NOTES FOR SUMMER STATISTICS INSTITUTE COURSE

COMMON MISTAKES IN STATISTICS – SPOTTING THEM AND AVOIDING THEM

Day 2: Important Details about Statistical Inference

"The devil is in the details" (anonymous)

MAY 23-26, 2016

Instructor: Martha K. Smith

CONTENTS OF DAY 2

I. MORE PRECISE DEFINITION OF SIMPLE RANDOM SAMPLE

In practice in applying statistical techniques, we're interested in *random variables* defined on the population under study.

Recall the examples mentioned yesterday:

1. In a medical study, the population might be all adults over age 50 who have high blood pressure.
2. In another study, the population might be all hospitals in the U.S. that perform heart bypass surgery.
3. If we're studying whether a certain die is fair or weighted, the population is all possible tosses of the die.

In these examples, we might be interested in the following random variables:

Example 1: The difference in blood pressure with and without taking a certain drug.

Example 2: The number of heart bypass surgeries performed in a particular year, or the number of such surgeries that are successful, or the number in which the patient has complications of surgery, etc.

Example 3: The number that comes up on the die.

Connection with Independent Random Variables:

If we take a *sample of units from the population*, we have a corresponding *sample of values of the random variable*.

In Example 1:

- The *random variable* is "difference in blood pressure with and without taking the drug."
    - Call this random variable Y (upper case Y)
- The sample of *units from the population* is a sample of adults over age 50 who have high blood pressure.
    - Call them person 1, person 2, etc.
- The corresponding sample of values of the random variable will consist of values we will call $y_1, y_2, ..., y_n$ (lower case y's), where
    - n = number of people in the sample;
    - y_1 = the difference in blood pressures (that is, the value of Y) for the first person in the sample;
    - y_2 = the difference in blood pressures (that is, the value of Y) for the second person in the sample;
    - etc.

We can look at this another way, in terms of n random variables $Y_1, Y_2, ..., Y_n$, described as follows:

- The random process for Y_1 is "pick the first person in the sample"; the value of Y_1 is the value of Y for that person – i.e., y_1.

- The random process for Y_2 is "pick the second person in the sample"; the value of Y_2 is the value of Y for that person – i.e., y_2.

- etc.

The difference between using the small y's and the large Y's is that *when we use the small y's we are thinking of a fixed sample of size n from the population, but when we use the large Y's, we are thinking of letting the sample vary (but always with size n).*

Note: The Y_i's are sometimes called *identically distributed*, because they have the same probability distribution (in this example, the distribution of Y).

Precise definition of simple random sample of a random variable:

> "*The sample $y_1, y_2, ... , y_n$ is a simple random sample*" means that the associated random variables $Y_1, Y_2, ... , Y_n$ are independent.

Intuitively speaking, "*independent*" means that *the values of any subset of the random variables $Y_1, Y_2, ... , Y_n$ do not influence the probabilities of the values of the other random variables in the list*.

Recall: We defined a random sample as one that is chosen by a random process.

- Where is the random process in the precise definition?

Note: To emphasize that the Y_i's all have the same distribution, the precise definition is sometimes stated as, "$Y_1, Y_2, ... , Y_n$ are independent, identically distributed," sometimes abbreviated as *iid*.

Connection with the initial definition of simple random sample

Recall the preliminary definition (from Moore and McCabe, *Introduction to the Practice of Statistics*) given in Simple Random Samples, Part 1:

> "*A simple random sample (SRS) of size n consists of n individuals from the population chosen in such a way that every set of n individuals has an equal chance to be the sample actually selected."*

Recall Example 3 above: We toss a die; the number that comes up on the die is the value of our random variable Y.

- In terms of the preliminary definition:
  - The population is all possible tosses of the die.
  - A simple random sample is n different tosses.
- The different tosses of the die are *independent events* (i.e., what happens in some tosses has no influence on the other tosses), which means that in the precise definition above, the random variables $Y_1, Y_2, \ldots, Y_n$ are indeed independent: The numbers that come up in some tosses in no way influence the numbers that come up in other tosses.

Compare this with example 2: The population is all hospitals in the U.S. that perform heart bypass surgery.

- Using the preliminary definition of simple random sample of size n, we end up with n *distinct* hospitals.
- This means that when we have chosen the first hospital in our simple random sample, *we cannot choose it again to be in our simple random sample*.
- Thus the events "Choose the first hospital in the sample; choose the second hospital in the sample; ... ," are *not* independent events: The choice of first hospital restricts the choice of the second and subsequent hospitals in the sample.
- If we now consider the random variable Y = the number of heart bypass surgeries performed in 2008, then it follows that the random variables $Y_1, Y_2, \ldots, Y_n$ are *not* independent.

The Bottom Line: *In many cases, the preliminary definition does* ***not*** *coincide with the more precise definition.*

More specifically, the preliminary definition allows sampling *without* replacement, whereas the more precise definition requires sampling *with* replacement.

The Bad News: *The precise definition is the one used in the mathematical theorems that justify many of the procedures of statistical inference.* (More detail later.)

The Good News:

1. *If the population is large enough, the preliminary definition is close enough for all practical purposes.*
2. *In many cases where the population is not "large enough," there are alternate theorems giving rise to alternate procedures using a "finite population correction factor" that will work.*
   - Unfortunately, the question, "How large is large enough?" does not have a one-size-fits-all answer; in particular, it depends on the procedure.
   - However, the answer is relative to sample size –we only need to worry for samples that are large relative to population size
3. *In many cases, even if the population is not large enough, there are alternate procedures (known as permutation or randomization or resampling tests) that are applicable.*

Consequent Problems with Small Populations:

1) Using a "large population" procedure with a "small population" is a **common mistake.**

2) One more difficulty in selecting an appropriate sample, which leads to one more source of uncertainty.

II. WHY RANDOM SAMPLING IS IMPORTANT

Recall the Myth:

"A random sample will be representative of the population".

A slightly better explanation (partly true but partly Urban Legend):

"Random sampling prevents bias by giving all individuals an equal chance to be chosen."

- The element of truth: Random sampling *does* eliminate *systematic* bias.
- A practical rationale: This explanation is often the best plausible explanation that is acceptable to someone with little mathematical background.
- However, this statement could easily be misinterpreted as the myth above.

An additional, very important, reason why random sampling is important, at least in frequentist statistical procedures, which are those most often taught (especially in introductory classes) and used:

The Real Reason: *The mathematical theorems that justify most parametric frequentist statistical procedures apply only to truly (suitably) random samples*.

The next section elaborates.

III. OVERVIEW OF FREQUENTIST HYPOTHESIS TESTING

Type of Situation where a Hypothesis Test is used:

- We *suspect* a certain pattern in a certain situation.
- But we realize that natural variability or imperfect measurement might produce an apparent pattern that isn't really there.

Basic Elements of Most Frequentist Hypothesis Tests:
Most commonly-used ("parametric"), frequentist hypothesis tests involve the following four elements:

i. *Model assumptions*

ii. *Null and alternative hypotheses*

iii. A *test statistic*

This is something that
   a. Is *calculated by a rule from a sample*;
   b. Is a measure of the strength of the pattern we are studying; and
   c. Has the property that, *if* the null hypothesis is true, *extreme values of the test statistic are rare, and hence cast doubt on the null hypothesis*.

iv. A *mathematical theorem* saying,

"If the model assumptions and the null hypothesis are both true, then the *sampling distribution* of the test statistic has a certain particular form."

Note:

- The *sampling distribution* is the probability distribution of the test statistic, when considering *all* possible suitably random samples of the same size. (More later.)
- The exact details of these four elements will depend on the particular hypothesis test.
- In particular, the form of the sampling distribution will depend on the hypothesis test.

Illustration: Large Sample z-Test for the mean, with two-sided alternative

The above elements for this test are:

1. *Model assumptions*: We are working with simple random samples of a random variable Y that has a normal distribution with known standard deviation.

2. *Null hypothesis*: "The mean of the random variable Y is a certain value μ_0."

Alternative hypothesis: "The mean of the random variable Y is not μ_0." (This is called the *two-sided alternative*.)

3. *Test statistic*: $\bar{y}$ (the *sample mean* of a simple random sample of size n from the random variable Y).

Before discussing item 4 (the mathematical theorem), we first need to:

1. Clarify terminology
2. Discuss sampling distributions

1. **Terminology and common confusions:**

- The ***mean of the random variable*** Y is also called the *expected value* or the *expectation* of Y.
  - It's denoted E(Y).
  - It's also called the *population mean*, often denoted as μ.
  - It's what we do *not* know in this example.

- A ***sample mean*** is typically denoted $\bar{y}$ (read "y-bar").
  - It's calculated from a sample $y_1, y_2, \ldots, y_n$ of values of Y by the familiar formula $\bar{y} = (y_1 + y_2 + \ldots + y_n)/n$.

- The sample mean $\bar{y}$ is an *estimate* of the population mean μ, but they are usually *not* the same.
  - Confusing them is a **common mistake**.

- Note the articles: "*the* population mean" but "*a* sample mean".
  - There is only one *population* mean associated with the random variable Y.
  - However, a *sample* mean depends on the sample chosen.
  - Since there are many possible samples, there are many possible sample means.

Illustration:

http://wise.cgu.edu/sdmmod/sdm_applet.asp

or

https://istats.shinyapps.io/sampdist_cont/

Putting this in a more general framework:

- A *parameter* is a constant associated with a population.
  - In our example: The population mean μ is a parameter.
  - When applying statistics, parameters are usually unknown.
  - However, the goal is often to gain some information about parameters.
- To help gain information about (unknown) parameters, we use *estimates* that are calculated from a sample.
  - In our example: We calculate the sample mean $\bar{y}$ as an estimate of the population mean (parameter) μ.

"Variance" is another common example where parameter and estimate might be confused:

- The *population variance* (or "the variance of Y") is a parameter, usually called σ^2 (or Var(Y)).
- If we have a sample from Y, we can calculate the *sample variance*, usually called s^2.
- A sample variance is an estimate of the population variance.
- Different samples may give different estimates.
- *Confusing population and sample variance is a* ***common mistake.***

2. **Sampling Distribution:**

Although we apply a hypothesis test using a single sample, we need to step back and consider *all* possible suitably random samples of Y of size n, in order to understand the test. In our example:

- For each simple random sample of Y of size n, we get a value of $\bar{y}$.
- We thus have a *new* random variable $\overline{Y}_n$:
  - The associated random process is "pick a simple random sample of size n"
  - The *value* of $\overline{Y}_n$ is the *sample mean* $\bar{y}$ for this sample
- Note that
  - $\overline{Y}_n$ stands for the new random variable
  - $\bar{y}$ stands for the value of $\overline{Y}_n$, for a particular sample of size n.
  - $\bar{y}$ (the value of $\overline{Y}_n$) depends on the sample, and typically varies from sample to sample.
- The distribution of the new random variable $\overline{Y}_n$ is called the *sampling distribution of* $\overline{Y}_n$ (or the *sampling distribution of the mean*).
- Note: $\overline{Y}_n$ is an example of an *estimator*: a random variable whose values are estimates.

Now we can state the theorem that the large sample z-test for the mean relies on:

4. The *theorem* states: *If* the model assumptions are all true (i.e., *if* Y is normal *and* all samples considered are simple random samples), and *if in addition* the mean of Y is indeed μ_0 (i.e., *if* the null hypothesis is true), then

   i. The sampling distribution of $\overline{Y}_n$ is normal

   ii. The sampling distribution of $\overline{Y}_n$ has mean μ_0

   iii. The sampling distribution of $\overline{Y}_n$ has standard deviation $\sigma/\sqrt{n}$, where σ is the standard deviation of the original random variable Y.

Check that this is consistent with what the simulation shows. (Try sample size n = 25)

Also note:

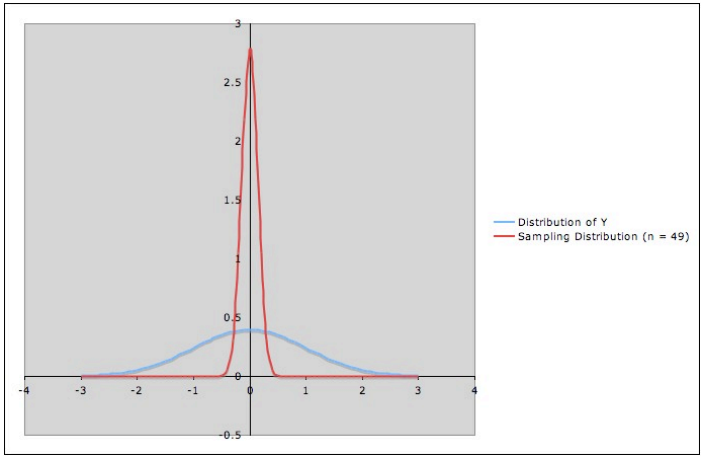
- $\sigma/\sqrt{n}$ is smaller than σ (if n is larger than 1)
- The larger n is, the smaller $\sigma/\sqrt{n}$ is.
- ***Why is this nice?***

More Terminology: σ is called the *population standard deviation* of Y; it is *not* the same as the *sample standard deviation* s, although s is an estimate of σ.

The following chart and picture summarize the conclusion of the theorem and related information:

	Considering the population ↓	Considering *one* sample ↓	Considering *all* suitable samples ↓
	Random variable Y (population distribution)	Related quantity calculated from a sample $y_1, y_2, \ldots, y_n$	**Random variable $\overline{Y}_n$ (sampling distribution)**
Type of Distribution	Y *has* a normal distribution	The sample is *from* the (normal) distribution of Y	$\overline{Y}_n$ *has* a normal distribution
Mean	*Population mean* μ ($\mu = \mu_0$ if null hypothesis true)	*Sample mean* $\overline{y} = (y_1 + y_2 + \ldots y_n)/n$ $\overline{y}$ is an estimate of the population mean μ	*Mean of the sampling distribution of* $\overline{y}$ -- it's also μ. ($\mu = \mu_0$ if null hypothesis true)
Standard deviation	*Population standard deviation* σ	*Sample standard deviation* $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(\overline{y} - y_i)^2}$ s is an estimate of the population standard deviation σ	*Sampling distribution standard deviation* $\sigma/\sqrt{n}$

↑ From the Theorem



- Which column of the chart corresponds to the blue distribution?
- Which column of the chart corresponds to the red distribution?
- How could you tell without looking at the legend?

The roles of the model assumptions for this hypothesis test (large sample z-test for the mean)**:**

Recall:

The theorem has *three assumptions*:

Assumption 1: Y has a normal distribution (*a model assumption).*

Assumption 2: All samples considered are simple random samples (*also a model assumption).*

Assumption 3: The null hypothesis is true (*assumption for the theorem, but not a model assumption).*

The theorem also has *three conclusions*:

Conclusion 1: The sampling distribution of $\overline{Y}_n$ is normal

Conclusion 2: The sampling distribution of $\overline{Y}_n$ has mean μ_0

Conclusion 3: The sampling distribution of $\overline{Y}_n$ has standard deviation $\sigma/\sqrt{n}$, where σ is the standard deviation of the original random variable Y.

The following chart shows *which assumptions each conclusion depends upon*:

	Assumptions of Theorem ↓		
Conclusions ***about Sampling Distribution*** (Distribution of $\overline{Y}_n$) ↓	1. Y normal (model assumption)	2: Simple random samples (model assumption)	3: Null hypothesis true (not a model assumption)
1: Normal	√	√	
2: Mean μ_0			√
3: Standard deviation $\sigma/\sqrt{n}$		√	

↑
Corresponds to third column in table on p. 18

Note that the model assumption that the sample is a simple random sample (in particular, that the Y_i's as defined earlier are independent) *is used to prove:*

1. that the sampling distribution is normal *and*

2. (even more importantly) that the standard deviation of the sampling distribution is $\sigma/\sqrt{n}$.

<u>This illustrates a general phenomenon that *independence conditions are usually very important in statistical inference.*</u>

Consequences (More detail later):

1. *If* the conclusion of the theorem is true, the sampling distribution of $\overline{Y}_n$ is narrower than the original distribution of Y

- In fact, conclusion 3 of the theorem gives us an idea of just how narrow it is, depending on n.

- *This will allow us to construct a useful hypothesis test*.

2. The only way we know the conclusion is true is if we know the hypotheses of the theorem (the model assumptions and the null hypothesis) are true.

3. Thus: *If the model assumptions are <u>not</u> true, then we <u>do not know</u> that the theorem is true, so we <u>do not know</u> that the hypothesis test is valid.*

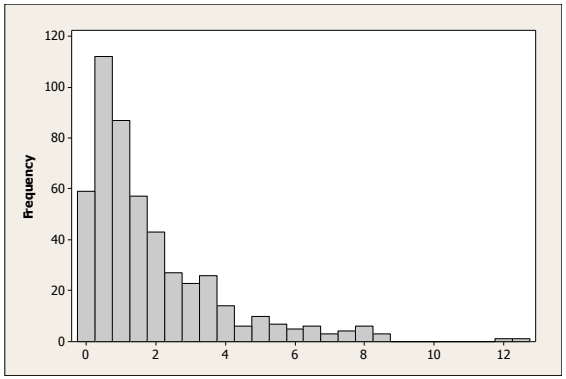
In the example (large sample z-test for a mean), this translates to:

If the sample is not a simple random sample, <u>or</u> if the random variable is not normal, then the reasoning establishing the validity of the test breaks down.

QUIZ: Would this test be reasonable to use in the following situations? Why or why not?

1. The histogram of values of Y for the sample is

a.

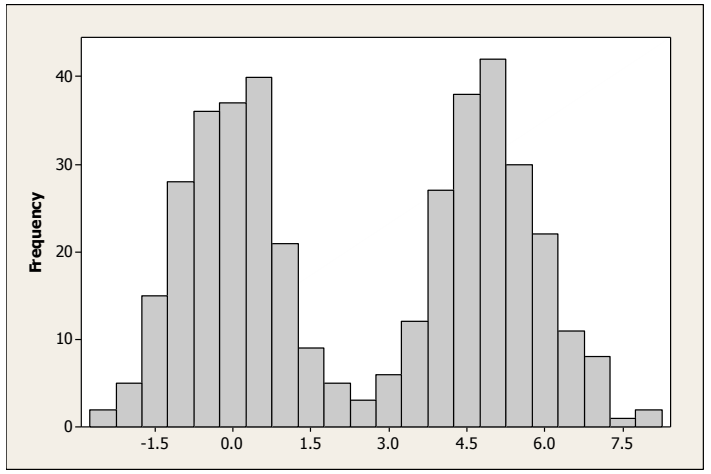


s

Reasonable to use test? _____

Why or why not? _____

b.



Reasonable to use test? _____

Why or why not? _____

2. Y is a random variable that can only take on values between 0 and 1.

Reasonable to use test? _____

Why or why not? _____

3. The sample consists of people who responded to a request on a website.

Reasonable to use test? _____

Why or why not? _____

Importance of Model Assumptions in General:

Different hypothesis tests have different model assumptions.

- Some tests apply to random samples that are not simple.
- For many tests, the model assumptions consist of *several* assumptions.
- If *any one* of these model assumptions is not true, we *do not know* that the test is valid. (*The bad news)*

Robustness:

Many techniques are ***robust*** *to some departures from at least some model assumptions*. (*The good news)*

- "Robustness" means that if the particular assumption is *not too far* from true, then the technique is still approximately valid.
- For example, the large sample z-test for the mean is *somewhat* robust to departures from normality. In particular, for *large enough* sample sizes, the test is very close to accurate.
    - Unfortunately, how large is large enough depends on the distribution of the random variable Y. *(More bad news)*
- Robustness depends on the particular procedure; there are no "one size fits all" rules.
    - Unfortunately, I have not been able to find easily available compilations of robustness results for a variety of procedures.
    - Unfortunately, some literature on robustness is incorrect.

Caution re terminology: "Robust" is used in other ways – for example, "a finding is robust" could be used to say that the finding appears to be true in a wide variety of situations, or that it has been established in several ways.

A **very common mistake** in using statistics*: Using a hypothesis test without paying attention to whether or not the model assumptions are true and whether or not the technique is robust to possible departures from model assumptions is a.*

- ***Unfortunately, the process of peer review often does not catch these and other statistical mistakes***.
- This is one of many reasons why the "results" published in peer-reviewed journals are often false or not well substantiated.
- Compounding the problem, journals rarely retract articles with mistakes. For examples and discussion, see
  - Allison et al, A Tragedy of Errors, Nature 530, 4 February 2016, 28 – 29, http://www.nature.com/news/reproducibility-a-tragedy-of-errors-1.19264
  - Kamoun, S. and C. Zipfel, Scientific record: Class uncorrected errors as misconduct, Nature 531, 10 March 2016, 173, http://www.nature.com/nature/journal/v531/n7593/full/531173e.html
  - Gelman and commenters, http://andrewgelman.com/2016/02/22/its-too-hard-to-publish-criticisms-and-obtain-data-for-replication/
- Resources for finding (at least some) retractions and other reports of errors include
  - PubPeer (https://pubpeer.com/)
  - Retraction Watch (http://retractionwatch.com/)
  - Authors sometimes post errors or corrections on their own website.
  - bioRxiv (http://biorxiv.org/) is a biology preprint server that also allows moderated public comments on posted papers. It classifies replication studies as Confirmatory or Contradictory.

IV: FREQUENTIST CONFIDENCE INTERVALS

Before continuing the discussion of hypothesis tests, it will be helpful first to discuss the related concept of confidence intervals.

The General Situation:

- We're considering a *random variable* Y.
- We're interested in a certain *parameter* (e.g., a proportion, or mean, or regression coefficient, or variance) associated with the random variable Y (i.e., associated with the population)
- We *don't* know the value of the parameter.
- *Goal 1*: We'd like to estimate the unknown parameter, using data from a sample. (A ***point estimate***)
- *But since estimates are always uncertain, we also need:*
- *Goal 2*: We'd like to get some sense of how good our estimate is. (Typically achieved by an ***interval estimate***)

The first goal is usually easier than the second.

Example: If the parameter we're interested in estimating is the *mean of the random variable* (i.e., the population mean, which we call ____), we can estimate it using a *sample mean* (which we call ____).

The rough idea for achieving the second goal (getting some sense of how good our estimate is):

> We'd like to get a *range of plausible values for the unknown parameter*. ("What we want.") We will call this range of plausible values a *confidence interval*.

The usual method for calculating a confidence interval has a lot in common with hypothesis testing:

- It involves the sampling distribution
- It depends on model assumptions.

A little more specifically:

- Although we typically have just one sample at hand when we do statistics, *the reasoning used in classical frequentist inference depends on thinking about all possible suitable samples of the same size n*.
- Which samples are considered "suitable" will depend on the particular statistical procedure to be used.
- Each confidence interval procedure has *model assumptions* that are needed to ensure that the reasoning behind the procedure is sound.
- The model assumptions determine (among other things) which samples are "suitable."
- The procedure is applicable only to "suitable" samples.

Illustration: Large Sample z-Procedure for a Confidence Interval for a Mean

- The parameter we want to estimate is the population mean $\mu = E(Y)$ of the random variable Y.
- The model assumptions for this procedure are: The random variable is normal, and samples are simple random samples.
    - So in this case, "suitable sample" means "simple random sample".
    - For this procedure, we also need to know that Y is normal, so that both model assumptions are satisfied.

Notation and terminology:

- We'll use σ to denote the (population) standard deviation of Y.
- We have a simple random sample, say of size n, consisting of observations $y_1, y_2, \ldots, y_n$.
    - For example, if Y is "height of an adult American male," we take a simple random sample of n adult American males; $y_1, y_2, \ldots, y_n$ are their heights.
- We use the sample mean $\bar{y} = (y_1 + y_2 + \ldots + y_n)/n$ as our estimate of the population mean μ.
    - This is an example of a ***point estimate*** -- a numerical estimate with no indication of how good the estimate is.

More details:

- To get an idea of how good our estimate is, we use the concept of *confidence interval*.
  - This is an example of an ***interval estimate***.
- To understand the concept of "confidence interval" in detail, we need to consider *all possible simple random samples of size n from Y*.
  - In the specific example, we consider all possible simple random samples of n adult American males.
  - For each such sample, the heights of the men in the sample of people constitute our simple random sample of size n from Y.
- We consider the *sample means* $\bar{y}$ *for all possible simple random samples of size n from Y*.
  - This amounts to defining a new random variable, which we will call $\overline{Y}_n$ (read "Y-bar sub n", or "Y-sub-n- bar").
  - We can describe the random variable $\overline{Y}_n$ briefly as "sample mean of a simple random sample of size n from Y", or more explicitly as: "pick a simple random sample of size n from Y and calculate its sample mean".
  - Note *that each value of* $\overline{Y}_n$ is an estimate of the population mean μ.
    - e.g. each simple random sample of n adult American males gives us an estimate of the population mean μ.

- This new random variable $\overline{Y}_n$ has a distribution, called a *sampling distribution* (since it arises from considering varying samples).
  - The values of $\overline{Y}_n$ are all the possible values of sample means $\bar{y}$ of simple random samples of size n of Y – i.e, the values of our *estimates* of μ.
  - *The sampling distribution (distribution of* $\overline{Y}_n$*) gives us information about the variability (as samples vary) of our estimates* of the population mean μ.
  - A mathematical theorem tells us that *if* the model assumptions are true, then:

    1. The sampling distribution is normal
    2. The mean of the sampling distribution is also μ.
    3. The sampling distribution has standard deviation $\sigma/\sqrt{n}$

  - Use these conclusions to compare and contrast the shapes of the distribution of Y and the distribution of $\overline{Y}_n$
    - What is the same? _____
    - What is different? _____
    - How do the standard deviations compare? _____
  - The chart (best read one column at a time) and picture below summarize some of this information.