LSE School of Public Policy

PART 2
PRE-ARRIVAL MATHS & STATISTICS MODULES 5-7

This is the second batch of the pre-arrival modules designed to prepare you for the courses in your core curriculum. Modules 5 through 7 follow up and delve deeper into some of the themes covered in the previous batch. All three of these new modules will prepare you to understand academic studies that supply the evidence for ‘evidence-based policymaking’.

As you make your way through these modules, you may encounter difficult concepts you won’t understand, at least not at first. This is to be expected and is nothing to worry about. Most of the information in these pre-arrival modules will be reviewed i again when you start taking your core courses.

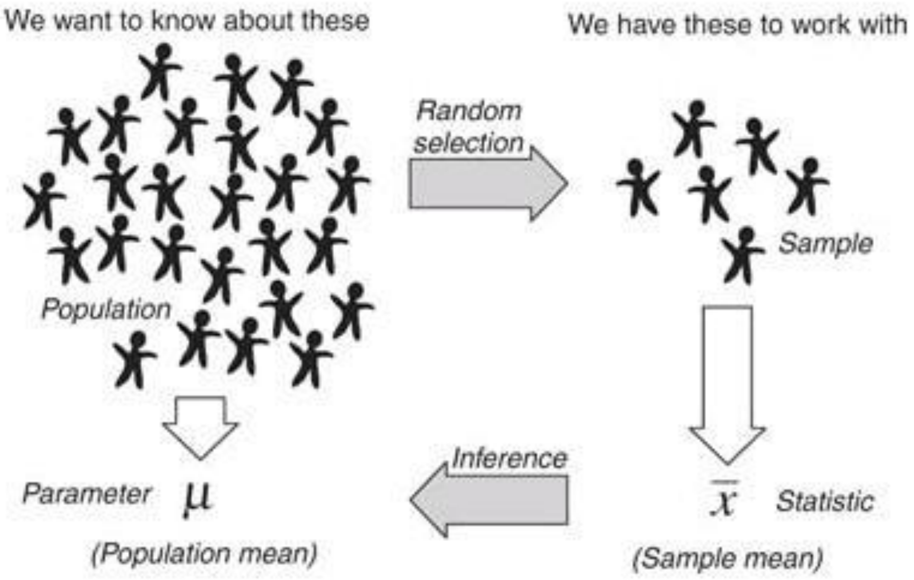
You should study these modules at your own pace, taking care to understand the various concepts you encounter. Feel free to review the modules you need and skip the others. Key concepts or terms are put in italics. Practice problems can be found at the end of each module which you can check using the answer key provided in a third pdf.

Table Of Contents

Module 5: Statistical Inference & Central Limit Theorem

Statistical inference is a procedure, in which we use what we know about the data from a sample to infer what is likely to be true about the whole population. This process of statistical inference involves some degree of uncertainty because we try to learn about the population using only a sample of data from that population. For instance, we might be interested in the average age of students in higher education institutions in the UK. Currently there are 2.34 million of them. But we may only have access to the age data of a small subset (sample) of these students, for instance, 1,500 of them. Then, we would need to make an inference about the average age of all students in higher education in UK using the age data of the students in our sample. Or, we might be interested in the level of support for Brexit in UK amongst all British voters, which is close to 46 million. We may have access to the data on the views of only a subset of these voters. Then we would infer the level of support for Brexit in the entire British electorate using the views of only a subset of these voters.

Figure 1: Population vs. Sample



As the Figure 1 illustrates, we are interested in learning about a population. Most of the time, this is the mean value of some variable, called μ . If we have access to data from a sample of this population that is randomly selected,[1] then, using the sample mean, called $\bar{X}$, , we can make an inference about the population mean, μ

Since we try to learn about a population using only a subset of data from that population, the process of statistical inference necessarily involves some uncertainty. But we also know how to measure this uncertainty thanks to the a very important tool of inference called 'The Central Limit Theorem'. Here is an example to understand this very important theorem.

5.1 An Example: Flipping Coins

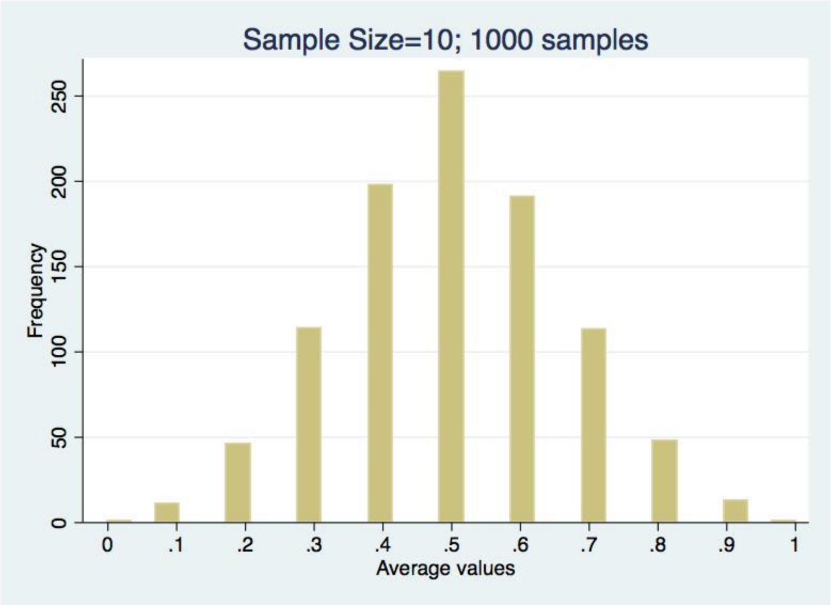
Let's think about the following experiment: Suppose we have a fair coin, and we flip it 10 times. Every time it comes up heads, we code it as 1. Every time it comes up tails, we code it as 0. After 10 flips, we calculate the average of these 0's and 1's. This average will be a number between 0 and 1. If its a fair coin, we should expect it to come up heads 5 times and tails 5 times so the average would be .5. But this is not necessarily the case in any set of 10 flips. Purely by chance, it may come up as tails each time we flip it. Then the average would be 0. Or it may come up as heads 10 times in a row. Then the average would be 1. Or it may be some other number between 0 and 1 depending on how many times we get tails and heads.

Now suppose we repeat this exercise 1000 times. In other words, we flip the same coin 10 times, and after each set of 10 flips, we calculate the average of those 10 flips. And we do this 1000 times. After having done this 1000 times, we plot the distribution of these

[1] Random selection means that every unit in the populatoin has the same probability of being selected into the sample.

1000 average values using a histogram. Figure 2 shows the histogram for this exercise. While .5 is the most frequent outcome, there are also cases of averages different than .5. For instance, there are close to 200 cases of averages of .4 and .6. These are cases with 6 tails and 4 heads, and 4 tails and 6 heads, respectively. As the cases become less likely, they also become less frequent. The cases where in all 10 flips we get tails or in all 10 flips we get heads are very rare. These are the cases with averages of 0 and 1.

Figure 2: Flipping Coins

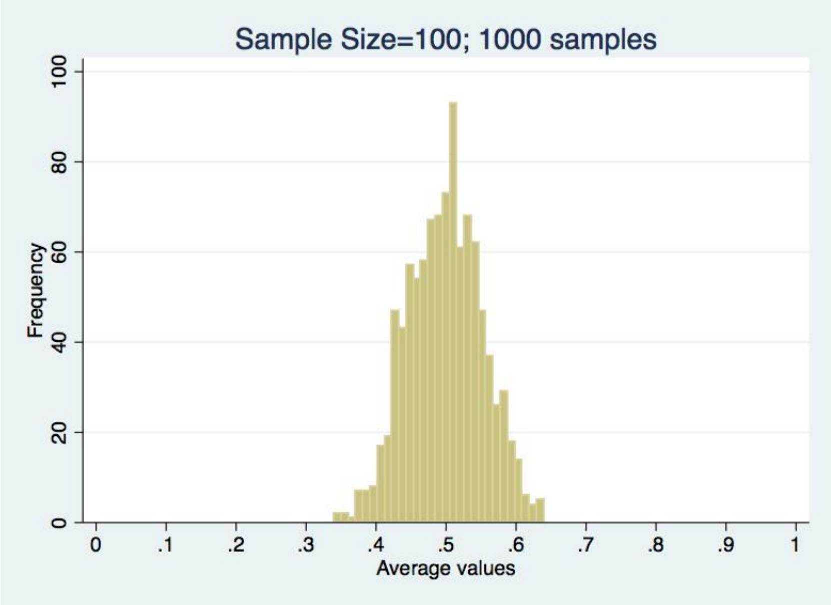


Note that this experiment is an exercise in inference. Our population is the cases of all possible coin flips in the world. We are interested in the average value of those flips, where head is 1 and tail is 0. And our sample is the set of flips we do. We call this our sample because we do not flip the coin an infinite number of times, but we do it only 10 times. So our sample size is 10. And because we do this 1000 times, we have 1000 such samples.

Now suppose we repeat the same exercise, but this time, we flip the coin 100 times.

And we calculate the average of these 100 flips. And we repeat the same exercise 1000 times. We should expect the average to be .5 again but now we will observe values closer to .5 since we flip the coin 100 times and then take the average. Figure 3 shows the histogram for this exercise. .5 is the most frequent outcome but we also observe averages other than .5. As we get further away from .5, those average values become less and less frequent as it becomes increasingly unlikely that we get e.g. only 30 heads and 70 tails. Note that now we have a sample size of 100 and we have 1000 such samples.

Figure 3: Flipping Coins



Suppose we repeat the same exercise and continue to increase our sample size to 1,000, then to 10,000 and to 100,000. Figures 4, 5 and 6 show the histograms of the sample averages for these three different sample sizes. As the sample size grows, the averages are more and more closely centered around .5.

As the sample size grows, we observe another pattern in the distribution of these averages, which is highlighted in Figure 7, which shows the distribution of the average

Figure 4: Flipping Coins

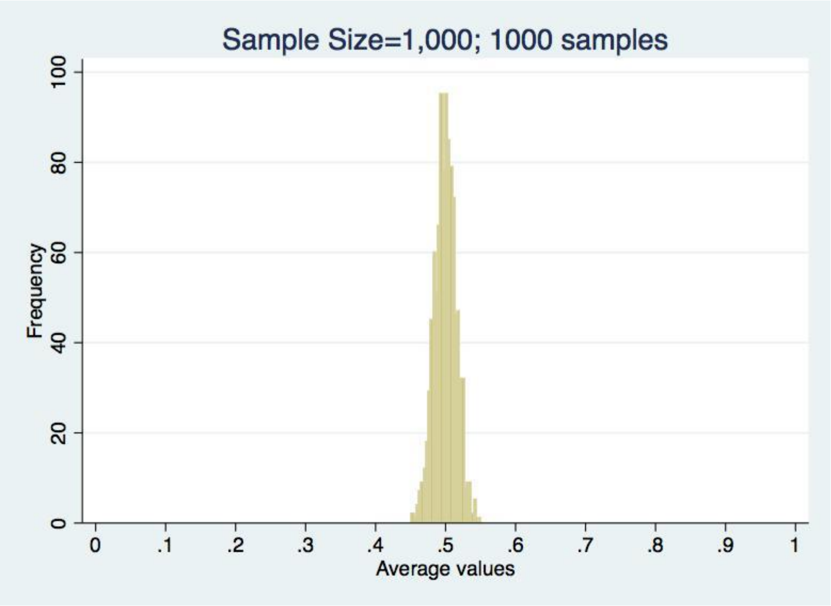


Figure 5: Flipping Coins

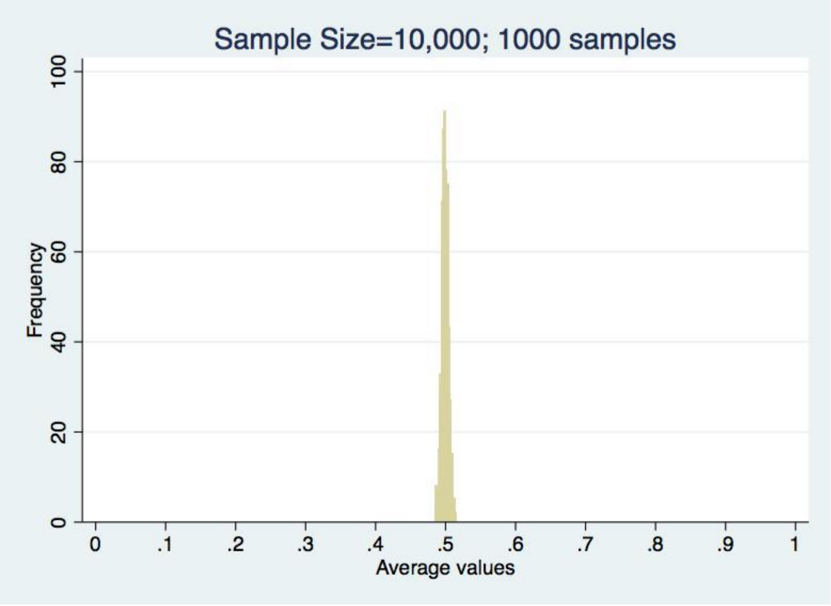


Figure 6: Flipping Coins

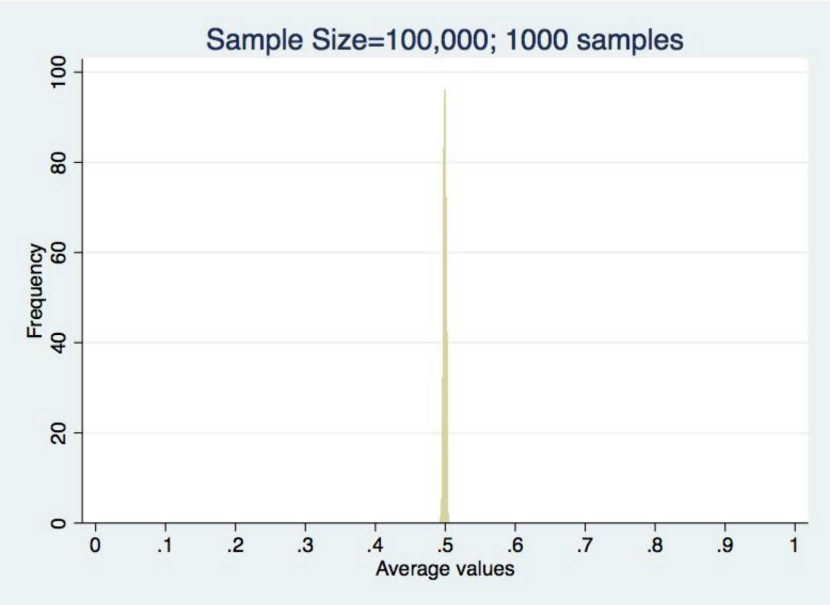
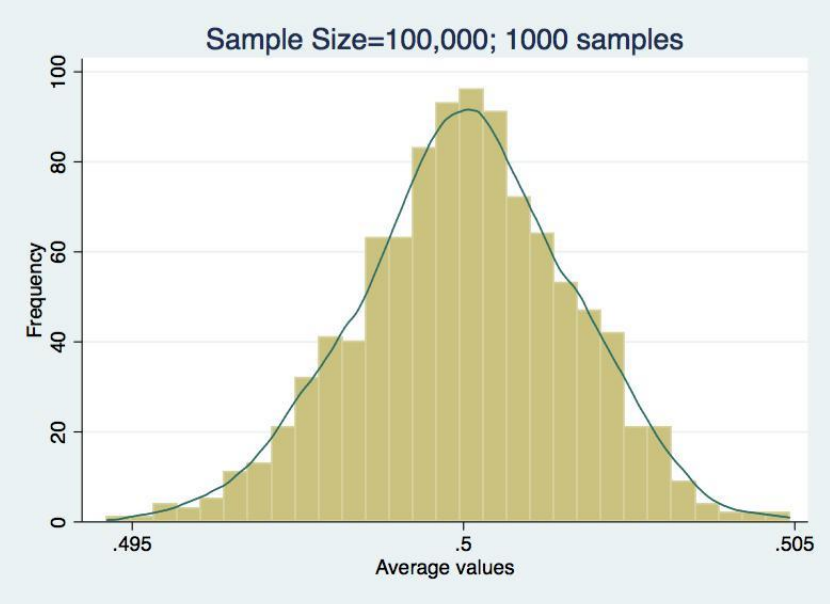


Figure 7: Flipping Coins



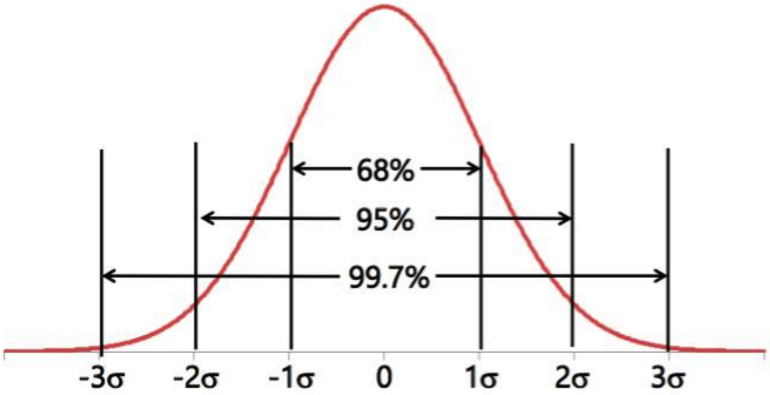
values with a sample size of 100,000 just like in Figure 6 but more closely, only displaying the frequency of values that actually occur. As you can see, the distribution of these averages has a bell-shape! The central limit theorem also tells us that this distribution has a mean, which is equal to the population mean, and its standard deviation is equal to $\frac{\sigma}{\sqrt{n}}$ where σ is the standard deviation of the population, and n is the sample size. Most of the time, we will not know the standard deviation of the population, but we can estimate it pretty accurately with the standard deviation of the sample, which we denote by s. The standard deviation of the sample means $\frac{s}{\sqrt{n}}$ is called *standard error of the mean*, or simply *standard error*.

The great beauty of the central limit theorem is that no matter what the distribution of the underlying variable is, the sample means will be normally distributed. To get back to our example of coin flips, the distribution of the variable that denotes heads and tails is clearly not normal, it takes the value of 0 50% of the time, and it takes the value of 150% of the time. But the distribution of the sample means will be normal.

Why is this theorem so useful? For even modest samples drawn from an unknown distribution, the distribution of the sample mean is approximately normal and we know a lot about the normal distribution. Most importantly, we know where observations are expected to be. Normal distribution is *symmetrical around the mean* such that the mode, median, and mean of the distribution are the same. Another property of the normal distribution is that it has a predictable area under the curve within specific distances of the mean. Starting from the mean, and going one standard deviation within each direction will capture 68% of the area under the curve. Going two standard deviations will capture a little more than 95% of the area. Going three standard deviations will capture over 99% of the area. Figure 8 shows one such distribution with a mean of 0.

So even if we have data from only one sample, we can calculate the mean from that

Figure 8: Normal Distribution



sample and construct an interval by adding two standard errors to the mean and subtracting two standard errors from the mean. The resulting interval will contain the true population mean 95% of the time. To understand why, think about how we would get different sample mean values each time we flip the coin 100,000 times. Each time we flip the coin 100,000 times, we can construct this interval using the sample mean and standard error values we get. And because these sample means are normally distributed, we know that we will get sample mean values that are within two standard errors of the true population mean 95 percent of the time. Therefore, the intervals that we construct by adding and subtracting two standard errors of the mean will also include the true population mean 95 percent of the time.

At this point, you may ask yourself: But why don't we add three standard errors to the mean to construct an interval that contains the true mean 99 percent of the time instead of 95? There is nothing wrong with that except the fact that the resulting interval would be wider. We would have more certainty about where the population mean is but the range of possible values of the true mean would also be larger.

So far, we only focused on learning about a single variable in a population, to understand the logic of statistical inference. We might be also interested in relationships between two variables in a population. The same logic of inference also applies to learning about relationships between two variables in a population. To understand how that works, it is useful to introduce a new tool: hypothesis testing.

Practice Questions for Module 5

1. Please fill in the blanks in the following statement:

Statistical inference is a procedure, in which we use what we know about the datafrom a __________to infer what is likely to be true about the__________ .

2. For sufficiently large samples, the sample means of a variable that is not normallydistributed will be normally distributed.

   a) True

   b) False

3. The population has a mean of 14 and a standard deviation of 3. The sample size is$n$=1,000. What is the mean of the distribution of the sample means from that population?

4. The population has a mean of 120 and a standard deviation of 12. The sample sizeis n=16. What is the standard error of the mean?

Module 6: Hypothesis Testing & Regression

Classical hypothesis testing begins with a null hypothesis, which is a statement about the population that we are interested in. For instance, we can have a hypothesis about a single variable, such as ‘the coin is fair’, or 55 percent of British electorate support Brexit. Or it can be about a relationship between two variables. For instance, a labour training program has no effect on wages. Or the political regime of a country has no effect on economic growth.

So if the null hypothesis is that political regime has no effect on economic growth; the hypothesis that it has an effect is the *alternative hypothesis*. We reject the null hypothesis if we have evidence ‘beyond a reasonable doubt’ against it. If we do reject the null hypothesis, that means we are fairly confident that type of political regime has an effect on economic growth.

One common element across a number of hypothesis tests is the *p-value*, where *p* stands for probability. P-values are a useful tool to describe the results of a hypothesis test. Suppose we are interested in whether there is a relationship between two variables, *X* and *Y*: When we test whether there is a relationship between *X* and *Y* with a sample of data, we compare the relationship we observe between X and Y in the sample with what we would expect to see if there were no relationship between *X* and *Y* in the entire population. The more different the observed relationship is from what we would expect if there were no relationship in the population, the more confident we are that there is a relationship between two variables.

The *p-value*, which ranges between 0 and 1, is a tool that tells us the probability of observing the relationship in a sample purely by chance, i.e. it is the likelihood that we observe the empirical relationship between two variables in a sample if there were no relationship between these two variables in the population. Therefore, the lower the *p-value*, the more confident we are that there is a relationship.

6.1 Statistical Significance

Another way to refer to a relationship between two variables which has a low *p-value* or is unlikely to be due to chance is to say that the relationship between the two variables is *statistically significant*. There are different standards according to which a *p-value* indicates statistical significance. The most common one is to call relationships with a *p-value* of less than .05 significant. Some use a more stringent standard of .01, some use a less stringent standard of .1. These different critical levels for *p-value* are called *significance levels*. More recently, scholars have started to simply report the p-value and let the readers evaluate the significance of the relationship.

Going back to the discussion of null and alternative hypotheses, if our null hypothesis is that there is no relationship between *X* and *Y*, p-value and the level of statistical significance of the relationship between two variables tell us how confidently we can reject a null hypothesis of no relationship between *X* and *Y*.

Before we move on, there are two critical points that need to be emphasized with respect to p-values and statistical significance: First, low p-values or a statistically significant relationship does not mean that there is necessarily a strong relationship between two variables. Second, the fact that there is a significant relationship between two variables does not mean that the relationship is *causal*. We will discuss strategies to establish a causal relationship between variables in the next module.

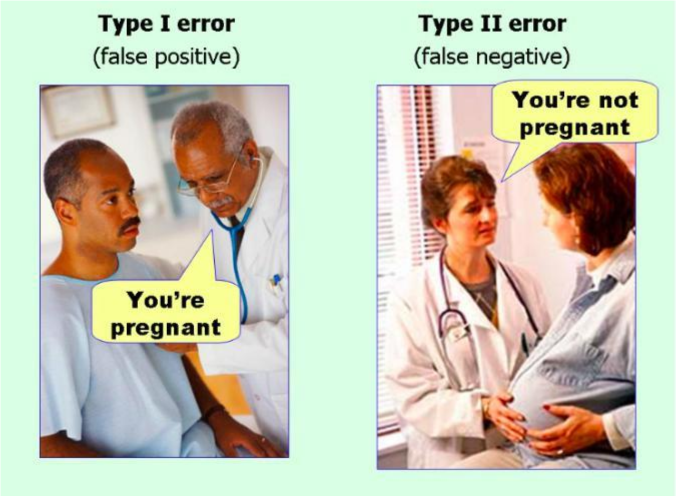
6.2 Hypothesis Testing: Mistakes

There are two different types of error one can make in testing hypotheses and deciding whether to reject or fail to reject a null hypothesis: A *Type-I* error means that you reject the null hypothesis when it is in fact true, e.g., we conclude the coin is not fair when in fact it is fair.

The probability of a *Type-I* error is denoted by the Greek letter α.

A *Type-II* error occurs when we fail to reject the null hypothesis although it is false, e.g., we do not conclude the coin is unfair when it is indeed unfair. The probability of a *Type-II* error is denoted by the Greek letter β. Below is a figure that illustrates the two types of errors with an example that hopefully makes it easier for you to remember!

Figure 9: Hypothesis Testing: Mistakes



Type-II errors are such that we fail to reach a conclusion about a significant relationship between variables although there is indeed a significant relationship. But in the case of *Type-I* errors, we do wrongfully conclude that there is a relationship. Therefore, generally, *Type-I* errors are considered more serious.

We should also note that there's a trade-off between the probability of *Type-I* and *Type-II* errors. As we want to have stronger evidence to reject a null hypothesis, this makes it less likely that we reject a true null, but at the same time, it increases our chances of failing to reject a false null. Remember that when we do the hypothesis test, i. e. engage in statistical inference, we do not know whether the null is true or not.

6.3 Hypothesis Testing: Back to the Coin Toss Example

Let’s go back to the coin toss example to understand the logic of hypothesis testing. The null hypothesis might be that the probability of heads is one-half, i.e. the coin is fair. Formally this would be stated as $H_0: E(X) = 0:5$. The alternative is usually the negation of the null. In our example, the typical alternative, denoted by H_1 is: H_1: $E(X) \neq 0:5$.

Suppose we wanted to determine whether a coin is fair. We toss it 10 times. It comes up heads 9 times. In light of this result, we would be inclined to reject the null hypothesis that the coin is fair. In fact, the probability of a fair coin coming up as heads 9 more times out of 10 tosses is about 1 percent. The logic of hypothesis testing for various types of hypotheses follows the exact same reasoning we just had regarding the coin toss example.

We construct a test *statistic* that summarizes the relationship between our observed data and the null hypothesis. In the case of the coin toss example, it would be the prob-ability of getting 9 or 10 heads out of 10 times. Then we construct a *critical region* or *rejection region* based on our null hypothesis. In our example, this means that we decide in advance for how many numbers of heads we would reject the hypothesis that the coin is fair. If the test statistic lies in this critical region, we will reject the null hypothesis. So if we decide that the rejection region is 9 or more heads, having observed 9 heads out of 10 tosses, we reject the null hypothesis of a fair coin.

Let’s see how hypothesis testing works for some of the most commonly used types of hypotheses.

6.4 Hypothesis Testing: Testing for Means of the Population

6.4.1 When the Variance is Known

Suppose we observe a random sample of 40 professional basketball players. And we want to know whether the average height of all professional basketball players is 2 meters or 200cms. Our null hypothesis is then H_0: $\mu = \mu_0$ = 200. Our alternative hypothesis is that $H_1 : \mu \neq 200$. Using the heights data of 40 players, we calculate the sample mean $\bar{X}$. Suppose, $\bar{X}$ = 205. From the central limit theorem, we know that the sample means follow the normal distribution and has variance $var(\bar{X}) = \frac{\sigma_x^2}{n}$.

In statistics, the notation for a normally distributed variable X is X ~ N (mean, variance). To test our hypothesis, we start from the assumption that the null is true, such that $\bar{X} \sim N(\mu_0, \frac{\sigma_x^2}{n})$.

Note that we do this because we want to understand how likely it is for us to observe the mean height in our sample if indeed it comes from a distribution with a mean of 200 cms. Now, we need to find out how likely it is to observe this sample mean.

How do we do that?

The special case of a normal distribution with a mean of 0 and variance of 1 is called *the standard normal distribution*. A variable *Z* with a standard normal distribution is de-noted $Z \sim N(0, 1)$. as *Z* ~ N(0,1).

If a variable has a normal distribution with a mean μ and standard deviation σ, then the $Z = \frac{X-\mu}{\sigma}$ has a standard normal distribution. Why is this a useful result? It is useful because now we can use the distribution of the variable *Z* when we need to find out the likelihood of observing specific values of a variable that is normally distributed, such as the sample means!

To see how this works, let's look at the so-called *Z-table*, which shows the likelihood of observing values less than or equal to specific values of *z*.

Figure 10: Probabilities and Z-Table

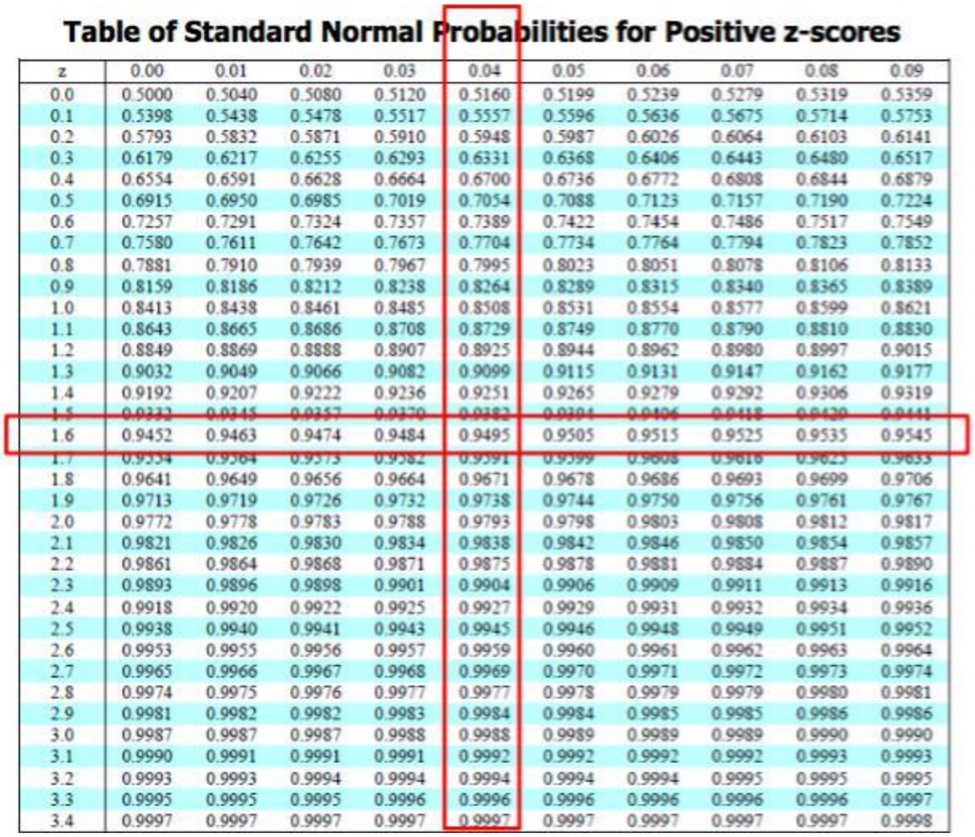
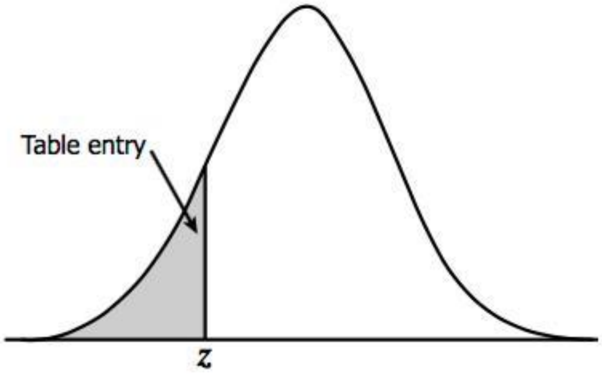


Table of Standard Normal Probabilities for Positive z-scores

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998

Figure 11: Standard Normal Distribution



Looking at the values in Table 10, we can see that the probability of observing a value less than or equal to 1.64 is about 95 percent. Because the distribution of z-scores is symmetric, we can also use this table to find out the probability of z scores less than or equal to a negative value. For instance, the probability of z-values less than 1.5 is equal to 0.9332. This means that the probability of observing z scores less than or equal to -1.5 is 1-0.9332=0.0668. Second, the likelihood of observing values less than -z and larger than z are equal to each other. For instance, as you can read from the table, the probability of observing a value less than -1.5 is 0.0668. The probability of observing z values larger than 1.5 is also 0.0668.

Going back to the example, we know that has $z = \frac{\bar{X} - \mu_0}{\sigma_x/\sqrt{n}}$ a standard normal distribution. Suppose we know in advance the standard deviation of the heights of all the professional basketball players. Let's say it is σ_x=10 cm. Then we can plug in this into the formula for the z-statistic and get $z = \frac{\bar{X} - \mu_0}{\sigma_x/\sqrt{n}} = \frac{205 - 200}{10/\sqrt{40}} = 3.16.$ The chances of observing a z statistic of 3 or larget or -3 and smaller is about .0016. We calculate this by looking at Table 10 and reading the value at row '3.1' and column '0.06'. The value is 0.9992. So, 2 (1-0.9992)=0.0016. This means that the chances of observing the sample mean of 205 if the true population mean of heights is 200 (the *p-value*) is very small. So we reject quite confidently the null hypothesis that the average height of these players is 200 cms.

How do we decide the level of probability below which we reject the null hypothesis? There is no fixed r ule. There are different thresholds used for this purpose. The most commonly used threshold level is .05. In other words, if the z score is such that the chances of observing values less than -z or larger than z together is less than .05, we reject the null hypothesis at 5% significance level. Otherwise, we 'fail to reject' it.

6.4.2 When the Variance is Not Known

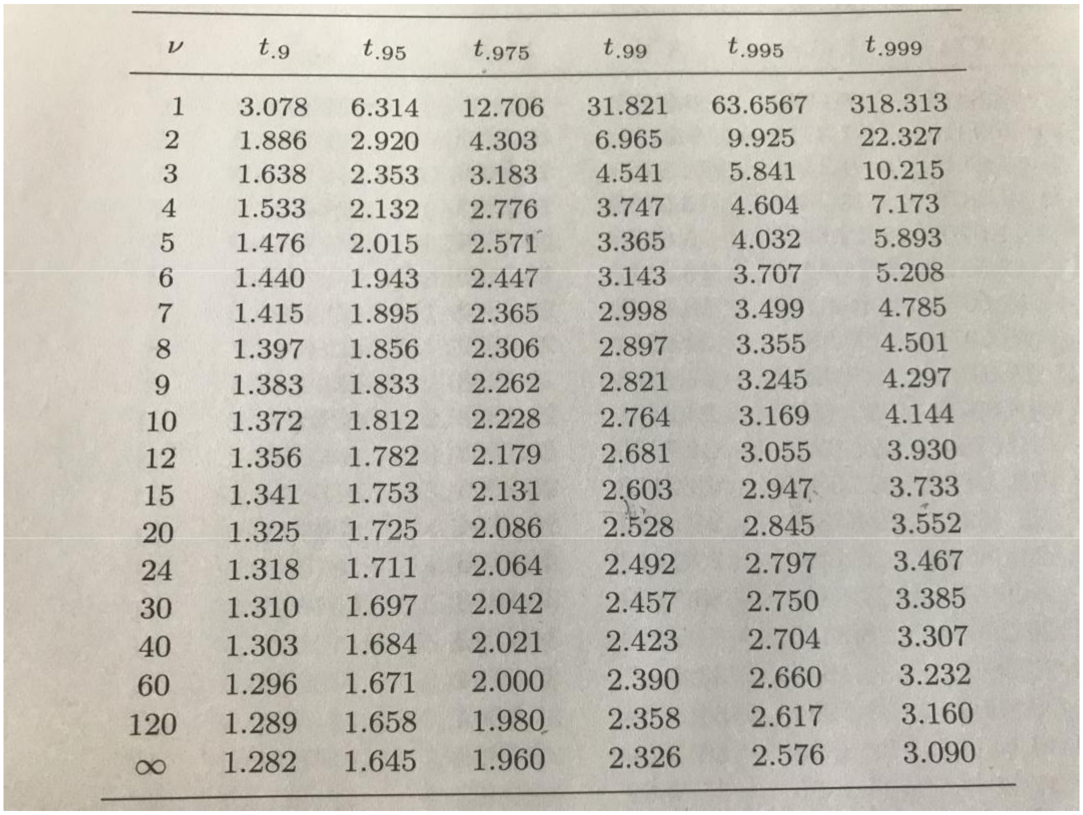
Most of the time, we will not know the population variance. If we don't have access to the whole population data, it is likely that we will not know the variance of the values in that population. In those cases, we need to use an estimate of the population variance in order to compute the test statistic and perform the hypothesis test. The estimate of the population variance is the same variance $s_X^2 = \frac{1}{n-1}\sum_n (X_i - \bar{X})^2.$

Similar to before, now we form t-statistic: $t = \frac{X - \mu_0}{s_X/\sqrt{n}}$

The logic of the hypothesis test is the same. The only difference is that now the distribution of the statistic we compute follows a slightly different distribution, the *t-distribution*, which is symmetric like the normal distribution but has heavier tails. The probability of observing a t-statistic greater than or equal to a particular value depends on the *degrees of freedom*. Degrees of freedom captures the idea that as the amount of data we use for inference increases, we have more confidence that the pattern we observe in this sample reflects what is true about the underlying population. The degrees of freedom for the t-statistic in testing for the population mean is equal to the number of observations (n) - 1. So in our example, the degrees of freedom is 100-1=99.

In Table 12, we can see the probabilities of obtaining values less than or equal to spe-cific t-scores. For instance, with 10 degrees of freedom, there is a 90 percent probability of obtaining a value less than or equal to 1.372. There is a 95 percent probability of obtaining a value less than or equal to 1.812. These t-scores depend on the degrees of freedom. For different degrees of freedom, we have different values for each probability. Note also that with larger degrees of freedom, the critical values for each probability become smaller.

Figure 12: Probabilities and t-Table



ν	$t_{.9}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$
1	3.078	6.314	12.706	31.821	63.6567	318.313
2	1.886	2.920	4.303	6.965	9.925	22.327
3	1.638	2.353	3.183	4.541	5.841	10.215
4	1.533	2.132	2.776	3.747	4.604	7.173
5	1.476	2.015	2.571	3.365	4.032	5.893
6	1.440	1.943	2.447	3.143	3.707	5.208
7	1.415	1.895	2.365	2.998	3.499	4.785
8	1.397	1.856	2.306	2.897	3.355	4.501
9	1.383	1.833	2.262	2.821	3.245	4.297
10	1.372	1.812	2.228	2.764	3.169	4.144
12	1.356	1.782	2.179	2.681	3.055	3.930
15	1.341	1.753	2.131	2.603	2.947	3.733
20	1.325	1.725	2.086	2.528	2.845	3.552
24	1.318	1.711	2.064	2.492	2.797	3.467
30	1.310	1.697	2.042	2.457	2.750	3.385
40	1.303	1.684	2.021	2.423	2.704	3.307
60	1.296	1.671	2.000	2.390	2.660	3.232
120	1.289	1.658	1.980	2.358	2.617	3.160
∞	1.282	1.645	1.960	2.326	2.576	3.090

Suppose the standard deviation in our sample is equal to s=9. Then, the t-statistic is

$t = \frac{205 - 200}{9/\sqrt{40}} = 3.51.$. Looking at Table 12, we see that it does not report probabilities for 9= 40

100 degrees of freedom but we see that the probability of obtaining values less than -3.16 or larger than 3.16 are .002 for 120 degrees of freedom, and the probability of obtaining values less than -3.232 or larger than 3.232 are .002 for 60 degrees of freedom. So the crit-ical values for obtaining the same probabilities with 100 degrees of freedom must be a number in between. Therefore, the chances of observing values less than -3.51 or greater then 3.51 should be less than 0.002, which means that the p-value is very low. We can con-fidently reject the null hypothesis that the average height of basketball players is equal to 200cms.

6.5 Difference of Means Test

Suppose we want to know whether there is a relationship between being a professional basketball player and height. Suppose, in addition to the sample of data we have for the players, we also have a random sample of data from 100 basketball fans. Let's assume that the average height of the fans is 175 cms and the standard deviation is 12 cms. To see if this difference of heights is significant, we conduct a *difference of means* test.

Let's call the mean height of professional players $\bar{X}_1$ and the mean height of fans $\bar{X}_2$.

Let's also call the standard deviation of the heights of players s_1 and the standard deviation of the fans s_2. n_1 is the sample size of players, which is 40, and n_2 is the sample size of fans, which is 100. We construct the following t-statistic:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{(s_1^2/n_1) + (s_2^2/n_2)}}$$

If we plug in all the values in, we get

$$t = \frac{205 - 175}{\sqrt{(81/40) + (144/100)}} = 16.12$$

The degrees of freedom for the difference of means test is n_1 - 1 + n_2 - 1 = 138. Looking at Table 12, it is clear that the chances of observing this statistic value if there was no significant difference in means is very very small so we confidently reject that there is no difference.

Can we conclude based on this evidence that there is a causal relationship between playing professional basketball and height? In other words, can we conclude that play-ing professional basketball makes you taller? We certainly can't, for a number of reasons:

First, despite the statistically significant relationship between playing professional bas-ketball and height, this might also be because taller individuals are more likely to play and succeed in professional basketball. This is an example of *reverse causality*. Second, it is possible that those who are professional basketball players are more likely to be male (simply because there are more male players) and because men are taller than women, basketball players might be taller than the general population. Or, players are younger than the general population, and therefore on average taller than the population of basketball fans. These are examples of *omitted variable bias*: Our difference of means test omits the effect of gender and age on likelihood of playing basketball and height.

So how do we run statistical analyses that allow us to establish the causal effect of an independent variable X on a dependent variable Y? Until recently, the main tool for this purpose used to be the *regression analysis*, which we will discuss at the end of this module. We should keep in mind that in the last twenty years or so, researchers have started to utilize a new set of tools in addition to regression analysis to establish causal relationships, which we will discuss in the next module.

6.6 Confidence Intervals

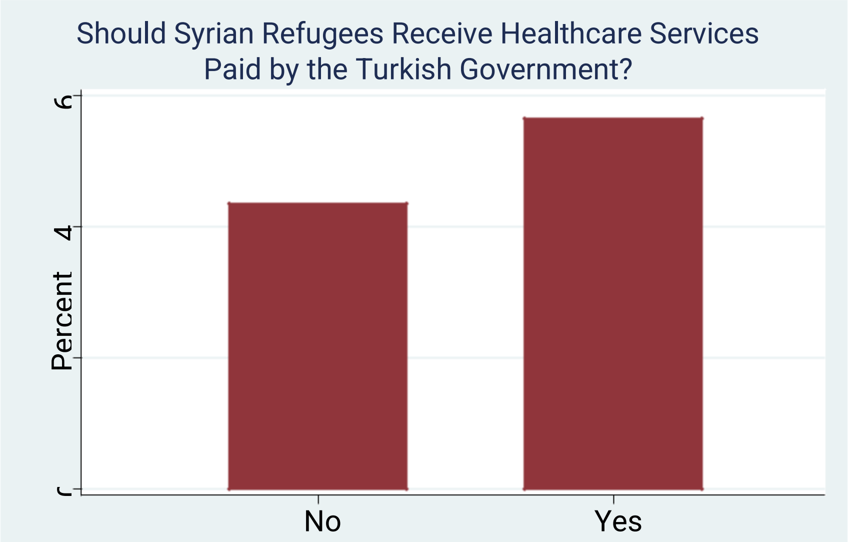
Before we move on to regression analysis, let's discuss one last concept widely used in statistical analysis: confidence intervals. Confidence intervals report the range of values where we think the true parameter lies. While the sample mean gives us a point estimate for the population mean, we usually want to convey how certain we are about this estimate. An x%-confidence interval for the population mean is a range for which we can say with x% confidence that it contains the true population mean. The most widely used confidence interval is 95 % confidence interval. When it comes to inference about population mean of a variable X, in general, we can say with 95% confidence that the interval

$\bar{X} \pm 2\frac{s_x}{\sqrt{n}}$ contains μ. Let's see how the idea of confidence interval helps in practice in survey research:

6.6.1 Example: Turkish Public Opinion on Syrian Refugees

In a survey conducted in 2014, a sample of 1221 Turkish citizens were asked whether they support Turkish government paying for Syrian refugees' healthcare.[2] 43.5% of the **sample** (531 people) don't support it while 56.5% of the **sample** (690 people) support it. Figure 13 shows the histogram of the answers.

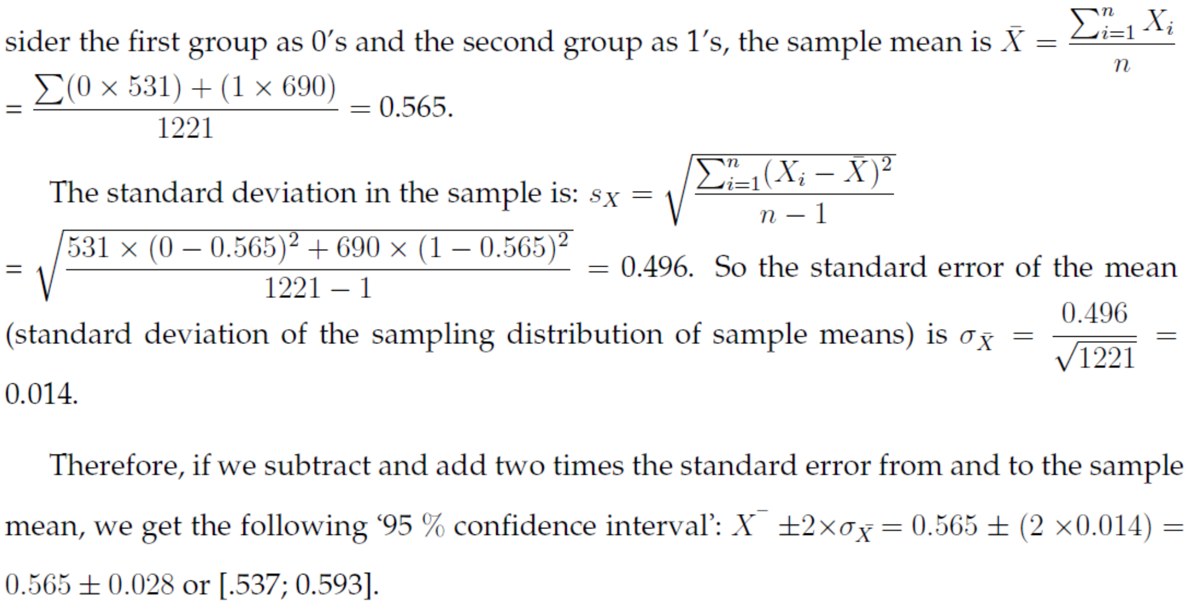
Figure 13: Turkish Citizens' Views on Government Support for Syrians in Turkey



Number of Respondents: 1221
Source: Sinmazdemir et al. 2016. Refugees, Xenophobia, and Domestic Conflict: Evidence from a
Survey Experiment in Turkey

The question that we are interested in is as follows: What percentage of the Turkish population supports the government to pay for the healthcare of Syrian refugees? If we con -

[2] The details of the survey can be found in the following article: Getmansky, Anna, Tolga Sinmazdemir, and Thomas Zeitzoff. 2018. 'Refugees, Xenophobia, and Domestic Conflict: Evidence from a Survey Experiment in Turkey' *Journal of Peace Research*, Volume: 55, Issue: 4. There are currently about 3.5 million Syrian refugees in Turkey who have fled the civil war in Syria.



sider the first group as $0'$s and the second group as $1'$s, the sample mean is $\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$ $= \frac{\sum(0 \times 531) + (1 \times 690)}{1221} = 0.565.$

The standard deviation in the sample is: $s_X = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}}$ $= \sqrt{\frac{531 \times (0 - 0.565)^2 + 690 \times (1 - 0.565)^2}{1221 - 1}} = 0.496.$ So the standard error of the mean (standard deviation of the sampling distribution of sample means) is $\sigma_{\bar{X}} = \frac{0.496}{\sqrt{1221}} =$ 0.014.

Therefore, if we subtract and add two times the standard error from and to the sample mean, we get the following '95 $\%$ confidence interval': X^{-} $\pm 2 \times \sigma_{\bar{X}} = 0.565 \pm (2 \times 0.014) =$ 0.565 ± 0.028 or [.537; 0.593].

Using the Central Limit Theorem, we can say with 95% confidence that the percentage of Turkish population that support their government paying for healthcare for Syrian refugees is between 53.7% and 59.3%.

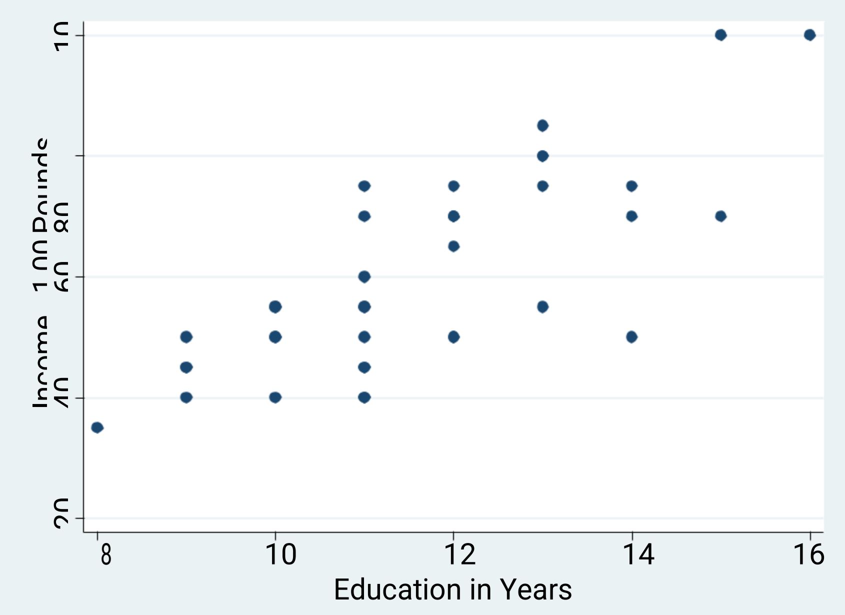
Remember this is only a sample of 1221 Turkish citizens. The population of Turkey is about 80 million! So there is some uncertainty around our estimate of 56.5% of Turkish citizens supporting their government to pay for Syrians' healthcare. But despite that, it is quite remarkable that we can have a fairly reasonable range of values for the level of support in a population of 80 million based only on a single sample from that population, which is 1221 people, or about 0.0015 percent of the total population!

6.7 Regression

6.7.1 Bivariate Regression

To understand how regression works, let's start with an example with only one independent variable. A regression of this type is called bivariate regression, meaning that we have a two-variable regression, with one dependent and one independent variable. The main idea of bivariate regression is that we are 'fitting the best line' through a scatterplot of data. Suppose we have income and education data for a sample of 40 professionals living in London. Figure 14 displays a scatter plot of this data. Education is measured in years of education completed. Income is measured in thousands of pounds. Each dot on the plot denotes one of the individuals in the sample. There is large variation in income levels: While some individuals have low income, below 40,000 pounds, some have much higher incomes, for instance 100,000 pounds. It is clear that there is a strong, positive relationship between years of education and income. The key question is how to estimate the impact of each additional years of education on income and how significant this impact is. This is where regression analysis comes in.

Figure 14: Scatterplot of Data: Income and Education



Mathematically, regression analysis involves the representation of a relationship be-tween two variables X and Y. In the case of bivariate regression, we represent this relationship with the following equation:

$$Y_i = \alpha + \beta X_i + \epsilon_i$$

Plugging in the names of the variables we use in our example, this becomes:

$$\text{income}_i = \alpha + \beta\ \text{education}_i + \epsilon_i$$

Here, the equation represents how we model the determinants of income. It is simply a function of education and some other random factors, which we summarize with ϵ . α is the intercept, which in substantive terms captures the income of an individual with no education.[3] $\beta$is the parameter that captures the relationship between education and income. If β is positive, higher education is associated with higher income. If it turns out to be negative, higher education is associated with lower income, which is very unlikely.

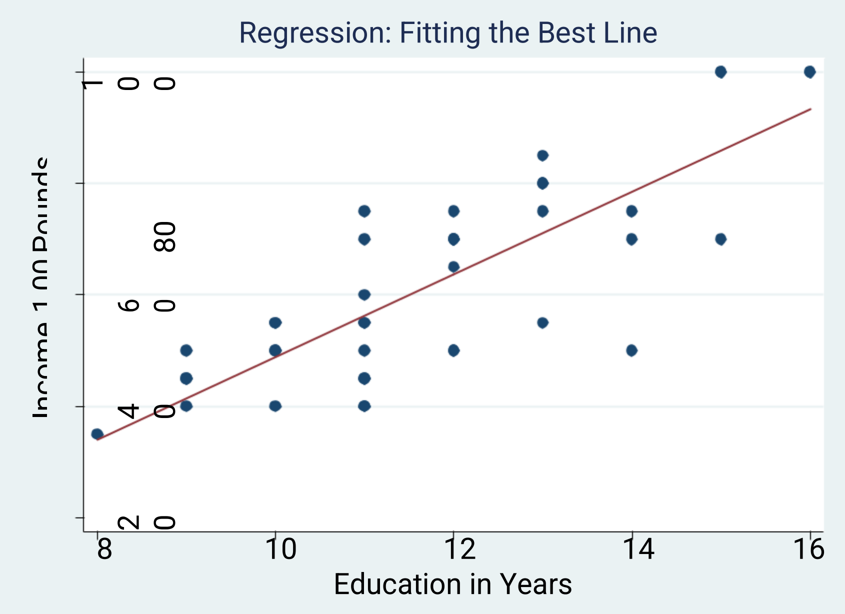
In a regression analysis, the goal is to draw a line that describes the relationship be-tween the independent and the dependent variable as closely as possible. In this example, our goal is to draw the line that comes as close as possible to all the points in our scatterplot of income and education values of individuals in our sample. Mathematically, this is the line that minimizes the distance between the line and all the data points. It is not possible to draw a line that goes through all the data points, there is no straight line that can accomplish that! Therefore, the value of income for each observation on the line will not be the same as the actual income of an individual in our sample. The vertical distance between the income value on the line and the actual income of an individual is called the *residual*. The best fitting line through the scatter plot of data is found by minimizing the sum of the squares of these residuals.[4] This type of regression is called *ordinary least squares (OLS) regression*, a term you will see in many of the research papers you will read throughout your studies in the SPP programme.

[3] To see why this is the case, think of the value of income if education is equal to 0.

[4] We can not use the sum of the residuals because some of these residuals will be negative, and some will be positive, and they would cancel each other out.

Figure 15 shows the same scatter plot of data together with the estimated regression line.

Figure 15: Example of Bivariate Regression: Income and Education

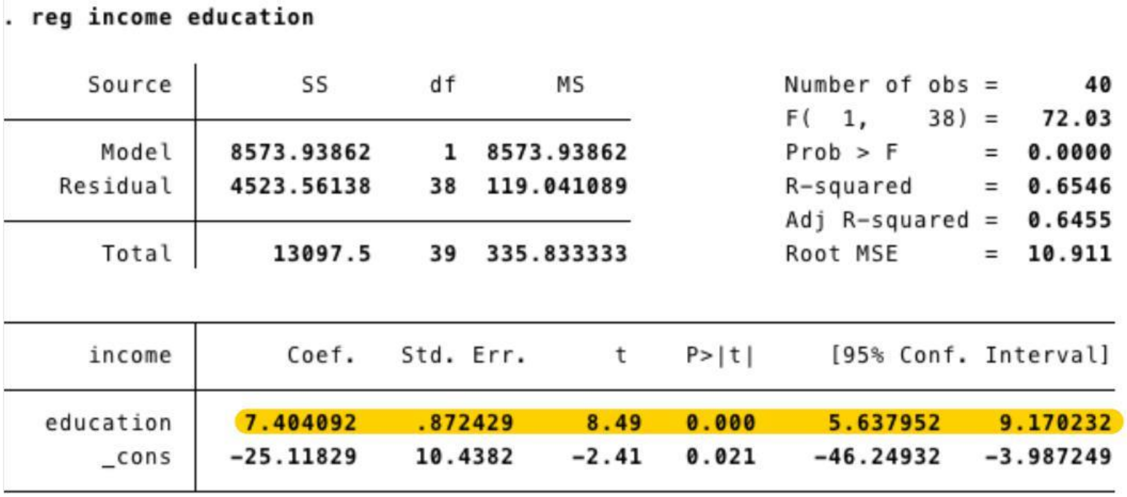


Remember that a line is summarized by its slope and its intercept. Using statistics software, it is very easy to compute the slope (β) and the intercept (α) of this regression line. If we indeed run a regression analysis of income and education using our sample of data on a statistics software, this is what we would get:

There is a lot of information on this table but for now, please only focus on the row highlighted in yellow. If you look at Figure 16, the coefficient next to education is 7.40. This is equal to the slope of the best fitting line plotted in Figure 15, and is our estimate of β. The substantive meaning of this slope is that each additional year of education is associated with a 7400 pound increase in income.

Regression is just another tool of statistical inference, which we use to learn about

Figure 16: Bivariate Regression Table: Income and Education



```
. reg income education

      Source |       SS       df       MS              Number of obs =      40
-------------+------------------------------           F(  1,    38) =   72.03
       Model |  8573.93862     1  8573.93862           Prob > F      =  0.0000
    Residual |  4523.56138    38  119.041089           R-squared     =  0.6546
-------------+------------------------------           Adj R-squared =  0.6455
       Total |     13097.5    39  335.833333           Root MSE      =  10.911

------------------------------------------------------------------------------
      income |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   education |   7.404092    .872429     8.49   0.000     5.637952    9.170232
       _cons |  -25.11829    10.4382    -2.41   0.021    -46.24932   -3.987249
------------------------------------------------------------------------------
```

a population using a sample of data from that population. In this case, our population is all professionals living in London. We try to learn about the effect of education on income using data only on 40 of those professionals. As it is the case with all other cases of inference, regression also involves some uncertainty around the parameter estimates. Next to the estimated effect of education on income (β), you see the estimated standard error around it. At the right end of the row, you see the 95 % confidence interval of the slope parameter. You also see the t-statistic and the associated p-value of the null hypothesis that education has no effect. As you see, the *p-value* is very small so we can confidently reject the hypothesis that education has no effect.

Can we interpret this as the causal effect of education on income? Not really because there can be other factors that affect both an individual's level of education and income. One such factor is family income. It is plausible to think that those with higher family income end up being more educated, and also have higher incomes themselves perhaps because of their better family connections. Therefore, the estimated impact of education on income may also include the effect of this 'omitted variable' of family income.

6.7.2 Multiple Regression Analysis

How do we take into account the effect of additional independent variables in the regres-sion analysis even if our main interest is on the effect of another independent variable? We simply add them to the regression analysis. These additional independent variables are commonly referred to as control variables. Regressions with more than one indepen-ent variable are called multiple regression. The equation of a multiple regression with two independent variables looks as follows:

$$Y_i = \alpha + \beta_1 X_i + \beta_2 Z_i + \epsilon_i$$

Plugging in the names of the variables we use in the analysis, this becomes:

$$income_i = \alpha + \beta_1 \ \ education_i + \beta_2 \ \ famincome_i + \epsilon_i$$

Suppose, we have data also on the family income of these individuals in our sample. Figure 17 shows the scatterplot of family income and individual income. There is a strong, positive relationship between the two. When we include family income in the regression analysis, we get the results in Figure 18.

There is no need to go too much into specifics of how these parameter estimates are calculated. At this stage, it is sufficient for you to know that multiple regression analysis estimates the impact of an independent variable *X* on the dependent variable *Y* by only using those components of *X* and *Y* that can not be accounted or explained by the third variable *Z*. Therefore, in our example, the coefficient estimate for education, β_1 tells us the impact of education on income *controlling for the effect of family income*. As you see, once family income is added as a control, the effect of education on income drops drasti-cally. Now the impact of an additional year of education on income is about 1,627 pounds. Similarly, we can interpret the coefficient estimate for family income, β_2 as the impact of

Figure 17: Scatterplot of Data: Income and Family Income

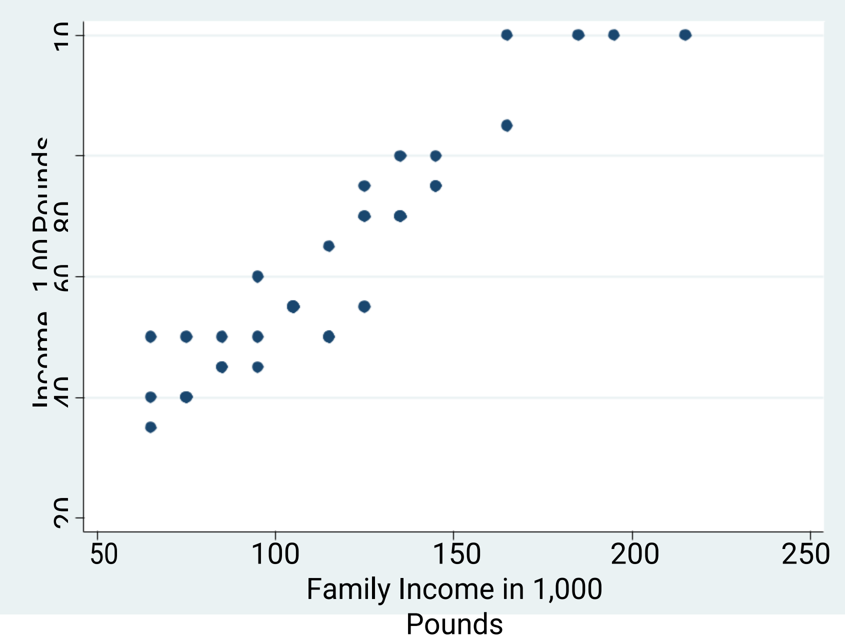
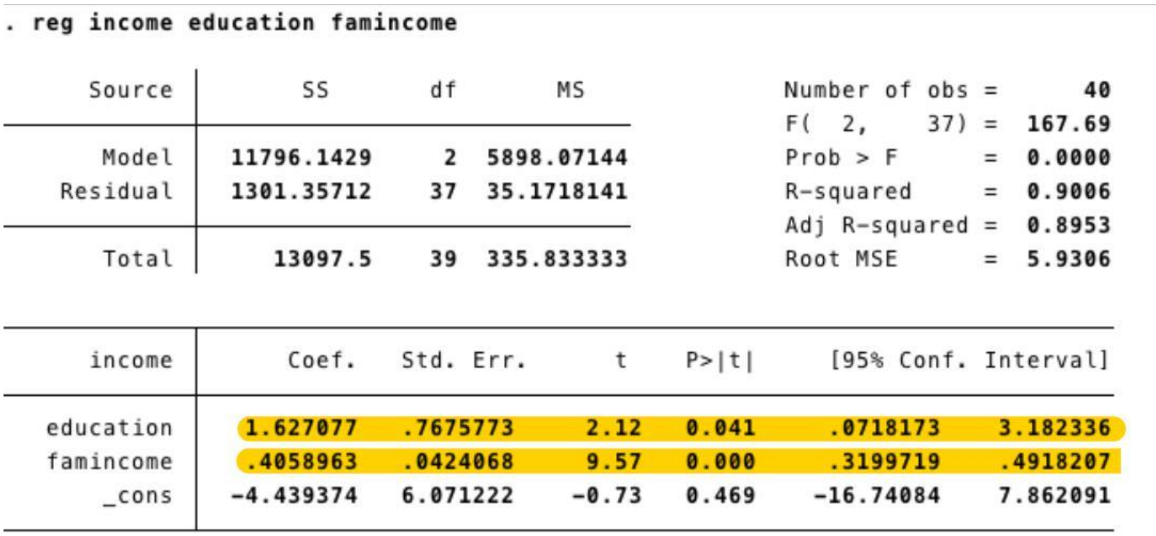


Figure 18: Multiple Regression Table: Income, Education and Family Income



```
. reg income education famincome

      Source |       SS           df       MS      Number of obs   =        40
-------------+----------------------------------   F(2, 37)        =    167.69
       Model |  11796.1429         2  5898.07144   Prob > F        =    0.0000
    Residual |  1301.35712        37  35.1718141   R-squared       =    0.9006
-------------+----------------------------------   Adj R-squared   =    0.8953
       Total |     13097.5        39  335.833333   Root MSE        =    5.9306

------------------------------------------------------------------------------
      income |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   education |   1.627077   .7675773     2.12   0.041     .0718173    3.182336
   famincome |   .4058963   .0424068     9.57   0.000     .3199719    .4918207
       _cons |  -4.439374   6.071222    -0.73   0.469    -16.74084    7.862091
------------------------------------------------------------------------------
```

family income on individual income, controlling for the effect of education. Each additional 1,000 pounds of family income is associated with an increase in income of about 406 pounds.

Most of the time, it is hard to measure and control for all the necessary variables in a regression. Therefore, scholars have started using more innovative methods that go beyond multiple regression with increasing frequency. We discuss these methods in the next module.

Practice Questions for Module 6

1. The higher the p-value of the test for a relationship between two variables, the more confident we are that there is a significant relationship.

   True
   False

2. Please fill in the blanks in the following statement:

   Type-I error occurs when w____________a___________null hypothesis. Type-II erroroccurs when we_______a___________null hypothesis.

3. Suppose you have a sample of 25 students from the MPA programme, and their mean exam score is 67. The standard deviation of their scores is 10. Using this sample ofdata, please test the hypothesis that the mean test score of all students in the programmeis 60. Use the significance level of 0.05.

4. In the latest YouGov survey in UK, 811 out of 1763 British adults say that their most favorite outcome of Brexit negotiations is to remain in the European Union. You can ac- cess the survey results at the following link: https://d25d2506sfb94s.cloudfront.net/cumulus_uploads/document/ete4gzwp4j/UCL_Brexit_190529_w.pdf. Pleasecalculate the 95% confidence interval for the percentage of British adults in UK, whose most favorite outcome is to remain in the EU.

Module 7: Strategies for Figuring Out Whether X Caused (A Change In) Y

There are five basic strategies used by researchers to figure out whether an independent variable X *caused* a change in the dependent variable Y. These strategies are natural experiments, matching, instrumental variable regression, difference-in-difference, and regression discontinuity. [5] As we discuss each of these methods, we will also learn about examples of research in political science that uses these methods. The citations of each of these papers is included in the References section at the end.

The best way to think about how these methods work is to understand their connec-tion to the *experimental method*, which we briefly summarize below.

7.1 The Experimental Method

There are three features of the experimental method:

1.Researchers compare the response of subjects in the *treatment group* to the response of subjects in the *control group*. The treatment group consists of those who have received an intervention, a program etc. whose effect on the response of the subjects the researcher cares about. The control group consists typically of those who have not received any intervention. For instance, in an experiment to determine the effectiveness of a new drug, the treatment group would be those who receive the new drug, and the control group would be those who don’t.

2. Subjects are *randomly assigned* to treatment and control groups. This is done through lottery, drawing numbers from an urn or any other mechanism that ensures that who is in the treatment and control groups is unrelated to the characteristics of the subjects.

3. The application of the treatment is under the control of the researcher.

Thanks to random assignment, the researcher ensures that the treatment and the control

[5] Note that you will learn about each of these methods in courses that are part of your core curriculum in greater detail. Here, our goal is to familiarize you with the logic of inference these methods rely on.

groups are not systematically different from each other and therefore, in the aggregate are identical to each other except the fact that those in the treatment group have received the treatment and those in the control group have not. Therefore, if the researcher observes a difference between the two groups' responses or values of the dependent variable, that difference must be because of the treatment assigned by the researcher.

7.2 Natural Experiments

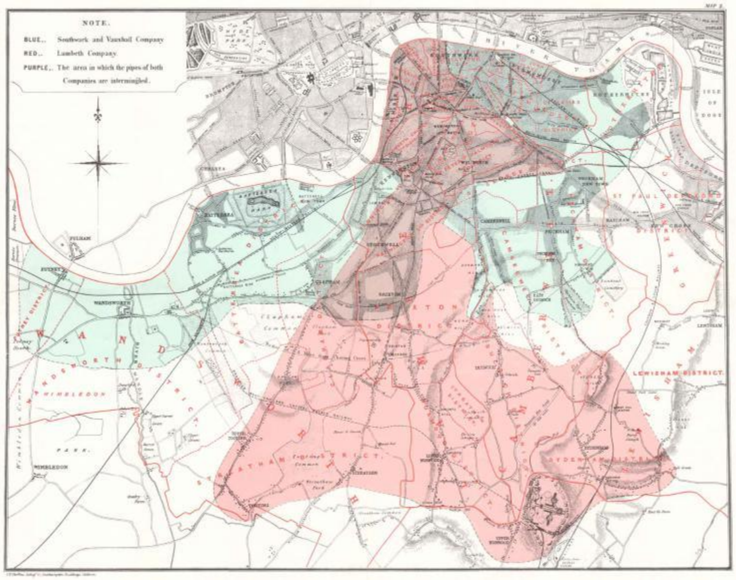
Natural experiment is a research design that is inspired by the experimental method. One difference is that the data in natural experiments are observational, i.e. it is collected from naturally occurring phenomena in the world around us, as opposed to the experimental data, which -as we discussed above- is generated under conditions controlled by the researcher.

Despite the data being observational, it is possible that sometimes, the researcher can reasonably claim that the variation in the treatment or the independent variable across units of observation are 'as-if-random'. Based on this claim, one can show that the units in the treatment and the control group are similar to each other on all relevant third factors Z, except the fact that some are treated, and some are not. Therefore, any difference in the average value of the dependent variable Y between the treatment and the control group must be due to the 'causal effect' of the treatment.

The earliest example of a natural experiment comes from where you will study next year, the city of London. London suffered from a cholera epidemic in 1853-1854. The dominant theory to explain these cholera outbreaks at the time was the theory of 'bad air" (miasma). John Snow, an anaesthesiologist, believed that cholera is an infectious disease carried through the water.

To prove his point, Snow conducted the following analysis: Large areas of London were served by two water companies: the Lambeth, and the Southwark and Vauxhall. In 1852, the Lambeth moved its intake pipe upstream on the Thames, obtaining a supply of water free from the sewage of London. The overlapping area in Figure 19 with houses served by both companies became the focus of Snow's natural experiment. Snow obtained records on cholera deaths throughout London and gathered information on the company that provided water to the house of each deceased as well as the total number of houses served by each company in each district of the city.

Figure 19: London, the Site of First Natural Experiment


NOTE.
BLUE.. Southwark and Vauxhall Company
RED.. Lambeth Company.
PURPLE.. The area in which the pipes of both Companies are intermingled.
LEWISHAM DISTRICT

Among houses served by Southwark and Vauxhall, the death rate from cholera was 315 per 10,000. Among houses served by Lambeth, the death rate was only 37 per 10,000. This dramatic difference provided compelling evidence for the impact of water supply source on deaths from cholera.

This is compelling evidence because as Snow explains, 'The mixing of the (water) supply is of the most intimate kind. The pipes of each Company go down all the streets, and into nearly all the courts and alleys. A few houses are supplied by one Company and a few by the other, according to the decision of the owner or occupier at that time when the Water Companies were in active competition. In many cases a single house has a supply different from that on either side. Each company supplies both rich and poor, both large houses and small; there is no difference either in the condition or occupation of the persons receiving the water of the different Companies. . . . It is obvious that no experiment could have been devised which would more thoroughly test the effect of water supply on the progress of cholera than this......the move of the Lambeth company's water pipe meant that more than three hundred thousand people of all ages and social strata were divided into two groups without their choice, and, in most cases, without their knowledge [italics added]; one group being supplied with water containing the sewage of London, and, amongst it, whatever might have come from the cholera patients, the other group having water quite free from such impurity' (Snow 1855, 75).

In short, the evidence is compelling because Snow could convincingly argue that only the water source distinguished the houses in the treatment group from the houses in the control group. He therefore could make the inference that the difference in the cholera death rates between the two groups was *due* to the effect of the water supply.

Natural experiments have become one of the standard toolkits for scholars doing re-search to establish causal effects of policies or institutions on political, economic and so-cial outcomes. For instance, Jones and Olken (2005) study the effect of political leaders on economic performance by comparing growth before and after a leader dies due to natu-ral causes, unrelated to the economic conditions in the country. They find that leaders do affect growth especially in autocratic settings. To examine the impact of access to arms on violent crime, Dube et al. (2013) focus on the aftermath of the expiration of the U.S. Federal Assault Weapons Ban in 2004, as a result of which there was a gun spillover to Mexican municipios near the states of Texas, Arizona, and New Mexico, while no such spillover happened in municipios near California, which maintained its state-level ban. The authors find that municipios closer to non-California border did indeed experience an increase in homicides.

7.3 Matching

Matching tries to control for variables that affect both the treatment and the dependent variable by matching treated units to one or more control units that are as similar as possible to the treated unit on the values of the control variables. There are various ways to match the treated and control units with each other. But most importantly, this method relies on the assumption that there are no confounding variables that the researcher is unable to measure or unaware of.

Gilligan and Sergenti (2008) study whether UN interventions in conflicts after the Cold War cause peace. Which conflicts UN intervenes is not randomly assigned, therefore, a simple comparison of conflicts with and without UN intervention would not give an un-biased estimate of the effect of UN interventions. Instead, they match cases of UN inter-ventions to cases with no UN interventions, on a number of confunding variables that affect both the likelihood of UN intervention and the likelihood of the duration of peace or the end of conflict. These factors include battle deaths from the conflict, ethnic frac-tionalization, population size, how mountainous the country is, the number of military personnel in the conflict country, and political regime of the country. The authors find that UN interventions prolong peace after the conflict ends but do not have an effect on ending on-going conflicts.

Lyall (2010) examines the effect of the ethnic identity of the soldiers conducting counterinsurgency operations in Chechenya on the subsequent insurgent attacks during the Second Chechen War (2000-2005). Having matched operations in villages conducted by pro-Russian Chechen units with those conducted by Russian-only units, he finds that op-erations by Chechen units reduce attacks relative to operations by Russian-only units. He matches these operations on a number of demographic, spatial and conflict-related con-founding variables, including the population of the settlement, its religious orientation, poverty level, and how geographically isolated the village is.

7.4 Instrumental Variable Regression

An alternative method to estimate causal effects looks for a variable that explains who is in the treatment and who is in the control group but has no other relation to the dependent variable. Such a variable is called an instrument. If the researcher has access to such an instrumental variable, she can estimate the effect of the treatment on the dependent variable by examining the effect of the instrument on the dependent variable, and the effect of the instrument on the treatment variable. For instance, Acemoglu et al. (2001) estimate the effect of political institutions on economic performance in former colonies, using the settler mortality rates in these former colonies as an instrument for the quality of institutions. Settler mortality is a ‘good’ instrument to the extent that it has an effect on the quality of institutions, and has no effect on economic performance other than through its effect on institutions.

Several other recent papers in political science and economics have used instrumental variable regression. For instance, Miguel et al. (2004) study the effect of growth shocks on the likelihood of civil war in Africa. They use rainfall growth as an instrument for economic growth. Rainfall is a ‘good’ instrument to the extent that it is related to growth rates in the sample of African countries the authors study (which it does) and the only way rainfall affects likelihood of civil war is through its effect on growth rates. The au-thors find that negative growth shocks significantly increase the likelihood of civil war in Africa: A five-percentage point negative growth shock increases the likelihood of a civil war the following year by one-half.

Nunn and Wantchekon (2011) study the causes of high levels of mistrust in the African continent today and found that it is caused by the transatlantic and Indian Ocean slave trades in the past: Using survey data on trust levels among various ethnic groups, they show that individuals from groups most affected by the slave trade have lower trust lev-els. The authors use the distance of an individual’s ethnic group from the coast to the slave trade as an instrument. This distance variable is a good instrument since it is corre-lated with the exposure of an ethnic group to the slave trade. This is because the purchase of slaves happened at the coast. And the distance of a group from the coast is also plau-sibly uncorrelated to any other characteristics of the current members of the same ethnic group that affects their levels of trust.

7.5 Difference-in-Difference

This approach starts with the idea of comparing the value of the dependent variable be-fore and after the application of the treatment. But instead of making this comparison only for the treatment group, one compares the change in the value of the dependent variable in the treatment group before and after the treatment to the change in the value of the dependent variable in the control group before and after the treatment (Remember, the treatment is applied only to the treatment group but we still calculate the change in dependent variable in the control group as well.). The key feature of this method is that the researcher needs to have control units that are not necessarily identical to the treat-ment units but the treatment and control units should only change with the treatment, and there should not be other changes that affect the two groups differently.

One of the well-known examples of this type of research design is Card and Krueger (1994), who study the effect of miminum wage on employment in the states of New Jersey and Pennslyvania. To do this, they utilize the fact that New Jersey raised state minimum wage from $ 4.25 to $ 5.05 in April 1992, while the minimum wage in the neighbouring Pennslyvania remained the same both before and after April 1992. The authors compare the change in the employment rate in the state of New Jersey from February 1992 to November 1992 to the change in the employment rate in the state of Pennslyvania from February 1992 to November 1992.

The key assumption is that the employment trends would be the same in both states, in other words, nothing else that can affect the employment rate has changed differently between the two states in the same time period. Under this assumption, the authors' comparison gives the effect of the minimum wage on employment rate. The conclusion of the authors is that the rise in the minimum wage does not reduce employment.

Scheve and Stasavage (2012) study the effect of mass mobilization for warfare on inheritance taxation in a sample of countries from Europe and North America. They com-pare the change in the top rates of inheritance taxes in countries before and after mass mobilization for warfare in major wars (including the two World Wars) to the change in the top rates of inheritance taxes in countries that did not participate in these wars before and after the mass mobilizations in the participating countries. The authors find strong evidence of mass mobilizations having a strong positive effect on the inheritance taxes.

7.6 Regression Discontinuity

This method is based on the idea that in some real-world cases, rigid rules decide who gets into a particular program or receive an intervention. These rules are rigid in the sense that units with values on a variable or score that are slightly above the cutoff receive the treatment while those slightly below do not. Hence, one can reasonably argue that among those who are slightly above and below the cutoff, and therefore who is assigned into the treatment and who is assigned into the control group is 'as-if-random'.

The first application of this method was by Thistlethwaite and Campbell (1960) who estimated the effect of a National Merit Award on subsequent student performance by comparing students whose test scores were just high enough to win the award to those whose scores fell barely short. The most common application of this method in political science is to examine the impact of election results on a number of political and economic outcomes of interest.

For instance, Hainmueller and Eggers (2009) compare the wealth of deceased mem-bers of the House of Commons in UK to those candidates who narrowly lost elections. They find that serving in office doubles the wealth of Conservative MPs, while it has no significant effect on the wealth of Labor MPs. The key assumption of this method for this study is that Conservative (Labor) MPs who narrowly won are similar to the Conserva-tive (Labor) candidates who narrowly lost.

Another research area that uses regression discontinuity examines the impact of poli-cies and institutions on outcomes in cases where the policy or institution is determined by the population size of subnational units in a country. The pioneering example in this area is Pettersson-Lidbom (2012) who look at the effect of the size of the municipal council on the municipal spending in Sweden and Finland by comparing spending in cities above and below the population threshold that determines the size of the municipal council.

Practice Questions for Module 7

1. Please fill in the blanks in the following statement:
   In an experiment to determine the effectiveness of a new drug, the____groupwould be those who receive the new drug, and the_______group would be those whodon’t.

2. Why is it important that subjects are randomly assigned to the treatment and control groups in an experiment? Briefly explain.

3. What type of data are used in natural experiments?

4. The legal drinking age in US is 21. Suppose we are interested in establishing the causal effect of legal access to alcohol on death rates in US. How can we use this information to construct a research design to establish this causal effect? Briefly describe.

References

Acemoglu, Daron, Simon Johnson, James Robinson. 2001.The Colonial Origins of Com-parative Development: An Empirical Investigation. American Economic Review 91(5): 1369-1401.

Card, David and Alan, Krueger. 1994. Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania. American Economic Review, 84 (4): 772-793.

Dube, Arindrajit, Oeindrila Dube, and Omar Garca-Ponce. 2013. Cross-Border Spillover: US Gun Laws and Violence in Mexico. American Political Science Review 107(3): 397-417.

Eggers, Andrew, Jens Hainmueller. 2009. MPs For Sale? Returns to Office in Post-War British Politics. American Political Science Review 103 (4): 513-533.

Gilligan, Michael J., and Ernest J. Sergenti. 2008. Evaluating UN peacekeeping with matching to improve causal inference. Quarterly Journal of Political Science: 89-122.

Jones, Benjamin F., and Benjamin A. Olken. 2005. Do leaders matter? National lead-ership and growth since World War II. The Quarterly Journal of Economics: 835-864.

Lyall, Jason. 2010. Are coethnics more effective counterinsurgents? Evidence from the second Chechen war. American Political Science Review 104(1): 1-20.

Miguel, Edward, Shanker Satyanath, and Ernest Sergenti. Economic shocks and civil conflict: An instrumental variables approach. Journal of Political Economy 112(4): 725-753.

Pettersson-Lidbom, Per. 2012. Does the Size of the Legislature Affect the Size of Gov-ernment: Evidence from Two Natural Experiments. Journal of Public Economics 98(34):269278. Snow, John. 1855. On the Mode of Communication of Cholera. London: John Churchill, New Burlington Street.

Stasavage, David, Kenneth Scheve. 2012. Democracy, War and Wealth: Lessons from Two Centuries of Inheritance Taxation. American Political Science Review 106 (1): 81-102.

Thistlethwaite, D. L., and Campbell, D. T. 1960. Regression-discontinuity analysis: An alternative to the ex post facto experiment. Journal of Educational Psychology 51(6), 309-317.