Aplicacion de tecnicas de estimacion y prueba de hipotesis

Caso: tendencias socio-economicas de algunas lineas de carrea de Ingenieria de sistemas

Alvarez Bautista Burga Casanova Cuyate

Facultad de Ingenieria Industrial y de Sistemas
Universidad Nacional de Ingenieria

Octubre 2022

Tabla de Contenido

Tabla de Contenido

Problematica

- Empiricamente se observa que en el mundo la precarizacion del trabajo se acrecenta cada vez mas, de igual modo con la evolucion de personas casadas y los salarios promedio de los jovenes
- Con motivo de generar informacion util para la prediccion de estas tendencias socio-economicas se ha procedio a realizar un analisis estadistico sobre 9 hipotesis planteadas

Tabla de Contenido

Objetivos del trabajo

General

Generar informacion relevante para la prediccion de tendencias socio-economicas en el mundo tomando como referencia datos provenientes de distintos paises.

Hipotesis especificas

- La distribucion de ingresos de sigue la ley normal
- La eleccion de especialidad es causa de la formacion de 2 grupos en la poblacion.
- La distribución de ingresos de los trabajadores en software no sigue una distribución exponencial
- En paises desarrollados existe una mayor cantidad de mujeres con puestos de trabajos relacionados a ingenieria que en paises en via de desarrollo
- Los cientificos de datos poseen un mejor distribucion de ingresos que los ingenieros de datos
- El sector *(publico / privado)* al que pertenece un trabajador es causa de la diferencia de salarios

Hipotesis especificas

- Las personas que trabajan una cantidad de horas superior a la media tienen una mejor destribucion de ingresos que aquellas que no lo hacen
- Las personas de mediana edad poseen una mejor distribucion de ingreso que las personas jovenes
- El promedio de ingresos de la poblacion mexicana es mayor que la peruana

Cada hipotesis tiene asociado el objetivo de comprobar o rechazar la suposicion

Tabla de Contenido

Acerca de la normalidad de la distribucion de ingesos

Ever no te olvides del texto

Sobre la eleccion de especialidad

Se desea analizar principalmente:

1. Si la eleccion de la especializacion es causa de la formacion de 2 clusters en la poblacion
2. Que tan confiable resulta usar **ANOVA** [1]

[1] *Analysis of Variance*

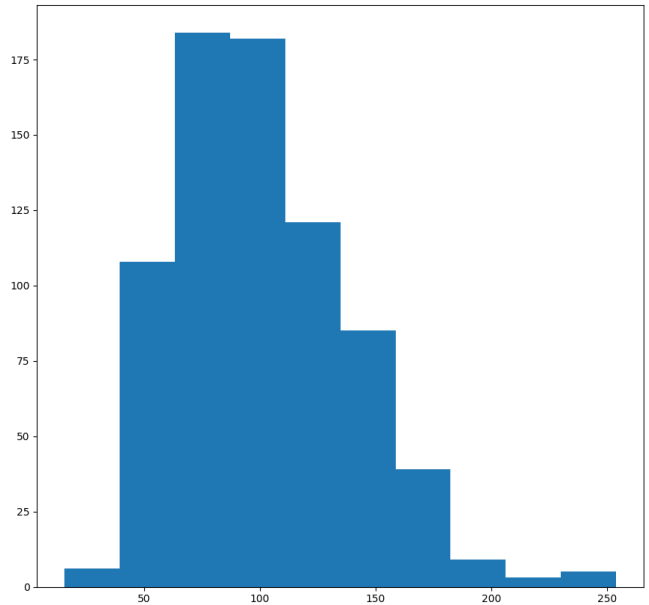
Sobre los demas objetivos

Se consideraron las hipotesis de tal forma que brinde información relevante para el análisis del mercado laboral para los egresados de la carrera de *Ingenieria de sistemas*

Tabla de Contenido

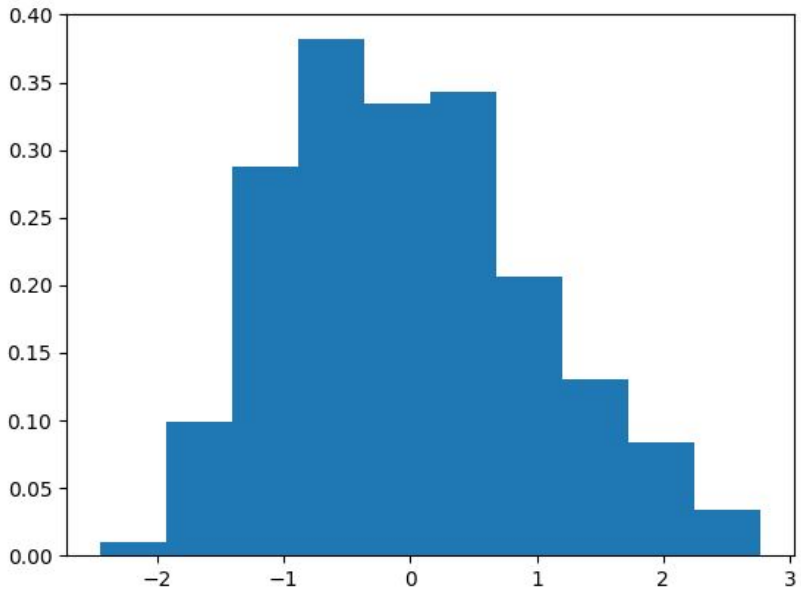
Hipotesis 1

Figura: Data sin estandarizar



Hipotesis 1

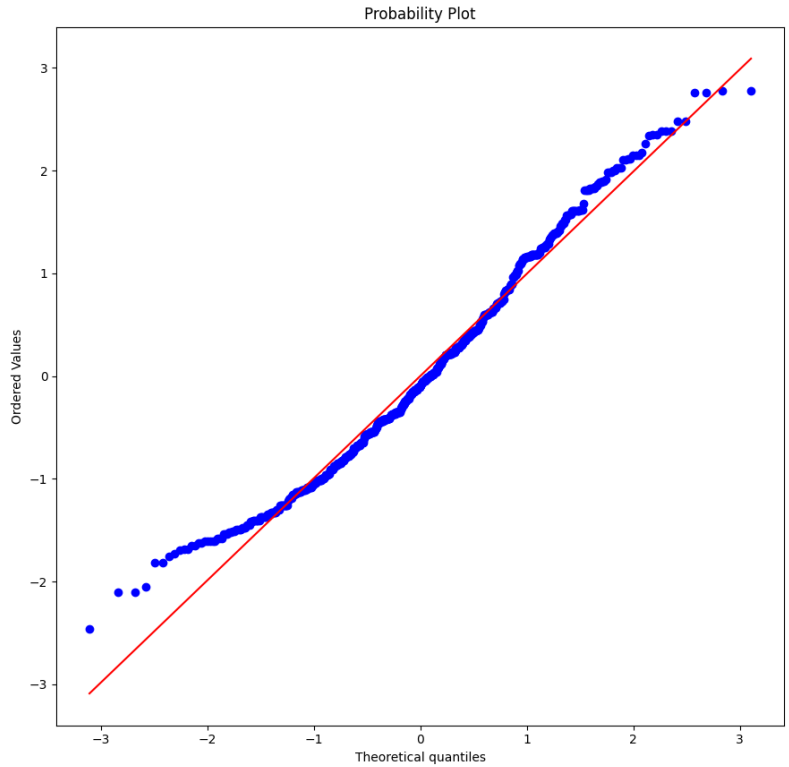
Figura: Data estandarizada y sin outliers



Puede parecer una distribucion Normal

Hipotesis 1

Figura: Grafica Q-Q



Se aleja de la distribucion normal

Se aplicó el test de *Jarque-Bera*, para comprobar si la muestra presenta una **curtosis** y **asimetria** correspondientes a una ley normal.
El estadistico de *Jarque Bera* es asintoticamente un estimador de una *Chi-Cuadrado* (χ^2_n) y toma como hipotesis nula que los datos de la muestra siguen la ley normal

Test de Jarque-Bera

$$\mathbf{JB} = \frac{n}{6}(S^2 + \frac{1}{4}(K-3)^3)$$

Siendo n los grados de libertad

Estimadores de momentos centrales

- Tercer Momento Central

$$S = \frac{\hat{\mu}_3}{\hat{\sigma}^3}$$

- Cuarto Momento Central

$$K = \frac{\hat{\mu}_4}{\hat{\sigma}^4}$$

Adicionalmente, se usara el test de *Kolmogorov-Smirnov*, donde se plantea que la distribucion de ingresos en la poblacion de ciencia de datos no sigue la ley normal y se comparará con la funcion acumulada teoria de esta

Conclusiones hipotesis 1

El test K-S y el de Jarque-Bera muestran los siguientes p-values.

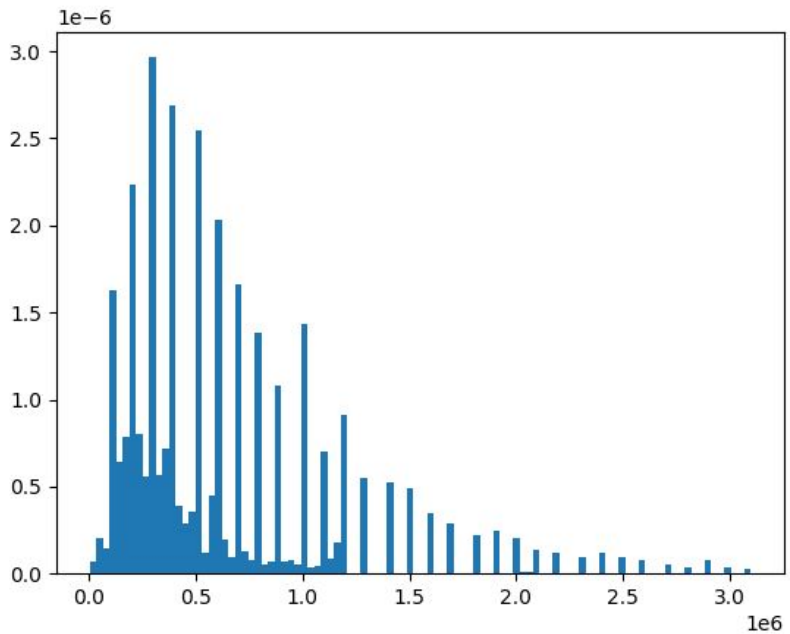
```
Desarrollo de la primera hipótesis
Jarque_beraResult(statistic=24.54482110101632, pvalue=4.6790729723023006e-06)
KstestResult(statistic=0.06189822129196375, pvalue=0.007064124010629 26)
```


```
Desarrollo de la primera hipótesis
Jarque_beraResult(statistic=24.54482110101632, pvalue=4.6790729723023006e-06)
KstestResult(statistic=0.06189822129196375, pvalue=0.007064124010629 26)
```

Hipotesis 2

Antes de realizar cualquier tecnica de inferencia es necesario conocer la forma de las distribuciones, incluso antes de analizar la varianza

Figura: Distribución de ingresos de ingenieros de software en la India



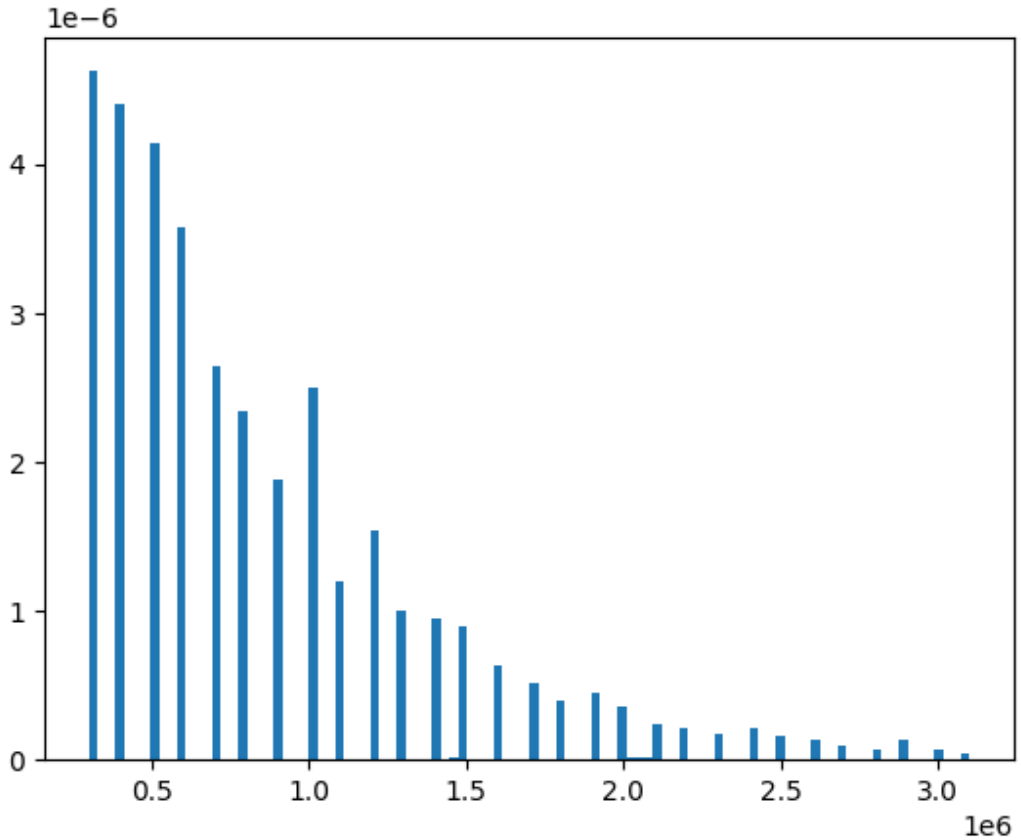
se puede notar como existen *2 grupos en la poblacion*

Distribución de los trabajadores en software

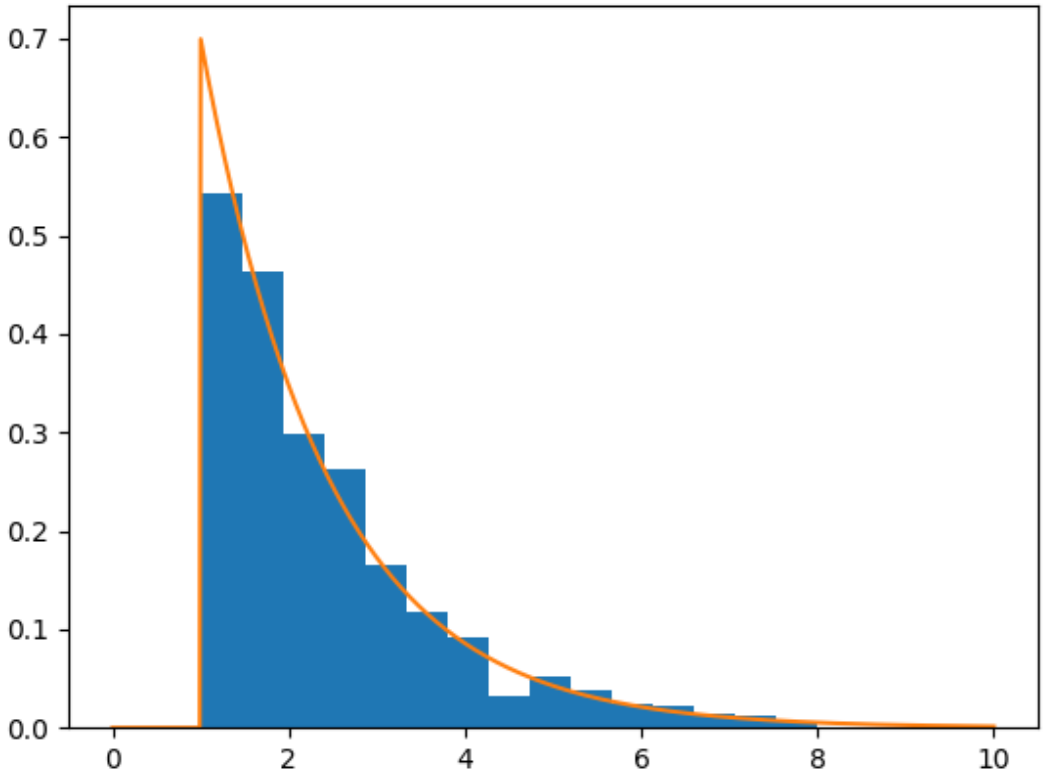
Aplicacion del test **Kolmogórov-Smirnov**

En este caso se va a comprar la funcion de distribucion acumulada observada con la de la distribucion teoria de una exponencial

Separando grupos aparentes



Ajustando Curva



Funciones acumuladas

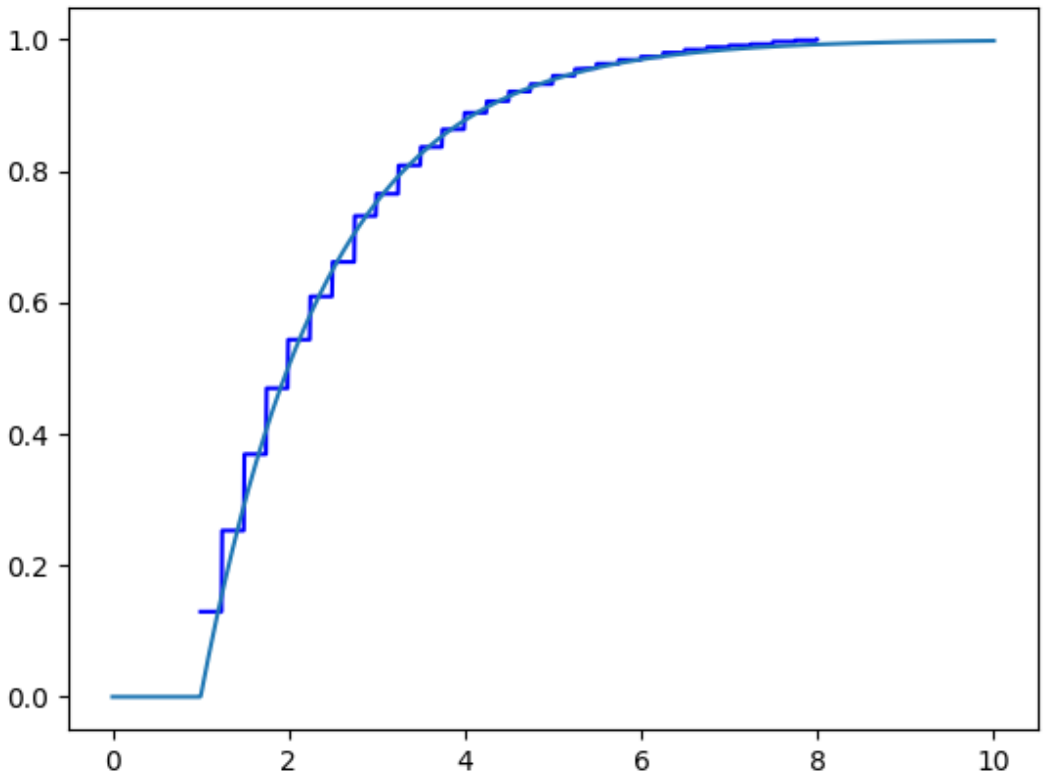
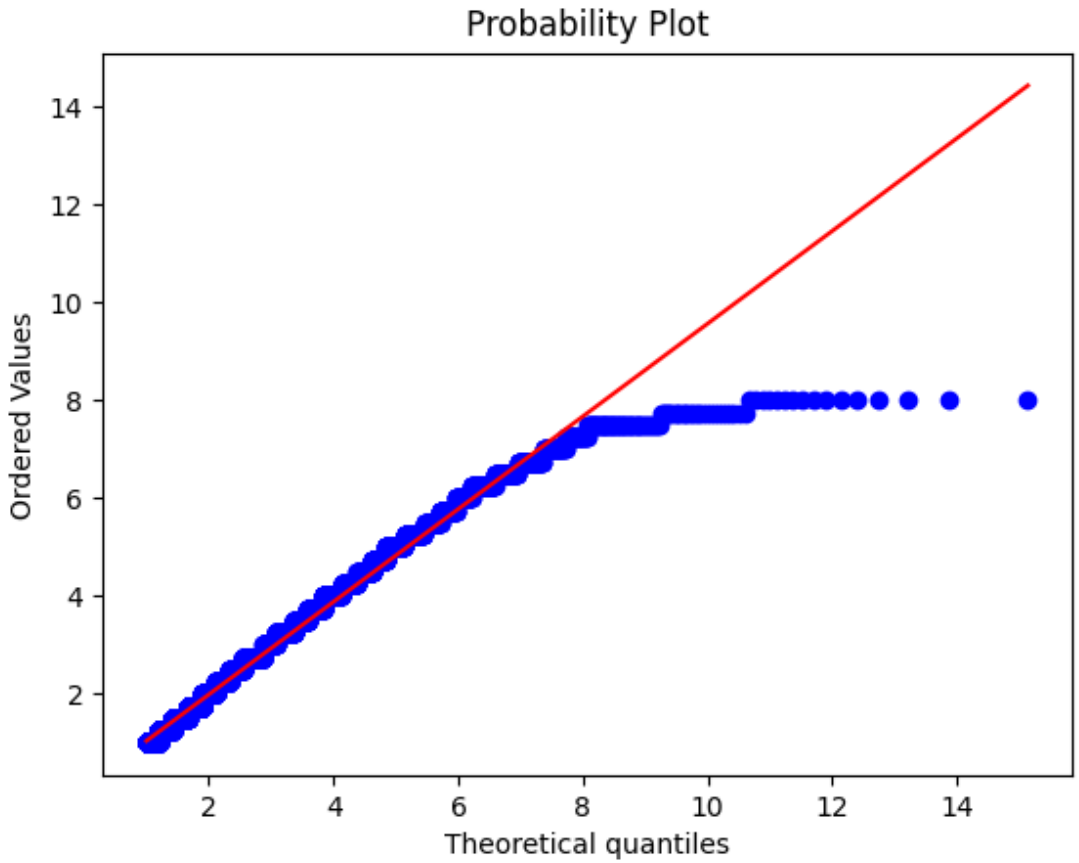


Grafico P-P


Probability Plot
Ordered Values
Theoretical quantiles
2
4
6
8
10
12
14

Conclusiones

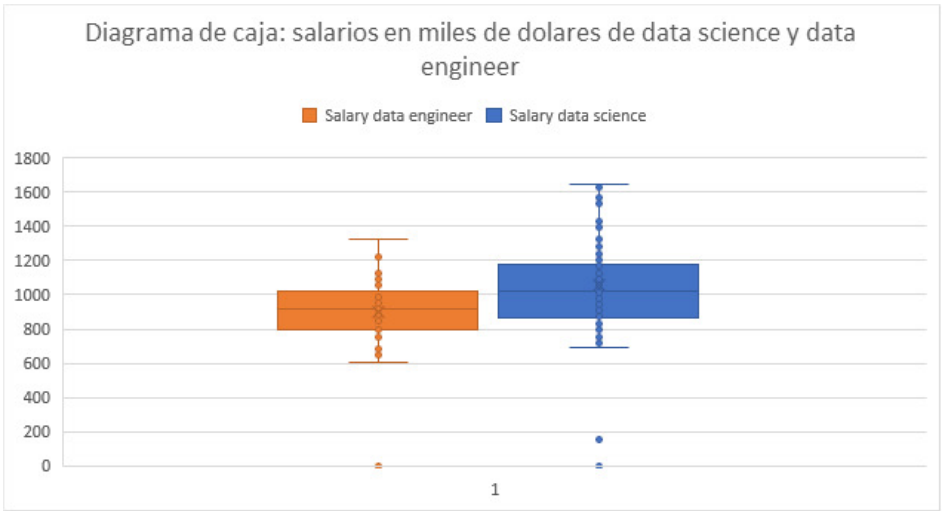
De acuerdo al p-value obtenido no se puede rechazar la hipótesis nula

```
Length of list:  11128
KstestResult(statistic=0.129449368008136, pvalue=1.4083532442224077e-201)
```

Comparacion de salarios DC y DI

X_1 : Salario de cientifico de datos $\rightarrow \overline{X_1} = 1061{,}79389312977, {\sigma_1}^2 = ?$

X_2 : Salario de ingeniero de datos $\rightarrow \overline{X_1} = 916{,}603773584, {\sigma_2}^2 = ?$



Test de hipotesis

Como se puede observar las alturas de ambas son diferentes, por lo que se puede considerar que existe una diferencia significativa entre ambas varianzas poblacionales.

Luego:
Estadisitico de prueba es:

$$t = \frac{\overline{X_1} - \overline{X_2} - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \sim t_{(v)}$$

$$v = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{\left(\frac{S_1}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{S_2}{n_2}\right)^2}{n_2 - 1}}$$

Con un nivel de significancia de: $\alpha = 0{,}05$
Reemplazamos datos:
t = 4,74240775380578
v = 149
Region crítica:

$$t_{(149;\, 0{,}95)} = 1{,}65514453379796$$

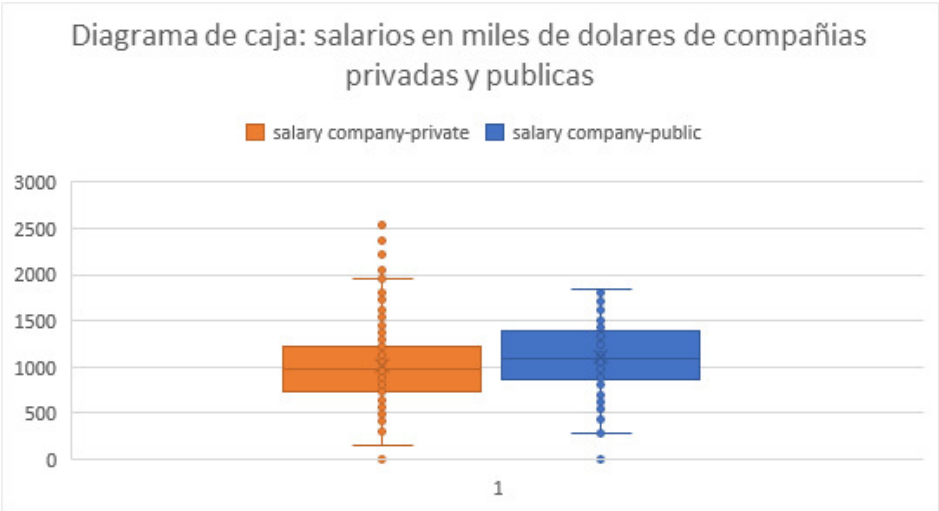
Decisión:
Se rechaza H_0 (hipotesis nula) al ser el valor critico es menor que el valor del estadistico de prueba.

Conclusión:
Con un NS de 5% en informacion de las muestras, no existe evidencia suficiente para afirmar que los salarios de los data science son mayores que los de data engineer. Se podria afirmar que este salario tambien depende del tipo de empresa en donde se trabaje, ubicacion de la mepresa en que se trabaja, al sector en que se necesite uno de estos tipos de profesionales, etc.

Comparacion de salarios publico y privado

$$X_1 : \text{Salario publico} \rightarrow \overline{X_1} = 1110{,}3886, {\sigma_1}^2 = ?$$

$$X_2 : \text{Salario privado} \rightarrow \overline{X_p rivado} = 1020{,}8170, {\sigma_2}^2 = ?$$



Test de hipotesis

Luego:

$H_0: \mu_1 \leq \mu_2 \rightarrow \mu_1 - \mu_2 \leq 0$

$H_1: \mu_1 > \mu_2 \rightarrow \mu_1 - \mu_2 > 0$

Estadisitico de prueba es:

$$t = \frac{\overline{X_1} - \overline{X_2} - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \sim t_{(v)}$$

$$v = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{\left(\frac{S_1}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{S_2}{n_2}\right)^2}{n_2 - 1}}$$

Con un nivel de significancia de: $\alpha = 0{,}05$

Reemplazamos datos:

t = 2,93398952326531

v = 436

Region crítica:

$$t_{(436;\, 0{,}95)} = 1{,}64835599316749$$

Decisión:

Se rechaza H_0 (hipotesis nula) al ser el valor critico menor que el valor del estadistico de prueba.

Conclusión:

Con un NS de 5% en informacion de las muestras, no existe evidencia suficiente para afirmar que los salarios en compañias publicas son mayores que en las privadas. Se podria afirmar que este salario de acuerdo al tipo de empresa , depende a la ubicacion de la empresa, al sector que pertenezca, etc.

Distribucion de ingresos segun edad

H0	P1=P2
H1	P1<P2
RC	Z<-1.645
P1	0.161470865
P2	0.3536413
Pc	0.240809557
Qc=1-Pc	0.759190443
ES	0.004812733
Z	-39.92958611

El Z está en la regió critica, por lo tanto H0 se rechaza. Las personas de mediana edad poseen una mejor distribución de ingreso que las personas jóvenes

Distribucion de proporciones ingresos segun edad

H0	Pper=Pmex
H1	Pper<Pmex
RC	Z<-1.645
Pper	0.064516129
Pmex	0.051321928
Pc	0.051928783
Qc=1-Pc	0.948071217
ES	0.040800752
Z	0.323381307

El Z no está en la región critica, por lo tanto H0 se acepta al nivel 5% y concluir que son iguales las proporciones de salarios mayores a 2500

Comprobar que las personas que trabajan una cantidad mayor que la media perciben mejores ingresos

hours-per-week >50k	
Media	45.4036333
Error típico	0.18992852
Mediana	40
Moda	40
Desviación estándar(s1)	11.2731815
Varianza de la muestra	127.084621
Curtosis	4.29634675
Coeficiente de asimetría	0.64024536
Rango	98
Mínimo	1
Máximo	99
Suma	159957
Cuenta(n1)	3523
Nivel de confianza(95.0%)	0.37238102

hours-per-week (<=50k)	
Media	38.8528593
Error típico	0.11901377
Mediana	40
Moda	40
Desviación estándar (s2)	12.5608872
Varianza de la muestra	157.775888
Curtosis	2.91139626
Coeficiente de asimetría	0.24283729
Rango	98
Mínimo	1
Máximo	99
Suma	432782
Cuenta(n2)	11139
Nivel de confianza(95.0%)	0.23328806

Para muestras grandes:

$$ET = \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

$H_0 : \mu_1 = \mu_2$
$H_1 : \mu_2 > \mu_2$

$$Z = \frac{\overline{X_1} - \overline{X_2}}{ET}$$