Aplicacion de test no parametricos

Caso: Distribucion de notas de estudiantes de la acadmia Trilce en los ultimos 3 años

Alvarez Bautista Burga Casanova Cuyate

Facultad de Ingenieria Industrial y de Sistemas
Universidad Nacional de Ingenieria

Diciembre 2022

Tabla de Contenido

Tabla de Contenido

Objetivos del trabajo

General

Conocer si el ciclo de repaso util para los postulantes a la Universidad Nacional de Ingenieria

Objetivos especificos

- Probar si alguna de las medianas de la distribución de los últimos años para los simulacros escolares difiere con un nivel de significancia del 5 % con la prueba de Kruskal-Wallis.
- Comparar el desempeño de los estudiantes antes de tomar un ciclo de repaso y despues de este.
- Comparar la evolucion del desempeño de aquellos que tomaron mas de un simulacro.
- Comparar la potencia entre el test de *Wilcoxon* y el *Test de signos*
- 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

Tabla de Contenido

Test de Kruskal-Wallis

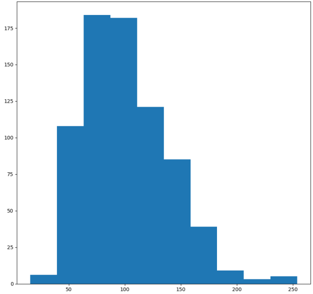
Test de Signos

Test de Wilcoxon

Si bien, el test de signos puede cumplir la misma funcion que el de **Wilcoxon**, este ultimo tiene mayor potencia al momento de detectar diferencia de medias.

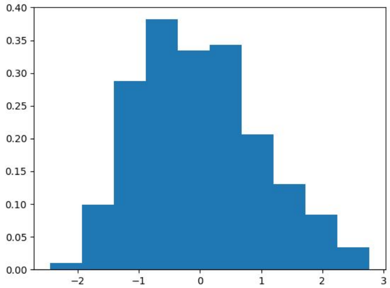
Tabla de Contenido

Figura: Data sin estandarizar



Hipotesis 1

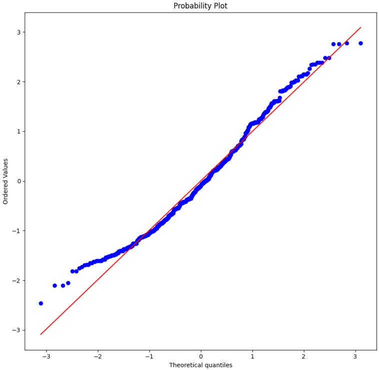
Figura: Data estandarizada y sin outliers



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

$$H_1 : \text{La distribucion de ingresos } \textbf{NO sigue} \text{ la ley normal}$$

$$H_0 : \text{La distribucion de ingresos } \textbf{SIGUE} \text{ la ley normal}$$

Conclusiones hipotesis 1

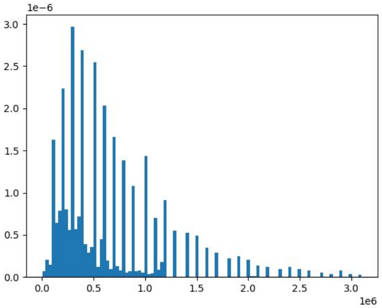
El test K-S y el de Jarque-Bera muestran los siguientes p-values.

```
Desarrollo de la primera hipótesis
Jarque_beraResult(statistic=24.54482110101632, pvalue=4.679072972302300e-06)
KstestResult(statistic=0.061898221291961375, pvalue=0.007064124010629260)
```


```
Desarrollo de la primera hipótesis
Jarque_beraResult(statistic=24.54482110101632, pvalue=4.679072972302300e-06)
KstestResult(statistic=0.061898221291961375, pvalue=0.007064124010629260)
```

Hipotesis 2

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



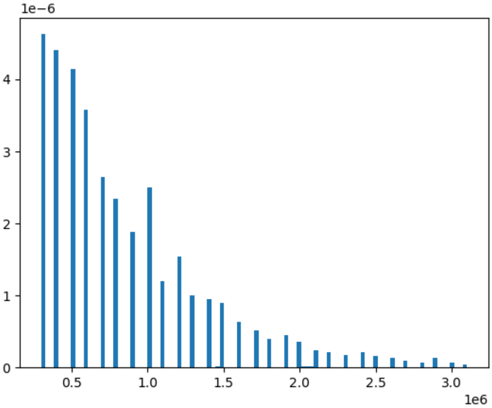
se puede notar como existen *2 grupos en la poblacion*

Distribución de los trabajadores en software

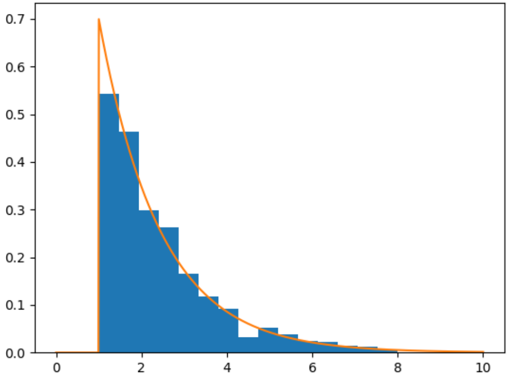
Aplicacion del test **Kolmogórov-Smirnov**

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

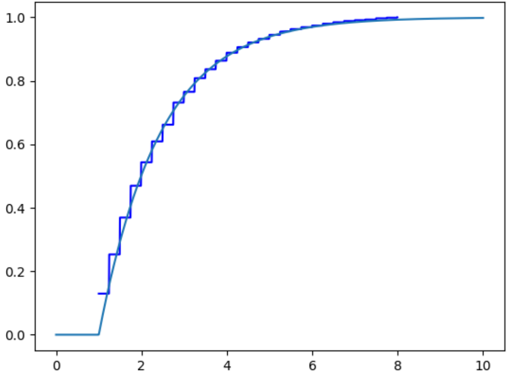
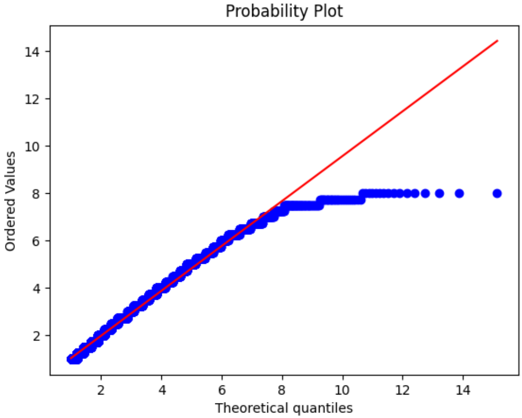


Grafico P-P


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

Hipotesis 3

Se aplicó el test de *Kruskal-Wallis* con la finalidad de:

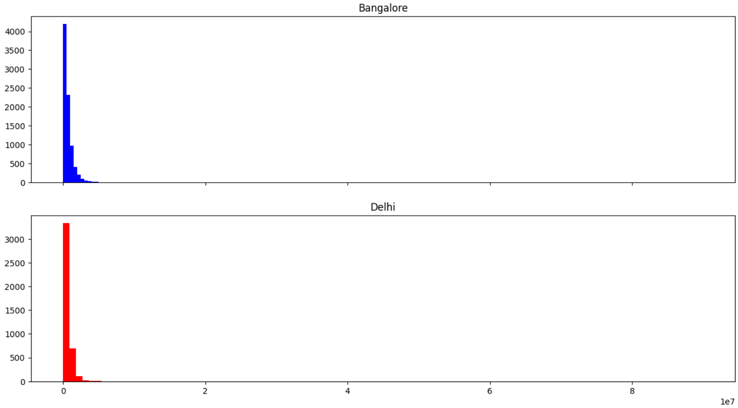
- Verificar si las muestras de Delhi y Bangalore provienen de poblacines distintas
- Comprobar si las 2 poblaciones difieren significativamente

Para esto se realizo un procedimiento similar al de la hipotesis anterior

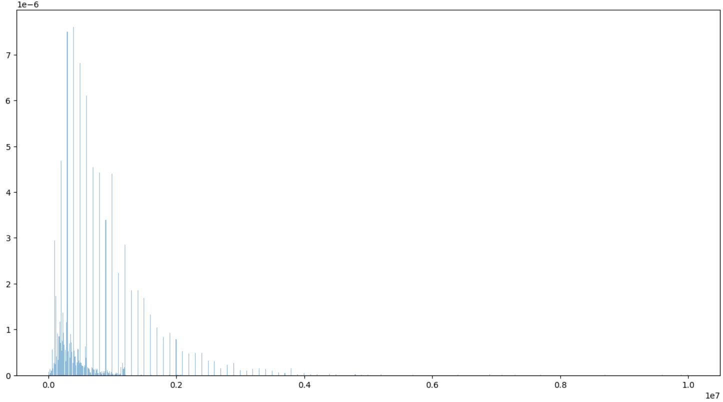
Consideracines del test

El test de Kruskal-Wallis es el sustituto no parametrico del *."One way ANOVA"*, en el cual se necesita un factor independiente

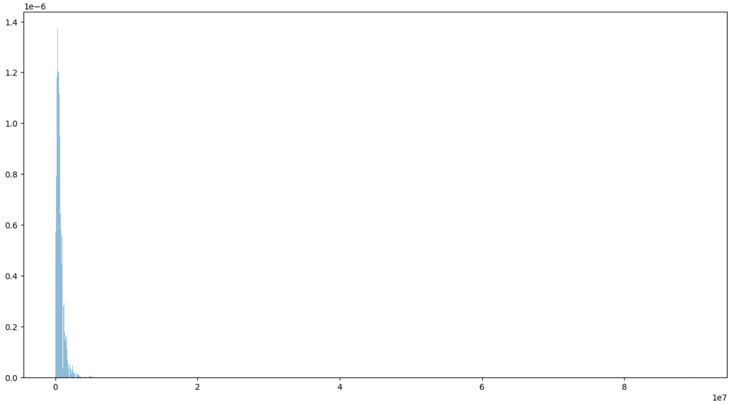
Graficas de ambas ciudades



Graficas de Bangalore sin filtrar



Graficas de Delhi sin filtrar



Resultados test

```
Kuskal Wallis Test
KruskalResult(statistic=41.02997485512918, pvalue=1.4991231167533473e-10)
```

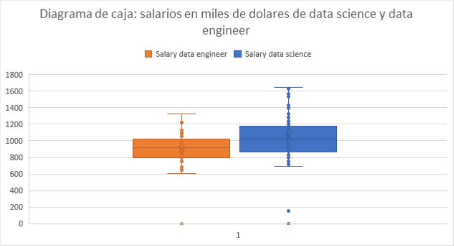
Conclusiones

1. Aparecen 2 grupos en la poblacion de Bangalore
2. Las distribuciones siguen la ley exponencial, solo varia su parametro de escalamiento

Comparacion de salarios DC y DI

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

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



Test de hipotesis

como se puede observar, las alturas de ambos diagras difieren significativamente, por lo que se consideran poblaciones con varianza diferentes

Estadistico de prueba:

$$t = \frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \sim t_{(\upsilon)}$$

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

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

Conclusiones

reemplazando datos:
$t = 4{,}7424$
$\upsilon = 149$
Region critica:
$t_{(149,0,95)} = 1{,}655144$
Se rechaza la hipotesis nula al ser el valor critico menor que el estadistico de prueba
Conclusion

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.

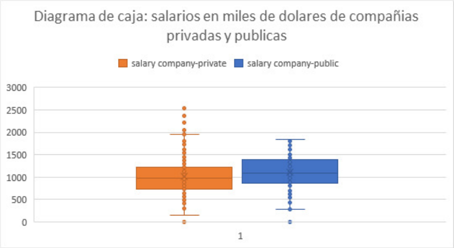
Conclusión:

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

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

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



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				*hours-per-week (<=50k)*	
Media	45.4036333			Media	38.8528593
Error típico	0.18992852			Error típico	0.11901377
Mediana	40			Mediana	40
Moda	40			Moda	40
Desviación estándar(s1)	11.2731815			Desviación estándar (s2)	12.5608872
Varianza de la muestra	127.084621			Varianza de la muestra	157.775888
Curtosis	4.29634675			Curtosis	2.91139626
Coeficiente de asimetría	0.64024536			Coeficiente de asimetría	0.24283729
Rango	98			Rango	98
Mínimo	1			Mínimo	1
Máximo	99			Máximo	99
Suma	159957			Suma	432782
Cuenta(n1)	3523			Cuenta(n2)	11139
Nivel de confianza(95.0%)	0.37238102			Nivel de confianza(95.0%)	0.23328806

Para muestras grandes:

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

$$H_0 : \mu_1 = \mu_2$$

$$H_1 : \mu_1 > \mu_2$$

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

Dado un error maximo permitido $\alpha = 0{,}05$, con H_1 indicando una cola unilateral hacia la derecha $\rightarrow RC = Z > 1{,}645$

$$Z_{cal} = \frac{45{,}4036}{38{,}8528} = 29{,}2267$$

Como $Z_{cal} \in RC$ Se debe rechazar H_0 y concluir que aquellas personas que ganan mas de 50K trabaja en promedio mas que las personas que ganan menos de 50K

Comparar proporciones de trabajadores segun sexo

Sean p_1 y p_2 la proporcion de trabajadores femeninos en *Estados Unidos* y *Mexico* respectivamente
Siendo $n_1 = 14662,\ x_1 = 4927$ y $n_2 = 308,\ x_2 = 69$ las cantidades totales y de poblacion femenina en ambos paises
Dando como resultados:

$$\hat{p} = \frac{4927 + 69}{14662 + 308} = 0{,}3245$$

$$\text{Error tipico: } \overline{p_1} - \overline{p_2}$$

$$ET = \sqrt{\frac{\hat{p}(1-\hat{p})}{n_1} + \frac{\hat{p}(1-\hat{p})}{n_2}} = 0{,}02695$$

Por **TCL**:

$$H_0 : p_1 = p_2$$

$$H_1 : p_1 > p_2$$

$$Z = \frac{\overline{p}_1 - \overline{p}_2}{ET} \sim N(0,1)$$

Dado un error maximo permitido $\alpha = 0{,}05$, con H_1 indicando una cola unilateral hacia la derecha $\rightarrow RC = Z > 1{,}645$

$$Z_{cal} = \frac{p_1 - p_2}{ET} = 4{,}15530$$

Como $Z_{cal} \in RC$ Se debe rechazar H_0 y concluir que existe una mayor proporcion de mujeres trabajando en ciencia de datos en Estados Unidos que en Mexico; sin embargo, ese analisis no puede ser tan confiable debido a la diferencia del tamaño de las muestras

Otras conclusiones

- Se logra visualizar la formacion de 2 grupos en la poblacion de Bangalore independientemente de la variable de analisis
- La distribucion teoria a la que mejor se aproximan los ingresos es la exponencial, la cual deriva de la distribución *Gamma*
- Los test no parametricos son muy suceptibles a :
  1. outliers
  2. Distribuciones con ligeras desviaciones de las teoricas
  3. Gran cantidad de data