Inferential Statistic

Preprint · June 2014

DOI: 10.13140/RG.2.2.31465.36963

1 author:

Diego Farren
University of Hamburg
34 PUBLICATIONS **136** CITATIONS

SEE PROFILE

Inferential statistic

using Stata

Diego Farren
Universität zu Köln
Wirtschafts- und Sozialwissenschaftliche Fakultät
Institut für Soziologie und Sozialpsychologie (ISS)
diegofarren@gmail.com

June 26, 2014

1 Introduction

The intention of this small text is to help you understand the essential concepts behind statistical inference, which you need to apply every time when running OLS. Through simulation we can create a fictional population and take samples from it in order to prove some of the abstract concepts and assumptions we rely on. The most essential concept you need to understand is **sampling distribution**. Everything we test depends on our estimation of this distribution.

The text discuses the concrete meaning of sampling distribution and also meaning and estimation of confidence intervals and hypothesis tests. Simple means are described, because the concepts are easier to explain with means, but everything that is shown applies also for regression coefficients. Throughout the text some more advance commands of Stata than the ones you need for this course are being used. You do not necessarily need to understand these commands in detail but only what it is being done. Anyway it is a good idea to replicate everything I show. All commands are given and since it is a simulation, no data set is needed.

2 Inferential statistics

Inferential statistic is all about making a connection between sample and population. Because we usually do not have access to the whole population, we make some estimations at the level of the sample, like calculating the mean of age in the sample, and would like to be able to say that this estimated value also represents the population. For this purpose we need to know how accurate our estimator is and the sampling distribution would allow us to check this.

The value we want to know from the population, e.g. average age of the population, is a **parameter**. Parameters are not known for us and that is why we need to estimate them. An **estimator** is a formula, like a mean. It is defined even before you have data to generate a value with it. Once you have applied this formula to a concrete sample, the value you get is an **estimate**. Estimators are variables, because depending on what information you put inside of them, i.e. depending on the sample, you are going to estimate different values. Estimates on the other side, are no variables but a concrete value (see Figure 1).

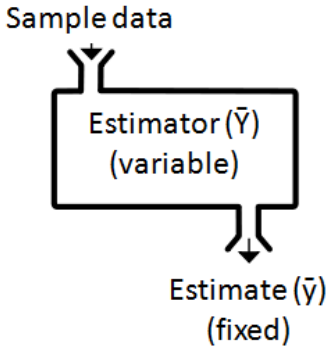

Figure 1: *Estimator and estimate*

Sampling distribution is the distribution of an estimator. This means that if you would make many estimations with this formula (e.g. if you would calculate means for many same sized samples coming from the same population) and would graph the distribution of these values, you would generate a sampling distribution. The sampling distribution has as compounding units no longer people, but estimates or in this case means. To put all these abstract concepts into visual representation, let us start by generating a fictional population.

3 Distribution of a population

```
. clear
. set obs 10000
obs was 0, now 10000
. set seed 666
. gen age = rnormal(40,10)
. save population.dta, replace
file population.dta saved
```

Box 1: *Commands for generating a fictional population*

The commands in Box 1 generate a fictional population[1]. `set obs 10000` tells Stata to create 10,000 empty rows for what comes next. `gen age = rnormal(40,10)` generates a variable called age which is normally distributed with mean 40 and standard deviation 10. This means that our fictional population has mean age 40 and the average distance to the mean is ± 10 years. Before running this command I wrote `set seed 666` so that you can replicate what I do and get exactly the same results (i.e. this command fixes the starting point of the randomness in the generation of the data, which makes it replicable).

Once the data is generated (see Box 2), we run `list in 1/5` to see the first five rows of our data set in the results window (always a good idea for keeping an eye to what we have done). Now we would like to see the descriptive statistics of *age* so we use `sum age` to get a table in the results window and `hist age, norm` to get a graph (see Figure 2).

[1]Note as general rule that all lines in the results window in Stata (presented here inside Boxes) that start with '·' represent commands, and lines without '·' represent results. Also note that each time a '>' appears in this Boxes, it represents a line brake in a command and this you have to erase when writing the commands in your do-file and put the whole command in one line.

```
. use population.dta, replace
. list in 1/5

     +----------+
     |      age |
     |----------|
  1. | 55.04719 |
  2. | 45.05816 |
  3. | 46.49934 |
  4. |  50.1361 |
  5. | 30.41114 |
     +----------+

. sum age
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
         age |      10000    39.99012    9.932762    6.291093   71.46303

. hist age, norm
(bin=40, start=6.2910933, width=1.6292984)
```

Box 2: *Results from describing a fictional population*

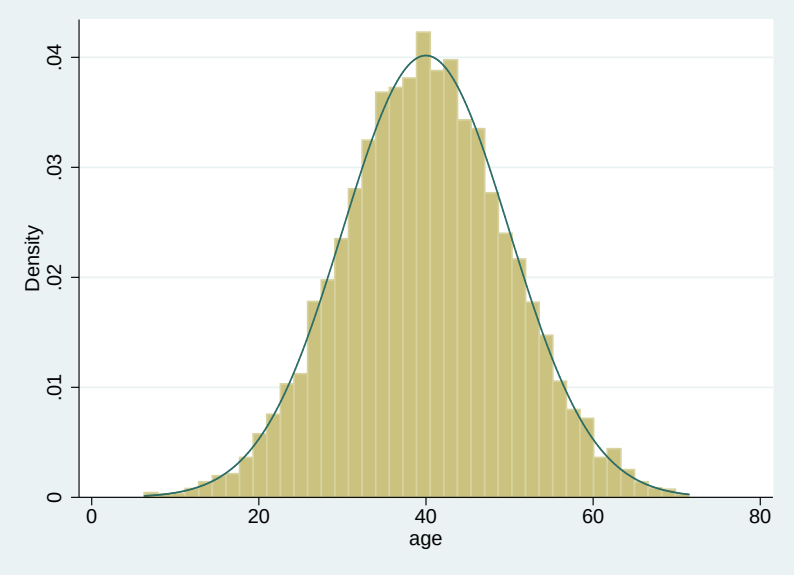

Figure 2: *Population distribution of age*

In the previous we generated a 'population' in the data set 'population.dta' and described it. Again, keep in mind that this is fictional and in real life if we would have access to the complete population no samples would be needed. In this case each row represents a fictional person. Each value in each cell of the column 'age' represents the age of one of this fictional persons. The descriptive statistics showed us that the mean age of this population is exactly 39.99012 with a standard deviation of exactly 9.932762. We also know that the youngest person in this population is 6.291093 years old and the oldest 71.46303. Now we want to generate an empirical sampling distribution, also something fictional for pedagogical purposes.

4 Empirical sampling distribution

Before we start generating an empirical sampling distribution, note that if we would know the parameters of the population (as we do), then it would not be necessary to generate this empirical sampling distribution. If we know the population distribution (in this case: $X_i \sim \mathrm{N}(\mu, \sigma^2)$) and the values of its parameters, we also know the exactly sampling distribution ($\bar{X} \sim \mathrm{N}(\mu, \frac{\sigma^2}{n})$).[2] Anyway we do this for pedagogical purposes.

```
. program define simulation
  1.         use population.dta, clear
  2.         sample 100, count
  3.         sum age
  4. end
. set seed 666
. simulate "simulation" mean=r(mean) sd=r(sd) N=r(N), reps(5000) saving(samplin
> g.dta) replace
command:      simulation
statistics:   mean       = r(mean)
              sd         = r(sd)
              N          = r(N)
. save sampling.dta, replace
file sampling.dta saved
```

Box 3: *Commands for generating a sampling distribution*

The commands in Box 3 might seem complicated but what we did is just telling Stata to take a random sample size 100 (defined where it says `sample 100, count`) from our population, estimate the mean for this sample (actually estimate all descriptive statistics through `sum age`), and save this information in one row in the data set called 'sampling.dta' (`saving(sampling.dta), replace` saves the program we just created and `save sampling.dta, replace` saves the generated data set). We also pushed Stata to repeat this procedure 5,000 times (by writing `reps(5000)` in the simulation). So **the newly generated data set 'sampling.dta' contains 5,000 rows, each one of them representing the descriptive statistics of one different but same sized sample coming from the same population** (we also saved the standard deviation of each sample for further purposes and the sample size only to prove that it is always 100).

The next step would be to take a look at the newly generated sampling distribution data set (see Box 4). As we did with the population data set, we start by looking at the five first rows in the sampling data set with the command `list 1/5` and continue by analyzing the descriptive statistics of the variable *mean* through `sum` and `hist` (see Figure 3).[3]

[2] See the appendix.

[3] The `gen id = _n` command generates running counter of the observations, i.e. a variable that numerates all observations from the first to the last with unique numbers in logical order. In this case from one till 5,000 because we have 5,000 observations. This is done for further purposes.

```
. use sampling.dta, replace
(simulate: simulation)
. gen id = _n
. save, replace
file sampling.dta saved
. list id mean sd N in 1/5
```

	id	mean	sd	N
1.	1	42.38844	10.10429	100
2.	2	41.48432	8.707458	100
3.	3	39.46537	10.80167	100
4.	4	40.80586	9.87238	100
5.	5	39.96799	9.846917	100

```
. sum mean sd N
```

Variable	Obs	Mean	Std. Dev.	Min	Max
mean	5000	39.9971	.9946097	36.91653	43.50111
sd	5000	9.88972	.6864858	7.382634	12.54134
N	5000	100	0	100	100

```
. hist mean, normal
(bin=36, start=36.916534, width=.18290499)
```

Box 4: *Results from describing the sampling distribution*

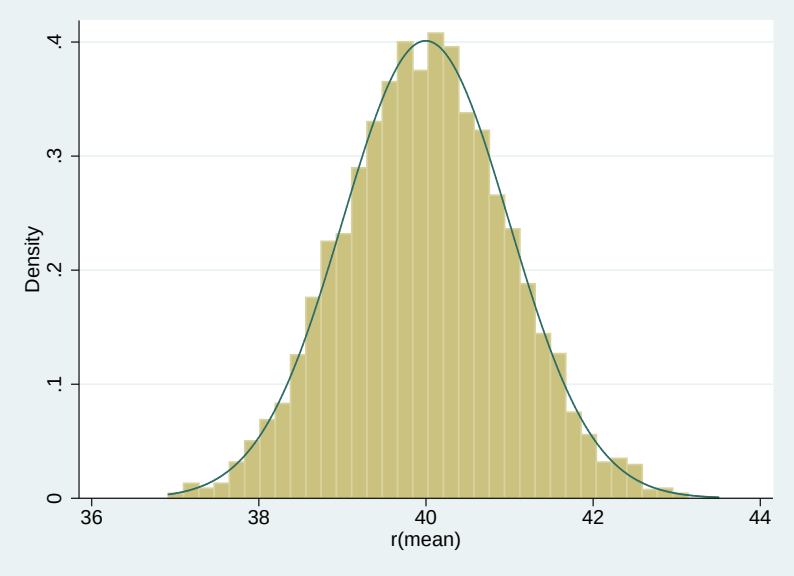

Figure 3: *Sampling distribution*

Keep in mind that the newly generated data set `sampling.dta` represents the distribution of the formula with which we calculate mean values (the estimator). Each row in this data set contains summary information of a different sample, i.e. each row already resumes information from 100 units (because we set the sample size to 100). The first thing to note is that the average value of the variable *mean* in the sampling distribution data set (i.e. 39.9971), although not exactly the same as the average value of age in the population (i.e. 39.99012), is quite similar.[4] This is so because the expected value of the population is the same as the expected value of the sampling distribution (see the appendix for proofs), and although a small difference exist we keep now

[4]If we would have all possible samples size 100 coming from the same population, the average of the sampling distribution of the mean of age would be exactly the same as the value of the average age in the population. The number of possible samples size 100 coming from a population size 10,000 is $10,000^{100}$, which gives a huge number. Note that the sampling distribution represents the distribution of all possible samples with replacement and considering order, because we assume the units to be independent from each other.

track of the expected value of the sampling distribution (which in an ideal scenario would be exactly the same as the one of the population).

Note also that the standard deviation of *age* in the population is quite different from the standard deviation of the variable *mean* in the sampling distribution. This is to be expected because the first one represents the average distance of age to the average value of age in the population, i.e. we are talking about age of persons here and in our fictional population (see Box 2) people can be as young as 6.291093 and as old as 71.46303 which is a very wide interval. While in the second case, i.e. in the sampling distribution, we are talking about the distribution of estimated means or average values of age. Now think a little about the meaning of this. The distribution of age in the population might be quite wide, but if you estimate average values each time, they should always be close to the real mean (at least if your estimator is unbiased, which the formula for estimating means is). So now the values of the estimated averages of age lie between 36.91653 and 43.50111 which is a much tighter interval.

```
. use sampling.dta, replace
(simulate: simulation)
. sum mean
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
        mean |       5000     39.9971    .9946097   36.91653   43.50111
. dis r(mean)-1.96*r(sd) " ; " r(mean)+1.96*r(sd)
38.047668 ; 41.946537
. gen emp_ci = (mean>38.047668 & mean<41.946537)
. tab emp_ci
     emp_ci |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |        253        5.06        5.06
          1 |      4,747       94.94      100.00
------------+-----------------------------------
      Total |      5,000      100.00
```

Box 5: *Confidence intervals from sampling distribution*

Before we continue with how to estimate the sampling distribution when only one sample is at hand, let us discuss a little about the meaning of 95% confidence interval. As we have seen the sampling distribution is conformed by many point estimates (many means). Relying on the confidence intervals formula, we can convert these point estimates into intervals. By doing so and specifying 95%, the intervals we generate will include the true parameter inside of them in 95% of the cases. The formula for 95% confidence intervals when the distribution is known and normal is: $\mu \pm 1.96 \cdot \sigma_{\bar{X}}$.

In box 5 I have generated a new variable *emp_ci* which gives a one to all estimated confidence intervals that contain the true parameter[5] and a zero otherwise. Next we see the frequencies of this variable and you can note that almost exactly 5% of the estimated confidence intervals do not contain the true parameter[6].

[5] Again note that the true parameter is known for us and is the average age of the population. Anyway we consider here the average of the sampling distribution, or the *mean of means* as our true parameter, because as already said if the sampling distribution were constructed with all possible samples, these two values would be exactly the same. Anyway in this case they are still almost equal. But strictly speaking the true parameter is the average of age in the population.

[6] Again note that this value would be exactly 5% if we would have access to all possible samples sized 100 coming from the same population.

5 Estimated sampling distribution

All the previous is very nice, but in real life you do not have access to the whole population (otherwise inferential statistic would not be needed) and you also cannot collect data from 5,000 different samples (actually it would then be easier to get information of the whole population). So how could we get an idea of how this sampling distribution looks like with the information of just one sample? We have to estimate it.

It can be shown that the sample mean is an unbiased estimator of the average age of the population (its expected value) and also of the *mean of means* in the sampling distribution (also its expected value, see the appendix for proofs). It can also be shown that by dividing the sample standard deviation by the square root of n, we get an unbiased estimator of the standard deviation of the sampling distribution (again see appendix for proofs).

```
. use sampling.dta, replace
(simulate: simulation)
. list id mean sd N in 1/5
```

	id	mean	sd	N
1.	1	42.38844	10.10429	100
2.	2	41.48432	8.707458	100
3.	3	39.46537	10.80167	100
4.	4	40.80586	9.87238	100
5.	5	39.96799	9.846917	100

```
. gen se = sd/sqrt(N)
. save, replace
file sampling.dta saved
. list id mean se in 1/5
```

	id	mean	se
1.	1	42.38844	1.010429
2.	2	41.48432	.8707458
3.	3	39.46537	1.080167
4.	4	40.80586	.987238
5.	5	39.96799	.9846917

```
. sum mean se
```

Variable	Obs	Mean	Std. Dev.	Min	Max
mean	5000	39.9971	.9946097	36.91653	43.50111
se	5000	.988972	.0686486	.7382634	1.254134

Box 6: *Standard errors*

In box 6 we have generated a new variable called *se* for *standard errors*. It represents our estimation of the standard deviation of the sampling distribution. The known formula is: $Se(\bar{X}) = \frac{S^*}{\sqrt{n}}$, where S^* stands for the corrected standard deviation estimated with each sample. We list the first five cases in order to show that in all of them the values are quite similar. Also note that these values should estimate the standard deviation of a sampling distribution of samples sized 100 which is known for us and is equal .9946097 (see in box 4 the standard deviation of the variable *mean*). Finally note that the mean of all standard errors is .988972 which is quite close to the real value.

Now let us keep track of only one sample as we would in real life. The first sample we have produces a mean value of 42.38844 and a standard error of 1.010429. These would be our estimates. And because in real life we do not know the distribution of the population or the

sampling distribution, we have to just trust these values. But these are only point estimates and we want to describe the whole distribution so that we know how much *probability* lies between standard errors. Because we use an estimated standard deviation to generate a new estimate, i.e. the standard error, we need to take these uncertainties into account when describing our estimated sampling distribution. This is done by relying on the t-distribution instead of the normal one.[7]

```
. use sampling.dta, replace
(simulate: simulation)
. dis invttail(99,.025)
1.984217
. gen lb = mean-1.984217*se
. gen ub = mean+1.984217*se
. save, replace
file sampling.dta saved
. list id mean se lb ub in 1
```

	id	mean	se	lb	ub
1.	1	42.38844	1.010429	40.38353	44.39335

Box 7: *Sample confidence intervals*

In box 7 you can see that we first ask for the critical values of a t-distribution for 95% of its probabilities. Degrees of freedom need to be taken into account and they equal the number of observations (i.e. 100) minus the number of estimated parameters (in this case 1, the mean). Now you *adapt* this distribution to your data by putting the estimated mean in the center of it and multiplying the critical values by the estimated standard error. This gives us the confidence intervals for the mean.

Note that the confidence intervals we get here have a different interpretation as the ones we estimated previously with the whole sampling distribution. When looking at the formula for the confidence intervals in abstract or at the empirical sampling distribution, it is right to say that the values that come out of it represent the probability that the parameter lies between them. But when looking at one sample, you cannot talk about probabilities anymore. You do not know where the true parameter lies. And the estimates you get are no longer variables but constant values. So the parameter lies or lies note inside the estimated confidence intervals, but you can never now which is the case.

As a matter of fact, in our sample 1 the estimated confidence intervals are 40.38353 and 44.39335 and we know that the true parameter is 39.9971 (or strictly speaking 39.99012). So in this case to say that with a 95% probability the true parameter lies between 40.38353 and 44.39335 is wrong. You could say with a 95% confidence or "if many samples from equal size were taken from the same population, in 95% of the cases the estimated confidence interval would contain the true parameter" (see figure 4). So although we cannot know if the parameter lies inside our estimated confidence interval, the wide of this interval still tells us something about the accuracy of our estimator.

[7] If we would have used the population standard deviation in order to estimate the standard error (i.e. $Se(\bar{X}) = \frac{\sigma}{\sqrt{n}}$) instead of the corrected estimation of the standard deviation (i.e. $Se(\bar{X}) = \frac{S^*}{\sqrt{n}}$), we could rely on a normal distribution (that is why we previously used 1.96 to construct the confidence intervals).

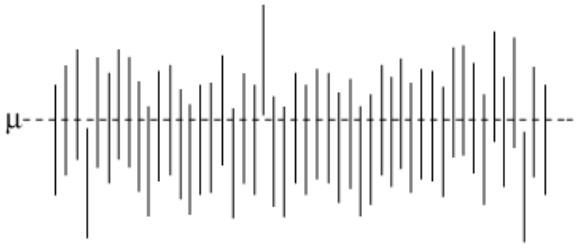

Figure 4: *Confidence intervals distribution*

6 Hypothesis test

In the previous we saw how to *adapt* the t-distribution to the estimates obtained from a sample. I.e. we adapt this theoretical distribution so that the estimated mean is in the center of it and 95% of the probabilities of this adapted distribution lie between $\pm Se(\bar{X}) \cdot CV_{df,\alpha/2}$, where $Se(\bar{X})$ is the estimated standard error and $CV_{df,\alpha/2}$ is the critical value of a t-distribution with given degrees of freedom (df) and significance level (α).

For hypothesis testing we kind of do the same but the other way around, i.e. we keep the theoretical t-distribution as it is and try to *insert* our estimates in it. By doing so we are again relying on the t-distribution being a good representation of the sampling distribution. But the t-distribution is standardized, i.e. has mean zero and standard deviation 1.[8] Because in this exercise we know the sampling distribution, we can compare the standardized empirical sampling distribution with the theoretical t-distribution to prove that they are quite the same.

```
. use sampling.dta, replace
(simulate: simulation)
. sum mean
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
        mean |       5000     39.9971    .9946097   36.91653   43.50111
. gen zmean = (mean-r(mean))/r(sd)
. save, replace
file sampling.dta saved
```

Box 8: *Commands for generating a standardized empirical sampling distribution*

In box 8 you see the commands for generating a standardized version of the variable *mean*. In box 9 you see some commands for describing the distribution of this variable and comparing it to the theoretical t-distribution with 99 degrees of freedom (see figure 5).

[8] Standardizing a variable means generating a new variable by subtracting to each value of the original variable its mean and dividing this result by the standard deviation, i.e. $Z_i = \frac{X_i - \mu}{\sigma/\sqrt{n}}$. Now because we usually do not know these parameters, we modify the formula to $Z_i = \frac{X_i - \bar{X}}{Se(X)}$.

```
. use sampling.dta, replace
(simulate: simulation)
. list id mean zmean in 1/5

     +-------------------------------+
     |  id       mean       zmean    |
     |-------------------------------|
  1. |   1   42.38844      2.4043    |
  2. |   2   41.48432    1.495279    |
  3. |   3   39.46537   -.5346179    |
  4. |   4   40.80586    .813143    |
  5. |   5   39.96799   -.0292656    |
     +-------------------------------+

. sum zmean

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
       zmean |       5000    1.24e-10           1  -3.097263   3.523001

. hist zmean, normal name(zmean,replace)
(bin=36, start=-3.0972633, width=.18389624)
. twoway function y=tden(99,x), range(-1.98 1.98) color(dknavy) name(tdist,repl
> ace)|| ///
> function y=tden(99,x), range(-4 -1.98) recast(area) color(dknavy) || ///
> function y=tden(99,x), range(1.98 4) recast(area) color(dknavy) ///
> xtitle("{it: x}") ///
> ytitle("Density") ///
> legend(off) xlabel(-4 -1.98 0 1.98 4)
. graph combine zmean tdist, cols(1) name(combined,replace) //both graphs toget
> her
```

Box 9: *Standardized empirical sampling distribution description*

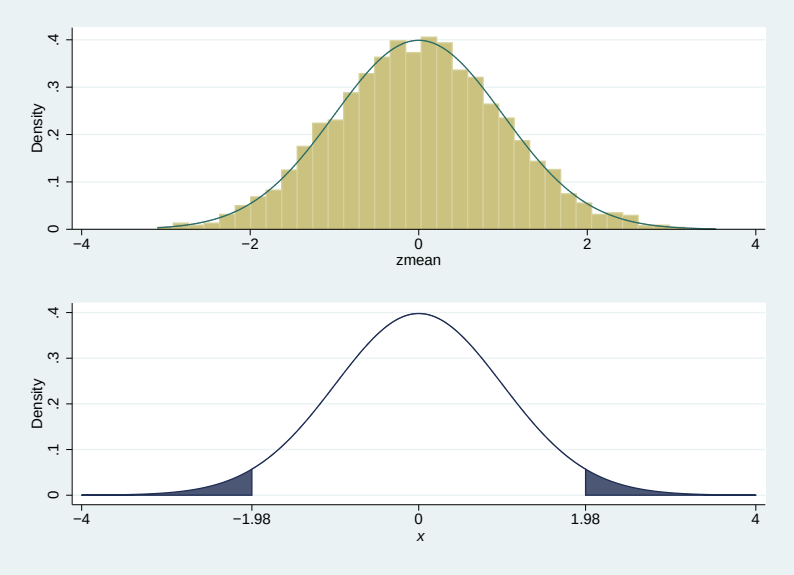

Figure 5: *Theoretical t-distribution and empirical sampling distribution compared*

Again in real life we could not access this standardized empirical sampling distribution, so we need to rely on the theoretical t-distribution. But in order to estimate a sampling distribution that allows us to test hypotheses, we need values for μ and σ. The last one we estimate through the $Se(\bar{X})$ as we saw. And μ we have to hypothesize. So what you are doing when you state as your null hypothesis, e.g. that the average age of the population is equal 41, is defining your μ. So you are assuming that the true parameter equals 41 and the sampling distribution lies around this value. And if you standardize the sampling distribution, 41 is still going to be in the middle of the distribution but represented by a 0. Now you just need to standardize the only value of *age* you have, i.e. the sample mean, and insert it into your hypothesized distribution.

An example should clarify things a little. If you look at the first sample again, the value of *mean* was 42.38844 and the value of *se* was 1.010429 (see box 7). You assume that $\mu = 41$ and trust that your value of *se* is a good approximation to the standard deviation of the sampling distribution. So you have everything you need to *insert* your sample mean into the hypothesized sampling distribution. For this we need to standardize our value (we now call the standardized value t and not z because we will insert it into a t-distribution), i.e. $t_{age} = \frac{42.38844-41}{1.010429} = 1.3741094$ and this value is smaller than the critical value of a two-sided t-distribution with $\alpha = 5\%$, i.e. 1.984217. This result means that if your null hypothesis were true and your sample comes from a population centered around 41, the probability of getting a value of 42.38844 is not unusual enough to reject this hypothesis (see figure 6).

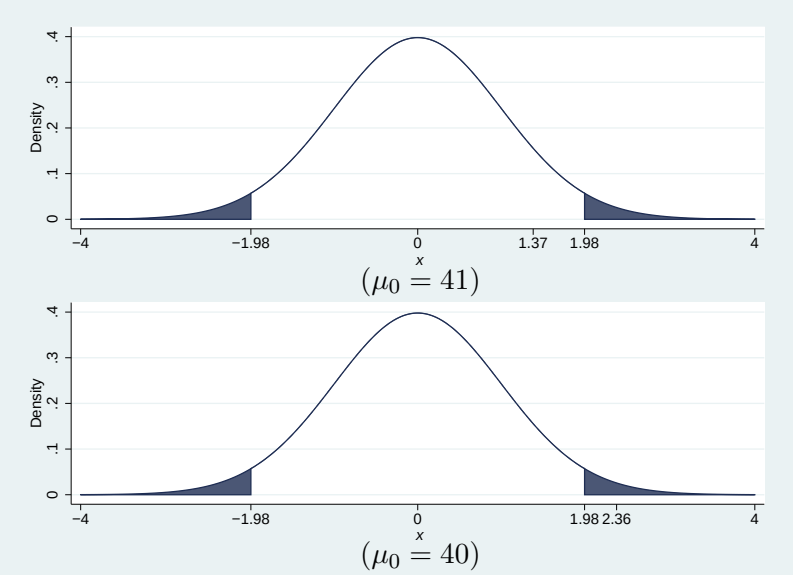

Figure 6: *Theoretical t-distribution standardized with* $\mu = 41$ *and* $\mu = 40$

Note that if we would have tested if the population parameter is equal to 40, which we know is almost the true parameter, then we would have rejected the null hypothesis ($t_{age} = \frac{42.38844-41}{1.010429} = 2.3637881$) and this would have been a mistake because we know 40 is right. This mistake is known as the type I error (see figure 6).

7 Appendix: proofs

Here I try to show you why the sample mean is an unbiased estimator of the parameter μ and also why the standard error is an unbiased estimator of the standard deviation of the sampling distribution. Before I start, keep in mind that these kind ob abstract subjects are of no real relevance for the exam. Anyway, understanding this makes obviously easier to understand what is being done. Those of you who find this kind of abstract things exiting, I recommend you to do the Minor in Statistik und Ökonometrie.

First of all, we need to clarify that each age of each individual in the sample is a random variable, because before we ask each individual about his age, almost any outcome is possible. But because we assume a distribution of age in the population (in this text we know this distribution), we have useful information about the possible outcomes and their probability.[9] This assumption about the distribution is made for all the individuals, because they all come from the same population distribution (i.e. from the same data generating process).

If we also make an assumption about the relation between variables (i.e. individuals), that the outcome of one of them does not influence the outcome of another one (an assumption that we usually make with cross-sectional data), things get even simpler. Both previous assumptions can be summarized by saying that all these random variables (age of each person) are **independent and identically distributed (iid)**, i.e. each random variable (age) has the same probability distribution as the others and all are mutually independent.

Because we have assumed a normal distribution of age in the population, we know that for each random variable age it's probability distribution is $X_i \sim \mathrm{N}(\mu, \sigma^2)$, where μ represents the expected value of the population probability distribution (think of this as a weighted average of all possible outcomes) and σ^2 its variance.

Our purpose is now to estimate the parameter μ through the values of age that we have. If we keep on thinking about each age as random variable with the same probability distribution and independent from each other, we have:

$$\mathrm{E}[\bar{X}] = \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right],$$

where X_i represents the random variable age of each person in our sample. By relying on some basic properties of probabilities and sums[10], we get:

$$\mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}[X_i] = \frac{1}{n}n\mu = \mu,$$

so that the sample mean is an unbiased estimator of μ.

In the case of the standard deviation of the sampling distribution, we look at the variance of the random variables. We have assumed that each variable age comes from the same probability distribution with variance σ^2 and that they are independent.[11] We want to get an unbiased

[9] Strictly speaking the probability of a concrete value for a continuous variable is zero, so that we look at intervals.

[10] The expected value of a constant is the constant, i.e. $\mathrm{E}[c] = c$ where 'c' stays for constant. The expected value of a fraction (or multiplication) is equal to the expected value of the numerator divided by the expected value of the denominator, i.e. $\mathrm{E}\left[\frac{X}{Y}\right] = \frac{\mathrm{E}[X]}{\mathrm{E}[Y]}$. The expected value of a sum of variables is equal the sum of the expected values, i.e. $\mathrm{E}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n}\mathrm{E}[X_i]$. The sum of a constant is equal the size of the sum multiplied by the constant, i.e. $\sum_{i=1}^{n}(c) = nc$ where 'c' stays for constant.

[11] The assumption of independence makes the calculation of the sums of variances much easier. Usually the sum of the variances of two random variables is: $\mathrm{Var}[X+Y] = \mathrm{Var}[X] + \mathrm{Var}[Y] + 2\mathrm{Cov}[X,Y]$, but with independent observations the covariances are equal to zero and can be disregarded so that $\mathrm{Var}[X+Y] = \mathrm{Var}[X] + \mathrm{Var}[Y]$.

estimate of the variance of the sample average $\bar{X}$ (seen as estimator), and by relying on some basic properties of probabilities an sums[12] we get:

$$\mathrm{Var}[\bar{X}] = \mathrm{Var}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n},$$

so that the population variance divided by n gives an unbiased estimate of the variance of the sampling distribution.

The last step would be to also get an unbiased estimate for σ^2. This is specially complicated and I proceed without much explanation. Just believe that the following rules apply:

$$\mathrm{Var}[X] = \mathrm{E}[X^2] - \mathrm{E}[X]^2 \text{ and } \sum_{i=1}^{n}(X_i - \bar{X})^2 = \sum_{i=1}^{n} X_i^2 - n\bar{X}^2$$

The population variance is given by:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{n}(X_i - \mu)^2$$

We could try to estimate it by replacing μ with the sample mean (which we proved is an unbiased estimator of μ):

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

Now we want to prove that this estimator is also unbiased. Using the previously mentioned rules we get:

$$\mathrm{E}[S^2] = \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2\right] = \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2\right] = \mathrm{E}[X_i^2] - \mathrm{E}[\bar{X}^2]$$

and because we know that $X_i \sim \mathrm{N}(\mu, \sigma^2)$, so that $\mathrm{E}[X_i^2] = \mathrm{Var}[X_i] + \mathrm{E}[X_i]^2 = \sigma^2 + \mu^2$ and we also know that $\bar{X} \sim \mathrm{N}(\mu, \frac{\sigma^2}{n})$, so that $\mathrm{E}[\bar{X}^2] = \mathrm{Var}[\bar{X}] + \mathrm{E}[\bar{X}]^2 = \frac{\sigma^2}{n} + \mu^2$, making the necessary substitutions we get:

$$\mathrm{E}[S^2] = \sigma^2 + \mu^2 - \frac{\sigma^2}{n} - \mu^2 = \frac{n-1}{n}\sigma^2$$

so that S^2 is biased but we know this bias and can correct for it:

$$\mathrm{E}\left[\frac{n}{n-1}S^2\right] = \frac{n}{n-1}\frac{n-1}{n}\sigma^2 = \sigma^2$$

I.e. by multiplying S^2 with $\dfrac{n}{n-1}$ we get ride of the bias and get the corrected estimator for the variance:

$$S^{*2} = \frac{n}{n-1}S^2 = \frac{n}{n-1}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2\right] = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

[12] The variance of a constant is zero, but the variance of a random variable multiplied by a constant is equal to the constant squared multiplied by the variance of the random variable, i.e. $\mathrm{Var}[cX] = c^2\mathrm{Var}[X]$ where c stays for constant. Also rules about how to sum independent variances are needed and were previously given.