

TEORÍA ESTADÍSTICA: APLICACIONES Y MÉTODOS

en conclusión $f(x, \theta)$ es de la forma (1.2.1) con $d(\theta) = \ln \theta$, $T(x) = x$, $c(\theta) = \exp\{-\theta\}$ y $h(x) = \frac{I_{\{0,1,\dots\}}(x)}{x!}$.

Ejemplo 1.2.2. La distribución Gamma con parámetro de forma k conocida pertenece a la familia exponencial uniparamétrica puesto que

$$\begin{aligned} f(x, \theta) &= \frac{x^{k-1} e^{-x/\theta}}{\theta^k \Gamma(k)} I_{(0,\infty)}(x) \\ &= \exp\left\{-\frac{x}{\theta}\right\} \theta^{-k} \frac{x^{k-1} I_{(0,\infty)}(x)}{\Gamma(k)}, \end{aligned}$$

el cual es de la forma (1.2.1) con $d(\theta) = -1/\theta$, $T(x) = x$, $c(\theta) = \theta^{-k}$ y $h(x) = \frac{x^{k-1} I_{(0,\infty)}(x)}{\Gamma(k)}$.

Nótese que esta representación de la familia exponencial no es única, puesto que al definir $d(\theta) = 1/\theta$ y $T(x) = -x$, también se puede concluir que la distribución Gamma con k fijo pertenece a la familia exponencial uniparamétrica.

En casi toda la teoría estadística, se tratan más de una variable aleatoria que en la práctica, se observan los datos que corresponden a realizaciones de estas variables. En este caso, se examina la pertenencia de la familia exponencial de la función de densidad conjunta, y tenemos el siguiente resultado.

Resultado 1.2.1. Si X_1, \dots, X_n son variables aleatorias independientes e idénticamente distribuidas con función de densidad común perteneciente a la familia exponencial uniparamétrica, entonces la función de densidad conjunta $f(x_1, \dots, x_n)$ también pertenece a la familia exponencial uniparamétrica.

Proof.

$$\begin{aligned} f(x_1, \dots, x_n, \theta) &= \prod_{i=1}^n f(x_i, \theta) \\ &= \prod_{i=1}^n h(x_i) c(\theta) \exp\{d(\theta) T(x_i)\} \\ &= c(\theta)^n \left[\prod_{i=1}^n h(x_i) \right] \exp\left\{d(\theta) \sum_{i=1}^n T(x_i)\right\} \end{aligned} \quad (1.2.2)$$

el cual es de la forma (1.2.1). \square

Usando el anterior resultado junto con el Ejemplo 1.2.1 donde se mostró que la distribución Poisson pertenece a la familia exponencial, podemos afirmar que cuando se tiene n variables independientes e idénticamente distribuidas con distribución Poisson, la función conjunta también pertenece a la familia exponencial.

Otra utilidad del resultado es que cuando se necesita ver que una función de densidad conjunta pertenece a la familia exponencial, basta ver para la función de densidad marginal.

La estadística $\sum_{i=1}^n T(x_i)$ en (1.2.2) será de gran interés en los futuros capítulos, y estamos interesados en conocer sus propiedades como la esperanza y la varianza, y para eso solo necesitamos $E(T(X))$ y $Var(T(X))$ para X con la misma distribución que X_1, \dots, X_n . El siguiente resultado nos provee las respectivas fórmulas.

Resultado 1.2.2. *Si X es una variable aleatoria con función de densidad perteneciente a la familia exponencial de la forma (1.2.1), entonces*

1. $E\left(\frac{\partial d(\theta)}{\partial \theta} T(X)\right) = -\frac{\partial}{\partial \theta} \ln c(\theta)$
2. $Var\left(\frac{\partial d(\theta)}{\partial \theta} T(X)\right) = -\frac{\partial^2}{\partial \theta^2} \ln c(\theta) - E\left(\frac{\partial^2 d(\theta)}{\partial \theta^2} T(X)\right)$

Ahora, las distribuciones con dos parámetros, como la distribución $N(\mu, \sigma^2)$ puede considerarse como distribución uniparamétrica cuando μ o σ^2 se considera fijo conocido. Y se puede ver que en ambos casos, la distribución pertenece a la familia exponencial uniparamétrica.

1.2.2 Familia exponencial multi-paramétrica

Las distribuciones Gamma, normal y beta dependen de dos parámetros, y para estas distribuciones, se puede generalizar la definición de la familia exponencial uniparamétrica para distribuciones dependientes de más de un parámetro. La definición correspondiente se da a continuación.

Definición 1.2.2. *Una distribución de probabilidad pertenece a la familia exponencial multi-paramétrica si la función de densidad se puede escribir de la forma*

$$f_X(x, \boldsymbol{\theta}) = c(\boldsymbol{\theta})h(x) \exp\{d(\boldsymbol{\theta})'T(x)\} \quad (1.2.3)$$

donde $T(x)$ y $d(\boldsymbol{\theta})$ son funciones vectoriales, $h(x)$ y $c(\boldsymbol{\theta})$ son funciones reales.

Ejemplo 1.2.3. *La distribución Gamma con parámetro de forma k y parámetro de escala θ pertenece a la familia exponencial multi-paramétrica pues*

$$\begin{aligned} f_X(x) &= \frac{x^{k-1}e^{-x/\theta}}{\theta^k \Gamma(k)} I_{(0,\infty)}(x) \\ &= \exp\left\{-\frac{x}{\theta} + (k-1) \ln x\right\} \frac{1}{\theta^k \Gamma(k)} I_{(0,\infty)}(x) \\ &= \exp\left\{\left(-\frac{1}{\theta}, k-1\right) \begin{pmatrix} x \\ \ln x \end{pmatrix}\right\} \frac{1}{\theta^k \Gamma(k)} I_{(0,\infty)}(x) \end{aligned}$$

el cual es de la forma (1.2.3) con $\boldsymbol{\theta} = (\theta, k)'$, $d(\boldsymbol{\theta}) = (-\frac{1}{\theta}, k-1)'$, $c(\boldsymbol{\theta}) = \frac{1}{\theta^k \Gamma(k)}$, $T(x) = (x, \ln x)'$ y $h(x) = I_{(0,\infty)}(x)$.

Nótese que al igual a la familia exponencial uniparamétrica, la representación de la familia exponencial multi-paramétrica tampoco es única, pues en el ejemplo anterior también se puede tomar $\eta(\theta) = (\frac{1}{\theta}, k-1)'$ y $T(x) = (-x, \ln x)'$.

Las distribuciones normal, gamma, exponencial, chi-cuadrado, beta, Bernoulli, binomial, binomial negativa, multinomial, Poisson y geométrica todas pertenecen a la familia exponencial. También la distribución Weibull pertenece a la familia exponencial cuando el parámetro de forma es conocida. Por el otro lado, las distribuciones Cauchy, Laplace, uniforme y Weibull cuando el parámetro de forma es desconocida no pertenecen a la familia exponencial.

La razón por la que las distribuciones de la familia uniforme no pertenecen a la familia exponencial va más allá de los objetivos de este libro, pues se necesita conocimientos sobre teoría estadística basada en la teoría de la medida. La afirmación de que la distribución uniforme no pertenece a la familia exponencial porque no se puede escribir de la forma (1.2.3) no es una razón válida.

1.3 Ejercicios

[1.1]

1. (Captura-recaptura) Suponga que un piscicultor necesita conocer el número de peces que tiene en un estanque, pero por limitación de los recursos humanos y económicos no puede sacar todos los peces del estanque y contar uno por uno, usted como conocer de la ciencia estadística, ¿cómo le puede ayudar? (Ayuda: Usa la distribución hipergeométrica).
2. Un vendedor de seguros realiza en promedio 10 visitas por semana a los posibles clientes, él por experiencia sabe que la probabilidad de que un cliente compre el seguro en una visita es de 0.2, y suponga que el resultado de una visita no se ve afectado por resultados de visitas anteriores.
 - (a) ¿Cuál es la probabilidad de que el vendedor en una semana venda más de dos seguros?
 - (b) Suponga que el vendedor ha tenido semanas consecutivas con muy malas ventas, y que perderá el trabajo si en la próxima semana no vende por lo menos un seguro. Por lo tanto el vendedor decide aumentar el número de visitas a la semana para tratar al menos un seguro, ¿por lo menos cuántas visitas debe realizar la próxima semana para que la probabilidad de vender al menos un seguro sea superior a 90%?
3. Demuestre el Resultado 1.1.6. (Ayuda: primero encuentra $m_X(t)$, luego calcular $E(X)$ y $Var(X)$, para encontrar $m_X(t)$ usar la expansión de Taylor de e^x .)
4. Demuestre el Resultado 1.1.12.

5. Para las distribuciones exponencial y chi-cuadrado, escriba cuál es el parámetro y cuál es el espacio paramétrico.
6. Sea $X \sim N(\mu, \sigma^2)$, y sea $Z_1 = \frac{X-\mu}{\sigma}$ y $Z_2 = \frac{\mu-X}{\sigma}$, compruebe que Z_1 tiene distribución normal estándar, y Z_2 también, es decir, la forma de estandarizar una variable con distribución normal no es única.
7. Sea X_1, \dots, X_n variables aleatorias independientes e idénticamente distribuidas con distribución $N(\mu, \sigma^2)$, encuentre la distribución de la variable $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$.
8. Demuestre el Resultado 1.1.15.
9. Calcule las siguientes probabilidades:
 - (a) $P(X > 2)$, $P(X < -1)$ y $P(1 < X < 3)$ donde $X \sim N(1.5, 4)$.
 - (b) $P(\frac{X-2Y}{3} > 4)$ donde $X \sim N(1.5, 4)$ y $Y \sim N(-1, 2)$ son variables independientes.
10. (a) Calcule el percentil 5%, 10% y 98% de una variable con distribución normal estándar.
 (b) Calcule el percentil 5% y 90% de una variable con distribución $N(1, 3)$.
 (c) Encuentre valores a y b tales que $P(a < Z < b) = 0.98$.
 (d) Encuentre valores a y b tales que $P(a < X < b) = 0.95$ donde $X \sim N(-2, 5)$.
11. Encuentre los percentiles 5%, 98% de una variable con distribución chi-cuadrado con 8 y 15 grados de libertad, respectivamente.
12. Encuentre los percentiles 5%, 98% de una variable con distribución t-student con 10 y 20 grados de libertad, respectivamente.
13. Encuentre los percentiles 5%, 98% de una variable con distribución F_5^8 y F_8^{12} , respectivamente.
14. Demuestre las partes 1 y 2 del resultado 1.1.16 usando la definición 1.1.11.
15. Demuestre que las siguientes distribuciones pertenecen a la familia exponencial identificando las funciones $d(\theta)$, $T(x)$, $c(\theta)$ y $h(x)$:
 - (a) Binomial con n conocido.
 - (b) Exponencial.
 - (c) Distribución normal con media μ conocido.
 - (d) Distribución normal con varianza σ^2 conocida.
 - (e) La función de densidad conjunta de X_1, \dots, X_n , donde $X_i \sim_{iid} N(\mu, \sigma^2)$ con μ conocido.
 - (f) La función de densidad conjunta de X_1, \dots, X_n , donde $X_i \sim_{iid} P(\lambda)$.

16. La función de densidad de la distribución Weibull está dada por:

$$f(x) = \alpha \beta x^{\beta-1} \exp\{-\alpha x^\beta\} I_{(0,\infty)}(x)$$

donde $\alpha > 0$ y $\beta > 0$. Demuestre que cuando β es conocido, la distribución pertenece a la familia exponencial uniparamétrica, identifica $d(\theta)$, $T(x)$, $c(\theta)$ y $h(x)$.

17. Considera la distribución $N(\mu, \sigma^2)$,
- (a) describa cómo es la forma del vector de parámetros y el correspondiente espacio paramétrico
 - (b) demuestre que esta distribución pertenece a la familia exponencial multiparamétrica, identifica $d(\theta)$, $T(x)$, $c(\theta)$ y $h(x)$.
18. Demuestre que la distribución Beta pertenece a la familia exponencial multiparamétrica, identifica $d(\theta)$, $T(x)$, $c(\theta)$ y $h(x)$.
19. Sean X_1, \dots, X_n variables aleatorias con $X_i \sim_{iid} N(\mu_1, \sigma^2)$ para $i = 1, \dots, n$ y Sean Y_1, \dots, Y_m variables con $Y_j \sim_{iid} N(\mu_2, \sigma^2)$ para $j = 1, \dots, m$, además las variables X_i son independientes de las variables Y_j , demuestre que la función de densidad conjunta de $X_1, \dots, X_n, Y_1, \dots, Y_m$ pertenece a la familia exponencial con tres parámetros.
20. Comprobar que en una distribución normal estándar $z_p = -z_{1-p}$, corrobora lo anterior con los percentiles dados en la Tabla 1.1.

gl	p									
	0.005	0.01	0.025	0.05	0.1	0.9	0.95	0.975	0.99	0.995
1	-63.66	-31.82	-12.71	-6.31	-3.08	3.08	6.31	12.71	31.82	63.66
2	-9.92	-6.96	-4.30	-2.92	-1.88	1.88	2.92	4.30	6.96	9.92
3	-5.84	-4.54	-3.18	-2.35	-1.64	1.64	2.35	3.18	4.54	5.84
4	-4.60	-3.75	-2.78	-2.13	-1.53	1.53	2.13	2.78	3.75	4.60
5	-4.03	-3.36	-2.57	-2.01	-1.47	1.47	2.01	2.57	3.36	4.03
6	-3.71	-3.14	-2.45	-1.94	-1.44	1.44	1.94	2.45	3.14	3.71
7	-3.50	-3.00	-2.36	-1.89	-1.41	1.41	1.89	2.36	3.00	3.50
8	-3.36	-2.90	-2.31	-1.86	-1.40	1.40	1.86	2.31	2.90	3.36
9	-3.25	-2.82	-2.26	-1.83	-1.38	1.38	1.83	2.26	2.82	3.25
10	-3.17	-2.76	-2.23	-1.81	-1.37	1.37	1.81	2.23	2.76	3.17
11	-3.11	-2.72	-2.20	-1.80	-1.36	1.36	1.80	2.20	2.72	3.11
12	-3.05	-2.68	-2.18	-1.78	-1.36	1.36	1.78	2.18	2.68	3.05
13	-3.01	-2.65	-2.16	-1.77	-1.35	1.35	1.77	2.16	2.65	3.01
14	-2.98	-2.62	-2.14	-1.76	-1.35	1.35	1.76	2.14	2.62	2.98
15	-2.95	-2.60	-2.13	-1.75	-1.34	1.34	1.75	2.13	2.60	2.95
16	-2.92	-2.58	-2.12	-1.75	-1.34	1.34	1.75	2.12	2.58	2.92
17	-2.90	-2.57	-2.11	-1.74	-1.33	1.33	1.74	2.11	2.57	2.90
18	-2.88	-2.55	-2.10	-1.73	-1.33	1.33	1.73	2.10	2.55	2.88
19	-2.86	-2.54	-2.09	-1.73	-1.33	1.33	1.73	2.09	2.54	2.86
20	-2.84	-2.53	-2.09	-1.72	-1.33	1.33	1.72	2.09	2.53	2.84
21	-2.83	-2.52	-2.08	-1.72	-1.32	1.32	1.72	2.08	2.52	2.83
22	-2.82	-2.51	-2.07	-1.72	-1.32	1.32	1.72	2.07	2.51	2.82
23	-2.81	-2.50	-2.07	-1.71	-1.32	1.32	1.71	2.07	2.50	2.81
24	-2.80	-2.49	-2.06	-1.71	-1.32	1.32	1.71	2.06	2.49	2.80
25	-2.79	-2.49	-2.06	-1.71	-1.32	1.32	1.71	2.06	2.49	2.79
26	-2.78	-2.48	-2.06	-1.71	-1.31	1.31	1.71	2.06	2.48	2.78
27	-2.77	-2.47	-2.05	-1.70	-1.31	1.31	1.70	2.05	2.47	2.77
28	-2.76	-2.47	-2.05	-1.70	-1.31	1.31	1.70	2.05	2.47	2.76
29	-2.76	-2.46	-2.05	-1.70	-1.31	1.31	1.70	2.05	2.46	2.76
30	-2.75	-2.46	-2.04	-1.70	-1.31	1.31	1.70	2.04	2.46	2.75
40	-2.70	-2.42	-2.02	-1.68	-1.30	1.30	1.68	2.02	2.42	2.70
50	-2.68	-2.40	-2.01	-1.68	-1.30	1.30	1.68	2.01	2.40	2.68
60	-2.66	-2.39	-2.00	-1.67	-1.30	1.30	1.67	2.00	2.39	2.66
70	-2.65	-2.38	-1.99	-1.67	-1.29	1.29	1.67	1.99	2.38	2.65
80	-2.64	-2.37	-1.99	-1.66	-1.29	1.29	1.66	1.99	2.37	2.64
90	-2.63	-2.37	-1.99	-1.66	-1.29	1.29	1.66	1.99	2.37	2.63

Tabla 1.3: Algunos percentiles de la distribución t student.

gl	p									
	0.005	0.01	0.025	0.05	0.1	0.9	0.95	0.975	0.99	0.995
1	0.00	0.00	0.00	0.00	0.02	2.71	3.84	5.02	6.63	7.88
2	0.01	0.02	0.05	0.10	0.21	4.61	5.99	7.38	9.21	10.60
3	0.07	0.11	0.22	0.35	0.58	6.25	7.81	9.35	11.34	12.84
4	0.21	0.30	0.48	0.71	1.06	7.78	9.49	11.14	13.28	14.86
5	0.41	0.55	0.83	1.15	1.61	9.24	11.07	12.83	15.09	16.75
6	0.68	0.87	1.24	1.64	2.20	10.64	12.59	14.45	16.81	18.55
7	0.99	1.24	1.69	2.17	2.83	12.02	14.07	16.01	18.48	20.28
8	1.34	1.65	2.18	2.73	3.49	13.36	15.51	17.53	20.09	21.95
9	1.73	2.09	2.70	3.33	4.17	14.68	16.92	19.02	21.67	23.59
10	2.16	2.56	3.25	3.94	4.87	15.99	18.31	20.48	23.21	25.19
11	2.60	3.05	3.82	4.57	5.58	17.28	19.68	21.92	24.72	26.76
12	3.07	3.57	4.40	5.23	6.30	18.55	21.03	23.34	26.22	28.30
13	3.57	4.11	5.01	5.89	7.04	19.81	22.36	24.74	27.69	29.82
14	4.07	4.66	5.63	6.57	7.79	21.06	23.68	26.12	29.14	31.32
15	4.60	5.23	6.26	7.26	8.55	22.31	25.00	27.49	30.58	32.80
16	5.14	5.81	6.91	7.96	9.31	23.54	26.30	28.85	32.00	34.27
17	5.70	6.41	7.56	8.67	10.09	24.77	27.59	30.19	33.41	35.72
18	6.26	7.01	8.23	9.39	10.86	25.99	28.87	31.53	34.81	37.16
19	6.84	7.63	8.91	10.12	11.65	27.20	30.14	32.85	36.19	38.58
20	7.43	8.26	9.59	10.85	12.44	28.41	31.41	34.17	37.57	40.00
21	8.03	8.90	10.28	11.59	13.24	29.62	32.67	35.48	38.93	41.40
22	8.64	9.54	10.98	12.34	14.04	30.81	33.92	36.78	40.29	42.80
23	9.26	10.20	11.69	13.09	14.85	32.01	35.17	38.08	41.64	44.18
24	9.89	10.86	12.40	13.85	15.66	33.20	36.42	39.36	42.98	45.56
25	10.52	11.52	13.12	14.61	16.47	34.38	37.65	40.65	44.31	46.93
26	11.16	12.20	13.84	15.38	17.29	35.56	38.89	41.92	45.64	48.29
27	11.81	12.88	14.57	16.15	18.11	36.74	40.11	43.19	46.96	49.64
28	12.46	13.56	15.31	16.93	18.94	37.92	41.34	44.46	48.28	50.99
29	13.12	14.26	16.05	17.71	19.77	39.09	42.56	45.72	49.59	52.34
30	13.79	14.95	16.79	18.49	20.60	40.26	43.77	46.98	50.89	53.67
40	20.71	22.16	24.43	26.51	29.05	51.81	55.76	59.34	63.69	66.77
50	27.99	29.71	32.36	34.76	37.69	63.17	67.50	71.42	76.15	79.49
60	35.53	37.48	40.48	43.19	46.46	74.40	79.08	83.30	88.38	91.95
70	43.28	45.44	48.76	51.74	55.33	85.53	90.53	95.02	100.43	104.21
80	51.17	53.54	57.15	60.39	64.28	96.58	101.88	106.63	112.33	116.32
90	59.20	61.75	65.65	69.13	73.29	107.57	113.15	118.14	124.12	128.30

Tabla 1.4: Algunos percentiles de la distribución χ^2 .

n	m									
	1	2	3	4	5	6	7	8	9	10
1	39.86	49.50	53.59	55.83	57.24	58.20	58.91	59.44	59.86	60.19
2	8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.38	9.39
3	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.24	5.23
4	4.54	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.94	3.92
5	4.06	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.32	3.30
6	3.78	3.46	3.29	3.18	3.11	3.05	3.01	2.98	2.96	2.94
7	3.59	3.26	3.07	2.96	2.88	2.83	2.78	2.75	2.72	2.70
8	3.46	3.11	2.92	2.81	2.73	2.67	2.62	2.59	2.56	2.54
9	3.36	3.01	2.81	2.69	2.61	2.55	2.51	2.47	2.44	2.42
10	3.29	2.92	2.73	2.61	2.52	2.46	2.41	2.38	2.35	2.32
11	3.23	2.86	2.66	2.54	2.45	2.39	2.34	2.30	2.27	2.25
12	3.18	2.81	2.61	2.48	2.39	2.33	2.28	2.24	2.21	2.19
13	3.14	2.76	2.56	2.43	2.35	2.28	2.23	2.20	2.16	2.14
14	3.10	2.73	2.52	2.39	2.31	2.24	2.19	2.15	2.12	2.10
15	3.07	2.70	2.49	2.36	2.27	2.21	2.16	2.12	2.09	2.06
16	3.05	2.67	2.46	2.33	2.24	2.18	2.13	2.09	2.06	2.03
17	3.03	2.64	2.44	2.31	2.22	2.15	2.10	2.06	2.03	2.00
18	3.01	2.62	2.42	2.29	2.20	2.13	2.08	2.04	2.00	1.98
19	2.99	2.61	2.40	2.27	2.18	2.11	2.06	2.02	1.98	1.96
20	2.97	2.59	2.38	2.25	2.16	2.09	2.04	2.00	1.96	1.94
21	2.96	2.57	2.36	2.23	2.14	2.08	2.02	1.98	1.95	1.92
22	2.95	2.56	2.35	2.22	2.13	2.06	2.01	1.97	1.93	1.90
23	2.94	2.55	2.34	2.21	2.11	2.05	1.99	1.95	1.92	1.89
24	2.93	2.54	2.33	2.19	2.10	2.04	1.98	1.94	1.91	1.88
25	2.92	2.53	2.32	2.18	2.09	2.02	1.97	1.93	1.89	1.87
26	2.91	2.52	2.31	2.17	2.08	2.01	1.96	1.92	1.88	1.86
27	2.90	2.51	2.30	2.17	2.07	2.00	1.95	1.91	1.87	1.85
28	2.89	2.50	2.29	2.16	2.06	2.00	1.94	1.90	1.87	1.84
29	2.89	2.50	2.28	2.15	2.06	1.99	1.93	1.89	1.86	1.83
30	2.88	2.49	2.28	2.14	2.05	1.98	1.93	1.88	1.85	1.82
40	2.84	2.44	2.23	2.09	2.00	1.93	1.87	1.83	1.79	1.76
50	2.81	2.41	2.20	2.06	1.97	1.90	1.84	1.80	1.76	1.73
60	2.79	2.39	2.18	2.04	1.95	1.87	1.82	1.77	1.74	1.71
70	2.78	2.38	2.16	2.03	1.93	1.86	1.80	1.76	1.72	1.69
80	2.77	2.37	2.15	2.02	1.92	1.85	1.79	1.75	1.71	1.68
90	2.76	2.36	2.15	2.01	1.91	1.84	1.78	1.74	1.70	1.67

Tabla 1.5: Algunos percentiles 0.9 de la distribución F .

n	m									
	12	14	16	18	20	24	30	40	60	100
1	60.71	61.07	61.35	61.57	61.74	62.00	62.26	62.53	62.79	63.01
2	9.41	9.42	9.43	9.44	9.44	9.45	9.46	9.47	9.47	9.48
3	5.22	5.20	5.20	5.19	5.18	5.18	5.17	5.16	5.15	5.14
4	3.90	3.88	3.86	3.85	3.84	3.83	3.82	3.80	3.79	3.78
5	3.27	3.25	3.23	3.22	3.21	3.19	3.17	3.16	3.14	3.13
6	2.90	2.88	2.86	2.85	2.84	2.82	2.80	2.78	2.76	2.75
7	2.67	2.64	2.62	2.61	2.59	2.58	2.56	2.54	2.51	2.50
8	2.50	2.48	2.45	2.44	2.42	2.40	2.38	2.36	2.34	2.32
9	2.38	2.35	2.33	2.31	2.30	2.28	2.25	2.23	2.21	2.19
10	2.28	2.26	2.23	2.22	2.20	2.18	2.16	2.13	2.11	2.09
11	2.21	2.18	2.16	2.14	2.12	2.10	2.08	2.05	2.03	2.01
12	2.15	2.12	2.09	2.08	2.06	2.04	2.01	1.99	1.96	1.94
13	2.10	2.07	2.04	2.02	2.01	1.98	1.96	1.93	1.90	1.88
14	2.05	2.02	2.00	1.98	1.96	1.94	1.91	1.89	1.86	1.83
15	2.02	1.99	1.96	1.94	1.92	1.90	1.87	1.85	1.82	1.79
16	1.99	1.95	1.93	1.91	1.89	1.87	1.84	1.81	1.78	1.76
17	1.96	1.93	1.90	1.88	1.86	1.84	1.81	1.78	1.75	1.73
18	1.93	1.90	1.87	1.85	1.84	1.81	1.78	1.75	1.72	1.70
19	1.91	1.88	1.85	1.83	1.81	1.79	1.76	1.73	1.70	1.67
20	1.89	1.86	1.83	1.81	1.79	1.77	1.74	1.71	1.68	1.65
21	1.87	1.84	1.81	1.79	1.78	1.75	1.72	1.69	1.66	1.63
22	1.86	1.83	1.80	1.78	1.76	1.73	1.70	1.67	1.64	1.61
23	1.84	1.81	1.78	1.76	1.74	1.72	1.69	1.66	1.62	1.59
24	1.83	1.80	1.77	1.75	1.73	1.70	1.67	1.64	1.61	1.58
25	1.82	1.79	1.76	1.74	1.72	1.69	1.66	1.63	1.59	1.56
26	1.81	1.77	1.75	1.72	1.71	1.68	1.65	1.61	1.58	1.55
27	1.80	1.76	1.74	1.71	1.70	1.67	1.64	1.60	1.57	1.54
28	1.79	1.75	1.73	1.70	1.69	1.66	1.63	1.59	1.56	1.53
29	1.78	1.75	1.72	1.69	1.68	1.65	1.62	1.58	1.55	1.52
30	1.77	1.74	1.71	1.69	1.67	1.64	1.61	1.57	1.54	1.51
40	1.71	1.68	1.65	1.62	1.61	1.57	1.54	1.51	1.47	1.43
50	1.68	1.64	1.61	1.59	1.57	1.54	1.50	1.46	1.42	1.39
60	1.66	1.62	1.59	1.56	1.54	1.51	1.48	1.44	1.40	1.36
70	1.64	1.60	1.57	1.55	1.53	1.49	1.46	1.42	1.37	1.34
80	1.63	1.59	1.56	1.53	1.51	1.48	1.44	1.40	1.36	1.32
90	1.62	1.58	1.55	1.52	1.50	1.47	1.43	1.39	1.35	1.30

Tabla 1.6: Algunos percentiles 0.9 de la distribución F .

n	m									
	1	2	3	4	5	6	7	8	9	10
1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54	241.88
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08
50	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07	2.03
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99
70	3.98	3.13	2.74	2.50	2.35	2.23	2.14	2.07	2.02	1.97
80	3.96	3.11	2.72	2.49	2.33	2.21	2.13	2.06	2.00	1.95
90	3.95	3.10	2.71	2.47	2.32	2.20	2.11	2.04	1.99	1.94

Tabla 1.7: Algunos percentiles 0.95 de la distribución F .

n	m									
	12	14	16	18	20	24	30	40	60	100
1	243.91	245.36	246.46	247.32	248.01	249.05	250.10	251.14	252.20	253.04
2	19.41	19.42	19.43	19.44	19.45	19.45	19.46	19.47	19.48	19.49
3	8.74	8.71	8.69	8.67	8.66	8.64	8.62	8.59	8.57	8.55
4	5.91	5.87	5.84	5.82	5.80	5.77	5.75	5.72	5.69	5.66
5	4.68	4.64	4.60	4.58	4.56	4.53	4.50	4.46	4.43	4.41
6	4.00	3.96	3.92	3.90	3.87	3.84	3.81	3.77	3.74	3.71
7	3.57	3.53	3.49	3.47	3.44	3.41	3.38	3.34	3.30	3.27
8	3.28	3.24	3.20	3.17	3.15	3.12	3.08	3.04	3.01	2.97
9	3.07	3.03	2.99	2.96	2.94	2.90	2.86	2.83	2.79	2.76
10	2.91	2.86	2.83	2.80	2.77	2.74	2.70	2.66	2.62	2.59
11	2.79	2.74	2.70	2.67	2.65	2.61	2.57	2.53	2.49	2.46
12	2.69	2.64	2.60	2.57	2.54	2.51	2.47	2.43	2.38	2.35
13	2.60	2.55	2.51	2.48	2.46	2.42	2.38	2.34	2.30	2.26
14	2.53	2.48	2.44	2.41	2.39	2.35	2.31	2.27	2.22	2.19
15	2.48	2.42	2.38	2.35	2.33	2.29	2.25	2.20	2.16	2.12
16	2.42	2.37	2.33	2.30	2.28	2.24	2.19	2.15	2.11	2.07
17	2.38	2.33	2.29	2.26	2.23	2.19	2.15	2.10	2.06	2.02
18	2.34	2.29	2.25	2.22	2.19	2.15	2.11	2.06	2.02	1.98
19	2.31	2.26	2.21	2.18	2.16	2.11	2.07	2.03	1.98	1.94
20	2.28	2.22	2.18	2.15	2.12	2.08	2.04	1.99	1.95	1.91
21	2.25	2.20	2.16	2.12	2.10	2.05	2.01	1.96	1.92	1.88
22	2.23	2.17	2.13	2.10	2.07	2.03	1.98	1.94	1.89	1.85
23	2.20	2.15	2.11	2.08	2.05	2.01	1.96	1.91	1.86	1.82
24	2.18	2.13	2.09	2.05	2.03	1.98	1.94	1.89	1.84	1.80
25	2.16	2.11	2.07	2.04	2.01	1.96	1.92	1.87	1.82	1.78
26	2.15	2.09	2.05	2.02	1.99	1.95	1.90	1.85	1.80	1.76
27	2.13	2.08	2.04	2.00	1.97	1.93	1.88	1.84	1.79	1.74
28	2.12	2.06	2.02	1.99	1.96	1.91	1.87	1.82	1.77	1.73
29	2.10	2.05	2.01	1.97	1.94	1.90	1.85	1.81	1.75	1.71
30	2.09	2.04	1.99	1.96	1.93	1.89	1.84	1.79	1.74	1.70
40	2.00	1.95	1.90	1.87	1.84	1.79	1.74	1.69	1.64	1.59
50	1.95	1.89	1.85	1.81	1.78	1.74	1.69	1.63	1.58	1.52
60	1.92	1.86	1.82	1.78	1.75	1.70	1.65	1.59	1.53	1.48
70	1.89	1.84	1.79	1.75	1.72	1.67	1.62	1.57	1.50	1.45
80	1.88	1.82	1.77	1.73	1.70	1.65	1.60	1.54	1.48	1.43
90	1.86	1.80	1.76	1.72	1.69	1.64	1.59	1.53	1.46	1.41

Tabla 1.8: Algunos percentiles 0.95 de la distribución F .

n	m									
	1	2	3	4	5	6	7	8	9	10
1	647.79	799.50	864.16	899.58	921.85	937.11	948.22	956.66	963.28	968.63
2	38.51	39.00	39.17	39.25	39.30	39.33	39.36	39.37	39.39	39.40
3	17.44	16.04	15.44	15.10	14.88	14.73	14.62	14.54	14.47	14.42
4	12.22	10.65	9.98	9.60	9.36	9.20	9.07	8.98	8.90	8.84
5	10.01	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68	6.62
6	8.81	7.26	6.60	6.23	5.99	5.82	5.70	5.60	5.52	5.46
7	8.07	6.54	5.89	5.52	5.29	5.12	4.99	4.90	4.82	4.76
8	7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.36	4.30
9	7.21	5.71	5.08	4.72	4.48	4.32	4.20	4.10	4.03	3.96
10	6.94	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.78	3.72
11	6.72	5.26	4.63	4.28	4.04	3.88	3.76	3.66	3.59	3.53
12	6.55	5.10	4.47	4.12	3.89	3.73	3.61	3.51	3.44	3.37
13	6.41	4.97	4.35	4.00	3.77	3.60	3.48	3.39	3.31	3.25
14	6.30	4.86	4.24	3.89	3.66	3.50	3.38	3.29	3.21	3.15
15	6.20	4.77	4.15	3.80	3.58	3.41	3.29	3.20	3.12	3.06
16	6.12	4.69	4.08	3.73	3.50	3.34	3.22	3.12	3.05	2.99
17	6.04	4.62	4.01	3.66	3.44	3.28	3.16	3.06	2.98	2.92
18	5.98	4.56	3.95	3.61	3.38	3.22	3.10	3.01	2.93	2.87
19	5.92	4.51	3.90	3.56	3.33	3.17	3.05	2.96	2.88	2.82
20	5.87	4.46	3.86	3.51	3.29	3.13	3.01	2.91	2.84	2.77
21	5.83	4.42	3.82	3.48	3.25	3.09	2.97	2.87	2.80	2.73
22	5.79	4.38	3.78	3.44	3.22	3.05	2.93	2.84	2.76	2.70
23	5.75	4.35	3.75	3.41	3.18	3.02	2.90	2.81	2.73	2.67
24	5.72	4.32	3.72	3.38	3.15	2.99	2.87	2.78	2.70	2.64
25	5.69	4.29	3.69	3.35	3.13	2.97	2.85	2.75	2.68	2.61
26	5.66	4.27	3.67	3.33	3.10	2.94	2.82	2.73	2.65	2.59
27	5.63	4.24	3.65	3.31	3.08	2.92	2.80	2.71	2.63	2.57
28	5.61	4.22	3.63	3.29	3.06	2.90	2.78	2.69	2.61	2.55
29	5.59	4.20	3.61	3.27	3.04	2.88	2.76	2.67	2.59	2.53
30	5.57	4.18	3.59	3.25	3.03	2.87	2.75	2.65	2.57	2.51
40	5.42	4.05	3.46	3.13	2.90	2.74	2.62	2.53	2.45	2.39
50	5.34	3.97	3.39	3.05	2.83	2.67	2.55	2.46	2.38	2.32
60	5.29	3.93	3.34	3.01	2.79	2.63	2.51	2.41	2.33	2.27
70	5.25	3.89	3.31	2.97	2.75	2.59	2.47	2.38	2.30	2.24
80	5.22	3.86	3.28	2.95	2.73	2.57	2.45	2.35	2.28	2.21
90	5.20	3.84	3.26	2.93	2.71	2.55	2.43	2.34	2.26	2.19

Tabla 1.9: Algunos percentiles 0.975 de la distribución F .

n	m									
	12	14	16	18	20	24	30	40	60	100
1	976.71	982.53	986.92	990.35	993.10	997.25	1001.41	1005.60	1009.80	1013.17
2	39.41	39.43	39.44	39.44	39.45	39.46	39.46	39.47	39.48	39.49
3	14.34	14.28	14.23	14.20	14.17	14.12	14.08	14.04	13.99	13.96
4	8.75	8.68	8.63	8.59	8.56	8.51	8.46	8.41	8.36	8.32
5	6.52	6.46	6.40	6.36	6.33	6.28	6.23	6.18	6.12	6.08
6	5.37	5.30	5.24	5.20	5.17	5.12	5.07	5.01	4.96	4.92
7	4.67	4.60	4.54	4.50	4.47	4.41	4.36	4.31	4.25	4.21
8	4.20	4.13	4.08	4.03	4.00	3.95	3.89	3.84	3.78	3.74
9	3.87	3.80	3.74	3.70	3.67	3.61	3.56	3.51	3.45	3.40
10	3.62	3.55	3.50	3.45	3.42	3.37	3.31	3.26	3.20	3.15
11	3.43	3.36	3.30	3.26	3.23	3.17	3.12	3.06	3.00	2.96
12	3.28	3.21	3.15	3.11	3.07	3.02	2.96	2.91	2.85	2.80
13	3.15	3.08	3.03	2.98	2.95	2.89	2.84	2.78	2.72	2.67
14	3.05	2.98	2.92	2.88	2.84	2.79	2.73	2.67	2.61	2.56
15	2.96	2.89	2.84	2.79	2.76	2.70	2.64	2.59	2.52	2.47
16	2.89	2.82	2.76	2.72	2.68	2.63	2.57	2.51	2.45	2.40
17	2.82	2.75	2.70	2.65	2.62	2.56	2.50	2.44	2.38	2.33
18	2.77	2.70	2.64	2.60	2.56	2.50	2.44	2.38	2.32	2.27
19	2.72	2.65	2.59	2.55	2.51	2.45	2.39	2.33	2.27	2.22
20	2.68	2.60	2.55	2.50	2.46	2.41	2.35	2.29	2.22	2.17
21	2.64	2.56	2.51	2.46	2.42	2.37	2.31	2.25	2.18	2.13
22	2.60	2.53	2.47	2.43	2.39	2.33	2.27	2.21	2.14	2.09
23	2.57	2.50	2.44	2.39	2.36	2.30	2.24	2.18	2.11	2.06
24	2.54	2.47	2.41	2.36	2.33	2.27	2.21	2.15	2.08	2.02
25	2.51	2.44	2.38	2.34	2.30	2.24	2.18	2.12	2.05	2.00
26	2.49	2.42	2.36	2.31	2.28	2.22	2.16	2.09	2.03	1.97
27	2.47	2.39	2.34	2.29	2.25	2.19	2.13	2.07	2.00	1.94
28	2.45	2.37	2.32	2.27	2.23	2.17	2.11	2.05	1.98	1.92
29	2.43	2.36	2.30	2.25	2.21	2.15	2.09	2.03	1.96	1.90
30	2.41	2.34	2.28	2.23	2.20	2.14	2.07	2.01	1.94	1.88
40	2.29	2.21	2.15	2.11	2.07	2.01	1.94	1.88	1.80	1.74
50	2.22	2.14	2.08	2.03	1.99	1.93	1.87	1.80	1.72	1.66
60	2.17	2.09	2.03	1.98	1.94	1.88	1.82	1.74	1.67	1.60
70	2.14	2.06	2.00	1.95	1.91	1.85	1.78	1.71	1.63	1.56
80	2.11	2.03	1.97	1.92	1.88	1.82	1.75	1.68	1.60	1.53
90	2.09	2.02	1.95	1.91	1.86	1.80	1.73	1.66	1.58	1.50

Tabla 1.10: Algunos percentiles 0.975 de la distribución F .

Chapter 2

Estimación puntual

2.1 Introducción

REVISAR

Desde la revolución estadística de Pearson y Fisher, la inferencia estadística busca encontrar los valores que parametrizan a la distribución desconocida de los datos. El primer enfoque, propuesto por Pearson, afirmaba que si era posible observar a la variable de interés en todos y cada uno de los individuos de una población, entonces era posible calcular los parámetros de la distribución de la variable de interés; por otro lado, si sólo se tenía acceso a una muestra representativa, entonces era posible calcular una estimación de tales parámetros. Sin embargo, Fisher discrepó de tales argumentos, asumiendo que las observaciones están sujetas a un error de medición y por lo tanto, así se tuviese acceso a toda la población, es imposible calcular los parámetros de la distribución de la variable de interés.

Del planteamiento de Fisher resultaron una multitud de métodos estadísticos para la estimación de los parámetros poblacionales. Es decir, si la distribución de X está parametrizada por $\theta \in \Theta$, con Θ el espacio paramétrico inducido por el comportamiento de la variable de interés, el objetivo de la teoría estadística inferencial es calcular una estimación $\hat{\theta}$ del parámetro θ por medio de los datos observados. En este enfoque, los parámetros se consideran cantidades fijas y constantes. El anterior será el enfoque que seguiremos a lo largo del desarrollo de este libro, aunque el lector debe tener conocimiento de que ése no es el único enfoque¹ que los estadísticos utilizan en términos de inferencia acerca de los parámetros de una distribución.

¹En términos de inferencia estadística existen por lo menos el enfoque clásico, propuesto por Fisher y desarrollado en este libro; el enfoque bayesiano (Gelman, Carlin, Stern & Rubin 2004); el enfoque no paramétrico (Conover n.d.) y el enfoque de inferencia en poblaciones finitas (Gutiérrez 2009).

2.2 Conceptos básicos

Muchos de los estudios estadísticos están enfocados en estudiar característica o características de una población objetiva, por ejemplo, una empresa productora puede estar interesado en conocer el gasto promedio semanal en alimentos de las familias de estrato socioeconómico bajo con el fin de diseñar una estrategia de mercadeo para promover la demanda en el mercado. Es claro que en la ciudad, hay una gran cantidad de familias de este perfil, y por consiguiente resulta prácticamente imposible saber el gasto promedio semanal de cada una de estas familias. En casos como este, la solución es inferir acerca de la característica de la población usando información obtenida de un subconjunto o una muestra de la población, y técnicas de esta rama constituyen la inferencia estadística. A continuación se presentan algunos conceptos básicos para poder estudiar la teoría de la inferencia estadística.

Definición 2.2.1. *Una muestra aleatoria de tamaño n es un conjunto constituyente por n variables aleatorias independientes e idénticamente distribuidas, X_1, \dots, X_n .*

Una muestra aleatoria es utilizada para lograr el objetivo de estimar un parámetro poblacional desconocido, y esto se logra usando una función o funciones de las variables aleatorias de la muestra, conocidas como estadísticas. La definición formal se enuncia a continuación.

Definición 2.2.2. *Una estadística es una función de variables aleatorias de una muestra aleatoria que no contiene parámetros desconocidos.*

Algunas estadísticas comunes basadas en una muestra aleatoria X_1, \dots, X_n son:

- El promedio muestral o la media muestral, definido como $\bar{X} = \sum_{i=1}^n X_i/n$.
- Las varianzas muestrales, definida como $S_n^2 = \sum_{i=1}^n (X_i - \bar{X})^2/n$ y $S_{n-1}^2 = \sum_{i=1}^n (X_i - \bar{X})^2/(n-1)$.
- Mínimo y máximo de la muestra, definidos como $X_{(1)} = \min\{X_1, \dots, X_n\}$ y $X_{(n)} = \max\{X_1, \dots, X_n\}$. Nótese que en general la estadística $X_{(1)}$, al igual que $X_{(n)}$, no es ninguna de las variables X_1, \dots, X_n . Considere un experimento aleatorio con $\Omega = \{a, b, c\}$ y se definen dos variables aleatorias X_1 y X_2 sobre Ω con $X_1(a) = X_1(b) = 1$, $X_1(c) = 2$ y $X_2(a) = 0$, $X_2(b) = X_2(c) = 2$. Si denotamos al mínimo de X_1 y X_2 como Y , se puede ver que $Y(a) = 0$, $Y(b) = 1$ y $Y(c) = 2$, y claramente Y no es ninguna de las variables X_1 y X_2 . De esta forma si la función de densidad de donde proviene la muestra es f_X , la función de densidad de $X_{(1)}$ y la de $X_{(n)}$ no

corresponden a f_X . Veamos

$$\begin{aligned} P(X_{(1)} > x) &= P(X_1 > x, \dots, X_n > x) \\ &= P(X_1 > x) \cdots P(X_n > x) \\ &= (1 - F_{X_1}(x)) \cdots (1 - F_{X_n}(x)) \\ &= (1 - F_X(x))^n \end{aligned}$$

donde F_X denota la función de distribución común de la muestra. Por otro lado, $P(X_{(1)} > x) = 1 - F_{X_{(1)}}(x)$, de esta forma, tenemos que

$$F_{X_{(1)}}(x) = 1 - (1 - F_X(x))^n,$$

y podemos hallar la función de densidad para encontrar la función de densidad de $X_{(1)}$ como

$$f_{X_{(1)}}(x) = n f_X(x) (1 - F_X(x))^{n-1}. \quad (2.2.1)$$

Usando lo anterior, podemos encontrar que la función de densidad de $X_{(1)}$ en una muestra con distribución exponencial de media θ , ésta corresponde a

$$f_{X_{(1)}}(x) = \frac{n}{\theta} e^{-nx/\theta} I_{(0,\infty)}(x).$$

Usando un razonamiento análogo, se puede ver que para la estadística máximo $X_{(n)}$, se tiene que

$$F_{X_{(n)}}(x) = F_X(x)^n \quad (2.2.2)$$

y

$$f_{X_{(n)}}(x) = n f_X(x) F_X(x)^{n-1}. \quad (2.2.3)$$

Una vez definido el concepto de estadística, podemos definir lo que es un estimador. Aunque en muchos contextos, lo que se desea estimar es el parámetro de una distribución, θ , en algunos casos, lo que nos interesa es una función del parámetro $g(\theta)$. Por ejemplo, suponga que en la línea de atención al cliente de una empresa, para describir el tiempo de espera para ser atendido por un asesor se emplea una distribución $Exp(\theta)$. En este caso, θ describe el tiempo de espera promedio que es el parámetro de interés; otra cantidad que puede resultar interesante es por ejemplo, la probabilidad de que un cliente tenga que esperar menos de 5 minutos antes de ser atendida, la cual es $1 - e^{-5/\theta}$, que es una función del parámetro θ . Por esta razón, a continuación se presenta la definición general para un estimador de $g(\theta)$.

Definición 2.2.3. Una estadística T cuyos valores son utilizados para estimar una función del parámetro $g(\theta)$ es un estimador de $g(\theta)$ y las realizaciones del estimador se llaman estimaciones y se denotan por t .

Nótese que un estimador es una función de variables aleatorias, de tal manera, que cuando las variables aleatorias se cambian de valor, el estimador también. Por lo tanto, cuando la muestra aleatoria cambia, el valor que toma el estimador,

es decir, la estimación también cambia. Por lo tanto, un mismo estimador puede producir diferentes estimaciones si se cambia la muestra aleatoria. Y de lo anterior, debe quedar claro que un estimador es aleatorio, mientras que una estimación es un número, puesto que es la realización numérica del estimador. En la literatura estadística, se acostumbra denotar a los estimadores con letras mayúsculas, y a las estimaciones, minúsculas. De esta forma, las realizaciones de \bar{X} se denotan como \bar{x} , las de S_n^2 como s_n^2 y análogamente para cualquier otro estimador.

Pongamos un ejemplo, suponga que se desea conocer la cantidad promedio de dinero que se gasta semanalmente una familia de estrato 2 en la compra de arroz. Un estadístico A seleccionó una muestra de 20 hogares, y utilizó como estimador el promedio muestral y tuvo como resultado $\bar{x} = 5300$ pesos; ahora, otro estadístico B seleccionó una muestra de 30 personas, y utilizó el mismo estimador, es decir, \bar{X} y tuvo como resultado $\bar{x} = 4700$. En este caso, los dos usaron el mismo estimador, sin embargo, las dos estimaciones que obtuvieron no son iguales.

2.3 Estimaciones puntuales

El tópico de estimación puntual consiste en encontrar estimadores para estimar un cierto parámetro θ o una función de este $g(\theta)$. Dada la definición de estimador, es claro que cualquier estadística puede ser un estimador, y por consiguiente cada persona puede usar cualquier estadística para estimar según su antojo.

Toma el ejemplo de estimar el gasto promedio semanal en alimentos de familias de estrato socioeconómico bajo, si la muestra aleatoria es de tamaño 10, y los valores numéricos (en miles de pesos) son: 50, 62, 53, 65, 70, 64, 60, 58, 62 y 65, una forma razonable de estimar es tomar el promedio muestral \bar{X} como estimador del media poblacional, y arroja como resultado $\bar{x} = 60900$ pesos, y parece ser una estimación aceptable; pero este no es la única forma de estimar, podemos definir otras estadísticas, por ejemplo: $\sum_{i=1}^n X_i$, $(X_{(N)} + X_{(1)})/2$, u otras estadísticas no tan lógicas como $\exp\{X_1 + \cdots + X_n\}$, $\sum_{i=1}^n X_i^2$ o cualquier otra estadística que se nos viene a la mente. Sin embargo, en la literatura estadística existen, por lo menos, dos métodos estándares que nos ayudan a construir estimadores: el método de máxima verosimilitud y el método de momentos que se estudiará a continuación. Los estudiaremos a continuación.

2.3.1 Método de máxima verosimilitud

Suponga que se desea estimar $g(\theta)$ basada en una muestra observada x_1, \dots, x_n . La idea del método de máxima verosimilitud se basa en encontrar el valor de $g(\theta)$ que maximiza la probabilidad de observar la muestra x_1, \dots, x_n . Esto es, el valor de $g(\theta)$ que hace más creíble a la muestra observada, y de allí viene el nombre de máxima verosimilitud.

Introducimos este método con un ejemplo muy sencillo, suponga que la alcaldía local está interesado en conocer el número promedio de homicidios men-

θ	5	6	7	8	9	10
Pr	3.1×10^{-5}	2.1×10^{-4}	6.8×10^{-4}	1.3×10^{-3}	1.5×10^{-3}	1.3×10^{-3}

Tabla 2.1: Probabilidad de observar una muestra 11,9,7 provenientes de una distribución $P(\theta)$ para diferentes valores de θ .

suales ocurridos en la localidad de Usaquén de Bogotá. Dadas las características de esta variable de estudio, se puede pensar que ésta sigue una distribución Poisson con el parámetro desconocido θ . Como θ es la esperanza de la distribución Poisson, entonces lo que interesa a la alcaldía es conocer el valor de θ . Además, suponga que durante los últimos tres meses, el número de homicidio fue: 11, 9 y 7 respectivamente.

En la anterior situación, θ es el parámetro del modelo probabilístico que rige en la población, donde solo están disponibles las realizaciones de tres variables, que al suponer que el tiempo no es un factor importante, constituyen una muestra aleatoria de tamaño 3, y los denotamos por X_1, X_2, X_3 . Ahora podemos hacer la siguiente pregunta: ¿cuál es la probabilidad de que la muestra aleatoria tenga como realización los valores 11, 9 y 7? Es decir, ¿cuál es la probabilidad de observar lo que realmente sucedió? Es claro que esta probabilidad depende del parámetro desconocido θ . Tenemos

Si $\theta = 6$, entonces

$$P(X_1 = 11, X_2 = 9, X_3 = 7) = \frac{e^{-6}6^{11}}{11!} \frac{e^{-6}6^9}{9!} \frac{e^{-6}6^7}{7!} = 2.1 \times 10^{-4}.$$

Si $\theta = 8$, entonces

$$P(X_1 = 11, X_2 = 9, X_3 = 7) = \frac{e^{-8}8^{11}}{11!} \frac{e^{-8}8^9}{9!} \frac{e^{-8}8^7}{7!} = 1.3 \times 10^{-3}.$$

Si calculamos esta misma probabilidad para otros valores de θ , podemos obtener la Tabla 2.1, donde se observa que la probabilidad es más grande cuando $\theta = 9$.² Ahora, como ya se observó los valores 11, 9 y 7, es natural pensar que la probabilidad de asociada a estos valores fuera grande, y esto nos conduce a que el valor más plausible para θ debe ser 9.

El anterior razonamiento induce el método de máxima verosimilitud, para estudiar este método, primero se da la siguiente definición.

Definición 2.3.1. Dadas n variables aleatorias X_1, \dots, X_n , la función de verosimilitud se define como la función de densidad conjunta de las n variables, y se denota por $L(x_1, \dots, x_n, \theta)$.

Aunque en este texto, se trabaja solamente muestras aleatorias, la definición de la función de verosimilitud presentada anteriormente es válida en cualquier

²El parámetro θ puede tomar cualquier valor positivo, no necesariamente entero, aquí se consideró solo algunos valores para θ para introducir el método de máxima verosimilitud, más no es un procedimiento riguroso.

conjunto de variables aleatorias. En particular, cuando las n variables conforman una muestra aleatoria, la función de verosimilitud queda expresado como

$$L(x_1, \dots, x_n, \theta) = f(x_1, \theta) \cdots f(x_n, \theta) = \prod_{i=1}^n f(x_i, \theta)$$

donde f es la función de densidad común para las n variables. También nótese que cuando solo se dispone de una variable aleatoria X , la función de verosimilitud es simplemente la función de densidad de X , esto es, $f_X(x)$.

Dada la definición de la función de verosimilitud, el método de máxima verosimilitud para un parámetro θ consiste en encontrar el valor de θ que maximice esta función, lo denotaremos por $\hat{\theta}_{MV}$. Cuando la función de verosimilitud es una función continua de θ y además derivable, entonces podemos usar la primera y la segunda derivada para encontrar el estimador de máxima verosimilitud. Lo ilustramos con el siguiente ejemplo:

Ejemplo 2.3.1. *Dada una muestra aleatoria X_1, \dots, X_n con distribución $P(\theta)$, el estimador de máxima verosimilitud de θ es el promedio muestral \bar{X} . Para verificar esta afirmación, se calcula primero la función de verosimilitud:*

$$\begin{aligned} L(x_1, \dots, x_n, \theta) &= f(x_1, \theta) \cdots f(x_n, \theta) \\ &= \frac{e^{-\theta} \theta^{x_1}}{x_1!} I_{\{0,1,\dots\}}(x_1) \cdots \frac{e^{-\theta} \theta^{x_n}}{x_n!} I_{\{0,1,\dots\}}(x_n) \\ &= \frac{e^{-n\theta} \theta^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n x_i!} \prod_{i=1}^n I_{\{0,1,\dots\}}(x_i) \end{aligned}$$

Obsérvese que encontrar el valor de θ para maximizar la anterior expresión de $L(x_1, \dots, x_n, \theta)$ es equivalente a maximizar $e^{-n\theta} \theta^{\sum_{i=1}^n x_i}$, pues ésta es la parte que depende de θ . Ahora, encontrar el valor que maximiza una función es equivalente a encontrar el valor que maximiza el logaritmo natural de esta función, pues la función logarítmica es estrictamente creciente. Por lo tanto, basta encontrar el valor de θ que maximiza

$$L'(\theta) = \ln(e^{-n\theta} \theta^{\sum_{i=1}^n x_i}) = -n\theta + \sum_{i=1}^n x_i \ln \theta$$

En la distribución $P(\theta)$, el espacio paramétrico es $(0, \infty)$, y la función $L'(\theta)$ es función derivable, entonces la forma de hallar el máximo de la función será resolver la ecuación $\frac{\partial L'(\theta)}{\partial \theta} = 0$. En este caso, la solución es $\theta = \sum x_i / n$. Ahora se calcula la segunda derivada de $L'(\theta)$ evaluado en la anterior solución, tenemos que

$$\left. \frac{\partial^2 L'(\theta)}{\partial \theta^2} \right|_{\theta = \sum x_i / n} = \left. \frac{-\sum x_i}{\theta^2} \right|_{\theta = \sum x_i / n} = \frac{-n^2}{\sum x_i}$$

las observaciones x_i provienen de distribuciones tipo Poisson, lo cual garantiza que toman valores no negativos, de donde concluimos que la anterior expresión es

negativo, lo cual verifica que la solución hallada $\theta = \sum x_i/n$ es efectivamente el máximo.

En conclusión, el estimador de máxima verosimilitud del parámetro θ de una distribución Poisson es el promedio muestral \bar{X} , y lo anterior también concuerda con el razonamiento, pues θ es el valor esperado o el media poblacional, y es lógico estimar el media poblacional con el promedio muestral.

Veamos una aplicación del anterior estimador.

Ejemplo 2.3.2. Suponga que una organización internacional de derechos humanos necesita conocer el número de muertes violentas que ocurren mensualmente en una determinada ciudad, y para eso se seleccionó 15 de los 63 barrios donde los resultados son 1, 1, 5, 5, 2, 3, 3, 6, 4, 3, 2, 3, 2, 3 y 4. Si denotamos el número de muertes violentas mensuales en un barrio de la ciudad por X , entonces X toma valor en $\{0, 1, \dots\}$, por consiguiente, una distribución apropiada para X puede ser la distribución Poisson(θ). Y por el anterior ejemplo, tenemos que la estimación de máxima verosimilitud para θ es el promedio muestral, esto es, $\hat{\theta}_{MV} = \bar{x} = 3.13$.

Aparte de estimar el parámetro θ que puede ser interpretado como el número promedio de muertes violentas mensuales que ocurren en un barrio de la ciudad, podemos estimar otras cantidades que permiten a la organización tener una mejor idea acerca de la ciudad. Por ejemplo, ¿cuál es la probabilidad de que en un barrio no ocurra ninguna muerte violenta durante un mes? Usando propiedades de la distribución Poisson, tenemos que esta probabilidad es igual a $P(X = 0) = e^{-\theta}$. Dado que ya se obtuvo una estimación para θ , podemos utilizarlo para estimar $e^{-\theta}$ como $e^{-3.1} = 0.045$. Más aún, podemos afirmar que esta estimación es de máxima verosimilitud (la teoría se verá más adelante).

No solo podemos hacer inferencia al nivel de los barrios sino, también al nivel de la ciudad. Dado que ésta está compuesta por 63 barrios, entonces el número de muertes violentas mensuales en la ciudad es la suma de los 63 barrios. Si denotamos con Y_i el número de muertes violentas mensuales en el i -ésimo barrio, entonces $Y = \sum_{i=1}^{63} Y_i$ denota el número de muertes violentas en la ciudad. Y usando el Resultado 1.1.9, podemos ver que $Y \sim P(63\theta)$, por lo tanto, el número promedio mensual de muertes violentas es 63θ , y una estimación de ésta será $63 \times 3.13 \approx 197$. Y podemos usar esta estimación para estimar cantidades como probabilidad de que en un mes en la ciudad ocurra menos de 100 muertes violentas u otras probabilidades de interés.

En muestras provenientes de distribuciones como $Exp(\theta)$ o $Ber(\theta)$, el procedimiento para encontrar el estimador $\hat{\theta}_{MV}$ es similar al ejemplo anterior y se puede ver fácilmente que $\hat{\theta}_{MV} = \bar{X}$, de nuevo se encuentra que el estimador de máxima verosimilitud del media poblacional es el promedio muestral en estas dos distribuciones.

Ejemplo 2.3.3. Suponga que una empresa de EPS para mascotas que cuenta con sedes en diferentes ciudades en Colombia tiene vendedores que hace visita a clientes potenciales, estos son, los hogares que tienen mascota, para ofrecer los productos.

Es claro que para fijar metas de venta, el gerente de la empresa debe conocer el rendimiento de los vendedores y así fijar un número de visitas que éstos deben realizar con el fin de lograr la meta de venta. En este caso, el gerente necesita conocer cuál es la probabilidad de que un vendedor logra obtener una venta exitosa en una visita.

Para economizar los recursos, el gerente hace seguimiento a 18 visitas, y en cada visita denota el éxito con 1 y fracaso con 0, de esta forma la muestra observada está constituida por 18 números de la forma 0 y 1, y por el contexto del problema, podemos identificar la distribución Bernoulli, y así estimar la probabilidad de éxito en cada ensayo usando \bar{X} , que en este caso corresponde al número de éxitos dividido por el número total de visitas. Así que si en las 18 visitas registradas el número de ventas exitosas es 3, la probabilidad estimada de que un vendedor de esta empresa logre una venta exitosa en una visita será $\hat{p}_{MV} = 3/18 \approx 0.167$.

Ahora, suponga que un vendedor en un día promedio realiza 5 visitas, y el gerente está interesado en conocer qué tan probable es que en las 5 visitas el vendedor logre por lo menos una venta exitosa. Si denotamos el número de ventas exitosas en las 5 visitas por X , podemos calcular esta probabilidad como

$$P(X > 1) = 1 - P(X = 0) = 1 - (1 - p)^5.$$

Para encontrar una estimación de esta probabilidad, podemos pensar en usar la estimación de máxima verosimilitud de p encontrada anteriormente, de esta forma, tenemos que una estimación de $P(X_1)$ será $1 - (1 - 0.167)^5 = 0.598$. Análogamente, también podemos estimar la probabilidad de vender un seguro en las cinco visitas. En este caso, esta probabilidad está dada por $5p(1 - p)^4$, que puede ser estimada como $5 * 0.167 * (1 - 0.167)^4 = 0.402$.

La pregunta interesante es ¿se puede afirmar que estas dos últimas estimaciones siguen siendo de máxima verosimilitud? Esta pregunta la responderemos más adelante.

Ejemplo 2.3.4. En muchas aerolínea, se puede comprar tiquetes por medio de llamadas telefónicas atendidas por operadores de la aerolínea. Si un cliente debe esperar mucho tiempo en la línea para ser atendida, es más probable que el cliente desiste, y la aerolínea perdería un cliente potencial. Por lo tanto la aerolínea desea conocer el rendimiento de los operadores que atienden estas llamadas. Para eso, se observa aleatoriamente 20 llamadas y se registró el tiempo transcurrido antes de que fueron atendidas por un operador, estos tiempos en minutos son 0.13, 0.06, 0.50, 0.41, 1.44, 0.60, 0.22, 1.08, 0.78, 0.92, 2.73, 0.83, 0.19, 0.21, 1.75, 0.79, 0.02, 0.05, 2.30 y 1.03. Dado el contexto, se desea estimar el tiempo promedio que debe estar un cliente antes de ser atendido, es decir, el media poblacional. Para eso, necesitamos, en primer lugar, suponer una distribución adecuada para los datos. Los datos a la mano son del tipo continuo, además solo toma valores positivos, de donde podemos proponer una distribución exponencial, gamma o una distribución normal³.

³Aunque una distribución normal toma valores en todos los números reales, pero se concentra alrededor de la media, por lo tanto, una muestra de valores positivos también pueden provenir de la distribución normal.

Una forma de verificar la distribución de los datos es observar el histograma, el cual está dada en la Figura 2.1. donde podemos ver que la forma de las barras se asemejan a la función de densidad de una distribución exponencial. Otra forma de ver la distribución de los datos es usando las gráficas de QQ plot, y los presentamos en la Figura 2.2 para la distribución exponencial y la distribución normal⁴. Podemos ver que una vez más, la distribución exponencial parece ser apropiada para los datos. Entonces el problema se convierte en estimar el parámetro θ de una distribución $Exp(\theta)$, puesto que la esperanza de la distribución es θ . Y como se observó anteriormente, $\hat{\theta}_{MV} = \bar{X}$, podemos tener que la estimación de máxima verosimilitud de θ es $\bar{x} = 0.8$ minuto.

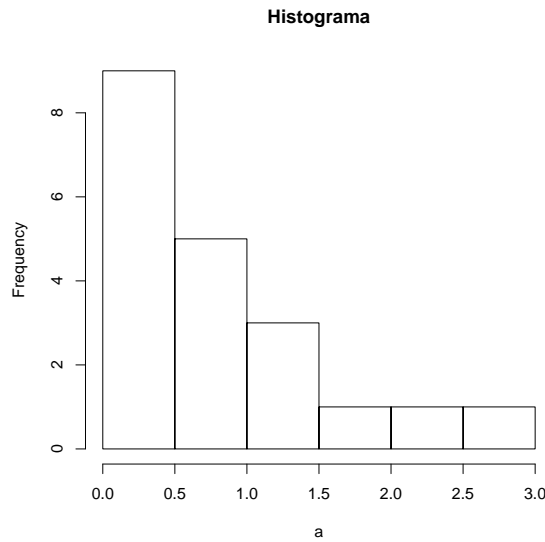


Figure 2.1: Histograma de los datos del Ejemplo 2.2.4.

Ahora suponga que los directivos de la aerolínea ha observado que si un cliente tiene que esperar más de 2 minutos, con toda seguridad cuelga la llamada; el 30% de los clientes que tienen que esperar entre 1 minuto y medio y 2 minutos cuelgan la llamada y ningún cliente cuelga antes del minuto y medio. Entonces podemos estimar el porcentaje de clientes que cuelgan antes de ser atendidos, esto es, clientes potenciales que la aerolínea pierde. Para eso, debe estimar el porcentaje de llamadas que necesitan más de 2 minutos para ser atendidas, ésta se puede expresar como $P(X > 2)$ donde X denota el tiempo de espera de una llamada y $X \sim Exp(\theta)$. Entonces debe estimar $P(X > 2) = e^{-2/\theta}$, el cual es una función de θ , y como ya se ha encontrado una estimación de θ dada por $\hat{\theta}_{MV} = 0.8$, podemos simplemente estimar $P(X > 2)$ como $e^{-2/0.8} = 0.08$, esto es, se estima que el 8%

⁴La inversa de la función de distribución de la distribución Gamma es difícil de hallar y por consiguiente no es posible encontrar el QQ plot para verificar que un conjunto de datos provienen de una distribución Gamma.

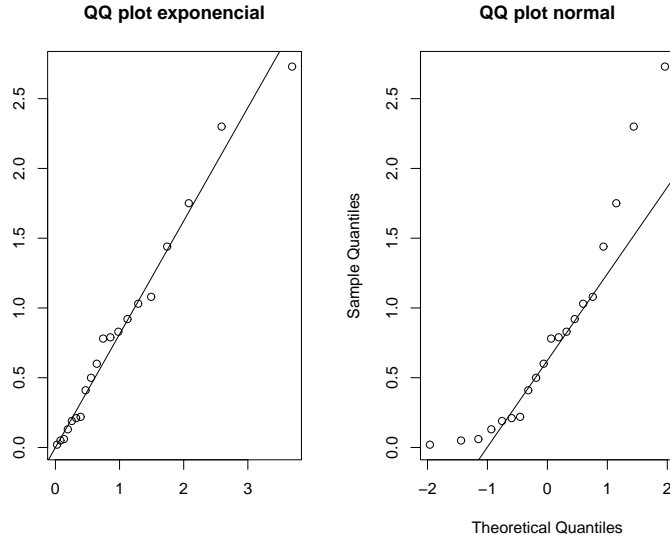


Figure 2.2: QQ plot para verificar la distribución de los datos del Ejemplo 2.2.4.

de llamadas necesitan más de 2 minutos para ser atendidas, y por consiguiente estos 8% de clientes cuelgan antes de ser atendidos. Ahora, para estimar el porcentaje de llamadas que necesitan entre 1 minuto y medio y 2 minutos para ser atendidas como $e^{-1.5/0.8} - e^{-2/0.8} = 0.07$, es decir 7% de llamadas requieren entre 1.5 y 2 minutos para ser atendidas, y por consiguiente $7\% \times 0.3 = 0.021 = 2.1\%$ de clientes cuelgan la llamada antes de ser atendida, y sumando los 8% hallado anteriormente, podemos afirmar que se estima que la aerolínea pierde el 10.1% de los clientes potenciales por no ser atendidos oportunamente. Más adelante, se verá que esta estimación sigue siendo de máxima verosimilitud.

Las distribuciones consideradas anteriormente tienen solo un parámetro desconocido, para distribuciones que tienen dos parámetros desconocidos, el procedimiento es levemente distinto como lo ilustra el siguiente ejemplo con la distribución normal.

Ejemplo 2.3.5. Dada una muestra aleatoria X_1, \dots, X_n con distribución $N(\mu, \sigma^2)$, el estimador de máxima verosimilitud del vector de parámetros $\theta = (\mu, \sigma^2)'$ es $(\bar{X}, S_n^2)'$. Tenemos las siguientes expresiones para la función de verosimilitud:

$$\begin{aligned} L(\theta, x_1, \dots, x_n) &= \frac{1}{\sqrt{\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(x_1 - \mu)^2\right\} \cdots \frac{1}{\sqrt{\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(x_n - \mu)^2\right\} \\ &= (\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2\right\} \end{aligned}$$

Para facilitar la maximización de L , se calcula $\ln(L)$:

$$\ln(L) = -\frac{n}{2} \ln(\pi) - \frac{n}{2} \ln(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2.$$

Ahora para obtener valores de μ y σ^2 que maximicen a $\ln(L)$, se procede a resolver las dos siguientes ecuaciones:

$$\frac{\partial \ln(L)}{\partial \mu} = 0 \quad (2.3.1)$$

y

$$\frac{\partial \ln(L)}{\partial \sigma^2} = 0 \quad (2.3.2)$$

Resolviendo la ecuación (2.3.1), se obtiene la solución de $\mu = \bar{x}$ y resolviendo (2.3.2), se obtiene la solución de $\sigma^2 = \sum (x_i - \mu)^2 / n$, donde al reemplazar $\mu = \bar{x}$, se tiene que $\sigma^2 = \sum (x_i - \bar{x})^2 / n = s_n^2$.

Ahora debemos verificar que las anteriores soluciones halladas efectivamente maximicen la función $\ln(L)$, dado que esta función tiene dos argumentos, es necesario hacer uso de la matriz Hessiana. La matriz se calcula de la siguiente manera:

$$\begin{aligned} H(\ln(L)) &= \begin{bmatrix} \frac{\partial^2 \ln(L)}{\partial \mu^2} & \frac{\partial^2 \ln(L)}{\partial \mu \partial \sigma^2} \\ \frac{\partial^2 \ln(L)}{\partial \sigma^2 \partial \mu} & \frac{\partial^2 \ln(L)}{\partial (\sigma^2)^2} \end{bmatrix} \\ &= \begin{bmatrix} \frac{-n}{\sigma^2} & \frac{n\mu - \sum x_i}{\sigma^4} \\ \frac{n\mu - \sum x_i}{\sigma^4} & \frac{n}{2\sigma^4} - \frac{\sum (x_i - \mu)^2}{\sigma^6} \end{bmatrix}. \end{aligned}$$

Ahora reemplazamos las soluciones halladas $\mu = \bar{x}$ y $\sigma^2 = s_n^2$, se tiene que la matriz Hessiana es:

$$H = \begin{bmatrix} \frac{-n^2}{\sum (x_i - \bar{x})^2} & 0 \\ 0 & \frac{-n^3}{2(\sum (x_i - \bar{x})^2)^2} \end{bmatrix}.$$

Obsérvese que la matriz Hessiana es una matriz diagonal con valores negativos en la diagonal, lo cual demuestra que es definida negativa, con eso se concluye que las soluciones halladas efectivamente maximizan la función $\ln(L)$. En conclusión los estimadores de máxima verosimilitud del vector de parámetros $\theta = (\mu, \sigma^2)'$ es $(\bar{X}, S_n^2)'$.

A continuación, se presenta una aplicación del anterior ejemplo.

Ejemplo 2.3.6. Suponga que una fábrica de vidrios tiene una línea de producción de láminas de vidrio templado de grosor de 3 cm, para controlar la calidad de los vidrios producidos por esta línea, se selecciona 12 láminas para inspección. Estas 12 láminas midieron (en cm) 3.56, 3.36, 2.99, 2.71, 3.31, 3.68, 2.78, 2.95, 2.82, 3.45, 3.42 y 3.15. Estos datos son, aparentemente continuos, y podemos pensar que una distribución normal puede ser apropiado para los datos. Podemos, en primer lugar, observar la forma del histograma de estos datos presentado la Figura 2.3, donde aparentemente no se observa una forma similar a la función de densidad de una distribución normal.

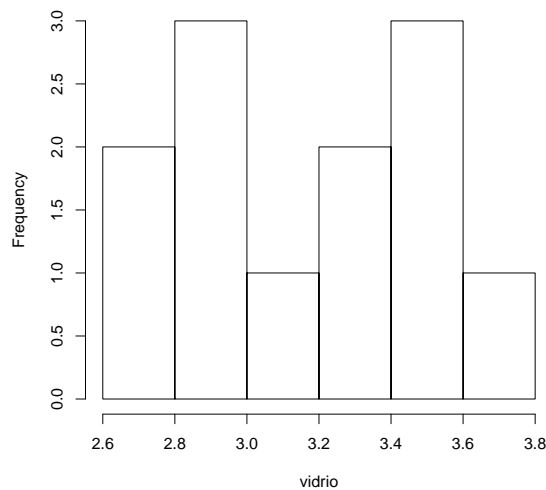


Figure 2.3: Histograma de los datos del Ejemplo 2.2.5.

Sin embargo, como el número de datos es relativamente pequeño, el histograma puede no reflejar la distribución verdadera de los datos, y por esta razón, usamos la gráfica de QQ plot para ver qué tan adecuado es la distribución normal. El comando en R está dado a continuación

```
> vidrio<-c(3.56, 3.36, 2.99, 2.71, 3.31,3.68, 2.78, 2.95,
2.82, 3.45, 3.42 ,3.15)
> qqnorm(vidrio,main="QQ plot para distribución normal",xlab=
"Cuantiles teoricos",ylab="Cuantiles muestrales")
> qqline(vidrio)
```

Esta gráfica se muestra en la Figura 2.4 donde podemos ver que una distribución normal parece ser apropiada. Por lo tanto, usando el anterior ejemplo, podemos estimar el grosor promedio de las láminas de esta línea como $\hat{\mu}_{MV} = \bar{x} = 3.18$ cm y la varianza estimada en este caso es $\hat{\sigma}_{MV}^2 = s_n^2 = 0.097$ cm². Sin

embargo, es difícil dar interpretación práctica a la varianza puesto que la unidad de ésta es la unidad de los datos al cuadrado, por esta razón en la práctica se usa con más frecuencia σ como la medida de dispersión, en este caso, tenemos que $\hat{\sigma} = \sqrt{0.097 \text{ cm}^2} = 0.31 \text{ cm}$.

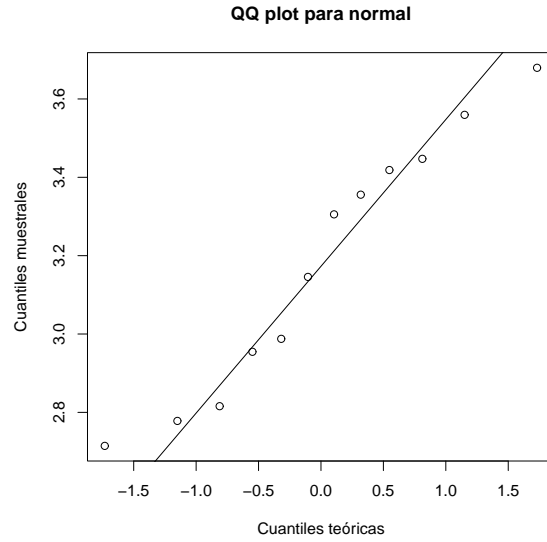


Figure 2.4: QQ plot para verificar la distribución normal de los datos del Ejemplo 2.2.5.

Otra cantidad interesante que se quiere conocer es σ/μ , que puede ser visto como una medida de dispersión poblacional que está libre de las unidades de medición y por consiguiente es útil en la práctica para comparar dos poblaciones. Esta cantidad se puede ver como una función del vector de parámetros (μ, σ^2) , y por consiguiente, puede ser estimada por $\frac{\sqrt{S_n^2}}{\bar{X}}$ que en este ejemplo da como resultado 9.7%.

Ahora, suponga que las láminas de grosor entre 2.8 cm y 3.2 cm son vendidos al mercado, las de grosor menor de 2.8 cm son desechados y las de grosor mayor de 3.2 son usados como materia prima para futuras producciones. Usando las estimaciones de μ y σ podemos estimar las proporciones de láminas que serán vendidas, desechadas y usadas como materia prima. Si denotamos el grosor de una lámina como X , tenemos que la proporciones de láminas que serán vendidas es igual a

$$\begin{aligned} P(2.8 < X < 3.2) &= P\left(\frac{2.8 - \mu}{\sigma} < \frac{X - 3.18}{0.31} < \frac{3.2 - \mu}{\sigma}\right) \\ &= \Phi\left(\frac{3.2 - \mu}{\sigma}\right) - \Phi\left(\frac{2.8 - \mu}{\sigma}\right). \end{aligned}$$

Usando $\hat{\mu}_{MV} = 3.18$ y $\hat{\sigma} = 0.31$, podemos estimar la proporción de láminas que serán vendidas como $\Phi\left(\frac{3.2 - 3.18}{0.31}\right) - \Phi\left(\frac{2.8 - 3.18}{0.31}\right)$ el cual es igual a 0.42, es decir, se estima que solo el 42% de las láminas producidas serán vendidas.

Análogamente, se puede encontrar que el 11% serán desechadas y el 47% serán usadas como materia prima. Es claro que según los datos muestrales y las estimaciones obtenidas de estos, el uso de esta línea de producción no parece ser muy rentable, puesto que menos que la mitad de las láminas producidas pueden ser vendidas. Para aumentar la proporción de láminas que son aptos para la venta, la fábrica debe mejorar la línea de producción con la ayuda de los expertos para

- disminuir el grosor promedio de las láminas, puesto que en la muestra se observó un promedio de 3.18 cm, y se podrá pensar que el grosor real de las láminas es superior al valor especificado de 3 cm. Suponga que después de una mejora de la línea de producción, el promedio muestral fuera $\bar{x} = 3.05$ y $\hat{\sigma}$ se mantiene igual, se puede ver que en este caso, la proporción estimada de láminas para venta se aumentará a 48%.
- estabilizar las láminas en término del grosor, de esta forma, la estimación de σ será más pequeña y la proporción de láminas para venta se incrementará. Suponga que después de una mejora de la línea de producción, $\hat{\sigma} = 0.2$ cm y $\hat{\mu}$ se mantiene igual, se puede ver que la proporción estimada de láminas para venta se aumentará a 51%.

Finalmente, si se puede lograr que μ sea más cercano a 3 cm y al mismo disminuir el valor de σ , la proporción de láminas para venta será aún mayor, y la línea de producción será más rentable.

En el anterior ejemplo, el estimador de máxima verosimilitud de μ es \bar{X} , y esto es válido aún cuando la varianza poblacional σ^2 es conocida; por otro lado, cuando μ es conocido, el estimador de máxima verosimilitud de σ^2 ya no es S_n^2 sino $\sum_{i=1}^n (X_i - \mu)^2 / n$, esto es, se mide la dispersión tomando las diferencia entre cada variable con la media poblacional μ (Ejercicio XXXXXX).

El problema de maximizar una función puede, en algunos casos, resultar complicado, y peor aun, puede no encontrar solución explícita, un ejemplo de estos es la distribución gamma cuando ambos parámetros son desconocidos. Para estos casos, es necesario usar métodos numéricos para encontrar el máximo de la función de verosimilitud.

Ahora, para las distribuciones con dos parámetros como normal o gamma, cuando uno de los dos parámetros es fijo conocido, entonces solo habrá necesidad de estimar el otro parámetro y el procedimiento es similar al presentado anteriormente, y lo ilustramos en el siguiente ejemplo.

Ejemplo 2.3.7. Dada una muestra aleatoria X_1, \dots, X_n con distribución Gamma con parámetro de forma k conocido y parámetro de escala θ desconocido, se tiene que el estimador de máxima verosimilitud de θ es $\frac{\sum_{i=1}^n X_i}{nk}$. Para la verificación,

primero calculamos $\ln(L)$ que es la función que se necesita maximizar:

$$\begin{aligned}\ln(L) &= \ln \left(\frac{\prod x_i^{k-1} e^{-\sum x_i/\theta}}{\theta^{nk} \Gamma(k)^n} \right) \\ &= (k-1) \sum \ln(x_i) - \sum x_i/\theta - nk \ln \theta - n \ln(\Gamma(k))\end{aligned}$$

cuya primera derivada parcial con respecto a θ está dada por:

$$\frac{\partial \ln(L)}{\partial \theta} = \frac{\sum x_i}{\theta^2} - \frac{nk}{\theta},$$

el cual al igualar a 0, se obtiene la solución de $\theta = \frac{\sum x_i}{nk}$. Ahora, al calcular la segunda derivada de $\ln(L)$ y evaluar en la anterior solución, se tiene que

$$\frac{\partial^2 \ln(L)}{\partial \theta^2} = \frac{-nk}{\theta^2} < 0,$$

con lo cual se concluye que el estimador de máxima verosimilitud de θ es $\frac{\sum X_i}{nk}$.

Teniendo en cuenta que la distribución exponencial es un caso particular de la distribución Gamma cuando $k = 1$, entonces del anterior ejemplo se puede concluir que el estimador de máxima verosimilitud del parámetro θ de una distribución exponencial es $\bar{X} = \frac{\sum X_i}{n}$, estimador que también se puede obtener maximizando directamente la función de verosimilitud.

Cuando la función de verosimilitud no es función continua o derivable del parámetro θ , el problema de maximizar no se puede llevar a cabo haciendo el uso de la derivada de la manera habitual, los dos siguientes ejemplos ilustran tales situaciones.

Ejemplo 2.3.8. *En situaciones donde el característica de interés es el tamaño de una población N , puede ser por ejemplo, la cantidad de cierto tipo de animales en un determinado lugar (puede ser un bosque). Una forma de determinar N es en primer lugar identificar R de los N individuos ($R < N$), en el caso de los animales, puede ser conveniente capturar R de ellos y marcarlos de alguna forma. Después de eso, se espera que los R individuos se mezclen bien con los otros $N - R$, y se seleccionan aleatoriamente n individuos ($n < N$), y se cuenta cuántos de los R individuos fueron seleccionados. Para ver cómo este procedimiento nos puede ayudar a encontrar el valor de N , primero identifiquemos el contexto en término de las distribuciones probabilísticas.*

Sea X la variable aleatoria que denota el número de los R individuos que fueron seleccionados en la muestra de tamaño n , entonces podemos ver que $X \sim Hg(n, R, N)$ donde n y R son conocidos, y se quiere encontrar el estimador de

máxima verosimilitud de N . En este caso la función de verosimilitud es la misma función de densidad de la variable X , esto es:

$$L(N, x) = \frac{\binom{R}{x} \binom{N-R}{n-x}}{\binom{N}{n}} I_{\{0,1,\dots,n\}}(x).$$

En esta función, el argumento N toma valores discretos y no se puede derivar L con respecto a N para hallar el máximo. La forma de encontrar el valor de N que maximiza a L es encontrar para qué valores de N , la función L es creciente y para qué valores de N es decreciente. Para lograr este fin, se despeja el valor de N en $L(N)/L(N-1) > 1$, como sigue:

$$\begin{aligned} \frac{L(N)}{L(N-1)} &= \frac{\binom{N-R}{n-x} \binom{N-1}{n}}{\binom{N}{n} \binom{N-R-1}{n-x}} > 1 \\ \frac{(N-R)!(N-1)!(N-n)!(N-R-1-n+x)!}{(N-R-1)!N!(N-n-1)!(N-R-n+x)!} &> 1 \\ (N-R)(N-n) &> N(N-R-n+x) \\ Rn/x &> N. \end{aligned}$$

Análogamente, se obtiene que $L(N)/L(N-1) < 1$ cuando $N > Rn/x$. Lo anterior indica que la función L es creciente para valores de N menores que Rn/x y decreciente para valores mayores que Rn/x . Pero no podemos afirmar que Rn/X es el estimador de máxima verosimilitud para L puesto que esta cociente puede no ser entero, y en este caso la estimación no ubicaría dentro del espacio paramétrico de N . Lo que sí se puede afirmar es que el estimador de máxima verosimilitud de N es $[Rn/X]$ o $[Rn/X] + 1$ donde $[\cdot]$ es la función parte entera. Shao (2003) afirma que $\hat{N}_{MV} = [Rn/X]$. Esto es verdadero puesto que hemos encontrado anteriormente que $L(N) < L(N-1)$ si $N > Rn/x$, entonces podemos concluir que como $[Rn/x] + 1 > Rn/x$, entonces

$$L([Rn/x] + 1) < L([Rn/x]).$$

Y podemos afirmar que el estimador de máxima verosimilitud de N es $[Rn/X]$.

Otra aplicación interesante de la distribución hipergeométrica es el caso donde se conoce el tamaño poblacional N , y se desea estimar el número de individuos que tienen cierta característica basada en una muestra de tamaño n , por ejemplo, se conoce que en un estanque hay N peces, y se sabe que una parte de ellas están infectados por un tipo de parásito, y quiere saber cuántos peces tienen dicho parásito basado en la observación de una muestra de tamaño n . En estos casos, estamos interesados en estimar R con N y n conocidos. Un razonamiento análogo al caso de estimar N conduce al siguiente estimador de R , (para más detalles, consulte Zhang (2009))

$$\hat{R}_{MV} = \begin{cases} \frac{x(N+1)}{n} - 1 & \text{si } \frac{x(N+1)}{n} \text{ es entero} \\ \left[\frac{x(N+1)}{n} \right] & \text{si } \frac{x(N+1)}{n} \text{ no es entero} \end{cases}$$

Retomando el problema de estimar el tamaño de un subgrupo considerado en el Capítulo 1, donde se supone que en una ciudad existen 2396 empresas que pueden clasificar en empresas grandes, medianas o pequeñas según el número de empleados. Si en una muestra aleatoria simple sin reemplazos de tamaño 200 se encuentran 28 empresas grandes, para tener una estimación del número total de empresas grandes en la ciudad, calculamos en primer lugar $28 * (2396 + 1)/200 = 335.58$, este no es entero, de donde concluimos que la estimación de máxima verosimilitud del número de empresas grandes en la ciudad es de 335.

Finalmente, consideramos las distribuciones donde los valores que toma la variable aleatoria X depende del parámetro θ , por ejemplo, las distribuciones del tipo uniforme. Para este tipo de distribuciones, el procedimiento para encontrar $\hat{\theta}_{MV}$ en general se puede resumir en los siguientes pasos:

- (1) Calcular la función de verosimilitud, sin omitir las funciones indicadoras, pues éstas depende de θ .
- (2) Encontrar el rango de valores de θ donde la función de verosimilitud no toma el valor 0. Generalmente este rango depende del máximo y/o el mínimo de la muestra: $x_{(n)}$ y $x_{(1)}$.
- (3) Dentro del rango encontrado en el paso anterior, mediante empleo de derivadas o simplemente observación directa, buscar el valor de θ que maximice la función de verosimilitud.

Ilustramos el anterior procedimiento en el siguiente ejemplo.

Ejemplo 2.3.9. Dada una muestra aleatoria X_1, \dots, X_n proveniente de una población con distribución uniforme continua sobre el intervalo $[0, \theta]$, se quiere encontrar el estimador de máxima verosimilitud del parámetro θ . En primer lugar, obsérvese que en la función de verosimilitud L existe un término de la función indicadora que depende de θ , puesto que $L(\theta) = \theta^{-n} \prod_{i=1}^n I_{[0, \theta]}(x_i)$.

Tenemos que

$$L \neq 0 \Leftrightarrow 0 \leq x_i \leq \theta \text{ para todo } i = 1, \dots, n \Leftrightarrow \theta \geq x_{(n)}$$

donde $x_{(n)} = \max\{x_1, \dots, x_n\}$. Entonces se concluye que el rango de valores de θ para que la función de verosimilitud sea diferente de 0 es $[x_{(n)}, \infty)$. Ahora, observe que dentro de este rango, $L = \theta^{-n}$, que es una función decreciente de θ , entonces para valores pequeños de θ , L toma valores grandes. Pero el valor más pequeño que puede tomar θ dentro del rango $[x_{(n)}, \infty)$ es $x_{(n)}$, entonces se concluye que $\hat{\theta}_{MV} = X_{(n)}$.

Ahora, retomando situaciones donde la cantidad que se quiere estimar es una función del parámetro, $g(\theta)$. Textos como Mood, Graybill & Boes (1974) establecen que el estimador de máxima verosimilitud de $g(\theta)$ es simplemente $g(\hat{\theta}_{MV})$ siempre y cuando g es una función uno a uno.

Otra forma de ver esto es mediante la reparametrización de la función $L(\theta)$. Por ejemplo, suponga que se desea estimar $\lambda = e^{-\theta}$ en una muestra proveniente de una distribución $P(\theta)$. Podemos escribir la función de verosimilitud en término de λ como

$$L(\lambda) = \frac{\lambda^n (-n \ln \lambda)^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n x_i} \prod_{i=1}^n I_{\{0,1,\dots\}}(x_i),$$

de donde

$$\frac{\partial \ln L(\lambda)}{\partial \lambda} = \frac{n}{\lambda} + \frac{\sum_{i=1}^n x_i}{\ln \lambda} \frac{1}{\lambda},$$

igualando la anterior expresión a 0 y resolviendo para λ , se tiene que $\hat{\lambda}_{MV} = e^{-\bar{X}}$. Ahora recordando que $\hat{\theta}_{MV} = \bar{X}$, lo cual coincide con la conclusión dada anteriormente.

Aunque lo planteado es válido para el caso cuando la función g es una función uno a uno, existe el siguiente resultado que establece la invarianza del estimador de máxima verosimilitud para cualquier función g .

Resultado 2.3.1. *Dada una muestra aleatoria X_1, \dots, X_n proveniente de una población con distribución $f(x, \theta)$ donde θ es el vector de parámetros, y suponga que $T = T(X_1, \dots, X_n)$ es el estimador de máxima verosimilitud de θ , y g es una función del vector de parámetros, entonces el estimador de máxima verosimilitud de $g(\theta)$ es la estadística $g(T)$.*

Proof. Ver Casella&Berger Pg. 172, 321

□

Dado el anterior resultado, podemos ver que todas las estimaciones de los Ejemplos 2.2.2, 2.2.3, 2.2.4 y 2.2.6 son estimaciones de máxima verosimilitud.

Otra situación interesante es cuando se dispone de dos muestras aleatorias independientes, esto es, cualquier conjunto de variables de la primera muestra es independiente de cualquier conjunto de variables de la segunda, y el objetivo es estimar los parámetros concernientes a las dos muestras.

Consideramos, en primer lugar, dos muestras provenientes de distribuciones normales.

Suponga que se tienen dos muestras aleatorias independientes X_1, \dots, X_{n_X} y Y_1, \dots, Y_{n_Y} provenientes de $N(\mu_X, \sigma_X^2)$ y $N(\mu_Y, \sigma_Y^2)$, respectivamente. Y se desean estimar alguno o algunos de los parámetros μ_X, μ_Y, σ_X^2 y σ_Y^2 . En primer lugar, calculamos la función de verosimilitud, ésta está dada por

$$L = (2\pi\sigma_X^2)^{-n_X/2} (2\pi\sigma_Y^2)^{-n_Y/2} \exp \left\{ -\frac{1}{2\sigma_X^2} \sum_{i=1}^{n_X} (x_i - \mu_X)^2 - \frac{1}{2\sigma_Y^2} \sum_{j=1}^{n_Y} (y_j - \mu_Y)^2 \right\}$$

En el caso de que las dos muestras provenientes de la misma distribución, esto es, si $\mu_X = \mu_Y = \mu$ y $\sigma_X^2 = \sigma_Y^2 = \sigma^2$, entonces el proceso de la estimación de

máxima verosimilitud de μ y σ^2 se llevan a cabo, simplemente, usando conjuntamente las variables de las dos muestras. Y tenemos que

$$\hat{\mu}_{MV} = \frac{\sum_{i=1}^{n_X} X_i + \sum_{j=1}^{n_Y} Y_j}{n_X + n_Y}, \quad (2.3.3)$$

y

$$\hat{\sigma}_{MV}^2 = \frac{\sum_{i=1}^{n_X} (X_i - \hat{\mu}_{MV})^2 + \sum_{j=1}^{n_Y} (Y_j - \hat{\mu}_{MV})^2}{n_X + n_Y}.$$

Ejemplo 2.3.10. Retomamos el Ejemplo 2.2.6, donde se disponía una muestra de 12 láminas, ahora suponga que se selecciona una nueva muestra de 10 láminas de la misma línea de producción con grosor 3.56, 3.17, 2.98, 2.95, 3.03, 2.87, 3.58, 3.73, 2.83 y 3.43. Dado que las dos muestras son productos de una misma línea de producción, entonces podemos afirmar que las dos muestras provienen de una misma distribución normal $N(\mu, \sigma^2)$, y podemos estimar el grosor promedio de esta línea como 3.2 cm y la desviación estándar como 0.31 cm.

Otra situación que puede surgir en la práctica es cuando las dos muestras provienen de distribuciones con la misma esperanza, pero diferentes varianzas, esto es, $\mu_X = \mu_Y = \mu$ y $\sigma_X^2 \neq \sigma_Y^2$. Supongamos, en primer lugar, que σ_X^2 y σ_Y^2 son conocidas, y tenemos que la función de verosimilitud está dada por

$$L(\mu) = (2\pi\sigma_X^2)^{-n_X/2} (2\pi\sigma_Y^2)^{-n_Y/2} \exp \left\{ -\frac{1}{2\sigma_X^2} \sum_{i=1}^{n_X} (x_i - \mu)^2 - \frac{1}{2\sigma_Y^2} \sum_{j=1}^{n_Y} (y_j - \mu)^2 \right\}$$

de donde

$$\frac{\partial \ln L(\mu)}{\partial \mu} = \frac{\sum_{i=1}^{n_X} (x_i - \mu)}{\sigma_X^2} + \frac{\sum_{j=1}^{n_Y} (y_j - \mu)}{\sigma_Y^2}.$$

Igualando la anterior derivada a cero y despejando μ , se encuentra que la solución está dada por $\mu = \frac{\sigma_Y^2 n_X \bar{x} + \sigma_X^2 n_Y \bar{y}}{n_X \sigma_Y^2 + n_Y \sigma_X^2}$. Ahora, es claro que

$$\frac{\partial^2 \ln L(\mu)}{\partial \mu^2} = -\frac{n_X}{\sigma_X^2} - \frac{n_Y}{\sigma_Y^2} < 0,$$

y en conclusión, se tiene que

$$\begin{aligned} \hat{\mu}_{MV} &= \frac{\sigma_Y^2 n_X \bar{X} + \sigma_X^2 n_Y \bar{Y}}{n_X \sigma_Y^2 + n_Y \sigma_X^2} \\ &= \frac{n_X \bar{X} + n_Y \bar{Y} \frac{\sigma_X^2}{\sigma_Y^2}}{n_X + n_Y \frac{\sigma_X^2}{\sigma_Y^2}}. \end{aligned} \quad (2.3.4)$$

Nótese que la anterior expresión se asemeja a un promedio ponderado, donde entre más grande sea la varianza poblacional de la segunda población σ_Y^2 , menos peso

tiene las variables de la muestra correspondiente. Esto es muy natural, puesto que en una distribución normal, una varianza grande indica que los valores de la distribución tienden a tomar valores muy alejados a la media. Entonces, si $\sigma_X^2/\sigma_Y^2 < 1$, los valores de las variables Y_1, \dots, Y_{n_Y} tienden a estar más lejos de μ que las variables X_1, \dots, X_{n_X} , lo cual las hacen menos confiables, y por esta razón se les asigna un peso menor. Cuando $\sigma_X^2 = \sigma_Y^2$, la anterior estimación de μ se reduce a (2.3.3).

Ejemplo 2.3.11. *Considera la fábrica vidrios del Ejemplo 2.2.6, y suponga que hay, en total, dos líneas de producción de láminas de vidrio templado de 3 cm, y además por ajuste inapropiado de temperatura, la línea A tiene una desviación estándar del 0.6 cm, mucho mayor que la línea B cuya desviación estándar es del 0.3 cm. Si se desea estimar el grosor promedio de las láminas de vidrio del grosor nominal del 3 cm, se debe seleccionar una muestra de las láminas de la línea A, y una muestra de la línea B. Suponga que el grosor de 10 láminas de cada línea corresponde a 3.80, 2.81, 2.98, 2.97, 3.69, 2.77, 3.08, 2.98, 2.37, 3.00 y 2.87, 3.48, 2.65, 3.38, 2.75, 2.99, 2.81, 2.54, 2.84, 2.79, respectivamente, entonces asignando un peso mayor a las observaciones de la línea B según la teoría expuesta anteriormente, se tiene que $\hat{\mu}_{MV} = 2.955$ cm.*

Finalmente, consideramos el caso donde se supone que las dos varianzas poblacionales coinciden, $\sigma_X^2 = \sigma_Y^2 = \sigma^2$ y $\mu_X \neq \mu_Y$. En este caso, tenemos que los estimadores de máxima verosimilitud de μ_X , μ_Y y σ^2 son \bar{X} , \bar{Y} y $[\sum_{i=1}^{n_X}(X_i - \bar{X})^2 + \sum_{j=1}^{n_Y}(Y_j - \bar{Y})^2]/(n_X + n_Y)$, respectivamente (Ejercicio 16), esto es, las observaciones de las dos muestras se utilizan separadamente para estimar las medias poblacionales, mientras que la varianza se estima usando las muestras conjuntamente con la misma ponderación. En general, cuando no se puede asumir que $\sigma_X^2 = \sigma_Y^2$, los estimadores de máxima verosimilitud de μ_X y μ_Y siguen siendo \bar{X} y \bar{Y} , respectivamente.

Ejemplo 2.3.12. *Suponga que se desea comparar dos institutos de capacitación en término de calificación obtenida por sus respectivos alumnos. Las calificaciones (sobre 100 puntos) para 15 alumnos del centro A es: 75, 87, 83, 73, 74, 88, 88, 74, 64, 92, 73, 87, 91, 83 y 84; y para 13 alumnos del centro B es: 64, 85, 72, 64, 74, 93, 70, 79, 79, 75, 66, 83 y 74. Dado que el objetivo es comparar los dos institutos, no se puede asumir la igualdad entre las dos medias poblacionales, de esta forma, las dos medias poblacionales se estiman mediante las medias muestrales dadas por $\hat{\mu}_A = 81$ y $\hat{\mu}_B = 75$. Observamos que la estimación para el media poblacional del centro A es superior a la del centro B, en capítulo 3 y 4 se estudiará herramientas que nos permite concluir si esta diferencia es significativa o puede considerarse como insignificante dado que las estimaciones no se pueden tratar como si fueran los valores verdaderos de los parámetros.*

En algunas situaciones, se desea comparar dos poblaciones en término de la dispersión tal como lo muestra el siguiente ejemplo.

Ejemplo 2.3.13. *Suponga que se desea estudiar el precio por metro cuadrado de viviendas en el centro del capital de los países Colombia y Ecuador. La información*

disponible para el caso de Colombia basada en 80 viviendas es: $\bar{x} = 700$ (miles de pesos colombianos) y $s_{x,n} = 95$ (miles de pesos colombianos); y para el caso ecuatoriano basado en 50 viviendas es: $\bar{y} = 1023$ (bolívars venezolanos) y $s_{y,m} = 300$ (bolívars venezolanos).

Dados los anteriores datos, no se puede comparar la dispersión de los dos países usando directamente las desviaciones estándares, puesto que éstas tienen unidades diferentes. Una alternativa es calcular los respectivos coeficientes de variación dados por $\rho_x = 95/700 = 13.57\%$ y $\rho_y = 300/1023 = 29.32\%$. Estos coeficientes de variación están libres de unidad de los datos originales y pueden ser usados directamente para comparar la dispersión. Y podemos concluir que el precio de la vivienda del centro del capital de país vecino es mucho más inestable que el caso colombiano.

2.3.2 Método de los momentos

Otro método común para encontrar estimadores de un parámetro es el método de los momentos, para estudiar este método, primero introducimos algunas definiciones útiles.

Definición 2.3.2. Dada una variable aleatoria X , se define el k -ésimo momento de X como $\mu_k = E(X^k)$.

La anterior definición es al nivel poblacional, cuando se dispone de una muestra aleatoria, se definen los momentos muestrales como sigue.

Definición 2.3.3. Dada una muestra aleatoria X_1, \dots, X_n , se define el k -ésimo momento muestral como $M_k = \sum_{i=1}^n X_i^k / n$.

Nótese que dada una muestra aleatoria, los momentos muestrales son variables aleatorias, más aun, son estadísticas. Y podemos utilizarlos para estimar los respectivos momentos poblacionales. Ahora, si logramos escribir a los parámetros desconocidos en término de los momentos poblacionales, podemos obtener fácilmente estimadores de los parámetros simplemente reemplazando los momentos poblacionales por los muestrales. Los estimadores obtenidos de esta manera se llaman estimadores de momentos, se denotará por $\hat{\theta}_{mom}$. En particular, tenemos el siguiente resultado que es válido en muestras provenientes de cualquier distribución.

Resultado 2.3.2. Dada una muestra aleatoria X_1, \dots, X_n con esperanza común μ y varianza común σ^2 , se tiene que $\hat{\mu}_{mom} = \bar{X}$ y $\hat{\sigma}_{mom}^2 = S_n^2$.

Proof. En primer lugar, nótese que la esperanza μ es simplemente el primer momento poblacional, es decir, $\mu = \mu_1$, el cual se estima con el primer momento muestral M_1 , entonces se tiene que $\hat{\mu}_{mom} = M_1 = \bar{X}$.

Ahora, la varianza poblacional σ^2 se puede escribir en término de los dos primeros momentos poblacionales, $\sigma^2 = \mu_2 - (\mu_1)^2$, que se estimará con $M_2 - (M_1)^2$, esto es: $\frac{1}{n} \sum_{i=1}^n X_i^2 - \bar{X}^2$, y con un poco de operación algebraica, se puede ver que ésta es $S_n^2 = \frac{1}{n} \sum_{i=1}^n X_i^2 - \bar{X}^2$. En conclusión, $\hat{\sigma}_{mom}^2 = S_n^2$. \square

Una aplicación inmediata del anterior resultado es en una muestra proveniente de una distribución normal donde los parámetros μ y σ^2 corresponden a la esperanza y la varianza de la distribución.

Ejemplo 2.3.14. *Dada una muestra aleatoria X_1, \dots, X_n proveniente de una distribución $N(\mu, \sigma^2)$, el estimador de momentos de μ y σ^2 es simplemente la media y la varianza muestral: \bar{X} y $S_n^2 = \sum (X_i - \bar{X})^2/n$. Nótese que en este caso, los estimadores de momentos coincide con los de máxima verosimilitud.*

Otra utilidad del Resultado 2.2.2 es la siguiente forma para encontrar un estimador de momentos.

- (a) Cuando hay que estimar un sólo parámetro θ , escribir a θ en término de la media poblacional μ : $\theta = g(\mu)$, y al estimar μ con la media muestral: \bar{X} , se obtiene un estimar de momentos para θ . Esto es, $\hat{\theta}_{mom} = g(M_1)$.⁵ (También se puede escribir θ en término de la varianza poblacional σ^2 , y al estimar σ^2 con S_n^2 , se tiene un estimador de momentos de θ .)
- (b) Cuando hay que estimar dos parámetros θ_1 y θ_2 , escribir a cada uno de ellos en término de la media y la varianza poblacional $\theta_1 = g_1(\mu, \sigma^2)$ y $\theta_2 = g_2(\mu, \sigma^2)$, y al estimar μ y σ^2 con la media muestral \bar{X} y la varianza muestral S_n^2 , se obtiene un estimar de momentos para θ_1 y θ_2 . Esto es, $\hat{\theta}_{1,mom} = g_1(\bar{X}, S_n^2)$ y $\hat{\theta}_{2,mom} = g_2(\bar{X}, S_n^2)$.

Nota: La parte (a) nos ilustra que el estimador de momentos al igual que el estimador de máxima verosimilitud, puede no ser único. Un ejemplo es cuando la muestra aleatoria proviene de la distribución $P(\lambda)$, se sabe que $\lambda = E(X)$, entonces un estimador de momentos de λ es, naturalmente, \bar{X} ; pero también se sabe que $\lambda = Var(X)$, de esta manera, se tiene otro estimador de λ que es S_n^2 .

De esta forma, para los datos del Ejemplo 2.2.2, podemos tener dos estimaciones de momentos para el número promedio de muertes violentas por ciudad: $\bar{x} = 3.13$, la cual coincide con la estimación de máxima verosimilitud, y la otra estimación de momentos corresponde a $s_n^2 = 1.98$, que es muy diferente a la estimación obtenida usando \bar{x} . La pregunta natural ahora es ¿cuál de las dos estimaciones es mejor? es decir, cuál valor se acerca más al valor verdadero de λ . No podemos responder esta pregunta directamente, puesto que no conocemos el valor de λ . En la siguiente sección, se introduce conceptos que nos permiten evaluar la calidad de los estimadores. A pesar de que hasta ahora no tenemos herramientas para escoger entre las dos estimaciones, un simple ejercicio de simulación nos permite escoger de forma empírico. Se simula, en primer lugar, muestras provenientes de una distribución $P(5)$ y $P(15)$ con tamaño de muestra $n = 5, \dots, 300$, y en cada muestra simulada, se calcula las dos estimaciones de momentos \bar{x} y s_n^2 , y se

⁵Este método es válido siempre y cuando se puede escribir al parámetro en término de la media poblacional. En caso donde esto no es posible, por ejemplo en distribuciones donde no existe la media poblacional o ésta no depende del parámetro, se debe recurrir a momentos poblacionales superiores.

observa cuál es más cercano al valor verdadero del parámetro. Los resultados se visualizan en la gráfica superior de la Figura 2.5, donde la línea negra horizontal denota el valor verdadero del parámetro λ . El comando en R de estas simulaciones es como sigue

```
> set.seed(123)
> n<-5:300
> est.mean<-rep(NA,length(n))
> est.var<-rep(NA,length(n))
> for(i in 1:length(n)){
+ x<-rpois(n[i],5)
+ est.mean[i]<-mean(x)
+ est.var[i]<-var(x)*(n[i]-1)/n[i]
+ }
>
> est1.mean<-rep(NA,length(n))
> est1.var<-rep(NA,length(n))
> for(i in 1:length(n)){
+ y<-rpois(n[i],15)
+ est1.mean[i]<-mean(y)
+ est1.var[i]<-var(y)*(n[i]-1)/n[i]
+ }
>
>
> par(mfrow=c(2,1))
> plot(n,est.var,xlab="n",ylab="Estimación",main="Población P(5)",
type="l",col="blue",ylim=c(2,9))
> abline(5,0)
> lines(n,est.mean,col="red")
> legend(200,9.2,c("Media","Varianza"),lty=c(1,1),col=c("red","blue"),box.col=0)
>
> plot(n,est1.var,xlab="n",ylab="Estimación",main="Población P(15)",
type="l",col="blue")
> abline(15,0)
> lines(n,est1.mean,col="red")
> legend(200,45,c("Media","Varianza"),lty=c(1,1),col=c("red","blue"),box.col=0)
```

Se puede ver claramente, de la Figura 2.5, que la media \bar{X} comparado con S_n^2 estima mejor el parámetro de la distribución, puesto que las estimaciones de \bar{X} parecen estar más cercanas del valor de λ en ambas gráficas. Y por consiguiente, podemos intuir que para los datos del Ejemplo 2.2.2, la estimación $\bar{x} = 3.13$ debe ser más cercano al valor verdadero de λ . En la siguiente sección, se discutirá métodos formales acerca de escogencia entre estimadores, y se verá que el estimador \bar{X} tiene mejores propiedades que S_n^2 .

En los siguientes ejemplos, se ilustra la forma general de encontrar estimadores de momentos para distribuciones con dos parámetros desconocidos.

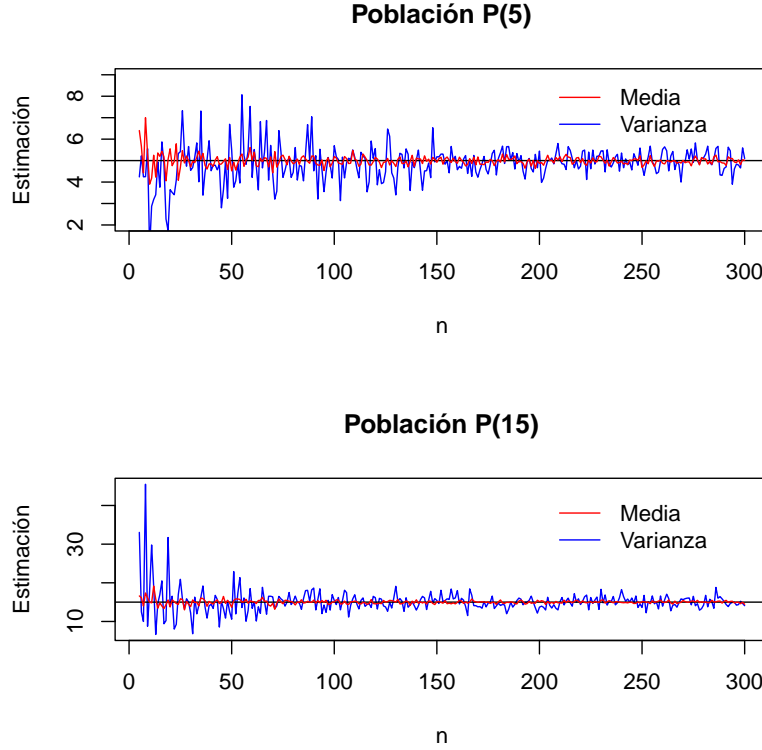


Figure 2.5: Comparación entre la media y la varianza muestral como estimador de λ en muestras provenientes de dos distribuciones Poisson.

Ejemplo 2.3.15. Dada una muestra aleatoria X_1, \dots, X_n proveniente de una distribución gamma con parámetro de forma k y parámetro de escala θ , entonces la media y la varianza poblacional están dadas por $\mu = k\theta$ y $\sigma^2 = k\theta^2$, de donde podemos expresar k y θ en término de μ y σ^2 como $k = \frac{\mu^2}{\sigma^2}$ y $\theta = \frac{\sigma^2}{\mu}$. Por lo tanto, los estimadores de momentos serán

$$\hat{k}_{mom} = \frac{\bar{X}^2}{S_n^2} \quad (2.3.5)$$

y

$$\hat{\theta}_{mom} = \frac{S_n^2}{\bar{X}} \quad (2.3.6)$$

Como una aplicación de lo anterior, suponga que un instituto técnico enfrenta problemas financieros y plantea un aumento en el costo de las matrículas. Es claro que ante un aumento del valor de matrícula, los estudiantes que no tengan la

capacidad de pago pueden retirar del instituto representando una pérdida económica para éste. Por lo anterior es necesario que el instituto conozca el nivel de ingreso de los estudiantes con el fin de decidir sobre el aumento de las matrículas que no cause la deserción estudiantil y a la vez pueda representar una ganancia económica para el instituto.

Con el fin de conocer el nivel de ingreso de los estudiantes, el instituto planeó una encuesta donde 127 estudiantes del instituto suministraron el valor de sus ingresos mensuales. Es usual pensar que una distribución Gamma puede ser apropiada para variables como "ingreso" puesto que en primer lugar, esta variable toma valores positivos, y en segundo lugar, es altamente no simétrico, ya que para la mayoría de la población la variable ingreso toma valores intermedios o bajos, pero existe una minoría de la población que tiene ingresos bastantes altos. Para observar el comportamiento del ingreso en los 127 datos muestrales, observamos el histograma presentado en la Figura 2.6 donde se puede ver estas características reflejas, y finalmente asumimos la distribución Gamma como la distribución poblacional.

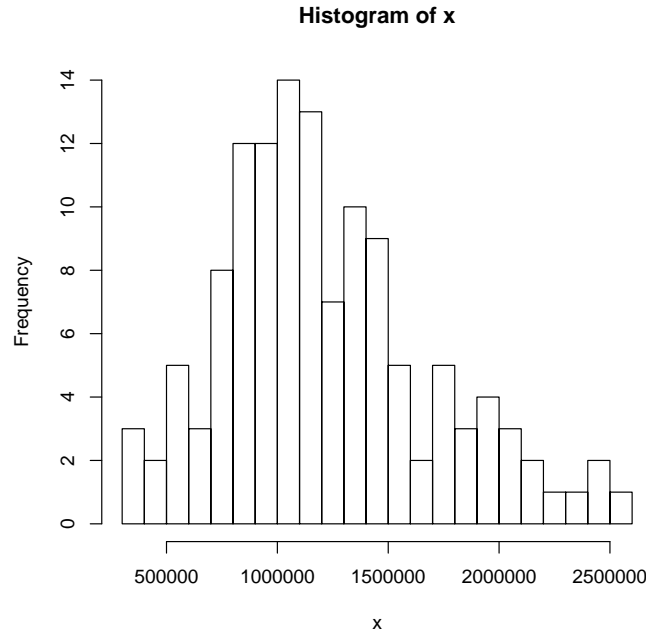


Figure 2.6: Histograma de los datos del Ejemplo 2.2.15.

Para estimar los parámetros de la distribución Gamma, calculamos los valores de las estadísticas \bar{X} y S_n , éstas son \$1220195 pesos y \$479670.8 pesos, y usando éstas calculamos los parámetros estimados usando (2.3.5) y (2.3.6), y tenemos que $\hat{k} = 6.96$ y $\hat{\theta} = 175094.6$. Para visualizar el ajuste de la distribución

$\text{Gamma}(6.96, 175094.6)$ a los datos, graficamos la función de densidad de esta distribución usando los siguientes códigos en R y la gráfica resultante se muestra en la Figura 2.7 donde podemos ver que la función de densidad se asemeja bastante al histograma de los datos presentado anteriormente.

```
> gama.densidad<-function(x){
+ fx<-dgamma(x,shape=k,scale=the)
+ }
>
> hist(x,breaks=20,freq=F)
> curve(gama.densidad(x),add=T)
## donde x contiene los 127 datos muestrales
```

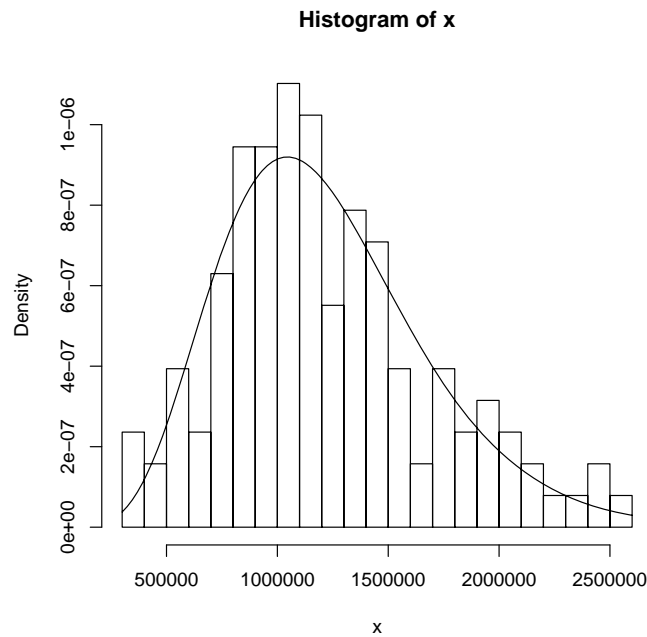


Figure 2.7: Función de densidad de la distribución Gamma estimada de los datos del Ejemplo 2.2.15.

Ahora, suponga que el instituto conoce que con el aumento los estudiantes con ingreso superior a \$2000000 pesos no tendrá dificultad para pagar el costo de la nueva matrícula, entonces una estimación del porcentaje de estos estudiantes puede ayudar a los directivos del instituto a prever los efectos del aumento. Como la distribución poblacional identificada es la distribución $\text{Gamma}(6.96, 175094.6)$, podemos calcular este porcentaje como $P(X > 1500000)$ donde $X \sim \text{Gamma}(6.96, 175094.6)$, esta probabilidad se puede calcular usando el comando

```
> pgamma(2000000,shape=6.96, scale=175094.6,lower.tail = F)
[1] 0.06111038
```

Y tenemos que aproximadamente el 6.1% de estudiantes de este instituto tiene un ingreso mensual superior a \$2000000 pesos. Dado que este porcentaje es bastante pequeño, los directivos deberían plantear un aumento menos drástico para evitar el efecto de deserción estudiantil.

Ejemplo 2.3.16. *Dada una muestra aleatoria X_1, \dots, X_n proveniente de una distribución $Beta(a, b)$, entonces la esperanza y la varianza de la distribución poblacional están dadas por*

$$\mu = \frac{a}{a+b} \quad (2.3.7)$$

y

$$\sigma^2 = \frac{ab}{(a+b+1)(a+b)^2}$$

Para encontrar los estimadores de momentos de a y b se debe escribir a estos en términos de μ y σ^2 . Para eso tenemos que

$$\begin{aligned} \sigma^2 &= \mu \frac{b}{(a+b-1)(a+b)} \\ &= \mu(1-\mu) \frac{1}{a+b-1} \end{aligned}$$

De donde

$$a+b = \frac{\mu(1-\mu)}{\sigma^2} - 1$$

Reemplazando lo anterior en (2.3.7) tenemos que

$$a = \mu(a+b) = \frac{\mu^2(1-\mu)}{\sigma^2} - \mu$$

y

$$b = \frac{\mu(1-\mu)}{\sigma^2} - 1 - a = (1-\mu) \left(\frac{\mu(1-\mu)}{\sigma^2} - 1 \right)$$

Y utilizando los principios del método de los momentos de estimar μ y σ^2 con \bar{X} y S_n^2 , tenemos los siguientes estimadores de momentos de a y b

$$\hat{a}_{mom} = \frac{\bar{X}^2(1-\bar{X})}{S_n^2} - \bar{X}$$

y

$$b = (1-\bar{X}) \left(\frac{\bar{X}(1-\bar{X})}{S_n^2} - 1 \right)$$

Como una aplicación de la distribución Beta, suponga que una almacén de cadena de ropa femenina investiga acerca de las prendas que son devueltas por

los clientes después de la venta. En el caso de que se observe un gran porcentaje de devoluciones, los directivos de la almacén estudiarán las causas que puede ser mala calidad de las prendas por parte del proveedor, precios muy altos comparado con productos de la misma categoría o inclusive vendedores muy hábiles pueden disuadir a los clientes aún cuando no quieren realizar la compra y éstos pueden arrepentir posteriormente y efectuar la devolución de la compra.

Con el fin de llevar a cabo la investigación, los directivos disponen de porcentaje de prendas devueltas en un mes para diferentes sucursales, estos datos son: 0.7%, 0.14%, 19.7%, 0.1%, 12.4%, 1.1%, 0.5%, 18.9%, 5.0%, 0.3%, 0.6%, 5.4%, 6.7% y 0.9%. Dado que los datos son porcentajes y los porcentajes siempre están dentro del intervalo $[0, 1]$, podemos pensar que una distribución Beta puede ser apropiada para describir a estos datos, observemos el histograma de los datos para verificar si esta distribución es adecuada o no.

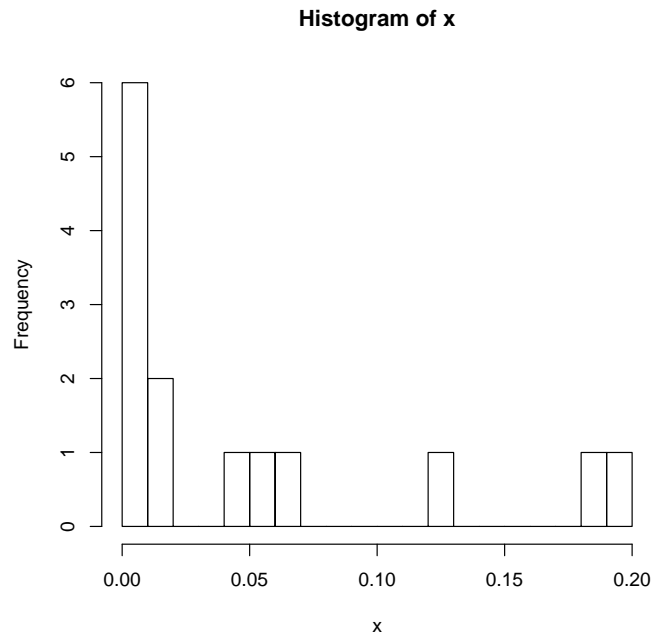


Figure 2.8: Histograma de los datos del Ejemplo 2.2.16.

Podemos ver que la gráfica es altamente no simétrica, de donde se descarta la distribución normal para describir a los datos. Sin embargo, podríamos pensar que una distribución exponencial puede ser apropiada dado que el histograma de los datos se asemejan a la función de densidad de esta distribución. Tenemos una herramienta útil para verificar si una distribución exponencial puede ser apropiada para los datos que es la gráfica QQ plot. Esta gráfica para los datos de este ejemplo se muestra en la Figura 2.9 donde se observa que la mayoría de datos

están situados por debajo de la recta, lo cual indica que la distribución exponencial no describe bien el comportamiento de los datos, y por consiguiente también está descartada.

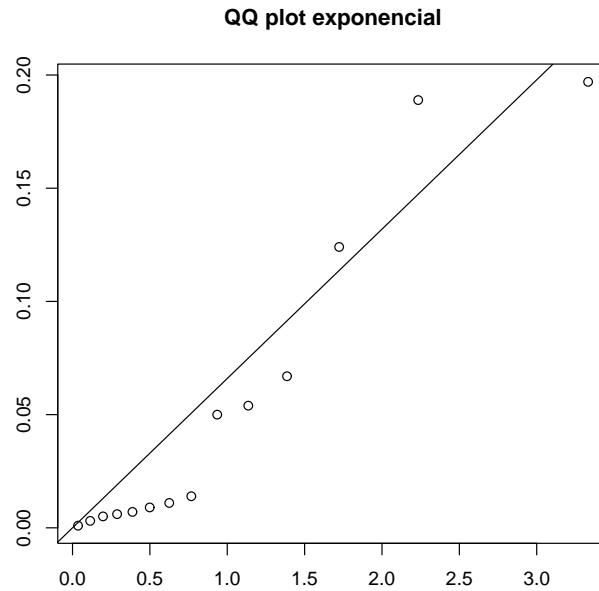


Figure 2.9: QQ plot para verificar la distribución exponencial para los datos del Ejemplo 2.2.16.

Dado lo anterior, podemos intentar utilizar la distribución Beta para describir este conjunto de datos. Primero calcularemos las estimaciones de los parámetros de la distribución Beta mediante el siguiente código.

```
> x<-c(0.7, 1.4, 19.7, 0.1, 12.4, 1.1, 0.5, 18.9, 5.0, 0.3, 0.6, 5.4, 6.7, 0.9)/100
> n<-length(x)
> va<-var(x)*(n-1)/n
> bar<-mean(x)
> a<-bar^2*(1-bar)/va-bar
> a
[1] 0.5447021
> b<-(1-bar)*(bar*(1-bar)/va-1)
> b
[1] 9.80242
```

De lo anterior, tenemos las estimaciones de 0.03 y 0.46 para los parámetros de la distribución Beta. Podemos visualizar la forma de la función de densidad Beta

con estos parámetros y ver que tenga aspectos similares con el histograma de los datos. Lo anterior se puede llevar a cabo usando los siguientes códigos.

```
> hist(x,breaks=20,freq=F)
> curve(dbeta(x,a,b),add=T)
```

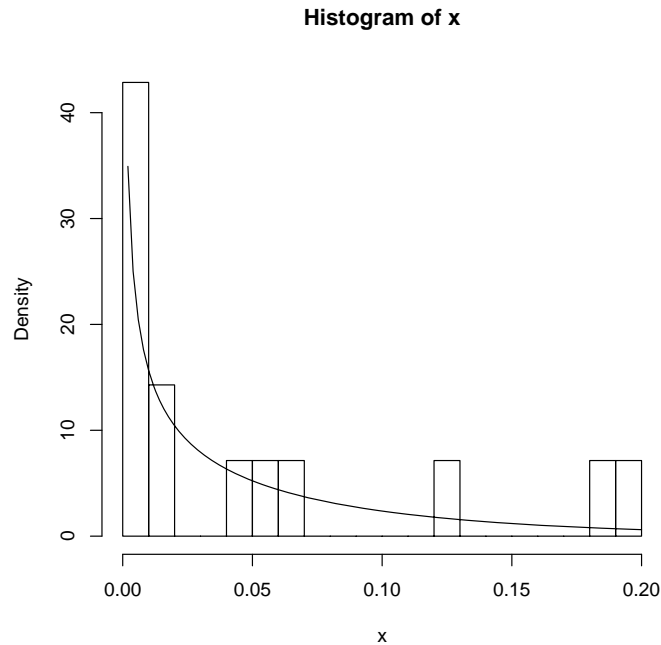


Figure 2.10: Histograma de los datos (a) y la función de densidad estimada (b) de los datos del Ejemplo 2.2.16.

Y como se puede observar en la Figura 2.10, la función de densidad de la distribución $Beta(0.03, 0.46)$ parece ser apropiada para los datos observados. Ahora, si se quiere estimar el porcentaje promedio de prendas devueltas, esto es la esperanza de la distribución poblacional, podemos utilizar simplemente $\bar{x} = 0.052 = 5.2\%$, o equivalentemente la expresión (2.3.7) dada por $\hat{a}/(\hat{a} + \hat{b}) = 0.052 = 5.2\%$.

En los ejemplos anteriores, se vio que en la distribución normal, los estimadores de momentos coinciden con los de máxima verosimilitud, situación que también ocurre en la distribución Poisson, exponencial y Bernoulli⁶. Sin embargo, en muestras provenientes de algunas distribuciones del tipo uniforme, el estimador de momentos puede no coincidir con el estimador de máxima verosimilitud, como lo ilustra el siguiente ejemplo.

⁶En general, el estimador de momentos no es único, por ejemplo en la distribución Poisson, el estimador de momentos \bar{X} coincide con el obtenido con el método de máxima verosimilitud, pero puede haber otros estimadores de momentos diferentes que \bar{X}

Ejemplo 2.3.17. Dada una muestra aleatoria X_1, \dots, X_n proveniente de una distribución uniforme continua sobre $(0, \theta)$. Para encontrar el estimador de momentos de θ , se tiene en cuenta que $\mu = \theta/2$, de donde $\theta = 2\mu$, entonces se concluye que $\hat{\theta}_{mom} = 2\bar{X}$, el cual es diferente que el estimador de máxima verosimilitud dada por $\hat{\theta}_{MV} = X_{(n)}$. En la siguiente sección, se estudiará cuál de estos dos estimadores es mejor. Sin embargo, podemos realizar un pequeño estudio simulación: se simulan muestras de tamaño 5, \dots , 300 de las distribuciones $Unif(0,3)$ y $Unif(0,5)$, en cada muestra simulada se calculan el estimador de momentos y de máxima verosimilitud. Estas simulaciones se pueden llevar a cabo modificando levemente los códigos en R presentados en la página 90. Las estimaciones resultantes se observan en la Figura 2.6, donde la línea negra horizontal denota el valor verdadero del parámetro. Podemos ver que con el estimador de máxima verosimilitud siempre se obtuvo valores más cercanos al parámetro sin importar el tamaño muestral, aunque las estimaciones de máxima verosimilitud parecen estar por debajo del θ verdadero, situación que no sucede con las estimaciones de momentos. Este hecho se confirmará en la siguiente sección mediante desarrollos teóricos.

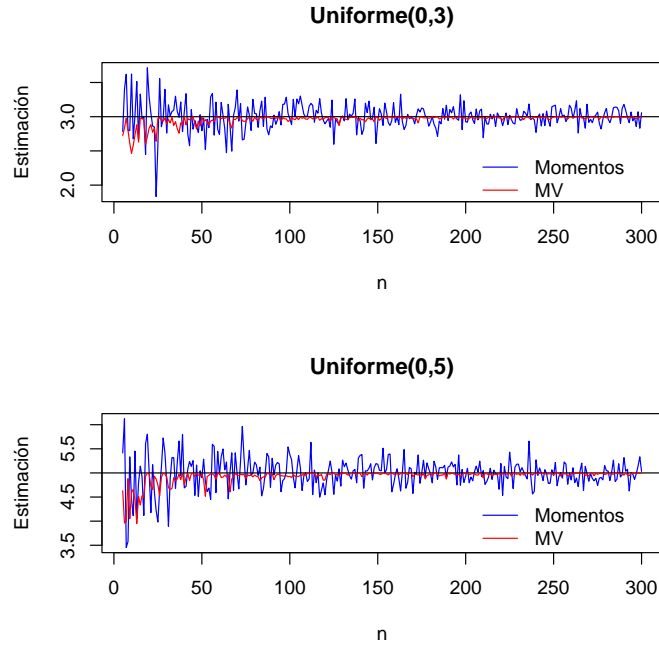


Figure 2.11: Comparación entre el estimador de máxima verosimilitud y el de momentos en muestras provenientes de distribuciones $Unif(0,3)$ y $Unif(0,5)$.

Análogo al método de máxima verosimilitud, en el método de los momentos también podemos garantizar la invarianza del estimador obtenido para cualquier

función del parámetro $g(\theta)$. Suponga que el tiempo de llegada de un bus a cierta estación contada desde las ocho de la mañana puede seguir una distribución $Unif(0, \theta)$, con esta distribución, estamos suponiendo que el tiempo de llegada tomar valor en cualquier intervalo de longitud fijo de $(0, \theta)$ con la misma probabilidad. Se ha visto en el ejemplo anterior que $\hat{\theta}_{mom} = 2\bar{X}$, y si la cantidad que se desea estimar es la probabilidad de que el bus llegue en menos de un minuto, entonces estamos interesados en estimar $p = 1/\theta$, y p puede ser escrito en función del primer momento, puesto que $p = 1/(2\mu)$, de esta forma, aplicando los principios del método de los momentos, se tiene que $\hat{p}_{mom} = 1/(2\bar{X})$.

Para concluir el método de los momentos, damos dos ejemplos interesantes acerca de la distribución uniforme.

Ejemplo 2.3.18. *Suponga que una muestra aleatoria X_1, \dots, X_n proviene de la distribución $U[-\theta, \theta]$, en este caso la esperanza es 0, y no depende del parámetro θ , por lo tanto, no hay forma de escribir a θ en función de la esperanza. Pero podemos recurrir a la varianza de la distribución uniforme, notando que la varianza es $\theta^2/3$, de donde se tiene que $\theta = \sqrt{3\sigma^2}$, (la solución $\theta = -\sqrt{3\sigma^2}$ claramente no puede ser el parámetro de la distribución, puesto que θ debe ser positivo), en conclusión un estimador de momentos de θ es $\sqrt{3S_n^2}$. Por otro lado, se puede ver que el estimador de máxima verosimilitud de θ está dado por $\max\{-X_{(1)}, X_{(n)}\}$ (Ejercicio XXXXXX).*

En la Figura 2.7, se observa resultados de muestras de tamaño 5, \dots , 300 simulados de una distribución $Unif[-3, 3]$ y en cada muestra simulada se calculan el estimador de momentos y de máxima verosimilitud. Podemos observar que en muestras pequeñas los resultados obtenidos con el estimador de máxima verosimilitud casi siempre están por debajo del parámetro causando el problema de subestimación, a medida que las muestras se hacen grandes el método de máxima verosimilitud parece funcionar mejor; por otro lado, aunque los valores obtenidos con el método de los momentos no parecen tener el problema de subestimación que sí lo tiene el de máxima verosimilitud, los valores obtenidos con el método de los momentos son muy dispersos, y hay muestras donde la estimación puede estar realmente lejos del parámetro.

En el anterior ejemplo se vio que en algunas ocasiones en el método de los momentos puede no ser útil evaluar el primer momento sino usando el segundo momento (o equivalentemente la varianza de la distribución poblacional). En el siguiente ejemplo, ilustramos un caso donde el procedimiento estándar del método de los momentos arroja dos soluciones y se debe tener en cuenta el espacio paramétrico de los parámetros de interés para escoger la solución apropiada.

Ejemplo 2.3.19. *Suponga que una muestra aleatoria X_1, \dots, X_n proviene de la distribución $U[\theta_1, \theta_2]$, donde θ_1 y θ_2 son desconocidos. Para estimar estos parámetros vía el método de los momentos, necesitamos escribirlos en término*

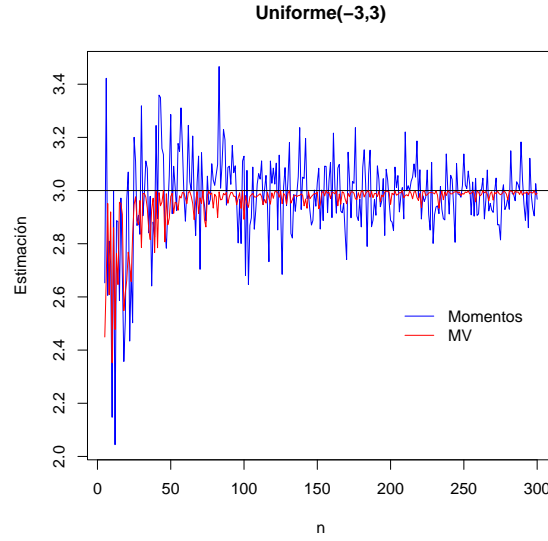


Figure 2.12: Comparación entre el estimador de máxima verosimilitud y el de momentos en muestras provenientes de la distribución $Unif(-3, 3)$.

de la media μ y la varianza σ^2 . Usando el Resultado 1.1.11, tenemos que

$$\begin{cases} \mu = \frac{\theta_1 + \theta_2}{2} \\ \sigma^2 = \frac{(\theta_2 - \theta_1)^2}{12}, \end{cases}$$

de donde tenemos dos soluciones para θ_1 y θ_2 , estas son $\begin{cases} \theta_1 = \mu - \sqrt{3}\sigma \\ \theta_2 = \mu + \sqrt{3}\sigma \end{cases}$ ó

$\begin{cases} \theta_1 = \mu + \sqrt{3}\sigma \\ \theta_2 = \mu - \sqrt{3}\sigma \end{cases}$. Nótese que en la segunda solución $\theta_1 > \theta_2$, no cumple con el supuesto de una distribución $U[\theta_1, \theta_2]$, y por consiguiente, usaremos la primera solución, y los estimadores de momentos de θ_1 y θ_2 serán $\bar{X} - \sqrt{3}S_n$ y $\bar{X} + \sqrt{3}S_n$, respectivamente. Se puede ver fácilmente que los estimadores de máxima verosimilitud de θ_1 y θ_2 son $X_{(1)}$ y $X_{(2)}$. (Ejercicio XXXX)

2.3.3 Método de mínimos cuadrados

El método de mínimos cuadrados es un método muy común en la teoría del modelamiento estadístico, aquí se hace una breve introducción utilizando este método para estimar la esperanza de una distribución poblacional.

Definición 2.3.4. Dado un conjunto de datos x_1, \dots, x_n , una medida de dispersión es una función que satisface las siguientes propiedades

1. $D(\mathbf{x} + \mathbf{1}_n c) = D(\mathbf{x})$

2. $D(-\mathbf{x}) = D(\mathbf{x})$

donde \mathbf{x} es el vector conformado por los n datos, $\mathbf{1}_n$ es el vector de unos de tamaño n y c es cualquier número real.

Ejemplo 2.3.20. Las varianzas S_n^2 , S_{n-1}^2 y las respectivas desviaciones estándares son medidas de dispersión. Puesto que en primer lugar, si se traslada a un conjunto de datos agregando una constante c S_n^2 y S_{n-1}^2 no cambian de valor; en segundo lugar estas varianzas tampoco cambian de valor si se multiplica por -1 a todos los datos.

Ahora introducimos el método de mínimos cuadrados para estimar la esperanza de una distribución poblacional μ basándonos en una muestra X_1, \dots, X_n . Dado que μ es el media poblacional, se espera que los valores de la muestra observada esté cercanos a μ , por lo tanto, podemos proponer estimar μ como el valor que más se acerque a los datos. Una forma de medir la distancia entre dos puntos es la distancia euclidiana, así podemos estimar μ como aquel que minimiza la cantidad

$$Q = \sum_{i=1}^n (x_i - \mu)^2.$$

De lo anterior, es claro el origen del nombre de estimador de mínimos cuadrados. Derivando Q con respecto a μ e igualando a cero conduce a la solución de $\mu = \bar{X}$, de donde tenemos que el estimador de la esperanza poblacional bajo cualquier distribución está dado por

$$\hat{\mu}_{MC} = \bar{X}.$$

2.4 Propiedades de estimadores puntuales

En la anterior sección, se observó que para estimar un parámetro θ , el método de máxima verosimilitud y el de momentos pueden conducir a estimadores diferentes, más aun, se puede crear muchos otros tipos de estimadores para θ , pues un estimador es simplemente una estadística que es usada para estimar. Por ejemplo dada una muestra aleatoria X_1, \dots, X_{20} con media poblacional μ desconocido, un estimar razonable para μ es la media muestral \bar{X} , sin embargo, alguien puede querer usar la estadística $\sum X_i$ para estimar μ , otro puede preferir algo como $\exp\{\sum X_i\}$, otra persona puede inventar su propio estimador, entonces ¿cómo se puede escoger el mejor estimador entre un conjunto de estimadores? ¿qué aspectos, propiedades se debe tener en cuenta para esa escogencia? El objetivo de este capítulo es introducir conceptos que contestan estas preguntas.

Muestra	T_1	T_2	T_3
1	4.1	5.5	5.1
2	4.3	5.6	5.0
3	5.6	5.4	4.8
4	5.3	5.5	4.9
5	4.5	5.4	5.2
6	4.7	5.6	5.0
7	5.7	5.5	4.9
promedio	4.88	5.5	4.99
desviación	0.64	0.08	0.13

Tabla 2.2: Valores de tres estimadores en 7 muestras diferentes.

2.4.1 Error cuadrático medio

Consideramos la siguiente situación hipotética. Suponga que para estimar un parámetro θ , se disponen de tres estimadores T_1 , T_2 y T_3 , y además suponga que las respectivas estimaciones en 7 muestras observadas de la población son los valores dados en la Tabla 2.2.

Y además suponga que el valor verdadero de θ es 5, ¿cuál estimador es mejor dadas las anteriores estimaciones? Para responder esta pregunta, observamos lo siguiente con respecto a los tres estimadores

- Los valores que toma T_1 en promedio están cerca del 5, pero estos están muy alejados entre sí, es decir, tienen una dispersión grande. Esta dispersión grande es una propiedad indeseada del estimador, pues generalmente en la práctica, tenemos solo una muestra, una dispersión grande entre los valores de T_1 implica que hay mayor probabilidad de que T_1 toma un valor alejado del parámetro en una muestra.
- Los valores que toma T_2 están alrededor del 5.5, muy por encima del valor verdadero de θ , 5, esta situación se llama la sobreestimación. Por otro lado, en término de la dispersión, se observa que los valores están altamente concentrados.
- Los valores que toma T_3 , en primer lugar, están alrededor del 5, además de tener una dispersión pequeña. Lo anterior indica que en todas las muestras, el valor de T_3 está cercano del valor de θ . Y podemos concluir que el mejor estimador de los tres es el T_3 .

La anterior situación nos ilustra que un buen estimador T debe tener dos propiedades

1. Los valores que toma T en promedio debe ser cercano al parámetro θ . Teniendo en cuenta que la esperanza de una variable puede ser interpretado como un promedio ponderado de todos los valores que toma la variable, podemos concluir que T debe cumplir con $E(T) = \theta$,

2. La varianza de T debe ser pequeña.

Ahora damos la siguiente definición que describe la propiedad 1.

Definición 2.4.1. Dada una muestra aleatoria X_1, \dots, X_n proveniente de una distribución con parámetro desconocido θ , y sea T un estimador de θ , se define el sesgo de T como $B_T = E(T) - \theta$.

Cuando $B_T = 0$ o equivalentemente $E(T) = \theta$, se dice que el estimador T es insesgado para θ . Cuando $B_T > 0$ o equivalentemente $E(T) > \theta$, se dice que T sobreesima a θ , es decir, en promedio la estimación obtenida usando T es mayor que θ . Análogamente se dice que T subestima a θ cuando $B_T < 0$.

Dada la anterior definición, en primera instancia, se necesita estimadores con sesgo pequeño, y si es posible, insesgados. Adicionalmente, se espera que un buen estimador tenga varianza pequeña. De esta forma, si entre dos estimadores T_1 y T_2 , T_1 tiene sesgo y varianza ambos menores que T_2 , podemos concluir fácilmente que T_1 es mejor que T_2 . Pero cuando T_1 tiene sesgo menor, pero varianza mayor que T_2 , no es fácil determinar cuál es mejor. En este caso, podemos usar el siguiente criterio que combina tanto al sesgo como a la varianza de un estimador.

Definición 2.4.2. Dada una muestra aleatoria X_1, \dots, X_n proveniente de una distribución con parámetro desconocido θ , y sea T un estimador de θ , se define el error cuadrático medio de T como $ECM_T = E(T - \theta)^2$.

Nótese que en la anterior definición, cuando un estimador T es insesgado para θ , se tiene que $ECM_T = E(T - E(T))^2$, esto es, el error cuadrático medio es la varianza del estimador T .

Más aun, el criterio del error cuadrático medio combina al sesgo y la varianza, tenemos que

$$\begin{aligned} ECM_T &= E[T - E(T) + E(T) - \theta]^2 \\ &= E[(T - E(T))^2 + 2(T - E(T))(E(T) - \theta) + (E(T) - \theta)^2] \\ &= E[(T - E(T))^2] + 2(E(T) - \theta) \underbrace{E[T - E(T)]}_{\text{igual a 0}} + (E(T) - \theta)^2 \\ &= Var(T) + B_T^2 \end{aligned}$$

Entonces un buen estimador debe tener el error cuadrático medio pequeño, y para los estimadores insesgados, se necesita que la varianza sea pequeña.

Ahora, al principio del capítulo, afirmaba que es natural estimar la media poblacional μ con la media muestral \bar{X} , ¿qué tan buena es esta idea? El siguiente resultado nos permite examinar el comportamiento de \bar{X} como estimador de μ .

Resultado 2.4.1. Sea una muestra aleatoria X_1, \dots, X_n proveniente de una distribución con media μ y varianza σ^2 , entonces

1. si se considera a \bar{X} como el estimador de μ , se tiene que \bar{X} es insesgado para μ , es decir, $E(\bar{X}) = \mu$ y además $Var(\bar{X}) = \sigma^2/n$.

2. si se considera a S_n^2 y S_{n-1}^2 como estimadores de σ^2 , se tiene que S_n^2 es sesgado para σ^2 donde el sesgo es $-\frac{\sigma^2}{n}$; y S_{n-1}^2 es insesgado para σ^2 .

Proof. La demostración de la parte 1 es trivial, y se deja como ejercicio. Para ver la parte 2, tenemos que

$$\begin{aligned} E(S_n^2) &= \frac{1}{n} E \left(\sum_{i=1}^n (X_i - \bar{X})^2 \right) \\ &= \frac{1}{n} E \left(\sum_{i=1}^n X_i^2 - n\bar{X}^2 \right) \\ &= \frac{1}{n} \left\{ \sum_{i=1}^n [Var(X_i) + (E(X_i))^2] - nE(\bar{X})^2 \right\} \\ &= \frac{1}{n} \left\{ n\sigma^2 + n\mu^2 - n \left[\frac{\sigma^2}{n} + \mu^2 \right] \right\} \\ &= \frac{n-1}{n} \sigma^2, \end{aligned}$$

de donde se concluye que S_n^2 es sesgado para σ^2 con sesgo $-\frac{\sigma^2}{n}$. Con respecto a S_{n-1}^2 , al observar que

$$S_{n-1}^2 = \frac{n}{n-1} S_n^2,$$

se tiene que $E(S_{n-1}^2) = \sigma^2$ y por consiguiente es insesgado para σ^2 . \square

Nótese en primer lugar que en el anterior resultado, no se ha especificado la distribución de probabilidad en la población, entonces podemos aplicarlo para muestras que provienen de cualquier distribución de probabilidad. En particular, en muestras provenientes de la distribución $Exp(\theta)$, $P(\theta)$, $Bernoulli(\theta)$, $N(\theta, \sigma^2)$, el estimador de máxima verosimilitud coincide con el de momentos \bar{X} , usando el anterior resultado, podemos concluir que \bar{X} es insesgado para el parámetro θ en cualquier de estas cuatro distribuciones.

Por otro lado, podemos ver que \bar{X} , en primer lugar, es un estimador insesgado para μ ; en segundo lugar observe que $Var(\bar{X})$ es inversamente proporcional al tamaño muestral n , es decir, a medida que la muestra crece, las estimaciones son más concentrados alrededor de μ . En la Figura 2.7, se muestra un estudio de simulación donde se simuló muestras de tamaños $1, \dots, 300$, provenientes de distribución normal y exponencial, y en cada muestra se calcula el promedio muestral. Se observa que entre más grande sea el valor de n , más concentrado están las estimaciones alrededor de la media poblacional $\mu = 5$.

Otra observación interesante es que en el contexto del resultado anterior, la variable X_1 , vista como un estimador de μ también es insesgado, puesto que por definición de μ , se tiene que $E(X_1) = \mu$; la misma conclusión se tiene para X_i para

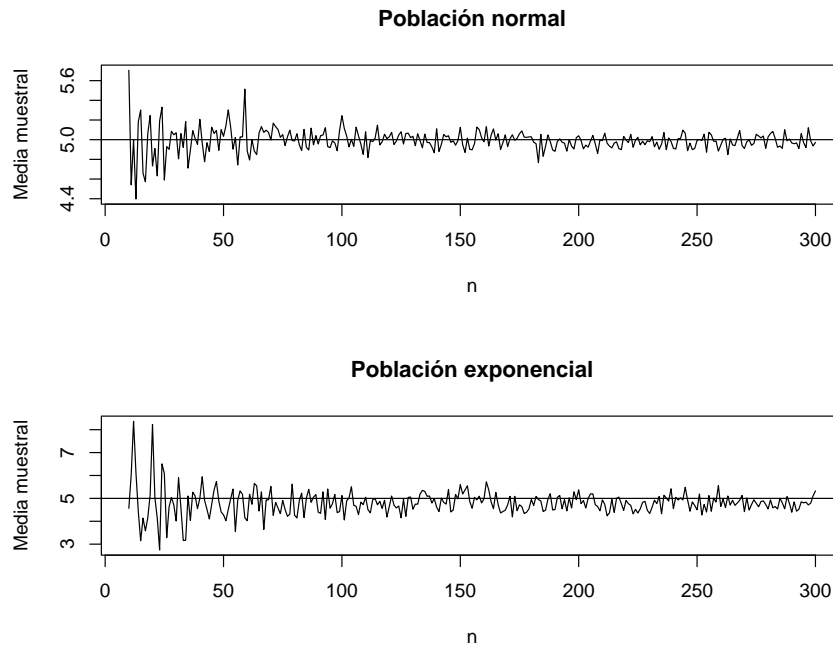


Figure 2.13: Relación entre la estimación de μ y el tamaño muestral n .

$i = 2, \dots, n$. Es decir, el estimador insesgado para un parámetro puede no ser único ⁷. Pero la varianza de X_i con $i = 1, \dots, n$ es σ^2 , que es mayor que $Var(\bar{X})$, de donde se concluye que estos no son tan buenos estimadores como \bar{X} .

Ahora, revisamos los estimadores S_n^2 y S_{n-1}^2 como estimadores de σ^2 . Aunque S_n^2 resulta ser sesgado para σ^2 , el sesgo se hace pequeño cuando el tamaño muestral crece, más aun,

$$\lim_{n \rightarrow \infty} B_{S_n^2} = \lim_{n \rightarrow \infty} -\frac{\sigma^2}{n} = 0.$$

Estimadores sesgados que cumplen la propiedad $\lim_{n \rightarrow \infty} B_{S_n^2} = 0$ son llamados asintóticamente insesgados. Para ilustrar los estimadores S_n^2 y S_{n-1}^2 en término de estimación, podemos simular muestras provenientes de una distribución normal con tamaños de muestral $n = 2, 20, 30, 50, 300$, y en cada muestra calculamos los dos estimadores. El comando utilizado en R es

```
> set.seed(1)
> n<-c(2,10,30, 50,100, 300, 1000,5000)
> var1<-rep(NA,length(n))
> var2<-rep(NA,length(n))
```

⁷De hecho, si tomamos cualquier subconjunto de la muestra aleatoria, el promedio muestral de este subconjunto será un estimador insesgado para μ

```

> for(k in 1:length(n))
+ {
+   data<-rnorm(n[k],5,9)
+   var1[k]<-var(data)
+   var2[k]<-(n[k]-1)*var(data)/n[k]
+ }
> plot(var1,type="b", col=4,ylim=c(min(var2),130),xlab="Tamaño de muestra", ylab="Estimac
> lines(var2,type="b", col=2, pch=2)
> abline(h=81)
> axis(1, 1:length(n), n)
> legend(3,120,c("Inssegado","Sesgado"), col=c(4,2), lty=c(1,1),pch=c(1,2))

```

Y como resultado, obtenemos la Figura 2.8, donde podemos observar que las estimaciones del estimador sesgado S_n^2 siempre son inferiores que los del estimador insesgado S_{n-1}^2 y en segundo lugar la diferencia entre los dos estimadores se hace cada vez más pequeña y los valores de ambos estimadores se acercan al parámetro poblacional a medida que el tamaño muestral crece. Por otro lado, aunque S_n^2 subestima la varianza poblacional, en la gráfica podemos observar que en la muestra simulada del tamaño 300, 1000 y 5000, las estimaciones de S_n^2 estuvieron por encima de la varianza poblacional, esto no es ninguna contradicción, ya que el concepto de un sesgo negativo, o equivalentemente la subestimación de un estimador indica que promediando todos los valores del estimador, da un valor inferior al parámetro, mas no indica que todos los valores del estimadores son inferiores que el parámetro.

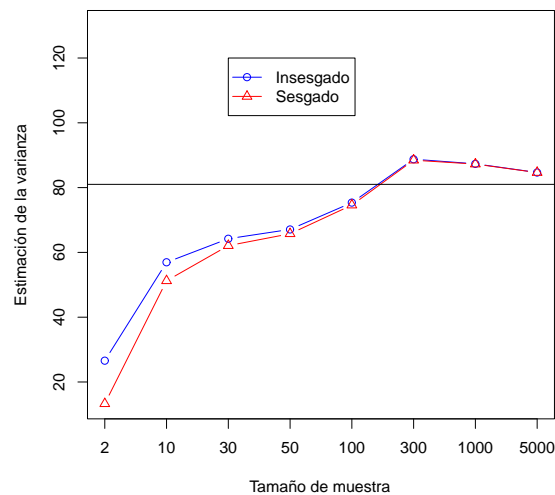


Figure 2.14: Ilustración de las estimaciones de S_n^2 y S_{n-1}^2 como estimadores de σ^2

Finalmente, de la parte 2 del Resultado 2.3.1, también se puede concluir que los estimadores obtenidos mediante el método de máxima verosimilitud o el de momentos pueden ser sesgados, puesto que una muestra proveniente de la distribución normal, se ha visto que cuando μ es desconocido, $\hat{\sigma}_{MV}^2 = \hat{\sigma}_{mom}^2 = S_n^2$, y ésta es sesgada para σ^2 . Sin embargo, en la demostración del Resultado 2.3.1, se vio que en algunos casos, una pequeña modificación al estimador de máxima verosimilitud o el de momentos puede corregir el sesgo y obtener un estimador insesgado.

El Resultado 2.3.1 es válida para muestras provenientes de cualquier distribución, sin embargo, cuando la muestra proviene de una distribución normal, existe el siguiente resultado que nos permite ver que S_n^2 es sesgado para σ^2 . Lo presentamos pues ésta es de gran utilidad para la teoría desarrollada en los capítulos siguientes.

Resultado 2.4.2. Sea X_1, \dots, X_n una muestra aleatoria proveniente de $N(\mu, \sigma^2)$, y sea $Y = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\sigma^2}$, entonces se tiene que $Y \sim \chi_{n-1}^2$.

Proof. En primer lugar, consideramos la variable $\sum_{i=1}^n (X_i - \mu)^2$, tenemos

$$\begin{aligned} \sum_{i=1}^n (X_i - \mu)^2 &= \sum_{i=1}^n (X_i - \bar{X} + \bar{X} - \mu)^2 \\ &= \sum_{i=1}^n (X_i - \bar{X})^2 + 2(\bar{X} - \mu) \underbrace{\sum_{i=1}^n (X_i - \bar{X})}_{=0} + n(\bar{X} - \mu)^2 \\ &= \sum_{i=1}^n (X_i - \bar{X})^2 + n(\bar{X} - \mu)^2 \end{aligned}$$

Dividiendo σ^2 en ambos lados, se tiene que

$$\underbrace{\sum_{i=1}^n \frac{(X_i - \mu)^2}{\sigma^2}}_A = \underbrace{\sum_{i=1}^n \frac{(X_i - \bar{X})^2}{\sigma^2}}_Y + \underbrace{\frac{n(\bar{X} - \mu)^2}{\sigma^2}}_B.$$

Si podemos suponer que las variables S_n^2 y \bar{X} son independientes, podemos concluir que $\sum_{i=1}^n \frac{(X_i - \bar{X})^2}{\sigma^2}$ y $\frac{n(\bar{X} - \mu)^2}{\sigma^2}$ también son independientes. Entonces existe la siguiente relación entre las funciones generadora de momentos de las variables A , Y y B : $m_A(t) = m_Y(t)m_B(t)$, de donde se obtiene que

$$m_Y(t) = m_A(t)/m_B(t). \quad (2.4.1)$$

Ahora, $\frac{X_i - \mu}{\sigma}$ son variables con distribución normal estándar para $i = 1, \dots, n$, entonces $\sum_{i=1}^n \frac{(X_i - \mu)^2}{\sigma^2}$ tiene distribución χ_n^2 con función generadora de momentos $(1 - 2t)^{-n/2}$. Por el otro lado $\bar{X} \sim N(\mu, \sigma^2/n)$, entonces $\frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \sim$

$N(0,1)$, de donde se tiene que $\frac{n(\bar{X} - \mu)^2}{\sigma^2} \sim \chi_1^2$ cuya función generadora de momentos es $(1 - 2t)^{-1/2}$. Reemplazando lo anterior en (2.4.1), se tiene que $m_Y(t) = (1 - 2t)^{-(n-1)/2}$, lo cual indica que $Y \sim \chi_{n-1}^2$. \square

Para completar la demostración del anterior resultado, es necesario probar la independencia entre \bar{X} y S_n^2 . Tenemos el siguiente resultado.

Resultado 2.4.3. *Dada X_1, \dots, X_n una muestra aleatoria proveniente de $N(\mu, \sigma^2)$, se tiene que \bar{X} y S_n^2 son independientes,*

Proof. La demostración de este resultado es tomada de Casella & Berger (2002). Se probará que \bar{X} y $\sum_{i=1}^n (X_i - \bar{X})^2$ son independientes. Tenemos

$$\begin{aligned} \sum_{i=1}^n (X_i - \bar{X})^2 &= (X_1 - \bar{X})^2 + \sum_{i=2}^n (X_i - \bar{X})^2 \\ &= \left[\sum_{i=1}^n (X_i - \bar{X}) - \sum_{i=2}^n (X_i - \bar{X}) \right]^2 + \sum_{i=2}^n (X_i - \bar{X})^2 \\ &= \left[\sum_{i=2}^n (X_i - \bar{X}) \right]^2 + \sum_{i=2}^n (X_i - \bar{X})^2. \end{aligned}$$

De lo anterior, se observa que $\sum_{i=1}^n (X_i - \bar{X})^2$ puede verse como una función de las variables $X_2 - \bar{X}, \dots, X_n - \bar{X}$, por lo tanto, basta ver que estas variables son independientes de \bar{X} . Sin embargo, las variables X_1, \dots, X_n tienen distribución $N(\mu, \sigma^2)$, y la presencia de estos dos parámetros complican un poco los cálculos, por lo que se trabajará con las variables estandarizadas, Z_1, \dots, Z_n , donde el promedio está dada por

$$\begin{aligned} \bar{Z} &= \frac{1}{n} \sum_{i=1}^n Z_i \\ &= \frac{1}{n} \sum_{i=1}^n \frac{X_i - \mu}{\sigma} \\ &= \frac{1}{n\sigma} \sum_{i=1}^n X_i - \frac{\mu}{\sigma} \\ &= \frac{\bar{X}}{\sigma} - \frac{\mu}{\sigma}, \end{aligned}$$

además $Z_i - \bar{Z} = \frac{X_i - \bar{X}}{\sigma}$ para todo $i = 2, \dots, n$. Por lo tanto, para ver que las variables $X_2 - \bar{X}, \dots, X_n - \bar{X}$ son independientes de \bar{X} , basta ver que $Z_2 - \bar{Z}, \dots, Z_n - \bar{Z}$ son independientes de \bar{Z} .

Para esto, utilizamos la función de densidad conjunta de las variables Z_1, \dots, Z_n dada por

$$f(z_1, \dots, z_n) = (2\pi)^{-n/2} \exp \left\{ -\frac{1}{2} \sum_{i=1}^n z_i^2 \right\},$$

Ahora, se define la transformación $Y_1 = \bar{Z}$, y $Y_i = Z_i - \bar{Z}$ para $i = 2, \dots, n$, con jacobiano igual a n^{-1} . Podemos ver que $Z_1 = Y_1 - \sum_{i=2}^n Y_i$ y $Z_i = Y_i + \bar{Y}$ para $i = 2, \dots, n$. Usando el teorema de transformación, se tiene que la función de densidad conjunta de Y_1, \dots, Y_n está dada por

$$\begin{aligned} f(y_1, \dots, y_n) &= n(2\pi)^{-n/2} \exp \left\{ -\frac{1}{2} (y_1 - \sum_{i=2}^n y_i)^2 \right\} \exp \left\{ -\frac{1}{2} \sum_{i=2}^n (y_i + \bar{y})^2 \right\} \\ &= n(2\pi)^{-n/2} \exp \left\{ -\frac{n}{2} y_1^2 \right\} \exp \left\{ \sum_{i=2}^n y_i^2 + \left[\sum_{i=2}^n y_i \right]^2 \right\}, \end{aligned}$$

la cual es producto entre dos funciones, una que depende solo de y_1 y la otra de y_i con $i = 2, \dots, n$ (ver el teorema 4.6.11 de Casella & Berger (2002)). Entonces podemos concluir que Y_2, \dots, Y_n y Y_1 son independientes y el resultado queda demostrado.

Existen otra forma de probar esta independencia utilizando el denominado teorema de Basu, sin embargo, no hemos introducido algunos conceptos necesarios para este teorema, por esta razón, será presentado más adelante. \square

Usando el Resultado 2.3.2 y propiedades de la distribución χ^2 , se tiene que $E(\sum_{i=1}^n \frac{(X_i - \bar{X})^2}{\sigma^2}) = n - 1$, de donde

$$E(S_n^2) = E \left(\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n} \right) = \frac{n-1}{n} \sigma^2. \quad (2.4.2)$$

Es decir, el estimador S_n^2 es sesgado para σ^2 , mientras que S_{n-1}^2 es insesgado para σ^2 .

Ahora, en muestras aleatorias provenientes de una distribución exponencial, Poisson o normal, el estimador de máxima verosimilitud es igual al estimador de momentos, pero no siempre es así, como ocurre en muestras provenientes de distribuciones uniformes continuas. Considera una muestra proveniente de $Unif[0, \theta]$, el estimador de máxima verosimilitud de θ es $\hat{\theta}_{MV} = X_{(n)}$ y el de momentos está dado por $\hat{\theta}_{mom} = 2\bar{X}$. Para saber cuál de estos dos estimadores es mejor, comparemos los dos estimadores en término del sesgo y la varianza en el siguiente ejemplo.

Ejemplo 2.4.1. Sea X_1, \dots, X_n una muestra aleatoria proveniente de una distribución uniforme continua sobre $[0, \theta]$, el estimador de máxima verosimilitud de θ es $\hat{\theta}_{MV} = X_n$ y el estimador de momentos es $\hat{\theta}_{mom} = 2\bar{X}$. Primero revisamos el desempeño de los estimadores en término del sesgo, es decir, calcularemos la

esperanza de ambos estimadores. Para calcular $E(X_{(n)})$ es necesario conocer la función de densidad de probabilidad o la función de distribución de X_n . Para eso, usamos la propiedad (2.2.2), de donde para $x \in [0, \theta]$ tenemos:

$$F_{X_{(n)}}(x) = \frac{x^n}{\theta^n}.$$

Dada la función de distribución de $X_{(n)}$, podemos obtener la función de densidad dada por

$$f_{X_{(n)}}(x) = \frac{nx^{n-1}}{\theta^n} I_{[0, \theta]}(x).$$

Ahora calculamos $E(X_{(n)})$ como

$$E(X_{(n)}) = \int_0^\theta \frac{nx^{n-1}}{\theta^n} dx = \frac{n}{n+1} \theta. \quad (2.4.3)$$

Lo anterior concluye que $X_{(n)}$ como estimador de θ , es sesgado. Más aun, subestima a θ , hecho que se había observado en la Figura 2.5. También nótese que en la expresión (2.4.3), cuando el tamaño de la muestra $n \rightarrow \infty$, el sesgo tiende a cero, esto es, $X_{(n)}$ es un estimador asintóticamente insesgado. Ahora, miramos cómo es el sesgo del estimador de momentos, tenemos

$$E(2\bar{X}) = 2E(\bar{X}) = 2\frac{\theta}{2} = \theta,$$

pues la esperanza de \bar{X} es igual a la esperanza de la distribución (ver Resultado 2.3.1). En conclusión, el estimador $2\bar{X}$ es insesgado para θ . En el término del sesgo, el estimador $2\bar{X}$ es mejor que $X_{(n)}$, aunque cuando n es grande, los dos son muy similares. Ahora miramos cuál es mejor en término de la varianza. Tenemos

$$\begin{aligned} \text{Var}(X_{(n)}) &= E(X_{(n)})^2 - (EX_{(n)})^2 \\ &= \int_0^\theta \frac{nx^{n+1}}{\theta^n} dx - \left(\frac{n\theta}{n+1}\right)^2 \\ &= \frac{n\theta^2}{n+2} - \frac{n^2\theta^2}{(n+1)^2} \\ &= \frac{n\theta^2}{(n+2)(n+1)^2}. \end{aligned}$$

Por el otro lado,

$$\text{Var}(2\bar{X}) = 4\text{Var}(\bar{X}) = 4\frac{\theta^2}{12n} = \frac{\theta^2}{3n}.$$

Algunas operaciones algebraicas indica que $\text{Var}(X_{(n)})$ es mucho más pequeña que la de $2\bar{X}$, y este aspecto ventajoso de $X_{(n)}$ puede recompensar con su sesgo que de todas formas es despreciable para valores grandes de n . Por lo tanto, se recomienda usar $X_{(n)}$ para estimar θ . Nótese que la anterior observación con respecto a la varianza también es reflejada en la Figura 2.5.

Hasta este punto, hemos concluido que en muchas situaciones, se prefiere, en primera instancia a los estimadores insesgados, (o por lo menos asintóticamente insesgados) y entre ellos, aquel que tiene menor varianza. Una pregunta interesante que surge ahora es si se dispone de un estimador insesgado para una función del parámetro $g(\theta)$, cómo podemos modificarlo para que sigue siendo insesgado, pero con varianza menor. Para eso necesitamos el concepto de suficiencia de un estimador.

2.4.2 Suficiencia

El concepto de la suficiencia de un estimador está ligado con la idea de reducción de datos. Una muestra aleatoria provee información acerca del parámetro desconocido que se desea estimar, pero esta información está contenida en un conjunto de variables aleatorias, si hay una manera de encontrar una función de estas variables, que contiene la misma cantidad de información para el propósito de estimación, se lograría la reducción de datos. Una variable que logra esta reducción y que además es usado para estimar el parámetro es un estimador suficiente para el parámetro. Siguiendo a esta idea, es natural pensar que toda la información contenida en la muestra X_1, \dots, X_n está contenida en el estimador suficiente (T), entonces una vez conocido el valor que toma T , la muestra ya no provee ninguna información acerca del parámetro.

La definición rigurosa de un estimador suficiente se presenta a continuación.

Definición 2.4.3. *Dada una muestra aleatoria X_1, \dots, X_n con función de densidad $f(x_i, \theta)$, y sea $T = T(X_1, \dots, X_n)$ un estimador de θ , se dice que T es suficiente para θ si la distribución condicional de X dado valores de T no depende de θ*

En algunos textos, establecen que un estimador T es suficiente para θ si $P(X_1 = x_1, \dots, X_n = x_n | T = t)$ no depende de θ , la cual no es de todo riguroso, puesto que cuando las variables X_i con $i = 1, \dots, n$ son continuas, la anterior probabilidad condicional (cuando está bien definida) siempre es igual a cero, que no depende de θ ; por el otro lado, también el estimador T como función de las variables de la muestra también puede ser continua, entonces $P(T = t) = 0$ y no se puede definir la esperanza condicional. Claro que cuando las variables X_i y T son discretas, podemos usar esta definición sin problema e ilustramos el forma de verificar que un estimador sea suficiente en el siguiente ejemplo.

Ejemplo 2.4.2. *Sea X_1, \dots, X_n una muestra aleatoria con distribución $P(\lambda)$, se ha visto que el estimador de máxima verosimilitud y de momentos de λ está dada por \bar{X} . Además éste es insesgado para θ por el Resultado 2.3.1. Ahora veamos que también es un estimador suficiente para λ . Como la distribución Poisson es discreta, entonces para verificar la suficiencia de \bar{X} podemos verificar*

que $P(X_1 = x_1, \dots, X_n = x_n | \bar{X} = x)$ no depende de λ , tenemos:

$$\begin{aligned}
& P(X_1 = x_1, \dots, X_n = x_n | \bar{X} = x) \\
&= \frac{P(X_1 = x_1, \dots, X_n = x_n, \bar{X} = x)}{P(\bar{X} = x)} \\
&= \frac{P(X_1 = x_1, \dots, X_n = x_n, \sum_{i=1}^n X_i = nx)}{P(\sum_{i=1}^n X_i = nx)} \\
&= \frac{P(X_1 = x_1, \dots, X_{n-1} = x_{n-1}, X_n = nx - \sum_{i=1}^{n-1} x_i)}{P(\sum_{i=1}^n X_i = nx)} \\
&= \frac{P(X_1 = x_1) \cdots P(X_{n-1} = x_{n-1}) P(X_n = nx - \sum_{i=1}^{n-1} x_i)}{P(\sum_{i=1}^n X_i = nx)} \\
&= \frac{e^{-n\lambda} \lambda^{x_1} \cdots \lambda^{x_{n-1}} \lambda^{nx - \sum_{i=1}^{n-1} x_i}}{e^{-n\lambda} (n\lambda)^{nx}} \\
&= \frac{x_1! \cdots x_{n-1}! (nx - \sum_{i=1}^{n-1} x_i)!}{(nx)!} \\
&= \frac{n^{nx} x_1! \cdots x_{n-1}! (nx - \sum_{i=1}^{n-1} x_i)!}{(nx)!},
\end{aligned}$$

claramente la anterior probabilidad condicional no depende de λ , de donde se concluye que \bar{X} es suficiente para λ . Utilizando un razonamiento completamente análogo, se puede ver que $\sum_{i=1}^n X_i$ también es suficiente para θ .

Ahora, como se vio en el anterior ejemplo, utilizar la definición para demostrar que un estimador es suficiente puede resultar un poco tedioso, pues el cómputo de una probabilidad condicional, en general, no es sencillo. El siguiente teorema, conocido como el criterio de factorización de Fisher-Neyman, es útil para verificar que un estimador es suficiente para el parámetro desconocido.

Resultado 2.4.4. Dada una muestra aleatoria X_1, \dots, X_n con función de densidad $f(x_i, \theta)$, y sea $T = T(X_1, \dots, X_n)$ un estimador de θ , entonces T es suficiente para θ , si y solo si, se puede factorizar la función de verosimilitud $L(\theta, x_1, \dots, x_n)$ como

$$L(\theta, x_1, \dots, x_n) = g(t(x_1, \dots, x_n), \theta) h(x_1, \dots, x_n)$$

Proof. Se hará la prueba para cuando las variables X_1, \dots, X_n y T son discretas, la demostración es como sigue:

(\Leftarrow) Primero supongamos que se tiene la factorización $L(\theta, x_1, \dots, x_n) = g(t(x_1, \dots, x_n), \theta) h(x_1, \dots, x_n)$, y veamos que T es suficiente para θ , es decir, veamos que $P(X_1 = x_1, \dots, X_n = x_n | T = t)$ no depende de θ . En primer lugar, si $t \neq T(x_1, \dots, x_n)$ entonces la probabilidad vale 0 y por consiguiente no depende

de θ . Ahora si $t = T(x_1, \dots, x_n)$, tenemos:

$$\begin{aligned} P(X_1 = x_1, \dots, X_n = x_n | T = t) &= \frac{P(X_1 = x_1, \dots, X_n = x_n, T = t)}{P(T = t)} \\ &= \frac{P(X_1 = x_1, \dots, X_n = x_n)}{P(T = t)} \\ &= \frac{g(t(x_1, \dots, x_n), \theta)h(x_1, \dots, x_n)}{P(T = t)} \end{aligned}$$

Al definir A como el conjunto de valores de x_1, \dots, x_n que son enviados al valor t mediante la variable T , tenemos que

$$\begin{aligned} P(X_1 = x_1, \dots, X_n = x_n | T = t) &= \frac{g(t(x_1, \dots, x_n), \theta)h(x_1, \dots, x_n)}{\sum_A P(X_1 = x_1, \dots, X_n = x_n)} \\ &= \frac{g(t(x_1, \dots, x_n), \theta)h(x_1, \dots, x_n)}{\sum_A g(t(x_1, \dots, x_n), \theta)h(x_1, \dots, x_n)} \\ &= \frac{g(t, \theta)h(x_1, \dots, x_n)}{\sum_A g(t, \theta)h(x_1, \dots, x_n)} \\ &= \frac{h(x_1, \dots, x_n)}{\sum_A h(x_1, \dots, x_n)}, \end{aligned}$$

el cual no depende del valor θ .

(\Rightarrow) Ahora supongamos que T es suficiente para θ , veamos que se tiene la factorización $L(\theta, x_1, \dots, x_n) = g(t(x_1, \dots, x_n), \theta)h(x_1, \dots, x_n)$ para algunas funciones g y h . Tenemos que

$$\begin{aligned} L(\theta, x_1, \dots, x_n) &= P(X_1 = x_1, \dots, X_n = x_n) \\ &= P(X_1 = x_1, \dots, X_n = x_n, T = t(x_1, \dots, x_n)) \\ &= \underbrace{P(T = t(x_1, \dots, x_n))}_g \underbrace{P(X_1 = x_1, \dots, X_n = x_n | T = t(x_1, \dots, x_n))}_h, \end{aligned}$$

la primera probabilidad no depende de θ por la suficiencia de T , y la segunda probabilidad depende de $t(x_1, \dots, x_n)$ y de θ , y hemos logrado obtener la factorización de $L(\theta, x_1, \dots, x_n)$.

La prueba para cuando X_1, \dots, X_n y T son continuas es más complicado, el lector puede consultarlo en Lehmann & Romano (2005, pg.20). \square

Ahora, retomamos el Ejemplo 2.3.2. utilizando el criterio de factorización para

ilustrar la utilidad del resultado. Tenemos:

$$\begin{aligned} L(\lambda, x_1, \dots, x_n) &= \frac{e^{-n\lambda} \lambda^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n x_i!} \prod_{i=1}^n I_{\{0,1,\dots\}}(x_i) \\ &= \underbrace{e^{-n\lambda} \lambda^{n\bar{x}}}_{g(\bar{x}, \lambda)} \underbrace{\frac{\prod_{i=1}^n I_{\{0,1,\dots\}}(x_i)}{\prod_{i=1}^n x_i!}}_{h(x_1, \dots, x_n)}, \end{aligned}$$

con la anterior expresión se logra escribir la función de verosimilitud en forma del Resultado 2.3.4., de donde se concluye que \bar{X} es suficiente para λ . Nótese que la anterior factorización no es única, pues también se tiene que:

$$L(\lambda, x_1, \dots, x_n) = \underbrace{e^{-n\lambda} \lambda^{\sum x_i}}_{g(\sum x_i, \lambda)} \underbrace{\frac{\prod_{i=1}^n I_{\{0,1,\dots\}}(x_i)}{\prod_{i=1}^n x_i!}}_{h(x_1, \dots, x_n)},$$

de donde se concluye que también $\sum_{i=1}^n X_i$ es suficiente para λ .

Utilizando este criterio, se puede verificar fácilmente que en muestras provenientes de las distribuciones $Exp(\theta)$, $Bernoulli(\theta)$, $N(\theta, \sigma^2)$ con σ^2 conocida, las estadísticas \bar{X} y $\sum_{i=1}^n X_i$ son suficientes para θ .

El criterio de factorización presentado en el Resultado 2.3.4. cubre solamente a las distribuciones con un parámetro desconocido, también existe la versión general para distribuciones con más de un parámetro. Dado que en mayoría de los casos, no se trabaja con distribuciones con más de dos parámetros, se presenta únicamente la versión para distribuciones con dos parámetros.

Resultado 2.4.5. *Dada una muestra aleatoria X_1, \dots, X_n con función de densidad $f(x_i, \theta_1, \theta_2)$, y sea $T_1 = T_1(X_1, \dots, X_n)$ y $T_2 = T_2(X_1, \dots, X_n)$ son estimadores de θ_1 y θ_2 , entonces T_1 y T_2 son suficientes para θ_1 y θ_2 , si y solo si, se puede factorizar la función de verosimilitud $L(\theta, x_1, \dots, x_n)$ como*

$$L(\theta, x_1, \dots, x_n) = g(t_1(x_1, \dots, x_n), \theta_1, t_2(x_1, \dots, x_n), \theta_2) h(x_1, \dots, x_n)$$

La utilidad del resultado se ilustra en el siguiente ejemplo.

Ejemplo 2.4.3. *La distribución Beta se usa generalmente para modelar probabilidades y la función de densidad de probabilidad está dada por:*

$$f(x) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1} I_{(0,1)}(x).$$

Dada una muestra aleatoria de tamaño n , la función de verosimilitud está dada

por

$$\begin{aligned} L(\alpha, \beta) &= \frac{\Gamma(\alpha + \beta)^n}{\Gamma(\alpha)^n \Gamma(\beta)^n} \prod_{i=1}^n x_i^{\alpha-1} \prod_{i=1}^n (1-x_i)^{\beta-1} \prod_{i=1}^n I_{(0,1)}(x_i) \\ &= \frac{\Gamma(\alpha + \beta)^n}{\Gamma(\alpha)^n \Gamma(\beta)^n} \underbrace{\left(\prod_{i=1}^n x_i \right)^{\alpha-1} \left(\prod_{i=1}^n (1-x_i) \right)^{\beta-1}}_{g(t_1, t_2, \alpha, \beta)} \underbrace{\prod_{i=1}^n I_{(0,1)}(x_i)}_{h(x_1, \dots, x_n)}. \end{aligned}$$

Y podemos concluir que las estadísticas $\prod_{i=1}^n X_i$ y $\prod_{i=1}^n (1 - X_i)$ son suficientes para α y β .

Para las distribuciones pertenecientes a la familia exponencial, siempre podemos encontrar estadísticas suficientes para el parámetro, el resultado se da a continuación.

Resultado 2.4.6. Dada una muestra aleatoria X_1, \dots, X_n proveniente de una distribución $f(x, \theta)$ perteneciente a la familia exponencial, es decir,

$$f(x, \theta) = h(x)c(\theta) \exp\{d(\theta)T(x)\},$$

entonces la estadística $\sum_{i=1}^n T(X_i)$ es una estadística suficiente para θ .

Proof. El resultado es trivial usando el criterio de factorización de Fisher-Neyman. Por (1.2.2), tenemos que la función de verosimilitud de una muestra aleatoria con función de densidad perteneciente a la familia exponencial está dada por

$$L(\theta, x_1, \dots, x_n) = c(\theta)^n \left[\prod_{i=1}^n h(x_i) \right] \exp \left\{ d(\theta) \sum_{i=1}^n T(x_i) \right\}.$$

Al tomar $\sum_{i=1}^n T(X_i)$ como la estadística T y $c(\theta)^n \exp \{d(\theta) \sum_{i=1}^n T(x_i)\}$ como la función $g(t(x_1, \dots, x_n), \theta)$, y el restante como $h(x_1, \dots, x_n)$, se tiene que $\sum_{i=1}^n T(X_i)$ es suficiente para θ . \square

Para ilustrar la utilidad del resultado, consideramos una muestra proveniente de la distribución $Ber(p)$, esta distribución pertenece a la familia exponencial, puesto que

$$\begin{aligned} f(x, p) &= p^x (1-p)^{1-x} I_{\{0,1\}}(x) \\ &= \left(\frac{p}{1-p} \right)^x I_{\{0,1\}}(x) \\ &= (1-p) I_{\{0,1\}}(x) \exp \left\{ x \ln \frac{p}{1-p} \right\}, \end{aligned}$$

entonces $T(x) = x$, y por el anterior resultado, se tiene que $\sum_{i=1}^n T(X_i) = \sum_{i=1}^n X_i$ es una estadística suficiente para p . Teniendo en cuenta que una estadística

suficiente resume toda la información contenida en una muestra acerca de un parámetro θ , lo anterior nos indica que en un conjunto de observaciones del tipo 0 y 1 provenientes de $Ber(p)$, para el efecto de estimación de p , basta con observar la suma de las observaciones, de esta forma podemos reducir un gran volumen de datos en solo un dato, y la estimación obtenida de p no se ve afectado ⁸.

Por otro lado, como la presentación de una función de densidad de la familia exponencial no es única, entonces podemos encontrar diferentes estadísticas suficientes para un mismo parámetro. En efecto, la densidad de la distribución $Ber(p)$ también puede escribir como:

$$f(x, p) = (1 - p)I_{\{0,1\}}(x) \exp \left\{ \frac{x}{n} \left[n \ln \frac{p}{1-p} \right] \right\},$$

de esta manera, $T(x) = x/n$, así también se probó que $\bar{X} = \sum_{i=1}^n X_i/n$ es una estadística suficiente para p .

Ahora, en distribuciones biparamétricas también podemos encontrar fácilmente estadísticas suficientes si éstas pertenecen a la familia exponencial. El siguiente resultado es el análogo al Resultado 2.3.6. para distribuciones biparamétricas.

Resultado 2.4.7. *Dada una muestra aleatoria X_1, \dots, X_n proveniente de una distribución $f(x_i, \theta_1, \theta_2)$ perteneciente a la familia exponencial biparamétrica de la forma*

$$f(x_i, \theta_1, \theta_2) = c(\theta)h(x) \exp\{d(\theta)'T(x)\},$$

donde $\theta = (\theta_1, \theta_2)$, $d(\theta) = (d_1(\theta), d_2(\theta))'$ y $T(x) = (T_1(x), T_2(x))'$, entonces las estadística $\sum_{i=1}^n T_1(X_i)$ y $\sum_{i=1}^n T_2(X_i)$ son estadísticas suficientes para θ_1 y θ_2 .

Proof. La prueba es análogo al caso para distribuciones uniparamétricas, usando el criterio de factorización de Fisher-Neyman. Tenemos que la función de verosimilitud está dada por

$$\begin{aligned} & L(\theta_1, \theta_2, x_1, \dots, x_n) \\ &= c(\theta)^n \left\{ \prod_{i=1}^n h(x_i) \right\} \exp \left\{ d(\theta)' \sum_{i=1}^n T(x_i) \right\} \\ &= c(\theta_1, \theta_2)^n \left\{ \prod_{i=1}^n h(x_i) \right\} \exp \left\{ (d_1(\theta), d_2(\theta))' \sum_{i=1}^n \begin{pmatrix} T_1(x_i) \\ T_2(x_i) \end{pmatrix} \right\} \\ &= \underbrace{\left\{ \prod_{i=1}^n h(x_i) \right\}}_{h(x_1, \dots, x_n)} \underbrace{c(\theta_1, \theta_2)^n \exp \left\{ d_1(\theta) \sum_{i=1}^n T_1(x_i) + d_2(\theta) \sum_{i=1}^n T_2(x_i) \right\}}_{g(t_1, \theta_1, t_2, \theta_2)}, \end{aligned}$$

de esta forma, tenemos que $\sum_{i=1}^n T_1(X_i)$ y $\sum_{i=1}^n T_2(X_i)$ son suficientes para θ_1 y θ_2 . \square

⁸Vea el Ejemplo 2.2.3. donde la estimación de p se llevó a cabo usando solamente la suma de las observaciones.

Ilustramos la aplicación del resultado en el siguiente ejemplo.

Ejemplo 2.4.4. Sea X_1, \dots, X_n una muestra aleatoria con distribución $N(\mu, \sigma^2)$, el anterior resultado servirá para encontrar estadísticas suficientes para μ y σ^2 . La función de densidad pertenece a la familia exponencial biparamétrica pues se puede escribir de la forma

$$f(x, \mu, \sigma^2) = \exp \left\{ \left(\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2} \right) \begin{pmatrix} x \\ x^2 \end{pmatrix} \right\} \exp \left\{ -\frac{\mu^2}{2\sigma^2} \right\} (2\pi\sigma^2)^{-1/2},$$

entonces $T_1(x) = x$ y $T_2(x) = x^2$, entonces el resultado anterior indica que las estadísticas $\sum_{i=1}^n X_i$ y $\sum_{i=1}^n X_i^2$ son suficientes y completas para μ y σ^2 .

Volviendo al tópico de la evaluación de la calidad de un estimador, una inquietud que había surgido al tener en cuenta que un buen estimador debe ser insesgado con varianza pequeña es dado un estimador insesgado, cómo construir otro insesgado con varianza menor. El siguiente teorema de Rao-Blackwell⁹ afirma que al combinar un estimador insesgado con una estadística suficiente, se puede lograr un estimador insesgado con una varianza menor.

Resultado 2.4.8. Sea X_1, \dots, X_n una muestra aleatoria con función de densidad $f(x_i, \theta)$, si T_1 es un estimador insesgado para una función de θ , $g(\theta)$, y T_2 es suficiente para θ , entonces el estimador $T = E(T_1|T_2)$ es insesgado para $g(\theta)$ y tiene varianza menor que T_1 .

Proof. En primer lugar $E(T) = E(E(T_1|T_2)) = E(T_1) = g(\theta)$. Ahora, en término de varianza, tenemos

$$\begin{aligned} \text{Var}(T_1) &= \text{Var}(E(T_1|T_2)) + E(\text{Var}(T_1|T_2)) \\ &= \text{Var}(T) + E(\text{Var}(T_1|T_2)) \\ &\geq \text{Var}(T) \end{aligned}$$

□

La importancia de este teorema radica en que para estimar una función de un parámetro desconocido $g(\theta)$ si tenemos un estimador insesgado T_1 podemos, con base en este, construir un mejor estimador que T_1 , siempre y cuando se dispone de un estimador suficiente para θ .

Para un mejor entendimiento del teorema revisamos, en primer lugar, el concepto de la esperanza condicional. Lo más importante que hay que aclarar que la expresión $E(T_1|T_2)$ en el resultado anterior no es una constante, sino una variable aleatoria. Para ilustrar esto, considere el siguiente ejemplo:

⁹El teorema fue establecido por estadístico hindú Calyampudi Radhakrishna Rao y americano David Blackwell

Ejemplo 2.4.5. Dadas variables aleatorias X e Y con función de densidad de probabilidad conjunta dada por:

$$f(x, y) = \begin{cases} e^{-y} & \text{si } 0 < x < y \\ 0 & \text{en otro caso} \end{cases},$$

Para calcular $E(X|Y)$ primero recordamos que ésta es una función con dominio igual al rango de Y y a cada valor y lo envía a la esperanza $E(X|Y = y)$. Entonces dado y , para calcular $E(X|Y = y)$, primero se calcula la función de densidad condicional $f_{X|Y}(x|y) = f(x, y)/f_Y(y)$, en nuestro caso,

$$f_{X|Y}(x|y) = \begin{cases} y^{-1} & \text{si } 0 < x < y \\ 0 & \text{en otro caso} \end{cases},$$

de donde para un valor particular que toma la variable Y , se puede calcular $E(X|Y = y)$ como

$$E(X|Y = y) = \int_{-\infty}^{\infty} x f_{X|Y}(x|y) dx,$$

nótese que la anterior esperanza condicional es un número, función de y . En nuestra caso, tenemos que $E(X|Y = y) = y/2$. Entonces $E(X|Y)$ envía cada valor y a $y/2$, es decir $E(X|Y) = Y/2$, la cual claramente es una variable aleatoria.

En general, calcular una esperanza condicional $E(X|Y)$ puede implicar cálculos tediosos, pero en algunos casos puede ser trivial como lo indica el siguiente resultado.

Resultado 2.4.9. Si X e Y son variables aleatorias, y X puede escribir como una función de Y , entonces $E(X|Y) = X$.

Aclarado el concepto de la esperanza condicional, volvemos al teorema de Rao Blackwell, en la demostración no se utilizó el hecho de que T_2 sea una estadística suficiente, por lo tanto, podemos intuir que al condicionar T_1 en cualquier otra estadística, digamos S , también se puede mejorar la calidad del estimador, en el término de que $Var(T_1|S)$ será menor que $Var(T_1)$. Lo anterior es cierto, pero puede suceder que $T_1|S$ dependa del parámetro θ y deja de ser un estimador (Vea Casella & Berger (2002, Ejemplo 7.3.18, pg. 343) para un ejemplo donde $T_1|S$ depende de θ). Por esta razón, se necesita que el condicionamiento sea sobre una estadística suficiente para garantizar que la resultante esperanza condicional no dependa del parámetro y pueda ser usado como un estimador.

2.4.3 Estimadores UMVUE

El teorema de Rao Blackwell plantea la posibilidad de un proceso continuo de construcción de estimadores insesgados con varianzas cada vez menores, entonces

la inquietud que surge ahora es si podemos construir estimadores de varianza cada vez menor o podemos encontrar un estimador insesgado T de tal forma que ya no existe ningún otro estimador insesgado con varianza menor que $Var(T)$. Si existe alguna cota inferior para la varianza de los estimadores, y se encuentra un estimador insesgado T con varianza igual a esta cota, se podrá concluir que no habrá otro estimador insesgado con varianza más pequeño que ésta, y se podrá afirmar que T es la mejor de todos los estimadores insesgados. Esta cota existe efectivamente y se denomina la cota de Cramer Rao, y no solo es la cota inferior para la varianza de los estimadores insesgados, sino también puede ser cota inferior para la varianza de todos los estimadores. Para estudiar la cota de Cramer Rao, introducimos algunos conceptos preliminares.

Definición 2.4.4. Dada X una variable aleatoria con función de densidad $f(x, \theta)$, donde θ es el parámetro de la distribución, y además existe $\frac{\partial}{\partial \theta} \ln f(x, \theta)$, entonces se define la información contenida en X acerca de θ como

$$I_X(\theta) = E \left\{ \left[\frac{\partial}{\partial \theta} \ln f(X, \theta) \right]^2 \right\}.$$

Nótese que en la anterior definición, $f(X, \theta)$ no es la función de densidad $f(x, \theta)$, sino la variable X transformada a través de la función f , es decir, $f(X, \theta)$ es una variable aleatoria, y por consiguiente, tiene sentido calcular la esperanza. Ahora en algunas situaciones, existe una definición equivalente que mide esta cantidad de información, y puede resultar más fácil el cálculo.

Resultado 2.4.10. En la anterior definición, si además existe $\frac{\partial^2}{\partial \theta^2} \ln f(x, \theta)$, entonces se tiene que

$$I_X(\theta) = -E \left\{ \frac{\partial^2}{\partial \theta^2} \ln f(X, \theta) \right\}.$$

Proof. Tenemos:

$$\begin{aligned} -E \left\{ \frac{\partial^2}{\partial \theta^2} \ln f(X, \theta) \right\} &= -E \left\{ \frac{\partial}{\partial \theta} \left[\frac{1}{f(X, \theta)} \frac{\partial f(X, \theta)}{\partial \theta} \right] \right\} \\ &= -E \left\{ -\frac{1}{f^2(X, \theta)} \left[\frac{\partial f(X, \theta)}{\partial \theta} \right]^2 + \frac{1}{f(X, \theta)} \frac{\partial^2 f(X, \theta)}{\partial \theta^2} \right\} \\ &= E \left\{ \frac{1}{f^2(X, \theta)} \left[\frac{\partial f(X, \theta)}{\partial \theta} \right]^2 \right\} - E \left\{ \frac{1}{f(X, \theta)} \frac{\partial^2 f(X, \theta)}{\partial \theta^2} \right\}, \end{aligned}$$

La última esperanza vale cero, puesto que si X es una variable continua,

$$\begin{aligned} E \left\{ \frac{1}{f(X, \theta)} \frac{\partial^2 f(X, \theta)}{\partial \theta^2} \right\} &= \int_{\mathbb{R}} \frac{1}{f(x, \theta)} \frac{\partial^2 f(x, \theta)}{\partial \theta^2} f(x, \theta) dx \\ &= \int_{\mathbb{R}} \frac{\partial^2 f(x, \theta)}{\partial \theta^2} dx \\ &= \frac{\partial^2}{\partial \theta^2} \int_{\mathbb{R}} f(x, \theta) dx \\ &= \frac{\partial^2}{\partial \theta^2} (1) = 0. \end{aligned}$$

Y si X es discreta, suponga que los valores que toma son x_1, x_2, \dots , entonces,

$$\begin{aligned} E \left\{ \frac{1}{f(X, \theta)} \frac{\partial^2 f(X, \theta)}{\partial \theta^2} \right\} &= \sum_i \frac{1}{f(x_i, \theta)} \frac{\partial^2 f(x_i, \theta)}{\partial \theta^2} P(X = x_i) \\ &= \sum_i \frac{\partial^2 f(x_i, \theta)}{\partial \theta^2} \\ &= \frac{\partial^2}{\partial \theta^2} \sum_i f(x_i, \theta) \\ &= \frac{\partial^2}{\partial \theta^2} (1) = 0. \end{aligned}$$

En conclusión,

$$-E \left\{ \frac{\partial^2}{\partial \theta^2} \ln f(X, \theta) \right\} = E \left\{ \frac{1}{f^2(X, \theta)} \left[\frac{\partial f(X, \theta)}{\partial \theta} \right]^2 \right\}.$$

Ahora,

$$\begin{aligned} I_X(\theta) &= E \left\{ \left[\frac{\partial}{\partial \theta} \ln f(X, \theta) \right]^2 \right\} \\ &= E \left\{ \left[\frac{1}{f(X, \theta)} \frac{\partial f(X, \theta)}{\partial \theta} \right]^2 \right\} \\ &= E \left\{ \frac{1}{f^2(X, \theta)} \left[\frac{\partial f(X, \theta)}{\partial \theta} \right]^2 \right\} \\ &= -E \left\{ \frac{\partial^2}{\partial \theta^2} \ln f(X, \theta) \right\}, \end{aligned}$$

y el resultado queda demostrado. \square

Las anteriores definiciones introducen la información contenida en una variable, sin embargo, cuando tenemos disponible una muestra aleatoria, es necesario definir la información contenida en una muestra aleatoria acerca de algún parámetro.

Definición 2.4.5. Dada X_1, \dots, X_n variables aleatorias con función de densidad $f(x_i, \theta)$, donde θ es el parámetro de la distribución, y además existe $\frac{\partial}{\partial \theta} \ln \prod_{i=1}^n f(x_i, \theta)$, entonces se define la información contenida en la muestra aleatoria acerca de θ como

$$I_{X_1, \dots, X_n}(\theta) = E \left\{ \left[\frac{\partial}{\partial \theta} \ln \prod_{i=1}^n f(X_i, \theta) \right]^2 \right\}.$$

Recordemos que en una muestra aleatoria, las variables tienen la misma distribución de probabilidad, además son independiente, entonces es natural pensar que cada variable debe aportar la misma cantidad de información, es decir, la información contenida en una muestra de tamaño n debe ser igual a n veces la información contenida en cualquier variable de la muestra. El siguiente resultado confirma esta intuición.

Resultado 2.4.11. Dada X_1, \dots, X_n una muestra aleatoria, entonces

$$I_{X_1, \dots, X_n}(\theta) = nI_X(\theta),$$

donde $I_X(\theta) = I_{X_i}(\theta)$, con $i = 1, \dots, n$. Es decir, en una muestra aleatoria, cada variable aporta la misma cantidad de información, y la cantidad total de información en la muestra es la suma de la información en cada variable.

Proof.

$$\begin{aligned} I_{X_1, \dots, X_n}(\theta) &= E \left\{ \left[\frac{\partial}{\partial \theta} \ln \prod_{i=1}^n f(X_i, \theta) \right]^2 \right\} \\ &= E \left\{ \left[\sum_{i=1}^n \frac{\partial}{\partial \theta} \ln f(X_i, \theta) \right]^2 \right\} \\ &= E \left\{ \sum_{i=1}^n \left[\frac{\partial}{\partial \theta} \ln f(X_i, \theta) \right]^2 \right\} + \\ &\quad \underbrace{E \left\{ \sum_{\substack{i,j=1 \\ i \neq j}}^n \left[\frac{\partial}{\partial \theta} \ln f(X_i, \theta) \frac{\partial}{\partial \theta} \ln f(X_j, \theta) \right] \right\}}_{=0, \text{ por la independencia entre } X_i \text{ y } X_j} \\ &= \sum_{i=1}^n E \left\{ \left[\frac{\partial}{\partial \theta} \ln f(X_i, \theta) \right]^2 \right\} \\ &= \sum_{i=1}^n I_X(\theta) = nI_X(\theta). \end{aligned}$$

□

Ilustramos el cálculo de la información contenida en una muestra en el siguiente ejemplo, y posteriormente presentamos cómo este concepto resulta útil en la definición de la cota de Cramer Rao.

Ejemplo 2.4.6. Sea X_1, \dots, X_n una muestra aleatoria proveniente de la distribución $N(\mu, \sigma^2)$, la información contenida en la muestra acerca de μ es n/σ^2 . Para verificar esta afirmación, calculamos la información acerca de μ en una variable X con distribución $N(\mu, \sigma^2)$. Tenemos:

$$\begin{aligned} I_X(\mu) &= -E \left\{ \frac{\partial^2}{\partial \mu^2} \ln f(X, \theta) \right\} \\ &= -E \left\{ \frac{\partial^2}{\partial \mu^2} \left[-\frac{1}{2} \ln 2\pi\sigma^2 - \frac{1}{2\sigma^2} (X - \mu)^2 \right] \right\} \\ &= -E \left\{ \frac{\partial}{\partial \mu} \left[\frac{X - \mu}{\sigma^2} \right] \right\} \\ &= -E \left\{ -\frac{1}{\sigma^2} \right\} \\ &= \frac{1}{\sigma^2}. \end{aligned}$$

Ahora, usando el Resultado 2.3.4, se tiene que $I_{X_1, \dots, X_n}(\mu) = n/\sigma^2$.

Nótese que esta información, en primer lugar, depende del tamaño n de manera que entre más grande sea la muestra, hay mayor información acerca de μ ; en segundo lugar, entre más pequeña sea la varianza σ^2 , la cantidad de información acerca de μ también incrementa, esto es natural, puesto que si σ^2 es pequeña, los datos de la muestra están muy concentrados alrededor de μ , entonces estos datos aportan más información que otros datos con más dispersión.

Cuando se introdujo el concepto de una estadística suficiente, su interpretación es que contiene toda la información de la muestra acerca de algún parámetro, esta información se puede entender como la información de Fisher, y el siguiente resultado provee la respectiva sustentación.

Resultado 2.4.12. Dada una muestra aleatoria X_1, \dots, X_n con función de densidad $f(x_i, \theta)$, y sea T un estimador suficiente para θ , entonces la información contenida en T acerca de θ es la misma información contenida en la muestra aleatoria acerca de θ

Proof. Lo que probaremos es $I_T(\theta) = I_{X_1, \dots, X_n}(\theta)$. Tenemos, en primer lugar,

$$\begin{aligned} f(X_1, \dots, X_n | T) &= \frac{f(X_1, \dots, X_n, T)}{f(T)} \\ &= \frac{g(T, \theta)h(X_1, \dots, X_n)}{f(T)} \quad \text{usando el criterio de factorización,} \end{aligned}$$

de donde tenemos que

$$f(T) = \frac{g(T, \theta)h(X_1, \dots, X_n)}{f(X_1, \dots, X_n | T)}.$$

Ahora usando la función de densidad de T calculamos la información contenida en T como

$$\begin{aligned}
I_T(\theta) &= E \left\{ \left[\frac{\partial}{\partial \theta} \ln f(T) \right]^2 \right\} \\
&= E \left\{ \left[\frac{\partial}{\partial \theta} \ln \frac{g(T, \theta) h(X_1, \dots, X_n)}{f(X_1, \dots, X_n | T)} \right]^2 \right\} \\
&= E \left\{ \left[\frac{\partial}{\partial \theta} (\ln g(T, \theta) + \ln h(X_1, \dots, X_n) - \ln f(X_1, \dots, X_n | T)) \right]^2 \right\} \\
&= E \left\{ \left[\frac{\partial}{\partial \theta} \ln g(T, \theta) \right]^2 \right\},
\end{aligned}$$

pues $h(X_1, \dots, X_n)$ no depende de θ , y tampoco $f(X_1, \dots, X_n | T)$ por la definición de suficiencia de T .

Ahora, calculamos la información contenida en la muestra acerca de θ , tenemos

$$\begin{aligned}
I_{X_1, \dots, X_n}(\theta) &= E \left\{ \left[\frac{\partial}{\partial \theta} \ln f(X_1, \dots, X_n) \right]^2 \right\} \\
&= E \left\{ \left[\frac{\partial}{\partial \theta} \ln g(T, \theta) h(X_1, \dots, X_n) \right]^2 \right\} \\
&= E \left\{ \left[\frac{\partial}{\partial \theta} (\ln g(T, \theta) - \ln h(X_1, \dots, X_n)) \right]^2 \right\} \\
&= E \left\{ \left[\frac{\partial}{\partial \theta} \ln g(T, \theta) \right]^2 \right\}.
\end{aligned}$$

Y podemos concluir que $I_T(\theta) = I_{X_1, \dots, X_n}(\theta)$ de donde se concluye que la información contenida en T con respecto a θ es la misma información contenida en la muestra X_1, \dots, X_n . \square

Ahora, como se mencionaba anteriormente, el concepto de información de Fisher permite encontrar una cota inferior para la varianza de los estimadores, este se enuncia en la famosa desigualdad de Cramer Rao, y se presenta a continuación:

Resultado 2.4.13. *Dada X_1, \dots, X_n variables aleatorias con distribución de probabilidad $f(x_i, \theta)$, y T es un estimador para $g(\theta)$, si*

$$\frac{\partial}{\partial \theta} E(T) = \int \frac{\partial}{\partial \theta} t(x_1, \dots, x_n) f(x_1, \dots, x_n) dx_1 \cdots dx_n \quad (2.4.4)$$

y $Var(T) < \infty$, entonces

$$Var(T) \geq \frac{(\frac{\partial}{\partial \theta} E(T))^2}{I_{X_1, \dots, X_n}(\theta)}.$$

Y $\frac{(\frac{\partial}{\partial \theta} E(T))^2}{I_{X_1, \dots, X_n}(\theta)}$ es llamado la cota de Cramer Rao.

Proof. La demostración del resultado se basa en el hecho de que el coeficiente de correlación entre dos variables es siempre menor o igual a 1. Entonces para las variables T y $\frac{\partial}{\partial \theta} \ln \prod_{i=1}^n f(X_i, \theta)$, se tiene que

$$1 \geq Corr(T, \frac{\partial}{\partial \theta} \ln \prod_{i=1}^n f(X_i, \theta)) = \frac{Cov(T, \frac{\partial}{\partial \theta} \ln \prod_{i=1}^n f(X_i, \theta))}{\sqrt{Var(T)Var(\frac{\partial}{\partial \theta} \ln \prod_{i=1}^n f(X_i, \theta))}},$$

el cual es equivalente a

$$1 \geq \frac{Cov(T, \frac{\partial}{\partial \theta} \ln \prod_{i=1}^n f(X_i, \theta))^2}{Var(T)Var(\frac{\partial}{\partial \theta} \ln \prod_{i=1}^n f(X_i, \theta))}$$

de donde se tiene

$$Var(T) \geq \frac{Cov(T, \frac{\partial}{\partial \theta} \ln \prod_{i=1}^n f(X_i, \theta))^2}{Var(\frac{\partial}{\partial \theta} \ln \prod_{i=1}^n f(X_i, \theta))}. \quad (2.4.5)$$

Ahora

$$\begin{aligned} Var(\frac{\partial}{\partial \theta} \ln \prod_{i=1}^n f(X_i, \theta)) &= E \left\{ \left[\frac{\partial}{\partial \theta} \ln \prod_{i=1}^n f(X_i, \theta) \right]^2 \right\} - \underbrace{\left(E \left[\frac{\partial}{\partial \theta} \ln \prod_{i=1}^n f(X_i, \theta) \right] \right)^2}_{=0} \\ &= E \left\{ \left[\frac{\partial}{\partial \theta} \ln \prod_{i=1}^n f(X_i, \theta) \right]^2 \right\} \\ &= I_{X_1, \dots, X_n}(\theta), \end{aligned}$$

puesto que

$$\begin{aligned} E \left[\frac{\partial}{\partial \theta} \ln \prod_{i=1}^n f(X_i, \theta) \right] &= \sum_{i=1}^n \int_{\mathbb{R}} \frac{\partial}{\partial \theta} \ln f(x_i, \theta) f(x_i, \theta) dx_i \\ &= \sum_{i=1}^n \int_{\mathbb{R}} \frac{1}{f(x_i, \theta)} \frac{\partial f(x_i, \theta)}{\partial \theta} f(x_i, \theta) dx_i \\ &= \sum_{i=1}^n \frac{\partial}{\partial \theta} \int_{\mathbb{R}} f(x_i, \theta) dx_i \\ &= \sum_{i=1}^n \frac{\partial}{\partial \theta} (1) = 0, \end{aligned}$$

si las variables X_1, \dots, X_n son continuas. Cuando son discretas, se tiene análogamente.

Por otro lado,

$$\begin{aligned}
& Cov(T, \frac{\partial}{\partial \theta} \ln \prod_{i=1}^n f(X_i, \theta)) \\
&= E \left[T \frac{\partial}{\partial \theta} \ln \prod_{i=1}^n f(X_i, \theta) \right] \quad \text{pues } E(\frac{\partial}{\partial \theta} \ln \prod_{i=1}^n f(X_i, \theta)) = 0 \\
&= \int_{\mathbb{R}^n} t(x_1, \dots, x_n) \frac{\partial \ln f_{X_1, \dots, X_n}(x_1, \dots, x_n)}{\partial \theta} f_{X_1, \dots, X_n}(x_1, \dots, x_n) dx_1 \cdots dx_n \\
&= \int_{\mathbb{R}^n} t(x_1, \dots, x_n) \frac{\partial}{\partial \theta} f_{X_1, \dots, X_n}(x_1, \dots, x_n) dx_1 \cdots dx_n \\
&= \frac{\partial}{\partial \theta} \int_{\mathbb{R}^n} t(x_1, \dots, x_n) f_{X_1, \dots, X_n}(x_1, \dots, x_n) dx_1 \cdots dx_n \\
&= \frac{\partial}{\partial \theta} E(T).
\end{aligned}$$

Reemplazando las anteriores expresiones en (2.4.5), se tiene el resultado. \square

La condición en (2.4.4) es parte de las condiciones denominadas condiciones de regularidad, y se debe garantizar que esta condición se cumple para tener la validez de la desigualdad de Cramer Rao. Para las distribuciones pertenecientes a la familia exponencial, esta condición siempre se tiene, sin embargo, para distribuciones donde el rango de la variable depende del parámetro como la distribución uniforme, la desigualdad de Cramer Rao puede no ser cierto. El lector puede consultar Casella & Berger (2002, Ejemplo 7.3.13, pg.339) para ver un ejemplo donde ocurre esta situación.

Ahora, la cota de Cramer Rao dada en el anterior resultado no asume muchas condiciones como la independencia acerca de las variables, y tampoco características especiales del estimador T . Sin embargo, la mayor utilidad de la cota de Cramer Rao se da cuando las variables constituyen una muestra aleatoria y el estimador T sea insesgado para $g(\theta)$. En este caso $\frac{\partial}{\partial \theta} E(T) = \frac{\partial}{\partial g(\theta)} \theta = g'(\theta)$ y la desigualdad se convierte en

$$Var(T) \geq \frac{g'(\theta)}{nI_X(\theta)}.$$

Y en el caso cuando lo que se desea estimar es simplemente el parámetro de la distribución θ , $g(\theta) = \theta$, y la desigualdad se convierte en

$$Var(T) \geq \frac{1}{nI_X(\theta)}. \quad (2.4.6)$$

Ahora, si un estimador insesgado T tiene varianza igual a la cota de Cramer Rao, entonces cualquier otro estimador insesgado necesariamente tendrá varianza

más grande que T , es decir, T será el mejor entre todos los estimadores insesgados, y existe un nombre especial para estos estimadores que se presenta en la siguiente definición.

Definición 2.4.6. *Dada X_1, \dots, X_n una muestra aleatoria con función de densidad $f(x_i, \theta)$, y T un estimador insesgado para θ , si $\text{Var}(T) \leq \text{Var}(T^*)$ para todo θ para cualquier otro estimador T^* insesgado de θ , entonces se dice que T es un estimador insesgado de varianza uniformemente mínima, UMVUE¹⁰.*

Con respecto a estimadores UMVUE, tenemos varios comentarios.

- Como lo mencionado anteriormente, cuando un estimador insesgado tiene varianza igual a la cota de Cramer Rao, entonces este es un UMVUE, pero no necesariamente sucede lo contrario, o por lo menos, en la literatura estadística no hay un resultado que lo establezca. Es decir, un UMVUE no necesariamente tiene varianza igual a la cota de Cramer Rao.
- Una forma para verificar que un estimador sea UMVUE es ver que la varianza es igual a la cota de Cramer Rao, además hay que ver que es un estimador insesgado. En otras palabras, un estimador sesgado que tenga varianza igual a la cota de Cramer Rao no es UMVUE.
- Un estimador UMVUE tiene la varianza más pequeña entre todos los estimadores insesgados, pero puede existir un estimador sesgado T^* con varianza aún más pequeña. En este caso, se debe escoger entre el UMVUE y T^* , pues puede suceder que la ganancia en la varianza sea considerable y que T^* sea asintóticamente insesgado y de esta forma, corregir el sesgo aumentando el tamaño de muestra n .

Consideramos un ejemplo para ilustrar los estimadores UMVUE, en la sección anterior se mencionaba que en una muestra proveniente de la distribución $P(\theta)$, existen dos estimadores de momentos \bar{X} y S_n^2 , que mediante estudios de simulación, se vio que \bar{X} es mejor que S_n^2 , la razón teórica es que \bar{X} es UMVUE para θ , lo mostramos en el siguiente ejemplo.

Ejemplo 2.4.7. *Dada X_1, \dots, X_n una muestra aleatoria con distribución $P(\theta)$, el estimador \bar{X} como estimador de θ es UMVUE. En primer lugar, se vio en el Resultado 2.3.1. \bar{X} siempre es insesgado para el media poblacional, el cual en la distribución Poisson es igual al parámetro θ . Entonces tenemos que \bar{X} es insesgado para θ . Resta verificar que la varianza de \bar{X} es mínima entre todos los estimadores insesgados. Para eso, podemos calcular la cota de Cramer Rao, y ver que ésta es igual a $\text{Var}(\bar{X})$.*

Por el Resultado 2.3.1, $\text{Var}(\bar{X}) = \sigma^2/n$ donde σ^2 es la varianza de la distribución de probabilidad, que en este ejemplo es la distribución $P(\theta)$ cuya varianza es θ . En conclusión

$$\text{Var}(\bar{X}) = \frac{\theta}{n}.$$

¹⁰Por su sigla en inglés *Uniformly minimum variance unbiased estimator*.

Ahora calculamos la cota de Cramer Rao, tenemos

$$\begin{aligned}
I_{X_1, \dots, X_n}(\theta) &= nI_X(\theta) \\
&= -nE \left\{ \frac{\partial^2}{\partial \theta^2} \ln f(X, \theta) \right\} \\
&= -nE \left\{ \frac{\partial^2}{\partial \theta^2} [-\theta + X \ln \theta - \ln X!] \right\} \\
&= -nE \left\{ \frac{\partial}{\partial \theta} \left[-1 + \frac{X}{\theta} \right] \right\} \\
&= -nE \left\{ -\frac{X}{\theta^2} \right\} \\
&= \frac{n}{\theta},
\end{aligned}$$

de donde se concluye que la cota de Cramer Rao es θ/n , que es igual a la varianza de \bar{X} . En conclusión, \bar{X} es UMVUE para θ .

Por otro lado, de nuevo usando el Resultado 2.3.1, el otro estimador de momentos S_n^2 es insesgado para la varianza poblacional que en una distribución Poisson también corresponde al parámetro θ . Este insesgamiento se puede corroborar con la Figura 2.5 donde muestran las estimaciones obtenidas usando \bar{X} y S_n^2 en muestras de distribución Poisson, donde es claro que las estimaciones de S_n^2 estuvieron siempre alrededor del valor de θ , sin embargo como \bar{X} es UMVUE, siempre será mejor que S_n^2 como estimador de θ .

Como se vio en el anterior ejemplo, para ver que la varianza de un estimador sea igual a la cota de Cramer Rao, se necesita calcular la información de Fisher contenida en la muestra, y esto puede ser tedioso. Sin embargo, existe un resultado, que requiere tal vez menos operaciones algebraicas, que nos permite saber cuándo un estimador tiene la varianza igual a la cota de Cramer Rao.

Resultado 2.4.14. Dada X_1, \dots, X_n una muestra aleatoria con distribución de probabilidad $f(x_i, \theta)$, y $T = T(X_1, \dots, X_n)$ un estimador de $g(\theta)$. Si se tiene la siguiente factorización

$$\frac{\partial}{\partial \theta} \ln \prod_{i=1}^n f(x_i, \theta) = K(\theta)(t(x_1, \dots, x_n) - g(\theta)),$$

entonces la varianza de T es igual a la cota de Cramer Rao.

Proof. La condición

$$\frac{\partial}{\partial \theta} \ln \prod_{i=1}^n f(x_i, \theta) = K(\theta)(t(x_1, \dots, x_n) - g(\theta))$$

es equivalente a

$$\frac{\partial}{\partial \theta} \ln \prod_{i=1}^n f(x_i, \theta) = \alpha t(x_1, \dots, x_n) + \beta$$

para algunos constantes α y β que no depende de la muestra aleatoria, sino únicamente del parámetro θ . Y en consecuencia, tenemos que

$$\frac{\partial}{\partial \theta} \ln \prod_{i=1}^n f(X_i, \theta) = \alpha T + \beta.$$

Ahora recordando propiedades del coeficiente de correlación, tenemos que

$$\text{Corr}(T, \frac{\partial}{\partial \theta} \ln \prod_{i=1}^n f(X_i, \theta)) = 1,$$

y retomando la demostración del Teorema 2.3.12, tenemos que la varianza de T es igual a la cota de Cramer Rao.

□

Ahora, volvemos al Ejemplo 2.3.7 para ilustrar la utilidad del anterior resultado. Tenemos:

$$\begin{aligned} \frac{\partial}{\partial \theta} \ln \prod_{i=1}^n f(x_i, \theta) &= \frac{\partial}{\partial \theta} \ln \left\{ \frac{e^{-n\theta} \theta^{\sum x_i}}{\prod x_i!} \prod I_{\{0,1,\dots\}}(x_i) \right\} \\ &= \frac{\partial}{\partial \theta} \left\{ -n\theta + \sum_{i=1}^n x_i \ln \theta \right\} \\ &= -n + \frac{\sum_{i=1}^n x_i}{\theta} \\ &= \frac{-n\theta + n\bar{x}}{\theta} \\ &= \frac{n}{\theta} (\bar{x} - \theta), \end{aligned}$$

la anterior expresión es de la forma $K(\theta)(t(x_1, \dots, x_n) - \theta)$ donde $K(\theta) = n/\theta$ y $t(x_1, \dots, x_n) = \bar{x}$. Así se concluye que el estimador \bar{X} tiene varianza igual a la cota de Cramer Rao, teniendo en cuenta que es también insesgado para θ , se puede concluir que \bar{X} es UMVUE para θ .

Finalmente, resaltamos el procedimiento para comprobar que un estimador sea UMVUE, primero se debe demostrar que el estimador es insesgado y luego comprobar que la varianza del estimador es igual a la cota de Cramer Rao. Para este último hay dos formas de hacerlo: (1) usar directamente la definición, en este caso debe calcular la varianza del estimador y también calcular la cota de Cramer Rao; (2) usar el Resultado 2.3.13, en este caso no es necesario el cálculo de la varianza del estimador, ni la cota de Cramer Rao.

2.4.4 Completez

Otro concepto útil en la construcción de un buen estimador es el concepto de la completez. Este concepto, a diferencia de los conceptos como el sesgo, varianza o

suficiencia, el concepto de completez carece de una interpretación clara y de fácil entendimiento, pero resulta ser muy útil, como se verá más adelante, este concepto nos permite encontrar los estimadores UMVUE.

Definición 2.4.7. Dada una muestra aleatoria X_1, \dots, X_n con función de densidad $f(x_i, \theta)$, y T una estadística, se dice que T es completo para θ , si para cualquier función $g(\cdot)$, el hecho de $E(g(T)) = 0$ para todo θ , implica que $g(T) = 0$.

En algunos casos, no es muy complicado demostrar que un estimador es completo, como lo ilustra el siguiente ejemplo.

Ejemplo 2.4.8. Dada una muestra aleatoria X_1, \dots, X_n una muestra aleatoria con distribución $Ber(p)$, entonces el estimador $\sum_{i=1}^n X_i$ es un estimador completo para p . Para ver eso, tomamos una función $g(\cdot)$ cualquiera, y supongamos que $E(g(\sum_{i=1}^n X_i)) = 0$, y veamos que $g(\sum_{i=1}^n X_i) = 0$. Recordando que la distribución de $\sum_{i=1}^n X_i$ tiene distribución $Bin(n, p)$, tenemos:

$$\begin{aligned} 0 &= E(g(\sum_{i=1}^n X_i)) \\ &= \sum_{i=0}^n g(i) \binom{n}{i} p^i (1-p)^{n-i} \\ &= (1-p)^n \sum_{i=0}^n \binom{n}{i} g(i) \left(\frac{p}{1-p}\right)^i, \end{aligned}$$

de donde se tiene que $\sum_{i=0}^n \binom{n}{i} g(i) \left(\frac{p}{1-p}\right)^i = 0$, nótese que el lado izquierdo de la igualdad es un polinomio en $\frac{p}{1-p}$ de grado n . Recordando que un polinomio es igual a 0 si cada coeficiente del polinomio es 0, entonces se puede concluir que $\binom{n}{i} g(i) = 0$ para todo $i = 0, \dots, n$, de donde se tiene que $g(i) = 0$ para $i = 0, \dots, n$. Ahora la estadística $\sum_{i=1}^n X_i$ toma valores $0, \dots, n$, de donde se concluye finalmente que $g(\sum_{i=1}^n X_i) = 0$.

En el anterior ejemplo, la muestra aleatoria proviene de una distribución discreta, cuando se trata de una distribución continua, la forma de proceder es diferente, como se ilustra en el siguiente ejemplo.

Ejemplo 2.4.9. Sea X_1, \dots, X_n una muestra aleatoria con distribución uniforme sobre $[0, \theta]$, el estimador por el método de máxima verosimilitud es $X_{(n)}$ el máximo de la muestra, veamos que este estimador es completa. Para cualquier función g , tenemos

$$\begin{aligned} 0 &= E(g(X_{(n)})) \\ &= \int_0^\theta g(x) \frac{nx^{n-1}}{\theta^n} dx, \end{aligned}$$

utilizando el teorema fundamental de cálculo, se tiene que

$$0 = \frac{ng(\theta)}{\theta}$$

para todo $\theta > 0$, es decir, $g(\theta) = 0$ para todo $\theta > 0$, y como la estadística $X_{(n)}$ toma valores positivos, entonces $g(X_{(n)}) = 0$.

Para distribuciones pertenecientes a la familia exponencial, es muy fácil encontrar una estadística completa. Para distribuciones en la familia exponencial uniparamétrica tenemos el siguiente resultado.

Resultado 2.4.15. *Dada una muestra aleatoria X_1, \dots, X_n proveniente de una distribución $f(x, \theta)$ perteneciente a la familia exponencial, es decir,*

$$f(x, \theta) = h(x)c(\theta) \exp\{d(\theta)T(x)\},$$

entonces la estadística $\sum_{i=1}^n T(X_i)$ es una estadística completa para θ .

La versión equivalente para distribuciones en la familia exponencial biparamétrica se da a continuación.

Resultado 2.4.16. *Dada una muestra aleatoria X_1, \dots, X_n proveniente de una distribución $f(x_i, \theta_1, \theta_2)$ perteneciente a la familia exponencial biparamétrica de la forma*

$$f(x_i, \theta_1, \theta_2) = c(\boldsymbol{\theta})h(x) \exp\{d(\boldsymbol{\theta})'T(x)\},$$

donde $\boldsymbol{\theta} = (\theta_1, \theta_2)$, $d(\boldsymbol{\theta}) = (d_1(\boldsymbol{\theta}), d_2(\boldsymbol{\theta}))'$ y $T(x) = (T_1(x), T_2(x))'$, entonces las estadísticas $\sum_{i=1}^n T_1(X_i)$ y $\sum_{i=1}^n T_2(X_i)$ son estadísticas completas para θ_1 y θ_2 .

Los dos anteriores resultados, en conjunto con los resultados 2.3.7 y 2.3.8, nos permiten encontrar fácilmente estadísticas que sean a la vez suficientes y completas para distribuciones pertenecientes a la familia exponencial.

Una utilidad de las estadísticas completas es que, anteriormente se había visto que para un parámetro puede haber más de un estimador insesgado, pero cuando el estimador insesgado es función se combina con completez, sí se tiene la unicidad, como lo ilustra el siguiente resultado.

Resultado 2.4.17. *Dada una muestra aleatoria X_1, \dots, X_n una muestra aleatoria con parámetro θ , si T es una estadística completa, $g_1(T)$ y $g_2(T)$ son estimadores insesgados de θ , entonces $g_1(T) = g_2(T)$.*

Proof. Por hipótesis, se tiene que $E(g_1(T)) = E(g_2(T)) = 0$, de donde, $E((g_1 - g_2)(T)) = 0$, por definición de estimador completo, se sigue que $(g_1 - g_2)(T) = 0$, es decir, $g_1(T) = g_2(T)$. \square

En la demostración del Resultado 2.3.3, se mencionó el teorema de Basu, este teorema necesita los conceptos de suficiencia y completez de una estadística y también el concepto de estadística auxiliar, esto es, estadística cuya distribución no depende del parámetro de la distribución. Dado este concepto, presentamos a continuación el teorema de Basu.

Resultado 2.4.18. Si T es una estadística suficiente y completa, entonces $T(X)$ es independiente de toda estadística auxiliar.

Proof. La demostración para el caso discreto se encuentra en Casella & Berger (2002, pg. 287). \square

El Resultado 2.3.3 puede ser fácilmente demostrado usando el teorema de Basú, puesto que en primer lugar, en una muestra proveniente de la distribución $N(\mu, \sigma^2)$, \bar{X} es una estadística suficiente y completa para μ . Por otro lado, $\frac{\sum (X_i - \bar{X})^2}{\sigma^2} \sim \chi_{n-1}^2$, de donde podemos tener que la distribución de S_n^2 no depende de μ , al igual que la distribución de S_{n-1}^2 , de esta forma, podemos concluir que \bar{X} y S_n^2 (o S_{n-1}^2) son independientes.

Las mayores importancias de las propiedades de suficiencia y completéz, a parte de ser propiedades deseables para los estimadores, es que nos permite construir estimadores UMVUE, como lo ilustra el siguiente resultado.

Resultado 2.4.19. Dada una muestra aleatoria X_1, \dots, X_n con parámetro θ , si T_1 es un estimador insesgado para una función del parámetro $g(\theta)$ y T_2 es suficiente y completo para θ , entonces se tiene que $E(T_1|T_2)$ es el único estimador UMVUE para $g(\theta)$.

El anterior resultado nos brinda una herramienta poderosa para encontrar estimadores UMVUE, que consiste en los siguientes pasos:

- (1) Encontrar un estimador insesgado para $g(\theta)$, la función del parámetro θ que se desea estimar. Nótese que cuando $g(\theta)$ es el media poblacional, entonces un estimador insesgado es el promedio muestral.
- (2) Encontrar un estimador suficiente y completo para θ , que para distribuciones pertenecientes a la familia exponencial resulta bastante útil.

Ahora aplicamos las anteriores herramientas a un problema de estimación.

Ejemplo 2.4.10. Dada una muestra aleatoria X_1, \dots, X_n con distribución $P(\theta)$, se quiere encontrar el estimador UMVUE para θ . Para eso, primero se debe encontrar un estimador insesgado para θ , que es \bar{X} por ser θ el media poblacional. En segundo lugar, la distribución Poisson pertenece a la familia exponencial con $T(x) = x$, entonces los Resultado 2.3.6 y 2.3.15 establecen que $\sum_{i=1}^n X_i$ es una estadística suficiente y completa para θ . De esta manera el estimador UMVUE es $E(\bar{X} | \sum_{i=1}^n X_i) = \bar{X}$.

Para distribuciones con dos parámetros, existe la siguiente generalización,

Resultado 2.4.20. Dada una muestra aleatoria X_1, \dots, X_n una muestra aleatoria con parámetro $\theta = (\theta_1, \theta_2)$, si T_1, T_2 son estimadores insesgados para θ_1, θ_2 , y S_1, S_2 son suficientes y completos para θ_1, θ_2 , entonces se tiene que $E(T_1, T_2 | S_1, S_2)$ son UMVUE para θ_1 y θ_2 , además se tiene la unicidad.

Ejemplo 2.4.11. Sea X_1, \dots, X_n una muestra aleatoria con distribución $N(\mu, \sigma^2)$, en el ejemplo 2.3.9, se vio que estimadores suficientes y completos para μ y σ^2 son $\sum_{i=1}^n X_i$ y $\sum_{i=1}^n X_i^2$, y también se vio anteriormente que \bar{X} es insesgado para μ y S_{n-1}^2 insesgado para σ^2 , entonces el resultado anterior indica que las estadísticas $E(\bar{X}, S_{n-1}^2 | \sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2)$ son UMVUE único para μ y σ^2 . Para calcular la esperanza condicional, se tiene en cuenta que en primer lugar \bar{X} es función de $\sum_{i=1}^n X_i$ y $S_{n-1}^2 = (\sum_{i=1}^n X_i^2 - n\bar{X}^2)/(n-1)$ es función de $\sum_{i=1}^n X_i$ y $\sum_{i=1}^n X_i^2$, entonces se tiene que los UMVUE para μ y σ^2 son \bar{X} y S_{n-1}^2 .

Nótese que por la forma como están definidos S_n^2 y S_{n-1}^2 , se tiene que $\text{Var}(S_n^2) < \text{Var}(S_{n-1}^2)$, es decir, el estimador S_n^2 tiene una varianza más pequeña que la del estimador UMVUE, sin embargo, nótese que como estimador de σ^2 , S_n^2 es sesgado. Lo anterior muestra el hecho de que puede un estimador UMVUE tiene varianza más pequeña sólo entre los estimadores insesgados, puede existir un estimador sesgado con varianza aún más pequeña que la de UMVUE.

Hasta este punto del libro, se ha expuesto varios aspectos que se debe tener en cuenta al momento de escoger un estimador, más aún, al momento de escoger uno de varios posibles estimadores. Y al momento de realizar comparaciones, es necesario calcular la esperanza y la varianza de los estimadores, y estos cálculos pueden ser muy difíciles cuando los estimadores toman formas complicadas, por ejemplo, cuando se trata de una muestra proveniente de $\text{Unif}[\theta_1, \theta_2]$ se vio anteriormente que los estimadores de momentos de los parámetros son $\bar{X} - \sqrt{3}S_n$ y $\bar{X} + \sqrt{3}S_n$, respectivamente, mientras que los estimadores de momentos son las estadísticas de orden $X_{(1)}$ y $X_{(n)}$, respectivamente. Para tener una idea en este caso cuál método arrojó mejores estimadores, deberíamos calcular el sesgo y la varianzas de estos estimadores, pero es claro que no es nada trivial calcular estas cantidades para los estimadores de momentos. En casos como lo anterior, no tenemos otra alternativa que utilizar las simulaciones.

El lector puede notar que la teoría expuesta anteriormente es válida para cuando se quiere estimar una función del parámetro $g(\theta)$. Sin embargo, poco se ha discutido acerca de la evaluación de estos estimadores. La razón es que, en primer lugar, en general no es fácil determinar el sesgo de estos estimadores. Por ejemplo, considera el Ejemplo 2.2.3, donde en un conjunto de ensayos del tipo Bernoulli, el parámetro de interés es la probabilidad de éxito p que se estima mediante \bar{X} , pero si la cantidad que se desea estimar se cambia a $3p(1-p)^4$, aunque es fácil encontrar el estimador de máxima verosimilitud $3\bar{X}(1-\bar{X})^4$, no es fácil determinar si este es insesgado para $3p(1-p)^4$, y por consiguiente tampoco se puede determinar si este es el estimador UMVUE para $3p(1-p)^4$. Lo anterior indica que en general no se puede establecer que para cualquier función g , si T es un estimador insesgado para θ , entonces $g(T)$ es insesgado para $g(\theta)$, esto es, no se tiene en general que $E(g(T)) = g(\theta)$ suponiendo que $E(T) = \theta$.

Es claro que cuando la función g es una función lineal, esto es $g(\theta) = a\theta + b$, para constantes a y b , entonces $E(g(T)) = E(aT + b) = aE(T) + b$ y si T es insesgado para θ , $g(T)$ también lo es para $g(\theta)$. Cuando g toma otras formas, podemos calcular $E(g(T))$ e inclusive $\text{Var}(g(T))$ de manera aproximada usando