Tema 3: PROBABILIDAD - VI

La binomial para n grande. La distribución normal.

Biología sanitaria 2017/18. Universidad de Alcalá

M. Marvá. Actualizado: 2017-10-25

¿Qué sucede con la binomial cuando n se hace grande y p es moderado?

GEOGEBRA

de Moivre se preguntó en el s. XVII por la distribución binomial $X \sim B(n, p)$ cuando:

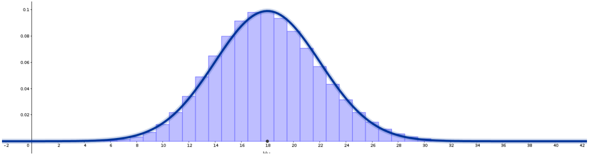
- n grande
- $0 << p << 1$

Descubrió que:

- $P(X = x) \sim 0$ conforme n crece, **pero** $\sum_{x=0}^{n} P(X = x) = 1$.
- Por eso tiene más sentido preguntarse por $P(a \leq X \leq b)$.
- Hay una función que aproxima muy bien los valores de $P(X = x)$:

$$P(X = x) \approx \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

  donde $\mu = np$ y $\sigma = \sqrt{np(1-p)}$ son el valor esperado y la desviación típica de la variable binomial X.



Es la **distribución normal** y su gráfica se llama **campana de Gauss**

Ejemplo: ¿Cálculos binomiales con n grande? por ejemplo, si $X \sim B(1000, 1/3)$, entonces

$$p(300 \leq X \leq 600) =$$

(5380129512302970792342860314398076093260983661754716812535419320384176 34-
3609477230096905514049952288012212015575634780572227086681747192217702384-
1388834078862818305292827189119107103178359318263504456079428051202542871-
0057520719013026163045323479364373120439874930282205964524878195309766641-
5581328389619244466997099160050918442442994709646536946855069475887091250-
1261038176288874223838233563647349000425977778847344543917779777066698319-
3455513109769679618748784337123436134 4) / (5440620656299615789672655389926 5-
2002454906186317756447598732661821740163517162510099085725864447762125065-
7521386306923275518331613582793875989064712995694131800906276536299604436-
2747910656989352855572021299943129264565753729345450125990377491937723230-
0619826389086561483764219916476911859039246195438739185526859491266906292-
2750684766699127147812989317221806327610589473958215472998746982572094405-
7123829647164184009829798469726353318878484190617720755807908358138637977-
58107)

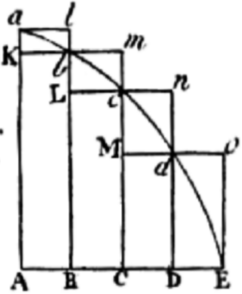
$$= 0.9889$$

Newton, a la vez, trabajaba en calcular el área que hay bajo la curva que define una función.

LEMMA II.

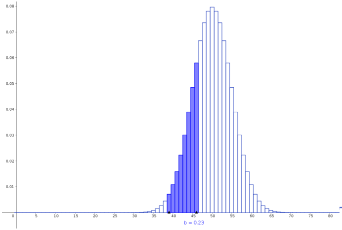
If in any figure AacE (Pl. 1. Fig. 6.) *terminated by the right lines* Aa, AE, *and the curve* acE, *there be inscrib'd any number of parallelograms* Ab, Bc, Cd, *&c. comprehended under equal bases* AB, BC, CD, *&c. and the sides* Bb, Cc, Dd, *&c. parallel to one side* Aa *of the figure; and the parallelograms* aKbl, bLcm, cMdn, *&c. are compleated. Then if the breadth of those parallelograms be suppos'd to be diminished, and their number to be augmented* in infinitum: *I say that the ultimate ratio's which the inscrib'd figure* AKbLcMdD, *the circumscribed figure* AalbmcndoE, *and curvilinear figure* AabcdE, *will have to one another, are ratio's of equality.*

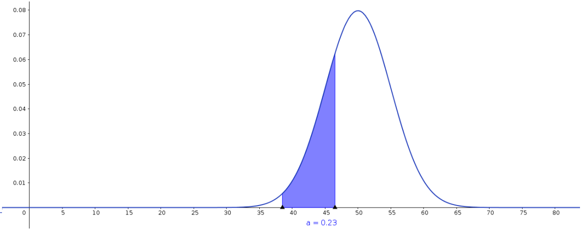
Fig. 6.
p. 42.



Demostró que eso es posible y que se puede hacer como límite de la suma de rectángulos cada vez más pequeños que aproximan dicho área

IDEA: aproximar la probabilidad (suma) binomial por la normal (integral)





Es decir, si $X \sim B(100, 1/2)$ e $Y \sim N(50, 5)$, entonces

$$P(39 \leq X \leq 46) = P(X = 39) + \cdots + P(X = 46) \approx \int_{39}^{46} f_{50,\,5}(x)dx$$

donde $f_{50,\,5}(x)$ es la normal con $\mu = 50$ y $\sigma = 5$.

Una función continua sirve para calcular probabilidades si

Definición:

Una función $f(x)$ es una **función de densidad de probabilidad** si cumple estas condiciones:

- Es no negativa: $f(x) \geq 0$ para todo x
- El área total bajo la gráfica de $f(x)$ es 1: $\displaystyle\int_{-\infty}^{+\infty} f(x)dx = 1$

Distribuciones continuas

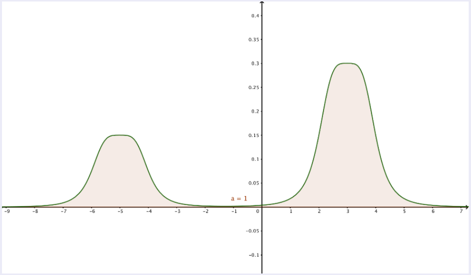
- Si $f(x)$ que cumple las propiedades enteriores se dice que f define una **variable aleatoria continua** X con función de densidad f.
- En tal caso la probabilidad de que el valor de X pertenezca a un intervalo (a, b) se **define así**

$$P(a \leq X \leq b) = \text{área bajo la gráfica de } f = \int_a^b f(x)\,dx.$$

Ejemplo: La función $f(x) = \frac{\sqrt{2}}{9\pi}\left(\frac{6}{((x-3)^4+1)} + \frac{3}{((x+5)^4+1)}\right)$ es una función de densidad continua. Comprueba que su integral de $-\infty$ a ∞ es 1 con Wolfram Alpha:

```
integrate f(x) = (sqrt(2) / (9 pi)) * (6 / ((x - 3)^4 + 1) + 3 / ((x + 5)^4 + 1)) from -oo to oo
```

La gráfica de esta función es:



Usemos *Wolfram Alpha* para calcular tres probabilidades:

- $P(-6 < x < 4) \approx 0.26034$. Usa este enlace
- $P(1 < x < 5) \approx 0.642484$. Usa este enlace
- $P(-2 < x < 0) \approx 0.0043287$. Usa este enlace

Media y varianza de una variable aleatoria continua.

- Para un **variable aleatoria discreta** con valores $x_1, \ldots, x_k$ y probabilidades $p_1, \ldots, p_k$ era:

$$\mu = E(X) = \sum_{i=1}^{k} x_i \cdot p_i$$

$$\sigma^2 = Var(X) = \sum_{i=1}^{k} (x_i - \mu)^2 \cdot p_i$$

- Para una **variable aleatoria continua** con densidad $f(x)$ se tiene:

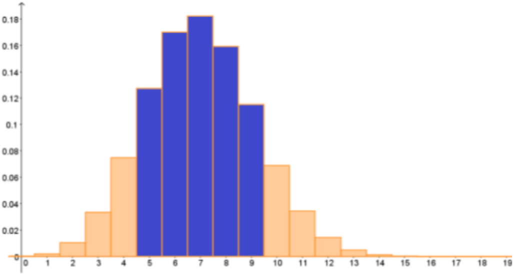
$$\mu = E(X) = \int_{-\infty}^{\infty} x \cdot f(x)\, dx$$

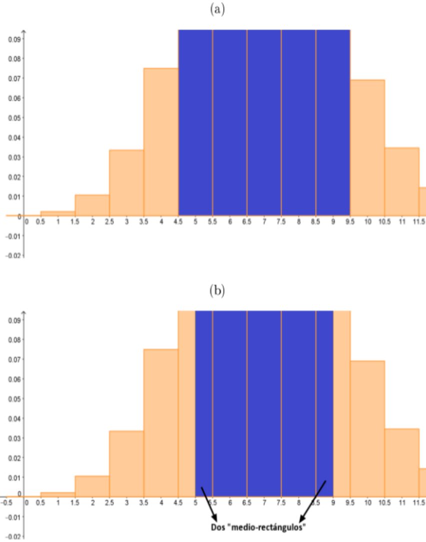
$$\sigma^2 = Var(X) = \int_{-\infty}^{\infty} (x - \mu)^2 \cdot f(x)\, dx$$

- Para pasar de discreto a continuo cambia el sumatorio por una integral y la probabilidad p_i por el *diferencial de probabilidad* $dp = f(x)\, dx$ 0(sección 5.4.2 del libro)
- No hay sorpresa: la media de la curva normal es μ y su varianza es σ^2.
- Calcula la media y la varianza de la función del ejemplo anterior

Aproximar v.a. binomial por normal: se usa n pequeño para facilitar visualización

Suponer X~B(21, 1/3). Calcular $P(5 \leqslant X \leqslant 9)$





Se puede aproximar la binomial por

$$Y \sim N\left(21/3, \sqrt{(21 \cdot 1/3 \cdot 2/3)}\right)$$

De modo que

$$P(5 \leqslant X \leqslant 9) \approx P(4.5 \leqslant Y \leqslant 9.5)$$

Sección 5.6.2 del libro

En definitiva

$$P(5 \leq X \leq 9) \approx \int_{4.5}^{9.5} \frac{1}{\sqrt{21\frac{1}{3}\frac{2}{3}}\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-21\frac{1}{3}}{\sqrt{21\frac{1}{3}\frac{2}{3}}}\right)^2} dx$$

Cálculo aproximado con la normal

```
mu = 21*1/3
sigma = sqrt(21*1/3*2/3)
pnorm(9.5, mu, sigma) - pnorm(4.5, mu, sigma)
```

```
## [1] 0.75284
```

Calculo exacto con la binomial

```
pbinom(9, 21, 1/3) - pbinom(4, 21, 1/3)
```

```
## [1] 0.7540638
```

#####

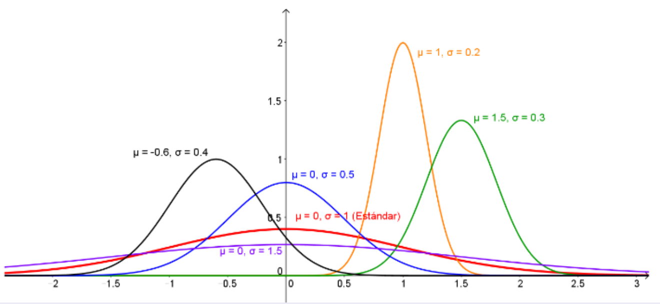
El Teorema Central del Límite (1ª versión).

- Explica como **aproximar una binomial** $X \sim B(n, p)$ **por normal** $X \sim N(\mu, \sigma)$

SOBRE LA NORMAL

Esa función es la **función normal con media μ y desviación típica σ**:

$$f_{\mu,\sigma}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$



La gráfica de $f_{\mu,\sigma}$ se llama **campana de Gauss**

Observa que

- Es simétrica respecto de μ
- Cuanto menor es σ el área bajo la curva está menos dispersa.

Ejemplo: considera $X_1 = N(5, 0.5)$, entonces

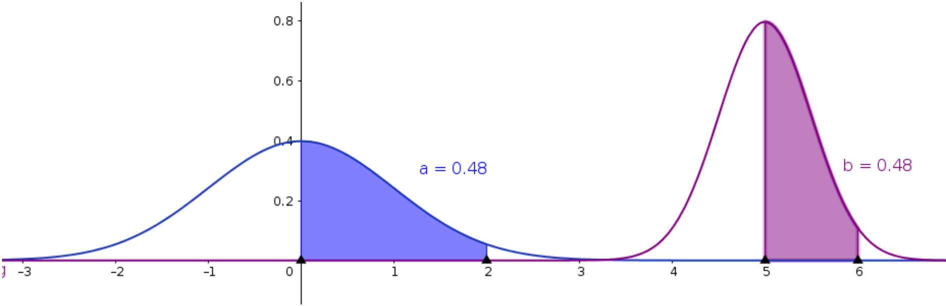
$$P(5 < X < 6) = P(\left(\frac{5-5}{0.5} < \frac{X_1 - 5}{0.5} < \frac{6-5}{0.5}\right) = P(0 < Z < 2)$$

```r
pnorm(6, mean = 5, sd = 0.5)-pnorm(5, mean = 5, sd = 0.5)
```

```
## [1] 0.4772499
```

```r
pnorm(2, mean = 0, sd = 1)  -pnorm(0, mean = 0, sd = 1)
```

```
## [1] 0.4772499
```



Regla 68 - 95 - 99

Si X es una variable normal de tipo $N(\mu, \sigma)$ entonces se cumplen estas aproximaciones:

$$\begin{cases} P(\mu - \sigma < X < \mu + \sigma) \approx 0.683, \\ \\ P(\mu - 2\sigma < X < \mu + 2\sigma) \approx 0.955 \\ \\ P(\mu - 3\sigma < X < \mu + 3\sigma) \approx 0.997 \end{cases}$$

Tipificación

Si $X \sim N(\mu, \sigma)$, entonces la variable que se obtiene mediante la transformación de **tipificación:**

$$Z = \frac{X - \mu}{\sigma}$$

es una variable normal de tipo $N(0, 1)$, la **normal estándar** a la que siempre llamaremos Z.

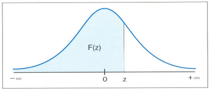
- La tipificación permite reducir cualquier observación de una normal $N(\mu, \sigma)$ a una *escala universal* que nos proporciona la distribución Z:

$$P(a < X < b) = P\left(\frac{a - \mu}{\sigma} < Z < \frac{b - \mu}{\sigma}\right)$$

LA TABLA DE LA NORMAL

Tabla de distribución normal $N(0, 1)$

$F(z) = P(Z \leq z)$



z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2,0	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,4	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,5	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,6	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964
2,7	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974
2,8	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981
2,9	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986
3,0	0,9987	0,9987	0,9987	0,9988	0,9988	0,9989	0,9989	0,9989	0,9990	0,9990
3,1	0,9990	0,9991	0,9991	0,9991	0,9992	0,9992	0,9992	0,9992	0,9993	0,9993
3,2	0,9993	0,9993	0,9994	0,9994	0,9994	0,9994	0,9994	0,9995	0,9995	0,9995
3,3	0,9995	0,9995	0,9995	0,9996	0,9996	0,9996	0,9996	0,9996	0,9996	0,9997
3,4	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9998
3,5	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998
3,6	0,9998	0,9998	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999
3,7	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999
3,8	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999
3,9	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000

284

Función de distribución.

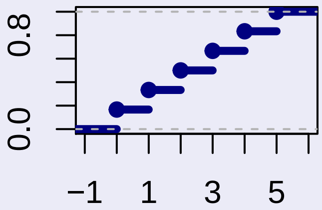
- La función de distribución de una variable aleatoria X (¡discreta o continua!) es:

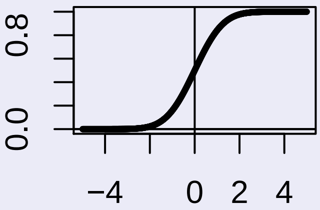
$$F(k) = P(X \leq k)$$

- Esto se traduce de forma diferente para v.a. discreta (izquierda) y continua (derecha) :

$$F(x_k) = \sum_{i=1}^{k} P(X = x_i) \qquad\qquad F(k) = \int_{-\infty}^{k} f(x)\,dx$$

- La gráfica de F también es diferente en el caso discreto (izq) y continuo (dcha)





- Para un intervalo $[a, b]$ se cumple esta **propiedad** que hace que F sea más útil que f:

$$P(a < X < b) = F(b) - F(a)$$

Cálculos con la distribución normal en R.

- La función `pnorm` calcula la función de distribución de $X \sim N(\mu, \sigma)$.

```
P(X < b) = pnorm(b, mean = mu, sd = sigma)
```

- La probabilidad de un intervalo $[a, b]$ en R:

```
pnorm(b, mean = mu, sd = sigma) - pnorm(a, mean = mu, sd = sigma)
```

- El *problema inverso* se resuelve con `qnorm`: dada una probabilidad p nos da b tal que:

$$P(X < b) = p$$

  Es el percentil, que usaremos mucho al hacer inferencia.

- La función `rnorm` crea una muestra de tamaño n de una variable $X \sim N(\mu, \sigma)$:

```
muestra = rnorm(n, mean = mu, sd = sigma)
```

¡Atención! `mean(muestra)` no es `mu` y `sd(muestra)` no es `sigma`: ¿ves por qué? Cuando queramos conseguir eso usaremos la función `mvrnorm` de la librería `MASS` (ver tutoriales).

Ejemplo: llegas a un apeadero desierto, no tienes reloj ni forma de averiguar la hora, y sólo sabes que. . .

La distribución uniforme en el intervalo [a, b]

- Esta distribución representa la idea intuitiva de que *todos los puntos del intervalo son igual de probables*. (¡Pero, cuidado, esa idea es inútil! ¿Ves por qué?)
- Es más acertado pensar que ninguna parte del intervalo es más probable que otra
- La función de densidad apropiada es constante en $[a, b]$ y vale 0 fuera:

$$f(x) = \begin{cases} \dfrac{1}{b-a} & \text{si } a \leq x \leq b \\ 0 & \text{en otro caso} \end{cases}$$

- La media y la deviación típica de la variable uniforme son

$$\mu = \frac{a+b}{2} \qquad \sigma^2 = \frac{(b-a)^2}{12}$$

- En prácticas veremos cómo usar la función `runif` para simulaciones.