**Revisión Estadística Descriptiva**

La base de datos Censo\_Universitario.Rdata contiene las siguientes variables medidas en 227540 alumnos universitarios

|  |  |
| --- | --- |
| Variable | Descripción |
| universidad | Universidad |
| facultad | Facultad |
| especialidad | Especialidad |
| sexo | Sexo |
|  | Hombre |
|  | Mujer |
| pre | Tipo de preparación para la universidad |
|  | Por su cuenta |
|  | Profesor particular |
|  | Academia |
|  | Centro Pre-Universitario |
| modalidad | Modalidad de ingreso a la universidad |
|  | Examen de admisión |
|  | Primeros puestos |
|  | Convenio universitario |
|  | Centro Pre-universitario |
|  | Traslado externo |
|  | Graduado o titulado |
|  | Disposiciones especiales |
|  | Otra |
| trabajo | Trabaja actualmente? |
|  | Si |
|  | No |
| tipo | Tipo de universidad |
|  | Pública |
|  | Privada |

Lea este conjunto de datos mediante la función load(),

**load**("Censo\_Universitario.Rdata")

esto generará un objeto d que contendrá los datos.

## Pregunta 1

Se desea comparar la distribución de frecuencias relativas del tipo de preparación para la universidad (pre) entre alumnos de universidades públicas (tipo=="Pública") y privadas (tipo=="Privada"). Con este fin construya la tabla de distribución de frecuencias para cada grupo y represéntela gráficamente. Presente un análisis comparativo de sus resultados.

## Solución

Al ser el tipo de preparación para la universidad una variable cualitativa, obtendremos primero una tabla de frecuencias para los alumnos de universidades públicas y privadas.

Cuadro 1: Distribución de frecuencias del tipo de preparación para la universidad en alumnos de universidades públicas

Tipo de

Número

Proporción

Número acumulado

Proporción acumulada

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| preparación de alumnos de alumnos | | | de alumnos | de alumnos |
| Por su cuenta | 10570 | 0.131 | 10570 | 0.131 |
| Profesor particular | 381 | 0.005 | 10951 | 0.136 |
| Academia | 44742 | 0.555 | 55693 | 0.690 |
| Centro Pre-Universitario | 24979 | 0.310 | 80672 | 1.000 |

Cuadro 2: Distribución de frecuencias del tipo de preparación para la universidad en alumnos de universidades privadas

Tipo de

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| preparación de alumnos de alumnos | | | de alumnos | de alumnos |
| Por su cuenta | 79428 | 0.541 | 79428 | 0.541 |
| Profesor particular | 3708 | 0.025 | 83136 | 0.566 |
| Academia | 40607 | 0.276 | 123743 | 0.843 |
| Centro Pre-Universitario | 23125 | 0.157 | 146868 | 1.000 |

Número

Proporción

Número acumulado

Proporción acumulada

Un gráfico apropiado para representar la frecuencia del tipo de preparación para la universidad es un gráfico de barras, el cuál presentamos a continuación

**Distribución de frecuencias del tipo de preparación para la universidad en alumnos de universidades públicas**

0.555

0.5

0.6

0.31

Frecuencia relativa

0.3

0.4

0.131

0.1

0.2

0.005

0.0

Por su cuenta Profesor particular Academia Centro Pre−Universitario

Tipo de preparación

**Distribución de frecuencias del tipo de preparación para la universidad en alumnos de universidades privadas**

0.541

0.5

0.6

0.276

Frecuencia relativa

0.3

0.4

0.157

0.2

0.025

0.0

0.1

Por su cuenta Profesor particular Academia Centro Pre−Universitario

Tipo de preparación

Podemos observar que el tipo de preparación más común para una universidad pública es por una academia, con una proporción del 55.5 %; mientras que para una universidad privada es por cuenta propia, con una proporción del 54.1 %.

También, podemos notar que el 31.0 % de lo alumnos de universidades públicas se preparo para la universidad en un centro pre-universitario, mientras que esta proporción es del 15.7 % para alumnos de universidades privadas.

Los códigos en R usados para obtener las tablas de frecuencia y los gráficos se presentan a continuación.

*# Tabla de frecuencias* **library**(DescTools) **Freq**(d**$**pre[d**$**tipo**==**"Pública"])

## level freq perc cumfreq cumperc

|  |  |  |  |
| --- | --- | --- | --- |
| ## 1 | Por su cuenta 10’570 13.1% | 10’570 | 13.1% |
| ## 2 | Profesor particular 381 0.5% | 10’951 | 13.6% |
| ## 3 | Academia 44’742 55.5% | 55’693 | 69.0% |
| ## 4 | Centro Pre-Universitario 24’979 31.0% | 80’672 | 100.0% |

**Freq**(d**$**pre[d**$**tipo**==**"Privada"])

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ## | level | freq | perc cumfreq | cumperc |
| ## 1 | Por su cuenta | 79’428 | 54.1% 79’428 | 54.1% |
| ## 2 | Profesor particular | 3’708 | 2.5% 83’136 | 56.6% |
| ## 3 | Academia | 40’607 | 27.6% 123’743 | 84.3% |
| ## 4 | Centro Pre-Universitario | 23’125 | 15.7% 146’868 | 100.0% |

alumnos de u

*# Gráficos de barras*

f.1 = **prop.table**(**table**(d**$**pre[d**$**tipo**==**"Pública"]))

f.2 = **prop.table**(**table**(d**$**pre[d**$**tipo**==**"Privada"]))

**par**(mfrow=**c**(2,1)) **barplot**(f.1,

main="Distribución de frecuencias del tipo de preparación para la universidad **\n**en ylab="Frecuencia relativa",

xlab="Tipo de preparación", ylim=**c**(0,1.1**\*max**(**c**(f.1,f.2))))

pos.j = **barplot**(f.2,plot = FALSE)

**text**(pos.j,f.1,

labels=**round**(f.1,3), pos=3)

**barplot**(f.2,

main="Distribución de frecuencias del tipo de preparación para la universidad **\n**en ylab="Frecuencia relativa",

xlab="Tipo de preparación", ylim=**c**(0,1.1**\*max**(**c**(f.1,f.2))))

pos.j = **barplot**(f.2,plot = FALSE)

**text**(pos.j,f.2,

labels=**round**(f.2,3), pos=3)

alumnos de u

## 

## Pregunta 2

Elabore un diagrama de cajas (boxplot) para comparar el tiempo de desplazamiento de casa a la universidad, entre los alumnos de universidades públicas y privadas. A partir de sus resultados, realice una comparación

en términos de tendencia central, dispersión y asimetría.

## Solución

Presentamos un diagrama de cajas del tiempo de desplazamiento de casa a la universidad, entre los alumnos de universidades públicas y privadas. .

# Diagrama de cajas del tiempo de desplazamiento de casa a la universi

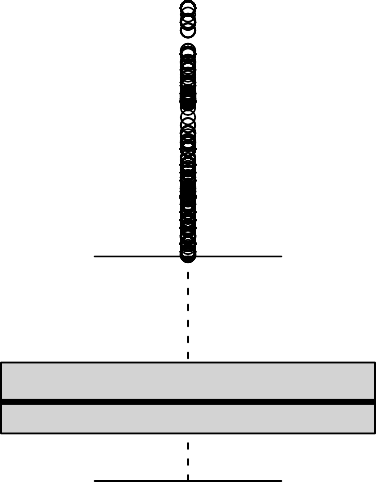
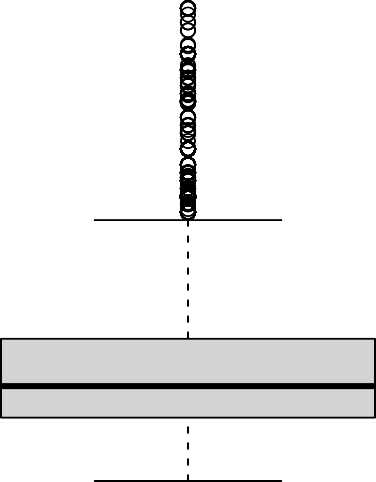
**por tipo de universidad**

Tiempo en minutos

200

300

### Pública Privada



0

50

100

Tipo de universidad

Tendencia central: En promedio, considerando la mediana, el mayor tiempo de desplazamiento de la casa a la universidad se dá para los alumnos de una universidad pública.

Dispersión: Considerando el rango intercuartil, se observa que similar entre alumnos de universidades públicas y privadas, siendo ligeramente mayor para las universidades públicas.

Asimetría: Se observa una asimetría positiva alta en el tiempo desplazamiento de los alumnos, tanto para universidades públicas como privadas.

Código en R:

**boxplot**(tiempo**~**tipo,data=d,

main="Diagrama de cajas del tiempo de desplazamiento de casa a la universidad por tipo de universidad",

ylab="Tiempo en minutos", xlab="Tipo de universidad")

## Pregunta 3

Para un nuevo estudio se van a considerar solamente la información de mujeres universitarias cuyas horas de trabajo por semana hayan estado dentro del 25 % más altas de su grupo, y de hombres universitarios cuyas horas de trabajo por semana hayan estado dentro del 20 % más altas de su grupo.

¿Qué número de horas de trabajo por semana debe superar como mínimo una mujer para que sea incluida en el estudio? ¿Qué número de horas de trabajo por semana debe superar como mínimo un hombre para que sea incluida en el estudio? Justifique sus respuestas.

## Solución

Para que una mujer universitaria participe en el estudio, su número de horas de trabajo por semana debe ser como mínimo el cuantil 0.75 que es igual a 48 horas.

Para que un hombre universitario participe en el estudio, su número de horas de trabajo por semana debe ser como mínimo el cuantil 0.80 que es igual a 48 horas.

Cálculo en R:

**quantile**(d**$**horas[d**$**sexo**==**"Mujer"],0.75,na.rm = TRUE)

## 75%

## 48

**quantile**(d**$**horas[d**$**sexo**==**"Hombre"],0.80,na.rm = TRUE)

## 80%

## 48

## Pregunta 4

Para la variable ingreso monetario mensual del hogar considere que

Hasta 500 corresponde a un ingreso en el intervalo [0*,* 500].

Más de 20000 corresponde a un ingreso en el intervalo (20000*,* 25000].

En los otros casos, asuma que es un intervalo abierto a la izquierda y cerrado a la derecha, por ejemplo, para De 501 a 1000 corresponde a un ingreso en el intervalo (500*,* 1000].

1. Construya un polígono de frecuencias para la distribución de los ingresos mensuales de los hogares de alumnos de universidades públicas y otro sobre el mismo eje *x* para los alumnos de universidades privadas. Interprete sus resultados. **Sugerencia**: en R para que un segundo gráfico se muestre sobre un primer gráfico elaborado previamente: utilice la función plot para crear el primer gráfico y luego use la función lines para crear el segundo.
2. Calcule las estadísticas necesarias y realice una comparación en términos de tendencia central, dispersión y asimetría del ingreso mensual de los hogares de alumnos por el tipo de universidad a la que asisten.
3. Algunos autores han propuesto la siguiente medida de asimetría robusta ante valores atípicos para una distribución

*A*0 = *q*0*,*875 *−* 2*q*0*,*5 + *q*0*,*125 *,*

*q*0*,*875 *− q*0*,*125

donde *qp* denota el cuantíl 0 *< p <* 1 de la distribución. Calcule e interprete esta medida para la variable ingreso monetario mensual del hogar para para los alumnos de universidades públicas y privadas.

## Solución parte a)

En este caso obtenemos la siguiente tabla de frecuencias para la variable edad considerando las condiciones de la pregunta.

Cuadro 3: Distribución de frecuencias del ingreso monetario mensual del hogar de alumnos de universidades públicas

Tipo de

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| preparación de alumnos de alumnos | | | de alumnos | de alumnos |
| [0,500] | 13523 | 0.168 | 13523 | 0.168 |
| (500,1000] | 28006 | 0.347 | 41529 | 0.515 |
| (1000,1500] | 18500 | 0.229 | 60029 | 0.744 |
| (1500,2000] | 9843 | 0.122 | 69872 | 0.866 |
| (2000,3000] | 6066 | 0.075 | 75938 | 0.941 |
| (3000,4000] | 2373 | 0.029 | 78311 | 0.971 |
| (4000,5000] | 1227 | 0.015 | 79538 | 0.986 |
| (5000,10000] | 784 | 0.010 | 80322 | 0.996 |
| (10000,15000] | 199 | 0.002 | 80521 | 0.998 |
| (15000,20000] | 89 | 0.001 | 80610 | 0.999 |
| (20000,25000] | 62 | 0.001 | 80672 | 1.000 |

Número

Proporción

Número acumulado

Proporción acumulada

Cuadro 4: Distribución de frecuencias del ingreso monetario mensual del hogar de alumnos de universidades privadas

Tipo de

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| preparación de alumnos de alumnos | | | de alumnos | de alumnos |
| [0,500] | 8527 | 0.058 | 8527 | 0.058 |
| (500,1000] | 26901 | 0.183 | 35428 | 0.241 |
| (1000,1500] | 27093 | 0.184 | 62521 | 0.426 |
| (1500,2000] | 22695 | 0.155 | 85216 | 0.580 |
| (2000,3000] | 20795 | 0.142 | 106011 | 0.722 |
| (3000,4000] | 13171 | 0.090 | 119182 | 0.811 |
| (4000,5000] | 10486 | 0.071 | 129668 | 0.883 |
| (5000,10000] | 10715 | 0.073 | 140383 | 0.956 |
| (10000,15000] | 3539 | 0.024 | 143922 | 0.980 |
| (15000,20000] | 1633 | 0.011 | 145555 | 0.991 |
| (20000,25000] | 1313 | 0.009 | 146868 | 1.000 |

Número

Proporción

Número acumulado

Proporción acumulada

**Gráfico de polígono de frecuencias del ingreso monetario mensual de los hogares de los alumnos por tipo de universidad**

Frecuencia relativa

0.00 0.05 0.10 0.15 0.20 0.25 0.30 0.35

Pública Privada

0 5000 10000 15000 20000 25000

Ingreso monetario mensual del hogar (soles)

Código en R

*# Tablas de frecuencias* **Freq**(d**$**ingreso[d**$**tipo**==**"Pública"]) **Freq**(d**$**ingreso[d**$**tipo**==**"Privada"]) *# Marcas de clase*

M = **c**(250,750,1250,1750,2500,3500,4500,7500,12500,17500,22500)

por tipo de u

*# Frecuencias relativas* f1=**prop.table**(**table**(d**$**ingreso[d**$**tipo**==**"Pública"])) f2=**prop.table**(**table**(d**$**ingreso[d**$**tipo**==**"Privada"])) *# Abcisas para el gráfico*

x.pol = **c**(0,M,25000)

*# Ordenadas para el gráfico*

y.pol\_1 = **c**(0,f1,0) y.pol\_2 = **c**(0,f2,0) *# Gráfico*

**plot**(x.pol,y.pol\_1,type="b",col=2,pch=16,lwd=2, xlab="Ingreso monetario mensual del hogar (soles)", ylab="Frecuencia relativa",

main="Gráfico de ojiva del ingreso monetario mensual de los hogares**\n** de los alumnos **lines**(x.pol,y.pol\_2,type="b",col=4,pch=16,lwd=2) **legend**("topright",pch=15,col=**c**(2,4),legend=**c**("Pública","Privada"))

Se observa una clara asimetría positiva para ambas distribuciones, siendo más acentuada para el caso de universidades públicas.

También, notamos que es más frecuente la ocurrencia de ingresos altos para los hogares de alumnos de universidades privadas que de públicas.

## Solución parte b)

Como se observa una distribución asimétrica para ambos tipos de universidad, para las comparaciones en términos de tendencia central, dispersión y asimetría consideramos la mediana, el rango intercuartil y el coeficiente de asimetría de Fisher.

Cuadro 5: Estadísticas descriptivas del ingreso por tipo de universidad

Tipo de universidad

|  |  |  |  |
| --- | --- | --- | --- |
| Estadística | Pública | Privada |  |
| Mediana | 978.70 | 1740.43 |  |
| Rango intercuartil | 905.49 | 2290.54 |  |
| Coeficiente de asimetría de Fisher | 6.35 | 3.15 |  |

Tendencia central: En promedio, considerando la mediana, el mayor ingreso del hogar se presenta en los alumnos de universidad privada con 1740.43 soles. Mientras que en el caso de los hogares de los alumnos de universidades públicas es de 978.70 soles.

Dispersión: Considerando el rango intercuartil, se observa que el ingreso de los hogares de los alumnos de universidades privadas tiene una mayor dispersión con un RIC de 2290.54 soles mientras que el caso de la universidades públicas el RIC es de 905.49.

Asimetría: Se observa una asimetría positiva para ambos tipos de universidad, siendo la más alta para las universidades públicas.

Código en R

*# Limites de los intervalos*

L = **c**(0,500,1000,1500,2000,3000,4000,5000,10000,15000,20000,25000)

*# Frecuencias acumuladas*

*# (adicionamos 0 para el primer limite inferior)* F1=**c**(0,**cumsum**(**prop.table**(**table**(d**$**ingreso[d**$**tipo**==**"Pública"])))) F2=**c**(0,**cumsum**(**prop.table**(**table**(d**$**ingreso[d**$**tipo**==**"Privada"])))) *# Mediana*

**approx**(F1,L,0.5)**$**y **approx**(F2,L,0.5)**$**y *# RIC*

**approx**(F1,L,0.75)**$**y**-approx**(F1,L,0.25)**$**y **approx**(F2,L,0.75)**$**y**-approx**(F2,L,0.25)**$**y *# Marcas de clase*

M = **c**(250,750,1250,1750,2500,3500,4500,7500,12500,17500,22500)

*# Frecuencias* f1=**table**(d**$**ingreso[d**$**tipo**==**"Pública"]) f2=**table**(d**$**ingreso[d**$**tipo**==**"Privada"])

*# Repetir la marca de clase tantas veces como la frecuencia*

x1 = **rep**(M,times=f1) x2 = **rep**(M,times=f2)

*# Coeficiente de asimetría de Fisher*

**Skew**(x1) **Skew**(x2)

## Solución parte c)

Observamos, que por la definición del coeficiente, se obtendrá *A*0 = 0 en caso se tenga una distribución simé- trica, *A*0 *<* 0 para distribuciones asimétricas negativos y *A*0 *>* 0 para distribuciones asimétricas positivas.

Para los hogares de alumnos de universidades públicas se obtiene que *A*0 = 0*,*3056857 y para universidades privadas que *A*0 = 0*,*4971492. Por lo tanto, bajo esta medidas se tiene una mayor asimetría en los ingresos de los hogares de alumnos de universidades privadas.

Las diferencias con los resultados obtenidos con la asimetría de Fisher se deben a la presencia de valores atípicos en la distribución de frecuencias de universidades públicas (pocos hogares tienen ingresos altos).

Código en R:

*# Limites de los intervalos*

L = **c**(0,500,1000,1500,2000,3000,4000,5000,10000,15000,20000,25000)

*# Frecuencias acumuladas*

*# (adicionamos 0 para el primer limite inferior)* F1=**c**(0,**cumsum**(**prop.table**(**table**(d**$**ingreso[d**$**tipo**==**"Pública"])))) F2=**c**(0,**cumsum**(**prop.table**(**table**(d**$**ingreso[d**$**tipo**==**"Privada"])))) *# A\_0 Universidades públicas*

q\_0.875=**approx**(F1,L,0.875)**$**y q\_0.125=**approx**(F1,L,0.125)**$**y q\_0.500=**approx**(F1,L,0.500)**$**y

(q\_0.875-2**\***q\_0.500**+**q\_0.125)**/**(q\_0.875**-**q\_0.125)

*# A\_0 Universidades públicas* q\_0.875=**approx**(F2,L,0.875)**$**y q\_0.125=**approx**(F2,L,0.125)**$**y q\_0.500=**approx**(F2,L,0.500)**$**y

(q\_0.875-2**\***q\_0.500**+**q\_0.125)**/**(q\_0.875**-**q\_0.125)

## Pregunta 5

Evalúe la veracidad o falsedad de cada una de las siguientes afirmaciones. Justifique su respuesta (presente sus códigos y resultados en R como parte de su justificación)

1. La variable tiempo en minutos de desplazamiento de casa a la universidad presenta una menor dispersión que el gasto anual en matrícula.
2. La mediana del tipo de preparación para la universidad es Academia.
3. El tercer cuartil del número de horas de trabajo a la semana es la misma para hombres y mujeres.
4. Se aplicó la función summary a la variable número de horas de trabajo a la semana y a

partir de los resultados obtenidos, se llegó a la conclusión que aproximadamente el 50 % de todos los alumnos trabaja más de 36 horas.

1. Para representar la tendencia central de la variable gasto anual en matrícula es adecuado usar la mediana.

## Solución parte a)

Para poder comparar la dispersión en variable medidas en diferentes unidades debemos usar el coeficiente de variabilidad

**CoefVar**(d**$**tiempo)

## [1] 0.6104214

**CoefVar**(d**$**gasto) ## [1] 5.595054

Conclusión: Verdadero. Se obtiene una menor dispersión con el tiempo. Observación: Si usan otra medida de dispersión, se considera puntaje 0.

## Solución parte b)

La variable preparación para la universidad es cualitativa nominal, por lo tanto, no se puede calcular la mediana en este caso.

Conclusión: Falso.

Observación: No es necesario realizar cálculo para contestar esta pregunta.

## Solución parte c)

Calculamos el tercer cuartil para cada grupo

**summaryFull**(horas**~**sexo,d)

## Hombre Mujer

## N 4.387e+04 3.414e+04

## NA’s 7.524e+04 7.429e+04

## N.Total 1.191e+05 1.084e+05

## Mean 3.460e+01 3.303e+01

## Median 4.000e+01 3.600e+01

## 10% Trimmed Mean 3.462e+01 3.326e+01

## Geometric Mean 2.867e+01 2.714e+01

## Skew -5.805e-03 -8.965e-02

## Kurtosis -3.842e-01 -5.419e-01

## Min 1.000e+00 2.000e+00

## Max 9.900e+01 9.900e+01

## Range 9.800e+01 9.700e+01

## 1st Quartile 2.000e+01 2.000e+01

## 3rd Quartile 4.800e+01 4.800e+01

## Standard Deviation 1.712e+01 1.655e+01 ## Geometric Standard Deviation 2.023e+00 2.055e+00 ## Interquartile Range 2.800e+01 2.800e+01 ## Median Absolute Deviation 1.483e+01 1.779e+01 ## Coefficient of Variation 4.947e-01 5.010e-01 ## attr(,"class")

## [1] "summaryStats"

## attr(,"stats.in.rows") ## [1] TRUE

## attr(,"drop0trailing") ## [1] TRUE

Se observa que el tercer cuartil es 48 horas tanto para hombre como para mujeres. Conclusión: Verdadero.

## Solución parte d)

**summary**(d**$**horas)

Soles

500

1000

|  |  |  |  |
| --- | --- | --- | --- |
| ## | Min. 1st Qu. Median | Mean 3rd Qu. | Max. NA’s |
| ## | 1.00 20.00 36.00 | 33.91 48.00 | 99.00 149530 |

Conclusión: Falso. Si bien la mediana es 36, esta fue calculada solamente sobre los alumnos que trabajan, por lo tanto, la afirmació es falsa.

Observación: Considere como máximo 0.50 puntos si no toma en cuenta que la mediana se ha calculado solamente sobre los alumnos que trabajan.

## Solución parte e)

Para el gasto anual en matrícula se obtiene el coeficiente de asimetría de Fisher y el diagrama de cajas

**Skew**(d**$**gasto)

## [1] 10.77741

**boxplot**(d**$**gasto,main="Diagrama de cajas del gasto anual en matrícula", ylab="Soles")

# Diagrama de cajas del gasto anual en matrícula



1500

2000

Conclusión: Verdadero. Se observa una asimetría positiva tanto en el coeficiente de asimetría de Fisher como en el diagrama de cajas.

0