Summarizing Data: Measures of Center & Rank

Engineering Statistics
Section 1.3

Josh Engwer

TTU

27 January 2016

PART 0: PRELIMINARIES

NOTATION FOR SAMPLES & POPULATIONS

SAMPLE STATISTICS & POPULATION PARAMETERS

SORTED SAMPLES & ROUNDING NOTATION

Descriptive Statistics

Recall that Statistics consists of two broad branches:

- Descriptive Statistics
- Statistical Inference

The remainder of this chapter focuses squarely on Descriptive Statistics:

Definition

Descriptive Statistics is the organization, summary, visualization and presentation of data that conveys useful information about the data.

Descriptive Statistics involves:

- Data Visualization (Section 1.2)
- Numerical Summaries (this section and the next)

Notation for Samples

For methods & procedures, it's helpful to have consistent notation for samples:

Definition

(Notation for a Single Univariate Sample)

Sample as a whole is denoted by x.
The **sample size** (i.e. # data points) is denoted by n.
Each data point is denoted by a corresponding subscript: $x_1, x_2, \ldots, x_{n-1}, x_n$

Definition

(Notation for Two Univariate Samples)

Samples as a whole are denoted by x & y.
The sample sizes are denoted by n & m or n_1 & n_2
The data points are denoted by subscripts:

$$x_1, x_2, \ldots, x_n \; \& \; y_1, y_2, \ldots, y_m \quad \text{OR} \quad x_1, x_2, \ldots, x_{n_1} \; \& \; y_1, y_2, \ldots, y_{n_2}$$

For 3+ samples (rare), run thru upper-end of lowercase alphabet as needed:

$$x, y, z, w, v, u$$

Notation for Samples (Examples)

- Student Heights (in ft) $x: \quad 6.1, 3.9, 5.6, 4.0, 5.9, 5.9$
  - Sample Size $n_1 =$ (# data points in sample x) $= 6$
  - Data points $x_1 = 6.1$, $x_2 = 3.9$, $x_3 = 5.6$, $x_4 = 4.0$, $x_5 = 5.9$, $x_6 = 5.9$

- Student Weights (in lb) $y: \quad 205, 135, 183$
  - Sample Size $n_2 =$ (# data points in sample y) $= 3$
  - Data points $y_1 = 205$, $y_2 = 135$, $y_3 = 183$

- Student Eye Colors Hazel, Blue, Brown, Hazel
  - Sample Size $n_3 =$ (# data points in sample of categorical data) $= 4$
  - Sample & Data points of categorical data are <u>not</u> labeled.

There are two ways to write out a sample:

- As a list of comma-separated values
  - $x: \quad 6.1, 3.9, 5.6, 4.0, 5.9, 5.9$
  - Hazel, Blue, Brown, Hazel
- As a list of space-separated values
  - $x: \quad 6.1 \quad 3.9 \quad 5.6 \quad 4.0 \quad 5.9 \quad 5.9$
  - Hazel Blue Brown Hazel

Sample Statistics & Population Parameters

Definition

(Sample Statistic)

A **statistic** of a sample is a meaningful characteristic of a the sample.
Statistics are denoted by certain ”decorations” of the letter for the sample.

Definition

(Population Parameter)

A **parameter** of a population is a meaningful characteristic of the population.
Parameters are often (but not always) denoted by lower-case Greek letters.

Definition

(Notation for Populations)

A population itself is never denoted by a letter.
However, the size of a finite population is denoted by N.
For two finite populations, their sizes are denoted N & M OR N_1 & N_2

Sorted Samples

With discrete numerical data, it's important for some sample statistics that the sample is sorted in ascending order.

As it happens, there's corresponding notation for a sorted sample:

Definition

(Sorted Samples)

Given a sample with n data points $\quad x: \; x_1, x_2, \ldots, x_{n-1}, x_n$

Then the corresponding **sorted sample** is $\quad x: \; x_{(1)}, x_{(2)}, \ldots, x_{(n-1)}, x_{(n)}$

where the data points are sorted in ascending order:

$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n-1)} \leq x_{(n)}$$

$x_{(1)}$ is the **smallest data point** in the sample.
$x_{(n)}$ is the **largest data point** in the sample.

EXAMPLE: Given sample $x: \; 5, 4, 8 \quad \Longrightarrow \quad x_1 = 5, \; x_2 = 4, \; x_3 = 8$

Then, the sorted sample is $x: \; 4, 5, 8 \quad \Longrightarrow \quad x_{(1)} = 4, \; x_{(2)} = 5, \; x_{(3)} = 8$

Rounding Numbers (Compact Notation)

It is convenient to have mathematical notation for **rounding numbers**.

Always Round Down:	$\lfloor 3 \rfloor = 3$	$\lfloor 3.1 \rfloor = 3$	$\lfloor 3.5 \rfloor = 3$	$\lfloor 3.9 \rfloor = 3$
Always Round Up:	$\lceil 3 \rceil = 3$	$\lceil 3.1 \rceil = 4$	$\lceil 3.5 \rceil = 4$	$\lceil 3.9 \rceil = 4$
Round to Nearest Integer:	$[\![3]\!] = 3$	$[\![3.1]\!] = 3$	$[\![3.5]\!] = 4$	$[\![3.9]\!] = 4$

$\lfloor x \rfloor$ is called the **floor function**.

$\lceil x \rceil$ is called the **ceiling function**.

PART I

PART I: MEASURES OF CENTER FOR DISCRETE NUMERICAL DATA

MEAN, MEDIAN, TRIMMED MEAN

Mean of a Sample

Definition

(Mean of a Discrete Numerical Sample)

Given a sample with n data points $\quad x: \;\; x_1, x_2, \ldots, x_n$
Then its **mean**, denoted $\bar{x}$, is the average of the sample.

$$\bar{x} := \frac{1}{n}\sum_{k=1}^{n} x_k = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

NOTE: $\bar{x}$ is pronounced "x bar"

REMARK: The symbol $:=$ translates to "is defined to be".

Median of a Sample

Definition

(Median of a Discrete Numerical Sample)

Given a sample with n data points $\quad x: \quad x_1, x_2, \ldots, x_n$

Then its **median**, denoted $\widetilde{x}$, is the middle value of the sorted sample.

$$\widetilde{x} := \begin{cases} x_{([n+1]/2)} & , n \text{ odd} \\ \dfrac{x_{(n/2)} + x_{(1+[n/2])}}{2} & , n \text{ even} \end{cases} = \begin{cases} \text{Middle data point in sorted sample} & , n \text{ odd} \\ \text{Average of the two middle data points in sorted sample} & , n \text{ even} \end{cases}$$

NOTE: $\widetilde{x}$ is pronounced "x tilde" OR "x twiddle"

For instance, given sample $\quad x: \quad 6, 4, 5, 7, 1, 7, 2$

First, sort the sample: $\quad x: \quad 1, 2, 4, \mathbf{5}, 6, 7, 7$

Then, since sample size $(n = 7)$ is odd, the median is:

$$\widetilde{x} = x_{([n+1]/2)} = x_{(4)} = \boxed{\mathbf{5}}$$

Median of a Sample

Definition

(Median of a Discrete Numerical Sample)

Given a sample with n data points $\quad x: \quad x_1, x_2, \ldots, x_n$

Then its **median**, denoted $\widetilde{x}$, is the middle value of the sorted sample.

$$\widetilde{x} := \begin{cases} x_{([n+1]/2)} & , n \text{ odd} \\ \dfrac{x_{(n/2)} + x_{(1+[n/2])}}{2} & , n \text{ even} \end{cases} = \begin{cases} \text{Middle data point in sorted sample} & , n \text{ odd} \\ \text{Average of the two middle data points in sorted sample} & , n \text{ even} \end{cases}$$

NOTE: $\widetilde{x}$ is pronounced "x tilde" OR "x twiddle"

For instance, given sample $\quad y: \quad 6, 4, 5, 7, 9, 1, 7, 2$

First, sort the sample: $\quad y: \quad 1, 2, 4, \mathbf{5}, \mathbf{6}, 7, 7, 9$

Then, since sample size $(n = 8)$ is even, the median is:

$$\widetilde{y} = \frac{y_{(n/2)} + y_{(1+[n/2])}}{2} = \frac{y_{(4)} + y_{(5)}}{2} = \frac{\mathbf{5}+\mathbf{6}}{2} = \boxed{\mathbf{5.5}}$$

Trimmed Mean of a Sample

Definition

(Trimmed Mean of a Discrete Numerical Sample)

Given a sample with n data points $\quad x: \ x_1, x_2, \ldots, x_n$
Then its $p\%$ **trimmed mean**, $\bar{x}_{tr(p\%)}$, is the mean of the dataset resulting from eliminating the smallest $p\%$ and largest $p\%$ of the sorted sample.

$\bar{x}_{tr(10\%)} :=$ Mean of sorted sample x with largest 10% & smallest 10% removed
$\bar{x}_{tr(25\%)} :=$ Mean of sorted sample x with largest 25% & smallest 25% removed

Relevant trimming percentages tend to be moderate: between 5% & 25%

$\bar{x}_{tr(10\%)}$ is spoken as "the 10% trimmed sample mean."

REMARK:
For simplicity, the trimming percentage will always evenly divide sample size n.
In other words, the expression $np/100$ will always be an integer.

Otherwise, **interpolation** would be needed which complicates matters!!

Interpolation is formally encountered in **Numerical Analysis**. (MATH 4310)

Mean, Median, Trimmed Means of a Population

The mean, median, and trimmed means can be computed for populations.

Definition

(Notation for Mean, Median, Trimmed Means of a Population)

The **population mean** is denoted by μ. ("mew")
The **population median** is denoted by $\widetilde{\mu}$. ("mew tilde" or "mew twiddle")

The $p\%$ **trimmed population mean** is denoted by $\mu_{tr(p\%)}$.
The 10% **trimmed population mean** is denoted by $\mu_{tr(10\%)}$.
The 25% **trimmed population mean** is denoted by $\mu_{tr(25\%)}$.

REMARKS:

μ is the lower-case Greek letter mu.

Computing μ, $\widetilde{\mu}$, etc for finite populations is not practical due to their enormity.

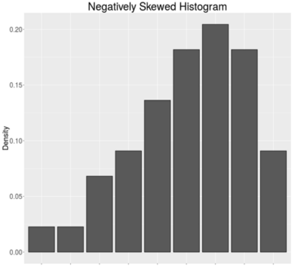
Computing μ, $\widetilde{\mu}$, etc for infinite populations will be encountered in Chapter 4.

Mean, Median and Skewness of a Sample

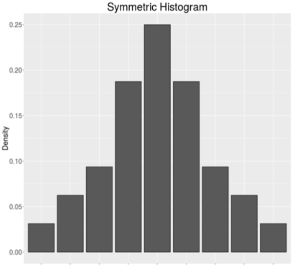
Proposition

Given a sample with n data points $x: \quad x_1, x_2, \ldots, x_n$ *Then:*

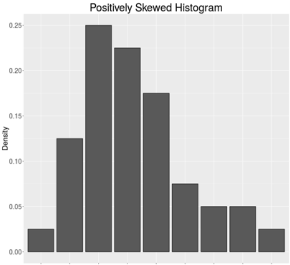
- *If $\bar{x} < \tilde{x}$, then the sample is negatively skewed.*
- *If $\bar{x} = \tilde{x}$, then the sample is symmetric.*
- *If $\bar{x} > \tilde{x}$, then the sample is positively skewed.*



$\overline{x} < \widetilde{x}$



$\overline{x} = \widetilde{x}$



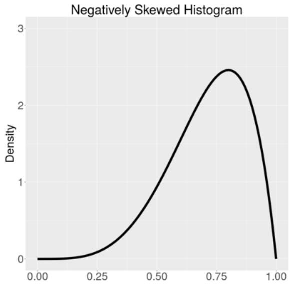
$\overline{x} > \widetilde{x}$

Mean, Median and Skewness of a Population

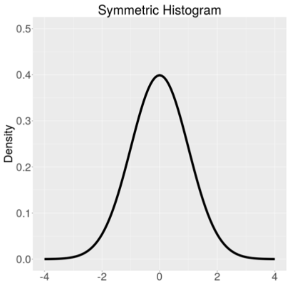
Proposition

Given a population with mean μ and median $\widetilde{\mu}$. Then:

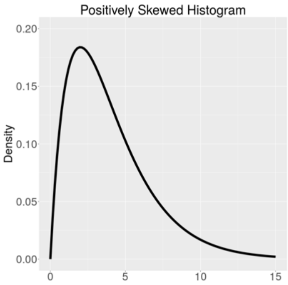
- *If $\mu < \widetilde{\mu}$, then the population is negatively skewed.*
- *If $\mu = \widetilde{\mu}$, then the population is symmetric.*
- *If $\mu > \widetilde{\mu}$, then the population is positively skewed.*



$\mu < \widetilde{\mu}$



$\mu = \widetilde{\mu}$



$\mu > \widetilde{\mu}$

Measures of Center & Their Sensitivity to Outliers

The sample mean, median, and trimmed mean are examples of **statistics**.
The sample mean, median, and trimmed mean are all **measures of center**.
This means they indicate "central locations" of the sample in unique ways.

A question that's important for many situations is:

HOW SENSITIVE ARE THESE MEASURES OF CENTER TO **OUTLIERS**??

- The mean, $\bar{x}$, is extremely sensitive to outliers.
- Lightly-trimmed means (e.g. $\bar{x}_{tr(5\%)}$) are largely sensitive to outliers.
- Heavily-trimmed means (e.g. $\bar{x}_{tr(25\%)}$) are largely insensitive to outliers.
- The median, $\widetilde{x}$, is almost completely insensitive to outliers.

So, if a subdivision has all its houses priced in the $100,000's and, later, a ten million-dollar house is built there, then:

- the <u>mean</u> house price will <u>increase substantially</u>...
- ...but the <u>median</u> house price will only <u>increase slightly</u>.

PART II: MEASURES OF RANK FOR DISCRETE NUMERICAL DATA

PERCENTILES, QUARTILES, HINGES

Percentiles of a Sample

Definition

(Percentile of a Discrete Numerical Sample)

Given a sample with n data points $\quad x: \quad x_1, x_2, \ldots, x_n$

Then its **p-th percentile**, denoted $x_{p/100}$, is the smallest data point such that $p\%$ of the sample is less than or equal to that data point:

$$x_{p/100} := x_{(\lceil np/100 \rceil)} = \left(\left\lceil \frac{np}{100} \right\rceil \right) \text{-th data point in } \underline{\text{sorted}} \text{ sample}$$

e.g. $(37\% \text{ of sample } x) \leq x_{0.37} \equiv (37^{th} \text{ percentile of sample } x)$

e.g. $(98\% \text{ of sample } y) \leq y_{0.98} \equiv (98^{th} \text{ percentile of sample } y)$

REMARK: The symbol $\equiv$ translates to ”represents” OR ”is represented by”.

Software packages (e.g. MATLAB, R, SPSS, SAS, Minitab) may define percentiles slightly differently.

Quartiles of a Sample

Definition

(Quartiles of a Discrete Numerical Sample)

Given a sample with n data points $x: \quad x_1, x_2, \ldots, x_n$ Then:

(1) $x_{Q1} := x_{0.25} \equiv 1^{st}$ **quartile** of sample x

- i.e. $(25\%$ of sample $x) \leq (1^{st}$ quartile of sample $x)$

(2) $x_{Q2} := x_{0.50} \equiv 2^{nd}$ **quartile** of sample x

- i.e. $(50\%$ of sample $x) \leq (2^{nd}$ quartile of sample $x)$
- 2^{nd} quartile, x_{Q2}, is never used since it's exactly or very close to median, $\widetilde{x}$.

(3) $x_{Q3} := x_{0.75} \equiv 3^{rd}$ **quartile** $(75^{th}$ percentile) of sample x

- i.e. $(75\%$ of sample $x) \leq (3^{rd}$ quartile of sample $x)$

REMARK: Software packages (e.g. MATLAB, R, SPSS, SAS, Minitab) may define quartiles slightly differently.

Hinges of a Sample

Definition

(Hinges of a Discrete Numerical Sample)

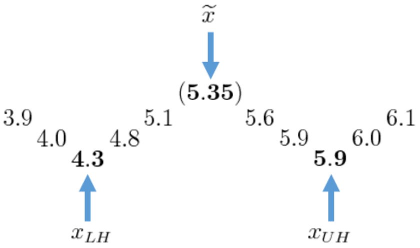
Given a sample with n data points $\quad x: \quad x_1, x_2, \ldots, x_n \quad$ Then:

(1) its **lower hinge**, x_{LH}, is the median of the lower half of sorted sample.

(2) its **middle hinge**, x_{MH}, is exactly the median of entire sample: $\quad x_{MH} = \widetilde{x}$

(3) its **upper hinge**, x_{UH}, is the median of the upper half of sorted sample.

For instance, given sample $\quad x: \quad 3.9, 4.0, 4.3, 4.8, 5.1, 5.6, 5.9, 5.9, 6.0, 6.1$



Hinges of a Sample

Definition

(Hinges of a Discrete Numerical Sample)

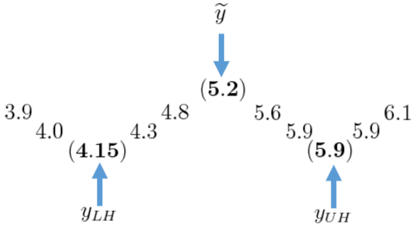
Given a sample with n data points $\quad x: \; x_1, x_2, \ldots, x_n \quad$ Then:

(1) its **lower hinge**, x_{LH}, is the median of the lower half of sorted sample.

(2) its **middle hinge**, x_{MH}, is exactly the median of entire sample: $\quad x_{MH} = \widetilde{x}$

(3) its **upper hinge**, x_{UH}, is the median of the upper half of sorted sample.

For instance, given sample $\quad y: \; 5.9, 3.9, 5.9, 4.8, 5.6, 4.0, 6.1, 4.3$



Hinges of a Sample

Definition

(Hinges of a Discrete Numerical Sample)

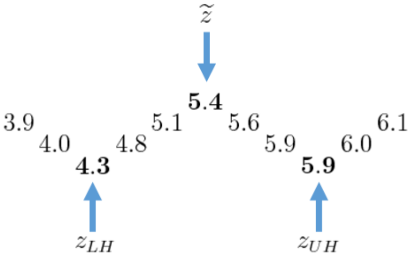
Given a sample with n data points $\quad x: \quad x_1, x_2, \ldots, x_n \quad$ Then:

(1) its **lower hinge**, x_{LH}, is the median of the lower half of sorted sample.

(2) its **middle hinge**, x_{MH}, is exactly the median of entire sample: $\quad x_{MH} = \widetilde{x}$

(3) its **upper hinge**, x_{UH}, is the median of the upper half of sorted sample.

For instance, given sample $\quad z: \quad 3.9, 4.0, 6.1, 4.3, 5.9, 4.8, 5.1, 6.0, 5.6, 5.4, 5.9$



Textbook Logistics for Section 1.3

- Difference(s) in Terminology:

TEXTBOOK TERMINOLOGY	SLIDES/OUTLINE TERMINOLOGY
Measures of **Location** (Mean, Median, ...)	Measures of **Center**
Measures of **Location** (Percentiles, Hinges, ...)	Measures of **Rank**

- Difference(s) in Notation:

CONCEPT	TEXTBOOK NOTATION	SLIDES/OUTLINE NOTATION
Arbitrary Trimming Percentage	$100\alpha\%$	$p\%$
Trimmed Mean	$\overline{x}_{tr(12.5)}$	$\overline{x}_{tr(12.5\%)}$

Textbook Logistics for Section 1.3

- Ignore ”Categorical Data and Sample Proportions” section (pg 34)
  - Sample proportions were encountered with freq. tables & histograms.
  - This numerical look at sample proportions will be encountered in Ch6.

Fin.