ST 517 Note Outline 4: Distributions Part 2 (Important Examples)

Notes for Lecture 4.1: Famous Discrete Distribution—Binomial

Example: A multiple choice test has 20 questions. Suppose that the chance a particular student will answer each question correctly is *p*, and that the answers to each question are independent of each other.

a. What is the probability that the student answers all 20 questions correctly?

b. What is the probability that the student answers 0 questions correctly?

c. What is the probability that the student answers 18 out of 20 questions correctly?

- How many ways are there to answer 18 out of 20 questions correctly?
  - This question can be answered by considering the number of **<u>combinations</u>** (unordered groups) of size *r* that can be formed from *n* individuals:

  - So, the number of ways to answer 18 out of 20 questions correctly is:

- Bringing this all together: There are _____ terms that each have probability _____; thus the probability of getting 18 out 20 questions correct is:

- Example illustrates first famous distribution: the Binomial Distribution

The Binomial Distribution:

- PMF: $P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$
    - Notation:
- X can take integer values
- Depends on two parameters:
- Conditions for using Binomial Distribution:
- Mean of a Binomial random variable:
- Variance of a Binomial random variable:

Example: Using the Binomial distribution to make decisions

a. What is the probability of getting 70 “heads” out of 100 flips of a fair coin?

b. Now imagine you watch someone flip a coin 100 times and get 70 “heads;” would you believe that the coin was biased?

c. What is the probability of getting 7 “heads” out of 10 flips of a fair coin?

d. Imagine you watch someone flip a coin 10 times and get 7 “heads;” would you believe that the coin was biased?

Example: Suppose that about 10% of Americans are left-handed.
Let X = the number of left-handed Americans in a random sample of 12 Americans.

a. What are the mean and standard deviation of the number of left-handed Americans in the sample? Interpret each of these values.

b. What is the probability that the sample contains at most 2 left-handed Americans?

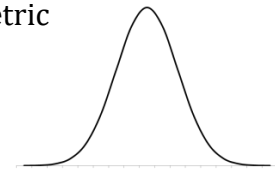
Notes for Lecture 4.2: Famous Continuous Distribution—Normal

The Normal Distribution

- Also known as the Gaussian distribution
- Most common continuous distribution since many random variables are well modeled by normal distributions

- PDF: $f(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}$
    - Notation:
- *X* can take any values
- Depends on two parameters:

- Bell shaped, unimodal, symmetric



- If you know μ and σ, you know everything there is to know about that normal distribution
    - About 68% of possible values are within 1 standard deviation of the mean
    - About 95% of possible values are within 2 standard deviations of the mean
    - About 99.7% of possible values are within 3 standard deviations of the mean
        - Distribution is about 6 standard deviations wide overall
        - Good guess for minimum value:
        - Good guess for maximum value:
        - More extreme values could occur, but would be really (really) surprising

Example: A professor teaches two sections of the same class, grading each on a curve so that a student's letter grade is based on their performance relative to that of their classmates. Grades on an exam in the first section have a mean of 85 with a standard deviation of 5. Grades on the same exam in the second section also have a mean of 85 but with a standard deviation of 3. The distributions of scores in both sections can be well modeled by a normal distribution. Suppose you are in this class and score 90 points on this exam. Which section would you rather be in (i.e. for which section would your letter grade for this exam be higher)?

Standard Scores

- Often values from very different scales are hard to compare
- Need a common measure of relative standing
- Look at how many standard deviations away from the mean the value would be:

$$Z = \frac{X - \mu}{\sigma}$$

- Standard scores can be calculated for values from any distribution, but they are commonly associated with the normal distribution
- If X follows a normal distribution with mean μ and standard deviation σ, then the standard score Z follows a normal distribution with mean 0 and standard deviation 1, called the **standard normal distribution**

Example: Using the Normal distribution to make decisions
One way to help classify dinosaur remains is by taking measurements of bones found. For example, a paleontologist could measure the width of a skull from the tip of the snout to a point at the back of the skull (in millimeters [mm]). Suppose a paleontologist is excavating in an area where two species of dinosaurs are known to have roamed. For the first species, skull widths can be well-modeled by a normal distribution with a mean of 619 mm with a standard deviation of 3.4 mm. For the second species, skull widths can be well-modeled by a normal distribution with a mean of 641 mm with a standard deviation of 10 mm. The paleontologist finds a skull with a width of 629 mm. Which species do you think it belongs to?

Notes for Lecture 4.3: Probabilities & Percentiles for a Normal Distribution

- Recall that we can find probabilities by integrating the pdf *f(x)*:

$$P(a < X < b) = \int_a^b f(x)dx$$

- Recall that for the normal distribution with mean μ and standard deviation σ:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}$$

- Integrals using the normal distribution must be approximated, which was difficult in the days before computers.
    - Not to mention that there are an infinite number of possible normal distributions one could encounter.
    - To deal with this people would "standardize" each normal distribution using the standard score. There is a table which then gives probabilities based on the standard normal distribution:

$$P(Z \leq z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-z^2/2}\, dz$$

    - However, tables are old-fashioned and only used in educational settings.
    - Fortunately for us, modern computing has improved to the point where we can use technology calculate probabilities for us.

The Normal Distribution: Finding Probabilities

- Focus on conceptual understanding of what probability represents
- Probability represents area under the curve; the percent of the distribution that is covered by the event of interest
- Exams: Communicate understanding through writing the integral or providing a well-labeled picture of the distribution
    - Need to include mean, standard deviation, and value of interest
- Quizzes, lecture examples: Use technology to calculate probabilities

Using Technology to Find Probabilities under the Normal Distribution

- Graphing calculator (e.g. TI-83 or 84)
  - Function: `normalcdf(`
  - Syntax: `normalcdf(LB, UB, mean, std_dev)`
- Software, e.g.
  - SAS:
```
DATA temp; prob_x = cdf('normal',UB,mean,std_dev);
    PROC PRINT; var prob_x; run;
```
  - Excel: `norm.dist(UB, mean, std_dev, TRUE)`
- Online calculators
  - E.g. stattrek.com/online-calculator/normal.aspx
  - Fill in: Value of x (or z) you are interested in, Mean, & Standard deviation
  - Click "Calculate" and computer will provide the Cumulative probability (e.g. area below the entered value of x)

Example: Scores on a standardized math test follow a normal distribution with a mean of 430 and a standard deviation 40. Janice scored 480; what percent of students scored below her?

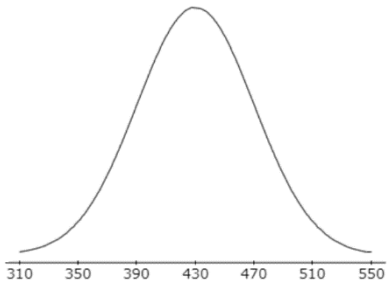
Writing the integral for this example:

$$P(X < 480) = \int_{-\infty}^{480} \frac{1}{\sqrt{2\pi(40^2)}} e^{-(x-430)^2/2(40^2)}\, dx$$

Note: $Z = \frac{X-\mu}{\sigma} = \frac{480-430}{40} = 1.25$, so this is equivalent to:

$$P(Z < 1.25) = \int_{-\infty}^{1.25} \frac{1}{\sqrt{2\pi}} e^{-(1.25^2)/2}\, dz$$

Picture for this example:



Using the technology for this example:

- Graphing calc: `normalcdf(-1000,480,430,40)`
- SAS:
```
DATA temp; prob_x = cdf('normal', 480, 430, 40);
    PROC PRINT; var prob_x; run;
```
- Excel: `norm.dist(480,430,40,TRUE)`
- Online calculator:

Standard score (z)	480			Normal random variable (x)	480
Cumulative probability P(Z ≤ z)		⇒	Calculate ⇒	Cumulative probability: P(X ≤ 480)	0.894
Mean	430			Mean	430
Standard deviation	40			Standard deviation	40

- From each of these:

The Normal Distribution: Finding Percentiles

- **Percentile** = value of variable that divides the distribution so that a specified percentage is below that value
  - Ex: 75th percentile is value of X such that 75% of area is less than x

- To calculate a percentile, you need to work backwards from the provided probability to solve for x:
  - Ex: 75th percentile

$$0.75 = P(X \leq x) = \int_{-\infty}^{x} f(y)dy$$

- Finding percentiles under a normal distribution: Focus on conceptual understanding
- As before:
  - Exams: Communicate understanding through writing the integral or providing a well-labeled picture
  - Quizzes, lecture examples: Use technology to calculate percentiles

Using Technology to Calculate Percentiles under a Normal Distribution

- Graphing calculator (e.g. TI-83 or 84)
  - Function: `invNorm(`
  - Syntax: `invNorm(proportion to left, mean, std_dev)`
- Software, e.g.
  - SAS uses $N(0,1)$ so it returns a z-score! Need to solve z-score formula for x:

```
DATA temp; z=probit(proportion); x = (z*std_dev) + mean;
PROC PRINT; var x; run;
```

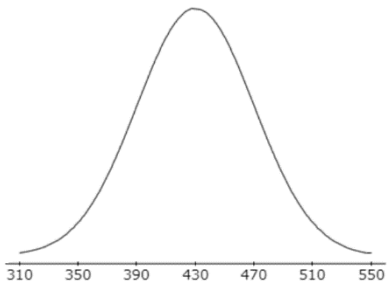
  - Excel: `norm.inv(proportion, mean, std_dev)`
- Online calculators
  - E.g. stattrek.com/online-calculator/normal.aspx
  - Fill in Cumulative probability, Mean, & Standard deviation
  - Click "Calculate" and computer will provide the value of x that is the appropriate percentile

Example: The principal of a high school wants give an award to students who score in the top 10% of the standardized mathematics test [recall: scores $\sim N(430,40)$]. What raw score has the top 10% above it?

Writing the integral for this example:

$$0.90 = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi(40^2)}} e^{-(y-430)^2/2(40^2)}\, dy$$

Picture for this example:



Using the technology for this example:

- Graphing calc: `invNorm(0.9,430,40)`
- SAS:
```
DATA temp; z = probit(proportion); x = (z*std_dev) + mean;
PROC PRINT; var x; run;
```
- Excel: `norm.inv(0.9,430,40)`
- Online calculator:

Standard score (z)				Normal random variable (x)	481.262
Cumulative probability P(Z ≤ z)	0.9	⇒	Calculate	⇒ Cumulative probability: P(X ≤ 481.262)	0.9
Mean	430			Mean	430
Standard deviation	40			Standard deviation	40

- From each of these:

Example: A proposal before the Board of Education has specified certain designations for elementary schools depending on how their students do on the Evaluation of Scholastic Scoring. This test is given to all students in all public schools in grades 2 through 4.
Schools that score in the top 20% are labeled excellent. Schools in the bottom 25% are labeled "in danger" and schools in the bottom 5% are designated as failing. Previous data suggests that scores on this test are approximately normal with a mean of 75 and a standard deviation of 5.

a. What is the probability that a randomly selected school will score below 70?

b. What is the probability that a randomly selected school will score above 83?

c. What is the probability that a randomly selected school will score between 80 and 85?

d. What is the score cut-off required for schools to be labeled excellent?

Lecture 4.4: Distributions for Sample Statistics

Recall the advantages of random samples:

- Avoids bias in process of selecting participants
- Observations in sample are independent and identically distributed (iid)
    - Independent = value for any individual is not affected by values for any other individuals
    - Identically distributed = all values follow same pattern, including mean and variance
- Statistics that result have a predictable long run pattern

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

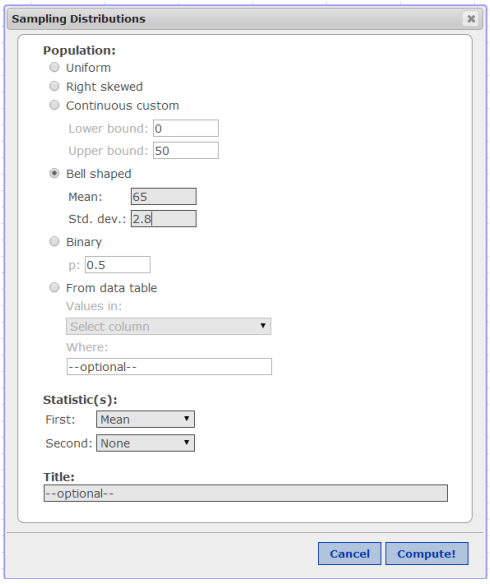
- The value of a sample statistic (such as a sample mean) depends on

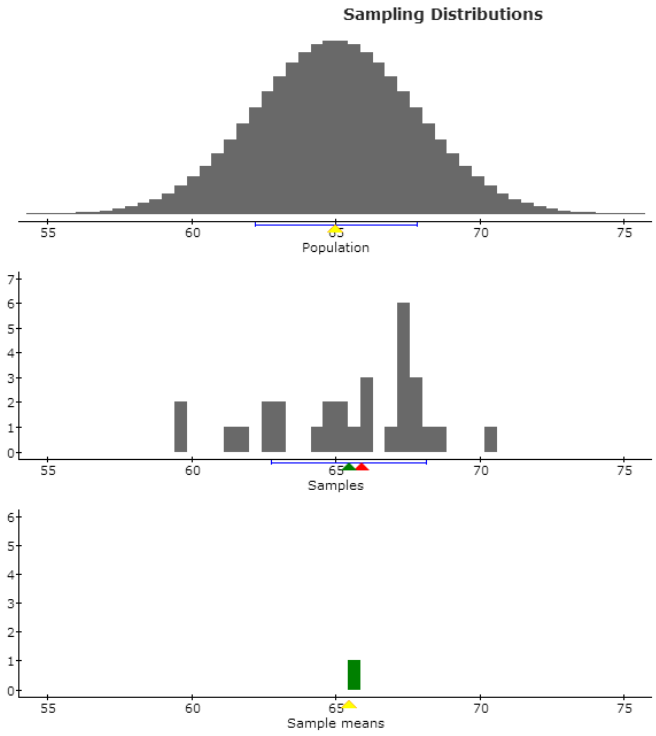
- _____ is the variation in sample statistics that results from selecting different random samples.
    - The pattern of this variability is _____, **<u>if</u>** we have a **random sample**.
    - This predictability makes _____ possible.
- The distribution of possible values of a sample statistic is 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'





Population	
Mean	65
Median	65
Std. dev.	2.8

Samples	
Sample size	30
Mean	65.4508
Median	65.8936
Std. dev.	2.6861

Sample means	
# of Samples	1
Mean	65.4508
Median	65.4508
Std. dev.	

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

- Standard deviation of the sampling distribution of $\bar{Y}$ 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 4.5: 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:
- 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 3.6.]

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 4.6: 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:

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

- 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

- 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 4.7: Additional Examples

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

a. We take a random sample of 10 NC Students. What is the probability 8 or more of them are in-state students?

b. What does your answer to part (a) indicate about the shape of the distribution used to calculate the probability?

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

d. 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. (Additional Example 2, continued) For a random sample of 11 people, what is the probability 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 aircraft company buys metal rods that are part of an assembly in their planes. Specifications for the rods require an average length of 3000mm with a standard deviation of 20mm. The supplier states that rod length is normally distributed. A manager for the aircraft company is worried 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 that the calculated average would be the result of sampling variability?

Additional Example 4: A manufacturing company 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 sample 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.