2º de Bachillerato

Matemáticas Aplicadas a las Ciencias Sociales II

Contenidos

Estadística inferencial: Muestras. Tipos de muestreo. Distribuciones muestrales.

1. Muestras aleatorias

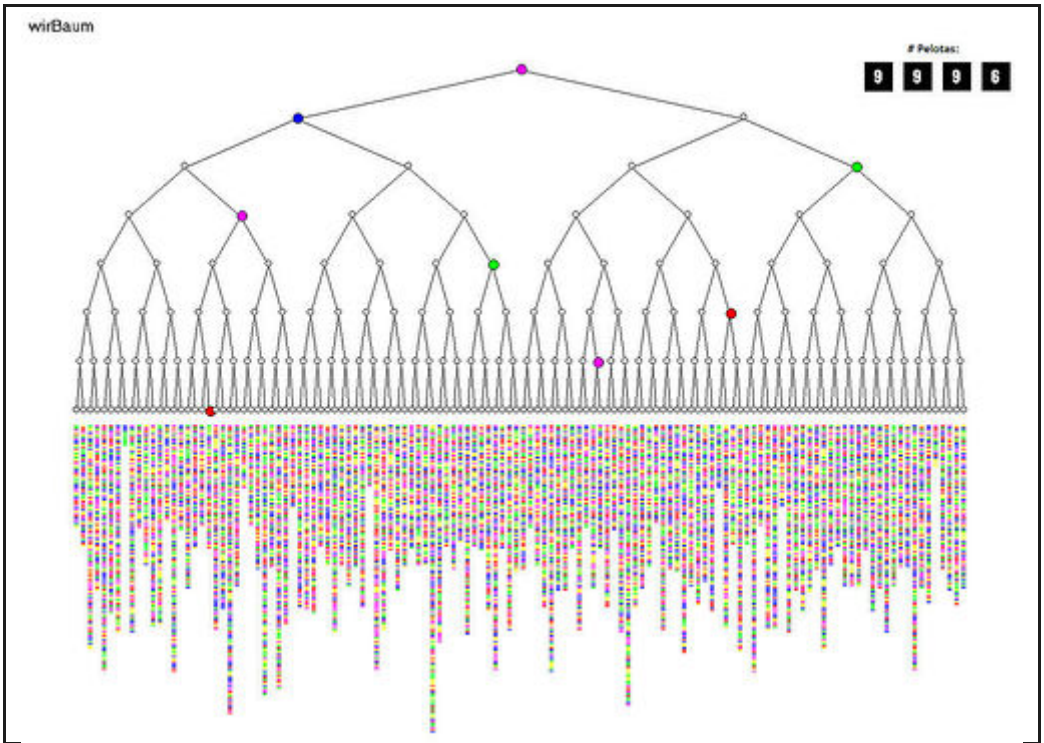


Imagen de eliazar bajo licencia Creative Commons

A nuestros amigos de la empresa TisBet Survey les gusta hacer las cosas bien y por eso tienen mucho cuidado a la hora de elaborar las encuestas, no vaya a ser que les pase lo que le ocurrió a la revista americana "Literary Digest" en 1936, anunciando la derrota de Franklin D. Roosevelt del Partido Demócrata, como presidente de EE.UU. después de haber escogido una muestra de 2,3 millones de votantes. (Mira más abajo en el apartado "Curiosidad").

Cuando nuestros amigos tienen que realizar alguna encuesta, toman una muestra representativa y para ello, sus elementos han tenido que ser elegidos de forma aleatoria, al azar.

Esta forma de elegir a los individuos de la muestra, es lo que determina que estemos hablando de un **muestreo aleatorio**.

Curiosidad

Roosevelt. Imagen de Wikimedia Commons

En 1936, la revista "**Literary Digest**" utilizó una muestra significativa de 2,3 millones de "votantes", con la cual determinó que la población americana votaría al Partido Republicano. Una semana antes del día de las elecciones presidenciales, se informaba que Alf Landon del Partido Republicano era, de lejos, mucho más popular que Franklin D. Roosevelt del Partido Demócrata.

Al mismo tiempo, **George Gallup** realizó una encuesta mucho más pequeña, a sólo 5000 personas, pero con mejores bases científicas, utilizando muestras demográficas representativas.

Gallup predijo correctamente la victoria arrolladora de Roosevelt. Al poco tiempo, el Literary Digest dejó de funcionar, mientras que los sondeos empezaron a incrementarse.

En 1948, la organización de George Gallup tuvo un gran error cuando predijo que Thomas Dewey derrotaría a Harry S. Truman en las elecciones de 1948, por una diferencia de entre el 5% y el 15%. Gallup consideró que el error se debió principalmente a que terminó los sondeos tres semanas antes de la jornada electoral y mucha gente que votaba a Dewey se abstuvo de hacerlo porque creía que iba a ser el ganador de todas formas.

Es a partir de este año 1948, cuando los estadísticos prestan todos sus conocimientos para la elaboración de encuestas.

Respecto a la muestra de 2,3 millones de personas que tomó la revista, el error fue que se basó solamente en direcciones de abonados al teléfono y listas de propietarios de automóviles, con lo que de aleatoria tenía muy poco.

Comprueba lo aprendido

En el instituto de la hija del director de TisBet Survey, quieren saber el número de horas que están conectados a Internet sus 1200 alumnos y alumnas.

Para ello realizan una encuesta a una muestra de 100 de ellos.

Determina con verdadero o falso, si las siguientes formas de elegir las muestras son aleatorias.

a) En la muestra tiene que haber alumnado con buenas notas, alumnado con todo aprobado y alumnado con asignaturas suspensas.

Sugerencia

○ Verdadero ○ Falso

Falso
La muestra sería sesgada porque puede ser que alumnado con buenas notas utilice mucho Internet para estudiar y alumnado con malas notas lo utilice mucho para chatear.

b) Para la muestra se eligen los 100 alumnos y alumnas que primeros lleguen al instituto.

Sugerencia

○ Verdadero ○ Falso

Falso
La muestra no es representativa porque el llegar primero al Centro puede ser por responsabilidad o bien por vivir cerca del instituto.

c) La muestra está formada por 100 alumnos y alumnas seleccionados por sorteo, al azar.

Sugerencia

○ Verdadero ○ Falso

Verdadero
Esta muestra sí es representativa. Este muestreo se llama aleatorio.

Un **muestreo** se dice que es **aleatorio** o **probabilístico** cuando todos los individuos de la muestra se eligen al azar, de modo que todos tienen la misma probabilidad de ser elegidos.

Sólo este método de muestreo nos asegura la representatividad de la muestra extraída y es, por tanto, el más recomendable.

En este apartado veremos distintos tipos de muestreo aleatorio y para empezar te dejo con un magnífico vídeo sobre Matemática electoral de la serie de TV Más por Menos.

1.1. Muestreo aleatorio simple

Imagen de Freddy the boy bajo licencia Creative Commons

Como viste en el tema anterior, la empresa Aventruzez le había encargado un estudio a nuestros amigos de TisBet Survey.

Los trabajadores de TisBet Survey deciden hacer una encuesta y pasársela a una muestra de la población de las Alpujarras que sea lo más significativa posible, es decir, que les sirva para que los resultados que obtengan sean fiables para el informe que tienen que hacer a Aventruzez.

A la hora de seleccionar la muestra, escogen el método de muestreo aleatorio simple, por ser el más riguroso y científico además de ser es el más sencillo donde se basan todos los demás.

Para ello seleccionan cada elemento de la muestra uno por uno de forma aleatoria, teniendo en cuenta que todos los elementos de la población tienen la misma probabilidad de ser incluidos en la muestra.

En este apartado estudiaremos este tipo de muestreo y veremos algunas técnicas para obtener una muestra aleatoria de una población finita.

Importante

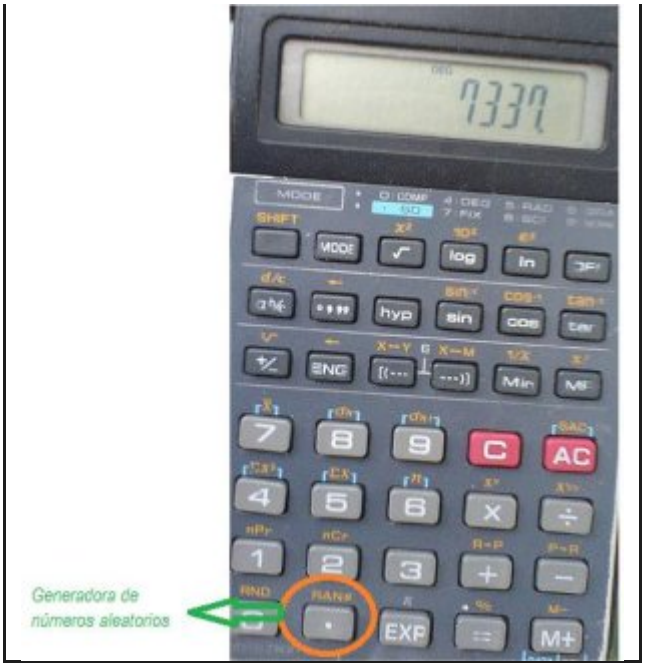
Un **muestreo** es **aleatorio simple** cuando listamos todos los elementos de la población y seleccionamos aleatoriamente los ***n*** elementos de la muestra.

El muestreo aleatorio simple es adecuado cuando la población es homogénea respecto a la característica que se estudia.

Una de las técnicas para obtener una muestra aleatoria de una población finita es la obtención de números aleatorios.

Para ello puedes utilizar la tecla que está marcada en la imagen de la izquierda.

Al pulsar sobre ella obtienes un número al azar comprendido entre 0,000 y 0,999.



Cada número que obtengas lo multiplicas por el número de elementos de la población (N), obteniendo un número decimal cuya parte entera está comprendida entre 0 y N-1.

Si tomas la parte entera del número obtenido multiplicando N por el número aleatorio y le sumas 1, ya tienes un número elegido al azar entre 1 y N.

En el siguiente vídeo se explica como generar números aleatorios con una calculadora científica.

Ejercicio resuelto

En el barrio donde está la oficina de TisBet Survey, se acaba de establecer una empresa nueva que se dedica a realizar encuestas en los centros educativos.

Las primeras encuestas las han realizado en el instituto "Benito V.", y han tratado sobre el uso de redes sociales por parte de alumnos y alumnas.

Empiezan en una clase de 2º de Bachillerato para analizar el número de horas que están conectados a la red Tuenti. En este grupo hay 30 alumnos y deciden coger una muestra de 5, tomando un alumno de cada fila de mesas según están sentados.

Imagen de Medellín Digital bajo licencia Creative Commons

1. ¿Es este tipo de muestreo, aleatorio simple? Razona tu respuesta.

Al día siguiente, cuando se entera el gerente de Tisbet Survey de cómo han escogido la muestra, se desplaza al instituto y escoge una muestra numerando a los 30 alumnos y generando como hemos visto antes con la calculadora, cinco números aleatorios.

Los alumnos que tienen asignados esos números son los que hacen la encuesta.

2. Si los números aleatorios obtenidos son: 0,224; 0,251; 0,058; 0,496 y 0,560, ¿cuáles son los números de los alumnos que tienen que hacer la muestra?

3. ¿Qué empresa obtendrá unos datos más fiables?

Mostrar retroalimentación

1. Claramente se ve que el muestreo no es aleatorio ya que no se han escogido a los alumnos y alumnas al azar.

2. Para obtener los números de los alumnos y alumnas, multiplico cada número aleatorio conseguido por el número de elementos de la población, en este caso 30 y del resultado + 1, tomo la parte entera del número.

Es decir, tendríamos lo siguiente:

30·0,224 +1=6,72+1=7,72, cuya parte entera es 7.

30·0,251+1= 7,53+1=8,53, cuya parte entera es 8.

30·0,058+1=1,74+1=2.74, cuya parte entera es 2.

30·0,496+1= 14,88+1=15,88, cuya parte entera es 15.

30·0,560+1= 16,8+1=17,8, cuya parte entera es 17.

La solución sería: 2, 7, 8, 15 y 17.

3. La empresa TisBet Survey obtendrá unos datos más fiables porque ha escogido los elementos de la muestra al azar.

Comprueba lo aprendido

Blanco

En la clase de 2ª de Bachillerato hay 25 alumnos y queremos extraer una muestra de tamaño 4.

Si numeramos a los alumnos desde el número 1 hasta el 25, encuentra los que serán seleccionados si consideramos los siguientes números aleatorios:

1. Para el número aleatorio 0,92, seleccionamos al alumno número [ ] .
2. Para el número aleatorio 0,963, seleccionamos al alumno número [ ] .
3. Para el número aleatorio 0,159, seleccionamos al alumno número [ ] .
4. Para el número aleatorio 0,479, seleccionamos al alumno número [ ] .

Enviar

Importante

El **muestreo con reemplazamiento** es aquel en el que un elemento puede ser seleccionado más de una vez en la muestra. Para ello se extrae un elemento de la población, se observa y se devuelve a la población, por lo que de esta forma se pueden hacer infinitas extracciones de la población aún siendo ésta finita.

El **muestreo sin reemplazamiento** es el que se realiza sin devolver los elementos extraídos a la población hasta que no se hayan extraído todos los elementos de la población que conforman la muestra.

Ejercicio resuelto

En el instituto "Benito V." estudia el hijo del gerente de TisBet Survey.

Hay un total de 560 alumnos y se ha escogido una muestra de 28 para conocer si tienen internet en casa.

a) ¿Qué significa elegir a 28 de 560? ¿Qué proporción de la población estamos entrevistando?

b) A la hora de obtener conclusiones sobre la población, ¿a cuántos alumnos de la población total representa cada uno de los de la muestra?

c) ¿Qué tipo de muestreo interesa más en este caso, con reemplazamiento o sin reemplazamiento?

Mostrar retroalimentación

a) Para calcular la proporción de alumnos que estamos entrevistando, dividimos el tamaño de la muestra entre el de la población: $\frac{28}{560} = 0{,}05$, lo que quiere decir que estamos pasando la encuesta al **5%** de la población.

b) Para calcular el número de individuos que representa cada uno de los elementos de la muestra, dividimos el número de individuos de la población entre los de la muestra: $\frac{560}{28} = 20$, lo que significa que cada uno de los elementos de la muestra representa a **20** alumnos del instituto.

c) Para elegir la muestra entre los 560 alumnos del instituto, si vamos a preguntar por el hecho de que posean internet en casa, no nos interesa preguntarle dos veces a la misma persona, luego una vez elegido un elemento de la muestra no queremos volverlo a seleccionar. Realizaríamos pues un muestreo aleatorio sin reposición o **sin reemplazamiento**.

Importante

1. Llamamos **factor de elevación** al cociente entre el tamaño de la población y el tamaño de la muestra, $k = \frac{N}{n}$. Representa el número de elementos que hay en la población por cada elemento de la muestra.

2. Llamamos **factor de muestreo** al cociente entre el tamaño de la muestra y el tamaño de la población, $\frac{1}{k} = \frac{n}{N}$. Si se multiplica por 100, obtenemos el porcentaje de la población que representa la muestra.

Comprueba lo aprendido

Blanco

En el equipo de fútbol de la ciudad donde tiene su oficina la empresa TisBet Survey acaban de llegar a los 10000 socios.

El presidente del club le encarga a nuestros amigos de esta empresa que hagan una encuesta sobre la bebida que suelen beber cuando van a un partido de fútbol, para así poder dar un mejor servicio en los bares del estadio durante el descanso del partido.

La empresa decide hacer una encuesta a una muestra de 400 aficionados.

a) El factor de muestreo en esta encuesta es de [ ] , es decir, un [ ] % de la población.

b) El factor de elevación es de [ ] , es decir, cada elemento de la muestra representa a [ ] socios.

c) Respecto al muestreo con reemplazamiento o sin reemplazamiento, el tipo de muestreo más recomendable es el de [ ] [ ] .

Enviar

1.2. Muestreo aleatorio estratificado

Imagen de Miguel Vera bajo licencia Creative Commons

A veces puede interesarnos hacer otro tipo de muestreo que no sea el aleatorio simple.

Imagina que queremos hacer un estudio para saber a qué dedican su tiempo libre las personas que viven en tu ciudad. Todos sabemos que los ancianos no realizan el mismo tipo de actividades que los jóvenes, ni tampoco que las personas de mediana edad, como por ejemplo tus padres. Nos interesaría entonces que toda esta información que tenemos de antemano nos ayude a construir una muestra más significativa. De hecho, nos interesa que todos esos colectivos estén representados en nuestra muestra.

A los colectivos que hemos definido, en este caso por edad, los llamaremos **estratos**. Lo que haremos será dividir nuestra muestra de manera que haya representantes de todos los estratos.

A este tipo de muestreo lo llamamos **muestreo estratificado**.

Importante

Decimos que un **muestreo es aleatorio estratificado** cuando dividimos la población en subgrupos o estratos homogéneos y en cada uno de ellos tomamos muestras aleatorias simples.

Para hacer los subgrupos, seguiremos el criterio de formarlos de tal manera que haya la máxima homogeneidad en relación con la variable a estudio dentro de cada estrato y la máxima heterogeneidad entre los estratos.

Respecto al reparto del tamaño de la muestra en los diferentes estratos o subpoblaciones, consideraremos distintos criterios de afijación: afijación igual si todos los estratos tienen el mismo número de elementos en la muestra y afijación proporcional si cada estrato tiene un número de elementos en la muestra proporcional a su tamaño.

Ejercicio resuelto

Imagen de IES Bajo Guadalquivir bajo licencia Creative Commons

En el instituto "Benito V." se quiere pasar una encuesta tomando una muestra de 28 alumnos considerando el curso al que pertenecen.

El número de alumnos por niveles es el siguiente:

Curso	Nº de alumnos/as
1º ESO	120
2º ESO	120
3º ESO	100
4º ESO	90
1º Bach.	70
2º Bach.	60

a) ¿Qué tipo de muestreo tendríamos que utilizar?

b) ¿Cómo sería la muestra si hacemos la encuesta solamente en el primer ciclo de Secundaria?

c) Siguiendo el criterio del curso y considerando todo el centro educativo, ¿cómo sería la muestra?

Mostrar retroalimentación

a) Al tener dividida la población en subgrupos, el muestreo que tenemos que hacer es el aleatorio estratificado.

b) Como la muestra es de 28 alumnos y hay el mismo número de alumnos en 1º y en 2º, la muestra estaría formada por 14 alumnos de cada nivel, escogidos de forma aleatoria.

c) En este caso el número de alumnos varía dependiendo del curso, por lo que habría que considerar un muestreo estratificado proporcional.

- 1º de ESO tendría en la muestra a $\frac{120}{560}\cdot 28 = 6$ alumnos.
- 2º de ESO tendría en la muestra a $\frac{120}{560}\cdot 28 = 6$ alumnos.
- 3º de ESO tendría en la muestra a $\frac{100}{560}\cdot 28 = 5$ alumnos.
- 4º de ESO tendría en la muestra a $\frac{90}{560}\cdot 28 = 4{,}5 \approx 4$ alumnos.

- 1º de Bachillerato tendría en la muestra a $\frac{70}{560}\cdot 28 = 3{,}5 \approx 4$ alumnos.
- 2º de Bachillerato tendría en la muestra a $\frac{60}{560}\cdot 28 = 3$ alumnos.

En el siguiente vídeo puedes ver un ejemplo de muestreo aleatorio estratificado.

Comprueba lo aprendido

Blanco

En el barrio donde se encuentra la oficina de TisBet Survey, viven unos 1500 niños y jóvenes, 7500 adultos y 1000 ancianos.

El ayuntamiento de la localidad se está planteando la construcción de un parque de atracciones, para lo cual decide encargarle a TisBet Survey que haga una encuesta sobre los gustos de ocio y aventura a una muestra de 200 individuos elegidos al azar.

Para ello se utiliza un muestreo estratificado.

¿Cuál será el tamaño muestral correspondiente a cada estrato?

La muestra está formada por ☐ niños y jóvenes, ☐ adultos y ☐ ancianos.

Enviar

El tamaño de cada uno de los estratos debe ser proporcional a la cantidad de individuos de cada uno de ellos.

N.º de niños y jóvenes: $\frac{200}{10000}\cdot 1500 = 30$.

N.º de adultos: $\frac{200}{10000}\cdot 7500 = 150$.

N.º de ancianos: $\frac{200}{10000}\cdot 1000 = 20$.

1.3. Muestreo aleatorio por conglomerados y sistemático

Imagen de amiramora bajo licencia Creative Commons

El equipo de baloncesto de la ciudad le ha encargado a TisBet Survey que haga un estudio de la altura de todos los alumnos y alumnas de Secundaria.

En lugar de hacer un muestreo de todos los chicos y chicas de la ciudad, la empresa se plantea elegir algunos barrios, ya que con respecto a la altura, los barrios son como "pequeñas poblaciones" comparables a la ciudad.

En este caso ¿podemos simplificar la elección de la muestra al elegir los barrios sin perder precisión?

La respuesta es que en este caso, podríamos elegir barrios y analizar las alturas de los estudiantes de cada barrio sin perder precisión.

El método que nos permite hacer esto es el **muestreo por conglomerados**.

La población se divide en unidades o grupos, llamados conglomerados (generalmente son unidades o áreas en los que se ha dividido la población), que deben ser lo más representativas posible de la población, es decir, deben representar la heterogeneidad de la población objeto del estudio y ser entre sí homogéneos.

El motivo para realizar este muestreo es que a veces resultaría demasiado costoso realizar una lista completa de todos los individuos de la población objeto del estudio, o que cuando se terminase de realizar la lista no tendría sentido la realización del estudio.

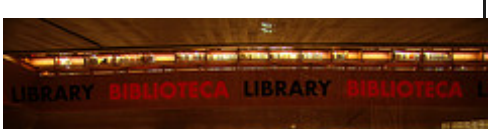
El principal inconveniente que tiene es que si los conglomerados no son homogéneos entre sí, la muestra final puede no ser representativa de la población.

Importante

Decimos que un **muestreo es aleatorio por conglomerados** cuando dividimos la población en conjuntos o conglomerados.

Se eligen al azar unos pocos de estos conglomerados y la muestra estará formada por todos los elementos de ellos o por muestras aleatorias simples de éstos.

Comprueba lo aprendido



A la empresa TisBet Survey le han encargado que haga un estudio sobre el estado de los libros de todas las bibliotecas municipales de la ciudad.

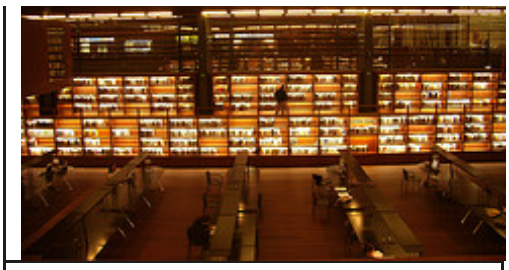


Imagen de microlito bajo licencia Creative Commons

ciudad.

¿Qué tipo de muestreo deben utilizar?

Deben utilizar el muestreo estratificado.

Sugerencia

Verdadero Falso

Falso
En este ejemplo no tiene sentido.

Deben utilizar el muestreo por conglomerados.

Sugerencia

Verdadero Falso

Verdadero
Aquí sí tiene sentido.

Imagen de joaoa bajo licencia Creative Commons

Volvamos al ejemplo del instituto "Benito V.".

Recuerda que queríamos tomar una muestra de 28 alumnos de los 560 que hay en el centro educativo y que el factor de elevación era de $k = \frac{560}{28} = 20$.

Numeramos a los alumnos del 1 al 560. Elegimos entonces un número ***x*** al azar entre 1 y 20 y ese será el primer alumno seleccionado, el que ocupa el lugar ***x***.

Luego tomamos el ***x+20***, ***x+2·20*** y así sucesivamente.

Este tipo de muestreo se le llama **sistemático** y no es aleatorio porque todas las muestras no son igualmente probables.

- El **muestreo sistemático** es equivalente al **muestreo aleatorio** si los elementos se encuentran numerados de manera aleatoria.

- El **muestreo sistemático** puede considerarse un caso particular del **muestreo por conglomerados**, estando cada uno de ellos formado por los siguientes elementos que ocupan en la lista el lugar:

Primer conglomerado: 1, 1 + k, 1 + 2·k, 1 + 3·k, 1 + 4·k, . . .

Segundo conglomerado: 2, 2 + k, 2 + 2·k, 2 + 3·k, 2 + 4·k, . . .

. . .

k-ésimo conglomerado: 20, 2·k, 3·k, 4·k, . . . n·k.

Seleccionar una muestra sistemática equivale a seleccionar al azar un único conglomerado. Para ello es necesario que cada uno de los conglomerados definidos tenga una composición similar a la población.

- También puede considerarse como un caso particular de **muestreo estratificado** con un número de estratos igual a ***n***, cada uno de ellos con ***k*** elementos de manera que en cada estrato se elige un único elemento.

En el muestreo estratificado el elemento seleccionado en cada estrato es aleatorio, mientras que en el sistemático se elige de forma aleatoria al primer elemento quedando los restantes determinados por el factor de elevación ***k***.

Importante

Decimos que un **muestreo es aleatorio sistemático** cuando seleccionamos los ***n*** elementos de la muestra de ***k*** en ***k*** siendo ***k*** el ***factor de elevación***:

$$k = \frac{N}{n}$$

N es el número de elementos de la población y ***n*** el tamaño de la muestra.

Ejercicio resuelto

Imagen de palm_z bajo licencia Creative Commons

Este año la vuelta ciclista a España tiene la meta en la ciudad donde está la oficina de TisBet Survey y además la organización quiere realizar un control antidopaje y han pedido ayuda a esta empresa para ver cómo deberían seleccionar a los ciclistas.

¿Cómo crees que deberían hacerlo?

Mostrar retroalimentación

Se elige al azar un ciclista al llegar a la meta y se le hace pasar el control.

A continuación se selecciona de *k* en *k* a los siguientes corredores a medida que pasan por la llegada, siendo *k* el factor de elevación.

Si el listado de los ciclistas se ha realizado al azar, el muestreo sistemático equivale al muestreo aleatorio simple.

Para saber más

Responde a las cuestiones que aparecen en la siguiente página de José Álvarez.

Enunciado

En un IES, en el que se imparte ESO y Bachillerato, hay 400 matrículas en ESO, de las que 175 son de alumnas,

Cuestiones

Contestemos a las siguientes cuestiones sobre esa situación:

1. El número total de alumnas que hay en el Instituto es... ?
2. En Bachillerato, el porcentaje de alumnos respecto del de alumnas es... ?

Vamos a realizar un muestreo aleatorio sistemático. Para ello numeramos todas las matriculas de la 1 a la 750.

3. El coeficiente de elevación es... ?

Para elegir el origen de entre los 15 primeros, se ha generado el número aleatorio x=0.568

4. El número de matrícula del origen es... ?

Queremos realizar un muestreo aleatorio estratificado, teniendo en cuenta el sexo y el nivel educativo.

5. En la muestra, estudiantes de la ESO habrá... ?
6. Alumnas de Bachillerato en la muestra habrá... ?

2. Muestras no aleatorias

A veces, hay situaciones donde el hacer un estudio muestral aleatorio o probabilístico resulta excesivamente costoso y se acude a métodos no aleatorios o no probabilísticos, aún siendo conscientes de que no sirven para realizar generalizaciones pues no se tiene certeza de que la muestra extraída sea representativa, ya que no todos los sujetos de la población tienen la misma probabilidad de ser elegidos.

En general se seleccionan a los sujetos siguiendo determinados criterios procurando, en la medida de lo posible, que la muestra sea representativa.

En este apartado veremos este tipo de muestreos no aleatorios.

En el siguiente vídeo puedes ver un resumen de lo que llevamos hasta ahora y lo que vamos a ver.

MUESTREO ERRÁTICO O CASUAL

Imagen de victorsmaria bajo licencia Creative Commons

Ha llegado el momento de votar en las elecciones municipales y a TisBet Survey le han encargado una encuesta "en boca de urna" que son las que se suelen hacer en momentos de elecciones.

El encargado de hacer la encuesta escoge el método de muestreo errático o casual, entrevistando a personas que acaban de votar al salir del colegio electoral.

Este método es de bajo costo, no necesita personal muy cualificado y se sacan conclusiones rápidamente, pero carece de validez externa y fiabilidad, además de que depende de los criterios arbitrarios del entrevistador para seleccionar a los entrevistados.

Comprueba lo aprendido tiple

En el instituto "Benito V." se va a realizar una encuesta para conocer los gustos lectores del alumnado.

Para ello se entrevistan a 1 de cada 5 alumnos que salen de la biblioteca el lunes.

¿Qué tipo de muestreo se debe hacer?

- Muestreo aleatorio simple
- Muestreo aleatorio sistemático
- Muestreo aleatorio estratificado
- Muestreo casual

Falso

Falso

Falso

Verdadero

Solution

1. Incorrecto
2. Incorrecto
3. Incorrecto
4. Opción correcta

MUESTREO INTENCIONADO O RACIONAL

El departamento de Matemáticas del instituto "Benito V." quiere hacer un estudio sobre las dificultades de aprendizaje que tienen los alumnos en esta asignatura, para ello es el propio departamento, ya que él mejor que nadie conoce a su alumnado, el que selecciona la muestra.

Este método tiene la ventaja de la rapidez y que sirve para la formulación de hipótesis, pero tiene el riesgo de que la muestra no sea representativa y carece de validez externa y fiabilidad para intentar generalizar hacia grupos mayores.

MUESTREO POR CUOTAS

La Facultad de Matemáticas le ha encargado a TisBet Survey que haga un estudio a sus alumnos para ver el número de horas que estudian al día y para ello le piden que cojan una muestra dentro del colectivo de estudiantes varones y que tienen coche propio.

Imagen de jlastras bajo licencia Creative Commons

Este tipo de muestreo es equivalente al muestreo estratificado, pero en este caso al entrevistador se le dan unos criterios de selección.

Son entrevistas dirigidas que proporcionan una información rápida en ausencia de muestras estratificadas pero no son muestras representativas ya que dependen de la elección que haga el entrevistador.

MUESTREO BOLA DE NIEVE

El índice de robos ha aumentado de forma considerable en el barrio donde está la oficina de TisBet Survey y parece que gran parte de ellos son para conseguir dinero para comprar droga, ya que el consumo ha aumentado en los últimos meses.

Para analizar el tema la empresa TisBet Survey decide hacer un estudio del número de drogadictos que hay en el barrio.

Como escoger una muestra no es fácil ya que los individuos son difíciles de identificar, la técnica consiste en localizar algunos drogadictos, los cuales conducen a otros y así sucesivamente va creciendo la "bola de nieve".

Este método está recomendado para hacer estudios sociológicos y analizar problemas psicopedagógicos.

El inconveniente es que necesita entrevistadores profesionales bien entrenados y que la interpretación de los resultados tiene problemas de fiabilidad.

Imagen de matesymas bajo licencia Creative Commons

Tipos de muestreo no aleatorio:

- **Muestreo errático o casual.**
- **Muestreo intencionado o racional.**
- **Muestreo por cuotas.**
- **Muestreo bola de nieve.**

Todos ellos ofrecen mayor rapidez en la obtención de datos que los muestreos aleatorios pero las muestras no son representativas con lo que la fiabilidad es menor que en los muestreos aleatorios.

Bola De Nieve from **moesmo**

3. Distribución muestral de proporciones

Existen ocasiones en las cuales no estamos interesados en la media de la muestra, sino que queremos investigar en la muestra la proporción de artículos defectuosos, la proporción de alumnos suspendidos o la proporción de pacientes que con una terapia determinada pierden el miedo a volar, por poner algunos ejemplos.

Para dar respuesta a estas situaciones vamos a utilizar lo que llamamos **distribución muestral de proporciones**.

Imagen de Kevin bajo licencia Creative Commons

Importante

Si te acuerdas de la unidad anterior, cuando en una población estudiábamos una determinada característica o variable que sólo podía tomar dos valores, éxito o fracaso, la población seguía una **distribución binomial**.

En esta población, la proporción de individuos que poseen esa característica la llamamos p y en todas las muestras de tamaño n que podamos extraer de la población, llamaremos $\hat{p}$ al porcentaje de individuos que tengan esa característica.

Los distintos valores de $\hat{p}$ que dependen de las muestras elegidas, dan lugar a una variable aleatoria que se representa por $\hat{P}$ y que se llama **estadístico**.

Llamamos **distribución muestral de proporciones** a la distribución de los valores de $\hat{P}$.

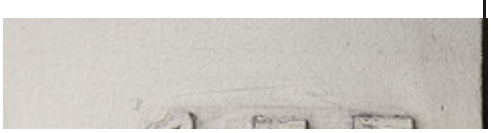
La variable aleatoria $\hat{P}$ tiene las siguientes características:

a) La media es: $\mu = p$.

b) La desviación típica es: $\sigma = \sqrt{\frac{p \cdot q}{n}}$, siendo $q = 1 - p$.

c) Para muestras donde **n≥30**, la distribución de $\hat{P}$ se aproxima a una distribución **normal** $N(p, \sqrt{\frac{p \cdot q}{n}})$.

Ejercicio resuelto



El profesor de Matemáticas lleva a la clase una bolsa con 4 tarjetas y en cada una de ellas aparece una cifra: 1, 2, 3, 5.

a) ¿Cuál es la proporción p de cifras pares?

b) Ahora vamos a coger dos tarjetas de las cuatro, pero cogemos una, la miramos y la volvemos a poner en la bolsa, es decir, estamos obteniendo muestras con reemplazamiento de tamaño dos. ¿Cuál será la proporción de cifras pares para cada una de las muestras?

c) Calcula la media y la desviación típica de la

Imagen de jfgornet bajo licencia Creative Commons

distribución muestral de proporciones.

Mostrar retroalimentación

a) La proporción de cifras pares es $\frac{1}{4}=0,25$, ya que el único par es el número 2 de las cuatro cifras que tengo.

b) La proporción de cifras pares de cada una de las muestras podemos verla en la siguiente tabla, teniendo en cuenta que podemos formar 4^2=16 muestras de tamaño 2.

Muestras	1-1	1-2	1-3	1-5	2-1	2-2	2-3	2-5	3-1
Proporción (p)	0	0,5	0	0	0,5	1	0,5	0,5	0

Por ejemplo, si en la primera tarjeta sacamos un 3 y en la segunda un 2, estamos en el caso 3-2. en este caso la proporción de n.º pares es $\frac{1}{2}$, pues de dos tarjetas, una es par.

Fíjate que hemos creado una nueva variable aleatoria, la variable que mide la proporción de tarjetas pares en una muestra de tamaño 2. Esta variable toma valores 0, 0,5 ó 1, y además, podemos saber su frecuencia, pues el 0 aparece 9 veces, el 0,5 6 veces y 1 una sola vez:

p_i	0	0,5
f_i	9	6

c) Para calcular la media de proporciones, recuerda que si X es una variable estadística que toma los valores x_1, x_2, x_3, ..., x_n con frecuencias absolutas f_1, f_2, f_3, ..., f_n , respectivamente, la media de la variable X viene dada por la siguiente expresión:

$$\overline{x}=\frac{x_1\cdot f_1+x_2\cdot f_2+\ldots+x_n\cdot f_n}{f_1+f_2+\ldots+f_n}=\frac{\sum_{i=1}^{n}x_i\cdot f_i}{\sum_{i=1}^{n}f_i}=\frac{\sum_{i=1}^{n}x_i\cdot f_i}{N}$$

En este caso tenemos

$$\overline{p}=\mu(p)=\frac{0\cdot 9+0,5\cdot 6+1\cdot 1}{16}=0,25$$

Para calcular la desviación típica, recuerda que tenemos que hacer la raíz cuadrada de la varianza, que es la media aritmética de los cuadrados de las desviaciones respecto de la media:

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

En nuestro caso tenemos que:

$$\sigma(p) = \sqrt{\frac{0^2 \cdot 9 + 0{,}5^2 \cdot 6 + 1^2 \cdot 1}{16} - (0{,}25)^2} = 0{,}3062$$

Como puedes observar, la media de la distribución muestral de proporciones coincide con la proporción de la población y la desviación típica con el valor señalado en el apartado "Importante" anterior.

Comprueba lo aprendido — Blanco

En un test de 5 preguntas, la variable "número de aciertos" con probabilidad p=0,50 se distribuye de la siguiente forma:

X_i	0	1	2	3	4	5
$f(x_i)$	0,031	0,156	0,312	0,312	0,156	0,031

Definimos la variable $P_i = \frac{X_i}{n}$, "Proporción de aciertos con probabilidad p".

El estadístico proporción (P) se distribuye mediante el modelo Binomial: B(n,p).

La distribución de la proporción de aciertos queda de la siguiente forma:

X_i	0	1	2	3	4	5
P_i	0	0,20	0,40	0,60	0,80	1,00
$f(x_i)$	0,031	0,156	0,312	0,312	0,156	0,031

Calcula:

a) La media y la desviación típica de P.

b) La probabilidad de que se acierten el 40% de las preguntas.

c) La probabilidad de que se acierten como máximo el 60% de las preguntas.

a) μ(p) = ☐

σ(P) = ☐

b) P(P_i = *0,40)* = ☐

c) P($P_i \leq 0,60$) = [ ]

Enviar

a) μ(p) = p; $\sigma(p)=\sqrt{\frac{pq}{n}}$, siendo $q=1-p=0,50$.

b) P($P_i = 0,40$) = P($X_i = 2$)

c) P($P_i \leq 0,60$) = P($X_i \leq 3$) = $P(X_i = 0) + P(X_i = 1) + P(X_i = 2) + P(X_i = 3)$

En el siguiente vídeo se te explica la distribución muestral de medias y proporciones.

Ejercicio resuelto

Imagen de wlappe bajo licencia Creative Commons

La empresa TisBet Survey, le ha encargado a una imprenta que le haga tarjetas de visita.

Esta imprenta suele imprimir un 3% de tarjetas defectuosas.

Si han encargado 500 tarjetas:

a) ¿Cuál es la probabilidad de que reciban más del 5% de tarjetas defectuosas?

b) ¿Cuál es la probabilidad de que reciban menos de un 1% de tarjetas defectuosas?

Mostrar retroalimentación

La distribución muestral de proporciones tiene como media y desviación típica:

Media	Desviación Típica
$\hat{p}$ = μ(p) = p = 0,03	$\sigma(p)=\sqrt{\frac{pq}{n}}=\sqrt{\frac{0,03\cdot 0,97}{500}}=0,0076$

Como el tamaño de la muestra es superior a 30, la distribución muestral se distribuye según la normal *N(0,03 ; 0,0076).*

Aplicando las propiedades del cálculo de probabilidades en la distribución normal, obtenemos:

a) P($\hat{p}$ >0,05) = 1 - P($\hat{p}$ ≤ 0,05) = 1 - $P(\frac{\hat{p}-0,03}{0,0076}\leq\frac{0,05-0,03}{0,0076})$ = 1 - P(Z ≤ 2,63) = 1-0,9957 = 0,0043.

b) P($\hat{p}$ <0,01) = $P(\frac{\hat{p}-0,03}{0,0076}<\frac{0,01-0,03}{0,0076})$ = P(Z< -2,63)= 1 - P(Z< 2,63)= 1 - 0,9957= 0,0043

Comprueba lo aprendido

Blanco

El gerente de la empresa TisBet Survey tiene pánico a volar.

Le han recomendado a un psicólogo que afirma que con su terapia para tratar "el miedo a volar en avión" se recupera el 80% de los pacientes.

Si seleccionamos al azar a 50 pacientes que han acudido a su consulta durante los 6 últimos meses por este tema, ¿cuál es la probabilidad de que al menos el 75% se hayan recuperado y puedan tomar aviones?

La media de la distribución muestral de proporciones es $\hat{p}$ = μ(p) = p = [ ] ; y la desviación típica

$\sigma(p)=\sqrt{\frac{pq}{n}}=\sqrt{\frac{0,80\cdot 0,20}{50}}=0,0032$.

Como el tamaño de la muestra es superior a 30, la distribución muestral se distribuye según la normal [ ] . *(Separa la media de la desviación típica con un ; y sin espacio)*

La probabilidad que se pide es:

P($\hat{p}$ ≥ [ ])= [ ] - P($\hat{p}$ < [ ]) = 1 - $P(Z<\frac{0,75-0,80}{0,0032})$ = 1 - P(Z< [ ]) = 1-(1-P(Z< [ ]))≈ [ ] .

Enviar

Para saber más

Para terminar este apartado te dejo con esta presentación de eraperez.

Distribucion muestral de proporciones from **eraperez**

4. Distribución muestral de la media

Cuando estamos haciendo un estudio sobre una población, por ejemplo, la satisfacción de los usuarios del metro de Sevilla valorados en una puntación entera de 0 a 100, lo que solemos hacer es escoger una muestra de 30 personas, por ejemplo, y reducir el estudio a esa muestra.

Pero si para esa muestra calculamos su media $\overline{x}$ y su desviación típica **s**, obtendremos dos valores que pueden estar o no estar próximos a los valores de la media **μ** y de la desviación típica **σ** de la población, pues lo hemos hecho sólo para una muestra. Si repetimos el proceso sucesivamente con más muestras, lo lógico es que obtengamos valores distintos para esa media y esa desviación típica, ¿verdad?:

Muestra 1: $\overline{x}_1$ y $\mathbf{s_1}$

Muestra 2: $\overline{x}_2$ y $\mathbf{s_2}$

Muestra 3: $\overline{x}_3$ y $\mathbf{s_3}$

.....

¿Con cuál nos quedamos?

Los distintos valores de las medias de cada muestra $\overline{x}_i$ dan lugar a una nueva variable aleatoria, que por ejemplo, podemos representar así: $\overline{X}$. La distribución de los valores de $\overline{X}$ se llama **distribución en el muestreo de la media**.

Esta nueva variable estadística, llamada **Estadístico**, tendrá su propia media $\mu_{\overline{X}}$ y su propia desviación típica $\sigma_{\overline{X}}$.

Veamos este vídeo del profesor de Bioestadística D. Francisco Javier Barón López de la Facultad de Medicina de la Universidad de Málaga que nos aclara bastante esta situación. Por cierto, las aplicaciones de las que habla al principio de vídeo serán las que veamos en la próxima unidad.

Importante

La **distribución en el muestreo de la media** $\overline{X}$ tiene las siguientes características:

a. Media: Tiene la misma media que la población **μ.**

b. Desviación típica: La desviación típica de esta distribución es $\frac{\sigma}{\sqrt{n}}$, siendo n el tamaño de las muestras.

c. Si la población no sigue una distribución normal, pero **n ≥ 30**, la distribución de las medias muestrales se aproxima a una distribución **normal**, esta aproximación será mejor cuanto mayor sea n.

Ejercicio resuelto

La empresa Sal Marina, S.A. comercializa sal que empaqueta en bolsas de 500 gramos. Se sabe que los pesos reales de las bolsas siguen una distribución normal de media 498 gr.y desviación típica 8 gr.

Si se toma al azar una muestra de 30 bolsas:

a) ¿Cuál es la distribución en el muestreo de la media?

b) Halla la probabilidad de que la media de la muestra sea mayor de 500 gr.

Mostrar retroalimentación

a) Como n ≥ 30, la distribución de las medias muestrales se aproxima a una distribución normal $N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$. En concreto:

$$N\left(498, \frac{8}{\sqrt{30}}\right) = N(498\ ;\ 1{,}46)$$

b) Nos están pidiendo la P($\overline{X}$ >500). Como $\overline{X}$ sigue una $N(498\ ;\ 1{,}46)$, vamos a calcular su valor tipificando la variable:

$$P(\overline{X}>500) = P\left(Z>\frac{500-498}{1{,}46}\right) = P\left(Z>\frac{2}{1{,}46}\right) =$$

$$= P(Z>1{,}37) = 1-P(Z\leq 1{,}37) = 1-0{,}9147 = 0{,}0853$$

O sea, 8,53%, por lo que en 9 de cada 100 veces (casi 1 de cada 10 veces) ocurrirá que la media de la muestra es mayor de 500 gr.

Comprueba lo aprendido

Blanco

La misma empresa empaqueta también bolsas de 1 kilo, aunque la realidad marca que el peso medio de esas bolsas es 1,01 kg y la desviación típica 9 gramos.

Si sacamos una muestra de 50 bolsas de sal de un kilo, su peso medio expresado en kilogramos se distribuye según una normal N([ ] ; [ ]) *(usa tres decimales para la desviación típica)*.

Además, en esa muestra, la probabilidad de que el peso medio sea menor que 900 gramos es [ ] .

Enviar

900 gr = 0,9 kg.

$$P(\overline{X}<0{,}9) = P\left(Z<\frac{0{,}9-1{,}01}{1{,}273}\right) = P(Z<-0{,}09) = 1-0{,}5359 = 0{,}4641$$

4.1. Distribución de las sumas muestrales

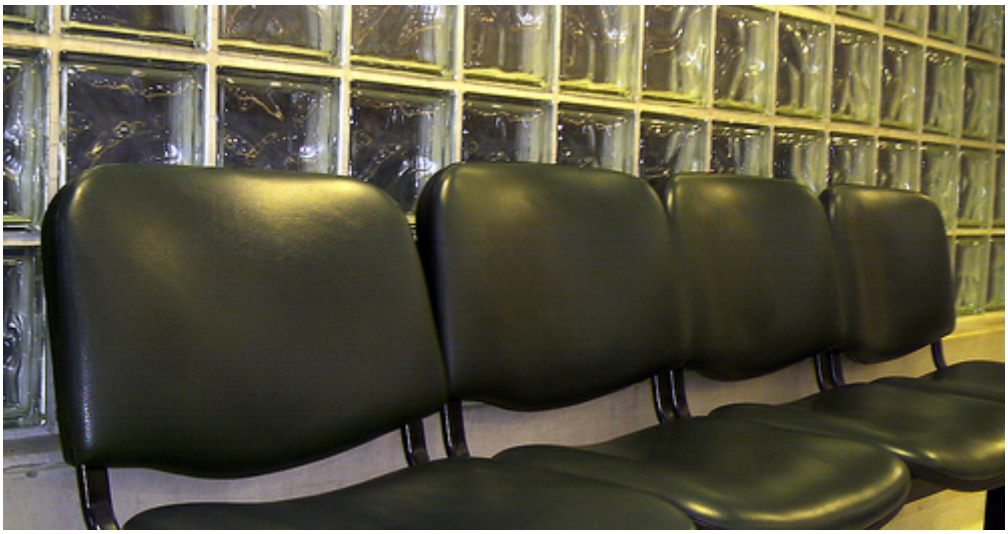


Imagen de medea_material con licencia Creative Commons

Nuestra empresa TisBet Survey ha realizado un trabajo sobre el tiempo que duran las consultas de un determinado médico. Tras el estudio han concluido que las consultas duran una media de 8 minutos con una desviación típica de 2,3 minutos.

Supongamos que una mañana nuestro médico tiene 38 pacientes citados. ¿Cuál es la probabilidad de que termine antes de que transcurran 5 horas?

Fíjate que, a diferencia del apartado anterior, ahora no me piden la media de una consulta, sino que me están pidiendo cuánto duran de media las 38 consultas. Si tomamos una muestra de 38 personas y realizamos el experimento obtenemos:

1º La suma de todos los tiempos, que es t_1. Por ejemplo t_1=308 minutos.

2º Si repetimos el proceso obtendremos t1, t2, t3,... estos valores dan lugar a una nueva variable aleatoria **T**.

T es un estadístico cuyos valores siguen la **distribución de las sumas muestrales** de la variable "Tiempo medio de una consulta"

La variable aleatoria T tiene las siguientes características:

a. Media: n·μ, donde n es el número de individuos de la muestra.

b. Desviación típica : $\sigma \cdot \sqrt{n}$

c. Si la población no sigue una distribución normal, pero **n ≥ 30**, la distribución de T se aproxima a una **normal** $N(n \cdot \mu\ ,\ \sigma \cdot \sqrt{n})$.

En nuestro ejemplo, T="Tiempo que tarda en atender a 38 pacientes" sigue una distribución normal:

$$N(38 \cdot 8\ ;\ 2{,}3 \cdot \sqrt{38}) \ = \ N(304\ ;\ 14{,}18)$$

.

Por lo tanto, nos están pidiendo la probabilidad de que T<300. Calculémosla, tipificando la variable T.

$$P(T>300) = P\left(Z>\frac{300-304}{12,18}\right) = P(Z>-0,33) = P(Z\leq 0,33) = 0,6293$$

Podemos llegar a la conclusión que 6 de cada 10 veces ocurrirá esto, es decir, la consulta tardará más de 5 horas.

Reflexiona

Imagen de marcelo träsel con licencia Creative Commons

Recuerdas nuestro ejemplo de la empresa Sal Marina, S.A. del punto anterior:

La empresa Sal Marina, S.A. comercializa sal que empaqueta en bolsas de 500 gramos. Se sabe que los pesos reales de las bolsas siguen una distribución normal de media 498 gr.y desviación típica 8 gr.

Si las bolsas de 500 gr. se empaquetan en cajas de 50 y 100 unidades.

a) ¿Qué distribución siguen los pesos de las cajas de 50 unidades?

b) ¿Qué distribución siguen los pesos de las cajas de 100 unidades?

c) Si se coge al azar una caja de 50 unidades, ¿cuál es la probabilidad de que el peso de las 50 bolsas de sal supere los 25 kg.?

d) Si se coge al azar una caja de 100 unidades, ¿cuál es la probabilidad de que el peso de las 100 bolsas sea mayor de 50 kg.?

Mostrar retroalimentación

a) Como la distribución de las bolsas de sal siguen una distribución normal, la de la suma de 50 bolsas también con una media 50·μ y una varianza de $\sqrt{50}\cdot\sigma$. Si llamamos T_{50} a la variable de la suma de los pesos medios de 50 bolsas de sal, esta variable sigue una distribución normal N(24900 ; 56,57).

b) Como la distribución de las bolsas de sal siguen una distribución normal, la de la suma de 100 bolsas también con una media 100·μ y una varianza de $\sqrt{100}\cdot\sigma$. Si llamamos T_{100} a la variable de la suma de los pesos medios de 100 bolsas de sal, esta variable sigue una distribución normal N(49800, 80).

c) Traducido a probabilidad lo que nos está pidiendo el ejercicio es $P(T_{50} > 25000)$. Vamos a tipificar la variable para calcular el valor de esta probabilidad con ayuda de las tablas, ya que T_{50} sigue una distribución N(24900 ; 56,57):

$$P\left(Z>\frac{25000-24900}{56,57}\right) = P(Z>1,77) = 1-P(Z\leq 1,77)=1-0,9616=0,0384$$

Es decir, aproximadamente 4 de cada 100 muestras superarán estos 25 Kg.

d) Ahora nos están pidiendo es la $P(T_{100}>50000)$. De nuevo sabemos que T_{100} sigue una distribución N(49800, 80), tipificamos la variable para resolver esta probabilidad:

$$P\left(Z>\frac{50000-49800}{80}\right) = P(Z>2,5) = 1-P(Z\leq 2,5)=1-0,9938 = 0,0062$$

Ahora vemos que podemos afirmar que sólo 6 de cada 1000 cajas de 100 bolsas superará los 50 kg.

4.2. Distribución de la diferencia de las medias

Imagen de Anton Fomkin con licencia Creative Commons

Vemos la importancia que tiene el cálculo de la distribución de las medias. Pero aún hay más.

Veamos ahora un ejemplo que nos va a permitir comparar dos tipos de bombillas de bajo consumo de distintas marcas, pero de similares características.

Las bombillas de la marca A tienen una vida media de 11000 horas y una desviación típica de 1500 horas, mientras que las bombillas de la marca B tienen una vida media de 10000 horas y una desviación típica de 750 horas.

Supongamos que nos piden comparar las dos marcas. Cuál es la probabilidad de que una bombilla de la marca A dure mil horas más que una bombilla de la marca B.

Sean $\bar{x}_A$ y $\bar{x}_B$ sus respectivas medias muestrales.

Para comparar las dos marcas se considera la diferencia de sus medias, que se estima con la diferencia de las medias muestrales: $\bar{x}_A - \bar{x}_B$

Si se realiza este mismo proceso para otras muestras formadas por n_A bombillas de la marca A y n_B bombillas de la marca B. Se irán obteniendo sus respectivas diferencias de las medias:

$\bar{x}_{A1} - \bar{x}_{B1}$, $\bar{x}_{A2} - \bar{x}_{B2}$, $\bar{x}_{A3} - \bar{x}_{B3}$, ...

Estos distintos valores dan lugar a una variable aleatoria que se representa por $\overline{X}_A - \overline{X}_B$.

La distribución de los valores de $\overline{X}_A - \overline{X}_B$ se llama distribución en el muestreo de la diferencia de las medias.

Importante

La distribución en el muestreo de la diferencia de las medias $\mu_A - \mu_B$ tiene las siguientes características:

a. Media: La media de la distribución es la diferencia de medias: $\mu_A - \mu_B$

b. Desviación típica: La desviación típica es $\sigma = \sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}$

c. Si la población no sigue una distribución normal, pero **$n_A \geq 30$ y $n_B \geq 30$**, la distribución $\overline{X}_A - \overline{X}_B$ se aproxima a una **normal** $N\left(\mu_A - \mu_B, \sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}\right)$

Ejercicio resuelto

Vamos a resolver el problema con el que habíamos comenzado este apartado.

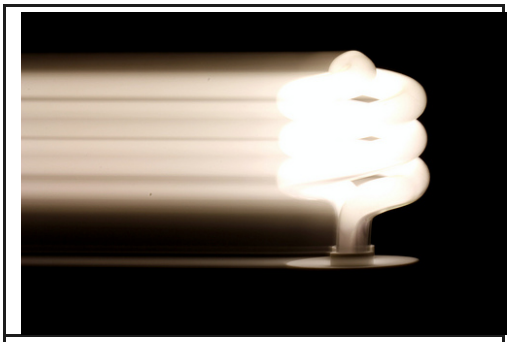


Imagen de reillybutler con licencia Creative Commons

Supongamos que cogemos muestras de 50 bombillas de cada tipo. La variable estadíastica $\overline{X}_A - \overline{X}_B$ sigue una distribución N(1000 ; 237,17), pues la diferencia entre las medias de A y B es 1000 y la desviación típica aplicando la fórmula sería:

$$\sigma = \sqrt{\frac{1500^2}{50} + \frac{750^2}{50}} = \sqrt{56250} = 237{,}17$$

Para calcular la probabilidad de que una bombilla de la marca A dure 1000 horas más que una bombilla de la marca B sólo tendremos que calcular $P(\bar{X}_A - \bar{X}_B > 1000)$. Como sabemos que la variable tiene una distribución normal, vamos a calcular su valor tipificando la variable:

$$P(\overline{X}_A - \overline{X}_B > 1000) = P\left(Z > \frac{1000-1000}{237{,}17}\right) = P(Z>0) = 1 - P(Z \leq 0) = 1 - 0{,}5 = 0{,}$$

Es decir, una de cada dos bombillas de la clase A durará 1000 horas más que una del tipo B.

Si calculamos la probabilidad de que una de las bombillas de la marca B dure más que una de la marca A, tendremos que calcular $P(\overline{X}_A - \overline{X}_B < 0)$

Tipificando la variable tenemos:

$$P(\overline{X}_A - \overline{X}_B < 0) = P\left(Z < \frac{0-1000}{237{,}17}\right) = P(Z<-4{,}32) = P(Z>4{,}32) = 1 - P(Z \leq 4{,}32) = 1 - 1 =$$

En conclusión, según el estudio que acabamos de hacer, una bombilla de la marca A durará más que una bombilla de la marca B.

Reflexiona

Una empresa ha contratado a TisBet Survey para que haga un estudio sobre la diferencia de los sueldos entre sus empleados. Están interesados en conocer si hay diferencia entre el salario medio de un hombre y de una mujer.

Para ello se han tomado una muestra aleatoria de 60 mujeres y de 40 hombres obteniendo los resultados que figuran en la tabla inferior:

Imagen de amboo who? con licencia Creative Commons

	Hombres	Mujeres
Media	1230 €	1025 €
Desviación Típica	160 €	75 €

Ayuda a nuestros amigos calculando la distribución en el muestreo de la diferencia de medias muestrales y la probabilidad de que la diferencia de las medias entre los salarios de los hombres y de las mujeres sea superior a 200 €.

Mostrar retroalimentación

Como en ambos casos n >30, podemos concluir que la distribución de la diferencia de las medias sigue una distribución Normal. Los parámetros de esta variable serán:

Media: Es igual a la diferencia de las medias 1230-1025=205 €

Desviación típica: $\sigma = \sqrt{\frac{\sigma_H^2}{n_H} + \frac{\sigma_M^2}{n_M}} = \sqrt{\frac{160^2}{40} + \frac{75^2}{60}} = \sqrt{733{,}75} = 27{,}09$

Por lo tanto $\overline{X}_H - \overline{X}_M \sim N(205 \ ; \ 27{,}09)$

b) Tendremos que calcular $P(\overline{X}_H - \overline{X}_M \geq 200)$.

Como $\overline{X}_H - \overline{X}_M \sim N(205 \ ; \ 27{,}09)$, para calcular esta probabilidad tipificaremos la variable y usaremos las tablas para ver su valor.

$$P(\overline{X}_H - \overline{X}_M \geq 200) = P\left(Z > \frac{200-205}{27{,}09}\right) = P\left(Z > \frac{-5}{27{,}09}\right) = P(Z > -0{,}18) = P(Z < 0{,}18) = 0{,}5714$$

Por lo tanto podemos afirmar que casi el 58 por ciento de las veces que hagamos el muestreo en esta empresa, el salario medio de los hombres superará al de las mujeres en más de 200 €.

5. Teorema central del límite

¿Ya te has familiarizado con esto de las distribuciones muestrales?

Parece todo un poco extraño, ¿verdad? Una distribución dentro de otra, un parámetro se convierte en variable, la media de las medias,... Uff, ¡qué lío!

Imagen de kadluba bajo licencia Creative Commons

Bueno, no tanto es así, y es que como te puedes imaginar, esto no sale de la chistera de un mago cual conejo saltarín, sino que todo tiene un fundamento, y eso precisamente es lo que vamos a ver en este apartado, aunque ya algo te han adelantado en el vídeo del apartado 2.

Lo que acabas de ver en los apartados anteriores de este tema es importantísimo, pues ten en cuenta que la distribución de la variable X se asocia a toda la población y conocer una característica de toda la población es complicado. De muestras no hay excesivos problemas, pero de toda la población... Pues bien, en la siguiente unidad, vamos a ver cómo se hace el salto de pasar la información de una muestra a toda la población, pero para ello son esenciales los resultados que acabas de ver.

De momento, sabemos la distribución que sigue la muestra (la media, la suma o diferencia de medias o la proporción) a partir de la distribución de la población, y en la próxima unidad, utilizando lo que acabamos de estudiar en estos apartados, veremos el paso contrario, que es lo verdaderamente interesante.

Por cierto, ¿no te ha resultado curioso que todo se aproxime a una normal? Sí, otra vez la distribución normal.

Como vamos a ver en el siguiente teorema, bajo ciertas condiciones, las cosas pueden parecerse mucho a una normal, y como recordarás, calcular probabilidades en una normal era bastante fácil.

¡Ah! este teorema es uno de los más importantes, o el que más quizás, de la estadística y la probabilidad, y todavía no lo hemos dicho, pero se llama **Teorema Central del Límite**.

El **Teorema Central del Límite** nos indica que si tenemos una serie de variables aleatoria independientes (el valor de una no influye en la otra) e idénticamente distribuidas (todas las variables tienen la misma distribución y por tanto, los mismos parámetros), la distribución de la **suma** de esas **variables** (si el número de variables que se suman es suficientemente grande) se **aproxima a una Distribución Normal**.

Lo precisamos todavía más, porque no se aproxima a una normal cualquiera sino que podemos saber a cuál:

Si tenemos n variables aleatorias X_1, X_2, X_3,..., X_n todas ellas independientes entre sí y todas ellas con media μ y desviación típica σ, la suma de esas variables genera una nueva variable aleatoria que se aproximará a una distribución Normal de media $n \cdot \mu$ y desviación típica $\sqrt{n} \cdot \sigma$

$$S_n = X_1 + X_2 + ... + X_n = \Sigma X_i \sim N(n \cdot \mu, \sqrt{n} \cdot \sigma)$$

La coletilla de si el número es suficientemente grande no es ninguna tontería, pues lo que en realidad se aproxima es el límite de la suma de las variables cuando el número de variables tiende a infinito. O sea, que para que esto funcione, el número de variables o de datos que se han de sumar tiene que ser grande. A efectos prácticos nos vale con que al menos haya 30 datos, es decir, para aplicar este teorema tiene que cumplirse que **n ≥ 30**.

Ojo, fíjate que en ningún momento estamos diciendo que la variable X tenga que ser una distribución normal, sino que sea lo que sea la distribución de la variable en la población, la suma de muchas observaciones se va a aproximar a una normal

la suma de muchas observaciones se va a aproximar a una normal.

Si ya de por sí la **población** de partida sigue una distribución **normal**, ese resultado se cumple siempre, sea el tamaño el que sea. **No importa el valor de n.**

Ejercicio resuelto

Según el estudio realizado por TisBet Survey, el número de días que un habitante de Lanjarón sale a pasear en bicicleta por la sierra sigue una distribución Normal de media 6,7 días y desviación típica 2,1. ¿Es muy probable que si juntamos 10 personas, entre las 10 hagan más de 80 días?

Mostrar retroalimentación

Piensa un momento, fíjate que la respuesta que de una persona es una variable aleatoria que sigue una distribución normal N(6,7;2,1), pues esa persona está dentro de la localidad objeto de estudio. Como le preguntamos a 10 personas, tenemos 10 variables aleatorias, todas con la misma distribución normal, que vamos a llamar X_1, X_2, X_3,... X_{10}.

Puesto que queremos saber una probabilidad sobre la suma de las respuestas que han dado las diez personas, hemos de encontrar la distribución de la variable $S_{10} = X_1 + X_2 + ...+X_{10}$, y para ello, aplicamos el teorema que acabamos de ver. (Llamamos S_{10} a la suma de las 10 variables para abreviar un poco la expresión y no escribir continuamente $X_1+X_2+...$)

En primer lugar, vemos que se cumplen las condiciones del teorema:

- Las variables son independientes, pues la encuesta se habrá hecho de forma que la respuesta de una persona no influya en la de otra. (¡Ves lo importante que hacer bien la encuesta!)
- Las variables están idénticamente distribuidas, pues todas las personas están dentro de la misma población y por tanto, todas siguen una distribución normal N(6,7;2,1).
- El tamaño de la muestra no es superior a 30, pero como la población ya de por sí es normal, no importa el tamaño.

Entonces aplicando el resultado anterior, podemos afirmar que: $S_{10} \sim N(10 \cdot 6{,}7 \ ; \ \sqrt{10} \cdot 2{,}1)$, o lo que es lo mismo, $S_{10} \sim N(67 \ ; \ 6{,}64)$

Ahora ya podemos calcular la probabilidad de la misma forma que lo hacíamos en el tema 4 de la unidad anterior:

$P(S_{10} > 80) = P(Z > 1{,}96) = 1 - 0{,}9750 = 0{,}025$.

Así que no, no es demasiado probable que entre las 10 personas salgan más de 80 días a pasear en bicicleta.

Sobre un examen de 10 preguntas tipo test, el número de preguntas acertadas sigue una media de 3,3 y una desviación típica de 1,49. ¿Qué distribución sigue la suma de las preguntas acertadas de 40 exámenes? ¿Habrá más de 100 preguntas acertadas entre todos?

Mostrar retroalimentación

Podemos entender que los 40 exámenes corresponden a 40 alumnos distintos, así, llamamos X_1 al número de preguntas acertadas en el primer examen, X_2 al del segundo y así sucesivamente X_{40} al n.º de respuestas acertadas en el cuadragésimo examen. Todas ellas tienen una distribución de probabilidad de la que sabemos su media es 3,3 y su desviación típica 1,49.

Imagen de knittymarie bajo licencia Creative Commons

- Las variables son independientes, pues el número de respuestas de un examen no influye en el otro.
- Las variables son idénticamente distribuidas; tienen todas la misma media y la misma distribución.
- La muestra tiene un tamaño suficiente, pues n = 40 y por tanto mayor que 30.

Por tanto, podemos aplicar el Teorema Central del Límite y obtenemos entonces que la variable aleatoria resultante de sumar las puntuaciones de las cuarenta variables sigue una distribución normal de media $\mu = 40 \cdot 3{,}3 = 132$ y una desviación típica $\sigma = \sqrt{40} \cdot 1{,}49 = 9{,}42$

$$S_{40} = X_1 + X_2 + ... + X_{40} \sim N(132 ; 9{,}42)$$

La segunda pregunta nos pide la probabilidad de que entre todos sumen más de 100 preguntas acertadas, luego tenemos que calcular $P(S_{40} > 100)$.

Como siempre, el primer paso tipificar:

$$P(S_{40}>100) = P(Z>\frac{100-132}{9{,}42}) = P(Z>-3{,}4)$$

Y por último terminamos calculando la probabilidad haciendo el cambio pertinente con las reglas de cálculo de probabilidades en la distribución normal y buscando la probabilidad en la tabla de probabilidades de la normal N(0,1).

$$P(S_{40}>100) = P(Z>\frac{100-132}{9{,}42}) = P(Z>-3{,}4) = P(Z<3{,}4) = 0{,}99966$$

Así que, es prácticamente seguro que se van a superar esas 100 preguntas acertadas.

Comprueba lo aprendido

tiple

1) Una variable aleatoria tiene una media de 12 unidades y una desviación típica de 2 unidades. La suma de las puntuaciones de 20 observaciones tiene aproximadamente una distribución:

- N(12 ; 40)
- N(240; 28,28)
- N(240; 40)
- No se puede determinar.

No es correcto ninguno de los dos parámetros

Aunque los cálculos son correctos, no se puede aplicar el teorema al no cumplir la condición del tamaño de la muestra.

No es correcto, en todo caso, estaría mal el cálculo de la desviación típica.

Correcto, no cumple el requisito del tamaño.

Solution

1. Incorrecto
2. Incorrecto
3. Incorrecto
4. Opción correcta

2) El tiempo que se tarda en encontrar aparcamiento en un parking sigue una determinada distribución de probabilidad en la que la media es 5 minutos y la desviación típica es 0,8. La distribución del tiempo total invertido por 30 coches elegidos al azar de entre los que entran en el parking sigue una distribución:

- N(5 ; 0,8)
- N(150 ; 26,8)
- N(150; 24)
- No puede determinarse.

No es correcto, lee bien el enunciado.

Muy bien.

No es correcto, hay que hacerle la raíz cuadrada al número de datos.

Falso. Sí se puede, pues se supone que los coches se eligen al azar (independientes), el parking es el mismo (igual distribución) y hay un número suficiente de datos n = 30.

Solution

1. Incorrecto
2. Opción correcta
3. Incorrecto
4. Incorrecto

3) Las sardinas que llegan a una lonja tienen un peso que se distribuye según una normal N(197 ; 48), donde el peso se expresa en gramos. Un distribuidor las empaqueta en bolsas de 12 unidades. ¿Qué distribución sigue el peso de las bolsas de sardinas?

Imagen de FreeCat bajo licencia Creative Commons

- N(197 ; 489)
- N(2364 ; 166,3)
- N(2364 ; 576)
- No se puede determinar.

No es correcto. Ese es el peso de una sardina.

¡Perfecto!

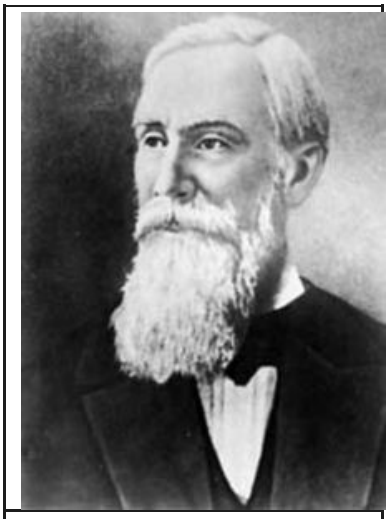
No es correcto. Repasa el cálculo.

Falso. Sí se puede pues la población de partida ya es de por sí normal, así que no importa el tamaño de la muestra.

Solution

1. Incorrecto
2. Incorrecto
3. Opción correcta
4. Incorrecto

Curiosidad



PL Chebishev. Imagen en Wikimedia Commons bajo licencia Creative Commons

Como ya sabes, la distribución normal tuvo su precedente en la binomial que Bernouilli desarrolló es su conocido Teorema Áureo, una función que años más tarde Poisson bautizaría como Ley de los Grandes Números. En 1733, el matemático francés **De Moivre** hizo una generalización de dicho teorema. No sólo fue el primero en obtener la característica forma de campana de la función, sino que también conjeturó, en 1733, el Teorema Central del Límite. Éste fue un resultado que, a pesar de no haber sido demostrado de forma rigurosa, fue aceptado durante mucho tiempo.

Una primera formulación clara del teorema no apareció hasta 1812, fecha en que **Laplace** llevó a cabo los primeros intentos de demostración. Aunque realmente el primero en iniciar un estudio riguroso fue el matemático ruso **P. L. Chebyshev**, siendo sus alumnos Markov, y especialmente Lyapunov, quienes resolvieron la cuestión de manera definitiva en 1901. Aunque la demostración completa del enunciado, tal y como lo conocemos actualmente, vino de la mano del matemático finlandés **Jarl Waldemar Lindeberg** (1876-1932) en 1930. Dicho enunciado dice: "La suma de un gran número de variables aleatorias independientes sigue aproximadamente una distribución normal".

Pocas conjeturas se han mantenido como ciertas durante tanto tiempo y con el absoluto beneplácito de la comunidad matemática como la del Teorema Central del Límite. Ello fue debido, por una parte, a que nadie dudaba de su veracidad, y por otra, a la enorme utilidad que representaba su afirmación. La mayoría de los fenómenos que se dan en la naturaleza y en las sociedades humanas siguen una distribución normal. Se aplica de la misma forma para establecer sondeos electorales o sondeos petrolíferos. La psicología se sustenta como ciencia gracias a las medidas que establece en torno a parámetros como determinadas percepciones sensoriales o cocientes intelectuales. Todas las teorías que se construyen en torno a dichos resultados se vendrían abajo si el Teorema Central del Límite no fuera cierto.

Actividad de lectura

En los apartados anteriores has visto que la distribución de la media muestral es $\overline{X} \sim N(\mu\,;\frac{\sigma}{\sqrt{n}})$ ¿Por qué es esto así? Pues fácil, a partir del Teorema Central del Límite.

Fíjate que acabamos de ver que la suma de las n variables, con las condiciones del Teorema, sigue una distribución Normal:

$$\Sigma X_i \sim N(n\mu\,;\sqrt{n}\sigma)$$

La media y la desviación típica de cualquier variable aleatoria cumplen que si la variable se multiplica por un número real, estos parámetros quedan también multiplicados por

dicho número. Por ejemplo, imagina una variable aleatoria A que tenga de media 2 y desviación típica 0,4. La media de la variable aleatoria 3·A sería 3·2, o sea, 6, y la desviación típica 3·0,4 o lo que es lo mismo 1,2.

Bien, pues si recuerdas la fórmula de la media, ésta era $\overline{X}=\frac{\Sigma X_i}{n}$, es decir, la nueva variable ΣX_i la dividimos entre el tamaño de la muestra "n", o lo que es lo mismo, la multiplicamos por $\frac{1}{n}$.

Si aplicamos la propiedad que acabamos de ver, la media muestral cumplirá que:

Media	Desviación típica
$\frac{1}{n}\cdot n\cdot\mu = \frac{n\cdot\mu}{n} = \mu$	$\frac{1}{n}\cdot\sqrt{n}\cdot\sigma = \frac{\sqrt{n}}{n}\cdot\sigma = \frac{\sigma}{\sqrt{n}}$

Luego efectivamente, se cumple que $\overline{X}\sim N(\mu\,;\frac{\sigma}{\sqrt{n}})$.

Las poblaciones normales

En el caso de que los datos vengan de una población normal, hemos dicho que es irrelevante el número de datos. No es necesario que haya 30 o más. Esto es debido a la **propiedad reproductiva** de la distribución Normal.

Esta propiedad nos dice que si hay dos variables aleatorias independientes que siguen distribuciones de probabilidad normales, la suma de ellas es también una distribución Normal con media la suma de las medias y varianza la suma de las varianzas.

Si suponemos que nuestra variable X~N(μ,σ), como cada X_i también tiene esa distribución, la suma de las n observaciones tendrá:

Media	Varianza
μ+μ+...+μ = n·μ	$\sigma^2+\sigma^2+...+\sigma^2 = n\cdot\sigma^2$

Y por tanto la desviación típica sería $\sqrt{n\sigma^2} = \sqrt{n}\cdot\sigma$.

Es por esto entonces por lo que se cumple que $\Sigma X_i\sim N(n\mu\,;\sqrt{n}\sigma)$ sea n el valor que sea y por tanto también, $\overline{X}\sim N(\mu\,;\frac{\sigma}{\sqrt{n}})$.

Esta propiedad de la reproductividad, es la que se utiliza para justificar también la distribución muestral de la suma de medias y diferencia de medias.

Mostrar retroalimentación

Si quieres saber cómo se obtiene la distribución de la proporción muestral, sigue este enlace.

Y en general, si quieres saber más sobre el Teorema Central del Límite, definición formal, propiedades, etc., sigue este otro enlace.

5.1.Teorema de De-Moivre

No hace mucho, a las oficinas de TisBet Survey llegó la Concejalía de Urbanismo del Ayuntamiento de Sevilla para que les hiciera un estudio sobre la satisfacción de los ciudadanos con el servicio de transporte público. No sé si lo sabes, pero hace escasamente dos años (abril de 2009) se inauguró la primera línea de Metro, y una de las peticiones a TisBet es que analicen el nivel de uso y satisfacción que tienen los ciudadanos y ciudadanas de este nuevo medio de transporte.

Imagen de dezpiadoz bajo licencia Creative Commons

La primera pregunta del estudio es clara, ¿Ha utilizado usted el Metro alguna vez? Y la respuesta es evidente, sí o no. Sólo dos posibilidades.

¿Te suena esto de dos posibilidades? Dos posibilidades..., éxito-fracaso... Exacto: Bernoulli y la binomial.

Fíjate bien. Hacemos 100 encuestas y para cada una de ellas, definimos una variable X_i (i=1, 2,... 100) a la que le damos el valor 1 si la respuesta es afirmativa y 0 en caso contrario. Además, llegamos a la conclusión de que la probabilidad de que una persona haya utilizado el Metro es de 0,31, luego la probabilidad de éxito es 0,31.

La suma de todas las puntuaciones de las variables nos da el número de respuestas afirmativas, o lo que es lo mismo, el número de éxitos obtenidos en las 100 repeticiones. Y esto si lo recuerdas, nos daba pie a una distribución binomial, en concreto, a la distribución binomial B(100; 0,31). Por tanto, si definimos una variable aleatoria S que sea la suma de los puntos de las 100 encuestas, obtenemos que:

$$S=\Sigma X_i \sim B(100;\ 0,31)$$

Todo esto de la binomial lo repasamos en el tema 3 de la unidad anterior, y como recordarás, calcular probabilidades era fácil pero un poco pesado. Imagínate que queremos calcular la probabilidad de que más de 60 personas hayan cogido el Metro. Tendríamos que calcular P(S=61), P(S=62), P(S=63),... y así hasta P(S=100). Una auténtica pesadez.

¡Vamos a utilizar el Teorema Central del Límite!

Como siempre, primero hemos de ver que se cumplen las condiciones:

- Las variables X_i i=1,... 100 son independientes, pues la muestra la hacemos de forma que la respuesta de uno no influye en la de otro.
- Las variables X_i i=1,... 100 están idénticamente distribuidas, pues todas son variables de Bernoulli con probabilidad de éxito 0,31.
- El tamaño es suficiente, pues hay 100 variables y tiene que haber más de 30.

Luego las hipótesis se cumplen.

Por otro lado, cada variable X_i hemos dicho que es una Bernoulli de parámetro 0,31. Por tanto, la media es $\mu = p = 0,31$ y la desviación típica $\sigma = \sqrt{p(1-p)} = \sqrt{0,31 \cdot 0,69} = \sqrt{0,2139} = 0,462$

Por tanto aplicando el Teorema Central del Límite, obtenemos que:

$$S = \Sigma X_i \sim N(100 \cdot 0,31\ ;\ \sqrt{100} \cdot 0,462) = N(31\ ;\ 4,62)$$

Y aquí sí que es fácil calcular, por ejemplo, la probabilidad de que más de 60 personas usen el Metro. Bastaría con calcular P(S > 60).

Fíjate que entonces, siempre que el número de datos sea suficiente, una Binomial se puede aproximar por una Normal, cuyos parámetros serán los que se obtienen al aplicar el Teorema Central del Límite, esto es, n·p y $\sqrt{n}\sigma$.

Si todavía no te lo crees, observa la siguiente escena. Ve aumentando el número de pruebas y verás que la forma de la distribución cada vez se parece más a la gráfica de una Normal.

Aunque si te fijas bien, hay una pequeña pega. Si el valor de la probabilidad de éxito (control "probabilidad") lo haces muy pequeño o muy grande, ya no se parece tanto a la gráfica de la normal cuando aumentamos n (tiene que ser n mucho más grande para que se parezca). Así que además

tenemos que poner una restricción para la probabilidad. El que no sea ni muy grande ni muy chico el valor de "p" lo vemos comprobando que n·p y n·(1-p) es mayor o igual que 5.

Importante

Si X es una variable aleatoria discreta que tiene una distribución de probabilidad Binomial de parámetros n y p (X~B(n,p)), podemos aproximarla a una distribución de probabilidad Normal siempre que:

1. $n \geq 30$
2. $n \cdot p \geq 5$
3. $n \cdot (1-p) \geq 5$

Además la distribución normal a la que se aproxima es:

$$N(n \cdot p\,;\sqrt{n \cdot p \cdot (1-p)})$$

A este resultado se le conoce como Teorema de De-Moivre

Ejercicio resuelto

Lanzamos un dado 50 veces. ¿Cuál es la probabilidad de que más de 10 veces salga un 1?

En este vídeo tienes la respuesta:

Imagen de Charly Morlock bajo licencia Creative Commons.

En el vídeo no se calcula directamente la probabilidad de X mayor que 10, sino que amplía un poquito el recinto para que el 10 quede dentro de él. Esto se hace para corregir el error que se produce al hacer la aproximación ya que en la normal, al ser variable continua la probabilidad del valor 10 sería cero, pero en la binomial no, al ser discreta. Así que, en las aproximaciones de binomial a normal, debes tener en cuenta esto, que hay que ampliar un poquito el recinto.

Para saber más

Aquí tienes otro vídeo en el que se resuelve una cuestión mediante la aproximación de una binomial a una normal. Es un poco largo pero está explicado con todo lujo de detalles y seguro que si lo ves lo vas a entender todo a la perfección:

Comprueba lo aprendido tiple

En una comarca, el 25% de las viviendas tienen conexión a Internet de banda ancha. Elegimos al azar 80 viviendas y definimos la variable aleatoria X = Número de viviendas con conexión a Internet.

Imagen de infoventas.vistahermosa bajo licencia Creative Commons.

1) La variable aleatoria X es discreta y tiene una distribución de probabilidad.

- B(80 ; 0,25)
- B(80; 0,75)
- No es una distribución binomial

Correcto

La probabilidad de éxito es 0,25, pues el éxito es que la casa tenga conexión a Internet.

Sí lo es, pues contamos el número de éxitos en una serie de repeticiones (80)

Solution

1. Opción correcta
2. Incorrecto
3. Incorrecto

2) Esa variable se puede aproximar a una normal.

- No es posible.
- Sí porque n ≥ 30.
- Sí porque n ≥ 30, $n \cdot p \geq 5$ y $n \cdot q \geq 5$

Repasa lo anterior.

Falta algo más.

Muy bien

Solution

1. Incorrecto
2. Incorrecto
3. Opción correcta

3) La distribución Normal a la que se aproxima es:

- N(20 ; 15)
- N(80; 15)
- N(20 ; 3,873)
- N(15 ; 20)

El segundo parámetro es la desviación típica, no la varianza.

Repásalo todo.

Muy bien.

Te has liado con los parámetros.

Solution

1. Incorrecto
2. Incorrecto
3. Opción correcta
4. Incorrecto

4) La probabilidad de que al menos 12 casas tengan conexión a Internet es:

Sugerencia

- 0,9857
- 0,0143
- 0,9808
- 0,0192

Perfecto.

Repasa cómo se calculaban probabilidades en una normal.

Recuerda que hay que ampliar un poquito el recinto.

Repasa cómo se calculaban probabilidades en una normal.

Solution

1. Opción correcta
2. Incorrecto
3. Incorrecto
4. Incorrecto

6. Ejemplo de selectividad

Ejercicio resuelto

En selectividad pueden aparecer apartados de problemas que hagan referencia a las distintas formas de hacer un muestreo en una población. Aquí tienes un ejemplo:

JUNIO 2011

EJERCICIO

a) (1 punto) Una población de tamaño 1000 se ha dividido en 4 estratos de tamaño 150, 400, 250 y 200. Utilizando muestreo aleatorio estratificado con afijación proporcional se han seleccionado 10 individuos del tercer estrato, ¿cuál es el tamaño de la muestra?

b) (1.5 puntos) El peso de los individuos de una población se distribuye según una ley Normal de desviación típica 6 kg. Calcule el tamaño mínimo de la muestra para estimar, con un nivel de confianza del 95%, el peso medio en la población con un error no superior a 1 kg.

Ejercicio tomado de http://www.iesayala.com/selectividadmatematicas/

Mostrar retroalimentación

Los apartados son independientes. Para hacer el apartado b) es necesario estudiar los contenidos de la unidad 6. Veamos cuál sería la respuesta al apartado a):

Sabemos que en un muestreo aleatorio estratificado con afijación proporcional, si hay "k" estratos y que el número de elementos de cada estrato es $N_1, N_2, ..., N_k$, y si $n_1, n_2, ..., n_k$ son los elementos de cada una de las muestras de los estratos, el tamaño total de la muestra $n = n_1 + n_2, +...+ n_k$ y se calculan eligiendo los números $n_1, n_2, ..., n_k$ proporcionales a los tamaños de los estratos $N_1, N_2, ..., N_k$, es decir

$$\frac{n_1}{N_1} = \frac{n_2}{N_2} = ... = \frac{n_k}{N_k} = \frac{n}{N}$$

En nuestro caso $\frac{n_1}{150} = \frac{n_2}{400} = \frac{10}{250} = \frac{n_4}{200} = \frac{n}{1000}$

De $\frac{10}{250} = \frac{n}{1000}$, tenemos $n = \frac{1000.10}{250} = 40$, luego el tamaño de la muestra es "n = 40".

Ejercicio resuelto

SEPTIEMBRE 2011

Sea X una variable aleatoria Normal de media 50 y desviación típica 4. Se toman muestras de tamaño 16.

(a) (1 punto) ¿Cuál es la distribución de la media muestral?

(b) (1'5 puntos) ¿Cuál es la probabilidad de que la media muestral este comprendida entre 47'5 y 52'5?

Ejercicio tomado de http://www.iesayala.com/selectividadmatematicas/

Mostrar retroalimentación

(a) y (b)

Sabemos que si una variable aleatoria X sigue una normal $N(\mu, \sigma)$, la distribución muestral de medias $\bar{X}$ sigue una normal $N(\mu, \frac{\sigma}{\sqrt{n}})$.

¿Cual es la probabilidad de que la media muestral este comprendida entre 47'5 y 52'5?

Datos: X sigue una normal N(50,4), luego $\mu = 50 = \bar{x}$ y $\sigma = 4$; n = 16

Me están pidiendo $p(47'5 < \bar{X} < 52'5)$ = { tipifico $Z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$ } = $p(\frac{47'5-50}{4/\sqrt{16}} < Z < \frac{52'5-50}{4/\sqrt{16}}) \cong p(-2'5 < Z < 2'5)$ =

= $p(Z < 2'5) - p(Z < -2'5) = p(Z < 2'5) - [1 - p(Z \leq 2'5)] = 2.p(Z < 2'5) - 1$= {mirando en las tablas de la N(0,1) } =

= 2.0'9798 - 1 = 0'9596.

Resumen

Un **muestreo** se dice que es **aleatorio** o **probabilístico** cuando todos los individuos de la muestra se eligen al azar, de modo que todos tienen la misma probabilidad de ser elegidos.

Sólo este método de muestreo nos asegura la representatividad de la muestra extraída y es, por tanto, el más recomendable.

El **muestreo con reemplazamiento** es aquel en el que un elemento puede ser seleccionado más de una vez en la muestra. Para ello se extrae un elemento de la población, se observa y se devuelve a la población, por lo que de esta forma se pueden hacer infinitas extracciones de la población aún siendo ésta finita.

El **muestreo sin reemplazamiento** es el que se realiza sin devolver los elementos extraídos a la población hasta que no se hayan extraído todos los elementos de la población que conforman la muestra.

Si te acuerdas de la unidad anterior, cuando en una población estudiábamos una determinada característica o variable que sólo podía tomar dos valores, éxito o fracaso, la población seguía una **distribución binomial**.

En esta población, la proporción de individuos que poseen esa característica la llamamos *p* y en todas las muestras de tamaño *n* que podamos extraer de la población, llamaremos $\hat{p}$ al porcentaje de individuos que tengan esa característica.

Los distintos valores de $\hat{p}$ que dependen de las muestras elegidas, dan lugar a una variable aleatoria que se representa por $\hat{P}$ y que se llama **estadístico**.

Llamamos **distribución muestral de proporciones** a la distribución de los valores de $\hat{P}$.

La variable aleatoria $\hat{P}$ tiene las siguientes características:

a) La media es: $\mu = p$.

b) La desviación típica es: $\sigma = \sqrt{\frac{p \cdot q}{n}}$, siendo $q = 1 - p$.

c) Para muestras donde **n≥30**, la distribución de $\hat{P}$ se aproxima a una distribución **normal** $N(p, \sqrt{\frac{p \cdot q}{n}})$.

Importante

La **distribución en el muestreo de la media** $\overline{X}$ tiene las siguientes características:

a. Media: Tiene la misma media que la población **μ.**

b. Desviación típica: La desviación típica de esta distribución es $\frac{\sigma}{\sqrt{n}}$, siendo n el tamaño de las muestras.

c. Si la población no sigue una distribución normal, pero **n ≥ 30**, la distribución de las medias muestrales se aproxima a una distribución **normal**, esta aproximación será mejor cuanto mayor sea n.

Importante

La variable aleatoria T tiene las siguientes características:

a. Media: n·μ, donde n es el número de individuos de la muestra.

b. Desviación típica : $\sigma \cdot \sqrt{n}$

c. Si la población no sigue una distribución normal, pero **n ≥ 30**, la distribución de T se aproxima a una **normal** $N(n \cdot \mu , \sigma \cdot \sqrt{n})$.

Importante

El **Teorema Central del Límite** nos indica que si tenemos una serie de variables aleatoria independientes (el valor de una no influye en la otra) e idénticamente distribuidas (todas las variables tienen la misma distribución y por tanto, los mismos parámetros), la distribución de la **suma** de esas **variables** (si el número de variables que se suman es suficientemente grande) se **aproxima a una Distribución Normal**.

Lo precisamos todavía más, porque no se aproxima a una normal cualquiera sino que podemos saber a cuál:

Si tenemos n variables aleatorias X_1, X_2, X_3,..., X_n todas ellas independientes entre sí y todas ellas con media μ y desviación típica σ, la suma de esas variables genera una nueva variable aleatoria que se aproximará a una distribución Normal de media $n \cdot \mu$ y desviación típica $\sqrt{n} \cdot \sigma$

$$S_n = X_1 + X_2 + ... + X_n = \Sigma X_i \sim N(n \cdot \mu , \sqrt{n} \cdot \sigma)$$

La coletilla de si el número es suficientemente grande no es ninguna tontería, pues lo

que en realidad se aproxima es el límite de la suma de las variables cuando el número de variables tiende a infinito. O sea, que para que esto funcione, el número de variables o de datos que se han de sumar tiene que ser grande. A efectos prácticos nos vale con que al menos haya 30 datos, es decir, para aplicar este teorema tiene que cumplirse que **$n \geq 30$**.

Ojo, fíjate que en ningún momento estamos diciendo que la variable X tenga que ser una distribución normal, sino que sea lo que sea la distribución de la variable en la población, la suma de muchas observaciones se va a aproximar a una normal.

Si ya de por sí la **población** de partida sigue una distribución **normal**, ese resultado se cumple siempre, sea el tamaño el que sea. **No importa el valor de n.**