Dealing with ambiguity (e.g., how wide is a curve)
Dealing with precise syntax (e.g., words and their order matters)
Population (the truth)
Sample (our observed data)
Random Variable
Types of data (well-defined, but ambiguous)
Histogram (its interpretation and uses)
Probability from histogram
Distribution (its interpretation and uses)
Probability from distribution
Named distributions
The random variable "template" associated with each distribution
Standardization (for Normal and other distributions)
Percentile/quantile (for sample and/or dist)
Boxplots (uses and interpretation)
Derivation and application of Binomial and Poisson.

Time to quantify some of our qualitative ideas.

In prev. chapters we played with histograms of sample/data and distributions of (random) variables (cont. and discrete/categ.).
Hists and dists are pillars of statistics. Hists describe the data/sample, while dists describe the population. One question we often ask is This: how likely is it that my data/hist came from, say, a normal dist with params $\mu = \ldots$, $\sigma = \ldots$?

One way to compare histograms with distributions is in terms of their summary measures. For example, we can compare the "location" of the Normal distr. ($\mu$) with the "location" of the histogram.
Or the "width" of the former ($\sigma$) with the width of the histogram.

So, we need some measure of location and width for both histograms and distributions (for any dist, not just Normal).

<table>
<thead>
<tr>
<th>hist/sample</th>
<th>dist/pop</th>
</tr>
</thead>
<tbody>
<tr>
<td>measure of location $\mu$</td>
<td>$\mu$</td>
</tr>
<tr>
<td>measure of width $\sigma$</td>
<td>$\sigma$</td>
</tr>
</tbody>
</table>

The first comparison between sample and pop. will happen soon (when we do qq-plots), and then more fully in the 2nd half of 310.
Measures of location for hist/sample:

- Sample mean: \( \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \)

- Sample median: \( \tilde{x} \) = middle of the ordered data.

Measures of spread for hist/sample:

- Sample Range (same units as \( \bar{x} \))

- Sample Variance: \( s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \)

\( s \sim \text{"average" (typical) deviation.} \)

\[ \sum_{i=1}^{n} (x_i - \bar{x}) = \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \bar{x} = n \bar{x} - \bar{x} \sum_{i=1}^{n} 1 = 0. \]

**Example:** \( x = \{1, 3, 8\} \text{ cm} \)

\( \bar{x} = \frac{1}{3} (1 + 3 + 8) = 4 \text{ cm} \)

\( S^2 = \frac{1}{3-1} \left[ (1-4)^2 + (3-4)^2 + (8-4)^2 \right] = \frac{1}{2} (9 + 1 + 16) = 13 \text{ cm}^2 \)

Again: A lot of statistics is about explaining/understanding this variance. Recall, if there were no variance/change, we wouldn't say that we even have any data.
In short, we will use the following summary measures for 
locating and spread of data:

**Sample mean**: \( \bar{x} = \frac{1}{n} \sum x_i \)

**Sample variance**: \( s^2 = \frac{1}{n-1} \sum (x_i - \bar{x})^2 \)

Because \( \sum (x_i - \bar{x}) = 0 \)

“Funny Average”

Then \( s \) will be another measure of spread, and it's even better than \( s^2 \), because \( s \) has the same physical dimension as \( x \) itself. So, we can write things like \( \bar{x} \pm s \) as a way of summarizing a histogram.

Important: Interpretation of \( \bar{x} \) is **typical \( x \)**

""" **\( s \)**""" typical deviation of \( x \)

\( \text{vs } s^2 \)

In some problems where the \( \frac{1}{n-1} \) is not important, one focuses on \( s_{xx} = \sum (x_i - \bar{x})^2 \), i.e. just the numerator of \( s^2 \).

Finally, note that all of these measures have the word “Sample,” reminding you that they pertain to sample/data, not population/distribution.

For the mathematically-inclined: If you think of \( x_i \) as the components of an \( n \)-vector, then after you “center” each component (i.e. subtract \( \bar{x} \)), \( s \) is proportional to the magnitude of that vector.
Another way of computing $s^2$.

\[ S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \]

\begin{align*}
&= \frac{1}{n-1} \sum_{i=1}^{n} (x_i^2 - 2x_i \bar{x} + \bar{x}^2) \\
&= \frac{1}{n-1} \left[ \sum_{i=1}^{n} x_i^2 - 2 \frac{\sum x_i}{n} \bar{x} + n\bar{x}^2 \right] \\
&= \frac{1}{n-1} \left[ \sum_{i=1}^{n} x_i^2 - 2n(\bar{x})^2 + n(\bar{x})^2 \right] \\
&= \frac{1}{n-1} \left[ \sum_{i=1}^{n} x_i^2 - n(\bar{x})^2 \right] \\
&= \frac{1}{n-1} \left[ n(\frac{1}{n} \sum x_i^2) - n(\bar{x})^2 \right] = \frac{n}{n-1} \left[ (\bar{x}^2) - (\bar{x})^2 \right]
\end{align*}

In summary:

\[ S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \quad \text{"Defining formula"} \]

\[ S^2 = \frac{n}{n-1} \left[ \frac{\sum x_i^2}{n} - (\bar{x})^2 \right] \quad \text{"Computational formula"} \]

Sometimes more useful,

always faster (1 vs. 2 loops)

Not too important.

**Example**

\[ x = c(1, 3, 8) \rightarrow x^2 = c(1, 9, 64) \rightarrow \overline{x^2} = \frac{74}{3} \]

\[ S^2 = \frac{3}{2} \left[ \frac{74}{3} - 16 \right] = \frac{3}{2} \frac{74 - 48}{3} = \frac{26}{2} = 13 \quad \text{(same as above).} \]
Keep the "big picture" in mind: We are looking for

\[ \overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \]
\[ \overline{x} \sim \text{"typical } x\text{"} \]

**Sample mean**

**Sample variance \((s^2)\)**

\[ s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2 \]

**Sample Std. dev. \((s)\)**

\[ s = \frac{1}{n-1} (\overline{x^2} - \overline{x}) \]

\[ s \sim \text{"typical deviation in } x\text{"} \]

Now, we need to come up with corresponding things in the *pop.*

So, switch to distributions \((p(x), f(x))\), no data/sample!
1) Distribution mean (or Expected Value)

\[ \mu_X = E[X] = \sum_x x \cdot p(x) \]

\[ = \int x \cdot f(x) \, dx \]

Motivation: Even though we are now in the realm of math (p(x), f(x)), not data, just to motivate the definition of E(x), consider the following "population": \{3, 2, 2, 1, 3, 2, 3, 1, 2, 2\}.

\[ \text{mean} = \frac{1}{10} \left[ 3 \cdot 3 + 2 \cdot 2 + \cdots \right] \]
\[ = \frac{1}{10} \left[ 3(3) + 5(2) + 2(1) \right] \]
\[ = \frac{3}{10} \cdot 3 + \frac{5}{10} \cdot 2 + \frac{2}{10} \cdot 1 = \sum_x p(x) \cdot x \]

\[ \text{distinct values of } x \]

Compare:

Sample mean: \[ \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i , \quad \sim \text{typical } x \text{ (in dist.)/pop} \]

Distr. mean (Expected Value): \[ \mu_X = E[X] = \sum x \cdot p(x) \]
\[ = \int_{-\infty}^{\infty} x \cdot f(x) \, dx \sim \text{typical deviation in } x. \]

The book drops the \( x \) on \( \mu_X \), but then \( \mu \) can be confused with the parameter of the normal distribution.

Next time: mean (\( \mu_X = E[X] \)) of our named dists (below).

\( E[X] \) does not mean that \( E \) is a function of \( x \). In fact, \( E \) is \( \sum_x p(x) \) or \( \int_{-\infty}^{\infty} dx \), and so it is not a function of \( x \). \( E[X] \) simply means that you need \( p(x) \) or \( f(x) \) to find it. See examples next lecture; there is no \( x \)-dependence in \( \mu_X = E[X] \).
Consider the adjacent histogram.

a) Compute/find the sample mean $x_{\text{bar}}$. Show work.

b) Find the sample std. dev. $s$, using the defining formula.

c) Draw the mean and the sample std. dev. on the histogram.

For the exponential distribution with parameter $\lambda$, find the mean.

Hint: You may use this integral:

\[ \int_0^\infty ye^{-y} \, dy = 1. \]