Continuous data – numerical and graphical summaries

Low Birth Weight Data

- Information on 100 low birth weight infants born in two teaching hospitals in Boston, Massachusetts

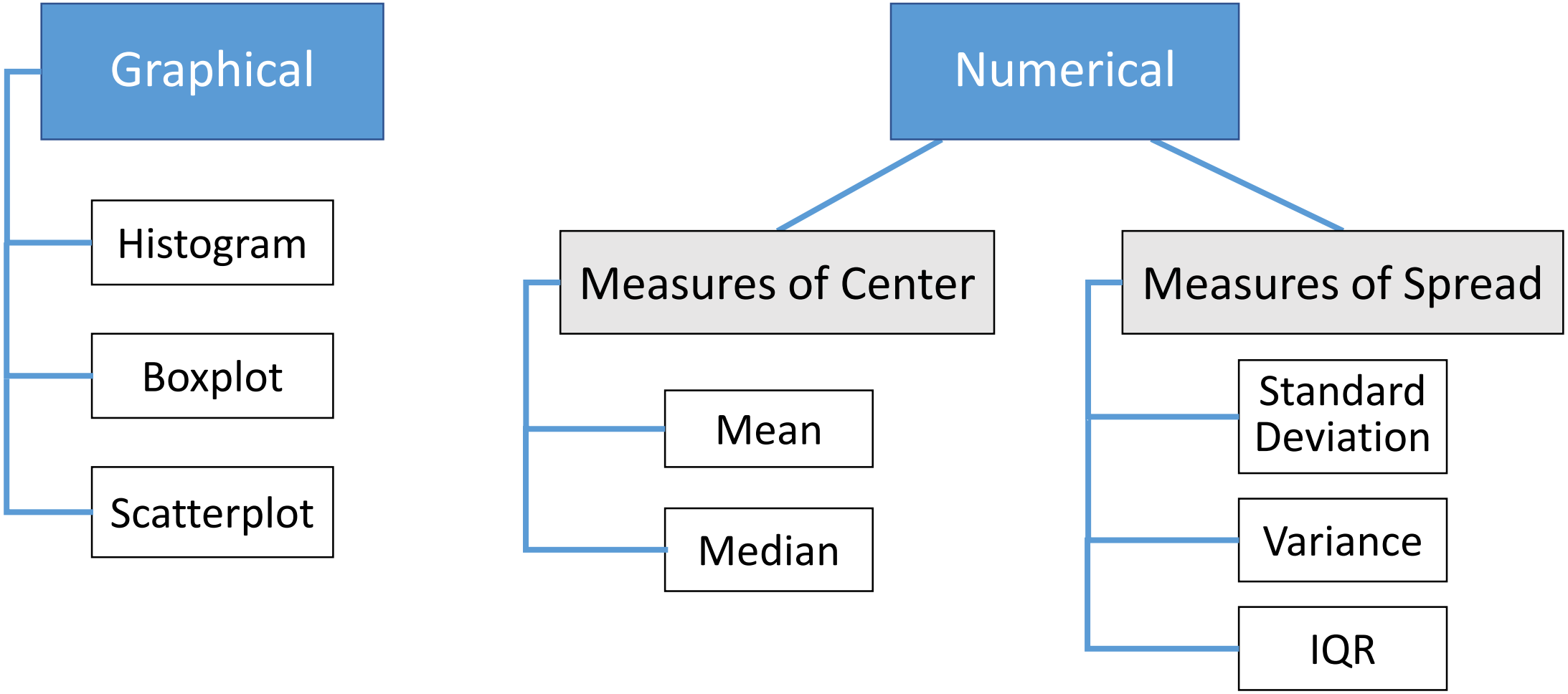
Variable	Description
sex	Sex of the baby (Male, Female)
gestage	Gestational age at time of birth (weeks)
length	Length of the baby (cm)
birthwt	Birth weight of the baby (g)
headcirc	Baby's head circumference (cm)
apgar	Apgar score (integers, min=0, max=10). This is a scoring system used for assessing the clinical status of a newborn. 7 or higher is generally considered normal, 4-6 is low, and 3 or below is critically low.

Find the dataset (lowbwt.xlsx) and the full data dictionary (lowbwt Data Dictionary.pdf) in the Data Module on the Canvas site

Fundamental Rule of Data Analysis

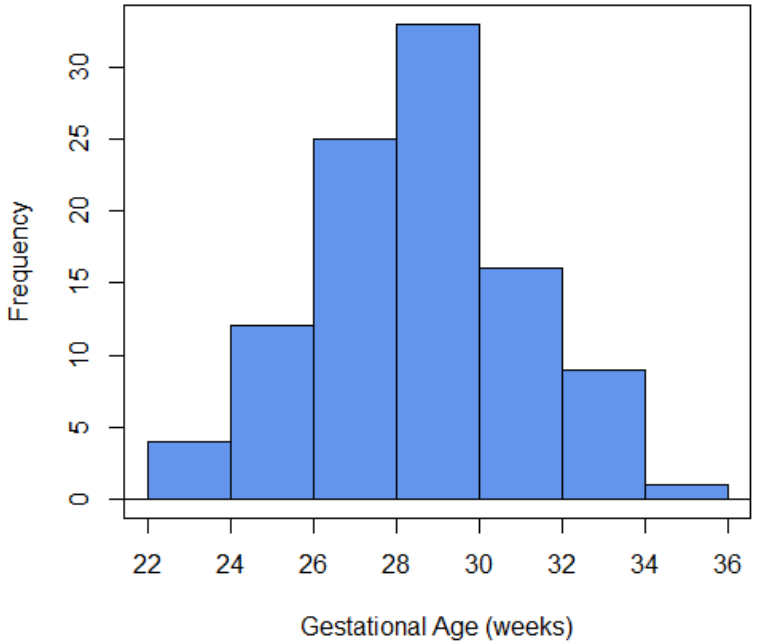
Different types of data require different statistical analyses.

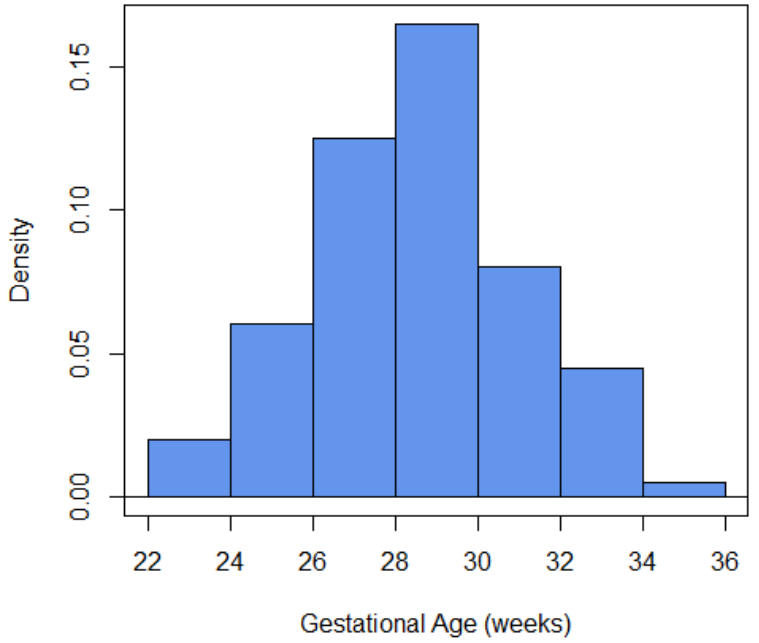
Summaries of Continuous Data



Histogram

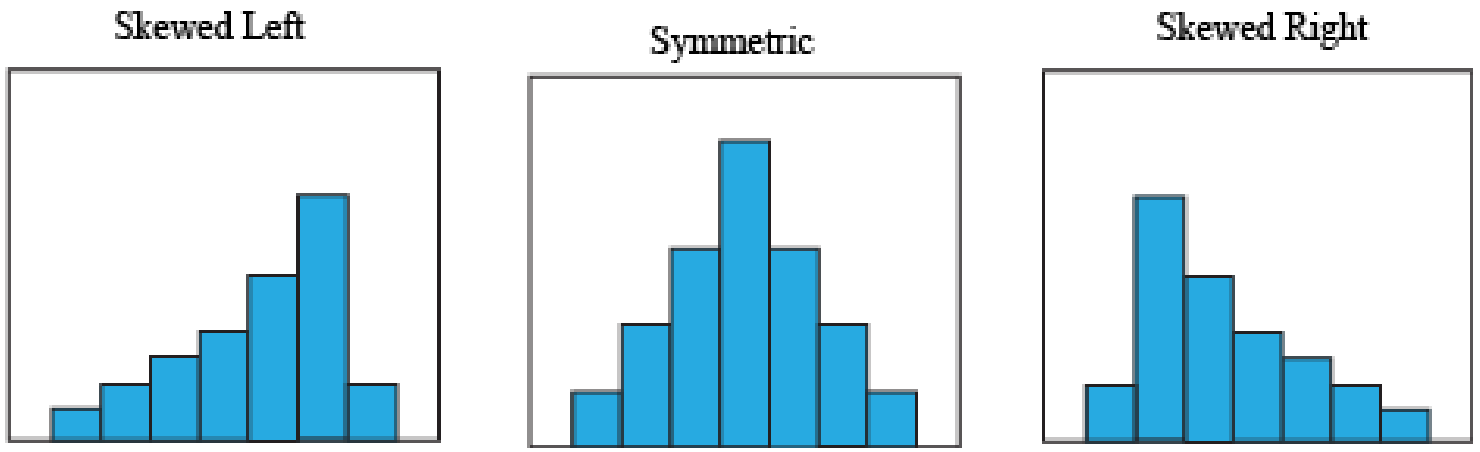
- Intervals displayed on horizontal axis
- Height of bar represents either the number of observations (frequency) or proportion of observations (density) in that interval





Histogram

- “Approximately symmetric” or “approximately normally distributed”
  - Don’t be too critical
- “Skewed left” = left tail is longer
- “Skewed right” = right tail is longer

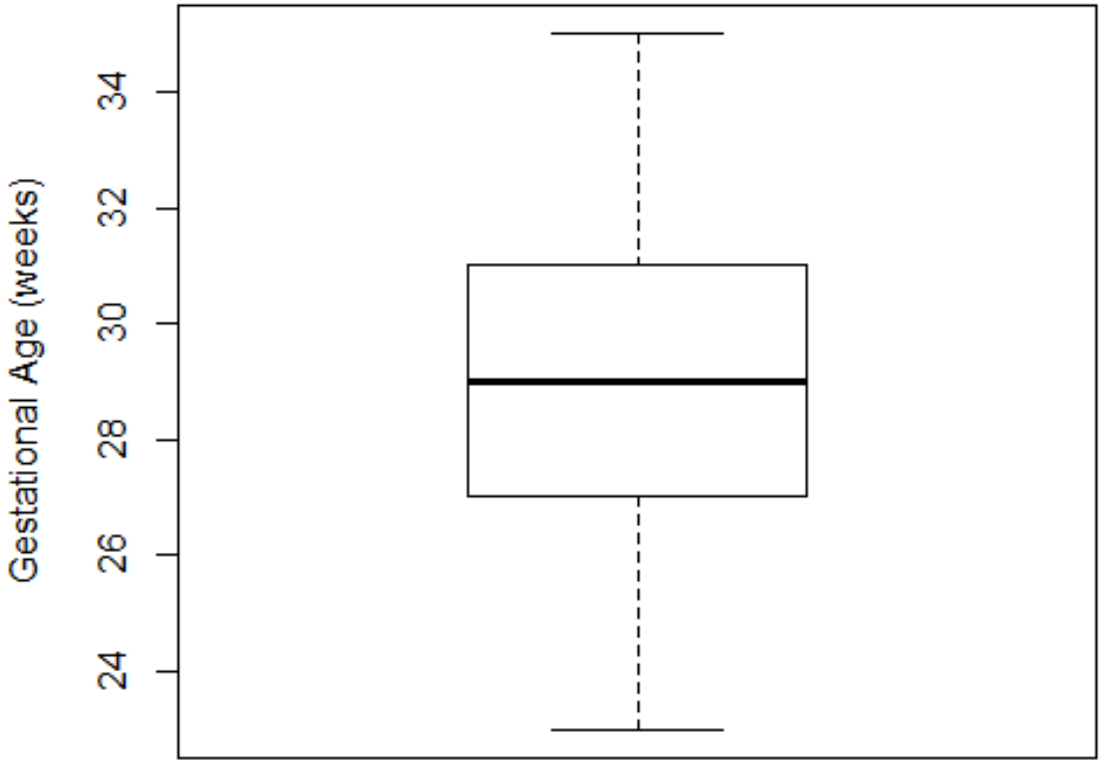


Boxplot

- Splits data into **percentiles**
  - Specifically, the 0^{th}, 25^{th}, 50^{th}, 75^{th}, and 100^{th} percentiles
- p^{th} percentile = the value that is greater than or equal to p% of the observations

Percentiles for gestational age:

0%	25%	50%	75%	100%
23	27	29	31	35



Boxplot

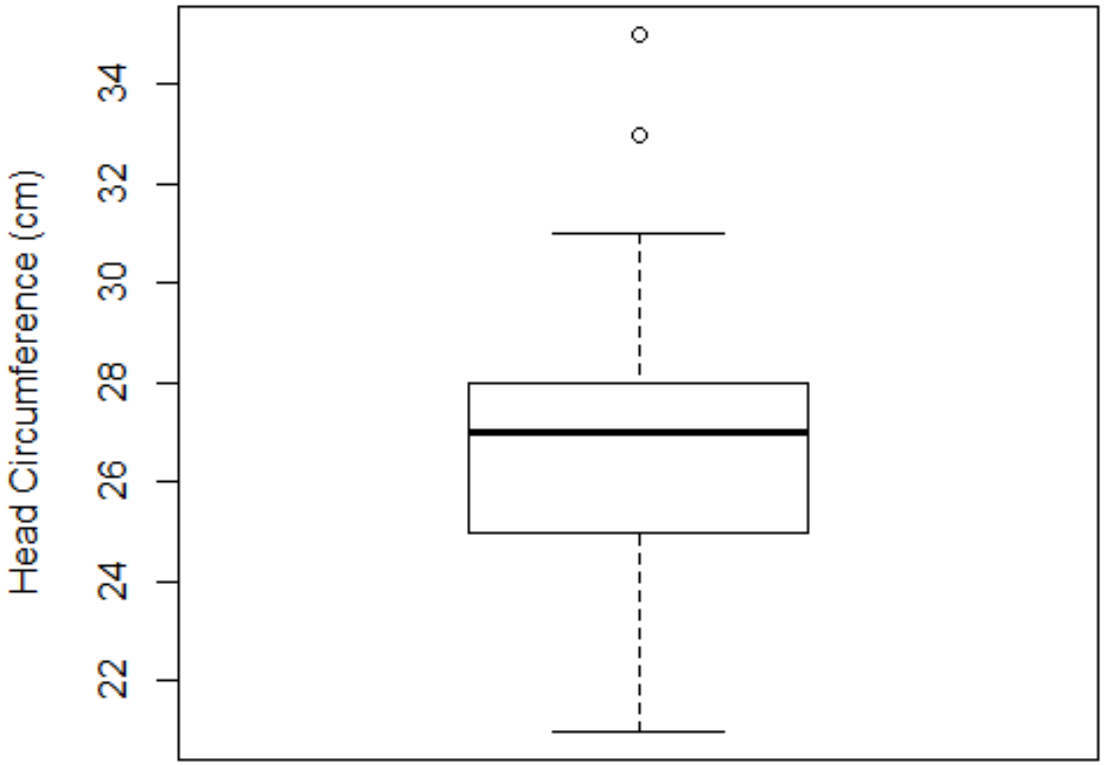
- Also indicates when there are **outliers**. An outlier is:

  a value greater than

  75^{th} percentile + 1.5 x (75^{th} – 25^{th})

  or a value less than

  25^{th} percentile – 1.5 x (75^{th} – 25^{th})

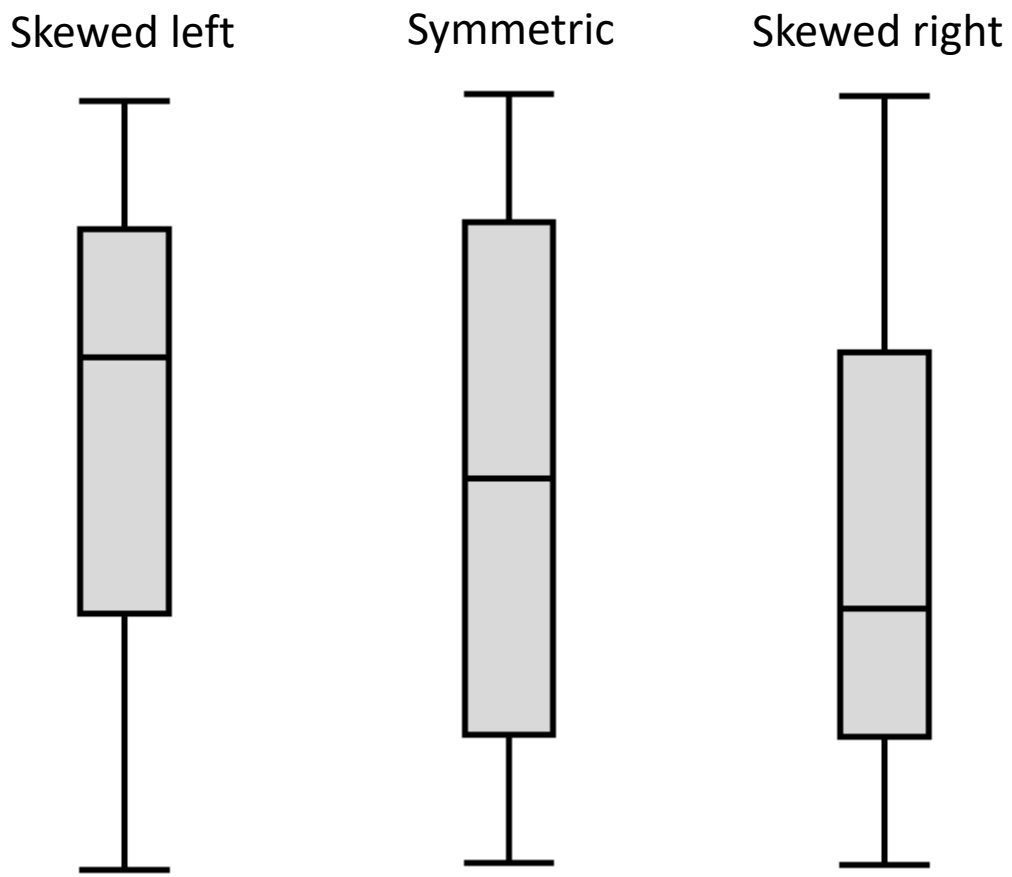


Percentiles for head circumference:

0%	25%	50%	75%	100%
21	25	27	28	35

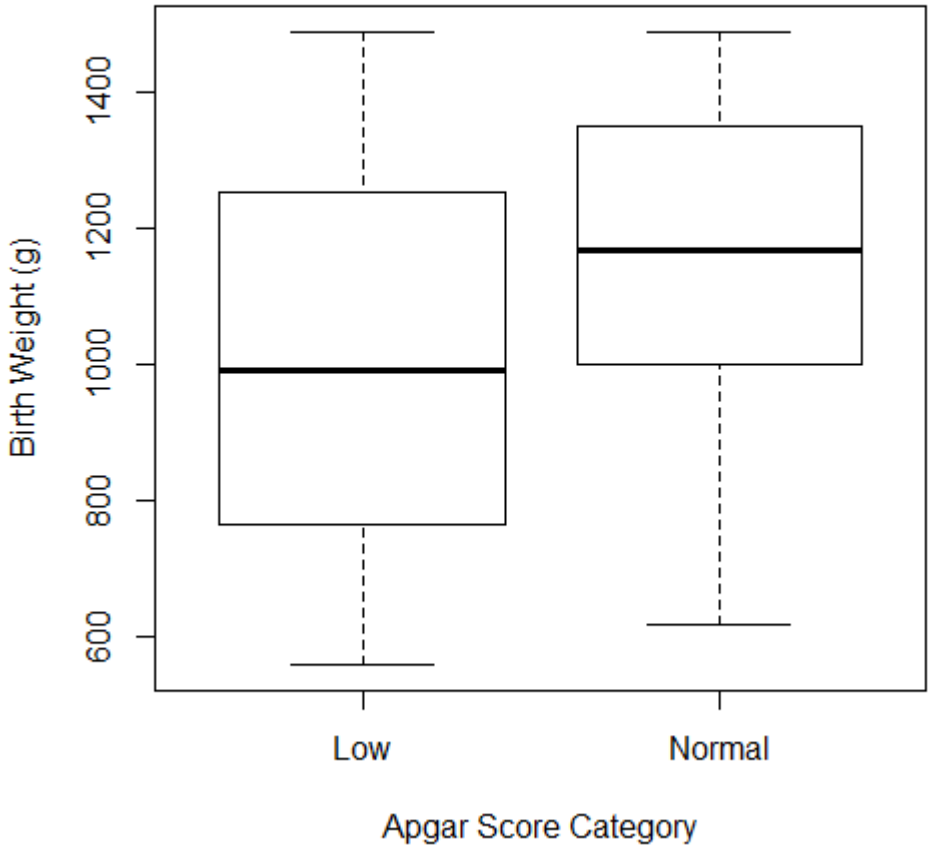
Boxplot

- Can sometimes show skew
- Long bottom whisker/larger bottom half of box = skewed left
- Long top whisker/larger top half of box = skewed right
- Distribution of observations is much more visible in a histogram (boxplot can “hide” weird features)



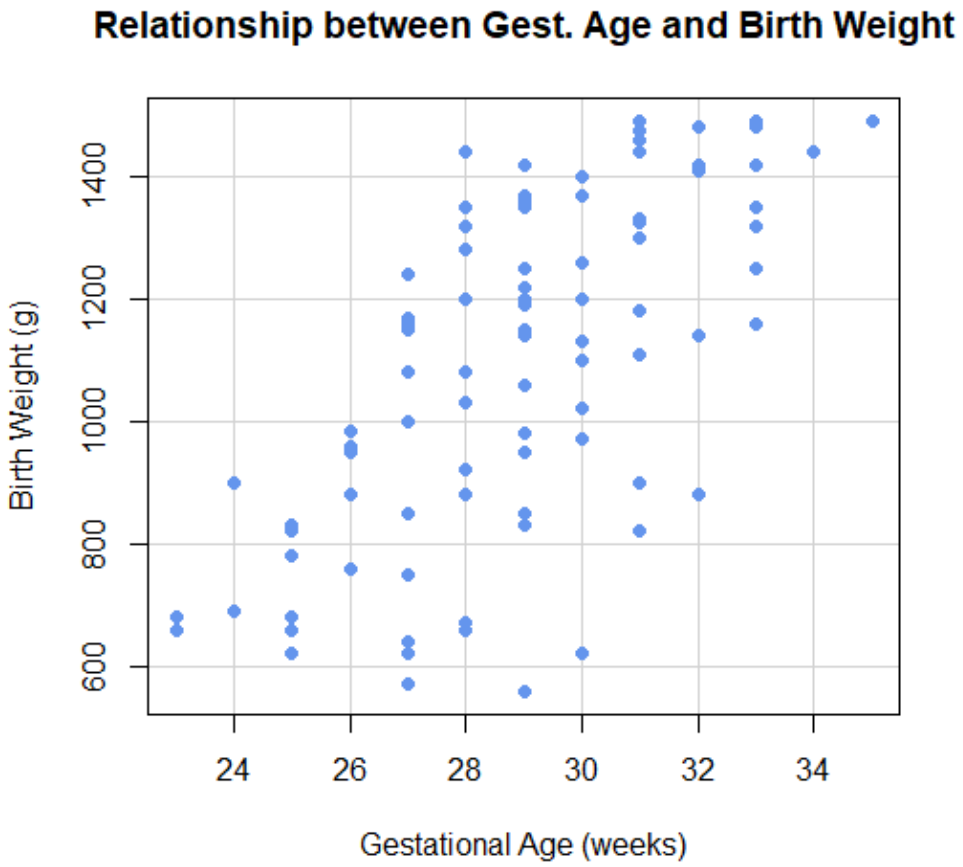
Side-by-side Boxplots

- Boxplots are particularly good for summarizing a continuous variable **split across two or more groups**



Scatterplot

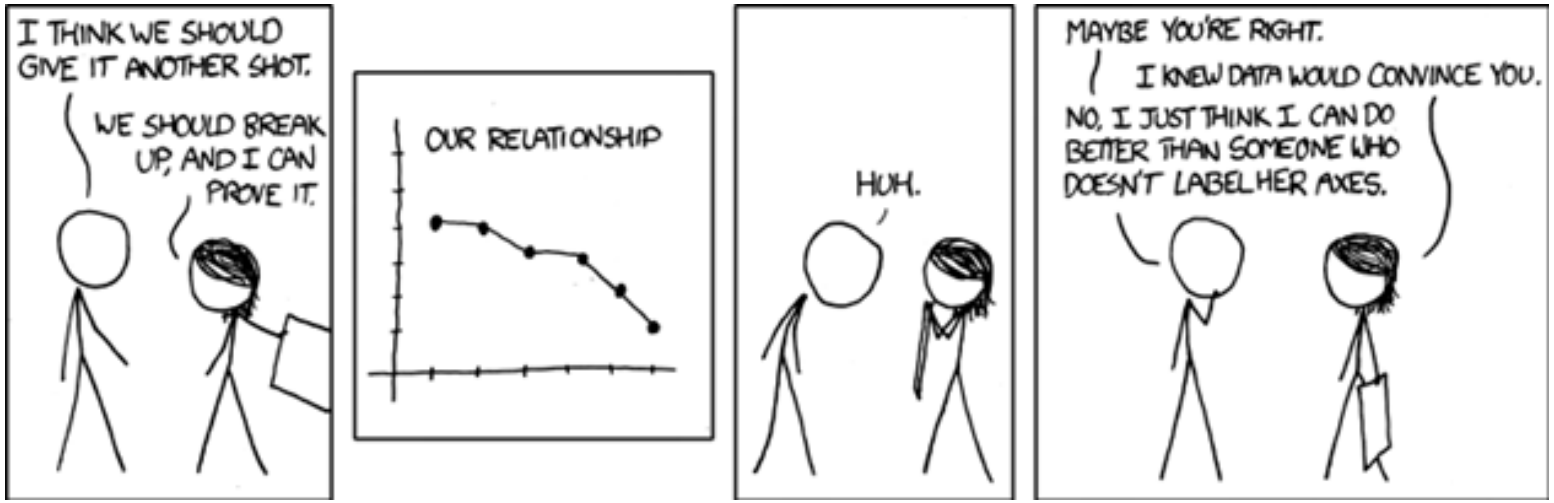
- Summarizes the relationship between two continuous variables
- Each point represents a pair of values for one subject
- Reveals whether variables are positively associated, negatively associated, or not associated



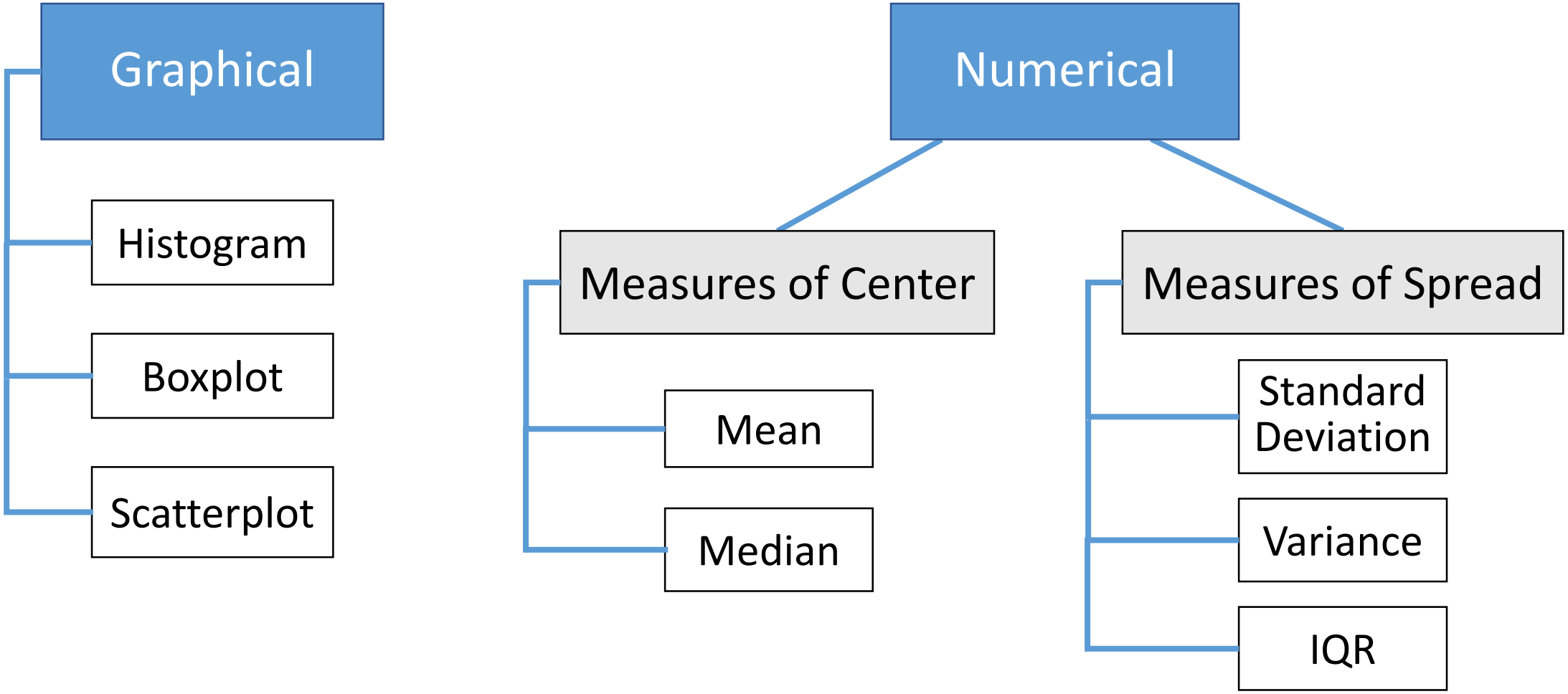
Interpretation: Gestational age and birth weight appear to be positively associated in this dataset.

Plot Details

- Don’t forget axis labels and units!!!
- Titles can also be helpful
- Legends are necessary for scatterplots containing points from two or more subgroups



Summaries of Continuous Data



Mean

- The average value, often denoted $\bar{x}$ in a sample or μ in a population
- Sum all measurements and divide by the total number of observations

Means of continuous variables in low birth weight data:

	mean	n
birthwt	1098.85	100
gestage	28.89	100
headcirc	26.45	100
length	36.82	100
momage	27.73	100
sbp	47.08	100

Median

- Order observations from smallest to largest; median is the observation in the middle (or the average of the two in the middle if there are an even number of observations)
- Also known as the 50th percentile

Medians of continuous variables in low birth weight data:

	50%	n
birthwt	1155	100
gestage	29	100
headcirc	27	100
length	38	100
momage	28	100
sbp	47	100

Mean vs. Median

- Mean is used much more than median in scientific literature
- Mean is very sensitive to unusual observations or **outliers**
  - We say that the median is **robust** to outliers

Variable
1
2
3
4
5

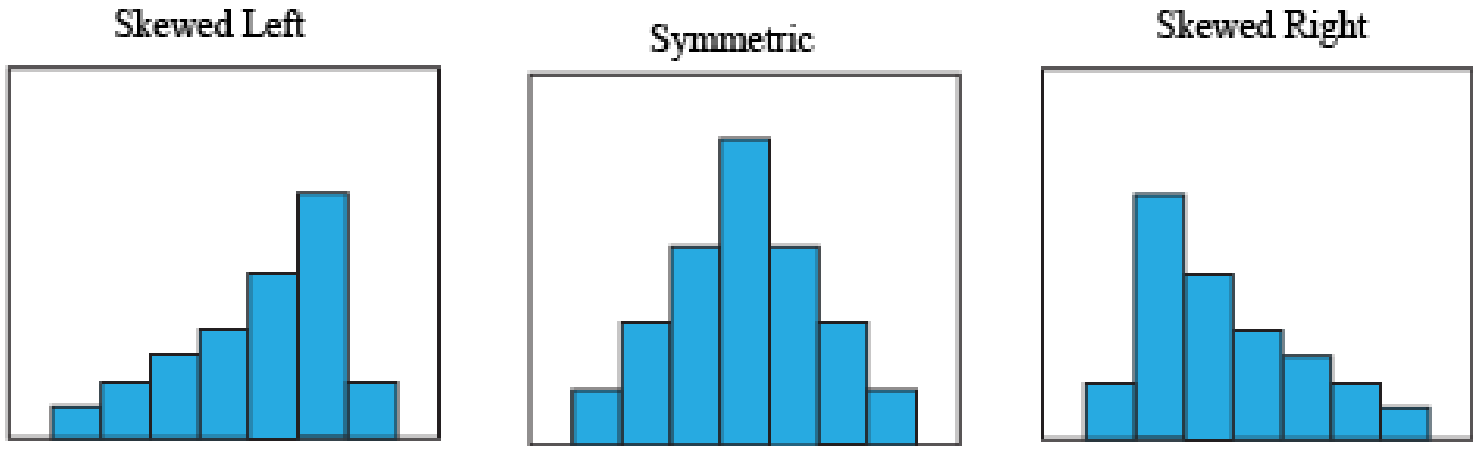
median =
mean =

Variable
1
2
3
4
50

median =
mean =

Mean vs. Median

- A long tail pulls the mean out in that direction

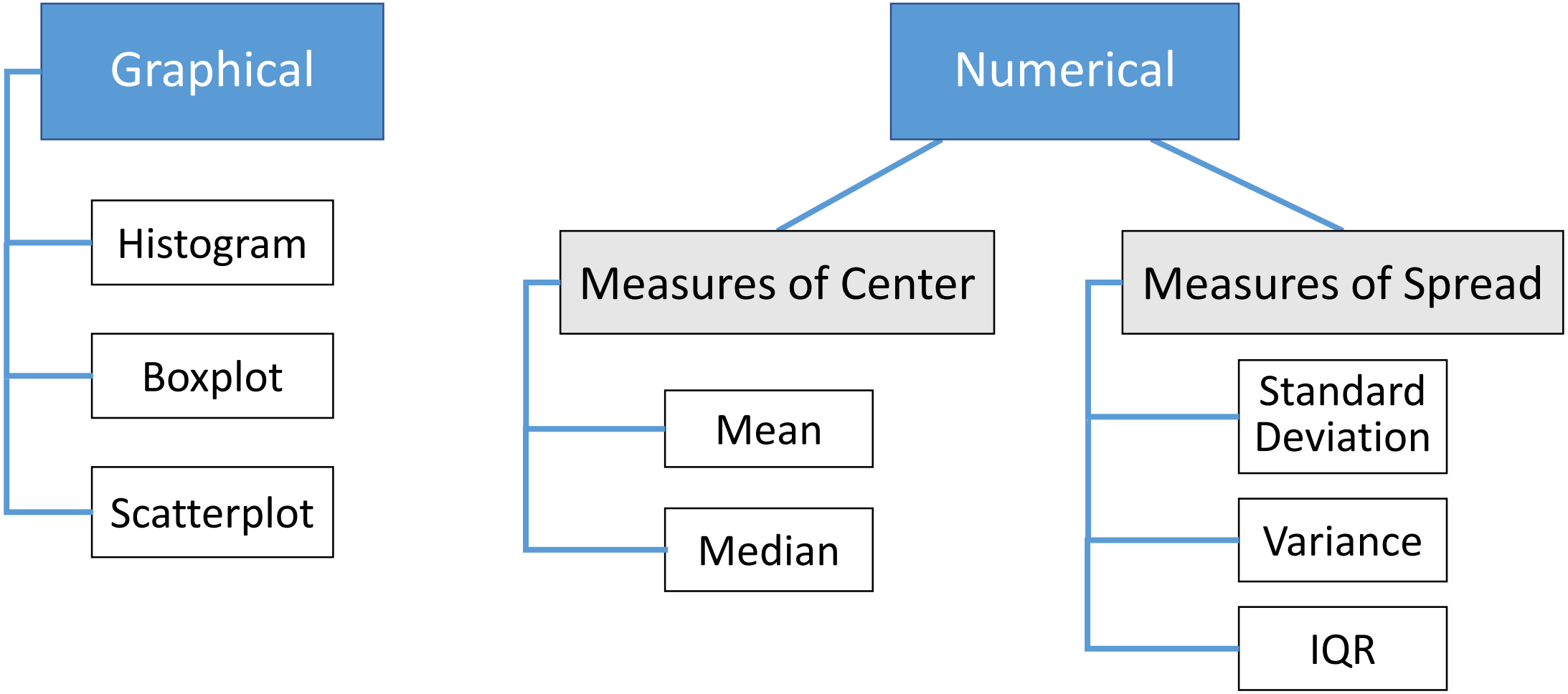


mean < median

mean = median

mean > median

Summaries of Continuous Data

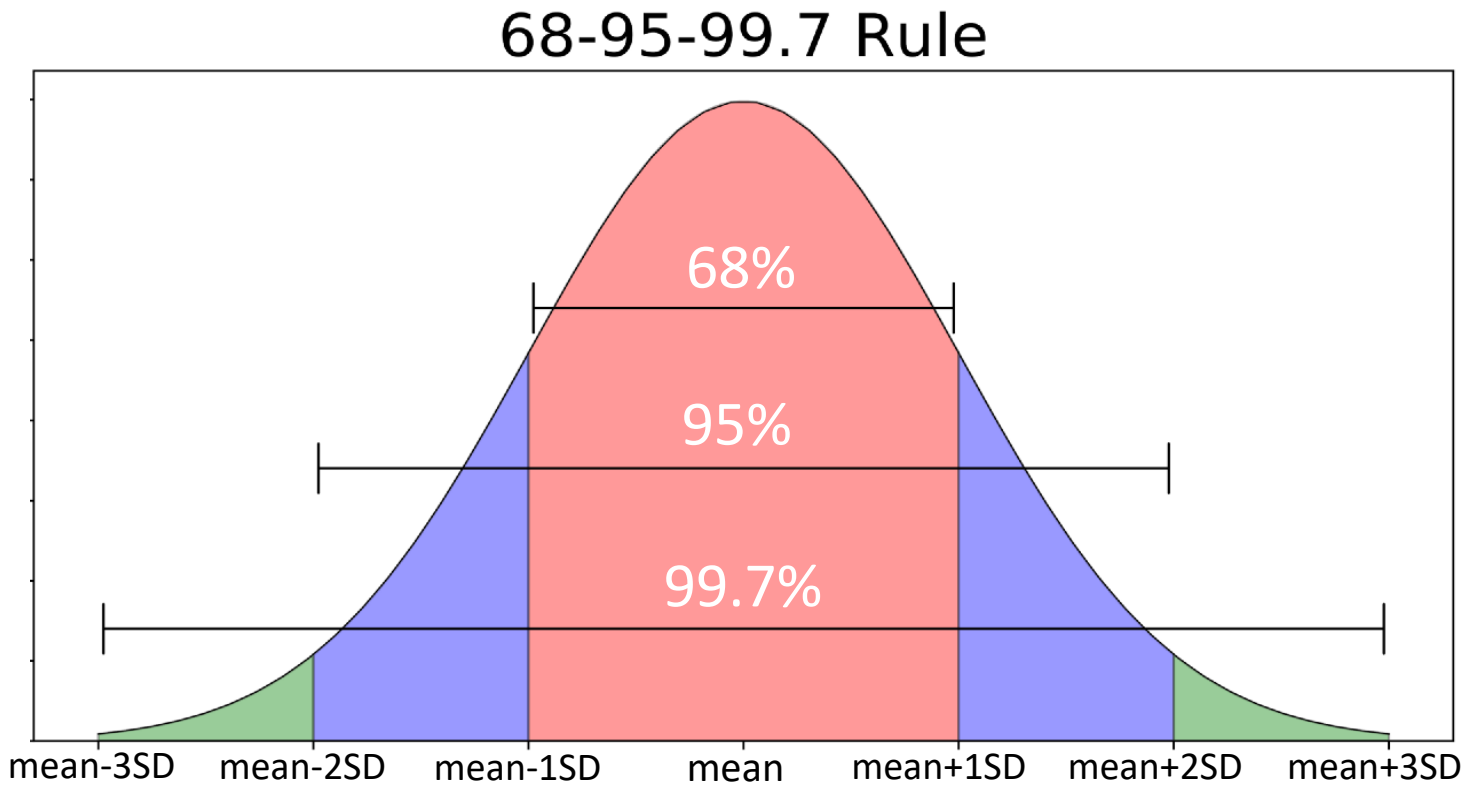


Standard Deviation

- Measures how far observations are away from the mean
  - Larger = observations are more spread out
- Usually denoted s in a sample or σ in a population
- Same units as the mean
- Calculate the distance each observation is from the mean, square it, add all of those up, divide by the total number of observations minus one, then take the square root
  - Don’t worry about the formula

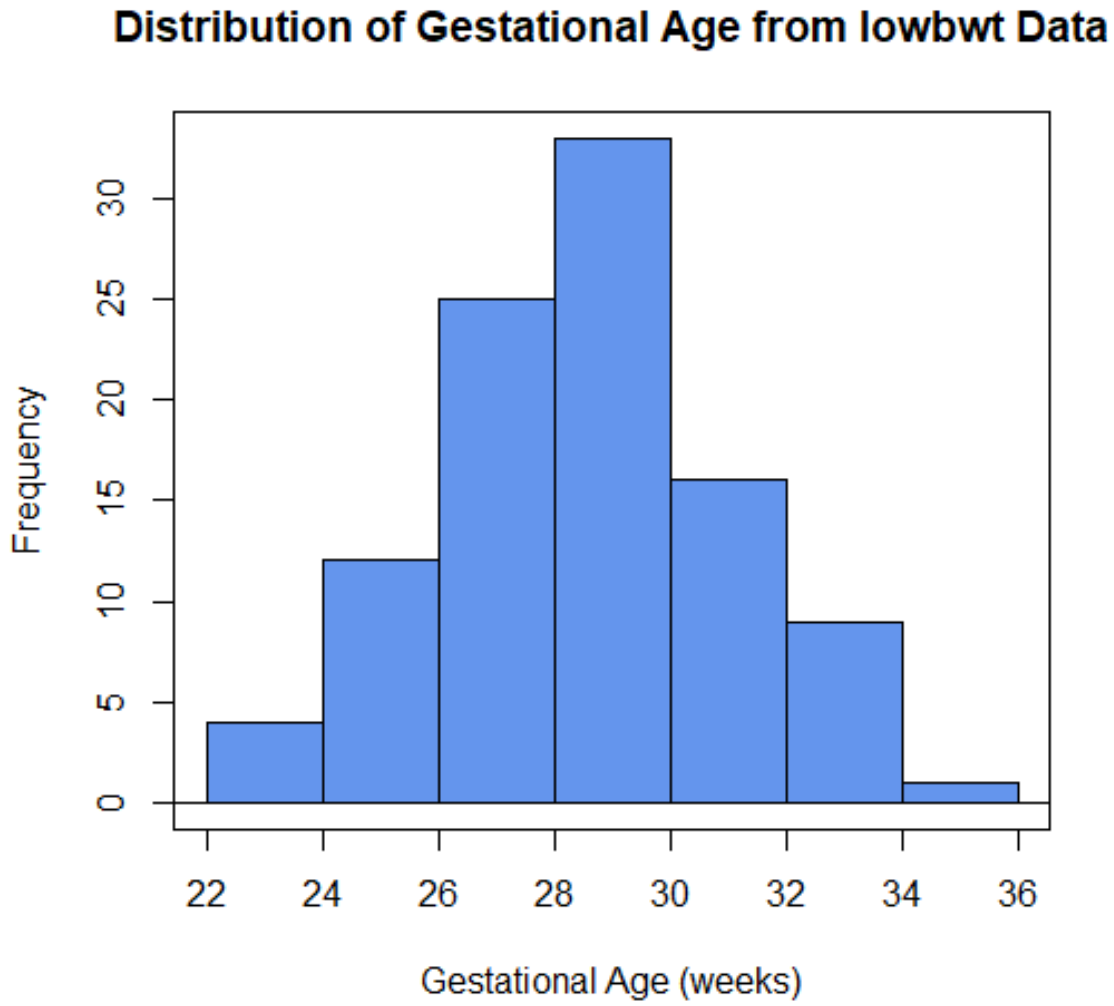
Standard Deviation

- About 68% of observations fall within 1 SD of the mean
- About 95% of observations fall within 2 SD of the mean
- About 99.7% of observations fall within 3 SD of the mean



Standard Deviation Example

- Gestational age
  - Mean = 28.9 weeks
  - SD = 2.5 weeks



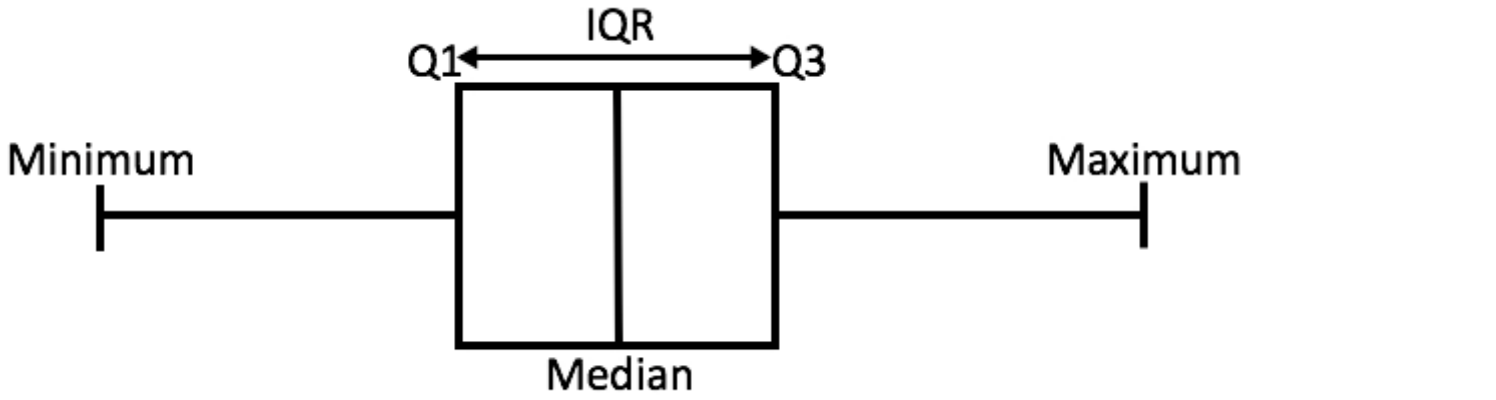
Variance

- Variance = SD^2
- Usually denoted s^2 in a sample or σ^2 in a population
- Can be thought of as an average of squared distances from the mean
  - Larger = observations are more spread out
- Units are problematic – calculated in terms of $units^2$
  - Example: Gestational age
    - Mean = 28.89 weeks
    - SD = 2.53 weeks
    - Variance = 6.42 $weeks^2$

Interquartile Range

- First quartile: 25% of observations are less than this cutoff ⟵ Q1
- Second quartile: 50% of observations are less than this cutoff ⟵ median
- Third quartile: 75% of observations are less than this cutoff ⟵ Q3

$$IQR = Q3 - Q1$$

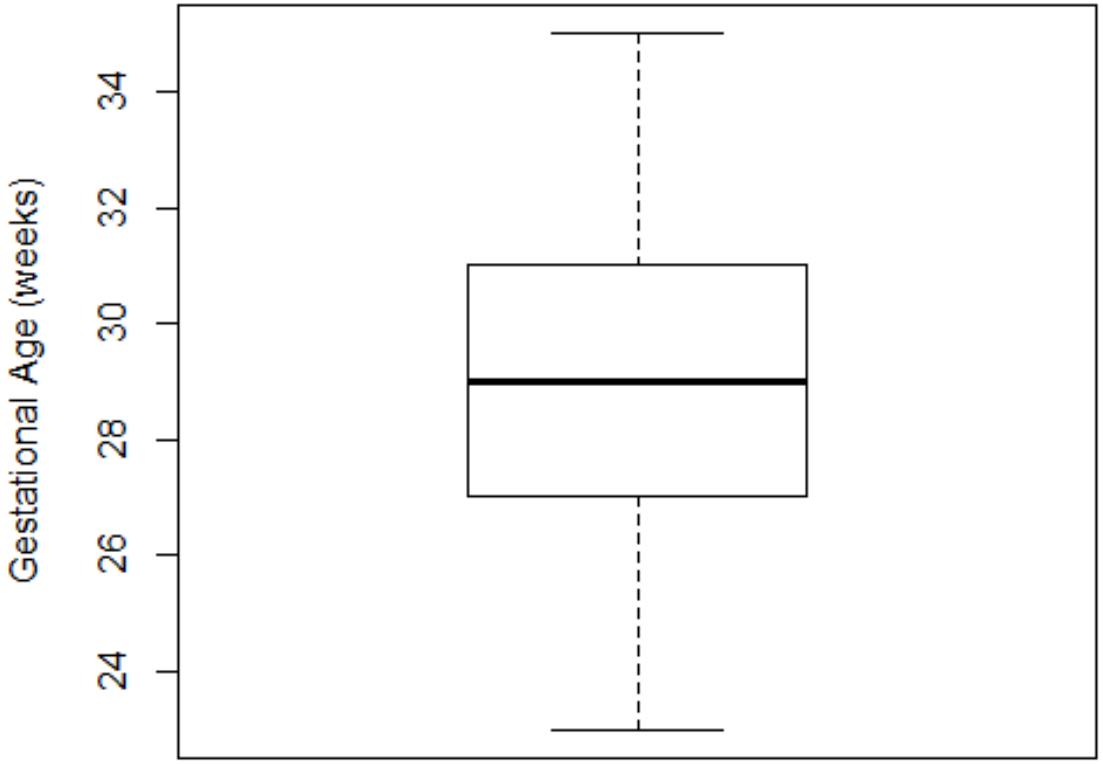


Interquartile Range Example

Percentiles for gestational age:

0%	25%	50%	75%	100%
23	27	29	31	35

IQR = Q3 – Q1
= 31 – 27
= 4 weeks



Important Points

- Construction of histograms, boxplots, and scatterplots
- Measures of center (mean, median) and when you would prefer one over the other
- Measures of spread (standard deviation, variance, IQR) and when you would prefer one over the others
- Describe the distribution of a continuous variable
  - A specific measure of center
  - A specific measure of spread
  - Description of shape (assess skew/symmetry, direction of association if comparing to another variable, etc.)