STAT 311: Confidence Intervals for Proportions

Y. Samuel Wang

Summer 2016

Logistics

- Homework has been posted

What proportion of WA voters support Hillary Clinton for president?

Recent poll showed that 49% of voters would vote for Clinton, 37% for Trump and the remainder were not sure.

- There is some true % of voters (the parameter), even though we don't know precisely what it is
- Gather a sample of WA voters and ask them whether or not they support Hillary Clinton
  - Record 1 if they answer yes, 0 if they answer no
  - Each individual voter is a Bernoulli random variable
  - The count of supporters is a Binomial random variable (under what assumptions)
  - The proportion of supporters is the average of all the Bernoulli random variables

Sampling distribution of proportions

If X_i is the outcome of any given individual

$$S_n = X_1 + X_2 \ldots X_n$$

and S_n is the total count, then

$$E(S_n) = np$$

and

$$Var(S_n) = np(1-p)$$

Proportions

Suppose we are interested in the proportion

$$\hat{p} = \frac{1}{n} S_n$$

so by the rules of expectation and variance

$$E(\hat{p}) = \frac{1}{n} E(S_n) = p$$

and

$$Var(\hat{p}) = \frac{1}{n^2} Var(S_n) = p(1-p)/n$$

Proportions

Furthermore, since the proportion is also an average of a bunch of Bernoulli random variables, we can apply the Central Limit Theorem. Putting everything together implies that as n becomes large, we have-

$$\hat{p} \sim \mathcal{N}(p, p(1-p)/n)$$

Proportions

Furthermore, since the proportion is also an average of a bunch of Bernoulli random variables, we can apply the Central Limit Theorem. Putting everything together implies that as n becomes large, we have-

$$\hat{p} \sim \mathcal{N}(p, p(1-p)/n)$$

So we can use the framework we've built up around the Normal distribution to help describe the sampling distribution of $\hat{p}$

Normal Probabilities

In particular-

$$\begin{aligned} .95 &= P(p - 1.96\sigma \leq \hat{p} \leq p + 1.96\sigma) \\ &= P(1.96\sigma \leq \hat{p} - p \leq 1.96\sigma) \\ &= P(-\hat{p} - 1.96\sigma \leq -p \leq -\hat{p} + 1.96\sigma) \\ &= P(\hat{p} + 1.96\sigma \geq p \geq \hat{p} - 1.96\sigma) \end{aligned} \tag{1}$$

Normal Probabilities

In particular-

$$\begin{aligned} .95 &= P(p - 1.96\sigma \leq \hat{p} \leq p + 1.96\sigma) \\ &= P(1.96\sigma \leq \hat{p} - p \leq 1.96\sigma) \\ &= P(-\hat{p} - 1.96\sigma \leq -p \leq -\hat{p} + 1.96\sigma) \\ &= P(\hat{p} + 1.96\sigma \geq p \geq \hat{p} - 1.96\sigma) \end{aligned} \tag{1}$$

So there is a 95% chance that when I form sample a $\hat{p}$ and form the interval $(\hat{p} - 1.96\sigma, \hat{p} + 1.96\sigma)$, it will contain the true parameter p

Confidence intervals

So there is a 95% chance that when I form sample a $\hat{p}$ and form the interval $(\hat{p} - 1.96\sigma, \hat{p} + 1.96\sigma)$, it will contain the true parameter p

Confidence intervals

So there is a 95% chance that when I form sample a $\hat{p}$ and form the interval $(\hat{p} - 1.96\sigma, \hat{p} + 1.96\sigma)$, it will contain the true parameter p

... But I don't know what p is, so I can't calculate σ. σ is the standard deviation, but when we use $\hat{p}$ to estimate σ we use the term **standard error**. This is sometimes denoted by $se(\hat{p})$ and $1.96se(\hat{p})$ is sometimes called the **margin of error**.

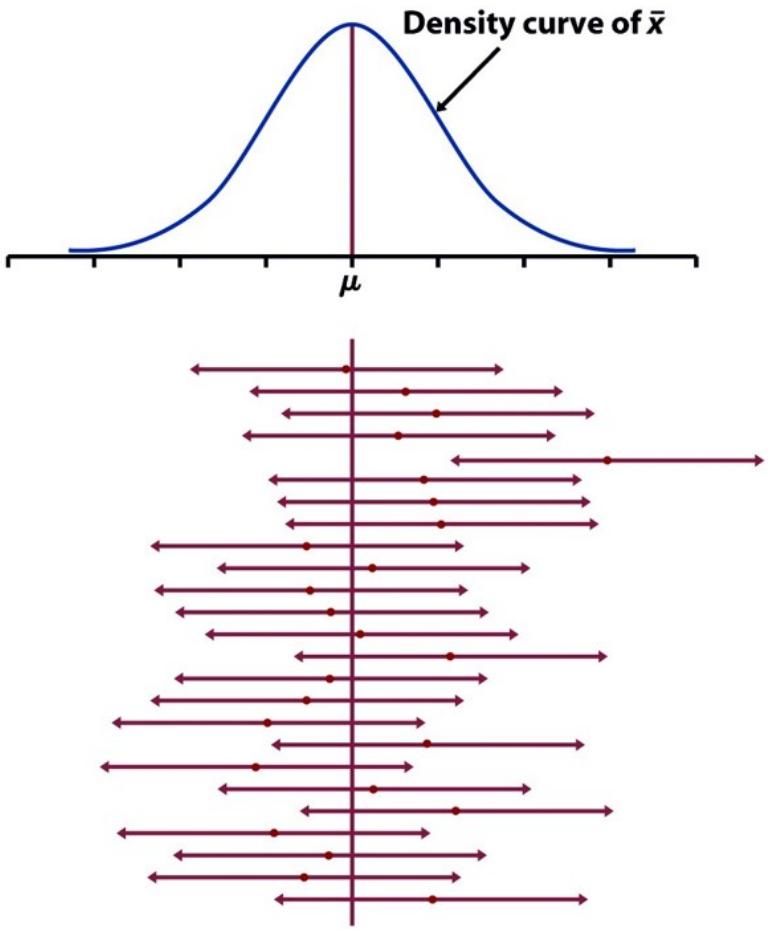
Confidence Interval

Confidence statement: I am 95% confident that the true probability lies between $\hat{p} - \sqrt{\hat{p}(1-\hat{p})/n}$ and $\hat{p} + \sqrt{\hat{p}(1-\hat{p})/n}$

Formal Meaning: Approximately 95% of the times I repeat this procedure, the confidence interval I produce will contain the true parameter

Ways to interpret: Roughly 95% of the $\hat{p}$ will be less than the margin of error away from the true parameter

Confidence Interval



Confidence Interval

- We created 95% confidence intervals, but in general, we can make intervals for any confidence level. We would use a different multiplier (not 1.96).
- We need a large enough sample size for the CLT to hold. A good rule of thumb is for the number of success and failures to both be greater than 10.
- The individuals need to be randomly selected from the population (no undercoverage or biased sampling)

Two sample proportions

Suppose, I am now interested in the difference between the proportion of people in WA who support Hillary Clinton and the proportion of people in TX who support Hillary Clinton

$$p_{wa} - p_{tx}$$

Two sample proportions

Suppose, I am now interested in the difference between the proportion of people in WA who support Hillary Clinton and the proportion of people in TX who support Hillary Clinton

$$p_{wa} - p_{tx}$$

I can still estimate that quantity with

$$\hat{p}_{wa} - \hat{p}_{tx}$$

and form confidence intervals like before, but what is the distribution of that statistic

Two sample proportions

Assuming that $\hat{p}_{wa}$ and $\hat{p}_{tx}$ are independent, then we can simply use the rules of expectation and variance

$$E(\hat{p}_{wa} - \hat{p}_{tx}) = p_{wa} - p_{tx},$$

$$Var(\hat{p}_{wa} - \hat{p}_{tx}) = p_{wa}(1 - p_{wa})/n_{wa} + p_{tx}(1 - p_{tx})/n_{tx})$$

Two sample proportions

Assuming that $\hat{p}_{wa}$ and $\hat{p}_{tx}$ are independent, then we can simply use the rules of expectation and variance

$$E(\hat{p}_{wa} - \hat{p}_{tx}) = p_{wa} - p_{tx},$$

$$Var(\hat{p}_{wa} - \hat{p}_{tx}) = p_{wa}(1 - p_{wa})/n_{wa} + p_{tx}(1 - p_{tx})/n_{tx})$$

As a rule of thumb, when the number of successes and failures are

both larger than 10 for both samples, we can use the CLT and the distribution of the statistic is approximately normal

Two sample proportions Confidence Intervals

We can form a confidence interval for the difference between two proportions-

$$(\hat{p}_{wa} - \hat{p}_{tx}) \pm 1.96\sqrt{\frac{p_{wa}(1-p_{wa})}{n_{wa}} + \frac{p_{tx}(1-p_{tx})}{n_{tx}}}$$

and the same interpretation and meaning from the one sample confidence interval still holds