

Capítulo 16

Definición de normalidad en Estadística y medidas de descripción de datos

Jesús Reynaga Obregón

Este capítulo muestra cómo sintetizar series de datos numéricos de tal manera que usando cifras representativas sea posible comprender las características generales de dichas series. A las cifras representativas que caracterizan al conjunto de datos se les conoce como medidas de resumen para variables cuantitativas.

A fin de comprender e interpretar de manera adecuada el promedio y desviación estándar, se requiere conocer una de las más importantes distribuciones de probabilidad denominada distribución normal. Las características básicas de ella se tratan a continuación.

Distribución normal

La distribución normal o distribución de Gauss representa la forma en la que se distribuyen en la naturaleza los diversos valores numéricos de las variables continuas, como pueden ser estatura, peso, etc.

Para el caso de una variable de origen biológico, como es la distancia interpupilar (DIP) en los adultos sanos, se sabe que existen muchos individuos con una DIP cercana a 61.5 mm. También hay muchos con una DIP de 61 o 62 mm pero ya no son tan numerosos como los de 61.5 mm. Asimismo es posible encontrar personas con DIP de 58 o 65 mm, pero la frecuencia de este tipo de valores es muy escasa. La forma en que se distribuyen naturalmente los valores numéricos de la DIP se ilustra en la figura 16-1.

Cuando se calcula la desviación estándar para una serie de datos no siempre es evidente el significado del resultado obtenido, y menos aún si no se compara con la desviación estándar de otra serie diferente de datos.

Para muchas personas podría tener significado el hecho de que el promedio de peso de un grupo de 300 personas fue de 80 kg, pues de acuerdo con la definición del promedio, imaginarían que si todos los individuos tuvieran el mismo peso éste sería de 80 kg; sin embargo, para quienes no tienen conocimiento de las características básicas del modelo de la curva normal podría carecer de significado que les mencionaran que la desviación estándar del peso de las mismas personas fue de 5 kg.

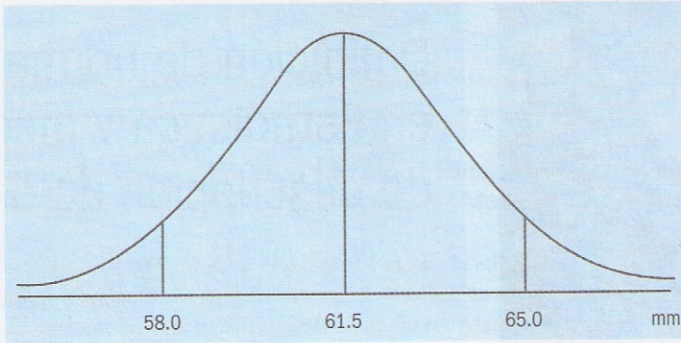


Figura 16-1 Distribución de la distancia interpupilar.

Interpretar la desviación estándar y comprender lo que significa en relación con los datos cuantitativos que se estén manejando sólo es posible a la luz del conocimiento del modelo de la curva normal.

Propiedades de la curva normal

1. La curva normal es un polígono de frecuencias en forma de campana para el que están calculadas sus áreas en función de los diversos valores del eje horizontal o abscisa (figura 16-2).
2. En la abscisa se encuentran valores de tipo cuantitativo continuo, denominados genéricamente como valores z , cuyas magnitudes en teoría pueden ir de izquierda a derecha desde $-\infty$ hasta $+\infty$ (desde menos infinito hasta más infinito).
3. El **promedio** de todos los valores z de la abscisa equivale a cero, pues la mitad son negativos y la mitad positivos. En el sitio de la abscisa que corresponde al cero, es decir al promedio, se encuentra la parte más alta de la curva. En este sitio también se encuentra la mediana de todos los valores z de la abscisa, pues 50% de ellos está antes del cero y el 50% restante se encuentra después.

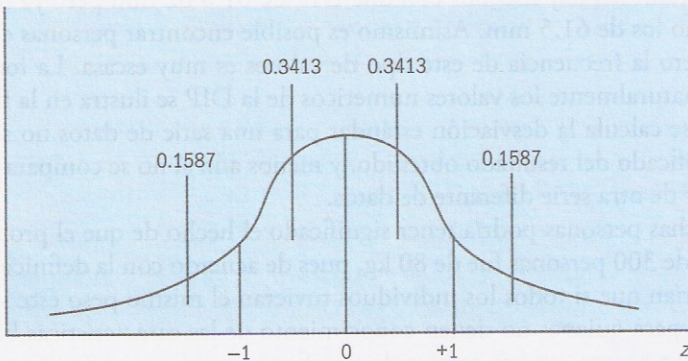


Figura 16-2 Algunas áreas bajo la curva de la distribución normal.

4. La curva es simétrica alrededor del promedio; esto es, hay una mitad izquierda que es reflejo de la mitad derecha.
5. En la abscisa existen segmentos unitarios de igual longitud y de tamaño 1. Los segmentos a la izquierda del promedio tienen signo negativo y los segmentos a la derecha del promedio tienen signo positivo. Tales segmentos, denominados **desviaciones estándar**, pueden dividirse en fracciones infinitamente pequeñas y continuas.
6. La curva es asintótica, es decir, sus extremos en teoría nunca tocan la abscisa. Por ello, la longitud de ésta podría ser infinitamente larga; sin embargo, se acostumbra graficar sólo hasta la distancia de tres segmentos a la izquierda y a la derecha del promedio.
7. Toda el área bajo la curva vale 1. Por lo anterior, el área a la izquierda del promedio vale 0.5, y el área a la derecha del promedio también vale 0.5.
8. El área que se encuentra sobre el segmento de la abscisa que va desde el promedio hasta el valor z de $+1$ vale 0.3413; por simetría, el área que se encuentra sobre el segmento que va desde el promedio hasta el valor z de -1 de la abscisa también vale 0.3413.

Por lo anterior, el área que se encuentra por arriba del amplio segmento que va desde el valor z de -1 hasta el valor z de $+1$ equivale a 0.6826; es decir, a la suma de $0.3413 + 0.3413$.

9. El área que se encuentra sobre el segmento de la abscisa que va más allá del valor z de $+1$ vale 0.1587; por simetría, el área que se encuentra sobre el segmento que va más allá (hacia menos infinito) del valor z de -1 de la abscisa también vale 0.1587.
10. Para cualquier segmento de la abscisa, y aun para fracciones de segmento, se encuentran calculadas las áreas correspondientes, como en el cuadro 16-1.

Cuadro 16-1 Fragmento de la tabla de áreas bajo la curva de la distribución normal.

(A) Valor z	(B) Área entre el promedio y el valor z	(C) Área más allá del valor z
0.00	0.0000	0.5000
0.25	0.0987	0.4013
0.50	0.1915	0.3085
0.75	0.2734	0.2266
1.00	0.3413	0.1587
1.25	0.3944	0.1056
1.50	0.4332	0.0668
1.65	0.4505	0.0495
1.75	0.4599	0.0401
1.96	0.4750	0.0250
2.00	0.4772	0.0228
2.58	0.4950	0.0050

Aprovechamiento de las propiedades de la curva normal para la interpretación de la desviación estándar

Al principio de este capítulo se comentó que si se desconocen las características básicas del modelo de la curva normal, carece de significado la mención de que el valor de la desviación estándar del peso de 300 personas fue de 5 kg.

Una vez que se han comprendido las propiedades principales de la curva normal es posible entender el significado del valor de la desviación estándar del peso de las 300 personas si se hacen suposiciones como las siguientes:

Suponiendo que al graficar el peso de los 300 individuos con un polígono de frecuencias la gráfica resultante fuera muy parecida al modelo de la curva normal, como se muestra en la figura 16-3. Entonces podría decirse que:

- El área bajo la curva de valores de peso que contiene a los individuos vale 300 de manera semejante a la propiedad del modelo de la curva normal de que su área vale 1.
- A la izquierda del promedio existen 150 individuos y a la derecha del promedio existen los otros 150.
- Así como en la curva normal existe un área de 0.3413 sobre el segmento que va desde el valor z de 0 hasta el valor z de +1, en la curva de valores x (es decir, kg de peso) habrá 0.3413 de 300, o sea que habrá 102 personas sobre el segmento que va desde el valor x de 80 kg hasta el valor x de 85 kg.

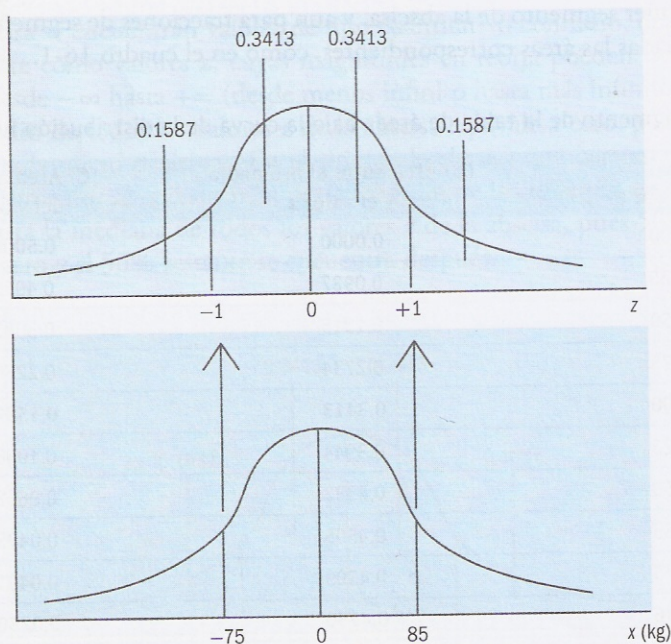


Figura 16-3 Correspondencia entre la curva de la distribución normal y la curva de la distribución del peso de 300 individuos.

- De acuerdo con el párrafo anterior, habrá 204 personas cuyo peso irá desde 75 hasta 85 kg.
- Al igual que en la curva normal, existe simetría alrededor del promedio; se puede considerar que en la curva de valores de peso habrá 102 personas sobre el segmento que va desde 80 hasta 75 kg de peso.
- En la curva de valores de peso habrá 0.1587 de las 300 personas; es decir, 48 personas con peso de 85 y más kg.
- De manera semejante a la curva normal, por simetría habrá un 0.1587 de las 300 personas; es decir 48 personas, con peso de 75 y menos kg.

Como es evidente, una vez que se conocen las características del modelo de la curva normal, la interpretación del resultado de la desviación estándar que se haya calculado para una serie de datos es mucho más fácil y brinda una gran cantidad de información sobre la manera en que se distribuyen los valores.

Para confirmar que la comprensión del significado de la desviación estándar brinda una importante cantidad de información, considere el ejemplo del recuadro 16-1.

Recuadro 16-1 Ejemplo de aplicación de las propiedades de la distribución normal.

Se aplicó un mismo examen escrito a dos grupos de 90 alumnos cada uno. En un caso se imprimió el examen en hojas de color amarillo paja y en otro caso en hojas de color marrón. Se midió con cronómetro el tiempo, en minutos y fracciones, que tardaron los alumnos en entregar el examen y se calculó el promedio y la desviación estándar para ambos grupos; los resultados se muestran en el cuadro 16-2.

A continuación se muestran algunas interpretaciones a partir de los valores de las desviaciones estándar:

- Los alumnos a quienes se aplicó el examen impreso en hojas color paja entregaron el examen en tiempos más homogéneos, pues 0.6826 de ellos (es decir, 61 alumnos) lo entregaron entre 40 y 50 min luego de haberlo iniciado.
- Los alumnos a quienes se aplicó el examen impreso en hojas color marrón entregaron el examen en tiempos más heterogéneos, pues 0.6826 de ellos (61 alumnos) lo entregaron entre 30 y 60 min después de haberlo iniciado.
- En el grupo paja, 0.1587 más lento, los alumnos (es decir, 14) entregaron su examen luego de 50 min, mientras que en el grupo marrón la misma cantidad de alumnos lo hizo después de 60 min.

Cuadro 16-2 Promedio y desviación estándar del tiempo de entrega del examen.

Grupo	Promedio	Desviación estándar
Color paja	45'	5'
Color marrón	45'	15'

Transformación de valores x a valores z ; uso de la tabla de áreas bajo la curva

Como se observa en el recuadro 16-1, hay correspondencia entre las áreas de la curva normal y las de la serie de datos cuantitativos continuos que se esté manejando, siempre y cuando se haya comprobado que esta última, al ser graficada con un polígono de frecuencias, muestra un parecido razonable con el perfil de la curva normal.

Tal correspondencia ha permitido mencionar las áreas que se encuentran sobre segmentos completos de la abscisa; es decir, sólo se han mencionado áreas por arriba o más allá de desviaciones estándar enteras. Sin embargo, ¿cómo podría responderse a la cuestión de cuántos alumnos de cada grupo tardaron 47 o más minutos en entregar su examen?

En este caso se aprecia que no hay coincidencia entre el valor z de +1 y el valor x de 47 min y, por ello, deja de ser útil el método de comparación analógica de las gráficas utilizado en páginas anteriores.

La respuesta estriba en el uso de una fórmula para transformar cualquier valor x en su correspondiente valor z y en hacer uso de la tabla de áreas bajo la curva normal.

Una vez que se han calculado tanto el promedio como la desviación estándar para una serie de datos cuantitativos continuos, el valor z que, en la abscisa de la curva normal corresponde a un determinado valor x de la abscisa de los datos que se están manejando, se encuentra con la fórmula para transformar valores x a valores z :

$$z = \frac{x - \bar{x}}{s}$$

Para responder a la pregunta recién planteada se hacen las siguientes sustituciones: para el grupo al que se aplicó el examen en hojas color paja se tiene que

$$\bar{x} = 45' \text{ y } s = 5';$$

el valor z que se desea conocer es el correspondiente a un valor x de 47; entonces:

$$z = \frac{47 - 45}{5} = \frac{2}{5} = 0.4$$

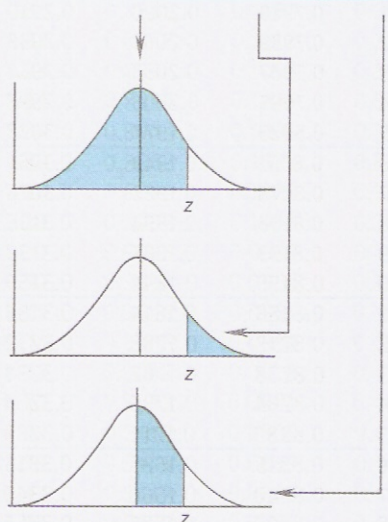
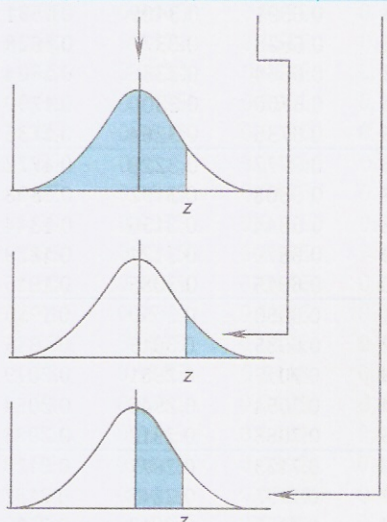
El valor z obtenido, en este caso 0.4, debe localizarse en la primera columna de la tabla de áreas bajo la curva —utilice la tabla detallada de áreas bajo la curva normal que se muestra en el cuadro 16-3.

Una vez localizado tal valor, se busca en la segunda columna cuál es el área que en la curva normal se encuentra más allá de dicho valor z ; en este caso es de 0.3446.

Como el área encontrada (0.3446) es una proporción del área total, entonces la misma proporción se aplica al total de alumnos del grupo para saber cuántos tardaron más de 47 minutos en entregar el examen.

Así, luego de efectuar la operación $0.3446 \times 90 = 31$, puede responderse a la pregunta con el señalamiento de que hubo en este grupo 31 alumnos que tardaron 47 o más minutos en entregar su examen. Desde luego, al conocer las propiedades básicas de la curva normal, también es factible decir que hubo 59 alumnos que tardaron 47 o menos minutos en entregar su examen.

Cuadro 16-3 Áreas bajo la curva de la distribución normal.

Valor z (A)	Área desde el extremo opuesto hasta el valor z (B)	Área en el mismo extremo más allá del valor z (C)	Área entre el promedio y el valor z (D)	Valor z (A)	Área desde el extremo opuesto hasta el valor z (B)	Área en el mismo extremo más allá del valor z (C)	Área entre el promedio y el valor z (D)
							
$Z = \frac{x - \bar{X}}{s}$				$Z = \frac{x - \bar{X}}{s}$			
0.00	0.5000	0.5000	0.0000	0.21	0.5832	0.4168	0.0832
0.01	0.5040	0.4960	0.0040	0.22	0.5871	0.4129	0.0871
0.02	0.5080	0.4920	0.0080	0.23	0.5910	0.4090	0.0910
0.03	0.5120	0.4880	0.0120	0.24	0.5948	0.4052	0.0948
0.04	0.5160	0.4840	0.0160	0.25	0.5987	0.4013	0.0987
0.05	0.5199	0.4801	0.0199	0.26	0.6026	0.3974	0.1026
0.06	0.5239	0.4761	0.0239	0.27	0.6064	0.3936	0.1064
0.07	0.5279	0.4721	0.0279	0.28	0.6103	0.3897	0.1103
0.08	0.5319	0.4681	0.0319	0.29	0.6141	0.3859	0.1141
0.09	0.5359	0.4641	0.0359	0.30	0.6179	0.3821	0.1179
0.10	0.5398	0.4602	0.0398	0.31	0.6217	0.3783	0.1217
0.11	0.5438	0.4562	0.0438	0.32	0.6255	0.3745	0.1255
0.12	0.5478	0.4522	0.0478	0.33	0.6293	0.3707	0.1293
0.13	0.5517	0.4483	0.0517	0.34	0.6331	0.3669	0.1331
0.14	0.5557	0.4443	0.0557	0.35	0.6368	0.3632	0.1368
0.15	0.5596	0.4404	0.0596	0.36	0.6406	0.3594	0.1406
0.16	0.5636	0.4364	0.0636	0.37	0.6443	0.3557	0.1443
0.17	0.5675	0.4325	0.0675	0.38	0.6480	0.3520	0.1480
0.18	0.5714	0.4286	0.0714	0.39	0.6517	0.3483	0.1517
0.19	0.5753	0.4247	0.0753	0.40	0.6554	0.3446	0.1554
0.20	0.5793	0.4207	0.0793				

(continúa)

Cuadro 16-3 Áreas bajo la curva de la distribución normal (Continuación).

Valor z (A)	Área desde el extremo opuesto hasta el valor z (B)	Área en el mismo extremo más allá del valor z (C)	Área entre el promedio y el valor z (D)	Valor z (A)	Área desde el extremo opuesto hasta el valor z (B)	Área en el mismo extremo más allá del valor z (C)	Área entre el promedio y el valor z (D)
0.41	0.6591	0.3409	0.1591	0.81	0.7910	0.2090	0.2910
0.42	0.6628	0.3372	0.1628	0.82	0.7939	0.2061	0.2939
0.43	0.6664	0.3336	0.1664	0.83	0.7967	0.2033	0.2967
0.44	0.6700	0.3300	0.1700	0.84	0.7995	0.2005	0.2995
0.45	0.6736	0.3264	0.1736	0.85	0.8023	0.1977	0.3023
0.46	0.6772	0.3228	0.1772	0.86	0.8051	0.1949	0.3051
0.47	0.6808	0.3192	0.1808	0.87	0.8078	0.1922	0.3078
0.48	0.6844	0.3156	0.1844	0.88	0.8106	0.1894	0.3106
0.49	0.6879	0.3121	0.1879	0.89	0.8133	0.1867	0.3133
0.50	0.6915	0.3085	0.1915	0.90	0.8159	0.1841	0.3159
0.51	0.6950	0.3050	0.1950	0.91	0.8186	0.1814	0.3186
0.52	0.6985	0.3015	0.1985	0.92	0.8212	0.1788	0.3212
0.53	0.7019	0.2981	0.2019	0.93	0.8238	0.1762	0.3238
0.54	0.7054	0.2946	0.2054	0.94	0.8264	0.1736	0.3264
0.55	0.7088	0.2912	0.2088	0.95	0.8289	0.1711	0.3289
0.56	0.7123	0.2877	0.2123	0.96	0.8315	0.1685	0.3315
0.57	0.7157	0.2843	0.2157	0.97	0.8340	0.1660	0.3340
0.58	0.7190	0.2810	0.2190	0.98	0.8365	0.1635	0.3365
0.59	0.7224	0.2776	0.2224	0.99	0.8389	0.1611	0.3389
0.60	0.7257	0.2743	0.2257	1.00	0.8413	0.1587	0.3413
0.61	0.7291	0.2709	0.2291	1.01	0.8438	0.1562	0.3438
0.62	0.7324	0.2676	0.2324	1.02	0.8461	0.1539	0.3461
0.63	0.7357	0.2643	0.2357	1.03	0.8485	0.1515	0.3485
0.64	0.7389	0.2611	0.2389	1.04	0.8508	0.1492	0.3508
0.65	0.7422	0.2578	0.2422	1.05	0.8531	0.1469	0.3531
0.66	0.7454	0.2546	0.2454	1.06	0.8554	0.1446	0.3554
0.67	0.7486	0.2514	0.2486	1.07	0.8577	0.1423	0.3577
0.68	0.7517	0.2483	0.2517	1.08	0.8599	0.1401	0.3599
0.69	0.7549	0.2451	0.2549	1.09	0.8621	0.1379	0.3621
0.70	0.7580	0.2420	0.2580	1.10	0.8643	0.1357	0.3643
0.71	0.7611	0.2389	0.2611	1.11	0.8665	0.1335	0.3665
0.72	0.7642	0.2358	0.2642	1.12	0.8686	0.1314	0.3686
0.73	0.7673	0.2327	0.2673	1.13	0.8708	0.1292	0.3708
0.74	0.7704	0.2296	0.2704	1.14	0.8729	0.1271	0.3729
0.75	0.7734	0.2266	0.2734	1.15	0.8749	0.1251	0.3749
0.76	0.7764	0.2236	0.2764	1.16	0.8770	0.1230	0.3770
0.77	0.7794	0.2206	0.2794	1.17	0.8790	0.1210	0.3790
0.78	0.7823	0.2177	0.2823	1.18	0.8810	0.1190	0.3810
0.79	0.7852	0.2148	0.2852	1.19	0.8830	0.1170	0.3830
0.80	0.7881	0.2119	0.2881	1.20	0.8849	0.1151	0.3849

(continúa)

Cuadro 16-3 Áreas bajo la curva de la distribución normal (*Continuación*).

Valor z (A)	Área desde el extremo opuesto hasta el valor z (B)	Área en el mismo extremo más allá del valor z (C)	Área entre el promedio y el valor z (D)	Valor z (A)	Área desde el extremo opuesto hasta el valor z (B)	Área en el mismo extremo más allá del valor z (C)	Área entre el promedio y el valor z (D)
1.21	0.8869	0.1131	0.3869	1.61	0.9463	0.0537	0.4463
1.22	0.8888	0.1112	0.3888	1.62	0.9474	0.0526	0.4474
1.23	0.8907	0.1093	0.3907	1.63	0.9484	0.0516	0.4484
1.24	0.8925	0.1075	0.3925	1.64	0.9495	0.0505	0.4495
1.25	0.8944	0.1056	0.3944	1.65	0.9505	0.0495	0.4505
1.26	0.8962	0.1038	0.3962	1.66	0.9515	0.0485	0.4515
1.27	0.8980	0.1020	0.3980	1.67	0.9525	0.0475	0.4525
1.28	0.8997	0.1003	0.3997	1.68	0.9535	0.0465	0.4535
1.29	0.9015	0.0985	0.4015	1.69	0.9545	0.0455	0.4545
1.30	0.9032	0.0968	0.4032	1.70	0.9564	0.0446	0.4554
1.31	0.9049	0.0951	0.4049	1.71	0.9564	0.0436	0.4564
1.32	0.9066	0.0934	0.4066	1.72	0.9573	0.0427	0.4573
1.33	0.9082	0.0918	0.4082	1.73	0.9582	0.0418	0.4582
1.34	0.9099	0.0901	0.4099	1.74	0.9591	0.0409	0.4591
1.35	0.9115	0.0885	0.4115	1.75	0.9599	0.0401	0.4599
1.36	0.9131	0.0869	0.4131	1.76	0.9608	0.0392	0.4608
1.37	0.9147	0.0853	0.4147	1.77	0.9616	0.0384	0.4616
1.38	0.9162	0.0838	0.4162	1.78	0.9625	0.0375	0.4625
1.39	0.9177	0.0823	0.4177	1.79	0.9633	0.0367	0.4633
1.40	0.9192	0.0808	0.4192	1.80	0.9641	0.0359	0.4641
1.41	0.9207	0.0793	0.4207	1.81	0.9649	0.0351	0.4649
1.42	0.9222	0.0778	0.4222	1.82	0.9656	0.0344	0.4656
1.43	0.9236	0.0764	0.4236	1.83	0.9664	0.0336	0.4664
1.44	0.9251	0.0749	0.4251	1.84	0.9671	0.0329	0.4671
1.45	0.9265	0.0735	0.4265	1.85	0.9678	0.0322	0.4678
1.46	0.9279	0.0721	0.4279	1.86	0.9686	0.0314	0.4686
1.47	0.9292	0.0708	0.4292	1.87	0.9693	0.0307	0.4693
1.48	0.9306	0.0694	0.4306	1.88	0.9699	0.0301	0.4699
1.49	0.9319	0.0681	0.4319	1.89	0.9706	0.0294	0.4706
1.50	0.9332	0.0668	0.4332	1.90	0.9713	0.0287	0.4713
1.51	0.9345	0.0655	0.4345	1.91	0.9719	0.0281	0.4719
1.52	0.9357	0.0643	0.4357	1.92	0.9726	0.0274	0.4726
1.53	0.9370	0.0630	0.4370	1.93	0.9732	0.0268	0.4732
1.54	0.9382	0.0618	0.4382	1.94	0.9738	0.0262	0.4738
1.55	0.9394	0.0606	0.4394	1.95	0.9744	0.0256	0.4744
1.56	0.9406	0.0594	0.4406	1.96	0.9750	0.0250	0.4750
1.57	0.9418	0.0582	0.4418	1.97	0.9756	0.0244	0.4756
1.58	0.9429	0.0571	0.4429	1.98	0.9761	0.0239	0.4761
1.59	0.9441	0.0559	0.4441	1.99	0.9767	0.0233	0.4767
1.60	0.9452	0.0548	0.4452	2.00	0.9772	0.0228	0.4772

(continúa)

Cuadro 16-3 Áreas bajo la curva de la distribución normal (Continuación).

Valor z (A)	Área desde el extremo opuesto hasta el valor z (B)	Área en el mismo extremo más allá del valor z (C)	Área entre el promedio y el valor z (D)	Valor z (A)	Área desde el extremo opuesto hasta el valor z (B)	Área en el mismo extremo más allá del valor z (C)	Área entre el promedio y el valor z (D)
2.01	0.9778	0.0222	0.4778	2.41	0.9920	0.0080	0.4920
2.02	0.9783	0.0217	0.4783	2.42	0.9922	0.0078	0.4922
2.03	0.9788	0.0212	0.4788	2.43	0.9925	0.0075	0.4925
2.04	0.9793	0.0207	0.4793	2.44	0.9927	0.0073	0.4927
2.05	0.9798	0.0202	0.4798	2.45	0.9929	0.0071	0.4929
2.06	0.9803	0.0197	0.4803	2.46	0.9931	0.0069	0.4931
2.07	0.9808	0.0192	0.4808	2.47	0.9932	0.0068	0.4932
2.08	0.9812	0.0188	0.4812	2.48	0.9934	0.0066	0.4934
2.09	0.9817	0.0183	0.4817	2.49	0.9936	0.0064	0.4936
2.10	0.9821	0.0179	0.4821	2.50	0.9938	0.0062	0.4938
2.11	0.9826	0.0174	0.4826	2.51	0.9940	0.0060	0.4940
2.12	0.9830	0.0170	0.4830	2.52	0.9941	0.0059	0.4941
2.13	0.9834	0.0166	0.4834	2.53	0.9943	0.0057	0.4943
2.14	0.9838	0.0162	0.4838	2.54	0.9945	0.0055	0.4945
2.15	0.9842	0.0158	0.4842	2.55	0.9946	0.0054	0.4946
2.16	0.9846	0.0154	0.4846	2.56	0.9948	0.0052	0.4948
2.17	0.9850	0.0150	0.4850	2.57	0.9949	0.0051	0.4949
2.18	0.9854	0.0146	0.4854	2.58	0.9951	0.0049	0.4951
2.19	0.9857	0.0143	0.4857	2.59	0.9952	0.0048	0.4952
2.20	0.9861	0.0139	0.4861	2.60	0.9953	0.0047	0.4953
2.21	0.9864	0.0136	0.4864	2.61	0.9955	0.0045	0.4955
2.22	0.9868	0.0132	0.4868	2.62	0.9956	0.0044	0.4956
2.23	0.9871	0.0129	0.4871	2.63	0.9957	0.0043	0.4957
2.24	0.9875	0.0125	0.4875	2.64	0.9959	0.0041	0.4959
2.25	0.9878	0.0122	0.4878	2.65	0.9960	0.0040	0.4960
2.26	0.9881	0.0119	0.4881	2.66	0.9961	0.0039	0.4961
2.27	0.9884	0.0116	0.4884	2.67	0.9962	0.0038	0.4962
2.28	0.9887	0.0113	0.4887	2.68	0.9963	0.0037	0.4963
2.29	0.9890	0.0110	0.4890	2.69	0.9964	0.0036	0.4964
2.30	0.9893	0.0107	0.4893	2.70	0.9965	0.0035	0.4965
2.31	0.9896	0.0104	0.4896	2.71	0.9966	0.0034	0.4966
2.32	0.9898	0.0102	0.4898	2.72	0.9967	0.0033	0.4967
2.33	0.9901	0.0099	0.4901	2.73	0.9968	0.0032	0.4968
2.34	0.9904	0.0096	0.4904	2.74	0.9969	0.0031	0.4969
2.35	0.9906	0.0094	0.4906	2.75	0.9970	0.0030	0.4970
2.36	0.9909	0.0091	0.4909	2.76	0.9971	0.0029	0.4971
2.37	0.9911	0.0089	0.4911	2.77	0.9972	0.0028	0.4972
2.38	0.9913	0.0087	0.4913	2.78	0.9973	0.0027	0.4973
2.39	0.9916	0.0084	0.4916	2.79	0.9974	0.0026	0.4974
2.40	0.9918	0.0082	0.4918	2.80	0.9974	0.0026	0.4974

(continúa)

Cuadro 16-3 Áreas bajo la curva de la distribución normal (*Continuación*).

Valor z (A)	Área desde el extremo opuesto hasta el valor z (B)	Área en el mismo extremo más allá del valor z (C)	Área entre el promedio y el valor z (D)	Valor z (A)	Área desde el extremo opuesto hasta el valor z (B)	Área en el mismo extremo más allá del valor z (C)	Área entre el promedio y el valor z (D)
2.81	0.9975	0.0025	0.4975	2.91	0.9982	0.0018	0.4982
2.82	0.9976	0.0024	0.4976	2.92	0.9982	0.0018	0.4982
2.83	0.9977	0.0023	0.4977	2.93	0.9983	0.0017	0.4983
2.84	0.9977	0.0023	0.4977	2.94	0.9984	0.0016	0.4984
2.85	0.9978	0.0022	0.4978	2.95	0.9984	0.0016	0.4984
2.86	0.9979	0.0021	0.4979	2.96	0.9985	0.0015	0.4985
2.87	0.9979	0.0021	0.4979	2.97	0.9985	0.0015	0.4985
2.88	0.9980	0.0020	0.4980	2.98	0.9986	0.0014	0.4986
2.89	0.9981	0.0019	0.4981	2.99	0.9986	0.0014	0.4986
2.90	0.9981	0.0019	0.4981	3.00	0.9987	0.0013	0.4987

Por otra parte, para el grupo al que se aplicó el examen en hojas color marrón se tiene que:

$$\bar{x} = 45' \text{ y } s = 15'$$

En vista de que el valor z que se desea conocer es el correspondiente a un valor x de 47, entonces el cálculo es como sigue:

$$z = \frac{47 - 45}{15} = \frac{2}{15} = 0.13$$

El valor z obtenido, en este caso 0.13, debe localizarse en la primera columna de la tabla de áreas bajo la curva. Una vez localizado tal valor, se busca en la segunda columna cuál es el área que en la curva normal se encuentra **más allá** de dicho valor z ; en este caso es 0.3446.

Como el área encontrada (0.4483) es una proporción del área total, entonces la misma proporción se aplica al total de alumnos del grupo para saber cuántos tardaron más de 47 minutos en entregar el examen.

Así, luego de efectuar la operación $0.4483 \times 90 = 40$, puede responderse a la pregunta con el señalamiento de que hubo en este grupo 40 alumnos que tardaron 47 o más minutos en entregar su examen. Desde luego, al conocer las propiedades básicas de la curva normal, también se puede decir que hubo 50 alumnos que tardaron 47 o menos minutos en entregar su examen.

Cuadro 16-4 Peso en kilogramos de un grupo de 20 niños de un año de edad.

9.1	9.4	8.9	9.6	10.5	8.8	9.4	9.2	9.0	8.1
9.3	8.8	9.5	9.7	9.2	9.4	9.6	9.0	9.4	9.8

Medidas de descripción de datos (cálculo e interpretación de medidas de tendencia central y de dispersión)

Mediana y percentiles

Cuando se desea sintetizar una serie de datos cuantitativos discretos, como el número de embarazos, el número de convulsiones o el número de dientes con caries, debe utilizarse la mediana y los percentiles. Estas medidas de resumen, a diferencia del promedio y la desviación estándar, son apropiadas para sintetizar las variables cuantitativas discretas.

Con el siguiente ejemplo debe quedar claro que el promedio y la desviación estándar no son medidas de resumen propias para sintetizar a las variables cuantitativas discretas: ¿qué significaría que el promedio de hijos de un grupo de madres fue de 2.75 hijos? (¿querría decir que en promedio cada una tuvo dos hijos completos y otro más al que le faltó un brazo?); ¿qué significaría 1.5 dientes perdidos (¿que se ha perdido un diente y la mitad de otro?).

A diferencia del promedio y la desviación estándar —que sólo deben usarse para sintetizar variables cuantitativas continuas—, la mediana y los percentiles pueden utilizarse para resumir tanto variables cuantitativas discretas como variables cuantitativas continuas.

La serie simple de valores que se muestra en el cuadro 16-4 se utilizará como ejemplo para ilustrar el cálculo e interpretación de la mediana y algunos percentiles.

Mediana (o percentil 50)

Definición de la mediana. En una serie de valores ordenados de menor a mayor, o viceversa, es aquel valor que divide en dos partes de igual tamaño a toda la serie.

Procedimiento de cálculo de la mediana. El cuadro 16-5 muestra la serie que aparece en el cuadro 16-4 ya ordenada; además, se ha localizado el valor que la divide en dos partes de igual tamaño, de tal manera que en una parte queda 50% de los datos y en la otra el 50% restante.

Cuadro 16-5 Misma serie del cuadro 16-4, pero en orden y con la mediana localizada.

8.1	8.8	8.8	8.9	9.0	9.0	9.1	9.2	9.2	9.3
9.4	9.4	9.4	9.4	9.5	9.6	9.6	9.7	9.8	10.5

En vista de que la serie es par, no existe un valor que se ubique exactamente en el centro y la divide en dos partes. Por lo anterior, se considera que el promedio de los dos valores centrales que están colocados en las posiciones 10 y 11 corresponde al valor de la mediana; es decir, la mediana equivale al valor promedio de 9.3 y 9.4 (9.35).

Interpretación de la mediana. “La mitad de los niños tuvieron un peso igual o menor que 9.35 kg, y la otra mitad pesaron 9.35 kg o más.”

Percentiles (Pp)

Definición de percentil. En una serie de valores ordenados de menor a mayor, o viceversa, es aquel valor que divide en dos partes porcentualmente complementarias a toda la serie. Por ejemplo, el percentil 40 divide a la serie en una parte que contiene 40% de los valores iguales o inferiores a él y, al mismo tiempo, en otra parte que contiene 60% de los valores de la serie iguales o mayores a dicho percentil.

Procedimiento de cálculo de cualquier percentil. Debe ordenarse la serie y localizarse el valor que la divide en los porcentajes complementarios deseados.

Por ejemplo, para encontrar el valor del percentil 25 del ejemplo presentado en el cuadro 16-4, debe localizarse el valor que deje a una cuarta parte de los valores con menores o iguales magnitudes a dicho valor y a las tres cuartas partes restantes de los valores con magnitudes más grandes o iguales (cuadro 16-6).

En esta serie, entre los varios valores 9.0 se debe encontrar un valor en una posición tal que hasta dicho valor se encuentra 25% de los casos y, al mismo tiempo, a partir del mismo se ubica el 75% restante de los casos.

Por lo general, cualquier percentil se ubica en una posición localizada aplicando la siguiente fórmula:

$$\text{Lugar que ocupa el percentil buscado} = \frac{(P_{\text{buscado}})(n + 1)}{100}$$

Para el caso del percentil 25, a la posición $\frac{(P_{25})(20 + 1)}{100}$ le corresponde el lugar

$$\frac{(25)(21)}{100} = 5.25$$

Lo anterior significa que el percentil 25 se encuentra entre los lugares 5 y 6. En estos casos, por convención, se considera posible obtener un promedio de los valores que se encuentren en las posiciones adyacentes. Como ya se observó, la quinta posición está

Cuadro 16-6 Percentil 25 en la serie del cuadro 16-4.

8.1	8.8	8.8	8.9	9.0	9.0	9.1	9.2	9.2	9.3
9.4	9.4	9.4	9.4	9.5	9.6	9.6	9.7	9.8	10.5

Cuadro 16-7 Percentil 75 en la serie del cuadro 16-4.

8.1	8.8	8.8	8.9	9.0	9.0	9.1	9.2	9.2	9.3
9.4	9.4	9.4	9.4	9.5	9.6	9.6	9.7	9.8	10.5

ocupada por un valor de 9.0 y la sexta por un valor también de 9.0; por lo anterior, el promedio de ambos valores es igual a 9.0.

Interpretación de los percentiles (válida para el percentil 25 o P_{25}). Así, “25% de los niños tuvo un peso de 9.0 kg o menor, y el 75% restante tuvo el mismo peso o más”.

Como ejemplo adicional suponga que se desea encontrar el valor del percentil 75. Para ello debe localizarse el valor que deje tres cuartas partes de los valores con menores o iguales magnitudes al mismo y a la cuarta parte restante de los valores con magnitudes más grandes o iguales (cuadro 16-7).

En esta serie, entre los valores 9.5 y 9.6 se debe encontrar un valor en una posición tal que hasta dicho valor se encuentra 75% de los casos y, simultáneamente, a partir del mismo se ubica el 25% restante de los casos.

Utilizando la fórmula:

$$\text{Lugar que ocupa el percentil buscado} = \frac{(P_{\text{buscado}})(n + 1)}{100}$$

Para el caso del percentil 75, a la posición $\frac{(P_{75})(20 + 1)}{100}$ le corresponde el lugar

$$\frac{(75)(21)}{100} = 15.75$$

Lo anterior significa que el percentil 75 se encuentra entre los lugares 15 y 16. Como ya se dijo, en estos casos y por convención, se considera posible obtener un promedio de los valores que se encuentren en las posiciones adyacentes. La posición 15 está ocupada por un valor de 9.5, y la posición 16 por un valor también de 9.6; por lo anterior, el promedio de ambos valores es igual a 9.55.

Interpretación (válida para el percentil 75 o P_{75}). “El 75% de los niños tuvo un peso de 9.55 kg o menor, y el 25% restante tuvo peso de 9.55 kg o mayor.”

Rango intercuartílico (RIC)

Definición. El RIC es la diferencia entre los percentiles 75 y 25. El rango intercuartílico es una medida que abarca al 50% central de los valores de una serie ordenada de números, y es una medida de síntesis que expresa el grado de homogeneidad o heterogeneidad de dicho porcentaje de datos.

Procedimiento de cálculo del rango intercuartílico. En el ejemplo que deriva de los datos del cuadro 16-4, la diferencia $P_{75} - P_{25}$ es igual a $9.55 - 9.0 = 0.55$; lo anterior signi-

Cuadro 16-8 Peso al nacer de dos grupos de niños según duración de la gestación.

Número de niños	Duración de la gestación en semanas	P25	P75	Diferencia P75 – P25	Comentario
60	32	1 800	2 800	1 000	El 50% central de la serie de pesos de los 60 niños tiene una diferencia entre el mayor de los pesos y el menor de ellos de 1 000 g. Puede hablarse de gran heterogeneidad
2 709	39	2 884	3 132	248	El 50% central de la serie de pesos de los 2 709 niños tiene una diferencia entre el mayor de los pesos y el menor de ellos de 248 g. Puede hablarse de gran homogeneidad

fica que, específicamente refiriéndose al 50% central de los datos ya ordenados, la diferencia entre el mayor y el menor de los datos de dicho 50% central de los valores es 0.55 kg.

Interpretación del RIC. El RIC es una medida que permite comparar con facilidad la homogeneidad o heterogeneidad de dos series de datos semejantes; véase el ejemplo del cuadro 16-8.

Promedio y desviación estándar

Promedio

En el cuadro 16-9 se muestra una tabla de valores cuantitativos continuos que servirá para ilustrar el cálculo e interpretación del promedio y de la desviación estándar.

Definición del promedio. Es el valor que tendrían todos los datos de una serie numérica si ellos fueran de igual valor.

Fórmula de cálculo del promedio:

$$\bar{x} = \frac{\sum x}{n}$$

Cuadro 16-9 Peso en kg de un grupo de 20 niños de un año de edad.

8.1	8.8	8.8	8.9	9.0	9.0	9.1	9.2	9.2	9.3
9.4	9.4	9.4	9.4	9.5	9.6	9.6	9.7	9.8	10.5

Procedimiento de cálculo del promedio. Sumar todos los valores y dividir tal suma entre el número de valores que componen la serie de datos cuantitativos continuos. En este caso, la suma es 185.7 y, entonces, el promedio vale $\bar{x} = \frac{185.7}{20} = 9.285$ kg.

Interpretación del promedio. “Si todos los niños tuvieran peso igual, éste sería de 9.285 kg.”

Desviación estándar

Es la raíz cuadrada de la varianza. A su vez, la varianza equivale al promedio de las desviaciones o diferencias cuadráticas de cada valor de una serie respecto al promedio de dicha serie.

Fórmula de la desviación estándar

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

Procedimiento de cálculo de la desviación estándar:

- I. Obtener el promedio de la serie de valores. En este caso se usa el valor antes calculado, que había sido 9.285 kg.
- II. Calcular la desviación o diferencia de cada valor en relación con el promedio de la serie; es decir, obtener una serie de valores $(x - \bar{x})$.
- III. Elevar al cuadrado cada una de las anteriores desviaciones; es decir, obtener una serie de valores $(x - \bar{x})^2$.
- IV. Efectuar la suma de desviaciones cuadráticas; es decir, obtener el siguiente valor: $\sum(x - \bar{x})^2$.
- V. Dividir la suma anterior entre el número de valores menos uno; es decir, obtener el promedio de desviaciones cuadráticas o varianza: $\frac{\sum(x - \bar{x})^2}{n - 1}$
- VI. Obtener la raíz cuadrada del anterior promedio; es decir, obtener la desviación estándar: $s = \sqrt{\frac{\sum(x - \bar{x})^2}{n - 1}}$

Así, los cálculos para los pesos de los 20 niños son los siguientes:

Para el paso I el promedio ya calculado en párrafos anteriores vale 9.285 kg.

Para los pasos II, III y IV se recomienda utilizar una tabla auxiliar de cálculo como la que se muestra en el cuadro 16-10.

Para el paso V el promedio de desviaciones cuadráticas, o varianza, vale, entonces,

$$\frac{4.446}{20 - 1} = 0.234$$

Cuadro 16-10 Cálculo de la desviación estándar del peso de los 20 niños.

Niño	Cada uno de los valores x	Desviación de cada valor respecto al promedio (paso II)	Elevación al cuadrado de cada una de las desviaciones (paso III)
1	9.1	-0.185	0.034
2	9.4	0.115	0.113
3	8.9	-0.385	0.148
—	—	—	—
—	—	—	—
20	9.8	0.515	0.265
—	—	—	4.446
—	—	—	(paso IV)

Para el paso VI la desviación estándar, que equivale a la raíz cuadrada de la varianza, equivale a la raíz cuadrada de 0.234, o sea

$$\sqrt{0.234} = 0.484$$

Interpretación de la desviación estándar. La interpretación está condicionada a la suposición de que los valores tienen una distribución semejante a la de la curva normal.

Dicha interpretación puede ser realizada en múltiples sentidos, ya que se sabe que 0.6827 (68.27%) de los valores de una serie que se distribuye como la curva normal están agrupados alrededor del promedio si a éste se le resta una vez y también se le suma una vez el valor calculado para la desviación estándar.

Así, para los valores del promedio (9.285) y de la desviación estándar (0.484) calculados a los pesos de los 20 niños, se puede decir que 0.6827 (68.27%) de ellos tuvieron los siguientes pesos: 9.285 ± 0.484 kg; dicho de otra forma: 0.6827 (68.27%) de los 20 niños pesaron desde 8.801 hasta 9.769 kg.

A manera de otro ejemplo, si a un grupo de personas se les hubiera encontrado valores de glucosa que, en resumen, tuvieron promedio de 70 y desviación estándar de 10 mg por cada 100 ml de sangre, y si se supiera que la concentración de glucosa en esas personas tiene una distribución semejante a la de la curva normal, habría que concluir que 0.6827 (68.27%) de ellos tuvieron concentraciones de glucosa del siguiente tipo: 70 ± 10 ; es decir, 0.6827 (68.27%) tuvo valores desde 60 hasta 80 mg de glucosa por cada 100 ml de sangre.

Retomando el caso del peso de los 20 niños en el cuadro 16-8, como se sabe que la curva normal tiene un área que equivale a un total de 100%, entonces también puede decirse que hubo 15.865% de los niños que pesaron menos de 8.801 kg y que hubo otro 15.865% que pesaron más de 9.769 kg. La suma de estos dos porcentajes equivale a 31.73%; por otra parte, también puede decirse que 31.73% de los niños pesaron ya menos de 8.801 kg o más de 9.769 kg (si a 100% se le resta 68.27% quedan 31.73%). En la figura 16-4 se ilustra la última conclusión.

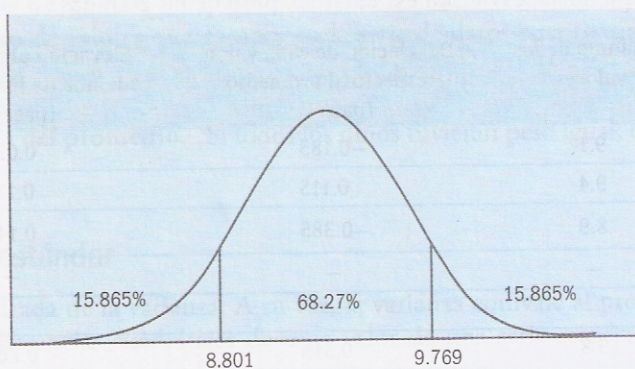


Figura 16-4 Distribución de los pesos de 20 niños.

Estadística paramétrica

Como se ha analizado en el presente capítulo, las distribuciones de datos cuantitativos continuos que tienen una distribución semejante a la de la curva normal pueden ser descritos perfectamente utilizando sólo dos medidas de resumen: el promedio y la desviación estándar. En efecto, habiendo calculado dichas medidas y habiéndose asegurado de la semejanza con la curva normal, es factible establecer una gran variedad de conclusiones acerca de la distribución de la variable que se esté manejando.

Se dice que el promedio y la desviación estándar son los parámetros de la distribución normal; esto es, son los valores que bastan para caracterizar a una distribución de datos cuantitativos continuos.

Debido a lo anterior, todo procedimiento estadístico que se base en el uso del promedio y de la desviación estándar para la obtención de conclusiones formará parte del campo de la estadística llamada paramétrica. De igual manera, los procedimientos estadísticos que no tienen su fundamento en el uso de dichos parámetros están integrados en el campo de la estadística no paramétrica, tal es el caso de las técnicas que utilizan los percentiles, las frecuencias o las series completas de datos sin ningún procedimiento de resumen.

Bibliografía

- Altman D. *Practical Statistics for Medical Research*. Londres: Chapman and Hall. 1991.
- Armitage E, Berry G. *Estadística para la investigación biomédica*. Madrid: Harcourt Brace. 1997.
- Daniel W. *Bioestadística: base para el análisis de las ciencias de la salud*. México: Limusa. 2006.
- Gerstman B. *Basic Biostatistics*. Massachusetts: Jones and Bartletts. 2008.
- Glantz S. *Primer of Biostatistics*, 6th ed. Washington, McGraw-Hill. 2002.
- Kirkwood B, Sterne, J. *Medical Statistics*, 2nd ed. Massachusetts, Blackwell Publishing. 2003.
- Motulsky H. *Intuitive Biostatistics*. 2nd ed. New York: Oxford University Press. 2010.