# Week 4 Day 5

Stat140-04

## How Important is Regular Exercise?

Social scientists are interested in knowing the the proportion of American adults who think exercise is an important part of daily life. In a recent poll of 1000 American adults, the number saying that exercise is an important part of daily life was 753.

### Part I: paramaeter v.s. statistic

(a) What is the population of interest and population parameter?

(b) Identify the observational units and variable for your sample. Is the variable categorical or numerical?

Observational units:

Variable: Type:

(c) Based on your data, do you think 0.75 is a plausible value for the probability an American thinks exercise is important? How are you deciding?

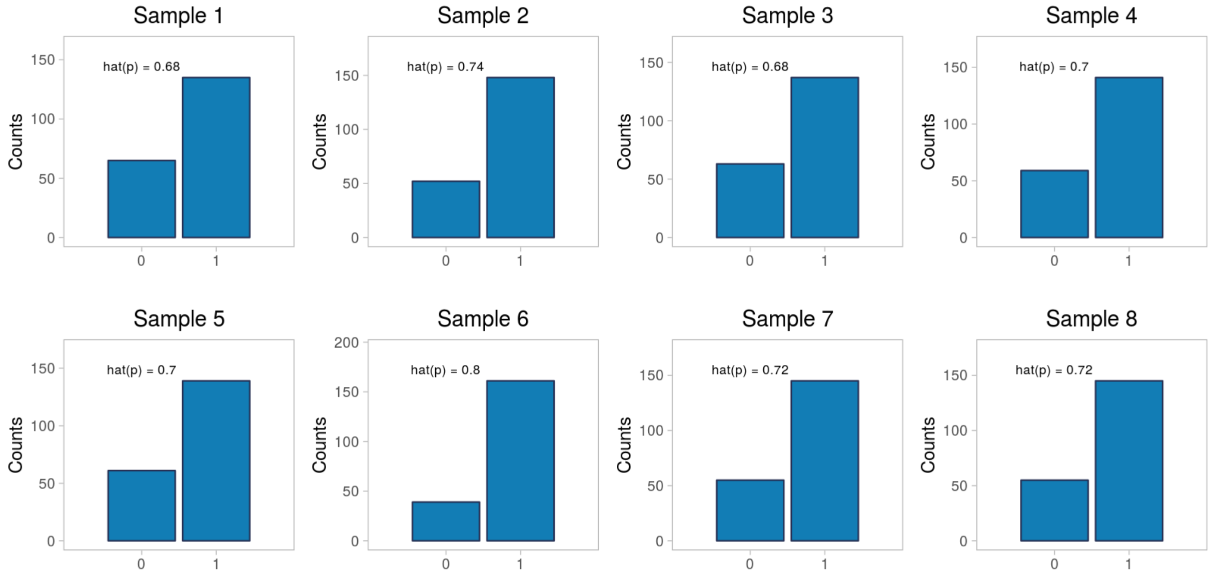
**Part II: Sample variability**

(d) Open the Shiny app (https://openintro.shinyapps.io/CLT_prop/) and set

- `Sample size (n)` = 200
- `Population Proportion (p)` = 0.75
- `Number of samples` = 1000

Click the `Samples` tab. The applet will randomly select 1000 (`Number of samples`) samples of 200 (`Sample size`) amercan, sort them, and report the proportion (`hat(p)`) of people who think exercise is important.

Note that we define `1` to be an American who thinks exercise is important and `0` to be an American who doesn't think exercise is important.



Did you get the same proportion (`hat(p)`) of people who think exercise is important each time? Why is it so?

**Part III: Sampling distribution**

(e) Click the `Sampling distribution` tab. Describe the behavior (shape, center, spread) of the resulting sampling distribution. Where does your observed sample statistics (0.753) fall in this distribution?

(f) If we had instead taken 1000 samples of size $n = 1000$, how do you think the distribution of the sample proportions would compare to the distribution where $n = 200$? Explain.

(g) Change the `Sample size` to be 1000 and display the sampling distribution. Is this what you expected in (f)?

(h) Now let's suppose the true population parameter is 0.5, meaning 50% of American adults think exercise is an important part of their daily life. How do you anticipate this will change the distribution of sample proportions? Explain.

(i) Create and describe the distribution of sample proportions in this case. You should set

- `Sample size (n)` = 200
- `Population Proportion (p)` = 0.50
- `Number of samples` = 1000

What is the primary difference in how the distribution of sample proportions has changed with this change in the `Pupulation proprotion`?

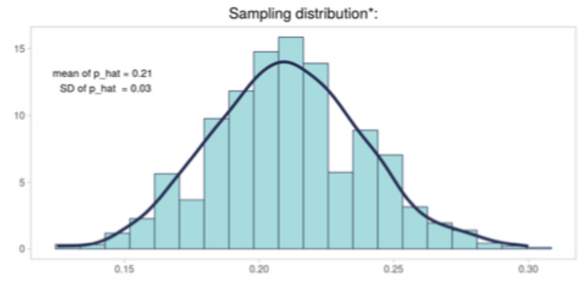
(j) Suppose we think the true population parameter is 0.2, meaning 20% of American adults think exercise is an important part of their daily life. Repeat the previous question (i) with

- `Sample size (n)` = 200
- `Population Proportion (p)` = 0.20
- `Number of samples` = 1000

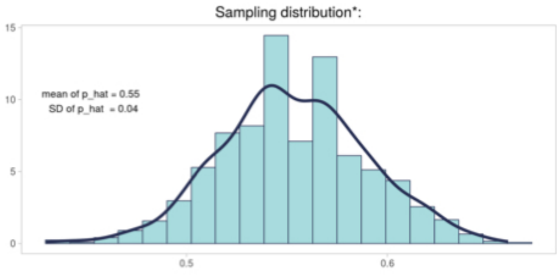
## Part IV: Compare and Contrast

(1) Below are simulation results for the four scenarios you have examined:

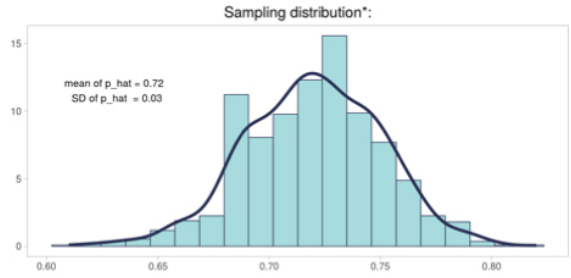
n=200, p = 0.2



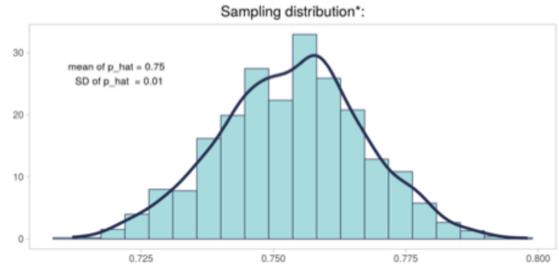
n=200, p = 0.5



n=200, p = 0.75



n=1000, p = 0.75



Summarize the differences you have noted between these distributions and also what characteristic(s) they all have in common.

Differences:

Similarities:

**Part IV: Central limit theorem**

For now, we will focus on using this mathematical model as an approximation to the distribution of sample proportions. The central limit theorem tells us that the theoretical mean of the distribution of sample proportions will be equal to the `Population Proportion` and the theoretical standard deviation (also known as the standard error) of the sampling distribution equals $SD(\hat{p}) = \frac{\sqrt{p(1-p)}}{n}$.

(m) Calculate the theoretical mean and standard deviation of the sampling distribution of sample proportions for each of these three cases.

| | Theoretical mean of $\hat{p}$ | Simulated mean of $\hat{p}$ |
|---|---|---|
| n = 200, p = 0.2 | | |
| n = 200, p = 0.5 | | |
| n = 200, p = 0.75 | | |
| n = 1000, p = 0.75 | | |

| | Theoretical standard deviation of $\hat{p}$ | Simulated standard deviation of $\hat{p}$ |
|---|---|---|
| n = 200, p = 0.2 | | |
| n = 200, p = 0.5 | | |
| n = 200, p = 0.75 | | |
| n = 1000, p = 0.75 | | |

How do the theoretical means and standard deviations compare to the simulated values (see the graphs in (l)) and to each other?

(n) Now let's set

- `Sample size (n)` $= 5$
- `Population Proportion (p)` $= 0.75$
- `Number of samples` $= 1000$

Use the central limit theorem to predict how the distribution of sample proportions will behave (shape, center, spread).

- Shape:
- Center:
- Spread:

(o) Use the applet to check your predictions. Discuss your observations.

Discussion: It is very important to keep in mind that the Central Limit Theorem does not come for free, i.e., the normal model is not always a valid approximation for the distribution of sample proportions. Whether it is valid will be determined by a combination of the sample size $n$ (larger samples result in more symmetric distributions) and the value of $p$ (values closer to 0 or 1 result in less symmetric distributions). A common guideline is $n \times p > 10$ and $n \times (1 - p) > 10$.

(p) Explain how these guidelines are consistent with your observations above.

**Part V: How unusual is an observation?**

We usually begin to think an observation is unusual when it lies more than two standard deviations above or below the mean of the distribution.

(q) Use the theoretical mean and SD values of $n = 200, p = 0.75$ from (m) to construct an interval of $\hat{p}$ that is two standard deviations within the mean.

(r) Write a one-sentence interpretation of the intervals

Roughly 95% of . . .