** is the difference between Q3 (the 75th percentile) and Q1 (the 25th percentile).
- The **variance** (s^2) and **standard deviation** (s) also measure variability.

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

An Important Reminder

- Recall that the <u>population</u> of measurements is a <u>complete</u> set of measurements. A <u>sample</u> is a <u>subset</u> of measurements selected from the population of interest.
- The <u>population mean</u> is denoted μ (mu); the <u>sample mean</u> is denoted $\bar{y}$.
- The <u>population standard deviation</u> is denoted σ (sigma); the <u>sample standard deviation</u> is denoted s.

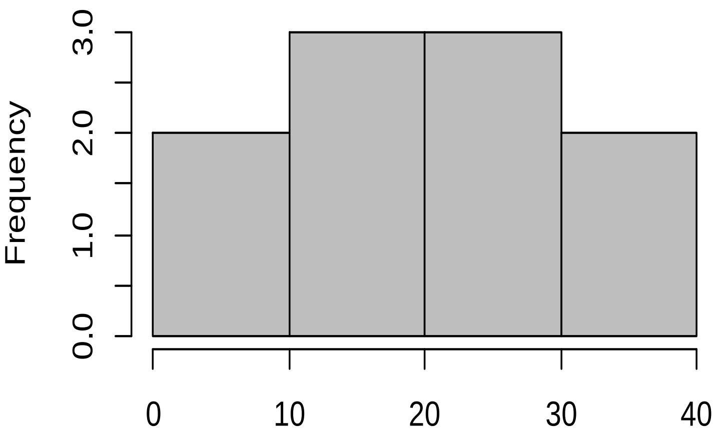
Graphs for a Single (Numerical) Variable

- Exploratory Data Analysis (EDA): We should spend more time looking at the data, and less time modeling. Advocated by the group at Bell Labs including John Tukey.
- Common graphics for a single numerical variable are histograms and boxplots.

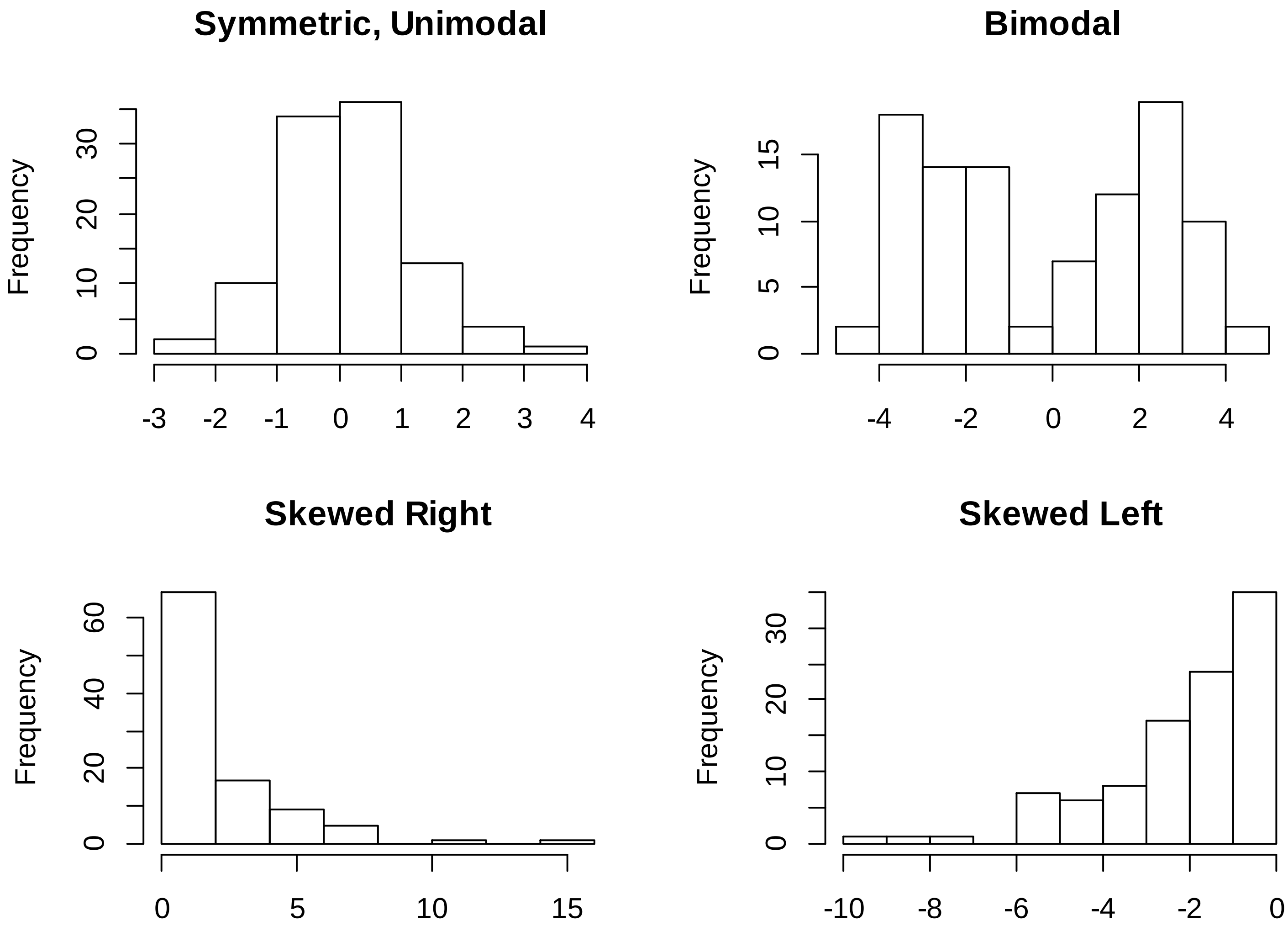
Histograms

- Start with some equally spaced intervals.
- Count the # of observations (or frequency) that fall into each interval.
- Relative frequency is the frequency divided by the total # of observations (n).
- Histogram is a graph of the frequencies or relative frequencies.

Interval	Freq	Rel Freq
0 - 9	2	2/10 = 0.2
10 – 19	3	3/10 = 0.3
20 – 29	3	3/10 = 0.3
30 - 40	2	2/10 = 0.2

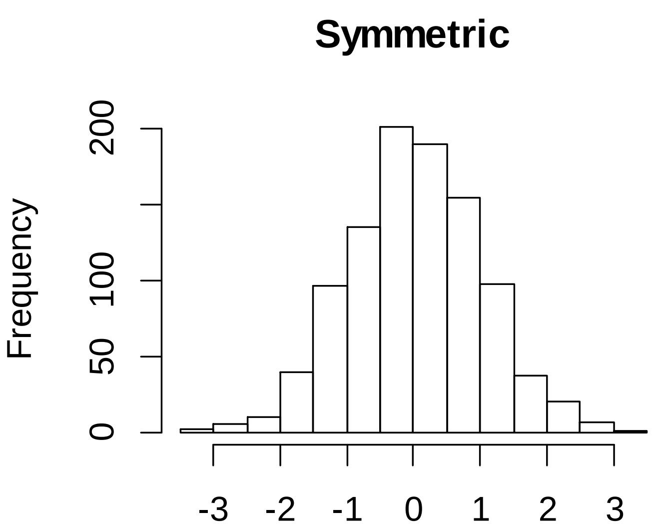


Some Example Histograms

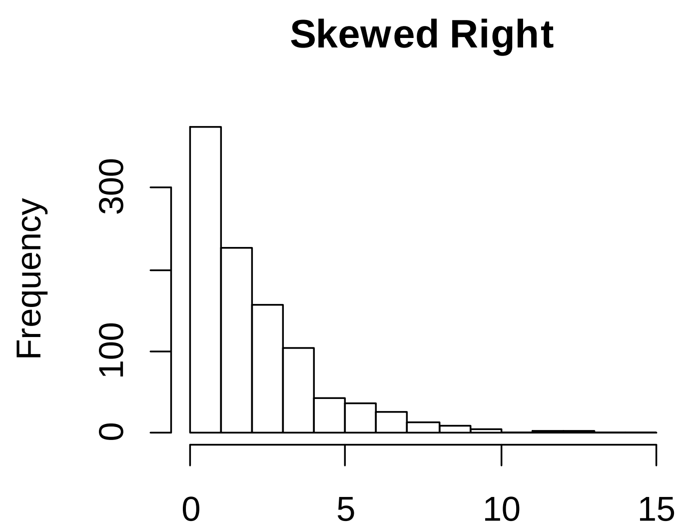


Mean vs Median

- For <u>symmetric</u> distributions, the sample mean and median will be close.
- For <u>skewed</u> distributions, the sample mean and median can be very different.

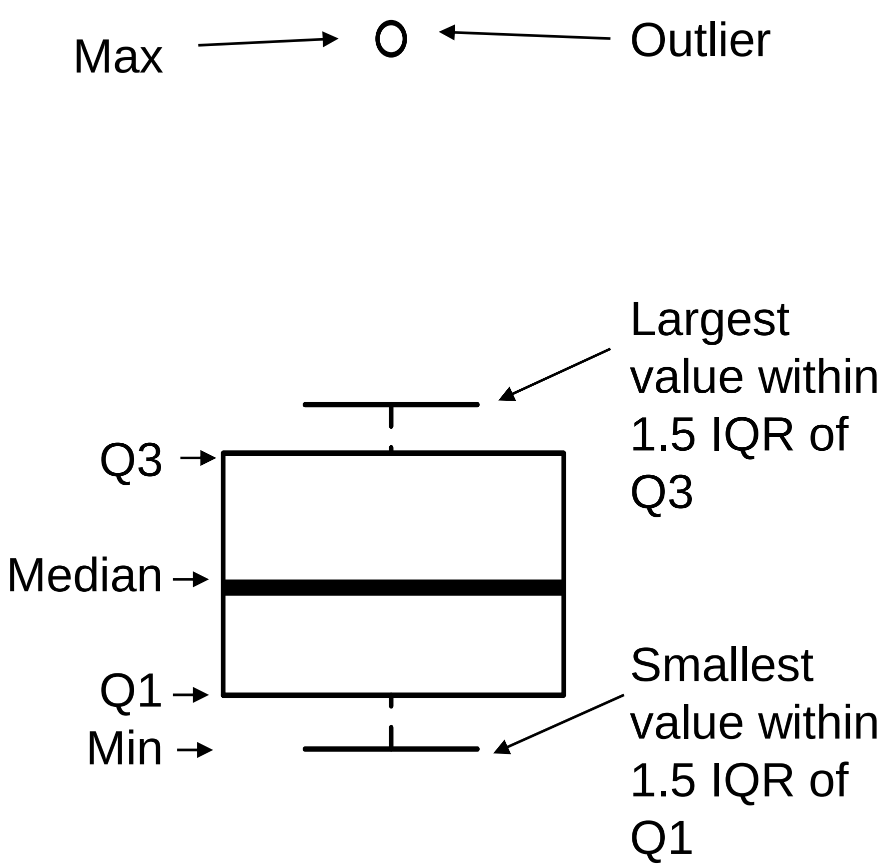


mean = 0.03, median = 0.03



mean = 2.08, median = 1.46

Boxplots



- The boxplot is a graph of the 5 number summary (min, Q1, median, Q3, max) with outliers marked.
- One definition of an **outlier** is a value that lies more than 1.5 IQR from Q1 or Q3. Recall that IQR = Q3 – Q1.

3. Random Variables

- **Probability** is a numerical quantity that expresses the likelihood of an event. Probabilities take values between 0 and 1.
- Relative frequency interpretation of probability: The probability of an event is interpreted as the relative frequency (proportion of times) the event occurs in an indefinitely long series of repetitions of the chance operation.
- **Example:** Single flip of a fair coin.

  P(Heads) = 0.5
- In a long series of tosses of a fair coin, we expect to get Heads about 50% of the time.

- A **random variable** (RV) Y is a variable whose value results from a measurement on some type of random process.
- A **probability distribution** for a RV is a description of the probabilities for all possible outcomes. Total probability equals 1.
- For discrete RVs, the distribution can be summarized as a table, graph or formula. Sum of probabilities must equal 1.
- For continuous RVs, the distribution is summarized as a formula to describe a curve. The area under the curve must equal 1.

Example of a Discrete RV

Let Y be a random variable that gives the outcome of a single roll of a fair die.

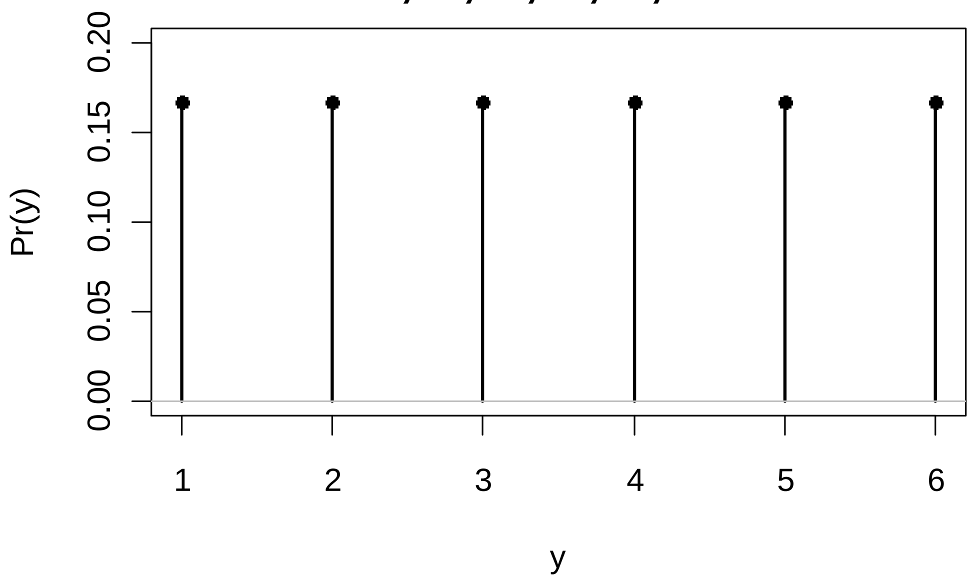
Table:

y	1	2	3	4	5	6
P(Y=y)	1/6	1/6	1/6	1/6	1/6	1/6

Formula:

P(Y=i) = 1/6 for i=1,2,3,4,5,6.

Graph:



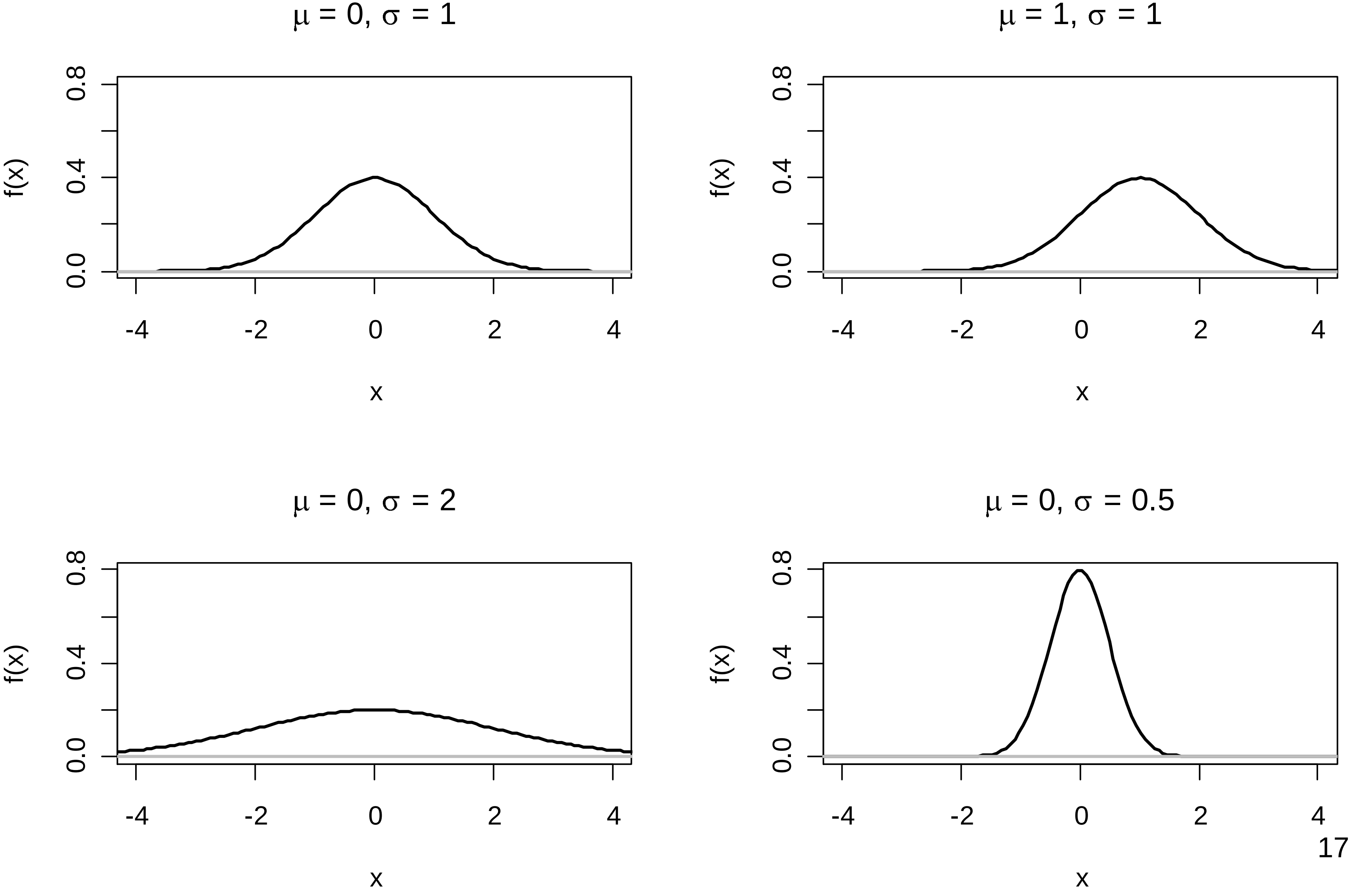
4. Continuous Example: the Normal (Gaussian) family of distributions

- Many populations can be described by a normal distribution.

- Each normal distribution is defined by it's mean (μ) and standard deviation (σ).
- If a variable Y follows a normal distribution with mean μ and standard deviation σ, then we write *Y~N(μ, σ).*
- All normal curves can be described by a single formula:

$$f(y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2}$$

Normal Distribution Examples



Example: Assume Z is normal with $\mu=0$, $\sigma=1$ ("Standard Normal", $Z \sim N(0, 1)$).

- Ex1: Find $P(Z \leq 1.31)$. (Use Table 1 in O&L)

R: pnorm(1.31)

- Ex2: Find $P(Z > 1.72)$

  R: 1-pnorm(1.72)
- Ex3: Find z such that $P(Z > z)=0.95$.

  (Use Table 1 from the inside out.)

  R: qnorm(0.05)

Standardizing Variables

- If Y has a normal distribution with mean μ and standard deviation σ ($Y \sim N(\mu, \sigma)$),
- Then $Z = (Y - \mu)/\sigma$ has a standard normal distribution ($Z \sim N(0, 1)$).
- Strategy for solving problems for non-standard normal distributions:
  - Standardize both sides (subtract mean and divide by standard deviation)
  - Calculate probabilities based on standard normal distribution using Table 1 or R function pnorm.

Example: Suppose $Y \sim N(\mu=5, \sigma=2)$.

- Ex4: Find $P(Y \leq 8)$

  R: pnorm((8-5)/2) or pnorm(8,mean=5,sd=2)

- Ex5: Find y such that $P(Y \leq y)=0.975$.

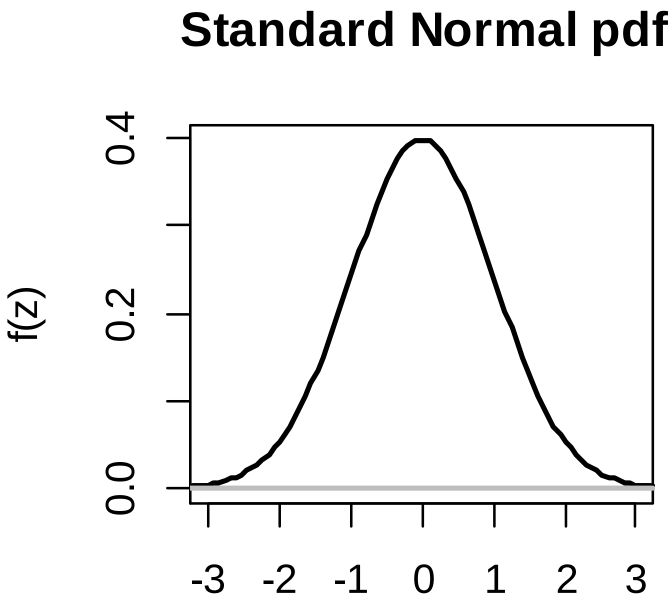
  R: 2*qnorm(0.975)+5 or qnorm(0.975,mean=5,sd=2)

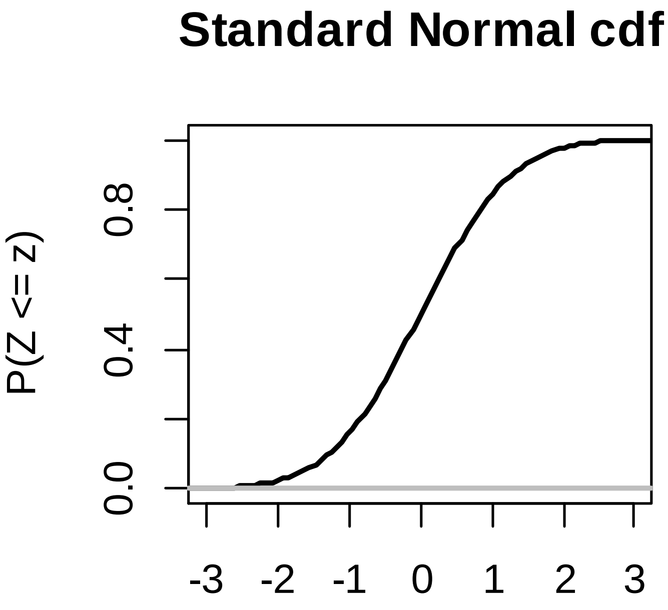
Normal Probabilities in Rcmdr

- Go to Distributions -> Continuous -> Normal.
- Enter the appropriate mean (μ) and standard deviation (σ).
- To find $P(Y \leq y)$, choose Normal Probabilities. Enter the value of y and make sure “Lower Tail” is selected.
- To find a percentile, choose Normal Quantiles. Enter the percentile (on the 0-1 scale) and make sure “Lower Tail” is selected.

Normal pdf vs cdf

- The normal probability density function (**pdf**) is like a smooth relative histogram. By definition the total area under the curve must equal one.
- The normal cumulative distribution function (**cdf**) gives the P(Y $\leq$ y).





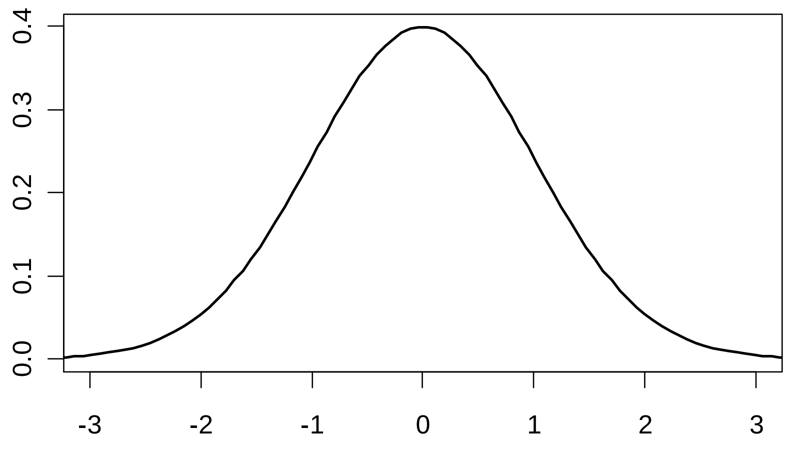
5. The Empirical Rule

For normal distributions (with sample mean $\bar{y}$ and sample standard deviation s):

Approx. 68% of the data lie within $\bar{y} \pm s$

Approx. 95% of the data lie within $\bar{y} \pm 2s$

Approx. 99.7% of the data lie within $\bar{y} \pm 3s$



Note: Doesn't work for skewed distributions

Ex: insect counts, blood hormone concentrations

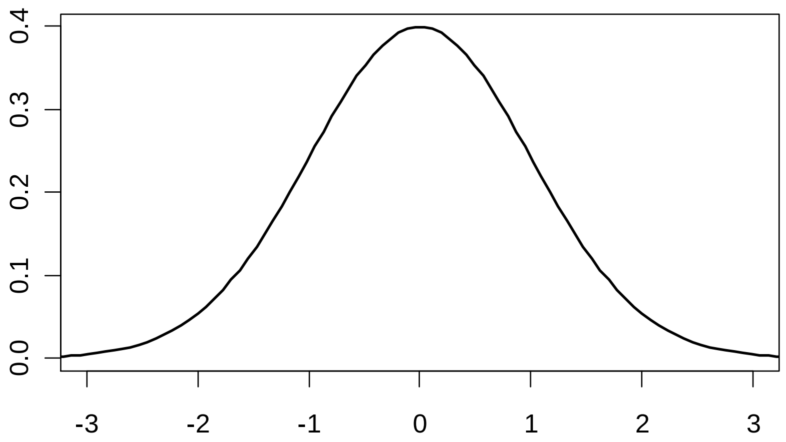
Chebyshev's Rule

For <u>any distribution</u>:

At least 75% of the data lie within $\bar{y} \pm 2s$

At least 88.8% of the data lie within $\bar{y} \pm 3s$

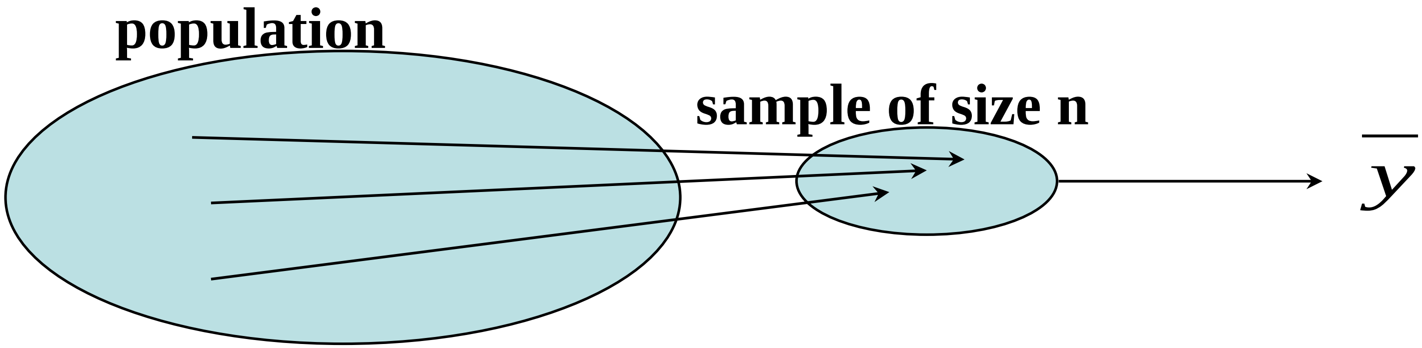
At least 93.75% of the data lie within $\bar{y} \pm 4s$



NOTES:

1. This is weaker than the Empirical rule (75% < 95%)
2. The general version of Chebyshev's rule is:
   At least $(1-1/k^2)$x 100% of the data lie within $\bar{y} \pm ks$

6. Sampling distribution of the sample mean



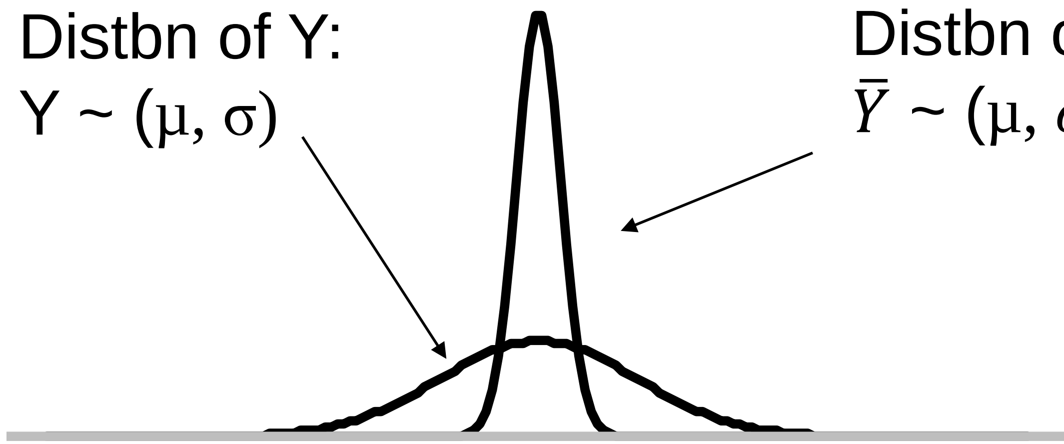
- We can imagine repeating the procedure (taking another sample of size n and finding another sample mean). Suppose we repeated this 1000 times. What distribution would these means have?
- In practice, we don't usually take repeat samples; we are imagining what would happen if we did, in order to better interpret the one sample we do take.

If $\bar{y}$ is the mean of a sample of size n taken from a population (of Y) with mean μ and standard deviation σ, then $\bar{Y}$ itself is a random variable.

Hence, there are two kinds of RV's being discussed here: (1) individual Y and (2) $\bar{Y}$. Neither of these is assumed to be normal (so far).

Population Distbn of Y: Y ~ (μ, σ)

Sampling Distbn of $\bar{Y}$: $\bar{Y}$ ~ $(\mu, \sigma/\sqrt{n})$



1. The mean $\bar{Y}$ of is μ.
2. The standard deviation of $\bar{Y}$ is $\sigma/\sqrt{n}$. (Population size N needs to be “very large”).
3. If the distribution of Y is **normal**, the distribution of $\bar{Y}$ will also be normal (for any n).
4. If the distribution of Y is **non-normal**, the distribution of $\bar{Y}$ will become close to normal as n gets large (**The Central Limit Theorem**).
5. The closer the distribution of Y is to normal, the smaller the n required for the distribution of $\bar{Y}$ to be approximately normal.

Try it:
http://onlinestatbook.com/stat_sim/sampling_dist/index.html