STAT 201 Chapter 7

Sampling Distribution

Sampling Distributions: Example 1

- There is a turkey farm in Texas. Inside the farm there are 20 coops, and in each coop there are 100 Jr. turkeys.

- We know the weight of a Jr. turkey living in this farm has mean 12 lbs and standard deviation 1 lb with normal distribution.

Sampling Distributions: Example 1

- The owner of this turkey farm only sells his Jr. turkey in the unit of coop.
- A manager from a local restaurant wants to buy a randomly selected coop of Jr. turkey. What is the probability that the manager can get at least 1300 lbs?

Sampling Distributions: Example 1

- Q: What is the probability that the manager can get at least 1300 lbs of Jr. turkey from a randomly selected coop?

- P(the manager can get at least 1300 lbs of Jr. turkey from a randomly selected coop)

=P(the **average weight of the Jr. turkeys in the randomly selected coop** bought by the manger is at least 1300/100=13 lbs)

Sampling Distributions: Example 1

- Denote X be the weight of a randomly selected Jr. turkey
- X is normally distributed with mean 12 and std 1
- What is the probability that **a randomly selected turkey** has at least 13 lbs?
- P(X>13) = 16%! Why? Empirical Rule!

Sampling Distributions: Example 1

- The probability that **a randomly selected Jr. turkey** has at least 13 lbs is 16%
- Is the probability the **average weight of the Jr. turkeys in the randomly selected coop** bought by the manger is at least 13 lbs also 16%?
- What is your opinion?

Sampling Distributions

- A **sampling distribution** is the probability distribution that specifies probabilities for the possible values of the **statistics** that we are interested in.
- The statistics are unknown before sampling, which means they are random.
- The interesting statistics could be mean and proportion, etc.

Sampling Distributions

- This may be confusing:

- In last chapter, we talked about events and random variables in n trials (think as n=100, trial=turkey)
- Now, we're talking about m groups of n trials which yield m sample means or m sample proportions (think as m=20, group=coop, n =100, trial=turkey)

Sampling Distributions: Proportions

- The first sampling distribution we'll talk about is the sampling distribution for the **sample proportion**
- The idea is that there is some true population proportion out there, say p, but it isn't feasible to actually calculate it
  - We may not have enough time or money to poll the population
  - It may be infeasible to get a population measure

Sampling Distributions: Proportions

- Instead, we look at sample proportions, say $\hat{p}$, the proportion of X’s that have a certain characteristic among our sample

Sampling Distributions: Proportions

- Before: $x_1, x_2, x_3, \ldots, x_n$
  - And we'd find summary statistics of all x's
  - We would have one sample proportion, one $\hat{p}$
  - $\hat{p} = \frac{\text{number of x with desired trait}}{\text{total sample size}}$

Sampling Distributions: Proportions

- Now: We have m groups: $\{x_{1,1}, x_{1,2}, x_{1,3}, \ldots, x_{1,n}\}$, $\{x_{2,1}, x_{2,2}, x_{2,3}, \ldots, x_{2,n}\}$, ……, $\{x_{m,1}, x_{m,2}, x_{m,3}, \ldots, x_{m,n}\}$
  - And we'd find summary statistics of all x's for each group
  - We would have m sample proportions , one $\hat{p}$ for each group
  - $\widehat{p_1} = \frac{\text{number of x with desired trait in group 1}}{\text{total sample size } of\ group\ 1}$
  - $\widehat{p_2} = \frac{\text{number of x with desired trait in group 2}}{\text{total sample size of } group\ 2}$ …
  - $\widehat{p_m} = \frac{\text{number of x with desired trait in group } m}{\text{total sample size of group m}}$

Sampling Distributions: Proportions

- You could think of each group as a barrel and we're only interested in the proportion of each barrel; we are no longer interested in the individual responses
- The example below shows how we would summarize 4 groups of 10 observations, into four representative sample proportions

$\widehat{p_1}$: $x_{1,1}, x_{1,2}, \ldots, x_{1,10}$

$\widehat{p_2}$: $x_{2,1}, x_{2,2}, \ldots, x_{2,10}$

$\widehat{p_3}$: $x_{3,1}, x_{3,2}, \ldots, x_{3,10}$

$\widehat{p_4}$: $x_{4,1}, x_{4,2}, \ldots, x_{4,10}$

Sampling Distributions: Example 2

- 300 hippopotamuses are in a zoo. 3 hippos in a group and 100 groups.
- 3 hippos (a group) are randomly selected to see if they are hungry. We know that 60% of hippos are hungry. What is the sampling distribution of the sample proportion of hippos who are hungry for the sample size of three?
- Note: we aren't interested in "yes" or "no" for each individual, but the proportion among the three

Sampling Distributions: Example 2

Sampling Distributions: Example 2

- Let X* = hungry status of three hippos in each group
- Let X = the proportion of the three hippos that are hungry for each group
- We used to look at the possibilities of yes and no's: {YYY, YYN, YNY, NYY, NNY, NYN, YNN, NNN} and give a descriptive statistic of the frequency
- Now, we consider {3/3, 2/3, 1/3, 0/3} across repeated samples of hippo. Note that there are less possible outcomes here.

Sampling Distributions: Example 2

- Before: {YYY, YYN, YNY, NYY, NNY, NYN, YNN, NNN}
- Now: {3/3, 2/3, 1/3, 0/3}
  - X = the proportion of the three hippos that are hungry
  - We multiply the probability of seeing any proportion by the number of ways we can see it. (Why? Binomial) We can write out the discrete probability distribution

X (Call this p-hat)	How Many Ways	Probability
0/3	1	0.4*0.4*0.4 = 0.064
1/3	3	3*(0.6*0.4*0.4) = 0.288
2/3	3	3*(0.6*0.6*0.4) = 0.432
3/3	1	0.6*0.6*0.6 = .216

Sampling Distributions: Example 2

- Let’s check this is still a valid probability distribution

X (Call this p-hat)	How Many Ways	Probability
0/3	1	0.4*0.4*0.4 = 0.064
1/3	3	3*(0.6*0.4*0.4) = 0.288
2/3	3	3*(0.6*0.6*0.4) = 0.432
3/3	1	0.6*0.6*0.6 = .216

- All the probabilities are between 0 and 1
- The probabilities sum to 1
- X follows a valid discrete distribution!

Sampling Distributions: Example 2

X (Call this p-hat)	How Many Ways	Probability
0/3	1	0.4*0.4*0.4 = 0.064
1/3	3	3*(0.6*0.4*0.4) = 0.288
2/3	3	3*(0.6*0.6*0.4) = 0.432
3/3	1	0.6*0.6*0.6 = .216

- We know a discrete distribution, what is the mean?
- $\left(\frac{0}{3}\right) * .064 + \left(\frac{1}{3}\right) * .288 + \left(\frac{2}{3}\right) * .432 + \left(\frac{3}{3}\right) * .216 \; = 0.6$
- The **mean (expected proportion)** of hungry hippos is 0.6. Note, this gives us the same as the **population proportion** given in the question.

Sampling Distribution: Extending

- Now, we only have size three for each group. This is OK, but what if we have size one hundred, one thousand, or one million? Would you want to write out all the possible combinations (sample space)? That would be AWFUL!
- Luckily, someone else has figured out some properties of discrete distributions and we can use these properties to understand large size problems

Sampling Distribution: Mean and Std. Error

- The **mean** of the sampling distribution for a sampling proportion will always equal the population proportion:

$$\boldsymbol{\mu_{\hat{p}} = p}$$

- Even though we know the mean is the population proportion, we note that some $\hat{p}$ will be lower and some will be higher. (Why? It is random!)

Sampling Distribution: Mean and Std. Dev.

- The **standard error**, the standard deviation of the sample proportion, is

$$\boldsymbol{\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}}$$

- What if we increase/decrease group size n?

Sampling Distributions: Large Sample Size

- Now, we know the mean and standard deviation of the sample proportions.
- If sample size is large, we can calculate z-scores to find some probabilities associated with sample proportions just like we did before.

$$\boldsymbol{\mu_{\hat{p}} = p}$$

$$\boldsymbol{\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}}$$

$$z = \frac{observation - mean}{st.dev} = \frac{\hat{p} - \mu_{\hat{p}}}{\sigma_{\hat{p}}} = \frac{\hat{p} - p}{\sqrt{\frac{\boldsymbol{p(1-p)}}{\boldsymbol{n}}}}$$

Sampling Distributions: Example 2

- 3 hippos are randomly selected to see if they are hungry. We know that 60% of hippos are hungry.

- **Mean** = 0.6
- **Standard Error** = $\sqrt{\frac{\mathbf{0.6(1-0.6)}}{\mathbf{3}}}$ = 0.2828

Sampling Distributions: Example 2

- We know that most hippos are hungry, because 60% of the population is, but our samples can vary, which means some $\hat{p}$ will be lower/higher, since it is random.
- For instance, it's possible that we can randomly select three hippos and all of them are not hungry.

Sampling Distributions: Example 2

- What if there are 40 hippos in each group, what is the mean and standard error?

- **Mean** = 0.6
- **Standard Error** = $\sqrt{\frac{\mathbf{0.6(1-0.6)}}{\mathbf{40}}}$ = 0.077

Sampling Distributions: Example 2

- What is the probability that most hippos in a random sample of 30 are hungry? (Most: more than half)
- $P(\hat{p} > 0.5) = P\left(Z > \frac{0.5-0.6}{\sqrt{\frac{0.6(1-0.6)}{40}}}\right) = \mathrm{P(Z{>}{-1.29})} =$ $1-\mathrm{P(Z{<}{-1.29})} = 1-0.0985 = 0.9015$
- Note that we assume the sample size is large here.
- Say np=40*0.6=24>15 and nq=40*0.4=16>15

Sampling Distributions: Example 2

- $P(\hat{p} > 0.5)$=0.9015
  - This tells us that 90.15% of random samples of 40 hippos will have a sample proportion of hungry hippos greater than 0.5
  - We should expect 90.15% of our repeated random samples to have a sample proportion, $\hat{p}$, greater than 0.5

Sampling Distributions: Means

In turkey farm example, we are interested in the mean of weight of Jr. turkeys in each coop, called sample mean.

We write the mean of sample x's as $\bar{x}$

Sampling Distributions: Means

- Before: $x_1, x_2, x_3, \ldots, x_n$
  - And we'd find summary statistics of all x's
  - We would have one sample proportion, one $\bar{x}$
  - $\bar{x} = \dfrac{\text{the sum of x's}}{\text{the total sample size}} = \dfrac{\sum x}{\text{n}}$

Sampling Distributions: Means

- Now: We have m groups: $\{x_{1,1}, x_{1,2}, x_{1,3}, \dots, x_{1,n}\}$, $\{x_{2,1}, x_{2,2}, x_{2,3}, \dots, x_{2,n}\}$, ……, $\{x_{m,1}, x_{m,2}, x_{m,3}, \dots, x_{m,n}\}$
  - And we'd find summary statistics of all x's for each group
  - We would have m sample proportions , one $\bar{x}$ for each group
  - $\overline{x_1} = \frac{\text{the sum of x}'\text{s in group 1}}{\text{the total sample size}}$
  - $\overline{x_2} = \frac{\text{the sum of x}'\text{s in group 2}}{\text{the total sample size}}$ …………
  - $\overline{x_m} = \frac{\text{the sum of x}'\text{s in group } m}{\text{the total sample size}}$

Sampling Distributions: Means

- You could think of each group as a barrel and we're only interested in the mean of each barrel; we are no longer interested in the individual responses
- The example below shows how we would summarize 4 groups of 10 observations, into four representative sample means

$\overline{x_1}$ $\overline{x_2}$ $\overline{x_3}$ $\overline{x_4}$

$x_{1,1}, x_{1,2}, \dots, x_{1,10}$ $x_{2,1}, x_{2,2}, \dots, x_{2,10}$ $x_{3,1}, x_{3,2}, \dots, x_{3,10}$ $x_{4,1}, x_{4,2}, \dots, x_{4,10}$

Sampling Distribution: Mean and SD

- The mean of the sampling distribution for a sample mean will always equal the population mean: $\boldsymbol{\mu_{\bar{x}} = \mu_x}$
  - but we note that some $\bar{\boldsymbol{x}}$ will be lower and some will be higher
- The standard error, the standard deviation of the sample mean, is:

$$\boldsymbol{\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}}}$$

Sampling Distribution: Mean and SD

- $\boldsymbol{\mu_{\overline{x}} = \mu_x}$
  - but we note that some $\boldsymbol{\overline{x}}$ will be lower and some will be higher
- $\boldsymbol{\sigma_{\overline{x}} = \frac{\sigma_x}{\sqrt{n}}}$
- Aside:
    - What if we increase n?
        - The standard deviation shrinks
    - What if we decrease n?
        - The standard deviation grows

Sampling Distributions

- Now that we know the mean and standard deviation of the sample means we can calculate z-scores to find some probabilities associated with sample means just like we did before.

$$\boldsymbol{\mu_{\bar{x}} = \mu_x}$$

$$\boldsymbol{\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}}}$$

$$z = \frac{observation - mean}{st.dev} = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{\bar{x} - \mu_x}{\frac{\sigma_x}{\sqrt{n}}}$$

Sampling Distributions: Example 1

- X be the weight of a randomly selected Jr. turkey
- X is normally distributed with mean 12 and std 1
- What is the probability that the manager can get at least 1300 lbs of Jr. turkey from a randomly selected coop containing 100 Jr. turkeys?
- What is the probability that the manager can get a coop with average weight of Jr. turkeys is at least 13 lbs?

Sampling Distributions: Example 1

- **First of all**, we need to find the sampling distribution of means
- $\mu_{\bar{x}} = \mu_x$ = 12
  - but we note that some $\boldsymbol{\bar{x}}$ will be lower and some will be higher
- $\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}} = \frac{1}{\sqrt{100}} = 0.1$

Sampling Distributions: Example 1

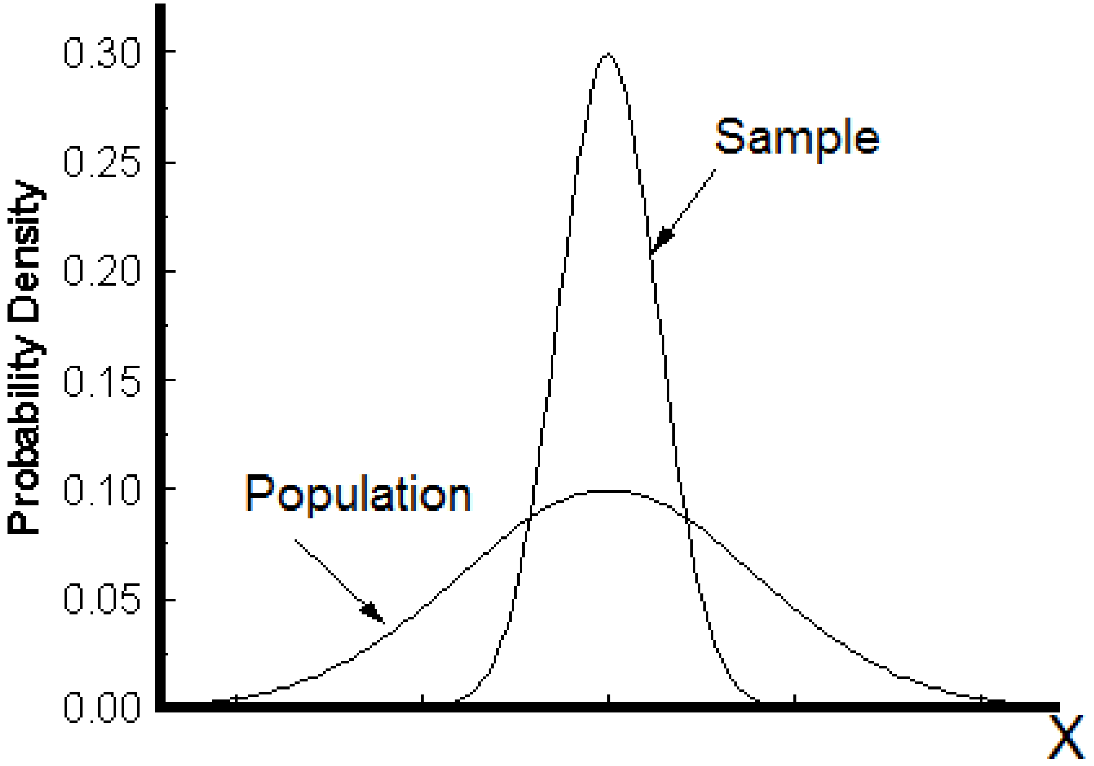
- P{the manager can get a coop with average weight of Jr. turkeys is at least 13 lbs} = P($\overline{\boldsymbol{x}}$ > 13)
- P($\overline{\boldsymbol{x}}$ < 13) = P(Z < $\frac{13-12}{0.1}$) = P(Z < 10) = 1
- P($\overline{\boldsymbol{x}}$ > 13) = 1 - P($\overline{\boldsymbol{x}}$ < 13) = 1 – 1 = 0

Sampling Distributions: Example 1

- The probability that **a randomly selected Jr. turkey** has at least 13 lbs is 16%
- Is the probability the **average weight of the Jr. turkeys in the randomly selected coop** bought by the manger is at least 13 lbs also 16%? No, it is 0%!
- Why?

Sampling Distribution: Graphs

- Sample vs. Population: the sampling distribution is narrower than the population distribution because grouping the data reduces the variation; pay attention to the standard error formula



Central Limit Theorem

- For random sampling with a large sample size n, the sampling distribution of the sample mean is approximately normal. (We say n=30 is large enough)

- Nice introduction in YouTube: https://www.youtube.com/watch?v=Pujol1yC1_A

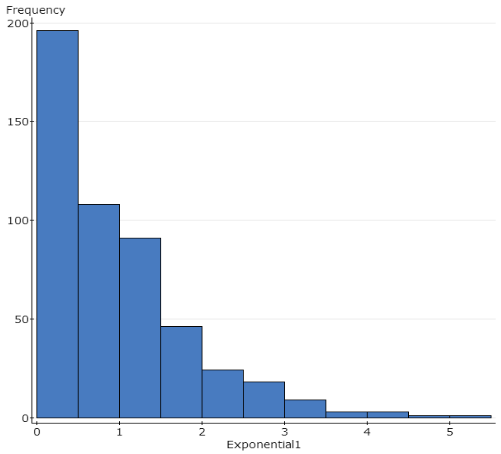
Central Limit Theorem

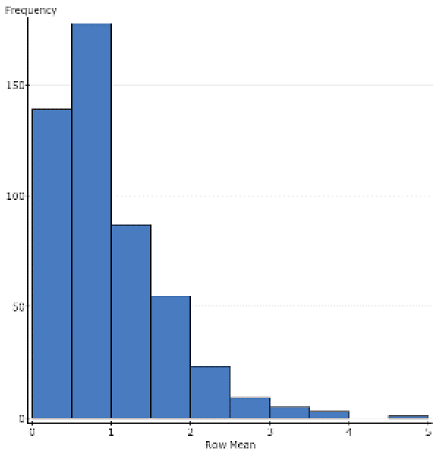
1) For any population the sampling distribution of $\bar{x}$ is bell-shaped when the sample size is large (n>30).
2) The sampling distribution of $\bar{x}$ is bell-shaped when the population distribution is bell-shaped, regardless of sample size.
3) We do not know the shape of the sampling distribution of $\bar{x}$ if the sample size is small and the population distribution isn’t bell-shaped.

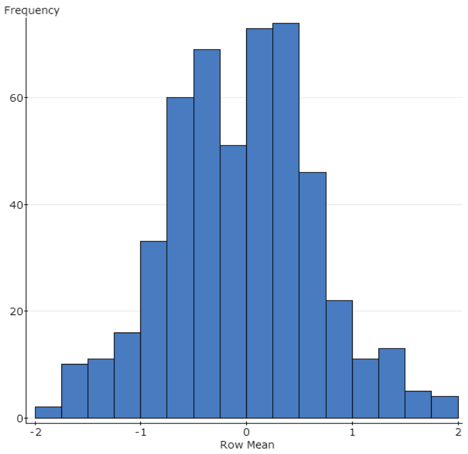
Central Limit Theorem

- For any population the sampling distribution of $\bar{x}$ is bell-shaped when the sample size is large (n>30).

Population $\quad$ $\bar{x}$ when n=2 $\quad$ $\bar{x}$ when n=30



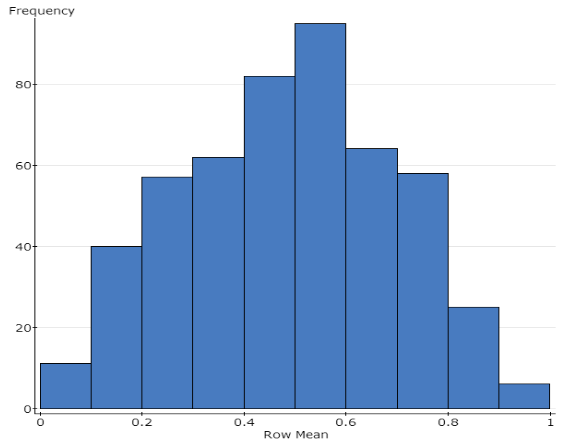




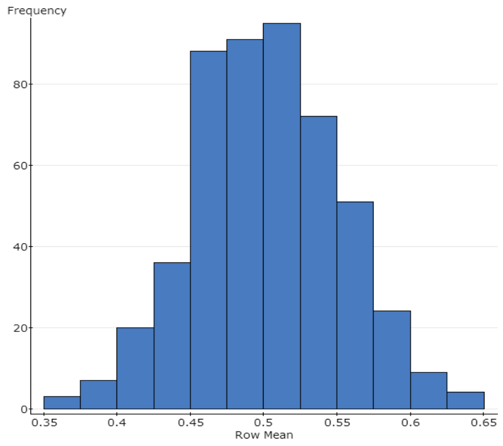
Central Limit Theorem

- The sampling distribution of $\bar{x}$ is bell-shaped when the population distribution is bell-shaped, regardless of sample size.

Population



$\bar{x}$ when n=2



$\bar{x}$ when n=30