A picture containing shape

Description automatically generated

Universidad Politécnica de Valencia

Escuela Técnica Superior de Ingeniería Informática

Statistical Study of the

StackOverflow’s Survey 2022

for the Iberian Peninsula.

Logo, company name

Description automatically generated

Author: Rodríguez Díaz, Gabriel  
Tutor: Zarzo Castelló, Manuel

**Distributions in sampling - Inference about one population**

**22. Assume that the variable YearsCode follows a normal distribution with average equal to the sample average and standard deviation equal to the sample value. If we take a random sample with 10 values and we compute its average, compute the confidence interval that would comprise the 95% of these values. (Solve theoretically)**

Assuming that the variable YearsCode follows a normal distribution

N(μ=14,2389 ; σ=8,80874) and taking 10 random values we have that N=10.

We know that ) therefore, using the Z-Table, the interval will be

. Note that 2 would be around 1.96.

That gives an interval of 1- =95% l

**23. Assume that YearsCode follows a normal distribution. Calculate the percentile 52, (Z52).**

According to Statgraphics, the percentile 52, Z52= 13.0

**23.1 Solve the contrast between H0: m = Z52 over H1 : m ≠ Z52 considering as size of the sample the observations and significance level of 10%.**

Considering n=494 and α=10%

Afterwards we obtain tn-1 from the table. Doing so, we get t494 = 0,999932

Finally, we should apply the following formula:

~tn-1 = ~t493 = 3,1259

Since 3,1259>0,999932 we don’t reject the null hypothesis for α=10%

**23.2 Solve the same hypothesis contrast with Statgraphics.**

Using Statgraphics you have to follow the next steps:

1. *Describir > Datos numéricos > Análisis de una variable*

*(Seleccionar YearsCode)> Marcar Test de hipótesis.*

1. Select **Prueba de hipótesis para YearsCode**
2. *Right click > Opciones de ventana*
3. Introduce 14,2389 in *Media/Mediana*
4. Introduce 10% in *Alfa.*

Once followed the previous steps, you just have to have a look at the results.

The following output will be shown:

Media Muestral = 14,2389

Mediana Muestral = 12,0

Desviación Estándar de la Muestra = 8,80874

Prueba t

Hipótesis Nula: media = 14,2389

Alternativa: no igual

Estadístico t = -0,0000847873

Valor-P = 0,999932

No se rechaza la hipótesis nula para alfa = 0,1.

**24. Assume that YearsCode follows a normal distribution X ≈ N(m;σ) with media and typical deviation equal to the sample´s.**

**If a 10 sample from this population is randomly taken and the variance is obtained, calculate the confidence interval for the 95% of the values.**

Assuming that the variable YearsCode follows a normal distribution N(μ=14,2389; σ=8,80874) and taking 10 random numbers, we have that N=10.

We know that (N-1) follows a Chi Square distribution with N-1 degrees of freedom, for this case, 9.

We compute the interval that comprises the 95% of the values of a Chi Square distribution Χ29 . In this case [3,3251;16,9189]

Let´s calculate the interval for which s2 belongs.

**25. Assume that YearsCode follows a normal distribution X ≈ N(m;σ) with media and typical deviation equal to the sample´s.**

**If a random sample of 12 data from the population is taken and the variance is obtained, what is the possibility of it being greater than 3 σ?**

Assuming that the variable YearsCode follows a normal distribution N(μ=14,2389; σ=8,80874) and taking 12 random numbers, we have that N=12.

We know that (N-1) follows a Chi Square distribution with N-1 degrees of freedom, we can compute the probability we are asked to in the following way:

Finally, using Statgraphics, we can compute the probability mentioned above.

1. *Describir > Ajustes de distribuciones > Distribuciones de probabilidad > Chicuadrado.*
2. G.L. (Grados de Libertad) with value N-1, in this case 11
3. Check in *Distribuciones acumuladas.*
4. *Right click > Opciones de ventana > Variable Aleatoria > 3,746*
5. The table will be computed and we take the value for *Área Cola Superior >*

Thanks to Statgraphics we have that

The probability of the variance being greater than 3σ is quite high.

**26. Using YearsCode, if two random samples of 14 values are taken, which is the probability of the variance of the second sample is three times the first´s?**

Even if the exercise indicates to use YearsCode, in the end, this will be irrelevant, since we will not take σ2 into account. Now we will see why.

Taking two random samples and we will have to calculate but , and so, in the end we have

Now we use the following formula:

As we mentioned before, the dataset used does not change the probability of the second sample´s variance being 3 times the first´s.

**27. Obtain a Confidence Interval of the 99% for the average of YearsCode.**

The average for YearsCode is μ=14,2389, we can find the Confidence Interval of 99% for this media with Statgraphics following the next steps:

1. *Describir > Datos numéricos > Análisis de una variable > YearsCode >* Check *Intervalos de Confianza*
2. *Right Click > Opciones de ventana > Intervalo de Confianza = 99%*
3. Take the confidence interval for *media.*

Now we have got the confidence interval for the average, which is

or

**27.1 What would happen in case YearsCode did not fit a normal distribution?**

This would mean that the calculus made are not valid.

**27.2 “If any value belonging to the confidence interval is taken and an hypothesis test is performed over the average, the conclusion will be always the same taking α=1%” Is it true? Why?**

For any value m inside the range the conclusion is not reject. Therefore, this is true.

**28. Using YearsCode, obtain with Statgraphics a confidence interval of 95% for the typical deviation of YearsCode. Calculate the Interval with a 99%**

The typical deviation for YearsCode is σ=8,80874, we can find the Confidence Interval of 99% for this media with Statgraphics following the next steps:

1. *Describir > Datos numéricos > Análisis de una variable > YearsCode >* Check *Intervalos de Confianza*
2. *Right Click > Opciones de ventana > Intervalo de Confianza = 95% or 99%*
3. Take the confidence interval for *desviación típica.*

Now we have got the confidence interval for the typical deviation, which is

For 95%

For 99%

**28.1 Which interpretation does it has?**

In this case, for α=5% the interpretation is just how wide we want the range to be.

**28.2 Which interval looks better? Which are the factors?**

For this case it does not really matter. If we were following an efficiency or economic criteria, like it may be used in a factory, we should take a smaller range, that is, a bigger alpha.

We may note that in industry the most followed criteria tend to be

**29. Indicate in a table the variance of YearsCode and the amount of data for each one of the variants of Country.**

|  |  |  |
| --- | --- | --- |
| **Country** | **YearsCode ( 2)** | **Observations (N)** |
| **Spain** | 78,7185 | 402 |
| **Portugal** | 73,1671 | 92 |

TABLE 14: Variance of YearsCode and Observations for the subset Country.

**29.1 Can we estate that the differences are statistically significant?**

Considering α=0.05 for these computations.

Let´s define the null hypothesis:

We have to compute the ratio

To do so, we have to compute the confidence interval and evaluate the ratio against the confidence interval

Now, using the table, we get the following result:

Since the value 1,075≤1,419 we cannot reject the null hypothesis, therefore there are not significant differences.

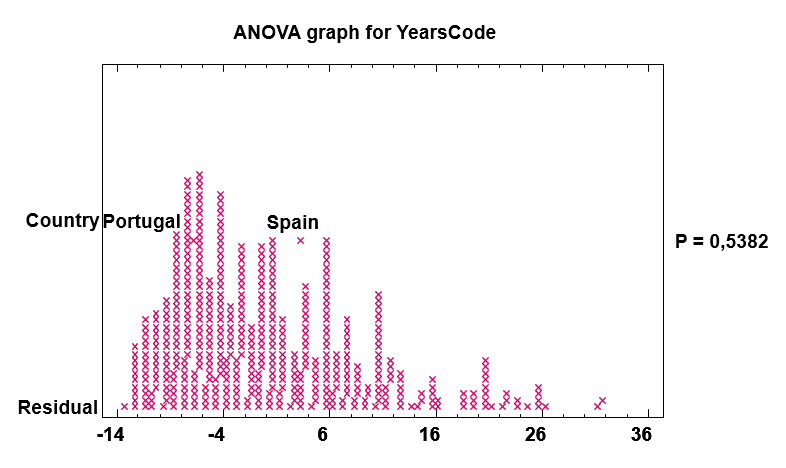
**ANOVA – Analysis of Variance**

**30. Make and ANOVA to study the effect of Country on the variable YearsCode.**

Firstable, it is important to remark that a 0,05 significance level was chosen. The main reasons are the following:

* The study does not require a high level of confidence. It is just a simple study to know the relationship between variables. This information is not relevant for any other purpose that may require a high precission.
* The sample size also matters. Using a huge sample size is more difficult to determine smaller effects. That is why an smaller significance level may be used. For this case, it is not a problem since the sample is small.
* It provides a better balance between making type I and type II errors. A 0.01 increases the possibility of accepting the null hypothesis when it is false.

Afterwards it was necessary to transform the data, because it was positive skewed.



PICTURE 31: Anova graph for YearsCode

As we can see, this graph is positively skewed, so some transformation will be performed.

To assess skewness, we can look at the Standarized Skewness values. Both Portugal and Spain have positive Std. Skewness values (5.359 and 7.56, respectively).

A picture containing text, screenshot, diagram, plot

Description automatically generated

PICTURE 32: Anova graph for transformed YearsCode

For this case the total Skewness Coefficient is -0,405 which makes it the best option.

Now, we can have a look at the ANOVA table of the model.

**Tabla ANOVA para YearsCode^0.25 por Country**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Source** | **Square Sum** | **Gl** | **Cuadrado Medio** | **Razón-F** | **Valor-P** |
| **Entre grupos** | 0,0098051 | 1 | 0,0098051 | 0,10 | 0,7462 |
| **Intra grupos** | 46,003 | 492 | 0,0935021 |
| **Total (Corr.)** | 46,0128 | 493 |

TABLE 14: ANOVA table for .

**30)** **[2 p.]** Realizar un análisis de la varianza (ANOVA) para estudiar el efecto de F2 en la variable **X1**.

- En caso de asimetría positiva fuerte, hay que utilizar la transformación más adecuada.

- Mostrar la tabla resumen del ANOVA del modelo.

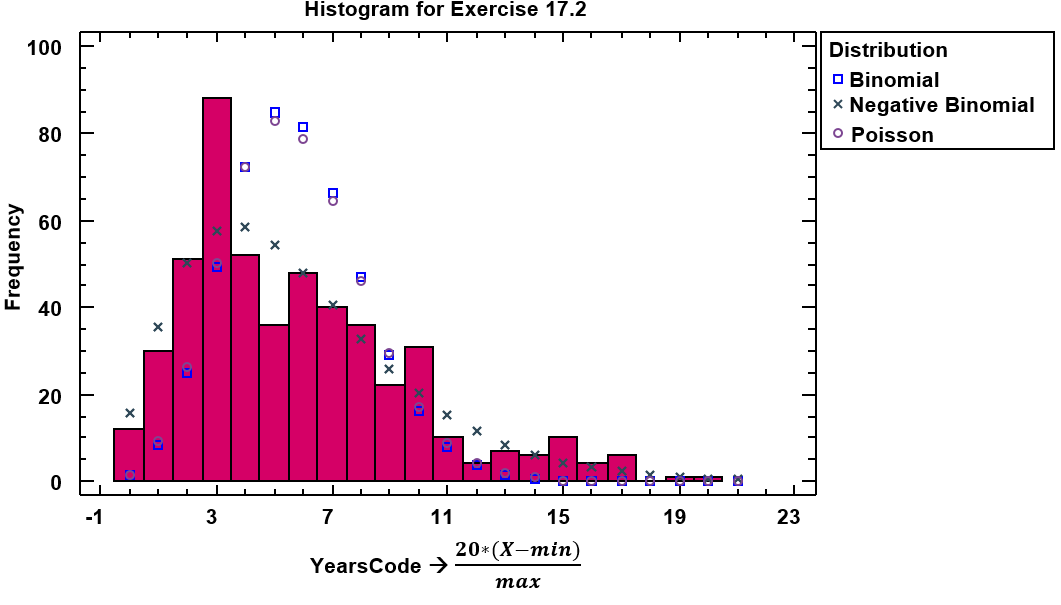
- Mostrar el gráfico de medias con intervalos LSD (eligiendo el  considerado más adecuado):

* Indica si todos los intervalos tienen la misma amplitud o no y por qué (1-4 líneas).
* ¿Las conclusiones que se derivan del gráfico son coherentes con la tabla resumen del ANOVA?

- Interpreta el análisis y explica las principales conclusiones derivadas del mismo (2-9 líneas).

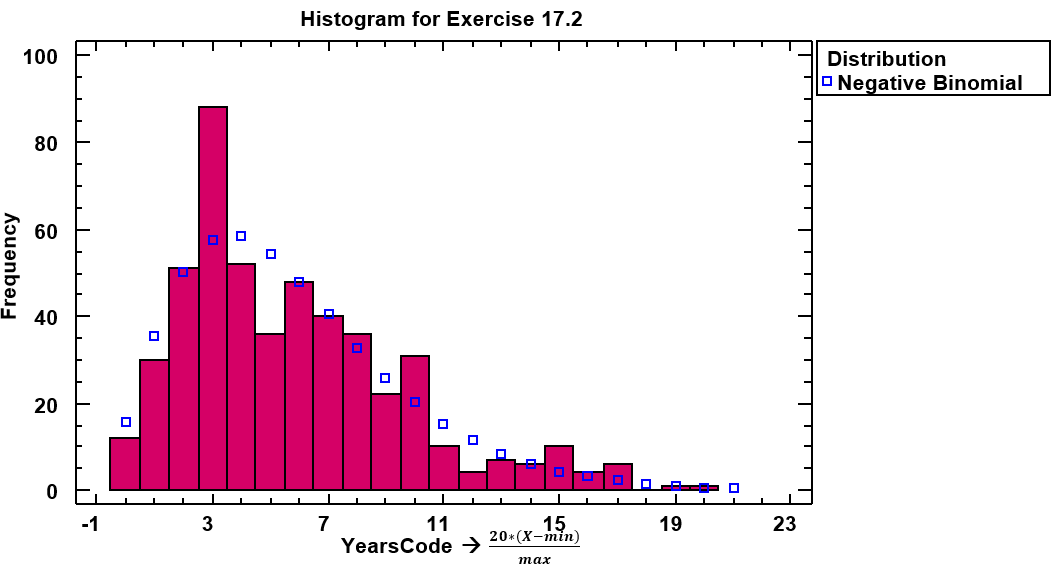
- Muestra el papel probabilístico normal de los residuos. A partir de este gráfico, indica si existen residuos anómalos que deberían descartarse del modelo. En caso afirmativo, repite el análisis y explica los cambios en las conclusiones si los hubiera.

**17.2 Take a continuous variable and make a new one using and round the results. Finally, answer the indicated questions.**



PICTURE 24: Different distributions compared with the histogram of the variable modified using the above formula

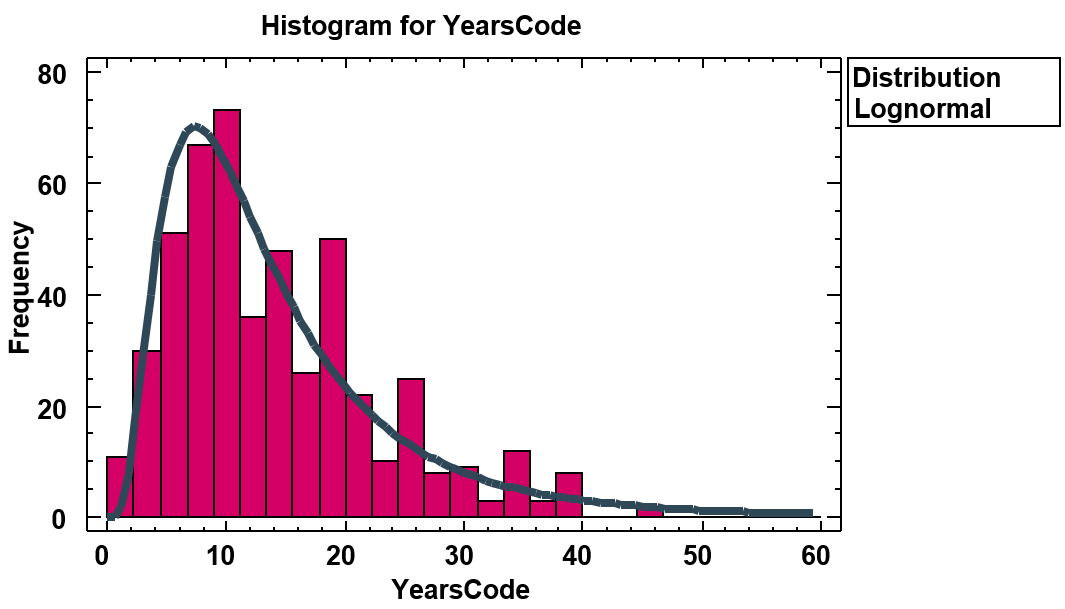
As we can see, none of the variables that were reviewed in class fit the histogram. Nevertheless, the Negative Binomial or the Pascal distribution fits more or less if we do not focus much on that extreme value on the left.



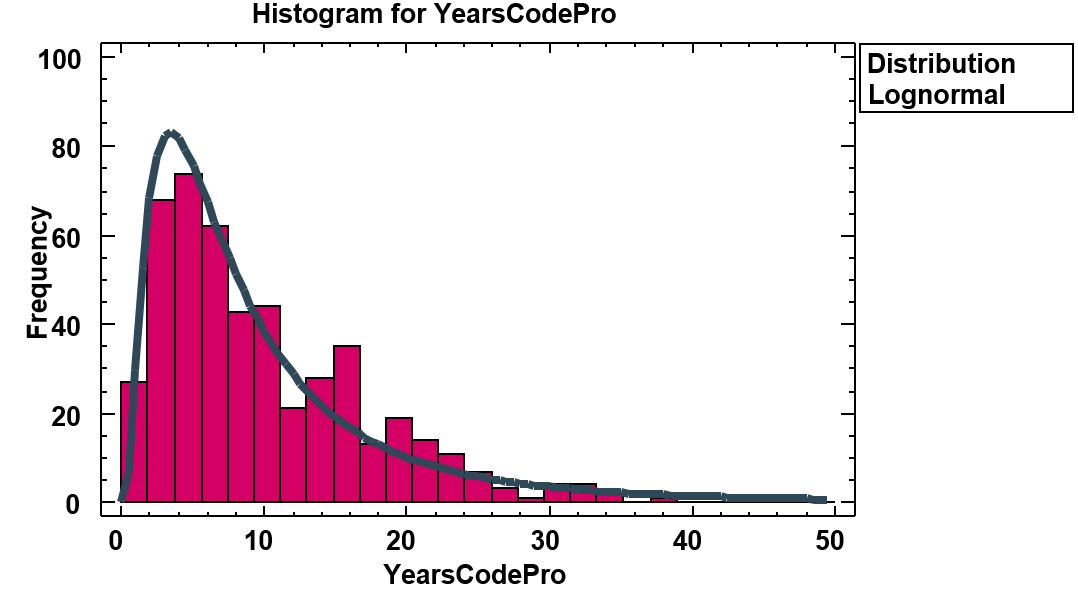
PICTURE 25: Negative Binomial compared with the histogram of the variable modified using the above formula

**18.1 For two continuous variables whose distribution is not normal, study the best fitting model of distribution. Place the graphic and discuss the conclusions.**

The chosen variables are YearsCode and YearsCodePro. Since they are more or less similar, it is interesting to appreciate the differences between both.



PICTURE 26: Histogram of YearsCode vs Lognormal distribution



PICTURE 26: Histogram of YearsCodePro vs Lognormal distribution

Clearly, both distributions fit properly into the lognormal distribution. It is important to remark that in the YearsCodePro distribution there are not many extreme values, which may make this variable even better than YearsCode.

We can also note, that the extreme values tend to concentrate in round values such as 10, 15 or 20. This may happen because people tend to answer this kind of numbers when they are not sure of the exact value or they are very close to these certain values.

|  |  |  |
| --- | --- | --- |
|  | **YearsCode** | **YearsCodePro** |
| **Media** | 14,5808 | 10,3679 |
| **Est. Dev.** | 11,0852 | 10,8424 |
| **Log. Media** | 2,45163 | 1,96926 |
| **Log. Est. Dev.** | 0,675393 | 0,859592 |

TABLE 11: Parameters for each of the studied variables

The Kolmogorov-Smirnov test is a statistical technique used to determine whether a sample of data follows a given distribution.

**DMAS** and **DMENOS** represent the maximum positive and negative discrepancies between the two cumulative distribution functions. On the other hand, **DN** represents the test statistic, which is calculated as the maximum absolute difference between **DMAS** and **DMENOS**.

Finally, the P-Value provides a quantitative measure of the strength of evidence against the null hypothesis.

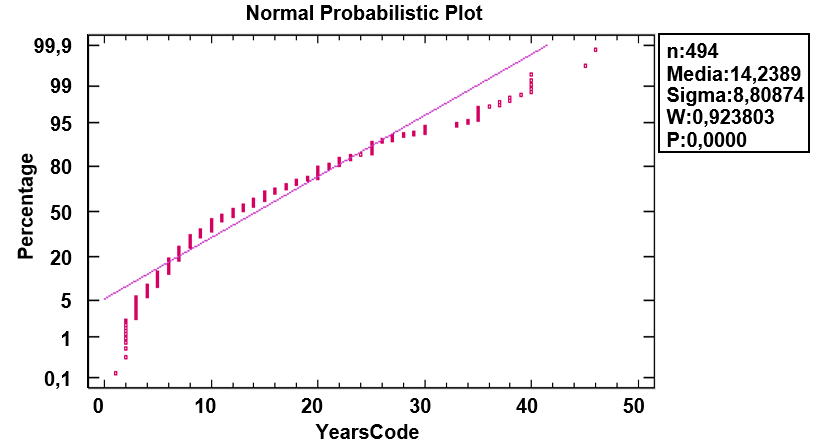
For this case, the Kolmogorov-Smirnov test, shows that the Lognormal distribution fits more or less the data. If we compare it, for instance with the triangular, these values are higher since triangular would have a P-Value of 0.

|  |  |  |
| --- | --- | --- |
|  | **YearsCode** | **YearsCodePro** |
| **DMAS** | 0,0341079 | 0,0579893 |
| **DMENOS** | 0,0667972 | 0,0812606 |
| **DN** | 0,0667972 | 0,0812606 |
| **P-Value** | **0,0243512** | **0,00348525** |

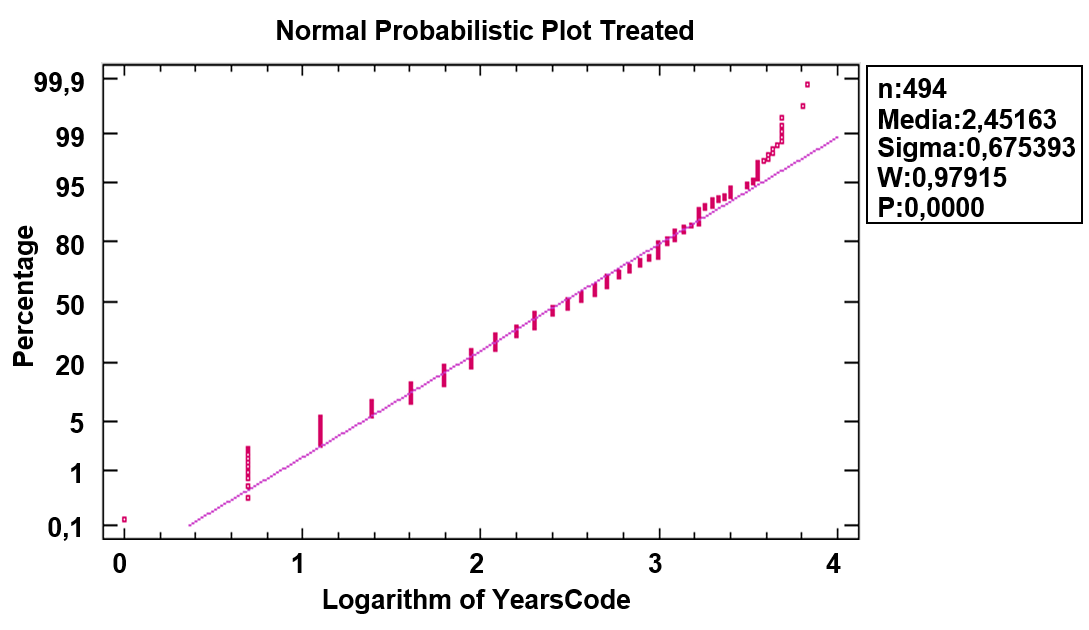
TABLE 11: Kolmogorov-Smirnov values for each of the studied variables

As we can see, the P-Values are quite low. This kind of tests are very interesting, because at first sight YearsCodePro seemed to fit better, but Kolmogorov-Smirnov is the mathematical proof that it was not like that. YearsCodePro is closer to zero than YearsCode.

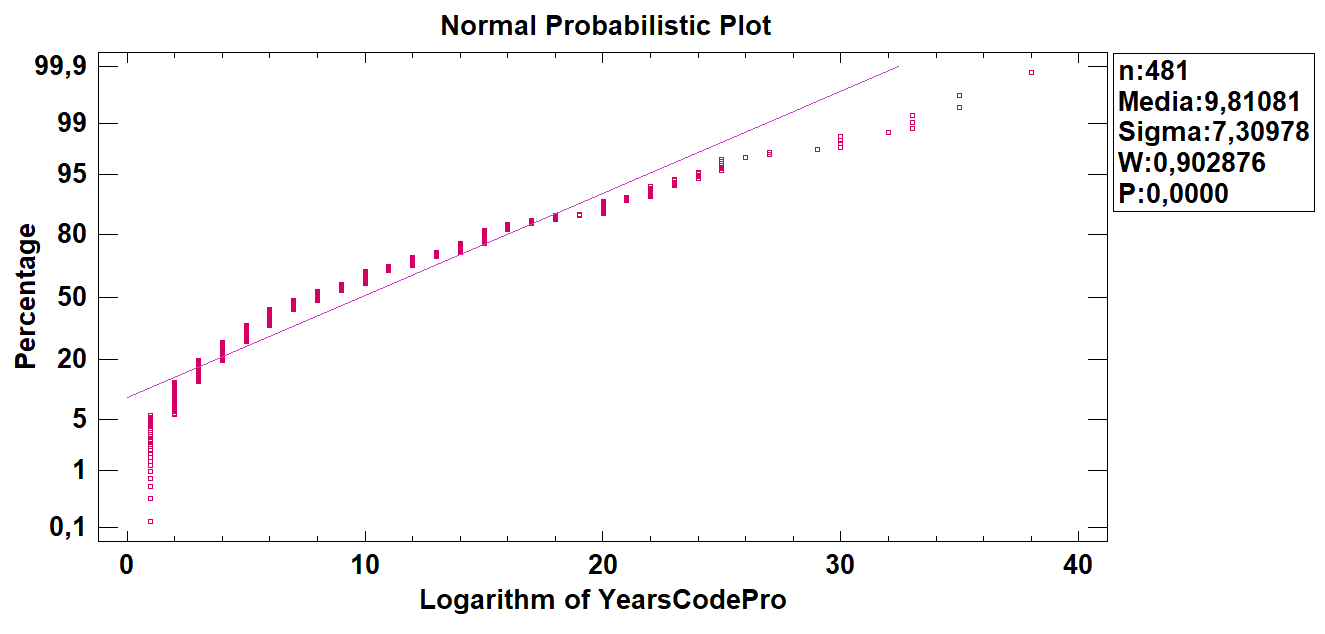
**19. For two continuous variables with assymetric positive distribution, study if any transformation is able to normaliza the data. Also, add the Normal Probabilistic Plot.**



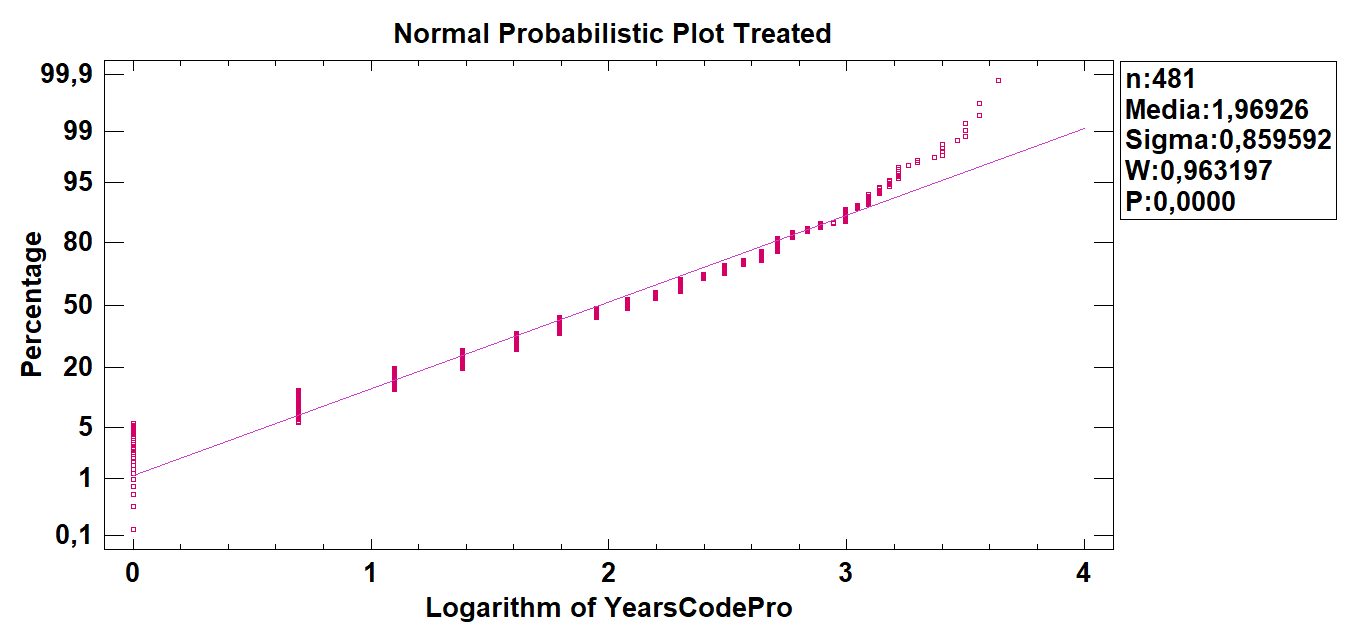
PICTURE 27: Normal Probabilistic Plot for YearsCode



PICTURE 28: Normal Probabilistic Plot for YearsCode with logarithm applied



PICTURE 29: Normal Probabilistic Plot for YearsCodePro with logarithm applied



PICTURE 30: Normal Probabilistic Plot for YearsCodePro with logarithm applied

The transformation chosen for both cases was the logarithmic one, since both were positively skewed.

In both cases, no transformation was successful. A possible explanation could be that the IT sector is not as consolidated as others. Many years ago, people used to study other engineerings and that is why they know how to program but they are not programmers by profession.

Also, the increasing demanding in the sector lead that many people started studying coding even if they already have degrees in other fields. That may explain the skewness of the graphic.

In my opinion, as the offer and demanding in the labor maket start stabilizing in the future, the tendency will be a normalization of this kind of data.

**19. Which of the variables follows a reasonably normal distribution? In case there is not any, argue the reasons.**

As it was stated, previously in this project, all the variables are quite skewed and none can be considered as normal.

The reasons for YearsCode and YearsCodePro are explained in the point of **18** this project.

Now, for the variable Salary, which can also be analyzed, the reason is simply how the labor market works, among other reasons.

If we talk about high salaries, in the IT sector you can earn salaries higher than 50000€ or even 100000€

When you are a highly skilled engineer with many years of experience, the lack of profiles like yours is a very good advantage while you are negociating a salary. There is an increasing demand of programmers with a low offer of professionals.

Also many people work for other countries with much higher income levels compared to Spain and Portugal.

On the other hand, if we take the low tail of the distribution, we can argue that a lot of people are starting to study Computer Science (So, they start with lower salaries). As we can see, for example, here at the ETSINF UPV, the Computer Science degree is the most demanded of the university and the cut-off marks for these kind of degrees have been increasing a lot.

Another reason may be that the medium salary in Portugal was around 23.9% lower than Spain’s in 2022 (Data from Datosmacro.expansion.com). This means that Portugal (Or Spain) is probably skewing the data.

This last hypothesis could be studied in the future with the current dataset.

**20. Generate 100 random values of N(m=15,****=4) and 100 more with N(m=3,****=3). Generate a new variable by summing the values by pairs.**

**20.1 Describe the process**

Firstable, I created the distributions in Statgraphics.

*Gráficos > Distribuciones de probabilidad > Normal > Insert the distributions*

Then I generated the data using this button  and moved it to Excel.

Finally, I computed the sum by pairs in Excel with the function *=ColA+ColB* and moved it back to a new column in Statgraphics.

**20.2 Based on asymmetry coefficient and Kurtosis, check if the variable is adjusted to a Normal Distribution.**

|  |  |
| --- | --- |
| **Standarized Skewness** | **0,404565** |
| **Standarized Kurtosis** | **-0.7497** |

TABLE 12: Standarized Skewness and Standarized Kurtosis for SUMA.

Based on the Skewness and Kurtosis values alone, the distribution appears to deviate from normality, primarily due to the negative Kurtosis value. However, the skewness is not strongly positive, so the deviation from normality may not be significant.

**20.3 Calculate the mean and the standard deviation for SUMA.**

|  |  |
| --- | --- |
| **Mean** | **18,3215** |
| **Standard Deviation** | **5,25152** |

TABLE 13: Mean and Standard Deviation for SUMA.

**20.4 Using theoretical calculus, which would be the expected mean and standard deviation?**

Being m1 the average for N(15,4) and m2 the average for N(3,3)

SUMA Average = = 9

Being ****the stdev for N(15,4) and ****the stdev for N(3,3)

Standard Deviation ****= == =

**20.5 Why do the theoretical values do not coincide with the observed ones?**

That is because we are generating random numbers for both of the normal distributions and saving them. Since we are only taking 100 random values that follow that kind of distribution, the average and the standard deviation are never going to be totally accurate. For instance, for the first normal distribution with average 15 and standard deviation 4 we obtain values of **18,3215** and **5,25152** respectively. That is because 100 values are not enough to obtain an accurate distribution. If we increase the number of values from 100 to 10000, then we obtain an average of **14,9885** and a standard deviation of **4,00597** for the first distribution, which are much more accurate. The parameters of the variable sum computed theoretically do not take into account the random values that we are generating, but Statgraphics does. That is why they do not match how we would expect them to. However, the more random values you generate, the more accurate they become. Since in this case we only have 100 values it is normal to get an average and standard deviation that are only an approximation of the theoretical ones.

**Note:** For the computations with 10000 values, I did the following:

I created a new column called “Random1 10k” on Statgraphics. Afterwards, I went to Excel where I used the following formula: *=DISTR.NORM.INV(ALEATORIO();15;4)*

I generated the 10000 registers and moved them to Statgraphics, where I computed the proper average and standard deviation.