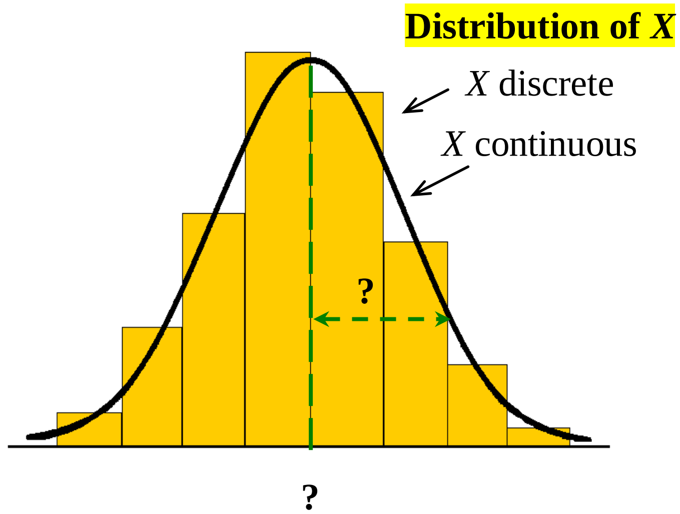
2.3 Summary Statistics – Measures of Center and Spread



POPULATION

Random Variable *X*, *numerical*

- ♦ True "center" = ???
- ♦ True "spread" = ???

parameters
("**population characteristics**")

☞ *unknown* fixed numerical values

☞ usually denoted by Greek letters, e.g., θ ("theta")

Statistical Inference

SAMPLE, size n

- ♦ Measures of center

  median, mode, mean

- ♦ Measures of spread

  range, variance, standard deviation

statistics
("**sample characteristics**")

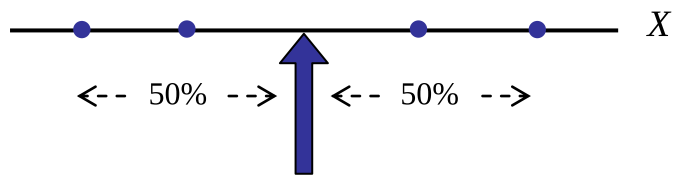
☞ *known* (or computable) numerical values obtained from sample data

☞ estimators of parameters, e.g., $\hat{\theta}$ usually denoted by corresponding Roman letters

Measures of Center

For a given numerical random variable X, assume that a random sample $\{x_1, x_2, \ldots, x_n\}$ has been selected, and *sorted* from lowest to highest values, i.e.,

$$x_1 \leq x_2 \leq \ldots \leq x_{n-1} \leq x_n$$



- **sample median** = the numerical "middle" value, in the sense that half the data values are smaller, half are larger.

  If n is odd, take the value in position # $\frac{n+1}{2}$.

  If n is even, take the average of the two closest neighboring data values, left (position # $\frac{n}{2}$) and right (position # $\frac{n}{2}$ + 1).

  Comments:

  ➢ The sample median is robust (insensitive) with respect to the presence of outliers.

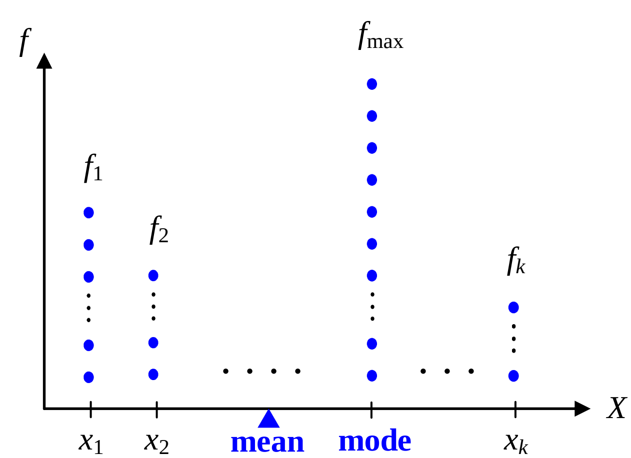
  ➢ More generally, can also define **quartiles** (Q_1 = 25% cutoff, Q_2 = 50% cutoff = **median**, Q_3 = 75% cutoff), or **percentiles** (a.k.a. **quantiles**), which divide the data values into any given p% vs. $(100 - p)$% split. Example: SAT scores

- **sample mode** = the data value with the largest frequency ($f_{\max}$)

  Comment: The sample mode is robust to outliers.

If present, *repeated* sample data values can be neatly consolidated in a **frequency table**, vis-à-vis the corresponding dotplot. (If a value x_i is not repeated, then its $f_i = 1$.)

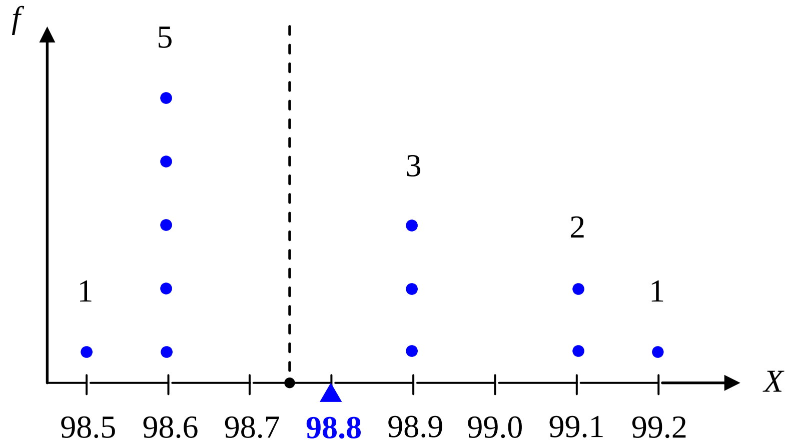
k distinct data values of X	*absolute* frequency of x_i	*relative* frequency of x_i
$\boldsymbol{x_i}$	$\boldsymbol{f_i}$	$\boldsymbol{p(x_i) = f_i / n}$
x_1	f_1	$f(x_1)$
x_2	f_2	$f(x_2)$
⋮	⋮	⋮
x_k	f_k	$f(x_k)$
	n	1



Example: $n = 12$ random sample values of X = "Body Temperature (°F)":

{98.5, 98.6, 98.6, 98.6, 98.6, 98.6, 98.9, 98.9, 98.9, 99.1, 99.1, 99.2}

x_i	f_i	$p(x_i)$
98.5	**1**	**1/12**
98.6	**5**	**5/12**
98.9	**3**	**3/12**
99.1	**2**	**2/12**
99.2	**1**	**1/12**
	$n = 12$	1



sample median $= \dfrac{98.6 + 98.9}{2} = 98.75$°F (six data values on either side)

sample mode = 98.6°F

sample mean $= \dfrac{1}{12}\big[\,(98.5)(1) + (98.6)(5) + (98.9)(3) + (99.1)(2) + (99.2)(1)\,\big]$

or, $= (98.5)\,\mathbf{\tfrac{1}{12}} + (98.6)\,\mathbf{\tfrac{5}{12}} + (98.9)\,\mathbf{\tfrac{3}{12}} + (99.1)\,\mathbf{\tfrac{2}{12}} + (99.2)\,\mathbf{\tfrac{1}{12}}$ = 98.8°F

- **sample mean** = the "weighted average" of *all* the data values

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} x_i\, f_i\,, \quad \text{where } f_i \text{ is the } \textit{absolute frequency} \text{ of } x_i$$

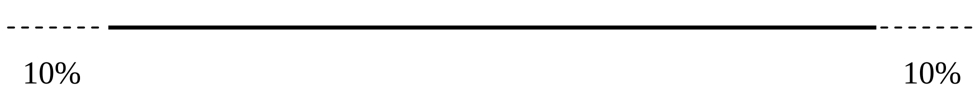
$$= \sum_{i=1}^{k} x_i\, p(x_i), \quad \text{where } p(x_i) = \frac{f_i}{n} \text{ is the } \textit{relative frequency} \text{ of } x_i$$

Comments:

➢ The sample mean is the **center of mass**, or "balance point," of the data values.

➢ The sample mean is sensitive to outliers. One common remedy for this…

Trimmed mean: Compute the sample mean after deleting a *predetermined* number or percentage of outliers from each end of the data set, e.g., "10% trimmed mean." Robust to outliers by construction.



Grouped Data – Suppose the original values had been “lumped” into categories.

Example: Recall the *grouped* “Memorial Union age” data set…

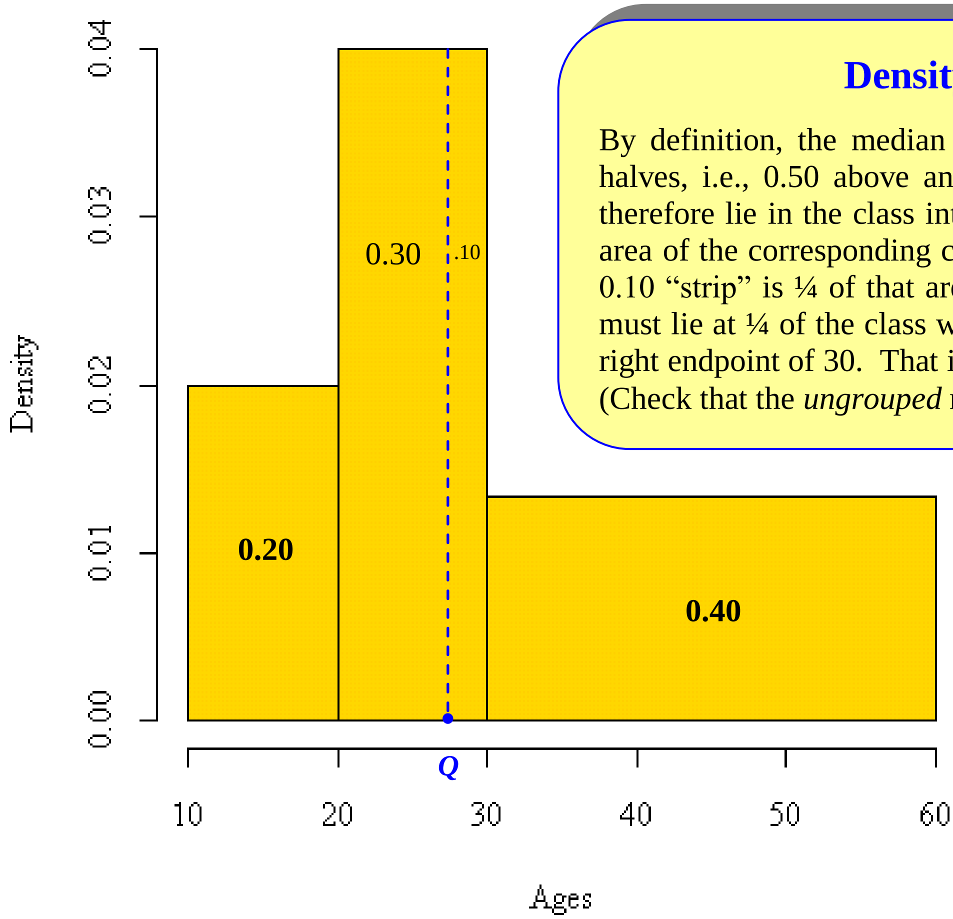
x_i	Class Interval	Frequency f_i	Relative Frequency $\frac{f_i}{n}$	Density (Rel Freq ÷ Class Width)
15	[10, 20)	**4**	0.20	0.02
25	[20, 30)	**8**	0.40	0.04
45	[30, 60)	**8**	0.40	0.013
		n = 20	1.00	

- **group mean**: Same formula as above, with x_i = *midpoint* of i^{th} class interval.

$$\bar{x}_{\text{group}} = \frac{1}{20}\left[\,(15)(4) + (25)(8) + (45)(8)\,\right] = 31.0 \text{ years}$$

Exercise: Compare this value with the *ungrouped* sample mean $\bar{x}$ = 29.2 years.

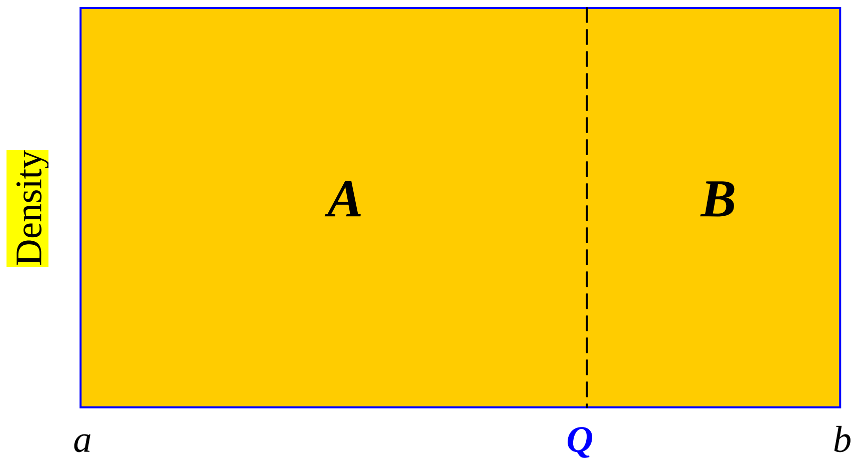
- **group median** (& other quantiles):



Density Histogram

By definition, the median ***Q*** divides the data set into equal halves, i.e., 0.50 above and below. In this example, it must therefore lie in the class interval [20, 30), and divide the 0.40 area of the corresponding class rectangle as shown. Since the 0.10 “strip” is ¼ of that area, it proportionally follows that ***Q*** must lie at ¼ of the class width 30 – 20 = 10, or 2.5, from the right endpoint of 30. That is, ***Q*** = 30 – 2.5, or ***Q* = 27.5 years**. (Check that the *ungrouped* median = 25 years.)

Formal approach ~



First, identify which class interval $[a, b)$ contains the desired quantile Q (e.g., median, quartile, etc.), and determine the respective left and right areas A and B into which it divides the corresponding class rectangle. Equating proportions for $\text{Density} = \dfrac{A+B}{b-a}$, we obtain

$$\text{Density} = \frac{A}{Q-a} = \frac{B}{b-Q},$$

from which it follows that

$$\boxed{Q = a + \frac{A}{\text{Density}}} \quad \text{or} \quad \boxed{Q = b - \frac{B}{\text{Density}}} \quad \text{or} \quad \boxed{Q = \frac{Ab+Ba}{A+B}}.$$

For example, in the grouped "Memorial Union age" data, we have $a = 20$, $b = 30$, and $A = 0.30$, $B = 0.10$. Substituting these values into any of the equivalent formulas above yields the median $Q_2 = 27.5$. ✓

Exercise: Now that Q_2 is found, use the formula again to find the first and third quartiles Q_1 and Q_3, respectively.

Note also from above, we obtain the useful formulas

$$\boxed{A = (Q-a) \times \text{Density}}$$

$$\boxed{B = (b-Q) \times \text{Density}}$$

for calculating the areas A and B, *when a value of Q is <u>given</u>!* This can be used when ***finding the area <u>between</u> two quantiles Q_1 and Q_2***. (See next page for another way.)

Alternative approach → First, form this column:

Class Interval	Frequency f_i	Relative Frequency $f_i\,/\,n$	Cumulative Relative Frequency $F_i = \frac{f_1}{n} + \frac{f_2}{n} + \ldots + \frac{f_i}{n}$
I_0	0	0	0
I_1	f_1	$f_1\,/\,n$	F_1
I_2	f_2	$f_2\,/\,n$	F_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$
I_i	f_i	$f_i\,/\,n$	F_{low} **< 0.5**
Q **= ?** in			**0.5**
$[a,\, b)$	f_{i+1}	$f_{i+1}\,/\,n$	F_{high} **> 0.5**
$\vdots$	$\vdots$	$\vdots$	$\vdots$
I_k	f_k	$f_k\,/\,n$	1
	n	1	

Next, identify F_{low} and F_{high} which bracket 0.5, and let $[a,\, b)$ be the class interval of the latter.

Then

$$\boxed{\mathbf{Q} = a + \left(\frac{0.5 - F_{low}}{F_{high} - F_{low}}\right)(b - a)} \quad \text{or} \quad \boxed{\mathbf{Q} = b - \left(\frac{F_{high} - 0.5}{F_{high} - F_{low}}\right)(b - a)}.$$

Again, in the grouped "Memorial Union age" data, we have $a = 20$, $b = 30$, $F_{low} = 0.2$, and $F_{high} = 0.6$ (why?). Substituting these values into either formula yields the median $Q_2 = 27.5$. ✓

To find Q_1, replace the 0.5 in the formula by 0.25; to find Q_3, replace the 0.5 in the formula by 0.75, etc.

Conversely, if a quantile $\mathbf{Q}$ *in an interval* $[a,\, b)$ *is* *<u>given</u>*, then we can solve for the cumulative relative frequency $F(\mathbf{Q})$ up to that quantile value:

$$\boxed{F(\mathbf{Q}) = F(a) + \left(\frac{F(b) - F(a)}{b - a}\right)(\mathbf{Q} - a)}.$$

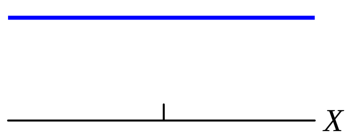
It follows that ***the relative frequency (i.e., area) <u>between</u> two quantiles $\mathbf{Q_1}$ and $\mathbf{Q_2}$ is equal to the <u>difference</u> between their cumulative relative frequencies: $F(\mathbf{Q_2}) - F(\mathbf{Q_1})$.***

Shapes of Distributions

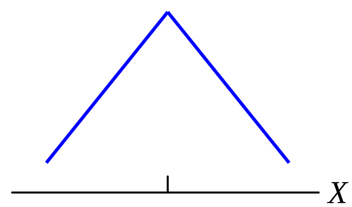
Symmetric distributions correspond to values that are spread equally about a "center."

mean = median

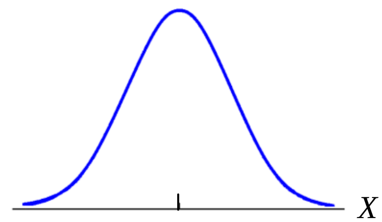
Examples: (Drawn for "smoothed histograms" of a random variable X.)



uniform



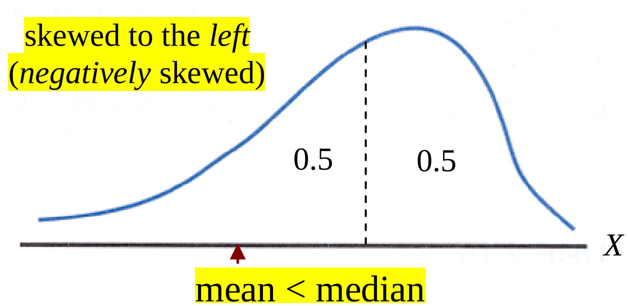
triangular

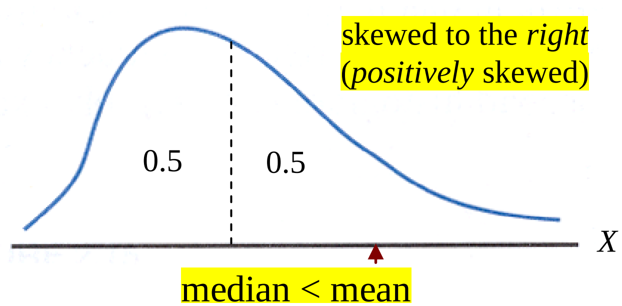


bell-shaped

Note: An important special case of the "bell-shaped" curve is the **normal distribution**, a.k.a. **Gaussian distribution**. Example: X = IQ score

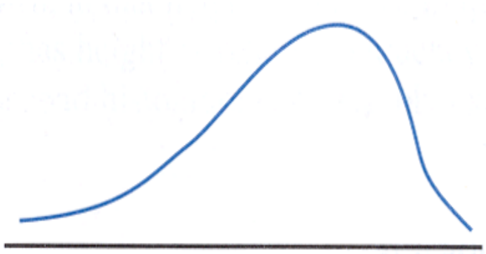
Otherwise, if more outliers of X occur on one side of the median than the other, the corresponding distribution will be **skewed** in that direction, forming a **tail**.



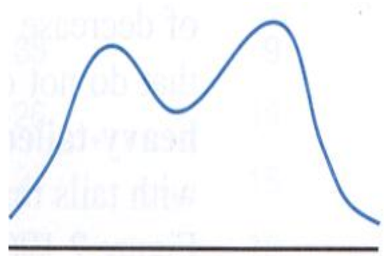


Examples: X = "calcium level (mg)" X = "serum cholesterol level (mg/dL)"

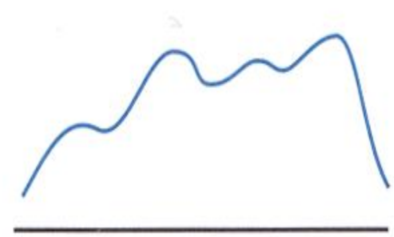
Furthermore, distributions can also be classified according to the number of "peaks":



unimodal



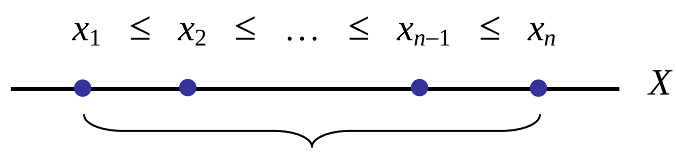
bimodal



multimodal

Measures of Spread

Again assume that a numerical random sample $\{x_1, x_2, \ldots, x_n\}$ has been selected, and *sorted* from lowest to highest values, i.e.,

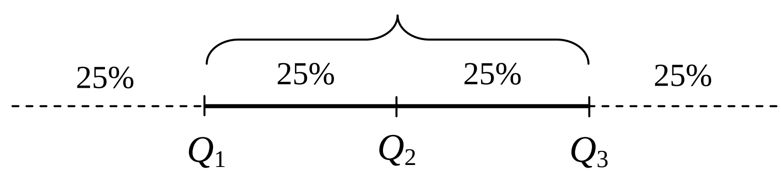


- **sample range** = $x_n - x_1$ (highest value – lowest value)

  Comments:

  - Uses only the two most extreme values. Very crude estimator of spread.
  - The sample range is extremely sensitive to outliers. One common remedy …

    Interquartile range (IQR) = $Q_3 - Q_1$. Robust to outliers by construction.



  - If the original data are grouped into k class intervals $[a_1, a_2), [a_2, a_3), \ldots, [a_k, a_{k+1})$, then the **group range** = $a_{k+1} - a_1$. A similar calculation holds for **group IQR**.

Example: The "Body Temperature" data set has a **sample range** = 99.2 – 98.5 = 0.7°F.

{98.5, 98.6, 98.6, 98.6, 98.6, 98.6, 98.9, 98.9, 98.9, 99.1, 99.1, 99.2}

x_i	f_i
98.5	**1**
98.6	**5**
98.9	**3**
99.1	**2**
99.2	**1**

$n = 12$

For a much less crude measure of spread that uses *all* the data, first consider the following…

Definition: $x_i - \bar{x}$ = **individual deviation** of the i^{th} sample data value from the sample mean

98.8

x_i	$x_i - \bar{x}$	f_i
98.5	**–0.3**	**1**
98.6	**–0.2**	**5**
98.9	**+0.1**	**3**
99.1	**+0.3**	**2**
99.2	**+0.4**	**1**

$n = 12$

Naively, an estimate of the spread of the data values might be calculated as the *average* of these $n = 12$ individual deviations from the mean. However, this will always yield zero!

FACT:

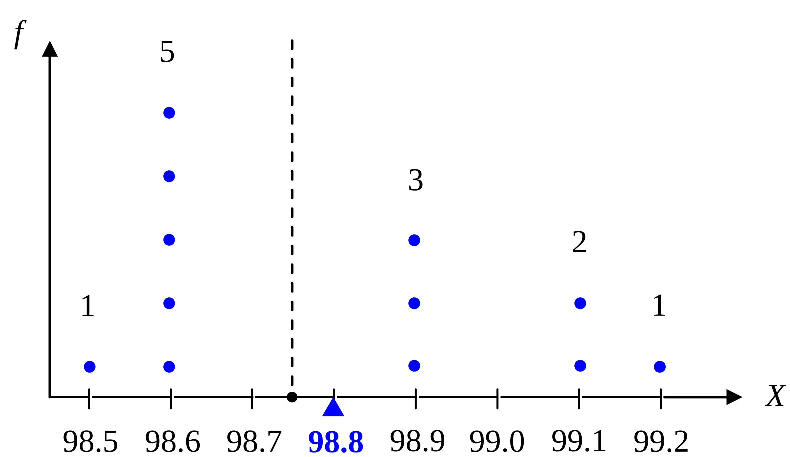
$$\sum_{i=1}^{k} (x_i - \bar{x})\, f_i = 0,$$

i.e., the sum of the deviations is always zero.

Check: In this example, the sum = (–0.3)(1) + (–0.2)(5) + (0.1)(3) + (0.3)(2) + (0.4)(1) = 0. ✓

Exercise: Prove this general fact algebraically.

Interpretation: The sample mean is the **center of mass**, or "balance point," of the data values.



Best remedy: To make them non-negative, *square* the deviations before summing.

- **sample variance**

$$\boxed{s^2 = \frac{1}{n-1}\sum_{i=1}^{k}(x_i - \bar{x})^2 f_i}$$

s^2 is <u>not</u> on the same scale as the data values!

- **sample standard deviation**

$$\boxed{s = +\sqrt{s^2}}$$

s is on the same scale as the data values.

<u>Example</u>:

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	f_i
98.5	–0.3	**+0.09**	**1**
98.6	–0.2	**+0.04**	**5**
98.9	+0.1	**+0.01**	**3**
99.1	+0.3	**+0.09**	**2**
99.2	+0.4	**+0.16**	**1**

$n = 12$

Then…

$$s^2 = \frac{1}{11}\left[\,(0.09)(1) + (0.04)(5) + (0.01)(3) + (0.09)(2) + (0.16)(1)\,\right] = 0.06\ (°\text{F})^2,$$

so that… $s = \sqrt{0.06} = 0.245°\text{F}$.

Body Temp has a small amount of variance.

<u>*Comments*</u>:

➢ $s^2 = \dfrac{\sum (x_i - \bar{x})^2 f_i}{n-1}$ has the important frequently-recurring form $\dfrac{\text{SS}}{\text{df}}$, where SS = "Sum of Squares" (sometimes also denoted S_{xx}) and df = "degrees of freedom" = $n - 1$, since the n individual deviations have a single constraint. (Namely, their sum must equal zero.)

➢ Same formulas are used for grouped data, with $\bar{x}_{\text{group}}$, and x_i = class interval midpoint.

Exercise: Compute s for the *grouped* and *ungrouped* Memorial Union age data.

➢ A related measure of spread is the **absolute deviation**, defined as $\frac{1}{n}\sum |x_i - \bar{x}|\, f_i$, but its statistical properties are not as well-behaved as the **standard deviation**. Also, see **<u>Appendix > Geometric Viewpoint > Mean and Variance</u>**, for a way to understand the "sum of squares" formula via the Pythagorean Theorem (!), *as well as a useful <u>alternate computational formula</u> for the **sample variance**.*

Typical "Grouped Data" Exam Problem

Age Intervals	Frequencies
[0, 18)	**-**
[18, 24)	**208**
[24, 30)	**156**
[30, 40)	**104**
[40, 60)	**52**
	520

Given the sample frequency table of age intervals shown above; answer the following.

1. Sketch the **density histogram**. (See Lecture Notes, page 2.2-6)
2. Sketch the graph of the **cumulative distribution**. (page 2.2-4)
3. What proportion of the sample is under 36 yrs old? (pages 2.3-5 bottom, 2.3-6 bottom)
4. What proportion of the sample is under 45 yrs old? (same)
5. What proportion of the sample is *between* 36 and 45 yrs old? (same)
6. Calculate the values of the following **grouped summary statistics**.

   Quartiles Q_1, Q_2, Q_3 and **IQR** (pages 2.3-4 to 2.3-6)

   Mean (page 2.3-4)

   Variance (page 2.3-10, second comment on bottom)

   Standard deviation (same)

Solutions at http://www.stat.wisc.edu/~ifischer/Grouped_Data_Sols.pdf