Statistical Inference with Confidence Intervals

Statistical Inference gives us the ability to answer specific questions about a population.

A statistic is a summary of a sample: $\bar{X}$, $\hat{p}$, s

Statistics represent parameters which are aspects of the population: μ, p, σ

Population (for some variable)
- Some mean μ
- Some std dev σ

} Presumed unknown

Sample
$X_1, X_2, \ldots, X_n$
- Some mean $\bar{X}$ (or $\hat{p}$)
- Some std dev s (or $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ ← estimation of s)

} We know the dist of $\bar{X}$, sort of

Central Limit Theorem (CLT)

Barring some conditions

Ex: n should be large ($n \geq 50$)

Need a pop without very extreme values

$\bar{X}$ is approximately distributed as

$$\bar{X} \sim \text{Normal}\left(\text{mean} = \mu, \text{stddev} = \frac{\sigma}{\sqrt{n}}\right)$$

(μ: pop mean; σ: pop std dev)

Confidence Interval: What is the value of the population mean?

For UAW: sent a survey asking members "Are you rent burdened?" → Spending > 30% of income on rent

- UAW: Union that represents our TAs, grader and academic student employees
- $n = 1{,}300$ $\hat{p} = 80\% = .8$

3 distributions at play

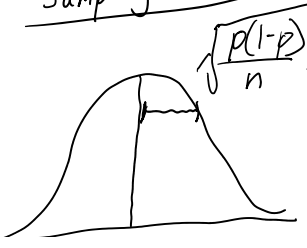
Population Distribution → dist of entire pop

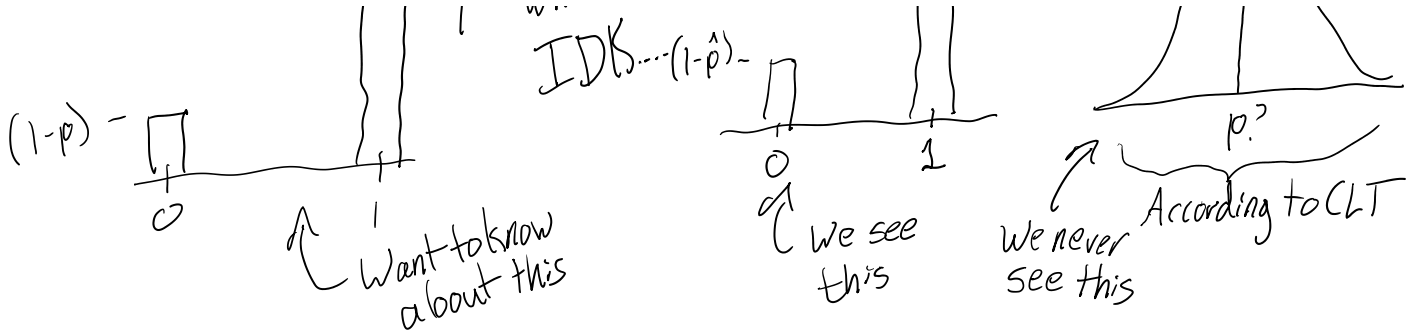
- p
- what is p? IDK... $(1-\hat{p})$

Sample distribution → dist of $X_1, X_2, \ldots, X_n$

- $\hat{p} = .8$

Sampling Dist → dist of $\hat{p}$





For a proportion the CLT requires

n large ($n \geq 50$) ✓

$n \times p$ $n \times (1-p) \geq 10$

estimate

$n \times \hat{p}$ $n \times (1-\hat{p}) \geq 10$ ✓

$1300 \times .8$, $1300 \times .2$

$\hat{p}$ follows a dist $N(p, \sqrt{\frac{p(1-p)}{n}})$

Estimate p we use $\hat{p}$, but $\hat{p}$ is wrong

We know $\hat{p}$ should be close

Thanks to CLT, we know how it should deviate from p.

Confidence Interval

Best estimate ± Range based on the variability of the sampling dist (via CLT)

$\hat{p} \pm Z \times$ Std Error of $\hat{p}$

↑ Based on standard normal distribution

Std error is like a standard deviation but for a statistic

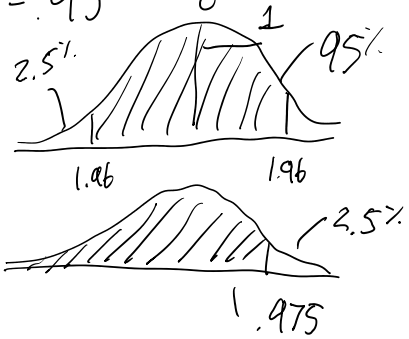
$\hat{p} \pm Z \cdot \sqrt{\frac{p(1-p)}{n}} \approx \hat{p} \pm Z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

???

Z is based on the level of confidence

I want a 95% CI, so I want a Z such that $P(-Z < z < Z) = .95$

$Z = 1.96$



$\hat{p} = .8$ $n = 1300$

$\hat{p} \pm Z_{.975} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

$.8 \pm 1.96 \cdot \sqrt{\frac{.8 \cdot (1-.8)}{1300}}$ ← std error

$1.27\% = .0127$

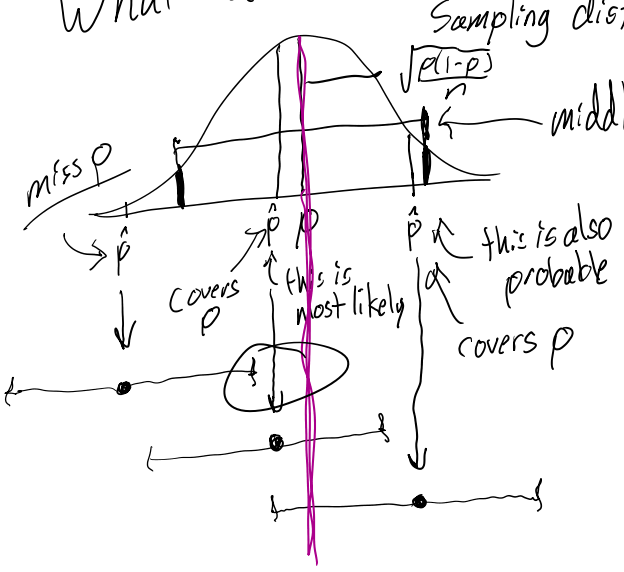
1.271% = .01271 ← 1.975

Margin of error ≈ 2.491%

95% Confidence Interval of (77.51%, 82.49%)

We are 95% confident that the proportion UAW members at UW facing rent burden is between 77.51% and 82.49%.

What does this mean?



$z_{.975}\sqrt{\frac{p(1-p)}{n}}$ ≈ MoE

A specific CI may or may not cover the true mean.

We never know which situation we're in... why is this useful

If $\hat{p}$ comes from the middle 95% of the sampling dist } happens 95% of the time
p will be in $\hat{p} \pm z_{.975}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

If $\hat{p}$ comes from outside the middle 95%; } happens 5% of the time.
p is not in $\hat{p} \pm z_{.975}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

When we say "We are 95% confident" we're saying

"The process we used to construct this interval should cover true mean 95% of the time".

I don't know μ (or p) specifically, but the process I used should produce an interval that covers it 95% of the time

I can

should produce an interval ... time.

Does not mean there a 95% chance p is between (77.5%; 82.5%)