Activity #14: Sampling distributions and the Central Limit Theorem

So far, this unit has focused on distributions of discrete and continuous random variables.
In this activity, we'll investigate **sampling distributions** – distributions of statistics.

Scenario: We want to know the average age of all cloned sheep that exist right now.
We don't know how many cloned sheep exist, but we are able to get samples of sheep delivered to us.

Unbeknownst to us, the entire population consists of 5 cloned sheep with ages 10, 11, 12, 13, 14 months.

1. Using R, I input the ages of the sheep with the code: `sheep <- c(10, 11, 12, 13, 14)`.
I then calculated population parameters that we do not know in this scenario: **μ = 12** and **σ = 1.414**.

To estimate μ, the unknown average age of all cloned sheep, we decide to do the following:

- Take a sample of n sheep from the population
- Calculate the average from each sample
- Repeat this process many times and sketch a distribution of the averages we calculate from our samples

Suppose we go through this process with a sample size of n=1 sheep. We first sample one sheep and then calculate it's "average" age. We then take another sample (possibly getting the same sheep) and calculate an average.

a) What possible averages could we get? Sketch a dotplot of those averages: ____________ (n=1)

10 11 12 13 14

b) Suppose we sample n=1 sheep 25,000 times. What would the distribution of all those sample means look like?

c) The mean of all 25,000 sample means should be: ____________

d) Suppose we decide to sample n = 2 sheep.
How many different averages could we get from samples of two sheep? ____________

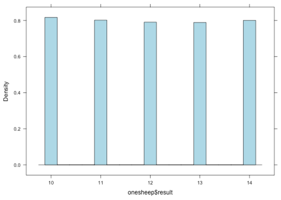
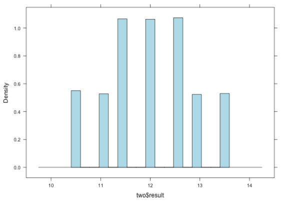
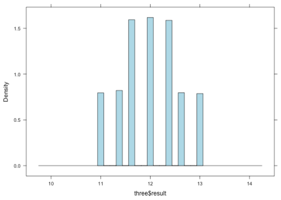
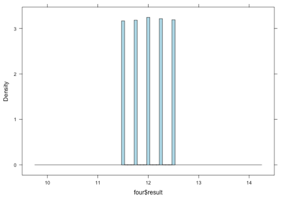
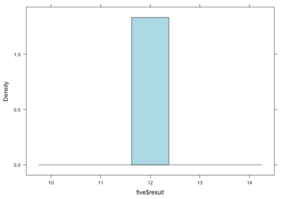
e) Sketch a dotplot of all possible averages we could get from n=2 sheep: ____________ (n=2)

10 11 12 13 14

f) Suppose we sample n=2 sheep 25,000 times. What would the distribution of all those sample means look like?

g) The mean of all 25,000 sample means should be: ____________

2. I simulated this process 25,000 times for sample sizes of n=1, 2, 3, 4, and 5 sheep. Fill-in-the-blanks:

| Sample Size | Distribution of 25,000 sample means | Mean of sample means $\mu_{\bar{X}}$ | Std. Deviation of sample means $\sigma_{\bar{X}}$ | Probability of a usual event: $P\left(11 \le \bar{X} \le 13\right)$ | Probability of an unusual event: $P\left(\bar{X} \le 11.5\right)$ |
|---|---|---|---|---|---|
| n = 1 |  | 11.989 | 1.4197 | __________ | __________ |
| n = 2 |  | 11.994 | 0.8678 | __________ | __________ |
| n = 3 |  | 12.001 | 0.5774 | __________ | __________ |
| n = 4 |  | 12.001 | 0.3526 | __________ | __________ |
| n = 5 |  | 12.000 | 0.000 | __________ | __________ |

3. From these simulations, let's generalize. If we repeatedly take samples of size *n* from a population and calculate the mean of each sample:

a) The expected value of our sample means (i.e., the mean of our means) = ______________________________

b) The standard deviation of our sample means is called the **standard error**.
If we take a larger sample, the size of our standard error……………………..… DECREASES INCREASES

c) If we take a larger sample, the probability of an **unusual** sample mean……… DECREASES INCREASES

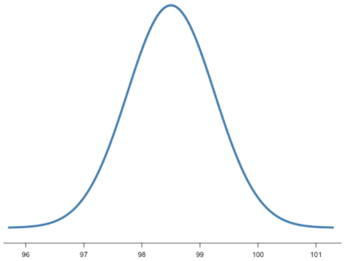
d) If we take a larger sample, the probability of an **usual** sample mean……..….. DECREASES INCREASES

If we repeatedly take samples of size *n* from a population with an unknown distribution and calculate the mean of each sample,

- $E\left(\bar{X}\right) = \mu_{\bar{x}} = \mu_x$ The mean of the sample means will equal the population mean
- $\sqrt{Var\left(\bar{X}\right)} = \sigma_{\bar{x}} = \dfrac{\sigma_x}{\sqrt{n}}$ The standard deviation of the sample means (the standard error) shrinks as the sample size increases
- We still don't know what shape the distribution of sample means will have (although, in this example, it looks like the distribution becomes unimodal and symmetric)

4. Suppose body temperatures for a population of interest follow a normal distribution with **μ = 98.5** and **σ = 0.75**.

a) Suppose we randomly select a single individual from this population. Use a computer (R, Wolfram Alpha, or the normal distribution applet*) to calculate:



$P\left(97.75 < X < 99.25\right) = P\left(-1 < Z < +1\right) =$ ____________________

$P\left(X < 98\right) =$ ____________________

b) Suppose we randomly select a sample of n=100 individuals from this population. Circle the correct symbol.

$P\left(97.75 < \bar{X} < 99.25\right)$ $<$ $=$ $>$ $0.683$

$P\left(\bar{X} < 98\right)$ $<$ $=$ $>$ $0.252$

c) Sketch the distribution of sample averages we'd get if we repeatedly sampled n=100 individuals.

d) Now calculate the probabilities for a sample of n=100 individuals:

$P\left(97.75 < \bar{X} < 99.25\right) =$ ____________________

$P\left(\bar{X} < 98\right) =$ ____________________

e) To calculate those probabilities, we assumed the sampling distribution had what kind of shape? ______________

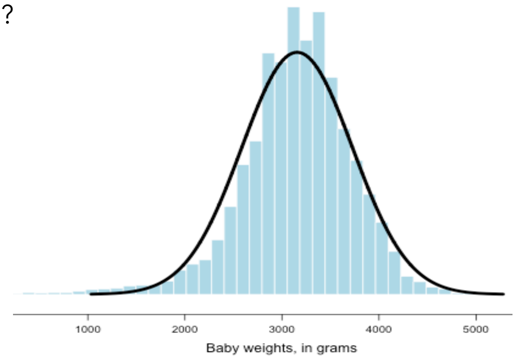
Applet: http://lock5stat.com/statkey/theoretical_distribution/theoretical_distribution.html#normal

To calculate the previous 2 probabilities, we needed to assume the sampling distribution was approximately normal. Is there a way we can know the shape of the distribution of sample means?

Scenario: Researchers collected data from 4,390 babies born to mothers in Georgia from 1980-1992. The birth weights of these babies approximated a normal distribution with a mean of 3156.3 grams and a standard deviation of 570.44.

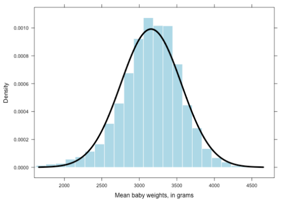
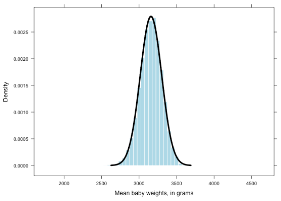
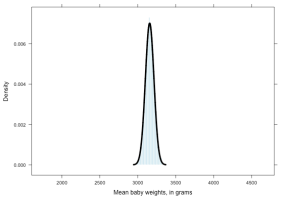
I had a computer randomly sample babies from this dataset and calculate the average weight for each sample.

I repeated this process 10,000 times to plot the sampling distributions.



Data: Adams MM, et. al. The relationship of interpregnancy interval to infant birthweight and length of gestation among low-risk women, Georgia. Paediatr Perinat Epidemiol. 1997 Jan;11 Suppl 1:48-62

5. Below, I've pasted results from my computer simulations. Fill-in-the-blanks to see if these simulated sampling distributions agree with the theory we've derived. Explain why the simulated results do not match the theory perfectly.

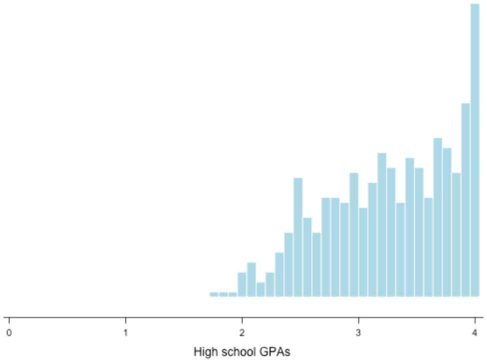
| Sample Size | Sampling distribution | Mean of sample means | Theoretical Mean | Standard deviation of sample means | Theoretical standard error |
|---|---|---|---|---|---|
| 2 |  | 3161.37 | __________ | 398.382 | __________ |
| 16 |  | 3154.85 | __________ | 141.975 | __________ |
| 100 |  | 3156.30 | 3156.3 | 56.933 | 57.044 |

It looks like these sampling distributions are approximately normal, but that might be because the population distribution was approximately normal. What happens if we start with a population that is not normally distributed?

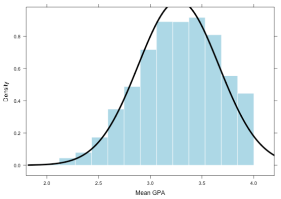
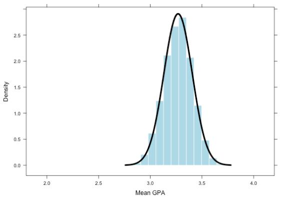
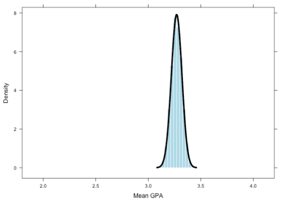
Scenario: The high school GPAs of 556 St. Ambrose freshmen in 2012 are displayed to the right. These GPAs are obviously not normally distributed (they have a negative skew).

The mean is 3.27 with a standard deviation of 0.5527.

I had a computer randomly sample GPAs from this dataset and calculate an average. I repeated this 10,000 times and graphed all 10,000 mean GPAs.



6. Fill-in-the-blanks. Do our theoretical results hold for populations that are not normally distributed?

| Sample Size | Sampling distribution | Mean of sample means | Theoretical Mean | Standard deviation of sample means | Theoretical standard error |
|---|---|---|---|---|---|
| 2 |  | 3.273 | __________ | 0.391 | __________ |
| 16 |  | 3.269 | 3.27 | 0.1371 | 0.1382 |
| 100 |  | 3.270 | 3.27 | 0.0504 | 0.0553 |

7. Under what conditions does it appear as though the distribution of the sample mean will be approximately normal?

8. Use the following applet to predict the distribution of various sample statistics under various conditions: http://www.onlinestatbook.com/stat_sim/sampling_dist/index.html.

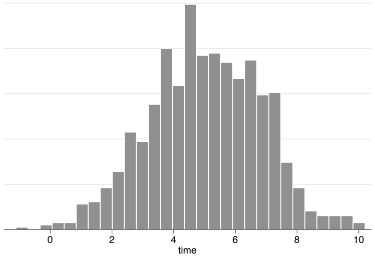
**Central Limit Theorem:**

If we repeatedly take samples of size *n* from a population with an unknown distribution and calculate the mean of each sample,

- $E\left(\bar{X}\right) = \mu_{\bar{x}} = \mu_x$ The mean of the sample means will equal the population mean
- $\sqrt{Var\left(\bar{X}\right)} = \sigma_{\bar{x}} = \dfrac{\sigma_x}{\sqrt{n}}$ The standard deviation of the sample means (the standard error) shrinks as the sample size increases

- The sampling distribution of sample means will approximate a normal distribution if:
  a) The population follows a normal distribution, or
  b) We repeatedly take large sample sizes (how large?)

Scenario: A (hypothetical) statistics professor often continues lecturing after the class period should have ended. Let X = the amount of time the professor lectures after class should have ended. Suppose students recorded X each day for several years and found X has a mean of 5 minutes and a standard deviation of 1.8 minutes.



9. Suppose we sample 1, 5, or 25 class days at random. Calculate the following probabilities:

| Sample 1 day: | Sample 5 days | Sample 25 days |
|---|---|---|
| **μ =** __________ and **σ =** ____________ | **μ =** __________ and **σ =** ____________ | **μ =** __________ and **σ =** ____________ |
| $P(X < 5.5) =$ ____________ | $P(\bar{X} < 5.5) =$ ____________ | $P(\bar{X} < 5.5) =$ ____________ |
| $P(X > 7) =$ ____________ | $P(\bar{X} > 7) =$ ____________ | $P(\bar{X} > 7) =$ ____________ |

10. Suppose we repeatedly sample 25 days and calculate the average time lecturing. What average represents the 10th percentile of this distribution?

11. Complete the following:

If you sample one day, $0.95 = P($__________________ $\leq \bar{X} \leq$ __________________$)$

If you sample 100 days, $0.95 = P($__________________ $\leq \bar{X} \leq$ __________________$)$

12. What sample size would we need in order for $0.95 = P\left(4.5 \leq \bar{X} \leq 5.5\right)$