SUMMARIZING DATA: PRELIMINARIES [DEVORE 1.3]

- **NOTATION FOR SAMPLES:** For methods & procedures, it's helpful to have consistent notation for samples:
- **NOTATION FOR A SINGLE 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: $x_1, x_2, \ldots, x_{n-1}, x_n$
- **NOTATION FOR TWO 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: $x_1, x_2, \ldots, x_n$ & $y_1, y_2, \ldots, y_m$ OR $x_1, x_2, \ldots, x_{n_1}$ & $y_1, y_2, \ldots, y_{n_2}$
- **NOTATION FOR THREE+ SAMPLES:** Run thru upper-end of lowercase alphabet as needed: x, y, z, w, v, u
- **NOTATION FOR SAMPLES (EXAMPLES):**
  - Student Heights (in ft) $x: \; 6.1, 3.9, 5.6, 4.0, 5.9, 5.9$
    * Sample Size $n_1 = (\#\text{ data points in sample } x) = 6$
    * 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) $y: \; 205, 135, 183$
    * Sample Size $n_2 = (\#\text{ data points in sample } y) = 3$
    * Data points $y_1 = 205$, $y_2 = 135$, $y_3 = 183$
  - Student Eye Colors Hazel, Blue, Brown, Hazel
    * Sample Size $n_3 = (\#\text{ data points in sample of categorical data}) = 4$
    * Sample & Data points of categorical data are <u>not</u> labeled.

- **SAMPLE STATISTICS:** 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.
- **POPULATION PARAMETERS:** A **parameter** of a population is a meaningful characteristic of the population.

  Parameters are often (but not always) denoted by lower-case Greek letters.

- **SORTED SAMPLES:** With <u>discrete numerical data</u>, it's important for some sample statistics that the sample is <u>sorted</u> in <u>ascending order</u>.

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

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

  where the data points are sorted in <u>ascending order</u>:

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

  $x_{(1)}$ is the **smallest data point** in the sample. $x_{(n)}$ is the **largest data point** in the sample.
- **SORTED SAMPLES (EXAMPLE):**

  Given sample $x: \; 5, 4, 8 \implies x_1 = 5, \; x_2 = 4, \; x_3 = 8$

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

- **ROUNDING (COMPACT NOTATION):** It is convenient to have mathematical notation for **rounding numbers**.

$$\begin{aligned}
\text{Always Round Down:} \quad & \lfloor 3 \rfloor = 3 \quad \lfloor 3.1 \rfloor = 3 \quad \lfloor 3.5 \rfloor = 3 \quad \lfloor 3.9 \rfloor = 3 \\
\text{Always Round Up:} \quad & \lceil 3 \rceil = 3 \quad \lceil 3.1 \rceil = 4 \quad \lceil 3.5 \rceil = 4 \quad \lceil 3.9 \rceil = 4 \\
\text{Round to Nearest Integer:} \quad & [\![3]\!] = 3 \quad [\![3.1]\!] = 3 \quad [\![3.5]\!] = 4 \quad [\![3.9]\!] = 4
\end{aligned}$$

SUMMARIZING DATA: MEASURES OF CENTER [DEVORE 1.3]

Throughout this page, assume the following discrete numerical sample $x: \; x_1, x_2, \ldots, x_n$

- **MEAN OF A SAMPLE:** The **mean**, denoted $\overline{x}$, is the average of the sample.

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

- **MEDIAN OF A SAMPLE:** The **median**, denoted $\widetilde{x}$, is the middle value of the sorted sample.

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

- **TRIMMED MEAN OF A SAMPLE:** The $p\%$ **trimmed mean**, $\overline{x}_{tr(p\%)}$, is the mean of the dataset resulting from eliminating the smallest $p\%$ and largest $p\%$ of the sorted sample.

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

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

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

  For simplicity, the trimming percentage will always evenly divide sample size n.

  In other words, the expression $np/100$ will always be an integer.

- **MEAN, MEDIAN, TRIMMED MEANS OF A POPULATION:**

  The **population mean** is denoted by μ. ("mew bar")

  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 $\mu_{tr(10\%)}$.

  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 IN SAMPLES:** (See the 1.3 Slides for histograms illustrating this)

  If $\overline{x} < \widetilde{x}$, then the sample is negatively skewed.

  If $\overline{x} = \widetilde{x}$, then the sample is symmetric.

  If $\overline{x} > \widetilde{x}$, then the sample is positively skewed.

- **MEAN, MEDIAN AND SKEWNESS IN POPULATIONS:** (See the 1.3 Slides for histograms illustrating this)

  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.

- **MEASURES OF CENTER AND THEIR SENSITIVITY TO OUTLIERS:**

  The mean, $\overline{x}$, is extremely sensitive to outliers.

  Lightly-trimmed means (e.g. $\overline{x}_{tr(5\%)}$) are largely sensitive to outliers.

  Heavily-trimmed means (e.g. $\overline{x}_{tr(25\%)}$) are largely insensitive to outliers.

  The median, $\widetilde{x}$, is almost completely insensitive to outliers.

SUMMARIZING DATA: MEASURES OF RANK [DEVORE 1.3]

Throughout this page, assume the following discrete numerical sample $\quad x: \; x_1, x_2, \ldots, x_n$

- **PERCENTILES OF A SAMPLE:** The **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:

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

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

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

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

- **QUARTILES OF A SAMPLE:**
  (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)

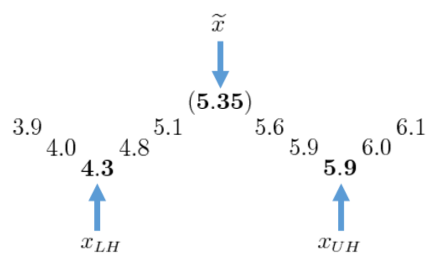
- **HINGES OF A SAMPLE:**
  (1) its **lower hinge**, x_{LH}, is the median of the lower half of sorted sample.

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

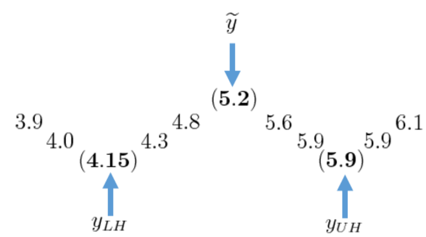
  (3) its **upper hinge**, x_{UH}, is the median of the upper half of sorted sample.

- **HINGES OF A SAMPLE (EXAMPLES):** [Parentheses around a value indicates it is not a data point in sample.]

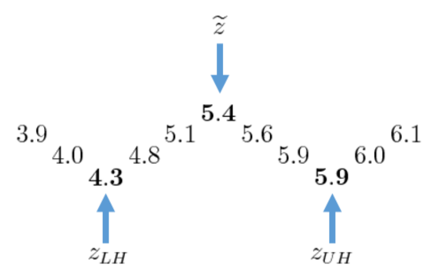
For instance, given sample $\quad x: \; 3.9, 4.0, 4.3, 4.8, 5.1, 5.6, 5.9, 5.9, 6.0, 6.1$



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



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



EX 1.3.1: Given the following sample of lifetimes of light bulbs (in years):

$$x: \ 8.5, \ 6.8, \ 7.7, \ 10.0, \ 11.3, \ 10.0, \ 9.9$$

a) Compute the sample mean, $\overline{x}$.

b) Compute the sample median, $\widetilde{x}$.

c) **Without visualizing the data**, identify the skewness of the sample.

d) Suppose the largest & smallest data points were removed from the sample. What trimming percentage achieves this?

e) Compute the trimmed mean using the trimming percentage found in the previous part.

f) Compute the 33^{rd} percentile of the sample, $x_{0.33}$.

g) Which data point is the smallest such data point that is greater than or equal to 25% of the sample?

h) Compute the lower hinge, x_{LH}, and upper hinge, x_{UH}, of the sample.

EX 1.3.2: Given the following sample of lifetimes of light bulbs (in years):

$$y: \; 8.5, \; 6.8, \; 7.7, \; 10.0, \; 11.3, \; 10.0, \; 9.9, 18.5$$

a) Compute the sample mean, $\overline{y}$.

b) Compute the sample median, $\widetilde{y}$.

c) **Without visualizing the data**, identify the skewness of the sample.

d) Compute the 25% trimmed sample mean, $\overline{y}_{tr(25\%)}$.

e) Compute the 3^{rd} quartile of the sample, y_{Q3}.

f) Which data point is the largest such data point that is less than or equal to 87% of the sample?

g) Compute the lower hinge, y_{LH}, and upper hinge, y_{UH}, of the sample.