****

**Universidad de Chile**

Facultad de Ciencias Sociales

Profesora: Catalina Canals

Ayudantes: Daniel Cifuentes, Camila Lizama,

Daniela Olivares, Sofía Villalobos

**Pauta Examen**

 **Estadística II, sección 2**

**Ejercicios**

1. **Recomendación de test a utilizar (Total: 1,2 pts. 0,6 pts. c/u)**

**En los siguientes casos: (1) identifique la hipótesis de investigación, (2) indique si tiene toda la información necesaria sobre los datos para decidir qué test utilizar, (3.a) si tiene toda la información necesaria sobre los datos para decidir qué test utilizar: sugiera qué test de hipótesis utilizar y por qué, indique si se cumple cada uno de los supuestos del test sugerido (si no se cumplen todos los supuestos, explique por qué recomienda utilizarlo a pesar de ello), y plantee las hipótesis nula y alternativa del test a realizar, 3.b) si NO tiene toda la información necesaria sobre los datos para decidir qué test utilizar, indique qué información requeriría para saber qué test recomendar, sugiera cómo podría obtenerla, e indique qué test recomendaría –y cuáles serían la hipótesis nula y la alternativa-en cada caso posible.**

1. **Unos investigadores están interesados en estimar el promedio de horas semanales trabajadas en Chile, a través de la CASEN 2013, y analizar si el promedio/mediana de horas efectivas trabajadas supera el máximo legal (45 hrs. semanales). Se disponen de los siguientes datos de la muestra:**

|  |
| --- |
| **Datos Muestra (n=200.000)** |
| **Valor P en prueba Kolmogorov-Smirnov de normalidad del promedio de horas semanales trabajadas** | **0,02** |

* Hipótesis de investigación: el promedio/mediana de horas efectivas trabajadas en Chile supera el máximo legal (45 hrs. semanales). **(0,1 pts.)**
* Elección del test: Test Z para medias de 1 cola, dado que se cumplen los supuestos para ello, la variable es cuantitativa, y la hipótesis de investigación (que corresponde en este caso a la alternativa) refiere a que el parámetro es mayor que un valor. **(0,2 pts.).**
* Supuestos: muestra probabilística (se cumple porque la CASEN es probabilística), variable cuantitativa (se cumple dado que las hrs. Refiere a una variable cuantitativa) y n es grande. **(0,2 pts.)**
* H0: el promedio de hrs. de trabajo en Chile =45, H1: el promedio de hrs. De trabajo en Chile >45. **(0,1 pts.)**
1. **Unos investigadores quieren comparar si el promedio/mediana de horas semanales trabajadas en Chile es igual en hombres y mujeres, a través de una muestra probabilística de 50 casos.**
* Hipótesis de investigación: el promedio/mediana de horas semanales trabajadas en Chile es igual en hombres y mujer. **(0,1 pts.)**
* Elección del test:
	1. No contamos con suficiente información. Dado que n es pequeño, la muestra es probabilística y la variable es cuantitativa, si la variable distribuye normal en ambos grupos podríamos hacer un test T de dos colas para medias. Para ello, debiéramos realizar un test Shapiro-Wilks, si rechazamos H0 (normalidad) en al menos uno de los grupos, no parece recomendable usar el test T. En ese caso podemos hacer un test Kolmogorov Smirnov de igualdad de distribuciones, y, en caso de tener hombres y mujeres igual distribución de las horas de trabajo, podríamos hacer un test U de Mann Whitney. **(0,2 pts.).**
	2. Contamos con suficiente información. Dado que n es pequeño, la muestra es probabilística y la variable es cuantitativa, si la variable distribuye normal en ambos grupos podríamos hacer un test T de dos colas para medias. Sin embargo, como T es robusta al no cumplimiento de normalidad para n>15, podemos hacer este test aun cuando no se cumpla el supuesto. **(0,2 pts.).**
* Supuestos: Test T: muestra probabilística, variable cuantitativa (se cumplen por enunciado) y normalidad de la variable (ya se describió como habría que averiguar si se cumple). Test U de Mann Whitney: muestra probabilística, variable cuantitativa (se cumplen por enunciado) e igualdad de distribuciones (ya se describió como averiguar si se cumple) **(0,2 pts.)**
* H0: Test T: el promedio de hrs. de trabajo en Chile es igual en hombres y mujeres, H1: el promedio de hrs. De trabajo en Chile es distinto en hombres y mujeres.Test U: la mediana de hrs. de trabajo en Chile es igual en hombres y mujeres, H1: la mediana de hrs. De trabajo en Chile es distinto en hombres y mujeres **(0,1 pts.)**
1. **Verdadero (V) o Falso (F) (Total: 3,2 pts., 0,4 pts. c/u)**

**En los siguientes casos, indique si las afirmaciones son verdaderas o falsas. En caso de ser falsas indique porqué lo son, e indique qué sería lo correcto. En caso de ser verdaderas, explique por qué lo son, agregue información que complemente la frase y/o explique qué quiere decir la afirmación. Sugerencia: sea breve, con 4 líneas por respuesta suele ser suficiente.**

1. V. **Si con una muestra probabilística de estudiantes, se estima un intervalo [203,205] de Puntaje SIMCE promedio, con 95% de confianza, esto implica que, con 5% de significancia rechazamos la hipótesis nula de que el Puntaje SIMCE promedio es igual a 200.**

Suponiendo que n es suficientemente grande como para que hacer inferencia mediante intervalos de confianza (IC) sea adecuado, si el parámetro propuesto por la hipótesis nula (el puntaje SIMCE es igual a 200 en este caso) no está incluido en el IC con 95% de confianza, esto implica que con 5% se rechaza H0. Esto, debido a que un IC con 95% de confianza es equivalente a un test Z de dos colas con 5% de significancia.

1. F. **Si se obtiene un Valor P mayor que la significancia, en todo test de hipótesis, se rechaza la hipótesis nula.**

Si se obtiene un Valor P MENOR que la significancia, en todo test de hipótesis, se rechaza la hipótesis nula.

1. F**. El Valor P indica cuál es la probabilidad de obtener el estadístico obtenido por el test.**

El valor P indica cuál es la probabilidad de obtener el estadístico obtenido por el test o uno más contradictorio con H0, siendo H0 cierta.

1. F**. Al realizar un Test T para medias (en un caso donde se cumplen sus supuestos y a través de una muestra de 18 casos), para comparar el promedio de notas de las mujeres y los hombres que estudian en la universidad de Chile y con la hipótesis alternativa que plantea que ambos promedios son diferentes, si T=2,5 el Valor P=0,2.**

En este caso el estadístico distribuye T con 16 gl, de modo P[T>2,5]=0,01, y por ende el Valor P=0,02.

1. F**. En una muestra probabilística de 30 casos, el promedio en un índice de actitudes hacia el cuidado del medioambiente (donde valores mayores indican mayor tendencia al cuidado del medioambiente, siendo 0 el valor mínimo y 10 el máximo) para los habitantes de zonas urbanas de la muestra es 3, y para los habitantes de zonas rurales es 7. Cumpliéndose los supuestos del test T, al aplicar el test cuya hipótesis alternativa establece que el promedio del índice es distinto entre las personas de la población que viven en zonas rurales y urbanas, se obtiene un Valor P=0,312 en un test T para muestras independientes. Lo anterior presentaría evidencia concluyente de que el promedio en el índice es igual para la población rural y la urbana.**

El no rechazo de H0 puede deberse a que H0 es cierta, pero también puede deberse a que, al ser el tamaño muestra tan pequeño, aun cuando H0 es falsa, no tenemos evidencia suficiente como para refutarla**.**

1. F**. Al realizar un ANOVA de una vía, en un caso donde se cumplen todos los supuestos y donde la variable dependiente es un índice de violencia en la comuna, y la variable independiente es el partido político del alcalde de la comuna, se obtiene un Valor P asociado al estadístico F=0,3, y un eta cuadrado =0,01, esto quiere decir que el partido político del alcalde explica el 30% de la varianza del índice de violencia en la comuna, y que se rechaza la hipótesis nula de que el promedio del índice de violencia es igual en todas las comunas.**

Debiera decir: “esto quiere decir que el partido político del alcalde explica el 1% de la varianza del índice de violencia en la comuna (dado que eta cuadrado=0,01), y que, con 95% de confianza, NO se rechaza la hipótesis nula de que el promedio del índice de violencia es igual en todas las comunas (dado que el valor p asociado al estadístico F>0,05)”

1. V**. Al analizar la correlación entre la cantidad de hijos y los ingresos familiares, la correlación es de -0,17, y el Valor P del test de hipótesis asociado es 0,03. Esto indica que, con 5% de significancia, existe relación lineal entre ambas variables.**

Dado que el valor P 0,05, con 95% de confianza se rechaza H0 que establece que la correlación es =0, y por ende se concluye que hay relación lineal. Además, esta relación es negativa y de baja magnitud.

1. F**. Para probar si la media de tiempo de viaje entre el hogar y el lugar de trabajos de una muestra es mayor que 1, podríamos utilizar un test de hipótesis.**

Los test de hipótesis solo son pertinentes cuando se quiere hacer inferencia, es decir, a partir de la muestra, sacar conclusiones sobre los parámetros poblacionales. Para saber si el valor muestral es =1, bastaría ver cuál es el promedio de tiempo de viaje en la muestra.

**Parte 3: Regresión lineal simple (Total: 1,6 pts.)**

**Un grupo de investigadores realiza un estudio para predecir un índice de consumo de drogas (que va de 0 a 10) en las personas con 12-35 años en Chile, a partir de la variable edad, obteniendo los siguientes resultados.**

**Observaciones: 1000**

**R cuadrado: 0,11**

**Prob >F: 0,001**

|  |  |  |
| --- | --- | --- |
| **Variable** | **Coeficiente** | **P>|t|** |
| **Edad** | **-0.1** | **0.01** |
| **Constante** | **0.3** | **0.1** |

1. **Refiérase al ajuste del modelo (interprete el R cuadrado y el test asociado al estadístico F) (0,4 pts.)**

El R cuadrado indica que la edad explica el 11% de la varianza del índice de consumo de drogas **(0,2 pts),** mientras que el Valor P asociado al estadístico F=0,001, indica que, con 95% de confianza, se rechaza H0 que establece que el efecto de la edad en tal índice es nulo **(0,2 pts.).** Dado lo anterior el ajuste del modelo es bajo (no explica mucho de la varianza), pero tampoco es nulo, en tanto la edad afecta el índice.

1. **Interprete estadísticamente el coeficiente asociado a la variable edad estimado por su modelo. Interprete sociológicamente su resultado (1,2 pts.).**

Al aumentar en 1 año la edad, el índice de consumo de drogas disminuye en 0,1 pts. **(0,4 pts.)**

Explicar hipótesis pertinente de porqué con la edad el consumo disminuye **(0,8 pts.)**

1. **Bonus: ¿Por qué cree que su modelo no explica toda la varianza del índice de consumo de drogas? (0,3 pts.)**

Explicar que esto se debe a que la relación consumo de drogas y edad no es lineal (al aumentar la edad aumenta el consumo hacia los 20-25 años, y después disminuye) (0,3 pts.).

Explicar que hay otros factores que probablemente influyen más en el consumo de drogas, sugiriendo en qué sentido lo hacen (0,3 pts.).