We can use R to find confidence intervals. We look at confidence intervals concerning proportions in this worksheet.

1 One sample

The New Yorker had an interesting article (John Cassidy 2006/04/03) about the poverty rate. One tidbit mentioned was the following:

> According the US Census bureau, the poverty rate in New York City during 2004 was 20.3 percent.

As the census is only done every 10 years, this is actually based on a sample. The sample size is large. We suppose 20.3 percent summarizes a random sample of size 50,000. What is a 95% confidence interval for the actual poverty rate based on the sample estimate?

We could compute the interval

$$[\hat{p} - a\mathsf{SE}(\hat{p}), \hat{p} + a\mathsf{SE}(\hat{p})]$$

but R allows us to do this more easily with the function `prop.test()`. We specify x, n and the confidence level. (The statistic $\hat{p} = x/n$.) For this problem we have

```
> n = 50000; phat = 0.203
> prop.test(phat * n, n, conf.level = 0.95)

        1-sample proportions test with continuity correction
data:  phat * n out of n, null probability 0.5
X-squared = 17640.61, df = 1, p-value < 2.2e-16
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.1994873 0.2065585
sample estimates:
    p
0.203
```

The confidence interval is found after

```
95 percent confidence interval:
 0.1995 0.2066
```

Problem 1:
Find a 99% confidence interval for the poverty rate.

Problem 2:
In Detroit, a sample of size 40,000 (say) had a poverty rate of 33.6 percent. Find a 99% confidence interval.

Problem 3:
In New Orleans, a sample of size 35,000 (say) had a poverty rate of 23.0 percent. This data was mentioned during the Katrina cleanup. Find a 99% confidence interval for the actual poverty rate. Does it contain New York's 20.3 percent?

2 Two samples

Sometimes we would like to compare to rates. For instance, is the *actual* poverty rate in New York City less than the *actual* poverty rate in New Orleans? Let's be clear, we have two samples which give *sample* rates of 20.3% and 23.0% so the answer looks like yes. But we can't simply conclude that!

Why? Sampling variation! Suppose the two actual rates are identical and we take a random sample from each town. One of the sample rates will be less than the other. So we can't conclude from that alone that the two cities have an *actual* difference.

What can we do? We need to investigate the randomness of the two sample rates. It is easier to consider their difference.

Generically speaking, let $\widehat{p}_1$ and $\widehat{p}_2$ summarize two random samples with x_1 and x_2 "successes" and n_1 and n_2 trials. Then the sample distribution of $\widehat{p}_1 - \widehat{p}_2$ is *normally distributed* provided both n_1 and n_2 are large enough. That's great. If we know the mean and standard deviation we can do computations.

Two facts: $\mathsf{E}(\widehat{p}_1 - \widehat{p}_2) = 0$ and

$$\mathsf{SD}(\widehat{p}_1 - \widehat{p}_2) = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}, \quad \text{so} \quad \mathsf{SE}(\widehat{p}_1 - \widehat{p}_2) = \sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n_1} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{n_2}}.$$

A $(1-\alpha)\cdot 100\%$ CI for $p_1 - p_2$ is

$$[(\widehat{p}_1 - \widehat{p}_2) - a\mathsf{SE}(\widehat{p}_1 - \widehat{p}_2), (\widehat{p}_1 - \widehat{p}_2) + a\mathsf{SE}(\widehat{p}_1 - \widehat{p}_2)].$$

The formulas are there to compute, but we'll let the computer do it. We only need to enter in the x values as a pair, along with the n values. (This is done using `c()` as indicated in the example.)

For instance, to find a 95% CI for the difference of population poverty rates for NYC and NOLA we have

```
> prop.test(x = c(0.203 * 50000, 0.23 * 35000), n = c(50000, 35000),
+     conf.level = 0.99)

        2-sample test for equality of proportions with continuity correction
data:  c(0.203 * 50000, 0.23 * 35000) out of c(50000, 35000)
X-squared = 89.0338, df = 1, p-value < 2.2e-16
alternative hypothesis: two.sided
99 percent confidence interval:
 -0.03444331 -0.01955669
sample estimates:
prop 1 prop 2
 0.203  0.230
```

The 99% CI does not contain 0, indicating that the population rates are different.

Problem 4:
The national poverty rate appears to be on the rise. Suppose in 2004 a random sample of 60,000 people found the sample rate to be 12.7 percent. In 2003 suppose a sample of 50,000 found the sample rate to be 12.4 percent.

Does a 99% CI for the difference of population rates include 0?

Problem 5:
The “poverty rate” is measured by a modification of a formula appearing in a 1963 paper by Mollie Orshansky. Oshansky carefully calculated what she though a family 4 would need to spend to eat in a day on an “economy plan” ($2.80 at the time). Then she saw that government figures indicated that 1/3 of household expenses went to food. From here she multiplied $2.80 times 365 times 3 to get a poverty level. Subsequently this level gets adjusted each year, but the basic reasoning remains.

Do you think this is a reasonable way to compute a “poverty rate?” Do you think it would work in all states? All Cities? Does the fact that the amount a family pays in tax is not included make sense?

Some additional problems:

Problem 6:
Two surveys of presidential approval ratings were recently released. An AP poll of 1,000 found the president at 36%. Another of 1,100 found the president at 34%. Does a 95% CI for the difference of the population proportions contain 0?

Problem 7:
Have attitudes about the war in Iraq changed amongst high school students? Suppose a random sample of 500 in 2003 found 42% in favor. A sample of size 450 in 2006 found 24% in favor. Is 0 in a 95% confidence interval? Find the margin or error for your CI.

Problem 8:
A test of a new allergy medicine is being done. To eliminate the *placebo* effect, a control group and treatment group are being used. The medication is not expected to work for everyone, and the sample shows it didn't. In particular the numbers are

	no. who got better	no. in trial
control	35	100
treatment	25	110

Find a 95% confidence interval for the difference of population proportions. Does it include 0?