Curso

Análisis Estadístico de Datos Climáticos

Distribuciones de Probabilidad

Mario Bidegain (FC) – Alvaro Diaz (FI)

Universidad de la República

Montevideo, Uruguay

2011

DISTRIBUCIONES DE PROBABILIDAD

¿Qué es una distribución de probabilidad?

Una variable aleatoria es aquella que toma un conjunto de valores numéricos asociados a los resultados de nuestra búsqueda que produce un proceso aleatorio.

Por ejemplo si el experimento es lanzar cuatro veces una moneda al aire y nuestro búsqueda es el número de caras, la variable aleatoria podrá tomar valores de 0, 1, 2, 3 y 4 caras.

Una distribución de probabilidad es una lista del total de valores que puede tomar una variable aleatoria con una probabilidad asociada.

Existen dos tipos de distribuciones de probabilidad, las distribuciones de probabilidad discretas y las distribuciones de probabilidad continuas.

Distribuciones Discretas

Las distribuciones de probabilidad discretas son aquellas en las que la variable aleatoria solo puede asumir ciertos valores claramente separados, y son resultado de un conteo.

Por ejemplo, el número de caras en dos lanzamientos de una moneda.

X	0	1	2
P(X)	0.25	0.50	0.25

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

Distribución Binomial,
Distribución Poisson,
Distribución Hipergeométrica.

Distribución Binomial

La distribución binomial fue desarrollada por Jakob Bernoulli (Suiza, 1654-1705), es la principal distribución de probabilidad discreta.

La binomial proviene de experimentos que solo tienen dos posibles resultados, a los que se les puede nombrar como éxito o fracaso. Los datos son resultado de un conteo, razón por la cual se clasifica como distribución discreta.

La binomial consiste de varias pruebas y en cada una la probabilidad de éxito es la misma, por lo que son independientes.

Para construir una distribución binomial es necesario conocer el número de pruebas que se repiten y la probabilidad de que suceda un éxito en cada una de ellas. Su función de densidad de probabilidad está dada por:

$$b(x;n,\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x} \qquad x = 0, 1, 2, ..., n$$

$\binom{n}{x} = \frac{n!}{x!(n-x)!}$ **son las combinaciones de n en x (elementos tomados de x en x)**

n es el número de pruebas
x es el número de éxitos
Θ es la probabilidad de obtener un éxito
1- Θ es la probabilidad de obtener un fracaso

Distribución Binomial (Ejemplo)

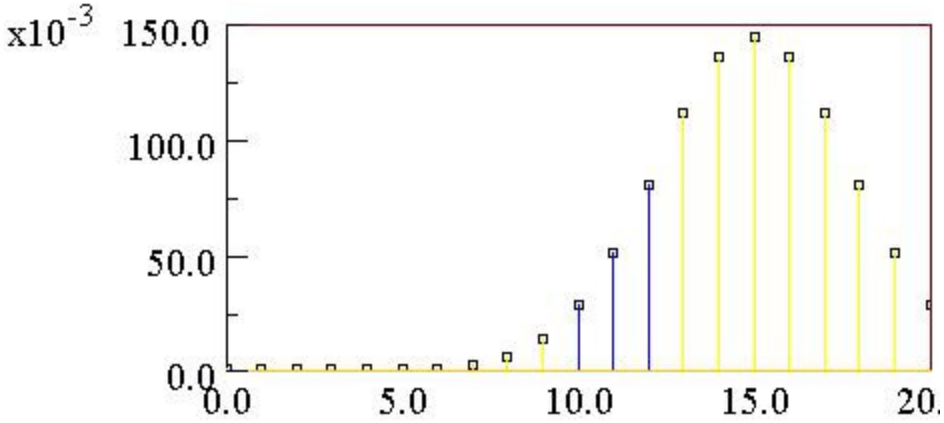
Por ejemplo, la distribución binomial se puede usar para calcular la probabilidad de tener 5 días despejados (sin nubes) en 30 días de un mes.
En realidad sólo se calcula la probabilidad de tener 5 días despejados, pero como es lógico si en 30 días de un mes tenemos 5 días despejados el resto deben ser días nublados o algo nubosos, 25 en este caso.

Por lo tanto debemos definir la variable "X: Número de días despejados obtenidos en 30 días".
En este caso se tiene que x = 5 y n = 30, **Θ** = 0.5 resulta:

b(5:30:0.5)= (30) $0.5^5(1-0.5)^{30-5}$= 0.0001327

Su media y su varianza son: $\mu = n\theta \qquad \sigma^2 = n\theta(1-\theta)$

μ= 30 . 0.5 = 15
σ = 15(1-0.5)= 7



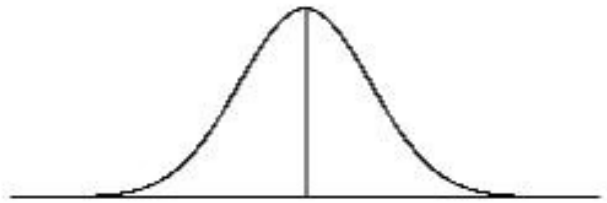
Distribuciones Continuas

Las distribuciones de probabilidad continuas son aquellas en las que la variable aleatoria puede asumir un número infinito de valores, que son resultado de una medición. Por ejemplo, el valor de la temperatura media del aire en intervalos dados de tiempo. Por supuesto que las variables aleatorias continuas dependen de la exactitud del instrumento de medición en este caso del termómetro.

También existen varios tipos de distribuciones continuas de probabilidad, las mas usadas son:

Distribución Normal o gausiana,
Distribución t de Student,
Distribución χ-cuadrado,
Distribución Gamma

Las distribuciones continuas son imposibles de tabular y por lo tanto se representan con curvas.



Curva de una distribución de probabilidad continua

Distribuciones continuas

Normal o gausiana

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:

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

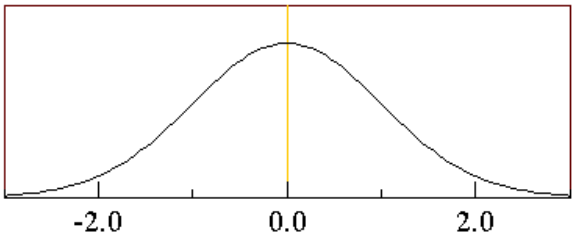
Donde μ es el valor medio
σ es la desviación estándar

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 dentro de un rango específico. Estos valores usualmente resultan de medir algo (medidas de longitud, de peso, de tiempo, de temperatura, etc.)

Características de la distribución de probabilidad normal

La distribución de probabilidad normal y su curva tiene las siguientes características:

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 se aproxima de manera asintótica al eje horizontal conforme se aleja de la media en cualquier dirección. Esto significa que la curva se acerca al eje horizontal conforme se aleja de la media, pero nunca lo llega a tocar.**



La familia de la distribución de probabilidad normal

La forma de la campana de Gauss depende de los parámetros μ y σ . La media indica la posición de la campana, de modo que para diferentes valores de la gráfica es desplazada a lo largo del eje horizontal.

Por otra parte, 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 estándar

Para facilitar los cálculos se decidió tabular la normal para diferentes probabilidades con variables que siguen la distribución normal. Pero, puesto que sería imposible tener una tabla para cada posible distribución normal, se elaboró la tabla de la *distribución normal estándar*, que es la distribución con media igual a cero y desviación estándar igual a uno.

De esta manera solo se tiene que transformar o estandarizar una distribución normal específica, se revisa la tabla, y se conoce la probabilidad. Para estandarizar los valores de una variable, se utiliza la siguiente fórmula:

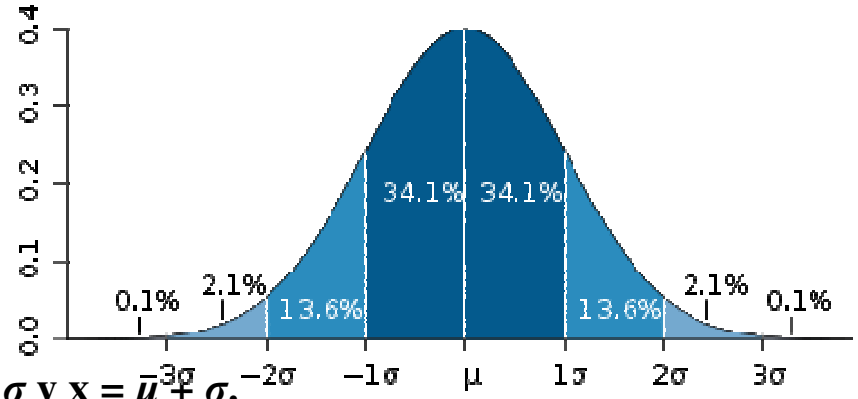
$$z = (x - \mu) / \sigma$$

Con esta fórmula podemos transformar cualquier distribución normal a la distribución normal estándar

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

Propiedades de la distribución normal

Algunas propiedades de la distribución normal son:



1) Es simétrica respecto de su media, μ;

2) La moda y la mediana son ambas iguales a la media, μ;

3) Los puntos de inflexión de la curva se dan para $x = \mu - \sigma$ y $x = \mu + \sigma$.

4) Las probabilidades en un entorno de la media son:

4.1 en el intervalo $[\mu - \sigma, \mu + \sigma]$ se encuentra comprendida, aproximadamente, el 68,26% de la distribución;

4.2 en el intervalo $[\mu - 2\sigma, \mu + 2\sigma]$ se encuentra, aproximadamente, el 95,44% de la distribución;

4.3 por su parte, en el intervalo $[\mu - 3\sigma, \mu + 3\sigma]$ se encuentra comprendida, aproximadamente, el 99,74% de la distribución.

Estas propiedades son de gran utilidad para el establecimiento de intervalos de confianza. Por otra parte, 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.

Normal o gausiana (Ejemplo)

Dados los datos de temperaturas medias (º C) para el mes de Enero de la Estación Meteorológica de Artigas. Se pide determinar la probabilidad de que la temperatura media del mes de Enero sea inferior a 26 ° C.

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

Número de datos: n = 30
Media = 25,4 °C
Desviación típica = 0.8 °C
Para la temperatura de 26 ºC, la variable tipificada será : ([26-25,4]/0.80) = 0,75.
En las tablas para un valor de z = 0,75, tenemos que la probabilidad de obtener una valor inferior a Z será 0,68.
Luego el 68 % de los años la temperatura 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.

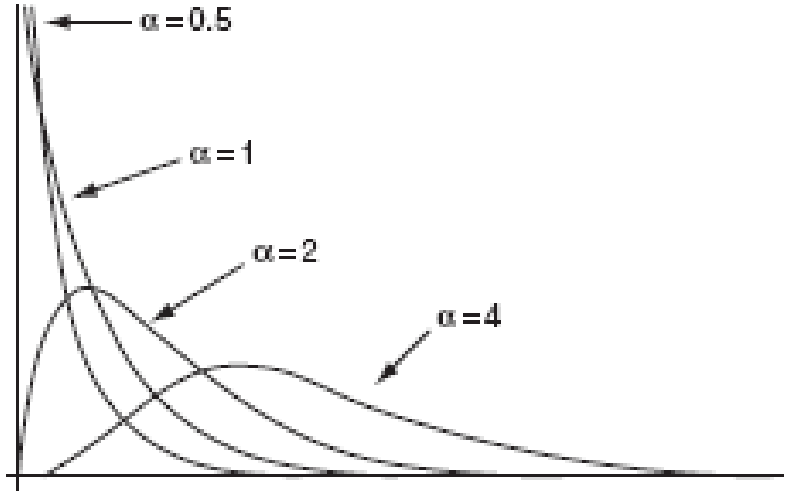
	0	**0,01**	**0,02**	**0,03**	**0,04**	**0,05**	**0,06**	**0,07**	**0,08**	**0,09**
0	0,5	0,504	0,508	0,512	0,516	0,5199	0,5199	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5596	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,591	0,5948	0,5987	0,5987	0,6064	0,6103	0,6141
0,3	0,6179	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 típicas de los variables climatológicas

- 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** tienen en general una distribución normal en nuestro País.
- La **velocidad del viento horaria y media diaria** no se ajusta a una distribución normal, nuevamente se emplean distribuciones de extremos (Gamma, Pearson, Weibull, etc.) para ajustar las distribuciones de velocidades de viento.
- Las estadísticas de fenómenos discontinuos como los **días con lluvia, con granizo, niebla, rocío, tormenta,** etc., obedecen a distribuciones discontinuas como la binomial.

Distribución Gamma

La distribuciones estadísticas de varias variables atmosféricas son sin lugar a dudas asimétricas, y sesgadas a la derecha. Es muy común que el sesgo ocurre 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 están físicamente restringidas a ser no-negativas. A pesar de que matemáticamente es posible ajustar una distribución gausiana en dichas situaciones, los resultados no son útiles.



Gamma distribution density functions for four values of the shape parameter, α.

Distribución Gamma (Cont.)

Existe una gran variedad de distribuciones continuas que están limitadas a la derecha por cero y están positivamente sesgadas. Una elección común usada para representar los datos de precipitación, es la distribución gamma. La distribución gamma esta definida por la PDF

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

Los dos parámetros de la distribución son α el parámetro de forma; y β el parámetro de escala. La cantidad Γ(α) es la función gamma.

Para α < 1 la distribución esta fuertemente sesgada a la derecha, con f(x)→ ∞ as x→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 (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 Gamma (Cont.)

Los dos parámetros de la distribución son α el parámetro de forma; y β el parámetro de escala.

Estos parámetros se pueden estimar mediante la aproximación de Thom (1958)

$$\gamma = \frac{1}{4A}\left[1 + \sqrt{\left(1 + \frac{4A}{3}\right)}\right]$$

$$A = \ln \bar{x} - \frac{\sum \ln x}{n}$$

$$\beta = \frac{\bar{x}}{\gamma}$$

Distribución Gamma (Ejemplo)

La distribución gamma se define a partir de los parámetros de forma (alfa) y de escala (beta).

Estos parámetros se pueden estimar mediante la aproximación de Thom (1958)

Se destaca que con valores iguales a cero no es posible el cálculo del valor A pues el logaritmo de cero es infinito. En el caso de que aparezcan valores nulos hay que crear una función mixta compuesta de la probabilidad del valor nulo y la probabilidad del valor no nulo: "q" y "p" = 1-q.

Ejemplo:

Con los datos de precipitación del mes de Julio se pide calcular los percentiles 20, 40, 60 y 80 , mediante el empleo de la ley de distribución Gamma.

0.0	44.8	3.2	2.8	0.0
0.0	8.7	2.5	68.6	5.6
9.4	10.0	8.2	71.2	4.0
6.0	2.8	37.1	9.7	37.9
0.0	12.3	16.7	72.9	2.6
10.5	3.9	4.8	13.8	

Solución.

El número de datos de la serie es de 29. Podemos observar que en algunos años durante el mes de Julio no hubo precipitación. Como con los valores iguales a cero no es posible el cálculo del valor A pues el logaritmo de cero es infinito. Hay que crear una función mixta compuesta de la probabilidad del valor nulo "q" y la del valor no nulo "p = 1-q".

Distribución Gamma (Ejemplo cont.)

Solución (cont.).

H(X) = q + p · G(X) Función mixta

q: probabilidad de que se presente un valor cero (sin precipitación) es fácil de calcular considerando los ceros existentes con respecto al total de datos. p = 1-q

Como del total de 29 datos tenemos 4 con cero, tenemos:

q = 4/29 = 0.1379 (13.79)

p = 1- q = 25/29 = 0.8620 (86.21)

Así eliminamos los ceros y hacemos los cálculos sólo para los 25 valores restantes (función G(X) que afecta a "p"), posteriormente al final consideraremos la función mixta (H).

H(X) = q + p · G(X) Función mixta

Suma de los 25 datos = 470

Media = 470/25 = 18.8

Las formulaciones a emplear son:

$$\gamma = \frac{1}{4A}\left[1 + \sqrt{\left(1 + \frac{4A}{3}\right)}\right]$$

$$A = \ln \bar{x} - \frac{\sum \ln x}{n}$$

Tomando el valor de A obtenemos el valor del parámetro alfa "ã" y el valor del parámetro de distribución beta "â" :

Alfa = 0.9109

Beta = 20.6393

Luego para calcular A es necesario calcular el logaritmo neperiano de todos los valores (los 25 no cero). Así:

ln (media) = 2.9338

Suma (lnx) = 57,11256

Luego A es igual a: A = 2,9338 – (57,11256/25) = 0.649

Distribución Gamma (Ejemplo cont.)

Solución (cont.):

Para calcular los percentiles se puede acudir al empleo de tablas o ábacos o emplear un programa de hojas de cálculo como el Excel. Si usamos el Excel hay que usar la función: [=DISTR.GAMMA.INV(probabilidad;alfa;beta)]. Los parámetros de la distribución gamma incompleta alfa y beta ya están calculados, sólo se necesita considerar las probabilidades. Así:
Percentil 20 es la probabilidad igual a 0,20 Como trabajamos con una función mixta :
$H(X) = q + p \cdot G(X)$ Siendo q la probabilidad de que se presente un valor cero (sin precipitación) y p = 1-q. Tenemos que: q = 4/29 = 0.1379 (13.79) ; y, p = 1- q = 25/29 = 0.8620 (86.21)
La precipitación que corresponde a una probabilidad del 0,2 será:

$$H(X) = q + p \cdot G(X) = 0{,}1379 + 0.8620 \cdot G(X) = 0.2\ (20\ \%)$$

Al valor de la probabilidad del 20 % para la función mixta le corresponde una probabilidad referida sólo a los valores no nulos de: G(X) = (0.2 – 0.1379)/0.8620 = 0,072.
No olvidemos que trabajamos sólo con los valores no nulos.
La función Excel a aplicar será: =DISTR.GAMMA.INV(0.072; 0.9109; 20.6393). Así:
Percentil 20 = 1,1 mm
Para el resto será:
G(X) = (0.4 – 0.1379)/0.8620 = 0.3040 . =DISTR.GAMMA.INV(0.3040; 0.9109; 20.6393) .
Percentil 40 = 6.3
G(X) = (0.6 – 0.1379)/0.8620 = 0.5360 . =DISTR.GAMMA.INV(0.536; 0.9109; 20.6393) .
Percentil 60 = 14
G(X) = (0.8 – 0.1379)/0.8620 = 0.768 . =DISTR.GAMMA.INV(0.768; 0.9109; 20.6393) .
Percentil 80 = 27.5

Distribuciones Conjuntas

Estudiaremos por ejemplo dos características de un mismo elemento (dirección y fuerza del viento, etc.).
De forma general, si se estudian sobre una misma población y se miden por las mismas unidades estadísticas una variable X y una variable Y, se obtienen series estadísticas de las variables X e Y.

Considerando simultáneamente las dos series, se suele decir que estamos ante una variable estadística bidimensional.

Vamos a considerar 2 tipos de tabulaciones:

1º) Para variables cuantitativas, que reciben el nombre de tabla de correlación.

2º) Para variables cualitativas, que reciben el nombre de tabla de contingencia.

I) Tablas de correlación.

Sea una población estudiada simultáneamente según dos caracteres X e Y; que representaremos genéricamente como (xi; yj ; nij), donde xi; yj, son dos valores cualesquiera y nij es la frecuencia absoluta conjunta del valor i-ésimo de X con el j-ésimo de Y.

Una forma de disponer estos resultados es la conocida como tabla de doble entrada o tabla de correlación, la cual podemos representar como sigue:

Distribuciones Conjuntas (cont.)

Y / X	y_1	y_2	.....	y_j	.....	y_s	$n_{i\cdot}$	$f_{i\cdot}$
x_1	n_{11}	n_{12}	.....	n_{1j}	.....	n_{1k}	$n_{1\cdot}$	$f_{1\cdot}$
x_2	n_{21}	n_{22}	.....	n_{2j}	.....	n_{2k}	$n_{2\cdot}$	$f_{2\cdot}$
. . .	. .	. .	. .... .	. .	. .... .	. .	. .	. .
x_i	n_{i1}	n_{i2}	.....	n_{ij}	.....	n_{ik}	$n_{i\cdot}$	$f_{i\cdot}$
. . .	. . .	. . .	. .... .	. . .	. .... .	. . .	. . .	. . .
x_r	n_{h1}	n_{h2}	.....	n_{hj}	.....	n_{hk}	$n_{h\cdot}$	$f_{h\cdot}$
$n_{\cdot j}$	$n_{\cdot 1}$	$n_{\cdot 2}$	.....	$n_{\cdot j}$	.....	$n_{\cdot k}$	N	
$f_{\cdot j}$	$f_{\cdot 1}$	$f_{\cdot 2}$	.....	$f_{\cdot j}$	.....	$f_{\cdot k}$		1

En este caso, n_{11} nos indica el número de veces que aparece x_1 conjuntamente con y_1; n_{12}, nos indica la frecuencia conjunta de x_1 con y_2, etc.

Distribuciones Conjuntas (cont.)

Cuando se estudian conjuntamente dos variables, surgen tres tipo de distribuciones:

Distribuciones conjuntas, distribuciones marginales y distribuciones condicionadas.

a) Distribución conjunta

-*La frecuencia absoluta conjunta*, viene determinada por el número de veces que aparece el par ordenado (x_i , y_j), y se representa por " **n_{ij}** **".**

b) Distribuciones marginales

Cuando trabajamos con más de una variable y queremos calcular las distribuciones de frecuencias de cada una de manera independiente, nos encontramos con las distribuciones marginales.

Variable X

x_i	$n_i.$	$f_i.$
x_1	$n_1.$	$f_1.$
x_2	$n_2.$	$f_2.$
x_3	n_3 .	f_3 .
x_4	$n_4.$	$f_4.$
	N	1

Variable Y

y_j	$n.j$	$f.j$
y_1	$n.1$	$f.1$
y_2	$n.2$	$f.2$
y_3	$n.3$	$f.3$
y_4	$n.4$	$f.4$
	N	1

Distribuciones Conjuntas (cont.)

***Frecuencia absoluta marginal**: el valor* ni representa el número de veces que aparece el valor xi de X, sin tener en cuenta cual es el valor de la variable Y. A ni. se le denomina frecuencia absoluta marginal del valor xi de X, de forma que:

$$n_{i.} = n_{i1} + n_{i2} + \ldots + n_{ik}$$

De la misma manera, la **frecuencia absoluta marginal** del valor yj de Y se denotará por n.j

$$n_{.j} = n_{1j} + n_{2j} + \ldots + n_{rj}$$

Frecuencia relativa marginal

La frecuencia relativa marginal de xi de X, viene dada por:

$$f_{i.} = \frac{n_{i.}}{N}$$

La frecuencia relativa marginal de yj de Y, viene dada por:

$$f_{.j} = \frac{n_{.j}}{N}$$

Distribuciones conjuntas y marginales (Ejemplo)

a) Distribución conjunta de la dirección y velocidad del viento

La frecuencia absoluta conjunta, viene determinada por el número de veces que aparece el par ordenado (rango velocidad, rumbo)

b) Distribuciones marginales de la dirección y velocidad del viento

La frecuencia absoluta marginal viene representada por la sumatoria para el rango de velocidad de todos los rumbos o para cada rumbo la sumatoria de todas los rangos de velocidad.

Rango	N	NNE	NE	ENE	E	ESE	SE	SSE	S	SSW	SW	WSW	W	WNW	NW	NNW	TOT
1 5	15	11	19	11	18	24	10	6	10	3	5	6	9	8	21	7	183
6 10	37	31	34	35	54	45	36	24	15	12	10	13	23	20	50	17	456
11 15	73	77	51	47	78	94	61	54	42	14	18	19	19	34	46	24	751
16 20	25	70	38	12	38	84	63	43	46	19	33	24	29	28	37	12	601
21 25	11	20	12	5	10	39	30	18	12	14	10	11	13	14	15	0	234
26 30	4	9	10	1	11	28	20	17	12	15	11	17	16	5	12	0	188
31 35	0	5	5	1	4	6	18	11	8	7	12	13	9	3	2	0	104
36 40	0	1	1	0	1	0	3	4	2	1	2	6	3	0	0	1	25
41 45	0	0	0	0	0	1	3	2	3	0	3	3	1	0	0	0	16
46 50	0	0	0	0	2	2	0	0	2	0	1	1	0	0	0	0	8
51 55	0	0	0	0	0	2	0	0	2	1	0	0	0	0	0	0	5
56 60	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1
Total	165	224	170	112	216	325	244	179	155	86	105	113	122	112	183	61	2572
Vel. Me	15.0	17.6	16.3	13.6	15.9	18.3	19.9	19.8	20.3	22.0	22.3	23.2	20.2	16.9	15.6	14.1	18.1

Distribuciones Conjuntas (cont.)

c) Distribuciones condicionadas

Consideremos a los n.j individuos de la población que representan la modalidad yj de la variable Y, y obsérvese la columna j-ésima de la tabla. Sus n.j elementos constituyen una población, que es un subconjunto de la población total. Sobre este subconjunto se define la distribución de X condicionada por y_j, que se representa por X / y_j ;su frecuencia absoluta se representa por ni / j , y su frecuencia relativa por f_i / j , para i = 1, 2, 3, …., r siendo

$$f_{i\,/\,j} = \frac{n_{ij}}{n_{.j}}$$

El razonamiento es análogo cuando condicionamos la variable Y a un determinado valor de X, es decir Y /xi

Estimación de parámetros

En general, de las variables observadas no conocemos la PDF. Podemos conocer la familia (normal, binomial, etc.) pero no los parámetros. Para **calcularlos** necesitaríamos tener **todos** los posibles valores de la variable, lo que no suele ser posible (p. ej. Clima). 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 para calcularla.

Estimación: Proceso por el que se trata de averiguar un parámetro de la población a partir del valor de un estadístico llamado estimador.

El problema se resuelve en base al conocimiento de la *"distribución muestral"* del estadístico que se use.

Por ejemplo en la media (μ). Si para cada muestra posible calculamos la media muestral ($\bar{x}$) obtenemos un valor distinto. $\bar{x}$ es un estadístico: es una variable aleatoria y sólo depende de la muestra, habrá por tanto una *pdf* para $\bar{x}$, llamada distribución muestral de medias. La desviación típica de esta distribución se denomina *error típico de la media*. Evidentemente, habrá una distribución muestral para cada estadístico, no sólo para la media, y en consecuencia un error típico para cada estadístico.

Si la distribución muestral de un estadístico estuviera relacionada con algún parámetro de interés, ese estadístico podría ser un estimador del parámetro.

Estimación de parámetros (cont.)

- Método de los momentos
- Método de la máxima verosimilitud:
- Método de estimación por intervalos de confianza:
- Método de los mínimos cuadrados: se verá en teoría de la Regresión

Método de los momentos

Consideremos una vez más una ley de probabilidad , dependiente de un parámetro desconocido θ y una muestra $(X_1, \ldots, X_n)$ de esta ley.

Sea f una función de R en R . Si es una variable aleatoria de ley P , la ley de f(x) depende también, en general, de θ y lo mismo sucede con su esperanza. Pero puede ser estimada por la media empírica de . $(f(X_1), \ldots, f(X_n))$ Si se expresa en función de E(f(x) , de aquí deduciremos un estimador de θ . En la mayor parte de los casos, f(x) es una potencia de X o X-E(f(x)). Las cantidades $\mathbb{E}[X^k]$ y $\mathbb{E}[(X - \mathbb{E}[X])^k]$ se llaman los *momentos* de X , de ahí el nombre del método.

Ejemplo de aplicación a la distribución gamma

Si X sigue una ley gamma de parámetros α y λ, su esperanza y su varianza valen:

$$\mathbb{E}[X] = \frac{a}{\lambda} \qquad Var[X] = \frac{a}{\lambda^2} .$$

Por tanto podemos expresar α y λ en función de $\mathbb{E}[X]$ $Var[X]$

$$a = \frac{\mathbb{E}[X]^2}{Var[X]} \qquad \lambda = \frac{\mathbb{E}[X]}{Var[X]} .$$

Estimación de parámetros (cont.)

Método de los momentos

Si se dispone de una muestra $(X_1, \ldots, X_n)$ de la ley gamma de parámetros α y λ, la media empírica $\overline{X}$ y la varianza empírica S^2 son estimadores consistentes de respectivamente $\mathbb{E}[X]$ y $Var[X]$ De aquí obtenemos dos estimadores consistentes de α y λ

$$A = \frac{\overline{X}^2}{S^2} \qquad \Lambda = \frac{\overline{X}}{S^2}\,.$$

Test de Bondad de ajuste Chi Cuadrado

El Test Chi - Cuadrado puede utilizarse para determinar la calidad del ajuste mediante distribuciones teóricas (como la distribución normal o la binomial) de distribución empíricas (o sea las obtenidas de los datos de la muestra).

La **prueba de Chi-cuadrado** es considerada como una prueba no paramétrica que mide la discrepancia entre una distribución observada y otra teórica (bondad de ajuste), indicando en qué medida las diferencias existentes entre ambas, de haberlas, se deben al azar en el contraste de hipótesis. También se utiliza para probar la independencia de dos variables entre sí, mediante la presentación de los datos en tablas de contingencia.

La fórmula que da el estadístico es la siguiente: $$\chi^2 = \sum_i \frac{(\text{observada}_i - \text{teorica}_i)^2}{\text{teorica}_i}$$

Cuanto mayor sea el valor de χ^2, menos verosímil es que la hipótesis sea correcta. De la misma forma, cuanto más se aproxima a cero el valor de chi-cuadrado, más ajustadas están ambas distribuciones.

Los grados de libertad vienen dados por :
gl= (r-1)(k-1). Donde r es el número de filas y k el de columnas.
•Criterio de decisión:
Se acepta H_0 cuando $\chi^2 < \chi_t^2(r-1)(k-1)$ En caso contrario se rechaza.
Donde t representa el valor proporcionado por las tablas, según el nivel de significación estadística elegido.

Test de Bondad de ajuste Chi Cuadrado (Ejemplo)

Sean 1000 valores de temperatura media horaria de las cuales:

38 horas han tenido una temperatura media de 0 °C	**3.8%**
144 horas han tenido una temperatura media de 1 °C	**14.4%**
342 horas han tenido una temperatura media de 2 °C	**34.2%**
287 horas han tenido una temperatura media de 3 °C	**28.7%**
164 horas han tenido una temperatura media de 4 °C	**16.4%**
25 horas han tenido una temperatura media de 5 °C	**2.5%**

$\mu = 2.47$

$\sigma = 1.11$

Media del intervalo Marca de Clase	Límites del intervalo $x \pm \frac{d}{2}$	Variable tipificada $z = \frac{x - \bar{x}}{\sigma}$	Area bajo la curva normal. 0 a z	Area del intervalo	Frec. esper.	Frec. obsda.
0 ºC	- 0,5	-2,675	0,49625	0,03485	3,4	3,8
	0,5	-1,775	0,46210			
1 ºC	0,5	-1,775	0,46210	0,15318	15,3	14,4
	1,5	-0,874	0,30892			
2 ºC	1,5	-0,874	0,30892	0,31972	32,0	34,2
	2,5	0,027	0,01080			
3 ºC	2,5	0,027	0,01080	0,31248	31,2	28,7
	3,5	0,928	0,32328			
4 ºC	3,5	0,928	0,32328	0,14304	14,3	16,4
	4,5	1,829	0,46632			
5 ºC	4,5	1,829	0,46632	0,03048	3,0	2,5
	5,5	2,730	0,49680			

Test de Bondad de ajuste Chi Cuadrado (Ejemplo)

a) Si queremos hallar la bondad del ajuste de la distribución normal del parágrafo 5.8, calcularemos la expresión

$$\chi^2 = \frac{(3,8-3,4)^2}{3,4} + \frac{(14,4-15,3)^2}{15,3} + \frac{(34,2-32,0)^2}{32,0} + \frac{(28,7-31,2)^2}{31,2} +$$

$$+ \frac{(16,4-14,3)^2}{14,3} + \frac{(2,5-3,0)^2}{3,0} = 0,842$$

y como $\sigma = 1,11$ y $\bar{x} = 2,47$, (k=2) se han obtenido a partir de los n=6 datos de la muestra (Sec. 5.8) el número de grados de libertad que es $\nu = n-1-k = 6-1-2 = 3$

La tabla 7.1.2 para 3 grados de libertad y $\chi^2 = 0,842$ da un valor aproximado de 0,17 que nos dice que la hipótesis de que la distribución χ^2 represente el fenómeno es aceptable a un nivel de confianza del 83% o a un nivel de significación del 17%, o del 0,17. No puede pues hablarse de la buena "calidad" del ajuste, aunque es aceptable.

Tabla de Chi Cuadrado

DISTRIBUCION DE χ^2

Grados de libertad	Probabilidad										
	0,95	0,90	0,80	0,70	0,50	0,30	0,20	0,10	0,05	0,01	0,001
1	0,004	0,02	0,06	0,15	0,46	1,07	1,64	2,71	3,84	6,64	10,83
2	0,10	0,21	0,45	0,71	1,39	2,41	3,22	4,60	5,99	9,21	13,82
3	0,35	0,58	1,01	1,42	2,37	3,66	4,64	6,25	7,82	11,34	16,27
4	0,71	1,06	1,65	2,20	3,36	4,88	5,99	7,78	9,49	13,28	18,47
5	1,14	1,61	2,34	3,00	4,35	6,06	7,29	9,24	11,07	15,09	20,52
6	1,63	2,20	3,07	3,83	5,35	7,23	8,56	10,64	12,59	16,81	22,46
7	2,17	2,83	3,82	4,67	6,35	8,38	9,80	12,02	14,07	18,48	24,32
8	2,73	3,49	4,59	5.53	7,34	9,52	11,03	13,36	15,51	20,09	26,12
9	3,32	4,17	5,38	6,39	8,34	10,66	12,24	14,68	16,92	21,67	27,88
10	3,94	4,86	6,18	7,27	9,34	11,78	13,44	15,99	18,31	23,21	29,59
	No significativo								Significativo		