Estadistica II
Tarea Examen Pruebas No Paramétricas

Profesor: Antonio Soriano Flores
Ayu. Yolanda Martínez Guerrero

28 de noviembre de 2019

Intrucciones: En cada uno de los ejercicios siguientes (1 al 8) utilice solo pruebas binomimales, se le pide lo siguiente:

- Identifique la prueba a realizar.
- Calcule el estadístico de prueba.
- Encuentre la region de rechazo si $\alpha = 0.06$
- Encuentre el p-value asociado y asumiendo un $\alpha = 0.06$ concluya adecuadamente

1. En las próximas elecciones existen solo 2 candidatos a la presidencia. Sea p las proporción de votos para el candidato 1 la cual se asume desconocida. Supongo que se observa una muestra aleatoria de votantes de tamaño 30 y se cuenta que 10 votarán por el candidato 1. Con esta información conteste lo siguiente:

   a) Contraste la hipótesis $H_0 : p \leq 0.5$ vs $H_1 : p > 0.5$

   b) Construya un intervalo al 95 % de confianza para p

   c) Con base en lo anterior, ¿Crees que se tenga evidencia para decir que el candidato 2 ganará la presidencia?

2. Veinte observaciones independientes sobre una variable aleatoria resultaron en los números

142	134	98	118	131
104	154	122	93	137
86	119	161	144	158
165	81	117	128	103

   Con los datos anteriores realice los siguientes ejercicios:

   a) Pruebe la hipótesis de que la mediana es 103 (prueba de dos colas)

b) Pruebe la hipótesis de que el cuartil superior es al menos 150.

c) Pruebe la hipótesis de que el tercer decil no es mayor que 100.

d) Encuentre un intervalo de confianza "aproximado" del 90 % para la mediana.

e) Encuentre un intervalo de confianza "aproximado" del 95 % para el primer decil.

3. Un fabricante de automóviles quiere dejar suficiente espacio para la cabeza en sus automóviles para alojar cómodamente a todos, pero el 5 % de las personas que conducen son muy altas. Estudios anteriores indican que el 95 % mide 70.3 pulgadas.Con intención de ver si el resultado del estudio sigue siendo válido, es seleccionada una muestra aleatoria de tamaño 100. Se encuentra que las doce personas más altas en muestra tienen las siguientes alturas.

72.6	70.0	71.3	70.5
70.8	76	70.1	72.5
71.1	70.6	71.9	72.8

¿Es razonable usar 70.3 como el percentil del $95th$? (Prueba de dos colas)

4. Seis estudiantes se pusieron a dieta en un intento por perder peso con los siguientes resultados:

Nombre	Abdul	Ed	Jim	Max	Phil	Ray
Peso anterior	174	191	188	182	201	188
Peso posterior	165	186	183	178	203	181

¿La dieta es un medio eficaz de pérdida de peso?

5. Fueron seleccionados aleatoriamente 135 ciudadanos, se les pidió expresar su opinión sobre la politica exterior de EEUU. 43 de los 135 se opusieron a la política exterior de EEUU. Después de varias semanas, durante las cuales recibieron un boletín informativo, se les preguntó de nuevo su opinión, y 37 se opusieron, de estas 37, 30 eran personas que en un principio no se oponían a la politica exterior de EEUU. ¿Es significativo el cambio en el número de personas que se oponian a la politica exterior de EEUU?

6. En cierta ciudad la tasa de mortalidad por cada $100,000$ ciudadanos debido a accidentes automovilísticos para los últimos quince años fue 17.3, 17.9, 18.4, 18.1, 18.3, 19.6, 18.6, 19.2, 17.7, 20.0, 19.0, 18.8, 19.3, 20.2, 19.9. ¿Hay alguna base para afirmar que la tasa de mortalidad aumenta?

7. Un fabricante calcula el costo promedio en dólares de la producción de un determinado artículo para cada uno de 44 meses con los promedios resultantes: 13.65, 13.41, 13.53, 13.23, 13.58, 13.43, 13.73, 13.40, 13.70, 13.58, 13.80, 13.40, 13.63, 13.69, 13.92, 13.68, 13.72, 13.42, 13.66, 13.98,

13.81, 13.60, 13.32, 13.45, 13.27, 13.26, 13.28, 13.29, 13.10, 13.09, 13.36, 13.40, 13.35, 13.53, 13.66, 13.10, 13.28, 13.33, 13.02, 13.09, 13.12, 13.16, 12.96, 12.95.
¿Hay una tendencia estadísticamente significativa en los promedios?

8. Un jugador de Grandes Ligas ha recopilado el siguiente registro de records de doce años

	1953	1954	1955	1956	1957	1958	1959
Num. de home runs	7	14	17	15	9	19	16
Promedio de bateo	.212	.232	.234	.210	.256	.261	0.252

	1960	1961	1962	1963	1964
Num. de home runs	17	22	17	13	10
Promedio de bateo	.247	.255	.241	.238	.235

¿Hay una correlación significante entre el numero de home runs que él bateo y el promedio de bateo para ese año?

PRUEBAS DE RANGO

9. (**Distribución aproximada de Mann-Whitney**)
Respaldados en el Teorema central del Limite podemos encontrar un estadístico con distribución aproximadamente normal. Para ello recordemos que en la prueba Mann-Whitney el estadístico de prueba es:

$$T = S - \frac{n(n+1)}{2}; \quad \text{Donde} : S = \sum_{i=1}^{n} R(X_i)$$

Luego entonces usando el Teorema del Limite Central obtenemos que:

$$Z = \frac{T - \mathbb{E}(T)}{\sqrt{\text{Var}(T)}} \sim_{aprox} N(0,1)$$

En clase probamos que bajo H_0 y que bajo el supuesto de que no hay empates se cumple que:

$$\mathbb{E}(T) = \frac{nm}{2}$$

Demuestre que:

$$\text{Var}(T) = \frac{nm(n+m+1)}{12}$$

(Hint: Recuerde que bajo H_0, $R(X_i)$ es tal que $\mathbb{P}\left(R(X_i) = k\right) = \frac{1}{N} = \frac{1}{n+m}$ para toda $k \in \{1, 2, \ldots, n+m\}$, pero que $R(X_i)$ no es independiente de $R(X_j)$ con $i \neq j$)
Finalmente concluya entonces que un estadístico alternativo para hacer la prueba de Mann-Whitney es utilizar:

$$Z = \frac{T - \frac{nm}{2}}{\sqrt{\frac{nm(n+m+1)}{12}}} \sim_{aprox} N(0,1)$$

10. En la prueba Kruskall-Wallis se definió:

$$R_j = \sum_{i=1}^{n_j} R(X_{ji})$$

Como la suma de los rangos asociados a la población j. Demuestre que bajo H_0 se tiene que:

$$\text{Var}(R_j) = \frac{n_j(n+1)(n-n_j)}{12}$$

11. En la prueba Kruskall-Wallis se definió el estadístico de prueba:

$$T = \frac{12}{n(n+1)} \sum_{j=1}^{k} \frac{\left(R_j - \frac{1}{2}n_j(n+1)\right)^2}{n_j}$$

Pruebe que otra forma equivalente de escribir este estadístico es:

$$T = \left(\frac{12}{n(n+1)} \sum_{j=1}^{k} \frac{R_j^2}{n_j} \right) - 3(n+1)$$

12. (**Simulación de la distribución de Kruskal-Wallis**) El programa ubicado en KruskallWallis.R tiene el código para calcular los cuantiles de la distribución Kruskall-Wallis cuando $k = 3$. Modifique el programa para que calcule cuantiles de esta distribución pero para $k = 4$ y encuentre los cuantiles con $\alpha = 0.1, 0.05, 0.025, 0.01$ para las siguientes configuraciones:

$$n_1 = 3, n_2 = 2, n_3 = 2, n_4 = 3$$
$$n_1 = 3, n_2 = 3, n_3 = 2, n_4 = 2$$
$$n_1 = 2, n_2 = 2, n_3 = 2.n_4 = 2$$

Compare sus cuantiles con los cuantiles aproximados usando la distirbución χ^2 con $4 - 1 = 3$ grados de libertad.
¿La aproximación de la distribución χ^2 es buena?

13. En la prueba Wilcoxon demuestre que cuando la muestra no tiene empates (no se eliminan observaciones), entonces la varianza del estadístico de prueba es:

$$\text{Var}(T) = \frac{n(n+1)(2n+1)}{24}$$

14. En la prueba Friedman se definió

$$R_j = \sum_{i=1}^{n} R(X_{ji})$$

Como la suma de los rangos asociados a la columna j. Demuestre que bajo H_0 se tiene que:

$$\text{Var}(R_j) = \frac{n(k+1)(k-1)}{12}$$

15. (**Simulación de la distribución de Fredman**)

    - Elabore un programa en R que simule observaciones del estadístico de prueba de Fredman , cuando $k = 3$ y $n = 4$

    $$T = \frac{12}{nk(k+1)} \sum_{i=1}^{k} R_j^2 - 3n(k+1)$$

    - Del programa elaborado simule 100,000 observaciones del estadístico T , haga un histograma y sobre ponga la distribución χ^2 con 2 grados de libertad ¿La aproximación fue buena?
    - Por medio de la función quantile de R calcule los cuantiles de la distribución de Fredman con $\alpha = 0.1, 0.05, 0.025, 0.001$

16. Demuestre que en la prueba Fredman se tiene la siguiente igualdad (Bajo el supuesto de que no hay empates en la asignación de rangos)

$$T = \frac{12}{nk(k+1)} \sum_{j=1}^{k} \left(R_j - \frac{n(k+1)}{2} \right)^2 = \frac{12}{nk(k+1)} \sum_{i=j}^{k} R_j^2 - 3n(k+1)$$

17. (Prueba Wilcoxon) Resuelva los problema 3 y 4 ubicados en la pagina 215 del libro "Practical NonParametric Statistics"de W.J. Conover (**Primera Edición**). Si tienen la Segunda Edición, resuelva los problemas 1 y 2 ubicados en la página 292

18. (Prueba Mann-Whitney - Tuckey) Resuelva los problema 1 y 6c) ubicados en la pagina 236 y 237 del libro "Practical NonParametric Statistics"de W.J. Conover (**Primera Edición**).

19. (Prueba ρ Spearman) Resuelva el problema 1 c) ubicado en la pagina 254 del libro "Practical NonParametric Statistics"de W.J. Conover (**Primera Edición**) Si tienen la Segunda Edición, resuelva el problema 1 c) ubicado en la página 261

20. (Prueba de Fredman) Resuelva el problema 1 ubicado en la pagina 274 del libro "Practical NonParametric Statistics"de W.J. Conover (**Primera Edición**)

21. (Prueba de Kuskall-Wallis) Resuelva el problema 1 ubicado en la pagina 263 del libro "Practical NonParametric Statistics"de W.J. Conover (**Primera Edición**) Si tienen la Segunda Edición, resuelva el problemas 1 ubicado en la página 237-238

22. (**Prueba ρ-Spearman**) La prueba ρ-Spearman esta diseñada para detectar correlación entre dos variables continuas. Es una prueba que se basa en el coeficiente de correlación de pearson definido como:

$$r = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \overline{x})^2 \sum_{i=1}^{n} (y_i - \overline{y})^2}}$$

La idea es contrastar las siguientes Hipótesis:

H_0 : Las variables no están asociadas $\quad vs \quad$ H_1 : Existe una correlación entre las variables

Los datos que recibimos en esta prueba son observaciones bivariadas

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$$

La idea de Spearman es muy sencilla, pues se limita a asignar rangos a cada una de las observaciones en sus respectivas poblaciones y luego sustituir en la formula del coeficiente de correlación de pearson los rangos de cada una de las observaciones.
Definamos entonces a $R(x_i)$ como el rango asociado a la observación x_i dentro de la población X y $R(y_i)$ como el rango asociado a la observación y_i dentro de la población Y. Luego entonces Spearman propone el siguiente estadístico de prueba.

$$\rho = \frac{\sum_{i=1}^{n} \left(R(x_i) - \overline{R(x)}\right)\left(R(y_i) - \overline{R(y)}\right)}{\sqrt{\sum_{i=1}^{n} \left(R(x_i) - \overline{R(x)}\right)^2 \sum_{i=1}^{n} \left(R(y_i) - \overline{R(y)}\right)^2}}$$

Donde:

$$\overline{R(y)} = \frac{\sum_{i=1}^{n} R(y_i)}{n} \qquad \overline{R(x)} = \frac{\sum_{i=1}^{n} R(x_i)}{n}$$

Demuestre que bajo el supuesto de que no hay empates al momento de asignar rangos entonces se cumple que:

$$\rho = \frac{\sum_{i=1}^{n} \left(R(x_i) - \frac{n+1}{2}\right)\left(R(y_i) - \frac{n+1}{2}\right)}{n(n^2-1)/12}$$

Demuestre además que:

$$\rho = 1 - \frac{6\sum_{i=1}^{n} \left(R(x_i) - R(y_i)\right)^2}{n(n^2-1)}$$

La interpretación de coeficiente de Spearman es igual que la del coeficiente de correlación de Pearson. Oscila entre -1 y +1, indicándonos asociaciones negativas o positivas respectivamente, 0 cero, significa no correlació. Luego entonces valores cercanos de ρ a 1 nos hablan de que las variables están correlacionadas positivamente mientras que si ρ toma un valor cercano a -1 nos habla de que las variables están correlacionadas negativamente. Debido a lo anterior se propone la siguiente región de rechazo:

Rechazar H_0 a un nivel de significancia α si $\rho > \rho_{1-\alpha/2}$ o si $\rho < \rho_{\alpha/2}$ donde ρ_α se refiere al cuantíl α de la distribución ρ-Spearman. También podemos llevar pruebas de una cola por ejemplo:

H_0 : Las variables no están asociados o están correlacionadas negativamente $\quad vs$

H_1 : Existe una correlación positiva entre las variables

En cuyo caso ahora la regla de decisión es rechazar si $\rho > \rho_{1-\alpha}$

H_0 : Las variables no están asociados o están correlacionadas positivamente $\quad vs$

H_1 : Existe una correlación negativa entre las variables

En cuyo caso ahora la regla de decisión es rechazar si $\rho < \rho_{\alpha}$

Existen tablas donde se tabulan los cuantiles exactos mas importantes de la distribución ρ (Ej, Conover Tabla 10). Sin embargo una ventaja que tenemos hoy día es que podemos simular la destribución exacta de ρ y obtener los cuantiles para cualquier valor de α

(Distribución exacta de ρ)
Bajo H_0 se espera que la asignación de rango en la población X no afecte la asignación de rangos de la población Y, luego entonces tenemos $n!$ formas de asignar un rangos en X y de la misma forma tenemos $n!$ formas para asignar rangos en la población Y. Por lo tanto tenemos un total de $(n!)^2$ de arreglos posibles,todos con la misma probabilidad de ocurrir. Finalmente podemos calcular la distribución de ρ simulando cada uno de estos posibles arreglos y luego calculando el estadístico ρ.
En Spearman.R se encuentra un programa en R que simula observaciones del estadístico ρ y obtiene los cuantiles exactos de la distribución.

(Distribución aproximada).
Cuando el tamaño de muestra es grande el estadístico de prueba puede ser sustituido por otro que tiene una distribución aproximada a la Normal estándar, para deducir esta aproximación recordemos que por el teorema del límite central se tiene que:

$$Z = \frac{\rho - \mathbb{E}(\rho)}{\sqrt{\text{Var}(\rho)}} \sim_{aprox} N(0,1)$$

Demuestre entonces que bajo H_0 (Independencia entre $R(x_i)$ y $R(y_i)$) y bajo el supuesto de que no hay empates, se tiene que:

$$\mathbb{E}(\rho) = 0 \qquad \text{Var}(\rho) = \frac{1}{n-1}$$

(Hint: Recuerde que bajo H_0 suponemos independencia entre las variables $R(x_i)$ y $R(y_j)$ para toda i y j pero que $R(x_i)$ no es independiente de $R(x_j)$ al igual que $R(y_i)$ no es independiente de $R(y_j)$. Ademas bajo el supuesto de no haber empates se tiene que $\mathbb{P}\left(R(X_i) = k\right) = \frac{1}{n}$ para toda k.)

Luego entonces concluya que:

$$Z = \rho\sqrt{n-1} \sim_{aprox} N(0,1)$$

De donde ahora la regla de decisión es rechazar H_0 (en la prueba de dos colas) si $Z > Z_{1-\alpha/2}$ o si $Z < Z_{\alpha/2}$, donde Z_α se refiere al cuantil α de una distribución Normal(0,1)