Summarizing data

Evidence-Based Medicine Lecture Series

Elizabeth G. Hill, PhD

Associate Professor of Biostatistics

30 August 2011

Outline

- Some basics
- Variable types
- Describing a continuous variable's distribution
  - Shape
  - Center
  - Dispersion
- Time-to-event variables
- Describing a categorical variable's distribution
- Describing joint distributions of two continuous variables

Some basics

TARGET (STUDY) POPULATION: The patients of clinical or scientific interest. If a study is well-designed, any conclusions drawn from study results generalize to the target population.

SAMPLE: The patients of clinical or scientific interest that participate in a study. In a well-designed study, these subjects are carefully selected from the target population and the sample is said to be *representative*.

VARIABLE: A quantity or trait of interest in the study population.

DATA: Quantitative measurements of variables obtained from subjects in the sample population.

Variable types

Depending on how its values are measured, a variable is (broadly) classified as either *continuous* or *categorical*.

- Continuous variables:
  - measured on a *continuum*
  - often have units of measure (e.g. pg/ml; kg/m^2)
  - typically reflect the quantity of that which is measured
- Categorical variables:
  - values represent discrete ‘levels’ or ‘classes’ (e.g. Male, Female)
  - value associated with class labels are not inherently meaningful (e.g. 1 = Male, 2 = Female)
  - can be *ordinal* (i.e. ordered) or *nominal*

Variable types (cont.)

In the article by Ohtsu et al. (JCO 2011), there are a number of continuous and categorical variables.

Variable	Continuous (units)	Categorical	
		Nominal	Ordinal
Sex		Nominal	
Age	Continuous (years)		
ECOG PS			Ordinal
Geographic region		Nominal	
Primary tumor site		Nominal	
Overall survival	Continuous (months)		
Response			Ordinal

Describing a continuous variable's distribution

When we collect data on study subjects, one objective is to describe how different variables are distributed in the target population.

For example, we might wish to summarize age in our population. We would therefore measure the ages of subjects in our sample, and infer to the target population.

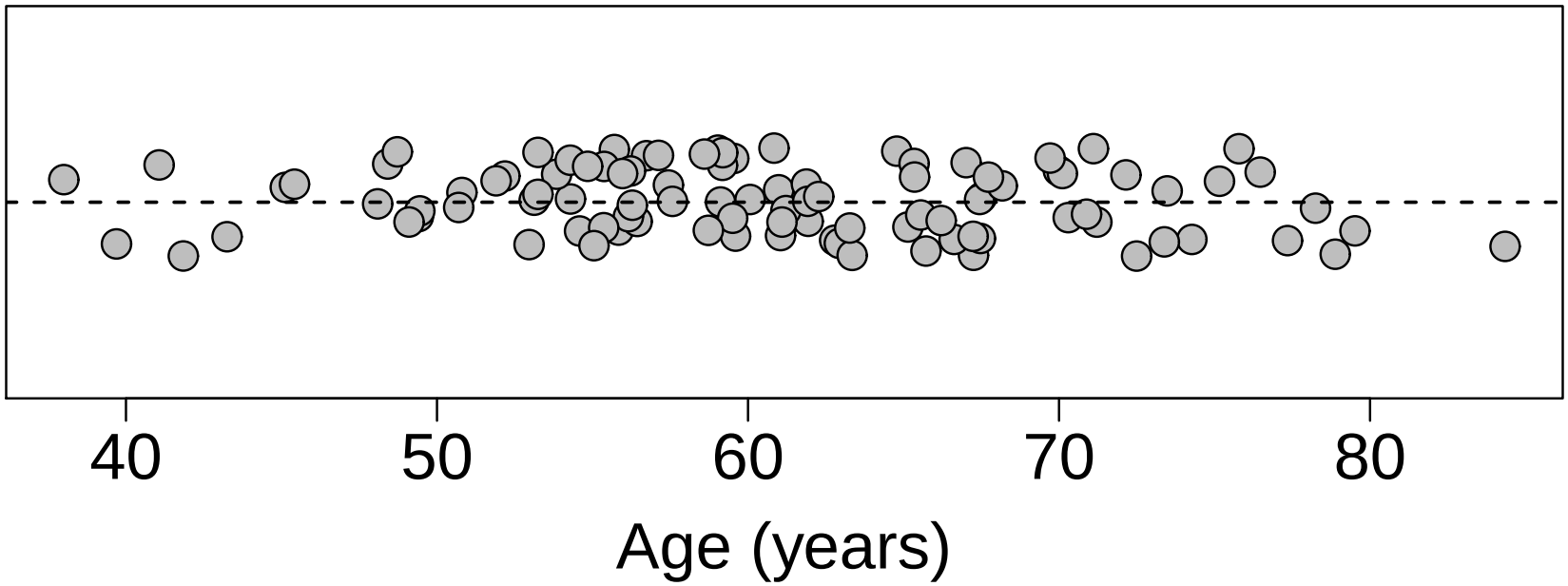
Q How do we generalize from sample data to target population?

A We use statistics.

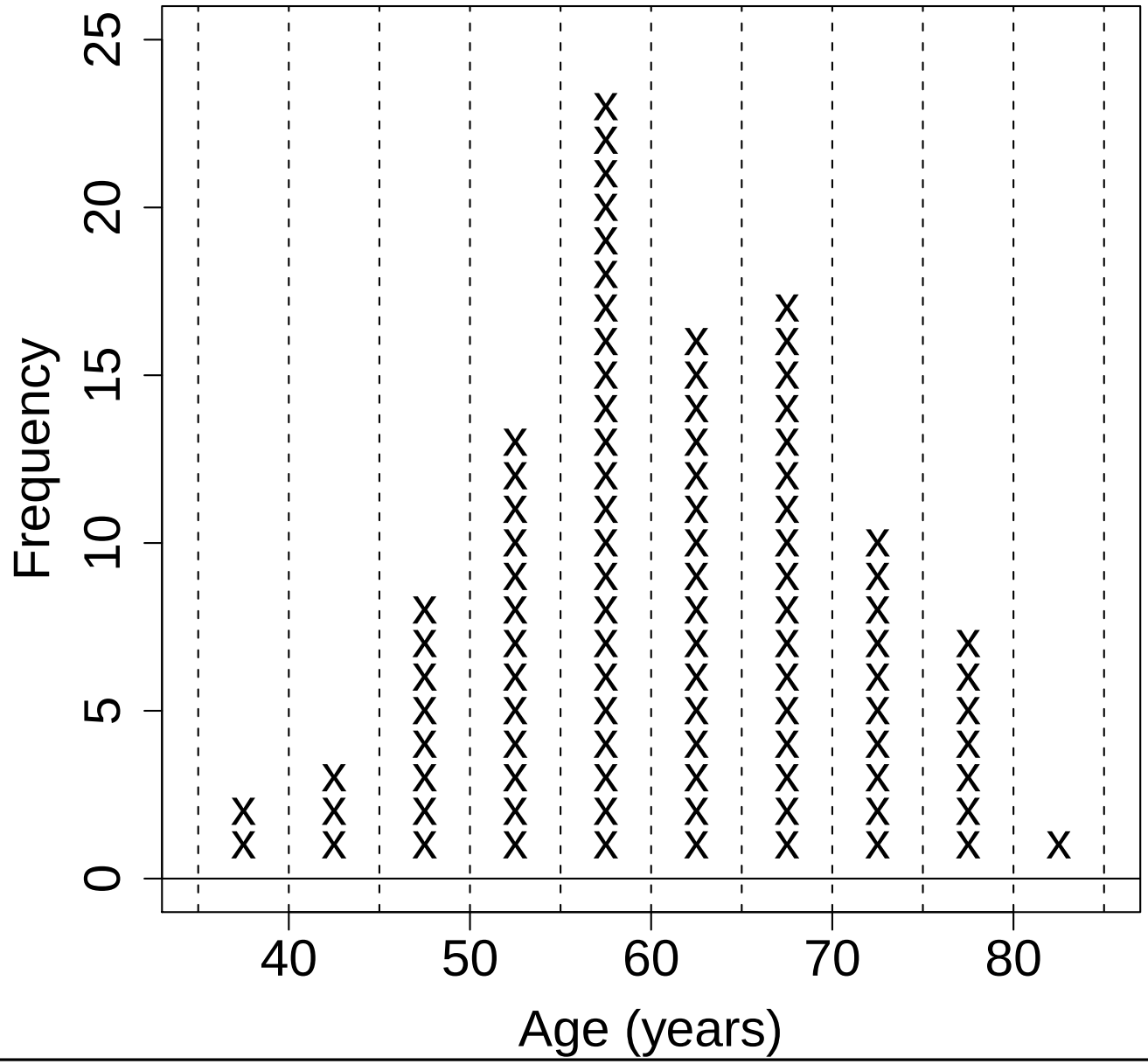
A *statistic* is a summary measure based on sample data. Statistics can be displayed numerically or graphically. What statistics should we construct to adequately summarize a variable's distribution?

Describing shape

Consider a plot of ages of 100 subjects in a study. The points are jittered to enhance visibility. If we group the data in five-year age bands (35-40, 40-45, ... , 75-80, 80-85) we begin to see the underlying structure of the distribution of age in our population.



Describing shape (cont.)

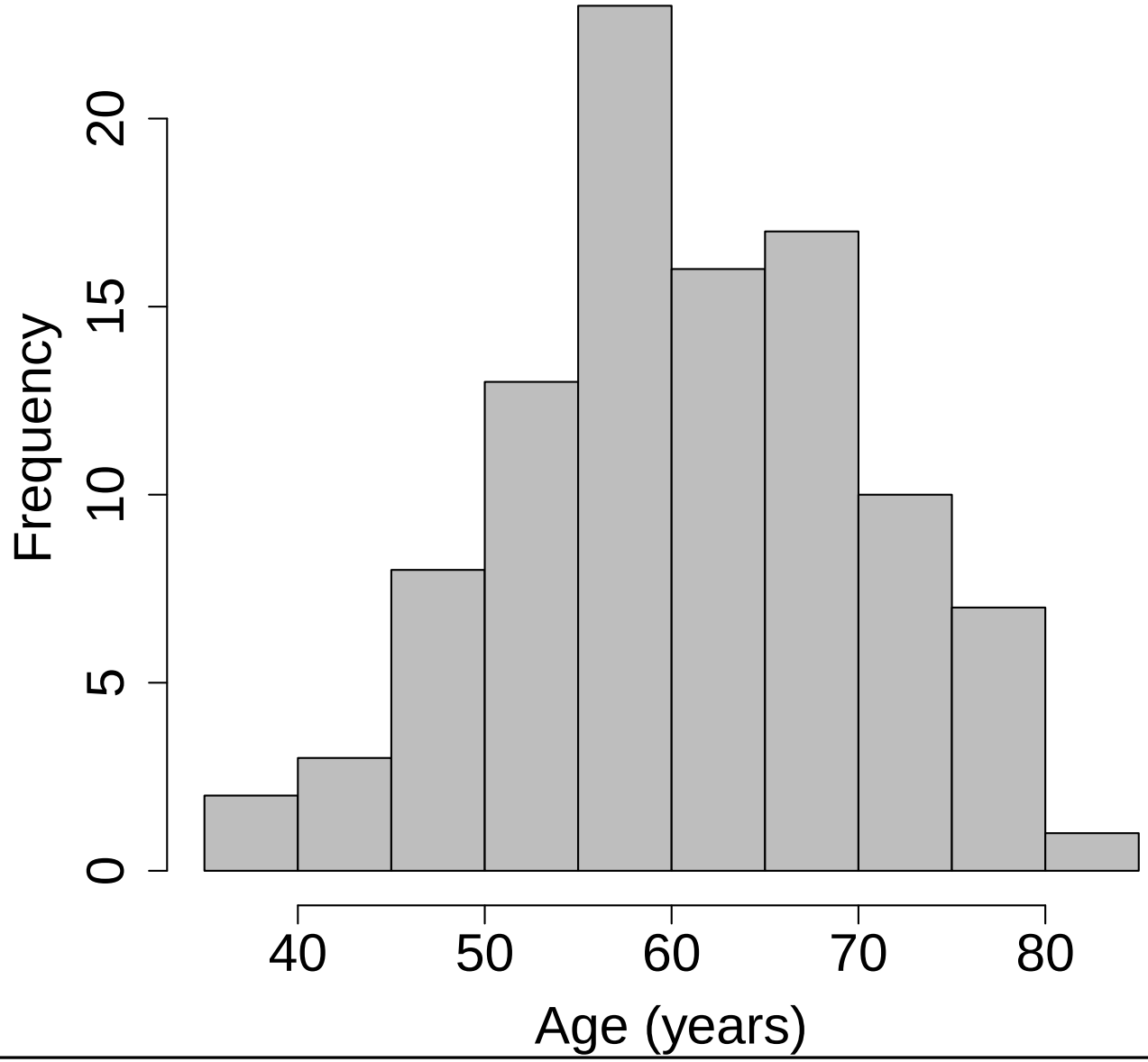


Describing shape with a histogram

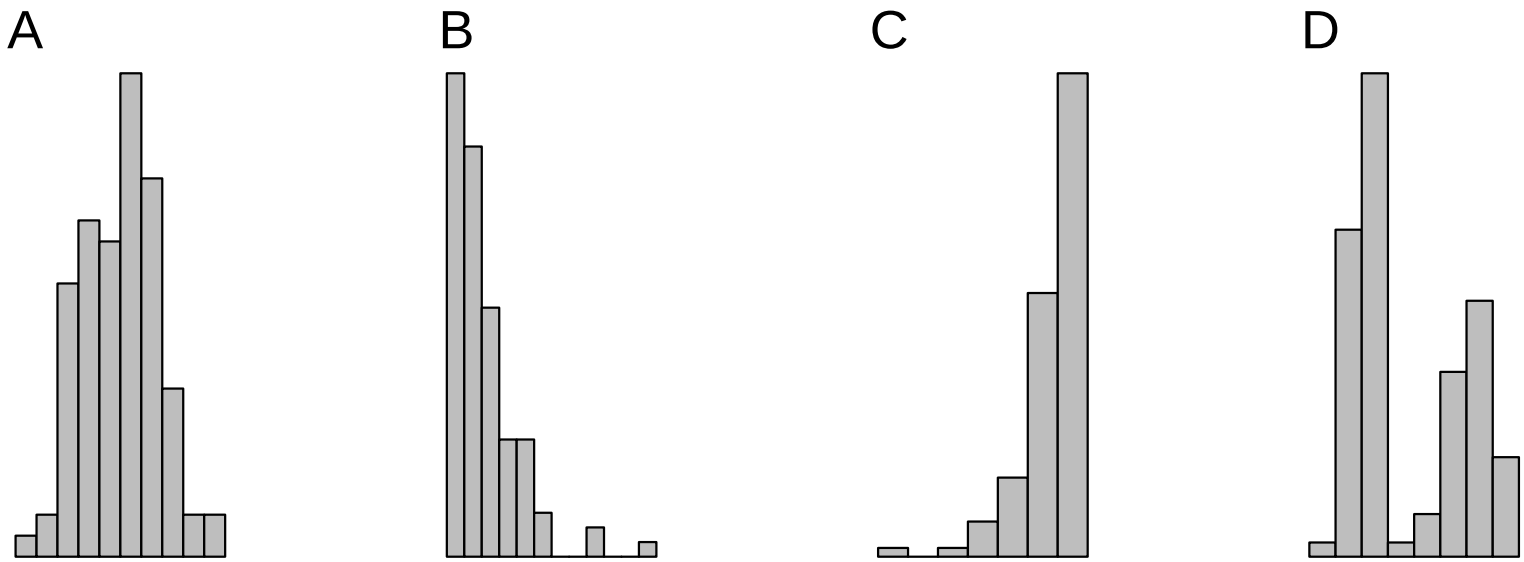
The figure shown on slide 8 illustrates the steps in constructing a *histogram*. The histogram is a useful graphical tool that succinctly depicts the distribution of a continuous variable. The individual groupings of values are called *bins*. We complete the histogram by plotting contiguous bars with heights showing the frequency (or proportion) of the measures occurring within each bin. For the figure shown on slide 8, the graph on slide 10 shows the completed histogram.

(*N.b. histogram* $\neq$ *bar chart*)

Describing shape with a histogram (cont.)



Shape terminology



A: Approximately symmetric and unimodal

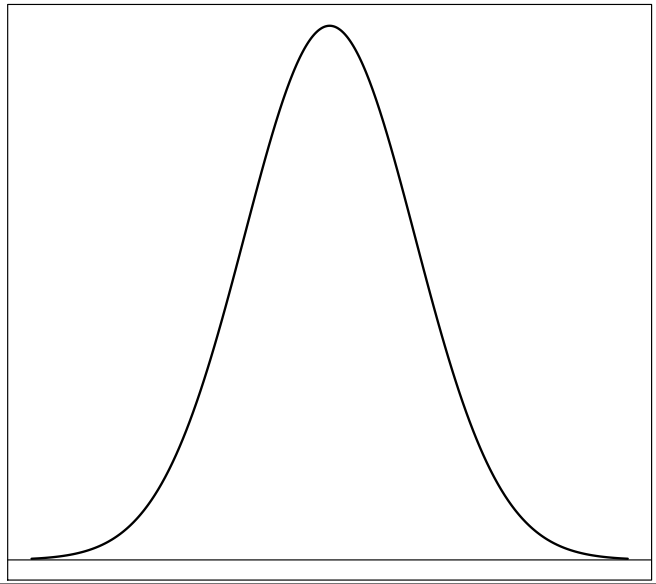
B: Positively skewed and unimodal

C: Negatively skewed and unimodal

D: Asymmetric and bimodal

The normal distribution

Symmetric unimodal distributions describe many variables we encounter in biologic and medical research. This distribution shape is so common, we call it the *normal* distribution. Normally distributed variables have many important properties in addition to symmetry and uni-modality (some of which we'll discuss later in this lecture). By examining the shape of a continuous variable's distribution with a histogram, we can gauge whether that variable is approximately normally distributed.



Describing central tendency

Beyond shape, an obvious summary measure for a continuous variable is one describing *central tendency*. A measure of central tendency answers the question:

"*Around what value do measures tend to aggregate?*"

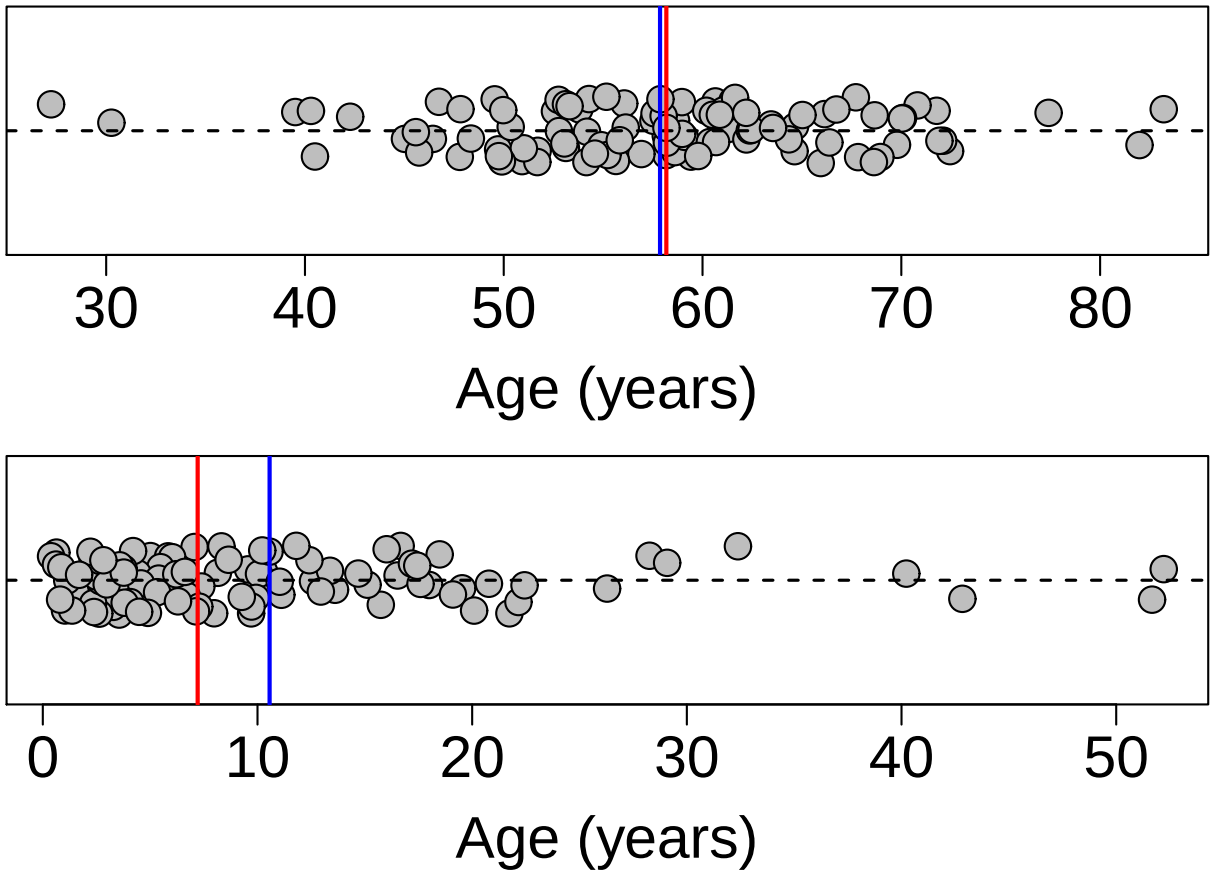
The two most common measures of the center of a continuous variable's distribution are

- sample mean (= arithmetic average)
- sample median (= 50th percentile)

If we know a patient is a member of the target population, our best guess of their value of the variable of interest is the measure of central tendency.

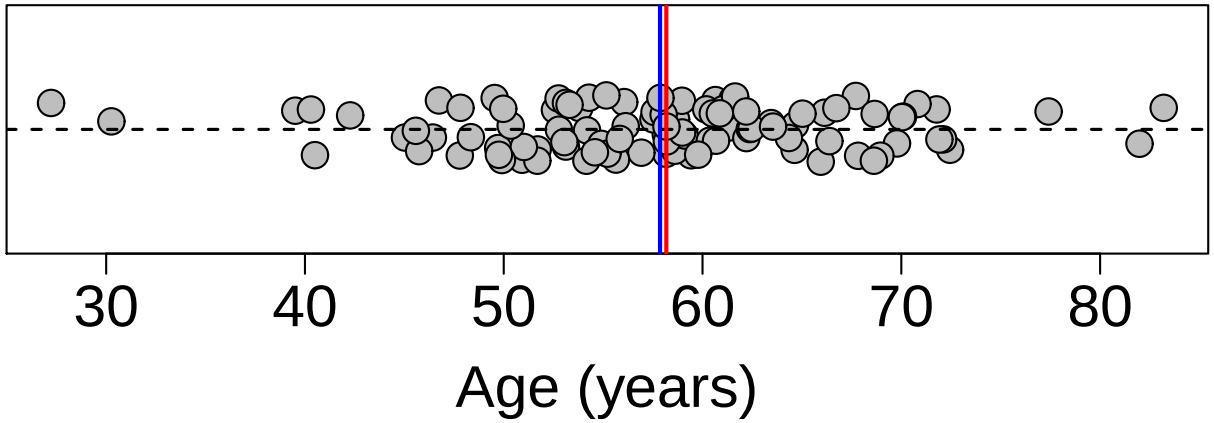
Example - Age central tendency

In the figures below, a sample of 100 ages are plotted for two different populations. The blue lines show the locations of the sample mean, and the red lines show the locations of the sample median.

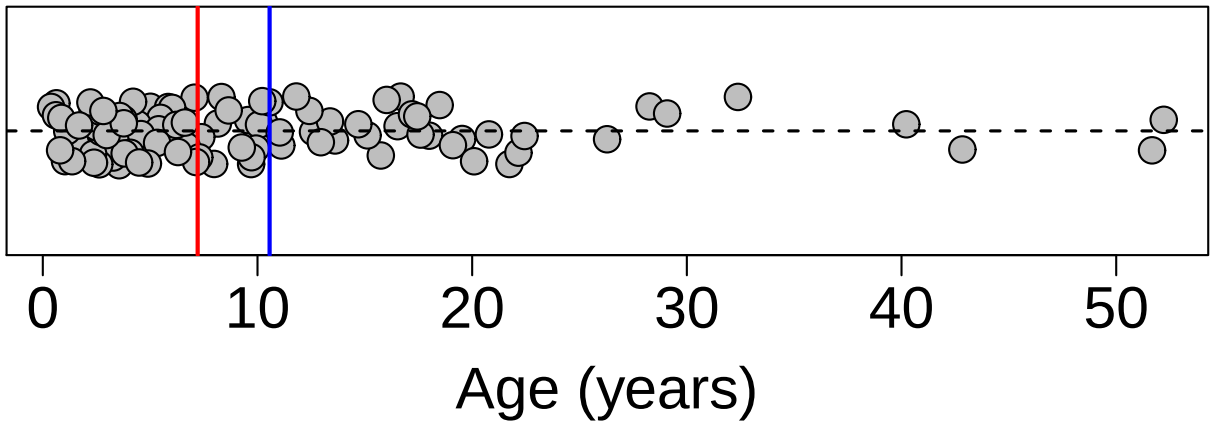


Example - Age central tendency (cont.)

In the first example, both the mean and median adequately describe the distribution's center.

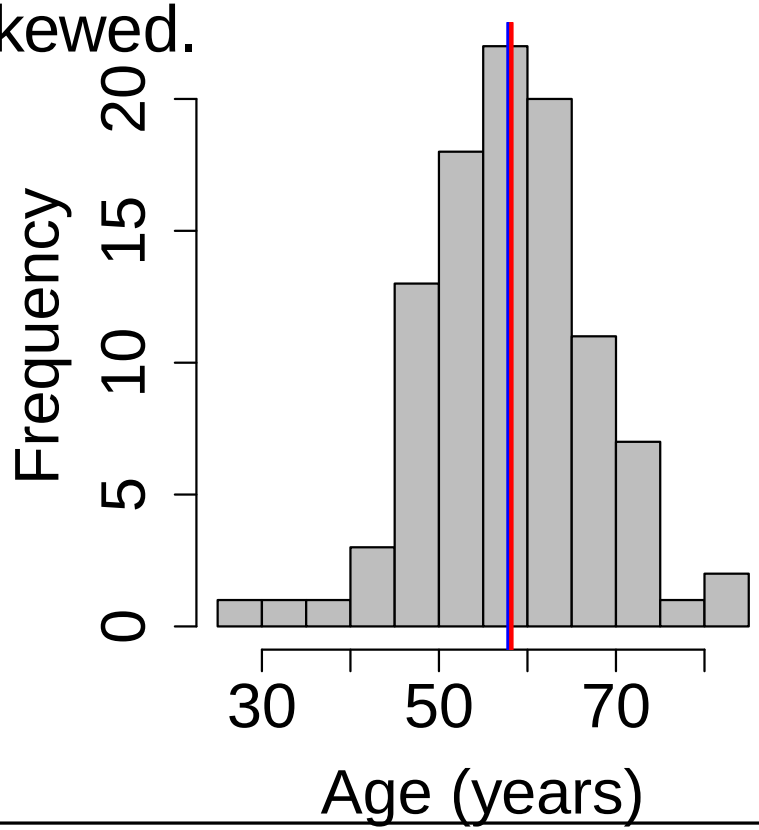


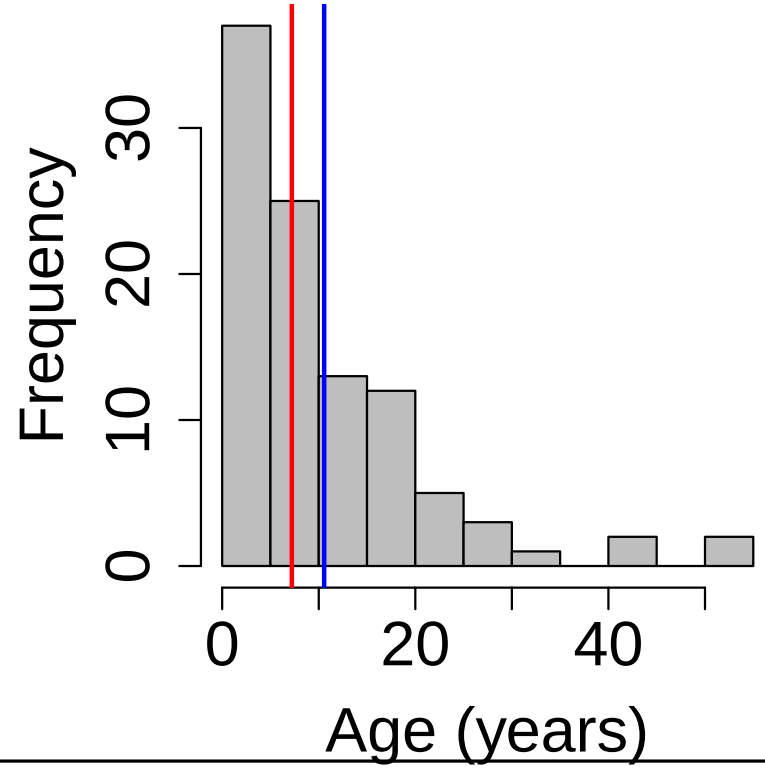
In the second example, the median better describes the distribution's center. The mean over-estimates the center.



Shape and central tendency

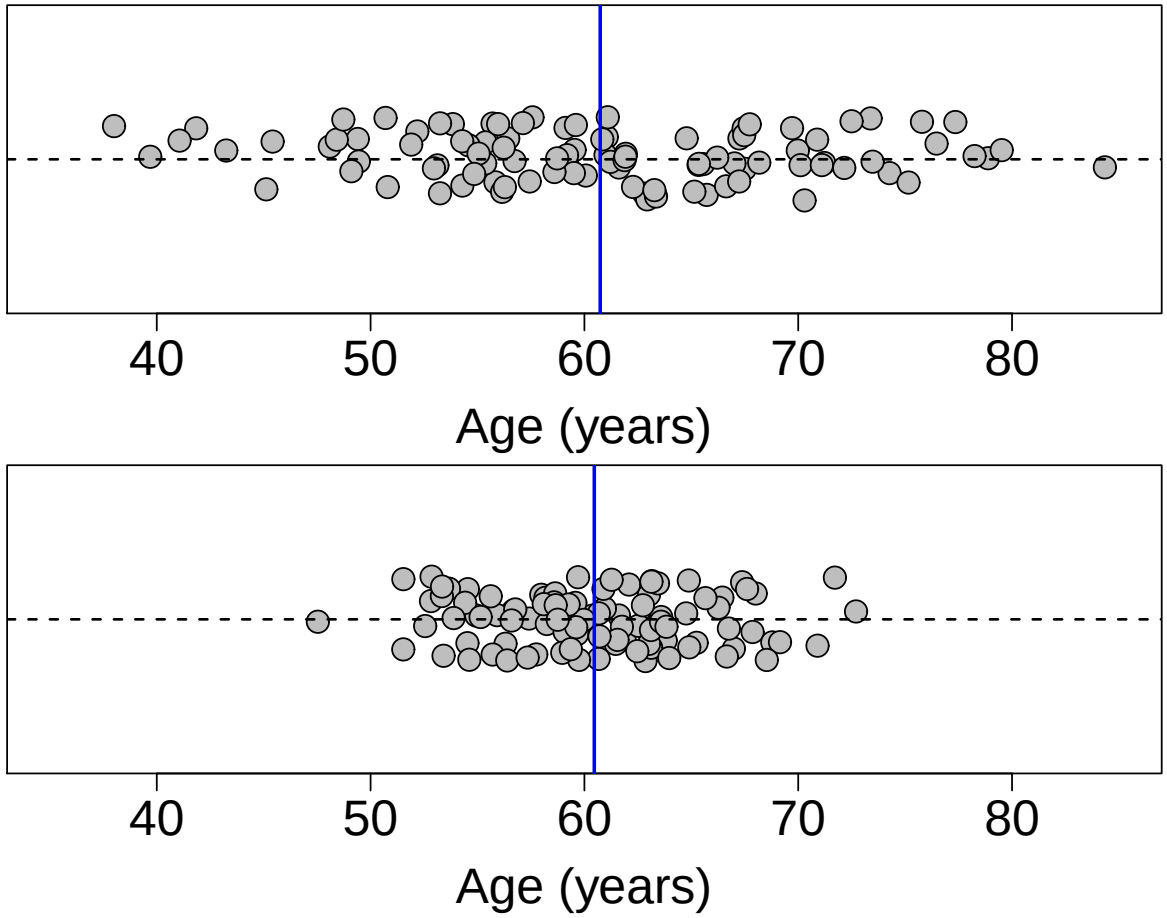
The difference between the distributions of age in the two figures can be seen better using histograms. Both the mean and median are adequate measures of central tendency for symmetric unimodal distributions. But only the median is an appropriate measure of central tendency if the data are skewed.





Describing dispersion

In the figures below, a sample of 100 ages are plotted for two different populations. Both distributions are centered at (approximately) the same value, but the ‘spread’ of the values differs.



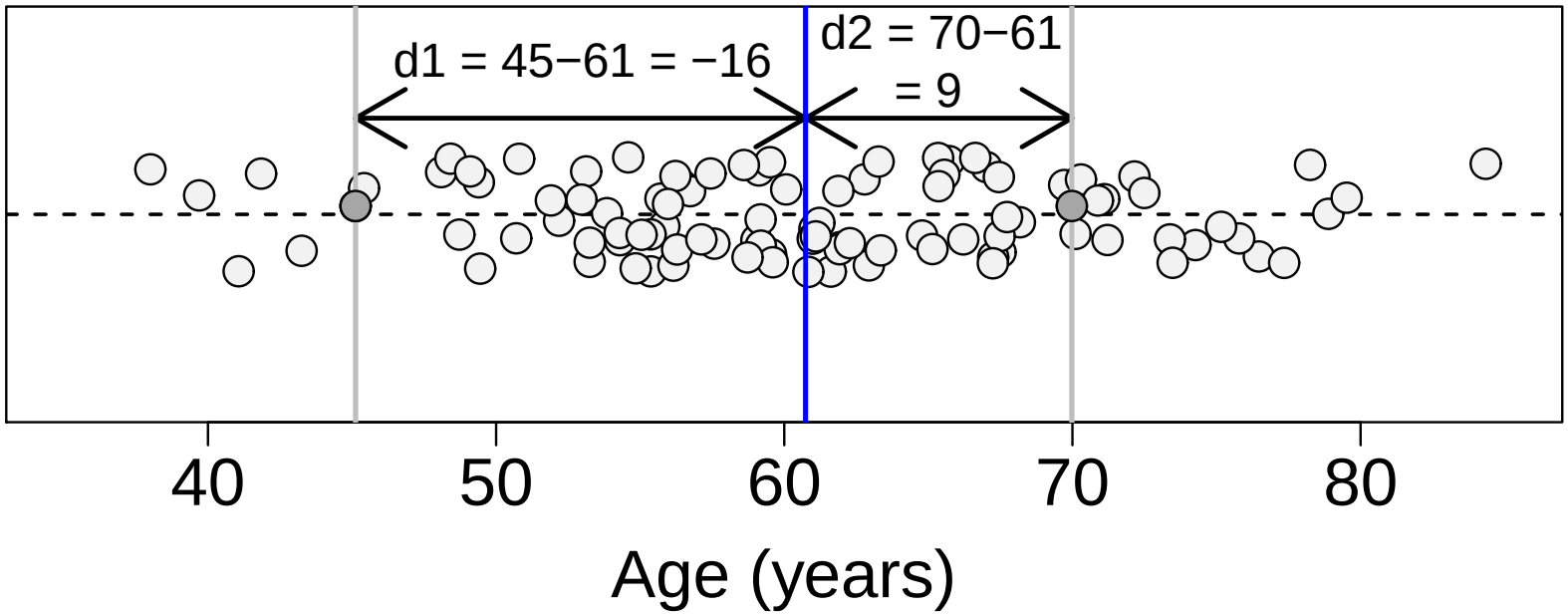
intentionally left blank

Standard deviation

Statistics that measure the spread of a distribution are called measures of *dispersion*. The most familiar measure of dispersion is called the *standard deviation* (abbreviated ‘SD’). Here is how we calculate SD.

1. Calculate the difference between each data point and the mean.
2. Sum the squared differences.
3. Standardize the sum by dividing by $n - 1$, where n is the number of data points. This quantity is called the *variance*.
4. The square root of the variance is the standard deviation.

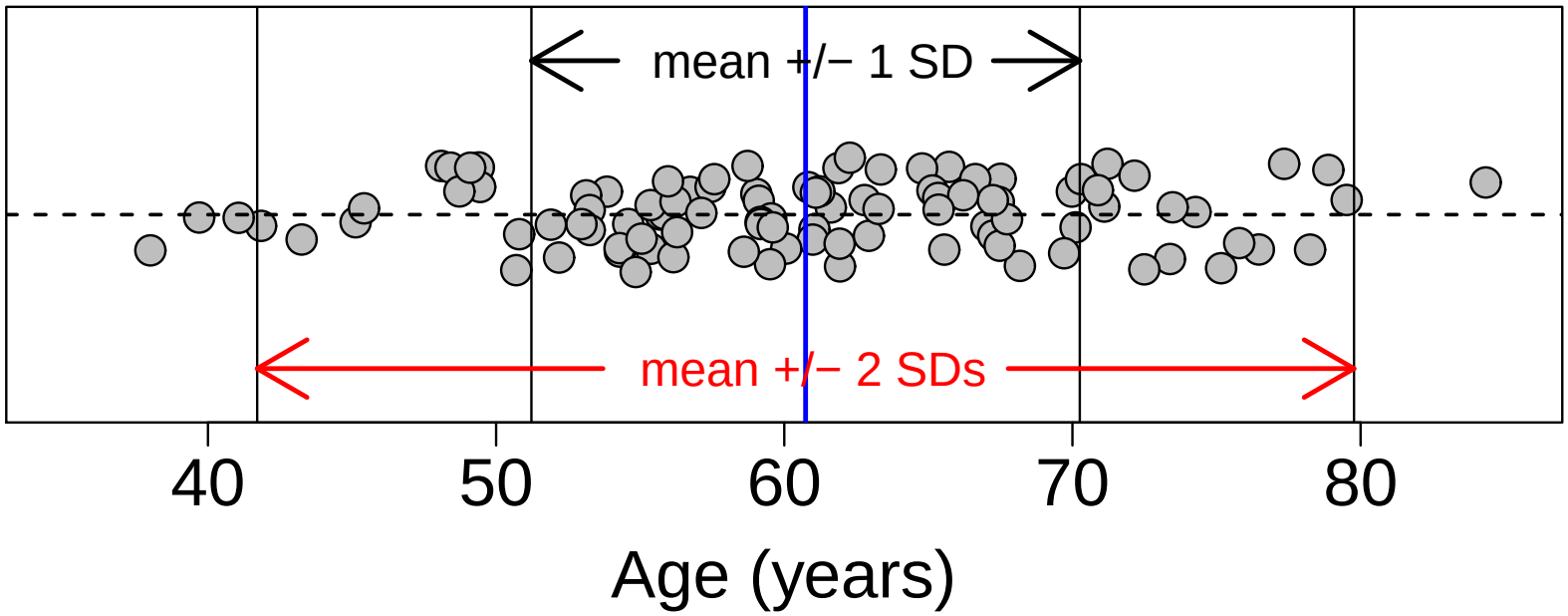
Standard deviation (cont.)



Variance = $\frac{d_1^2+d_2^2+...+d_{100}^2}{n-1}$ and SD = $\sqrt{\text{Variance}}$

For this example, SD = 9.5 years.

Standard deviation (cont.)



The vertical lines in this figure show the locations of the mean $\pm$ 1 SD and the mean $\pm$ 2 SDs. (The mean $\pm$ 3 SDs extends beyond the plotting region.) Of the 100 data points shown 68/100 = 68% fall within 1 SD of the mean, 96/100 = 96% fall within 2 SDs of the mean, and 100/100 = 100% of the data points are within 3 SDs of the mean.

SD and the normal distribution

This observation is not just a coincidence. If a variable is normally distributed:

- 68% of all values lie within 1 SD of the mean
- 95% of all values lie within 2 SDs of the mean
- $>$ 99% of all values lie within 3 SDs of the mean

The percents reported on slide 21 don't match these theoretical results exactly, because the figure on slide 21 is constructed from sample data.

Additional comments about SD

- The SD for a variable is often reported as “mean $\pm$ SD”
- Results regarding proportions of values within 1, 2 and 3 SDs of the mean are true only for normally distributed variables
- The SD is an appropriate measure of dispersion to report only for approximately normally distributed variables.
- What is wrong with this sentence?

  “Mean treatment duration was 6.8 ($\pm$ 5.1) months in the bevacizumab plus fluoropyrimidine-cisplatin group ...”
  (Ohtsu et al., JCO 2011, page 4)

Other measures of dispersion

If a variable is not approximately normally distributed, then a better measure of dispersion is

- Interquartile range (IQR) = 75th percentile - 25th percentile
- Range = max - min

Typically, what is reported in the literature is not the IQR or range, but the values used to construct that measure. For example, the range for age in the placebo arm of the trial is reported as 22 - 82 (Ohtsu et al., JCO 2011, page 3, Table 1).

Continuous variable's distribution - redux

For a unimodal continuous variable, the table below summarizes measures of central tendency and dispersion that are appropriate to report depending on the distribution's shape.

	Center		Dispersion		
Distribution type	Mean	Median	SD	IQR	Range
Normal	√	√	√	√	√
Non-normal		√		√	√

Time-to-event variables

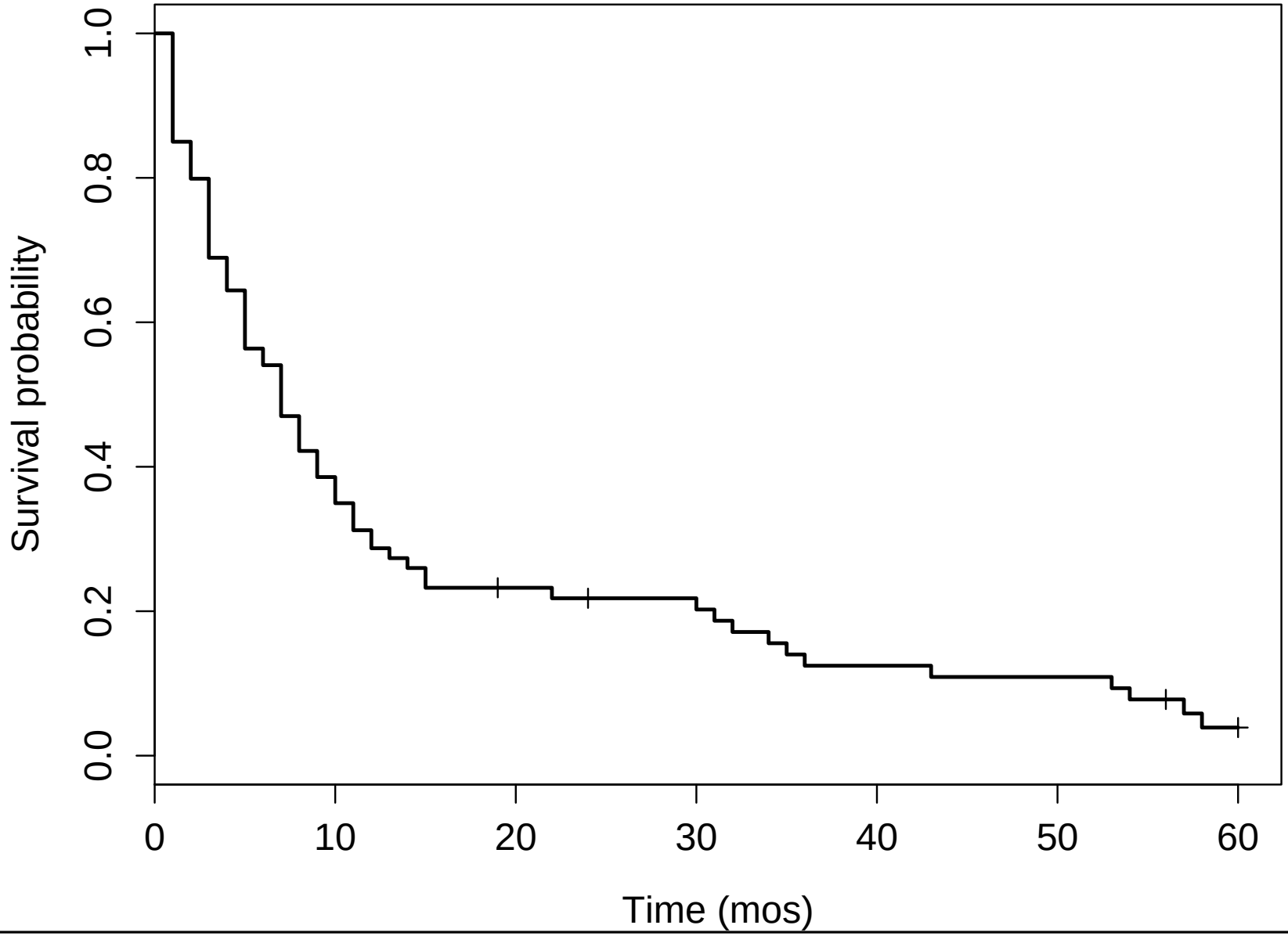
A *time-to-event* variable is a special type of continuous variable common in clinical research. Examples of time-to-event variables include:

- Time to death (measured from date of diagnosis or possibly date of randomization in a trial)
- Time to death typically called *overall survival*
- Time to disease progression (following baseline measure of disease burden)
- Time to disease progression typically called *progression free survival*
- Time to HIV seroconversion (following enrollment in an HIV vaccine trial)

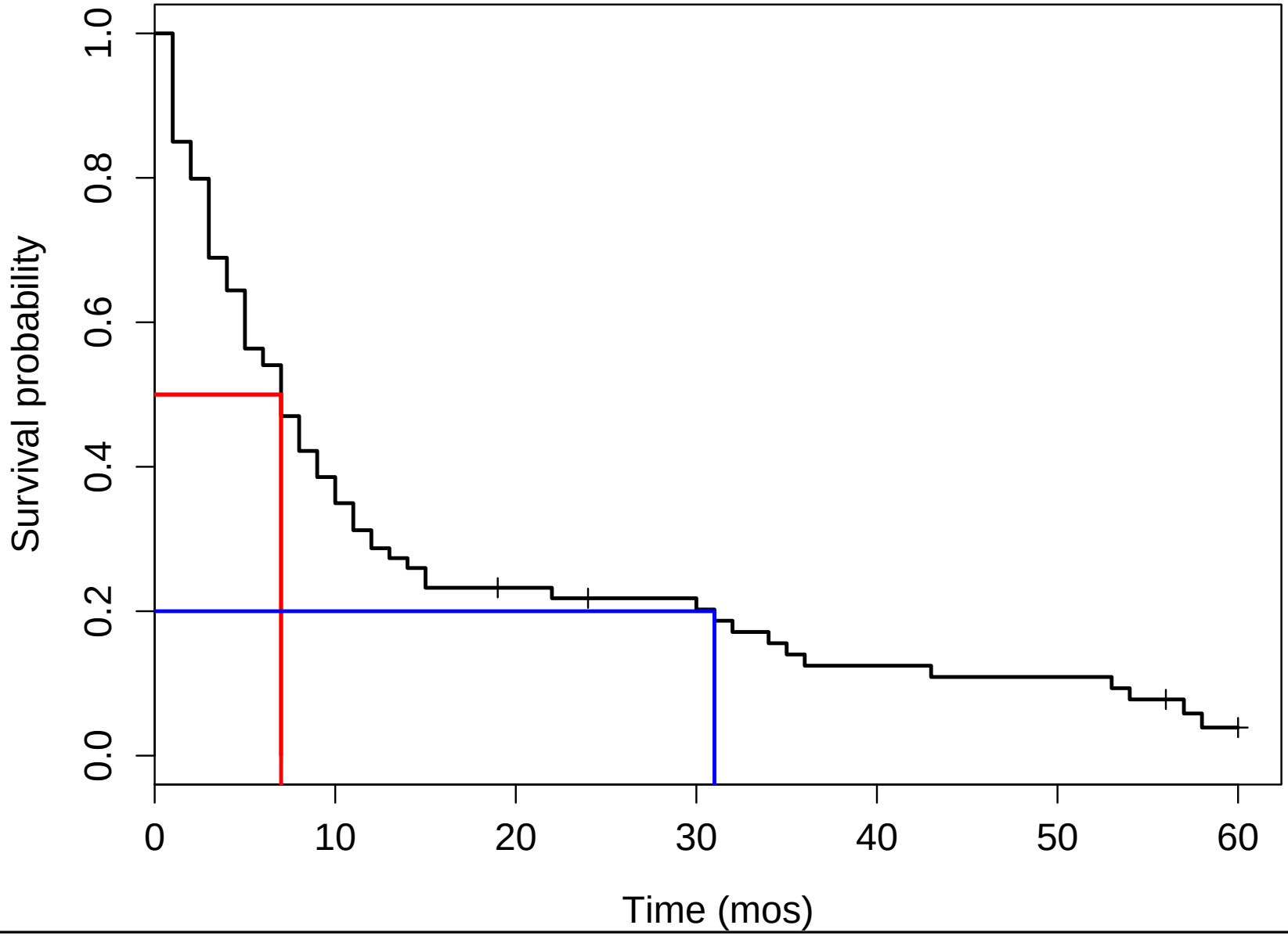
Properties of time-to-event variables

- Event times are subject to *censoring*
  - The event of interest is not observed for all subjects in the study
  - Subject has not yet had the event at the end of the study
  - Subject is lost to follow-up (drops out of the study)
  - Censoring means the event is only partially observed
- Event time distributions tend to be positively skewed
  - Some subjects have much longer event times than others
  - Distribution is non-normal

Kaplan-Meier survival curve



Kaplan-Meier survival curve



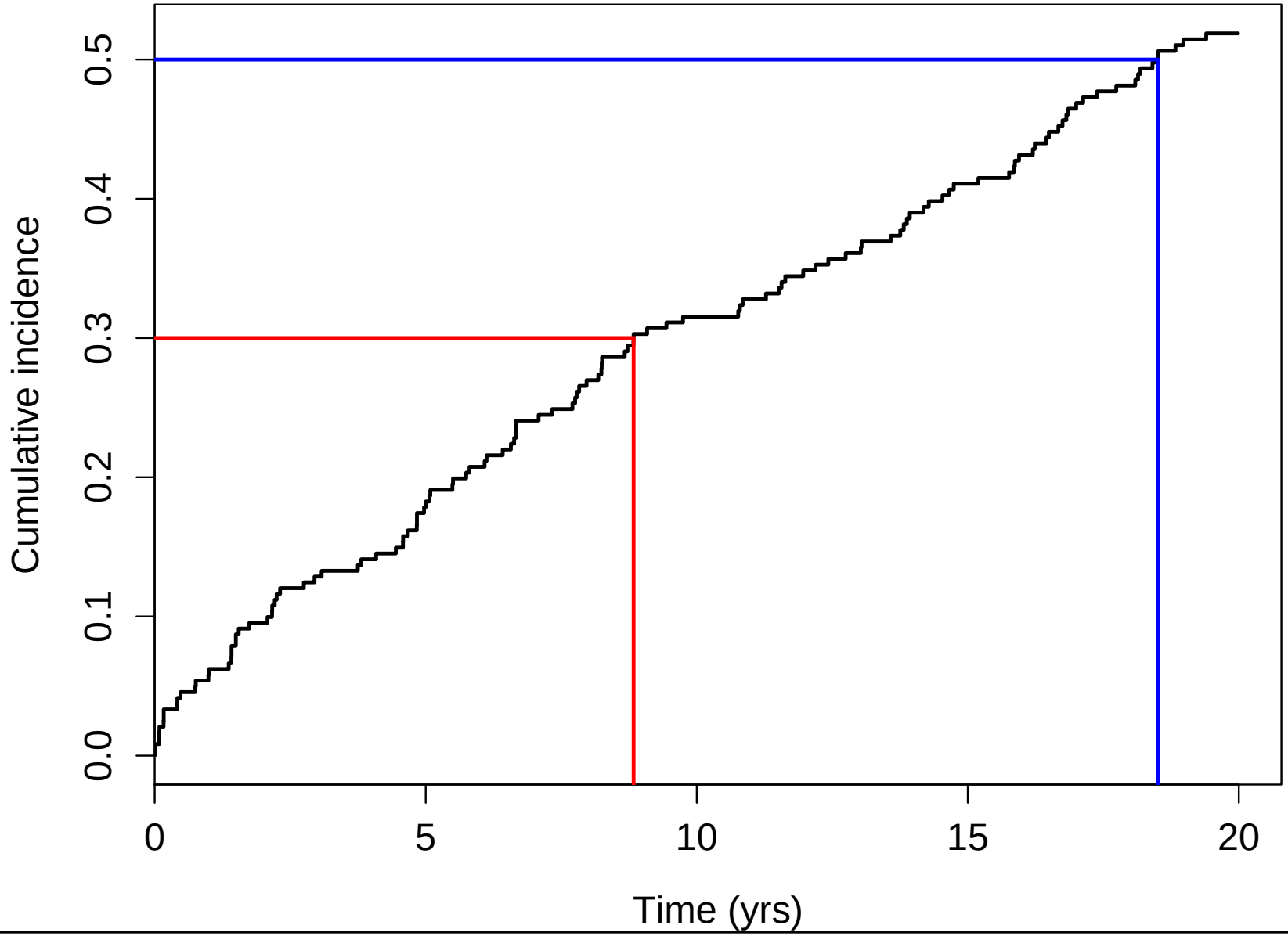
Kaplan-Meier survival curve

Assuming the ‘event’ is death on Slide 29, then

- 50% of patients survive to 7 months (median survival)
- 20% of patients survive to 31 months

In Ohtsu et al., JCO 2011, page 5, Figure 2A, the median overall survival (OS) for subjects in the placebo arm is 10.1 months while median OS in the intervention arm is 12.1 months. In Figure 2B, the median progression free survival (PFS) in the placebo arm is 5.3 months, while median PFS in the intervention arm is 6.7 months.

Kaplan-Meier cumulative incidence plot



Kaplan-Meier cumulative incidence plot

Assuming the ‘event’ is death on Slide 31, then

- 30% of patients died by 8.8 years
- 50% of patients died by 18.5 years

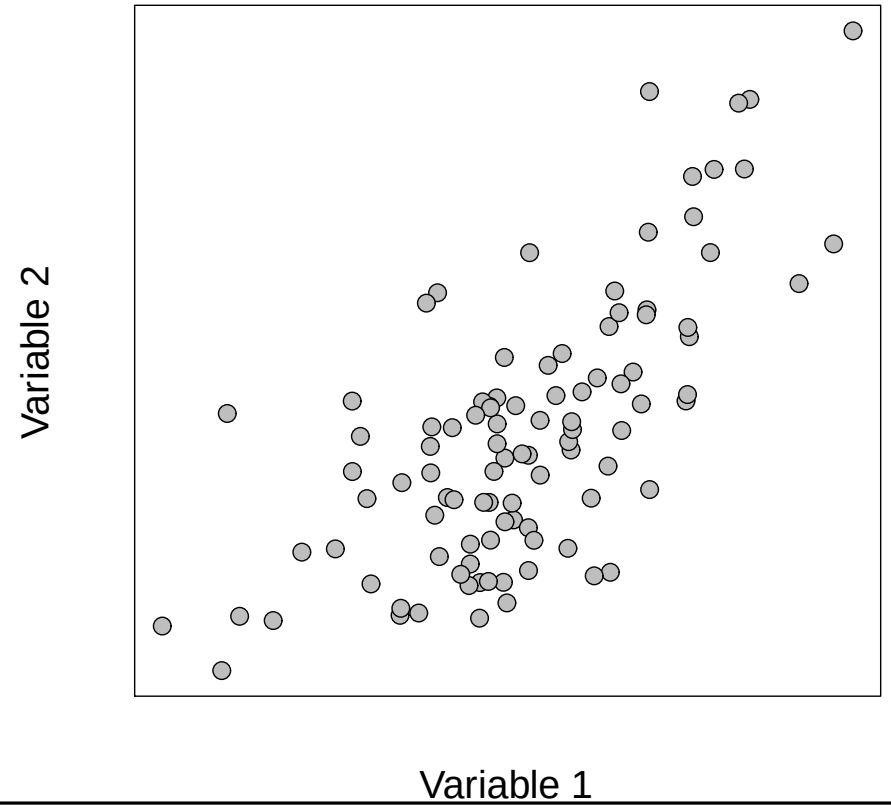
Describing a categorical variable's distribution

When describing a categorical variable's distribution, we generally report the frequency and percent of observations occurring in each level.

For example, in Table 1 of the Ohtsu article, the number of males in the bevacizumab arm is 257, which represents 66% of those randomized to that arm.

Joint distribution of two continuous variables

The *joint distribution* of two variables is a description of how they co-vary. For example, we might note that BMI and total cholesterol have the property that when one is large (or small) the other tends to be large (or small). A good way to visualize this co-variation is with a scatterplot.



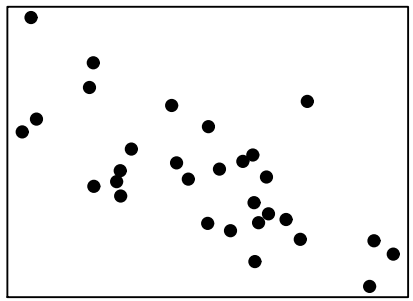
Correlation

Correlation measures the strength of the linear relationship between two continuous variables. We usually use the letter 'r' to represent sample correlation - that is, correlation calculated from data. We assume the sample correlation is a measure of some true underlying (but unknown) correlation represented by the Greek letter 'rho' (ρ). Here are some important properties of the sample correlation coefficient.

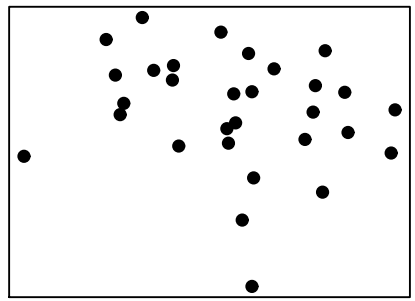
- r's range of values is -1 to 1.
- $r = 1 \Rightarrow$ observations lie on positively sloped line.
- $r = -1 \Rightarrow$ observations lie on negatively sloped line.
- r is a dimensionless measure (i.e. no units of measure).
- r measures the strength of the *linear* association.
- r tends to be close to zero if there is no linear association.

Picturing ρ and r

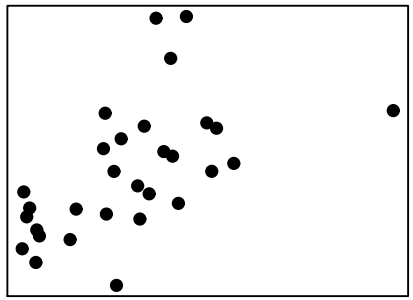
Each graph depicts a sample of 30 data points, (x, y), drawn from a population with the specified value of ρ. r is calculated based on the 30 data points.



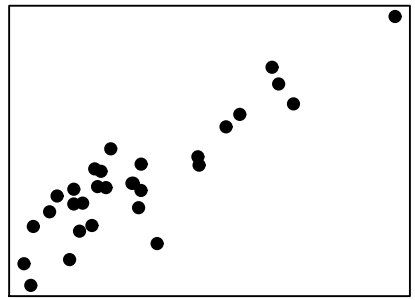
rho = −0.6 ; r = −0.691



rho = −0.05 ; r = −0.201



rho = 0.4 ; r = 0.556



rho = 0.9 ; r = 0.892