STAT 234 Lecture 18A
Confidence Intervals for Proportions
Sample Size Calculation
Section 8.1-8.2

Yibi Huang
Department of Statistics
University of Chicago

Confidence Intervals for Proportions

Confidence Intervals for Proportions

Suppose we are interested in the proportion p of individuals with some characteristic of a certain population. We may

1. Draw *simple random sample* of size n
2. Let X be the count of "successes" in the sample. (Here a "success" is an observation with the characteristic of interest)
3. Estimate the unknown true *population proportion* p with the *sample proportion* $\widehat{p} = X/n$

What is the *sampling distribution* of $\widehat{p}$?

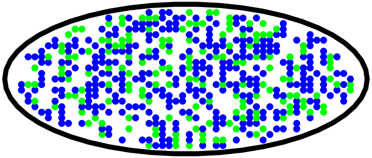
- The exact distribution of X is $Bin(n, p)$.
- Normal approximation to Binomial tells us that when n is sufficiently large

$$X \dot{\sim} N\left(np, \sqrt{np(1-p)}\right) \quad \text{and} \quad \widehat{p} = \frac{X}{n} \dot{\sim} N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$$

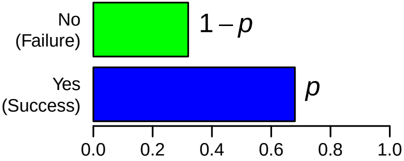
Population

● Yes (Success) ● No (Failure)

p = proportion of Yes (Success)
in the population



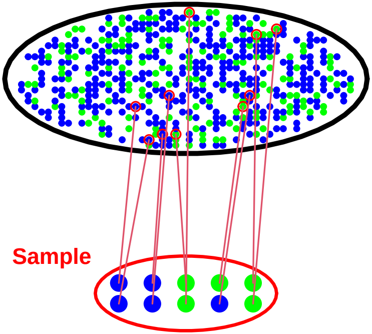
Population Distribution



Population

● Yes (Success) ● No (Failure)

p = proportion of Yes (Success) in the population



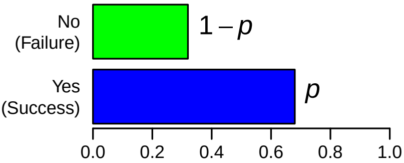
X = # of Yes in the sample = 5

n = sample size = 10

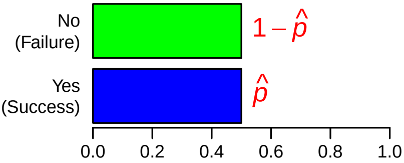
$\hat{p} = X/n$= proportion of Yes (Success) in the sample

= 5 /10 = 0.5

Population Distribution



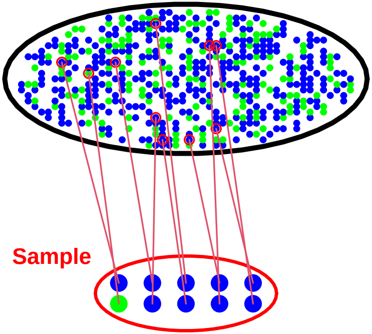
Distribution of the Sample



Population

● Yes (Success) ● No (Failure)

p = proportion of Yes (Success)
in the population

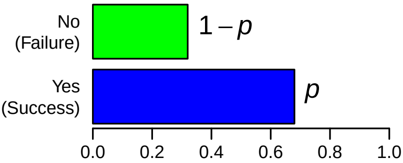


X = # of Yes in the sample = 9

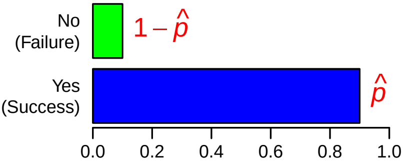
n = sample size = 10

$\hat{p} = X/n$= proportion of Yes (Success)
in the sample
= 9 /10 = 0.9

Population Distribution



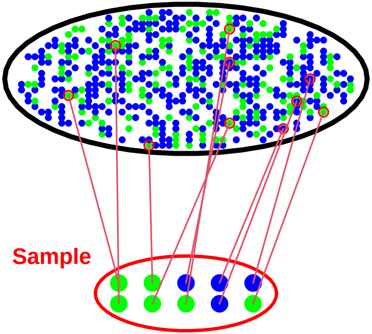
Distribution of the Sample



Population

● Yes (Success) ● No (Failure)

p = proportion of Yes (Success)
in the population

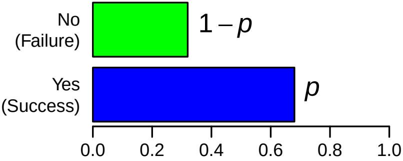


X = # of Yes in the sample = 4

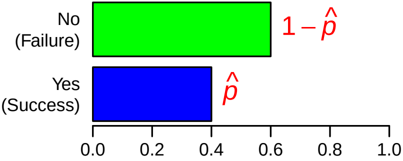
n = sample size = 10

$\hat{p} = X/n$= proportion of Yes (Success)
in the sample
= 4 /10 = 0.4

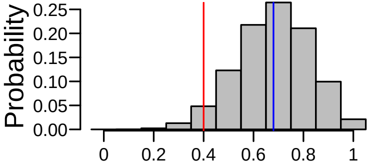
Population Distribution



Distribution of the Sample



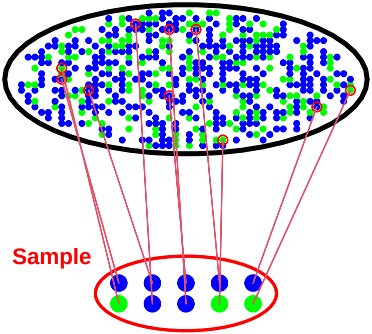
Sampling Distribution of $\hat{p}$



Population

● Yes (Success) ● No (Failure)

p = proportion of Yes (Success)
in the population

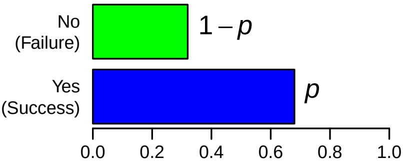


X = # of Yes in the sample = 7

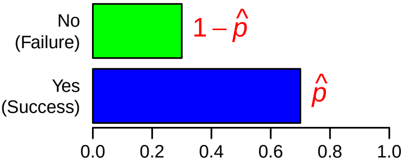
n = sample size = 10

$\hat{p} = X/n$= proportion of Yes (Success)
in the sample
= 7 /10 = 0.7

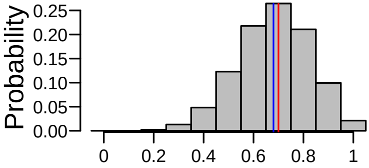
Population Distribution



Distribution of the Sample



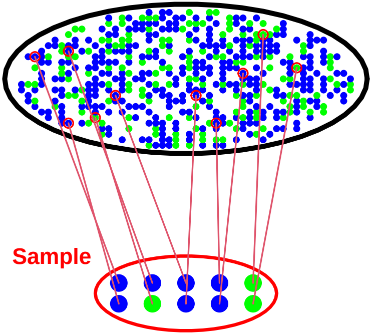
Sampling Distribution of $\hat{p}$



Population

● Yes (Success) ● No (Failure)

p = proportion of Yes (Success) in the population



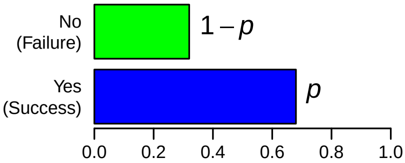
X = # of Yes in the sample = 7

n = sample size = 10

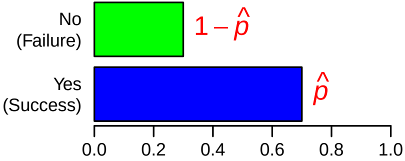
$\hat{p} = X/n$= proportion of Yes (Success) in the sample

= 7 /10 = 0.7

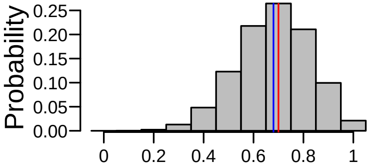
Population Distribution



Distribution of the Sample



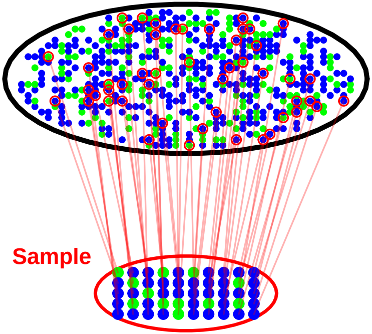
Sampling Distribution of $\hat{p}$



Population

● Yes (Success) ● No (Failure)

p = proportion of Yes (Success)
in the population

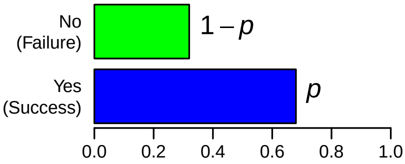


X = # of Yes in the sample = 38

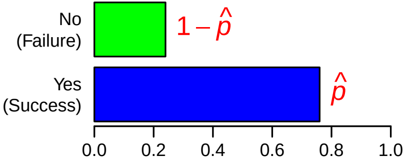
n = sample size = 50

$\hat{p} = X/n =$ proportion of Yes (Success)
in the sample
= 38 /50 = 0.76

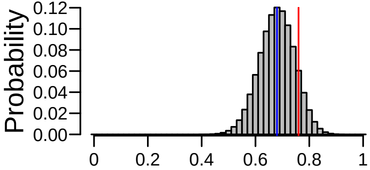
Population Distribution



Distribution of the Sample



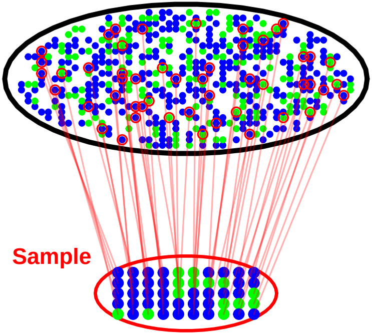
Sampling Distribution of $\hat{p}$



Population

● Yes (Success) ● No (Failure)

p = proportion of Yes (Success)
in the population

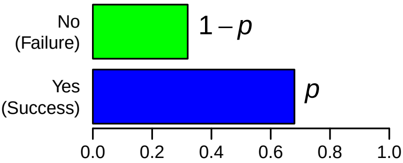


X = # of Yes in the sample = 36

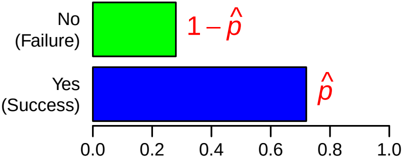
n = sample size = 50

$\hat{p} = X/n$= proportion of Yes (Success)
in the sample
= 36 /50 = 0.72

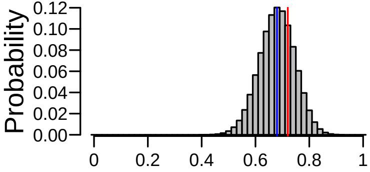
Population Distribution



Distribution of the Sample



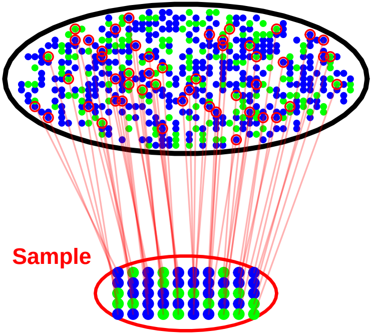
Sampling Distribution of $\hat{p}$



Population

● Yes (Success) ● No (Failure)

p = proportion of Yes (Success)
in the population

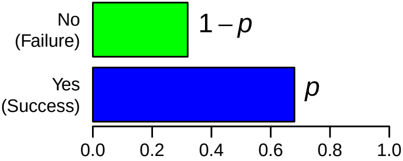


X = # of Yes in the sample = 32

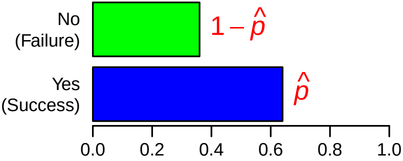
n = sample size = 50

$\hat{p} = X/n$= proportion of Yes (Success)
in the sample
= 32 /50 = 0.64

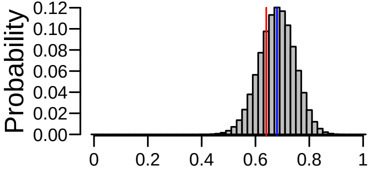
Population Distribution



Distribution of the Sample



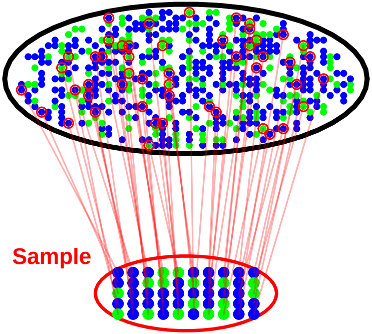
Sampling Distribution of $\hat{p}$



Population

● Yes (Success) ● No (Failure)

p = proportion of Yes (Success)
in the population

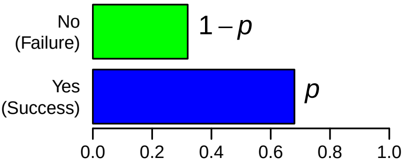


X = # of Yes in the sample = 34

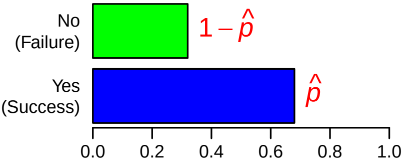
n = sample size = 50

$\hat{p} = X/n$= proportion of Yes (Success)
in the sample
= 34 /50 = 0.68

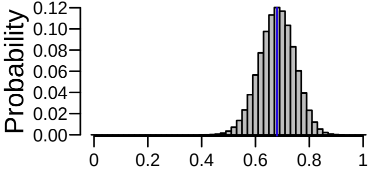
Population Distribution



Distribution of the Sample



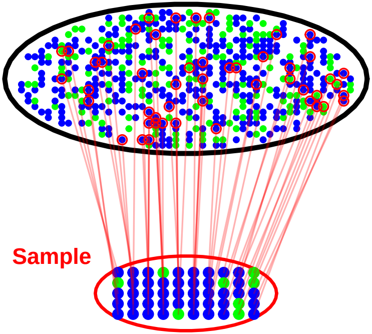
Sampling Distribution of $\hat{p}$



Population

● Yes (Success) ● No (Failure)

p = proportion of Yes (Success)
in the population

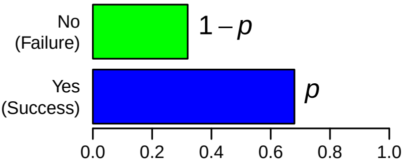


X = # of Yes in the sample = 42

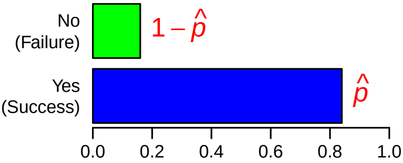
n = sample size = 50

$\hat{p} = X/n$= proportion of Yes (Success)
in the sample
= 42 /50 = 0.84

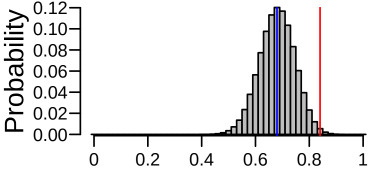
Population Distribution



Distribution of the Sample



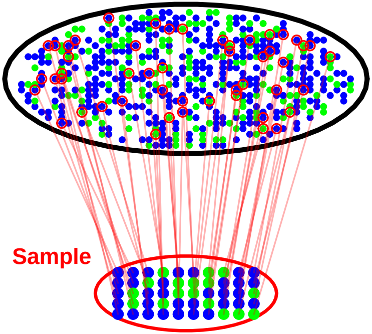
Sampling Distribution of $\hat{p}$



Population

● Yes (Success) ● No (Failure)

p = proportion of Yes (Success)
in the population

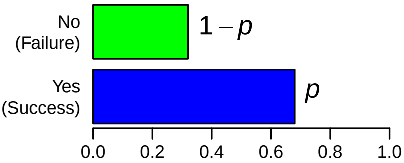


X = # of Yes in the sample = 33

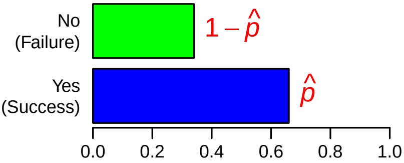
n = sample size = 50

$\hat{p} = X/n$= proportion of Yes (Success)
in the sample
= 33 /50 = 0.66

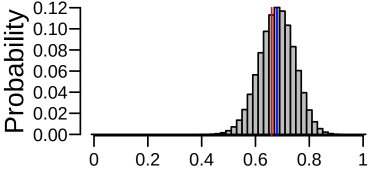
Population Distribution



Distribution of the Sample



Sampling Distribution of $\hat{p}$



Large-Sample Confidence Interval for p

An approximate $100(1-\alpha)\%$ CI for the population proportion p is

$$\widehat{p} \pm z_{\alpha/2}\text{SE} \quad \text{where} \quad \text{SE} = \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}$$

where

Confidence level	90%	95%	99%
α	0.1	0.05	0.01
$z_{\alpha/2}$	1.645	1.960	2.576

Remark: The exact SE should be $\sqrt{p(1-p)/n}$, but the unknown p is replaced with the estimate $\widehat{p}$. This large-sample CI is not very accurate, meaning the actual confidence level often falls below the nominal level.

Example: Side Effects of Pain Relievers (1)

Arthritis is a painful, chronic inflammation of the joints, so many arthritis patients rely on pain relievers, like Ibuprofen. However, Ibuprofen may induce side effects (like dizziness, muscle cramp, allergy, or even seizure) on some patients.

A study interviewed 440 arthritis patients taking Ibuprofen, and found 23 had experienced side effects. Suppose the 440 patients is a SRS from the population of arthritis patients taking Ibuprofen.

Find a 90% confidence interval for the population proportion p of arthritis patients who suffer some adverse symptoms.

Example: Side Effects of Pain Relievers (2)

The sample proportion is $\widehat{p} = \frac{23}{440} \approx 0.052$.

The $z_{\alpha/2}$ for a 90% CI is $z_{0.1/2} \approx 1.645$. So a 90%-CI for p is

$$\begin{aligned}\widehat{p} \pm z_{0.1/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} &\approx 0.052 \pm 1.645\sqrt{\frac{0.052 \times (1-0.052)}{440}} \\ &\approx 0.052 \pm 0.017 = (0.035, 0.069)\end{aligned}$$

Conclusion: With a 90% confidence, between 3.5% and 6.9% of arthritis patients taking this pain medication experience some adverse symptoms.

Why Not Using t for Proportions?

If the sample size n is large enough to apply normal approximation to binomial, n is usually a few hundreds or a few thousands.

The t_{n-1} distribution is very close to normal when $n > 99$, and therefore it is justified to do normal-based inference for proportions.

Sample Size Calculation

Margin of Error of a CI

So far, we introduced 3 CIs for the population mean

$$\overline{x} \pm \underbrace{z_{\alpha/2}\frac{\sigma}{\sqrt{n}}}_{\text{margin of error}}, \quad \overline{x} \pm \underbrace{z_{\alpha/2}\frac{s}{\sqrt{n}}}_{\text{margin of error}}, \quad \overline{x} \pm \underbrace{t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}}_{\text{margin of error}}$$

and one CI for the population proportion

$$\widehat{p} \pm \underbrace{z_{\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}}_{\text{margin of error}}$$

The ± values are called the *margin of error* of the CI, which is the half width of the CI.

Choosing a Sample Size (CI for Proportion)

How large the sample size n need to be to make the margin of error of a CI $\leq m$? Say $m = 0.03$?

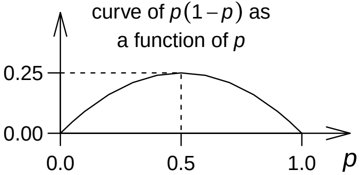
$$\text{margin of error} = z_{\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} \leq m \quad \Rightarrow \quad n \geq \left(\frac{z_{\alpha/2}}{m}\right)^2 \widehat{p}(1-\widehat{p})$$

But $\widehat{p}$ is UNKNOWN before we get the data.
Need to make a guess p^* for the unknown $\widehat{p}$.
How to choose p^*?

Choosing a Sample Size (CI for Proportion) (Cont'd)

How to make a guess p^* for the unknown $\hat{p}$ in a sample size calculation?

$$n \geq \left(\frac{z_{\alpha/2}}{m}\right)^2 p^*(1-p^*)$$



1. Conduct a small pilot study, or use prior studies or knowledge to get a range for possible values of p. Choose the p in the range that is closest to 0.5. E.g.,
   - if possible range of p is [0.1, 0.2], choose $p^* = 0.2$
   - if possible range of p is [0.85, 0.95], choose $p^* = 0.85$
   - if possible range of p is [0.3, 0.6], choose $p^* = 0.5$.
2. The most **conservative** approach is to choose $p^* = 0.5$ since the margin of error is the largest when $\widehat{p} = 0.5$.

Example – Sample Size Calculation for a Proportion

A 1993 survey reported that 72.1% of freshmen responding to a national survey were attending the college of their first choice. Suppose that $n = 500$ students responded to the survey.

1. Find a 95% CI for the proportion p of college freshmen attending their first choice college.
2. Suppose that given the CI, we want to conduct a survey which has a margin of error of 1% (i.e. $m = 0.01)$ with 95% confidence? How many people should we interview?

Example – Sample Size Calculation for a Proportion

The two-sided 95% confidence interval is:

$$\begin{aligned}\widehat{p} \pm z_{0.05/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} &= 0.721 \pm 1.96\sqrt{\frac{0.721(1-0.721)}{500}} \\ &= (0.682, 0.760)\end{aligned}$$

We have a good reason to believe p is in that range. For a sample size calculation, we choose the p closest to 0.5 in this range , $(0.682, 0.760)$. That is $p = 0.682$

$$\left(\frac{z_{0.05/2}}{m}\right)^2 p(1-p) = \left(\frac{1.96}{0.01}\right)^2 \times 0.682(1-0.682) \approx 8331.51$$

The required sample size is 8332. We need a much larger sample size than the original study because we want a smaller margin of error.

Sample Size Calculation (CI for a Mean)

How large the sample size n need to be to make the margin of error of a CI $\leq m$ for a CI for mean

$$\text{margin of error} = z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq m \quad \Rightarrow \quad n \geq \left(\frac{z_{\alpha/2}}{m}\right)^2 \sigma^2$$

But the population variance σ^2 is UNKNOWN before we get the data.
Need to make a guess for the unknown σ^2 based on prior studies or past experience, etc.