Stats in Practice #10, Comparing Populations

Kellin Rumsey

4/22/2019

1. The dataset

The CDI dataset contains demographic information for the 440 most populated counties in the United States. Data for a random sample of $n = 100$ of these counties can be found on the course webpage in a file called cdi_sample.csv.

Load the dataset

- Open Rstudio and open a new script.
- Load the dataset by typing: `cdi = read.csv('http://math.unm.edu/~knrumsey/cdi_sample.csv')`

2. Poverty and Crime

In 2000 (the year this CDI dataset is from), the average poverty rate was about 10%. For each of the $n = 100$ counties in this dataset, check to see if the poverty rate is higher than average by typing:

```
high_poverty <- (cdi$PercentBelowPoverty > 10)
```

Researchers hypothesize that there will be more serious crimes (per capita) in counties with high poverty rates than in counties with low poverty rates. We can use 2-sample t-procedures to test this hypothesis.

Lets divide this population into two sub-populations.

- Population 1: Counties with poverty rates higher than the national average.
- Population 2: Counties with poverty rates less than or equal to the national average.

1. State the hypotheses using mathematical notation.

Now we get the crimes per capita for each subgroup in R by typing:

```
#Get crimes per capita here
crimes_pc <- cdi$TotalSeriousCrimes/cdi$Population
#Split into two sub-groups here
x1 <- crimes_pc[high_poverty]
x2 <- crimes_pc[-high_poverty]
```

2. Make a histogram of crimes per capita for each population. The following R code will help. **Include this plot in your writeup**. Pay close attention to the x-axis. Does there seem to be a significant difference in the averages?

```
par(mfrow=c(1,2))
hist(x1, xlab='Crimes per capita', main='Population 1\nHigh Poverty')
hist(x2, xlab='Crimes per capita', main='Population 2\nLow Poverty')
```

3. Use the functions `length()`, `mean()` and `sd()` to get the sample size, mean and standard devition for each population.

4. Calculate the 2-sample t-procedure test statistic.

5. At the 1% significance level, what is your conclusion in terms of the data. (Use the table.)

6. You can also find the actual p-value (not just a range) in R by typing: `pt(t, df=?, lower.tail=FALSE)` where *t* is the computed test statistic and the ? should be the degrees of freedom that we use. What is the p-value using this approximation?

Two sample test using R.

We mention breifly in class that the degrees of freedom can be really difficult to find, so we use a simplified approximation. In R, we can find a much better approximation of the degrees of freedom and it's **way easier** than the process we went through above! Just type the following in R:

```r
t.test(x1, x2)
```

6. What is the degrees of freedom used by R?
7. What is the p-value of the test? This is more accurate than what we did. How does it compare to your answer?
8. The t.test function also gives a confidence interval. Interpret this interval in words.
9. Does the headline: "Increased poverty leads to significantly higher instances of violent crime" seem fair or misleading? Explain your answer using by discussing the physical significance of the lower and upper bounds.