Probabilidad y Estadística (93.24)
Trabajo Práctico Nº 7

Suma de variables aleatorias, distribución en el muestreo.

1. La probabilidad de dar en el blanco para un cañón es 0.8 (se supone constante y los disparos independientes). Calcular la probabilidad de que el número de blancos X satisfaga $70 < X < 90$ si se efectúan 100 disparos usando la distribución binomial y la aproximación normal a la distribución binomial.

2. Un sistema compuesto está formado por 100 componentes que funcionan independientemente. La probabilidad de que cualquier componente falle durante el período de operación es igual a 0.1. Para que el sistema completo funcione por lo menos deben funcionar 85 componentes. Calcular la probabilidad de que el sistema funcione correctamente usando la distribución binomial y la aproximación normal de la distribución binomial.

3. Una central telefónica A da servicio a 1800 usuarios de una central cercana B. Sería costoso y extravagante instalar 1800 líneas troncales de A a B. Es suficiente instalar un número N de líneas tan grande que, en condiciones ordinarias, solamente una de cada 100 llamadas en promedio no encuentre inmediatamente una línea troncal disponible. Supóngase que, en la hora más ocupada del día, cada usuario requiere una linea troncal de B durante un promedio de 2 minutos. En un momento fijo de la hora de máximo tráfico, puede suponerse la situación como un conjunto de 1800 ensayos independientes con una probabilidad $p = \frac{1}{30}$ (por lo de 2 min de cada 60 min) en que cada uno se requiere una línea. Determinar el número de líneas N a instalar entre A y B. Explicar claramente el planteo.

4. Cierto proceso de fabricación de componentes electrónicos produce partes con un porcentaje de defectos del 5 %. Las partes son enviadas en cajas de 400 unidades. Las cajas con 25 o más partes defectuosas son devueltas.

   a) ¿Cuál es la probabilidad de que se devuelva una caja?

   b) En un día particular se enviaron 500 cajas. ¿Cuál es la probabilidad de que se devuelvan 80 o más cajas?

   c) Se introduce un nuevo proceso de fabricación que se supone reduce el procentaje de defectos. El objetivo es reducir la probabilidad de devolución de una caja a 0.01. ¿Cuál debe ser el porcentaje de defecto para que se alcance este objetivo.

5. Suponga que 30 instrumentos electrónicos $D_1, D_2, \ldots, D_{30}$ se usan de la siguiente manera: tan pronto como D_1 falla empieza a actuar D_2, cuando éste falla empieza a actuar D_3 y así sucesivamente. Si la variable aleatoria asociada a la duración de cada instrumento tiene distribución exponencial de parámetro 0.1 1/hora y T es el tiempo total de operación de los 30 instrumentos, ¿cuál es la probabilidad de que T exceda 310 horas? Se supone que los tiempos de operación son variables aleatorias independientes. La variable aleatoria T tiene *distribución Gamma*. Calcule la probabilidad pedida usando la distribución Gamma y la aproximación normal de esta suma de variables aleatorias independientes.

6. Al sumar números, una computadora aproxima cada número al entero más próximo. Supongamos que todos los errores de aproximación son independientes y distribuidos uniformemente en el intervalo (-0.5, 0.5). Si se suman 1500 números, ¿cuál es la probabilidad de que, en valor absoluto, el error total exceda 15?

7. Suponga que X_i, $i = 1, 2, \ldots, 50$, son variables aleatorias independientes que tienen cada una distribución de Poisson con parámetro $\lambda = 0.3$. Sea $S = X_1 + \cdots + X_{50}$.

   a) Usando el teorema central del límite calcular $P(S \geq 18)$.

   b) Comparar la respuesta anterior con el valor exacto de esta probabilidad. Para calcular esta probabilidad considere que la suma de n variables aleatorias independientes con distribución de Poisson de parámetro λ tiene distribución de Poisson de parámetro $n\lambda$.

8. El número de accesos a un servidor web se supone (en cierto período del día) que es una variable aleatoria con distribución de Poisson con una media de 27 accesos por hora. Obtener la probabilidad de que haya al menos 90 accesos a lo largo de tres horas. Comparar los resultados obtenidos con la distribución de Poisson y con la aproximación normal.

9. Unos tambores etiquetados con un contenido de 30 litros son llenados con una solución proveniente de un gran tanque. El dispositivo de llenado de los tambores dispensa una cantidad por tambor que se supone una variable aleatoria con media de 30.01 litros y desvío estándar de 0.1 litros.

   a) ¿Cuál es la probabilidad de que la cantidad total de solución contenida en 50 tambores sea mayor a 1500 litros?

   b) Si la cantidad total de solución en el tanque es de 2401 litros, ¿cuál es la probabilidad de que puedan llenarse 80 tambores sin que se vacie el tanque?

   c) ¿Que cantidad de solución debe contener el tanque para que sea 0.9 la probabilidad de que puedan llenarse 100 tambores sin que el tanque se vacie?

10. Los tiempos de servicio para los clientes que llegan a una caja de un supermercado pueden suponerse variables aleatorias independientes con un promedio de 1.5 minutos y una dispersión de 1 minuto.

    a) Calcular la probabilidad aproximada de que pueda atender a 100 clientes en menos de 2 horas de tiempo total de servicio de esta caja.

    b) Calcular el número de clientes n tal que la probabilidad de dar servicio a todos en menos de 2 horas sea aproximadamente 0.1.

11. Generar 10 números al azar, sumarlos, restarle 5 a la suma y dividir por $\sqrt{\frac{10}{12}}$ (¿un número mágico tal vez?). Repetir esa generación 2000 veces, por ejemplo. Hacer un histograma de los 2000 números obtenidos y comparar con la distribución normal estándar. Para los números aleatorios usar la función de generación de números aleatorios (RAND por ejemplo) de cualquier planilla de cálculo, programa utilitario matemático, calculadora o su lenguaje de programación preferido.

    Este código en *Octave* realiza la generación solicitada:

```
R = rand(10,2000);S = (sum(R)-5)/sqrt(10/12);
hist(S)
SS = sort(S);F =(1:2000)/2000;
plot(SS,F,SS,normcdf(SS))
D = max(abs(F-normcdf(SS)))
```

En una ejecución de este código la distancia `D` fue `0.017934`. Aqui se compara la distribución empírica de los valores generados de `S` con la distribución normal estándar.

12. Una fábrica produce determinados artículos de tal manera que el 7 % resulta defectuoso. Se inspeccionan n de tales artículos, y se determina la frecuencia relativa de defectuosos f_D.

    a) ¿Cuál debería ser el tamaño de n de manera tal que la probabilidad de que f_D difiera de 0.07 en menos de 0.01 sea 0.98? Suponga válida la aproximación normal de la distribución binomial.

    b) Conteste la pregunta anterior si 0.07, la probabilidad de tener un artículo defectuoso, se sustituye por p que se supone desconocida. En este caso recuerde que si $p \in (0, 1)$ entonces $p\,(1-p) \leq 0.25$.

13. En un sistema están conectados 50 dispositivos. La probabilidad de que en el tiempo T un dispositivo cualquiera falle es 0.1. Utilizando la desigualdad de Tchebycheff estime la probabilidad de que la diferencia en valor absoluto entre el número de dispositivos que fallan y el promedio de los que fallan resulte: a) menor que 5 b) mayor que 5. Comparar el resultado con los que se obtienen si se utilizan el modelo binomial y su aproximación normal.

14. Se desea estimar la probabilidad p de ocurrencia de cierto suceso a partir de la frecuencia relativa de ocurrencia medida al realizar n repeticiones del experimento en que puede ocurrir el suceso de interés. Se pone como condición que la probabilidad de que el error (diferencia entre p y la frecuencia relativa) no supere el valor 0.05, sea mayor que 0.97. ¿Cuántas pruebas deben realizarse como mínimo? Comparar los resultados obtenidos a partir de usar:

    a) la desigualdad de Tchebycheff;

    b) la aproximación normal de la distribución binomial.

15. El peso de las cajas transportadas por una compañía de transportes se distribuye de manera aproximadamente normal de media 20 kg y desviación estándar 3 Kg. Calcular la probabilidad de que:

    a) el peso de una caja tomada al azar esté comprendido entre 19.7 y 20.6 kg;

    b) la media de una muestra de 100 cajas esté comprendida entre 19.7 y 20.6 kg.

16. Suponga que $X_1, X_2, \ldots, X_{40}$ representa una muestra aleatoria de mediciones independientes de la proporción de impurezas en muestras de una aleación. Se considera que la densidad de probabilidad de cada una de estas variables aleatorias es $f(x) = 3x^2 \;\; 0 < x < 1$ y $f(x) = 0$

si $x < 0$ o $1 < x$. Un comprador potencial rechaza una partida de mineral si la media de una muestra de tamaño 40 es superior a 0.8. Calcular la probabilidad de que la partida sea rechazada.

17. Se toman muestras independientes de tamaño 10 y 15 de una variable aleatoria con distribución normal de media 20 y varianza 3. ¿Cuál es la probabilidad de que el promedio de las dos muestras se diferencie (en valor absoluto) en más de 0.3?

18. Se considera la variable aleatoria X: número obtenido al arrojar un dado no cargado.

    a) Obtener la distribución de probabilidades de X, calcular su valor esperado y su varianza.

    b) Enumerar todas las posibles muestras de tamaño 2 que se pueden obtener de la variable X (son 36 resultados posibles). Calcular para cada una de ellas el promedio y determinar entonces la distribución de probabilidades de ese promedio (media muestral). Compare los valores medios de X y de la media muestra ¿qué concluye?, ¿qué resulta al comparar las varianzas?

    c) Analice la distribución en el muestreo de la varianza muestral S^2. Considere todas las muestras posibles de tamaño 2 y calcule para cada una la varianza muestral. Encuentre la distribución de probabilidades de la varianza muestral y compare su valor esperado con la varianza poblacional.

19. Genere 50 muestras cada una de tamaño 30 de una variable aleatoria con distribución exponencial de parámetro característico 2. Para cada muestra calcular la media muestral obteniéndose así 50 realizaciones del estadístico media muestral para muestras de tamaño 10. Agrupe esos datos y analice las frecuencias relativas. Repita el procedimiento pero ahora con un tamaño de muestra igual a 500. Comente los resultados.

```
R = rand(30,500);E = -(1/2)*log(1-R);
Xraya = mean(E);
Z =(Xraya-0.5)/(0.5/sqrt(30));
ZZ = sort(Z);
plot(ZZ,F,ZZ,normcdf(ZZ))
D = max(abs(F-normcdf(ZZ)));
```

En una ejecución de este código la distancia `D` fue `0.065007`. Aqui se compara la distribución empírica de los valores generados de `Z` con la distribución normal estándar.

20. En la fabricación de cierto tipo de cojinetes para motor se sabe que el diámetro promedio es de 5 cm con un desvío estándar de 0.005 cm. El proceso es controlado en forma periódica mediante la selección aleatoria de 64 cojinetes, midiendo sus correspondientes diámetros. El proceso se supone bajo control si la media muestral se encuentra entre dos límites especificados en el 95 % de las veces que se extraen las muestras para realizar el control del proceso. Determinar los valores de esos límites si son simétricos respecto de la media poblacional.

21. Los montos X de las facturas tienen una función densidad de probabilidad dada por $f_X(x) = 3/x^4$ para $x > 1$ y 0 para $x < 1$ (X en centenas de \$). Suponiendo que los montos de facturas diferentes son independientes ¿cuál es la probabilidad de que el monto de 300 facturas del mes supere \$ 42000?

22. Para facilitar el cobro de facturas, una empresa descuenta redondeando a la decena a favor del cliente. La empresa completa 1000 facturas por mes.

    a) ¿Cuál es el monto medio de descuento en el mes por esta causa?

    b) ¿Entre qué valores puede Ud. informar que estará ese valor, con probabilidad 0.95? Considere que el intervalo sea simétrico respecto del valor medio.

    c) ¿Cuál es el máximo descuento mensual tal que pueda asegurarse que no va a ser superado con probabilidad de 0.95?

ACTIVIDAD OPTATIVA (recomendable para fijar conceptos).
Actividades en Octave.
Una verificación del Teorema Central del Límite (TCL)

Dado que toda variable aleatoria con distribución binomial de parámetros n y p puede representarse como suma de n variables aleatorias independientes con distribución Bernoulli de parámetro p, es razonable la idea de aproximar dicha distribución utilizando la de una normal de media np y desviación $\sqrt{np(1-p)}$. Usando algún programa (por ejemplo, Excel, Octave, Matlab o R) represente las funciones de distribución binomial y la de su aproximación por la normal para distintos valores de n y p (por ejemplo $n = 10, 30, 60, 100, 400$ y $p = 0{,}01, 0{,}1, 0{,}25, 0{,}5, 0{,}75, 0{,}9, 0{,}99$ y obtenga, en forma aproximada, la máxima diferencia en valor absoluto entre la distribución calculada en forma exacta y en forma aproximada en el conjunto $0, 1, \ldots n$. Por ejemplo si $n = 400$ y p =0.5 esa máxima diferencia es 0.0199 (usando Octave).

El siguiente código en *Octave* muestra como obtener 4 gráficos de la distribución binomial para los parámetros (n, p) de estos pares: (5, 0.01), (5, 0.5), (50, 0.01) y (50, 0.5).

```
P1 = binopdf(0:5,5,0.01);P2=binopdf(0:5,5,0.5);
P3 = binopdf(0:50,50,0.01);P4=binopdf(0:50,50,0.5);
subplot(2,2,1),stem(0:5,P1),subplot(2,2,2),stem(0:5,P2)
subplot(2,2,3),stem(0:50,P3),subplot(2,2,4),stem(0:50,P4)
```

El comando `binopdf(x,n,p)` calcula los valores de la función de probabilidad de la variable aleatoria con distribución binomial para los elementos del vector `x` si los parámetros de la distribución son `n` y `p` (el número de repeticiones del experimento de Bernoulli y la probabilidad de éxito respectivamente). El comando de gráficos `stem` se usa para la representación de funciones de variable discreta.

Otra verificación del Teorema Central TCL

Una variable aleatoria con distribución de Poisson de parámetro λ siempre puede representarse como suma de n variables aleatorias Poisson independientes de parámetro λ/n. Por lo tanto es razonable que dicha distribución pueda ser aproximada por la de una normal de media λ y desviación $\sqrt{\lambda}$, sobre todo para grandes valores de λ. Usando algún programa (por ejemplo, Excel, Octave, Matlab o R) represente las funciones de distribución de Poisson y la de su aproximación por la normal para $\lambda = 1, 5, 10, 20, 50, 100$ y obtenga, en forma aproximada, la máxima diferencia en valor absoluto entre la función de distribución calculada en forma exacta y en forma aproximada en el conjunto $0, 1, \ldots 5\,\lambda$. Por ejemplo si $\lambda = 50$ esa máxima distancia es 0.03752 (calculado con Octave).

El siguiente código en *Octave* muestra como obtener 4 gráficos de la distribución de Poisson para el parámetro tomando los valores 1, 3, 5 y 10.

```
P1=poisspdf(0:10,1);P2=poisspdf(0:20,3);
P3=poisspdf(0:20,5);P4=poisspdf(0:40,10);
subplot(2,2,1),stem(0:10,P1),subplot(2,2,2),stem(0:20,P2)
subplot(2,2,3),stem(0:20,P3),subplot(2,2,4),stem(0:40,P4)
```

Se muestra a continuación una aproximación de la distribución de Poisson por la normal y la máxima distancia entre las funciones de distribución si el parámetro de la distribución de Poisson vale 50 y se analiza el intervalo el conjunto $0, 1, \cdots 250$.

```
L = 50;
x = 0:5*L;
poisson = poisscdf(x,L);
E = L;S = sqrt(L);
normal = normcdf((x-E)/S);
max(abs(poisson-normal))
ans = 0.037517
```

El comando `poisscdf(x,L)` calcula los valores de la función de distribución de la variable aleatoria con distibución de Poisson para los elementos del vector `x` si el parámetro de la distribución es `L`.

Otra verificación del Teorema Central del Límite por simulación

Sea $X = X_1 + X_2 + X_3 + X_4 + X_5$, con X_i variables aleatorias independientes con distribución uniforme en $(0, 1)$. Se desea saber si la distribución de X puede o no aproximarse por una normal. Para esto, se propone generar 200 copias simuladas de la variable aleatoria X. Se procede de la siguiente manera usando una planilla de cálculo (Excel, por ejemplo):

1. utilizando la función aleatorio() (o rand()) generar una matriz de 200 x 5 números.
2. Sumar por columnas, generando una 6ta columna, cuyos elementos serán copias simuladas de X.
3. Graficar estos valores en un histograma mediante la opción: Herramientas/Análisis de Datos/Histograma (en Excel)
4. Evaluar si dicho histograma posee o no forma de campana, y en consecuencia responder si la aproximación propuesta es adecuada o no.

En *Octave* se puede ensayar la ejecución de estas líneas:

```
N = 1000; Xs = sum(rand(5,N)); hist(Xs)
XX=sort(Xs);F=(1:N)/N;
mu=2.5; s=sqrt(5/12); FN=normcdf(XX,mu,s);
plot(XX,FN,XX,F)
max(abs(FN-F))
```

Con este código se genera el histograma para `N` valores simulados de la suma de 5 variables aleatorias uniformes independientes. La instrucción `rand(5,N)` genera una matriz $5 \times N$ con números pseudo-aleatorios de distribución uniforme en $(0, 1)$, mientras que `Xs` almacena la suma de cada columna de la matriz. De ese modo se obtienen N valores simulados de X, cuyo histograma es dado por la instrucción `hist(Xs)`.

El vector `XX` contiene las componentes de `Xs` ordenadas de menor a mayor. Si a cada valor del vector `XX` se le asigna la probabilidad 1/N se obtiene una variable aleatoria discreta, cuya función de distribución se denomina distribución empírica de la variable aleatoria X y que cuando N es grande, aproxima bien a la distribución de X. El vector `F` contiene los valores de esa distribución empírica en cada uno de los valores almacenados en el vector `XX`, en el mismo orden. Por otra parte `FN` es un vector que almacena los valores de la función de distribución normal de media $\mu = E(X)$ y desviación $\sigma = \sqrt{V(X)}$ evaluada en las componentes de `XX` y en el mismo orden (aproximación de X mediante una normal con mismas media y desviación). Finalmente `plot(XX,FN,XX,F)` produce los gráficos de `F`

y `FN` en una misma figura, y `max(abs(FN-F))` calcula el máximo apartamiento entre esos dos gráficos. Si desea, por ejemplo el gráfico de `F`, ejecute `plot(XX,F)`. Ejecutando este código podrá ver que los gráficos aparecen superpuestos a la vista. Para una ejecución esa distancia resultó 0.019441.

ACTIVIDAD OPTATIVA (recomendable para fijar conceptos).

Sobre distribuciones en el muestreo, usando *Statlets*, el laboratorio virtual de la Universidad de Rice:

`http://onlinestatbook.com/stat_sim/sampling_dist/index.html`

Haga click sobre el botón `Begin` y el applet de Java se ejecutará en una nueva ventana. Este applet de Java explora varios aspectos de la distribución en el muestreo de un estadístico o estimador. Cuando se ejecuta por primera vez puede verse un histograma de la distribución normal que es la distribución de probabilidades de la variable aleatoria definida en la población de la que se extrae la muestra. La media de la distribución de probabilidades de la variable definida en la población se indica con una pequeña linea azul y la mediana con una roja, como para la distribución normal son iguales las dos marcas están superpuestas. La línea roja que se extiende desde la media hacia cada lado corresponde al intervalo de semiamplitud igual al desvío estándar (poblacional en este caso)con centro en la media. El segundo histograma muestra los datos muestrales, inicialmente está en blanco. El tercer y cuarto histogramas muestran la distribución de probabilidades de los estadísticos obtenidos a partir de las muestra. El número de veces (replicaciones) que se extraen muestras de tamaño N se indica con el valor de la variable Reps=."

Operaciones básicas

Inicialmente la simulación corresponde a la extracción de una muestra de tamaño N = 5, se calcula la media y se representa su valor. Activando el botón `Animated sample` se pueden ver las cinco observaciones muestrales. Se calcula la media de esos cinco números y se representa ese valor en el tercer histograma. Es interesante repetir esta operación varias veces y observar como se va obteniendo una distribución de valores de la media muestral.Puede acelerarse el proceso tomando 5, 100, 1000, o 10000 muestras.

Elección del estadístico

Los siguientes estadísticos pueden ser calculados para cada muestra extraída: Media, Desvío estándar de la muestra (usando N en el denominador) Varianza de la muestra (usando N en el denominador) Estimador insesgado de la varianza (usando N-1 en el denominador) Valor absoluto del desvío respecto de la media. Rango.

Elección del tamaño muestra

El tamaño de la muestra puede elegirse en 2, 5, 10, 16, 20 or 25 en el menú. Observe no confundir el tamaño de la muestra (N) con el número de muestras de ese tamaño que son generadas para poder observar la distribución en el muestreo del estadístico elegido.

Comparación con la distribución normal

Pulsando el botón `Fit normal` puede verse la superposición del gráfico de la función densidad de la distribución normal sobre el histograma de la distribución en el muestreo analizada.