ST 517 Note Outline 4: Sampling Distributions

Lecture 4.1: Distributions for Linear Combinations of Variables

Recall: Transformations

- If a and b are constants, then $Y = a + bX$ is a linear transformation of the rv X
- Expectation: $E(a + bX) = a + b\,E(X) = a + b\mu$
- Variance: $V(a + bX) = b^2V(X) = b^2\sigma^2$
- We derived these properties in the case of a single random variable
  - They can be extended to multiple random variables

Propositions:

- Let $X_1,\ldots, X_n$ be rvs with $E(X_i) = \mu_i$ and $V(X_i) = \sigma_i^2$.
- Let Y be a **<u>linear combination</u>** of the X_i's: $Y = a_1X_1 + a_2X_2 + \ldots + a_nX_n$.

1. Expectation of a linear combination, for any $X_1,\ldots, X_n$:

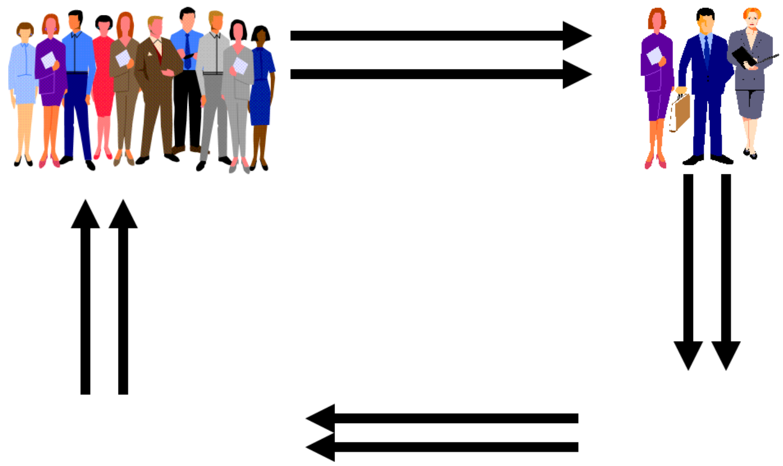
2. Variance of a linear combination, for <u>independent</u> $X_1,\ldots, X_n$:

3. If $X_1,\ldots, X_n$ are normally distributed rvs (possibly with different means and/or variances), then any linear combination of the X_i's also has a normal distribution.

- These will all be important when we discuss sampling distributions—an special type of distribution that is key for statistical inference

Lecture 4.2: Distributions for Sample Statistics

Recall the basic paradigm of statistics:



Recall the advantages of random samples:

- Unbiased – Preferences of person taking sample does not come into play
- Statistics that result have a predictable long run pattern

Connection between random sampling and random variables:

- The random variables $Y_1,\ldots,Y_n$ are said to form a random sample of size n if

Recall:

- Expected value $[E(Y)]$ represents population mean
  - Based on mathematical model for probability distribution
- Observed data:
  - Population mean often unknown
  - Sample mean $\left[\bar{Y} = \frac{\sum Y_i}{n}\right]$ estimates population mean

Important Points about Sampling

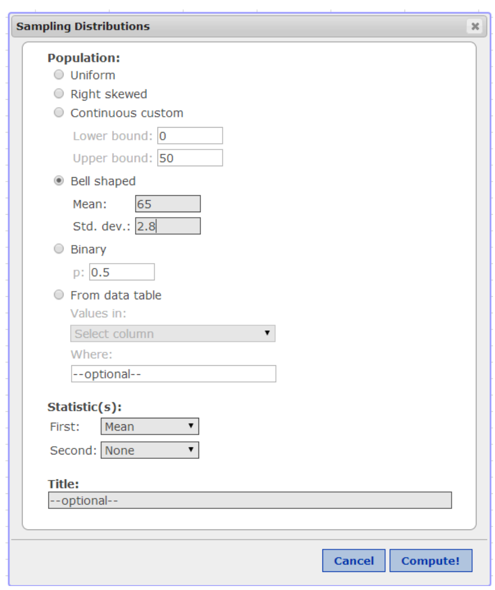
- The value of a sample statistic (such as a sample mean) depends on

- ____ is the variation in sample statistics that results from selecting different random samples.

  - The pattern of this variability is ____________________________, **if** we have a **random sample**.
  - This predictability makes ____________________________________ possible.

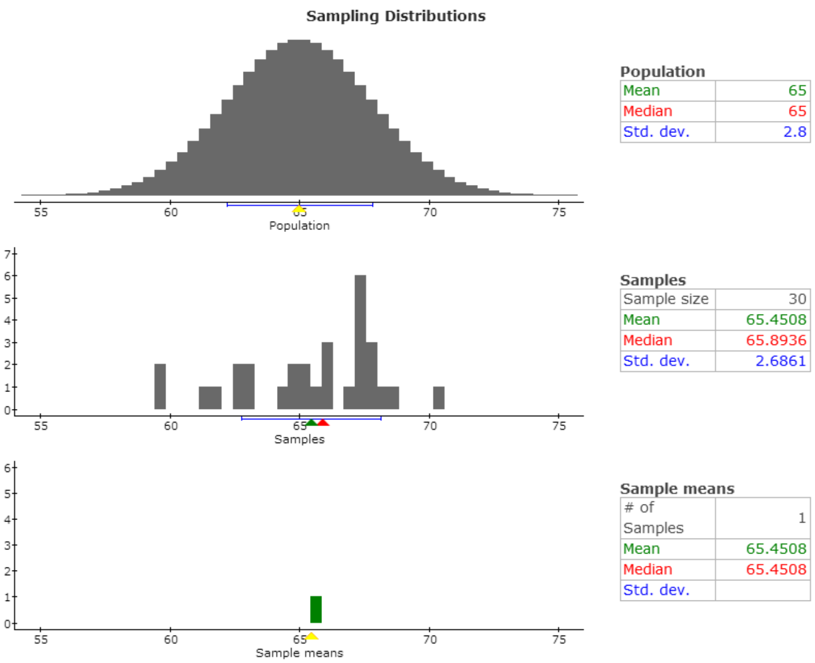
- The distribution of possible values of a sample statistic is called the

Simulating a sample



- Now use applet to experiment with sample means
- http://statcrunch.stat.ncsu.edu
  Applets>Sampling distribution

- Population:
  - Select 'bell shaped'
  - Mean = 65,
  - Std. dev. = 2.8
- Click 'Compute'



Experiment to study the standard deviation of the mean

- Set bell shaped population Mean 600 and standard deviation 100
    - n=25 standard deviation of sampling distribution ______
    - n=36 standard deviation of sampling distribution ______
    - n=100 standard deviation of sampling distribution ______
- Change to bell shaped population Mean 600 and standard deviation 200
    - n=25 standard deviation of sampling distribution ______
    - n=36 standard deviation of sampling distribution ______
    - n=100 standard deviation of sampling distribution ______

Important Points

- As sample size increased, variability in sample means

- As population standard deviation increased, variability in sample means

Standard Deviation of Sample Mean

- Fact: The standard deviation of the distribution of the sample mean $\bar{Y}$ can be predicted with the formula

$$\sigma_{\bar{Y}} = \frac{\sigma}{\sqrt{n}}$$

Example: We know that college age females have an average height of 65 inches with a standard deviation of 2.8 inches. We are going to take a random sample of 100 college age females. What would the standard deviation of the sample mean be in this situation?

Experiment to study the shape of the sampling distribution

- Right skewed population:

- Bell shaped population:

- Bi-modal population:

Important Point

- The normal distribution will be a good model for the sampling distribution of the sample mean under certain conditions.
    - Model is better if the parent population is
    - Model is better if the sample size is
        - This result is referred to as the Central Limit Theorem (CLT)

Lecture 4.3: Sampling Distribution of the Sample Mean

Summary: The Sampling Distribution of the Sample Mean

- Let $Y_1,\ldots,Y_n$ be a random sample from a population that has a $N(\mu, \sigma)$ distribution
- Distribution of the sample mean $\bar{Y}$:
    - Well modeled by
    - Mean =
    - Standard deviation =

Summary: The Central Limit Theorem (CLT)

- Let $Y_1,\ldots,Y_n$ be a random sample from *any* distribution with mean μ and standard deviation σ
- Condition needed:
    - Sample size must be
- Distribution of the sample mean $\bar{Y}$:
    - Well modeled by
    - Mean =
    - Standard deviation =
- CLT works for any shape population, so long as the sample is large
- CLT is about the sample mean not the individuals
- CLT allows us to use normal distribution as a model for distribution of sample mean

Example: Let $Y_1, \ldots, Y_n$ be a random sample from a distribution with mean μ and standard deviation σ. Show that $E(\bar{Y}) = \mu$ if $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$. [Hint: use the properties from Lecture 4.1.]

Example: For a certain population of adults, height (in inches) follows a normal distribution with a mean of 68 inches and a standard deviation of 3 inches. For parts (a) and (b), draw a well-labeled picture of the probability and use technology to find the value.

a. We randomly select a single person from this population. What is the probability that they are over 72 inches tall?

b. We select a random sample of 25 people from this population. What is the probability that the sample mean is greater than 72 inches?

c. Are your calculations in part (b) valid, even though the sample size is less than 30?

Lecture 4.4: Sampling Distribution of the Sample Proportion

Summarizing Categorical data with a Proportion

- Good way to summarize a categorical variable is with the proportion that fall into each category
- Often, we are interested in a particular category, so we focus on the proportion who have that value
  - E.g. What color car do you own? Vs. Do you have a red car?
- The sample proportion summarizes this information for the sample:

$$\hat{p} = \frac{number\ of\ \text{“}yes\text{"}(\text{"}successes\text{”})}{total\ number} = \frac{y}{n}$$

- The population proportion (p) summarizes this information for the population

Simulating a sample

- Applet to experiment with sample proportions
- http://statcrunch.stat.ncsu.edu
  Applets>Sampling distribution

Experiment to study the standard deviation of the proportion

- Set Population to Binary with p=0.5
- n=25 standard deviation of sampling distribution = ______________
- n=50 standard deviation of sampling distribution = ______________
- n=100 standard deviation of sampling distribution = ______________

Important Point

- As the sample size increased, variability in the sample proportions

Experiment to study the standard deviation of the proportion

- Keep n=100 constant
- p=0.10 standard deviation of sampling distribution = _______________
- p=0.30 standard deviation of sampling distribution = _______________
- p=0.50 standard deviation of sampling distribution = _______________
- p=0.70 standard deviation of sampling distribution = _______________
- p=0.90 standard deviation of sampling distribution = _______________

Important Point

- The standard deviation of the proportion depends on the population proportion

Standard Deviation of Sample Proportion

- Fact: the standard deviation of a sample proportion is given by $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$

Example: An accountant is reviewing the accounts for a small city's water and sewer service. The accountant knows that 20% of the city's water customers are commercial accounts (the remainder are residential accounts). The accountant picks 100 accounts at random from the city's water customers. What is the standard error of the proportion in this situation?

Experiment to study the shape of the sampling distribution

- At p=0.5 and n=10 the shape of the sampling distribution is ________________
- At p=0.5 and n=25 the shape of the sampling distribution is ________________
- At p=0.5 and n=100 the shape of the sampling distribution is ________________

- At p=0.1 and n=10 the shape of the sampling distribution is ________________
- At p=0.1 and n=25 the shape of the sampling distribution is ________________
- At p=0.1 and n=100 the shape of the sampling distribution is ________________

- At p=0.9 and n=10 the shape of the sampling distribution is ________________
- At p=0.9 and n=25 the shape of the sampling distribution is ________________
- At p=0.9 and n=100 the shape of the sampling distribution is ________________

Important Point

- The normal distribution will be a good model for the sampling distribution of the sample under certain conditions

Summary: Sampling Distribution of the Sample Proportion $\hat{p}$

- How are the possible values of the sample proportion expected to behave?
- Conditions:
- Shape:
- Center:
- Standard deviation of the sample proportion:

Example: An accountant is reviewing the accounts for a small city's water and sewer service. The accountant knows that 20% of the city's water customers are commercial accounts (the remainder are residential accounts). The accountant picks 100 accounts at random from the city's water customers.

a. Describe the sampling distribution of the sample proportion.

b. What is the probability that the proportion would be less than 14%?

Lecture 4.5: Additional Examples

Additional Example 1: Suppose the true proportion of in-state students at NC State is 80%.

a. What is the smallest sample size we could take so that the normal distribution would be a good model for the sampling distribution of $\hat{p}$?

b. Consider the sample size you solved for in part a. Based on a random sample of that size, what is the probability we would find a sample proportion of 77% or less? Draw a picture of this value and use technology to calculate it.

Additional Example 2: For a population of students, the number of hours of sleep in a typical night follows a normal distribution with a mean of 7.5 and a standard deviation of 0.75 hours.

a. What is the probability that randomly selected person from this population typically gets more than 9 hours of sleep? Draw a picture of this value and use technology to calculate it.

b. For a random sample of 11 people, what is the probability that the sample mean is between 7.5 and 7.73 hours? Draw a picture of this value and use technology to calculate it.

Additional Example 3: A large aircraft company buys metal rods that are part of an assembly in their airplanes. The specifications for the rods require that the lengths have an average of 3000mm and have a standard deviation of 20mm. The supplier of the rods also states that the rods will have a length that is normally distributed. A manager for the aircraft company was worried that the rods are not actually 3000 mm in length. In recent shipment of 25 rods the manager found that the average length was 3005 mm. The supplier said this would just be random sampling variability. If we assume the shipment is a random sample of all rods, is it reasonable to think the average would be the result of sampling variability?

Additional Example 4: Megacorp is a large manufacturing company which knows that 30% of its employees are members of a labor union. A junior executive was assigned to take a simple random sample of 300 employees. He reported that 40% of the sampled employees were members of a labor union. His department head was skeptical and stated "We know that 30% of all employees are in labor unions, so I don't see how you could take a random sample of 300 employees and find that 40% of them are in the unions. I think you either made an error in taking your sample or faked your data." The junior executive replied that this is just an example of random sampling variability. Who do you think is correct? Explain your reasoning.