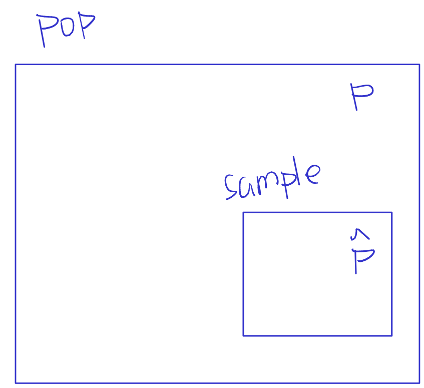
The Sampling Distribution of the Sample Proportion

$\hat{p}$

David Armstrong

UCI

Categorical Data



Binomial Distribution

Suppose we are interested in X, where X = the number of successes in n independent trials.

- Fixed number of trials
- Fixed probability of success
- Trials are independent
- $X \sim Binomial(n, p)$

  $\mathrm{E}(X) = np.$

  $\mathrm{VAR}(X) = np(1-p) = npq$

Normal Approximation to the Binomial

When you have a large SRS from a large population:

- Same assumptions as the Binomial Distribution
- Rule of Thumb

  Expected number of successes = $np \geq 10$

  Expected number of failures = $nq \geq 10$
- We would have to use MANY equations to find cumulative probabilities if we used the Binomial Distribution.
- We would only have to use 1 equation to find cumulative probabilities if we approximated the Binomial distribution with the Normal distribution.

Suppose we are interested in X:

- X = the number of successes in n independent trials.
- We say: "X follows an Approximately Normal distribution with mean of the number of successes equal to np and standard deviation of the number of successes equal to the square root of npq"
- $X \sim$ Approximately Normal$(\mu_X = np, \sigma_X = \sqrt{npq})$

  $\mathrm{E}(X) = np$.

  $\mathrm{VAR}(X) = np(1-p) = npq$
- To calculate probabilities we

  STANDARDIZE

  $$Z = \frac{X - \mu_X}{\sigma_X} = \frac{X - np}{\sqrt{npq}}$$

$p = 0.8$ $\quad n = 7$ $\qquad np = 7(0.8) = 5.6 < 10$

$q = 0.2$

Example: Eighty percent of all patrons at a local restaurant request the non-smoking section. Suppose we randomly select 7 customers.

Which statement below describe the correct distribution of X= number of patrons that request the non-smoking section?

$\boxed{\text{A. } X \sim B(7, 0.80)}$

~~B. $X \sim AN(5.6, 1.058)$~~

C. $X \sim N(5.6, 1.058)$

D. $X \sim B(5.6, 1.058)$

~~E. $X \sim AN(7, 0.80)$~~

What is the probability that at least 6 of the 7 customers selected will request the non-smoking section?

$$\begin{aligned} P(X \geq 6) &= 1 - P(X < 6) \\ &= 1 - P(X \leq 5) \\ &= 1 - \text{pbinom}(5, 7, 0.8) \\ &= 0.5767 \end{aligned}$$

Suppose we now take a larger random sample of 119 customers.

What are the mean and standard deviation of the number that request the non-smoking section?

$\mu = E(X) = np = 119(0.8) = 95.2$

$\sigma = \sqrt{npq} = \sqrt{119(0.8)(0.2)} = 4.3635$

Using the mean and standard deviation from above, what is the probability that at most 100 of the 119 customers selected will request the non-smoking section?

$X \sim AN(\mu_X = 95.2, \sigma_X = \sqrt{119(0.8)(0.2)})$

$$\begin{aligned} P\left(X \leq \frac{100}{119}\right) &= \text{pnorm}\left(\frac{100}{119}, 95.2, \sqrt{119(0.8)(0.2)}\right) \\ &= 5.2440 \times 10^{-104} \end{aligned}$$

Linear Transformation of the Normal Approximation

Let $X \sim$ Approximately Normal$\left(\mu_X = np, \sigma_X = \sqrt{npq}\right)$

Let $\hat{p} = \dfrac{X}{n}$.

- Calculate $E(\hat{p})$

$$E(\hat{p}) = E\left(\frac{x}{n}\right) = \frac{1}{n}E(x) = \frac{1}{n}(np) = p$$

$$E(\hat{p}) = p$$

- Calculate VAR$(\hat{p})$

$$\sigma_x^2 = npq$$

$$Var(\hat{p}) = Var\left(\frac{x}{n}\right) = \frac{1}{n^2} Var(x) = \frac{1}{n^2}(npq) = \frac{pq}{n}$$

- Calculate the standard deviation of $\hat{p}$

$$SD(\hat{p}) = \sqrt{Var} = \sqrt{\frac{pq}{n}}$$

The Distribution of a Sample Proportion

Suppose we are interested in $\hat{p} = \dfrac{X}{n}$ and you have a large SRS from a large population:

- Fixed number of trials
- Fixed probability of success
- Fixed probability of failure
- Trials are independent

- $\hat{p}$ = the sample proportion of successes in n independent trials.
- We say: "p-hat follows an Approximately Normal distribution with mean of the sample proportion of successes equal to p and standard deviation of the sample proportion of successes equal to the square root of pq over n"

Then: $\hat{p} \sim \text{Approximately Normal}\left(\mu_{\hat{p}} = p, \sigma_{\hat{p}} = \sqrt{\dfrac{pq}{n}}\right)$

- To calculate probabilities we **STANDARDIZE**

$$Z = \frac{\hat{p} - \mu_{\hat{p}}}{\sigma_{\hat{p}}} = \frac{\hat{p} - p}{\sqrt{\dfrac{pq}{n}}}$$

- Conservative Rule of Thumb

  Observed number of successes = $X \geq 10$

  Observed number of failures = $n - X \geq 10$

$p = 0.71 \quad \hat{p} = \frac{405}{600} = 0.675 \qquad x = 405 \quad n = 600$

Ex: The Associated Press reported that 71% of Americans ages 25 and older are overweight. A researcher wants to know whether the proportion of such individuals in his state that are overweight differs from the national
proportion. A random sample of 600 adults in his state results in 405 who are classified as overweight.

a. What is the sample proportion of overweight Americans?

$\hat{p} = 0.675$

b. Check and verify all of the assumptions and conditions.

Random $\rightarrow$ independent

n is constant

p is constant

$np = 600(0.71) = 426 \geq 10$

$nq = 600(0.29) = 174 \geq 10$

c. Describe the sampling distribution of the sample proportion for size 600 using the appropriate notation.

$$\hat{p} \sim AN\left(\mu_{\hat{p}} = p = 0.71,\ \sigma_{\hat{p}} = \sqrt{\frac{pq}{n}} = \sqrt{\frac{0.71(0.29)}{600}} = 0.0185\right)$$

d. Find the probability that at most 405 of the 600 sampled adults are classified overweight.

$$P\left(\hat{p} \leq \frac{405}{600}\right) = \text{pnorm}\left(0.675, 0.71, \sqrt{\frac{0.71(0.29)}{600}}\right)$$

$$= 0.0294$$

$p = 0.39 \quad \hat{p} = \frac{x}{n} = \frac{86}{200} \qquad x = 86 \quad n = 200$

Ex: According to the 2001 Youth Risk Behavior Surveillance by the Center for Disease Control and Prevention, 39% of the 10th-graders surveyed said that they watch three or more hours of television on a typical school day. Assume that this percentage is true for the current population of all 10th graders. Suppose in a random sample of 200 10th-graders, 86 watched three or more hours of television on a typical school day.

- Find the probability that 86 or more out of the 200 students watched three or more hours of television on a typical day.

$$\hat{p} \sim AN\left(\mu_{\hat{p}} = p = 0.39,\ \sigma_{\hat{p}} = \sqrt{\frac{pq}{n}} = \sqrt{\frac{(0.39)(0.61)}{200}}\right)$$

$$P\left(\hat{p} \geq \frac{86}{200}\right) = 1 - P\left(\hat{p} < \frac{86}{200}\right)$$

$$= 1 - \text{pnorm}\left(\frac{86}{200},\ 0.39,\ \sqrt{\frac{(0.39)(0.61)}{200}}\right)$$

$$= 0.1231$$

Ex: A nationwide survey by the University of Connecticut Center for Survey Research and Analysis found that 30% of men aged 18 to 29 had tattoos in 2002. Suppose this result holds true for the current population of all men in this age group.

$p = 0.30$

- Find the probability that in a random sample of 500 men aged 18 to 29, between 28.4% and 32.6% have tattoos.

$$\hat{p} \sim AN\left(\mu_{\hat{p}} = p = 0.30,\ \sigma_{\hat{p}} = \sqrt{\frac{pq}{n}} = \sqrt{\frac{(0.30)(0.70)}{500}} = 0.0205\right)$$

$$P(0.284 < \hat{p} < 0.326) = P(\hat{p} < 0.326) - P(\hat{p} < 0.284)$$

$$= \text{pnorm}\left(0.326,\ 0.30,\ \sqrt{\frac{(0.30)(0.70)}{500}}\right) - \text{pnorm}\left(0.284,\ 0.30,\ \sqrt{\frac{(0.30)(0.70)}{500}}\right)$$

$$= 0.6802$$