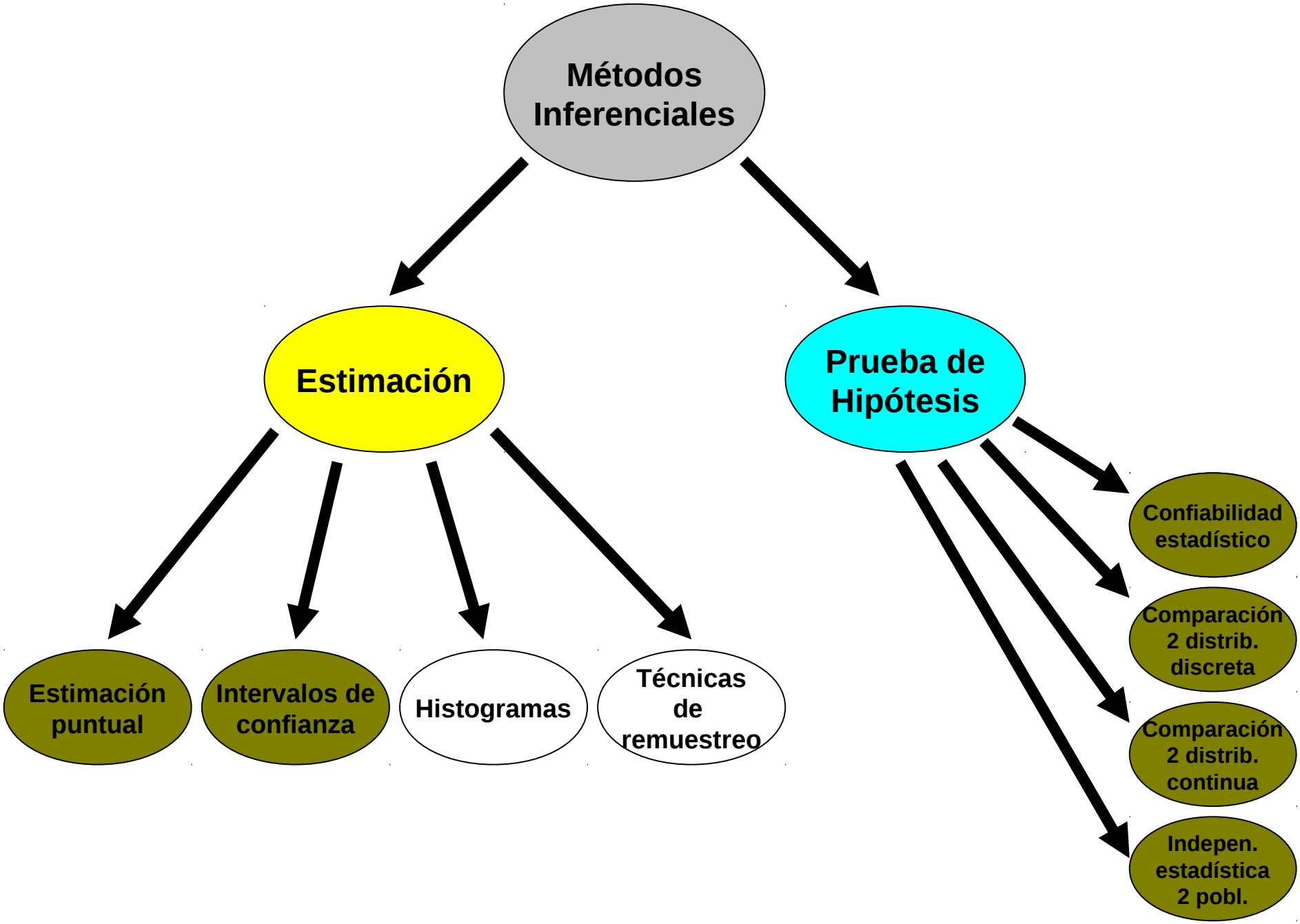
Inferencia Estadística: Prueba de Hipótesis

Inferencia Estadística:

Hemos estudiado cómo a partir de una muestra de una población podemos obtener una estimación puntual o bien establecer un intervalo más o menos aproximado para encontrar los parámetros que rigen la ley de probabilidad de una v.a. definida sobre la población. Es lo que denominábamos estimación puntual y estimación de intervalos de confianza respectivamente.

En la Prueba de Hipótesis queda implícita la existencia de dos teorías o Hipótesis (nula y alternativa) que de alguna manera reflejarán las ideas a priori que tenemos y que pretendemos contrastar con la “realidad”.


Métodos Inferenciales
Estimación
Prueba de Hipótesis
Estimación puntual
Intervalos de confianza
Histogramas
Técnicas de remuestreo
Confiabilidad estadístico
Comparación 2 distrib. discreta
Comparación 2 distrib. continua
Indepen. estadística 2 pobl.

Procedimiento general para la PH

1. Hipótesis
2. Nivel de significación
3. Estadístico de prueba
4. Zona de aceptación
5. Cómputos necesarios
6. Decisión
7. Conclusión

Procedimiento general para la PH

Hipótesis

"el ejercicio constante disminuye el nivel de colesterol en sangre"

"el ejercicio constante NO disminuye el nivel de colesterol en sangre"

Hipótesis estadísticas: Hipótesis nula H_0 Hipótesis alternativa H_1

$$H_0 : \theta = \theta_0$$

$$H_1 : \begin{cases} \theta > \theta_0 \\ \theta < \theta_0 \\ \theta \neq \theta_0 \end{cases}$$

Procedimiento general para la PH

Nivel de significación

PH se basa fundamentalmente en determinar si la diferencia que existe entre el valor del estadístico muestral y el valor del parámetro poblacional es lo suficientemente grande que no pueda atribuirse simplemente al azar, si no a la falsedad de la hipótesis nula.

0.05 **0.10** **0.01** **...**

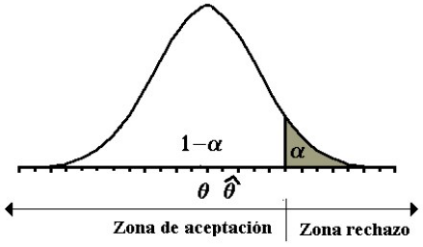
Zona de aceptación de H0

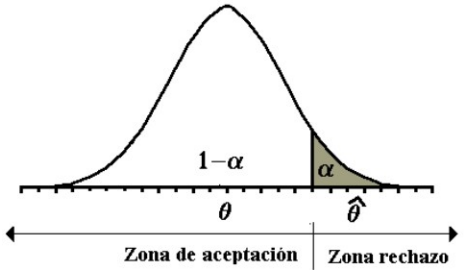
Zona de rechazo de H0

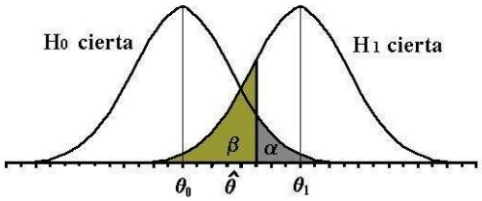
Procedimiento general para la PH

Nivel de significación: Errores de tipos I y II

	DECISIÓN	
CONDICIÓN REAL	Rechazar Ho	No Rechazar Ho
H_0 cierta	***Error (Tipo I)***	***Acierto***
H_0 falsa	***Acierto***	***Error (Tipo II)***







α nivel de significación

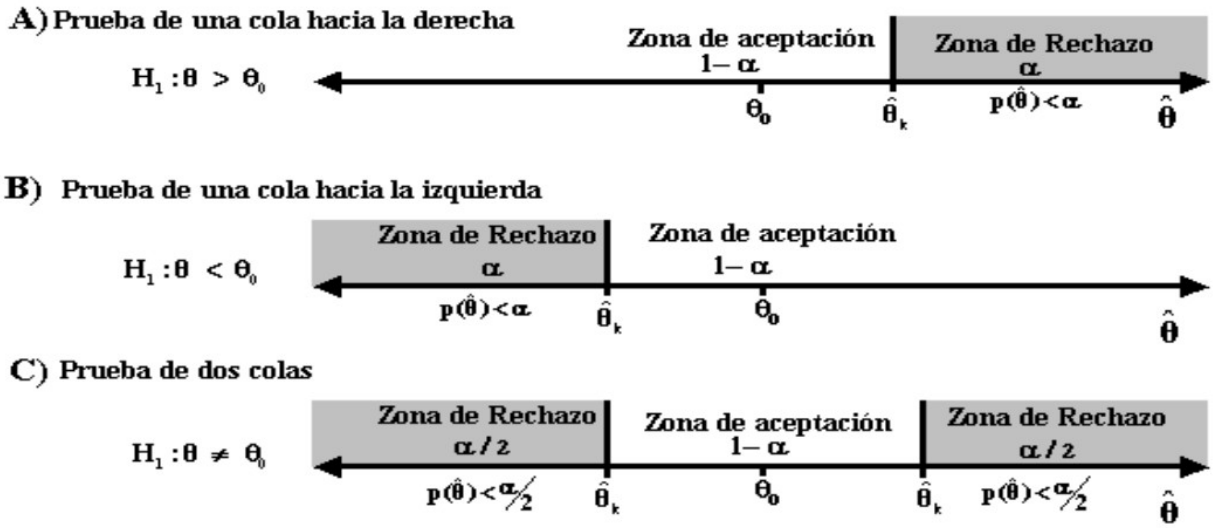
Procedimiento general para la PH

Estadístico de prueba

Parámetro	Estadístico de prueba	Estadísticos de prueba derivados
Media (μ)	$\bar{x}$	$z = (\bar{x} - \mu) / (\sigma / \sqrt{n})$ $z = (\bar{x} - \mu) / (s / \sqrt{n})$ $t = (\bar{x} - \mu) / (s / \sqrt{n})$
Diferencia de medias ($\mu_2 - \mu_1$)	$\bar{x}_2 - \bar{x}_1$	$Z = (\bar{x}_2 - \bar{x}_1) - (\mu_2 - \mu_1) \Big/ \sqrt{\frac{\sigma_2^2}{n_2} + \frac{\sigma_1^2}{n_1}}$ $Z = (\bar{x}_2 - \bar{x}_1) - (\mu_2 - \mu_1) \Big/ \sqrt{\frac{s_2^2}{n_2} + \frac{s_1^2}{n_1}}$ $T = (\bar{x}_2 - \bar{x}_1) - (\mu_2 - \mu_1) \Big/ \sqrt{\frac{s_2^2}{n_2} + \frac{s_1^2}{n_1}}$
Varianza	S^2	$\chi^2 = (n-1) S^2 / \sigma_0^2$
Razón de varianzas	S_2^2 / S_1^2	$F = (s_2^2 \sigma_2^2) / (s_1^2 \sigma_1^2)$

Procedimiento general para la PH

Zona de aceptación



$\alpha = 0{,}100$	$\pm z_{(0{,}900)} = 1{,}29$	$\pm t_{(0{,}90;\ 10)} = 1{,}372$
$\alpha = 0{,}050$	$\pm z_{(0{,}950)} = 1{,}65$	$\pm t_{(0{,}95;\ 10)} = 1{,}812$
$\alpha = 0{,}025$	$\pm z_{(0{,}975)} = 1{,}96$	$\pm t_{(0{,}975;\ 10)} = 2{,}228$
$\alpha = 0{,}010$	$\pm z_{(0{,}990)} = 2{,}33$	$\pm t_{(0{,}99;\ 10)} = 2{,}764$

Procedimiento general para la PH

Cómputos necesarios

Muestra de tamaño *n*

el valor medio muestral

la desviación estándar

el valor crítico

zona de aceptación

Procedimiento general para la PH

Decisión

Si el estadístico de prueba cae dentro de la región de rechazo, se considera que la diferencia entre el parámetro y el estadístico de prueba es significativa, y por lo tanto se rechaza H_0.

Si por el contrario el estadístico de prueba se ubica en la zona de aceptación se considera que la diferencia entre el parámetro y el estadístico de prueba no es significativa, en consecuencia se puede aceptar H_0 planteada.

Procedimiento general para la PH

Conclusión

Hipótesis de investigación

Hipótesis estadística

Conclusión estadística

Conclusión de investigación

PH: planteo de hipótesis

Cómo decidir la manera correcta de plantear las hipótesis?

Cómo decidir cuál es la hipótesis nula y cuál es la hipótesis alternativa?

PH: planteo de hipótesis

Algunos ejemplos:

Supongamos que debemos realizar un estudio sobre la altura media de los habitantes de cierto pueblo de España. Antes de tomar una muestra, lo lógico es hacer la siguiente suposición a priori, (hipótesis que se desea contrastar y que denotamos H0):

H0 : La altura media no difiere de la del resto del país.

Al obtener una muestra de tamaño n = 8, podríamos encontrarnos ante uno de los siguientes casos:

1. Muestra = {1,50 ;1,52; 1,48; 1,55; 1,60; 1,49; 1,55; 1,63}

2. Muestra = {1,65; 1,80; 1,73; 1,52; 1,75; 1,65; 1,75; 1,78}

En una Prueba de hipótesis se decide si la hipótesis nula puede ser rechazada o no a la vista de los datos suministrados por una muestra de la población. Para realizar el contraste es necesario establecer previamente una hipótesis alternativa que será admitida cuando la hipótesis nula sea rechazada. Normalmente, H1 es la negación de H0, aunque esto no es necesariamente así.

PH: planteo de hipótesis

Algunos ejemplos:

En los procesos judiciales donde hay alguien acusado de un delito, hay dos hipótesis: inocente (H0) y culpable (H1).

El fiscal público tiene interés en probar que el acusado es culpable. Para poder llegar a una decisión de culpable es necesario presentar suficientes evidencias que garanticen que la decisión es correcta.

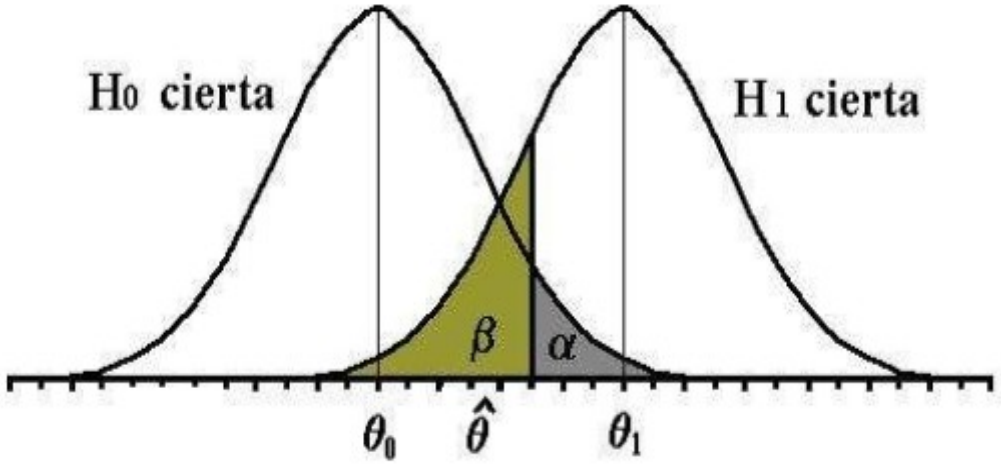
De no tenerse evidencias fuertes la hipótesis nula de inocencia no puede ser rechazada, pero esto no signica que se comprobó la inocencia del acusado, sino que no se logró acumular suficientes elementos para rechazar H0.

De hecho es posible que con nuevas investigaciones se determine la culpabilidad del acusado. Por el contrario habiéndose obtenido fuertes evidencias de culpabilidad, se acepta la hipótesis alternativa, decisión que es mucho más difícil de revertir.

El rechazo de H0 es un veredicto mucho más robusto que su no rechazo, puesto que es necesario acumular evidencia científica muy fuerte para poder rechazar una hipótesis nula. Por lo tanto la consecuencia de rechazar una hipótesis nula es un gran apoyo a la hipótesis alternativa.

PH: planteo de hipótesis

Si la región de no rechazo y la región de rechazo son complementarias, no se pueden disminuir los errores tipos I y II al mismo tiempo para un tamaño fijo de la muestra. Cuando uno decrece, el otro aumenta.



No es posible encontrar tests que hagan tan pequeños como queramos ambos errores simultáneamente. De este modo es siempre necesario privilegiar a una de las hipótesis, de manera que no será rechazada, a menos que su falsedad se haga muy evidente. En general, la hipótesis privilegiada es H0 que sólo será rechazada cuando la evidencia de su falsedad supere el umbral especificado (1-α)

PH: planteo de hipótesis

Al tomar α muy pequeño tendremos que β se puede aproximar a uno. Lo ideal a la hora de definir un test es encontrar un compromiso satisfactorio entre α y β (aunque siempre a favor de H0).

En el momento de elegir una hipótesis privilegiada podemos en principio dudar entre si elegir una dada o bien su contraria. Criterios a tener en cuenta en estos casos son los siguientes:

Simplicidad Científica

A la hora de elegir entre dos hipótesis científicamente razonables, tomaremos como H0 aquella que sea más simple.

Evaluar las consecuencias de una mala elección

Por ejemplo al juzgar el efecto que puede causar cierto tratamiento médico que está en fase de experimentación, en principio se ha de tomar como hipótesis nula aquella cuyas consecuencias por no rechazarla siendo falsa son menos graves, y como hipótesis alternativa aquella en la que el aceptarla siendo falsa trae peores consecuencias.

PH: planteo de hipótesis

Algunos ejemplos:

H0 : el paciente empeora o el tratamiento no tiene efecto
H1 : el paciente mejora

H0 : el ascensor se caerá
H1 : el ascensor funciona correctamemte

En este caso conviene esperar a que el ascensor sea usado muchas veces (n grande) para rechazar la hipótesis nula con alta significancia estadística ($\alpha \sim 0$) aún cuando para ello haya que aceptar β grande. Un error tipo I implica ir al hospital, y un error tipo II implica subir las escaleras...

PH: planteo de hipótesis

Algunos ejemplos:

H0 : el acusado de un crimen es inocente
H0 : todos los porotos de una bolsa son blancos
H0 : no existe vida en otros planetas

A veces la aceptación de una hipótesis es solamente temporal. Si se reúne mayor evidencia (n grande) es posible que deje de ser posible la aceptación de H0.
- obtengo más evidencias para el juicio
- miro mayor cantidad de porotos
- estudio mayor cantidad de planetas

No es lo mismo aceptar una hipótesis que afirmar la imposibilidad de rechazarla

PH: procedimiento de decisión

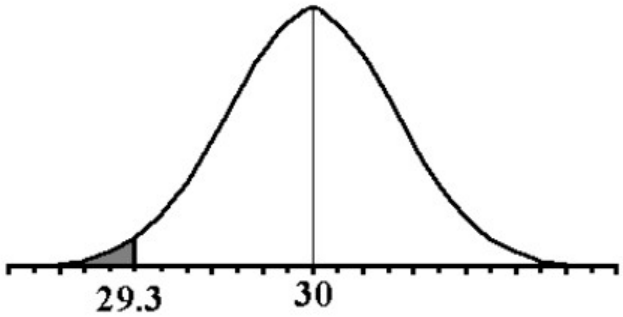
***Ejemplo*: con el propósito de determinar el efecto de una nueva dieta se forman varios lotes de 36 ratones con un peso aprox. 30g.**

a) El valor promedio de peso para cada grupo se considera una simple desviación fortuita de los 30 g.

b) Si el valor medido esta verdaderamente desviado del valor esperado 30 g.

H0: μ=30
supongamos medición con
un promedio de 29.3 con desviación 2 g.

$$P(\bar{X} \le 29.3) = P\left(Z \le \frac{29.3-30}{0.33}\right) = 0.0179$$



Modelo: (Z)

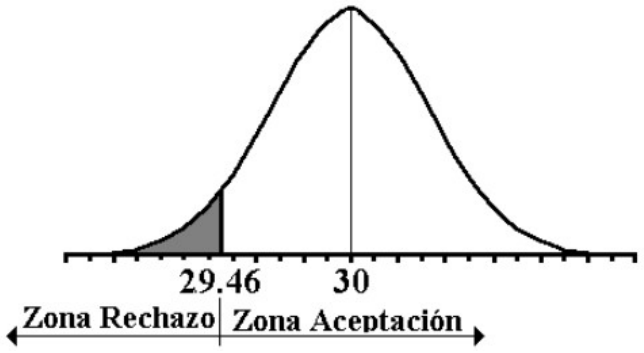
$$P(\bar{X} \le 29.9) = 0.382$$

$$P(\bar{X} \le 29.5) = 0.1151$$

$$P(\bar{X} \le \bar{x}) = 0.05 = P(\bar{Z} \le \bar{z})$$

Buscando en las tablas $z_{0.05}$

$$\bar{x} = \mu_x + z_{0.05} S_x/\sqrt{n} = 30 + (-1.64)2/\sqrt{36} = 29.46$$



PH para una media poblacional

PH para una media pobl. cuando la muestra proviene de una población distribuida normalmente y con varianza conocida

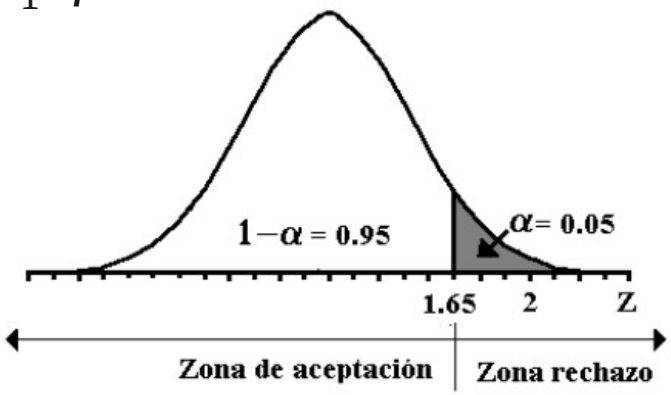
Ejemplo: Un médico afirma que el contenido de calcio en los huesos de mujeres que padecen osteoporosis, después de aplicarse cierto tratamiento es mayor al valor promedio observado, el cual es de 270 mg/g con una desviación de 120mg/g. Construyó una muestra de 36 individuos y obtuvo una media de 310mg/g.
La concentración de calcio se distribuye normalmente.

H_0: el tratamiento para la osteoporosis no tiene ningún efecto

H_1: el tratamiento para la osteoporosis aumenta los niveles de calcio en los huesos.

PH para una media pobl. cuando la muestra proviene de una población distribuida normalmente y con varianza conocida

1. **Formulación de la hipótesis:** $H_0: \mu = 270 \quad y \quad H_1: \mu > 270$
2. **Nivel de significación:** $\alpha = 0.05$
3. **Estadístico de prueba:** $Z = \dfrac{\bar{x} - \mu}{\sigma / \sqrt{n}}$
4. **Zona de aceptación:** $ZA = \left\{ Z / Z \leq z_{(1-\alpha)} \right\}$



5. **Cómputos necesarios:**

$$Z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} = \frac{310 - 270}{120 / \sqrt{36}} = \frac{40}{20} = 2 \qquad ZA = \left\{ Z / Z \leq z_{(0.95)} \right\} = \left\{ Z / Z \leq 1.65 \right\}$$

6. **Decisión:** **Como Z=2 > $z_{0.95}$= 1.65, entonces esta dentro de la zona de rechazo de H_0.**

7. **Conclusión:** **Podemos afirmar que se tiene un 95% de confianza que el tratamiento aplicado a los pacientes enfermos de osteoporosis aumenta el nivel de calcio en los tejidos óseos.**

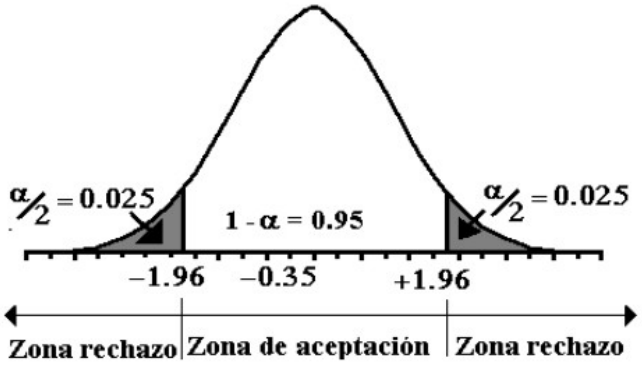
PH para una media pobl. cuando la muestra proviene de una población distribuida normalmente, con varianza desconocida y tamaño de muestra grande (n ≥ 30)

Ejemplo: Un médico entomólogo sospecha que en cierta zona endémica para el dengue el valor de la tasa reproductiva (R_0) ha cambiado en relación con el valor determinado hace 5 año, el cual era de 205 individuos. con un muestra de 40 hembras del mosquito, determinó R_0. La media de dicha muestra es 202.9 con una dispersión de 36.17. La variable se distribuye normalmente. Se quiere someter a PH no queriendo equivocarse en más del 5% de las veces.

H_0: la tasa neta de reproducción no ha cambiado.

H_1: la tasa neta de reproducción se modificó después de 5 años..

1. **Formulación de la hipótesis:** $H_0: \mu = 205 \quad y \quad H_1: \mu \neq 205$
2. **Nivel de significación:** $1-\alpha = 0.95$
3. **Estadístico de prueba:** $Z = \dfrac{\bar{x}-\mu}{s/\sqrt{n}}$
4. **Zona de aceptación:** $ZA = \left\{ Z / -z_{(1-\alpha/2)} < Z < +z_{(1-\alpha/2)} \right\}$



5. **Cómputos necesarios:**

$$Z = \frac{\bar{x}-\mu}{s/\sqrt{n}} = \frac{202.9-205}{36.17/\sqrt{40}} = \frac{-2.1}{5.719} = -0.37$$

$$ZA = \left\{ Z / -z_{(0.975)} < Z < +z_{(0.975)} \right\} =$$

$$¿ \left\{ Z / -1.96 < Z < +1.96 \right\}$$

6. **Decisión:** Como Z=-0.37 se encuentra dentro de la zona de aceptación de H_0.

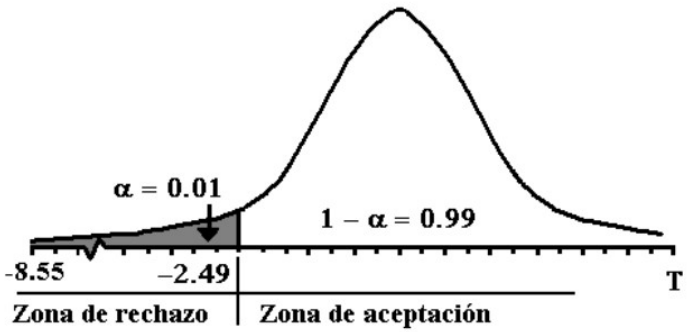
7. **Conclusión:** La sospecha del investigador fue rechazada con un 95% de confianza a la luz de la información proporcionada por la muestra.

PH para una media pobl. cuando la muestra proviene de una población distribuida Normalmente, con varianza desconocida y tamaño de muestra pequeño (n < 30)

Ejemplo: **Un fisiólogo desea verificar si el contenido de nitrógeno en las hojas jóvenes de la especia R. mangle, Es menor en las plantas que viven en zona ambientalmente protegida con relación a las que viven en un zona que está siendo afectada por la contaminación con fertilizantes y cuyo valor promedio es de 14.6 mg/g. El análisis de 25 hojas jóvenes de la zona protegida produjo una media muestral de 10.48 con una desviación estándar de 2.41. Si la concentración se distribuye normalmente, hay evidencia que las plantas tienen menos N_2? El error tipo I no debe ser mayor al 1%.**

H_0: la concentración de N_2 en las hojas de R. mangle en ambas regiones es la misma.
H_1: la concentración de N_2 en las hojas de R. mangle es menor en la región protegida.

1. **Formulación de la hipótesis:** $H_0: \mu = 14.6 \quad y \quad H_1: \mu < 14.6$
2. **Nivel de significación:** $1-\alpha = 0.99$
3. **Estadístico de prueba:** $T = \dfrac{\bar{x}-\mu}{s/\sqrt{n}}$
4. **Zona de aceptación:** $ZA = \left\{ T / -t_{(1-\alpha;n-1)} < T \right\}$



5. **Cómputos necesarios:**

$$T = \frac{\bar{x}-\mu}{s/\sqrt{n}} = \frac{10.48-14.6}{2.41/\sqrt{25}} = \frac{-4.12}{0.482} = -8.55 \qquad ZA = \left\{ T / -t_{(0.99;24)} < T \right\} = \left\{ T / -2.492 < T \right\}$$

6. **Decisión:** **Como t=-8.55<<-2.492, se encuentra dentro de la zona de rechazo de H_0.**

7. **Conclusión:** **Se puede afirmar con un 99% de confianza que la concentración de nitrógeno en las hojas de R. mangle en ambas regiones es diferente.**

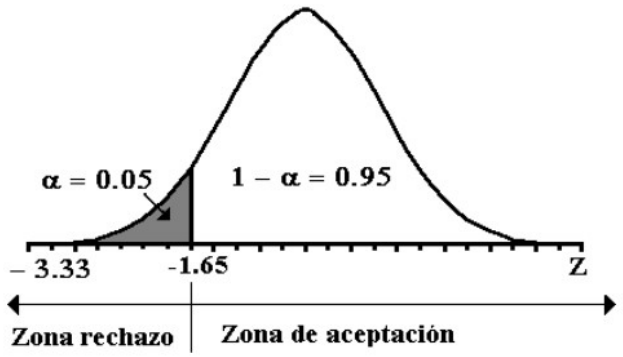
PH para una media pobl. cuando la muestra proviene de una población con distribución no normal y tamaño de muestra grande (n ≥ 30).

Ejemplo: En cierto nervio del cuerpo humano, los impulsos eléctricos viajan a una velocidad promedio de 4.3 m/s con una desviación de 1.2 m/s. Un fisiólogo observo que la veloc. promedio de conducción del impulso elect. en 45 individuos con una distrofia fue de 3.7 m/s. Se sospecha que el impulso eléctrico viaja a menor veloc., en el nervio estudiado, para personas enfermas con respecto a las sanas. Soporta ésta hipótesis los resultados obtenidos?

H_0: la veloc. del impulso nerv. es igual en los individuos con distrofia y en los normales.

H_1: la veloc. del impulso nerv. es menor en los individuos con distrofia que en los normales.

1. **Formulación de la hipótesis:** $H_0: \mu = 4.3 \quad y \quad H_1: \mu < 4.3$
2. **Nivel de significación:** $1-\alpha = 0.95$
3. **Estadístico de prueba:** $TLC \quad Z = \dfrac{\bar{x}-\mu}{\sigma/\sqrt{n}}$
4. **Zona de aceptación:** $ZA = \left\{ Z / -z_{(1-\alpha)} \leq Z \right\}$



5. **Cómputos necesarios:**

$$Z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} = \frac{3.7-4.3}{1.2/\sqrt{45}} = \frac{-0.6}{0.18} = -3.354 \qquad ZA = \left\{ Z / -z_{(0.95)} \leq Z \right\} = \left\{ Z / -1.65 \leq Z \right\}$$

6. **Decisión:** Como z=-3.354 < $-z_{0.95}$= -1.65, entonces esta dentro de la zona de rechazo de H_0.

7. **Conclusión:** Los datos soportan la suposición de que en los individuos con distrofia la veloc. de transmisión del impulso nervioso es menor a la observada en indiv. normales.

PH para dos medias poblacionales

PH para dos medias pobl. cuando las muestras provienen de poblaciones distribuidas normalmente y con varianza conocida.

$$H_0 : \mu_1 = \mu_2 \quad o \quad \mu_1 - \mu_2 = 0$$

$$H_1 : \begin{cases} \mu_1 \neq \mu_2 & o \quad \mu_1 - \mu_2 \neq 0 \\ \mu_1 > \mu_2 & o \quad \mu_1 - \mu_2 > 0 \\ \mu_1 < \mu_2 & o \quad \mu_1 - \mu_2 < 0 \end{cases}$$

PH para dos medias pobl. cuando las muestras provienen de poblaciones distribuidas Normalmente, con varianzas desconocidas y tamaños de muestras grandes (n_1,n_2 ≥ 30).

PH para dos medias pobl. cuando las muestras provienen de poblaciones distribuidas Normalmente, con varianzas desconocidas y tamaños de muestras pequeñas (n_1,n_2 ≥ 30).

PH para dos medias pobl. cuando las muestras proviene de poblaciones con distribución no normal y tamaño de muestras grandes (n_1,n_2 ≥ 30).

PH para dos varianzas poblacionales

Comparar mediante PH dos varianzas

Plantear la hipótesis

Tener un estadístico de prueba

$$H_0: \sigma_2^2 = \sigma_1^2 \quad o \quad H_0: \sigma_2^2/\sigma_1^2 = 1$$

$$H_1: \begin{cases} \sigma_2^2 \neq \sigma_1^2 & o \quad \sigma_2^2/\sigma_1^2 \neq 1 \\ \sigma_2^2 > \sigma_1^2 & o \quad \sigma_2^2/\sigma_1^2 > 1 \\ \sigma_2^2 < \sigma_1^2 & o \quad \sigma_2^2/\sigma_1^2 < 1 \end{cases}$$

$$F_0 = \frac{s_2^2}{s_1^2}$$

Distribución F

$$h(f) = d_1^{d_1/2} d_2^{d_2/2} \frac{\Gamma(d_1/2 + d_2/2)}{\Gamma(d_1/2)\Gamma(d_2/2)} \frac{f^{d_1/2-1}}{(d_1 f + d_2)^{d_1/2+d_2/2}}$$

para f>0, h(f)=0 para f≤0. Grados de libertad: $d_1=n_1-1$, $d_2=n_2-1$

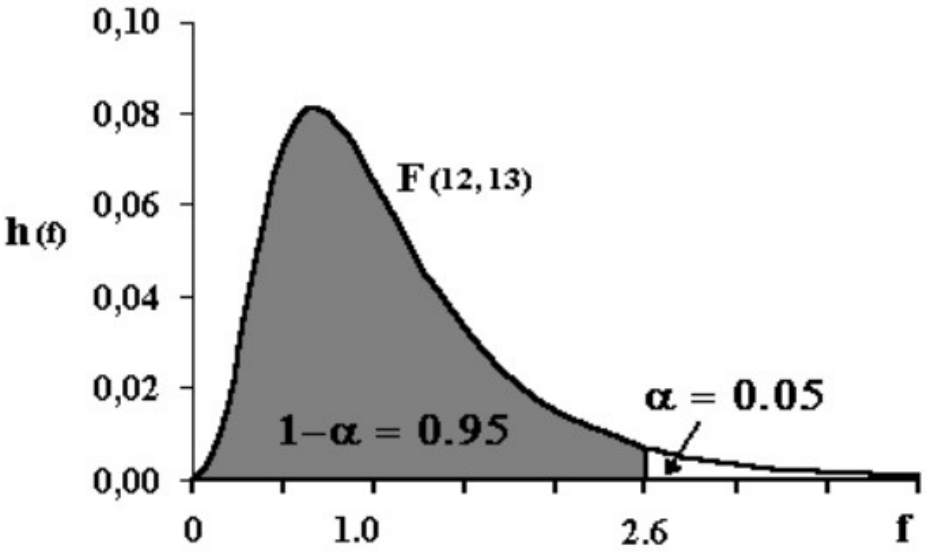
PH para dost varianzas poblacionales

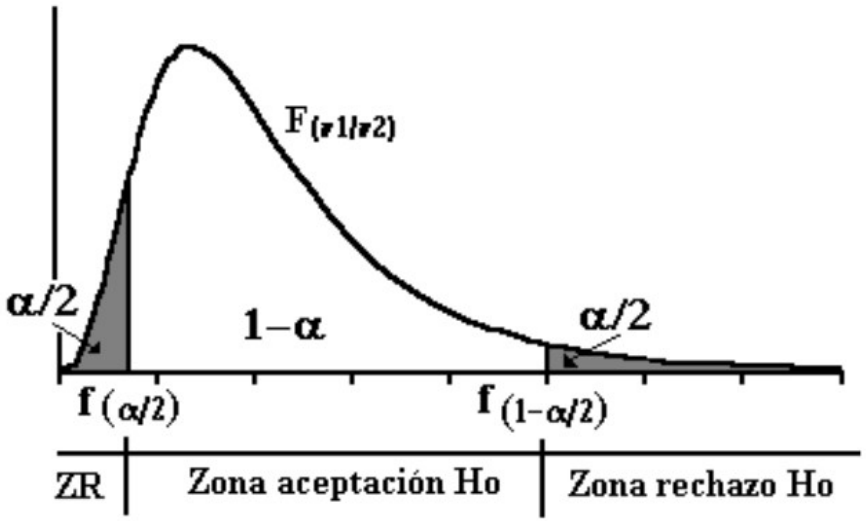
Ejemplo: $d_1=12$ $d_2=13$

0.95 área a la izq. de f=2.6

Si $F_0 < 2.6$ eso significa que su prob. de ocurrencia es > 0.05, entonces la dif. entre las dos varianzas es aleatoria.

Si F0 > 2.6 eso significa que su prob. de ocurrencia es < 0.05, entonces la dif. entre las dos varianzas no es aleatoria, son diferentes.





$$f_{(\alpha/2;d_1/d_2)} = \frac{1}{f_{(1-\alpha/2;d_2/d_1)}}$$

PH para dost varianzas poblacionales

Ejemplo: En un estudio taxonómico sobre una especie de insecto se quiere usar una característica morfológica del cuerpo para estimar el tamaño de los adultos. Se escogerá como carácterística aquella que tenga la menor variabilidad. Con éste proposito se midieron en 10 individuos la longitud del ala anterior y la longitud total del cuerpo. Con base en los resultados de la tabla, y sabiendo que las dos variables se distribuyen normalmente, esocja la que mejor estima el tamaño de los insectos.

Nº de Individuo	1	2	3	4	5	6	7	8	9	10
Alas anteriores (mm)	17,1	17	17,1	16,3	16,9	15,9	16,2	17,2	17,1	16,8
Tamaño del cuerpo (mm)	17,6	16,5	15,5	16,9	17,1	15,2	16,7	17,7	16,9	15,1

$$H_0 : \sigma_2^2 / \sigma_1^2 = 1 \quad y \quad H_1 : \sigma_2^2 / \sigma_1^2 \neq 1$$

PH para dost varianzas poblacionales

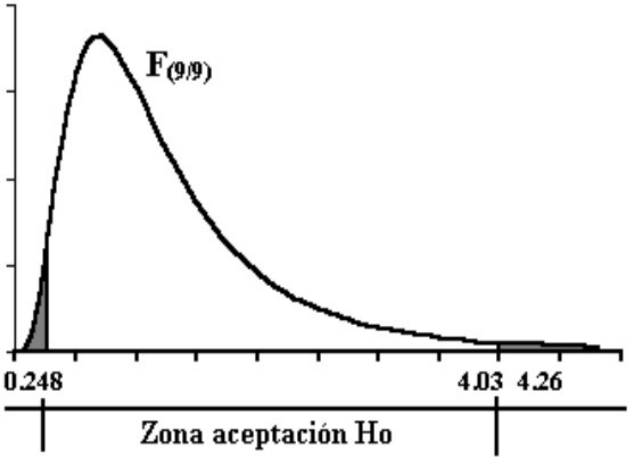
1. **Formulación de la hipótesis:** $H_0 : \sigma_2^2 / \sigma_1^2 = 1 \quad y \quad H_1 : \sigma_2^2 / \sigma_1^2 \neq 1$

2. **Nivel de significación:** $\alpha = 0.05$

3. **Estadístico de prueba:** $F_0 = \dfrac{s_2^2}{s_1^2}$

4. **Zona de aceptación:**

$$ZA = \left\{ F / -f_{(\alpha/2;\, n_2-1/n_1-1)} \leq F \leq f_{(1-\alpha/2;\, n_2-1/n_1-1)} \right\}$$



5. **Cómputos necesarios:**

$$F_0 = \frac{s_2^2}{s_1^2} = \frac{0.8907}{0.2093} = 4.26 \qquad f_{(0.025;9/9)} = \frac{1}{f_{(0.975;9/9)}} = 0.248$$

$$ZA = \left\{ F / -f_{(0.025;9/9)} \leq F \leq f_{(0.975;9/9)} \right\} =$$

$$¿ \left\{ F / -0.248 \leq F \leq 4.03 \right\}$$

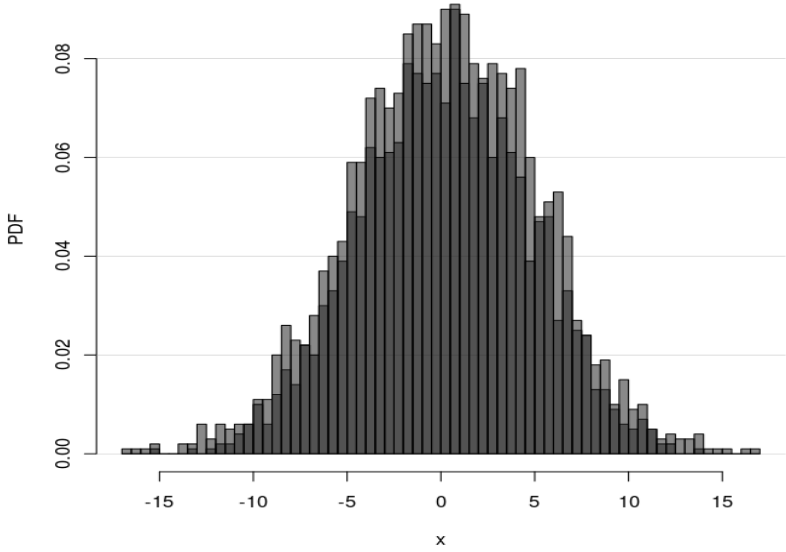
6. **Decisión:** **Como F_0=4.26 > 4.03, entonces esta dentro de la zona de rechazo de H_0.**

7. **Conclusión:** **Se puede afirmar con un 95% de confianza que las varianzas de las dos variables morfométricas son diferentes, siendo la longitud de las alas una variab. + homogénea.**

Otras aplicaciones de la PH:

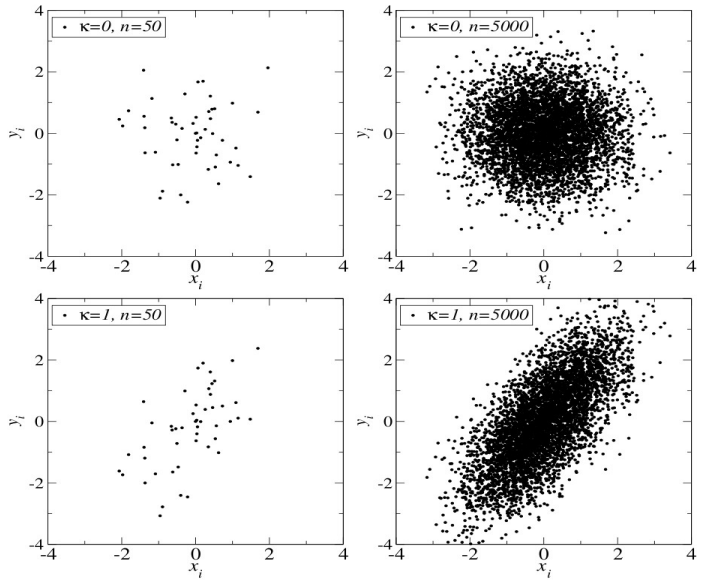
dos distribuciones son consistentes?

(chi2 bin., KS cont.)



Dos V.A. Son independientes?

(chi2, Coef. Correl. Pearson)



Método de Chi-cuadrado

Histogramas con distribuciones de probabilidades discretas

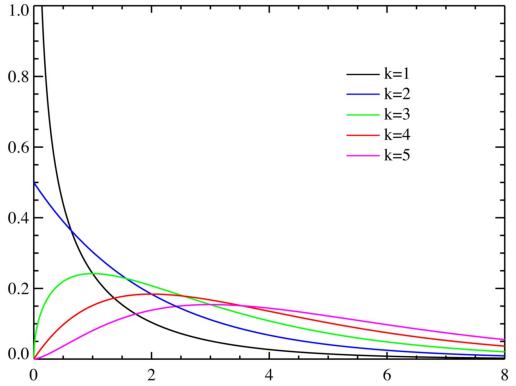
1. comparar un histograma con una función de prob. acum. discretizada.
2. comparar dos histogramas obtenidos de muestras diferentes.

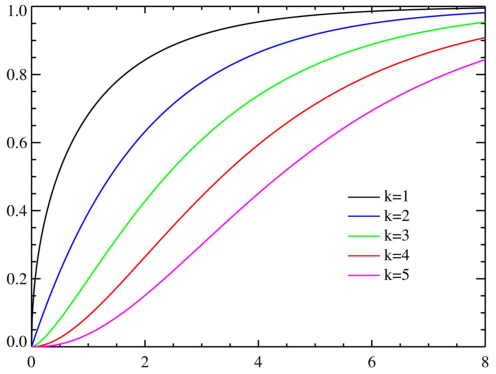
Caso 1: comparación de una muestra con una dist. teórica
La hipótesis nula es que la muestra tiene esa distrib. teórica.

$$\chi^2 = \sum_i \frac{\left(observada_i - teórica_i\right)^2}{teórica_i}$$

$$q(\chi^2,\nu) = \frac{1}{2^{\nu/2}\Gamma(\nu/2)}(\chi^2)^{(\nu/2)-1}e^{-\chi^2/2}$$

$$Q(\chi^2,\nu) = \frac{\gamma\left(\nu/2,\chi^2/2\right)}{\Gamma(\nu/2)}$$





$\chi^2 < \chi^2_{(1-\alpha,\nu)}$ H_0 no se rechaza (acepta)

Método de Chi-cuadrado

Ejemplo: supongamos que en una escuela las estadísticas de años pasados muestran que, la comisión tiende a aceptar 4 alumnos por 1 que rechaza. Este año una nueva comisión aceptó 275 y rechazó 55. Se puede decir que esta nueva comisión difiere de manera significativa con la razón de rechazo de la comisión anterior?

H0: no hay diferencias entre las comisiones.

330 alumnos en total

264 aceptados y 66 rechazados

$$\chi^2 = \frac{(275-264)^2}{264} + \frac{(55-66)^2}{66} = 0.4589 + 1.83 = 2.29$$

2x2 grados de libertad son ν=(filas-1)(columnas-1)=1x1=1

$$2.29 = \chi^2 < \chi^2_{(0.95,1)} = 3.841$$

Por lo tanto, la hipótesis nula no se rechaza, en consecuencia no hay diferencias entre la nueva comisión y la anterior.

Método de Chi-cuadrado

Caso 2: comparación de dos muestras
La hipótesis nula es que las dos muestras siguen la misma distribución.

$$\chi^2 = \sum_i \frac{\left(observada_{1,i} - observada_{2,i}\right)^2}{observada_{1,i} + observada_{2,i}}$$

Método de Kolmogorov-Smirnov

$$d_{\max} \equiv \max_{-\infty<x<\infty} \left| F_X(x) - F_{\hat{X}}(x) \right|$$

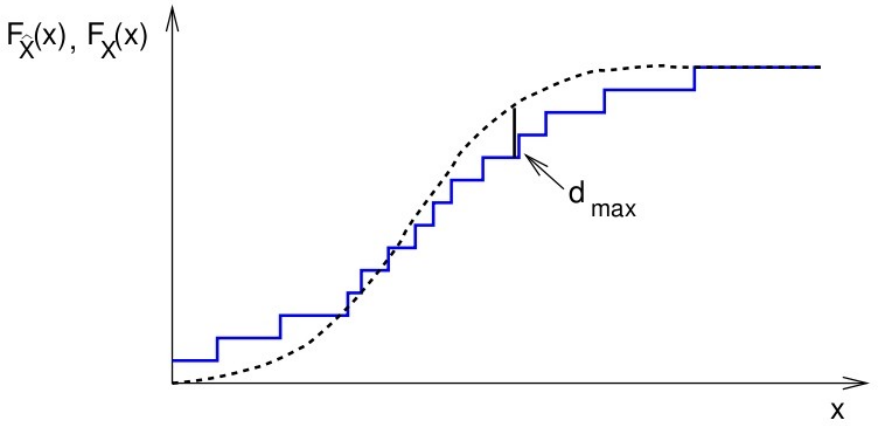
$$d_{\max} \equiv \max_{-\infty<x<\infty} \left| F_{\hat{X}_1}(x) - F_{\hat{X}_2}(x) \right|$$

$$Q_{KS}(x) = 2\sum_{j=1}^{\infty} (-1)^{j-1} e^{-2jx^2}$$

$$P(d_{\max} \le x) = 1 - Q_{KS}(x)$$

$$d_{\max}^{observ} > d_{\max}^{\alpha} \quad \textbf{rechaza } H_0$$

$$P(d_{\max} \le d_{\max}^{\alpha}) = 1 - \alpha$$



	Nivel de significación α							
n	**0.20**	**0.10**	**0.05**	**0.02**	**0.01**	**0.005**	**0.002**	**0.001**
1	0.90000	0.95000	0.97500	0.99000	0.99500	0.99750	0.99900	0.99950
2	0.68337	0.77639	0.84189	0.90000	0.92929	0.95000	0.96838	0.97764
3	0.56481	0.63604	0.70760	0.7 8456	0.82900	0.86428	0.90000	0.92065
4	0.49265	0.56522	0.62394	0.68887	0.73424	0.77639	0.82217	0.85047
5	0.44698	0.50945	0.56328	0.62718	0.66853	0.70543	0.75000	0.78137
6	0.41037	0.46799	0.51926	0.5774 1	0.61661	0.65287	0.69571	0.72479
7	0.38148	0.43607	0.48342	0.53844	0.57581	0.60975	0.65071	0.67930
8	0.35831	0.40962	0.45427	0.50654	0.54179	0.57429	0.61368	0.64098
9	0.33910	0.38746	0.43001	0.47960	0.51332	0.54443	0.58210	0.60846
10	0.32260	0.36866	0.40925	0.45562	0.48893	0.51872	0.55500	0.58042
***n* > 50**	$\frac{1.07}{\sqrt{n}}$	$\frac{1.22}{\sqrt{n}}$	$\frac{1.36}{\sqrt{n}}$	$\frac{1.52}{\sqrt{n}}$	$\frac{1.63}{\sqrt{n}}$	$\frac{1.73}{\sqrt{n}}$	$\frac{1.85}{\sqrt{n}}$	$\frac{1.95}{\sqrt{n}}$

Método de Kolmogorov-Smirnov

Ejemplo: Una investigación consiste en medir la altura de 100 niños de 5 años de edad. Se desea saber si las observaciones provienen le una población normal. El valor promedio de la muestra es 99.2 con desviación 2.85.

Serie de clases (talla en cm)	F	Fa
De 90 a 93	5	5
De 94 a 97	21	26
De 98 a 101	48	74
De 102 a 105	19	93
De 106 a 109	7	100
Total	100	

Hipótesis:

H_0: No hay diferencias entre los valores obs. y los teóricos de la distribución normal.

H_1: Los valores son diferentes.

Límites de clases	Valor Z de los límites	Area bajo la curva tipificada	Diferencias entre clases	Diferencias N (100) = F	Fa
90	-3.23	-0.4994			
93	-2.18	-0.4854	0.014	1.4	1.4
97	-0.77	-0.2794	0.206	20.6	22.0
101	0.63	0.2357	0.5151	51.5	73.5
105	2.04	0.4793	0.2436	24.4	77.9
109	3.44	0.4997	0.0200	2.0	99.9
Total				99.9	

Nivel de significación: $\alpha=0.05$

Zona de rechazo: $P > 0.05$ acepta H_0

$$d_{\max}^{observ} = 0.049$$

$$d_{\max}^{0.05} = \frac{1.36}{\sqrt{100}} = 0.136$$

$$d_{\max}^{observ} < d_{\max}^{\alpha}$$ No rechaza H_0

Rangos	1	2	3	4	5
f_t acumulada	1.4 / 100	22 / 100	73.5 / 100	97.9 / 100	99.9 / 100
f_{obs} acumulada	5 / 100	26 / 100	74 / 100	93 / 100	100 / 100
$f_t - f_{obs}$	-0.036	-0.04	-0.005	0.049	-0.001

Independencia Estadística

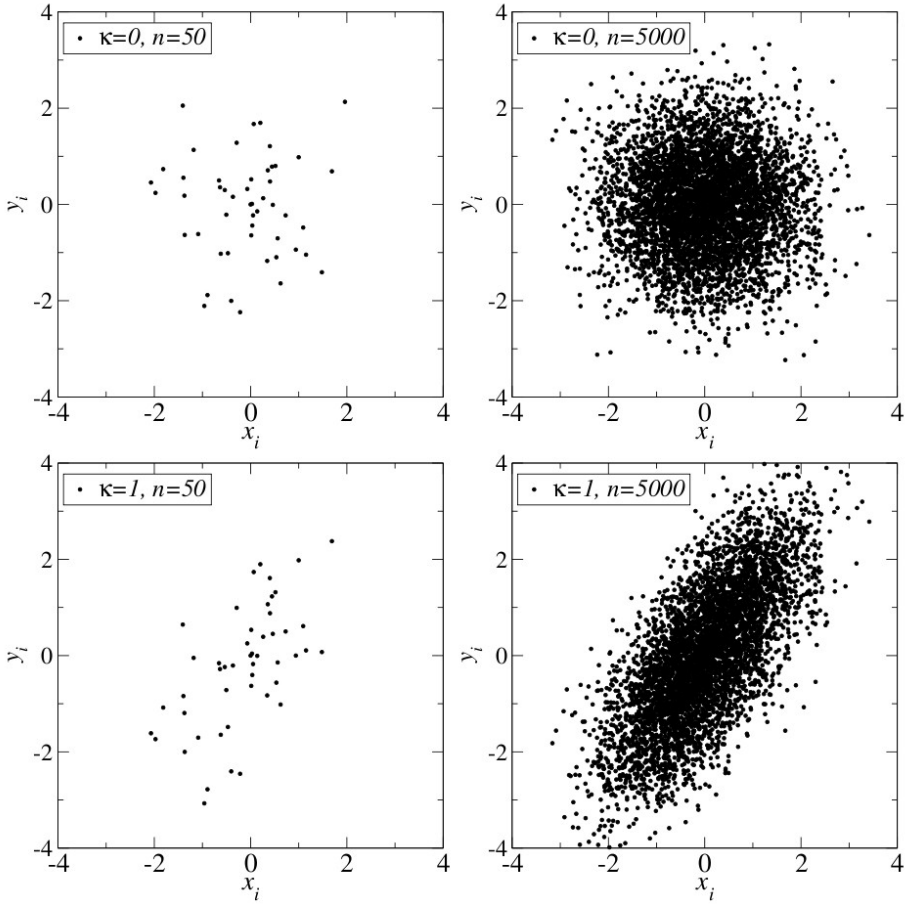
$$(x_i, y_i)$$

Cuando los valores de y_i dependerán de los valores de x_i?

Significancia Estadística

Potencia Estadística

Mientras más grande sea la potencia, más fácil será probar que la muestra es significativa.



Independencia Estadística

El método chi-cuadrado … el regreso

$$\{(x_0,y_0),(x_1,y_1),\ldots,(x_{n-1},y_{n-1})\} \Rightarrow \{h_{kl}\}$$

$$\hat{h}_k^{(x)}=\sum_l h_{kl} \quad ; \quad \hat{h}_l^{(y)}=\sum_k h_{kl} \Rightarrow \frac{\hat{h}_k^{(x)}}{n} \quad y \quad \frac{\hat{h}_l^{(y)}}{n} \Rightarrow n_{kl}=n\frac{\hat{h}_k^{(x)}}{n}\frac{\hat{h}_l^{(y)}}{n}=\frac{\hat{h}_k^{(x)}\hat{h}_l^{(y)}}{n}$$

$$\chi^2=\sum_{kl}\frac{(h_{kl}-n_{kl})^2}{n_{kl}}$$

$$\nu=k_xk_y-(k_x-1)-(k_y-1)+1$$

$$\nu=(k_x-1)(k_y-1)$$

$$p=1-Q(\chi^2,\nu)$$

$p<\alpha$ **La hipótesis nula será rechazada**

p(k=0, n=50) = 0.077
p(k=0, n=5000) = 0.457
p(k=1, n=50) = 0.140
p(k=1, n=5000) < 10^{-100}

Independencia Estadística

El coeficiente de correlación lineal de Pearson

$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2}\sqrt{\sum_i (y_i - \bar{y})^2}}$$

r(k=0, n=50) = 0.009
r(k=0, n=5000) = 0.009
r(k=1, n=50) = 0.653
r(k=1, n=5000) = 0.701

Independencia Estadística

Función de correlación

$$\hat{C}(\tau) = \frac{1}{n-\tau} \sum_{i=0}^{n-1-\tau} x_i x_{i+\tau} - \left(\frac{1}{n-\tau} \sum_{i=0}^{n-1-\tau} x_i \right) \times \left(\frac{1}{n-\tau} \sum_{i=0}^{n-1-\tau} x_{i+\tau} \right)$$

$$\hat{C}(\tau) = \frac{1}{n-\tau} \sum_{i=0}^{n-1-\tau} (x_i - \bar{x})(x_{i+\tau} - \bar{x})$$

$$C(\tau) = \frac{\hat{C}(\tau)}{\hat{C}(0)}$$