



Ready, Tech, Go:

If technology has revolutionized the teaching of statistics, why are we still teaching the same old course?

Paul Velleman - Cornell University

pfv2@cornell.edu

Dick De Veaux - Williams College

Dave Bock - Ithaca High School

5.4 2002

Assumptions for Inference (and the conditions that confirm or override them.)

Proportions (z)

- **One Proportion z**
 1. individuals are independent
 2. sample is sufficiently large
 - **Two Proportion z**
 1. samples are independent
 2. data in each sample are independent
 3. both samples are sufficiently large
- 1. SRS and <10% of population
 - 2. successes & failures ≥ 10
 - 1. (sorry – you have to think about this)
 - 2. both SRSs and < 10% of populations OR random allocation
 - 3. successes & failures ≥ 10 for both

Means (t)

- **One Sample t** (df = $n - 1$)
 1. individuals are independent
 2. population has a Normal model
 - **Matched Pairs** (df = $n - 1$)
 1. data are matched
 2. individuals are independent
 3. population of differences Normal
 - **Two independent samples** (df from formula, or smaller $n - 1$ if necessary)
 1. samples are independent
 2. data in each sample independent
 3. both populations are Normal
- 1. SRS & < 10% of the population
 - 2. histogram is unimodal and symmetric*
 - 1. (think about the design)
 - 2. SRS & < 10% OR random allocation
 - 3. histogram of differences is unimodal and symmetric (*less critical as n increases)
 - 1. (think about the design)
 - 2. SRSs & < 10% OR random allocation
 - 3. both histograms unimodal & symmetric*

Distributions (ChiSquare)

- **Goodness of Fit** (df = $cells - 1$; one variable, one sample compared to population model)
 1. data are counts
 2. data in sample are independent
 3. sample is sufficiently large
 - **Homogeneity** (df = $(r - 1)(c - 1)$; samples from many populations compared on one variable)
 1. data are counts
 2. data in samples are independent
 3. samples are sufficiently large
 - **Independence** (df = $(r - 1)(c - 1)$; sample from one population classified on two variables)
 1. data are counts
 2. data are independent
 3. sample is sufficiently large
- 1. (are they?)
 - 2. SRS & < 10% of the population
 - 3. all expected counts ≥ 5
 - 1. (are they?)
 - 2. SRSs & < 10% OR random allocation
 - 3. all expected counts ≥ 5
 - 1. (are they?)
 - 2. SRSs & < 10% of the population
 - 3. all expected counts ≥ 5

Regression coefficients (t, df = $n - 2$)

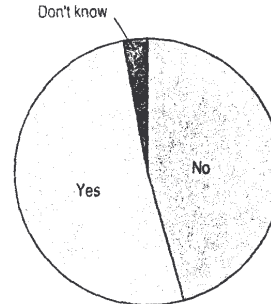
- **Association** between two measured variables ($\beta = 0$?)
 1. form of relationship is linear
 2. errors are independent
 3. variability of errors is constant
 4. errors have a Normal model
- 1. scatterplot looks approx linear
 - 2. no apparent pattern in residuals plot
 - 3. residuals plot has consistent spread
 - 4. histogram of residuals is approximately unimodal and symmetric*

A Confidence Interval for a Proportion **STEP-BY-STEP**

THE W's:

WHO	Adults in the United States
WHAT	Response to a question about marijuana
WHEN	August 2000
WHERE	United States
HOW	507 adults were randomly sampled and asked by the Gallup Poll.
WHY	Public opinion research

In August 2000, the Gallup Poll asked 507 randomly sampled adults the question "Do you think the possession of small amounts of marijuana should be treated as a criminal offense?" Of these, 47% responded "No."³ What can we conclude from this survey?



To answer this question we'll build a confidence interval for the proportion of all U.S. adults who would respond "No." There are four steps to building a confidence interval for proportions:

Think

Parameter Identify the *parameter* you wish to estimate. This is the value we hope to catch within our confidence interval.

Choose and state a confidence level.

Plan Check the conditions. We often cannot check assumptions. The conditions are the practical aspects of the data that we *can* check to be assured that the null model can be applied.

State the sampling distribution model for the statistic.

Choose your method.

We wish to find an interval that is likely with 95% confidence to contain the true proportion p , of U.S. adults who think possession of small amounts of marijuana should not be treated as a criminal offense.

✓ **Plausible independence condition:** Gallup phoned a random sample of U.S. adults. It is very unlikely that any of the respondents influenced each other.

✓ **Random condition:** Gallup drew a random sample from all U.S. adults. We do not have details of the randomization, but we assume that we can trust it.

✓ **10% condition:** Although sampling was necessarily without replacement, there are many more U.S. adults than were sampled. The sample is certainly less than 10% of the population.

✓ **Success/failure condition:**

$$n\hat{p} = 507 \times 47.0\% = 238.3 > 10 \text{ and}$$

$$n\hat{q} = 507 \times 53.0\% = 268.7 > 10$$

so our sample is large enough.

Under these conditions the sampling distribution of the proportion can be modeled by a Normal model.

We will find a **one-proportion z-interval**.

Show

Mechanics Construct the confidence interval.

We could informally use 2 for our critical value, but 1.96 is more accurate.

Reality Check

The confidence interval is centered at the sample proportion and about as wide as we might expect for a sample of 500.

Tell

Interpretation Tell what the confidence interval means, in the proper context.

We know: $n = 507$, $\hat{p} = .47$

$$\text{So } SE(\hat{p}) = \sqrt{\frac{pq}{n}} = \sqrt{\frac{.47 \times .53}{507}} = 0.022$$

Because the sampling model is Normal, for a 95% confidence interval, the critical value $z^* = 1.96$.

From these, we find the margin of error as

$$\text{m.o.e.} = z^* \times SE(\hat{p}) = 1.96 \times 0.022 = 0.43$$

So the 95% confidence interval is:

$$.47 \pm 0.043 \text{ or } (0.427, 0.513)$$

We can be 95% confident that between 42.7% and 51.3% of all U.S. adults think possession of small amounts of marijuana should not be treated as a criminal offense.

What Can Go Wrong?

Confidence intervals are powerful tools. Not only do they tell what we know about the parameter value, but—more important—they also tell what we *don't* know. But in order to use them effectively, you must be clear about what you say about them.

Don't Misstate What the Interval Means

- *Don't suggest that the parameter varies.* A statement like "There is a 95% chance that the true proportion is between 42.7% and 51.3%" sounds as though you think the population proportion wanders around and sometimes happens to fall between 42.7% and 51.3%. When you interpret a confidence interval, make it clear that *you* know that the population parameter is fixed and that it is the *interval* that varies from sample to sample.
- *Don't claim that other samples will agree with yours.* Keep in mind that the confidence interval makes a statement about the true population proportion. An interpretation such as "In 95% of samples of U.S. adults the proportion who think marijuana should be decriminalized will be between 42.7% and 51.3%" is just wrong. The interval isn't about sample proportions, but about the population proportion.
- *Don't be certain about the parameter.* Saying "Between 42.1% and 61.7% of sea fans are infected" asserts that the population proportion cannot be outside that interval. Of course, we can't be absolutely certain of that. (Just pretty sure.)
- *Don't forget: It's the parameter.* Don't say, "I'm 95% confident that \hat{p} is between 42.1% and 61.7%." Of course you are—in fact, we calculated that $\hat{p} = 51.8\%$ of the fans in our sample were infected. So we already *know* the sample proportion. The confidence interval is about the (unknown) population parameter, p .
- *Don't claim to know too much.* Don't say, "I'm 95% confident that between 42.1% and 61.7% of all the sea fans in the world are infected." You didn't sample from all the sea fans in the world. Just those of this type on the Las Redes Reef.
- *Do take responsibility.* Confidence intervals are about uncertainty. *You* are the one who is uncertain, not the parameter. You have to accept the responsibility and consequences of the fact that not all the intervals you compute will capture the true value. In fact, about 5% of the 95% confidence intervals you find will fail to capture the true value of the parameter.

You *can* say, "I am 95% confident that between 42.1% and 61.7% of the sea fans on the Las Redes Reef are infected."

There are about 500 species of sea fans. They are found on coral reefs throughout the world.

S.8.2 Not currently available.