Valersa, Jeffrey

```
Stat 145, Tue 19-Oct-2021 -- Tue 19-Oct-2021
Biostatistics
Spring 2021
```

Binary categorical vars?
Is Democrat?
$\bar{X}$ same as $\hat{p}$.

```
------------------------------
Tuesday, October 19th 2021
------------------------------
Due::   WW NormalDists due at 11 pm

------------------------------
Tuesday, October 19th 2021
------------------------------
Wk 8, Tu
Topic:: Inference for proportions
Read::  Lock5 6.1-6.3
```

$$\bar{X} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$$

Central Limit Theorem 5.2

population

In summary, here is the take-away from the **Central Limit Theorem**.

Suppose you have a random sample of size n that is either

- i.i.d., or
- an SRS, with the sample size n being no more than 10% of the size of the population.

In the case that

1. the variable under consideration is quantitative, having population mean μ and standard deviation σ, then the sampling distribution for the sample mean $\overline{X}$ is approximately $\mathsf{Norm}(\mu, \sigma/\sqrt{n})$ for n large enough.

2. the variable under consideration is binary categorical, having population proportion p, then the sampling distribution for the sample proportion $\widehat{p}$ is approximately $\mathsf{Norm}(p, \sqrt{p(1-p)/n})$ for n large enough.

Since

- null distributions
- randomization distributions
- bootstrap distributions

are all specialized versions of sampling distributions, then so long as the sample statistic in question is the sample's *mean* $\overline{X}$ or the sample *proportion* $\widehat{p}$, we can expect the CLT to apply to these as well.

```
Explorations using apps at
  https://shiny.calvin.edu:3838/scofield/samplingDists/
  https://shiny.calvin.edu:3838/scofield/cltProportions/

or, for means, use script  samplingDistOfSampleMeanExperiments.R
  require(mosaic)
  require(gridExtra)

  # Create a population
  mypop <- 50 - rexp(10000, rate=.15)            # left-skewed
  #mypop <- rgamma(10000, shape=1.6, rate=.1)    # right-skewed
  #mypop <- rnorm(10000, mean=25, sd=6)          # normal
  print(favstats(~mypop))

  # Simulate the sampling distribution for the sample mean
  sampleSize = 20
  manyMeans <- do(5000) * mean(~sample(mypop, sampleSize, replace=TRUE))
  print(favstats(~mean, data=manyMeans))

  p1 <- gf_density(~mypop) %>% gf_refine(scale_x_continuous(limits=c(0,55)))
  p2 <- gf_density(~mean, data=manyMeans) %>%
    gf_refine(scale_x_continuous(limits=c(0,55)))
  grid.arrange(p1, p2, nrow=2)
```

Code for experiments

Chapter 6

6.1 – 6.3 : univariate, binary categorical data (single-proportions)
6.4 – 6.6 : univariate, quantitative data (single means)
6.7 – 6.9 : bivariate binary categorical responses for 2 groups (2-proportions)

6.10 – 6.12 : bivariate data, quantitative response, 2 groups (2 means)

Single proportions: tasks
1. CI for p
2. Hyp. test on null / alt. hypotheses about p

Recall CI in the past

$$\hat{p} \pm z^* \left(SE_{\hat{p}} \right)$$

$\hat{p}$: sample est. of p

Note that

for 95% conf., $z^* = 1.96$
90% conf., $z^* = 1.645$
99% conf., $z^* = 2.576$

Ex. Suppose we ask 450 students if they are left-handed, 47 say yes.

So $\hat{p} = \dfrac{47}{450} = 0.104$

Q: Can I assume $\hat{p}$ has a nearly normal dist?

$$n\hat{p} = 450\left(\frac{47}{450}\right) = 47 \geq 10$$

$$n(1-\hat{p}) = 450\left(\frac{450-47}{450}\right) = 403 \geq 10$$

A: Yes. So $\hat{p} \sim \text{Norm}\left(p, \sqrt{\frac{p(1-p)}{\sqrt{n}}}\right)$

Use $\hat{p}$: $SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ $\quad q$

$= \sqrt{\frac{(.104)(0.896)}{450}}$

$= 0.0144$

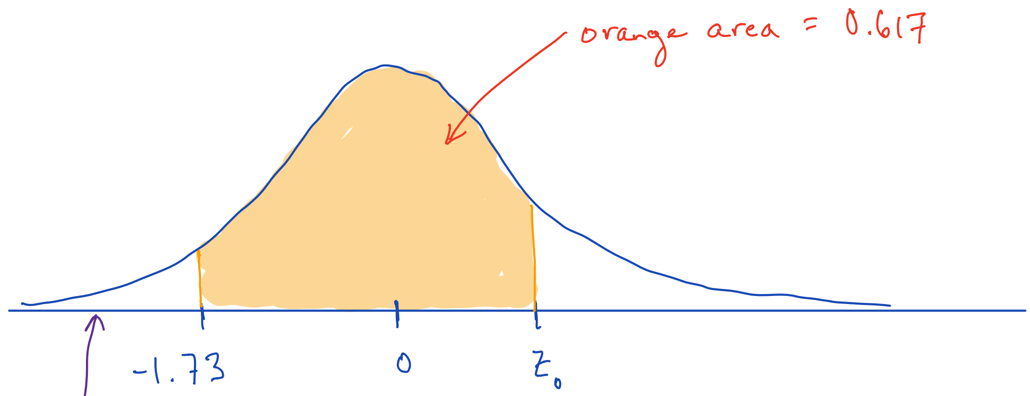
So, a 95% CI for p

$$\underset{\hat{p}}{\underset{\uparrow}{0.104}} \pm \underset{z^*}{\underset{\uparrow}{(1.96)}}\underset{SE_{\hat{p}}}{\underset{\uparrow}{(0.0144)}}$$

A problem like 8(f) on the current WebWork set might be:

"Determine z_0 so that $Pr(-1.73 < Z < z_0) = 0.617$."

I have depicted a standard normal distribution below with -1.73, z_0, and 0.617 displayed.



Use pnorm(-1.73), which behaves the same as pnorm(-1.73, mean = 0, sd = 1), to find the area A to the left of 1.73.

Then use qnorm(A + 0.617) to find z_0.

```
Chapter 6 overview
 - Scenarios are all ones we have discussed
    univariate (one population)
      proportion arising from binary categorical variable
      mean arising from quantitative variable
    2 populations
      difference of proportions
      difference of means investigated using
        two independent samples
        matched pairs

 - Deferred to later chapter: 2 quant vars

 - Can see Chapter 6 as something of a history lesson

 - Relies entirely on facts from Central Limit Theorem

Sections 1-3: single proportion

Confidence interval construction
 - review how done using bootstrapping (Ch. 3)
 - refining the z*-value
    in past, stats students used tables of Z-scores
      see https://www.math.arizona.edu/~jwatkins/normal-table.pdf
    compare with pnorm(), qnorm() calculations
 - formula for SE

Practice:
 - obtaining critical z* values for
    96% confidence
    90% confidence
    99% confidence
 - doing inference (CI and hypothesis testing) with datasets
    1. in 119 games of rock-paper-scissors, player did rock 66 times
    2. in 70 out of 120 soccer games, the home team won
    3. suppose that 42% of people have O+ blood.  sample shows 65 out of 192
```