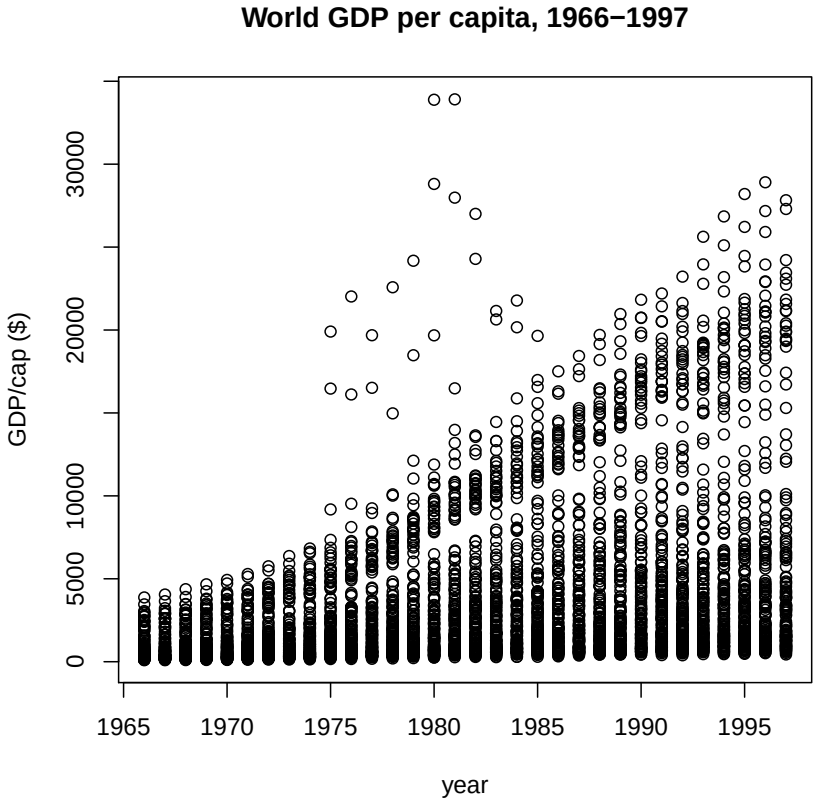
1 Motivation

Turnout and the two-tailed t-test

Up to now, we have used relatively imprecise methods to look at differences in turnout between the South and the rest of the Union across time. We've seen that turnout appears to have converged over time, but we stopped at a visual inspection without analyzing the extent of that convergence. Our final task with the turnout data will be to use the first statistical test we've learned, the two-tailed t-test, to quantify the extent to which turnout in the regions is different or similar.

Intro to World GDP per Capita

In the second part of this assignment, we will take our first foray into analyzing the World GDP per capita data by looking more closely at outliers between 1975 and 1985.



2 Exercises

Turnout

Exercise 1: Let's quantify the difference in turnout between Southern and non-Southern states in 1960. Assume that our hypothesis is that the gap is at least 20%.

1. Using the language of the two-tailed t-test, clearly state the null hypothesis, and the alternate hypothesis.
2. What are the steps you must take to test this hypothesis?

Exercise 2: Using the two-tailed t-test...

1. Test whether the gap between South and non-South in 1960 is at least 20%. Use the R-code from the 'pointers' section of this assignment. Show how you call that function (but don't include the code below), and show the resulting output.
2. Interpret the results of this test – can we reject the null hypothesis or not? Why?

Exercise 3:

1. For 1960, find the biggest integer value Δ for which we can reject the (null) hypothesis that the regions are samples from populations with means that are no more than Δ apart (*hint:* work your way down from the difference in sample means).
2. Do the same for 2000.

Exercise 4: *Answer using around 250 words:* In HW1 and HW2, it looked like the gap in turnout between the South and the rest decreased from 1960 onwards. Is this observation supported by the results of the preceding exercises? If we were to repeat exercise 3 for other elections between 1960 and 2000, what would we expect to see?

World GDP per capita: let's arbitrarily designate countries with GDP/cap > 15000 in 1980 as the outliers we are looking for.

Exercise 5: Using the wide form GDP data, identify which countries are the outliers between 1971 and 1985. State clearly which they are. (*hint:* there are 3 which meet our criteria.)

Exercise 6: Recreate the scatterplot at the top of this assignment. Add to this plot lines for the outlier countries you've identified. Be sure to color the lines so that we can see which country is which, and add a legend.

Exercise 7: Answer using between 250 and 750 words total:

- Briefly describe the pattern that you see.
- What do these countries share in common?
- Given the time period in which we see these patterns, what do you think might be the cause of the pattern?
- What information would you need to see whether your ideas are correct?

You'll need to read about these countries to answer this question. Remember to cite your sources. Given the casual nature of the question, Wikipedia is an acceptable source.

3 Administrative details

- This assignment is due at 11:59 PM, 4/29/14 via turnitin.
- You may work with one partner. If you do work with a partner, include your partner's name under yours in the heading. Remember that working with a partner does not mean copy from each other. Your submission **must be** your own work.
- Submit R code where specifically instructed to. Elsewhere, code is optional.
- See the syllabus for policies regarding late submissions and plagiarism.

4 Pointers

▶ **2 tailed t-test:** Copy-paste the following function into R – it is an implementation of the two-tailed t-test from the lecture notes:

```
t.test2 = function(mu1, sig1, n1, mu2, sig2, n2, Delta=0){
        Mu1 = max(c(mu1, mu2))
        Mu2 = min(c(mu1, mu2))
        T <- (Mu1 - Mu2 - Delta)/sqrt(sig1^2/n1 + sig2^2/n2)
        nu <- (sig1^2 + sig2^2)^2/(sig1^4/(n1-1)+sig2^4/(n2-1))
        pval <- 2*(1-pt(abs(T), nu))
        return(list(pval=as.vector(pval), T=as.vector(T), nu=as.vector(nu)))
}
```

To use it, just pass sample parameters. For example, to test for difference in means of 2 between two samples with $\mu_1 = 6, \sigma_1 = 1, N_1 = 20, \mu_2 = 3, \sigma_2 = 2, N_2 = 30$:

```
t.test2(6, 1, 20, 3, 2, 30, Delta=2)
```

▶ **Identifying particular countries:** In lecture, I showed how to find countries that had GDP/cap > \$8000 in 1981 as follows:

```
h81 = subset(gdp, gdp$Year == 1981 & GDPcapImp > 8000)
highCountries = h81$Country
gdp$high = ifelse(gdp$Country %in% highCountries, 1, 0)
```

That was for the long-form. For the wide-form, the following accomplishes the same:

```
gdp.wide$high = ifelse(gdp.wide$'1981' >= 8000, 1, 0)
```

Use whichever is appropriate for exercise 5.

▶ **Scatterplot with lines:** Exercise 6 is done most easily by using both the long-form and wide-form versions of 'gdp'. Roughly speaking, you should use the 'plot' command with years for x values and gdp values from the long-form data as the y values (if you get really stuck, look very carefully at the plot at the beginning of this HW). Then, add lines for each row in the (wide-form) subset that you found in exercise 1.