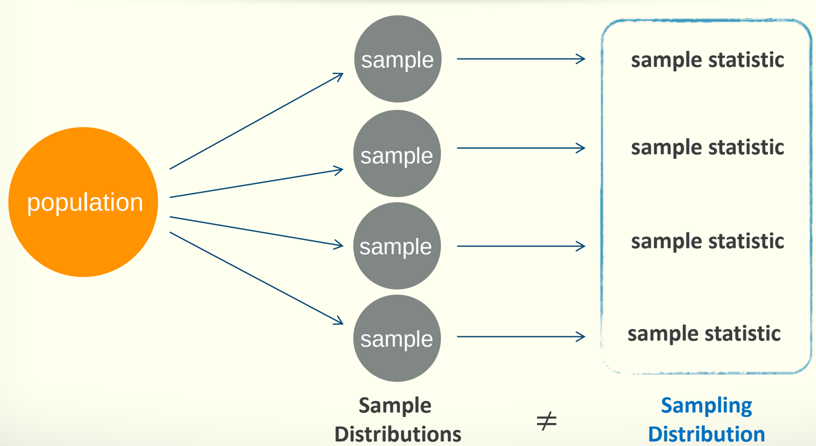
Statistical Inference

Inference for Categorical Variables

Behnam Bahrak
Spring 2020

Sample vs. Sampling Distributions



Example

	smoker or not categorical	% smoker numerical
Sample 1:	$x_{1,1}, x_{1,2}, \dots, x_{1,1000}$	$\hat{p}_1$
…	…	…
Sample 20:	$x_{20,1}, x_{20,2}, \dots, x_{20,1000}$	$\hat{p}_{20}$
…	…	…
Sample 50:	$x_{50,1}, x_{50,2}, \dots, x_{50,1000}$	$\hat{p}_{50}$

World Population
N = pop size
p

$$p = \frac{\text{\# of smokers in the world}}{N}$$

sampling distribution

$$mean(\hat{p}) \approx p$$

CLT for Proportions

- The distribution of sample proportions is nearly normal, centered at the population proportion, and with a standard error inversely proportional to the sample size.

$$\underbrace{\hat{p} \sim N}_{\text{shape}}\left(\underbrace{mean = p}_{\text{center}}, \underbrace{SE = \sqrt{\frac{p(1-p)}{n}}}_{\text{spread}}\right)$$

$$p = \frac{x_1 + x_2 + \cdots + x_N}{N} \quad \text{where} \quad x_i = \begin{cases} 1 & \text{if } i \text{ is a smoker} \\ 0 & \text{otherwise} \end{cases}$$

Conditions for the CLT

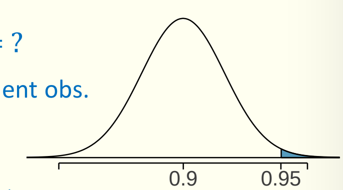
1. Independence: Sampled observations must be independent.
   - random sampling/assignment
   - if sampling without replacement, $n < 10\%$ of the population.

2. Sample size/skew: There should be at least 10 successes and 10 failures in the sample:
   - $np \geq 10$ and $n(1-p) \geq 10$
   - If p unknown, use $\hat{p}$

Example

- 90% of all plants species are classified as flowering plants. If you were to randomly sample 200 plants from the list of all known plant species, what is the probability that at least 95% of plants in your sample will be flowering plants?

$p = 0.9$, $n = 200$, $P\{\hat{p} > 0.95\} = ?$



1. random sample & <10% of all plants independent obs.
2. $200 \times 0.90 = 180$ and $200 \times 0.10 = 20$

$$\hat{p} \sim N\left(mean = 0.9 , SE = \sqrt{\frac{0.9 \times 0.1}{200}} \approx 0.0212\right)$$

$$P\{\hat{p} > 0.95\} = P\left\{Z > \frac{0.95 - 0.9}{0.0212}\right\} = P\{Z > 2.36\} \approx 0.0091$$

Example

- 90% of all plants species are classified as flowering plants. If you were to randomly sample 200 plants from the list of all known plant species, what is the probability that at least 95% of plants in your sample will be flowering plants?

$p = 0.9\ ,\qquad n = 200\ ,\qquad P\{\hat{p} > 0.95\} = ?$

Using Binomial distribution:

$$200 \times 0.95 = 190$$

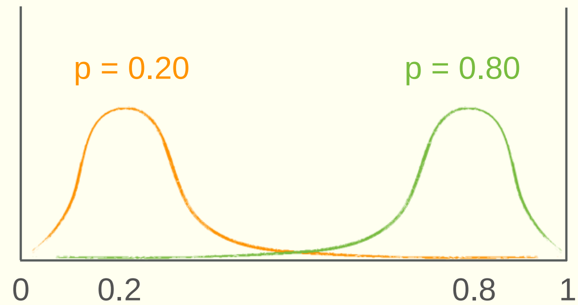
R

```
> sum(dbinom(190:200, 200, 0.90))
[1] 0.00807125
```


Success-Failure Condition

- There should be at least 10 successes and 10 failures in the sample.
- What if the success-failure condition is not met:
  - the center of the sampling distribution will still be around the true population proportion
  - the spread of the sampling distribution can still be approximated using the same formula for the standard error
  - the shape of the distribution will depend on whether the true population proportion is closer to 0 or closer to 1

Shape of the Sampling Distribution



Confidence interval for a proportion

- Two scientists want to know if a certain drug is effective against high blood pressure. The first scientist wants to give the drug to 1000 people with high blood pressure and see how many of them experience lower blood pressure levels. The second scientist wants to give the drug to 500 people with high blood pressure, and not give the drug to another 500 people with high blood pressure, and see how many in both groups experience lower blood pressure levels. Which is the better way to test this drug?

(a) All 1000 get the drug
(b) 500 get the drug, 500 don't

experimental design	
bad intuition	99
good intuition	571
total	**670**

Confidence interval for a proportion

➢ What percent of Americans have good intuition about experimental design?

parameter of interest	point estimate
Percentage of **all** Americans who have good intuition about experimental design.	Percentage of **sampled** Americans who have good intuition about experimental design.
p	$\hat{p}$
	571 / 670 ≈ 0.85

Estimating a Proportion

point estimate ± margin of error

$$\hat{p} \pm z^{\star} SE_{\hat{p}}$$

Standard error for a proportion, for calculating a confidence interval:

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

Example

- The GSS found that 571 out of 670 (~85%) of Americans answered the question on experimental design correctly. Estimate (using a 95% confidence interval) the proportion of all Americans who have good intuition about experimental design?

1. independence: 670 < 10% of Americans, and GSS samples randomly
Whether one American in the sample has good intuition about experimental design is independent of another.

2. sample size / skew: 571 successes, 670 - 571 = 99 failures
Since the success-failure condition is met, we can assume that the sampling distribution of the proportion is nearly normal.

Example

$$\hat{p} \pm z^* SE = 0.85 \pm 1.96 \sqrt{\frac{0.85 \times 0.15}{670}}$$

$$= 0.85 \pm 1.96 \times 0.0138$$

$$= 0.85 \pm 0.027$$

$$= (0.823, 0.877)$$

We are 95% confident that 82.3% to 87.7% of all Americans have good intuition about experimental design.

Example

- The margin of error for the previous confidence interval was 2.7%. If, for a new confidence interval based on a new sample, we wanted to reduce the margin of error to 1% while keeping the confidence level the same, at least how many respondents should we sample?

$$ME = 0.01 = 1.96\sqrt{\frac{0.85 \times 0.15}{n}}$$

$$n = \frac{1.96^2 \times 0.85 \times 0.15}{0.01^2} = 4898.04$$

→ at least 4899

Calculating the required sample size for desired ME

$$ME = z^{\star}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

- If there is a previous study that we can rely on for the value of $\hat{p}$ use that in the calculation of the required sample size
- If not, use $\hat{p} = 0.5$
  - if you don't know any better, 50-50 is a good guess
  - gives the most conservative estimate – highest possible sample size

Hypothesis test for a proportion

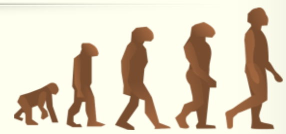
1. Set the hypotheses: $H_0 : p = null\ value$
   $H_A : p < \ or\ > \ or\ \neq\ null\ value$
2. Calculate the point estimate: $\hat{p}$
3. Check conditions:
   a) Independence: Sampled observations must be independent (random sample/assignment & if sampling without replacement, $n < 10\%$ of population)
   b) Sample size/skew: $np \geq 10$ and $n(1-p) \geq 10$
4. Draw sampling distribution, shade p-value, calculate test statistic: $Z = \frac{\hat{p} - p}{SE}, \quad SE = \sqrt{\frac{p(1-p)}{n}}$
5. Make a decision, and interpret it in context of the research question:
   - If p-value $< \alpha$, reject H_0; the data provide convincing evidence for H_A.
   - If p-value $> \alpha$, fail to reject H_0 the data do not provide convincing evidence for H_A.

$\hat{p}$ vs. p

	Confidence Interval	Hypothesis Test
Success-Failure Condition	$n\hat{p} \geq 10$ $n(1-\hat{p}) \geq 10$	$np \geq 10$ $n(1-p) \geq 10$
Standard Error	$SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$	$SE = \sqrt{\frac{p(1-p)}{n}}$

Example

- A 2013 Pew Research poll found that 60% of 1,983 randomly sampled American adults believe in evolution. Does this provide convincing evidence that majority of Americans believe in evolution?



$H_0: p = 0.5$ $\quad\quad \hat{p} = 0.6 \quad\quad n = 1983$

$H_A: p > 0.5$

1. independence: 1983 < 10% of Americans & random sample

Whether one American in the sample believes in evolution is independent of another.

2. sample size / skew: $1983 \times 0.5 = 991.5 > 10$

S-F condition met → nearly normal sampling distribution

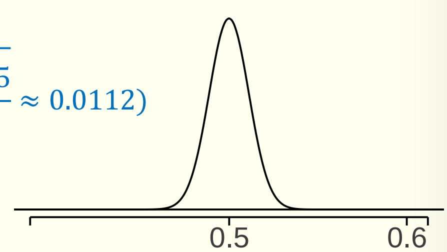
Example

$H_0: p = 0.5$ $\quad\quad \hat{p} = 0.6 \quad\quad n = 1983$

$H_A: p > 0.5$

$$\hat{p} \sim N\left(mean = 0.5\,, SE = \sqrt{\frac{0.5 \times 0.5}{1983}} \approx 0.0112\right)$$

$$Z = \frac{0.6 - 0.5}{0.0112} \approx 8.92$$



p-value $= P(Z > 8.92) \approx 0$ → Reject H_0