Clase 2: Nociones Avanzadas de Probabilidades y Estadística Básica

Pablo Barceló
DCC, Universidad de Chile

1. Probabilidades Avanzadas: Correlación de Variables

Notación: Cuando sea necesario denotaremos a $E(X)$ por μ_X y a $Var(X)$ por σ_X^2.

Probabilidad conjunta y marginal: Sean $X, Y : S \to \mathbb{R}$ dos variables aleatorias discretas. Definimos su *función de probabilidad conjunta* $f(x,y) : \mathbb{R} \times \mathbb{R} \to [0,1]$ de tal forma que para todo par $(r,t) \in \mathbb{R} \times \mathbb{R}$ se cumple que:

$$f(r,t) \; = \; Pr(X = r \cap Y = t) \; = \sum_{\{s \in S \mid X(s) = r \text{ y } Y(s) = t\}} Pr(s).$$

Note que $\sum_{(x,y) \in \mathbb{R} \times \mathbb{R}} f(x,y) = 1$.

Definimos también la *función de probabilidad marginal con respecto a X* como $f_X : \mathbb{R} \to [0,1]$ de tal forma que $f_X(r) = \sum_{t \in \mathbb{R}} f(r,t)$, para todo $r \in \mathbb{R}$. Análogamente definimos $f_Y : \mathbb{R} \to [0,1]$.

Esto generaliza fácilmente a n variables $X_1, \ldots, X_n$.

Variables independientes: X e Y son *independientes* si $f(r,t) = f_X(r) \cdot f_Y(t)$ para todo par $(r,t) \in \mathbb{R} \times \mathbb{R}$.

Ejercicio: Considere la siguiente tabla de probabilidad conjunta de las variables X e Y:

$f(x,y)$	0	100	200
100	.2	.1	.2
250	.05	.15	.3

¿Es cierto que estas variables son independientes? Respuesta: No. Es claro que $f_X(100) = 0{,}5$ y $f_Y(0) = 0{,}25$. Además, $f(100,0) = 0{,}2$. Esto quiere decir que $f(100,0) \neq f_X(100) \cdot f_Y(0)$.

En el caso de n variables aleatorias $X_1, \ldots, X_n$, tenemos que estas son independientes si para cualquier subconjunto A de $\{1, \ldots, n\}$ se cumple que:

$$Pr(\bigcap_{i \in A} X_i = r_i) \; = \; \prod_{i \in A} Pr(X_i = r_i),$$

donde los r_i's son reales cualesquiera.

Distribuciones condicionales: Definimos $Pr(X = x \mid Y = y) = Pr(X = x \cap Y = y)/Pr(Y = y)$.

Ejercicio: En la tabla anterior, ¿cuánto vale $Pr(X = 100 \mid Y = 100)$? Respuesta: Tenemos que $Pr(X = 100 \cap Y = 100) = 0,1$. Además. $Pr(Y = 100) = 0,25$. Entonces $Pr(X = 100 \mid Y = 100) = 2/5$.

Valor esperado: Uno podría definir el valor esperado $E(X,Y)$ de las variables aleatorias X e Y como $\sum_{(r,t) \in \mathbb{R} \times \mathbb{R}} (r,t) \cdot f(r,t)$. Esto entrega un valor en $\mathbb{R} \times \mathbb{R}$. Sin embargo, en la práctica tiene mayor utilidad calcular el valor esperado de una función h que toma las dos variables aleatorias X e Y y entrega un valor real. Por ejemplo, h podría computar el valor absoluto de la diferencia entre X e Y, o su promedio. En tal caso definimos

$$E(h(X,Y)) \;=\; \sum_{(r,t)\in\mathbb{R}\times\mathbb{R}} h(r,t)\cdot f(r,t).$$

Ejercicio: ¿Cuál es el valor esperado del promedio entre X e Y en la tabla anterior? Respuesta: $50 \cdot 0,2 + 100 \cdot 0,1 + 150 \cdot 0,2 + 125 \cdot 0,05 + 175 \cdot 0,15 + 225 \cdot 0,3$.

Ejercicio: 5 amigos compran entradas en asientos consecutivos 1-5 en la misma fila para un concierto. Asuma que estos asientos se reparten al azar entre los 5 amigos. Sean X e Y la variable que denota el asiento asignado al primer y segundo amigo, respectivamente. ¿Cuál es el valor esperado para el número de asientos que existen entre ellos? Respuesta: Note que (X,Y) solo puede tomar valores $r,t \in \{1,\ldots,5\}$ tal que $r \neq t$, y cada uno de estos valores ocurre con igual probabilidad $1/20$. La función de probabilidad conjunta f satisface entonces que $f(r,t)$ toma valor $1/20$ cada vez que $r,t \in \{1,\ldots,5\}$ y $r \neq t$, y toma valor 0 en todos los otros puntos. El número de asientos entre el primer y el segundo amigo puede definirse como $h(X,Y) = |X-Y|-1$. La tabla que define $h(X,Y)$ es la siguiente:

$h(x,y)$	1	2	3	4	5
1	–	0	1	2	3
2	0	–	0	1	2
3	1	0	–	0	1
4	2	1	0	–	0
5	3	2	1	0	–

Es fácil concluir entonces que $E(h(x,y)) = 1$.

Correlación: Cuando las variables X e Y no son independientes es conveniente medir su nivel de dependencia. Esto se define mediante la *covarianza*:

$$Cov(X,Y) \;=\; E((X-\mu_X)(Y-\mu_Y)).$$

Note que $Cov(X,X) = E((X-\mu_X)^2) = \sigma_X^2$.

Intuición: Si hay una alta correlación positiva entre las variables (por ejemplo, ambas difieren mucho de su media al mismo tiempo), entonces la covarianza será alta. Lo mismo si tienen una correlación negativa. Por otro lado, si no hay mayor correlación los productos tenderán a cancelarse y la covarianza se hará cada vez más pequeña.

Importante: $Cov(X,Y) = E(XY) - \mu_X \cdot \mu_Y$. Por tanto, si X e Y son independientes su covarianza es 0.

Ejercicio: Considere nuevamente la función de probabilidad conjunta dada por la siguiente tabla:

$f(x,y)$	0	100	200
100	.2	.1	.2
250	.05	.15	.3

¿Cuál es la covarianza de estas dos variables? Respuesta: Es posible demostrar que $\mu_X = 175$ y $\mu_Y = 125$. Además, $E(XY) = 10000 \cdot 0{,}1 + 25000 \cdot 0{,}15 + 20000 \cdot 0{,}2 + 50000 \cdot 0{,}3 = 23750$. Por tanto, $Cov(X,Y) = E(XY) - \mu_X \cdot \mu_Y = 23750 - 175 \cdot 125 = 1875$.

Importante: Si X e Y son variables aleatorias, entonces $Var(X+Y) = Var(X) + Var(Y) + 2Cov(X,Y)$.

Problema con la covarianza: Está afectada por las unidades de medida. Asuma por ejemplo que en la tabla anterior X e Y representan valores en dólares, y que ahora cambiamos la unidad de medida a pesos chilenos. Entonces la covarianza aumenta en aproximadamente 700^2.

Coeficiente de correlación: Corresponde a $\rho(X,Y) = Cov(X,Y)/(\sigma_X \cdot \sigma_Y)$.

Ejercicio: Calcule el coeficiente de correlación en el ejemplo anterior. Respuesta: $\rho = 0{,}301$.

Importante: Siempre se cumple lo siguiente:

1. Si X e Y son independientes entonces $\rho(X,Y) = 0$ (pero lo contrario no necesariamente es cierto).
2. $\rho(aX + b, cX + d) = \rho(X,Y)$.
3. $-1 \leq \rho(X,Y) \leq 1$.
4. $|\rho(X,Y)| = 1$ si y solo si Y es de la forma $aX + B$.

En conclusión, la correlación mide la *dependencia lineal* entre dos variables. En caso en que la correlación sea 0 no podemos conluir entonces que no haya correlación: esta podría existir pero ser altamente no lineal. Si $\rho = 0$ decimos que X e Y no están correlacionados.

2. Estadística: Nociones Básicas

La estadística sirve para trabajar en escenarios con información incompleta y variación. Los métodos estadísticos ayudan a predecir valores y ver si exceden ciertos límites que los hacen significativos (por ejemplo, ayudan a detectar casos críticos) o a diseñar tests para comparar entre diferentes productos (en tal caso pueden ayudar a decidir si el nuevo producto es mejor que el anterior). Aún más en general, la estadística sirve para organizar nuestros datos y obtener conclusiones valiosas a partir de ellos.

Típicamente tenemos una población que queremos observar y una muestra de esta población. Lo que observamos son valores (categóricos o numéricos) de los datos que participan en la muestra. A partir de estos valores queremos deducir algo acerca del valor de ciertos parámetros de una población.

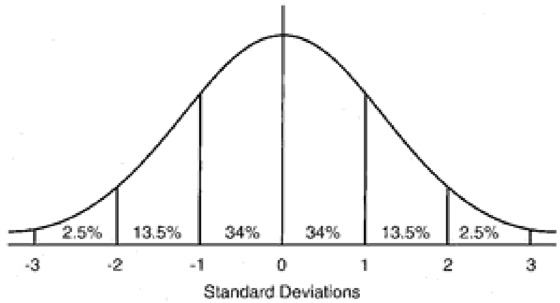
Distribución binomial: Representa n repeticiones de un experimento con dos salidas, las cuales ocurren con probabilidad p y $(1-p)$, respectivamente (por ejemplo, lanzar una moneda, sacar bolas rojas de un recipiente con bolas rojas y verdes, etc). Recordemos que si X representa el número de veces que ocurre la primera salida entonces $f(x) = Pr(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$. Recordemos además que $E(X) = np$ y $Var(X) = np(1-p)$.

Ejercicio: En el ejército es común testear a grupos de personas por ciertas enfermedades que ocurren raramente. Asuma que tal enfermedad ocurre con probabilidad 0,002 y que debemos testear a 1000 conscriptos. Una técnica usual para realizar esto (sin tener que realizar siempre los 1000 tests) es la siguiente: Dividimos a los conscriptos en 10 grupos de 100 personas, tomamos una muestra de sangre de cada persona en cada grupo y la combinamos. Por cada uno de esos grupos realizamos el test en la muestra combinada. Si la muestra da negativo quiere decir que ninguno de los 100 conscriptos tenía la enfermedad. De otra forma, uno de ellos la tiene y es preciso refinar el test. En el mejor caso realizaremos tan solo 10 tests y en el peor deberemos hacer 1010. ¿Cuál es el número esperado de tests que deberemos realizar? Respuesta: Sea Z_i variable que cuenta el número de personas que tienen la enfermedad en el grupo i, para $1 \leq i \leq 10$, e Y_i una variable que vale 1 si $Z_i > 0$ y vale 0 en caso contrario. Defina Y como el número de grupos en los que deberemos refinar el test. Entonces $Y = Y_1 + \cdots + Y_{10}$. Luego el número de tests que deberemos realizar es $100Y + 10$. Calcularemos, por tanto, el valor esperado de $100Y + 10$, lo que corresponde a $100 \cdot E(Y) + 10$. Note que $E(Y) = E(Y_1) + \cdots + E(Y_10)$. Por otro lado, para todo $1 \leq i \leq 10$ tenemos que $E(Y_i) = Pr(Y_i = 1) = Pr(Z_i > 0) = 1 - Pr(Z_i = 0) = 1 - 0,998^{100} = 0,181$. Por tanto, el número esperado de tests es $100 \times 10 \times 0,181 + 10 = 191$. ¿Cuál es la probabilidad de que tengamos que realizar 1010 tests? Esto corresponde a $Pr(Y_1 = 1 \cap \cdots \cap Y_{10} = 1) = Pr(Y_1 = 1) \cdots Pr(Y_{10} = 1)$. Pero $Pr(Y_i = 1) = 0,181$, por tanto $Pr(Y_1 = 1) \cdots Pr(Y_{10} = 1) = (0,181)^{10}$.

Distribución Normal: Esta es la distribución más utilizada y estudiada debido a su ubicuidad en variadas áreas. La distribución normal con media μ y desviación estándar σ se halla definida por la fórmula:

$$f(x) \; = \; \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

Asumamos que X distribuye normal con valor esperado μ y desviación estándar σ. Queremos calcular la probabilidad de que $X \leq x$. Esto no es fácil de hacer, sin embargo se halla calculado en el caso $\mu = 0$ y $\sigma^2 = \sigma = 1$. Esta distribución normal se llama *estándar*. A continuación vemos una representación gráfica de tal distribución, la que además muestra la probabilidad de que el valor x esté en un cierto rango:



En particular, es posible encontrar el valor $\Phi(z) = Pr(Z \leq z)$. Estos valores pueden encontrarse en la Figura **??**.

Ejercicio: ¿Cuánto valen $Pr(Z \leq 1,25)$, $Pr(Z > 1,25)$, $Pr(Z \leq -1,25)$, y $Pr(-0,38 \leq Z \leq 1,25)$? Respuesta: $Pr(Z \leq 1,25) = \Phi(1,25) = 0,8944$, $Pr(Z > 1,25) = 1 - \Phi(1,25) = 0,1056 = Pr(Z \leq -1,25)$, y $Pr(-0,38 \leq Z \leq 1,25) = \Phi(-0,38) + (1 - \phi(1,25)) = 1 - \Phi(0,38) + 1 - \Phi(1,25) = 0,352 + 0,1056$.

En muchos casos utilizamos también los *valores críticos*. Para un $0 \leq \alpha \leq 1$ definimos el *valor crítico* z_α como el valor que satisface $Pr(Z \geq z_\alpha) = \alpha$. Por ejemplo, para $\alpha = 0,05$ tenemos que $z_\alpha = 1,645$. Esto quiere decir que el 95 % de la probabilidad de la distribución normal está antes de 1,645.

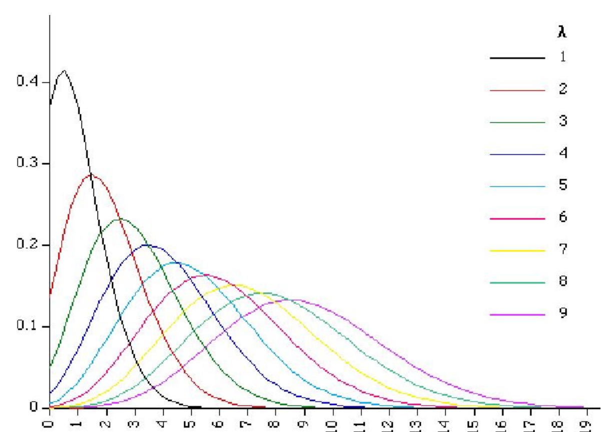
¿Cómo podemos utilizar esto entonces para computar probabilidades para nuestra distribución original? Defina $Z = (X - \mu)/\sigma$. Se puede demostrar que Z tiene distribución normal con valor esperado 0 y desviación 1. Por tanto, $Pr(X \leq x) = Pr(Z\sigma + \mu \leq x) = Pr(Z \leq (x - \mu)/\sigma)$, lo que sabemos calcular. Además, $Pr(y \leq X \leq x) = Pr(\frac{y-\mu}{\sigma} \leq Z \leq \frac{x-\mu}{\sigma})$.

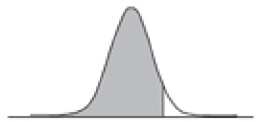
Ejercicio: Se ha sugerido que el tiempo que le toma a un conductor apretar el freno en una situación de emergencia distribuye normal con media 1,25 segundos y desviación estándar de 0,46 segundos. ¿Cuál es la probabilidad de que el tiempo de reacción esté entre 1 y 1,75 segundos? Respuesta: Queremos determinar $Pr(1 \leq X \leq 1,75) = Pr(\frac{1-1,25}{0,46} \leq Z \leq \frac{1,75-1,25}{0,46}) = 0,5675$. ¿Cuál es la probabilidad de que el tiempo de reacción se mayor o igual a 2? Respuesta: $1 - \Phi((2 - 1,25)/0,46) = 0,0516$.

Distribución Poisson: Representa los tiempos de llegada aleatorios, por ejemplo, de clientes, de inundaciones, etc. En particular, describe el número de tales llegadas en un período fijo de tiempo (por ejemplo, un minuto). Se define como

$$f(x) = \frac{e^{-\lambda}\lambda^x}{x!}, \qquad \text{para } x = 0, 1, 2, \ldots,$$

donde $\lambda > 0$ es un parámetro. Es posible demostrar que $E(X) = Var(X) = \lambda$. La siguiente figura muestra la distribución Poisson para diferentes valores de λ:





a	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

Figura 1: Tabla de valores para la distribución normal

Ejemplo: Asuma que el promedio de llegada de personas a un paradero es de 4,5 personas por hora, es decir, de 0,00125 por segundo. Es decir, la probabilidad de que una persona llegue en un segundo es de 0.00125. ¿Cuál es la probabilidad de que n personas lleguen en una hora? Podríamos tomar una binomial con 3600 valores posible (los segundos de una hora) y ver la posibilidad de que en n de ellos ocurra el evento (que llegue una persona). Es decir, debemos calcular $\binom{3600}{n}0,00125^n(1-0,00125)^{3600-n}$. Esto es bastante engorroso. Sin embargo, es posible aproximarlo con una Poisson de media $4,5$. Es decir, esto corresponde aproximadamente a $\frac{e^{-4,5}\cdot 4,5^n}{n!}$.