T-test for Difference in Means and Hypothesis Testing

Loading and preparing our dataset

In this seminar, we will load a file in comma separated format (`.csv`). Our data comes from the Quality of Government Institute. The `load()` function from last week works only for the native R file format. To load our csv-file, we use the `read_csv()` function.

```
library(readr)
world_data <- read_csv("QoG2012.csv")
```

Go ahead:

1. check the dimensions of `world_data` ,
2. find the column names of the variables of the dataset,
3. print the first six rows of the dataset,

The variables are only a small set from the full Quality of Governance dataset and store the following information:

Variable	Description
`h_j`	1 if free judiciary (justice system operates independently from the government)
`wdi_gdpc`	Per capita wealth in US dollars
`undp_hdi`	Human development index (higher values = higher quality of life)
`wbgi_cce`	Control of corruption index (higher values = more control of corruption)
`wbgi_pse`	Political stability index (higher values = more stable)
`former_col`	1 = country was a colony once
`lp_lat_abst`	absolute latitude of country's captial divided by 90

Let's change these into human-readable names.

Variable	Description
`h_j`	`judiciary`
`wdi_gdpc`	`gdp`
`undp_hdi`	`hdi`
`wbgi_cce`	`corruption`
`wbgi_pse`	`stability`
`former_col`	*leave unchanged*
`lp_lat_abst`	`abs_lat`

To do this we'll use the `rename` function from the `dplyr` library.

```
library(dplyr)
world_data <- rename(world_data,
                     judiciary = h_j,
                     gdp = wdi_gdpc,
                     hdi = undp_hdi,
                     corruption = wbgi_cce,
                     stability = wbgi_pse,
                     abs_lat = lp_lat_abst)
```

Take a look at the summary of our dataset:

```
summary(world_data)
```

One issue is that our categorical variables `judiciary` and `former_col` are being treated as numerical values.

Let's transform the numerical `0`, `1` values in the `judiciary` column into a categorical values using the function `factor`.

To associate useful labels with the `0` `1` values, we define the `labels` to be applied to each `level` of category.

```
factor( x =        # the data column to convert
        levels =   # vector of category values as in column provided
        labels =   # vector of category labels to apply )
```

```
world_data$judiciary <- factor(world_data$judiciary, levels=c(0,1), labels=c('co
```

Find out how many observations in our dataset (countries) have a free judiciary.

```
table(world_data$judiciary)
controlled       free
       105         64
```

T-test (one sample hypothesis test)

A knowledgeable friend declares that worldwide wealth stands at exactly 10 000 US dollars per capita today. We would like to know whether she is right and tease her relentlessly if she isn't.

So, first we take the mean of the wealth variable `gdp`.

```
mean(world_data$gdp)
```

```
[1] NA
```

R returns `NA` because for some countries we have no reliable information on their per capita wealth. `NA` means that the data is missing. We can tell the `mean()` function to estimate the mean only for those countries we have data for, we will, therefore, ignore the countries we do not have information for.

We do so by setting the argument `na.rm` to TRUE like so: `mean(dataset_name$var_name, na.rm = TRUE)`.

```
gdp_mean <- mean(world_data$gdp, na.rm = TRUE)
gdp_mean
```

```
[1] 10184.09
```

Wow, our friend is quite close. Substantially, the difference of our friends claim to our estimate is small but we could still find that the difference is statistically significant (it's a noticeable systematic difference).

Because we do not have information on all countries, our 10184.09 is an estimate and the true population mean – the population here would be all countries in the world – may be 10000 as our friend claims. We test this statistically.

In statistics jargon: we would like to test whether our estimate is statistically different from the 10000 figure (the null hypothesis) suggested by our friend. Put differently, we would like to know the probability that we estimate 10184.09 if the true mean of all countries is 10000.

Recall, that the standard error of the mean (which is the estimate of the true standard deviation of the population mean) is estimated as:

$$S.E. = \frac{\sigma}{\sqrt{n}}$$

Before we estimate the standard error, let's get `n` (the number of observations). That is not simply the number of rows in our sample because some of the data is missing. We take the number of non-missing observations below and explain the code after.

```
n <- length(world_data$gdp[!is.na(world_data$gdp)])
n
```

```
[1] 178
```

With the function `length(world_data$gdp)` we get all observations in the data, i.e. the number of rows in the dataset. The function `is.na()` checks whether an observation is missing and the operator `!` means not. Therefore the result of `!is.na(world_data$gdp)` is a vector storing `TRUE` if the data is not `NA`, and can be used to select only these rows from the data frame `world_data$gdp`.

Note. A quick way to check for the number of NA's is: `summary(world_data)`.

Now, let's take the standard error of the mean.

```
se <- sd(world_data$gdp, na.rm = TRUE) / sqrt(n)
```

We have enough information to construct our confidence interval. Our sample is large enough to assume that the sampling distribution is approximately normal. So, we can go 1.96 standard deviations to the left and to the right of the mean to construct our 95% confidence interval.

```
# lower bound
lb <- gdp_mean - 1.96 * se
# upper bound
ub <- gdp_mean + 1.96 * se
```

```
# results
lb
[1] 8375.531
```

```
gdp_mean
[1] 10184.09

ub
[1] 11992.65
```

So we are 95% confident that the population average level of wealth is between $8375.53 US dollars and $11992.65 US dollars. You can see that we are not very certain about our estimate and we most definitely cannot rule out that our friend is right. Note that as we have made use of the 95% confidence level, in a process of repeated sampling we can expect that the confidence interval that we calculate for each sample will include the true population value 95% of the time.

We can also estimate the t-test by hand. We subtract the claim of our friend (10000) from our estimated mean and divide the result by the standard error of the estimated mean:

```
t.value <- (gdp_mean - 10000) / se
t.value

[1] 0.1995059
```

Because our sample is large, the sampling distribution of the test statistic can be approximated by a standard normal distribution, and so if `t.value` was bigger than 1.96 would imply that we could reject the null hypothesis at the 5% level. This one is much too small.

Let's estimate the precise p-value by calculating how likely it would be to observe a t-statistic of 0.1995059 from a t-distribution with n - 1 (177) degrees of freedom.

Notice, that the function `pt(t.value, df = n-1)` is the cumulative probability that we get the t.value we put into the formula if the null is true. The cumulative probability is estimated as the interval from minus infinity to our `t.value`. So, 1 minus that probability is the probability that we see anything larger (in the right tale of the distribution). But we are testing whether the true mean is different from 10000 (including smaller). Therefore, we want the probability that we see a `t.value` in the right tale *or* in the left tale of the distribution. The distribution is symmetric. So we can just calculate the probability of seeing a t-value in the right tale and multiply it by 2.

```
# p-value calculation
2* ( 1 - pt(t.value, df = (n-1) ))
```

```
[1] 0.8420961
```

The p-value is way too large to reject the null hypothesis (the true population mean is 10000). If we specified an alpha-level of 0.05 in advance, we would reject it only if the p-value was smaller than 0.05. If we specified an alpha-level of 0.01 in advance, we would reject it only if the p-value was smaller than 0.01, and so on.

Let's verify this using the the t-test function `t.test()`. The syntax of using this data to test the estimate mean of a data sample is:

```
t.test(x, mu, alt, conf)
```

Arguments	Description
`x`	Here we enter the column we are examining (in other cases we can enter a formula as in the next example).
`mu`	Here, we set the null hypothesis. The null hypothesis is that the true population mean is 10000. Thus, we set `mu=10000`.
`alt`	There are two alternatives to the null hypothesis that the difference in means is zero. The difference could either be smaller or it could be larger than zero. To test against both alternatives, we set `alt = "two.sided"`.
`conf`	Here, we set the level of confidence that we want in rejecting the null hypothesis. Common choices confidence intervals are: 95%, 99%, and 99.9% (these correspond to alpha or significance levels: 0.05, 0.01, 0.001).

```
t.test(world_data$gdp, mu = 10000, alt = "two.sided", conf=0.95)
```

```
    One Sample t-test

data:  world_data$gdp
t = 0.19951, df = 177, p-value = 0.8421
alternative hypothesis: true mean is not equal to 10000
95 percent confidence interval:
  8363.113 12005.069
sample estimates:
mean of x
 10184.09
```

The results are similar. Therefore we can conclude that we are unable to reject the null hypothesis suggested by our friend that the population mean is equal to 10000. Let's move on to a t-test to test the difference between two estimated means.

T-test (difference in means)

We are interested in whether there is a difference in income between countries that have an independent judiciary and countries that do not have an independent judiciary. Put more formally, we are interested in the difference between two conditional means. Recall that a conditional mean is the mean in a subpopulation such as the mean of income given that the country was a victim of colonialization (conditional mean 1).

The t-test is the appropriate test statistic. Our interval-level dependent variable is `gdp` which is GDP per capita taken from the World Development Indicators of the World Bank. Our binary independent variable is `judiciary`.

Let's check the summary statistics of our dependent variable GDP per captia using the `summary()`. It returns several descriptive statistics as well as the number of NA observations (missing values). Missing values mean that we have no information on the correct value of the variable for an observation. Missing values may be missing for many different reasons. We need to be aware of missings because we cannot calculate with missings.

```
summary(world_data$gdp)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  226.2  1768.0  5326.1 10184.1 12976.5 63686.7      16
```

Someone claims that countries with free judiciaries are usually richer than countries with controlled judiciaries. We know from the output of the summary function that across all countries the average wealth is 10184.09 US dollars.

We use the `which()` function from last week again to identify the row-numbers of the countries in our dataset that have free judiciaries. The code below returns the row index numbers of countries with free judiciaires.

```
which(world_data$judiciary=="free")
```

```
 [1]   9  10  15  16  20  25  31  36  38  43  44  46  47  48  49  55  57
[18]  59  60  65  75  76  77  79  81  82  83  86  88  91  92  97 101 102
[35] 113 114 116 119 122 124 125 128 138 139 143 156 157 158 159 163 167
[52] 168 169 171 174 177 180 181 182 183 184 185 186 194
```

Now, all we need is to index the dataset like we did last week. We access the variable that we want (`gdp`) with the dollar sign and the rows in square brackets. The code below returns the per capita wealth of the countries with a free judiciary.

```
mean( world_data$gdp[which(world_data$judiciary=="free")], na.rm = TRUE)
```

```
[1] 17826.59
```

Now, go ahead and find the mean per capita wealth of countries with controlled judiciaries yourself.

(You should find this gives you the value: `5884.882` .

Finally, we run the t-test. In this case we are testing whether the hypothesis that GDP is dependant on judiciary.

To tell R to test this we define the *formula* using the form *dependent variable ~ independent variable*.

In this case our formula is: `world_data$gdp ~ world_data$judiciary` . This means the test compares the mean value of GDP according to judiciary value.

We set `mu = 0` to show that our null hypothesis is that the difference in means is 0. We would change this if we are testing the hypothesis that the means of the groups is a certain distance apart.

Our alternative is again `alt = "two.sided"` reflecting that we are just testing for a difference in means (either positive or negative difference).

```
t.test(gdp ~ judiciary, mu = 0, alt = "two.sided", conf = 0.95, data=world_data)
```

```
  Welch Two Sample t-test

data:  gdp by judiciary
t = -6.0094, df = 98.261, p-value = 3.165e-08
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -15885.06  -7998.36
sample estimates:
mean in group controlled       mean in group free
                5884.882                17826.591
```

Let's interpret the results you get from `t.test()` .

The first line is `data: dependent variable by independent variable`. This tells you that you are trying to find out whether there is a difference in means of your dependent variable by the groups of an independent variable. In our example: Do countries with independent judiciaries have different mean income levels than countries without independent judiciaries?

In the following line you see the t-value, the degrees of freedom and the p-value. Knowing the t-value and the degrees of freedom you can check in a table on t distributions how likely you were to observe this data, if the null-hypothesis was true. The p-value gives you this probability directly. For example, a p-value of 0.02 would mean that the probability of seeing this data given that there is no difference in incomes between countries with and without independent judiciaries *in the population*, is 2%. Here the p-value is much smaller than this: 3.165e-08 = 0.00000003156!

In the next line you see the 95% confidence interval because we specified `conf=0.95`. If you were to take 100 samples and in each you checked the means of the two groups, 95 times the difference in means would be within the interval you see there.

At the very bottom you see the means of the dependent variable by the two groups of the independent variable. These are the means that we estimated above. In our example, you see the mean income levels in countries were the executive has some control over the judiciary, and in countries were the judiciary is independent.

To view the relationship we can look at the box plot (to plot `gdp` vs `judiciary` we can again use the `~` *formula* notation.

```
boxplot(gdp ~ judiciary, data = world_data)
```

Exercises

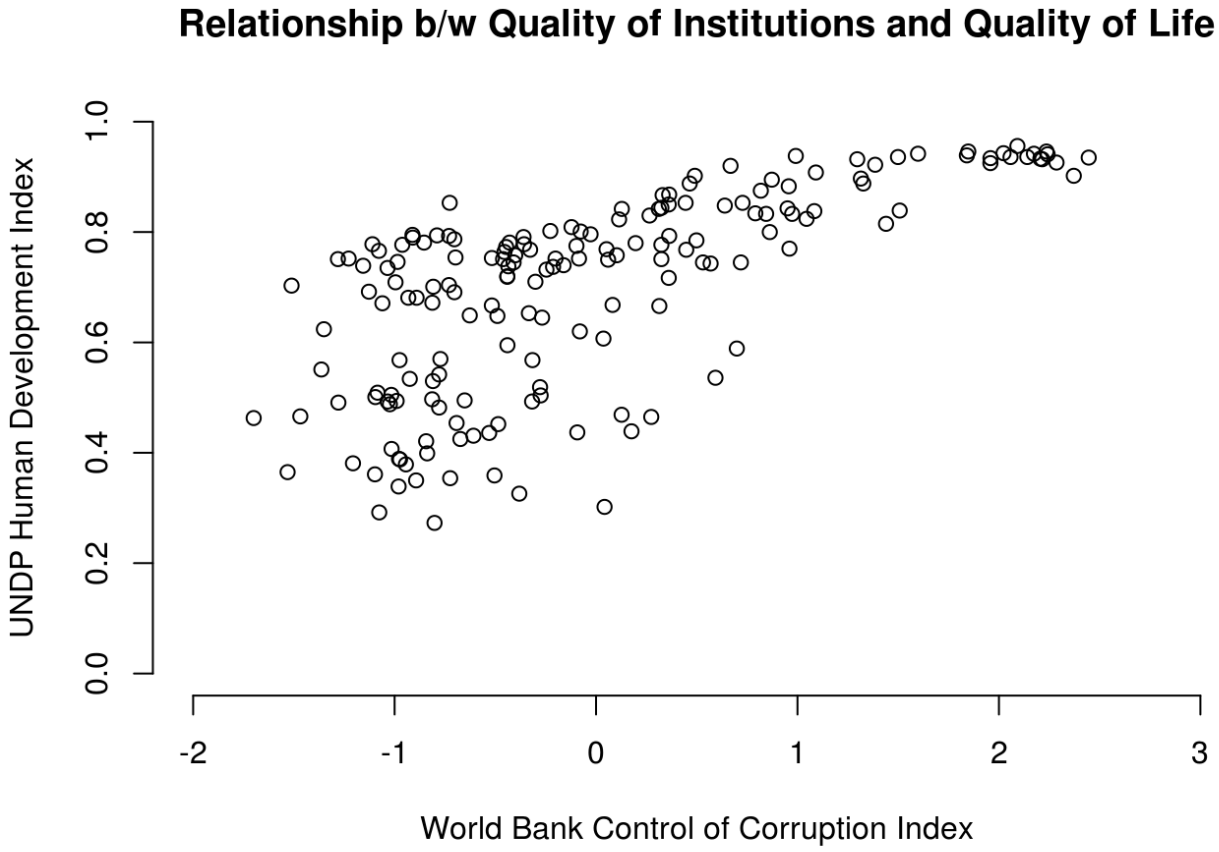
1. Turn former colonies into a factor variable and choose appropriate labels.
2. How many countries were former colonies? How many were not?
3. Find the means of political stability in countries that (1) were former colonies, (2) were not former colonies.
4. Is the the difference in means statistically significant?
5. In layman's terms, are countries which were former colonies more or less stable than those that were not?
6. Does the result of part 4. change if we choose an alpha level of 0.01?
7. How many countries have a recorded value for United Nations Development index variable `hdi`?
8. Check the claim that its true population mean is 0.85.

9. Discuss your findings in terms of the original claim. Interpret the t value, the p value, and the confidence interval.
10. We claim the difference in means in terms of political stability between countries that were former colonies and those that were not is 0.3. Check this hypothesis.
11. An angry citizen who wants to defund the Department of International Development (DFID) claims that countries that were former colonies have reached 75% of the level of wealth of countries that were not colonised. Check this claim.

Correlation

When we want to get an idea about how two continuous variables change together, the best way is to plot the relationship in a scatterplot. A scatterplot means that we plot one continuous variable on the x-axis and the other on the y-axis. Here, we illustrate the relation between the human development index `hdi` and control of corruption `corruption`.

```
# scatterplot
plot(hdi ~ corruption,
  data = world_data,
  xlim = c(xmin = -2, xmax = 3),
  ylim = c(ymin = 0, ymax = 1),
  frame = FALSE,
  xlab = "World Bank Control of Corruption Index",
  ylab = "UNDP Human Development Index",
  main = "Relationship b/w Quality of Institutions and Quality of Life"
  )
```



Sometimes people will report the correlation coefficient which is a measure of linear association and ranges from -1 to +1. Where -1 means perfect negative relation, 0 means no relation and +1 means perfect positive relation. The correlation coefficient is commonly used as as summary statistic. It's disadvantage is that it only provides information on linear relationships and if you only look at this statistic, without viewing the scatterplot, you may miss a non-linear relation.

We take the correlation coefficient like so:

```
cor(y = world_data$hdi, x = world_data$corruption, use = "complete.obs")
```

```
[1] 0.6821114
```

Argument	Description
`x`	The x variable that you want to correlate.
`y`	The y variable that you want to correlate.
`use`	How R should handle missing values. `use="complete.obs"` will use only those rows where neither `x` nor `y` is missing.

We use the `t.test` function to test a hypothesis about the mean value of a population based on a sample. Similarly we can use `cor.test` to test the correlation value against the null hypothesis that the population correlation is actually zero (variables are uncorrelated).

This can be called in the following way:

```
cor.test(~ hdi+ gdp, data=world_data, )
```

```
  Pearson's product-moment correlation

data:  hdi and gdp
t = 12.75, df = 170, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.6137236 0.7683987
sample estimates:
     cor
0.699152
```

In the case of correlation we use a one sided formula, `~ hdi + gdp` to test the correlation between `hdi` and `gdp`. This is because testing for correlation does not suppose which variable is dependent and which is independent.

We can interpret the result using similar arguments to interpreting a t.test result.

- The estimated correlation from the sample is `0.699`
- The 95% confidence limit for the population correlation is between `0.613` to `0.768`.
- The probability of finding a correlation value this far from zero under the null hypothesis (both variables independent and distributed normally) is `2.2e-16` (a very small number!).
- Therefore we find evidence to reject the null hypothesis at the 5% significance level, and accept the hypothesis that the variables are correlated.

Note in the correlation test we have options:
`alternative = "two.sided" "less"` or `"greater"`
depending on if we are testing if the correlation:
differs from zero, **is less than zero**, or **is greater than zero** respectively.

Exercises

Which columns in the dataset show a correlation that is significant at the 5% level?