Recall one of the main paradigms of statistical analysis is that

1. There is one population of interest, or there may be multiple populations of interest
   1.1. Widgets
   1.2. All Porsche’s made with a new bumper design
   1.3. All the paper airplanes made with the six treatments formed by combining the levels of the two factors:
      1.3.1. Design (dart, glider)
      1.3.2. Paper weight ( NB(light), printer(med), construction(heavy) )

2. There is a variable of interest—a random variable because its value will be different every time we observe it.
   2.1. X=Material strength of the widget
   2.2. X=yes or no: was there visible damage to the bumper in a 35 mph crash?
   2.3. X = distance the plane flew

3. We want to learn about the value of this variable for this population, but we cannot observe the population directly because
   3.1. It is too large (like the population of likely voters in the next Presidential election)—enumerative study.
   3.2. It is a theoretical or conceptual population—analytical study.

4. So we use a sample from the population to learn about the population.

5. Things we want to use the sample to do:
   5.1. Identify the probability distribution of the random variable for this population
   5.2. Estimate (with point or interval estimates) values of the parameters that give the probability distribution its shape
   5.3. Estimate the mean and variance of the population (called the first and second moments).
   5.4. Are there multiple populations? Or just one?
   5.5. Predict (with a point prediction or a prediction interval) the value of the random variable in the future:

6. Major point of sampling distributions chapter:
   6.1. Each sample is different, and will lead to slightly different conclusions.
   6.2. Sample statistics have probability distributions also, so sampling variability is predictable.
   6.3. We can use our knowledge of sampling distributions to quantify the uncertainty of our inference.
Example 5.19

Suppose we are concerned with the material strength of a widget, measured in mystery units (mu). Let’s say that the strength of a widget is the force required to squash the widget. “To squash” has been defined in a highly technical document. Thus the random variable of interest $X$ is defined as the strength of a widget. It is a random variable because every time we measure the strength of a widget, we will get a different number.

**Our goal:** We want to learn about the underlying true distribution of the strength of widgets. More specifically, we want to know the mean strength $\mu$, the median strength $\tilde{\mu}$, and the standard deviation $\sigma$.

**The underlying truth (that we are pretending we don’t know for the sake of creating an educational example):** It just so happens that widgets have a Weibull distribution. The pdf of the Weibull distribution is given by

$$f_X(x | \alpha, \beta) = \begin{cases} \frac{\alpha}{\beta^\alpha} x^{\alpha-1} e^{-(x/\beta)^\alpha}, & x \geq 0 \\ 0 & \text{otherwise} \end{cases}.$$  

For widgets, the parameter values are $\alpha = 2$ and $\beta = 5$, which so that the population mean is $\mu = 4.4311$ mu, the median is $\tilde{\mu} = 4.1628$ mu, and the variance is $\sigma^2 = 5.365$ mu$^2$. Therefore, here’s what the unknown population distribution looks like (this is Figure 5.6 in your textbook).

![Weibull distribution graph](image)

It just so happens that six different scientists at the same time decided to test the strength of widgets! Each scientist happened to choose a random sample of 10 widgets to squash, and then recorded the strength in mystery units.
Definition

**Random sample:** The random variables (RV’s) \( X_1, X_2, \ldots, X_n \) are said to form a random sample (RS) of size \( n \) if

1. The \( X_i \) are independent from one another.
2. The \( X_i \) are identically distributed (they have the same probability distribution, with the same values of the parameters)

Sometimes we write \( X_1, \ldots, X_n \overset{iid}{\sim} Weibull(\alpha, \beta) \) instead of saying that the \( X_i \) form a random sample.

Also, frequently it’s called a **simple random sample**. It depends on the book. I have mostly seen it called an SRS in undergraduate and high school texts.

<table>
<thead>
<tr>
<th></th>
<th>Scientist 1</th>
<th>Scientist 2</th>
<th>Scientist 3</th>
<th>Scientist 4</th>
<th>Scientist 5</th>
<th>Scientist 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6.1171</td>
<td>5.07611</td>
<td>3.4671</td>
<td>1.55601</td>
<td>3.12372</td>
<td>8.93795</td>
</tr>
<tr>
<td>2</td>
<td>4.61</td>
<td>6.79279</td>
<td>2.71938</td>
<td>4.56941</td>
<td>6.09685</td>
<td>3.92487</td>
</tr>
<tr>
<td>3</td>
<td>3.195</td>
<td>4.43259</td>
<td>5.88129</td>
<td>4.7987</td>
<td>3.41181</td>
<td>8.76202</td>
</tr>
<tr>
<td>4</td>
<td>0.6694</td>
<td>8.55752</td>
<td>5.14915</td>
<td>2.49759</td>
<td>1.65409</td>
<td>7.05569</td>
</tr>
<tr>
<td>5</td>
<td>1.8552</td>
<td>6.82487</td>
<td>4.99635</td>
<td>2.33267</td>
<td>2.29512</td>
<td>2.30932</td>
</tr>
<tr>
<td>6</td>
<td>5.2316</td>
<td>7.39958</td>
<td>5.86887</td>
<td>4.01295</td>
<td>2.12583</td>
<td>5.94195</td>
</tr>
<tr>
<td>7</td>
<td>2.7609</td>
<td>2.14755</td>
<td>6.05918</td>
<td>9.08845</td>
<td>3.20938</td>
<td>6.74166</td>
</tr>
<tr>
<td>8</td>
<td>10.2185</td>
<td>8.50628</td>
<td>1.80119</td>
<td>3.25728</td>
<td>3.23209</td>
<td>1.75468</td>
</tr>
<tr>
<td>9</td>
<td>5.2438</td>
<td>5.4951</td>
<td>4.21994</td>
<td>3.70132</td>
<td>6.84426</td>
<td>4.91827</td>
</tr>
<tr>
<td>10</td>
<td>4.559</td>
<td>4.04525</td>
<td>2.12934</td>
<td>5.50134</td>
<td>4.20694</td>
<td>7.26081</td>
</tr>
</tbody>
</table>
The image shows a screenshot of a software interface for analyzing distributions. Each widget displays quantile and moment data:

**Widget 1**
- Quantiles: 100.0% maximum = 10.2185, 99.5% = 10.2185, 97.5% = 10.2185, 95.0% = 8.08305, 75.0% = 5.46213, 50.0% = 4.3595, 25.0% = 2.53440, 10.0% = 1.39799, 5.0% = 0.6694, 0.0% = 0.6694
- Moments: Mean = 4.40135, StdDev = 2.642098, StdErrMean = 0.8320474, Upper 95% Mean = 6.2910237, Lower 95% Mean = 2.5110713, N = 10

**Widget 2**
- Quantiles: 100.0% maximum = 8.9575, 99.5% = 8.9575, 97.5% = 8.9575, 95.0% = 8.5633, 75.0% = 7.9726, 50.0% = 6.6453, 25.0% = 4.3157, 10.0% = 2.3373, 5.0% = 1.4427, 0.0% = 1.4427
- Moments: Mean = 5.927744, StdDev = 2.6519724, StdErrMean = 0.6320933, Upper 95% Mean = 7.4192189, Lower 95% Mean = 4.4571711, N = 10

**Widget 3**
- Quantiles: 100.0% maximum = 5.05918, 99.5% = 5.05918, 97.5% = 5.05918, 95.0% = 5.04139, 75.0% = 4.05815, 50.0% = 4.05815, 25.0% = 2.57187, 10.0% = 2.01819, 5.0% = 1.8119, 0.0% = 1.8119
- Moments: Mean = 4.229119, StdDev = 2.6519724, StdErrMean = 0.585932, Upper 95% Mean = 5.381635, Lower 95% Mean = 3.705228, N = 10

**Widget 4**
- Quantiles: 100.0% maximum = 3.99485, 99.5% = 3.99485, 97.5% = 3.99485, 95.0% = 3.5935, 75.0% = 3.07714, 50.0% = 2.45004, 25.0% = 1.83306
- Moments: Mean = 4.131573, StdDev = 2.1239899, StdErrMean = 0.671534, Upper 95% Mean = 5.050997, Lower 95% Mean = 2.612447, N = 10
Definition

Statistic:
A number describing the sample.
A function of a sample.
Any quantity that can be calculated from the data.

Important!!
Since the sample data consists of random variables, statistics are random variables! There is still the distinction between the random variable and a specific value of it.

What was the point of that demonstration? What did I want you to learn?

1. In different random samples, the observed (or realized) values \(x_1, x_2, \ldots, x_n\) of random variables \(X_1, X_2, \ldots X_n\) will be different. (Duh. Remember that’s why we call them random variables.)
2. Therefore, the values $\bar{x}, \tilde{x}, s$, and $s^2$ of sample statistics $\bar{X}, \tilde{X}, S$ and $S^2$ will be different.
3. The sample statistics are also random variables, because they are functions of random variables.
4. If we use these sample statistics to estimate population parameters $\mu, \tilde{\mu}$, and $\sigma$ or $S\sigma$, then different samples will yield different estimates.
5. The estimates are almost always wrong, and they will be wrong by different amounts and/or wrong in different directions for different samples.
6. The probability distributions of the statistics $\bar{X}, \tilde{X}, S$ and $S^2$ are called the sampling distributions of the statistics.

Why do we care about sampling distributions of statistics?

7. If we know how widely varied these statistics can be, we can quantify the degree of our wrongness! We (statisticians) call it quantifying uncertainty.

How do we find sampling distributions of statistics?

1. If we know the the population distribution (or if we know properties of it), we can derive the sampling distribution using rules of probability and mathematics! (Sometimes this is really easy. Sometimes it’s really hard, but some really smart person already did it and came up with a theorem that we use.)
2. If doing number one is too difficult or impossible mathematically, we can use simulations.

Example 5.20—Example of deriving the sampling distribution

Problem of interest: using the sample mean to estimate the population mean

A certain brand of MP3 player comes in three configurations: a model with 2GB memory, costing $80, a 4GB model costing $100, and an 8GB model costing $120. Suppose 20% of all shoppers choose the 2GB model, 30% the 4GB model, and 50% the 8GB model. Consider the random variable $X =$ amount spent by a randomly selected shopper.

1. Write the PMF of $X$ in tabular form.
2. Find $EX$, the expected value of $X$, which is the same as the population mean.
3. Find $Var(X)$, the variance of $X$.
4. Why would $EX$ be of interest to the merchant?
More real life:
Now pretend we don’t know

1. The PMF of \( X \)
2. \( \mu = EX \)
3. \( Var(X) \)

We use the mean from the purchases of two shoppers to estimate \( \mu \), the population mean, the long term average income from sales of these MP3 players. Let \( X_1 \) and \( X_2 \) represent the price paid by each shopper. Assume \( X_1, X_2 \sim f_X(x) \). How reliable is the sample mean \( \overline{X}_2 = \frac{1}{2}(X_1 + X_2) \) for estimating the population mean \( \mu \)? What if we used the mean of four shoppers: \( \overline{X}_4 = \frac{1}{4}(X_1 + X_2) \)? Let’s compare the distributions of \( X, \overline{X}_2, \overline{X}_4 \).

To do this, we have to again pretend we do know the probability distribution of \( X \).