Introductory Statistics Lectures

Estimating a population proportion

Confidence intervals for proportions

Anthony Tanbakuchi
Department of Mathematics
Pima Community College

(Compile date: Tue May 19 14:50:07 2009)

Contents

1 Estimating a population proportion

1.1 Introduction

Example 1. We want to estimate the proportion of people in the US who wear corrective lenses. Assuming our class data represents an unbiased sample of the US population, (1) what would our estimate be and (2) how precise is it?

```
R: summary(corrective_lenses)
 NO YES
  8  10
```

POINT ESTIMATES

Notation

p population proportion.

Note: **proportion, percentage, and probability can all be considered as p.**

$\hat{p}$ estimate of sample proportion with x successes in n trials.

$$\boxed{\hat{p} = \frac{x}{n}, \qquad \hat{q} = 1 - \hat{p}} \tag{1}$$

DEFINITION 1.1 POINT ESTIMATE.

A single value (or point) used to approximate a population parameter.

The sample proportion $\hat{p}$ is the best point estimate of the population proportion p.

Importance of proper sampling.

If a sample is not representative of the population, $\hat{p}$ will not be a useful estimate of p. Use proper sampling techniques!

Example 2. Point estimate of proportion of people who wear corrective lenses in the US using class data:

```
R: x = sum(corrective_lenses == "YES")
R: x
[1] 10
R: n = length(corrective_lenses)
R: n
[1] 18
R: p.hat = x/n
R: p.hat
[1] 0.55556
```

Question 1. How good is the estimate of p? How precise is the estimate?

Question 2. What do we need to know about $\hat{p}$ to determine the precision of the estimate?

1.2 Confidence intervals

CONFIDENCE INTERVAL. DEFINITION 1.2

is a range of values — an interval — used to estimate the true value of a population parameter. It provides information about the inherent sampling error of the estimate. (In contrast to point estimate.)

Just as we used the empirical rule to estimate an interval 95% of the data would fall within if the data's distribution was normal, we can construct a similar interval for a statistic given it's **sampling distribution**.

"We are 95% confident that the interval $(\hat{p}_{\text{L}}, \hat{p}_{\text{U}})$ actually contains the true value of p."

CONFIDENCE LEVEL. DEFINITION 1.3

is the probability that the confidence interval contains the true population parameter that is being estimated, if the estimation process is repeated a large number of times.

$$\boxed{\text{confidence level} = 1 - \alpha} \tag{2}$$

where α is the probability that the confidence interval will not contain the true parameter value.

Typical confidence levels

CL	α
99%	0.01
95%	0.05
90%	0.10

Most commonly used is 95%.

1.3 Confidence interval for p

USE

Often used to answer:

1. What is a reasonable estimate for the population proportion?
2. How much variability is there in the estimate for the population proportion?
3. Does a given target value fall within the confidence interval?

COMPUTATION

Sampling distribution of $\hat{p}$

If np and $nq \geq 5$ then p will have a normal distribution[1] and the CLT tells us that $\hat{p}$ **is approximately normally distribution** where:

$$\mu_{\hat{p}} = p \tag{3}$$

[1]Normal approximation of binomial.

$$\boxed{\sigma_{\hat{p}} = \sqrt{\frac{pq}{n}} \approx \sqrt{\frac{\hat{p}\hat{q}}{n}}} \tag{4}$$

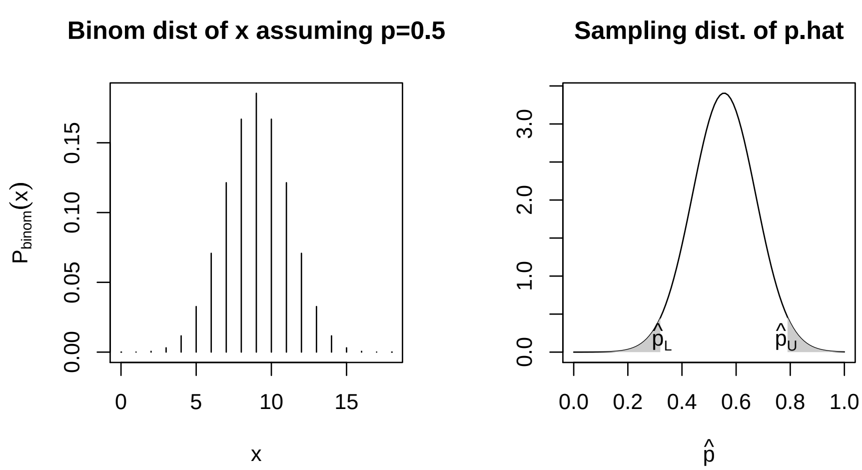
Definition 1.4

Confidence interval for p.

The confidence interval for p at the $(1-\alpha)$ confidence level is:

$$\hat{p}_{\mathrm{L}} < p < \hat{p}_{\mathrm{U}} \tag{5}$$

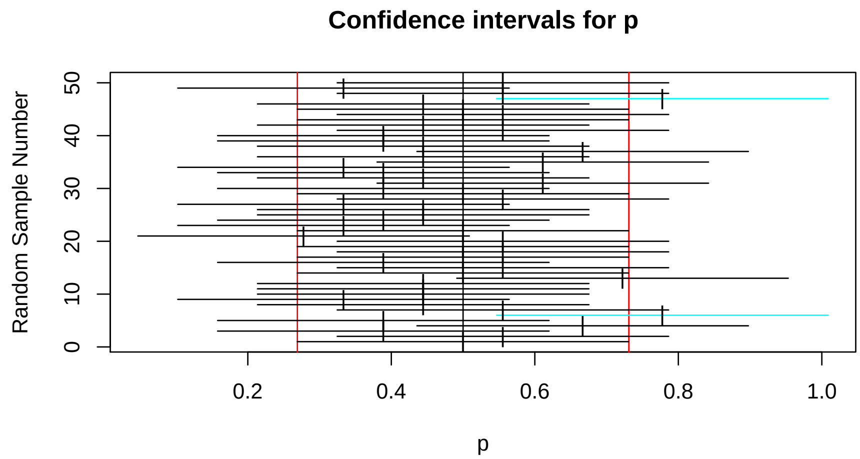
$$F^{-1}_{\mathrm{norm}}(\alpha/2) < p < F^{-1}_{\mathrm{norm}}(1-\alpha/2) \tag{6}$$



Sampling distribution for $\hat{p}$: If the requirements are met it will have a normal distribution with $\mu_{\hat{p}} \approx \hat{p} = 0.556$, $\sigma_{\hat{p}} \approx \sqrt{\frac{\hat{p}\hat{q}}{n}} = 0.117$. Total shaded area is $\alpha = 0.05$, each tail has an area of $\alpha/2 = 0.025$. Thus, 95% confidence interval for p is $(\hat{p}_{\mathrm{L}}, \hat{p}_{\mathrm{U}}) = (0.326, 0.785)$.

Variation in CI of p from sample to sample

Simulate study of corrective lens use 50 times with random sample size of 18 **assuming** true $p = 0.5$.



95% CI's, tick marks represent each point estimate $\hat{p}$.
In general, 95% of the confidence intervals will contain p.

Confidence intervals for p in R
To construct a CI $(\hat{p}_{\mathrm{L}}, \hat{p}_{\mathrm{U}})$ at $(1-\alpha)$ confidence level:
$\hat{p}_{\mathbf{L}} = \hat{p} - E$
$\hat{p}_{\mathbf{U}} = \hat{p} + E$
where E is the margin of error.
With the following requirements:
1. Simple random sample.
2. Satisfies binomial distribution.
3. Satisfies normal approximation to binomial.

Margin of error E. Definition 1.5
The confidence interval can be expressed in terms of the margin of error E:

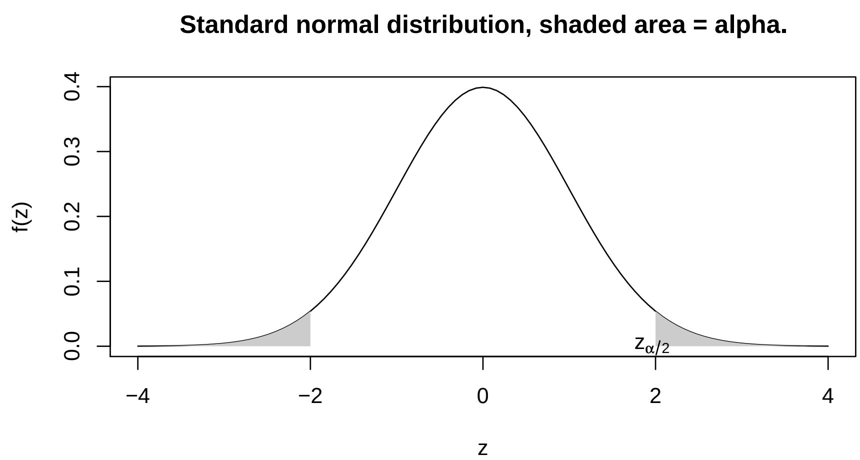
$$\boxed{\text{CI: } \hat{p} \pm E} \tag{7}$$

where the margin of error for $\hat{p}$ is:

$$\boxed{E = z_{\alpha/2} \cdot \sigma_{\hat{p}}} \tag{8}$$

or if the upper and lower values are known:

$$E = \frac{\text{upper} - \text{lower}}{2} = \frac{\hat{p}_{\mathrm{U}} - \hat{p}_{\mathrm{L}}}{2} \tag{9}$$



Definition 1.6 Critical value $z_{\alpha/2}$.

The critical value $z_{\alpha/2}$ is the value of z on the standard normal distribution with $\alpha/2$ **area to the RIGHT**.

Example 3. Find the critical value $z_{\alpha/2}$ for the 95% confidence interval.

```
R: alpha = 1 - 0.95
R: z.critical = qnorm(1 - alpha/2)
R: z.critical
[1] 1.9600
```

$z_{\alpha/2}$ for 95% CL

$$\boxed{z_{\alpha/2} = 1.96 \qquad \text{for } \alpha = 0.05} \tag{10}$$

Question 3. How does this differ from the Empirical Rule?

Example 4. Using our class data to estimate the 95% confidence interval for the proportion of people in the US who wear corrective lenses.

What's known:

```
R: alpha = 1 - 0.95
R: n
[1] 18
R: x
[1] 10
R: p.hat = x/n
```

```
R: q.hat = 1 - p.hat
R: sigma.p.hat = sqrt(p.hat * q.hat/n)
```

Finding the 95% CI

```
R: z.critical = qnorm(1 - alpha/2)
R: E = z.critical * sigma.p.hat
R: p.L = p.hat - E
R: p.U = p.hat + E
R: CI = c(p.L, p.U)
R: CI
[1] 0.32600 0.78511
```

Thus, we are 95% confident that the true proportion of people who wear corrective lenses lie somewhere between 32.6% and 78.5%, or in terms of the margin of error: $55.6\% \pm 23\%$. (A 2001 study estimated it at 56%.[2])

Question 4. What would our confidence interval be if we wanted a 100% confidence level?

Question 5. What would our confidence interval be if we wanted a 0% confidence level?

DETERMINING SAMPLE SIZE FOR DESIRED E

Determining sample size for desired E
To find the necessary sample size for a desired E, just solve for n.

$$E = z_{\alpha/2} \cdot \sigma_{\hat{p}}$$
$$E = z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}\hat{q}}{n}}$$

solving for n

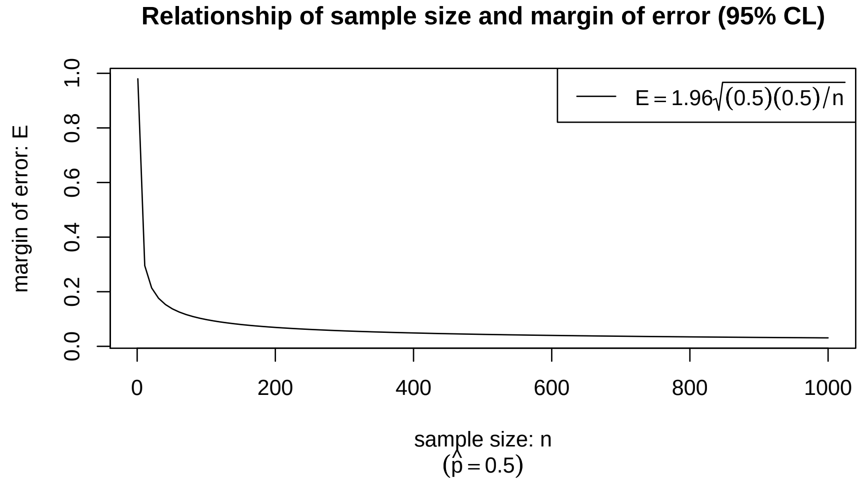
$$\boxed{n = \hat{p}\hat{q}\left(\frac{z_{\alpha/2}}{E}\right)^2} \tag{11}$$

If $\hat{p}$ and $\hat{q}$ are unknown[3], use 0.5 for both. Always round up!

[2]Source: Walker, T.C. and Miller, R.K. 2001 Health Care Business Market Research Handbook, fifth edition, Norcross (GA): Richard K. Miller & Associates, Inc., 2001. Study estimated about 160 million people in US wear glasses. 2001 population was estimated to be 286 million.

[3]Common, since you often determine n before doing the study to decide how big it needs to be. However, if an estimate of $\hat{p}$ can be found, use it.

Relationship of n and E



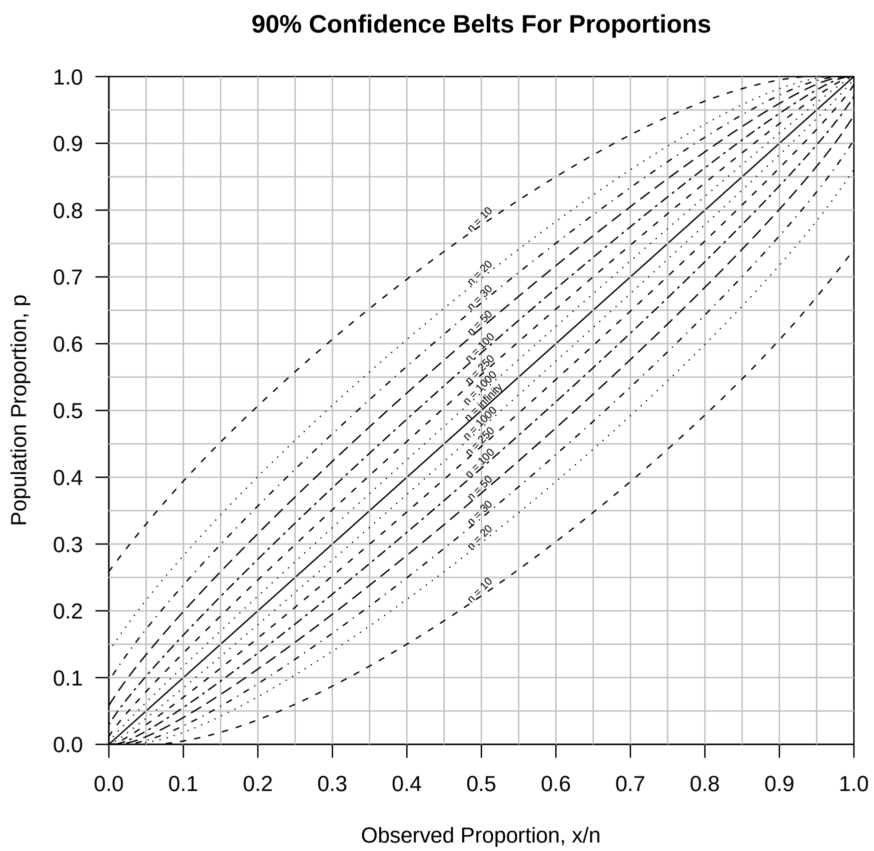
Example 5. You have been hired by the Clear Optical company[4] to design a study to estimate the proportion of the US population who wear corrective lenses. The desired margin of error is 1% (at the 95% confidence level). What is the minimum sample size you should use? (Assume we don't know $\hat{p}$ yet.)

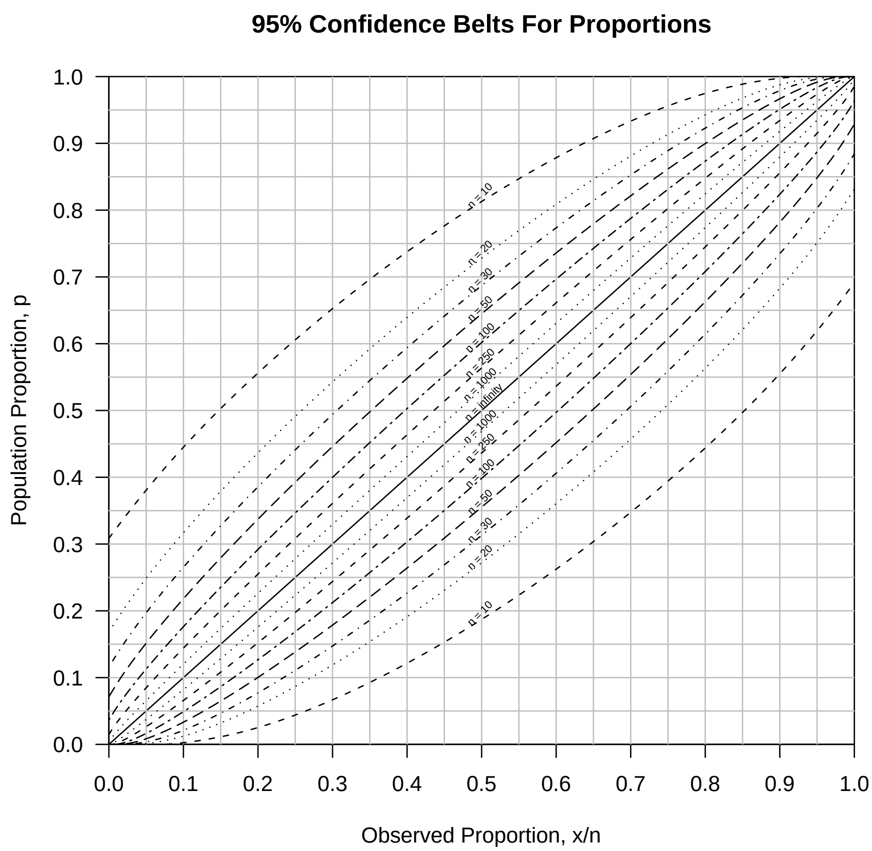
```
R: alpha = 1 - 0.95
R: E = 0.01
R: p.hat = 0.5
R: q.hat = 1 - p.hat
R: z = qnorm(1 - alpha/2)
R: z
[1] 1.9600
R: n = p.hat * q.hat * (z/E)^2
R: n
[1] 9603.6
```

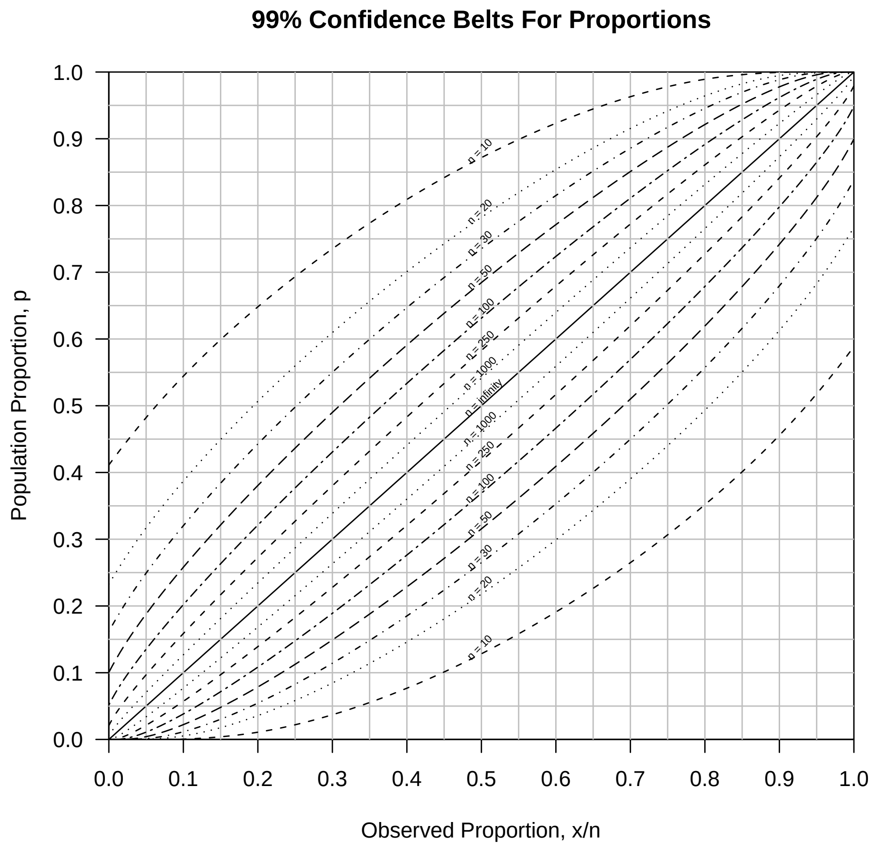
Thus, to attain the desired margin of error (at the 95% confidence level), a random sample of 9604 people should be used.

CONFIDENCE INTERVAL BELT GRAPHS

[4]Because you did so well in your statistics class.


90% Confidence Belts For Proportions
Population Proportion, p
Observed Proportion, x/n
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
n = 10
n = 20
n = 30
n = 50
n = 100
n = 250
n = 1000
n = infinity


95% Confidence Belts For Proportions
Population Proportion, p
Observed Proportion, x/n
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0


99% Confidence Belts For Proportions
Population Proportion, p
Observed Proportion, x/n
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0

1.4 Summary

- Understand: point estimate, confidence interval, confidence level.
- $\hat{p} = x/n$
- Confidence level $= 1 - \alpha$ (Assume 95% if unspecified.)
- The confidence interval for p is $\hat{p} \pm E$
  - Requirements: (1) Simple random samples, (2) Binomial Dist, (3) Normal approx to binom.
  - If the requirements are satisfied, the sampling distribution of $\hat{p}$ will be a normal distribution with mean and standard deviation $\mu_{\hat{p}}$ and $\sigma_{\hat{p}}$.
  - $E = z_{\alpha/2} \cdot \sigma_{\hat{p}}$ where $\sigma_{\hat{p}} \approx \sqrt{\hat{p}\hat{q}/n}$
  - Critical value: $z_{\alpha/2} =$ `qnorm(1-alpha/2)`
    (z with $\alpha/2$ area to the **right** on standard normal distribution.)
- To find required sample size given E: $n = \hat{p}\hat{q}\left(\frac{z_{\alpha/2}}{E}\right)^2$
  Let $\hat{p} = \hat{q} = 0.5$ if unknown.

Be accurate, don't use the Empirical rule from this point forward for actual calculations.

1.5 Additional examples

Additional examples

College officials want to estimate the percentage of students who carry a gun, knife, or other such weapon.

Question 6. How many randomly selected students must be surveyed in order to be 95% confident that the sample percentage has a margin of error of 1%? (Assume no available info.)

Question 7. If we use a sample size that is smaller than required, what do we expect to happen to the margin of error

Question 8. The above study was conducted with 500 students and 8 indicated that they carried a weapon. What is the actual 95% confidence interval?

Question 9. Why was our margin of error still less than 1% even though our sample size was too small?