APUNTES DE BIOESTADISTICA

LUIS VILLARROEL
GONZALO VALDIVIA

Generalmente la investigación científica en medicina está dirigida al estudio de una determinada población. Esta población habitualmente la componen personas con cierta patología o alguna cualidad de interés.

Generalmente no se puede estudiar toda la población, es necesario tomar una muestra de ésta, estudiarla e inferir que los resultados que se obtienen de la muestra son representativos de lo que se habría obtenido en la población, si se hubiese estudiado.

Este proceso requiere el uso de la estadística en dos etapas:

- primero, obtener una estadística descriptiva de los datos muestrales;
- segundo, hacer inferencias a la población mediante estadística analítica.

Ambas etapas requieren seguir pasos en forma rigurosa, de modo que los resultados tengan validez.

En este primer capítulo revisaremos los elementos necesarios para hacer una buena estadística descriptiva de los datos.

En el segundo, revisaremos los test estadísticos que nos permitan hacer inferencias a la población.

DEFINICIONES

Población y Muestra

Generalmente las inquietudes de investigación nacen del desconocimiento que se tiene de alguna población de interés: personas que mueren a causa de un infarto, población que sufre de colelitiasis, afectados de cáncer gástrico, niños con bajo peso de nacimiento, diabéticos, etc.

Población: "Es el Conjunto total de objetos o de personas, con algo en común, de interés en un estudio".

¿Por qué no estudiar la población completa?

i Problemas presupuestarios. Es de un alto costo hacer un censo.
ii Limitaciones de tiempo. Además de que un estudio prolongado necesariamente debe considerar cambios que se produzcan en las variables importantes a causa del tiempo.

iii Dificultad de acceso. No necesariamente dificultad geográfica, sino de identificación de los individuos que componen la población (ejemplo: población de portadores de VIH).

Debido a estos problemas, debemos conformarnos con trabajar con una **muestra** de la población de interés.

Muestra: "Es un subconjunto de la población en estudio. Será el subconjunto que es realmente observado".

El objetivo, entonces, es trabajar con una muestra de la población de interés, pero a la vez queremos ganar información sobre la población de la cual proviene. Es decir, deseamos que las características de la población se vean reflejadas en la muestra que obtengamos.

Para cumplir con lo anterior, la muestra seleccionada debe cumplir con ciertos requisitos.
La muestra debe:

- Ser una muestra **aleatoria**.
- Ser de un **tamaño mínimo.**
- Ser una muestra **representativa** de la población.

Una muestra es **aleatoria** cuando todas las personas u objetos de la población tienen la misma probabilidad de ser elegidos en la muestra.

Una muestra es de **tamaño mínimo adecuado** cuando las inferencias que se puedan hacer en base a ésta tienen un error de estimación acotado (generalmente, el error máximo aceptado es de 5%).

Una muestra es **representativa** de una población cuando la(s) característica(s) más importantes de la población está(n) presente(s) en la misma proporción o promedio en la muestra.

Es decir, si la población tiene 30% de hombres y 70% de mujeres, esta proporción se debe mantener en la muestra. Si la edad promedio de la población es 50 años, en la muestra se observa más o menos lo mismo, etc.

Si una muestra es **aleatoria** y de **tamaño adecuado**, entonces esta suele ser además **representativa** de la población de interés.

Nótese que la aleatoriedad y el tamaño mínimo son elementos controlables (existen métodos de selección aleatoria de los datos y podemos calcular el tamaño mínimo adecuado). En cambio, la representatividad es una **cualidad** de la muestra obtenida.

MÉTODOS DE SELECCIÓN DE UNA MUESTRA ALEATORIA.

1. Muestreo Aleatorio Simple

Es una muestra en que cada sujeto u objeto tiene la misma probabilidad de ser seleccionado en la muestra.

Las formas usuales de seleccionar una muestra aleatoria simple es mediante una tabla de números aleatorios o una lista de números aleatorios generada por un computador. También se puede recurrir a una tómbola o una bolsa con papeles numerados para este tipo de muestreo.

Si se desea obtener una muestra aleatoria representativa de los alumnos de un colegio que tiene 800 alumnos en educación básica y 400 en media, de modo de ESTIMAR la edad promedio de los alumnos de todo el colegio, ¿Es conveniente una muestra aleatoria simple?

2. Muestreo Estratificado

Es una muestra en que se divide primero la población en estratos o grupos separados y luego se obtiene una muestra aleatoria simple al interior de cada estrato.

El muestreo aleatorio estratificado es llamado proporcional (o con afijación proporcional), si los estratos están presentes en la muestra en igual proporción que en la población.

En ocasiones, si un estrato presenta mucha variabilidad (o dispersión), es recomendable hacer un muestreo proporcional al tamaño de la variabilidad de cada estrato. Esta variante se denomina afijación no proporcional.

3. Muestreo Sistemático.

Este método es útil cuando se cuenta con una población **ordenada** de alguna forma conocida (por ejemplo, por número de ficha, por fecha de ingreso al hospital, etc.).

¿Cómo se lleva a cabo un muestreo sistemático?

Si se tiene:

- N (es el tamaño de la población),
- n (es el tamaño de la muestra) y
- k =N/n,

Una muestra aleatoria sistemática es aquella donde se selecciona un sujeto al azar de entre los primeros k pacientes en la población ordenada, seleccionando luego cada k-ésimo dato hasta completar los "n" necesarios en la muestra.

Por ejemplo, si la población es de tamaño N=5.000 y se quieren n=200 casos en la muestra, se deben seguir los siguientes pasos:

i Calcular k. En este caso, k=5000/200 = 25.
ii Seleccionar un sujeto al azar (muestreo aleatorio simple) de entre los primeros 25 casos en la muestra ordenada.
iii Posteriormente, seleccionar un sujeto cada 25, contando desde el primer sujeto seleccionado, hasta llegar al n-ésimo sujeto.

TIPOS DE VARIABLE:

Una vez tomada la muestra, cada sujeto que la compone será caracterizado según ciertas cualidades o cantidades de interés. Cada una de estas características, como la edad, sexo, estado civil, peso, etc., se denominadas **variables**.

Variable: "Característica que puede tomar uno o más valores en los elementos de la población". Visto de otra forma, si al mantener constantes las condiciones experimentales no es posible predecir el valor de una variable, entonces se está frente a una variable aleatoria.

Nos abocaremos a estudiar sólo **variables aleatorias**, para las cuales no es posible anticipar su resultado, aún cuando se intente controlar los demás factores que puedan afectarlas.

Todas las variables, con la sola excepción de las usadas como variables de **identificación** (nombre, número de ficha clínica, etc.), se pueden clasificar en uno de los 3 grupos siguientes:

Nominal: Sólo podemos clasificar sus valores en clases (o categorías), entre las cuales no se puede establecer ningún ordenamiento sugerido por la magnitud de sus valores.
Ejemplos: Sexo, Estado Civil, Profesión, Actividad.

Ordinal: Sus valores se pueden clasificar en categorías y si bién no tienen magnitudes asociadas, se pueden ordenar las clases.
Ejemplos: Nivel Socioeconómico, Apgar, Puntaje Apache de Gravedad cardíaca.

Intervalar: Existe un orden natural en sus valores y es posible cuantificar la diferencia entre dos valores intervalares. Generalmente tienen unidad de medida.

Una variable intervalar es **discreta** cuando sólo puede tomar un conjunto numerable de valores (por ejemplo: número de hijos); o bien es **continua** si puede tomar cualquier valor en un intervalo (por ejemplo.: peso, talla, Indice de Masa Corporal , etc).

***Notas**:*

1. *Una variable intervalar puede transformarse en ordinal o nominal construyendo **rangos** para ésta. Por ejemplo, el peso del recién nacido (intervalar), puede expresarse también como:*
   Ordinal: Hasta 2000 grs, 2001-3000, 3001-4000, 4001-Más.
   Nominal: Bajo Peso (<3000 grs), No Bajo Peso (>=3000 grs).

2. *El tamaño muestral que se requiere para describir y analizar una variable intervalar suele ser mucho menor que el requerido para analizar una nominal u ordinal.*

Los Dos Tipos de Variables a describir

Un paso importante en el estudio del comportamiento de una o más poblaciones, luego de tomar una muestra aleatoria de cada una, consiste en describir adecuadamente estas muestras, de modo que las **medidas resumen** que obtengamos reflejen bien el comportamiento poblacional.

La forma de describir las variables muestrales depende del TIPO al que pertenezca cada variable, y para efecto de simplificar esta descripción basta con considerar dos grandes tipos:

1. **Variables Categóricas**. Incluye a todas las variables para las cuales no es posible (y no tiene sentido) obtener su promedio. Incluye a las nominales (sexo, profesión, etc.), las ordinales que sólo tienen categorías ordenadas (Nivel socioeconómico, grado de dolor, etc.) y las intervalares en rangos (como el peso de nacimiento en rangos).

2. **Variables Numéricas**. Incluye a todas las variables para las cuales tiene sentido obtener su promedio. Incluye a todas las intervalares (edad, peso, talla) y las ordinales promediables (apgar, puntaje apache, etc).

DESCRIPCIÓN DE VARIABLES CATEGORICAS (NO PROMEDIABLES)
En este caso, las medidas resúmenes más adecuadas son el **número de casos** y el **porcentaje** en que se presenta cada categoría de la variable.

Al usar un computador para obtener las medidas resumen, éstas se presentan en una **tabla de frecuencias.** Estas tablas son también útiles en presentaciones orales, aunque no así en publicaciones.

Tablas de Frecuencias
Estas tablas sirven para resumir en forma ordenada el número de casos y porcentaje obtenido para cada categoría de una variable. Aunque hay muchas formas de tabular resultados, la presentación habitual de la tabla es la siguiente:

Variable	Nro.de Casos	Frecuencia Relativa	Porcentaje	Porcentaje Acumulado
categoría-1	n1	n1/n	100*n1/n	100*n1/n
categoría-2	n2	n2/n	100*n2/n	100*(n1+n2)/n
:	:	:	:	:
categoría-k	nk	nk/n	100*nk/n	100*(n1+..+nk)/n
	n	1.0	100.0	

La frecuencia relativa y el porcentaje muestran la misma información en diferente escala: la frecuencia relativa varía entre 0 y 1 y el porcentaje entre 0 y 100%. El porcentaje acumulado se usa cuando la variable tabulada está medida al menos a nivel ordinal.

Ejemplo: La siguiente datos son de una muestra de 12 recién nacidos: 5 que presentaron alguna patología de nacimiento y 7 sanos. Obtenga tablas de frecuencias de las variables tabulatdas. Use rangos para el peso del recién nacido: Hasta 3000 grs.; 3001-3300 grs.; y más de 3300 grs.

Ident	Sexo (M=Masc F=Fem)	Patología (S=Si N=No)	Peso Nac. (Gramos)	Edad Gest (Semanas)
1	M	S	2500	36
2	M	N	3000	39
3	M	N	3050	41
4	M	S	2900	40
5	M	S	2800	39
6	M	S	2590	38
7	F	N	3080	39
8	F	N	3500	40
9	F	N	3320	40
10	F	S	3005	39
11	F	N	3270	40
12	F	N	3330	36

Nota: Cuando se tabula una variable en rangos no es necesario que éstos tengan igual longitud. A veces es más útil recurrir a intervalos de uso habitual en la literatura respecto al tema.

Para "k" intervalos de igual longitud, determine los valores mínimo y máximo de la variable y calcule: Longitud= (Máximo - Mínimo)/k. Por ejemplo, si se quieren 3 intervalos de igual largo para el Peso RN, la longitud es: (3500-2500)/3=333.3. Entonces, los intervalos son: 2500-2833; 2834-3167 y 3168-3500.

El problema de estos rangos es que se puede tener intervalos vacíos o con cantidades de datos muy desbalanceados.

Presentación Gráfica de Variables Categóricas.

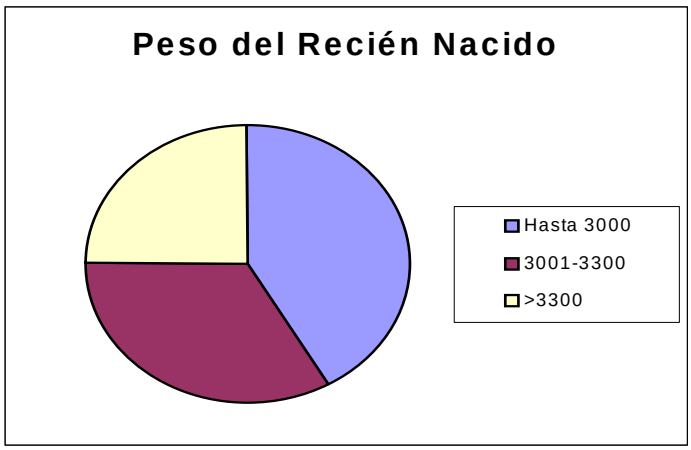
Las formas habituales de graficación de una tabla univariada son:

Barras Simples: Son gráficos de barras rectangulares cuya altura es proporcional al porcentaje que de casos en cada categoría o nivel de la variable. Si la variable tiene muchas categorías, una alternativa es hacer el gráfico con barras horizontales en vez de verticales.

Si el gráfico muestra una variable para una sola población también puede graficarse el número de casos en cada categoría. Si es de dos o más poblaciones debe graficarse el porcentaje para poder hacer comparaciones.

Gráfico Circular: También llamado Gráfico Sectorial o Torta, es un círculo dividido en porciones proporcionales al porcentaje de cada nivel respecto al total de datos. Cada porción se obtiene multiplicando las **frecuencias relativas** por 360□, obteniéndose los grados para cada porción de la torta.

Ejemplo: Gráfico de distribución porcentual del Peso RN obtenidos en la tabla de frecuencias del ejemplo previo.



DESCRIPCIÓN DE VARIABLES NUMERICAS (PROMEDIABLES)

Si la variable es intervalar u ordinal promediable, la mejor forma de describirla es mediante medidas que resuman la posición y dispersión de los datos.
Es decir, ahora necesitamos medidas que indiquen el **centro** u otras posiciones importantes de la distribución de la variable, además del grado de **variabilidad** respecto al valor central.

MEDIDAS DE POSICIÓN

Las medidas de posición tienen como objetivo resumir en un solo valor las mediciones obtenidas de una variable.
Las más importantes son las **medidas de tendencia central**, que tratan de ubicar el centro de la distribución, destacando el **promedio aritmético**, la **mediana** y la **moda.**

Promedio Aritmético

Este es el promedio de uso habitual en investigación en medicina. Se simboliza θ y se calcula como la suma de las mediciones de la variable dividido por el número de observaciones. Simbólicamente se escribe como:

$$\theta = \frac{\sum x}{n} = \frac{x_1+x_2+...+x_n}{n}$$

***Ejercicio**: Calcule la media aritmética de los pesos de nacimiento según patología. Según lo observado, ¿Existe alguna relación entre la patología y el peso?*

Mediana

La mediana es la observación justo al centro de la muestra, cuando ésta es ordenada en forma ascendente. Se simboliza generalmente como Md y su forma de cálculo es la siguiente:

1. Ordene los datos de menor a mayor.
2. Si el tamaño muestral n es impar, ubique la observación n/2 en la muestra ordenada. Este valor corresponde a la mediana.
3. Si n es par, promedie las dos observaciones al centro de la muestra ordenada. Este valor corresponde a la mediana.

La interpretación de la mediana de una variable es que el 50% de los casos muestrales tienen valores inferiores a la mediana y el otro 50% tiene valores superiores a ésta.

Una importante característica de la mediana es su poca sensibilidad ante valores extremos u "outliers". En cambio, el promedio puede sufrir cambios de importancia que la alejen del centro de los datos.

***Ejercicio**: Calcule la mediana del peso del recién nacido.*

Moda o Modo

La moda es el valor observado con mayor frecuencia en una variable y es utilizada generalmente cuando se tiene un gran conjunto de datos. Esta medida no es muy práctica cuando la variable es intervalar, dado lo difícil que ocurran al menos dos casos con un mismo valor.

MEDIDAS DE DISPERSIÓN

Las medidas de posición no son suficientes por sí solas para describir el comportamiento de una variable, ya que no nos dicen nada acerca de la variabilidad de los datos.

Las medidas de dispersión de uso habitual en medicina son el **rango**, la **varianza**, la **desviación estándar** y el **error estándar**.

Rango

Es la diferencia entre el valor máximo y mínimo de la variable. Por ejemplo, el peso del recién nacido tiene un rango de 3500-2500=1000 gr. Es decir, la diferencia entre el mínimo y el máximo es de 1000 gramos.
En ocasiones se opta por presentar los valores mínimo y máximo en vez del rango, ya que aportan más información sobre la dispersión de los datos.
El rango es muy sensible a outliers, ya que se construye justamente con los valores extremos. Además, el rango muestral siempre subestima al rango poblacional.

Varianza

Aunque no es la medida de dispersión más usada, es necesario calcularla para obtener la desviación estándar.
Si x1, x2,...,xn son las n observaciones muestrales de la variable X, la varianza, simbolizada S^2, se define como:

$$S^2 = \frac{\square(x_i - \theta)^2}{n-1}$$

Es decir, la varianza es una especie de promedio de las desviaciones cuadráticas de los datos con respecto al promedio. La razón por la que la varianza es poco utilizada es que el resultado queda expresado en la unidad de medida al cuadrado (por ejemplo, $kg^{\square}$, $mts^{\square}$, etc.), mientras que los datos y el promedio están expresados en la unidad de medida original.

Ejercicio: Calcule la varianza del peso del recién nacido.
(Peso Promedio: θ = 3028.75 grs)

Id	Peso	$x-\theta\square$	$\square(x_i - \theta)^2$
1	2500	-528.70	279576.56
2	3000	-28.75	826.56
3	3050	21.25	451.56

4	2900	-128.70	16576.56
5	2800	-228.70	52326.56
6	2590	-438.70	192501.56
7	3080	51.25	2626.56
8	3500	471.25	222076.56
9	3320	291.25	84826.56
10	3005	-23.75	564.06
11	3270	241.25	58201.56
12	3330	301.25	90751.56

□=1.001.306.25

Luego: S^2 = 1001306.25 / (12-1) = 91.027,84 grs^2

Desviación Estándar

Esta es la medida de dispersión de mayor uso en investigación científica y se deriva directamente de la varianza.

Si x1, x2,...,xn son las n observaciones muestrales de la variable X, la desviación estándar, simbolizada s, se define como:

$$s = \sqrt{\frac{\square(x_i - \theta)^2}{n-1}}$$

Nótese que si la varianza está en la unidad de medida al cuadrado, la desviación estándar está en la unidad de medida original de los datos.

Ejercicio *Calcule la desviación estándar del peso del recién nacido.*
s = √91027.84 = 301.7 grs.

En la descripción de los resultados de un estudio generalmente se mencionan tres valores:

- el número de casos (n),
- la media aritmética (θ) y
- la desviación estándar (s).

Por ejemplo, respecto al peso del recién nacido se dice que con n=12 casos, el promedio fue 3028.7 grs. y la desviación estándar 301.7 grs.

Generalmente se escribe: θ =3028.7 grs. ±□301.7 grs (n=12 casos)

El valor obtenido para s no quiere decir que todos los datos se sitúen entre θ - s y θ + s. Las reglas que sí se cumplen son:

1. Sin importar la distribución de los datos, al menos el 75% de los casos **siempre** se sitúa entre θ -2s y θ +2s.

2. Si la distribución de los datos es simétrica en torno al promedio, entonces:

-Aproximadamente el 68% de los casos se sitúa entre θ -s y θ +s
-Aproximadamente el 95% de los casos se sitúa entre θ -2s y θ +2s
-Aproximadamente el 99% de los casos se sitúa entre □-3s y θ +3s

Error Estándar
El error estándar es útil como medida de dispersión cuando se quieren presentar los resultados de una misma variable para diferentes grupos poblacionales, ya que es una dispersión **estandarizada** por el número de observaciones. El error estándar se calcula a partir de la desviación estándar, y se define como:

$$\text{Error Estándar} = \text{e.s.} = s/\sqrt{n}$$

Es decir, el error estándar es igual a la desviación estándar dividido por la raíz cuadrada del número de observaciones.

Se usa generalmente cuando la desviación estándar es muy grande y se quiere graficar el comportamiento del promedio de una variable en una o más poblaciones y sus respectivas variabilidades.

OTRAS MEDIDAS DE POSICIÓN: PERCENTILES

Para cualquier variable intervalar, un percentil de orden p (0<p<100%) es un valor muestral que :

- deja el p% de los datos bajo ese valor y
- el (100-p)% de los datos restantes sobre él.

El cálculo de percentiles requiere tener la muestra ordenada en forma ascendente según la variable a describir.

Por ejemplo, el percentil 20% de una variable X corresponde al valor en la muestra que deja un 20% de los valores observados bajo el percentil y el 80% restante sobre el percentil.

En general, para calcular un percentil en una muestra ordenada de tamaño "n", el valor Xp que corresponde a ese percentil se encuentra en la posición:

$$k = (n+1)*p/100$$

Es decir, si x(1),.., x(n) son los n valores ordenados de X, el percentil de orden p corresponde al valor en la posición x(k). Si k es un número entero, entonces x(k) queda perfectamente determinado. Si k tiene decimales hay que aproximarlo al entero más cercano. Si k tiene decimal 0.5 (3.5, 9.5, etc.), se promedian los valores superior e inferior a la posición k (3 y 4; 9 y 10, etc.).

***Ejemplo**: Calcule e interprete los percentiles 25 y 50 del peso del recién nacido.*

** Los valores ordenador de peso de nacimiento son:*
2500 2590 2800 2900 3000 3005 3050 3080 3270 3320 3330 3500

** Para calcular percentil 25: n=12 p=25. Luego, k = (12+1)*25/100 = 3.25*
De esta forma, el percentil 25 corresponde a x(3) = 2800 grs

** Para calcular percentil 50: n=12 p=50. Luego, k = (12+1)*50/100 = 6.5*
Así, el percentil 50 corresponde a (x(6)+x(7))/2 = 3027.5 grs

** Interpretación: "El 25% de los recién nacidos tienen peso de nacimiento inferior a 2800 grs, mientras que el 50% tiene peso inferior a 3027 grs."*

Algunos Percentiles Especiales

Los percentiles más utilizados en medicina son:

- los **cuartiles**, correspondientes a los percentiles 25%, 50% y 75%;
- los **deciles**, que dividen la muestra en grupos de 10%; y
- la **mediana**, que corresponde al percentil 50%, al segundo cuartil o al quinto decil.

Además, para construcción de patrones de normalidad se utilizan con frecuencia los percentiles 5%, 10%, 90% y 95%, de modo que datos muestrales que se sitúan entre los percentiles 5 y 95 se consideran "normales" y los casos bajo el percentil 5% o sobre el percentil 95% son considerados "anormales" o "patológicos".

NOTAS:

- Si los datos presentan una dispersión moderada, la presentación de los datos suele hacerse usando el **número de casos, promedio y desviación estándar**.

- Si los datos presentan mucha dispersión (o hay valores extremos u outliers), de modo que el promedio se ve distorsionado por estos valores, la presentación de los datos se hace usando el **número de casos, mediana y rango.**

- En ocasiones, se usa la **media geométrica** como alternativa al uso de la mediana, si hay mucha dispersión (actualmente esta opción se ha hecho muy popular), siempre acompañada del número de casos y del rango como medida de dispersión.

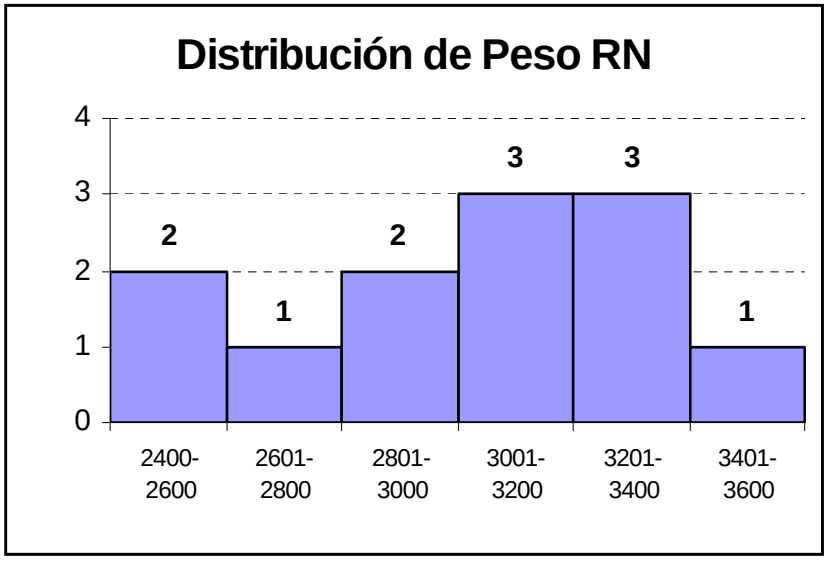
Representación Gráfica de Variables Promediables.

Histograma:
Un histograma es un gráfico de barras agrupadas que permite observar la **distribución** de una variable intervalar.

Si la variable es discreta (o discretizada), cada barra puede representar el porcentaje de casos que toma cada valor de la variable.

Si la variable es continua, cada barra representa un intervalo de valores. En este gráfico los intervalos **deben tener la misma longitud**, de modo que las barras muestren en forma **proporcional** el porcentaje que representa el intervalo en el total de datos.

Ejemplo: Construya un histograma para el peso RN. (Rangos de 2400-2600, 2601-2800, 2801-3000, 3001-3200, 3201-3400, 3401-3600 grs).



A medida que el tamaño muestral aumenta, es posible hacer intervalos más angostos de la variable, para observar mejor la distribución.. De esta forma, podremos observar el **grado de simetría** de los datos, con tres posibilidades:

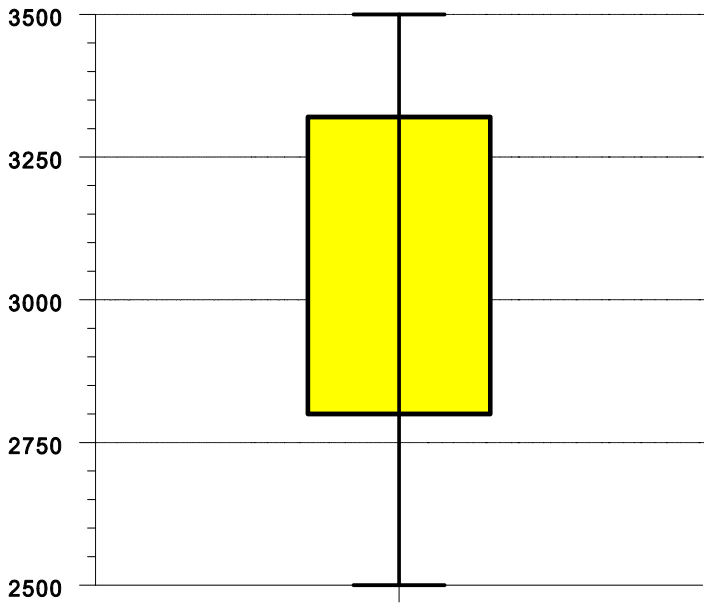
Simetría: Los datos se distribuyen en forma similar a ambos lados del centro. En este caso el promedio aritmético es igual a la mediana.

Asimetría a la izquierda: Los datos se concentran en menor proporción a la izquierda del punto central. En este caso el promedio es menor que la mediana.

Asimetría a la derecha: Los datos se concentran en menor proporción a la derecha del punto central. En este caso el promedio es mayor que la mediana.

Cajón con Bigotes (Box Plot)

Su objetivo es mostrar gráficamente medidas de posición, ya sea basado en el promedio y desviación estándar o en cuartiles. El gráfico siguiente muestra un box plot para el peso del recién nacido basado en percentiles.



Máximo = 3500 grs

Tercer Cuartil = 3320 grs

Mediana = 3027.5 grs

Primer Cuartil = 2800 grs

Mínimo = 2500 grs

El box plot es una caja en la que el borde inferior, la línea media y el borde superior corresponden al primer, segundo y tercer cuartil, respectivamente. Las lineas inferior y superior unen la caja con los valores mínimo y máximo, respectivamente.

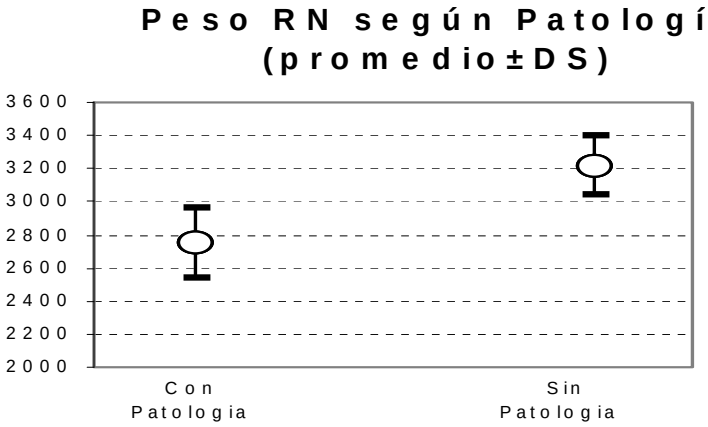
Si el gráfico se hace con promedio y desviación estándar, el borde inferior corresponde a θ -s, la línea media es θ y el borde superior es θ +s.

Gráfico de Promedio y Desviación Estándar (o Error Estándar)

Este gráfico es alternativo al cajón con bigotes hecho con el promedio y desviación estándar. Se usa principalmente cuando se grafica más de una población (o sea, más de un promedio ± desv.estándar), o cuando se grafica el promedio de una variable en el tiempo.

A continuación se muestra el gráfico del promedio y desviación estándar del peso del recién nacido según patología:

		θ -s	θ +s
Con patología:	n=5, θ =2759.0±210.8	2548.2	2969.8
Sin patología:	n=7 θ =3221.4±182.5	3038.9	3403.9



El gráfico de promedio ± error estándar se utiliza para representación gráfica cuando las desviaciones estándar de los datos son muy grandes y distorsionan la escala. A veces se grafica también θ□±2 (e.s), lo cual tiene como propiedad mostrar un **intervalo de confianza al 95%** para el promedio poblacional.

DESCRIPCIÓN DE DOS VARIABLES CATEGORICAS.

Este es el caso cuando se quiere describir simultáneamente dos variables nominales, ordinales no promediables e intervalares en rangos.

En esta situación el resultado se presenta generalmente en una **tabla de contingencia.** Al igual que en el caso de una variable categórica, las medidas resumen adecuadas son el número de casos y porcentaje, pero esta vez para cada combinación de niveles o categorías de las variables.

Por ejemplo, supongamos que un estudio busca determinar si existe relación entre fumar y cáncer pulmonar. Para esto, se tomaron 70 personas con cáncer y 380 sin cáncer y se observó en sus antecedentes si estas 450 personas eran fumadoras. La tabla resultante es la siguiente:

		Cáncer Pulmonar		
		Sí	**No**	**Total**
Fuma	**Sí**	30	120	150
	No	40	260	300
	Total	70	380	450

La tabla anterior permite observar los resultados del estudio, pero no incluye porcentajes. La pregunta es: ¿Qué porcentaje se debe calcular? ¿porcentaje de fumadores con cáncer o el porcentaje de enfermos que fuman?.

		Cáncer Pulmonar				
		Sí		**No**		**Total**
		n	**%**	**n**	**%**	
Fuma	**Sí**	30	42.9	120	31.6	150
	No	40	57.1	260	86.4	300
	Total	70	100	380	100	450

Nótese que la primera tabla muestra un mayor número de personas con cáncer en el grupo de no fumadores. De otra forma, el mayor número de fumadores se observa en el grupo sin cáncer. Ninguna de estas observaciones toma en cuenta el mayor número de no fumadores (o el mayor número de personas sin cáncer).

Presentación Gráfica.
La graficación de dos o más variables simultaneas generalmente muestra porcentajes, los cuales deben ser bien definidos, como en la tabla.

Gráfico Circular: En este caso se hace un gráfico para cada población. Es una buena alternativa a los gráficos de barras, principalmente en presentaciones.

Barras Agrupadas: Muestra los porcentajes en cada categoría de la variable en barras adyacentes, separado por cada población.

Barras Subdivididas Muestra una sola barra para cada población, todas de altura 100%, divididas en forma proporcional al porcentaje de cada

categoría de la variable. Es muy útil cuando se grafican muchas poblaciones.

Ejemplo: Construya un gráfico que muestre la relación entre patología de nacimiento y peso inferior a 3000 gramos.

DESCRIPCIÓN DE 2 VARIABLES PROMEDIABLES.

Cuando es de interés observar la relación entre dos variables numéricas, la medida resumen más utilizada es el Coeficiente de Correlación Lineal, que se simboliza "r". Hay dos métodos de obtener la correlación:

Correlación de Pearson: Llamado Coeficiente de Correlación Muestral de Pearson, se usa cuando las dos variables a relacionar con intervalares.

Correlación de Spearman: Llamado Coeficiente de Correlación por Rangos de Spearman, se usa cuando al menos una de las variables a relacionar es ordinal.

Como en este caso tenemos dos variables X e Y en una muestra de tamaño n, los datos son pares (x1,y1), (x2,y2), ..., (xn,yn). La forma de calcular el r de Pearson es la siguiente:

$$r = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sqrt{\sum (x-\bar{x})^2}\sqrt{\sum (y-\bar{y})^2}}$$

Ejemplo: Supongamos que se tomó una muestra de 5 madres, registrándose las variables edad materna y peso de sus hijos recién nacidos. Los datos son: (31,3500), (26,2990), (17,2800), (20,3000) y (28,3100).

El promedio de edad es $\bar{x}=24.4$ y de peso RN es $\bar{y}=3078$. El cálculo de la correlación de Pearson es el siguiente:

x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})(y-\bar{y})$	$(x-\bar{x})^2$	$(y-\bar{y})^2$
31	3500	6.6	422	2785.2	43.56	178084
26	2990	1.6	-88	-140.8	2.56	7744
17	2800	-7.4	-278	2057.2	54.76	77284
20	3000	-4.4	-78	343.2	19.36	6084
28	3100	3.6	22	79.2	12.96	484

Luego: $\sum (x-\bar{x})(y-\bar{y}) = 5124$ $\sum (x-\bar{x})^2=133.2$ $\sum (y-\bar{y})^2=269680$

Finalmente:

$$r = \frac{5124}{\sqrt{133.2*269680}} = 0.855$$

Para calcular la correlación de Spearman es necesario que al menos una de las variables sea ordinal. En este caso, es necesario calcular los **rangos** para cada variable por separado, es decir, el orden que tiene cada observación al interior de cada variable y luego calcular la correlación usando estos rangos en vez de los datos originales.

Por ejemplo, si calculamos la correlación de Spearman para la edad y peso RN, y los rangos de la edad son E1,E2,..E5 y los de peso son P1,P2,...,P5, el cálculo es:

$$r = \frac{\sum(E-\bar{E})(P-\bar{P})}{\sqrt{\sum(E-\bar{E})^2}\ \sqrt{\sum(P-\bar{P})^2}}$$

En este caso, el promedio de rangos de edad es $E=\bar{3}.0$ y de peso RN es $P=\bar{3}.\bar{0}$

E	P	$E-\bar{E}$	$P-\bar{P}$	$(E-\bar{E})(P-\bar{P})$	$(E-\bar{E})^2$	$(P-\bar{P})^2$
5	5	2	2	4	4	4
3	2	0	-1	0	0	1
1	1	-2	-2	4	4	4
2	3	-1	0	0	1	0
4	4	1	1	1	1	1

Luego: $\sum(E-\bar{E})(P-\bar{P}) = 9$ $\sum(E-\bar{E})^3=10$ $\sum(P-\bar{P})^3=10$

Finalmente:

$$r = \frac{9}{\sqrt{10*10}} = 0.900$$

Interpretación del Coeficiente de Correlación

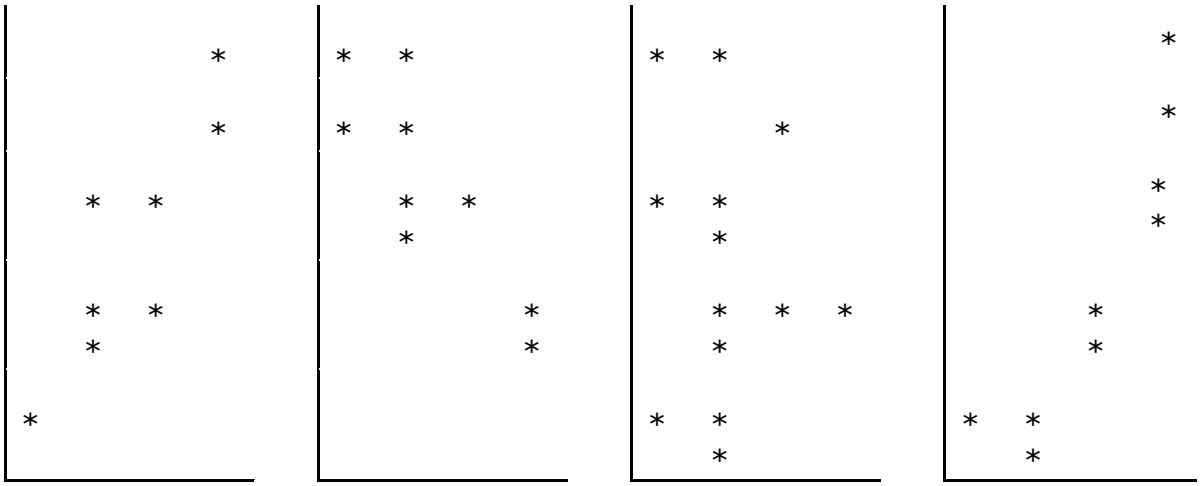
El coeficiente de correlación (Pearson o Spearman) varía **siempre** entre -1 y 1.

- Si r es cercano a 1, existe una asociación lineal **directa** entre X e Y.
- Si r es cercano a -1, existe una asociación lineal **inversa** entre X e Y.
- Si r es cercano a 0, **no existe** asociación lineal entre X e Y.

Algunos autores coinciden en valorar de la siguiente forma un coeficiente de correlación:

a. Si **r** está entre 0 y 0.25 (o -0.25) indica que **no hay** asociación lineal entre X e Y.
b. Si **r** está entre 0.25 y 0.50 (o entre -0.25 y -0.50) hay una **pobre** o **muy baja** asociación lineal entre X e Y.
c. Si **r** está entre 0.50 y 0.75 (o entre -0.50 y -0.75) hay una **buena** o **satisfactoria** asociación lineal entre X e Y.
d. Si **r** es mayor que 0.75 (o -0.75) hay una **muy buena** o **excelente** asociación lineal entre X e Y.

Representación Gráfica
la mejor forma de observar el grado de asociación entre X e Y es mediante un **gráfico de dispersión** (o **Scattergram**). La variable explicatoria X debe graficarse en el eje X o abscisa. La variable explicada Y debe graficarse en el eje Y u ordenada. Las posibilidades son las siguientes:



Asociación Lineal Directa (r +1) (r□0)

Asociación Lineal Inversa (r□-1)

Sin Asociación Lineal (r□0)

Sin Asociación Lineal

Nótese que el cuarto gráfico muestra una correlación cercana a cero (indicador de asociación lineal nula). Sin embargo, es claro que sí existe asociación entre X e Y, posiblemente de tipo exponencial.

ASOCIACIÓN DE UNA VARIABLE NO PROMEDIABLE Y UNA PROMEDIABLE

La asociación de una variable categórica y una numérica no requiere hacer cálculos adicionales. Se recurre a descripciones y gráficos ya vistos.

Medidas Resumen
Se obtienen medidas resumen de la variable numérica (n, $\bar{\theta}$□, s, Md, percentiles, etc.) para cada nivel de la variable categórica.

Por ejemplo, si interesa describir el peso del recién nacido según patología, las

medidas resumen pueden ser:

Con patología: n=5 θ =2759.0 ± 210.8
Sin patología: n=7 θ =3221.4 ± 182.5

Representación Gráfica
La representación gráfica en este caso son:

Box Plot: De la variable numérica, separado para cada nivel de la categórica.

θ□ **± s**: De la variable numérica, separado para cada nivel de la categórica.

θ□**± e.s.**: De la variable numérica, separado para cada nivel de la categórica.

ESTIMACIÓN DE PARÁMETROS POBLACIONALES.

La estadística descriptiva vista hasta ahora no sólo nos permite obtener un perfil del comportamiento de los datos muestrales; nos permite también obtener estimaciones de parámetros poblacionales, lo que generalmente es lo más importante.

Por una parte, en la población tenemos medidas de tendencia central, de posición y de dispersión que son fijas e invariables.
Estas medidas son llamadas **parámetros poblacionales** o simplemente **parámetros**. Por ejemplo, la talla promedio de la mujer chilena en la población es constante, así como su desviación estándar, cuartiles, etc.

Por otra parte, el cálculo de promedios, medianas, etc. obtenidos en una muestra son estimaciones de esos parámetros. Estas medidas son llamadas **parámetros estimados** o **estimadores**.
A diferencia de los parámetros poblacionales, los estimadores muestrales **no son únicos**, ya que varían al tomar distintas muestras de la misma población.
En su dimensión muestral, los estimadores son llamados **medidas resumen**, **estadígrafos** o **estadísticos**.

Los parámetros poblacionales habitualmente se simbolizan con una letra griega y sus estimadores con una letra latina. También es posible estimar distribuciones, conglomerados, etc.

Característica	**Parámetro**	**Estimador**
Media o Promedio	□	θ
Desv. estándar	□	s
Varianza	□²	s²
Error Estándar	□/√N	s/√n
Proporción	P ó □	p (frec.relativa)
Distribución	--	Histograma

Los estimadores muestrales también suelen representarse con la letra griega que representa al parámetro con un tilde ^ sobre ella. Por ejemplo: □ = □^ es un estimador de □.

Sesgo.
Se llama sesgo a la diferencia que existe entre un estimador y el parámetro al cual estima. Este sesgo (o error) se presenta cuando hay problemas en la selección de los sujetos que componen la muestra, la calidad de los instrumentos utilizados, la confiabilidad de las respuestas de personas encuestadas, etc. Evidentemente, mientras mayor es el sesgo, peor es la estimación del parámetro de interés. Mientras mayor es la precisión, menor es el sesgo cometido.

Cuando un estimador se "acerca" o "aproxima" cada vez más al parámetro al cual estima, a medida que el tamaño muestral aumenta, se denomina un estimador **insesgado**.

Finalmente, dado que una medida resumen obtenida en una muestra es al fin y al cabo un sólo valor destinado a estimar un parámetro, y dado además que este estimador no es único, suele llamarse un **estimador puntual**.