7. INFERENCIA ESTADISTICA

Dr. Edgar Acuña

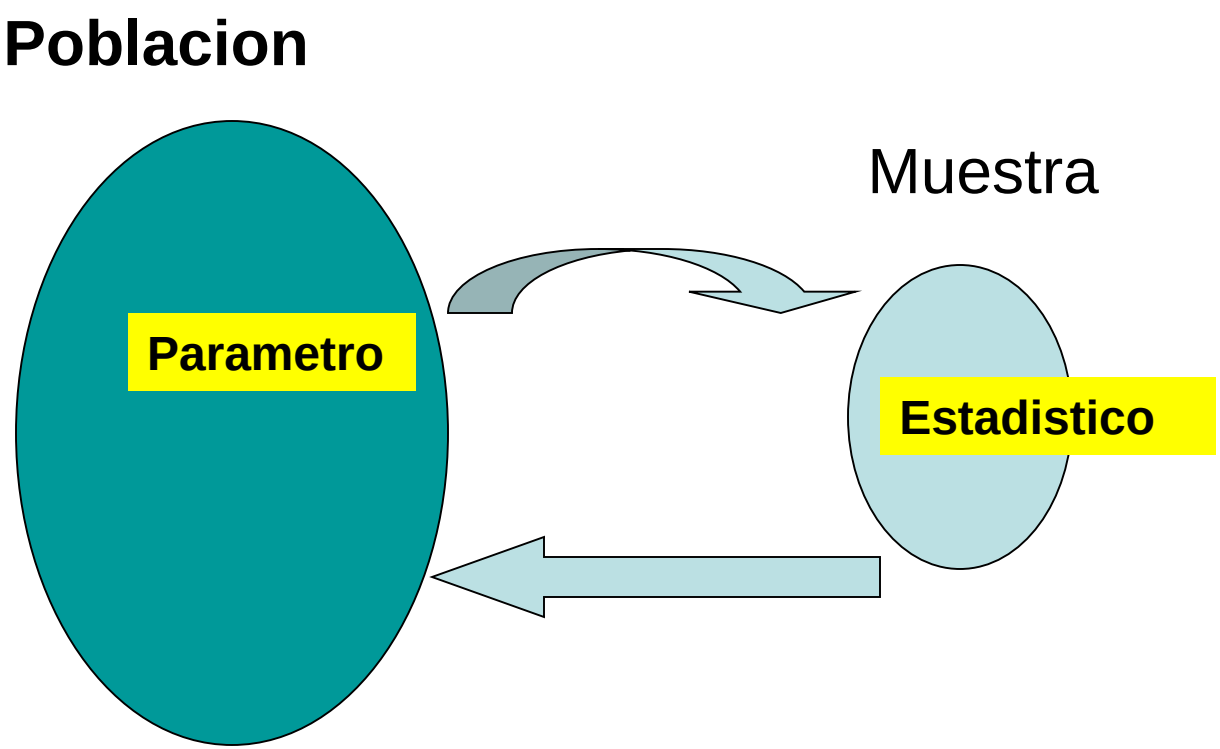
RECINTO UNIVERSITARIO DE MAYAGUEZ

INFERENCIA ESTADÍSTICA

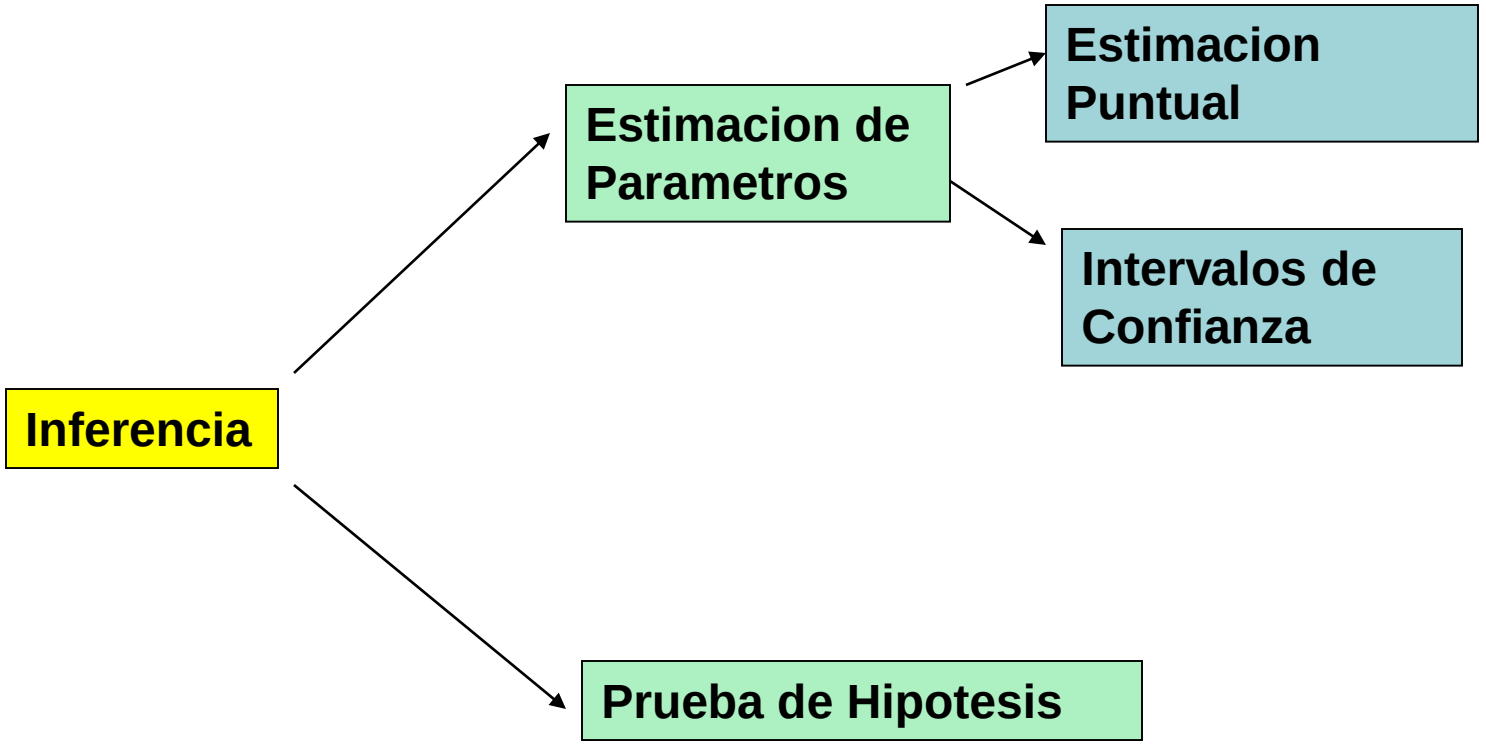
La Inferencia Estadística comprende los métodos que son usados para obtener conclusiones de la población en base a una muestra tomada de ella. Incluye los **métodos de estimación de parámetros** y **las pruebas de hipótesis**.

La Estimación de parámetros comprende a su vez la Estimación Puntual, en donde se estudian los diversos métodos de encontrar estimadores y las propiedades óptimas que deben tener éstos, y la Estimación por Intervalos de Confianza, en donde se estima un parámetro usando un intervalo centrado en un estimado del parámetro y de longitud igual a dos veces el error de estimación.

Inferencia Estadistica



Inferencia estadistica



Hipótesis Estadística

Una **Hipótesis Estadística** es una afirmación que se hace acerca de un parámetro poblacional. Por ejemplo, el tiempo de vida promedio para una persona diagnosticada con cáncer de pulmón es 180 días.

hipótesis nula, La afirmación que está establecida y que se espera sea rechazada después de aplicar una **prueba estadística** es llamada la ***hipótesis nula*** y se representa por *Ho*.

hipótesis alterna La afirmación que se espera sea aceptada después de aplicar una **prueba estadística** es llamada la ***hipótesis alterna*** y se representa por *Ha*.

Hipótesis Estadística (cont)

Una **prueba estadística** es una fórmula, basada en la distribución del estimador del parámetro que aparece en la hipótesis y que va a permitir tomar una decisión acerca de aceptar o rechazar una hipótesis nula.

La prueba estadistica no es ciento por ciento confiable y hay dos tipos de errores que se pueden cometer.

El ***error tipo I,*** se comete cuando se rechaza una hipótesis nula que realmente es cierta. Similar a que a una persona que se somete a una prueba no se le diagnostique con cáncer cuando realmente lo tiene (falso negativo).

El ***error tipo II*** que se comete cuando se acepta una hipótesis nula que realmente es falsa. Similar a que a una persona se le diagnostique con cáncer cuando realmente no lo tiene (falso positivo).

Hipótesis Estadística (cont)

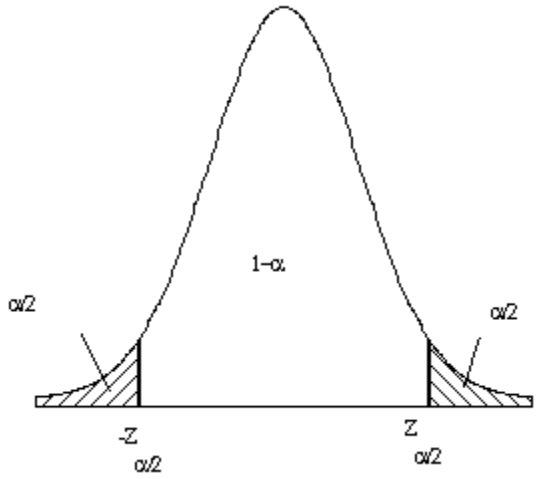
El **nivel de significación,** representada por α, es la probabilidad de cometer ***error tipo I***, y por lo general se asume que tiene un valor de .05 ó .01. También puede ser interpretado como el área de la región que contiene todos los valores posibles de la prueba estadística para los cuales la hipótesis nula es rechazada.

La probabilidad de cometer ***error tipo II***, es representado por β y al valor $1-\beta$ se le llama ***la potencia de la prueba***.

Una buena prueba estadística es aquella que tiene una potencia de prueba alta.

7.1 Inferencias acerca de la Media Poblacional (varianza conocida).

Supongamos que de una población normal con media desconocida μ y varianza conocida σ^2 se extrae una muestra de tamaño *n*, entonces de la distribución de la media muestral se obtiene que:



$$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

se distribuye como una normal estándar.

Luego $P(-Z_{a/2} < Z < Z_{a/2}) = 1 - \alpha$

Donde $Z_{\alpha/2}$ es el valor de la normal estándar tal que el área a la derecha de dicho valor es $\alpha/2$.

Inferencias acerca de la Media Poblacional (varianza conocida).

Sustituyendo la fórmula de *Z*, se obtiene:

$$P(\bar{x} - Z_{\alpha/2}\,\sigma/\sqrt{n} < \mu < \bar{x} + Z_{\alpha/2}\,\sigma/\sqrt{n}) = 1 - \alpha$$

Notar que los dos extremos del intervalo son aleatorios.

De lo anterior se puede concluir que un Intervalo de Confianza del 100 (1-α) % para la media poblacional μ, es de la forma:

$$\bar{x} - Z_{\alpha/2}\,\sigma/\sqrt{n}\ ,\ \bar{x} + Z_{\alpha/2}\,\sigma/\sqrt{n}$$

Inferencias acerca de la Media Poblacional (varianza conocida).

La siguiente tabla muestra los $Z_{\alpha/2}$ más usados.

Nivel de Confianza	$Z_{\alpha/2}$
90	1.645
95	1.96
99	2.58

Python tiene un comando **norm.interval** para hallar intervalos de confianza para μ.

Inferencias acerca de la Media Poblacional (varianza conocida)

Ejemplo 7.1. Un cardiólogo desea hallar un intervalo de confianza del 90% para el nivel colesterol promedio de todos los pacientes que presentan problemas cardíacos. Para esto asume que la distribución de los niveles de colesterol es normal con una desviación estándar σ = 13 y usa la siguiente muestra al azar de niveles de colesterol de 20 pacientes con problemas cardíacos.

217	223	225	245	238	216	217	226	202
233	235	242	219	221	234	199	236	248
218	224							

Laboratorio 20 en Python

En el Laboratorio 19, se crea una función IC_media, para hallar el intervalo de confianza para la media poblacional.

IC_media(colest,13,.90)
El interval de confianza es: (221.11859412022926, 230.68140587977075)

Interpretación: Hay un 90% de confianza de que el nivel de colesterol de todos los pacientes con problemas cardíacos caiga entre 221.12 y 230.68.

Inferencias acerca de la Media Poblacional (varianza conocida).

En la práctica si la media poblacional es desconocida entonces, es bien probable que la varianza también lo sea puesto que en el cálculo de σ^2 interviene μ. Si ésta es la situación, y si el tamaño de muestra es grande ($n > 30$, parece ser lo más usado), entonces σ^2 es estimada por la varianza muestral s^2 y se puede usar la siguiente fórmula para el intervalo de confianza de la media poblacional:

$$\overline{X} - z_{\alpha/2} s/\sqrt{n}, \overline{X} + z_{\alpha/2} s/\sqrt{n}$$

También se pueden hacer pruebas de hipótesis con respecto a la media poblacional μ. *Por conveniencia, en la hipótesis nula siempre se asume que la media es igual a un valor dado.*

Existen dos métodos para hacer la prueba de hipótesis: el método clásico y el método del "P-value".

En el método clásico, se evalúa la prueba estadística de Z y al valor obtenido se le llama ***Z calculado*** (*Zcalc*). *Por otro lado el nivel de significancia* α, definido de antemano determina una región de rechazo y una de aceptación. *Si Zcalc cae en la región de rechazo, entonces se concluye que hay suficiente evidencia estadística para rechazar la hipótesis nula basada en los resultados de la muestra tomada.*

Formulas para prueba de hipotesis de medias

Caso I

$H_o : \mu = \mu_0$

$H_a : \mu < \mu_0$

Caso II

$H_o : \mu = \mu_0$

$H_a : \mu \neq \mu_0$

Caso III

$H_o : \mu = \mu_0$

$Ha : \mu > \mu_0$

Prueba Estadística:

$$Z = \frac{\bar{x} - \mu_o}{\frac{\sigma}{\sqrt{n}}}$$

Decisión:

Si $Z_{cal} < -Z_{\alpha}$ *entonces*

se rechaza Ho

Si $|Zcal| > Z\alpha/2$ *entonces*

se rechaza Ho

Si $Zcal > Z_{\alpha}$ *entonces*

se rechaza Ho

Prueba de hipotesis usando “p-values”

El “P-value” llamado el nivel de significación observado, es el valor de α al cual se rechazaría la hipótesis nula si se usa el valor calculado de la prueba estadística. En la práctica un “P-value” cercano a 0 indica un rechazo de la hipótesis nula. Así un “P-value” menor que .05 indicará que se rechaza la hipótesis nula.

Fórmulas para calcular “P-value”: *Depende de la forma de la hipótesis alterna*

Si Ha: $\mu > \mu o$, entonces P-value = Prob ($Z > Zcalc$).

Si Ha: $\mu < \mu o$, entonces P-value = Prob ($Z < Zcalc$).

Si Ha: $\mu \neq \mu o$, entonces P-value = 2Prob ($Z > |Zcalc|$).

Los principales programas estadísticos, dan los “P-values” para la mayoría de las pruebas estadísticas.

El p-value de la prueba Z se puede hallar usando la funcion norm.cdf

Ejemplo

Ejemplo 7.3. En estudios previos se ha determinado que el nivel de colesterol promedio de pacientes con problemas cardíacos es 220. Un cardiólogo piensa que en realidad el nivel es más alto y para probar su afirmación usa la muestra del Ejemplo 7.1. ¿Habrá suficiente evidencia estadística para apoyar la afirmación del cardiólogo? Justificar su contestación.

Solución:

La hipótesis nula es *Ho:* $\mu = 220$ (el nivel de colesterol promedio es 220)
La hipótesis alterna es Ha: $\mu > 220$ (el cardiólogo piensa que el nivel promedio de colesterol es mayor de 220).

En el Laboratorio 19, se crea una función ztest para probar hipótesis acerca de la media poblacional, si se conoce la varianza.

Los resultados son los siguientes:
ztest(colest,220,13)

El P-values : 0.02119546913849634 8

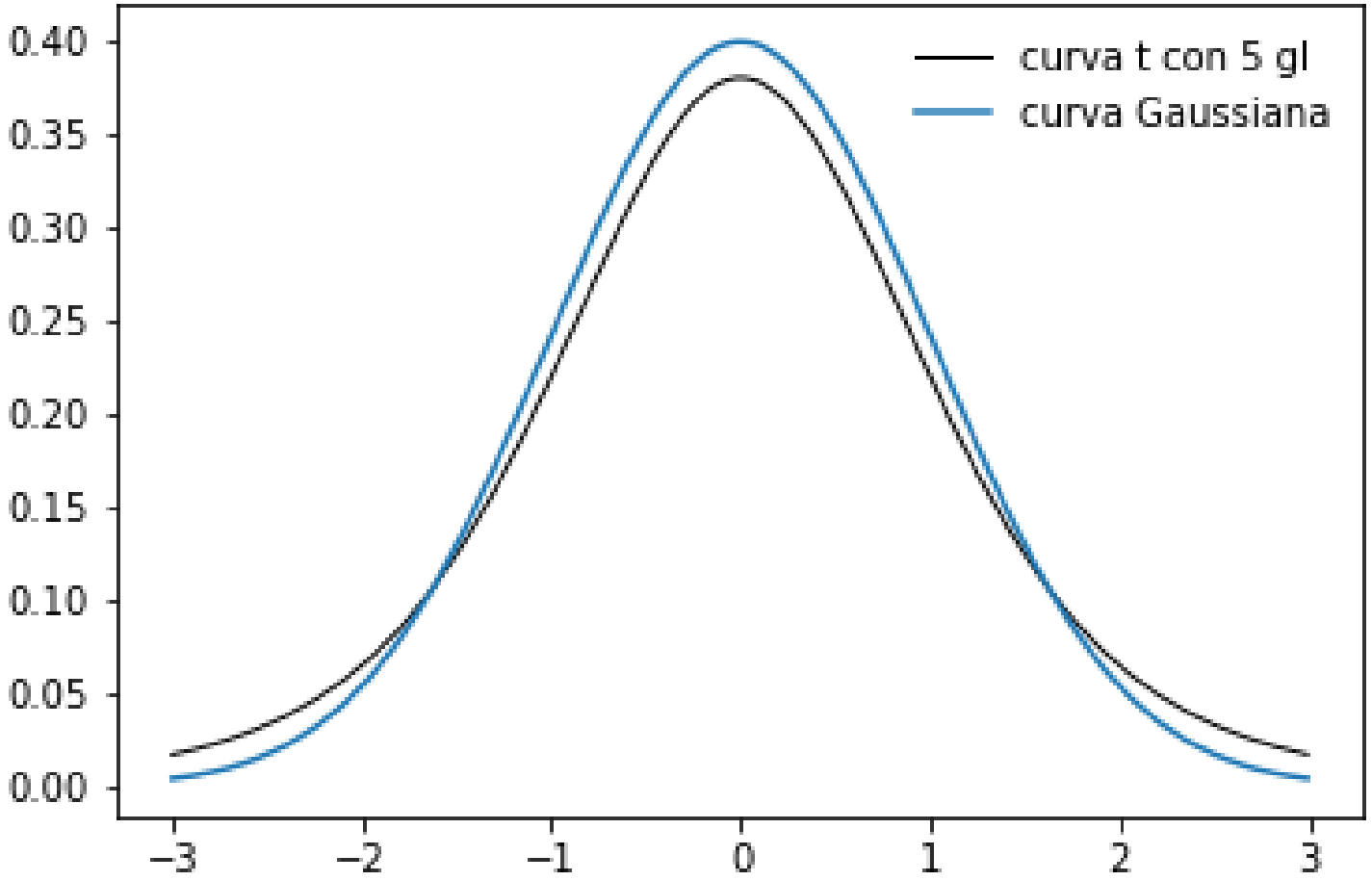
Interpretación: *El valor del "P-value" (el área a la derecha de 2.03) es .021 menor que el nivel de significación α=.05, por lo tanto; se rechaza la hipótesis nula y se concluye de que sí hay evidencia estadística de que el nivel de colesterol promedio de los pacientes con problemas cardíacos es mayor de 220. O sea los resultados apoyan lo que afirma el cardiólogo.*

7.2 Inferencias acerca de la Media Poblacional (Varianza Desconocida)

Supongamos que la población es normal con media y varianza desconocida y que se desea hacer inferencias acerca de μ, *basada en una muestra pequeña (n* < 30) tomada de la población. *En este caso la distribución de la media muestral* $\overline{X}$ ya no es normal, sino que sigue la distribución ***t de Student*.**

La distribución ***t de Student*** (W. Gosset, 1908) es bastante similar a la Normal Estándar, con la diferencia que se aproxima más lentamente al eje horizontal. El parámetro de esta distribución es llamado **grados de libertad**, y se puede notar que a medida que los grados de libertad aumentan, la curva de la ***t*** y la curva normal estándar se asemejan cada vez más. Por cada estimación de parámetro, calculada en forma independiente, que aparece en la formula del estadistico se pierde un grado de libertad con respecto al total de datos tomados.

Comparacion de la curva Normal y la curva t con 5gl.



Si de una población Normal con media μ y desviación estándar σ se extrae una muestra de tamaño *n*, entonces el estadístico:

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

se distribuye como una *t de Student* con *n*-1 grados de libertad.

Recordar que la desviación estándar *s* puede ser escrita en términos de $\bar{X}$

Un intervalo de confianza del 100 (1-α) % para μ *es de la forma:*

$$\bar{X} - t_{(n-1,\alpha/2)} s/\sqrt{n}, \bar{X} + t_{(n-1,\alpha/2)} s/\sqrt{n}$$

donde *s* es la desviación estándar muestral. Aquí $t_{(n-1,\alpha/2)}$ es un valor de t con n-1 grados de libertad y tal que el area a la derecha de dicho valor es $\alpha/2$.

En Python se usa la funcion t.Interval para hallar el intervalo de confianza para la media usando la t.

Ejemplo

Ejemplo 7.5. Los tiempos de sobrevivencia (en años) de 12 personas que se han sometido a un transplante de corazón son los siguientes:

3.1 .9 2.8 4.3 .6 1.4 5.8 9.9 6.3 10.4 0 11.5

Hallar un intervalo de confianza del 99 por ciento para el promedio de vida de todas las personas que se han sometido a un transplante de corazón.

Solución: Laboratorio 20

Python tiene una función t.Interval para hacer este intervalo de confianza, Pero podemos construir una función IC_mean_t

```
surv=[3.1,.9,2.8,4.3,.6,1.4,5.8,9.9,6.3,10.4,0,11.5]
IC_media_t(surv,0.99)
(1.1224942203449952, 8.377505779655003)
```

Con un 99% de confianza el tiempo de vida promedio de toda persona que se somete a un transplante de corazon caera entre 1.12 y 8.37 anos.

Prueba de hipotesis (varianza desconocida)

Caso I	Caso II	Caso III
$H_o : \mu=\mu_0$	$H_o : \mu=\mu_0$	$H_o : \mu=\mu$
$H_a : \mu<\mu_0$	$H_a : \mu \neq \mu 0$	$Ha : \mu>\mu 0$

Prueba Estadística

$$t = \frac{\bar{x} - \mu_o}{\frac{s}{\sqrt{n}}}$$

Esta prueba se distribuye como una t con n-1 grados de libertad

Si $t_{cal} < -t_{\alpha}$ *entonces*	*Si* $\lvert t_{cal} \rvert > t_{\alpha/2}$ *entonces*	*Si* $t_{cal} > t_{\alpha}$ *entonces*
se rechaza H_o	*se rechaza* H_o	*se rechaza* H_o

Si Ha: $\mu > \mu o$, *entonces P-value = Prob* ($t > tcalc$).

Si Ha: $\mu < \mu o$, *entonces P-value = Prob* ($t < tcalc$).

Si Ha: $\mu \neq \mu o$, *entonces P-value = 2Prob* ($t > |tcalc|$).

Ejemplo 7.6 Usando los datos del Ejemplo 7.5, un cardiocirujano afirma que el tiempo de vida promedio de las personas sometidas a transplante de corazón es mayor que 4 años. ¿A qué conclusión se llegará después de hacer la prueba de hipótesis?

Solución:

La hipótesis nula es *H0:* $\mu = 4$ (el tiempo de vida promedio de todas las personas que se han sometido a transplante de corazón es de 4 años) y la hipótesis alterna es *Ha:* $\mu > 4$ (el tiempo de vida promedio *es mayor que 4 años).*

Ver Laboratorio 21

Interpretación: *El valor del "P-value" (el área a la derecha de 0.64) es .267 mayor que el nivel de significación* $\alpha = .05$*, por lo tanto NO se rechaza la hipótesis nula y se concluye de que no hay evidencia de que el tiempo promedio de vida después del transplante haya aumentado de 4 años.*

7.3 Inferencia para Proporciones

Cuando estamos interesados en estimar la proporción *p* (o el porcentaje) de ocurrencia de un evento. Se necesita definir una variable aleatoria *X* que indique el número de veces que ocurre el evento en una muestra de tamaño *n* y con probabilidad de éxito, *p*. Se puede mostrar que cuando el tamaño de muestra es grande, tal que np > 5, entonces el estadístico

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}}$$

se distribuye aproximadamente como una normal estándar. Aquí *p* representa la proporción poblacional que se desea estimar, y $\hat{p} = \frac{x}{n}$ es la proporción muestral.

En **Python**, se tiene el modulo statsmodels que tiene una función proportions_ztest que hace prueba de proporciones

Inferencia para Proporciones

Intervalo de confianza (aproximado) del 100 (1-α) % para la proporción poblacional *p* es:

$$\hat{p} - Z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}, \hat{p} + Z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}$$

Pruebas de hipótesis:

Caso I	**Caso II**	**Caso III**
$Ho : p=p_0$	$Ho : p=p_0$	$Ho : p=p_0$
$Ha : p<p_0$	$Ha : p \neq p_0$	$Ha : p>p_0$

Prueba Estadística (Aproximada):

$$Z = \frac{(\hat{p} - p_0)}{\sqrt{\frac{p_0 q_0}{n}}}$$

Algunos autores usan $\hat{p}$ en lugar de p en la expression del denominador

Decisión

Si $Zcal < -Z\alpha$ entonces se rechaza H_o	**S**i $\|Zcal\| > Z\alpha/2$ entonces se rechaza H_o	**S**i $Zcal > Z\alpha$ entonces se rechaza H_o

Ejemplo

Ejemplo 7.7. En 1990 en un cierto país, se reportó que dos de cada 5 personas pensaban que debería incrementarse el poder nuclear. En una encuesta reciente hecha en1996 a 1225 personas se encontró que 478 de ellos pensaban que se debería aumentar el poder nuclear.¿Piensa Ud. que hay evidencia de que la opinión de la gente en 1996 ha cambiado con respecto al 1990? Justificar su contestación.

Solución:

Hay que probar la siguiente hipótesis:

$H_0 : p = .4$ (la proporción no cambió de 1990 a 1996).

$H_a : p \neq .4$ (la proporción cambió de 1990 a 1996).

Ejemplo (sol.)

proportions_ztest(478,1225,.4)

(-0.7028700551758291 2, 0.4821367341350553 1)

La prueba estadistica de Z es -0.702 y el pvalue es 0.4821.

Interpretación: Viendo que el "p-value" es .482 mucho mayor que .05 se llega a la conclusión de que no hay suficiente evidencia de que la proporción de personas a favor de un incremento del poder nuclear haya cambiado de 1990 a 1996.

.

Ejemplo

Ejemplo 7.8. El director de un hospital afirma que el 25 por ciento de los nacimientos que
ocurren allí son por cesárea. Un médico que trabaja en dicho hospital piensa que ese porcentaje es mayor. Para probar su afirmación recolecta información de los 25 nacimientos ocurridos durante una semana.

Partos

Cesárea normal cesárea normal normal normal normal cesárea normal
cesárea normal cesárea normal normal normal normal normal cesárea
normal normal cesárea normal normal cesárea normal

¿Habrá suficiente evidencia estadística para apoyar la afirmación del médico?

Solución:

Los datos son entrados en una columna llamada *partos*, luego se usará la funcion proportions_ztest de statsmodels

será considerado éxito que el parto sea normal y fracaso, que el parto sea por cesárea.

Ejemplo (cont.)

Luego las hipótesis planteadas son:
Ho: p =.25 (el 25% de los partos son por cesárea)
Ha: p >.25 (mass del 25% de los partos son por cesarea)

Python en la prueba estadistica de Z usa p estimado para estimar la desviacion estandar, No usa la po asumida en la hipotesis nula.

```
conteos=partos.count('cesarea')
trials=len(partos)
proportions_ztest(conteos,trials,value=.25,alternative='larger')
(0.750306309998475 7, 0.2265351215665391 5)
```

Interpretación: De acuerdo al "P-value" = 0.2265 > .05 no se rechaza la hipótesis nula. Por lo tanto, no hay evidencia suficiente para concluir que lo que afirma el médico es correcto.

7.5. Comparando la varianza de dos poblaciones

Supongamos que se tienen dos poblaciones normales con varianzas desconocidas σ_1^2 y σ_2^2

Si de la primera población se toma una muestra de tamaño *m* que tiene una varianza muestral s_1^2 y de la segunda población se toma una muestra, independiente de la primera, de tamaño *n* que tiene una varianza muestral s_2^2

Se puede mostrar que la razón

$$\frac{s_1^2 / \sigma_1^2}{s_2^2 / \sigma_2^2}$$

se distribuye como una F con *m*-1 grados de libertad en el numerador y *n*-1 en el denominador.

Caso I

Ho : $\sigma_1^2 = \sigma_2^2$

Ha : $\sigma_1^2 < \sigma_2^2$

Caso II

Ho : $\sigma_1^2 = \sigma_2^2$

Ha : $\sigma_1^2 \neq \sigma_2^2$

Caso III

Ho : $\sigma_1^2 = \sigma_2^2$

Ha : $\sigma_1^2 > \sigma_2^2$

Prueba Estadística:

$$F = \frac{s_1^2}{s_2^2}$$

con *m*-1 g.l. en el numerador y *n*-1 g.l en el denominador

Decisión:

Si $F_{cal} < F_{\alpha}$ entonces se rechaza Ho

Si $F_{cal} < F_{\alpha/2}$ o Fcal $> F_{1-\alpha/2}$ se rechaza Ho

Si $F_{cal} > F_{1-\alpha}$ entonces se rechaza Ho

Python no hace pruebas de igualdad de varianza de dos o más grupos, basada en la prueba de F, en su lugar usar la prueba de Barttlet o de Levene.

Ejemplo 7.11 En el siguiente ejemplo se trata de comparar las varianzas de los puntajes de aprovechamiento matematico en el examen del College Board, de los estudiantes de escuelas públicas y privadas. Los datos recolectados son:

Est	aprovech	escuela
1	580	pública
2	638	pública
3	642	privada
4	704	pública
5	767	privada
6	641	privada
7	721	privada
8	625	privada
9	694	pública
10	615	pública
11	617	pública
12	623	pública
13	689	privada
14	689	pública

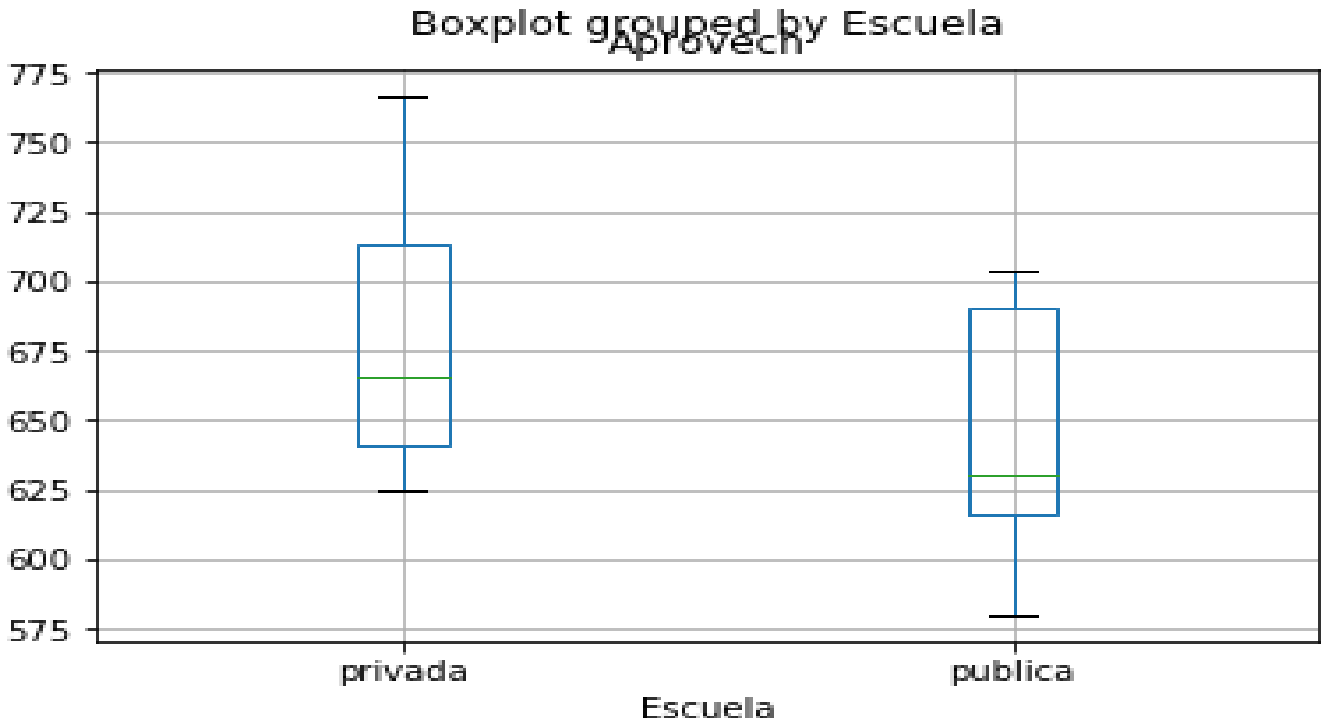
Resultados

Ho: Varianza de los puntajes de estudiantes de escuela pública es igual a la varianza de puntajes de los estudiantes de escuela privada.

Ha: Las varianzas no son iguales.

```
grupo1=aprovech.query('Escuela=="privada"')['Aprovech']
grupo2=aprovech.query('Escuela=="publica"')['Aprovech']
```

La prueba de F es: 1.50375642824 El p-value es: 0.600930704069. El p-value es mayor que .05 no hay suficiente evidencia para rechazar la hipotesis nula



Interpretación: El "P-value" de la prueba de F es .601 mucho mayor que .05, luego se acepta la hipótesis nula y se concluye que los puntajes en la prueba de aprovechamiento en las escuelas pública y privada tienen igual varianza. De las gráficas se puede ver que los "boxplots" de ambos grupos tienen aproximadamente el mismo alargamiento.

7.6 Comparación entre dos medias poblacionales usando muestras independientes

Supongamos que se tienen dos poblaciones distribuidas normalmente con medias desconocidas μ_1 y μ_2, respectivamente. Se puede aplicar una prueba *t de Student* para comparar las medias de dichas poblaciones basándonos en dos muestras independientes tomadas de ellas.

a) Si las varianzas de las poblaciones son iguales ($\sigma_1^2 = \sigma_2^2 = \sigma^2$) entonces se puede mostrar que:

$$t = \frac{(\bar{x} - \bar{y}) - (\mu_1 - \mu_2)}{s_p \sqrt{\frac{1}{m} + \frac{1}{n}}}$$

se distribuye como una *t* con $m + n - 2$ grados de libertad.

la varianza poblacional es estimada por una varianza combinada de las varianzas de las dos muestras tomadas.

$$s_p^2 = \frac{(m-1)s_1^2 + (n-1)s_2^2}{m+n-2}$$

Un intervalo de confianza del 100(1-α) % para la diferencia μ_1-μ_2 de las medias poblacionales será de la forma:

$$\bar{x} - \bar{y} \pm t_{(\alpha/2, n+m-2)} s_p \sqrt{\frac{1}{m} + \frac{1}{n}}$$

Las pruebas de hipótesis son:

Caso I	**Caso II**	**Caso III**
Ho : $\mu_1 = \mu_2$	Ho : $\mu_1 = \mu_2$	Ho : $\mu_1 = \mu_2$
Ha : $\mu_1 < \mu_2$	Ha : $\mu_1 \neq \mu_2$	Ha : $\mu_1 > \mu_2$

Prueba Estadística:

$$t = \frac{\bar{x} - \bar{y}}{s_p \sqrt{\frac{1}{m} + \frac{1}{n}}}$$ con m+n-2 grados de libertad

Decisión:

Si $t_{cal} < -t_{\alpha}$ entonces se rechaza Ho	Si $t_{cal} < t_{\alpha/2}$ o $t_{cal} > t_{1-\alpha/2}$ se rechaza Ho	Si $t_{cal} > t_{1-\alpha}$ se rechaza Ho

Ejemplo

Ejemplo 7.13. Se desea comparar si los estudiantes de escuelas privadas y públicas tienen igual rendimiento en la prueba de aprovechamiento matemático del College Board. Los datos aparecen en el Ejemplo 7.11.

Solucion

En Python se usa stats.ttest_ind

```
stats.ttest_ind(grupo1, grupo2, equal_var = True)
Ttest_indResult(statistic=1.336468097703087,
pvalue=0.2061849344443425)
```

Interpretación: El valor del “P-value” es .206 mayor que el nivel de significación α = .05, por lo tanto NO se rechaza la hipótesis nula y se concluye de que no hay evidencia de que los estudiantes de escuela pública tengan un rendimiento distinto que los de escuela privada en las pruebas de aprovechamiento. El número de grados de libertad de la t es 12.

b) Si las varianzas de las poblaciones no son iguales, entonces se usa una prueba aproximada de *t*, donde el número de grados de libertad es calculado aproximadamente.

La prueba de *t* aproximada está dada por:

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}}}$$

donde los grados de libertad *gl* son aproximados por la siguiente fórmula:

$$gl = \frac{(c_1 + c_2)^2}{\frac{c_1^2}{m-1} + \frac{c_2^2}{n-1}}$$

Con $c_1 = \frac{s_1^2}{m}$ y $c_2 = \frac{s_2^2}{n}$

Ejemplo

Ejemplo 7.14. Usando los datos del Ejemplo 7.12 (academic.uprm.edu/eacuna/gpasex.csv), probar si las estudiantes mujeres tienen mejor promedio académico que los varones.

```
gpasex=pd.read_csv("c://esma3016/gpasex.csv",sep=",")
grupo1=gpasex.query('genero=="mujer"')['gpa']
grupo2=gpasex.query('genero=="hombre"')['gpa’]
Ftest(grupo1,grupo2)
La prueba de F es: 0.323504885709 El p-value es: 0.044764470156
ttest, pvalue,dof=sms.ttest_ind(grupo1, grupo2, usevar = 'unequal',
alternative="larger")
print "ttest es:",ttest,"p-value es:",pvalue,"grados de libertad son:",dof
ttest es: 1.45280070785 p-value es: 0.0826394688551 grados de libertad son:
16.2824660949
```

Interpretación: Como el “P-value” es .083 > .05 aunque no por mucho, se concluye que no hay suficiente evidencia de que el promedio académico de las mujeres sea mayor que el de los hombres.

Comparando media de dos poblaciones usando muestras pareadas

En este caso se trata de comparar dos métodos o tratamientos, pero se quiere que las unidades experimentales (sujetos) donde se aplican los tratamientos sean las mismas, ó lo más parecidas posibles, para evitar influencia de otros factores en la comparación Por ejemplo, cuando se evalua la efectividad de un seminario o de la utilizacion de una droga tendria que hacerse las mediciones antes y despues en los mismos sujetos

Sea X_i el valor antes del tratamiento y Y_i el valor despues del tratamiento en el *i-ésimo* sujeto. Consideremos $d_i = X_i - Y_i$ la diferencia antes-despues del tratamiento en el *i-ésimo* sujeto.

Las inferencias que se hacen son acerca del promedio poblacional μ_d de las d_i. Si $\mu_d = 0$, entonces significa que no hay diferencia entre los dos tratamientos.

Intervalo de Confianza

Un intervalo de confianza del 100(1-α)% para la diferencia poblacional μ_d dada una muestra de tamaño n es de la forma

$$\left(\bar{d} - t(n\text{-}1,\alpha/2)\ \text{sd}/\sqrt{n}\ ,\ \bar{d} + t(n\text{-}1,\alpha/2)\ \text{sd}/\sqrt{n}\ \right)$$

donde $\bar{d}$, es media de las diferencias muestrales d_i y es la desviación estándar.

$$s_d = \sqrt{\frac{\sum_i (d_i - \bar{d})^2}{n-1}}$$

Pruebas de Hipótesis

Caso I
Ho : $\mu_d = 0$
Ha : $\mu_d < 0$

Caso II
Ho : $\mu_d = 0$
Ha : $\mu_d \neq 0$

Caso III
Ho : $\mu_{d-}=0$
Ha : $\mu_d > 0$

Prueba Estadística:

$t = \dfrac{\bar{d}}{\frac{s_d}{\sqrt{n}}}$ se distribuye con una *t* de Student con *n*-1 *gl*.

Decisión:

Si t<-*t*α entonces se rechaza Ho

Si | t |>*t*α/2 entonces se rechaza Ho

Si Tcal >*t*α entonces se rechaza Ho

En **Python se usa la función** ttest_rel del modulo stats de spicy

Ejemplo 7.15

Un médico desea investigar si una droga tiene el efecto de bajar la presión sanguínea en los usuarios. El médico eligió al azar 15 pacientes mujeres y les tomó la presión, luego les recetó la medicina por un período de 6 meses, y al final del mismo nuevamente les tomó la presión. Los resultados son como siguen (academic.uprm.edu/eacuna/ejemplo715.txt):

Sujetos															
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Antes	70	80	72	76	76	76	72	78	82	64	74	92	74	68	84
Después	68	72	62	70	58	66	68	52	64	72	74	60	74	72	74

Solución:

Sea μ_d que representa la media poblacional de las diferencias. Luego:

H_o: $\mu_d = 0$ (La droga no tiene ningún efecto)

H_a: $\mu_d > 0$ (La droga tiene efecto, la presión antes de usar la droga era mayor que después de usarla).

Ejemplo (Cont.)

stats.ttest_rel(antes,despues)

Ttest_relResult(statistic=array([3.10536049]), pvalue=array([0.00774944]))

Conclusion: Esta funcion da el p-value de la prueba de dos lados, el p-value la prueba de un solo dado es la mitad de este valor. O sea .0038, menor que 0.5. Se rechaza la hipotesis Nula. Hay suficiente evidencia estadistica para conluir que la medicina baja la presión.

Tambien se pueden aplicar la t de una sola muestra a la diferencias. Usando el modulo statsmodels seria:

sms.DescrStatsW(antes-despues).ttest_mean(0,"larger")

(array([3.10536049]), array([0.00387472]), 14.0)

Interpretación: Notando que el "P-value" es .0038 menor que .05, se rechaza la hipótesis nula y se llega a la conclusión de que, efectivamente la droga reduce la presión sanguínea.