Biost 517 / Biost 514 Applied Biostatistics I / Biostatistics I

Timothy A. Thornton, Ph.D.
Associate Professor of Biostatistics
University of Washington

Lecture 2:

Types of Variables; Descriptive Statistics: Histograms, Central Tendencies and Percentiles

Definitions

- **Population**: the complete set of individuals, objects or scores of interest for a study.
  - Often too large to sample in its entirety. It may be real or hypothetical (e.g. the results from an experiment repeated infinitely many times)
- **Sample**: A subset of the population collected for a study
  - A sample may be classified as random (each member has equal chance of being selected from a population) or convenience (what's available). Random selection attempts to ensure that sample is representative of the population. More on this later.

Definitions

- **Individuals/Subjects:** are the people or objects that are described by a set of data for a study
- **Variable**: any characteristic of an individual or study subject that is measured or observed.
  - A variable can take different values for different individuals
- **Distribution of a variable**: arrangement of observed or theoretical values of a variable and how often the variable takes these values.

Types of Variables

- **Binary**, a.k.a. dichotomous, Bernoulli: dead/alive
- **Categorical or Qualitative** variable: takes values that are intrinsically nonnumerical.
  - Nominal or unordered categorical: occupation, race, gender
  - Ordered categorical: cancer stage I/II/III/IV
- **Quantitative** variable: takes values that are intrinsically numerical
  - **Discrete**: 0/1/2 copies of a genetic variant
  - **Continuous**: blood pressure,
    - Interval continuous: have units that are of equal magnitude as well as rank order on a scale **without** an absolute zero, e.g., temperature in Fahrenheit scale
    - Ratio continuous: have units that are of equal magnitude as well as rank order on a scale **with** an absolute zero, e.g., heart rate

Types of Variables

- **Censored variable:** time to relapse for cancer patients in remission; some loss to follow-up occurs, or the study ends before observing outcome
- Will focused on censored variables at later in the class

All variables can have missing values; note that all censored values (e.g. “in remission 5+ years”) are partially missing.

Binary Variables

- Only two possible values, which can be either
  - Labels, e.g., “Male” or “Female”
  - Coded as numbers, e.g., 1 or 2
- It is usually practically advantageous to represent as “indicator variables”
  - Possible values 0 or 1
  - 1 indicates the quality named by the variable
  - E.g., MALE is 1 for males, 0 for females
    - contrast: a “sex” variable that has values 1 and 2
  - E.g., MARRIED is 1 for married, 0 for everything else (single, divorced, widowed)

Categorical Variables

- Have a finite number of possible values denoting qualities ,e.g., levels or groups. Each should have a label;
  - Occupation is nominal: laborer, clerical, professional, retired
  - Marital status is nominal: single, cohabiting, married, divorced, separated, widowed
  - Stage of cancer is ordinal: I, II, III, or IV
- The only generally-sensible mathematical operations for nominal categorical variables are enumeration – i.e. counting – and checking equivalence.

Categorical Variables

- For ordered categorical variables mathematical operations can be done (e.g. mean cancer stage of patients) but...
  - Spacing between categories is not well-defined, so results of addition/subtraction not well-defined
  - Means (or averages) can be used to identify (but not quantify) differences between distributions of categorical variables
-

Quantitative Variables

- Values precisely quantify some scientific measure;
- Discrete means no measurements are possible between adjacent levels, e.g. counts of fatalities, number of copies of variant allele
- Continuous means any measurement is possible, e.g. weight, height, to arbitrary accuracy
-

Quantitative Variables

- Due to measurement accuracy, quantitative values may only be recorded to some precision
  - Weights to the nearest pound
  - Money is only defined to the nearest $0.01
- Precision level of measurements is highly unlikely to make a difference in how data is summarized or analyzed – almost always treat it as 'truly' continuous
  - e.g., Age often given in years but we usually treat as continuous

Interval vs Ratio Measurements

- Two types of quantitative variables: Ratio and Interval Variables
- Ratio variables: variables that are measured in units that are of equal magnitude as well as rank order on a scale *with* an absolute zero
  - Both differences and ratios of interest
  - Age, height, weight have absolute zeroes
- Interval variables: variable that are measured in units that are of equal magnitude as well as rank order on a scale *without* an absolute zero
    - Temperature has different zeroes in Fahrenheit and Celsius
    - Only differences make clear sense
- Generally, differences make sense for all quantitative variables
- Ratios make sense if measurements are made relative to an absolute zero

Notation for Variables

- If measurements made on n subjects – n is the sample size
- Random variables are denoted by capital letters (e.g. X, Y)
- Subscripts on variables denote the variable measurements made on different subjects;
  - X_1 is the measurement of X on subject 1
  - X_5 is the measurement of X on subject 5
  - X_i is the measurement of X on subject i
- Vectors of variables $X_1, X_2, X_3, \ldots, X_n$ *are denoted by* $\boldsymbol{X}$,

Notation for Variables

- The ordering of the data (i.e. who is labeled subject 1, 2, etc) generally doesn't matter – but keeping each subject's outcomes and measurements for other variables together is essential!
- When we do need to order a variable, use subscript parentheses, i.e. from smallest to largest, the values of *Y* on *n* subjects are: $Y_{[1]} \leq Y_{[2]} \leq Y_{[3]} \leq \ldots \leq Y_{[n-1]} \leq Y_{[n]}$.

Frequency Distribution of a Variable

- A Frequency Distribution or **Histogram** for a continuous quantitative variable presents the counts of observations grouped within pre-specified classes or groups
- A Relative Frequency Distribution presents the corresponding proportions of observations within variable classes
- A Barchart presents the frequencies for a categorical variable

Example: Creatine kinase (CK) activity

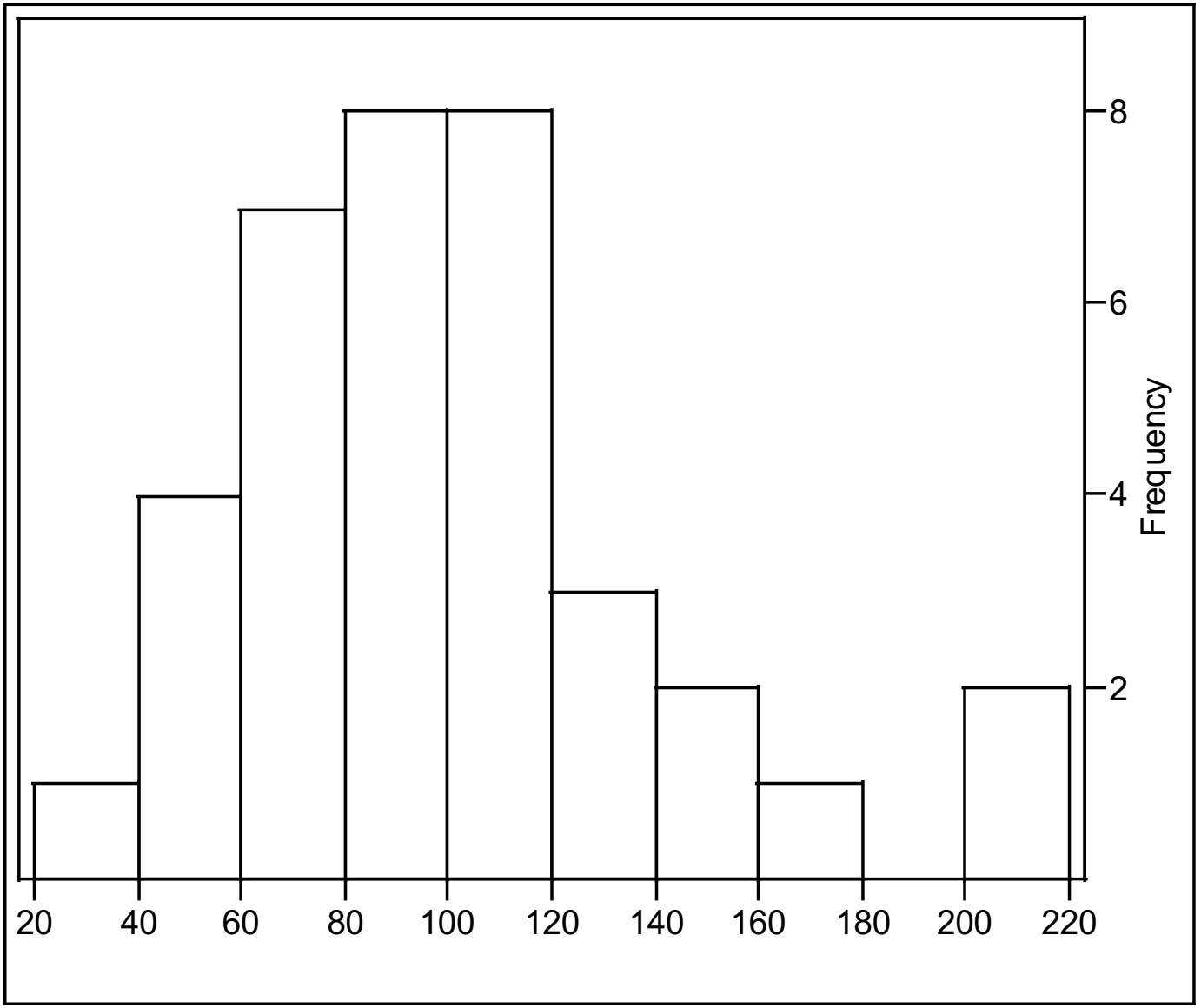
- **Creatine kinase** (**CK**) activity is greatest in striated muscle, heart tissue, and brain. The determination of **CK** activity is a proven tool in the investigation of skeletal muscle disease (muscular dystrophy) and is also useful in the diagnosis of myocardial infarction (MI) and cerebrovascular accidents.
- Data from 36 males with serum CK (in U/L) measurements

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

Relative Frequency Table

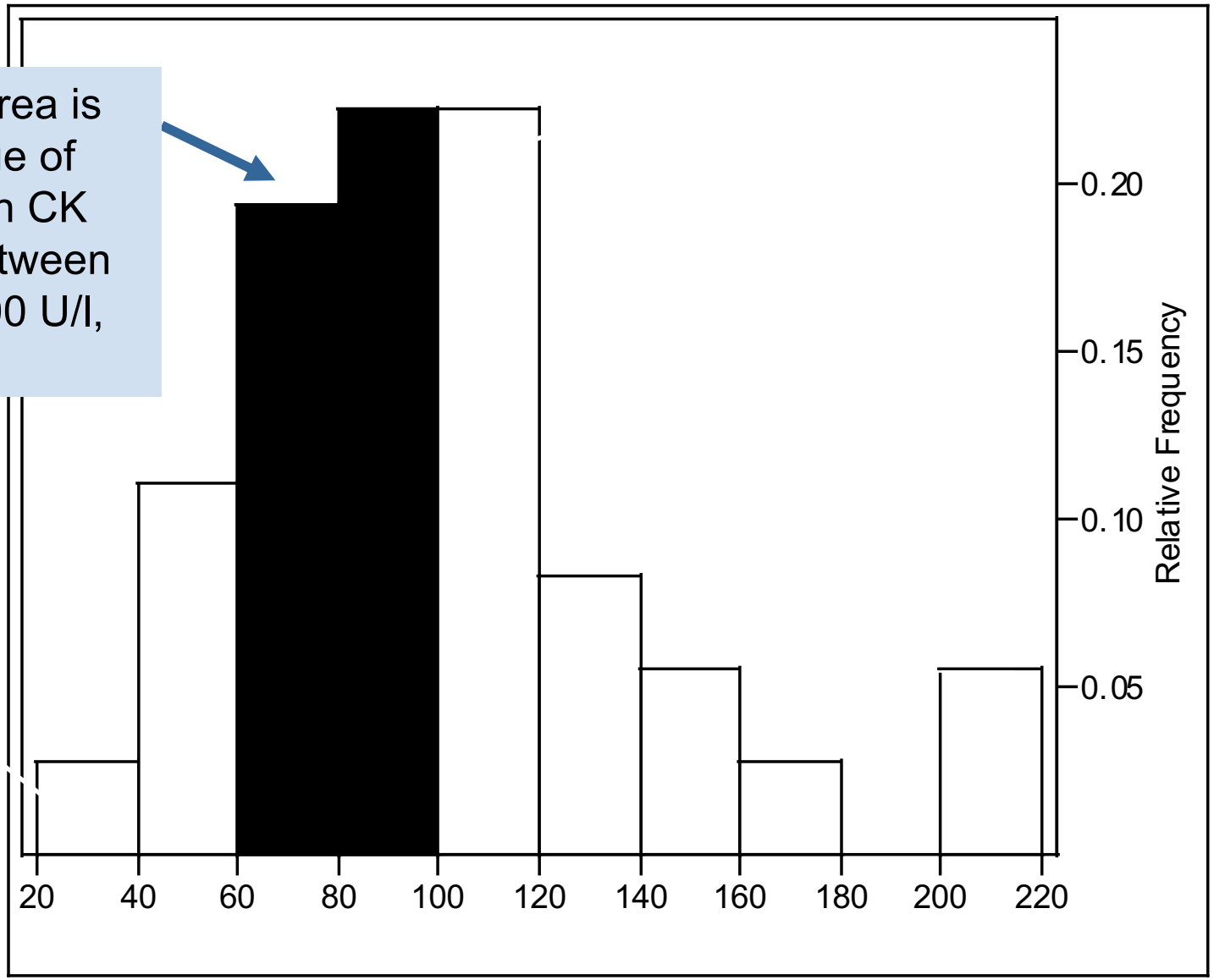
Serum CK (U/l)	Frequency	Relative Frequency	Cumulative Rel. Frequency
20-39	1	0.028	0.028
40-59	4	0.111	0.139
60-79	7	0.194	0.333
80-99	8	0.222	0.555
100-119	8	0.222	0.777
120-139	3	0.083	0.860
140-159	2	0.056	0.916
160-179	1	0.028	0.944
180-199	0	0.000	0.944
200-219	2	0.056	1.000
Total	**36**	**1.000**	

Frequency Distribution: CK



Relative Frequency Distribution

Shaded area is percentage of males with CK values between 60 and 100 U/l, i.e. 42%.

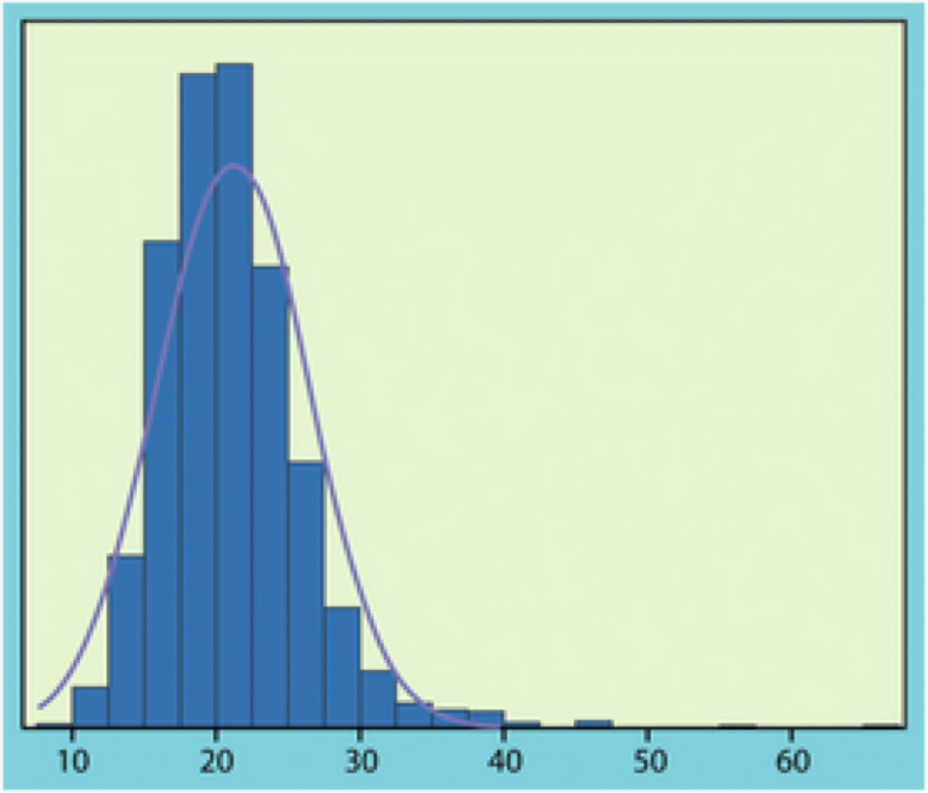


Interpreting Histograms

- Histograms provide a graphical display of the distribution: can discern the overall pattern and any significant deviations from that pattern.
- **Shape**: Is the distribution (approximately) symmetric or skewed?
- **Center**: Where is the “middle’” of the distribution?
- **Spread**: Where do most of the values fall? What are the smallest and largest values?
- **Outliers**: Are there any observations that lie outside the overall pattern? They could be unusual observations, or they could be mistakes. Investigate them!

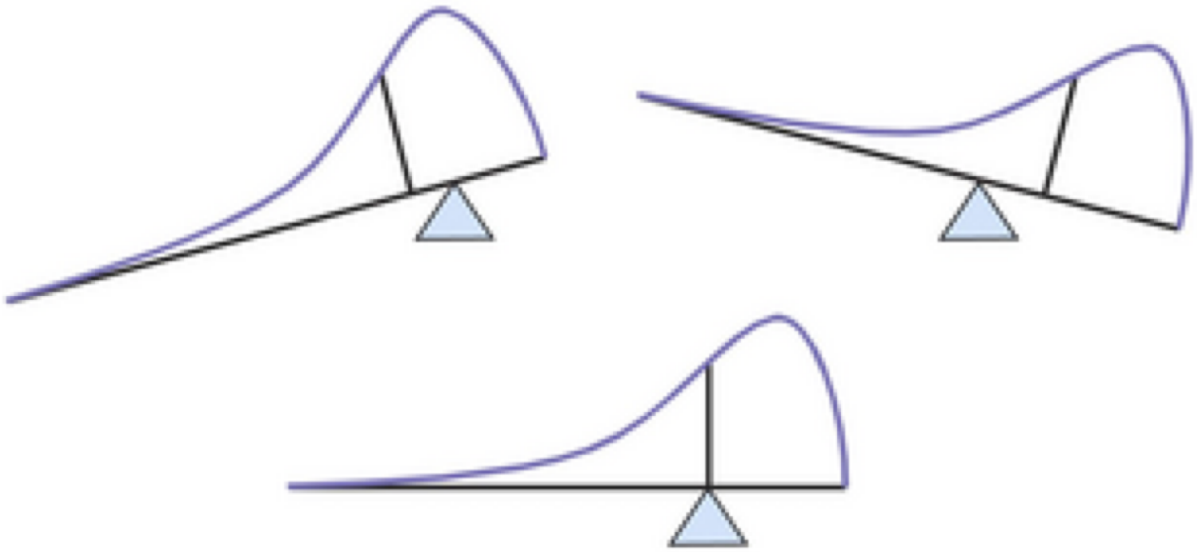
Density Curves

- If the distribution pattern is sufficiently regular, we can approximate it with a smooth curve.
- Area under the curve in a range of values indicates the proportion of values in that range.
- Come in a variety of shapes, but the “normal” family of familiar bell-shaped densities is commonly used. (More on this in later lectures).
- Density is only an approximation, but it simplifies analysis and is generally accurate enough for practical use.



Descriptive Statistics: Center of a Distribution

- Two commonly used descriptive statistics for the “center” of a distribution are: Mean and Median
- Median: The equal-areas point with 50% of the “mass” (or observations) on either side.
- Mean: The balancing point of the curve, if it were a solid mass, where larger values (in magnitude) have more weight.



Descriptive Statistics: Sample Mean

- Let $x_1, x_2, x_3, \ldots, x_n$ be the realised values of a variable ***X***, from a sample of size **n.** The **sample mean (or arithmetic mean)** is defined as the average of the numbers:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + x_3 \cdots + x_{n-1} + x_n}{n}$$

Descriptive Statistics: Sample Median

- If the sample data are arranged in increasing order, the median is
  - Ordered observation $(n+1)/2$ if n is an odd number
  - the value that is halfway between ordered observation $n/2$ and $n/2+1$, if n is an even number

Mean of Creatine Kinase (CK) activity

- In the CK measurement study, calculate the sample mean and sample median
- Sample mean for CK data:

$$\bar{x} = \frac{121 + 82 + 100 \cdots + 95 + 42}{36} = \frac{3538}{36} = 98.277$$

- So the sample mean is 98.28 U/L

Median of Creatine Kinase (CK) activity

- To obtain the sample median, we must first sort the data from smallest to largest.
- Since the number of samples is even, 36, the median is the midpoint (or average) between the 18th (36/2) and 19th (36/2+1) largest value in the sorted dataset

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

- So the median is **94.5 U/L**, which is the midpoint between 94 and 95,

Descriptive Statistics: Mode

- The **mode** is the most commonly occurring value.
- There can be multiple modes in a dataset
- Identify the mode (or modes) in the CK dataset:

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

Mean versus Median

- Large or small sample values can heavily influence the mean. Histogram will often be left or right-skewed.
- The median is not influenced by extreme sample values and is a better measure of centrality if the distribution is skewed.
- For a unimodal distribution (a distribution with only one mode), if mean=median=mode then the data are said to be symmetrical (e.g., data follow a **normal distribution**).
  - The exception is a bi-modal symmetric distribution, where the mean and median are the same but there are two modes.

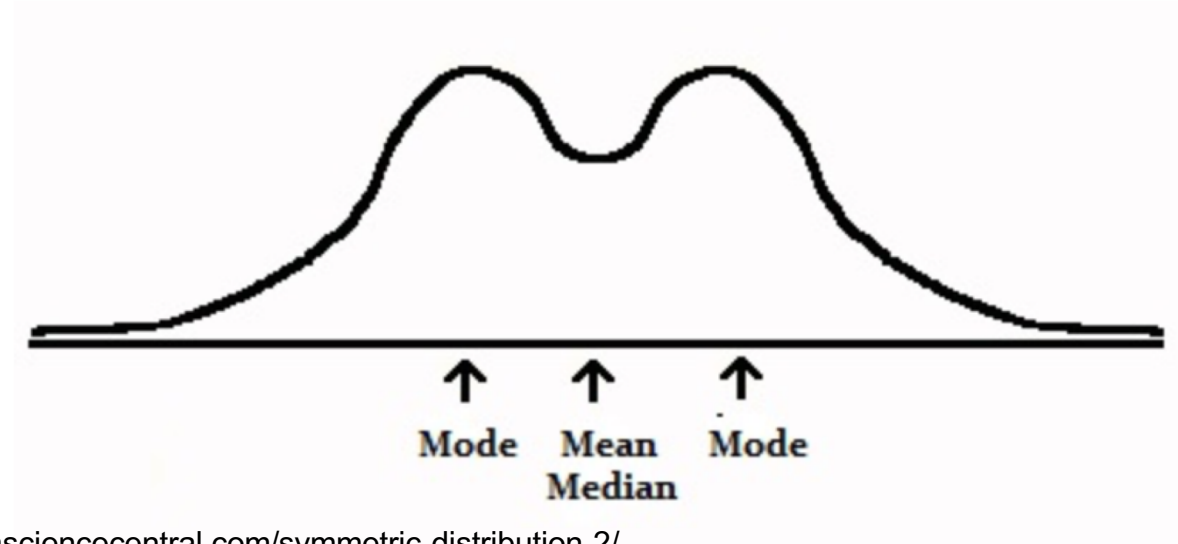


Figure available at:
https://www.statisticshowto.datasciencecentral.com/symmetric-distribution-2/

Mean versus Median

- In the CK measurement study
  - sample mean = 98.28.
  - median = 94.5

- The mean is larger than median indicating that histogram is right skewed. Mean is influenced by two large data values 201 and 203.

Descriptive Statistics: Percentiles

- A percentile has an intuitively simple meaning—for example, the 25th percentile is that value of a variable such that 25% of the observations are less than that value and 75% of the observations are greater.
- The *P*th *percentile* of a sample of n observations is that value of the variable with rank (*P* /100)(1 + n). If this rank is not an integer, it is often rounded to the nearest half rank.

Splitting data at evenly-spaced quantiles

- Any distribution of values along a continuum can be divided into evenly spaced quantiles
- Tertiles: Split at 33%, 66%
- Quartiles: Split at 25%, 50%, 75%
- Quintiles: Split at 20%, 40%, 60%, 80%:

Descriptive Statistics: Quartiles

- Distribution of values along a continuum divided into quarters are called *Quartiles*.
- The median is the **second quartile (Q2).** It is often called the 50th percentile, i.e., 50% of the observations are less than the median.
- The 25th percentile is the **first quartile (Q1).** This is the value such that 25% of the observations are at or below this values. To obtain Q1, subset the sorted data and only consider values that are less than or equal to the overall median. Q1 can be calculated to be the median of this subset of values between the smallest value and the overall median.
- Similarly, the 75th percentile is the **third quartile (Q3).** This is the value such that 75% of the observations are at or below this values. To obtain Q3, subset the sorted data and only consider values that are greater than or equal to the overall median. Q3 can be calculated to be the median of this subset of values between the overall median and the largest value.

Quartiles of creatine kinase (CK) activity

- To obtain the 25th percentile, we must find the median of the sorted values that are less than or equal to the median
- Since the number of samples in this subset is even, 18, then median is the midpoint (or average) between the 9^{th} (18/2) and 10^{th} (18/2+1) largest value in the sorted subset

25 42 48 57 58 60 62 64 67 68

70 78 82 83 84 92 93 94

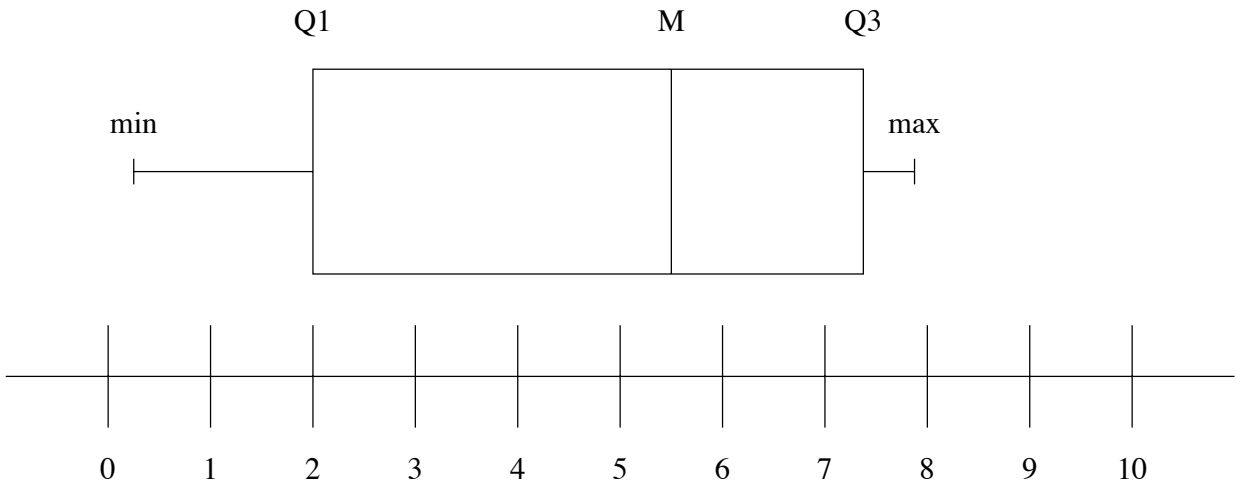
- So the 25^{th} percentile is **67.5 U/L**, which is the midpoint between 67 and 68
- Can similarly show that the 75^{th} percentile is CK is **118.5 U/L**

The Five-Number Summary and Boxplots

- A useful numerical summary of a distribution is given by

Min Q1 Median Q3 Max

- A **boxplot** is a graph of the five-number summary.
  - A central box spans the quartiles;
  - A line in the box marks the median;
  - Lines extends from the box out to the smallest and largest observations if there are no outliers; more on this later.
  - Useful for side-by-side comparison of several distributions.



Outliers and 1.5 x IQR Rule

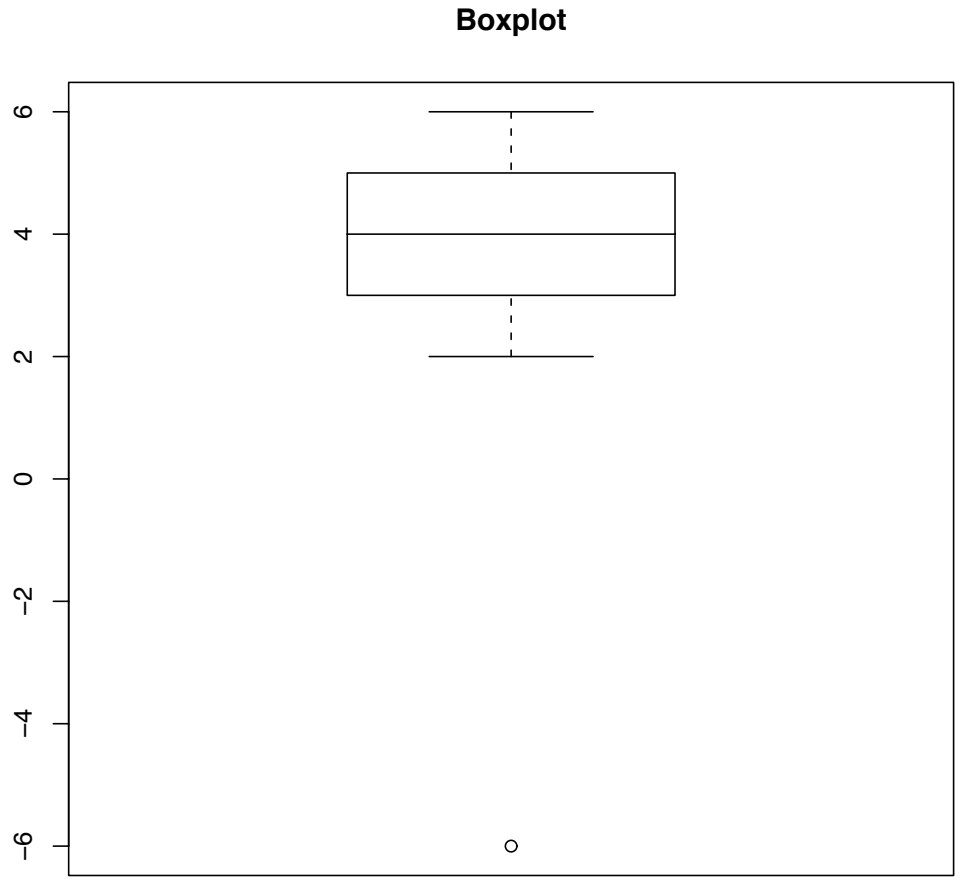
- Outliers are observations that lie outside of the regular pattern of the data.
- Outliers can be unusual observations, or they could be mistakes. Outliers should be investigated.
- The previously discussed **interquartile range (IQR)** is the third quartile minus the first quartile: Q3 - Q1
- A common rule for identifying outliers is the **1.5 x IQR rule**, where an observation is a suspected outlier if it lies more than 1.5 x IQR below Q1 or above Q3
- For most boxplots, the two "whiskers" will extend out to
  - the minimum value in the data or 1.5 x IQR below Q1, whichever is larger
  - the maximum value in the data or 1.5 x IQR above Q3, whichever is smaller

Boxplots with Outliers: Example

- Dataset (in increasing order): -6 2 3 4 4 4 5 5 6
- Five Number Summary:

Min = -6, Q1 = 3, Median = 4, Q3 = 5, Max = 6

- **So, IQR= (Q3-Q1)=2;**



Geometric Mean

- For a sample of size n, the geometric mean for a variable *X* is:

$$GeoMean(X) = \left(\prod_{i=1}^{n} X_i \right)^{1/n} = e^{\sum_{i=1}^{n} \log(X_i)/n}$$

$$\log(GeoMean(X)) = \sum_{i=1}^{n} \log(X_i) / n$$

- The geometric mean is equivalent to taking the mean of log-transformed data, and then exponentiating

- Note: all logarithms here are 'natural' – the default in scientific work. Base 10 logarithms (if needed) will be denoted $\log_{10}$.

Geometric Mean

- Geometric mean gives the central tendency or typical value of a set of numbers by using the product of their values (as opposed to the arithmetic mean which uses their sum).
- It is also often used for variables with values that are meant to be multiplied together or are exponential in nature, such as data on the growth of the human population or interest rates of a financial investment.
- Can be used for any positive quantitative variable;

Geometric Mean of creatine kinase (CK) activity

- The (arithmetic) mean of the log of the CK data is 4.503
- So the geometric mean is $e^{4.503} = 90.29$.