Universidad Panamericana

FACULTAD DE INGENIERÍA

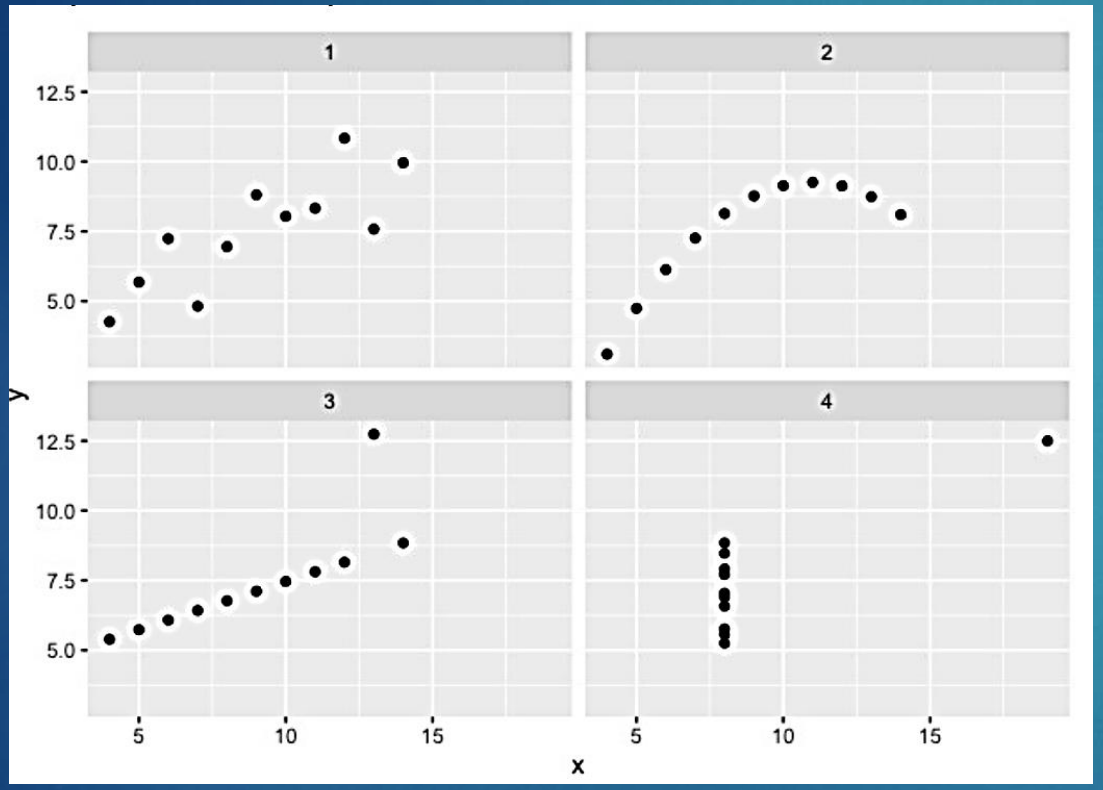
MAESTRÍA EN CIENCIA DE DATOS

Correlación entre variables

El análisis de correlación busca medir la fuerza de las relaciones entre dos variables por medio de un solo número denominado coeficiente de correlación.

Con frecuencia se supone que la distribución condicional $f(y|x)$ de Y, para valores fijos de X, es normal con media $\mu_{Y|x} = \beta 0 + \beta_1 x$ y varianza $\sigma_{Y|x}^2 = \sigma^2$, y que, de igual manera, X se distribuye de forma normal con media μ y varianza σ_x^2.

Identificar el tipo de relación o asociación entre dos variables, su dispersión y si existen datos que se comportan de manera atípica (*outliers*).



1.En el primer gráfico observamos una relación lineal, sin outliers, positiva y bastante fuerte.

2.En el segundo la relación no es lineal.

3.El tercer conjunto de datos muestra una relación lineal, positiva y muy fuerte, pero presenta un outlier que puede disminuir su valor de correlación porque escapa al comportamiento general.

4.En el último vemos que no existe realmente una relación entre “x” e “y”, sino que existe un outlier que puede confundir los resultados.

La correlación es un tipo de asociación entre dos variables, específicamente evalúa una tendencia (creciente o decreciente) en los datos.

Dos variables se correlacionan cuando muestran una tendencia creciente o decreciente.

La medida ρ de la asociación lineal entre dos variables X y Y se estima por medio del **coeficiente de correlación muestral** r, donde

$$r = b_1\sqrt{\frac{S_{xx}}{S_{yy}}} = \frac{S_{xy}}{\sqrt{S_{xx}\,S_{yy}}}.$$

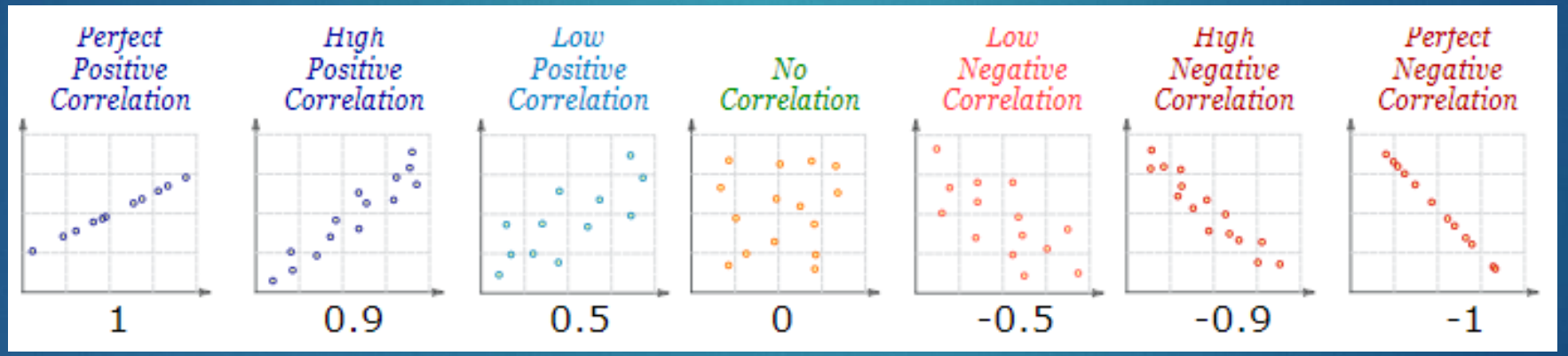
El coeficiente muestral de determinación r^2 expresa la proporción de la variación total de los valores de la variable *Y* que son ocasionados o explicados por una relación lineal con los valores de la variable aleatoria *X*.

$$r^2 = \frac{S_{xy}^2}{S_{xx}\,S_{yy}}$$

Se acostumbra hacer referencia al estimador ***r*** como coeficiente de correlación producto-momento de Pearson, o sólo como coeficiente de correlación muestral.

Coeficientes de correlación

- El coeficiente de correlación lineal de Pearson sirve para cuantificar tendencias lineales.
- El coeficiente de correlación de Spearman se utiliza para tendencias de aumento o disminución, no necesariamente lineales pero sí monótonas (las variables tienden a moverse en la misma dirección relativa, pero no necesariamente a un ritmo constante).



Coeficiente de correlación de Pearson

- Es el método de correlación **más utilizado, se asume que**:
  - la tendencia debe ser de tipo **lineal**.
  - no existen valores atípicos (**outliers**).
  - las variables deben ser **numéricas**. Si las variables son de tipo *ordinal,* no podremos aplicar la correlación de Pearson.
  - Se tienen **suficientes datos**(algunos autores recomiendan tener más de 30 puntos u observaciones).

Ejemplo

- **Se desea probar la efectividad de un medicamento sobre un grupo de pacientes. Para ello se han hecho mediciones en tres meses diferentes.**
  - **Verifique si los datos tienen distribución normal.**
  - **Si están correlacionados entre sí.**

Correlacion.r

Una **hipótesis estadística** es una aseveración respecto a una o más poblaciones

Prueba de Hipótesis

1. Hipótesis nula, H_0
2. Hipótesis alternativa, H_1
3. Estadístico de prueba y p
4. Región de rechazo
5. Conclusión

- Hipótesis nula, hipótesis que se desea desaprobar.
- El rechazo de H_0 conduce a la aceptación de una **hipótesis alternativa**,
- La hipótesis alternativa H_1 por lo general representa la *pregunta que se responderá o la teoría que se probará*.

- El rechazo de la hipótesis nula cuando es verdadera se denomina **error tipo I**.
- No rechazar la hipótesis nula cuando es falsa se denomina **error tipo II**.

	H_0 es verdadera	H_0 es falsa
No rechazar H_0	Decisión correcta	Error tipo II
Rechazar H_0	Error tipo I	Decisión correcta

Tipos de error

- La probabilidad de cometer un error tipo I, también llamada **nivel de significancia**, se denota con la letra griega α. Es el máximo riesgo tolerable de rechazar incorrectamente la hipótesis nula.
- La probabilidad de cometer un error tipo II, que se denota con β, es imposible de calcular a menos que se tenga una hipótesis alternativa especifica.

1. Los errores tipo I y tipo II están relacionados. Por lo general una disminución en la probabilidad de cometer uno da como resultado un incremento en la probabilidad de cometer el otro.
2. El tamaño de la región crítica y, por lo tanto, la probabilidad de cometer un error tipo I, siempre se puede reducir ajustando el (los) valor(es) crítico(s).
3. Un aumento en el tamaño de la muestra n reducirá α y β de forma simultánea.
4. Si la hipótesis nula es falsa, β es un máximo cuando el valor verdadero de un parámetro se aproxima al valor hipotético. Cuanto más grande sea la distancia entre el valor verdadero y el valor hipotético, más pequeña será β.

- La probabilidad de cometer ambos tipos de errores se puede reducir aumentando el tamaño de la muestra.
- La distribución muestral de la media muestral es aproximadamente normal cuando *n* es grande.

- Para determinar la probabilidad de cometer un error tipo I se utiliza la aproximación a la curva normal con:

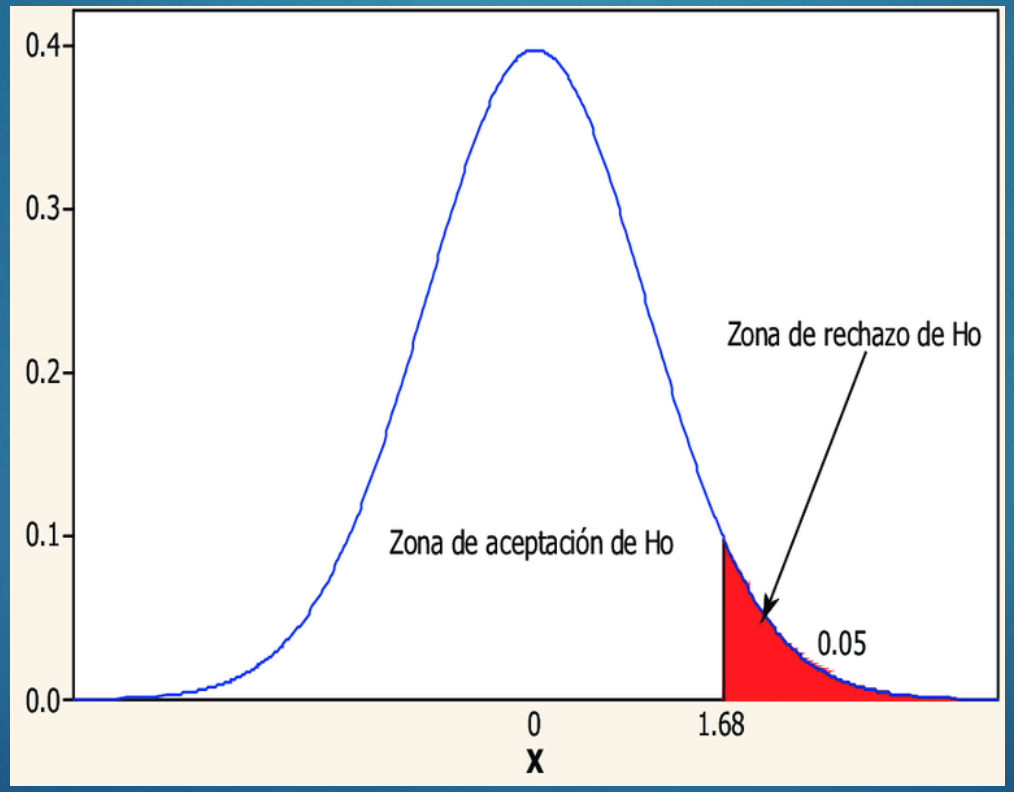
$$\mu_0 = np$$
$$\sigma = \sqrt{npq}$$

El número de desviaciones estándar a las que la media muestral está de μ_0 se pueden medir usando el estadístico estandarizado de prueba:

$$z = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

z tiene una distribución estándar normal aproximada cuando H_0 es verdadera y $\bar{x} = \mu_0$

- El nivel de significancia α es el área bajo la curva normal que se encuentra a la derecha del valor del estadístico de prueba z.



Ejemplo:

z= 1.68
α= 0.05

Ejercicio 1

El promedio semanal de ganancia para un grupo de trabajadoras es $670.00, ¿Los hombres de la misma posición tienen ganancias promedio más altas que las mujeres?

Una muestra aleatoria de n=40 trabajadores mostró un promedio de $725.00 y s=$102.00

Pruebe la hipótesis adecuada usando $\alpha = 0.01$

- Se quiere demostrar que el promedio semanal de ganancias para hombres es mayor que $670.00
- $H_0: \mu = 670$, $H_1: \mu > 670$
- $z = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{725-670}{102/\sqrt{40}} = 3.41$
- Región de rechazo: Los valores mayores a 670 llevarán al rechazo de H_0.

- Para controlar el riesgo de tomar una decisión incorrecta cuando el valor de α = 0.01, se establece el valor crítico que separe las regiones de rechazo y aceptación para que el área de la cola derecha sea exactamente 0.01

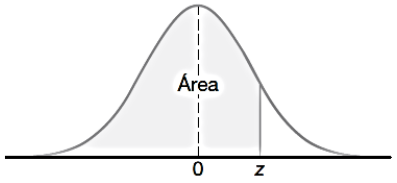
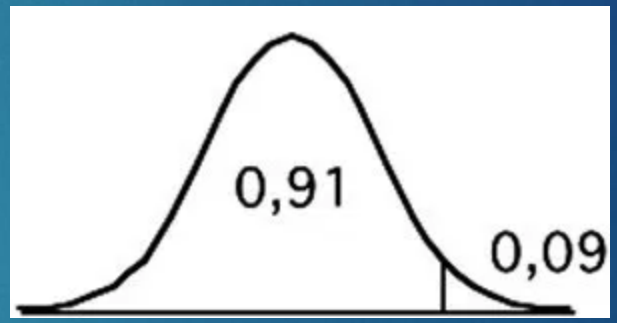


Tabla A.3 Áreas bajo la curva normal

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
−3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
−3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
−3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
−3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
−2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
−2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
−2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
−2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
−2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
−2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
−2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
−2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
−2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
−2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
−1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
−1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294

Hipotesis.r



Región de aceptación | Región de rechazo

Z=2.33

El estadístico de prueba z=3,41 es mayor que el valor crítico de rechazo z=2.33
Por lo tanto, se puede rechazar H0. El promedio semanal de ganancia para los hombres trabajadores es más alto que el de las trabajadoras de este problema.

Si se buscan desviaciones mayores o menores que μ_0, entonces la hipótesis alternativa es de dos colas.

- $H_1: \mu \neq \mu_0$
  - $\mu > \mu_0$
  - $\mu < \mu_0$

Ejercicio 2

- La producción diaria de una planta química ha promediado 880 toneladas en los últimos años. Al gerente de control de calidad le gustaría saber si el promedio ha cambiado en meses recientes. Se seleccionan al azar 50 días de los datos registrados, de donde el promedio de las *n*=50 producciones es $\bar{x} = 871$ toneladas y $\sigma = 21$ toneladas. Pruebe la hipótesis apropiada usando $\alpha = 0.05$
- H_0: $\mu = 880$, H_1: $\mu \neq 880$
- $z = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{871-880}{21/\sqrt{50}} = -3.03$
- Con $\alpha = 0.05$,

  $$\text{Los valores críticos que separan las regiones de rechazo y aceptación cortan}$$

  $$\text{áreas de aceptación de } \frac{\alpha}{2} = 0.025 \textit{ en las colas derecha e izquierda.}$$

  $$\textit{Los valores de } z = \pm 1.96$$
- H_0 se rechaza si $z > 1.96$ *o* $z < -1.96$
- 3.03<-1.96 , se rechaza la hipótesis nula H_0: $\mu = 880$
- Se concluye que el promedio ha cambiado
- La probabilidad (0.05) de rechazar H_0 cuando es verdadera es pequeña, por lo que se puede estar razonablemente seguro de que la decisión es correcta.

Potencia de una prueba estadística $(1-\beta)$

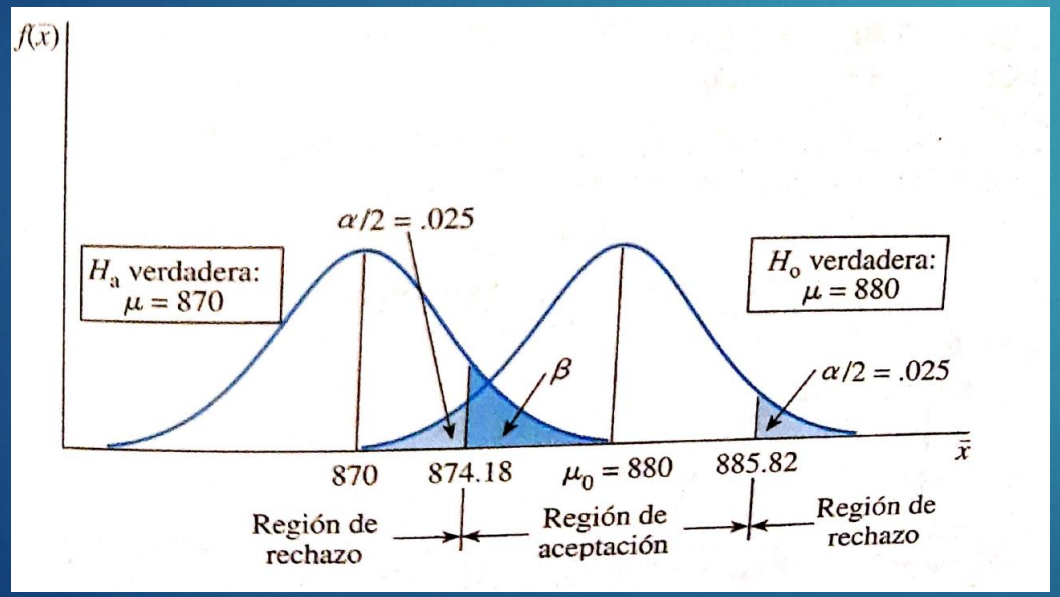
- La bondad de una prueba estadística se mide por el tamaño de las dos tasas de error: $\alpha \; y \; \beta$
- Una buena prueba es aquella para la que las dos tasas de error son pequeñas.

- Si β es la probabilidad de aceptar $H_o \; cuando \; H_1$ es verdadera (error tipo II).
- Entonces, una forma de evaluar una prueba es observar el complemento de un error tipo II: $1-\beta = P(rechazo \; de \; H_o \; cuando \; H_1 \; es \; verdadera)$
- $(1-\beta)$ mide la capacidad de la prueba para funcionar como se requiere.
- Idealmente se tendría significancia pequeña y potencia alta para una buena prueba estadística.

… Ejercicio 2

- Si se despeja $\bar{x}$ de $z = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \pm 1.96$
- $874.18 < \bar{x} < 885.82$
- Si se rechaza H_0 y se toma como $\mu = 870$
- Entonces β, *es el área bajo la curva normal del lado izquierdo, localizada entre* 874.18 *y* 885.82

La probabilidad de rechazar correctamente H_0, dado que $\mu = 870$, es 99.9999657%



$Potencia = 1 - \beta$ =3.431985e-05

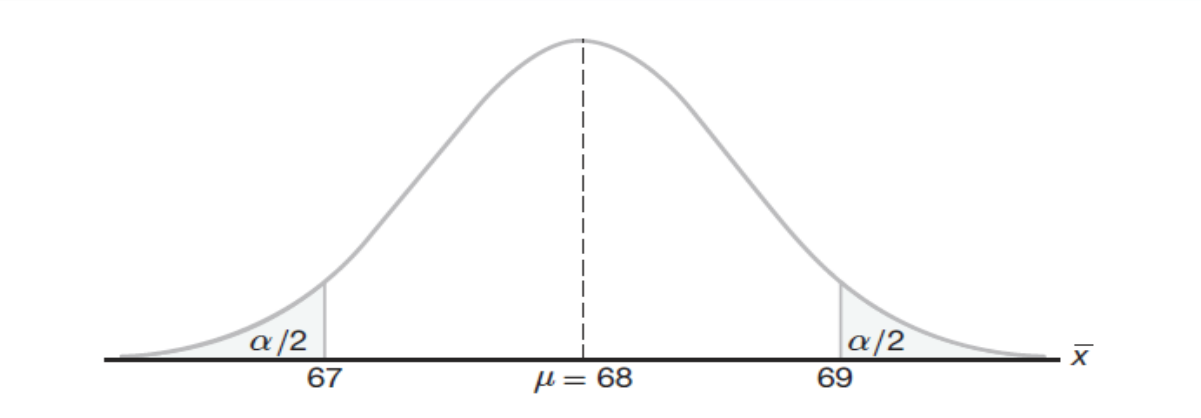
es la probabilidad de cometer error tipo II

	H_0 es verdadera	H_0 es falsa
No rechazar H_0	Decisión correcta	Error tipo II
Rechazar H_0	Error tipo I	Decisión correcta

Ejercicio 3

- Considere la hipótesis nula de que el peso promedio de estudiantes hombres en cierta universidad es de 68 kilogramos, contra la hipótesis alternativa de que es diferente a 68. Con $\sigma = 3.6$ y $n = 36$.
  - Para muestras grandes podemos sustituir s por σ si no disponemos de ninguna otra estimación de σ.
- $H_0: \mu = 68$, $H_1: \mu \neq 68$
- $\mu < 68$ o $\mu > 68$

- La probabilidad de cometer un error tipo I, o el nivel de significancia de la prueba, es igual a la suma de las áreas sombreadas en cada cola de la distribución.
- $\alpha = P(\bar{X} < 67\ cuando\ \mu = 68) + P(\bar{X} > 69\ cuando\ \mu = 68)$



Los valores z correspondientes a $\bar{x}_1 = 67$ y $\bar{x}_2 = 69$ cuando H_0 es verdadera son

$$z_1 = \frac{67 - 68}{0.6} = -1.67 \quad \text{y} \quad z_2 = \frac{69 - 68}{0.6} = 1.67.$$

Por lo tanto,

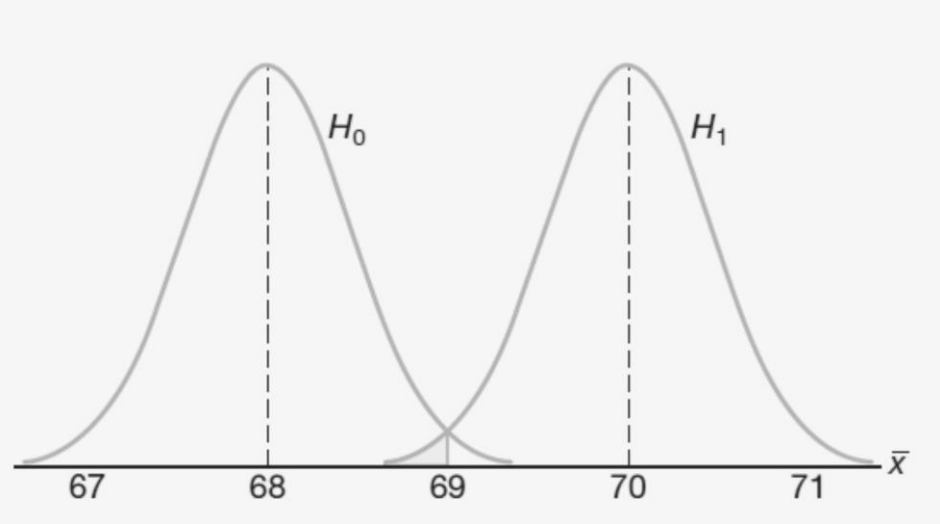
$$\alpha = P(Z < -1.67) + P(Z > 1.67) = 2P(Z < -1.67) = 0.0950.$$

Por consiguiente, 9.5% de todas las muestras de tamaño 36 nos conducirían a rechazar μ = 68 kilogramos cuando, de hecho, esta es verdadera. Para reducir α tenemos que elegir entre aumentar el tamaño de la muestra o ampliar la región de no rechazo.
Suponga que aumentamos el tamaño de la muestra a n = 64. Entonces $\sigma_{\bar{x}} = 3.6/8 = 0.45$.

$$z_1 = \frac{67 - 68}{0.45} = -2.22 \quad \text{y} \quad z_2 = \frac{69 - 68}{0.45} = 2.22$$

$\alpha = 0\ .0264$

Encontremos $\beta = P(67 \leq \bar{X} \leq 69$ cuando $\mu = 70)$



β =0.0132, probabilidad de cometer error tipo II

Ejercicio 4.
Se obtuvieron durante 132 días las concentraciones máximas de ozono (en partes por 10^9) en una determinada zona de Nueva York.
Estados Unidos fija como requerimiento un nivel máximo de 120 de ozono.
De los 132 días, 2 días presentaron niveles de ozono por encima de 120.
Verifique si la proporción de días con nivel de ozono mayor que el permitido es menor o igual que 0.05 y calcule un intervalo de confianza al 95%.

Hipótesis nula es H_0:p>0.05
Hipótesis alternativa es H_1:p≤0.05

Un intervalo de confianza es un rango de valores (calculado en una muestra) en el cual se encuentra el verdadero valor del parámetro con una probabilidad determinada.
Nivel de confianza 1 − α: probabilidad de que el verdadero valor del parámetro se encuentre en el intervalo.

```
binom.test( x = 2, # los 2 días con niveles ozono superiores
n = 132, # el total de días, los 132
p = 0.05, alternative = "less", # en relación a la H.alternativa
conf.level = 0.95)
```

Con p = 0.03658 menor de 0.05 se rechaza la hipótesis nula H_0.
Se acepta la hipótesis alternativa H_1.
Por lo tanto, se concluye que los días con nivel de ozono mayor que el permitido es menor o igual que el 5%.

Ejercicio 5

Se desea saber si la media de un conjunto de valores normales x es diferente a 0.

H0 : μ = 0

H1 : μ <=0

```
norm <- c( 3.2005, 0.2608, 1.5324, 1.92, 1.4173, 0.0164, -0.9709, 1.8213 )
med <- mean(norm);
sd <- sd(norm)
c( med, sd )
```

Hipotesis2.r

Encuentre el valor de p y de una interpretación a ese valor

```
tstat <- (med - 0) / (sd/sqrt(8)) # estadístico t
gl <- length(norm) - 1 # grados de libertad
tstat; gl

pval <- 2 * pt( -abs(tstat), gl ) # p-valor, función de distribución
pval
```

Ejercicio 6.
Se desea probar la efectividad de un medicamento sobre un grupo de pacientes. Para ello se han hecho mediciones en tres meses diferentes.

¿Tienen distribución normal los datos?

a) Contrastar, con nivel de significación α=0.05, si la media de los valores en el mes inicial m0 es 43.

- La normalidad se puede visualizar con los gráficos Q-Q (qqnorm(), qqline()).
- Para contrastarla se puede utilizar:
  - El test de Shapiro-Wilk con shapiro.test().

  Funciona bien con muestras pequeñas (menores a 50)

  - El test de Kolmogorov-Smirnov con ks.test().

  Contrasta distribuciones (no sólo la normal)

  - Corrección de Lillefors en KS, lillie.test() del paquete nortest.

Contraste de hipótesis
Hipótesis nula es H_0:μ=43
Hipótesis alternativa es H1:μ≠43

- La *prueba t para una muestra* se utiliza cuando se tiene una variable de medida y un valor esperado para la media y se supone normalidad de los datos (o muestra grande).

b) ¿Debemos aceptar o rechazar la diferencia de la media del mes inicial m0 según el género, para α=0.05?
Estamos ante un contraste para dos muestras independientes (hombres y mujeres). Para dos muestras independientes se debe comprobar el supuesto de normalidad y el supuesto de homocedasticidad.
Después se realiza el contraste sobre lo que queremos probar, en nuestro caso si la media de los hombres es distinta de la media de las mujeres para el mes inicial.

SUPUESTO DE HOMOCEDASTICIDAD (homogeneidad de varianzas).
En el contraste de homogeneidad de varianzas la hipótesis nula tiene varianza es constante (no varía) en los diferentes grupos.
Para contrastarla podemos utilizar el test F de Snedecor con var.test( ), que se aplica cuando sólo hay dos grupos.

CONTRASTE DE HIPÓTESIS: Se supone normalidad y homocedasticidad u homogeneidad de varianzas, podemos realizar nuestro contraste.

Definimos nuestras hipótesis.

Queremos probar si la media de los hombres es distinta de la media de las mujeres para el mes inicial. :

Hipótesis nula es $H0:\mu H = \mu M$

Hipótesis alternativa es $H1:\mu H \neq \mu M$

c) Los investigadores afirman que hay diferencia entre los valores tomados en el mes inicial m0 y en el tercer mes m3. ¿Tienen razón?

Para dos muestras dependientes se debe comprobar el supuesto de normalidad. Después se realiza el contraste sobre lo que queremos probar

CONTRASTE DE HIPÓTESIS: Se quiere probar si la media de los valores en el mes inicial m0 es distinta de la media en el tercer mes m3.

Hipótesis nula es $H0:\mu m0 = \mu m3$

Hipótesis alternativa es $H1:\mu m0 \neq \mu m3$

Ejercicio de clase 1

El conjunto de datos cancer.csv contiene los resultados de un estudio que mide las capacidades orales de enfermos de cáncer de garganta. Las medidas están tomadas inicialmente y a las 2, 4 y 6 semanas de tratamiento.

Además las variables edad, peso inicial y estado inicial del cáncer fueron medidas para cada paciente.

En el hospital, a un grupo se le administra un placebo (0) y al otro un tratamiento (1).

a. Determine si hay diferencias en la media de la capacidad oral de los enfermos en la segunda semana

(Variable TOTALCW2) según el grupo de edad (AGE) al que pertenece el enfermo

b. Determine si hay diferencias en la media de la capacidad oral de los enfermos en la segunda semana

(Variable TOTALCW2) según el tratamiento que se le da (TRT)

Ejercicio de clase # 2

Se ha lanzado un dado 275 veces, de las que 60 ha salido el 6.
Si el dado no está cargado, esperamos que el 6 salga 275/6 = 45.8333333 veces. ¿Es razonable pensar con 95% de confianza que el dado no esté trucado?

Ejercicio de clase # 3

Se estima que la proporción de adultos que vive en una pequeña ciudad que son graduados universitarios es p = 0.6. Para probar esta hipótesis se selecciona una muestra aleatoria de 15 adultos. Si el número de graduados en la muestra es cualquier número entre 6 y 12, no rechazaremos la hipótesis nula de que p = 0.6; de otro modo, concluiremos que p ≠ 0.6.

a) Evalúe α suponiendo que p = 0.6. Utilice la distribución binomial.
b) Evalúe β para las alternativas p = 0.5 y p = 0.7.
c) ¿Es éste un buen procedimiento de prueba?

Ejercicio de clase #4

El archivo notasTest.csv contiene calificaciones de dos exámenes (inicial y final) aplicados en dos clases diferentes. Se quiere saber si:

a. La media de las notas del examen inicial es 22.
b. La media de la clase A es distinta de la de la clase B en el examen inicial.
c. La media de la clase A se modifica en el examen final respecto del inicial.
d. La media de la clase B se modifica en el examen final respecto del inicial.
e. La media mejora en el examen final respecto del inicial.