Some Important Sampling Distribution (IV)

Huining Kang

HuKang@salud.unm.edu

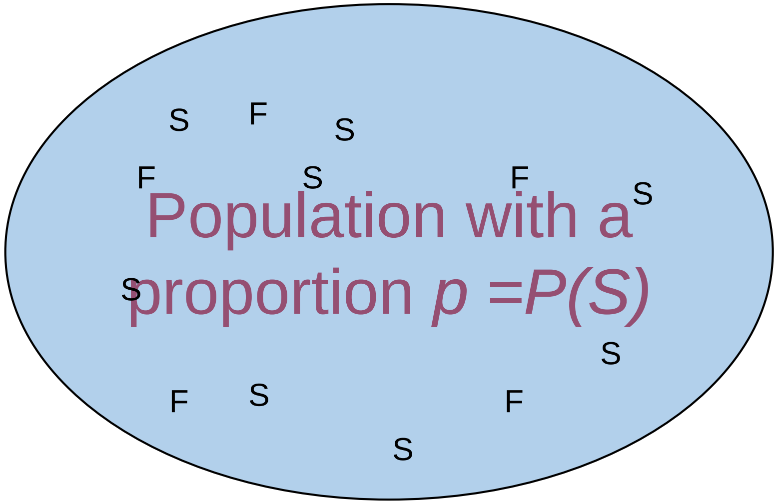
August 24, 2020

Outline

- Contents
  - Distribution of the sample proportion $\hat{p}$
    - Characteristics
    - Distribution
    - Application
  - Distribution of the difference between two sample proportions $\hat{p}_1 - \hat{p}_2$
    - Characteristics
    - Distribution
    - Application
- Learning objective
  - Understand the ideas
  - Know the characteristics (mean and variance) and the distributions of $\hat{p}$ and $\hat{p}_1 - \hat{p}_2$
  - Know how to calculate the probability related to these statistics using normal distribution

Distribution of sample proportion

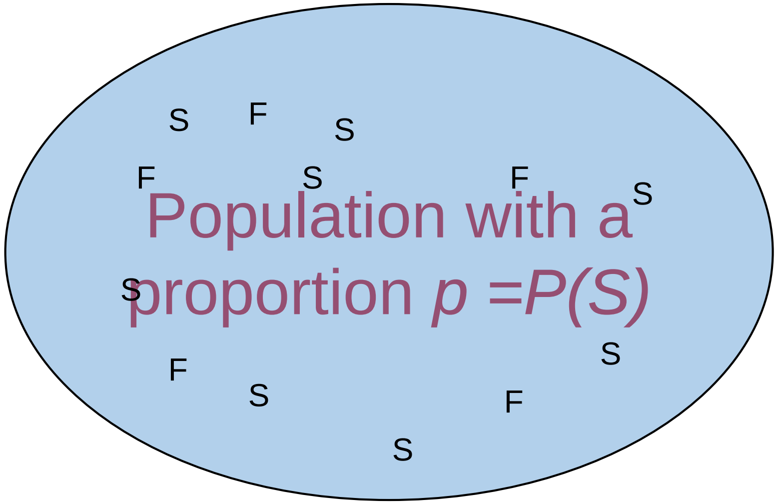
- Example 5.5.1 (page 150)
  - 35.7% of the U.S. adults aged 20 and over are obese.
  - We designate this population proportion as $p = .357$
  - If we randomly select 150 individuals from this population, what is the probability that the proportion in the sample who are obese will be .40 or greater.
- Model
  - Let
    - S="obese"
    - F ="non-obese"
  - Draw one at random and look at the obese status
    - $p(S) = p = .357$
    - $p(F) = q = 1 - p = .643$
    - Bernoulli trial



- Model (cont.)
  - Draw a sample of $n = 150$, and let $X =$ # of S's
    - $X \sim Binomial(n,\ p)$
    - Binomial distribution
- Solution
  - Find $P(X \geq .4 \times 150) = P(X \geq 60) = 1 - P(X \leq 59) = .155$
- Stata command
  - disp 1 – binomial(150, 59, .357)
  - .15531987

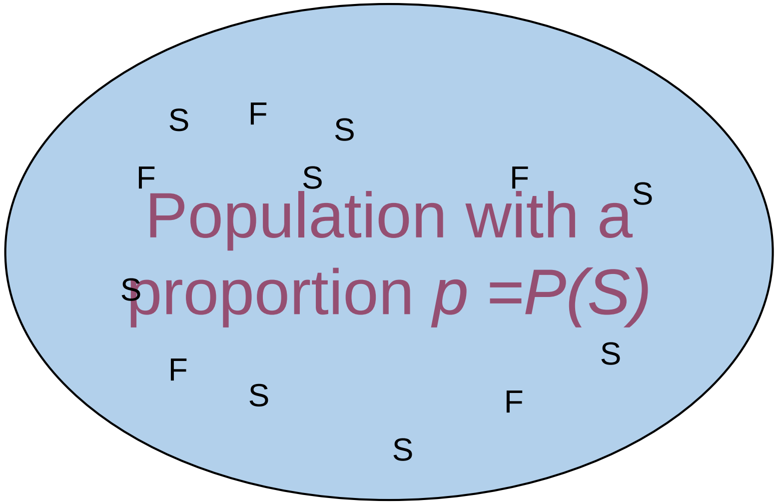
Distribution of sample proportion

- Population distribution
  - Take one element at random and denote $X = \begin{cases} 1, & if\ it\ is\ an\ S, \\ 0, & if\ it\ is\ an\ F. \end{cases}$
  - A simple Bernoulli trial
    - $p = P(X = 1)$
    - $q = P(X = 0) = 1 - p$
  - Characteristics
    - Mean $\mu = p$
    - Variance $\sigma^2 = p(1 - p)$
    - Standard deviation $\sigma = \sqrt{p(1 - p)}$



Distribution of sample proportion

- Sampling distribution
  - Take an SRS of size n, denote $X = \#$ of S's
  - Binomial distribution
    - $X \sim binomial(n, p)$
    - $n =$ sample size
    - $p =$ population proportion
  - Characteristics
    - Sample mean $\mu_X = np$
    - Variance $\sigma_X^2 = np(1-p)$
    - Standard deviation $\sigma_X = \sqrt{np(1-p)}$



Distribution of sample proportion

- Sampling distribution
  - Take an SRS of size n, denote $X =$ # of S's
  - Binomial distribution
    - $X \sim binomial(n, p)$
    - $n =$ sample size
    - $p =$ population proportion
  - Characteristics
    - Sample mean $\mu_X = np$
    - Variance $\sigma_X^2 = np(1-p)$
    - Standard deviation $\sigma_X = \sqrt{np(1-p)}$
- Normal Approximation
  - Let X_i be the ith sample unite, i.e.
    - $X_i = \begin{cases} 1, & if\ it\ is\ an\ S, \\ 0, & if\ it\ is\ an\ F. \end{cases}$
    - # S's $= \sum_{i=1}^{n} X_i$
  - Sample proportion is
    - $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}$
  - Characteristics
    - $\mu_{\hat{p}} = p$, $\sigma_{\hat{p}}^2 = \frac{p(1-p)}{n}$,
    - $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$
  - Sampling distribution of $\hat{p}$ is approximately normal if n is large ($np > 5$ and $n(1-p) > 5$)
    - $\hat{p} \sim N(p, \frac{p(1-p)}{n})$
  - Convert to Z
    - $Z = \frac{\hat{p}-p}{\sqrt{p(1-p)/n}} \sim N(0,1)$

Distribution of sample proportion

- Example 5.5.1 (cont.)
  - 35.7% of the U.S. adults aged 20 and over are obese.
  - We designate this population proportion as $p = .357$
  - If we randomly select 150 individuals from this population, what is the probability that the proportion in the sample who are obese will be .40 or greater.
- Solution 1
  - To find $P(\hat{p} \geq .4)$
  - $\mu_{\hat{p}} = p = .357$
  - $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{.357\times(1-.357)}{150}} = .03912$
  - $\hat{p} \sim N(.357, .03912^2)$
  - $P(\hat{p} \geq .4) = 1 - P(\hat{p} \leq .4) = 1 - \Phi\left(\frac{.4-.357}{.03912}\right) = .1358$
- Stata command
  - disp sqrt(.357*(1-.357)/150)
  - .03911956
  - disp 1 - normal((.4 - .357)/.03911956)
  - .135841654

- Normal Approximation
  - Let X_i be the ith sample unite, i.e.
    - $X_i = \begin{cases} 1, & if\ it\ is\ an\ S, \\ 0, & if\ it\ is\ an\ F. \end{cases}$
    - # S's = $\sum_{i=1}^{n} X_i$
  - Sample proportion is
    - $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}$
  - Characteristics
    - $\mu_{\hat{p}} = p$, $\sigma_{\hat{p}}^2 = \frac{p(1-p)}{n}$,
    - $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$
  - Sampling distribution of $\hat{p}$ is approximately normal if n is large ($np > 5$ and $n(1-p) > 5$)
    - $\hat{p} \sim N(p, \frac{p(1-p)}{n})$
  - Convert to Z
    - $Z = \frac{\hat{p}-p}{\sqrt{p(1-p)/n}} \sim N(0,1)$

Distribution of sample proportion

- Example 5.5.1 (cont.)
  - 35.7% of the U.S. adults aged 20 and over are obese.
  - We designate this population proportion as $p = .357$
  - If we randomly select 150 individuals from this population, what is the probability that the proportion in the sample who are obese will be .40 or greater.
- Solution 2
  - To find $P(\hat{p} \geq .4)$
  - $\mu_{\hat{p}} = p = .357$
  - $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{.357\times(1-.357)}{150}} = .03912$
  - Z value corresponding to $\hat{p} = .4$
    $z = \frac{.4-.357}{\sqrt{.357\times(1-.357)/150}} = 1.099$
  - $P(\hat{p} \geq .4) = P(Z \geq z) = 1 - P(Z \leq z) = 1 - \Phi(z) = 1 - \Phi(1.099) = 1.358$
- Stata command
  - disp (.4-.357)/sqrt(.357*(1-.357)/150)
  - 1.0991944
  - disp 1 - normal(1.0991944)
  - .13584164

- Normal Approximation
  - Let X_i be the ith sample unite, i.e.
    - $X_i = \begin{cases} 1, \ if\ it\ is\ an\ S, \\ 0, if\ it\ is\ an\ F. \end{cases}$
    - # S's = $\sum_{i=1}^{n} X_i$
  - Sample proportion is
    - $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}$
  - Characteristics
    - $\mu_{\hat{p}} = p$, $\sigma_{\hat{p}}^2 = \frac{p(1-p)}{n}$,
    - $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$
  - Sampling distribution of $\hat{p}$ is approximately normal if n is large ($np > 5$ and $n(1-p) > 5$)
    - $\hat{p} \sim N(p, \frac{p(1-p)}{n})$
  - Convert to Z
    - $Z = \frac{\hat{p}-p}{\sqrt{p(1-p)/n}} \sim N(0,1)$

Distribution of sample proportion

- Example 5.5.1 (cont.)
  - Results based on (exact) binomial distribution: $P(\hat{p} \geq .4)$=.15531987
  - Results based on (approx) normal distribution: $P(\hat{p} \geq .4)$=.13584164
- Normal Approximation
  - $\hat{p} \sim N(p, \frac{p(1-p)}{n})$
  - $Z = \frac{\hat{p}-p}{\sqrt{p(1-p)/n}} \sim N(0, 1)$
- Correction for continuity
  - Let use $\hat{p}_{obs}$ to denote .4, the observed sample proportion
  - When compute the probability of an event, replace $\hat{p}_{obs}$ with

$$\hat{p}_c = \begin{cases} \hat{p}_{obs} + \frac{.5}{n}, & if\ \hat{p}_{obs} < p \\ \hat{p}_{obs} - \frac{.5}{n}, & if\ \hat{p}_{obs} > p \end{cases}$$

  - Normal approximation (cont.)
    - $P\left(\hat{p} \geq .4 - \frac{.5}{150}\right) = 1 - P\left(\hat{p} \leq .4 - \frac{.5}{150}\right) = 1 - \Phi\left(\frac{.4 - \frac{.5}{150} - .375}{.03912}\right) = 1.553$
    - Stata command
      - disp 1 - normal((.4 -.5/150 - .357)/.03911956)
      - .15529484

(Closer to the exact result)

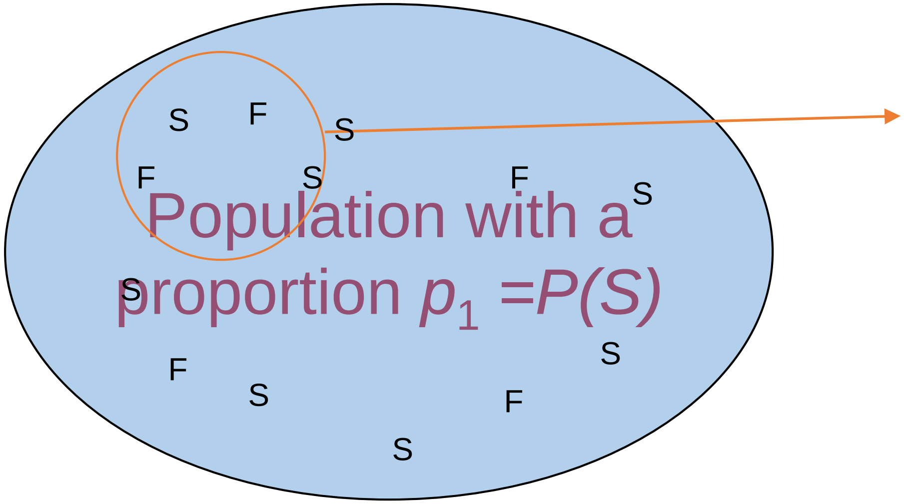
  - Most of software offer the option of correction for continuity.

Distribution of sample proportion

- Example 5.5.2 (page 152)
  - Blanche Mikhail [A-4] studied the use of prenatal care among low-income African-American women. She found that only 51 percent of these women had adequate prenatal care. Let us assume that for a population of similar low-income African-American women, 51 percent had adequate prenatal care. If 200 women from this population are drawn at random, what is the probability that less than 45 percent will have received adequate prenatal care?
- Solution
  - Population proportion: $p = .51$
  - Sample size $n = 200$ (large sample)
  - Sample proportion: $\hat{p} \sim N\left(\mu_{\hat{p}}, \sigma_{\hat{p}}^2\right)$, were $\mu_{\hat{p}} = p = .51$, $\sigma_{\hat{p}}^2 = \frac{p(1-p)}{n} = \frac{.51\times(1-.51)}{200} = .00125, \sigma_{\hat{p}} = \sqrt{.00125}$
  - Z statistic: $Z = \frac{\hat{p}-p}{\sigma_{\hat{p}}} = \frac{\hat{p}-.51}{\sqrt{.00125}} \sim N(0,1)$, z value: $z = \frac{.45-.51}{\sqrt{.00125}} = -1.70$
  - Find: $P(\hat{p} \leq .45) = P(Z \leq -1.70) = \Phi(-1.70) = 0.0448$
- Stata command
  - disp .51*(1-.51)/200
  - .0012495
  - disp (.45 - .51)/sqrt(.0012495)
  - -1.6973958
  - disp normal(-1.6973958)
  - .04481093

- With correction for continuity
  - $P\left(\hat{p} \leq .45 + \frac{.5}{200}\right) = .0519035$
- Based on Binomial Distribution
  - $P(\hat{p} \leq .45) = .05186767$

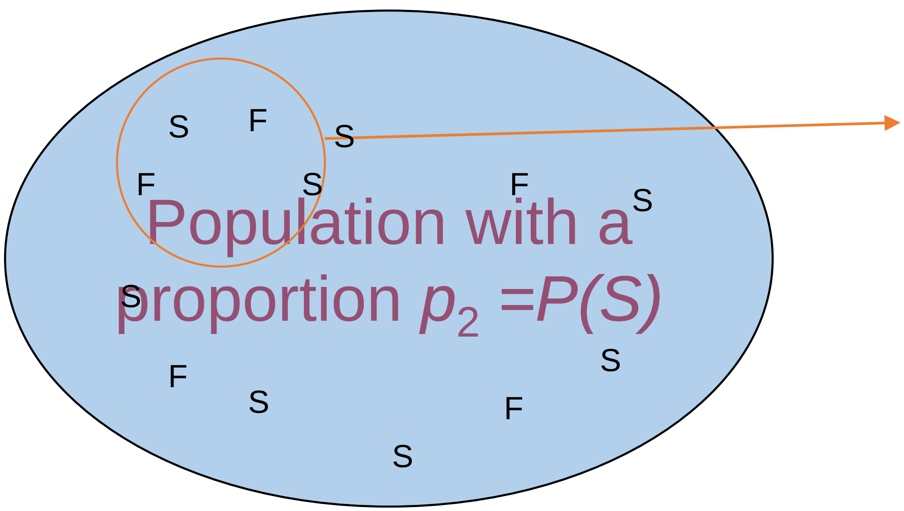
Distribution of the difference between two sample proportions



SRS of size n_1 yields

$$\hat{p}_1 = \frac{\#\, S's\ in\ Sample}{n_1}$$

Where $\hat{p}_1$ is an estimate of p_1



SRS of size n_2 yields

$$\hat{p}_2 = \frac{\#\, S's\ in\ Sample}{n_2}$$

Where $\hat{p}_2$ is an estimate of p_2

Distribution of $\hat{p}_1 - \hat{p}_2$ = ? $\quad \mu_{\hat{p}_1-\hat{p}_2} = ? \quad \sigma^2_{\hat{p}_1-\hat{p}_2} = ?$

Distribution of the difference between two sample proportions

- Characteristcs of $\hat{p}_1 - \hat{p}_2$
  - $\mu_{\hat{p}_1-\hat{p}_2} = p_1 - p_2$
  - $\sigma^2_{\hat{p}_1-\hat{p}_2} = \frac{p_1(1-P_1)}{n_1} + \frac{p_2(1-P_2)}{n_2}$
- Distribution of $\hat{p}_1 - \hat{p}_2$
  - When n_1 and n_2 are large, i.e., $n_1 p_1$, $n_2 p_2$, $n_1(1-p_1)$, $n_2(1-p_2)$ are all greater than 5

$$\hat{p}_1 - \hat{p}_2 \sim N(\mu_{\hat{p}_1-\hat{p}_2}, \sigma^2_{\hat{p}_1-\hat{p}_2})$$

$$\hat{p}_1 - \hat{p}_2 \sim N\left(p_1 - p_2, \frac{p_1(1-P_1)}{n_1} + \frac{p_2(1-P_2)}{n_2}\right)$$

- Convert to Z

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{p_1(1-P_1)}{n_1} + \frac{p_2(1-P_2)}{n_2}}} \sim N(0,1)$$

Distribution of the difference between two sample proportions

- Example 5.6.1 (Page 155)
  - Populations: (1) Self-identified Non-Hispanic White vs. (2) Hispanic origin
  - Population parameter – proportions of those who had experienced lower back pain: $p_1 = .28, p_2 = .21$
  - Sample sizes: $n_1 = n_2 = 100$
  - Statistic of interest: $\hat{p}_1 - \hat{p}_2$, where $\hat{p}_1$ and $\hat{p}_2$ are sample proportions
  - Question: $P(\hat{p}_1 - \hat{p}_2 \geq .10) = ?$
- Solution
  - $\mu_{\hat{p}_1 - \hat{p}_2} = p_1 - p_2 = .28 - .21 = .07$
  - $\sigma^2_{\hat{p}_1 - \hat{p}_2} = \frac{p_1(1-P_1)}{n_1} + \frac{p_2(1-P_2)}{n_2} = \frac{.28\times(1-.28)}{100} + \frac{.21\times(1-.21)}{100} = .003675$
  - Sampling distribution: $Z = \frac{(\hat{p}_1-\hat{p}_2)-(p_1-p_2)}{\sqrt{\frac{p_1(1-P_1)}{n_1}+\frac{p_2(1-P_2)}{n_2}}} \sim N(0,1)$
  - z value: $z = \frac{(.10)-(p_1-p_2)}{\sqrt{\frac{p_1(1-P_1)}{n_1}+\frac{p_2(1-P_2)}{n_2}}} = \frac{.10-.07}{\sqrt{.003674}} = .49$
  - $P(\hat{p}_1 - \hat{p}_2 \geq .10) = P(Z \geq .49) = 1 - P(Z \leq .49) = 1 - \Phi(.49) = .3121$
- Stata Command
  - disp .28*(1-.28)/100 + .21*(1-.21)/100
  - .003675
  - disp (.1 - .07)/sqrt(.003675)
  - .49487166
- Stata Command (cont.)
  - disp 1 - normal(.49)
  - .31206695

Distribution of the difference between two sample proportions

- Example 5.6.2 (Page 155)
  - Population1: US adult age 75 or older
    - Proportion of those who had lost all their natural teeth: $p_1 = .34$
  - Population2: US adult age 65 – 74
    - Proportion of those who had lost all their natural teeth: $p_2 = .26$
  - Sample sizes
    - $n_1 = 250$, $n_2 = 200$.
  - Statistic of interest: $\hat{p}_1 - \hat{p}_2$, where $\hat{p}_1$ and $\hat{p}_2$ are sample proportions
  - Question: $P(\hat{p}_1 - \hat{p}_2 \leq .05) = ?$
- Solution
  - $\mu_{\hat{p}_1 - \hat{p}_2} = p_1 - p_2 = .34 - .26 = .08$
  - $\sigma^2_{\hat{p}_1 - \hat{p}_2} = \frac{p_1(1-P_1)}{n_1} + \frac{p_2(1-P_2)}{n_2} = \frac{.34\times(1-.34)}{250} + \frac{.26\times(1-.26)}{200} = .0018596$
  - Sampling distribution: $Z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{p_1(1-P_1)}{n_1} + \frac{p_2(1-P_2)}{n_2}}} \sim N(0,1)$
  - z value: $z = \frac{(.05) - (p_1 - p_2)}{\sqrt{\frac{p_1(1-P_1)}{n_1} + \frac{p_2(1-P_2)}{n_2}}} = \frac{.05 - .08}{\sqrt{.0018596}} = -.69568315$
  - $P(\hat{p}_1 - \hat{p}_2 \leq .05) = P(Z \leq -.69568315\ ) = \Phi(-.69568315\ ) = .24331364$

- Stata Command
  - disp .34*(1-.34)/250 + .26*(1-.26)/200
  - .0018596
  - disp (.05 - .08)/sqrt(.0018596)
  - -.69568315
- Stata Command (cont.)
  - disp normal(-.69568315)
  - .24331364

Summary

- Contents
  - Distribution of the sample proportion $\hat{p}$
    - Characteristics
    - Distribution
    - Application
  - Distribution of the difference between two sample proportions $\hat{p}_1 - \hat{p}_2$
    - Characteristics
    - Distribution
    - Application
- Learning objective
  - Understand the ideas
  - Know the characteristics (mean and variance) and the distributions of $\hat{p}$ and $\hat{p}_1 - \hat{p}_2$
  - Know how to calculate the probability related to these statistics using normal distribution
- Note
  - It is important (also the goal) to understand the concepts and remember the distributions and characteristics of the sample proportion and the difference between two sample proportions.
  - The examples and exercises are designed to try to help you reach above goal

Homework

- Due date: Wed. Sep. 2nd
- Exercise 5.5.6 (page 153 , 10th Ed)
  - Please provide solutions with and without correction for continuity.
  - Use binomial distribution to check the result of Exercise
- Exercise 5.6.1 (page 156, 10th Ed)