Stat 201: Introduction to Statistics

Standard 24 – Sampling Distribution for the sample proportion

Recall Definitions from Ch 2

- **Statistic**: numerical summary of a sample
  - Mean($\bar{x}$), proportion($\hat{p}$),median, mode, standard deviation(s), variance(s^2), Q1, Q3, IQR, etc.
  - We use US alphabet letters to denote these
- **Parameter**: numerical summary of a population
  - Mean(μ_x), proportion(ρ), median, mode, standard deviation(σ), variance(σ^2), Q1, Q3, IQR, etc.
  - We usually don’t know these values
  - We use Greek letters to denote these

Sampling Distributions

- Intro: https://www.youtube.com/watch?v=DmZJ1blQOns

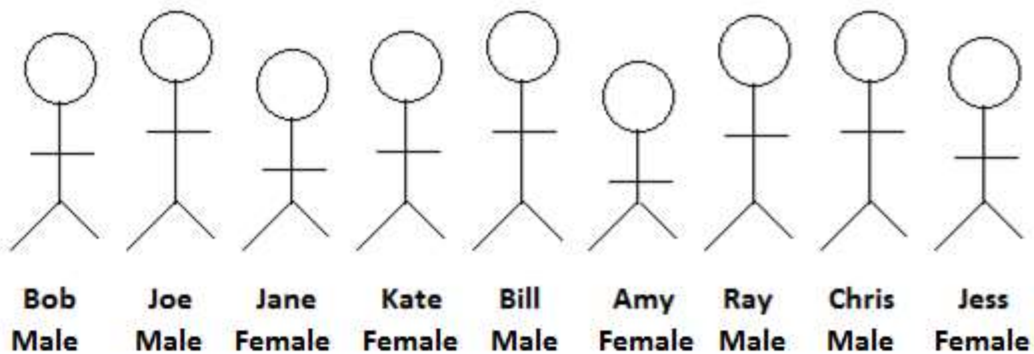
- A **sampling distribution** is the **probability distribution** that specifies probabilities for the possible values of the mean or proportion.
  - Proportions – consider the Binomial from Chapter 6
  - Means – consider the standard normal from Chapter 6
- A **sampling distribution** is a special case of a probability distribution where the outcome of an experiment that we are interested in is a sample statistic such as a **sample proportion($\hat{p}$)** or **sample mean ($\overline{x}$)**
  - It's the same as what we were doing before, but now instead of singular observations we're looking at groups

Sampling Distributions

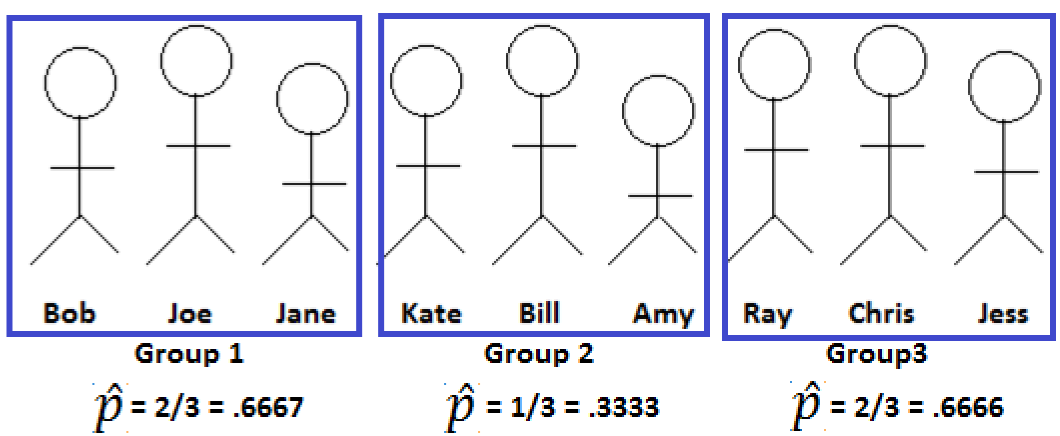
- This is confusing.
  - Remember, before we talked about events and random variables in n trials
  - Now, we're talking about m groups of n trials which yield m sample means or m sample proportions
    - $\overline{x_i} = \frac{\sum x}{n}\ for\ i = 1, 2, \ldots, m$
    - $\widehat{p_i} = \frac{x}{n}\ for\ i = 1, 2, \ldots, m$

Sampling Distributions

- Variable: Gender of Students
  - Before, we measured individuals:

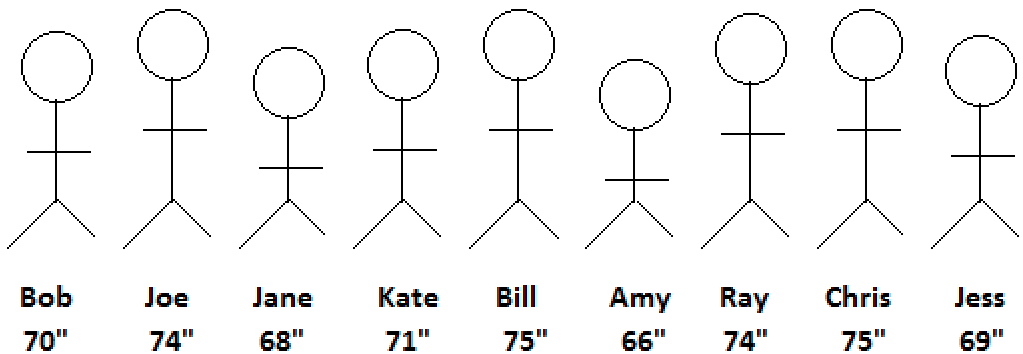


  - Now, we have one measurement across groups:

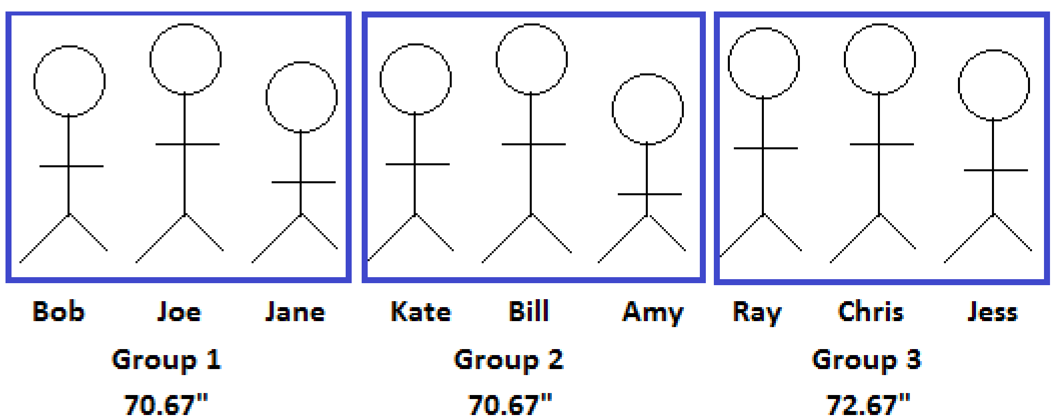


Sampling Distributions

- Variable: Heights of Americans
  - Before, we measured individuals:

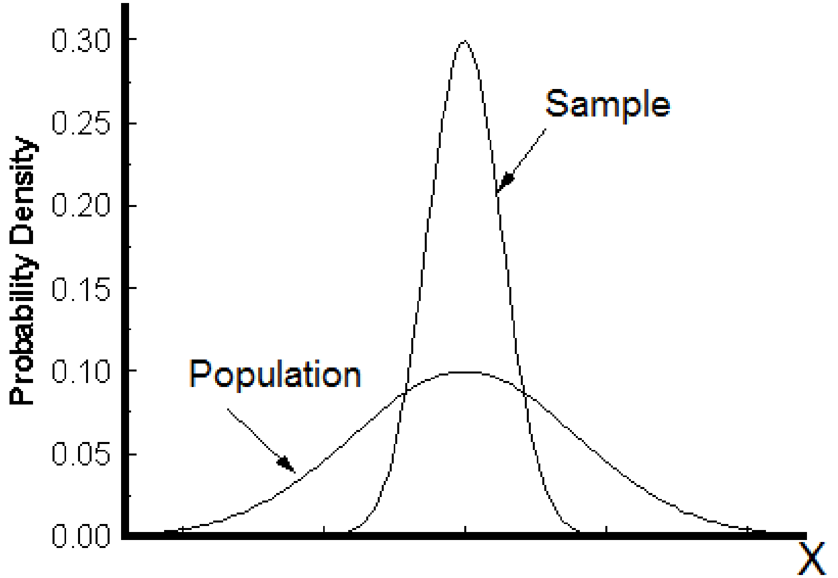


  - Now, we have one measurement across groups:



Sampling Distribution - Graphs

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



Sampling Distributions: Proportions

- This first sampling distribution we'll talk about is the **sampling distribution for the sample proportion $\hat{p}$**.

- The idea is that there is some **true population proportion out there, ρ,** but in most cases it isn't feasible to know 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

- We look at **sample proportions, $\boldsymbol{\hat{p}}$,** the proportion of observations in our sample that have a certain characteristic among our sample
  - Think "x out of n" then $\boldsymbol{\hat{p} = \frac{x}{n}}$

- We've looked at this before in the **descriptive statistics** but now we're going to talk about **all possible sample proportions from repeated random samples from the population** and their distribution (mean and standard deviation)

Sampling Distributions: Proportions

- **Before we had categorical observations:** $x_1, x_2, x_3, \dots, x_n$
  - We would summarize all x's with one **sample proportion, one $\boldsymbol{\hat{p}}$**
  - $\hat{p} = \dfrac{\text{number of x with desired trait}}{\text{total sample size}}$

    $= \text{the proportion of our sample with the desired trait}$

Sampling Distributions: Proportions

- **Now we have m groups of n subjects with categorical observations:** $\{x_{1,1}, x_{1,2}, x_{1,3}, \dots, x_{1,n}\}$,$\{x_{2,1}, x_{2,2}, x_{2,3}, \dots, x_{2,n}\}$, …, $\{x_{m,1}, x_{m,2}, x_{m,3}, \dots, x_{m,n}\}$
- Now, we find summary statistics for each group

  $\widehat{p_1}, \widehat{p_2}, \widehat{p_3}, \widehat{p_4}, \dots, \widehat{p_m}$

  - We 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 like we might have been before
- The example below shows how we could summarize 40 observations by splitting them into four representative sample proportions

$\widehat{p_1}$	$\widehat{p_2}$	$\widehat{p_3}$	$\widehat{p_4}$
$x_1, x_2, \dots, x_{10}$	$x_{11}, x_{12}, \dots, x_{20}$	$x_{21}, x_{22}, \dots, x_{30}$	$x_{31}, x_{32}, \dots, x_{40}$

Sampling Distribution – Mean and SD

- The **mean of the sampling distribution** for a sample proportion will always equal the population proportion: $\boldsymbol{\mu_{\hat{p}} = \rho}$
  - Even though we know the mean is the population proportion, we note that some $\hat{p}$ will be lower and some will be higher

Sampling Distribution – Mean and SD

- **Think about it this way:**
  - **Q:** If the population proportion of females in the United States is 51% what would you expect the number of females to be in a random sample of 100 Americans?
  - **A:** 51%, or 51 of 100, is our best guess; think of the binomial expectation.

- Later, we'll do this the other way around and we will call $\hat{p}$ the **point estimate for $\boldsymbol{\rho}$** since it's our best guess for the population proportion if we don't know it

Sampling Distribution – Mean and SD

- The **standard error**, the standard deviation of all possible sample proportions, is:

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

$$= \boldsymbol{St.Dev}(\widehat{p_1}, \widehat{p_2}, \widehat{p_3}, \widehat{p_4}, \dots, \widehat{p_m})$$

Sampling Distribution – Mean and SD

- **Think about it this way:**
  - **Q:** If our best guess for $\boldsymbol{\rho}$ is $\hat{p}$ we need a **measure of reliability** for our estimate
  - **A:** We'll talk more about this later, but our standard error calculator is a big part of this

- Recall: $\boldsymbol{\sigma_{\hat{p}}} = \sqrt{\frac{\boldsymbol{\rho(1-\rho)}}{\boldsymbol{n}}}$
- Later, in the case we don't know **ρ** we're estimating it with our **point estimate** $\hat{p}$
  - Consider:

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

Sampling Distribution – Mean and SD

- $\boldsymbol{\mu_{\hat{p}} = p}$
  - Even though we know the mean is the population proportion, we note that some $\hat{p}$ will be lower and some will be higher
- $\boldsymbol{\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}}$
  - Aside:
    - What if we increase n?
      - The standard deviation shrinks
    - What if we decrease n?
      - The standard deviation grows

Sampling Distribution:

- Now that we know the mean and standard deviation of the sample proportions 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 Distribution:

$$P(\hat{p} > c) = 1 - P\left(z < \frac{c - \mu_{\hat{p}}}{\sigma_{\hat{p}}}\right) = 1 - P\left(z < \frac{c - \rho}{\sqrt{\frac{\rho(1-\rho)}{n}}}\right)$$

$$P(\hat{p} < c) = P\left(z < \frac{c - \mu_{\hat{p}}}{\sigma_{\hat{p}}}\right) = P\left(z < \frac{c - \rho}{\sqrt{\frac{\rho(1-\rho)}{n}}}\right)$$

$$P(c_1 < \hat{p} < c_2) = P\left(z < \frac{c_2 - \mu_{\hat{p}}}{\sigma_{\hat{p}}}\right) - P\left(z < \frac{c_1 - \mu_{\hat{p}}}{\sigma_{\hat{p}}}\right)$$

$$= P\left(z < \frac{c_2 - \rho}{\sqrt{\frac{\rho(1-\rho)}{n}}}\right) - P\left(z < \frac{c_1 - \rho}{\sqrt{\frac{\rho(1-\rho)}{n}}}\right)$$

Sampling Distributions – Example 1

- The people over at Mars Candy tell us that the **population proportion** of blue M&M's is **p=1/6=0.1667**. I received a bag of M&M's from a stranger on Halloween that had **3 blue M&M's out of 25** – **is that weird?**

- Also, M&M's only have one m on them, so why aren't they just called m's?

Sampling Distributions – Example 1

- The people over at Mars Candy tell us that the population proportion of blue M&M's is $p=1/6=0.1667$. I received a bag of M&M's from a stranger on Halloween that had 3 blue M&M's out of 25 – is that weird?

- To answer this question we have to know something about the **center and spread for repeated random samples of size n = 25**. (This is another way of saying we need to know the **sampling distribution of the sample proportion.**)

Sampling Distributions – Example 1

- Let's find the sampling distribution mean:

- **The mean of all sample proportions of n=25**
  = $\mu_{\hat{p}} = p = 1/6 = .1667$

  – Some $\hat{p}$ will be lower and some will be higher but **the mean of all sample proportions of n=25 m&m's will be .1667**

Sampling Distributions – Example 1

- Let's find the sampling distribution st. deviation:

- **The st. deviation of all sample proportions of n=25**

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

  $\sqrt{\frac{.1667(1-.1667)}{25}}$=.0789

  – **The standard deviation of all sample proportions of n=25 m&m's is .0789**

Sampling Distributions – Example 1

- Let's find the sampling distribution:
- $\mu_{\hat{p}} = p = 1/6 = .1667$
  - Some $\hat{p}$ will be lower and some will be higher but **the mean of all sample proportions of n=25 m&m's will be .1667**

- $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{.1667(1-.1667)}{25}}$=.0789

Sampling Distributions – Example 1

- Is it weird to have only 3 blue m&m's or fewer in a bag of 25?
- $\hat{p} = \frac{3}{25} = .12$
- $P(\hat{p} < .12) = P\left(Z < \frac{.12-.1667}{\sqrt{\frac{.1667(1-.1667)}{25}}}\right) =$

  $\mathrm{P(Z<-.59197)} \approx \mathrm{P(Z<-.59)=.2776}$

Sampling Distributions – Example 1

- Is it weird to have only 3 blue m&m’s or fewer in a bag of 25?

- $P(\hat{p} < .12) = .2776$
- No, it’s not so weird because it happens about 25.14%, about a quarter of the time

Sampling Distributions – Example 1

- Is it weird to have only 3 blue m&m's or fewer in a bag of 25?

- $P(\hat{p} < .12) = .2776$
- No, it's not so weird because it happens about 25.14%, about a quarter of the time

Sampling Distributions – Example 2

- Say, we know that **16% of Americans approve of Congress (Gallup).**
- **What is the sampling distribution of the sample proportion** of Americans that approve of Congress for n=100?
  - Note, we aren't interested in the yes or no's individually but the proportion among the ten
  - Here, X=the proportion of the one hundred Americans in each group

Sampling Distributions - Example 2

- Say, we know that **16% of Americans approve of Congress (Gallup).**
- **What is the sampling distribution of the sample proportion** of Americans that approve of Congress for n=100?

  - n = sample size = **sample size of one hundred**= 100
  - p = population proportion = **16%** =.16

Sampling Distributions – Example 2

- Let's find the sampling distribution mean:

- **The mean of all sample proportions of n=100** = $\mu_{\hat{p}} = \rho = 16\% = .16$

  – Some $\hat{p}$ will be lower and some will be higher but **the mean of all sample proportions of n=100 will be .6**

Sampling Distributions – Example 2

- Let's find the sampling distribution st. error:

- **The st. deviation of all sample proportions of n=100**

= Standard Error = $\sigma_{\hat{p}} = \sqrt{\frac{\rho(1-\rho)}{n}}$

$$= \sqrt{\frac{.16(1-.16)}{100}} = .0367$$

Sampling Distributions – Example 2

- Let's find the sampling distribution :

$$\mu_{\hat{p}} = \rho = 16\% = .16$$

$$\sigma_{\hat{p}} = \sqrt{\frac{\rho(1-\rho)}{n}} = \sqrt{\frac{.16(1-.16)}{100}} = .0367$$

Sampling Distributions – Example 2

- The probability that **most**, of our sample of n=100, approve of Congress:

$$P(\hat{p} > .5) = P\left(z > \frac{.5 - .16}{.0367}\right) = P(Z > 9.26)$$
$$= 1 - P(Z \leq 9.26) \approx 1 - 1$$
$$= 0$$

Sampling Distributions – Example 2

- The probability that **less than 10%**, of our sample of n=100, approve of Congress:

$$P(\hat{p} < .1) = P\left(z < \frac{.1 - .16}{.0367}\right) = P(Z < -1.63)$$
$$= .0516$$

Sampling Distributions – Example 2

- The probability that **between 5 and 19 percent**, of our sample of n=100, approve of Congress:

$$P(.05 < \hat{p} < .19) = \mathrm{P}(\hat{p} < .19) - \mathrm{P}(\hat{p} < .05)$$
$$= \mathrm{P}\left(\mathrm{z} < \frac{.19 - .16}{.0367}\right) - \mathrm{P}\left(z < \frac{.05 - .16}{.0367}\right)$$
$$= P(Z < .82) - P(Z < -3.00)$$
$$= .7939 \ - .0013$$
$$= .7926$$

Central Limit Theorem: Proportions

- For random sampling with a **large sample size n, the sampling distribution of the sample proportion** is approximately a normal distribution
  - $n * p \geq 15 \; and \; n * (1 - p) \geq 15$

- Introduction:
  - https://www.youtube.com/watch?v=PujoI1yC1_A

Sampling Distribution for the Sample Proportion Summary

Shape of sample	Center of sample	Spread of sample
The shape of the distribution is bell shaped if $n * p \geq 15$ $\boldsymbol{and}$ $n * (1 - p) \geq 15$	$\mu_{\hat{p}} = p$	$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$

Sampling Distribution:

$$P(\widehat{p} > c) = 1 - P\left(z < \frac{c - \mu_{\widehat{p}}}{\sigma_{\widehat{p}}}\right) = 1 - P\left(z < \frac{c - \rho}{\sqrt{\frac{\rho(1-\rho)}{n}}}\right)$$

$$P(\widehat{p} < c) = P\left(z < \frac{c - \mu_{\widehat{p}}}{\sigma_{\widehat{p}}}\right) = P\left(z < \frac{c - \rho}{\sqrt{\frac{\rho(1-\rho)}{n}}}\right)$$

$$P(c_1 < \widehat{p} < c_2) = P\left(z < \frac{c_2 - \mu_{\widehat{p}}}{\sigma_{\widehat{p}}}\right) - P\left(z < \frac{c_1 - \mu_{\widehat{p}}}{\sigma_{\widehat{p}}}\right)$$

$$= P\left(z < \frac{c_2 - \rho}{\sqrt{\frac{\rho(1-\rho)}{n}}}\right) - P\left(z < \frac{c_1 - \rho}{\sqrt{\frac{\rho(1-\rho)}{n}}}\right)$$