6- SUMA DE VARIABLES ALEATORIAS Y TEOREMA CENTRAL DEL LÍMITE

6.1 – Suma de variables aleatorias independientes

Cuando se estudiaron las variables aleatorias bidimensionales se habló de una ***función de variable aleatoria bidimensional.*** En particular se nombró la suma de *n* variables aleatorias**,** pero no se dijo nada sobre la ***distribución*** de esa v.a. suma.
Es a menudo importante saber cuál es la distribución de una suma de variables aleatorias independientes.
Consideramos algunos ejemplos en el caso discreto

1- ***Suma de variables aleatorias independientes con distribución Poisson***

$$X \sim P(\lambda_1) \;;\; Y \sim P(\lambda_2) \;;\; X \text{ y } Y \text{ independientes} \quad \Rightarrow \quad X+Y \sim P(\lambda_1+\lambda_2)$$

Dem.)
Consideramos el evento $\{X+Y=n\}$ como unión de eventos excluyentes $\{X=k, Y=n-k\}$ $\quad 0 \le k \le n$, entonces

$$P(X+Y=n)=\sum_{k=0}^{n} P(X=k,Y=n-k)=\sum_{k=0}^{n} P(X=k)P(Y=n-k)=\sum_{k=0}^{n} e^{-\lambda_1}\frac{\lambda_1^{\,k}}{k!}e^{-\lambda_2}\frac{\lambda_2^{\,n-k}}{(n-k)!}=$$

X e *Y* independientes

$$=e^{-(\lambda_1+\lambda_2)}\sum_{k=0}^{n}\frac{\lambda_1^{\,k}}{k!}\frac{\lambda_2^{\,n-k}}{(n-k)!}=\frac{e^{-(\lambda_1+\lambda_2)}}{n!}\sum_{k=0}^{n}\frac{n!}{k!(n-k)!}\lambda_1^{\,k}\lambda_2^{\,n-k}=\frac{e^{-(\lambda_1+\lambda_2)}}{n!}(\lambda_1+\lambda_2)^n$$

Binomio de Newton

O sea *X+Y* tiene distribución Poisson con parámetro $\lambda_1+\lambda_2$

2- ***Suma de variables aleatorias binomiales independientes***

$$X \sim B(n_1,p) \;;\; Y \sim B(n_2,p) \;;\; X \text{ y } Y \text{ independientes} \quad \Rightarrow \quad X+Y \sim B(n_1+n_2,p)$$

Dem.)
Nuevamente consideramos el evento $\{X+Y=k\}$ como unión de eventos excluyentes $\{X=i, Y=k-i\}$ $\quad 0 \le i \le n_1$, entonces

$$P(X+Y=k)=\sum_{i=0}^{n_1} P(X=i,Y=k-i)=\sum_{i=0}^{n_1} P(X=i)P(Y=k-i)=\sum_{k=0}^{n_1}\binom{n_1}{i}p^i(1-p)^{n_1-i}\binom{n_2}{k-i}p^{k-i}(1-p)^{n_2-k+i}=$$

X e *Y* independientes

$$=p^k(1-p)^{n_1+n_2-k}\sum_{i=0}^{n_1}\binom{n_1}{i}\binom{n_2}{k-i}$$

En la expresión anterior si $j>r$ entonces $\binom{r}{j}=0$

Por último usamos la siguiente identidad combinatoria $\sum_{i=0}^{n_1}\binom{n_1}{i}\binom{n_2}{k-i}=\binom{n_1+n_2}{k}$

Y entonces

$$P(X+Y=k)=\binom{n_1+n_2}{k}p^k(1-p)^{n_1+n_2-k}$$

O sea $X+Y$ tiene distribución binomial con parámetros n_1+n_2 y p

<u>Observación</u>: en los dos casos anteriores se puede generalizar el resultado a n variables aleatorias independientes, usando el principio de inducción completa, es decir

1- Si $X_1, X_2,\ldots,X_n$ son n variables aleatorias independientes donde $X_i \sim P(\lambda_i)$ para todo $i=1,2,\ldots,n$ entonces $\sum_{i=0}^{n} X_i \sim P(\sum_{i=0}^{n}\lambda_i)$

2- Si $X_1, X_2,\ldots,X_n$ son n variables aleatorias independientes donde $X_i \sim B(n_i, p)$ para todo $i=1,2,\ldots,n$ entonces $\sum_{i=0}^{n} X_i \sim B(\sum_{i=0}^{n} n_i, p)$

<u>Suma de variables aleatorias normales independientes</u>

Si X e Y son dos variables aleatorias continuas independientes con densidades $g(x)$ y $h(y)$ respectivamente se puede probar (no lo demostraremos aquí) que la v.a. $Z=X+Y$ tiene densidad dada por $f_{X+Y}(z)=\int_{-\infty}^{\infty} g(z-y)h(y)dy$

Usando esto se puede demostrar el siguiente importante resultado:

Si X e Y son variables aleatorias independientes donde $X \sim N(\mu_1, \sigma_1{}^2)$ y $Y \sim N(\mu_2, \sigma_2{}^2)$ entonces $X+Y \sim N(\mu_1+\mu_2, \sigma_1{}^2+\sigma_2{}^2)$

Por inducción completa se puede generalizar este resultado a n variables:

Si $X_1, X_2,\ldots,X_n$ son n variables aleatorias independientes donde $X_i \sim N(\mu_i, \sigma_i^2)$ para todo $i=1,2,\ldots,n$ entonces $\sum_{i=0}^{n} X_i \sim N(\sum_{i=0}^{n}\mu_i, \sum_{i=1}^{n}\sigma_i^2)$

De lo anterior y del hecho que $X \sim N(\mu,\sigma^2) \Rightarrow aX+b \sim N(a\mu+b, a^2\sigma^2)$ tenemos:

Si $X_1, X_2,\ldots,X_n$ son n variables aleatorias independientes donde $X_i \sim N(\mu_i, \sigma_i^2)$ para todo $i=1,2,\ldots,n$ entonces $\sum_{i=0}^{n} a_i X_i \sim N(\sum_{i=0}^{n} a_i\mu_i, \sum_{i=1}^{n} a_i^2\sigma_i^2)$ donde $a_1, a_2,\ldots,a_n$ son números reales

Se dice que $\sum_{i=0}^{n} a_i X_i$ es una ***combinación lineal de variables aleatorias***.

Ejemplos:

1- La envoltura de plástico para un disco magnético está formada por dos hojas. El espesor de cada una tiene una distribución normal con media 1.5 milímetros y desviación estándar de 0.1 milímetros. Las hojas son independientes.

a) Determine la media y la desviación estándar del espesor total de las dos hojas.
b) ¿Cuál es la probabilidad de que el espesor total sea mayor que 3.3 milímetros?

Solución: Sean las variables aleatorias

X: "espesor de la hoja 1" e *Y*: "espesor de la hoja 2"

Entonces $X \sim N(1.5, 0.1^2)$; $Y \sim N(1.5, 0.1^2)$ y X e Y independientes

a) Si definimos la v.a. *Z*: "espesor total de las dos hojas" , entonces $Z = X + Y$

Por lo tanto $Z \sim N(1.5 + 1.5, \quad 0.1^2 + 0.1^2)$ es decir $Z \sim N(3, \quad 0.02)$

En consecuencia $E(Z) = 3$, $\sigma_Z = \sqrt{V(Z)} = \sqrt{0.02}$

b) Se pide calcular $P(Z > 3.3)$

$$P(Z > 3.3) = P\left(\frac{Z-3}{\sqrt{0.02}} > \frac{3.3-3}{\sqrt{0.02}}\right) = 1 - \Phi\left(\frac{3.3-3}{\sqrt{0.02}}\right) = 1 - \Phi(2.12132) = 1 - 0.983 = 0.017$$

2-Tengo tres mensajes que atender en el edificio administrativo. Sea X_i : " el tiempo que toma el *i*-ésimo mensaje" (i = 1, 2 ,3), y sea X_4 : " el tiempo total que utilizo para caminar hacia y desde el edificio y entre cada mensaje". Suponga que las X_i son independientes, normalmente distribuidas, con las siguientes medias y desviaciones estándar:

$$\mu_1 = 15\,\text{min}, \quad \sigma_1 = 4, \quad \mu_2 = 5, \quad \sigma_2 = 1, \quad \mu_3 = 8, \quad \sigma_3 = 2, \quad \mu_4 = 12, \quad \sigma_4 = 3$$

Pienso salir de mi oficina precisamente a las 10.00 a.m. y deseo pegar una nota en mi puerta que dice "regreso a las *t* a.m." ¿A qué hora *t* debo escribir si deseo que la probabilidad de mi llegada después de *t* sea 0.01?

Solución: Definimos la v.a. *Z*: "tiempo transcurrido desde que salgo de mi oficina hasta que regreso", entonces $T = X_1 + X_2 + X_3 + X_4$

Por lo tanto $T \sim N\left(\sum_{i=1}^{4} \mu_i, \quad \sum_{i=1}^{4} \sigma_i^2\right)$, y se pide hallar *t* tal que $P(T > t) = 0.01$

$$\sum_{i=1}^{4} \mu_i = 15 + 5 + 8 + 12 = 50 \quad \text{y} \quad \sum_{i=1}^{4} \sigma_i^2 = 4^2 + 1^2 + 2^2 + 3^2 = 30$$

Entonces $P(T > t) = 1 - \Phi\left(\frac{t-50}{\sqrt{30}}\right) = 0.01$, es decir $\Phi\left(\frac{t-50}{\sqrt{30}}\right) = 0.99$

Buscando en la tabla de la normal $\frac{t-50}{\sqrt{30}} = 2.33 \quad \Rightarrow \quad t = 2.33 \times \sqrt{30} + 50 = 62.7619$

3- El ancho del marco de una puerta tiene una distribución normal con media 24 pulgadas y desviación estándar de 1/8 de pulgada. El ancho de la puerta tiene una distribución normal con media 23.875 de pulgadas y desviación estándar de 1/16 de pulgadas. Suponer independencia.

a) Determine la distribución, la media y la desviación estándar de la diferencia entre el ancho del marco y de la puerta.

b) ¿Cuál es la probabilidad de que la diferencia entre el ancho del marco y de la puerta sea mayor que ¼ de pulgada?.
c) ¿Cuál es la probabilidad de que la puerta no quepa en el marco?.

<u>Solución</u>: Sean las variables aleatorias
X: “ancho del marco de la puerta en pulgadas”
Y: “ancho de la puerta en pulgadas”
Entonces $X \sim N(24,\ (1/8)^2)$, $Y \sim N(23.875,\ (1/16)^2)$, *X* e *Y* independientes

a) Se pide la distribución de *X-Y* , $E(X-Y)$, $\sigma_{X-Y} = \sqrt{V(X-Y)}$

$$E(X-Y) = E(X) - E(Y) = 24 - 23.875 = 0.125$$

$$V(X-Y) = V(X) + V(Y) = \left(\frac{1}{8}\right)^2 + \left(\frac{1}{16}\right)^2 = \frac{5}{256} \quad \therefore \ \sigma_{X-Y} = \frac{\sqrt{5}}{16}$$

Por lo tanto $X - Y \sim N\left(0.125,\ \left(\frac{\sqrt{5}}{16}\right)^2\right)$

b) Se pide la probabilidad $P(X-Y > 1/4)$

$$P(X-Y>1/4) = 1 - \Phi\left(\frac{0.25-0.125}{\frac{\sqrt{5}}{16}}\right) = 1 - \Phi\left(\frac{2\sqrt{5}}{5}\right) = 1 - \Phi(0.8944) = 1 - 0.8133 = 0.1867$$

c) Si la puerta no entra en el marco entonces se da el evento $\{X < Y\}$ o equivalentemente $\{X - Y < 0\}$, por lo tanto

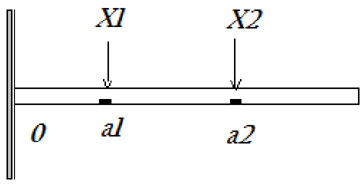
$$P(X-Y<0) = \Phi\left(\frac{0-0.125}{\frac{\sqrt{5}}{16}}\right) = \Phi\left(-\frac{2\sqrt{5}}{5}\right) = 1 - \Phi\left(\frac{2\sqrt{5}}{5}\right) = 0.1867$$

4- Supongamos que las variables aleatorias *X* e *Y* denotan la longitud y el ancho en cm, respectivamente, de una pieza.
Supongamos además que *X* e *Y* son independientes y que $X \sim N(2,\ 0.1^2)$, $Y \sim N(5,\ 0.2^2)$.
Entonces $Z = 2X + 2Y$ es una v.a. que representa el perímetro de la pieza.
Calcular la probabilidad de que el perímetro sea mayor que 14.5 cm.

<u>Solución</u>: tenemos que $Z \sim N\left(2\times 2 + 2\times 5,\ 2^2\times 0.1^2 + 2^2 \times 0.2^2\right)$, o sea $Z \sim N(14,\ 0.2)$
La probabilidad pedida es $P(Z > 14.5)$, entonces

$$P(Z>14.5) = 1 - \Phi\left(\frac{14.5-14}{\sqrt{0.2}}\right) = 1 - \Phi\left(\frac{\sqrt{5}}{2}\right) = 1 - \Phi(1.1180) = 1 - 0.8810 = 0.119$$

5- Si se aplican dos cargas aleatorias X_1 y X_2 a una viga voladiza como se muestra en la figura siguiente, el momento de flexión en 0 debido a las cargas es $a_1 X_1 + a_2 X_2$.



a) Suponga que X_1 y X_2 son v.a. independientes con medias 2 y 4 KLbs respectivamente, y desviaciones estándar 0.5 y 1.0 KLbs, respectivamente.

Si $a_1 = 5$ pies y $a_2 = 10$ pies, ¿cuál es el momento de flexión esperado y cuál es la desviación estándar del momento de flexión?

b) Si X_1 y X_2 están normalmente distribuidas, ¿cuál es la probabilidad de que el momento de flexión supere 75 KLbs?

Solución: Sea la v.a. Z: "momento de flexión en 0", entonces $Z = 5X_1 + 10X_2$

Por lo tanto

a) $E(Z) = 5E(X_1) + 10E(X_2) = 5\times 2 + 10\times 4 = 50$

$$V(Z) = 5^2 \times 0.5^2 + 10^2 \times 1^2 = 25\times 0.25 + 10\times 1 = \frac{65}{4} \quad \therefore\ \sigma_Z = \sqrt{\frac{65}{4}}$$

b) Si X_1 y X_2 están normalmente distribuidas, entonces $Z \sim N\left(50,\ \frac{65}{4}\right)$

Por lo tanto

$$P(Z > 75) = 1 - \Phi\left(\frac{75-50}{\sqrt{\frac{65}{4}}}\right) = 1 - \Phi\left(\frac{10\sqrt{65}}{13}\right) = 1 - \Phi(6.20) \approx 1 - 1 = 0$$

Promedio de variables aleatorias normales independientes

Si $X_1, X_2, \ldots, X_n$ son n variables aleatorias independientes donde $X_i \sim N(\mu, \sigma^2)$ para todo $i = 1,2,\ldots,n$ entonces la v.a. $\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$ tiene distribución normal con media μ y varianza $\frac{\sigma^2}{n}$

Dem.) Notar que $\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$ es un caso particular de combinación lineal de variables aleatorias donde $a_i = \frac{1}{n}$ para todo $i = 1,2,\ldots,n$

Además en este caso $\mu_i = \mu$ y $\sigma_i^2 = \sigma^2$ para todo $i = 1,2,\ldots,n$

Por lo tanto, $\overline{X}$ tiene distribución normal con esperanza $\sum_{i=1}^{n} \frac{1}{n}\mu_i = \sum_{i=1}^{n} \frac{1}{n}\mu = \frac{1}{n} n\mu = \mu$ y varianza

$$\sum_{i=1}^{n} \left(\frac{1}{n}\right)^2 \sigma_i^2 = \sum_{i=1}^{n} \left(\frac{1}{n}\right)^2 \sigma^2 = \left(\frac{1}{n}\right)^2 n\sigma^2 = \frac{\sigma^2}{n}$$

Es decir, $\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$

Observación: a $\overline{X}$ se lo llama ***promedio muestral*** o ***media muestral***

Ejemplos:
1) El diámetro interno de un anillo de pistón seleccionado al azar es una v.a. con distribución normal con media 12 cm y desviación estándar de 0.04 cm.
a) Si $\overline{X}$ es el diámetro promedio en una muestra de $n=16$ anillos, calcule $P(11.99 \le \overline{X} \le 12.01)$
b) ¿Qué tan probable es que el diámetro promedio exceda de 12.01 cuando $n=25$?

Solución:
a) Sean las variables aleatorias X_i : "diámetro del anillo *i*" $i=1,2,\dots,16$
Entonces $X_i \sim N(12, 0.04^2)$ para cada *i*.
Por lo tanto $\overline{X} \sim N\left(12, \dfrac{0.04^2}{16}\right)$. Entonces

$$P(11.99 \le \overline{X} \le 12.01) = P\left(\frac{11.99-12}{\sqrt{\frac{0.04^2}{16}}} \le \frac{\overline{X}-12}{\sqrt{\frac{0.04^2}{16}}} \le \frac{12.01-12}{\sqrt{\frac{0.04^2}{16}}}\right) =$$

$$= \phi\left(\frac{12.01-12}{\sqrt{\frac{0.04^2}{16}}}\right) - \phi\left(\frac{11.99-12}{\sqrt{\frac{0.04^2}{16}}}\right) = \phi(1) - \phi(-1) = 2\phi(1) - 1 =$$

$$= 2 \times 0.8413 - 1 = 0.6826$$

b) En este caso $\overline{X} \sim N\left(12, \dfrac{0.04^2}{25}\right)$, entonces

$$P(\overline{X} > 12.01) = 1 - \phi\left(\frac{12.01-12}{\sqrt{\frac{0.04^2}{25}}}\right) = 1 - \phi(1.25) = 1 - 0.8944 = 0.1056$$

2) Una máquina embotelladora puede regularse de tal manera que llene un promedio de μ onzas por botella. Se ha observado que la cantidad de contenido que suministra la máquina presenta una distribución normal con $\sigma = 1$ onza. De la producción de la máquina un cierto día, se obtiene una muestra de 9 botellas llenas (todas fueron llenadas con las mismas posiciones del control operativo) y se miden las onzas del contenido de cada una.
a) Determinar la probabilidad de que la media muestral se encuentre a lo más a 0.3 onzas de la media real μ para tales posiciones de control
b) ¿Cuántas observaciones deben incluirse en la muestra si se desea que la media muestral esté a lo más a 0.3 onzas de μ con una probabilidad de 0.95?

Solución:

a) Sean las variables aleatorias X_i : “contenido en onzas de la botella *i*” $i = 1,2,\ldots,9$

Entonces $X_i \sim N(\mu,1)$ para cada *i*.

Por lo tanto $\overline{X} \sim N\left(\mu, \frac{1}{9}\right)$. Se desea calcular

$$P(|\overline{X}-\mu| \le 0.3) = P(-0.3 \le \overline{X}-\mu \le 0.3) = P\left(-\frac{0.3}{\sigma/\sqrt{n}} \le \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \le \frac{0.3}{\sigma/\sqrt{n}}\right) =$$

$$= P\left(-\frac{0.3}{\sigma/\sqrt{n}} \le \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \le \frac{0.3}{\sigma/\sqrt{n}}\right) = P\left(-0.9 \le \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \le 0.9\right) = \Phi(0.9) - \Phi(-0.9) =$$

$$= 2\Phi(0.9) - 1 = 0.6318$$

b) Ahora se pretende que

$$P(|\overline{X}-\mu| \le 0.3) = P(-0.3 \le \overline{X}-\mu \le 0.3) = 0.95$$

Entonces

$$P(|\overline{X}-\mu| \le 0.3) = P\left(\frac{-0.3}{\sigma/\sqrt{n}} \le \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \le \frac{0.3}{\sigma/\sqrt{n}}\right) = P\left(-0.3\sqrt{n} \le \frac{\overline{X}-\mu}{1/\sqrt{n}} \le 0.3\sqrt{n}\right) = 0.95$$

Mediante la tabla de la acumulada de la normal estándar se tiene que

$$P\left(-0.3\sqrt{n} \le \frac{\overline{X}-\mu}{1/\sqrt{n}} \le 0.3\sqrt{n}\right) = 2\Phi\left(0.3\sqrt{n}\right) - 1 = 0.95 \;\Rightarrow\; \Phi\left(0.3\sqrt{n}\right) = 0.975 \;\Rightarrow\; \left(0.3\sqrt{n}\right) = 1.96$$

O sea $n \approx \left(\frac{1.96}{0.3}\right)^2 = 42.68$

Si tomamos $n = 43$, entonces $P(|\overline{X}-\mu| \le 0.3)$ será un poco mayor que 0.95

6.2 - Teorema central del límite

Acabamos de ver que la suma de un número finito *n* de variables aleatorias independientes que están normalmente distribuidas es una variable aleatoria también normalmente distribuida. Esta propiedad reproductiva no es exclusiva de la distribución normal. En efecto, por ejemplo, ya vimos que existen variables aleatorias discretas que la cumplen, es el caso de la Poisson y la Binomial.

En realidad, la propiedad que le da a la distribución normal el lugar privilegiado que ocupa entre todas las distribuciones es el hecho de que la suma de un número muy grande, rigurosamente un número infinito numerable, de variables aleatorias independientes con distribuciones ***arbitrarias*** (no necesariamente normales) es una variable aleatoria que tiene, aproximadamente, una distribución normal. Este es, esencialmente, el contenido del

<u>Teorema central del límite (T.C.L.)</u>:
Sean $X_1, X_2, \ldots, X_n$ variables aleatorias independientes con $E(X_i) = \mu$ y $V(X_i) = \sigma^2$ para todo $i = 1,2,\ldots,n$, es decir ***independientes idénticamente distribuidas***

Sea la v.a. $S_n = \sum_{i=1}^{n} X_i$ y sea $Z_n = \dfrac{S_n - n\mu}{\sqrt{n\sigma^2}}$.

Entonces $\lim_{n\to\infty} P(Z_n \le z) = \Phi(z)$, esto es $\lim_{n\to\infty} P\left(\dfrac{S_n - n\mu}{\sqrt{n\sigma^2}} \le z\right) = \dfrac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-x^2/2}\, dx$

Dem.) sin demostración

<u>Observaciones</u>:

1- Notar que $E(S_n) = E\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} E(X_i) = n\mu$ y $V(S_n) = V\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} V(X_i) = n\sigma^2$

Por lo tanto $Z_n = \dfrac{S_n - n\mu}{\sqrt{n\sigma^2}}$ es la v.a. S_n ***estandarizada***

2- Notar que $P\left(\dfrac{S_n - n\mu}{\sqrt{n\sigma^2}} \le z\right) = P\left(\dfrac{\frac{S_n - n\mu}{n}}{\frac{\sqrt{n\sigma^2}}{n}} \le z\right) = P\left(\dfrac{\overline{X} - \mu}{\sigma/\sqrt{n}}\right)$, por lo tanto también se puede enunciar el Teorema central del límite de la siguiente forma

Sean $X_1, X_2, \ldots, X_n$ variables aleatorias independientes con $E(X_i) = \mu$ y $V(X_i) = \sigma^2$ para todo $i = 1,2,\ldots,n$, es decir ***independientes idénticamente distribuidas***

Sea la v.a. promedio muestral $\overline{X} = \dfrac{1}{n}\sum_{i=1}^{n} X_i$ y sea $Z_n = \dfrac{\overline{X} - \mu}{\sigma/\sqrt{n}}$.

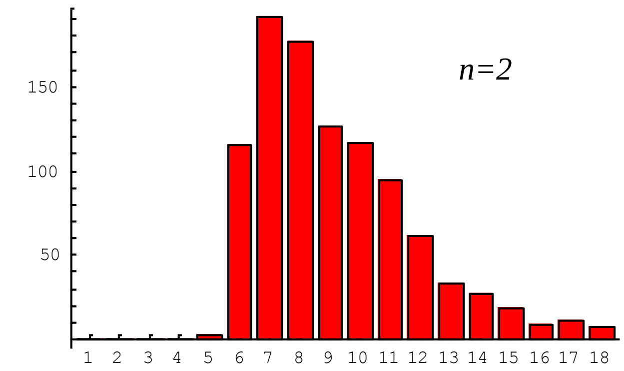
Entonces $\lim_{n\to\infty} P(Z_n \le z) = \Phi(z)$, esto es $\lim_{n\to\infty} P\left(\dfrac{\overline{X} - \mu}{\sigma/\sqrt{n}} \le z\right) = \dfrac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-x^2/2}\, dx$

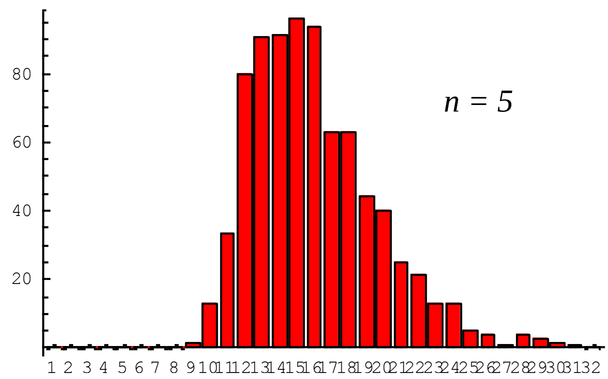
Donde $Z_n = \dfrac{\overline{X} - \mu}{\sigma/\sqrt{n}}$ es el ***promedio muestral estandarizado***

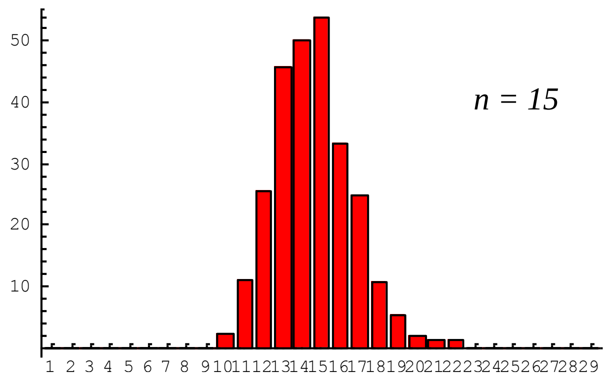
3- Aunque en muchos casos el T.C.L. funciona bien para valores de *n* pequeños , en particular donde la población es continua y simétrica, en otras situaciones se requieren valores de *n* mas grandes, dependiendo de la forma de la distribución de las X_i. En muchos casos de interés práctico, si $n \ge 30$, la aproximación normal será satisfactoria sin importar cómo sea la forma de la distribución de las X_i . Si $n < 30$, el T.C.L. funciona si la distribución de las X_i no está muy alejada de una distribución normal

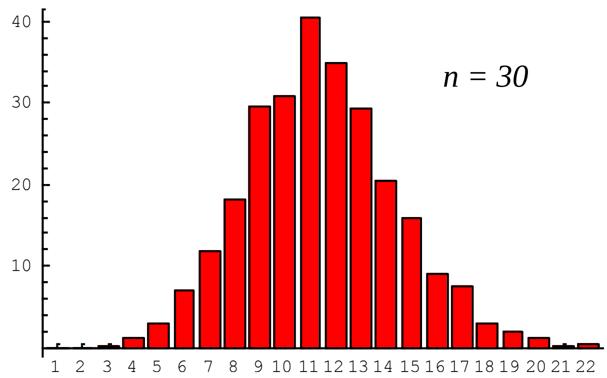
4- Para interpretar el significado del T.C.L., se generan (por computadora) *n* valores de una v.a. exponencial con parámetro $\lambda = 0.5$, y se calcula el promedio de esos *n* valores. Esto se repite 1000 veces, por lo tanto tenemos 1000 valores de la v.a. $\overline{X}$.

Hacemos un ***histograma de frecuencias*** de $\overline{X}$, esto es, tomamos un intervalo (a,b) donde "caen" todos los valores de $\overline{X}$, y lo subdividimos en intervalos mas chicos de igual longitud. La ***frecuencia de cada subintervalo*** es la cantidad de valores de $\overline{X}$ que caen en dicho subintervalo. Se grafican estas frecuencias obteniéndose los gráficos siguientes que se pueden considerar una aproximación a la verdadera distribución de $\overline{X}$.
Se observa que a medida que aumenta el valor de n los gráficos se van haciendo más simétricos, pareciéndose a la gráfica de una distribución normal.









<u>Ejemplos</u>:
1- Supóngase que 30 instrumentos electrónicos D_1, D_2, ...,D_{30}, se usan de la manera siguiente: tan pronto como D_1 falla empieza a actuar D_2. Cuando D_2 falla empieza a actuar D_3, etc. Supóngase que el tiempo de falla de D_i es una v.a. distribuida exponencialmente con parámetro $\lambda = 0.1$ por hora. Sea T el tiempo total de operación de los 30 instrumentos. ¿Cuál es la probabilidad de que T exceda 350 horas?

<u>Solución</u>:
Si X_i: "tiempo de falla del instrumento D_i" $i = 1,2,\ldots,30$

Entonces $X_i \sim Exp(0.1)$ para $i = 1,2,\ldots,30$

El tiempo total de operación de los 30 instrumentos es $T = \sum_{i=1}^{30} X_i$, donde

$$E(T) = E\left(\sum_{i=1}^{30} X_i\right) = 30 \times E(X_i) = 30 \times \frac{1}{0.1} = 300$$

$$V(T) = V\left(\sum_{i=1}^{30} X_i\right) = 30 \times V(X_i) = 30 \times \frac{1}{0.1^2} = 3000$$

Entonces por T.C.L. $\frac{T-300}{\sqrt{3000}} \sim N(0,1)$ ***aproximadamente*** pues $n = 30$

La probabilidad pedida es

$$P(T > 350) = P\left(\frac{T-300}{\sqrt{3000}} > \frac{350-300}{\sqrt{3000}}\right) \underset{\text{T.C.L.}}{\approx} 1 - \Phi\left(\frac{350-300}{\sqrt{3000}}\right) = 1 - \Phi(0.9128) = 1 - 0.81859 = 0.18141$$

2- Suponga que el consumo de calorías por día de una determinada persona es una v.a. con media 3000 calorías y desviación estándar de 230 calorías. ¿Cuál es la probabilidad de que el promedio de consumo de calorías diario de dicha persona en el siguiente año (365 días) sea entre 2959 y 3050?

<u>Solución</u>:

Definimos las variables aleatorias

X_i: "cantidad de calorías que una persona consume en el día *i*" $i = 1,2,\ldots,365$

Se sabe que $E(X_i) = 3000$ y $V(X_i) = 230^2$

Si $\overline{X} = \frac{1}{365}\sum_{i=1}^{365} X_i$ entonces $E(\overline{X}) = 3000$ y $V(\overline{X}) = \frac{\sigma^2}{n} = \frac{230^2}{365}$

La probabilidad pedida es

$$P\left(2959 \leq \overline{X} \leq 3050\right) = P\left(\frac{2959-3000}{230/\sqrt{365}} \leq \frac{\overline{X}-3000}{230/\sqrt{365}} \leq \frac{3050-3000}{230/\sqrt{365}}\right) \underset{\text{T.C.L.}}{\approx}$$

$$\approx \Phi\left(\frac{3050-3000}{230/\sqrt{365}}\right) - \Phi\left(\frac{2959-3000}{230/\sqrt{365}}\right) = \Phi(4.15) - \Phi(-3.40) \approx 1 - 0 = 1$$

<u>Aplicaciones del Teorema central del límite</u>

<u>Aproximación normal a la distribución binomial</u>

El Teorema central del límite se puede utilizar para aproximar las probabilidades de algunas variables aleatorias discretas cuando es difícil calcular las probabilidades exactas para valores grandes de los parámetros.

Supongamos que *X* tiene una distribución binomial con parámetros *n* y *p*. Para calcular $P(X \leq k)$ debemos hacer la suma $P(X \leq k) = \sum_{i=0}^{k} P(X = i)$ o recurrir a las tablas de la F.d.a. , pero para valores de *n* grandes no existen tablas, por lo tanto habría que hacer el cálculo en forma directa y muchas veces es laborioso.

Como una opción podemos considerar a X como suma de variables aleatorias más simples, específicamente, si definimos

$$X_i = \begin{cases} 1 \quad si \quad en \quad la \quad í-ésima \quad repetición \quad de \quad \varepsilon \quad ocurre \quad éxito \\ \\ \qquad 0 \qquad caso \quad contrario \end{cases} \qquad i = 1,2,\ldots,n$$

entonces cada X_i se la puede considerar $B(1,p)$, y además $X_1, X_2, \ldots, X_n$ son independientes

Podemos escribir $X = X_1 + X_2 + \ldots + X_n = \sum_{i=1}^{n} X_i$ y ***si n es grande*** entonces X tendrá ***aproximadamente*** una distribución normal con parámetros np y $np(1-p)$, es decir

$$Z_n = \frac{X - n\mu}{\sqrt{n\sigma^2}} = \frac{X - n.p}{\sqrt{n.p(1-p)}} \approx N(0,1) \quad \text{si } n \text{ es lo suficientemente grande}$$

Observaciones:

1- La aproximación normal a la distribución binomial funciona bien aun cuando n no sea muy grande si p no está demasiado cerca de cero o de uno. En particular la aproximación normal a la binomial es buena si n es grande , $np > 5$ y $n(1-p) > 5$, ***pero es más efectivo aplicar esta aproximación cuando*** $np > 10$ ***y*** $n(1-p) > 10$

2- Corrección por continuidad.

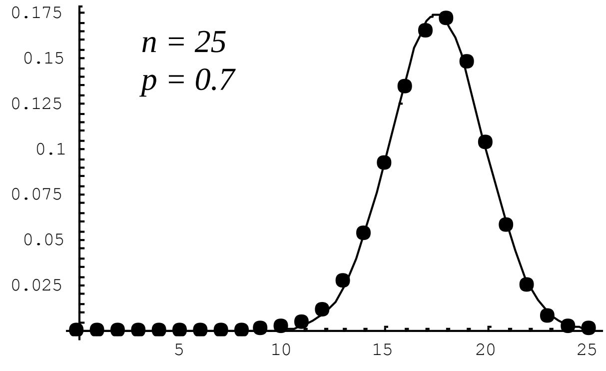
Acabamos de ver que si $X \sim B(n,p)$ entonces, para n suficientemente grande, podemos considerar que aproximadamente es $X \sim N[n.p, n.p(1-p)]$. El problema que surge de inmediato si deseo calcular, por ejemplo, la probabilidad de que $X = k$ (con k alguno de los valores posibles $0,1,2,\ldots,n$) es que la binomial es una distribución discreta y tiene sentido calcular probabilidades como $P(X = k)$ mientras que la normal es una distribución continua y, en consecuencia, $P(X = k) = 0$ puesto que para una variable aleatoria continua la probabilidad de que ésta tome un valor aislado es cero. Esto se resuelve si se considera $P(X = k) \approx P\left(k - \frac{1}{2} \le X \le k + \frac{1}{2}\right)$

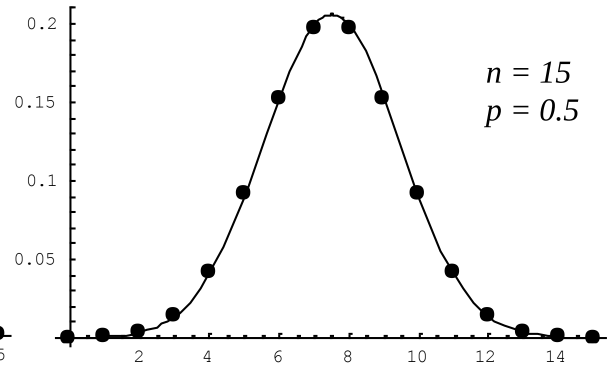
También se puede usar esta corrección para mejorar la aproximación en otros casos, específicamente en lugar de $P(X \le k)$ calculamos

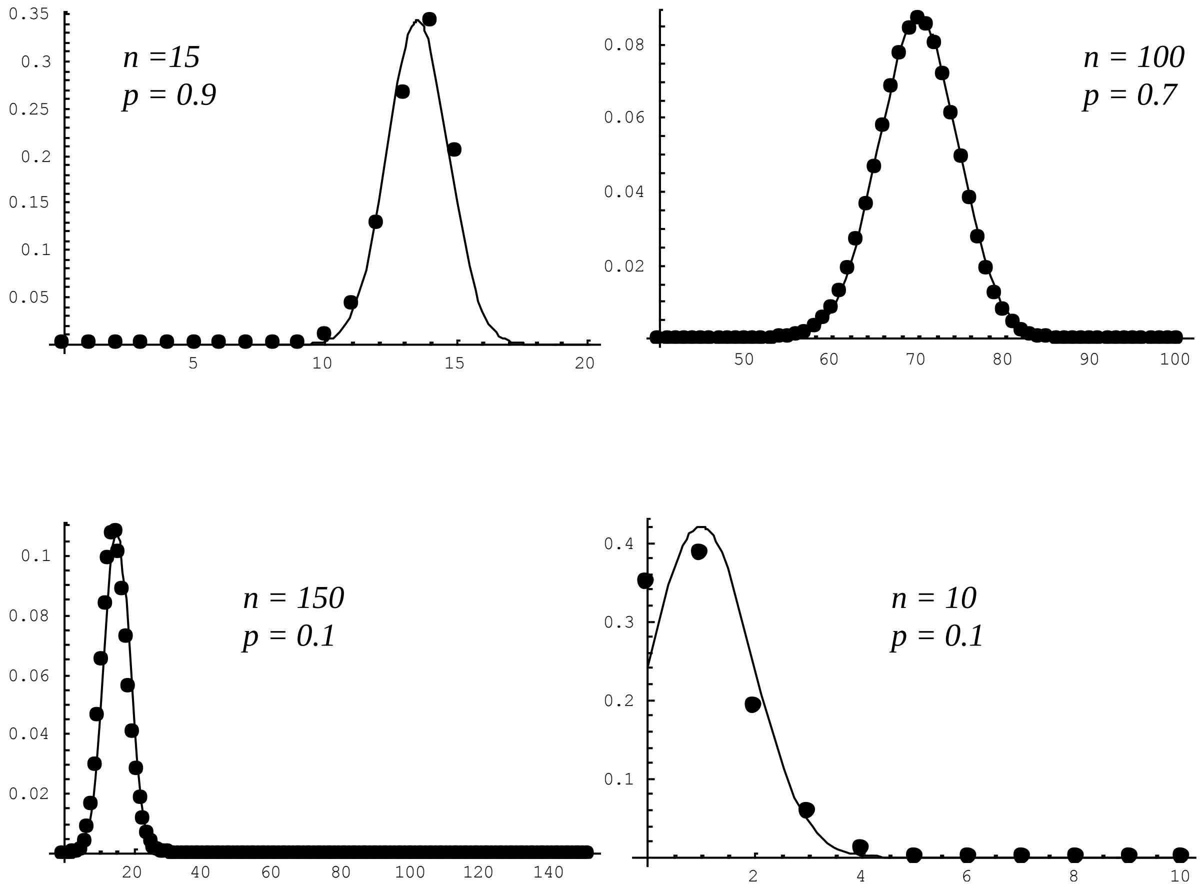
$$P(X \le k) \approx P\left(X \le k + \frac{1}{2}\right)$$

Y en lugar de $P(X \ge k) \approx P\left(X \ge k - \frac{1}{2}\right)$

En los gráficos siguientes se muestra para diferentes valores de n y p cómo aproxima la distribución $N(np,\ np(1-p))$ a la distribución $B(n,p)$







<u>Ejemplos</u>:

1- Sea $X\sim B(25,0.4)$. Hallar las probabilidades exactas de que $X \le 8$ y $X = 8$ y comparar estos resultados con los valores correspondientes encontrados por la aproximación normal.

<u>Solución</u>:

De la tabla de la $F.d.a.$ de la binomial encontramos $P(X \le 8) = 0.274$

Y $P(X = 8) = P(X \le 8) - P(X \le 7) = 0.274 - 0.154 = 0.120$

Ahora usamos la aproximación normal

$$P(X \le 8) \approx P(X \le 8.5) = P\left(\frac{X - np}{\sqrt{np(1-p)}} \le \frac{8.5-10}{\sqrt{25\times 0.4\times 0.6}}\right) \approx \Phi(-0.61) = 0.2709$$

corrección por continuidad

Observar que el valor aproximado está muy cercano al valor exacto para $P(X \le 8) = 0.274$

$$P(X = 8) \approx P(7.5 \le X \le 8.5) = P\left(\frac{7.5-10}{\sqrt{6}} \le \frac{X-10}{\sqrt{6}} \le \frac{8.5-10}{\sqrt{6}}\right) = P\left(-1.02 \le \frac{X-10}{\sqrt{6}} \le -0.61\right) =$$

$$= 0.2709 - 0.1593 = 0.1170$$

Nuevamente este valor aproximado está muy cerca del valor real de $P(X = 8) = 0.120$

2- Suponga que el 10% de todos los ejes de acero producidos por cierto proceso están fuera de especificaciones, pero se pueden volver a trabajar (en lugar de tener que enviarlos a la chatarra). Considere una muestra aleatoria de 200 ejes y denote por X el número entre ellos que estén fuera de especificaciones y se puedan volver a trabajar. ¿Cuál es la probabilidad (aproximada) de que X sea

a) a lo sumo 30?
b) menos de 30?
c) entre 15 y 25 (inclusive)?

Solución:
Sea la v.a. X: "número de ejes fuera de especificaciones"
Entonces $X \sim B(200,\ \ 0.1)$, además $np = 200 \times 0.1 = 20 > 5$ y $n(1-p) = 200 \times (1-0.1) = 180 > 5$

Por lo tanto podemos aplicar la aproximación normal a la binomial
a) la probabilidad pedida es $P(X \le 30)$

$$P(X \le 30) \approx P(X \le 30.5) = P\left(\frac{X-np}{\sqrt{np(1-p)}} \le \frac{30.5-20}{\sqrt{18}}\right) \approx \Phi\left(\frac{30.5-20}{\sqrt{18}}\right) = \Phi(2.474) = 0.993244$$

b) La probabilidad pedida es $P(X < 30)$
Al ser X una v.a. ***discreta*** con distribución binomial $P(X < 30) = P(X \le 29)$

$$P(X \le 29) \approx P(X \le 29.5) \approx \Phi\left(\frac{29.5-20}{\sqrt{18}}\right) = \Phi(2.2391) = 0.98745$$

c)

$$P(15 \le X \le 25) \approx P(14.5 \le X \le 25.5) \approx \Phi\left(\frac{25.5-20}{\sqrt{18}}\right) - \Phi\left(\frac{14.5-20}{\sqrt{18}}\right) =$$

$$= \Phi(1.2963) - \Phi(-1.2963) = 2\Phi(1.2963) - 1 = 2 \times 0.90147 - 1 = 0.80294$$

3- El gerente de un supermercado desea recabar información sobre la proporción de clientes a los que no les agrada una nueva política respecto de la aceptación de cheques. ¿Cuántos clientes tendría que incluir en una muestra si desea que la fracción de la muestra se desvíe a lo mas en 0.15 de la verdadera fracción, con probabilidad de 0.98?.

Solución:
Sea X: "número de clientes a los que no les agrada la nueva política de aceptación de cheques"
Entonces $X \sim B(n,\ \ p)$ donde p es desconocido y es la ***verdadera proporción*** de clientes a los que no les agrada la nueva política de aceptación de cheques. El gerente tomará una muestra de n clientes para ***"estimar"*** p con $\overline{X} = \dfrac{X}{n}$ ya que $\overline{X} = \dfrac{X}{n}$ es la proporción de clientes a los que no les agrada la nueva política de aceptación de cheques ***en la muestra de n clientes***. Si no se toman a ***todos los clientes,*** entonces $\overline{X} = \dfrac{X}{n}$ ***no será igual a p***.

La pregunta es cuál debe ser n para que $\overline{X} = \dfrac{X}{n}$ se aleje del verdadero p en menos de 0.15 con probabilidad 0.98 por lo menos, o sea para que $P\left(\left|\overline{X} - p\right| \le 0.15\right) \ge 0.98$
Entonces planteamos

$$P\left(\left|\overline{X} - p\right| \le 0.15\right) = P\left(-0.15 \le \overline{X} - p \le 0.15\right) = P\left(\frac{-0.15n}{\sqrt{np(1-p)}} \le \frac{X-np}{\sqrt{np(1-p)}} \le \frac{0.15n}{\sqrt{np(1-p)}}\right) \approx$$

$$\approx \Phi\left(\frac{0.15n}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{-0.15n}{\sqrt{np(1-p)}}\right) = 2\Phi\left(\frac{0.15n}{\sqrt{np(1-p)}}\right) - 1 \geq 0.98$$

Por lo tanto $\Phi\left(\frac{0.15n}{\sqrt{np(1-p)}}\right) \geq \frac{0.98+1}{2} = 0.99$

Además $\frac{0.15n}{\sqrt{np(1-p)}} = \frac{0.15\sqrt{n}}{\sqrt{p(1-p)}} \geq \frac{0.15\sqrt{n}}{\sqrt{0.5(1-0.5)}} = 0.3\sqrt{n}$

Entonces debe cumplirse que $0.3\sqrt{n} \geq 2.33$ o sea $n \geq \left(\frac{2.33}{0.3}\right)^2 = 60.3211$

O sea ***se debe tomar una muestra de al menos 61 clientes***

Aproximación normal a la distribución Poisson

Se puede probar aplicando Teorema central del límite que

> Si $X \sim P(\lambda)$ entonces para λ suficientemente grande $\frac{X-\lambda}{\sqrt{\lambda}}$ tiene aproximadamente distribución $N(0,1)$

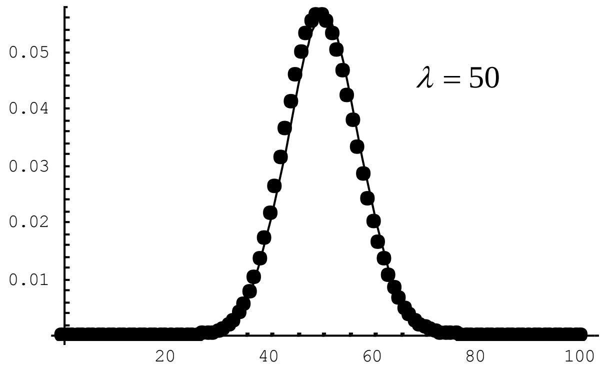
Es decir para λ suficientemente grande $\frac{X-\lambda}{\sqrt{\lambda}} \approx N(0,1)$
En la práctica si $\lambda \geq 30$ la aproximación es buena.

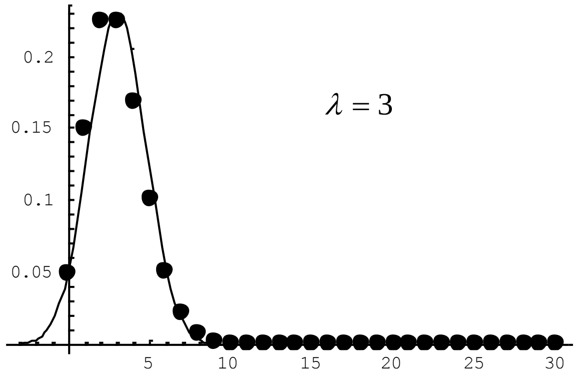
Observación: la demostración es sencilla si λ es igual a un número natural *n* pues, si consideramos las variables aleatorias $X_i \sim P(1)$ con $i = 1,2,\ldots,n$ independientes, entonces ya sabemos que $\sum_{i=1}^{n} X_i \sim P\left(\sum_{i=1}^{n} 1\right)$, es decir $\sum_{i=1}^{n} X_i \sim P(n)$

Pero además por T.C.L. si *n* es grande $\sum_{i=1}^{n} X_i$ tiene aproximadamente distribución normal con parámetros $n\mu = n \times 1 = n$ y $n\sigma^2 = n \times 1 = n$

O sea la distribución de $\sum_{i=1}^{n} X_i$ que es exactamente Poisson con parámetro *n*, se puede aproximar con una $N(n,n)$, por lo tanto $\frac{X-n}{\sqrt{n}} \approx N(0,1)$ aproximadamente para valores de *n* suficientemente grandes

En los gráficos siguientes se muestra para diferentes valores de λ cómo aproxima la distribución $N(\lambda,\ \lambda)$ a la distribución $P(\lambda)$





Ejemplo:

El número de infracciones por estacionamiento en cierta ciudad en cualquier día hábil tiene una distribución de Poisson con parámetro λ = 50. ¿Cuál es la probabilidad aproximada de que:

a) más de 35 y a lo sumo 70 infracciones se expidan en un día en particular?

b) el número total de infracciones expedidas durante una semana de 5 días sea más 225 y a lo sumo 275?

Solución:

Sea *X*: "número de infracciones por estacionamiento en cierta ciudad en cualquier día hábil"

Entonces $X \sim P(\lambda)$ donde $\lambda = 50$

Como $\lambda = 50$ entonces $\dfrac{X-50}{\sqrt{50}} \approx N(0,1)$ (aproximadamente)

a) la probabilidad pedida es

$$P(35 < X \le 70) \approx \Phi\left(\frac{70-50}{\sqrt{50}}\right) - \Phi\left(\frac{35-50}{\sqrt{50}}\right) = \Phi(2.8284) - \Phi(-2.12132) =$$

$$= 0.997599 - 0.017 = 0.9805$$

b) Sea *Y*: "número total de infracciones expedidas durante una semana de 5 días"

Entonces $Y \sim P(\lambda)$ donde $\lambda = 50 \times 5 = 250$

La probabilidad pedida es

$$P(225 < Y \le 275) \approx \Phi\left(\frac{275-250}{\sqrt{250}}\right) - \Phi\left(\frac{225-250}{\sqrt{250}}\right) = \Phi(1.5811) - \Phi(-1.5811) =$$

$$= 2\Phi(1.5811) - 1 = 2 \times 0.94295 - 1 = 0.8859$$