3. Teoría de Muestras y diseño.

3.1. Muestras de una población. Teoría de la estimación.

Como se comentó a principio de curso, la estadística estudia como se presentan uno o varios caracteres dentro de una población. En su apartado descriptivo la estadística estudia estos caracteres desde un punto de vista estático. Se limita e gráficas de los mismos y a medidas de centralización y dispersión.

Gracias a las probabilidades podemos dar respuestas a problemas más ambiciosos. Como estudiar las posibilidades de que individuos de una población tengan el valor de un carácter dentro de un intervalo de valores. Resolver hipótesis a cerca de valores de un parámetro. Ver si en dos poblaciones distintas un carácter tenga el mismo valor para ambas. Y otros muchos problemas.

Normalmente para ver el valor que toma un carácter dentro de una población hay que estimar su valor, y hacerlo de manera que sea fiable, además esto se suele hacer a través de una muestra de la población.

El ver como se hace la elección de una muestra, y como se estima un valor a través de ella corresponde a la teoría de muestras y a la inferencia estadística.

3.1.1. Poblaciones y muestras.

Para hacer un estudio estadístico a veces es posible tomar la población entera, así, para estudiar los resultados de un examen realizado por 30 alumnos, podemos tomar las notas de los 30 alumnos. Pero a veces no es posible, o no interesa tomar la población entera. Por ejemplo: Si queremos estudiar la intención de voto de los españoles, no tomaremos todos los españoles, simplemente haremos una encuesta entre un grupo de españoles.

Por lo tanto, en muchas ocasiones no tomamos la población entera, sólo tomamos una muestra o parte de la población, estudiamos la característica en dicha muestra, e inferimos los resultados en toda la población. Por ejemplo: Para estudiar la altura media de los españoles tomamos una muestra de 5000 individuos, calculamos la media de su altura, y este resultado lo inferimos en la media global de toda la población.

Podemos resumir diciendo:

Población. Conjunto de individuos o cosas que vamos a estudiar.
Característica. Carácter que vamos a estudiar en la población.
Muestra. Parte de la población donde se hará el estudio de la característica, dicho estudio nos servirá para inferir sus resultados en toda la población.

La necesidad de escoger una muestra puede ser por varios motivos:

- Población muy numerosa.
- Proceso de medición destructivo.
- Rapidez en la necesidad de conocer resultados.

- Imposibilidad de conocer o determinar la población entera.
- Otras causas.

A continuación veremos métodos y formas de escoger las muestras.

3.1.2. Elección de la muestra

Elegir una muestra debe de ser un proceso riguroso y meditado. Si la muestra está mal elegida la inferencia de los resultados no es válida. Por el contrario, una buena elección, nos darán resultados muy exactos.

La muestra nunca se obtendrá sesgada. Por ejemplo: No tiene sentido tomar una muestra consistente en jugadores de baloncesto para estudiar la altura de la población de un país. Tampoco tendría sentido estudiar el peso de la población japonesa tomando una muestra formada por luchadores de sumo.

Por lo tanto la muestra debe de ser lo más independiente posible, y nunca elegirla tratando de llegar a un resultado.

Un buen muestreo tratará siempre de buscar una muestra lo más representativa posible, con una muestra bien elegida, podemos con pocos individuos inferir un resultado muy exacto.

Tampoco nos valen muestras muy pequeñas; cuanto más grande sea posible, más representativa será.

3.1.3. Muestro aleatorio.

Consiste en tomar los individuos de la muestra de forma aleatoria, o sea, al azar, de manera que todos los individuos de la población tienen la misma probabilidad de estar en una muestra. Esto se consigue normalmente mediante la elección por sorteo.

3.1.4. Muestreo aleatorio simple.

Es la simple elección por sorteo de los elementos de la muestra; sorteo que se puede realizar por cualquier método. Lo normal es asignar un número a cada individuo de la población, y luego sortear los individuos de la muestra.

3.1.5. Muestro aleatorio sistemático.

En este caso se numeran los individuos de la población, se elige al azar uno de ellos, los restantes se obtienen sumando una cantidad fija al número del primer individuo elegido. Esto podría hacerse así:

Supongamos que la población tiene N individuos, y nos interesa una muestra de tamaño n.
En primer lugar determinamos el "salto" o **coeficiente de elevación.**
Este se llama h y se define como h = N/n Si no es entero lo aproximamos por defecto.
A continuación elegimos un número al azar entre uno y h, le llamamos n_1. Es el primer

individuo de la muestra que tomamos. A continuación tomamos los demás de la forma siguiente:

$n_2 = n_1 + h$
$n_3 = n_2 + h = n_1 + 2h$
$n_4 = n_3 + h = n_1 + 3h$
.....
$n_n = n_1 + (n-1)h$

Veamos un ejemplo:

Si la población es de 1000 elementos y queremos una muestra de 100 individuos N=1000 y n=100. Entonces h = 1000/100 = 10

Tomamos al azar un número entre uno y 10, por ejemplo el 3, entonces:
$n_1 = 3$, $n_2 = 3 + 10 = 13$, $n_3 = 23$, ... $n_{100} = 3 + 99.10 = 993$.

Esto es, un sencillo ejemplo de muestreo aleatorio sistemático.

3.1.6. Muestro aleatorio estratificado.

Este se realiza si la población está dividida en capas, o intervalos, agrupados por una característica común. Por ejemplo: Estudiar la altura de una población, pero con los individuos agrupados en varios intervalos según su edad. Así tenemos, individuos entre 0-10 años, entre 10-20 años, entre 20-30 años ...

Podemos formular el problema de la siguiente forma:

La población está formada por N individuos, la población está formada por capas de N_1, N_2, ... N_k individuos. La muestra tiene un tamaño n. ¿Cuántos individuos tomamos de cada capa?.

La solución es muy sencilla, simplemente mantenemos la proporción, o sea, si n_1, n_2, ... n_k son los individuos que tomamos para la muestra en cada estrato, se verificará la igualdad proporcional:

$N/n = N_1/n_1 = N_2/n_2 = ... N_k/n_k$
De donde $n_1 = n\,(N_1/N)$ $n_2 = n(N_2/N)$ ..... $n_k = n(N_k/N)$

Veamos un ejemplo: Una población tiene 10000 individuos, N = 10000, está dividida en 5 estratos o capas de 100, 1000, 8000, 500, 400 individuos respectivamente. Queremos una muestra estratificada de 500 individuos.

Si los tamaños de la muestra en cada estrato son n_1, n_2, n_3, n_4, n_5, tendremos que:
$N/n = (10000/500) = (100/n_1) = (1000/n_2) = (8000/n_3) = (500/n_4) = (400/n_5)$. De donde:
$n_1 = 500(100/10000) = 5$ individuos.
$n_2 = 500(1000/10000) = 50$ individuos.
$n_3 = 500(8000/10000) = 400$ individuos.
$n_4 = 500(500/10000) = 25$ individuos.
$n_5 = 500(400/10000) = 20$ individuos.

Elegido el tamaño de los estratos, elegimos los individuos de cada estrato por un muestreo aleatorio simple, o sistemático, según creamos necesario.

3.1.7. Técnicas aleatorias.

Una vez fijado el tamaño de la muestra, y decidido el muestro aleatorio, nos preguntamos cómo realizar tal muestreo. Esto se puede llevar a cabo de varias formas:

- Numerar la población, introducir en una urna todos los números asignados e ir sacando **sin reemplazamiento los números asignados.** Es importante que sea sin reemplazamiento para no repetir un individuo dos veces.
- Numerar la población, y sacar los números aleatorios con una calculadora.
- Hacer un programa de ordenador que simule la extracción aleatoria sin reemplazamiento.
- Hacer un programa que simule el muestreo sistemático.

3.2. Inferencia estadística.

Una vez que tenemos la población, y que con el fin de estudiar uno o varios caracteres obtenemos una muestra de la misma, hemos de inferir o pronosticar los valores de dichos caracteres de la población a partir de los valores obtenidos en la muestra. La inferencia estadística trata este problema.

Señalemos que normalmente al dar el resultado de un parámetro estimado, lo haremos expresando un nivel de confianza que corresponde a las posibilidades de que el valor estimado se aproxime al valor real. Entramos pues a la teoría del cálculo de probabilidades.

3.2.1. Tipos de inferencia estadística.

En cuanto al objetivo del estudio podemos considerar dos tipos de inferencia:

1. **Métodos paramétricos.** Tenemos una población conocida y estimamos sus parámetros (normalmente media y desviación típica).
2. **Métodos no paramétricos.** Desconocemos la población y hacemos hipótesis sobre ella.

En cuanto al tipo de información, podemos clasificarla en:

1. **Inferencia clásica.** Semejante a los métodos paramétricos. Se supone unos parámetros fijos a estudiar a través de la muestra.
2. **Inferencia bayesiana.** Se supone que los parámetros son variables aleatorias.

Si estudiamos la inferencia clásica, podemos ver los siguientes métodos de llevarla a cabo:

Estimación. Es la forma más clásica de hacer la inferencia. Consiste en tomar una muestra, y haciendo cálculos con el valor del carácter a estudiar dentro de la muestra, inferir el valor en la población mediante un estadístico. O sea a través del parámetro muestral inferir el parámetro poblacional.

La estimación a su vez puede ser **puntual**, si sólo damos el valor posible del parámetro poblacional a través del parámetro muestral.

La estimación también puede ser por **intervalos,** que consiste en a través de un parámetro muestral de terminar un intervalo de valores donde puede estar el parámetro poblacional con cierta probabilidad de ello, probabilidad que llamamos nivel de confianza.

Contraste de hipótesis. Es ver y comprobar si diversas suposiciones sobre un parámetro poblacional son ciertas a través de los parámetros muestrales obtenidos. Por ejemplo ver si la media toma un valor. Ver si la media de dos poblaciones es la misma. Ver si la desviación típica se ajusta a un valor.

3.2.2. Estadístico y estimador puntual.

Supongamos una población en la que queremos estudiar un carácter, por ejemplo las alturas

de sus individuos. Para ello tomamos una muestra de la población.

Llamamos **estadístico** de la muestra a cualquier función realizada con los datos obtenidos de la muestra de la población. Puede ser cualquier función, como la suma de las alturas, la suma de los cuadrados de las alturas, u otro que se nos ocurra.

Llamamos **estimador** de un parámetro poblacional a un estadístico cuyos valores se aproximen a los valores de un determinado parámetro poblacional. Por ejemplo la media muestral se duele acercar a la media poblacional.

3.2.3. Principales estimadores puntuales.

Normalmente los parámetros que normalmente estimamos de una población son la media y la desviación típica, o bien si hacemos un experimento varias veces podemos estimar la proporción de que un determinado suceso ocurra. Entonces:

Si en una muestra $x_{1,}x_{2,}x_3....x_n$ son los valores que se presentan, para estimar la media de la población que llamamos μ utilizamos la media muestral $\bar{x}=\frac{1}{n}\sum_{i=1}^{n} x_i$.

Para estimar la varianza poblacional σ^2 empleamos la cuasivarianza muestral que llamamos s^2 mientras que la varianza muestral le llamamos S^2 las definiciones son:

$s^2=\frac{n}{n-1}S^2=\frac{n}{n-1}\sum_{n=1}^{n}\frac{(x_i-\mu)^2}{n}=\sum_{n=1}^{n}\frac{(x_i-\mu)^2}{n-1}$ Expresión que si queremos calcular con la forma reducida del calculo de la varianza podemos emplear:

$$s^2=\frac{n}{n-1}\left(\sum_{n=1}^{n}\frac{x_i^2}{n}-\bar{x}^2\right)$$

Si al hacer n veces un experimento se obtiene el suceso A k veces, el estimador de la proporción poblacional, será la proporción muestral. O sea:

$$\bar{p}=\frac{k}{n}$$

3.2.4. Medias muestrales. Teorema central del límite.

Supongamos una población, donde vamos a estudiar una característica. Dicha característica tiene una media μ y una desviación típica σ . En el apartado anterior hemos expuesto que si tomamos una muestra de la población, la media muestral y la cuasivarianza son buenos estimadores puntuales de la media y varianza poblacional.

El teorema central del límite va un poco más allá, pues afirma que si consideremos una variable aleatoria que sea los valores de las medias muestrales, esta variable también tiene una media μ y una desviación típica de $\sigma/\sqrt{n}$. Pero es más, si $n>30$ las medias muestrales siguen una variable aleatoria $N(\mu,\sigma/\sqrt{n})$ aunque la población no sea normal.

El enunciado completo sería así:

Supongamos una población, y que una característica de la misma presenta una **media μ y una desviación típica σ**. Esta población puede seguir, o no, el esquema de una Normal, no es pues necesario que sea una distribución Normal.

Si obtenemos un serie de muestras aleatorias de dicha población, se puede demostrar lo siguiente:

- La distribución de la media de las muestras tiene la misma media que la de la población, o sea, **μ.**
- La desviación típica de la distribución de las medias muestrales es $\frac{\sigma}{\sqrt{n}}$. Donde n es el tamaño de la muestra.
- La distribución de las medias muestrales es una Normal si la muestras son grandes, esto se cumple a partir de muestras con tamaño superior a 30 unidades.

Este teorema tiene suma importancia y nos va a permitir muchas cosas en el muestreo, así podemos señalar que:

- La distribución de la población puede ser cualquiera, continua o discreta.
- Aunque **n** no llegase a 30, en muchos casos podrían seguir las medias muestrales un esquema de distribución Normal.
- Si la distribución de partida es Normal, las medias muestrales siguen una normal, independientemente del tamaño de la muestra.
- Si n es muy grande, la desviación típica de la distribución de las medias muestrales, al ser , $\frac{\sigma}{\sqrt{n}}$ disminuye. Por lo que las medias de las muestras se acercan con más precisión a la media real.

Este teorema, tiene importantes consecuencias prácticas. Así:

- **Control de medias muestrales.**

Si en una población con media **μ** y desviación **σ**, tomamos medias de tamaño **n**, la media de estas muestras sigue una Normal $N(\mu, \sigma/\sqrt{n})$ Podemos calcula por ejemplo la probabilidad de que la media de una muestra esté comprendida en un intervalo determinado.

- Control de la suma de los valores de las muestras.

Puesto que las medias muestrales siguen una Normal de media **μ**, la suma de los valores de la muestra, o sea, $\sum_{i=1}^{i=n} x_i$ seguirá una normal con media **nμ**. y la desviación típica será $n\frac{\sigma}{\sqrt{n}} = \sigma\sqrt{n}$. Por tanto esta suma sigue una. $N\left(n\mu, \sigma\sqrt{n}\right)$

- Inferir la media de la población con la media de la muestra.

Es la aplicación más importante, a partir de una muestra se podrán obtener conclusiones de

la media de la población a partir de la media de la muestra. Se verá más adelante.

3.2.5. Otras consecuencias del Teorema central del límite.

Además de lo anteriormente expuesto también podemos obtener otras dos consecuencias bastante interesantes.

- Si tenemos una población de tipo Normal donde conocemos μ pero la desviación típica es desconocida, en este caso el estadístico:

  $\frac{\bar{x}-\mu}{\frac{s}{\sqrt{n}}}$ sigue una distribución t de student con $n-1$ grados de libertad.

- Si tenemos una población que sigue una normal, con σ conocida. El estadístico:

  $\frac{(n-1)s^2}{\sigma^2}=\frac{nS^2}{\sigma^2}$ sigue una $\chi^2(n-1)$ de Pearson, con $n-1$ grado de libertad.

3.2.6. Ejercicios.

3.2.6.1. Ejercicio 1.

Las bolsas de azúcar envasadas por cierta máquina tienen una media μ = 500 g, y una desviación típica σ = 35 g. Las bolsas se empaquetan en cajas de 100 unidades.

1. Calcular la probabilidad de que la media de los pesos de las bolsas en un paquete, sea menor de 495 g.
2. Calcular un intervalo característico de la media de los pesos de las bolsas para un nivel de confianza del 95% . Nivel de riesgo del 5%.
3. Calcular la probabilidad de que una de las cajas de 100 bolsas supere los 51 kg.

Apartado 1. La media de los pesos de las bolsas en un paquete de 100 unidades sigue una normal con media μ = 500 gramos, y $\sigma=\frac{35}{\sqrt{100}}=\frac{35}{10}=3{,}5$. O sea, una N(500,3,5). Entonces:

Si X representa la media muestral, nos piden P(X < 495). Tipificando X, tendremos que calcular:

$$p\left(\frac{X-500}{3{,}5}<\frac{495-500}{3{,}5}\right)=p(Z<-1{,}43)=1-p(Z<1{,}43)=1-0{,}9236=0{,}0764$$

Apartado 2. Para una Normal N(500,3,5) tenemos que calcular el intervalo característico con una probabilidad de $1-\alpha=0{,}95$ o nivel crítico del 5%. Determinamos $Z_{\alpha/2}$ para una N(0,1) nivel de riesgo 0,05. Entonces:

$p(Z<Z_{\alpha/2})=1-0{,}05/2=1-0{,}025=0{,}975$ y da un valor de $Z_{\alpha/2}=1.96$

Entonces el intervalo característico tiene por extremos:

$$\mu-Z_{\alpha/2}.\sigma=500-3{,}5.1.96=500-6{,}88=493{,}12$$
$$\mu+Z_{\alpha/2}.\sigma=500+3{,}5.1.96=500+6{,}88=506{,}88$$

Apartado 3. Como a suma de los pesos de las bolsas dentro de las cajas es multiplicar por $n=100$ la medie de los pesos, Esta suma seguirá una Normal $N(n.\mu, \sigma.\sqrt{n})$, o sea, una Normal con media = 100.500 g = 50kg. Y una desviación = 35.10g =350g. O sea N(50000g,350g). Nos piden dentro de esta normal cual será:

$$P(X>51000)=p(\frac{X-50000}{350}>\frac{51000-50000}{350})=p(Z>2{,}86)=1-p(Z<2{,}86)=0{,}0021$$

3.2.6.2. Ejercicio 2.

Los pesos en kilogramos de un equipo de rugby siguen una Normal N(120, 16). Para un partido una alineación del equipo está formada por 15 jugadores.

1. Hallar la probabilidad de que el peso medio de una alineación supere 126 Kg
2. Calcular el intervalo característico para nivel crítico o de riesgo $\alpha = 0{,}10$.
3. ¿Cuál es la probabilidad de que la suma de los pesos se una alineación tomada al azar sea menor de 1900 kg?.
4. ¿Cuál es la probabilidad de que un jugador elegido al azar pese más de 130 kg?.

Aunque n = 15, y no se llegue a 30, como se parte de una población Normal, las medias de las muestras también siguen una distribución normal. En este caso será $N(120,16/\sqrt{15})$. Luego para las muestras $\mu=120, \sigma=4{,}13$

.

Apartado 1. Si X son las medias muestrales, se nos pide:

$$p(X>126)=p(\frac{X-120}{4{,}13}>\frac{126-120}{4{,}13})=p(Z>1{,}45)=1-p(Z<1{,}45)=1-0{,}9265=0{,}0735$$

Apartado 2. En la N(0,1) el valor de $Z_{\alpha/2}$ de manera que $p(Z<Z_{\alpha/2})=1-0{,}10/2$ determina el extremo de intervalo característico, y se llama valor crítico para el nivel de confianza $1-\alpha=0{,}90$. Mirando en la tabla de N(0,1) para $p(Z<Z_{\alpha/2})=1-0{,}10/2=0{,}95$ es $Z_{\alpha/2}=1{,}645$

El valor crítico para la $N(120,16/\sqrt{15})$ será $K_{\alpha/2}=1.645.16/\sqrt{15}=6{,}79$. Luego el intervalo característico será $\mu-K_{\alpha/2}=120-6.79=113{,}21$ $\mu+K_{\alpha/2}=120+6.79=126{,}79$

Apartado 3. La suma de los pesos sigue una:
$N(n.\mu, \sigma.\sqrt{n})=N(120.15, 16.\sqrt{15})=N(1800, 61{,}97)$ por tanto:

$$p(X<1900)=p(\frac{X-1800}{61{,}97}<\frac{1900-1800}{61{,}97})=p(Z<1{,}61)=0{,}9463$$

Apartado 4. Los jugadores individualmente siguen una $N(120,16)$ luego:

$$p(X>130)=p(\frac{X-120}{16}>\frac{130-120}{16})=p(Z>0{,}625)=1-p(Z<0{,}625)1-0{,}7342=0{,}2658$$

3.2.6.3. Ejemplo 3.

Un estudio nos hace saber que el gasto medio de una población en un fin de semana es de

100 euros, con una desviación típica de 20 euros. Se toman muestras al azar. ¿Qué tamaño debería de tener dichas muestras para que las medias del gasto de las mismas estuviesen entre 95 y 105 euros con un nivel de confianza del 95%. (Nivel de riesgo 0,05).

Como sabemos las muestras de tamaño n siguen una $N(100, 20/\sqrt{n})$ y se debe cumplir que si $\alpha = 0{,}05$ el intervalo característico es $[95\,105]$. Entonces:

$p(95 < X < 105) = 0{,}95$ por lo que $p(X < 105) = 1 - 0{,}05/2 = 0{,}975$ luego:

$$p\left(\frac{X-100}{20/\sqrt{n}} < \frac{105-100}{20/\sqrt{n}}\right) = 0{,}975$$

Buscando en la N(0,1) el valor crítico al 95% nos sale que :

$$\frac{105-100}{20/\sqrt{n}} = 1{,}96 = 0{,}25 \cdot \sqrt{n}$$ luego $\sqrt{n} = 7.84$ luego $n \geq 62$

3.2.6.4. Ejemplo 4.

Se estima que la duración media de una bombilla es de 1000 días, con una desviación típica de 25 días.

Se toma una muestra de 30 bombillas. ¿Cual es la probabilidad de que la cuasivarianza de su duración sea de mayor de 625.

Este test ya no corresponde a una normal, hay que aplicar la propiedad que decía que el estadístico $\frac{(n-1)s^2}{\sigma^2}$ sigue una $\chi^2(n-1)$ de Pearson. Nos piden:

$p(s^2 \geq 625) = p\left(\frac{(30-1).s^2}{25^2} > \frac{(30-1).625}{25^2}\right) = p(\chi^2(29) > 29)$ mirando en la tabla de la χ^2 de Pearson con 29 grados de libertad, nos sale que es $p = 0{,}48$

3.2.6.5. Ejemplo propuesto.

La longitud de los tornillos fabricados por una máquina sigue una N(25,2). Se toma una muestra de 30 tornillos. Calcular:

1. Probabilidad que la media de sus longitudes esté entre 24,5 y 25,5
2. Calcular un intervalo característico para las medias a un nivel crítico $\alpha = 0{,}10$
3. ¿Qué tamaño debería tener una muestra para que entre 24,5 y 25,5 la probabilidad sea del 95%?
4. Probabilidad de que la cuasivarianza de una muestra de 30 tornillos sea mayor que 5.

3.3. Estimación por intervalos de confianza.

La estimación puntual de un parámetro poblacional simplemente establece un estadístico para calcular el parámetro muestral e inferir el valor poblacional a partir del valor muestral.

La estimación por intervalos de confianza consiste en mediante un estadístico muestral construir un intervalo de valores de manera que si el 95% es el nivel de confianza, el parámetro poblacional esta dentro de este intervalo en un 95% de los intervalos construidos por el anterior método.

De esta forma si a fuese el resultado del estadístico, el intervalo suele ser de la forma $a \pm \lambda . b$ donde λ va a depender del nivel de confianza. En general b depende del tamaño de la muestra. El producto $\lambda . b$ representa el margen de error, ya que al estar el parámetro poblacional en este intervalo, del cual $\lambda . b$ es la mitad de su amplitud total, este valor representa el error o precisión del estimador muestral.

Normalmente vamos a estimar los siguientes parámetros.: La media poblacional μ con el estimador media muestral $\bar{x}$. La varianza poblacional σ^2 con la cuasivarianza muestral s^2 y la proporción muestral $\bar{p}$ para contar las veces que nos sale un suceso al repetir un experimento, o sea estimar p en una binomial $B(n, p)$.

Al nivel de confianza le vamos a llamar $1-\alpha$ El nivel crítico o riesgo le llamamos α . Si utilizamos una $N(0,1)$ para el intervalo de confianza, entonces $Z_{\alpha/2}$ será el nivel crítico para α . O sea $p(Z > Z_{\alpha/2}) = \alpha/2$ y $p(-Z_{\alpha/2} < Z < Z_{\alpha/2}) = 1-\alpha$.

$T_{\alpha/2}(n-1)$ Es lo mismo pero par una t – Student $T(n-1)$ con $n-1$ grado de libertad. $p(-T_{\alpha/2} < T < T_{\alpha/2}) = 1-\alpha$. También $p(T > T_{\alpha/2}) = \alpha/2$.

$\chi^2_{\alpha/2}(n-1)$ es el valor de la $\chi^2(n-1)$ de Pearson con $n-1$ grado de libertad de manera que $p(\chi^2 > \chi^2_{\alpha/2}(n-1)) = \alpha/2$

Por último $F_{\alpha/2}(m, n)$ tiene la misma interpretación para la F – Snedecor con m, n grados de libertad para el nivel de confianza de α .

3.3.1. Estimadores por intervalos para una sola población

3.3.1.1. Media de una normal, varianza conocida.

Se supone que estimamos la media de una normal cuya varianza se conoce. Con la terminología anterior, $\bar{x} \pm Z_{\alpha/2} . \frac{\sigma}{\sqrt{n}}$ es el intervalo de confianza a un nivel de confianza $1-\alpha$.

En este caso el error de la estimación es $Error = Z_{\alpha/2}.\frac{\sigma}{\sqrt{n}}$ al nivel $1-\alpha$ de confianza.

3.3.1.2. Ejemplo 1. Estimación media, varianza conocida.

Una población es una normal con varianza $\sigma^2 = 12$.

1. Una muestra de 25 individuos tiene una media de 5. Estimar la media con un intervalo al nivel crítico de 10%. ¿Cuál es el error?
2. Se quiere que el intervalo de confianza tenga un máximo error de 0,5 con un nivel de confianza del 95% ¿Cuál es el tamaño de la muestra?.
3. En una muestra de tamaño 30, se quiere que el error se como mucho 1 . Cual es el nivel de confianza.

Apartado 1.

Para el nivel crítico del 10% $p(Z > Z_{\alpha/2}) = 0{,}05$ luego $Z_{\alpha/2} = 1{,}645$. Luego el intervalo será:

$$5 \pm 1.645.\frac{\sqrt{12}}{\sqrt{25}} = 5 \pm 1{,}1396 \text{ o sea } [3.8603, 6{,}1396] .$$

El error es en este caso $Error = 1.645.\frac{\sqrt{12}}{\sqrt{25}} = 1{,}1396$

Apartado 2.

En un nivel de confianza del 95% $Z_{\alpha/2} = 1.96$. Entonces:

$$Error = 1.96.\frac{\sqrt{12}}{\sqrt{n}} = 0{,}5 \text{ Luego } \sqrt{n} = \frac{1.96.\sqrt{12}}{0{,}5} = 13{,}5792 \quad n \geq 185$$

Apartado 3.

En este caso:

$$Error = 1 = Z_{\alpha/2}.\frac{\sqrt{12}}{\sqrt{25}} \text{ De donde } Z_{\alpha/2} = \frac{5}{\sqrt{12}} = 1{,}1433 \text{ . Entonces:}$$

$$p(Z < 1{,}1433) = 0{,}8729 = 1 - \alpha/2 \quad \alpha = 2.(1 - 0{,}8729) = 0{,}2542$$

El nivel de riesgo es muy significativo para conseguir este intervalo, es posible que la media de la población no esté en este intervalo 5 ± 1 .

3.3.1.3. Media de una normal, varianza desconocida.

Ahora de la normal, no conocemos ni la media, ni la varianza. Entonces estimamos la media y varianza puntualmente. El intervalo de confianza tiene por extremos:

$\bar{x} \pm t_{\alpha/2}(n-1).\frac{s}{\sqrt{n}}$ Ahora el valor crítico lo calculamos con una t-Student.

Como antes $Error = t_{\alpha/2}.\frac{s}{\sqrt{n}}$.

3.3.1.4. Ejercicio 2. Estimación media, varianza desconocida.

En una muestra de 100 tornillos producidos en una fábrica se comprueba que su longitud media es de 50 mm. La varianza de la muestra es de 5 mm

Estimar puntualmente la media y desviación típica de la población dando un intervalo de confianza para la media al 95%. (Se supone una distribución Normal en la fabricación).

Solución.

El estimador de la media de la fábrica es la media de la muestra $\bar{x}=5$.

El estimador de la varianza es la cuasivarianza $s^2=\frac{n}{n-1}.S^2=\frac{100}{99}.5=5.0505$.

Para estimar la media calculamos en una t- student el valor crítico para $\alpha/2=0{,}025$ y mirando en la tabla $t_{\alpha/2}(99)=1{,}984$.

Entonces $Error=1{,}984.\frac{\sqrt{5{,}0505}}{\sqrt{100}}=0{,}4458$. Por lo tanto al 95% de confianza el intervalo para la media es $50\pm 0{,}4458=[49{,}55-50{,}44]$.

3.3.1.5. Estimación de la varianza. Media conocida.

En una población que sigue una variable normal, suponemos conocida la media poblacional μ . Veamos cual es un intervalo de confianza para la varianza.

Como de costumbre el estimador puntual de la varianza poblacional será la cuasivarianza muestral.

En este caso como se supone que la media de la población es conocida (lo cual es bastante raro, lo normal es desconocer ambas). Debemos calcular la cuasivarianza a partir de la media de la población, no de la media de la muestra. O sea:

$s^2=\frac{n}{n-1}.\left(\sum_{i=1}^{n}\frac{x_i^2}{n}-\mu^2\right)$. Entonces, calculada la cuasivarianza como estimador puntual de la varianza, miramos los valores críticos en una $\chi^2(n)$. Y miramos en este caso **dos valores críticos** que son: $\chi^2_{\frac{\alpha}{2}}$ y $\chi^2_{1-\frac{\alpha}{2}}$ Valores que buscamos en la tabla de la χ^2 con n grado de libertad. El intervalo de confianza, tiene una forma diferente al de la media ya que sus extremos son: $\frac{s^2.(n-1)}{\chi^2_{\frac{\alpha}{2}}}$ y $\frac{s^2.(n-1)}{\chi^2_{1-\frac{\alpha}{2}}}$.

Como puede verse el tipo de intervalo no se parece al de la media, nos limitamos a calcularlo, pero ya no hablamos del error a un nivel de confianza.

3.3.1.6. Estimación de la varianza, Media desconocida.

Se calcula exactamente igual que el anterior, sólo que ahora la cuasivarianza se obtiene a partir de la media de la muestra, no de la media de la población. O sea:

$s^2=\frac{n}{n-1}.\left(\sum_{i=1}^{n}\frac{x_1^2}{n}-\bar{x}^2\right)$. Además ahora e $\chi^2(n-1)$ o sea con $n-1$ grado.

3.3.1.7. Ejercicio 3. Estimación varianza. Media y varianza desconocidas.

En una población, que sigue una normal, se extrae una muestra de $n=30$ elementos. De dicha muestra se sabe que: $\sum_{i=1}^{n} x_i=120$ $\sum_{i=1}^{n} x_i^2=48300$. Dar un intervalo de confianza para la media y para la varianza al 90%

Solución. En esta muestra se tiene:

$$\bar{x}=\frac{120}{30}=40 \qquad S^2=\frac{48300}{30}-(40)^2=10 \qquad s^2=\frac{n}{n-1}.S^2=\frac{30}{29}.10=10,34482759$$

Para el intervalo de confianza de la media $t_{(\alpha/2)}(29)=t_{0,5}(29)=1,699$. Entonces:

$Error=t_{\alpha/2}.\frac{s}{\sqrt{n}}=1,699.\frac{\sqrt{10,3448}}{\sqrt{30}}=0,9976$.

El intervalo para media es $40\pm error=40\pm0,9976=[39,0023-40,9976]$.

Para la varianza, los valores críticos en la $\chi^2(29)$ al $\alpha/2=0,05$ y $1-\alpha/2=0,950$ son:

$\chi^2_{0,05}=42,577$ y $\chi^2_{0,95}=17,708$. entonces los extremos del intervalo de confianza para la varianza son:

$$\frac{s^2.(n-1)}{\chi^2_{0,05}}=\frac{300}{42.557}=7.0460 \text{ y } \frac{s^2.(n-1)}{\chi^2_{0,95}}=\frac{300}{17,708}=16.9414 .$$

3.3.1.8. Estimación de una proporción.

El caso siguiente es estimar el parámetro p de una binomial. La muestra consiste en este caso en repetir un experimento n veces, y contar las veces que ocurra un determinado suceso. La proporción de las veces que ocurra dicho suceso (probabilidad del mismo) la estimamos viendo la proporción que sale en el experimento.

Si de las n veces que hacemos el experimento obtenemos el suceso k veces, el estimador puntual de la proporción o probabilidad del suceso es $\bar{p}=\frac{k}{n}$. Como podemos ver otra interpretación es la de estimar el parámetro p de una binomial $B(n,p)$

Se puede demostrar que si el nivel de confianza es $1-\alpha$ el valor crítico puede se buscado en una $N(0,1)$ y es $Z_{\alpha/2}$ el error se puede calcular entonces por $error=Z_{\alpha/2}.\sqrt{\frac{\bar{p}.\bar{q}}{n}}$ luego el intervalo de confianza es $\bar{p}\pm Z_{\alpha/2}.\sqrt{\frac{\bar{p}.\bar{q}}{n}}$.

3.3.1.9. Ejemplo 4. Estimación de una proporción.

Se lanza un dado 100 veces . El número 1 sale 18 veces. Dar un estimador puntual de la proporción y construir un intervalo de confianza al 99%.

Solución. Para el estimador puntual utilizamos $\bar{p}=\frac{18}{100}=0,18$ y $\bar{q}=1-\bar{p}=0.82$.

Mirando en la N(0,1) como $\alpha=0,01$ $\alpha/2=0,005$ y $1-\alpha/2=0,9950$ luego:

$Z_{0,005}=2,575$ luego $error=2,575.\sqrt{\frac{0,18\,.\,0,82}{100}}=0,0989282695$ y el intervalo de confianza sale $[0,081-0,2789]$

Como vemos sale un intervalo de confianza demasiado grande. Vamos a suponer que queremos que el error sea muy pequeño con este mismo nivel de confianza, para ello suponiendo que la proporción estimada siga siendo la misma queremos que $error\leq 0,001$ para ello vamos aumentar n para lograrlo. O sea:

$$Error=0,001=2,575.\sqrt{\frac{0,18\,.\,0,82}{n}}=0,9892827 \quad \text{luego} \quad \sqrt{n}=\frac{0,9892827}{0,001}=989.28$$

Luego $n\geq 978681$ Lo que obliga a una repetición muy grande del experimento para obtener un intervalo significativo muy pequeño.

3.3.2. Estimación por intervalos para dos poblaciones independientes.

Se supone que ahora se tienen dos poblaciones, Se trata de estimar danto un intervalo de confianza la diferencia de las medias, o el cociente de las varianzas. También podemos estimar la diferencia entre dos proporciones.

3.3.2.1. Diferencia de medias. Desviaciones típicas conocidas.

Se tienen dos poblaciones que siguen una $N(\mu_x,\sigma_x)$ y $N(\mu_y,\sigma_y)$ con medias desconocidas y desviaciones típicas conocidas, a partir de dos muestras de tamaños n_x n_y estimamos las medias $\bar{x}$ y $\bar{y}$ de cada muestra. Entonces un intervalo de confianza para la diferencia de las medias a in nivel de confianza $1-\alpha$ podemos obtenerlo así:

Si $Z_{\alpha/2}$ es el valor crítico en una $N(0,1)$ entonces el error se calcula con la expresión:

$Error=Z_{\alpha/2}\cdot\sqrt{\frac{\sigma_x^2}{n_x}+\frac{\sigma_y^2}{n_y}}$ Y los extremos del intervalo son: $(\bar{x}-\bar{y})\pm Error$

3.3.2.2. Diferencia de medias. Desviaciones típicas desconocidas.

Es el mismo planteamiento que el caso anterior, pero ahora se supone que las desviaciones típicas de cada variable son desconocidas, aunque significativamente iguales. Como siempre estimaremos la varianza poblacional, con la cuasivarianza muestral. El valor crítico para el nivel de confianza dado se calcula también con una t-student de n_x+n_y-2 grados de libertad, y le llamamos $t_{\alpha/2}(n_x+n_y-2)$. El error del intervalo se calcula así:

$$Error=t_{\alpha/2}(n_x+n_y-2).\sqrt{\frac{(n_x-1).s_x^2+(n_y-1).s_y^2}{n_x+n_y-2}}\cdot\sqrt{\frac{1}{n_x}+\frac{1}{n_y}}$$

Como antes el intervalo será: $(\bar{x}-\bar{y})\pm Error$

3.3.2.3. Ejemplo 1. Intervalo para diferencias de medias muestrales.

Se tiene un grupo de 25 alumnos, la media de las notas de un examen es de 4,8, y varianza 3. Para un segundo grupo de 35 alumnos la media es de 5,2 y varianza 4. Dar al 90% un intervalo para la diferencia de las medias.

Solución. Suponemos que la media y varianza real son desconocidas. Por los datos dados de tiene que:

$n_x=25 \quad n_y=35 \quad \bar{x}=4{,}8 \quad \bar{y}=5{,}2 \quad s_x^2=\frac{24}{23}.3=3{,}13 \quad s_y^2=\frac{35}{34}.4=4{,}12$

Mirando $t_{0{,}05}(25+35-2)=1{,}671$ y

$Error=1{,}671.\sqrt{\frac{24.3{,}13+34.4{,}12}{25+35-2}}.\sqrt{\frac{1}{25}+\frac{1}{35}}=0{,}8428$ El intervalo al 90% es:

$(4{,}8-5{,}2)\pm 0{,}8428=[-1{,}2428 \ , \ 0{,}4428]$.

3.3.2.4. Diferencias de dos proporciones.

Se estiman ahora dos proporciones para dos experimentos en el primero se repite el experimento n_x veces, estimamos la proporción de un suceso con $\bar{p}_x$. El segundo experimento lo hacemos n_y veces con una proporción de $\bar{p}_y$. Queremos a un nivel de confianza de $1-\alpha$ estimar un intervalo para $\bar{p}_x-\bar{p}_y$. En primer lugar calculamos el nivel crítico en una $N(0,1)$ $Z_{\alpha/2}$ El error del intervalo se calcula así:

$Error=Z_{\alpha/2}\cdot\sqrt{\frac{\bar{p}_x.\bar{q}_x}{n_x}+\frac{\bar{p}_y.\bar{q}_y}{n_y}}$. El intervalo buscado es $(\bar{p}_x-\bar{p}_y)\pm Error$

3.3.2.5. Ejemplo 2. Diferencia de dos proporciones.

Se tienen dos dados el primero se lanza 100 veces y se obtienen 19 veces el uno. El segundo se lanza 200 veces y se obtiene el uno 35 veces. Estimar al 90% un intervalo de confianza para la diferencia de las proporciones:

Solución.

La estimaciones puntuales de las proporciones son: $\bar{p}_x=\frac{19}{100}=0{,}19 \quad \bar{q}_x=1-0{,}19=0{,}81$

$\bar{p}_y=0{,}175 \quad q_y=1-0{,}175=0{,}825$ El valor crítico $Z_{0{,}05}=1{,}645$. Luego:

$Error=1.645\cdot\sqrt{\frac{0{,}19\,.\,0{,}81}{100}+\frac{0{,}175\,.\,0{,}825}{1000}}=0{,}067$ El intervalo es:

$0{,}015\pm 0{,}067=[-0{,}052 \ , \ 0{,}082]$

3.3.2.6. Cociente de dos varianzas. Medias desconocidas.

En dos poblaciones normales tenemos una muestra en cada uno de ellas. Vamos a construir un intervalo de confianza para el cociente de las varianzas de la población. En este caso se supone que tanto la media como la varianza de las poblaciones se desconocen, aunque como en todos los casos son distribuciones normales independientes. Como de cada población tenemos una muestra, supondremos que s_x^2 y s_y^2 serán las cuasivarianzas para muestras de tamaños n_x, n_y respectivamente.

En este caso vamos a buscar en la tabla de la F-Snedecor, buscaremos un valor crítico con un nivel critico dado por $F_{\alpha/2}(n_x-1,n_y-1)$ o sea con n_x-1, n_y-1 grados de libertad.

También tenemos que buscar: $F_{1-\alpha/2}(n_x-1,n_y-1)$. Sin embargo como por una

propiedad de la F – Snedecor se tiene que: $F_{\alpha}(m,n)=\frac{1}{F_{1-\alpha}(n,m)}$ entonces podemos calcular el segundo valor crítico mediante $F_{1-\alpha/2}(n_x-1,n_y-1)=\frac{1}{F_{\alpha/2}(n_y-1.n_x-1)}$. Calculados los dos niveles críticos, el intervalo de confianza para le cociente de las varianzas a un nivel de confianza de $1-\alpha$ es:

$$\left[\frac{s_x^2}{s_y^2}.\frac{1}{F_{\alpha/2}(n_x-1,n_y-1)} \quad , \quad \frac{s_x^2}{s_y^2}.\frac{1}{F_{1-\alpha/2}(n_x-1,n_y-1)}\right]$$

3.3.2.7. Ejemplo 3. Comparación de dos varianzas.

Se estudian las calificaciones de dos grupos de alumnos de un centro. Para el primer grupo de 35 alumnos la varianza de es de 4.5. Para el segundo grupo de 30 alumnos es de 3.7. Determinar con un nivel de confianza del 90% un intervalo de confianza del cociente de las varianzas.

Solución. En primer lugar los datos que tenemos son:

$s_x^2=\frac{35}{34}.4{,}5=4{,}63235$ $\quad s_y^2=\frac{30}{29}.3{,}7=3{,}8276$.

Para la F-Snedecor los valores críticos son:

$F_{0{,}05}(34{,}29)=1{,}83$ y $F_{1-0{,}05}(34{,}29)=\frac{1}{F_{0{,}05}(29{,}34)}=\frac{1}{1{,}80}=0{,}55$. Entonces como:

$\frac{s_x^2}{s_y^2}=\frac{4{,}63235}{3{,}8276}=1{,}21$ El intervalo es $\left[1.21.\frac{1}{1{,}83} \ , \ 1.21.\frac{1}{0{,}55}\right]=[0{,}6612 \ , \ 2{,}2]$

3.3.2.8. Consideraciones finales.

En todas las estimaciones anteriores se ha supuesto que las poblaciones seguían una distribución normal. Si la muestra es pequeña es imprescindible que esto sea así. Sin embargo en muestras grandes se pueden considerar válidos los intervalos anteriormente construidos. De todas formas para muestras la desigualdad de Tchebycheff nos permite aproximar intervalos de confianza para la media de la población aunque no sea normal. Esta desigualdad dice que:

Si X es una variable de media μ y varianza σ^2 para cualquier número $k>0$ se tiene: $p(|X-\mu|\geq k\,\sigma)\leq\frac{1}{k^2}$. En el caso de ser X la variable aleatoria de las medias muestrales entonces el intervalo $\left(\bar{x}-k.\frac{\sigma}{\sqrt{n}},\bar{x}+k.\frac{\sigma}{\sqrt{n}}\right)$ es un intervalo de la media de la población con desviación típica conocida. Siendo su nivel de confianza $1-2.\frac{1}{k^2}$ ya que $\frac{\alpha}{2}=\frac{1}{k^2}$.

3.4. Contraste de hipótesis.

3.4.1. Consideraciones generales.

Consideremos una o varias poblaciones, podemos hacer suposiciones diversas sobre ellas. Ejemplos : La media de una población es $\mu=5$. Dos poblaciones tienen la misma media. La varianza de una población es distinta de la de otra. Dos variables aleatorias que representan a dos caracteres de una población son independientes. Una población se ajusta a un modelo estadístico.

En contraste de hipótesis es verificar afirmaciones como las dadas en los ejemplos anteriores. Esto se realiza tomado una muestra, y según los valores que tome un estadístico de la misma, vemos, si con una determinada probabilidad de equivocación, aceptamos o rechazamos la hipótesis.

Cuando hacemos un contraste de hipótesis normalmente establecemos lo siguiente:

Hipótesis nula o H_0 . Es la hipótesis que queremos comprobar.

Hipótesis alternativa, o complementaria, es H_1 .Normalmente es la contraria a H_0 .

Estadístico de prueba. Es una función tomada a partir de los valores de la muestra cuyo resultado nos dirá si debemos aceptar o rechazar la hipótesis. Este estadístico deberá seguir una variable aleatoria conocida.

Región de aceptación. Es un conjunto de valores para los cuales la probabilidad de que el estadístico tome uno de ellos es igual al nivel de confianza $1-\alpha$. Si establecemos la región de aceptación tal que la probabilidad de que el estadístico esté dentro de ella es $1-\alpha$ entonces ocurrirá lo esperado. En cuyo caso no hay razones para rechazar H_0 y aceptamos la hipótesis nula.

Región de rechazo o crítica. Son los valores complementarios a la región de aceptación. Entonces la probabilidad de que el estadístico esté en esta región es α . En este caso , si el estadístico toma aquí su valor no podemos aceptar la hipótesis nula y aceptamos la alternativa.

Nivel de confianza o de aceptación. Es el valor $1-\alpha$

Nivel de significación, riesgo, crítico o de rechazo. Es justamente α .

Supongamos una hipótesis H_0 y establecemos un nivel de confianza del 90%. Esto supone que fijamos una región de aceptación para la cual la probabilidad de que el estadístico está dentro de ella es del 90%. Evidentemente es lo más fácil . Cuando ocurra no tenemos motivos para rechazar la hipótesis nula ya que sale lo más probable. Sin embargo existe la posibilidad de aceptar la hipótesis nula sin que sea cierta. Si por el contrario es estadístico está fuera de la región de aceptación, ocurre algo poco probable lo que nos lleva a la conclusión de rechazar la hipótesis nula, aunque existe el riesgo de tener mala suerte y ser la hipótesis nula cierta. El contraste de hipótesis

no es pues algo exacto, es probabilístico . Sólo se habla de aceptar o rechazar con un riesgo dado por una probabilidad. Existe la posibilidad de que por culpa de las muestras rechacemos o aceptemos la hipótesis equivocada.

Cabe la tentación de sesgar o falsear intencionadamente un contraste para demostrar lo que queremos. Basta tomar una muestra escogida con un propósito. Por ejemplo para demostrar que en un sitio hay una intención de voto. Que un medicamento cura una enfermedad. Esto es algo que debemos evitar si queremos hace un estudio estadístico serio.

3.4.2. Contraste de hipótesis para una sola población.

3.4.2.1. Hipótesis sobre la media de una normal. Varianza conocida.

Se tiene una población que se supone normal, se conoce su varianza, pero no su media. Se establece como hipótesis nula una de estas tres:

$H_0 : \mu = \mu_0$ o bien $H_0 : \mu \geq \mu_0$ o bien $H_0 : \mu \leq \mu_0$ las tres hipótesis alternativas:
$H_1 : \mu \neq \mu_0$ o bien $H_1 : \mu < \mu_0$ o bien $H_1 : \mu > \mu_0$

Si deseamos hacer un contraste con un nivel de confianza de $1-\alpha$ precedemos así:

1. Calculamos para la muestra que tengamos el estadístico $T = \dfrac{\bar{x} - \mu_0}{\dfrac{\sigma}{\sqrt{n}}}$ Dicho estadístico es el mismo para cualquiera de las tres hipótesis, este estadístico sigue una $N(0,1)$. Esto es cierto aunque no se parta de una población normal si el tamaño de la muestra es mayor de 30.
2. En una normal $N(0,1)$ calculamos en el caso de la primera hipótesis $Z_{\alpha/2}$. Entonces como la probabilidad de que T esté en la región de aceptación. cosa que ocurre si $|T| < Z_{\alpha/2}$, es $1-\alpha$, si esto es así debemos de aceptar H_0 . Mejor aún, como dado que esto es lo más probable que ocurra, decimos que no tenemos razones para rechazar H_0 . Si por el contrario $|T| \geq Z_{\alpha/2}$ está ocurriendo lo más improbable, ya que la probabilidad de que esto ocurra es α . Por lo que no podemos aceptar H_0 y debemos de aceptar H_1 .
3. Para la segunda hipótesis si $T \geq Z_{1-\alpha}$ aceptamos H_0 en caso contrario la rechazamos.
4. Para la tercera hipótesis si $T \leq Z_{\alpha}$ aceptamos H_0 en caso contrario la rechazamos.

3.4.2.2. Hipótesis sobre la media de una normal. Varianza desconocida.

Es el mismo caso anterior pero ahora la varianza de la población es desconocida. Entonces hemos de estimarla con la muestra. Como de costumbre estimamos la varianza de la población con la cuasivarianza muestral.

Como se tienen las tres mismas hipótesis del caso anterior, es procedimiento es similar:

1. Calculamos el estadístico $T = \dfrac{\bar{x} - \mu_0}{\dfrac{s}{\sqrt{n}}}$, que es el mismo para cualquiera de las tres

hipótesis. Pero ahora T sigue una t-Student con $(n-1)$ grados de libertad, en vez de una normal. Aunque si n es grande la t-sudent coincide prácticamente con la normal. Si n es grande y no tenemos la tabla de la t-Student, buscaremos en la $N(0,1)$

2. Calculamos en dicha t-student con $(n-1)$ el valor $t_{\alpha/2}(n-1)$ Si $|T|<t_{\alpha/2}(n-1)$ aceptamos $H_0 : \mu=\mu_0$ En caso contrario la rechazamos.
3. Si $T\geq t_{1-\alpha}(n-1)$ aceptamos la segunda de las hipótesis nulas : $H_0 : \mu\geq\mu_0$.
4. Si $T\leq t_{\alpha}(n-1)$ aceptamos la tercera de las hipótesis nula : $H_1 : \mu<\mu_0$.

3.4.2.3. Ejemplo 1. Hipótesis sobre la media de una normal.

En una población una muestra de 30 personas tiene un peso medio $\bar{x}=83\,kg$. La varianza muestral es de $S^2=15$. Se establece un nivel de confianza $1-\alpha=0{,}90$ para todos los casos.

1. Ver si se puede aceptar H_0 que la media de la población sea de $\mu_0=85\,kg$.
2. Ver si se puede aceptar H_0 que la media $\mu_0<82\,kg$.
3. Ver si se puede aceptar H_0 que la media de la población $\mu_0>85\,kg$.
4. Establecer un intervalo de confianza para esta media muestral. Comprobar si para cualquier valor de la media $\mu=\mu_0$ dentro del intervalo de confianza se puede aceptar que la media de la población tome dicho valor.

Solución.

Vamos a hacer algunos cálculos antes de hacer los apartados.

$s^2=\frac{30}{29}.15=15{,}51725$. $\alpha=1-0{,}90=0{,}10$. $\alpha/2=0{,}05$. $1-\alpha/2=0{,}95$.

Mirando en $t(29)$ de student:

$t_{0{,}05}(29)=1{,}699$ $t_{0{,}10}(29)=1{,}311$ $t_{0.90}(29)=-1{,}311$ No viene este último en la tabla, pero por semejanza simétrica de la t-sudent ya que es muy parecida a la normal, podemos aunque el último valor no venga en la tabla, calcularlo, ya que:

$$p(t<t_{0{,}10})=1-p(t>t_{0{,}10})=1-0{,}10=0{,}90=p(t>-t_{0{,}10})=p(t>t_{0{,}90})$$.

Entonces $t_{0{,}90}(29)=-t_{0{,}10}(29)$

Apartado 1.

El estadístico del contraste es $T=\frac{\bar{x}-\mu_0}{\frac{s}{\sqrt{n}}}=\frac{83-85}{\sqrt{\frac{15{,}51725}{30}}}=-2.781$

Como $|T|=2{,}781>t_{0{,}05}(29)=1{,}699$ Rechazamos la hipótesis H_0 la media no puede ser 85 kg.

Apartado 2.

El estadistico $T=\frac{\bar{x}-\mu_0}{\frac{s}{\sqrt{n}}}=\frac{83-82}{\sqrt{\frac{15{,}51725}{30}}}=1{,}3904$

Como $T=1{,}3904>t_{0{,}10}(29)=1{,}311$ no podemos aceptar que la media sea menor que 82

kg.

Apartado 3.

El estadístico $T=\dfrac{\bar{x}-\mu_0}{\frac{s}{\sqrt{n}}}=\dfrac{83-85}{\sqrt{\frac{15{,}51725}{30}}}=-2.781$.

Como $T=-2{,}781<t_{0{,}90}(29)=-1{,}311$ Tampoco se puede aceptar que la media esté por encima de 85 kg.

Apartado 4.

Un intervalo de confianza al nivel del 90% tiene el siguiente error:

$Error=t_{0{,}05}.\dfrac{s}{\sqrt{n}}=1{,}699.\sqrt{\dfrac{15{,}51725}{30}}=1{,}222$. Luego los extremos del intervalo son:

$83\pm 1{,}222=[81{,}778\ ,\ 84{,}222]$

Si ahora hacemos un contraste de hipótesis donde μ_0 estuviese dentro del intervalo, entonces como en este caso $|\bar{x}-\mu_0|<Error$ se tendrá:

$|T|=\left|\dfrac{\bar{x}-\mu_0}{\frac{s}{\sqrt{n}}}\right|<\left|\dfrac{Error}{\frac{s}{\sqrt{n}}}\right|=\left|\dfrac{Z_{\alpha/2}.\frac{s}{\sqrt{n}}}{\frac{s}{\sqrt{n}}}\right|=Z_{\alpha/2}$ Entonces para estos valores de μ_0 siempre queda el estadístico dentro de la región de aceptación, y aceptaremos para todos ellos la hipótesis nula.

Así para $\mu_0=84$ aceptaríamos la hipótesis. Para $\mu_0=82$ también.

3.4.2.4. Hipótesis sobre varianzas. Media conocida.

Se trata de establecer hipótesis sobre los valores de la varianza en una distribución normal tanto como si se conoce o no la media de la misma. La forma de hacerlo es muy parecida en ambos casos.

Supongamos que se tiene la población, de la cual se conoce la media μ . se extrae una muestra. Para dicha muestra como en el caso de intervalos de confianza podemos calcular la cuasivarianza a partir de la media de la población. Su expresión es:

$$s^2=\frac{n}{n-1}.\left(\sum_{i=1}^{n}\frac{x_i^2}{n}-\mu^2\right)$$

Si verificásemos la hipótesis nula $H_0:\sigma=\sigma_0$. El estadístico será:

$T=\dfrac{(n-1)s^2}{\sigma_0^2}$ que sigue una $\chi^2(n)$ de Pearson. Para un nivel de confianza $1-\alpha$ la región de aceptación se calcula a partir de dicha $\chi^2(n)$ de Pearson. Para ello buscamos en la tabla los valores $\chi^2_{\alpha/2}(n)$ y $\chi^2_{1-\alpha/2}(n)$ si se verifica que $\chi^2_{1-\alpha/2}(n)<T<\chi^2_{\alpha/2}(n)$ Se aceptará la hipótesis H_0 .

Si verificásemos la hipótesis nula $H_0:\sigma\geq\sigma_0$ Entonces construimos el mismo estadístico T, buscamos el valor $\chi^2_{1-\alpha}(n)$. Si $T\geq\chi^2_{1-\alpha}(n)$ aceptamos la hipótesis.

Si verificásemos la hipótesis nula $H_0: \sigma \leq \sigma_0$. Con el mismo estadístico buscamos el valor $\chi^2_\alpha(n)$. Si $T \leq \chi^2_\alpha(n)$ aceptamos la hipótesis.

3.4.2.5. Hipótesis sobre varianzas. Media desconocida.

Si ahora tenemos una población que sigue una distribución normal donde no se conoce ni la media ni la varianza, y establecemos unas hipótesis semejantes al caso anterior . El cálculo es muy semejante, salvo que ahora la cuasivarianza se hace mediante la media de la muestra ya que no se conoce la media de la población. Luego:

$$s^2 = \frac{n}{n-1} \cdot \left(\sum_{i=1}^{n} \frac{x_i^2}{n} - \bar{x}^2 \right) \quad .$$

El estadístico es $T = \frac{(n-1)s^2}{\sigma_0^2}$ que sigue también una χ^2 de $n-1$ grados de libertad en vez de χ^2 con n grados , que era el caso anterior.

Todo lo demás es igual, tanto las hipótesis, como las regiones de aceptación o de rechazo.

3.4.2.6. Ejemplo 2. Hipótesis sobre la varianza.

Supuesto que la vida media en un país es una Normal, en una muestra de 30 personas sabemos que la media $\bar{x} = 79$ años $S^2 = 20$. ¿Se puede aceptar las hipótesis de que la vida media es de $\mu_0 = 80$ y $\sigma_0^2 = 22$ a un nivel $1 - \alpha = 0{,}90$

Solución: Tenemos en realidad dos contrastes de hipótesis, una para le media y otra para la varianza:

En primer lugar $s^2 = \frac{n}{n-1} . S^2 = 20 . \frac{30}{29} = \frac{600}{29}$ y $\frac{s}{\sqrt{n}} = \frac{s}{\sqrt{30}} = \frac{\sqrt{\frac{600}{29}}}{\sqrt{30}} = \sqrt{\frac{20}{29}}$

El estadístico para la media es:

$$T = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{79 - 80}{\sqrt{\frac{20}{29}}} = -1{,}2041$$ Como $t_{0,05}(29) = 1{,}699$ $\quad |T| = 1{,}2041 < 1{,}68 = t_{0,05}(29)$

No hay motivos para rechazar que la hipótesis de que sea $\mu_0 = 80$ al 90%

El estadístico para la varianza es:

$T = \frac{(n-1).s^2}{\sigma_0^2} = \frac{\frac{29.600}{29}}{22} = 27{,}27$ Mirando en la tabla $\chi^2_{0,05}(29) = 42{,}557$ y $\chi^2_{0,95}(29) = 17{,}708$ como T está entre ambos valores no tenemos motivo para rechazar la hipótesis, y claramente aceptamos la hipótesis

Como $(n-1).s^2 = 29 . \frac{600}{29} = 600$ entonces si cogemos el intervalo de confianza de la varianza que es:

$\frac{(n-1).s^2}{\chi^2_{0,05}}=\frac{600}{42,557}=14,96$ y $\frac{(n-1).s^2}{\chi^2_{0,95}}\frac{600}{17,708}=33,88$ Cualquier valor de este intervalo tomado como hipótesis como valor de σ_0^2 da validez a dicha hipótesis.

3.4.2.7. Contraste de hipótesis para una proporción.

Se trata de verificar que la proporción de un suceso en un determinado experimento aleatorio, toma un valor determinado valor.

De esta forma el experimento se hace n veces, la proporción de veces que obtenemos un suceso es $\bar{p}$. Se trata de verificar una de las hipótesis siguientes:

$H_0: p=p_0$ $H_0: p\geq p_0$ o bien $H_0: p\leq p_0$ con el nivel de confianza $1-\alpha$.

Para todos los casos el estadístico es:

$T=\frac{\bar{p}-p_0}{\sqrt{\frac{p_0.(1-p_0)}{n}}}$. que sigue una $N(0,1)$.

Para determinar la región de confianza, miramos los valores: $Z_{\alpha/2}$, Z_{α} y también $Z_{1-\alpha}$ en dicha $N(0,1)$. Entonces si:

$|T|<Z_{\alpha/2}$ Aceptamos $H_0: p=p_0$. En caso contrario la rechazamos.
$Z_{1-\alpha}\leq T$ Aceptamos $H_0: p\geq p_0$. En caso contrario la rechazamos.
$T\leq Z_{\alpha}$ Aceptamos $H_0: p\leq p_0$. En caso contrario la rechazamos.

3.4.2.8. Ejemplo 3. Hipótesis sobre la proporción.

Lanzamos un dado 1000 veces, nos sale 175 veces el número 1. ¿ Podemos aceptar que la hipótesis $H_o: p_0=\frac{1}{6}$ a un nivel de confianza $1-\alpha=0,80$?.

Solución.

En este caso la proporción de la muestra es $\bar{p}=\frac{175}{1000}=0,175$. El estadístico es:

$T=\frac{\bar{p}-p_0}{\sqrt{\frac{p_0.(1-p_0)}{n}}}=\frac{\frac{175}{1000}-\frac{1}{6}}{\sqrt{\frac{\frac{1}{6}.\frac{5}{6}}{1000}}}=0,7071$. Para este nivel de confianza $Z_{0,10}=1,281$

Como $|T|=0,7071<1,281=Z_{0,10}$. A este nivel de confianza aceptamos la hipótesis.

Es más siempre que $Z_{\alpha/2}>0,7071$ también aceptaríamos al hipótesis. Mirando la tabla se tiene que $p(Z<0,7071)=0,808=1-\alpha/2$ $\alpha/2=0,192$ $\alpha=0,384$ $1-\alpha=0,616$

O sea hasta en un nivel del 60% podemos aceptar la hipótesis.

3.4.3. Contraste de hipótesis para dos poblaciones.

3.4.3.1. Hipótesis sobre la igualdad de dos medias. Varianza conocida.

Se tienen dos poblaciones normales, no son conocidas las medias, pero sí las varianzas. Se extrae una muestra de cada población, el tamaño de las muestras es n_x, n_y , las media de cada muestra es $\bar{x}, \bar{y}$ y la varianza de cada población es σ_x^2, σ_y^2 . Vamos a constatar una de las siguientes hipótesis a un nivel de confianza $1-\alpha$.

$H_0: \mu_x - \mu_y = a$. Diferencia de las medias igual a un cierto valor.
$H_0: \mu_x - \mu_y \geq a$. Diferencias de las medias igual o mayor que un cierto valor
$H_0: \mu_x - \mu_y \leq a$. Diferencia de las medias menor o igual que un cierto valor.

El estadístico que vamos a utilizar es:

$T = \dfrac{\bar{x} - \bar{y} - a}{\sqrt{\dfrac{\sigma_x^2}{n_x} + \dfrac{\sigma_y^2}{n_y}}}$, que sigue también una $N(0,1)$ En dicha normal determinamos como siempre: $Z_{\alpha/2}$ Z_α $Z_{1-\alpha}$.

Entonces si:

$|T| < Z_{\alpha/2}$ Aceptamos $H_0: \mu_x - \mu_y = a$. En caso contrario la rechazamos.
$Z_{1-\alpha} \leq T$ Aceptamos . $H_0: \mu_x - \mu_y \geq a$ En caso contrario la rechazamos.
$T \leq Z_\alpha$ Aceptamos . $H_0: \mu_x - \mu_y \leq a$ En caso contrario la rechazamos.

3.4.3.2. Hipótesis de igualdad de medias. Varianzas desconocidas pero próximas.

En este caso no conocemos ni la media, ni la varianza de ninguna de las dos poblaciones. Se trata de verificar las mismas hipótesis del apartado anterior. En este caso de las muestras conocemos: $n_x, n_y, \bar{x}, \bar{y}, s_x^2, s_y^2$.

El estadístico que utilizamos es:

$T = \dfrac{\bar{x} - \bar{y} - a}{\sqrt{\dfrac{(n_x - 1).s_x^2 + (n_y - 1).s_y^2}{n_x + n_y - 2}} \cdot \sqrt{\dfrac{1}{n_x} + \dfrac{1}{n_y}}}$. que sigue una t-student con $n_x + n_y - 2$ grados de libertad. En dicha t – student con $n_x + n_y - 2$ buscamos las regiones críticas. Una vez determinado los valores $t_{\alpha/2}(n_x + n_y - 2)$, t_α y $t_{1-\alpha}$ se hace el test de costumbre . Así pues:

$|T| < t_{\alpha/2}$ Aceptamos $H_0: \mu_x - \mu_y = a$. En caso contrario la rechazamos.
$t_{1-\alpha} \leq T$ Aceptamos . $H_0: \mu_x - \mu_y \geq a$ En caso contrario la rechazamos.
$t \leq Z_\alpha$ Aceptamos . $H_0: \mu_x - \mu_y \leq a$ En caso contrario la rechazamos.

3.4.3.3. Ejemplo 1. Hipótesis sobre las medias.

Las medias de dos muestras de tamaño $n_x = 100, n_y = 200$ son $\bar{x} = 5{,}2, \bar{y} = 6$ si las cuasivarianzas son $s_x^2 = 10, s_y^2 = 12$. Contrastar la hipótesis de que las medias son iguales. . Nivel de confianza $1-\alpha = 0{,}9$

Solución.

$$T=\frac{\bar{x}-\bar{y}-a}{\sqrt{\frac{(n_x-1).s_x^2+(n_y-1).s_y^2}{n_x+n_y-2}}\cdot\sqrt{\frac{1}{n_x}+\frac{1}{n_y}}}=\frac{5{,}2-6}{\sqrt{\frac{99.10+199.12}{100+200-2}}\cdot\sqrt{\frac{1}{100}+\frac{1}{200}}}=-1.9401$$

La región crítica deberíamos buscarla en una t-student con $100+200-2=298$ grados de libertad. Entonces como $t_{0,05}(298)=1{,}65$ parecido al valor que nos saldría si buscamos en una $N(0,1)$ para el mismo nivel crítico.

Pero como $|T|=1{,}9401>1{,}65=t_{0,05}(298)$. No podemos aceptar la hipótesis de igualdad de medias.

Sin embargo como :

$\sqrt{\frac{(n_x-1).s_x^2+(n_y-1).s_y^2}{n_x+n_y-2}}\cdot\sqrt{\frac{1}{n_x}+\frac{1}{n_y}}.t_{\alpha/2}=0{,}41235\,.1.65=0{,}6803$. Entonces aceptaríamos la hipótesis $H_0:\mu_x-\mu_y=a$ cuando a sea un número que cumpla que: $|\bar{x}-\bar{y}-a|<0{,}6803$ o sea $|5.2-6-a|=|-0{,}8-a|<0{,}6803$ luego $-0{,}6803<-0{,}8-a<0{,}6803$. Esto dice que si a es un número que cumple la desigualdad: $-1{,}4803<a<-0.1197$ se aceptaría la hipótesis. Por lo tanto la diferencia de medias podría aceptarse dentro de este intervalo. O sea si establecemos que $\mu_x-\mu_y$ es igual a cualquier número entre -1,4803 y -0,1197 de aceptaría esta hipótesis al 90%.

3.4.3.4. Hipótesis sobre igualdad de varianzas.

Se tienen dos poblaciones Normales. Se establecen una de las siguientes hipótesis:

$H_o:\sigma_x^2=\sigma_y^2$. Varianzas de la población iguales.
$H_o:\sigma_x^2\geq\sigma_y^2$. Varianza de la primera mayor que la de la segunda.
$H_o:\sigma_x^2\leq\sigma_y^2$. Varianza de la primera menor que la de la segunda.

Si s_x^2, s_y^2 son las cuasivarianzas de dos muestras de tamaños n_x, n_y . Podemos establecer el estadístico:

$T=\frac{s_x^2}{s_y^2}$ El cual sigue una $F(n_x-1,n_y-1)$ de Snedecor con (n_x-1,n_y-1) grados de libertad.

Para constatar las hipótesis buscamos los valores $F_{\alpha/2}(n_x,-1,n_y-1)$ y $F_{1-\alpha/2}(n_x,-1,n_y-1)$. Como este último valor no aparece en la tabla, aplicamos la igualdad:

$F_{1-\alpha/2}(n_x,-1,n_y-1)=\frac{1}{F_{\alpha/2}(n_y-1,n_x-1)}$. También con los mismos grados de libertad caculamos los valores: F_α y $F_{1-\alpha}$. Entonces si:

$F_{1-\alpha/2}(n_x-1,n_y-1)<T<F_{\alpha/2}(n_x-1\,,n_y-1)$. Aceptamos $H_o:\sigma_x^2=\sigma_y^2$.
$F_{1-\alpha}\leq T$ Aceptamos $H_o:\sigma_x^2\geq\sigma_y^2$
$T\leq F_\alpha$. Aceptamos $H_o:\sigma_x^2\leq\sigma_y^2$

3.4.3.5. Ejemplo 2. Hipótesis sobre las varianzas.

Se tienen dos tipos de bombillas, de desea saber si la vida media de cada grupo coincide, así coma la varianza respecto a dicha vida media. Se toman dos muestras de 50 bombillas de cada grupo y se determina que la duración media en cada grupo es de 510 y 509 horas respectivamente y sus varianzas de 9 y de 10. Determinar si podemos aceptar las hipótesis dadas, al nivel $1-\alpha=0{,}90$

Solución. El estadístico para la comparación de vida media es:

$$T=\frac{\bar{x}-\bar{y}}{\sqrt{\frac{(n_x-1).s_x^2+(n_y-1).s_y^2}{n_x+n_y-2}}\cdot\sqrt{\frac{1}{n_x}+\frac{1}{n_y}}}=\frac{510-509}{\sqrt{\frac{49.50.\frac{9}{49}+49.50.\frac{10}{49}}{50+50-2}}\cdot\sqrt{\frac{1}{50}+\frac{1}{50}}}=1{,}6059$$

Como $t_{0,05}(50+50-2)=1{,}660$ Entonces $1{,}6059=T<1{,}660=t_{0,05}$

Debemos aceptar la hipótesis de que la vida media de ambos grupos es la misma.

Para la varianza, el estadístico es:

$$T=\frac{s_x^2}{s_y^2}=\frac{\frac{50}{49}.9}{\frac{50}{49}.10}=0{,}9$$

$F_{0,05}(49{,}49)=1{,}59$ y $F_{0,95}(49{,}49)=\frac{1}{F_{0,05}(49{,}49)}=\frac{1}{1{,}59}=0{,}62$

Como $F_{0,95}=0{,}62<0{,}9=T<1{,}59=F_{0,05}$

Aceptamos la hipótesis de que las dos varianzas son iguales.

3.4.4. Hipótesis sobre bondad de ajuste.

Este test se utiliza para ver si una población a través de una muestra sigue una determinada variable aleatoria.

Consideremos una población donde suponemos que un determinado carácter sigue una variable aleatoria, continua o discreta. Queremos comprobar que esto es cierto a través de una muestra.

La hipótesis nula es H_0 La población sigue la variable aleatoria X .

3.4.4.1. Ajuste a una distribución discreta.

Supongamos una población sobre la que hacemos una hipótesis de que siga una variable aleatoria discreta X. Esta variable tiene su función de probabilidad y si son $x_1, x_2, x_3, ... x_k$ los posibles valores que toma se tendrá que $p_i = p(X = x_i)$ será la función de probabilidad siendo lógicamente $\sum_{i=1}^{k} p_i = 1$.

En el test de ajuste tomamos una muestra de tamaño n y contamos la frecuencia experimental de las veces que se obtiene cada valor x_i en la muestra, estos valores les llamamos $n_1, n_2, n_3, ... n_k$ y su suma $n_1 + n_2 + n_3 + ... + n_k = n$

Por otra parte podemos calcular las frecuencia teórica para cada valor de la variable en la muestra. Estos valores se obtienen así: $e_i = p_i . n$. Evidentemente también:

$$\sum_{i=1}^{k} e_i = \sum_{i=1}^{k} n.p_i = n. \sum_{i=1}^{k} p_i = n$$

Es importante señalar que todas las frecuencias experimentales deben de ser mayores de cinco. En caso contrario se puede aumentar el tamaño de la muestra, lo cual obliga a modificar el muestreo, o bien se agrupan valores. Por ejemplo: Si $n_6 = 3$, lo agrupamos con n_5 o con n_7 de forma que la suma de ambos sea mayor que 5. Esto obliga a considerar la zona agrupada donde la variable X toma uno de los dos valores con probabilidad la suma de las probabilidades. Luego para esos dos valores agrupados la frecuencia teórica es $e_5 + e_6$ o bien $e_6 + e_7$.

Una vez realizado el muestreo el estadístico del tests es el siguiente:

$$T = \sum_{1}^{k} \frac{(n_i - e_i)^2}{e_i} = \sum_{i=1}^{k} \frac{n_i^2 - 2.\mathrm{n}_i . e_i + e_i^2}{e_i} = \sum_{i=1}^{k} \frac{n_i^2}{e_i} - \frac{2.\mathrm{n}_i . e_i}{e_i} + \frac{e_i^2}{e_i} = \sum_{i=1}^{k} \frac{n_i^2}{e_i} - \sum_{i=1}^{k} 2.\mathrm{n}_i + \sum_{i=1}^{k} e_i$$

$$T = \sum_{i=1}^{k} \frac{n_i^2}{e_i} - 2.\mathrm{n} + n = \sum_{i=1}^{k} \frac{n_i^2}{e_i} - n$$

Que será la expresión del estadístico.

Se demuestra que en general el estadístico sigue una variable aleatoria $\chi^2(k-1)$ de Pearson con $k-1$ grados de libertad . Observar que k no es el tamaño de la muestra, es el

número de valores que toma la variable aleatoria.

Podría ocurrir que para el cálculo de las frecuencias teóricas necesitásemos estimar p parámetros de la población, en este caso los grados de libertad serían $k-p-1$.

Entonces si χ^2_α es el valor crítico al nivel de confianza $1-\alpha$ si:

$T<\chi^2_\alpha(k-p-1)$ Aceptamos la hipótesis de sobre la variable X.
$T\geq\chi^2_\alpha(k-p-1)$ Rechazamos la hipótesis realizada.

3.4.4.2. Ejemplo 1. Ajuste una muestra. Distribución discreta.

Supongamos que lanzamos 600 veces un dado, y obtenemos la siguiente tabla de frecuencias. Queremos saber si el dado se ajusta a lo esperado, o sea la probabilidad de cada número es $\frac{1}{6}$ con un nivel del 90%.

Solución: Construimos la tabla siguiente:

x_i	$x_1=1$	$x_2=2$	$x_3=3$	$x_4=4$	$x_5=5$	$X_6=6$
p_i	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$
n_i	95	107	98	96	106	98
$e_i=n.p_i$	100	100	100	100	100	100
n_i^2	9025	11449	9604	9216	11236	9604
n_i^2/e_i	90,25	114,49	96,04	92,18	112,36	96,04

Si calculamos:

$T=\sum_{n=1}^{6}\frac{n_i^2}{e_i}-600=601{,}36-600=1{,}36$. Como tenemos 5 grados de libertad ya que son 6 los posibles valores del dado y no se ha estimado ningún parámetro. Entonces:

$T=1{,}36<9{,}236=\chi^2_{0,10}(5)$.

Por lo que debemos aceptar la hipótesis de que el dado está correcto.

3.4.4.3. Ejemplo 2. Ajuste de una muestra. Distribución discreta.

Un arquero lanza 4 flechas a un blanco, y repite el experimento 1000 veces, obteniendo la siguiente tabla de resultados. En dicha tabla x_i son los posibles resultados de una variable aleatoria que cuenta los aciertos en cada una de las veces que se tiran las cuatro flechas, por lo que es una binomial $B(4,p)$. p es la probabilidad de que una flecha de en el blanco, y $k=4$ son las flechas que se tiran cada una de las $n=1000$ veces que se repite el experimento.

n_i Es la frecuencia absoluta de cada uno de los posibles resultados en las 1000

repeticiones. Así $n_0=415$ ya que en 415 ocasiones de las 1000 el arquero no acierta ni sólo disparo.

x_i	0	1	2	3	4
n_i	415	410	150	23	2

Se quiere ver si los resultados se ajustan realmente a una binomial donde debe de ser estimado el parámetro p .

Solución:

Si vemos las veces que acierta en los 4000 = 4x1000 intentos de hacerlo, vemos que:

Aciertos $a=415x0+410x1+150x2+23x3+4x2=787$. Entonces estimamos la proporción de aciertos con $\bar{p}=\frac{787}{4000}=0,19675$. Y $\bar{p}$ sería la estimación puntual de p de $B(k\,,p)$

Un intervalo de confianza para la proporción al 90% de confianza sería el siguiente:

Al 90% $Z_{0,05}=1,645$, luego:

$Error=Z_{0,05}.\frac{\sqrt{\bar{p}.(1-\bar{p})}}{n}=1,645.\sqrt{\frac{0,19675.(1-0,19675)}{4000}}=0,0104$ y el intervalo es:

$0,1965\pm 0,0104=[\,0,1864\,,\,0,2070\,]$

Vemos que si tomásemos como el valor estimado de la proporción $\bar{p}=0,20$ estaríamos dentro del intervalo de confianza.

Pero es más, si hacemos la hipótesis $H_0: p=p_0=0,20$ también al 90% . Si tomamos el estadístico siguiente:

$T=\frac{\bar{p}-p_0}{\sqrt{\frac{p_0.(1-p_0)}{n}}}=\frac{0,19675-0,20}{\sqrt{\frac{0,20.0,80}{4000}}}=-0,4110$. Como $|T|=0,4110<1,645=Z_{0,05}$

Podemos aceptar la hipótesis H_0 . A partir de ahora aceptamos que la proporción es $p=0,20$ y vamos a comprobar si podemos ajustar la tabla a una $B(k=4,\,p=0,20)$

En la tabla vamos a añadir la probabilidades teóricas para una binomial $B(n=k\,,\,p=0,20)$, y las frecuencias teóricas correspondientes. Realizaremos el correspondiente test de ajuste de la muestra a la binomial.

x_i	0	1	2	3	4
n_i	415	410	150	23	2
p_i	0,4096	0,4096	0,1536	0,0256	0,0016
$e_i=n\,.\,p_i$	410	410	153	25	2

Pero vemos que las frecuencias para x_4 no llegan a 5, por lo que se ha de agrupar esa columna con la anterior, quedando la tabla:

x_i	0	1	2	$X \geq 3$
n_i	415	410	150	25
p_i	0,4096	0,4096	0,1536	0,0272
e_i	410	410	153	27
n_i^2	172225	168100	22500	625
n_i^2/e_i	420.06	410	147,0588	23,1481

Si calculamos el estadístico $T=\sum_{n=1}^{6}\frac{n_i^2}{e_i}-1000=1000{,}267947-1000=0{,}267947$

Hemos realizado el Test con $k=4$ (desde x_0 a x_3) además hemos estimado un parámetro, luego los grados de libertad de χ^2 son $4-1-1=2$ entonces para el test al 90% tenemos que $\chi^2_{0,10}=4{,}605$ evidentemente mayor que el estadístico, luego a esta nivel aceptamos que la muestra representa una binomial con los parámetros dados.

3.4.4.4. Ajuste a una distribución continua.

Lo mismo que hemos comprobado si una muestra se ajusta a una distribución teórica de tipo discreto se puede realizar el ajuste para variables aleatorias de tipo continuo.

Se tiene una población que suponemos que sigue un determinado tipo de variable aleatoria continua X. Queremos comprobar esta hipótesis a través de los tatos de una muestra con un nivel de confianza $1-\alpha$. Mientras en el otro caso los datos de la muestra se estudiaban tomando los posibles valores de la variable aleatoria, esto ahora no sirve. Al ser la variable aleatoria de tipo continuo se deben considerar intervalos de valores donde la variable puede tomar los valores.

Llamemos $x_0, x_1, x_2, x_3, \ldots . x_k$ a los extremos de los intervalos de la forma $[x_{k-1}, x_k]$ desde $k=1$ y terminando en k . Como antes llamamos $n_1, n_2, n_3, \ldots . n_k$ las frecuencias absolutas experimentales que toma la muestra en cada uno de los intervalos. Como antes $\sum_{i=1}^{k} n_i = n$ .Siendo n el número de elementos de la muestra. Entonces ahora definimos:
$p_1 = p(X < x_1)$, $p_2 = p(x_1 < X < x_2)$, $p_3 = p(x_2 < X < x_3) \ldots . p_{k-1} = p(x_{k-2} < X < p_{k-1})$ y
$p_k = p(X > x_{k-1})$.

Observar que tanto p_1 y p_k los tomamos par valores de X que no tienen en cuenta ni x_0 ni n_k ya que tomamos la probabilidad de las colas para que estén todos los valores que puede tomar X y sea $\sum_{i=1}^{k} p_i = 1$. El calcular todas estas probabilidades se hará tomando las tablas de la correspondiente distribución, como esto puede ser largo y tedioso se pueden buascar los datos en ordenador. En general es más sencillo con ordenador ya que por ejemplo $p(1 < X < 5)$ en una $N(4,3)$ se puede hacer utilizando una Hoja de Cálculo de forma directa sin necesidad siquiera de tipificar la variable. Es mucho más rápido.

Los pasos siguiente son semejantes a las variables discretas, o sea calculamos $e_i = n.p_i$ y también el estadístico:

$T = \sum_{i=1}^{k} \frac{n_i^2}{e_i} - n$. Aplicamos el mismo criterio que en las discretas, o sea:

Se demuestra que en general el estadístico sigue una variable aleatoria $\chi^2(k-1)$ de Pearson con $k-1$ grados de libertad . Observar que k no es el tamaño de la muestra, es el número de valores que toma la variable aleatoria.

Podría ocurrir que para el cálculo de las frecuencias teóricas necesitásemos estimar p parámetros de la población, en este caso los grados de libertad serían $k-p-1$.

Entonces si χ^2_α es el valor crítico al nivel de confianza $1-\alpha$ si:

$T < \chi^2_\alpha(k-p-1)$ Aceptamos la hipótesis de sobre la variable X.

$T \geq \chi^2_\alpha(k-p-1)$ Rechazamos la hipótesis realizada.

3.4.4.5. Ejemplo 3. Ajuste de una muestra. Distribución continua.

Supongamos que las notas de un examen de 40 alumnos tiene la siguiente tabla de frecuencias:

x_i	1	2	3	4	5	6	7	8	9	10
n_i	2	4	4	6	8	5	4	3	2	2

Estudiar si es posible ajustarla a un distribución Normal, con una confianza del 90%

En primer lugar vamos a estimar la media, y la varianza de la normal.

Si calculamos la media de la normal nos sale:

$\bar{x} = \frac{\sum_{i=1}^{10} x_i.n_i}{n} = \frac{206}{40} = 5{,}15 \qquad S^2 = \frac{\sum_{i=1}^{10} x_i^2.n_i}{n} - \bar{x}^2 = \frac{1280}{40} - 5{,}15^2 = 5{,}4775$ y

$s^2 = \frac{n}{n-1} S^2 = 5{,}617949$

Podemos hacer un contraste de hipótesis auxiliar con el objetivo de verificar la hipótesis:

$H_0 : \mu = 5$ Para ello construimos el estadístico:

$T = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{5{,}15-5}{\sqrt{\frac{5{,}617949}{40}}} = 0{,}4$. Como $|T| = 0{,}14 < 1{,}68 = T_{0{,}05}$. Podemos aceptar la hipótesis de que la normal tiene media $\mu = 5$.

Pasamos a hacer el test de ajuste de la muestra. Formamos la siguiente tabla, donde las probabilidades de la Normal se han calculado por ordenador con una hoja de cálculo. Donde

$p_1 = p(X < 1{,}5)$, $p_2 = p(1{,}5 < X < 2{,}5) \ldots . p_9 = p(8{,}5 < X < 9{,}5)$, $p_{10} = p(X > 9{,}5)$.

La variable X es una $N(5{,}5, s)$ Se ha tomado la varianza como la cuasivarianza dada por el estimador puntual.

La tabla es así:

x_i	1	2	3	4	5	6	7	8	9	10
n_i	2	4	4	6	8	5	4	3	2	2
p_i	0,0698	0,0758	0,1176	0,1530	0,1670	0,1530	0,1176	0,7588	0,4110	0,0288
e_i	2,8	3,04	4,71	6,12	6,88	6,12	4,71	3,04	1,64	1,15
n_i^2	4	16	16	36	64	25	16	9	4	4
n_i^2 / e_i	1,43	5,27	3,4	5,88	9,58	4,08	3,4	2,97	2,43	3,47

En este caso vemos que muchas de las frecuencias de la muestra no llegan a 5, había teóricamente que agruparlas, pero si calculamos el estadístico para la tabla actual y la agrupada casi no hay diferencia en los resultados. Podemos comprobar que en este caso:

$T = \sum_{n=1}^{6} \frac{n_i^2}{e_i} - 40 = 1{,}91$. La $\chi_{0,10}(40-2-1) = 49{,}5$. Por tanto aceptamos que nuestros datos se ajustan a una normal con un nivel de confianza del 90%.

3.4.5. Hipótesis sobre independencias de caracteres.

Una población presenta dos caracteres, Este test trata de estudiar su dependencia o independencia, veamos en que consiste este concepto. Si se tiene una población y en ella se estudian dos caracteres, podemos hablar de variables aleatorias X e Y para cada uno de ellos, así como de la variable aleatoria conjunta XY.

Esta variable XY podría darnos por ejemplo $p(X = a, Y = b)$ (X, Y discretas) o bien si X e Y continuas de $p(a < X < b, c < Y < d)$ Entonces si en cada caso:

$p(X = a, Y = b) = p(X = a).p(Y = b)$ (X, Y discretas) o bien:
$p(a < X < b, c < Y < d) = p(a < X < b).p(c < Y < d)$ (X, Y continuas)

Entonces se dice que estas dos variables aleatorias son independientes, concepto que en cierto modo coincide con el de sucesos independientes.

Podemos hacer un test para verificar a través de una muestra la independencia de dos caracteres.

Supongamos una muestra de n elementos, donde para el carácter X la muestra toma los valores $x_1, x_2, x_3 \ldots . x_k$ y para la variable Y la muestra toma los valores $y_1, y_2, \ldots y_h$. Llamamos n_{ij} las veces que en la muestra se da en valor (x_i, y_j) que serán las frecuencias experimentales. Consideremos la tabla siguiente

	y_1	y_2	y_3	.....	y_{h-1}	y_h	Frec. X
x_1	n_{11}	n_{12}	n_{13}		n_{1h-1}	n_{1h}	n_{x1}
x_2	n_{21}	n_{22}	n_{23}		n_{2h-1}	n_{2h}	n_{x2}
x_3	n_{31}	n_{32}	n_{33}		n_{3h-1}	n_{3h}	n_{x3}
.....							
x_{k-1}	n_{k-11}	n_{k-12}	n_{k-13}		n_{k-1h-1}	n_{k-1h}	n_{xk-1}
x_k	n_{k1}	n_{k2}	n_{k3}		n_{kh-1}	n_{kh}	n_{xk}
Frec. Y	n_{y1}	n_{y2}	n_{y3}		n_{yh-1}	n_{yh}	n

Donde $n_{xi}=\sum_{j=1}^{h} n_{ij}$ (frecuencias para la variable X) y $n_{yj}=\sum_{i=1}^{k} n_{ij}$ (Frec Y)

Realmente la tabla anterior es una tabla de doble entrada para las variables X e Y donde también aparecen las frecuencias para cada variable de forma independiente.

Si aceptamos que las variables son independientes entonces el valor esperado de las frecuencias para cada par de valores de la variable sería el producto de n por la probabilidad que tiene de salir el par (x_i, y_j) que si fuesen independientes las variables tiene la expresión:

$$e_{ij}=n.p_{ij}=n.p_{xi}.p_{xj}=n.\frac{n_{xi}}{n}.\frac{n_{yj}}{n}=\frac{n_{xi}.n_{yj}}{n}$$

Calculados estos valores teóricos podemos calcular el estadístico siguiente:

$$T=\sum_{1=1}^{k}\sum_{j=1}^{h}\frac{(n_{ij}-e_{ij})^2}{e_{ij}}=\sum_{i=1}^{k}\sum_{j=1}^{h}\frac{n_{ij}^2}{e_{ij}}-n$$ que sigue una $\chi^2(h-1).(k-1)$. entonces:

Si $T<\chi_{\alpha}^2(h-1).(k-1)$ entonces aceptamos la hipótesis de H_o de que las variables son independientes al nivel de aceptación de $1-\alpha$

3.4.5.1. Ejemplo de un análisis de independencia de valores:

En una muestra de $n=125$ personas, se analizan las preferencias por 5 marcas de un determinado producto alimenticio, la muestra está dividida en 5 grupos de personas según sus edades. El resultado de la encuesta arroja la siguiente tabla:

	A	B	C	D	E	Frecuencia edad.
X < 15	8	8	6	2	1	25

15 < X <20	6	7	6	4	2	25
20 < X < 30	4	6	6	6	3	25
30 < X < 40	2	4	6	6	7	25
40 < X <60	1	2	6	8	8	25
Frecuencia de gustos.	21	27	30	26	21	125

Verificar al 85%, 90%, 95% y 99% la hipótesis H_0 : El gusto por un producto y la edad son independientes.

Solución: Vamos a aplicar un test de independencia, para ello formamos la tabla con la frecuencias teóricas. Donde $e_{ij} = \frac{n_{xi} . n_{yj}}{n}$. La tabla será así:

	A	B	C	D	E	Frecuencia edad.
X < 15	4,20	5,40	6,00	5,20	4,20	25
15 < X <20	4,20	5,40	6,00	5,20	4,20	25
20 < X < 30	4,20	5,40	6,00	5,20	4,20	25
30 < X < 40	4,20	5,40	6,00	5,20	4,20	25
40 < X <60	4,20	5,40	6,00	5,20	4,20	25
Frecuencia de gustos.	21	27	30	26	21	125

A continuación formamos la tabla donde cada celda se obtiene mediante $\frac{n_{ij}^2}{e_{ij}}$

La tabla será la siguiente:

	A	B	C	D	E	Frecuencia edad.
X < 15	15,2381	11,8915	6,0000	0,7692	0,2381	34,0973
15 < X <20	8,5714	9,0741	6,0000	3,0769	0,9524	27,6748
20 < X < 30	3,8095	6,6667	6,0000	6,9231	2,1429	25,5421
30 < X < 40	0,9524	2,9630	6,0000	6,9231	11,6667	28,5051
40 < X <60	0,2381	0,7407	6,0000	12,3077	15,2381	34,5246
Frecuencia	28,8095	31,2963	30,0000	30,000	30,2381	150,3439

de gustos.						

El estadístico es $T=150{,}3429-125=25{,}3439$. Si miramos la tabla de la $\chi^2(16)$. Entonces:

$\chi^2_{0,15}(5-1,5-1)=21{,}7$
$\chi^2_{0,10}(5-1,5-1)=23{,}7$
$\chi^2_{0,05}(5-1,5-1)=26$
$\chi^2_{0,01}(5-1,5-1)=31$

Sólo en el último caso se está en la región de aceptación, por lo que a los niveles 85%, 90% y 95% se rechaza H_0 , sólo se aceptaría al nivel del 99%, por lo que en general rechazamos la hipótesis y las variables se consideran dependientas, luego los gustos si dependen de la edad.

3.4.6. Hipótesis sobre homogeneidad de muestras.

Se tiene una población, donde se estudia un determinado carácter , se toman varias muestras, se trata de determinar si para dicho carácter presentan resultados semejantes, y no hay diferencias significativas de una a otra muestra.

Para ello sean $n_1, n_2, n_3, n_4, \ldots\ldots, n_k$ los tamaños de cada muestra, y sea $y_1, y_2, y_3, \ldots\ldots y_k$ las veces que en cada muestra se presenta el carácter a estudiar.

Si todas las muestras representan a la misma población, la proporción total de las veces que se ha presentado el carácter es:

$p=\dfrac{y_1+y_2+y_3+\ldots+y_k}{n_1+n_2+n_3+\ldots.+n_k}$. Podemos llamar $e_i=n_i.\,p$ que son los valores esperados que se debe presentar el carácter en cada una de las muestras.

Podemos hacer una sencilla tabla donde aparecen los valores obtenidos y los valores esperados. La tabla es así:

Muestras	Se presenta el carácter	No se presenta	Tamaño muestra.
Primera muestra	y_1 $n_1.p$	n_1-y_1 $n_1.(1-p)$	n_1
Segunda muestra	y_2 $n_2.p$	n_2-y_2 $n_2.(1-p)$	n_2
.....			
Muestra - k	y_k $n_k.p$	n_k-y_k $n_k.(1-p)$	n_k

La Hipótesis es H_0 Todas las muestras son homogéneas y representan a la misma

población. El nivel de confianza es $1-\alpha$.

El estadístico es $T=\frac{1}{p.(1-p)}.\sum_{i=1}^{k}\frac{(y_i-n_i.p)^2}{n_i}$ que sigue una $\chi^2(k-1)$ y que por lo tanto si $T<\chi^2_\alpha(k-1)$ aceptaremos la hipótesis.

3.4.6.1. Caso de varios caracteres.

En el caso de querer estudiar varios caracteres tomando varias muestras, en las columnas de la tabla anterior ponemos los caracteres que vamos a considerar, y en las filas las muestras. Entonces podemos realizar un tests de independencia, o sea, con los mismos cálculos que si fuese un test de dependencia podemos verificar H_0 : Las muestras son homogéneas. Si haciendo el test de independencia el estadístico está dentro de la zona de aceptación, podemos aceptar que las muestras son homogéneas. Los grados de libertad serán (muestras-1).(carácteres -1)

3.4.6.2. Ejemplo para test de muestras homogéneas, Un sólo carácter.

Se tienen tres medicamentos que administrados queremos saber si igualmente eficaces, al nivel del 90%, para ello se han probado con una serie de paciente obteniendo el resultado siguiente:

Medicamentos	Curados	No curados	Total
A	10	20	30
B	20	35	55
C	8	12	20

Calculamos la proporción:

Calculamos $p=\frac{10+20+8}{30+55+20}=\frac{38}{105}=0{,}3619$

Calculamos la tabla de valores esperados:

Medicamentos	Curados	No curados	Total
A	10 30.p = 10,8571	20 (1-p).30 =19.1429	30
B	20 55.p = 19.9048	35 (1-p).55 = 35,0952	55
C	8 20.p = 7,2381	12 (1-p)= 12,7619	20

Para el estadístico, ponemos en la primera columna $\frac{(y_i-n_i.p)^2}{n_1}$ y la primara columna sale:

Medicamentos	Curados	No curados	Total
A	0,0245		30
B	0,0002		55
C	0,0290		20
Suma	0,0537		

Entonces $T=\frac{1}{p(1-p)}.0{,}0537=0{,}0946$ Mirando en la tabla de la $\chi^2(3-1)$, tenemos: $T=0{,}0946<\chi^2_{0,10}(2)=4{,}605$ por lo que se puede aceptar la hipótesis de que las muestras son homogéneas.

3.4.6.3. Test de homogeneidad para varios caracteres.

Se tienen dos dados, después de lanzar 40 veces el primero y 30 veces el segundo se obtiene la siguiente tabla de resultados:

Dados	1	2	3	4	5	6	Total
A	7	6	7	6	7	7	40
B	4	6	5	5	6	4	30
Total	11	12	12	11	13	11	70

Verificar al 90% que ambos dados son homogéneos.

Como cada una de las dos muestras presenta 6 caracteres el test a realizar es el de independencia.

Entonces hacemos la tabla esperada suponiendo la independencia:

La tabla se calcula con: $e_{ij}=\frac{n_{xi}.n_{yj}}{n}$ y es:

Dados	1	2	3	4	5	6	Total
A	6,2857	6,8571	6,8571	6,2857	7,4286	6,2857	40
B	4,7143	5,1429	5,1429	4,7143	5,5714	4,7143	30
Total	11	12	12	11	13	11	70

Calculamos el estadístico con los valores $\frac{n_{ij}^2}{e_{ij}}$. Es la siguiente :

Dados	1	2	3	4	5	6	Total

A	7,7955	5,2500	7,1458	5,7273	6,5962	7,7955	40,3102
B	3,3939	7,0000	4,8611	5,3030	6,4615	3,3939	30,4136
Total							70,7237

El estadístico es:

$T = 70{,}7237 - 70 = 0{,}7237$. Como los grados de libertad son (2-19.(6-1)=5. Entonces:

$T = 0{,}7237 < \chi^2_{0,10}(5) = 9{,}236$. Podemos al 90% aceptar la Hipótesis de que los dos dados son homogéneos.

3.4.7. Ejemplos de evaluación final.

Los siguientes ejemplos son referentes a toda la teoría de muestras, se deberán de presentar resueltos, salvo que se hagan de forma práctica con datos reales.

3.4.7.1. Ejemplo 1.

La siguiente tabla representa los tiempos de curación de dos medicamentos administrados a una muestra de 40 pacientes. Realizando todos los test al 90% . Comprobar:

Medicamento - A

Tiempo en días	1	2	3	4	5	6	Total
Pacientes curados	7	8	10	7	5	3	40

Medicamento - B

Tiempo en días	1	2	3	4	5	6	Total
Pacientes curados	7	6	9	8	6	4	40

1. Estimar la media y varianza de la eficacia en días de ambos medicamentos supuesto que se comporta dicha eficacia como una distribución normal cuya media y varianza deseamos estimar. Hacer la estimación por intervalos de confianza al 90%

2. Comprobar la hipótesis de que la media de la curación con ambos medicamentos es la misma.

3. Comprobar la hipótesis de que las varianzas en cada medicamento es la misma.

4. Comprobar la hipótesis de homogeneidad en las muestras.

3.4.7.2. Ejemplo 2.

Se lanza un dado 60 veces y se obtiene las siguientes tabla:

Dado 1

Resultado	1	2	3	4	5	6	Total
Frecuencia	12	9	9	10	11	9	60

1. Estimar dando un intervalo de confianza al 90% la proporción de números pares.
2. Verificar al 90% si el dado está trucado.

4. Índice.