

Intro to confidence intervals

95% confidence interval for a population proportion (denoted by π):

$$\left(P - 1.96\sqrt{\frac{P(1-P)}{n}}, P + 1.96\sqrt{\frac{P(1-P)}{n}} \right)$$

where n is the sample size and P is the sample proportion

If we were to take 100 different different samples and find the confidence interval using each each of them, approximately 95 of these intervals would actually enclose π within their limits.

Example

Say we have a random sample of 100 Duke students. Of these 100, 10 camped out for last year's Duke vs. UNC-CH game. Can we provide a 95% confidence interval for the true proportion of Duke students who camped out for the game?

$$\begin{aligned} &\left(P - 1.96\sqrt{\frac{P(1-P)}{n}}, P + 1.96\sqrt{\frac{P(1-P)}{n}} \right) \\ &\left(0.1 - 1.96\sqrt{\frac{0.1(0.9)}{100}}, 0.1 + 1.96\sqrt{\frac{0.1(0.9)}{100}} \right) \\ &\hspace{15em} (0.04, 0.16) \end{aligned}$$

Types of variables

variable: characteristic which changes from person to person in a study

We tend to classify variables as *qualitative* or *quantitative*

A *qualitative* variable measures levels/categories for which it would not make sense to describe the levels/categories according to a mathematical relation.

A *quantitative* variable has numerical values for which it does make sense to compare and contrast these values using a mathematical relation.

Nominal

- Can be represented by names of categories or by numbers used to represent these groupings
- Within each group, subjects are homogeneous in terms of the variable
- Groupings must be mutually exclusive
- Groupings should be exhaustive - all subjects fall into one of the categories
- Representations (by name or number) do not assume any ordered relationships between the categories
- Examples include: M&M colors, gender/ethnic groupings, yes/no survey responses

Ordinal

- Can be represented by names of categories or by numbers used to represent groupings
- Ordinal variables meet the homogeneous, exhaustive, mutually exclusive criteria
- They assume a rational **ordering** of categories
- Still cannot say that one is twice as high as another or make other such mathematical statements
- Examples: letter grades, movie ratings, S/M/L sizes, etc.

Discrete

- Have equality of counting units - each increase in the discrete variable is a jump equal in size to the previous
- Adding, subtracting, multiplying, and dividing make sense in the context of discrete variables
- As long as the counting procedure is accurate, the value is exact
- Examples: number of students per class, number of cars per family

Continuous

- Has an unlimited number of possible values between adjacent values
- Adding, subtracting, multiplying, and dividing make sense in this context
- Theoretically, the value can always be obtained with greater precision
- Examples: many scientific measurements such as distances, weights, etc.

How and why we summarize data

When you have a large amount of data, it's important to be able to be able to get a “general picture” of the population quickly and easily.

- Graphical display options include stem-and-leaf plots, bar graphs/histograms, boxplots
- Summary statistics include measures of central tendency, variability/spread, and measures of association between variables

Bar graphs

One way to represent qualitative data is to use a bar graph (analogous to a histogram for quantitative data).

- A bar is drawn for each category
- Height of bar represents number of members of that group or area of bar represents the relative frequency for that group
- Total area of the bars equals N (number of data points) or 100%
- Bars are generally drawn so that they don't touch each other
- If data is ordinal, arrange the bars to represent ordering
- If data is nominal, arrange the bars so categories are in ABC order

Histograms

A histogram is often a good way to represent quantitative data.

- Height of bar represents number of members of that class
- Very similar to bar graph, but bars on the histogram are drawn contiguously
- If the histogram is drawn such that the area of all the bars together is 1 (100%), then it can be useful to visualize a continuous curve that interpolates between the midpoints of the bars.

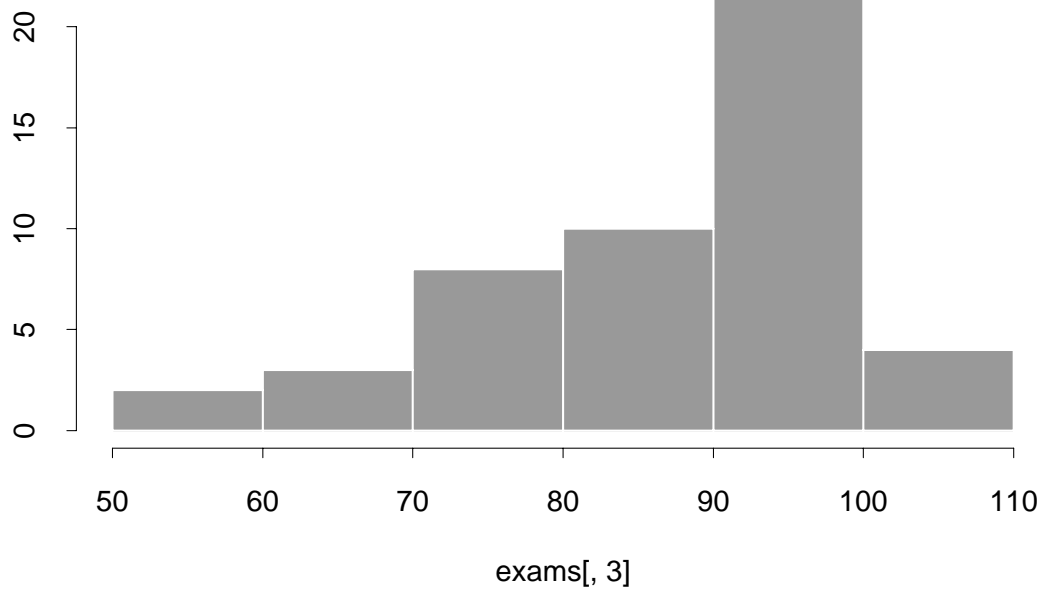
Example data set - final scores

Below are the 49 final exam scores from another class.

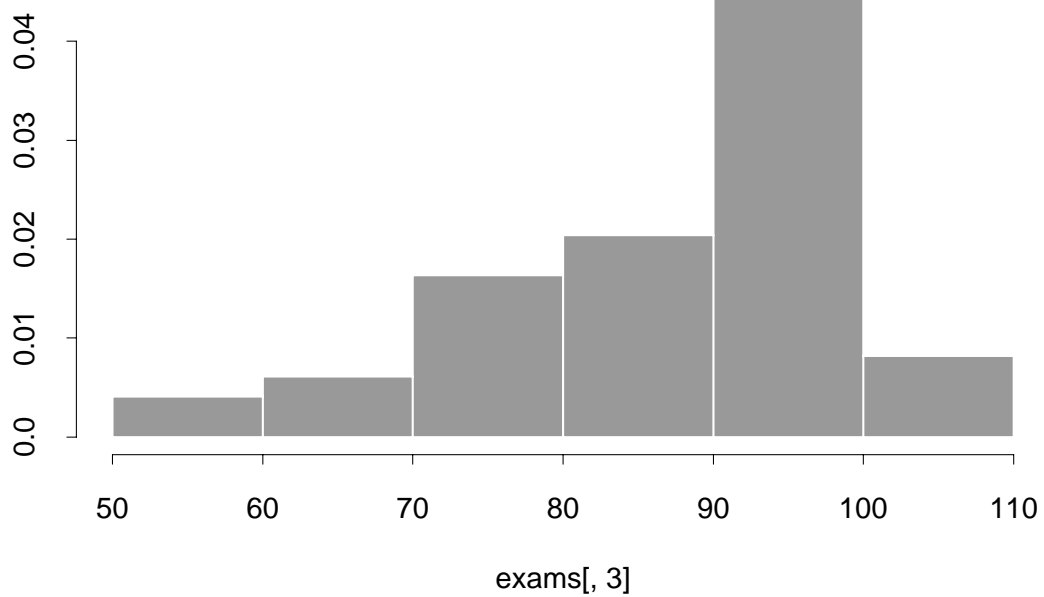
96.5 82.5 94.0 62.5 100.0 89.0 86.5 87.5 84.0 89.5 97.5 83.5 72.5 50.0
91.5 80.0 74.0 97.0 78.5 98.0 96.0 79.0 82.0 88.0 98.0 96.0 70.0 79.0
101.0 94.5 91.0 95.0 50.0 96.0 97.0 100.0 103.0 97.0 101.0 80.5 97.0
102.0 64.0 90.5 96.0 94.0 97.0 78.5 79.0

How can we make this data easier to understand or interpret?

Histogram of final grades



Scaled histogram of final grades



Stem-and-leaf plot of example scores

N = 49 Median = 91

Quartiles = 80, 97

Decimal point is 1 place to the right of the colon

```
5 : 00
6 : 24
7 : 02488999
8 : 00223467899
9 : 01144456666677777788
10 : 001123
```

Numerical summaries of data

Three characteristics to summarize:

- Central tendency: Where is the “center” of the data?
- Variability: How spread out are the data points?
- Shape: If we make a histogram of the data, how will it be shaped?

If we understand these 3 characteristics, we can get a good idea of how the data are distributed.

Ways to measure central tendency

Three commonly used statistics are:

- Mean (average): Think of it like the balance point of your histogram
- Median (50th percentile): What is the middle score?
- Mode: Which value occurs most often in the data set?

Mean

- Sum of data / total number of data points
- All the data is taken into account
- Influenced greatly by extreme values
- Mathematically very useful - use it to calculate variance, standard deviation, etc.

Median

- Middle value of the data set
- All values are not taken into account (except insofar as the general ordering of the values goes)
- Not influenced by extreme values
- Particularly useful when a distribution is *skewed*
- Not as useful mathematically as the mean

Mode

- Most frequently occurring value in the data set
- Can be thought of as a typical point in the data set
- All values are not taken into account
- In some cases, not as intuitively “central” as other measures
- Not as useful mathematically as the mean

Skewness

- Skewed to the right - right tail is long, $\text{mean} > \text{median}$
- Skewed to the left - left tail is long, $\text{mean} < \text{median}$
- In a mound-shaped symmetric distribution, $\text{mean} = \text{median} = \text{mode}$ (e.g. normal and t distributions)
- It's possible to have a symmetric, bimodal distribution with $\text{mean} = \text{median}$, and with modes on either side

Measuring dispersion

How much variation is there in the data? How are the data spread out across the different possible values?

- Range
- Inter-quartile range
- Mean squared deviation (MSD)
- Variance and standard deviation (SD)

Range

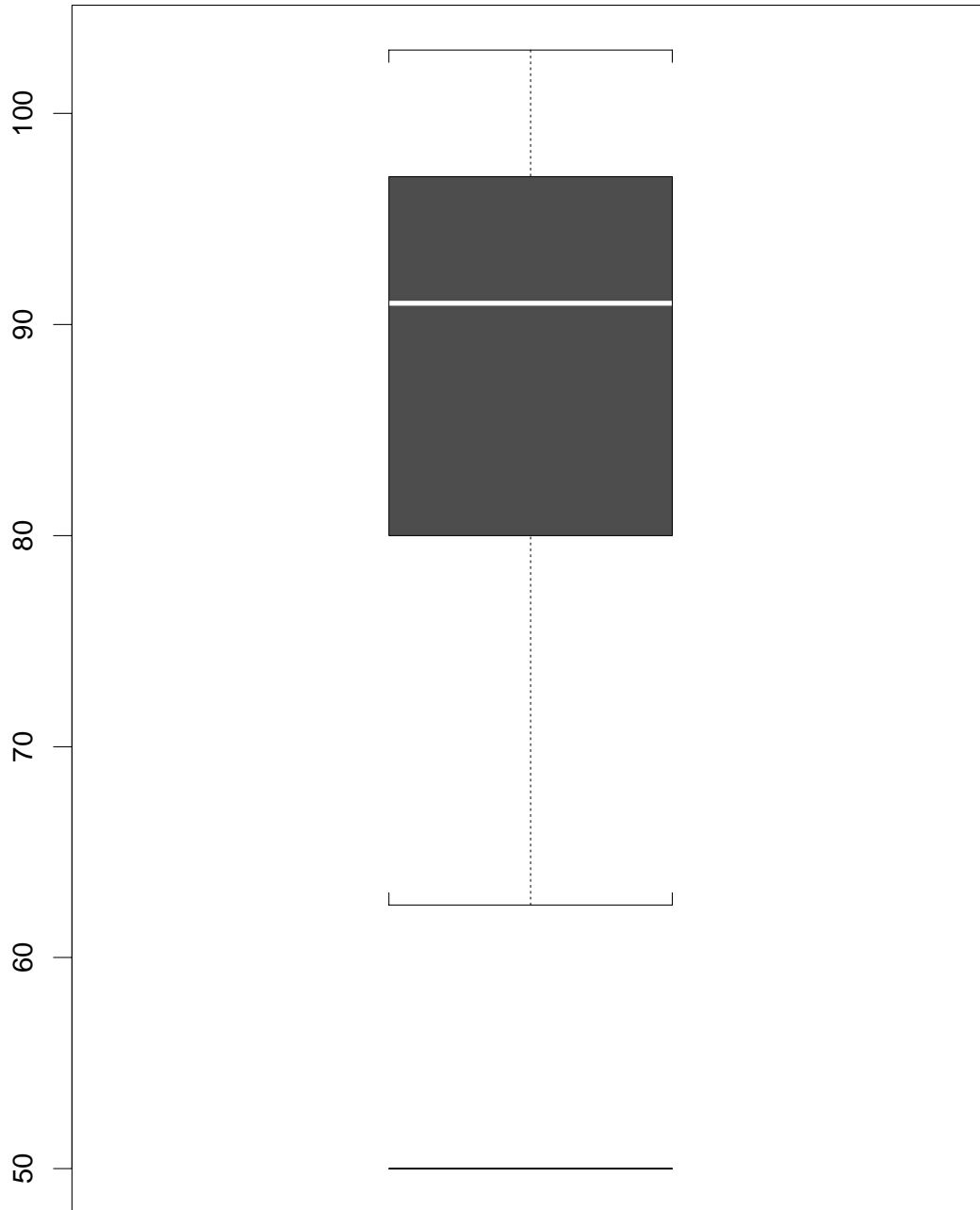
- Distance between the highest data value and lowest data value
- Easy to calculate!
- Easily affected by extreme values
- Doesn't use all the data

Inter-quartile range (IQR)

Quartiles are the data points that divide an ordered data set into quarters. So, the 1st quartile (Q1) is the data value that separates the bottom fourth of the data from the remainder; the 3rd quartile (Q3) separates the top fourth of the data from the remainder.

- Distance between 3rd quartile and 1st quartile
- Not as influenced by extreme values
- Tells where the middle half of the data is located

Boxplot of final grades



Boxplots

- Provides a clear visual representation of both central tendency and dispersion
- Edges of box give the first and third quartiles
- Line bisecting the box gives the median
- Ending points or lines of the boxplot give the minimum and maximum values (in the convention of our book)

Mean squared deviation (MSD)

- $\frac{1}{n} \sum (X - \bar{X})^2$
- Well-behaved mathematically
- Basis for obtaining variance

Variance and standard deviation

- Sample variance: $s^2 = \frac{1}{n-1} \Sigma (X - \bar{X})^2$
- Sample variance: $s^2 = \frac{n}{n-1} MSD$
- Sample standard deviation s is the square root of the sample variance
- The divisor changes to $n - 1$ because there are $n - 1$ degrees of freedom

Linear transformation

A linear transformation Y of the data X follows the form $Y = a + bX$, where a and b determine the transformation.

- $\bar{Y} = a + b\bar{X}$

- $s_Y = |b|s_X$