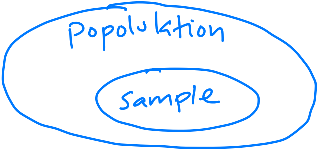
Lecture 6:
Sampling Distributions
STAT 310, Spring 2021

- A **parameter** is a numerical characteristic of the population (fixed number that is usually unknown).
- A **statistic** is a numerical characteristic of the sample (varies depending on sample).
- The statistic is also referred to as a **point estimate**, since it is our best guess at the value of a population parameter.



$$
\begin{tikzcd}
\text{population} \arrow[r] \arrow[d] & \text{sample} \arrow[d] \\
\text{parameter} & \text{statistic} \arrow[l, "\text{estimate}"']
\end{tikzcd}
$$

Notation for the proportion:

- p: population proportion
- $\hat{p}$: sample proportion

Notation for the mean:

- μ: population mean
- $\bar{x}$: sample mean

The sample size is denoted as n

Example

A recent Gallup poll found that 66% of Americans are dissatisfied with how the COVID-19 vaccination process is going in the U.S. The survey results were based on random sample of 4,098 adults.[1]

(a) What is the sample proportion?

$\hat{p} = 0.66$

(b) What is the sample size?

$n = 4,098$

(c) Describe the population proportion?

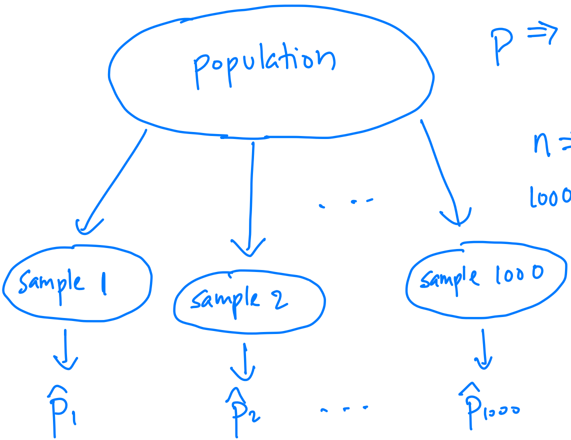
p : population proportion of Americans that are dissatisfied with vaccine rollout

(d) Suppose another poll is conducted with a different random sample of adults. Would you expect the sample proportion to be the same, or slightly different? slightly different

[1] https://news.gallup.com/poll/329552/two-thirds-americans-not-satisfied-vaccine-rollout.aspx

- **Sampling Error** refers to how much a statistic, such as the sample proportion, will vary from one random sample to the next.
- For example, the Gallup poll reported a sampling error of ± 2 percentage points. This means that the population proportion of Americans that are dissatisfied with the vaccine rollout is likely between 64% and 68%.

A **sampling distribution** is the distribution of a statistic when repeatedly taking random samples from a population.

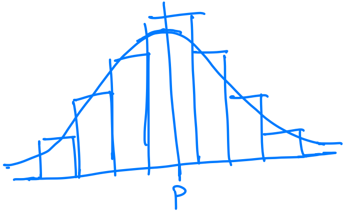


$p \Rightarrow$ population proportion

$n \Rightarrow$ sample size

$1000 \Rightarrow$ # of samples

Sampling distribution is histogram of 1000 sample proportions



- When n is large, histogram looks normal and centered around population proportion p
- The standard deviation of the 1000 sample proportions is called the <u>standard error</u> (SE). It measures the variability, or spread, of the estimates from sample to sample.

- In real-world applications, we never actually observe the sampling distribution, since we usually take a single random sample.
- However, it is useful to always think of a statistic, such as the sample proportion, as coming from such a hypothetical distribution.
- The concept of a sampling distribution is very important when trying to quantify sampling error.

Central Limit Theorem (CLT)

The sampling distribution for $\hat{p}$ follows an approximate normal distribution centered around the population proportion p, and with standard error $\sqrt{p(1-p)/n}$.

$$\hat{p} \sim N\left(\underbrace{p}_{\text{mean}}, \underbrace{\sqrt{\frac{p(1-p)}{n}}}_{\text{standard error (SE)}}\right)$$

Can verify CLT with computer simulation

Conditions for CLT

The following conditions should be met to apply the CLT:

- The data come from a simple random sample. This is called the **independence condition** since it implies that the individuals or cases in the data are unrelated.
- $np \geq 10$ and $n(1-p) \geq 10$. This is sometimes called the **success-failure condition** since np can be interpreted as the expected number of successes and $n(1-p)$ the expected number of failures.

Example

Suppose that the population proportion of Americans who support the expansion of solar energy is $p = 0.88$, and $n = 1000$ Americans are randomly sampled.

(a) What is the mean, or center, of the sampling distribution for $\hat{p}$?

$$\text{mean} = p = 0.88$$

(b) What is the standard error of the sampling distribution for $\hat{p}$?

$$SE = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.88(0.12)}{1000}} = 0.01$$

(c) What distribution does $\hat{p}$ follow?

$$\hat{p} \sim N\left(p, \sqrt{\frac{p(1-p)}{n}}\right) = N(0.88, 0.01)$$

(d) Are the conditions for the CLT satisfied? yes

- random sample ✓
- $np = 1000(0.88) = 880 \geq 10$ $\quad n(1-p) = 1000(0.12) = 120 \geq 10$

Example

$$z = \frac{\hat{p} - p}{SE}$$

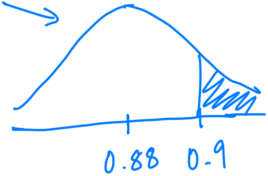
Suppose that population proportion of Americans who support the expansion of solar energy is $p = 0.88$, and $n = 1000$ Americans are randomly sampled. What is the probability that the sample proportion $\hat{p}$ will be greater than 0.9?

$$\text{By CLT } \hat{p} \sim N\left(p, \sqrt{\frac{p(1-p)}{n}}\right) = N(0.88, 0.01)$$

SE

mean

sampling dist for $\hat{p}$



$$z = \frac{0.9 - 0.88}{0.01} = 2$$

$$P(z > 2) = 1 - P(z < 2) = 1 - \text{pnorm}(2) = \boxed{0.023}$$