# Distribuciones de Probabilidad

Sesión 4

Dr. Carlos J. Vilalta (DAP)
Estadística - Curso propedéutico MAPP 2016
Centro de Investigación y Docencia económicas (CIDE)
Contacto: carlos.vilalta@cide.edu

# Lo que vamos a ver hoy

1. *Aplicación 1 de valor esperado y varianza*
2. Conceptos básicos de distribuciones de probabilidad
3. Distribución Binomial
4. *Aplicación 2 de la distribución Binomial*
5. Distribución de Poisson
6. *Aplicación 3 de la distribución de Poisson*
7. Repaso de lo que hemos hecho hasta el momento
8. Distribución Normal Estándar
9. *Aplicación 4 de la distribución Normal Estándar*

# Utilidad en Políticas Públicas

- Vilalta, C. 2016. A descriptive model of residential burglary in Mexico: A contextual approach. Status: In process.

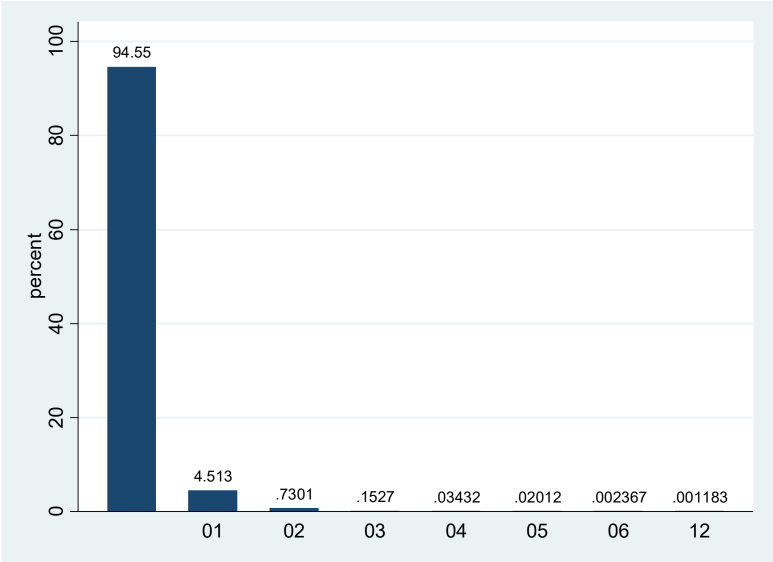
- Pregunta: ¿cuál es la probabilidad de ser víctima de un robo en vivienda?
  - Población: adultos en hogares en todo el país
  - Fuente: Encuesta Nacional de Victimización y Percepción sobre Seguridad Pública, 2015
  - Evidencia: reporte de delitos en encuesta en hogares (entrevista cara a cara)
    - Puede existir un efecto de "telescopio" de estos problemas sociales lo que lleva a una inflación en el reporte de frecuencias

- Veamos las probabilidades de este fenómeno

# Utilidad en Políticas Públicas

- La frecuencia relativa de cada valor posible nos da las probabilidades de la victimización de robo en vivienda:

| ap6_6_04 | Freq. | Percent | Cum. |
|---|---|---|---|
| | 79,898 | 94.55 | 94.55 |
| 01 | 3,814 | 4.51 | 99.06 |
| 02 | 617 | 0.73 | 99.79 |
| 03 | 129 | 0.15 | 99.94 |
| 04 | 29 | 0.03 | 99.98 |
| 05 | 17 | 0.02 | 100.00 |
| 06 | 2 | 0.00 | 100.00 |
| 12 | 1 | 0.00 | 100.00 |
| Total | 84,507 | 100.00 | |




- Hechos cuantitativos importantes en este problema:
  - No victimización en la muestra = 94.55% de adultos
  - Victimización en la muestra = 5.45% de adultos
  - Victimización repetida (2 o más veces) en la muestra = 795/84,507 = 0.94% de adultos
  - Número de víctimas en la muestra = 4,609
  - Número de robos en la muestra = 5,660

# Utilidad en Políticas Públicas

- ¿Cuál es el valor esperado o esperanza matemática (valor medio del fenómeno aleatorio o valor más probable) de la variable "victimización por robo en vivienda"?

| ap6_6_04 | Freq. | Percent | Cum. |
|---|---|---|---|
| | 79,898 | 94.55 | 94.55 |
| 01 | 3,814 | 4.51 | 99.06 |
| 02 | 617 | 0.73 | 99.79 |
| 03 | 129 | 0.15 | 99.94 |
| 04 | 29 | 0.03 | 99.98 |
| 05 | 17 | 0.02 | 100.00 |
| 06 | 2 | 0.00 | 100.00 |
| 12 | 1 | 0.00 | 100.00 |
| Total | 84,507 | 100.00 | |

Nota sobre 2 mediciones diferentes:

- Prevalencia: es la proporción de adultos que sufren el delito

Tasa de prevalencia = 4,609/84,507 = 5.45% = 5,545 adultos por cada 100 mil

- Incidencia: es el número de veces que ocurre un delito entre el número de adultos

Tasa de incidencia = 5,660/85,558 = 6.6% = 6,615 adultos por cada 100 mil

# Utilidad en Políticas Públicas

- ¿Cuál es el valor esperado o esperanza matemática de "victimización por robo en vivienda"?
  - Es decir: ¿cuántos robos (incidencia delictiva) y de qué forma distribuidos podríamos anticipar que sucedan en 2016?
  - Ojo: el valor esperado <u>no tiene que ser un valor</u> de la variable…

| ap6_6_04 | Freq. | Percent | Cum. |
|---|---|---|---|
| | 79,898 | 94.55 | 94.55 |
| 01 | 3,814 | 4.51 | 99.06 |
| 02 | 617 | 0.73 | 99.79 |
| 03 | 129 | 0.15 | 99.94 |
| 04 | 29 | 0.03 | 99.98 |
| 05 | 17 | 0.02 | 100.00 |
| 06 | 2 | 0.00 | 100.00 |
| 12 | 1 | 0.00 | 100.00 |
| Total | 84,507 | 100.00 | |

| x | f(x) | xf(x) | Robos esperados (n*xf(x)) |
|---|---|---|---|
| 0 | 0.945 | 0.000 | - |
| 1 | 0.045 | 0.045 | 3,814 |
| 2 | 0.007 | 0.015 | 1,234 |
| 3 | 0.002 | 0.005 | 387 |
| 4 | 0.000 | 0.001 | 116 |
| 5 | 0.000 | 0.001 | 85 |
| 6 | 0.000 | 0.000 | 12 |
| 12 | 0.000 | 0.000 | 12 |
| Suma = | 1.000 | | |
| VE ($\mu$) = | | 0.067 | 5,660 |

- Se esperan 0.067 robos por cada adulto…
- O… 5,660 en una muestra de procedimiento y tamaño idéntico
- O… 6,615 por cada cien mil adultos (la tasa de prevalencia)
- O… 5,505,952 robos (ya que VE = 0.067 y hay 82,206,973 adultos en 2016 según CONAPO)

# Utilidad en Políticas Públicas

- ¿Y la varianza o desviación estándar? En esta caso es <u>enorme y no tiene utilidad práctica política esta predicción</u>
  - Recordar el teorema de Chebyshev

| x | f(x) | x * f(x) | x - $\mu$ | $(x-\mu)^2$ | $(x-\mu)^2$ * f(x) |
|---|---|---|---|---|---|
| 0 | 0.945 | 0.000 | -0.067 | 0.004 | 0.004 |
| 1 | 0.045 | 0.045 | 0.933 | 0.871 | 0.039 |
| 2 | 0.007 | 0.015 | 1.933 | 3.737 | 0.027 |
| 3 | 0.002 | 0.005 | 2.933 | 8.603 | 0.013 |
| 4 | 0.000 | 0.001 | 3.933 | 15.469 | 0.005 |
| 5 | 0.000 | 0.001 | 4.933 | 24.335 | 0.005 |
| 6 | 0.000 | 0.000 | 5.933 | 35.201 | 0.001 |
| 12 | 0.000 | 0.000 | 11.933 | 142.397 | 0.002 |
| Suma = | 1.000 | | | | |
| VE ($\mu$) = | | 0.067 | | | |
| Varianza ($\sigma^2$) = | | | | | 0.097 |
| Desv. estándar ($\sigma$) = | | | | | 0.311 |

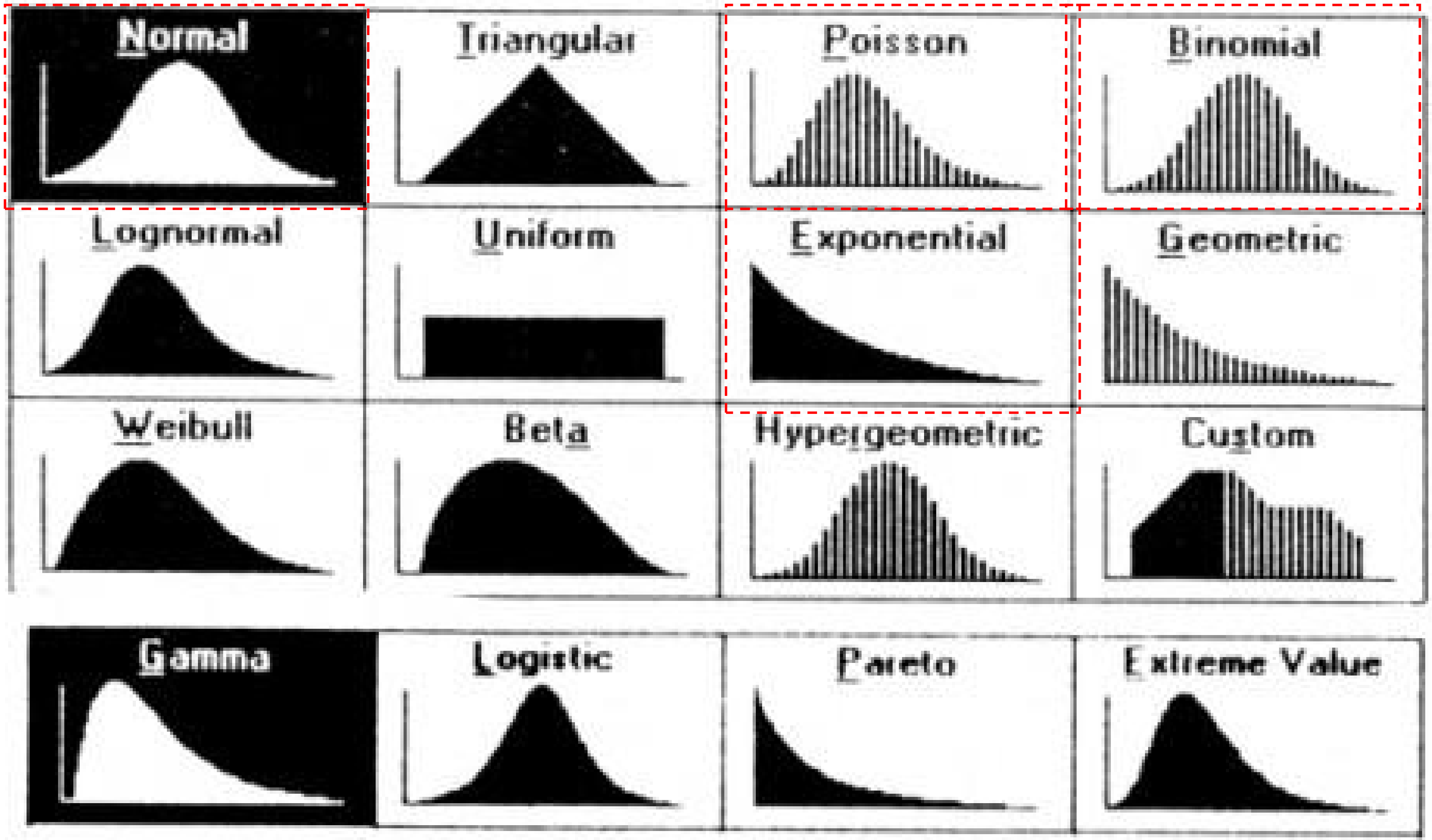
- Si VE = 0.067 robos por adulto y s = 0.31 robos por adulto, entonces…
- Podemos esperar entre 0 y 56 millones de robos… (con un 95% de probabilidad)
- Ya que $\mu + 2s = 0.067 + 0.311 + 0.311 = 0.689$
- Y… 0.689*82 millones = 56 millones

# Lecciones prácticas del ejemplo anterior

- Las probabilidades objetivas se obtienen sobre las frecuencias pasadas observadas y son mejorables conforme aumenta la evidencia
  - Cuantas más veces midamos algo, mejor será la predicción puesto que más variación habremos capturado sobre las probabilidades de todos los valores posibles de las variables aleatorias o fenómenos bajo estudio
- El valor esperado o esperanza matemática es el valor más probable de ocurrir dentro del rango posible de valores
  - En escenarios de incertidumbre es siempre la mejor estimación
  - Puede darnos la cifra de eventos más probables y cómo se distribuyen
- Los fenómenos altamente variables son difíciles de predecir porque tienen rangos de valores posibles muy amplios
  - Y en las PPs todo es muy variable
- SI hubiéramos utilizado un método clásico de probabilidad o equiprobable en vez del método de frecuencias relativas, el valor esperado y su variabilidad habrían sido imprecisos

# Las diferentes distribuciones

- Las más comunes




Barras: distribuciones para variables discretas
Áreas: distribuciones para variables continuas

# Conceptos básicos (1)

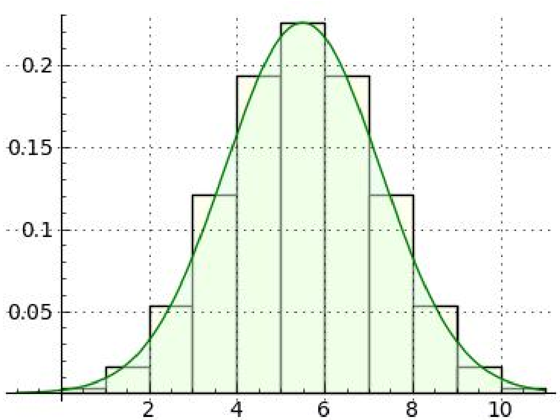
- Probabilidad:
  - La frecuencia con que esperamos que algo suceda si repetimos 1 mismo experimento o intervención de PP
- Variable aleatoria (X):
  - Instrumento numérico que se usa para describir resultados de experimentos o intervenciones de PP
  - Es una variable cuyos valores posibles se distribuyen a lo largo de un intervalo de valores
  - Cada resultado posible tiene un valor dentro de este rango de valores
    - Variable aleatoria discreta: números que sólo toman un rango finito de valores (ej. nominales y ordinales)
    - Variable aleatoria continua: números que pueden tomar un rango infinito de valores (ej. intervalos y continuas)

# Conceptos básicos (2)

- Distribuciones de probabilidades:
  - Sirven para saber (normalmente anticipar) como se distribuyen las probabilidades de los valores posible que puede tomar una variable aleatoria
- Función de probabilidades f(x)
  - Consiste de valores posibles: $x_1, x_2, \ldots, x_n$
  - Y de probabilidades posibles: $p_1, p_2, \ldots, p_n$
  - Entendiendo que:
    - $p(X = x_1) = p_1$, $p(X = x_2) = p_2$, . . . , $p(X = x_n) = p_n$
    - Y que $p_i \geq 0$, y que $p_1 + p_2 + \ldots + p_n = 1$

# Distribución binomial

- Un experimento o intervención o decisión (x) de tipo binomial puede realizarse un número indefinido (n) de veces pero sólo puede tener 2 tipos de resultados y sus probabilidades no cambian entre experimentos o intervenciones
- La distribución normal deriva de la binomial
- Aplicable a eventos que sólo pueden tener 2 resultados
  - Ej.: éxito o fracaso de una intervención, lo que es igual a p y 1-p
  - Nótese que los valores posibles en esta distribución son finitos, por ende, es aplicable solamente a variables discretas




Función de probabilidad

$$f(x) = \binom{n}{x} p^x (1-p)^{(n-x)}$$

$$\binom{n}{x} = \frac{n!}{x!(n-x)!}$$

# Utilidad en Políticas Públicas

- En 2014, 16 reclusos en promedio egresaron diariamente de las cárceles de la CMDX por cumplimiento de sus sentencias. En el CIDE tenemos estimaciones de que la tasa de reincidencia en CDMX es del 33%.
- Pregunta: ¿Cuántos de ellos podemos esperar que reincidan delictivamente, sean detenidos, procesados y sentenciados condenatoriamente?

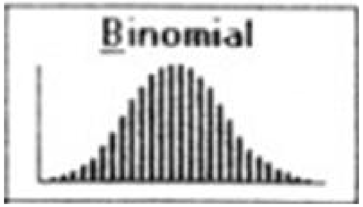
# Utilidad en Políticas Públicas

$$f(x) = \binom{n}{x} p^x (1-p)^{(n-x)}$$

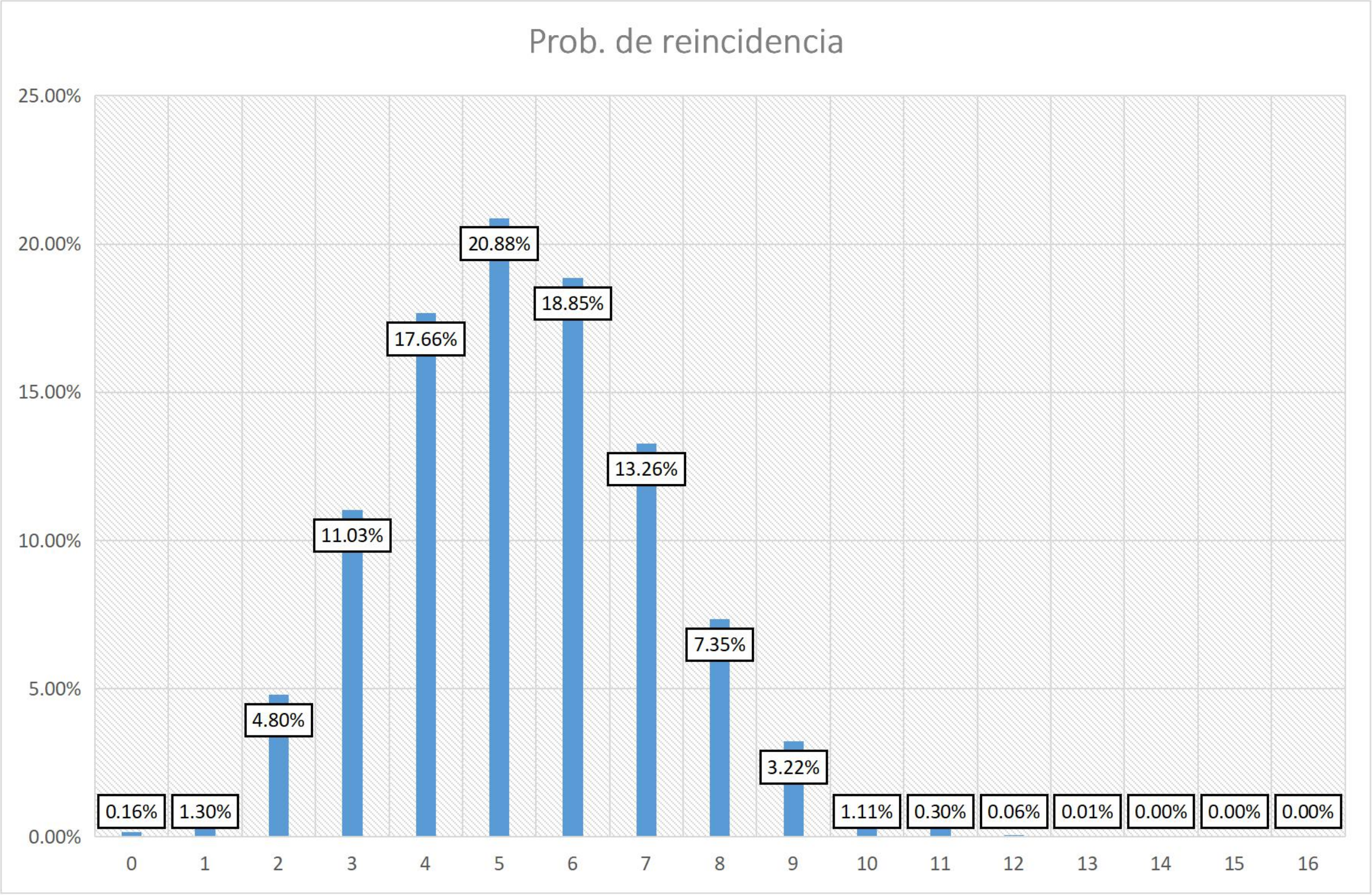
- De 16 lo más probable es que 5 regresen a la cárcel (todo constante)
- Pregunta: ¿cuál es la probabilidad de que regresen los 16?

| x | Resultados posibles | Probabilidades | función binomial | Prob. de reincidencia en % |
|---|---|---|---|---|
| 0 | 1 | 0.0016 | 0.0016 | 0.16% |
| 1 | 16 | 0.0008 | 0.0130 | 1.30% |
| 2 | 120 | 0.0004 | 0.0480 | 4.80% |
| 3 | 560 | 0.0002 | 0.1103 | 11.03% |
| 4 | 1820 | 0.0001 | 0.1766 | 17.66% |
| 5 | 4368 | 0.0000 | 0.2088 | 20.88% |
| 6 | 8008 | 0.0000 | 0.1885 | 18.85% |
| 7 | 11440 | 0.0000 | 0.1326 | 13.26% |
| 8 | 12870 | 0.0000 | 0.0735 | 7.35% |
| 9 | 11440 | 0.0000 | 0.0322 | 3.22% |
| 10 | 8008 | 0.0000 | 0.0111 | 1.11% |
| 11 | 4368 | 0.0000 | 0.0030 | 0.30% |
| 12 | 1820 | 0.0000 | 0.0006 | 0.06% |
| 13 | 560 | 0.0000 | 0.0001 | 0.01% |
| 14 | 120 | 0.0000 | 0.0000 | 0.00% |
| 15 | 16 | 0.0000 | 0.0000 | 0.00% |
| 16 | 1 | 0.0000 | 0.0000 | 0.00% |
| | | | | 100.00% |
| n = | 16 | | | |
| p (reincidencia) = | 0.33 | | | |
| q (no reincidencia) = | 0.67 | | | |

# Utilidad en Políticas Públicas




- Visto gráficamente:



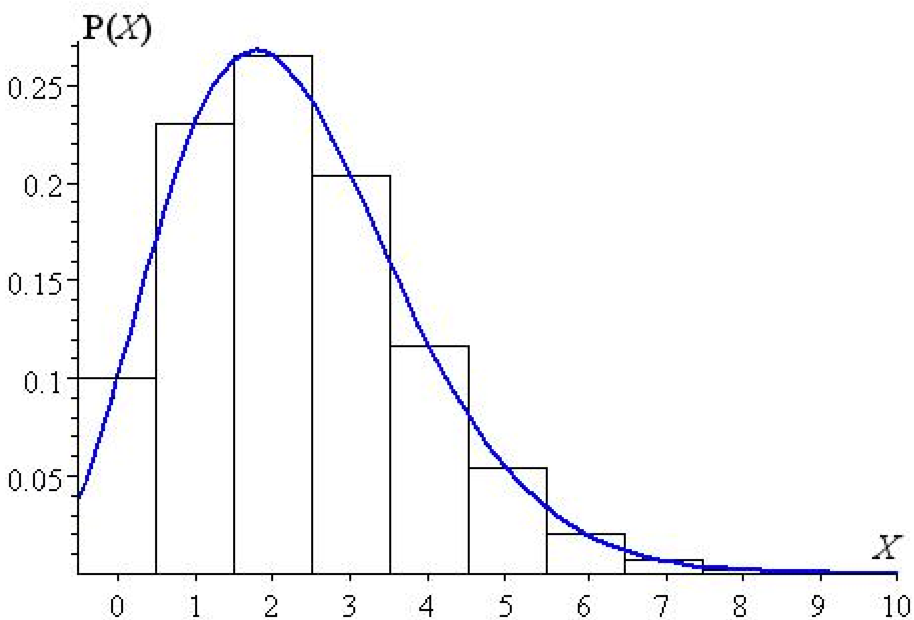
# Utilidad en Políticas Públicas

- De manera más general podemos estimar cuántos regresarán a la cárcel usando el valor esperado y la dispersión:
  - np = 16*(0.33) = 5
  - npq = 16*(0.33)*(0.67)= 4
  - √npq = 2

- Aproximando esta probabilidad binomial a una distribución normal estándar (o al revés), habría al menos un 95% de probabilidad de que regresaran:
  - Entre 2 y 9 reclusos
  - IC (95%) = 2-9
  - Nota: en el intervalo de 2 a 9 reclusos se halla el 97.1% de todas las probabilidades de reincidencia

# Distribución de Poisson

- Es utilizada cuando se hacen estimaciones de probabilidades para eventos discretos en intervalos de tiempo o bien cuando se trata de eventos poco frecuentes
- En muchas ocasiones se desconoce el valor del parámetro ($\mu$) por lo que se infiere a partir de los estadísticos muestrales (M) lo que ocasiona que las pruebas de hipótesis estén sesgadas en alguna dirección




Función de probabilidad

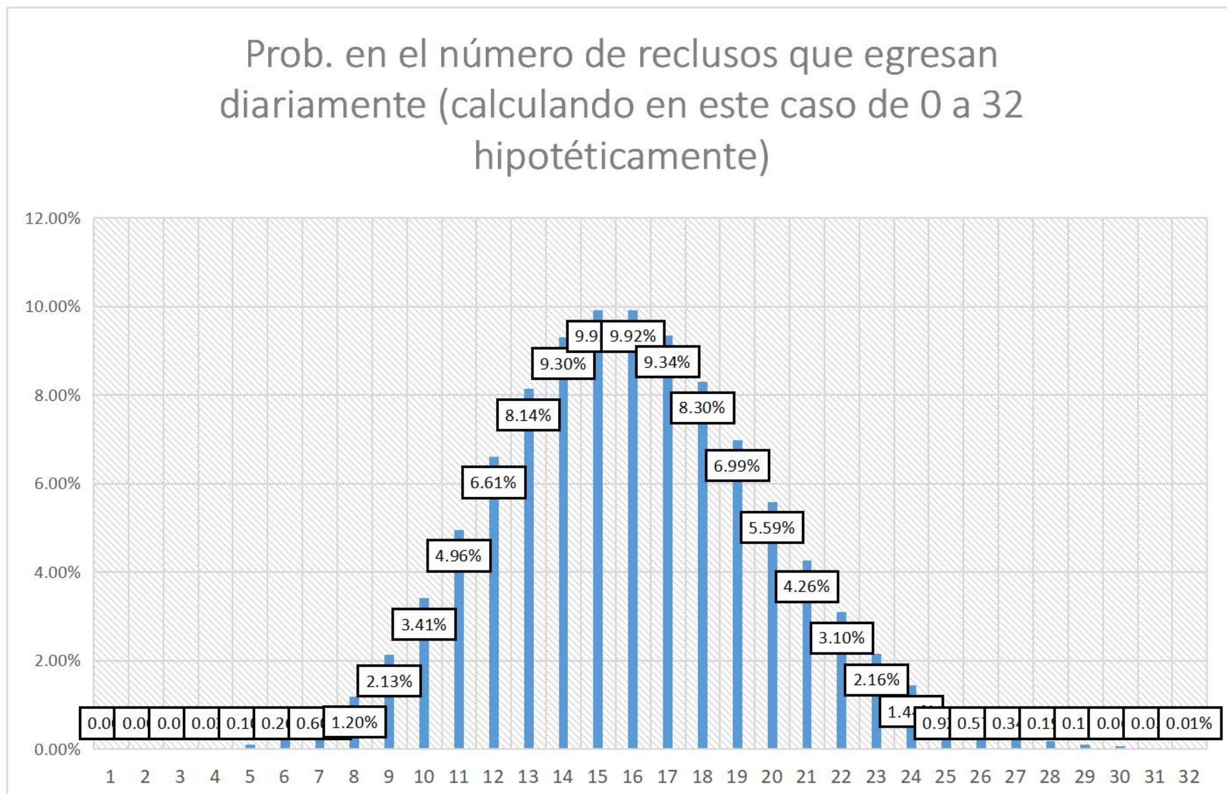
$$f(x) = \frac{\mu^x e^{-\mu}}{x!}$$

# Utilidad en Políticas Públicas

- Sobre el ejemplo anterior, ahora veamos cómo predecir el número de reclusos que podemos esperar que serán liberados diariamente
  - Pregunta: ¿cuál es la probabilidad de que sean liberados exactamente 10 (x) reclusos si la media diaria es que sean liberados 16 ($\mu$)?

$$f(x) = \frac{\mu^x e^{-\mu}}{x!} = \frac{16^{10} 2.71^{-16}}{10!} = 0.0341$$

¿Y cuál es la probabilidad de que hayan entre 10 y 12 egresos?

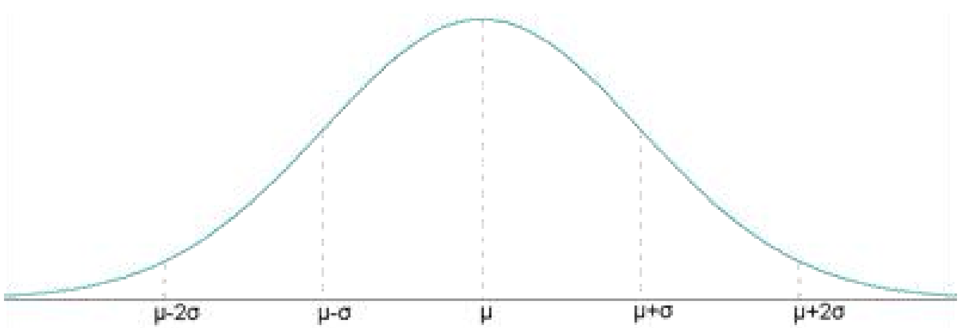


# Lo que hemos visto hasta el momento

- Las PP requieren claridad y precisión
- No obstante, la incertidumbre en las PP está siempre presente y puede tener o presentarse de muchas formas
- Estas formas pueden modelarse con distribuciones de probabilidad
  - Las distribuciones de probabilidad son los sistemas más elegantes y rigurosos con que contamos para reducir el error en la toma de decisiones
- Por lo que, en la teoría y práctica de las PP, las distribuciones de probabilidad nos ayudan a:
  - Predecir escenarios en situaciones de incertidumbre o con información limitada (o sea, siempre)
  - Tomar decisiones a través del procedimiento estándar de prueba de hipótesis nula
  - Dejar atrás el método favorito entre los políticos no profesionales de la adivinación

# Distribución Normal

- Es la que más se utiliza en las CCSS
- También se le denomina Gaussiana
- Sólo aplicable a variables continuas
- Una distribución de datos con media = 0 y desviación estándar = 1, se le llama a distribución normal estándar
  - En este caso la Media, Mediana y Moda = 0
- La diferencia entre la Distribución Normal N($\mu$,$\sigma$) y las distribuciones para variables discretas es que aquella no ofrece probabilidades directamente, sino que da un intervalo de valores probables de la variable aleatoria(PDF)




Función de densidad de probabilidades de la Normal Estándar (PDF)

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$$

# Utilidad en Políticas Públicas

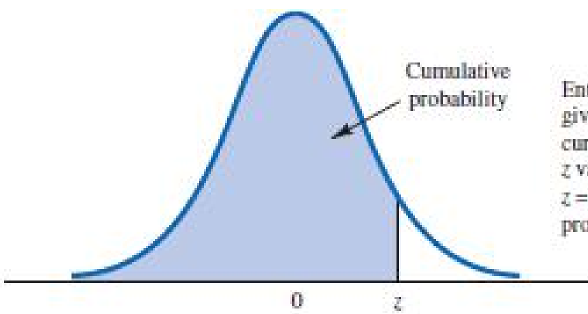
- En 2014 se robaron 15,568 vehículos en la CDMX. En 2015, 12,615. Pregunta: todo constante ¿cuántos vehículos robados podemos esperar en 2016?
  - Para estimarlo exactamente obtendríamos el valor esperado
  - Para estimar una probabilidad, tendríamos que preguntar sobre un valor exacto y ver el rango de valores posibles con puntuaciones Z
- Por ejemplo: ¿cuál es la probabilidad de que sucedan más de 15,000 (x) robos de vehículo?
  - $\mu$ de robos en años anteriores = 14,092
  - $\sigma$ de robos en años anteriores = 2,088

$$z = \frac{x-\mu}{\sigma} = \frac{15{,}000 - 14{,}092}{2{,}088} = 0.435$$

# Utilidad en Políticas Públicas

$$z = \frac{x-\mu}{\sigma} = \frac{15{,}000 - 14{,}092}{2{,}088} = 0.435$$

- Para z = 0.435, tenemos que la distribución normal estándar nos daría el siguiente área de probabilidades:




Entries in the table give the area under the curve to the left of the z value. For example, for z = 1.25, the cumulative probability is .8944.

| z | .00 | .01 | .02 | .03 | .04 | .05 | .06 | .07 | .08 | .09 |
|---|---|---|---|---|---|---|---|---|---|---|
| .0 | .5000 | .5040 | .5080 | .5120 | .5160 | .5199 | .5239 | .5279 | .5319 | .5359 |
| .1 | .5398 | .5438 | .5478 | .5517 | .5557 | .5596 | .5636 | .5675 | .5714 | .5753 |
| .2 | .5793 | .5832 | .5871 | .5910 | .5948 | .5987 | .6026 | .6064 | .6103 | .6141 |
| .3 | .6179 | .6217 | .6255 | .6293 | .6331 | .6368 | .6406 | .6443 | .6480 | .6517 |
| .4 | .6554 | .6591 | .6628 | .6664 | .6700 | .6736 | .6772 | .6808 | .6844 | .6879 |

- Por lo que la probabilidad de tener menos de 15,000 = 67.0%
- Y la probabilidad de que sean más de 15,000 = 1 – 0.67 = 33.0%

# Tarea

- En la siguiente clase veremos distribuciones muestrales en conexión con intervalos de confianza
- Anderson et al.: Resolver ejercicios (entregar a mano y con procedimientos):
  - Distribuciones Binomial y Poisson (Cap. 5): 15 y 16, 25 a 30, 38 a 45
  - Distribución Normal (Cap. 6): 8 a 20
- Entrega: jueves en el laboratorio