Análisis Estadístico de Datos Climáticos

Distribuciones paramétricas de probabilidad

(Wilks, cap. 4)

2015

Variables aleatorias (discretas y continuas)

Una variable aleatoria es aquella que toma un conjunto de valores numéricos asociados a los resultados de la realización de un proceso aleatorio.

Ejemplos:

1) si el experimento es lanzar cuatro veces una moneda al aire y nos interesa el número de caras, la variable aleatoria podrá tomar los valores: 0, 1, 2, 3 y 4.

Es una variable aleatoria discreta (toma un número finito de valores particulares, que suelen ser resultado de un conteo).

Variables aleatorias…

2) otra variable aleatoria: el número de días N que hay que esperar para que en una cierta localidad la precipitación diaria supere los 20 mm. N puede tomar los valores 1, 2, 3,...

En principio N no está acotado (puede tomar los infinitos valores de los números naturales) . Por ser un conjunto numerable, es una variable **discreta**

3) la variable aleatoria es la medida de la temperatura mínima diaria en una cierta localidad. En principio puede tomar un conjunto "infinito" de valores, pero no numerables. Es una variable aleatoria **continua**. (Habitualmente éstas son el resultado de mediciones.)

Distribuciones paramétricas de probabilidad

Una distribución paramétrica de probabilidad es una función matemática (que depende de uno o más parámetros) que permite asignar probabilidades a los valores, o intervalos de valores, que puede tomar una variable aleatoria.

Se usan muy ampliamente como alternativa a las distribuciones *empíricas*, que se construyen a partir de una muestra de datos.

Las razones para usar distribuciones paramétricas son:

- **mayor facilidad de manejo que con los datos originales**
- **posibilidad de suavizar e interpolar**
- **posibilidad de extrapolar**

Una distribución particular puede representar mejor o peor a un conjunto de datos (que son los valores que toma la variable aleatoria).

Existen dos tipos de distribuciones de probabilidad, discretas y continuas, según lo sea la variable aleatoria asociada.

Distribuciones Discretas

Por ejemplo, si se considera la variable aleatoria X = número de caras en dos lanzamientos de una moneda "no cargada";

X	0	1	2
P(X=x)	0.25	0.50	0.25

Hay varios tipos de distribuciones discretas de probabilidad, tales como:

binomial,
geométrica,
binomial negativa,
Poisson,

… y otras.

Distribución Binomial

Fue desarrollada por Jakob Bernoulli (Suiza, 1654-1705); es la principal distribución de probabilidad discreta.

Proviene de experimentos que solo tienen dos posibles resultados, a los que se les puede llamar "éxito" o "fracaso". Los experimentos suelen llamarse "ensayos o pruebas de Bernoulli".
Los datos son resultado de un conteo, por lo que es una distribución discreta.

La distribución binomial consiste en la realización reiterada de varias pruebas y se hacen 2 suposiciones:

1) en cada una la probabilidad de éxito es la misma (p), y

2) las pruebas son independientes entre sí.

Distribución binomial…

Para construir una distribución binomial es necesario conocer 2 parámetros: el número n de pruebas que se repiten y la probabilidad p de que suceda un éxito en cada una de ellas.

Su función de densidad de probabilidad está dada por:

$$f(k;n,p) = \Pr(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

con k = 0, 1, …, n

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

son las combinaciones de n en k
(n elementos tomados de k en k)

n es el número de pruebas
k es el número de éxitos
p es la probabilidad de obtener un éxito
1- p es la probabilidad de obtener un fracaso

Distribución Binomial (Ejemplo)

La distribución binomial se puede usar para calcular la probabilidad de tener exactamente 10 días despejados (sin nubes) en un conjunto *aleatorio* de 30 días. (P (X = 10)).

Es lo mismo que calcular la probabilidad de tener 20 días nublados o algo nubosos.

Definimos la variable "X: Número de días despejados obtenidos en 30 días". En este caso se tiene que x = 10 y n = 30 y suponiendo además que p = 0.5, (o sea que suponemos *en forma arbitraria* que es igualmente probable tener un día despejado que nublado o algo nuboso), resulta:

$f(10;30,0.5)= (30\ 10)\ 0.5^{10}(1-0.5)^{30-10}= 0.028$

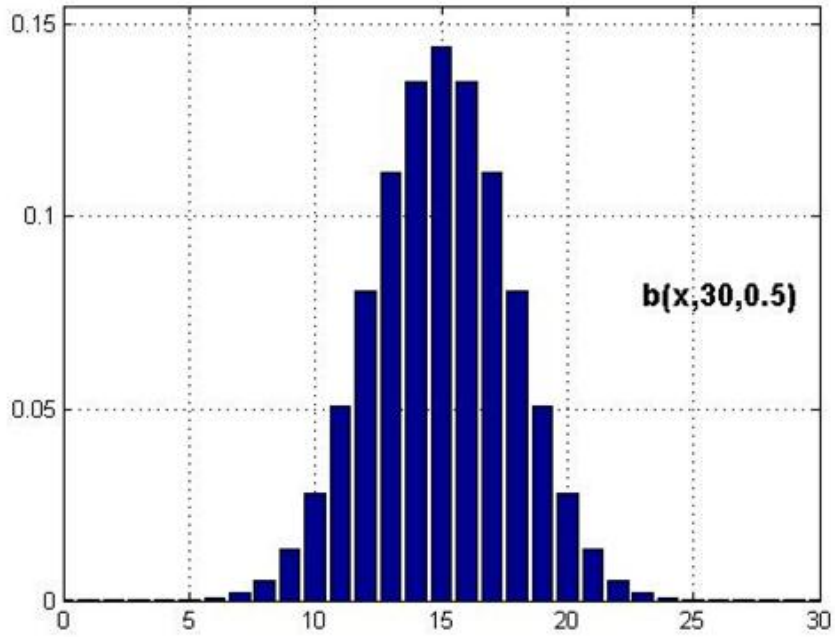
La media de la distribución binomial es np y su varianza es np (1-p)

$$\mathrm{E}[X] = np. \qquad \mathrm{Var}[X] = np(1-p)$$

En este ejemplo: $\mu = 30 * 0.5 = 15$ $\sigma^2 = 15 *(1-0.5)= 7.5$

Matlab: binopdf.m

binopdf(x,n,p)
binopdf(10,30,0.5) = 0.028
% para graficar:
x=[0:30];
bar(x,binopdf(x,30,0.5)), grid



La probabilidad de tener como máximo 10 días despejados, o sea
P(X<=10) = binocdf(10,30,0.5) =0.049

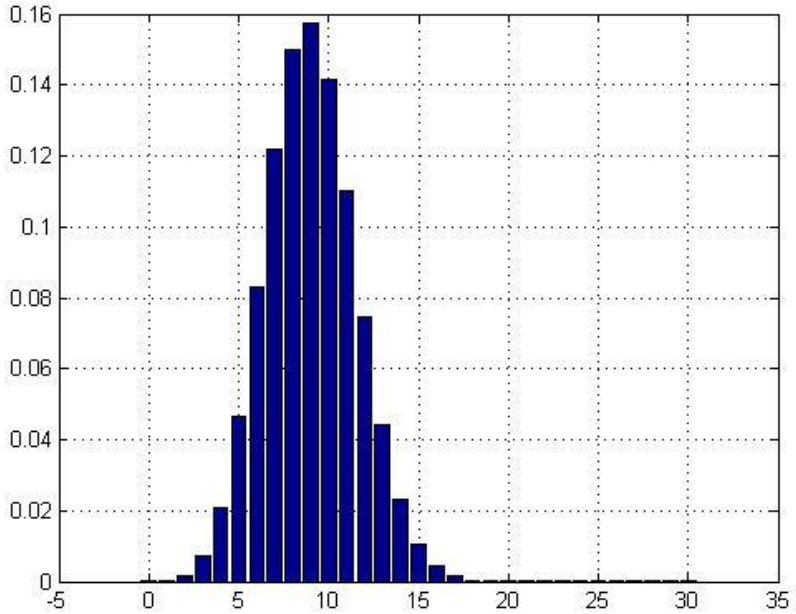
sum(binopdf(x,30,0.5)) **¿cuánto da?**

Si hubiera sido p = 0.3

binopdf(10,30,0.3) = 0.142

x=[0:30];

bar(x,binopdf(x,30,0.3)), grid



La distribución es asimétrica

La probabilidad de tener como máximo 10 días despejados es

P(X<=10) = binocdf(10,30,0.3) =0.730

Distribución geométrica

- **Como en la binomial, hay ensayos repetidos independientes entre sí, con 2 resultados posibles. La probabilidad de éxito es la misma (p) en todos los ensayos.**
- **Pero ahora la variable aleatoria X es *el número de ensayos que hay que realizar hasta que ocurra un éxito.***

$$\Pr(X = k) = (1-p)^{k-1}\, p$$

k = 1,2,….

El parámetro de esta distribución es p.

Matlab: geopdf (k , p)

- **Un ejemplo es la probabilidad de esperar x años hasta que una variable meteorológica supere un cierto valor umbral. Dependiendo del caso, la distribución geométrica podrá o no ajustarla adecuadamente.**

Distribución binomial negativa

Es similar a la geométrica, y con las mismas hipótesis, pero ahora X es *el número de fracasos que deben ocurrir antes que se observe el r-ésimo éxito.* Tiene 2 parámetros, p y r. Se tiene que:

$$P\{X=k\} = \binom{r+k-1}{k} p^r (1-p)^k \qquad \text{k= 0, 1, 2,...}$$

X+r es el "tiempo que hay que esperar" para que ocurran r éxitos.

Matlab: nbinpdf (k, r, p)

Distribución de Poisson

Describe el número de eventos discretos independientes que ocurren en una serie o secuencia (en general en el tiempo, pero puede ser en el espacio).

Se supone que hay independencia en la ocurrencia de eventos en intervalos disjuntos. Los eventos ocurren aleatoriamente, pero con un valor medio constante de ocurrencia.

Tiene un solo parámetro, λ , que representa la ocurrencia media de eventos.

$$P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

x = 0, 1, 2,…

Matlab: poisspdf (x, lambda)

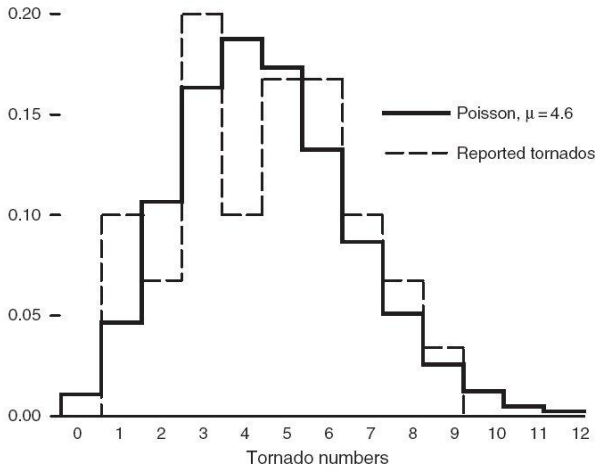
Ej: (Wilks, Sec 4.2.4):

Número de tornados por año en el estado de Nueva York (1959-1988)

TABLE 4.3 Numbers of tornados reported annually in New York state, 1959–1988.

1959	3	1969	7	1979	3
1960	4	1970	4	1980	4
1961	5	1971	5	1981	3
1962	1	1972	6	1982	3
1963	3	1973	6	1983	8
1964	1	1974	6	1984	6
1965	5	1975	3	1985	7
1966	1	1976	7	1986	9
1967	2	1977	5	1987	6
1968	2	1978	8	1988	5

$\lambda \sim 138/30 = 4.6$



"Ajustar la distribución de Poisson a estos datos proporciona una forma razonable de suavizar variaciones irregulares del histograma de datos, lo cual es deseable si las mismas no tienen un significado físico claro."

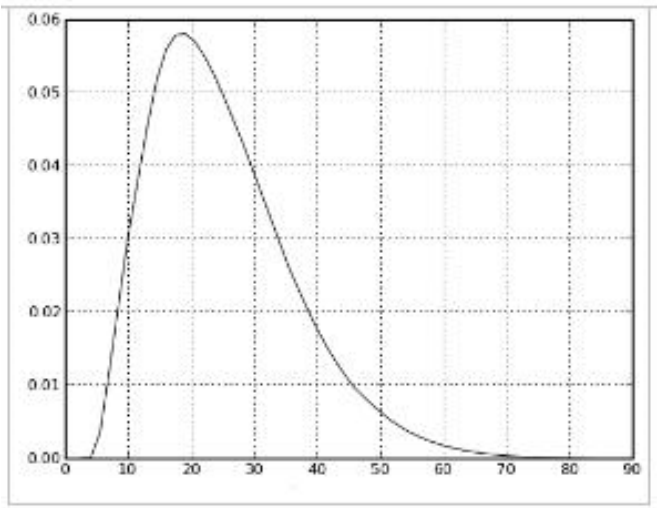
Distribuciones Continuas

Las distribuciones de probabilidad continuas son aquellas en las que la variable aleatoria es continua, o sea que puede asumir un número virtualmente infinito y no numerable de valores, que suelen ser resultado de una medición. Por ejemplo, el valor de la temperatura media del aire en intervalos dados de tiempo. Los valores de las variables aleatorias continuas dependen de la exactitud del instrumento de medición.

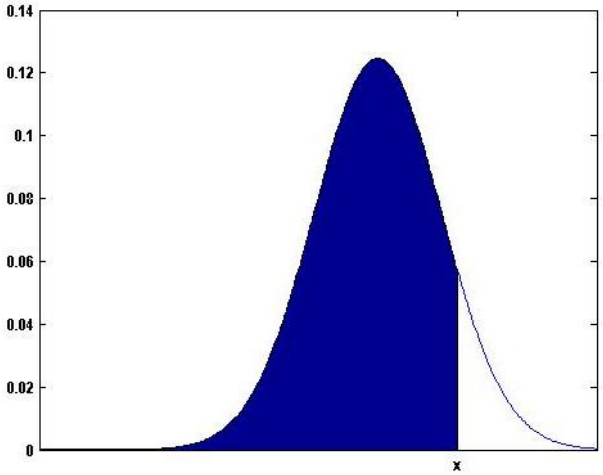
Algunas distribuciones continuas:

Normal o gausiana,
Log-normal
Gamma
t de Student
χ-cuadrado,
y otras.

.

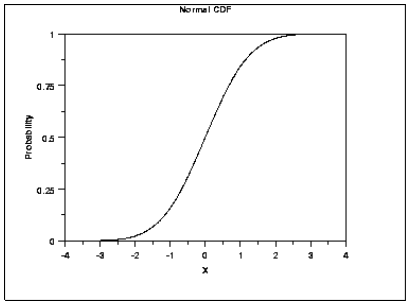


Función de densidad (o PDF) f(x) de una distribución de probabilidad continua



$$P(X \le x) = \int_{-\infty}^{x} f(u)\,du$$

Función de distribución acumulada (o CDF)



$$P(X=c) = 0$$

$$P(a \le X \le b) = \int_{a}^{b} f(x)\,dx$$

$$\int_{-\infty}^{+\infty} f(u)\,du = 1$$

El valor esperado (o valor medio, o media) de la distribución es:

$$\mu = \int x\, f(x)\, dx$$ **(si existe)**

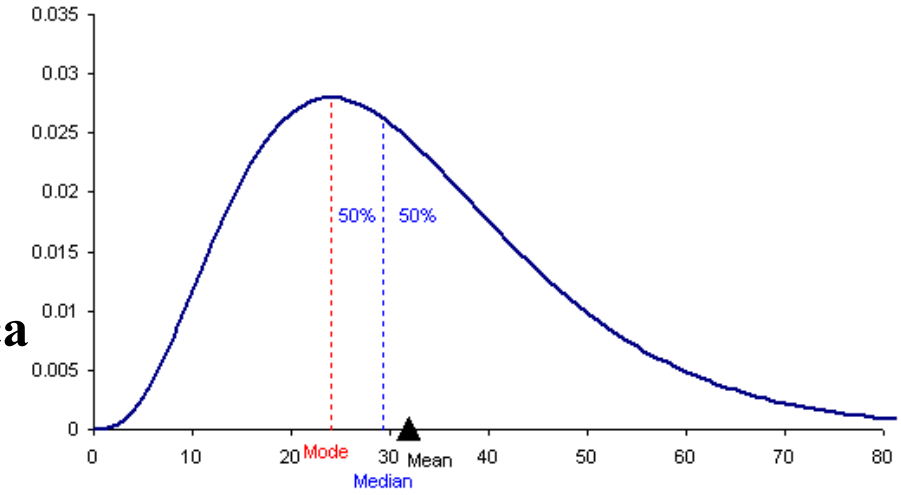
y la varianza es:

$$\mathrm{Var}(X) = \sigma^2 = \int (x-\mu)^2 f(x)\, dx = \int x^2 f(x)\, dx - \mu^2$$ **(si existe)**

La mediana es un valor m tal que

P(X<=m)=P(X>=m)= 0.5

Siempre existe, pero puede no ser única



Distribución gaussiana o normal

La distribución normal fue reconocida por primera vez por el francés Abraham de Moivre (1667-1754) y posteriormente, Carl Friedrich Gauss (1777-1855) formuló la ecuación de la curva; de ahí que también se la conozca, más comúnmente, como la "campana de Gauss".

La distribución de una variable normal está completamente determinada por dos parámetros, su media y su desviación estándar. La función de densidad de la curva normal está definida por la siguiente ecuación:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right], \quad -\infty < x < \infty.$$

Donde: μ es el valor medio

σ es la desviación estándar (σ > 0)

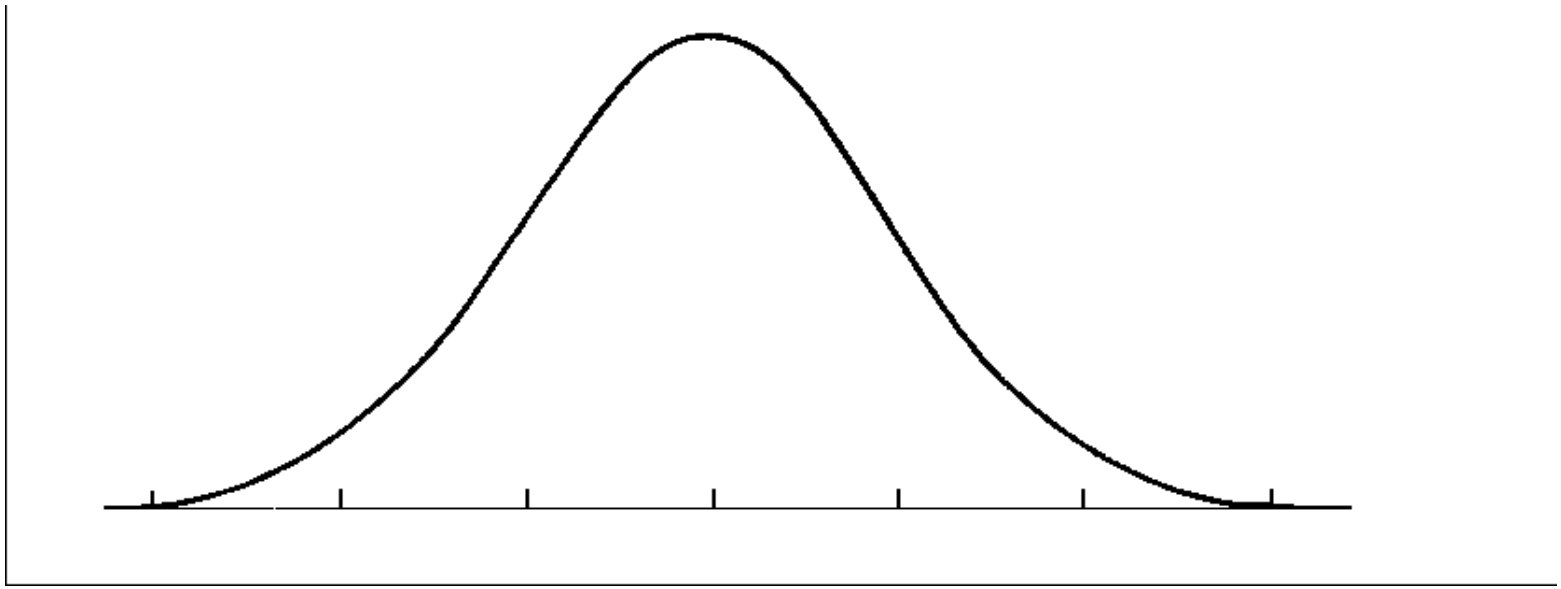
Es la distribución continua de probabilidad más importante de toda la estadística. Como vimos anteriormente, una variable aleatoria continua es la que puede asumir un número infinito de posibles valores que, usualmente resultan de medir alguna magnitud (medidas de longitud, de peso, de tiempo, de temperatura, etc.).

Características de la distribución de probabilidad normal

1. La curva normal tiene forma de campana. La media, la moda y la mediana de la distribución son iguales y se localizan en el centro de la distribución.

2. La distribución de probabilidad normal es simétrica alrededor de su media. Por lo tanto, la mitad del área bajo la curva está antes del punto central y la otra mitad después. El área total bajo la curva es igual a 1.

3. La curva normal tiende a 0 conforme se aleja de la media en ambas direcciones.

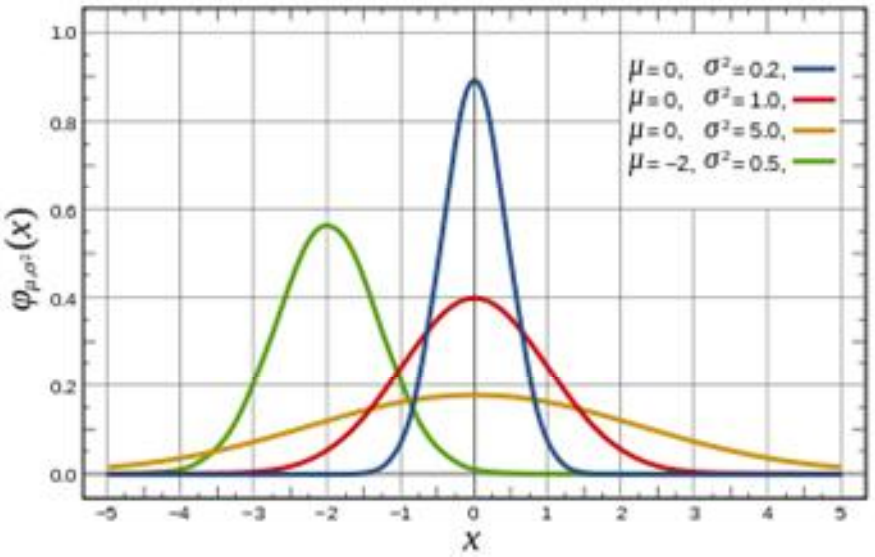


La familia de la distribución de probabilidad normal

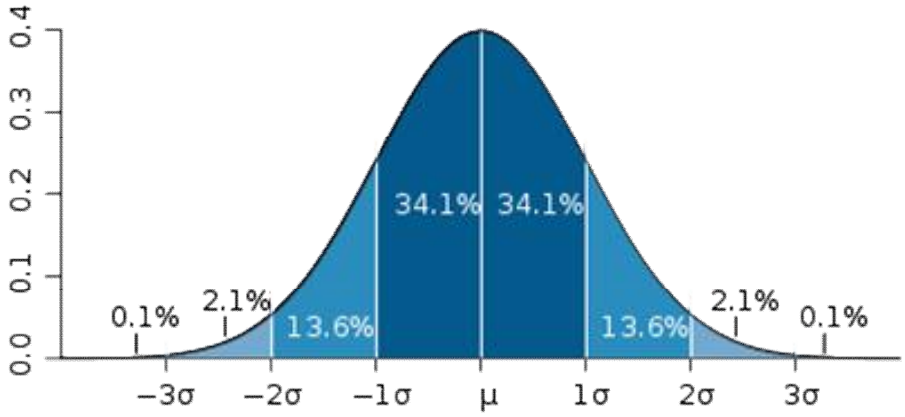
La forma de la campana de Gauss depende de los parámetros μ y σ .
Se suele designar como $N(\mu, \sigma^2)$

La media indica la posición de la campana, de modo que para diferentes valores de μ la gráfica es desplazada a lo largo del eje horizontal.

La desviación estándar determina el grado de achatamiento de la curva. Cuanto mayor sea el valor de σ , más se dispersarán los datos en torno a la media y la curva será más plana. Un valor pequeño de este parámetro indica, por tanto, una gran probabilidad de obtener datos cercanos al valor medio de la distribución.



Distribución normal (cont.)



Probabilidades en un entorno de la media:

- **en el intervalo $[\mu - \sigma, \mu + \sigma]$ se encuentra comprendido, aproximadamente, el 68,26% de la distribución;**
- **en el intervalo $[\mu - 2\sigma, \mu + 2\sigma]$ se encuentra, el 95,44% de la distribución;**
- **en el intervalo $[\mu - 3\sigma, \mu + 3\sigma]$ se encuentra el 99,74% de la distribución.**

El hecho de que prácticamente la totalidad de la distribución se encuentre a tres desviaciones típicas de la media justifica los límites de las tablas empleadas habitualmente en la normal estándar.

Distribución normal estándar

Es la que tiene media igual a cero y desviación estándar igual a uno, y se designa como N(0,1).

Los valores para la función acumulada de la N(0,1) están tabulados, y los valores correspondientes para cualquier otra $N(\mu, \sigma^2)$, se pueden obtener mediante transformaciones matemáticas sencillas: para estandarizar la $N(\mu, \sigma^2)$, se utiliza el cambio de variable:

$$z = (x - \mu) / \sigma$$

(x es de la $N(\mu, \sigma^2)$ y z es de la N(0,1)), y así se puede utilizar la tabla de la N(0,1).

Y a la inversa: $x = \sigma z + \mu$, para pasar de la N(0,1) a la $N(\mu, \sigma^2)$

Matlab: normcdf(X,mu,sigma), normpdf(X,mu,sigma)

Para $N(\mu, \sigma^2)$ se tiene que:

50 % de las observaciones están en el intervalo ($\mu \pm 0{,}68\ \sigma$)
95 % están en el intervalo ($\mu \pm 1{,}96\ \sigma$)
99 % están en el intervalo ($\mu \pm 2{,}58\ \sigma$)
99,9 % están en el intervalo ($\mu \pm 3{,}29\ \sigma$)

Distribución normal o gaussiana (Ejemplo)

Dados los datos de temperaturas medias (º C) para el mes de Enero de la estación de Artigas para 1971-2000, se pide *estimar* la probabilidad de que la temperatura media del mes de Enero sea inferior a 26 ° C, suponiendo que la distribución de las temperaturas se puede aproximar razonablemente por una $N(\mu,\sigma^2)$;

1971	24.2	**1986**	26.5
1972	24.8	**1987**	25.2
1973	25.0	**1988**	24.9
1974	25.2	**1989**	27.0
1975	24.7	**1990**	26.1
1976	25.3	**1991**	24.6
1977	24.9	**1992**	24.7
1978	24.9	**1993**	25.7
1979	26.1	**1994**	25.2
1980	25.8	**1995**	26.0
1981	24.8	**1996**	25.6
1982	24.6	**1997**	27.2
1983	26.1	**1998**	24.0
1984	25.6	**1999**	25.8
1985	26.0	**2000**	26.7

Comenzamos *estimando* ambos parámetros (μ y σ)

Media muestral = 25,4 °C
Desviación típica muestral = 0.8 °C

Para la temperatura de 26 ºC, el valor de la variable estandarizadada será :
([26-25,4]/0.80) = 0,75.

En la tabla de la N(0,1) para un valor de z = 0,75, tenemos que la probabilidad de obtener un valor inferior a z será 0,77.

O bien: normcdf(26, 25.4, 0.8) = 0.7734

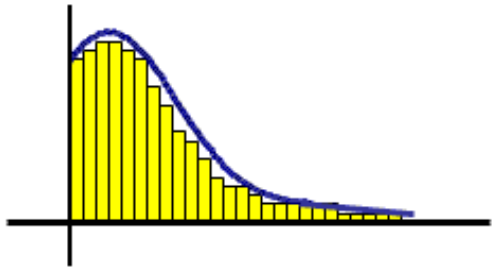
Luego, en estas hipótesis, se espera que el 77 % de los años la temperatura en enero en Artigas será inferior a 26 ºC

Normal o gausiana

Tabla: Valor de la variable tipificada. El valor correspondiente a la fila y la columna nos da la probabilidad de obtener una valor inferior a Z.

	0	**0,01**	**0,02**	**0,03**	**0,04**	**0,05**	**0,06**	**0,07**	**0,08**	**0,09**
0	0,5	0,504	0,508	0,512	0,516	0,5199	0,5199	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5596	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,591	0,5948	0,5987	0,5987	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6368	0,6443	0,648	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,67	0,6736	0,6736	0,6808	0,6844	0,6879
0,5	0,6915	0,695	0,6985	0,7019	0,7054	0,7088	0,7088	0,7157	0,719	0,7224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7422	0,7486	0,7517	0,7549
0,7	0,758	0,7611	0,7642	0,7673	0,7704	0,7734	0,7734	0,7794	0,7823	0,7852
0,8	0,7881	0,791	0,7939	0,7967	0,7995	0,8023	0,8023	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8289	0,834	0,8365	0,8389
1	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8531	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8749	0,879	0,881	0,883
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8944	0,898	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9115	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9265	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,937	0,9382	0,9394	0,9394	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9505	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9599	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9678	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9744	0,9756	0,9761	0,9767
2	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9798	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,983	0,9834	0,9838	0,9842	0,9842	0,985	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9878	0,9884	0,9887	0,989
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9906	0,9911	0,9913	0,9916
2,4	0,9918	0,992	0,9922	0,9925	0,9927	0,9929	0,9929	0,9932	0,9934	0,9936
2,5	0,9938	0,994	0,9941	0,9943	0,9945	0,9946	0,9946	0,9949	0,9951	0,9952
2,6	0,9953	0,9955	0,9956	0,9957	0,9959	0,996	0,996	0,9962	0,9963	0,9964
2,7	0,9965	0,9966	0,9967	0,9968	0,9969	0,997	0,997	0,9972	0,9973	0,9974
2,8	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9978	0,9979	0,998	0,9981
2,9	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9984	0,9985	0,9986	0,9986
3	0,9987	0,9987	0,9987	0,9988	0,9988	0,9989	0,9989	0,9989	0,999	0,999

Distribuciones sesgadas positivamente



La distribuciones estadísticas de varias variables atmosféricas suelen ser asimétricas, y con sesgo positivo.

Es muy común que el sesgo ocurra cuando existe un límite físico sobre la izquierda que está relativamente cerca del rango de datos. Los ejemplos mas comunes son la precipitación, la velocidad del viento, la humedad relativa, los cuales esta físicamente restringidos a ser no-negativos.

A pesar de que matemáticamente es posible ajustar una distribución gausiana en dichas situaciones, los resultados no son útiles.

A veces se realizan transformaciones de los datos para obtener una distribución simétrica.

También es posible intentar utilizar otras distribuciones para representar los datos.

Distribución Lognormal

Una de las transformaciones que se suele usar para obtener una distribución más simétrica (en el caso de datos positivos y sesgo positivo), es:

$y = \ln(x)$.

(No necesariamente es siempre la mejor transformación, pero se usa muy habitualmente (ver Wilks, p. 43-47)).

Si X es una variable aleatoria que toma valores positivos, tal que la variable $y=\ln(x)$ es $N(\mu_y, \sigma_y^2)$, se dice entonces que X tiene distribución lognormal.

La densidad para la variable original x es:

$$f(x) = \frac{1}{x\sigma_y\sqrt{2\pi}} \exp\left[-\frac{(\ln x - \mu_y)^2}{2\sigma_y^2}\right], \quad x > 0$$

con media, desviación estándar y mediana:

$$\mu_x = \exp\left[\mu_y + \frac{\sigma_y^2}{2}\right] \qquad \sigma_x^2 = (\exp[\sigma_y^2] - 1)\exp[2\mu_y + \sigma_y^2]. \qquad \text{m} = \exp(\mu_y)$$

La variable $z = \dfrac{\ln(x) - \mu_y}{\sigma_y}$ **es gaussiana estándar (media 0 y varianza 1).**

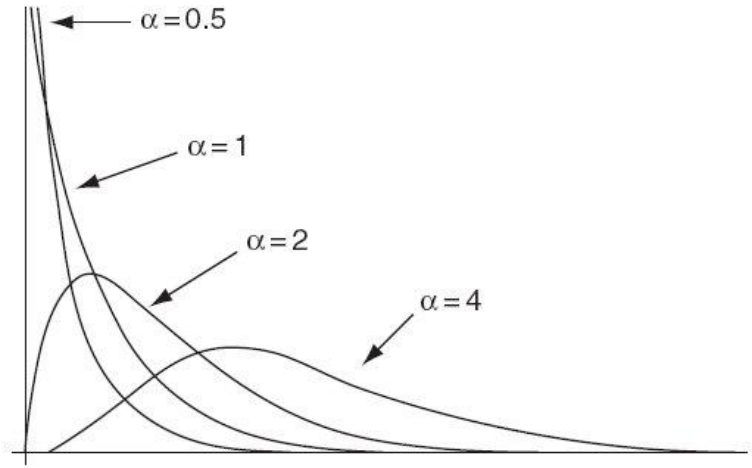
Distribución Gamma

Una elección común usada para representar en particular datos de precipitación, es la distribución gamma, que esta definida por la función de densidad (PDF):

$$f(x) = \frac{(x/\beta)^{\alpha-1} \exp(-x/\beta)}{\beta\Gamma(\alpha)}, \quad x, \alpha, \beta > 0.$$

siendo $$\Gamma(k) = \int_0^\infty t^{k-1} e^{-t}\, dt.$$

α es el parámetro de forma; β el parámetro de escala



Para **α < 1** la distribución esta fuertemente sesgada a la derecha, con $f(x) \to \infty$ si $x \to 0$.

Para **α = 1** la función corta el eje vertical en 1/β para x = 0 (Este caso especial de la distribución gamma es llamada la **distribución exponencial**).

Para **α >1** la distribución gamma comienza en el origen, f(0)=0.

Progresivamente mayores valores de α resultan en menos sesgo, y un desplazamiento de la probabilidad de densidad a la derecha. **Para valores de α muy grandes** (p.ej., mayores que 50 a 100), la distribución gamma se aproxima a la distribución normal en su forma.

El parámetro α es siempre adimensional.

El rol del parámetro de escala β es alargar o estrechar la función gamma a la derecha o a la izquierda.

Distribución chi-cuadrado (χ^2)

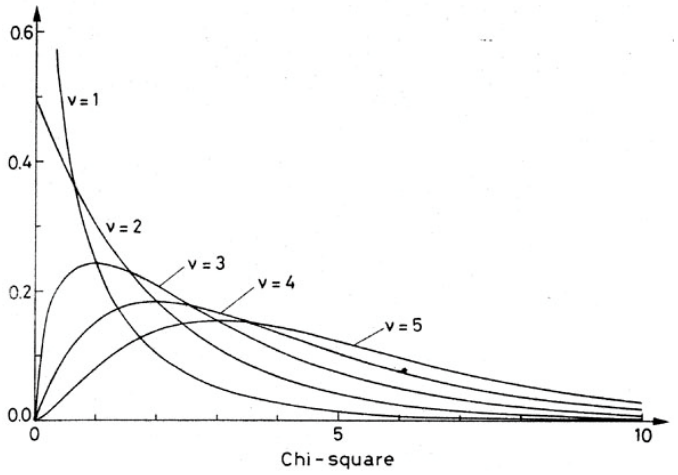
Es un caso particular de la dist. gamma:

$\beta = 2$ y $\alpha = \nu/2$, siendo ν un número natural.

La PDF de la $\chi_\nu^{\,2}$ es: $$f(x) = \frac{x^{(\nu/2-1)} \exp\left(-\frac{x}{2}\right)}{2^{\nu/2}\Gamma\left(\frac{\nu}{2}\right)}, \quad x > 0.$$

La distribución chi-cuadrado surge en forma independiente de la gamma, como la distribución de la suma de los cuadrados de ν variables aleatorias independientes gaussianas estándar, y es muy utilizada en pruebas de hipótesis.

ν es el *número de grados de libertad* de la distribución $\chi_\nu^{\,2}$.



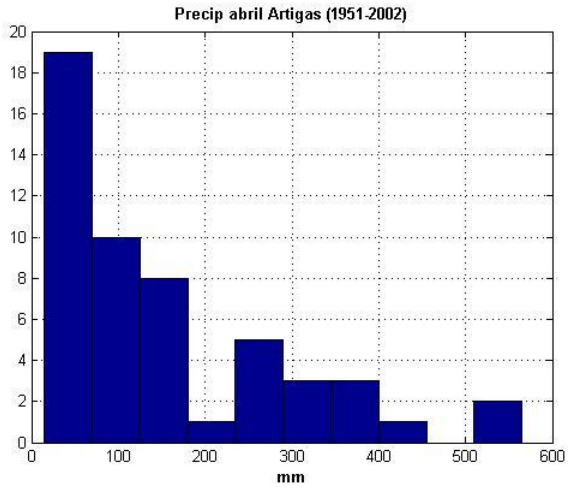
$E(\chi_\nu^2) = \nu$

$Var(\chi_\nu^2) = 2\,\nu$

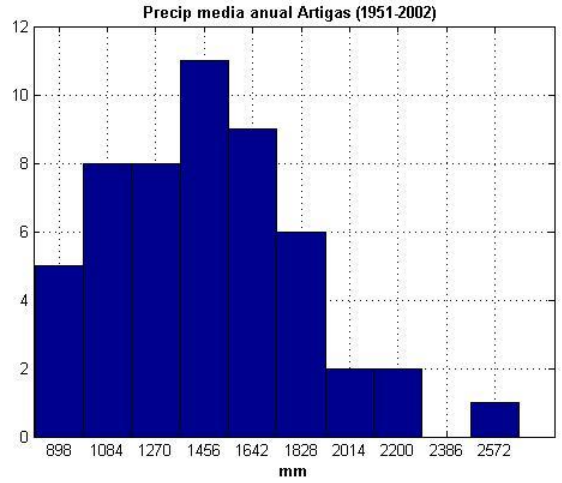
Aplicación del teorema del límite central

- **Teorema del límite central (una versión): si se tiene una serie infinita de variables aleatorias, independientes e idénticamente distribuidas (iid), con media y varianza finitas (μ y σ^2), entonces la variable aleatoria igual al promedio (o la suma) de n de ellas es *asintóticamente* gausiana, aunque la distribución original no lo fuera.**
- **Se aplica a variables climáticas (temperatura, precipitación, etc). La cantidad de casos necesarios para que se note esa tendencia depende de la variable climática. (Ver Wilks, p. 88.)**

Precip **abril** Artigas 1951-2002



Precip **acumulada anual** Artigas 1951-2002



Distribuciones de algunas variables climatológicas

Dependiendo de la localizacion geográfica, se puede decir, como orientación general, que:

- La **temperatura media** horaria suele tener una distribución normal en climas tropicales y una distribución algo mas asimétrica en latitudes medias. Las **temperaturas medias diarias** muestran una distribución casi normal. En cambio las **temperaturas máximas diarias** presentan una distribución asimétrica positiva principalmente en verano. Por el contrario las **temperaturas mínimas diarias** presentan un distribución asimétrica negativa sobre todo en invierno.
- La **humedad atmosférica** puede estar representado por varios índices (p. ej. humedad relativa), ninguno de los cuales se comporta como normal.
- La **precipitación diaria** no tiene una distribución normal. Usualmente se emplea una distribución de extremos (Gamma, etc.) para ajustar las distribuciones de lluvias diarias. La **precipitaciones acumuladas mensuales** en general no tienen una distribución normal en nuestro país.
- Las estadísticas de fenómenos discontinuos como los **días con lluvia, con granizo, niebla, rocío, tormenta,** etc., obedecen a distribuciones discretas como la binomial.

Estimación de parámetros

En general, no conocemos la PDF de las variables observadas . Podemos conocer o suponer la familia (normal, binomial, etc.) a la que pertenecen, pero no los valores de los parámetros de la distribución. Para calcularlos necesitaríamos tener todos los posibles valores de la variable, lo que en general no es posible. La inferencia estadística trata de cómo obtener información (*inferir*) sobre los parámetros a partir de subconjuntos de valores (*muestras*) de la variable.

***Estadístico*: variable aleatoria que sólo depende de la muestra aleatoria elegida ($x_1, x_2, \ldots, x_n$) para calcularla (es decir que no dependen de magnitudes desconocidas, como los parámetros que se quieren estimar)**

Ej: la media muestral $\bar{x} = \dfrac{\sum_{i=1}^{n} x_i}{n}$

***Estimador*: Es un estadístico que se usa para estimar un parámetro.**
El valor que tome el estimador dependerá de la muestra aleatoria, por lo que el estimador tendrá una distribución de probabilidad, que será su distribución muestral.

Es deseable que un estimador tenga algunas propiedades (que sea insesgado, de variancia mínima, etc), cosa que no siempre se puede lograr.

Estimación de parámetros (cont.)

Algunos métodos para estimar parámetros son:

- **Método de los momentos**
- **Método de la máxima verosimilitud**
- **Método de los mínimos cuadrados**

Ejemplo del método de los momentos

La media y la varianza son momentos de primer y segundo orden respectivamente

Ejemplo de aplicación a la distribución gamma

Si la variable aleatoria X sigue una distribución gamma de parámetros α y β, su valor esperado y su varianza valen:

$$E(X) = \alpha\beta \qquad Var(X) = \alpha\beta^2$$

Por tanto podemos expresar α y β como

$$\alpha = \frac{E(X)^2}{Var(X)} \qquad \beta = \frac{Var(X)}{E(X)}$$

donde E(X) y Var (X) se estiman a partir de la muestra (por medio de $\overline{X}$ y s^2)

Estimación de parámetros por el método de máxima verosimilitud

La idea es determinar, para una muestra de datos dada y para una distribución elegida adecuadamente, el conjunto de valores más probables de los parámetros, dados los datos que se observaron.
Para eso se define la función de verosimilitud, y se busca determinar los valores de los parámetros que la hacen máxima

La función de verosimilitud de los parámetros, para una sola observación *x*, es la PDF, pero debe interpretarse considerando a *x* como dato, y a los parámetros como variables o incógnitas.
Ej: para la distribución gaussiana:

$$\frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$

La función de verosimilitud para n observaciones *independientes* (x_i, i=1, 2, …, n) es el producto de las n funciones individuales:

$$\Lambda(\mu, \sigma) = \sigma^{-n}(\sqrt{2\pi})^{-n} \prod_{i=1}^{n} \exp\left[-\frac{(x_i-\mu)^2}{2\sigma^2}\right].$$

Tomando logaritmos y planteando las derivadas parciales respecto a los parámetros μ y σ, se obtiene:

$$\frac{\partial L(\mu,\sigma)}{\partial \mu} = \frac{1}{\sigma^2}\left[\sum_{i=1}^{n} x_i - n\mu\right]$$

$$\frac{\partial L(\mu,\sigma)}{\partial \sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(x_i-\mu)^2.$$

Anulando las derivadas, se obtiene:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\hat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\hat{\mu})^2}.$$

Para la distribución gausiana, es posible obtener una expresión analítica de los estimadores de máxima verosimilitud. Esto no es habitual para otras distribuciones, y se hace necesario resolver las ecuaciones iterativamente.

En Matlab, hay rutinas que estiman parámetros por máxima verosimilitud (MLE) para muchas distribuciones, dando además intervalos de confianza de los estimadores.
normfit, gamfit, binofit, etc, etc

Estimación de la varianza de la media en presencia de dependencia serial

La estimación de la varianza de la distribución muestral de la media de n observaciones independientes es:

$$\hat{\mathrm{Var}}[\bar{x}] = s^2/n$$

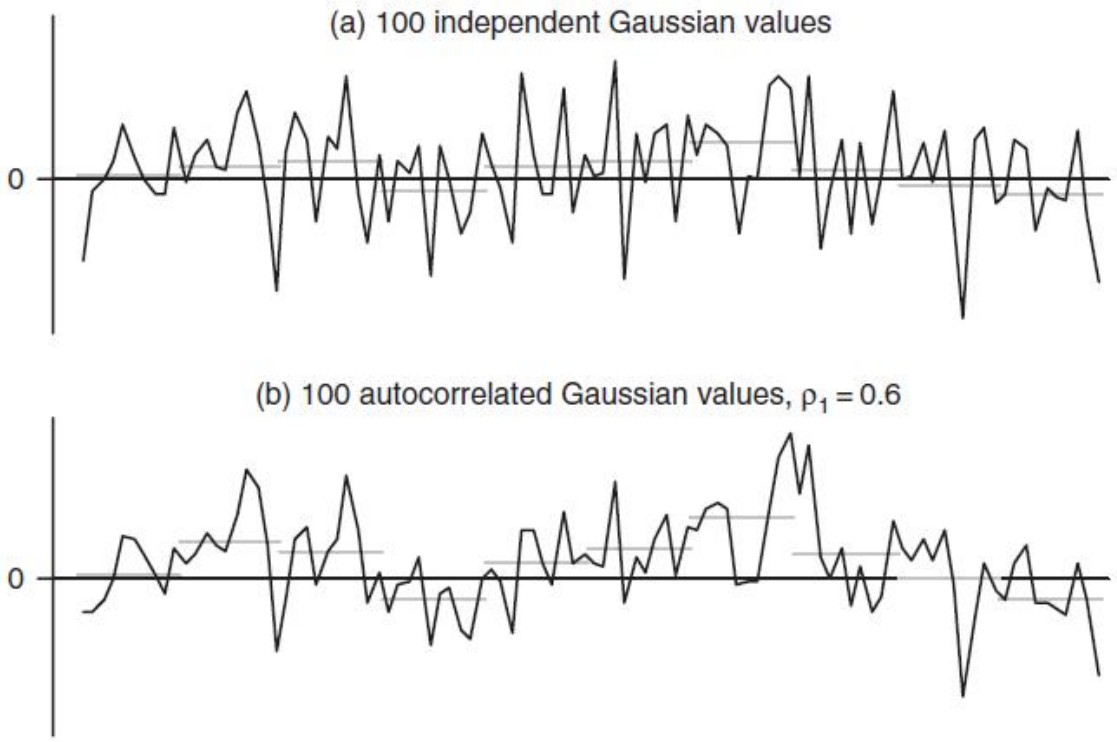
siendo

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

Los datos atmosféricos no suelen cumplir la condición de independencia debido a la presencia de persistencia (ej. temperatura media diaria).

En ese caso la fórmula anterior no es válida

Wilks, p. 144



Se aprecia que los promedios de n = 10 valores para la serie con autocorrelación $\rho_1 = 0.6$ están más dispersos alrededor del valor medio que para la otra serie con $\rho_1 = 0$.

Lo que se hace es, a partir de ciertas hipótesis definir un tamaño de muestra efectivo:

$$n' \cong n \frac{1 - \rho_1}{1 + \rho_1}.$$