

Manual de Estadística

para Ciencias e Ingenierías



Alfredo Sánchez Alberca
asalber@ceu.es
<https://aprendeconalf.es>

Índice de contenidos

Prefacio	3
Licencia	3
1 Introducción a la Estadística	4
1.1 La estadística como herramienta científica	4
1.1.1 ¿Qué es la estadística?	4
1.1.2 La variabilidad de nuestro mundo	4
1.2 Población y muestra	5
1.2.1 Población estadística	5
1.2.2 Inconvenientes en el estudio de la población	5
1.2.3 Muestra estadística	6
1.2.4 Determinación del tamaño muestral	6
1.2.5 Tipos de razonamiento	9
1.3 Muestreo	11
1.3.1 Modalidades de muestreo	11
1.3.2 Muestreo aleatorio simple	12
1.3.3 Variables estadísticas	12
1.3.4 Tipos de estudios estadísticos	14
1.3.5 La tabla de datos	15
1.3.6 Fases del análisis estadístico	15
2 Estadística Descriptiva	17
2.1 Distribución de frecuencias	17
2.1.1 Clasificación de la muestra	18
2.1.2 Recuento de frecuencias	19
2.2 Frecuencias muestrales	19
2.2.1 Tabla de frecuencias	20
2.2.2 Construcción de clases	21
2.3 Representaciones gráficas	22
2.3.1 Diagrama de barras	22
2.3.2 Histograma	27
2.3.3 Diagrama de sectores	30
2.3.4 La distribución Normal	31
2.4 Datos atípicos	41
2.4.1 Tratamiento de los datos atípicos	43
2.5 Estadísticos muestrales	43

2.6	Estadísticos de posición	44
2.6.1	Media aritmética	44
2.6.2	Mediana	47
2.6.3	Moda	50
2.6.4	¿Qué estadístico de tendencia central usar?	51
2.6.5	Medidas de posición no centrales	52
2.7	Estadísticos de dispersión	55
2.7.1	Recorrido	55
2.7.2	Rango intercuartílico	55
2.7.3	Diagrama de caja y bigotes	56
2.7.4	Varianza y desviación típica	59
2.7.5	Coeficiente de variación	62
2.8	Estadísticos de forma	63
2.8.1	Coeficiente de asimetría	63
2.8.2	Coeficiente de apuntamiento o curtosis	67
2.8.3	Distribuciones no normales	70
2.9	Transformaciones de variables	72
2.9.1	Transformaciones lineales	73
2.9.2	Transformación de tipificación y puntuaciones típicas	73
2.9.3	Variables clasificadoras o factores	76
3	Regresión	79
3.1	Distribución de frecuencias conjunta	79
3.1.1	Frecuencias conjuntas	79
3.1.2	Distribución de frecuencias bidimensional	80
3.1.3	Diagrama de dispersión	81
3.1.4	Distribuciones marginales	83
3.2	Covarianza	83
3.3	Regresión	87
3.3.1	Modelos de regresión simple	88
3.3.2	Residuos o errores predictivos	88
3.3.3	Ajuste de mínimos cuadrados	89
3.3.4	Coeficiente de determinación	90
3.3.5	Coeficiente de correlación lineal	92
3.3.6	Distintos grados de correlación	93
3.3.7	Fiabilidad de las predicciones de un modelo de regresión	93
3.4	Regresión no lineal	94
3.4.1	Transformación de modelos de regresión no lineales	94
3.4.2	Relación exponencial	95
3.5	Riesgos de la regresión	99
3.5.1	La falta de ajuste no significa independencia	99
3.5.2	Datos atípicos en regresión	99
3.5.3	La paradoja de Simpson	101

4 Relaciones entre variables cualitativas	103
4.1 Relación entre atributos ordinales	103
4.1.1 Coeficiente de correlación de Spearman	103
4.2 Relación entre atributos nominales	105
4.2.1 Frecuencias teóricas o esperadas	106
4.2.2 Coeficiente chi-cuadrado χ^2	106
4.2.3 Coeficiente de contingencia	107
5 Probabilidad	109
5.1 Experimentos y sucesos aleatorios	109
5.1.1 Espacio de sucesos	111
5.1.2 Unión de sucesos	111
5.1.3 Intersección de sucesos	112
5.1.4 Contrario de un suceso	113
5.1.5 Diferencia de sucesos	113
5.1.6 Álgebra de sucesos	114
5.2 Definición de probabilidad	114
5.2.1 Definición clásica de probabilidad	114
5.2.2 Definición frecuentista de probabilidad	115
5.2.3 Definición axiomática de probabilidad	116
5.2.4 Interpretación de la probabilidad	118
5.3 Probabilidad condicionada	119
5.3.1 Experimentos condicionados	119
5.3.2 Probabilidad condicionada	120
5.3.3 Probabilidad del suceso intersección	120
5.3.4 Independencia de sucesos	121
5.4 Espacio probabilístico	121
5.4.1 Árboles de probabilidad con variables dependientes	122
5.4.2 Árboles de probabilidad con variables independientes	122
5.5 Teorema de la probabilidad total	123
5.5.1 Teorema de la probabilidad total	124
5.6 Teorema de Bayes	126
5.7 Epidemiología	127
5.7.1 Prevalencia	127
5.7.2 Incidencia	128
5.7.3 Tasa de incidencia o Riesgo absoluto	128
5.7.4 Prevalencia vs Incidencia	128
5.7.5 Comparación de riesgos	129
5.7.6 Riesgo atribuible o diferencia de riesgos RA	129
5.7.7 Riesgo relativo RR	130
5.7.8 Odds	132
5.7.9 Odds ratio OR	132
5.7.10 Riesgo relativo vs Odds ratio	133

5.8	Tests diagnósticos	136
5.8.1	Sensibilidad y especificidad de un test diagnóstico	136
5.8.2	Valores predictivos de un test diagnóstico	138
5.8.3	Razón de verosimilitud de un test diagnóstico	139
6	Estimación de parámetros poblacionales	142
6.1	Distribuciones muestrales	142
6.1.1	Distribución de la media muestral para muestras grandes ($n \geq 30$)	145
6.1.2	Distribución de una proporción muestral para muestras grandes ($n \geq 30$)	147
6.2	Estimadores	147
6.3	Estimación puntual	149
6.4	Estimación por intervalos	154
6.4.1	Error de estimación	156
6.5	Intervalos de confianza para una población	157
6.5.1	Intervalo de confianza para la media de una población normal con varianza conocida	158
6.5.2	Intervalo de confianza para la media de una población normal con varianza desconocida	161
6.5.3	Intervalo de confianza para la media de una población no normal .	163
6.5.4	Intervalo de confianza para la varianza de una población normal .	164
6.5.5	Intervalo de confianza para una proporción	166
6.6	Intervalos de confianza para la comparación dos poblaciones	169
6.6.1	Intervalo de confianza para la diferencia de medias de poblaciones normales con varianzas conocidas	169
6.6.2	Intervalo de confianza para la diferencia de medias de dos poblaciones normales con varianzas desconocidas e iguales	170
6.6.3	Intervalo de confianza para la diferencia de medias de dos poblaciones normales con varianzas desconocidas y distintas	173
6.6.4	Intervalo de confianza para el cociente de varianzas	175
6.6.5	Intervalo de confianza para la diferencia de proporciones	177
7	Contrastes de hipótesis paramétricos	180
7.1	Hipótesis estadística y tipos de contrastes	180
7.1.1	Contraste de hipótesis	180
7.1.2	Tipos de contrastes de hipótesis	181
7.1.3	Hipótesis nula e hipótesis alternativa	181
7.1.4	Contrastes de hipótesis paramétricos	182
7.2	Metodología para realizar un contraste de hipótesis	183
7.2.1	Estadístico del contraste	183
7.2.2	Regiones de aceptación y de rechazo	184
7.2.3	Errores en un contraste de hipótesis	185
7.2.4	Riesgos de los errores de un contraste de hipótesis	186

7.2.5	Determinación de las regiones de aceptación y de rechazo en función del riesgo α	187
7.2.6	Riesgo β y tamaño del efecto	189
7.2.7	Potencia de un contraste	189
7.2.8	Cálculo del riesgo β y de la potencia $1 - \beta$	189
7.2.9	Relación del riesgo β y el tamaño del efecto δ	190
7.2.10	Relación entre los riesgos α y β	192
7.2.11	Relación de los riesgos de error y el tamaño muestral	193
7.3	Curva de potencia	195
7.3.1	p -valor de un contraste de hipótesis	196
7.3.2	Regla de decisión de un contraste	196
7.3.3	Pasos para la realización de un contraste de hipótesis	197
7.4	Contrastes paramétricos más importantes	197
7.5	Contraste para la media de una población normal con varianza conocida	198
7.6	Contraste para la media de una población normal con varianza desconocida	198
7.6.1	Determinación del tamaño muestral en un contraste para la media	200
7.7	Contraste para la media de una población con varianza desconocida y muestras grandes	201
7.8	Contraste para la varianza de una población normal	201
7.9	Contraste para proporción de una población	202
7.10	Contraste de comparación de medias de dos poblaciones normales con varianzas conocidas	203
7.11	Contraste de comparación de medias de dos poblaciones normales con varianzas desconocidas e iguales	204
7.12	Contraste de comparación de medias de dos poblaciones normales con varianzas desconocidas	205
7.13	Contraste de comparación de varianzas de dos poblaciones normales	206
7.14	Contraste de comparación de proporciones de dos poblaciones	207
7.15	Realización de contrastes mediante intervalos de confianza	208
8	Análisis de la Varianza	210
8.1	Análisis de la varianza de 1 factor	210
8.1.1	El contraste de ANOVA	210
8.1.2	Test de comparaciones múltiples y por parejas	214
8.2	ANOVA de dos o más factores	214
8.2.1	ANOVA de dos factores con dos niveles cada factor	215
8.2.2	ANOVA de dos factores con tres o más niveles en algún factor	223
8.2.3	ANOVA de tres o más factores	223
8.2.4	Factores fijos y Factores aleatorios	224
8.3	ANOVA de medidas repetidas	225
8.3.1	ANOVA de medidas repetidas + ANOVA de una o más vías	228
8.4	Ánálisis de la covarianza: ANCOVA	229

Prefacio

¡Bienvenida/os al manual de Estadística!

Este libro es una introducción a la Estadística básica y el cálculo de probabilidades para alumnos de grados de ciencias e ingenierías.

Este libro se complementa con los siguientes recursos:

- Colección de problemas resueltos
- Prácticas de Estadística con R

Licencia

Esta obra está bajo una licencia Reconocimiento – No comercial – Compartir bajo la misma licencia 3.0 España de Creative Commons. Para ver una copia de esta licencia, visite <https://creativecommons.org/licenses/by-nc-sa/3.0/es/>.

Con esta licencia eres libre de:

- Copiar, distribuir y mostrar este trabajo.
- Realizar modificaciones de este trabajo.

Bajo las siguientes condiciones:

- **Reconocimiento Debe reconocer los créditos de la obra de la manera especificada por el autor o el licenciador (pero no de una manera que sugiera que tiene su apoyo o apoyan el uso que hace de su obra).
- **No comercial No puede utilizar esta obra para fines comerciales.
- **Compartir bajo la misma licencia Si altera o transforma esta obra, o genera una obra derivada, sólo puede distribuir la obra generada bajo una licencia idéntica a ésta.

Al reutilizar o distribuir la obra, tiene que dejar bien claro los términos de la licencia de esta obra.

Estas condiciones pueden no aplicarse si se obtiene el permiso del titular de los derechos de autor.

Nada en esta licencia menoscaba o restringe los derechos morales del autor.

1 Introducción a la Estadística

1.1 La estadística como herramienta científica

1.1.1 ¿Qué es la estadística?

Definición 1.1 (Estadística). La *estadística* es una rama de las matemáticas que se encarga de la recogida, análisis e interpretación de datos.

El papel de la Estadística es extraer información de los datos para adquirir el conocimiento necesario para tomar decisiones.



Figura 1.1: Propósito de la Estadística

La estadística es imprescindible en cualquier disciplina científica o técnica donde se manejen datos, especialmente si son grandes volúmenes de datos, como por ejemplo en Física, Química, Medicina, Psicología, Economía o Ciencias Sociales.

Pero, *¿por qué es necesaria la Estadística?*

1.1.2 La variabilidad de nuestro mundo

El científico trata de estudiar el mundo que le rodea; un mundo que está lleno de variaciones que dificultan la determinación del comportamiento de las cosas.

La estadística actúa como disciplina puente entre la realidad del mundo y los modelos matemáticos que tratan de explicarla, proporcionando una metodología para evaluar las discrepancias entre la realidad y los modelos teóricos.

Esto la convierte en una herramienta indispensable en las ciencias aplicadas que requieran el análisis de datos y el diseño de experimentos.

1.2 Población y muestra

1.2.1 Población estadística

Definición 1.2 (Población). Una *población* es un conjunto de elementos definido por una o más características que tienen todos los elementos, y sólo ellos. Cada elemento de la población se llama *individuo*.

Definición 1.3 (Tamaño poblacional). El número de individuos de una población se conoce como *tamaño poblacional* y se representa como N .

Ejemplo 1.1. En unas elecciones generales a la presidencia del gobierno, la población serían todos los individuos del estado con derecho a voto. En el estudio de una enfermedad, la población sería todas las personas que tienen la enfermedad. Y en un proceso de control de calidad en la fabricación de un fármaco, la población estaría formada por todos los fármacos que se producen en la fábrica.

A veces, no todos los elementos de la población están accesibles para su estudio. Entonces se distingue entre:

- **Población Teórica:** Conjunto de elementos a los que se quiere extraer los resultados del estudio.
- **Población Estudiada:** Conjunto de elementos realmente accesibles en el estudio.

Ejemplo 1.2. En el caso del estudio de una enfermedad, la población teórica sería todas las personas que contraigan la enfermedad, incluso si aún no han nacido, mientras que la población estudiada se limitaría al número de personas enfermas que realmente podemos estudiar (obsérvese que incluso quedarían fuera las personas enfermas pero de las que no podemos conseguir información).

1.2.2 Inconvenientes en el estudio de la población

El científico estudia un determinado fenómeno en una población para comprenderlo, obtener conocimiento sobre el mismo, y así poder controlarlo. Pero, para tener un conocimiento completo de la población *es necesario estudiar todos los individuos de la misma*. Sin embargo, esto no siempre es posible por distintos motivos:

- El tamaño de la población es infinito, o bien es finito pero demasiado grande.
- Las pruebas a que se someten los individuos son destructivas.
- El coste, tanto de dinero como de tiempo, que supondría estudiar a todos los individuos es excesivo.

1.2.3 Muestra estadística

Cuando no es posible o conveniente estudiar todos los individuos de la población, se estudia sólo una parte de la misma.

Definición 1.4 (Muestra). Una *muestra* es un subconjunto de la población.

Definición 1.5 (Tamaño muestral). Al número de individuos que componen la muestra se le llama *tamaño muestral* y se representa por n .

Habitualmente, el estudio de una población se realiza a partir de muestras extraídas de dicha población.

Generalmente, el estudio de la muestra sólo aporta conocimiento aproximado de la población. Pero en muchos casos es *suficiente*.

1.2.4 Determinación del tamaño muestral

Una de las preguntas más interesantes que surge inmediatamente es: *¿cuántos individuos es necesario tomar en la muestra para tener un conocimiento aproximado pero suficiente de la población?*

La respuesta depende de varios factores, como la variabilidad de la población o la fiabilidad deseada para las extrapolaciones que se hagan hacia la población.

Por desgracia no se podrá responder hasta casi el final del curso, pero en general, cuantos más individuos haya en la muestra, más fiables serán las conclusiones sobre la población, pero también será más lento y costoso el estudio.

Ejemplo 1.3. Para entender a qué nos referimos cuando hablamos de un tamaño muestral suficiente para comprender lo que ocurre en la población, podemos utilizar el siguiente símil en que se trata de comprender el motivo que representa una fotografía.

Una fotografía digital está formada por multitud de pequeños puntitos llamados pixels que se dispone en una enorme tabla de filas y columnas (cuantas más filas y columnas haya se habla de que la foto tiene más resolución). Aquí la población estaría formada por todos y cada uno de los píxeles que forman la foto. Por otro lado cada pixel tiene un color y es la variedad de colores a lo largo de los pixels la que permite formar la imagen de la fotografía.

¿Cuántos píxeles debemos tomar en una muestra para averiguar la imagen de la foto?

La respuesta depende de la variabilidad de colores en la foto. Si todos los pixels de la foto son del mismo color, entonces un sólo pixel basta para desvelar la imagen. Pero, si la foto tiene mucha variabilidad de colores, necesitaremos muchos más pixels en la muestra para descubrir el motivo de la foto.

La imagen siguiente contiene una muestra pequeña de píxeles de una foto. ¿Puedes averiguar el motivo de la foto?

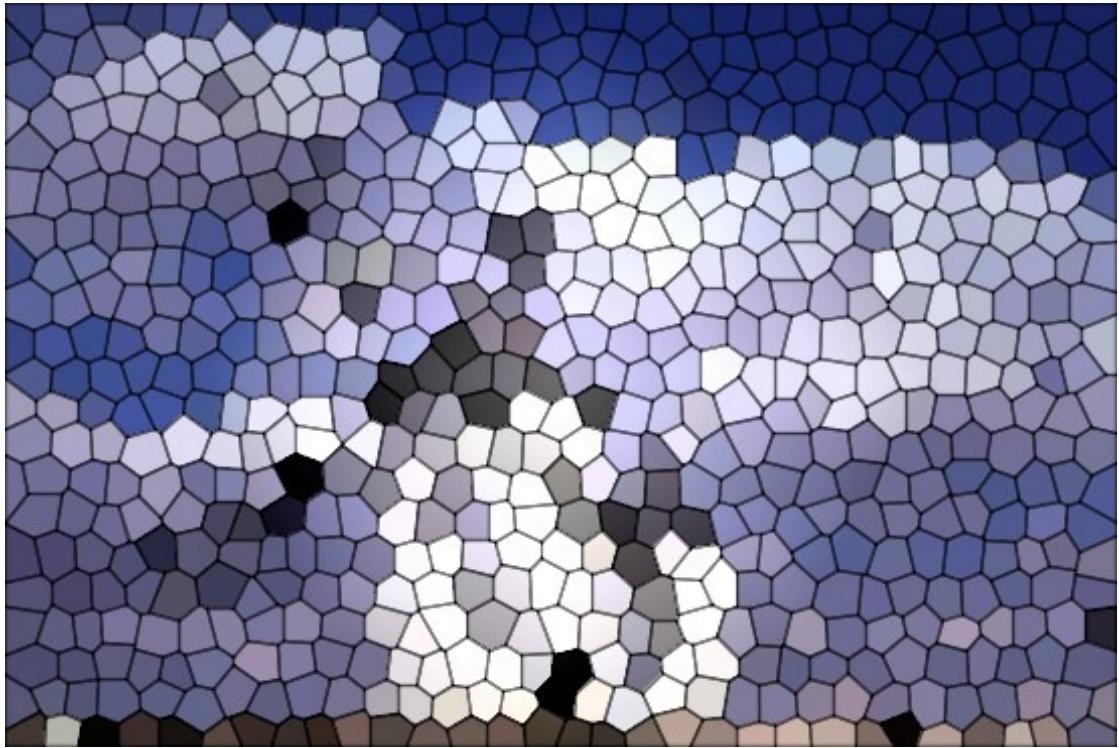


Figura 1.2: Muestra pequeña de píxeles de una foto.

¡Con una muestra pequeña es difícil averiguar el contenido de la imagen!

Seguramente no has podido averiguar el motivo de la fotografía, porque en este caso el número de píxeles que hemos tomado en la muestra es insuficiente para comprender toda la variabilidad de colores que hay en la foto.

La siguiente imagen contiene una muestra mayor de píxeles. ¿Eres capaz de adivinar el motivo de la foto ahora?

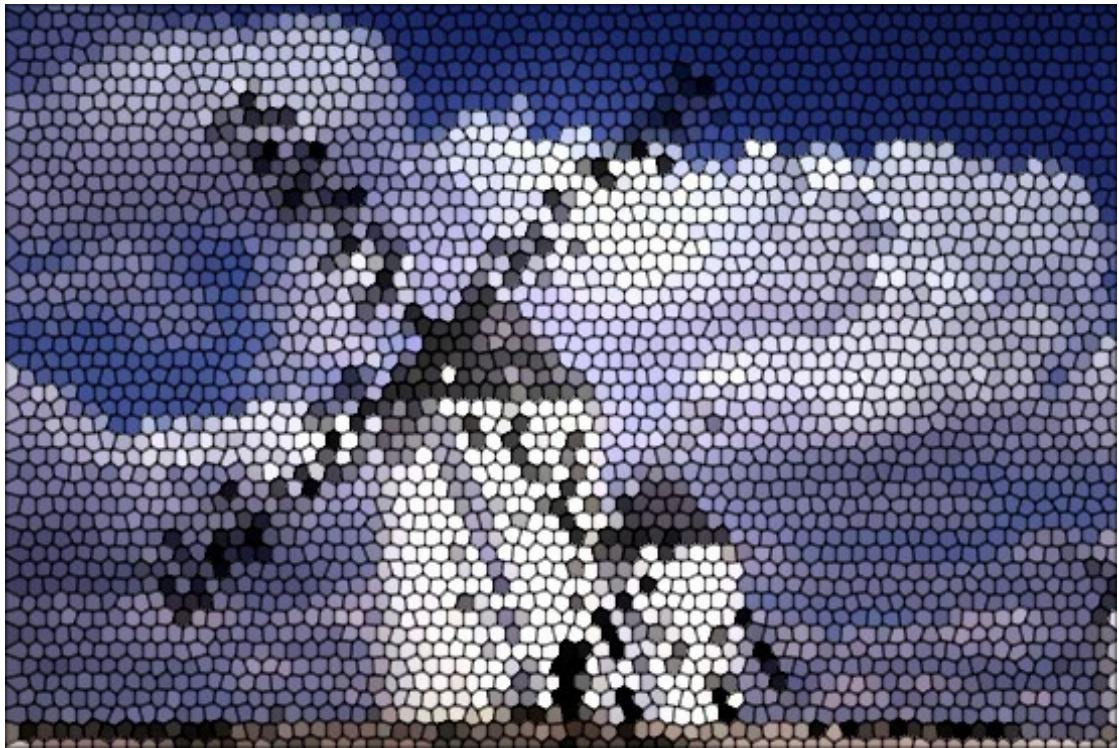


Figura 1.3: Muestra mayor de píxeles de una foto.

¡Con una muestra mayor es posible desvelar el motivo de la foto!

Y aquí está la población completa.



Figura 1.4: Población de píxeles de una foto.

Lo importante es que *¡No es necesario conocer todos los píxeles para averiguar la imagen!*

1.2.5 Tipos de razonamiento

Así pues, habitualmente realizaremos el estudio de la población a partir de muestras y luego trataremos de extrapolar lo observado en la muestra al resto de la población. A este tipo de razonamiento que saca conclusiones desde la muestra hacia la población se le conoce como *razonamiento induutivo*.

Población



De lo general
a lo particular

Deducción

Inducción

De lo particular
a lo general



Muestra

Figura 1.5: Tipos de razonamiento.

- *Características de la deducción:* Si las premisas son ciertas, garantiza la certeza de las conclusiones (es decir, si algo se cumple en la población, también se cumple en la muestra). Sin embargo, *no aporta conocimiento nuevo!*

- *Características de la inducción:* No garantiza la certeza de las conclusiones (si algo se cumple en la muestra, puede que no se cumpla en la población, así que ¡cuidado con las extrapolaciones!), pero *¡es la única forma de generar conocimiento nuevo!*

La estadística se apoya fundamentalmente en el razonamiento inductivo ya que utiliza la información obtenida a partir de muestras para sacar conclusiones sobre las poblaciones. A diferencia del razonamiento deductivo que va de lo general a lo particular, o en nuestro caso de la población a la muestra, el razonamiento inductivo no garantiza la certeza de las conclusiones, por lo que debemos ser cuidadosos a la hora de generalizar sobre la población lo observado en al muestra, ya que si la muestra no es representativa de la población o contiene sesgos, las conclusiones pueden ser erróneas.

1.3 Muestreo

Definición 1.6 (Muestreo). El proceso de selección de los elementos que compondrán una muestra se conoce como *muestreo*.

.
- **Muestreo No Aleatorio:** Los individuos se eligen de forma no aleatoria. Algunos individuos tienen más probabilidad de ser seleccionados que otros.

Sólo las técnicas aleatorias evitan el sesgo de selección, y por tanto, garantizan la representatividad de la muestra extraída, y en consecuencia la validez de las conclusiones.

Las técnicas no aleatorias no sirven para hacer generalizaciones, ya que no garantizan la representatividad de la muestra. Sin embargo, son menos costosas y pueden utilizarse en estudios exploratorios.

1.3.2 Muestreo aleatorio simple

Dentro de las modalidades de muestreo aleatorio, el tipo más conocido es el *muestreo aleatorio simple*, caracterizado por:

- Todos los individuos de la población tienen la misma probabilidad de ser elegidos para la muestra.
- La selección de individuos es con reemplazamiento, es decir, cada individuo seleccionado es devuelto a la población antes de seleccionar al siguiente (y por tanto no se altera la población de partida).
- Las sucesivas selecciones de un individuo son independientes.

La única forma de realizar un muestreo aleatorio es asignar un número a cada individuo de la población (*censo*) y realizar un sorteo aleatorio.

1.3.3 Variables estadísticas

Todo estudio estadístico comienza por la identificación de las características que interesa estudiar en la población y que se medirán en los individuos de la muestra.

Definición 1.7 (Variable estadística). Una *variable estadística* es una propiedad o característica medida en los individuos de la población.

Los *datos* son los valores observados en las variables estadísticas.



Figura 1.6: Variables estadísticas.

Estas características pueden ser de distintos tipos de acuerdo a su naturaleza y su escala:

- **Variables cualitativas o atributos:** Miden cualidades no numéricas. Pueden ser:

– **Nominales:** No existe un orden entre las categorías.

Ejemplo: El color de pelo o el sexo.

– **Ordinales:** Existe un orden entre las categorías. Ejemplo: El nivel de estudios o la gravedad de una enfermedad.

- **Variables cuantitativas:** Miden cantidades numéricas. Pueden ser:

– **Discretas:** Toman valores numéricos aislados (habitualmente números enteros).

Ejemplo: El número de hijos o el número de coches en una familia.

– **Continuas:** Pueden tomar cualquier valor en un intervalo real.

Ejemplo: El peso o la estatura.

Las variables cualitativas y discretas se conocen también con *variables categóricas* y sus valores *categorías*.

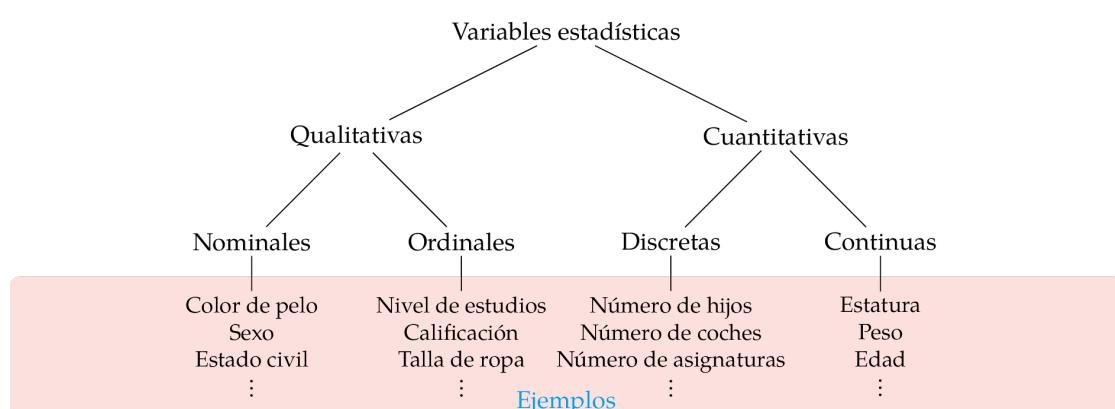


Figura 1.7: Tipos de variables estadísticas.

1.3.3.1 Elección del tipo de variable más apropiado

En ocasiones una característica puede medirse mediante variables de distinto tipo.

Ejemplo 1.4. Si una persona fuma o no podría medirse de diferentes formas:

- Fuma: si/no. (Nominal)

- Nivel de fumador: No fuma / ocasional / moderado / bastante / empedernido. (Ordinal)
- Número de cigarros diarios: 0,1,2,... (Discreta)

En estos casos es preferible usar variables cuantitativas a cualitativas. Dentro de las cuantitativas es preferible usar las continuas a las discretas y dentro de las cualitativas es preferible usar ordinales a nominales pues aportan más información.



Figura 1.8: Cantidad de información de los tipos de variables estadísticas.

De acuerdo al papel que juegan en el estudio las variables también pueden clasificarse como:

- **Variables independientes:** Variables que supuestamente no dependen de otras variables en el estudio. Habitualmente son las variables manipuladas en el experimento para ver su efecto en las variables dependientes. Se conocen también como *variables predictivas*.
- **Variables dependientes:** Variables que supuestamente dependen de otras variables en el estudio. No son manipuladas en el experimento y también se conocen como *variables respuesta*.

Ejemplo 1.5. En un estudio sobre el rendimiento de los alumnos de un curso, la inteligencia de los alumnos y el número de horas de estudio diarias serían variables independientes y la nota del curso sería una variable dependiente.

1.3.4 Tipos de estudios estadísticos

Dependiendo de si se manipulan las variables independientes existen dos tipos de estudios:

- **Experimentales:** Cuando las variables independientes son manipuladas para ver el efecto que producen en las variables dependientes.

Ejemplo 1.6. En un estudio sobre el rendimiento de los estudiantes en un test, el profesor manipula la metodología de estudio para crear dos o más grupos con metodologías de estudio distintas.

- **No experimentales:** Cuando las variables independientes no son manipuladas. Esto no significa que sea imposible hacerlo, sino que es difícil o poco ético hacerlo.

Ejemplo 1.7. En un estudio un investigador puede estar interesado en el efecto de fumar sobre el cáncer de pulmón. Aunque es posible, no sería ético pedirle a los pacientes que fumasen para ver el efecto que tiene sobre sus pulmones. En este caso, el investigador podría estudiar dos grupos de pacientes, uno con cáncer de pulmón y otro sin cáncer, y observar en cada grupo cuántos fuman o no.

Los estudios experimentales permiten identificar causas y efectos entre las variables del estudio, mientras que los no experimentales sólo permiten identificar relaciones de asociación entre las variables.

1.3.5 La tabla de datos

Las variables a estudiar se medirán en cada uno de los individuos de la muestra, obteniendo un conjunto de datos que suele organizarse en forma de matriz que se conoce como tabla de datos__.

En esta tabla cada columna contiene la información de una variable y cada fila la información de un individuo.

Ejemplo 1.8. La siguiente tabla contiene información de las variables Nombre, Edad, Sexo, Peso y Altura de una muestra de 6 personas.

Nombre	Edad	Sexo	Peso(Kg)	Altura(cm)
José Luis Martínez	18	H	85	179
Rosa Díaz	32	M	65	173
Javier García	24	H	71	181
Carmen López	35	M	65	170
Marisa López	46	M	51	158
Antonio Ruiz	68	H	66	174

1.3.6 Fases del análisis estadístico

Normalmente un estudio estadístico pasa por las siguientes etapas:

1. El estudio comienza por el diseño previo del mismo en el que se establezcan los objetivos del mismo, la población, las variables que se medirán y el tamaño muestral requerido.

2. A continuación se seleccionará una muestra representativa del tamaño establecido y se medirán las variables en los individuos de la muestra obteniendo la tabla de datos. De esto se encarga el *Muestreo*.
3. El siguiente paso consiste en describir y resumir la información que contiene la muestra. De esto se encarga la *Estadística Descriptiva*.
4. La información obtenida es proyectada sobre un modelo matemático que intenta explicar el comportamiento de la población y el modelo se valida. De todo esto se encarga la *Estadística Inferencial*.
5. Finalmente, el modelo validado nos permite hacer predicciones y sacar conclusiones sobre la población de partida con cierta confianza.

1.3.6.1 El ciclo estadístico

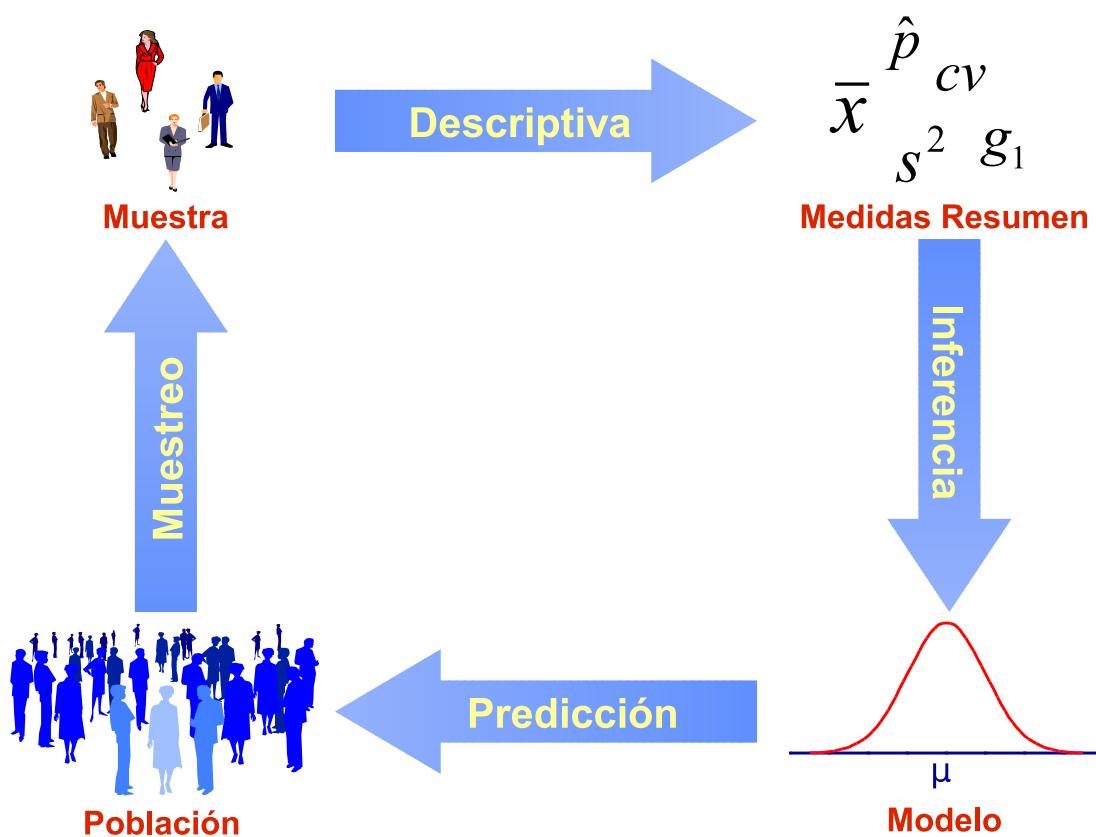


Figura 1.9: El ciclo estadístico.

2 Estadística Descriptiva

La estadística descriptiva es la parte de la estadística encargada de representar, analizar y resumir la información contenida en la muestra.

Tras el proceso de muestreo, es la siguiente etapa de todo estudio estadístico y suele consistir en:

1. Clasificar, agrupar y ordenar los datos de la muestra.
2. Tabular y representar gráficamente los datos de acuerdo a sus frecuencias.
3. Calcular medidas que resuman la información que contiene la muestra (*estadísticos muestrales*).

Interpretación

No tiene poder inferencial, por lo que nunca deben sacarse conclusiones sobre la población a partir de las medidas resumen que aporta la Estadística Descriptiva.

2.1 Distribución de frecuencias

El estudio de una variable estadística comienza por medir la variable en los individuos de la muestra y clasificar los valores obtenidos.

Existen dos formas de clasificar estos valores:

- **Sin agrupar:** Ordenar todos los valores obtenidos en la muestra de menor a mayor. Se utiliza con atributos y variables discretas con pocos valores diferentes.
- **Agrupados:** Agrupar los valores en clases (intervalos) y ordenar dichas clases de menor a mayor. Se utiliza con variables continuas y con variables discretas con muchos valores diferentes.

2.1.1 Clasificación de la muestra

Consiste colocar juntos los valores iguales y ordenarlos si existe un orden entre ellos.



Figura 2.1: Clasificación de la muestra.

2.1.2 Recuento de frecuencias



Figura 2.2: Recuento de frecuencias

2.2 Frecuencias muestrales

Definición 2.1 (Frecuencias muestrales). Dada una muestra de tamaño n de una variable X , para cada valor de la variable x_i observado en la muestra, se define

- **Frecuencia Absoluta n_i :** Es el número de veces que el valor x_i aparece en la muestra.

- **Frecuencia Relativa** f_i : Es la proporción de veces que el valor x_i aparece en la muestra.

$$f_i = \frac{n_i}{n}$$

- **Frecuencia Absoluta Acumulada** N_i : Es el número de valores en la muestra menores o iguales que x_i .

$$N_i = n_1 + \cdots + n_i = N_{i-1} + n_i$$

- **Frecuencia Relativa Acumulada** F_i : Es la proporción de valores en la muestra menores o iguales que x_i .

$$F_i = \frac{N_i}{n}$$

2.2.1 Tabla de frecuencias

Al conjunto de valores observados en la muestra junto a sus respectivas frecuencias se le denomina **distribución de frecuencias** y suele representarse mediante una **tabla de frecuencias**.

Valores de X	Frecuencia Absoluta	Frecuencia Relativa	Frecuencia Absoluta Acumulada	Frecuencia Relativa Acumulada
x_1	n_1	f_1	N_1	F_1
\vdots	\vdots	\vdots	\vdots	\vdots
x_i	n_i	f_i	N_i	F_i
\vdots	\vdots	\vdots	\vdots	\vdots
x_k	n_k	f_k	N_k	F_k

Ejemplo 2.1 (Variable cuantitativa y datos no agrupados). El número de hijos en 25 familias es:

1, 2, 4, 2, 2, 2, 3, 2, 1, 1, 0, 2, 2, 0, 2, 2, 1, 2, 2, 3, 1, 2, 2, 1, 2

La tabla de frecuencias del número de hijos en esta muestra es

x_i	n_i	f_i	N_i	F_i
0	2	0.08	2	0.08
1	6	0.24	8	0.32
2	14	0.56	22	0.88
3	2	0.08	24	0.96
4	1	0.04	25	1
\sum	25	1		

Ejemplo 2.2 (Variable cuantitativa y datos agrupados). Se ha medido la estatura (en cm) de 30 universitarios obteniendo:

179, 173, 181, 170, 158, 174, 172, 166, 194, 185, 162, 187, 198, 177, 178, 165, 154, 188, 166, 171, 175, 182, 167, 169, 172, 186, 172, 176, 168, 187.

La tabla de frecuencias de la estatura en esta muestra es

x_i	n_i	f_i	N_i	F_i
(150, 160]	2	0.07	2	0.07
(160, 170]	8	0.27	10	0.34
(170, 180]	11	0.36	21	0.70
(180, 190]	7	0.23	28	0.93
(190, 200]	2	0.07	30	1
\sum	30	1		

2.2.2 Construcción de clases

Cada intervalo de agrupación de datos se denomina **clase** y el centro del intervalo se llama **marca de clase**.

A la hora de agrupar los datos en clases hay que tener en cuenta lo siguiente:

- El número de intervalos no debe ser muy grande ni muy pequeño. Una regla orientativa es tomar un número de intervalos próximo a \sqrt{n} o $\log_2(n)$.
- Los intervalos no deben solaparse y deben cubrir todo el rango de valores. Es indiferente si se abren por la izquierda y se cierran por la derecha o al revés.
- El valor más pequeño debe caer dentro del primer intervalo y el más grande dentro del último.

Ejemplo 2.3 (Variable cualitativa). Los grupos sanguíneos de una muestra de 30 personas son:

A, B, B, A, AB, 0, 0, A, B, B, A, A, A, AB, A, A, A, B, 0, B, B, B, A, A, A, 0, A, AB, 0.

La tabla de frecuencias del grupo sanguíneo en esta muestra es

x_i	n_i	f_i
0	5	0.16
A	14	0.47
B	8	0.27
AB	3	0.10
\sum	30	1

Advertencia

Obsérvese que en este caso las frecuencias acumuladas no tienen sentido al no existir un orden entre los valores de la variable.

2.3 Representaciones gráficas

La tabla de frecuencias también suele representarse gráficamente. Dependiendo del tipo de variable y de si se han agrupado o no los datos, se utilizan distintos tipos de gráficos:

- Diagrama de barras
- Histograma
- Diagrama de líneas o polígonos.
- Diagrama de sectores.

2.3.1 Diagrama de barras

Un **diagrama de barras** consiste en un conjunto de barras, una para cada valor o categoría de la variable, dibujadas sobre unos ejes cartesianos.

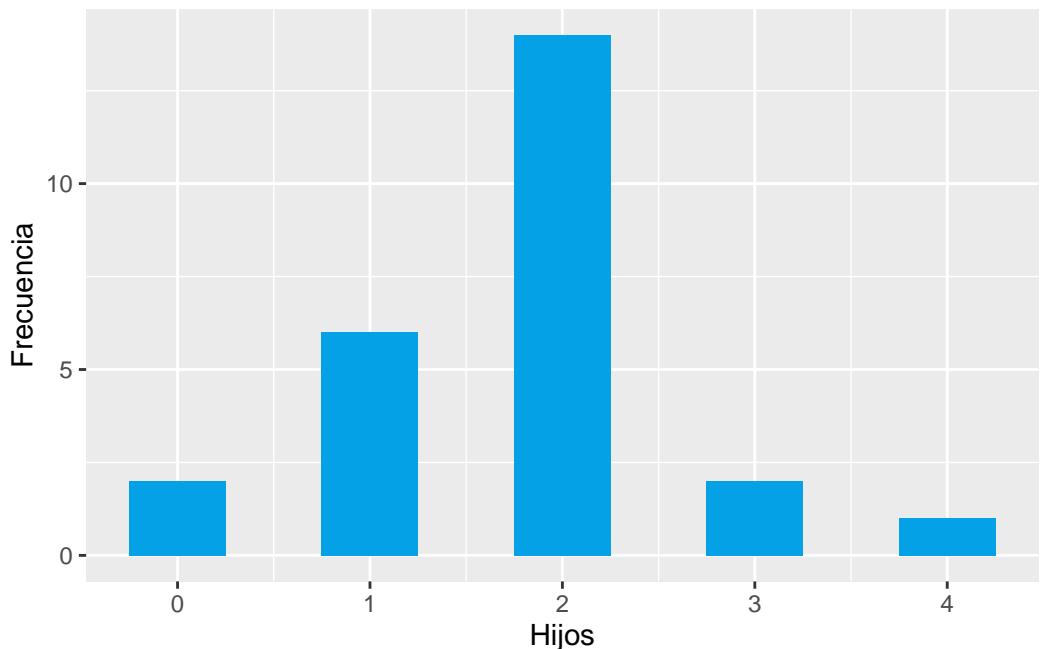
Habitualmente los valores o categorías de la variable se representan en eje X , y las frecuencias en el eje Y . Para cada valor o categoría se dibuja una barra con la altura correspondiente a su frecuencia. La anchura de la barra no es importante pero las barras deben aparecer claramente separadas unas de otras.

Dependiendo del tipo de frecuencia representada en el eje Y se tienen diferentes tipos de diagramas de barras.

En ocasiones se dibuja un polígono, conocido como **polígono de frecuencias**, uniendo mediante segmentos los puntos más altos de cada barra.

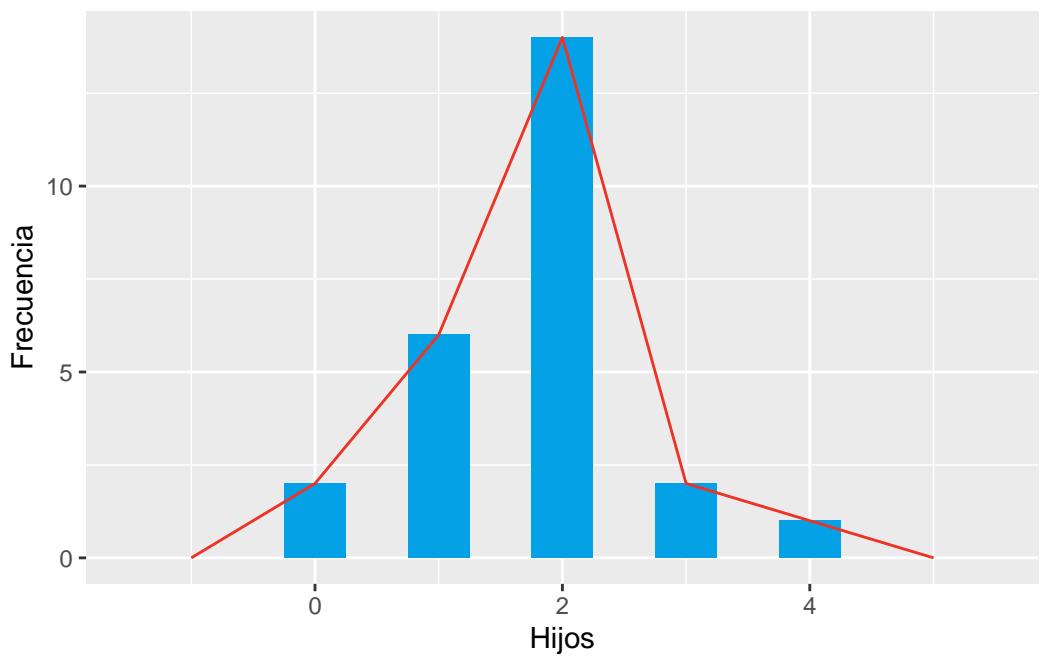
Ejemplo 2.4. El diagrama de barras que aparece a continuación muestra la distribución de frecuencias absolutas del número de hijos en la muestra anterior.

```
df <- read.csv("datos/hijos-coches.csv")
p <- ggplot(df, aes(x=Hijos)) +
  geom_bar(fill=blueceu, width = 0.5) +
  ylab("Frecuencia")
p
```



Y a continuación se muestra el polígono de frecuencias.

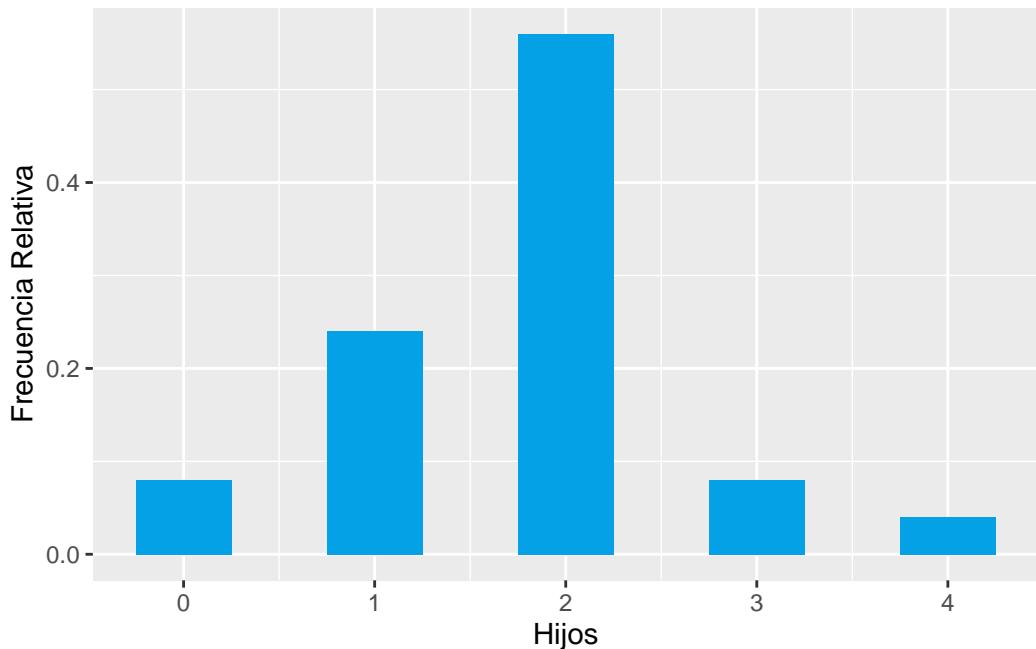
```
p <- p +
  geom_freqpoly(bins=5, col=redceu)
p
```



El diagrama de barras que aparece a continuación muestra la distribución de frecuencias relativas del número de hijos en la muestra anterior.

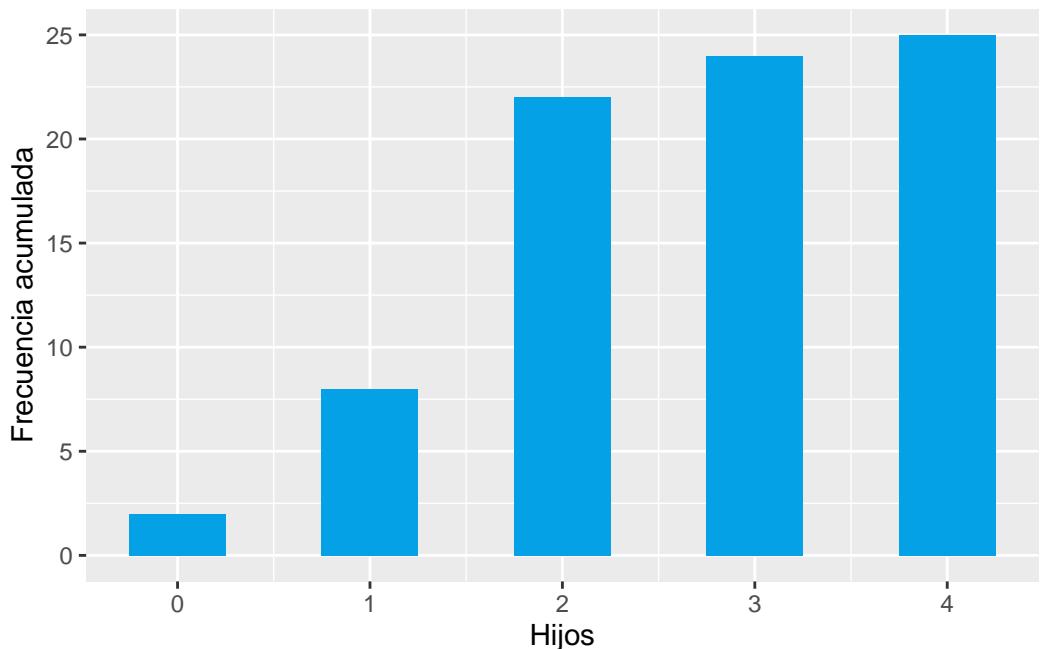
```
p <- ggplot(df, aes(x=Hijos)) +  
  geom_bar(aes(y = ..prop..), fill=blueceu, width = 0.5) +  
  ylab("Frecuencia Relativa")  
p
```

Warning: The dot-dot notation (`..prop..`) was deprecated in ggplot2 3.4.0.
i Please use `after_stat(prop)` instead.



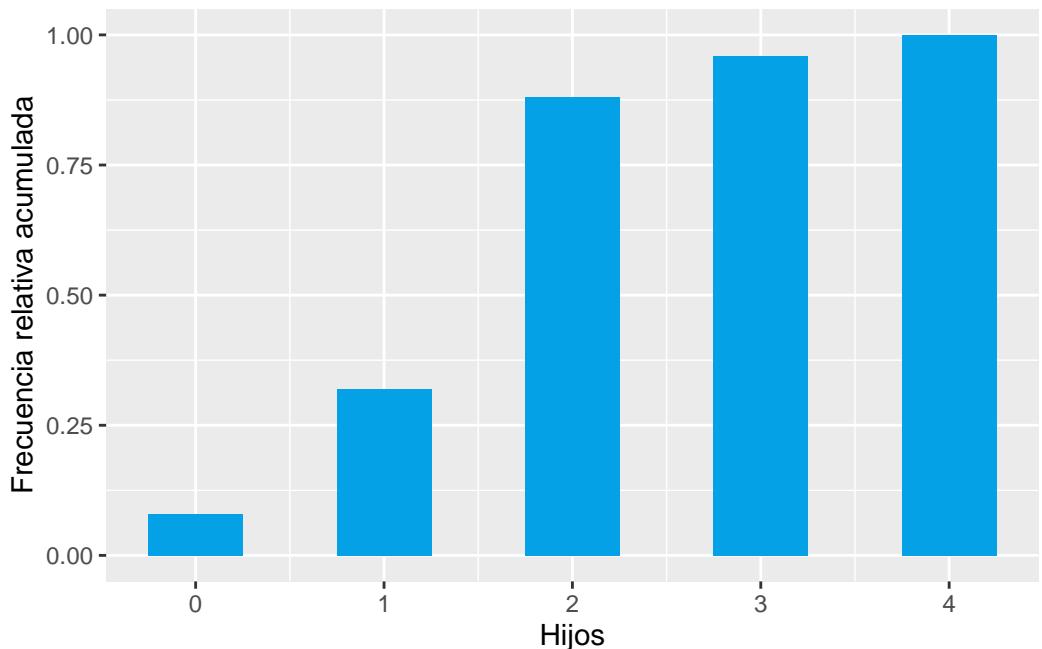
El diagrama de barras que aparece a continuación muestra la distribución de frecuencias absolutas acumuladas del número de hijos en la muestra anterior.

```
p <- ggplot(df, aes(x=Hijos)) +  
  geom_bar(aes(y = cumsum(..count..)), fill=blueceu, width = 0.5) +  
  ylab("Frecuencia acumulada")  
p
```



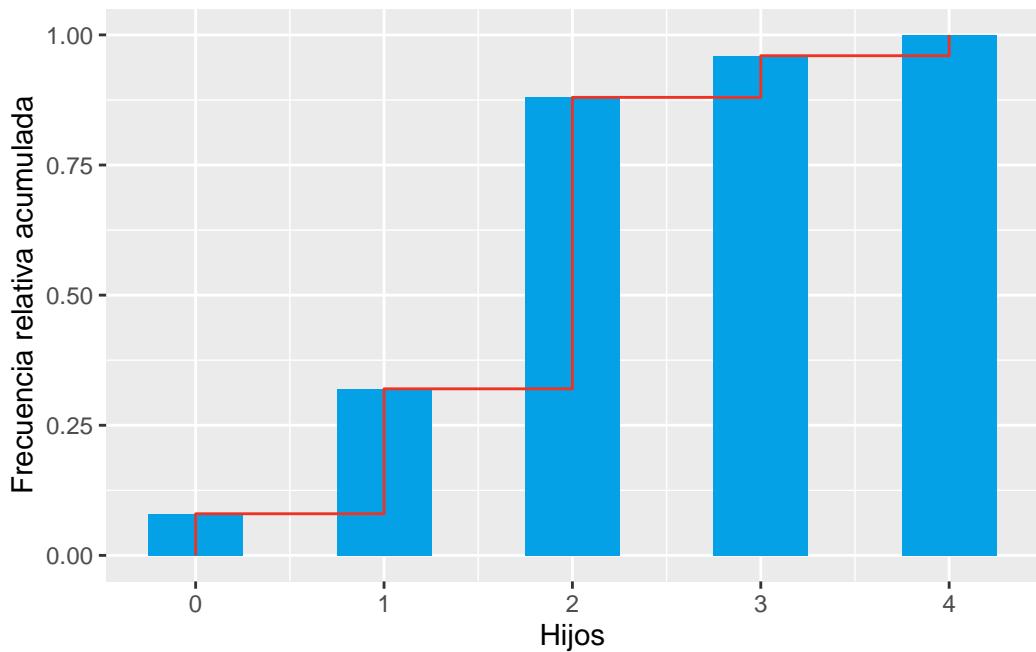
Y el diagrama de barras que aparece a continuación muestra la distribución de frecuencias relativas acumuladas del número de hijos en la muestra anterior.

```
p <- ggplot(df, aes(x=Hijos)) +  
  geom_bar(aes(y = cumsum(..prop..)), fill=blueceu, width = 0.5) +  
  ylab("Frecuencia relativa acumulada")  
p
```



Finalmente, el último diagrama muestra el polígono de frecuencias relativas acumuladas.

```
df.freq <- count(df, Hijos) %>%
  mutate(f = n / sum(n), N = cumsum(n), F = N / sum(n))
x <- unlist(lapply(df.freq$Hijos, rep, 2))
F <- c(0, head(unlist(lapply(df.freq$F, rep, 2)), -1))
df2 <- data.frame(Hijos = x, F = F)
p <- ggplot(df, aes(x=Hijos)) +
  geom_bar(aes(y = cumsum(..prop..)), fill=blueceu, width = 0.5) +
  geom_line(data = df2, aes(x=Hijos, y=F, group=1), col=redceu) +
  ylab("Frecuencia relativa acumulada")
p
```



2.3.2 Histograma

Un *histograma* es similar a un diagrama de barras pero para datos agrupados.

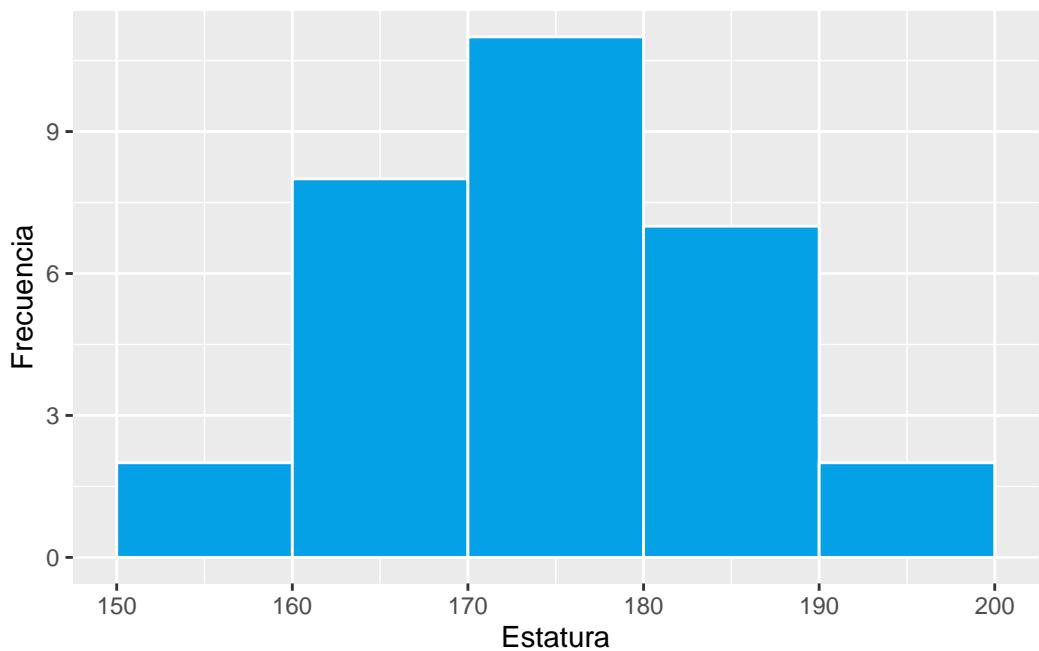
Habitualmente las clases o intervalos de agrupación se representan en el eje *X*, y las frecuencias en el eje *Y*. Para cada clase se dibuja una barra de altura la correspondiente frecuencia. A diferencia del diagrama de barras, la anchura del la barra coincide con la anchura de las clases y no hay separación entre dos barras consecutivas.

Dependiendo del tipo de frecuencia representada en el eje *Y* existen distintos tipos de histogramas.

Al igual que con el diagrama de barras, se puede dibujar un *polígono de frecuencias* uniendo los puntos centrales más altos de cada barra con segmentos.

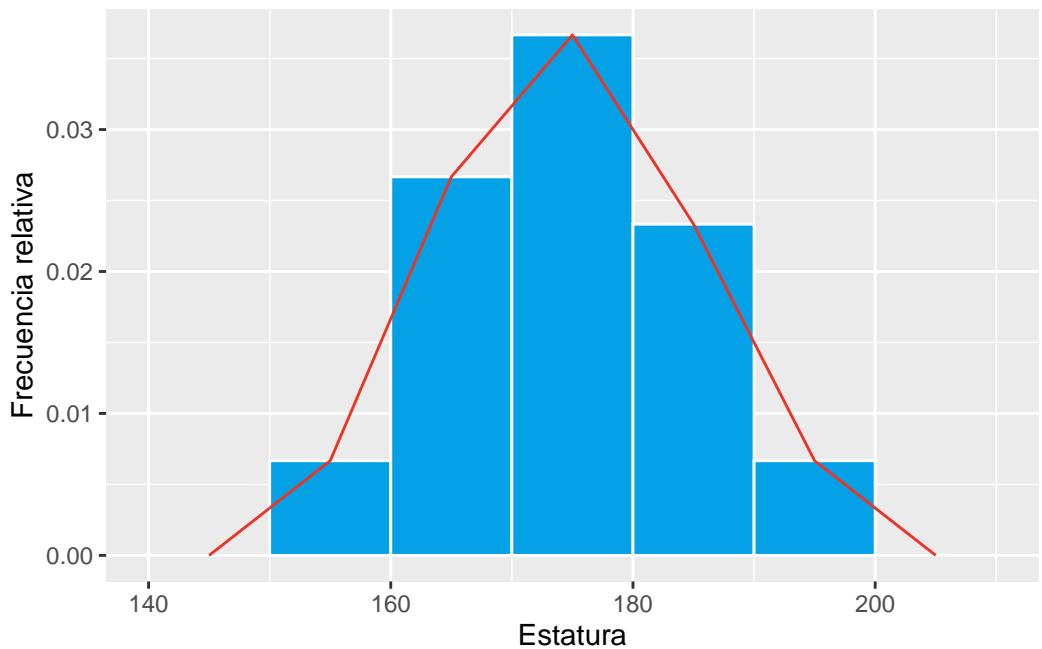
Ejemplo 2.5. El siguiente histograma muestra la distribución de frecuencias absolutas de las estaturas.

```
df <- read.csv("datos/estatura-peso.csv")
p <- ggplot(df, aes(x=Estatura)) +
  geom_histogram(breaks = seq(150, 200, 10), col="white", fill=blueceu) +
  ylab("Frecuencia")
p
```



El siguiente histograma muestra la distribución de frecuencias relativas con el polígono de frecuencias.

```
breaks <- seq(150, 200, 10)
p <- ggplot(df, aes(x=Estatura)) +
  geom_histogram(aes(y = ..density..), breaks = breaks, col="white", fill=blueceu) +
  geom_freqpoly(aes(y = ..density..), col=redceu, breaks = breaks) +
  ylab("Frecuencia relativa")
p
```



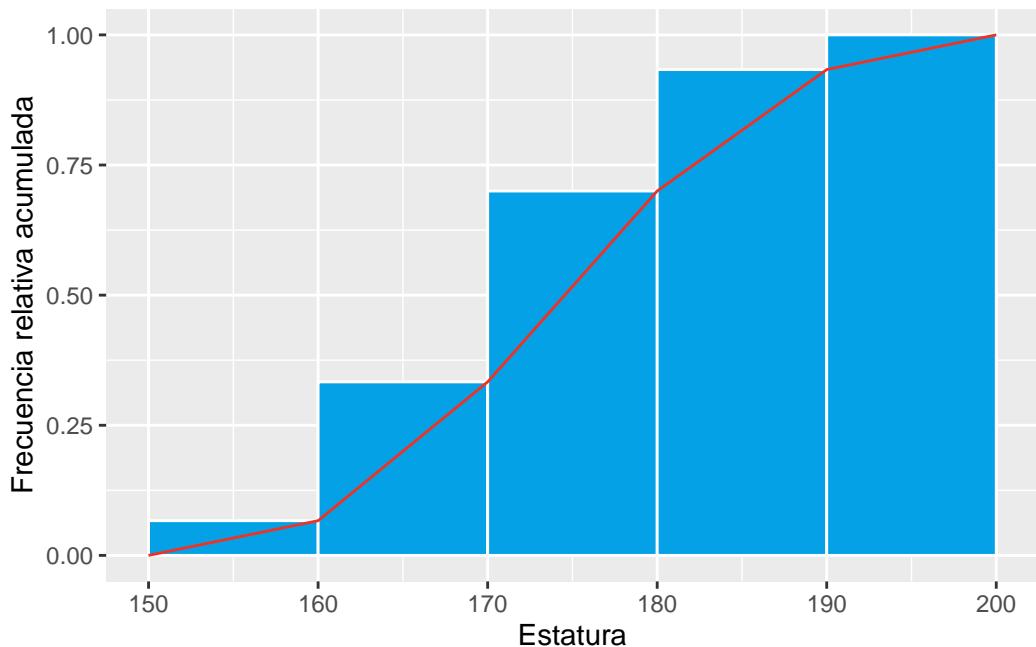
El polígono de frecuencias acumuladas (absolutas o relativas) se conoce como **ojiva**.

Ejemplo 2.6. El histograma y la ojiva siguientes muestran la distribución de frecuencias relativas acumuladas de estaturas.

```

breaks <- seq(150, 200, 10)
p <- ggplot(df, aes(x=Estatura)) +
  geom_histogram(aes(y = cumsum(..count..)/sum(..count..)), breaks = breaks, col="white")
  ylab("Frecuencia relativa acumulada")
df.p <- ggplot_build(p)$data[[1]]
x <- c(df.p$xmin[1], df.p$xmax)
y <- c(0, df.p$ymax)
df2 <- data.frame(x, y)
p <- p +
  geom_line(data = df2, aes(x = x, y = y, group = 1), col="redceu")
p

```



Obsérvese que en la ojiva se unen los vértices superiores derechos de cada barra con segmentos, en lugar de los puntos centrales, ya que no se consigue alcanzar la frecuencia acumulada correspondiente a la clase hasta que no se alcanza el final del intervalo.

2.3.3 Diagrama de sectores

Un *diagrama de sectores* consiste en un círculo dividido en porciones, uno por cada valor o categoría de la variable. Cada porción se conoce como *sector* y su ángulo o área es proporcional a la correspondiente frecuencia del valor o categoría.

Los diagramas de sectores pueden representar frecuencias absolutas o relativas, pero no pueden representar frecuencias acumuladas, y se utilizan sobre todo con atributos nominales. Para atributos ordinales o variables cuantitativas es mejor utilizar diagramas de barras, ya es más fácil percibir las diferencias en una dimensión (altura de las barras) que en dos dimensiones (áreas de los sectores).

Ejemplo 2.7. El diagrama de sectores siguiente muestra la distribución de frecuencias relativas de los grupos sanguíneos.

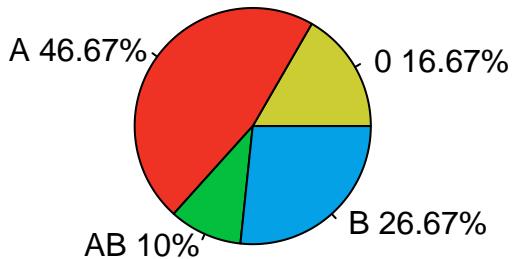
```
df <- read.csv("datos/grupo-sanguineo.csv")
tab <- table(df[["Grupo.Sanguineo.Hijo"]])
labels <- names(tab)
pctg <- round(tab/sum(tab)*100, 2)
labels <- paste(labels, pctg) # add percents to labels
```

```

labels <- paste(labels,"%",sep="") # ad % to labels
pie(tab, main="Distribución de los grupos sanguíneos", labels=labels, col=c(greenceu, re

```

Distribución de los grupos sanguíneos



2.3.4 La distribución Normal

Las distribuciones con diferentes propiedades presentan formas distintas.

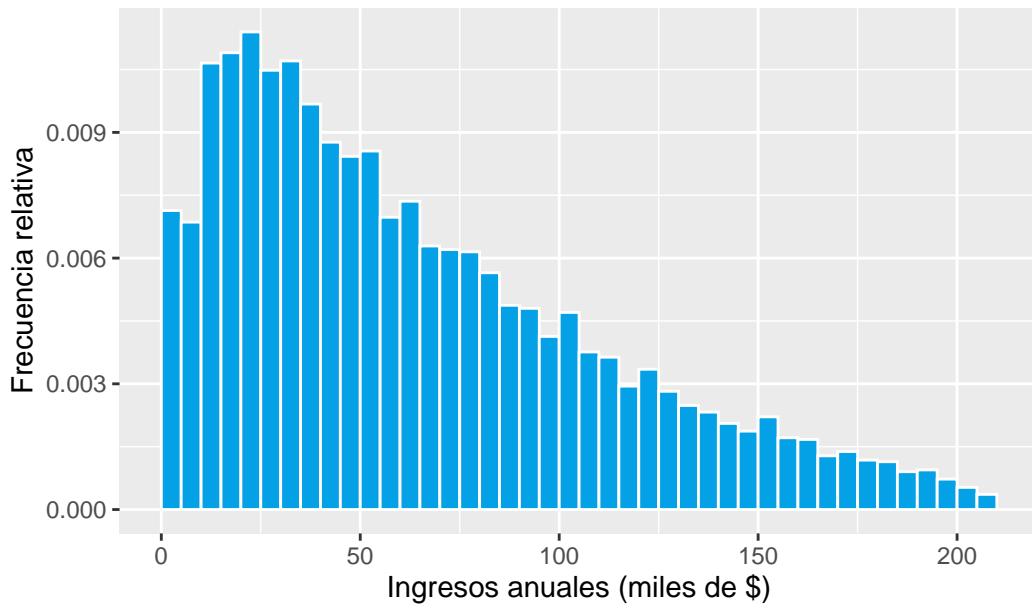
Ejemplo 2.8 (Distribución de los ingresos familiares).

```

income <- seq(2500,207500,5000)/1000
counts <- c(4235, 4071, 6324, 6470, 6765, 6222, 6354, 5743, 5203, 5002, 5078, 4140, 4367
breaks <- seq(0,210000,5000)/1000
df <- data.frame(Ingresaos = rep(income, counts))
p <- ggplot(df, aes(x=Ingresaos)) +
  geom_histogram(aes(y = ..density..), breaks = breaks, col = "white", fill = blueceu)
  xlab("Ingresos anuales (miles de $)") +
  ylab("Frecuencia relativa") +
  ggtitle("Distribución de ingresos familiares en USA")
p

```

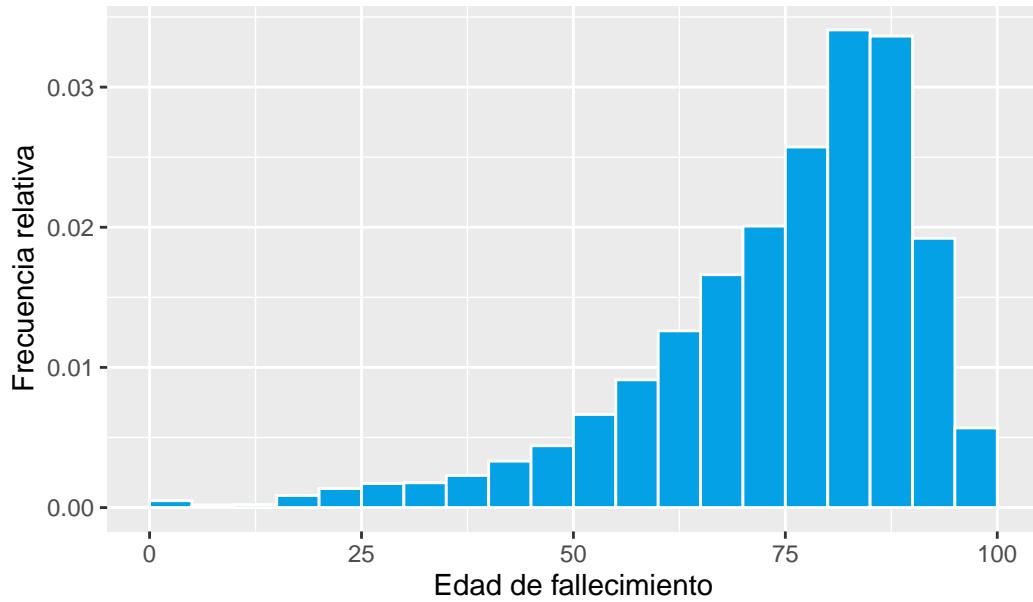
Distribución de ingresos familiares en USA



Ejemplo 2.9 (Distribución de la edad de fallecimiento).

```
counts <- c(65, 116, 69, 78, 319, 501, 633, 655, 848, 1226, 1633, 2459, 3375, 4669, 6152
breaks <- seq(0,100,5)
df <- data.frame(Edad = rep(breaks, counts))
p <- ggplot(df, aes(x=Edad)) +
  geom_histogram(aes(y = ..density..), breaks = breaks, col = "white", fill = blueceu)
  xlab("Edad de fallecimiento") +
  ylab("Frecuencia relativa") +
  ggtitle("Distribución de la edad de fallecimiento de hombres australianos.")
p
```

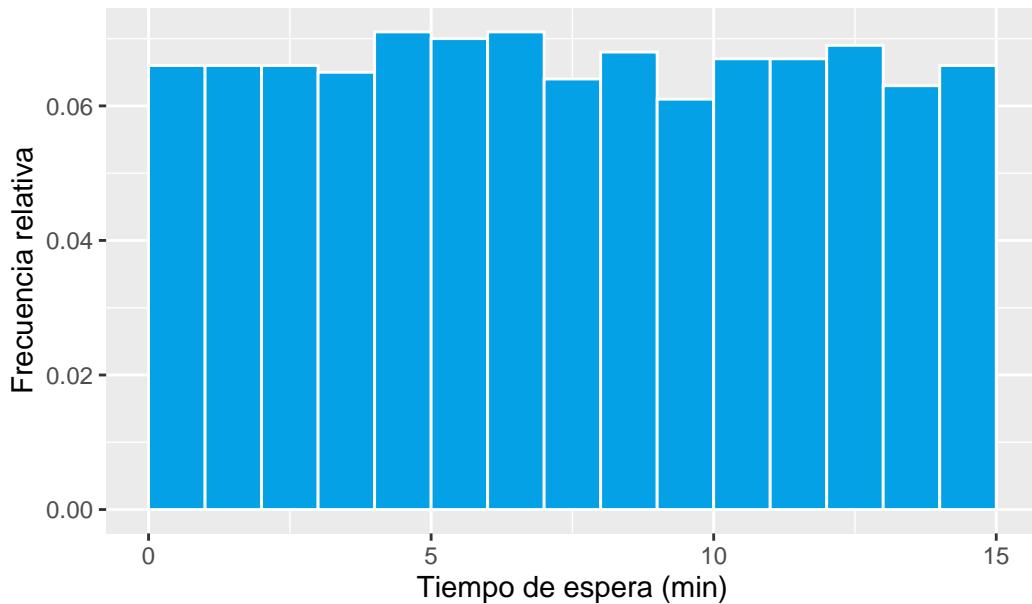
Distribución de la edad de fallecimiento de hombres australianos



Ejemplo 2.10 (Distribución del tiempo de espera del metro).

```
set.seed(123)
time <- runif(1000, min = 0, max = 15)
breaks <- seq(0, 15)
df <- data.frame(Tiempo = time)
p <- ggplot(df, aes(x=Tiempo)) +
  geom_histogram(aes(y = ..density..), breaks = breaks, col = "white", fill = blueceu)
  xlab("Tiempo de espera (min)") +
  ylab("Frecuencia relativa") +
  ggtitle("Distribución del tiempo de espera del metro.")
p
```

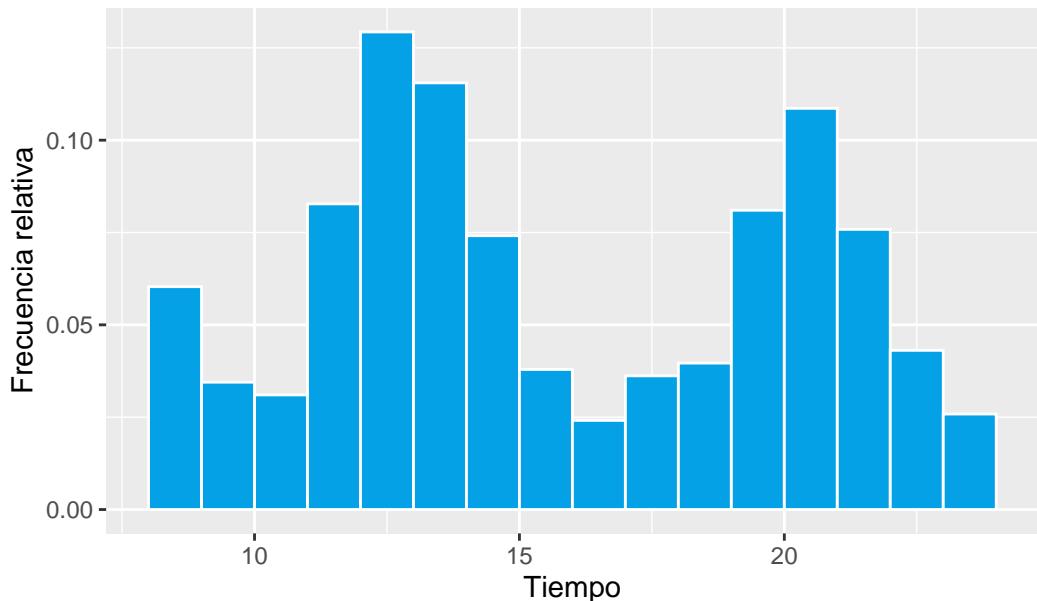
Distribución del tiempo de espera del metro.



Ejemplo 2.11 (Distribución del tiempo de llegada de clientes a un restaurante).

```
counts <- c(35, 20, 18, 48, 75, 67, 43, 22, 14, 21, 23, 47, 63, 44, 25, 15)
breaks <- seq(8.5,23.5,1)
df <- data.frame(Tiempo = rep(breaks, counts))
breaks <- seq(8,24)
p <- ggplot(df, aes(x=Tiempo)) +
  geom_histogram(aes(y = ..density..), breaks = breaks, col = "white", fill = blueceu)
  xlab("Tiempo") +
  ylab("Frecuencia relativa") +
  ggtitle("Distribución del tiempo de llegada de clientes a un restaurante")
p
```

Distribución del tiempo de llegada de clientes a un restaurante

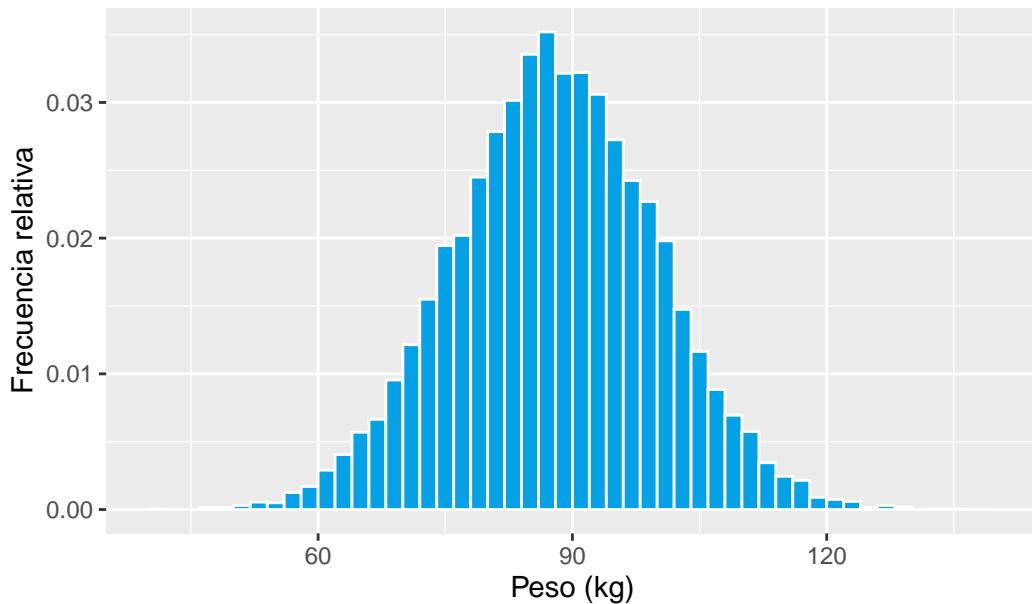


Las distribuciones con forma de campana se presentan muy a menudo en las variables biológicas.

Ejemplo 2.12 (Distribución del peso de los hombres).

```
set.seed(123)
df <- data.frame(Peso = rnorm(10000, mean = 88, sd = 12))
breaks <- seq(40, 140, 2)
p <- ggplot(df, aes(x = Peso)) +
  geom_histogram(aes(y = ..density..), breaks = breaks, col = "white", fill = blueceu)
  xlab("Peso (kg)") +
  ylab("Frecuencia relativa") +
  ggtitle("Distribución del peso de los hombres")
p
```

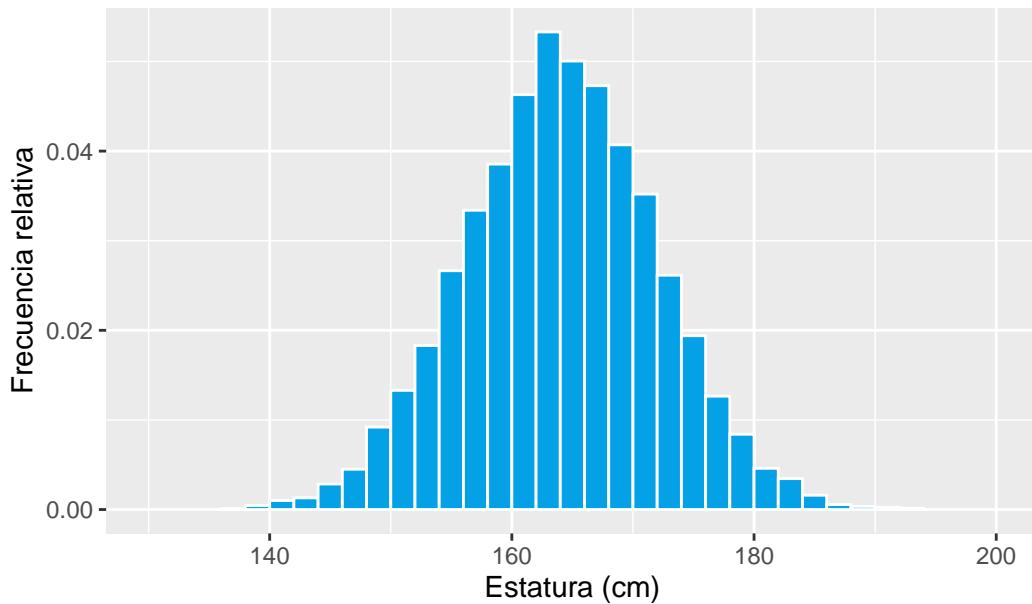
Distribución del peso de los hombres



Ejemplo 2.13 (Distribución de la estatura de las mujeres).

```
set.seed(1234)
df <- data.frame(Estatura = rnorm(10000, mean = 164, sd = 8))
breaks <- seq(130, 200, 2)
p <- ggplot(df, aes(x = Estatura)) +
  geom_histogram(aes(y = ..density..), breaks = breaks, col = "white", fill = blueceu)
  xlab("Estatura (cm)") +
  ylab("Frecuencia relativa") +
  ggtitle("Distribución de la estatura de las mujeres")
p
```

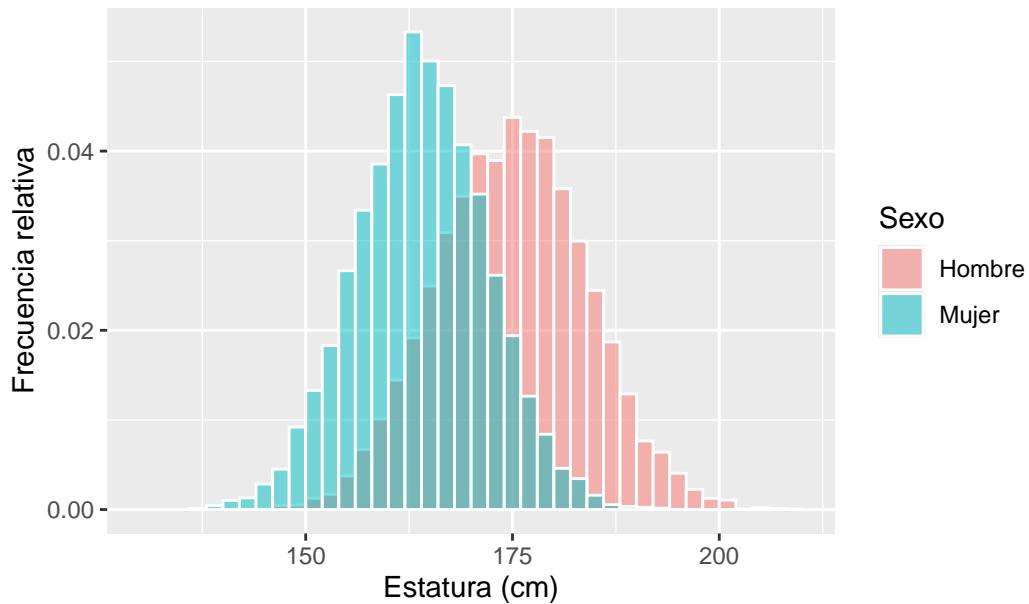
Distribución de la estatura de las mujeres



Ejemplo 2.14 (Distribución de la estatura según el sexo).

```
set.seed(1234)
n <- 10000
mujeres <- rnorm(n, mean = 164, sd = 8)
hombres <- rnorm(n, mean = 175, sd = 9)
df <- data.frame(Estatura = c(mujeres, hombres), Sexo = c(rep("Mujer",n), rep("Hombre", n)))
breaks <- seq(130, 210, 2)
p <- ggplot(df, aes(x = Estatura, fill = Sexo)) +
  geom_histogram(aes(y = ..density..), breaks = breaks, position = "identity", col = "black")
  xlab("Estatura (cm)") +
  ylab("Frecuencia relativa") +
  ggtitle("Distribución de estaturas según sexo")
p
```

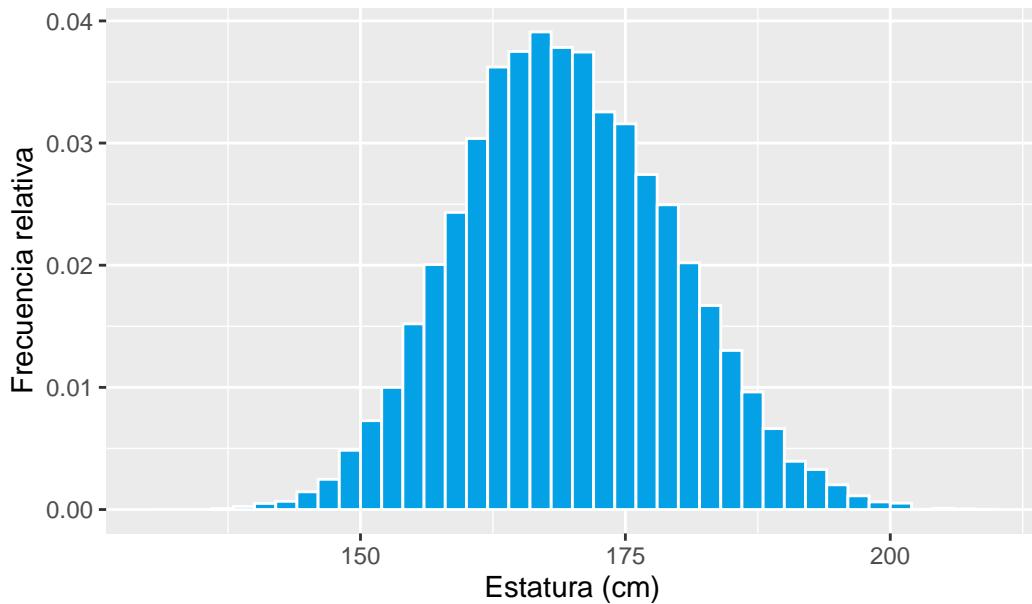
Distribución de estaturas según sexo



Ejemplo 2.15 (Distribución de la estatura de hombres y mujeres).

```
p <- ggplot(df, aes(x = Estatura)) +  
  geom_histogram(aes(y = ..density..), breaks = breaks, col = "white", fill = blueceu)  
  xlab("Estatura (cm)") +  
  ylab("Frecuencia relativa") +  
  ggtitle("Distribución de estaturas de hombres y mujeres")  
p
```

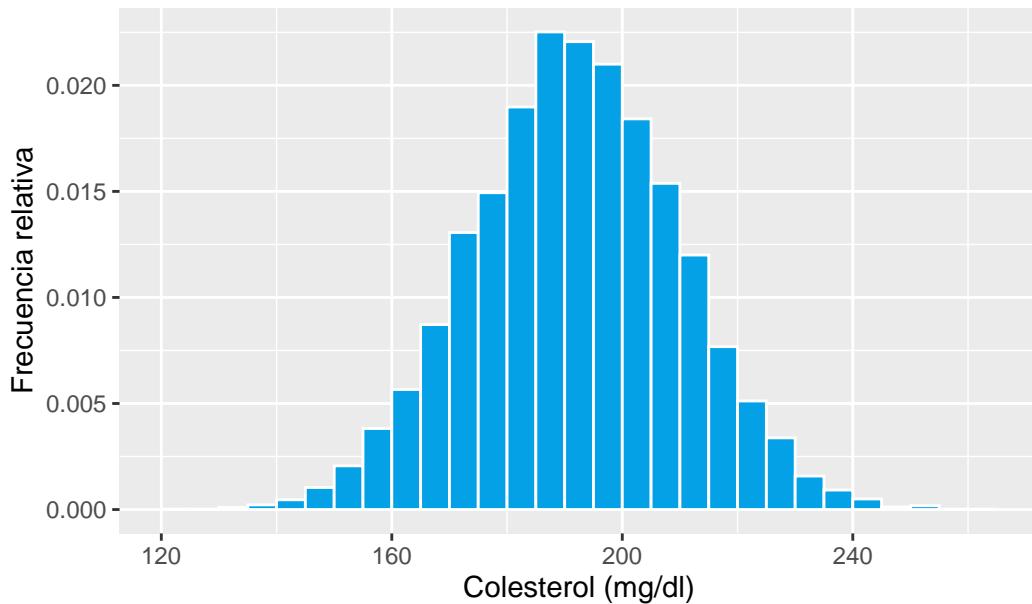
Distribución de estaturas de hombres y mujeres



Ejemplo 2.16 (Distribución del colesterol).

```
set.seed(123)
df <- data.frame(Colesterol = rnorm(10000, mean = 192, sd = 18))
breaks <- seq(120, 265, 5)
p <- ggplot(df, aes(x = Colesterol)) +
  geom_histogram(aes(y = ..density..), breaks = breaks, col = "white", fill = blueceu)
  xlab("Colesterol (mg/dl)") +
  ylab("Frecuencia relativa") +
  ggtitle("Distribución del colesterol")
p
```

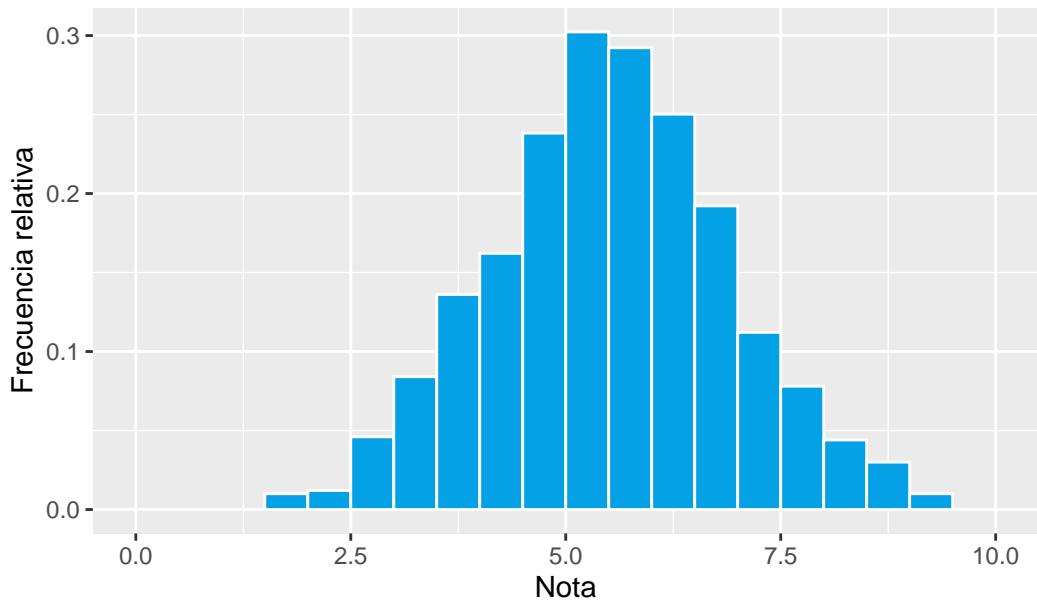
Distribución del colesterol



Ejemplo 2.17 (Distribución de notas).

```
set.seed(123)
df <- data.frame(Nota = rnorm(1000, mean = 5.5, sd = 1.4))
breaks <- seq(0, 10, 0.5)
p <- ggplot(df, aes(x = Nota)) +
  geom_histogram(aes(y = ..density..), breaks = breaks, col = "white", fill = blueceu)
  xlab("Nota") +
  ylab("Frecuencia relativa") +
  ggtitle("Distribución de notas de Estadística")
p
```

Distribución de notas de Estadística



La distribución con forma de campana aparece tan a menudo en la Naturaleza que se conoce como *distribución normal* o *distribución gaussiana*.

2.4 Datos atípicos

Uno de los principales problemas de las muestras son los **datos atípicos**, que son valores de la variable que se diferencian mucho del resto de los valores en la muestra.



Figura 2.4: Dato atípico.

Gauss bell

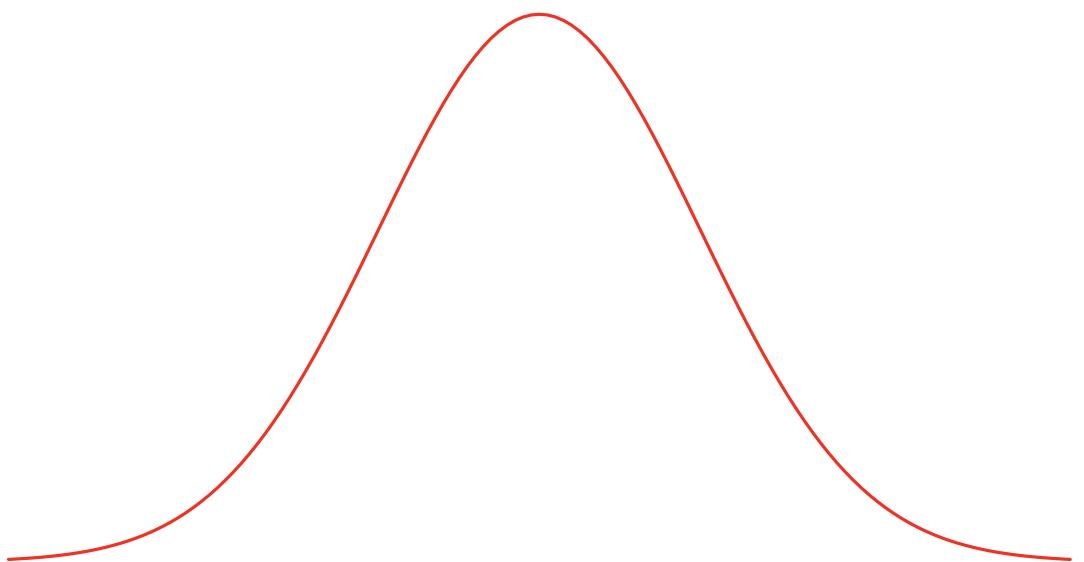


Figura 2.3: Campana de Gauss.

Es muy importante detectar los datos atípicos antes de realizar cualquier análisis de los datos, pues suelen distorsionar los resultados.

Aparecen siempre en los extremos de la distribución, y pueden detectarse con un diagrama de caja y bigotes (tal y como veremos más adelante).

2.4.1 Tratamiento de los datos atípicos

Cuando trabajemos con muestras grandes, los datos atípicos tienen menor influencia y pueden dejarse en la muestra.

Cuando trabajemos con muestras pequeñas tenemos varias opciones:

- Eliminar el dato atípico si se trata de un error.
- Sustituir el dato atípico por el menor o el mayor valor de la distribución que no es atípico si no se trata de un error y el dato atípico no concuerda con la distribución teórica.
- Dejar el dato atípico si no es un error, y cambiar el modelo de distribución teórico para adecuarlo a los datos atípicos.

2.5 Estadísticos muestrales

La tabla de frecuencias sintetiza la información de la distribución de valores de la variable estudiada en la muestra, pero en muchas ocasiones es insuficiente para describir determinados aspectos de la distribución, como por ejemplo, cuáles son los valores más representativos de la muestra, cómo es la variabilidad de los datos, qué datos pueden considerarse atípicos, o cómo es la simetría de la distribución.

Para describir esos aspectos de la distribución muestral se utilizan unas medidas resumen llamadas **estadísticos muestrales**.

De acuerdo al aspecto de las distribución que miden, existen diferentes tipos de estadísticos:

Estadísticos de Posición: Miden los valores en torno a los que se agrupan los datos o que dividen la distribución en partes iguales.

Estadísticos de Dispersión: Miden la heterogeneidad de los datos.

Estadísticos de Forma: Miden aspectos de la forma que tiene la distribución de los datos, como la simetría o el apuntamiento.

2.6 Estadísticos de posición

Pueden ser de dos tipos:

Estadísticos de Tendencia Central: Determinan valores alrededor de los cuales se concentran los datos, habitualmente en el centro de la distribución. Estas medidas suelen utilizarse como valores representativos de la muestra. Las más importantes son:

- Media aritmética
- Mediana
- Moda

Estadísticos de Posición no centrales: Dividen la distribución en partes con el mismo número de datos. Las más importantes son:

- Cuartiles.
- Deciles.
- Percentiles.

2.6.1 Media aritmética

Definición 2.2 (Media aritmética muestral \bar{x}). La *media aritmética muestral* de una variable X es la suma de los valores observados en la muestra dividida por el tamaño muestral

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

A partir de la tabla de frecuencias puede calcularse con la fórmula

$$\bar{x} = \frac{\sum x_i n_i}{n} = \sum x_i f_i$$

En la mayoría de los casos, la media aritmética es la medida que mejor representa a la muestra.



Advertencia

No puede calcularse para variables cualitativas.

Ejemplo 2.18 (Datos no agrupados). Utilizando los datos de la muestra del número de hijos en las familias, la media aritmética es

$$\bar{x} = \frac{1 + 2 + 4 + 2 + 2 + 2 + 3 + 2 + 1 + 1 + 0 + 2 + 2}{25} + \\ + \frac{0 + 2 + 2 + 1 + 2 + 2 + 3 + 1 + 2 + 2 + 1 + 2}{25} = \frac{44}{25} = 1.76 \text{ hijos.}$$

o bien, desde la tabla de frecuencias

x_i	n_i	f_i	$x_i n_i$	$x_i f_i$
0	2	0.08	0	0
1	6	0.24	6	0.24
2	14	0.56	28	1.12
3	2	0.08	6	0.24
4	1	0.04	4	0.16
\sum	25	1	44	1.76

$$\bar{x} = \frac{\sum x_i n_i}{n} = \frac{44}{25} = 1.76 \text{ hijos} \quad \bar{x} = \sum x_i f_i = 1.76 \text{ hijos.}$$

Esto significa que el valor que mejor representa el número de hijos en las familias de la muestra es 1.76 hijos.

Ejemplo 2.19 (Datos agrupados). Utilizando los datos de la muestra de estaturas, la media es

$$\bar{x} = \frac{179 + 173 + \dots + 187}{30} = 175.07 \text{ cm.}$$

o bien, desde la tabla de frecuencias utilizando las marcas de clase x_i :

X	x_i	n_i	f_i	$x_i n_i$	$x_i f_i$
(150, 160]	155	2	0.07	310	10.33
(160, 170]	165	8	0.27	1320	44.00
(170, 180]	175	11	0.36	1925	64.17
(180, 190]	185	7	0.23	1295	43.17
(190, 200]	195	2	0.07	390	13
\sum		30	1	5240	174.67

$$\bar{x} = \frac{\sum x_i n_i}{n} = \frac{5240}{30} = 174.67 \text{ cm} \quad \bar{x} = \sum x_i f_i = 174.67 \text{ cm.}$$

Obsérvese que al calcular la media desde la tabla de frecuencias el resultado difiere ligeramente del valor real obtenido directamente desde la muestra, ya que los valores usados en los cálculos no son los datos reales sino las marcas de clase.

2.6.1.1 Media ponderada

En algunos casos, los valores de la muestra no tienen la misma importancia. En este caso la importancia o *peso* de cada valor de la muestra debe tenerse en cuenta al calcular la media.

Definición 2.3 (Media ponderada muestral \bar{x}_p). Dada una muestra de valores x_1, \dots, x_n donde cada valor x_i tiene asociado un peso p_i , la *media ponderada muestral* de la variable X es la suma de los productos de cada valor observado en la muestra por su peso, dividida por la suma de todos los pesos

$$\bar{x}_p = \frac{\sum x_i p_i}{\sum p_i}$$

A partir de la tabla de frecuencias puede calcularse con la fórmula

$$\bar{x}_p = \frac{\sum x_i p_i n_i}{\sum p_i}$$

Ejemplo 2.20. Supóngase que un estudiante quiere calcular una medida que represente su rendimiento en el curso. La nota obtenida en cada asignatura y sus créditos son

Asignatura	Créditos	Nota
Matemáticas	6	5
Economía	4	3
Química	8	6

La media aritmética vale

$$\bar{x} = \frac{\sum x_i}{n} = \frac{5 + 3 + 6}{3} = 4.67 \text{ puntos.}$$

Sin embargo, esta nota no representa bien el rendimiento académico del alumno ya que no todas las asignaturas tienen la misma importancia ni requieren el mismo esfuerzo para aprobar. Las asignaturas con más créditos requieren más trabajo y deben tener más peso en el cálculo de la media.

Es más lógico usar la media ponderada como medida del rendimiento del estudiante, tomando como pesos los créditos de cada asignatura

$$\bar{x}_p = \frac{\sum x_i p_i}{\sum p_i} = \frac{5 \cdot 6 + 3 \cdot 4 + 6 \cdot 8}{6 + 4 + 8} = \frac{90}{18} = 5 \text{ puntos.}$$

2.6.2 Mediana

Definición 2.4 (Mediana muestral Me). La *mediana muestral* de una variable X es el valor de la variable que está en el medio de la muestra ordenada.

La mediana divide la distribución de la muestra en dos partes iguales, es decir, hay el mismo número de valores por debajo y por encima de la mediana. Por tanto, tiene frecuencias acumuladas $N_{Me} = n/2$ y $F_{Me} = 0.5$.

 Advertencia

No puede calcularse para variables nominales.

Con datos no agrupados pueden darse varios casos:

- Tamaño muestral impar: La mediana es el valor que ocupa la posición $\frac{n+1}{2}$.
- Tamaño muestral par: La mediana es la media de los valores que ocupan las posiciones $\frac{n}{2}$ y $\frac{n}{2} + 1$.

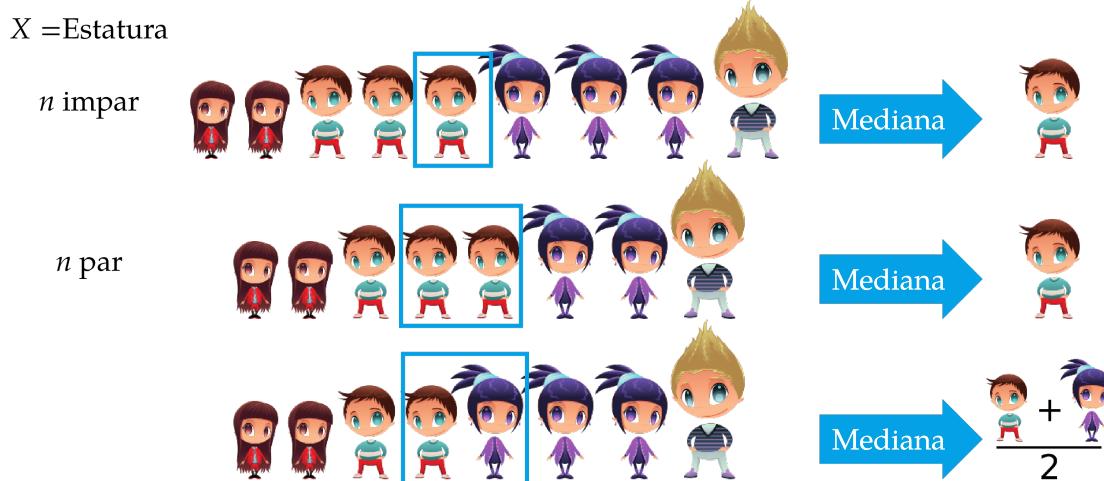


Figura 2.5: Cálculo de la mediana con datos no agrupados.

:::{#exm-mediana-datos-no-agrupados} Utilizando los datos del número de hijos de las familias, el tamaño muestral es 25, que es impar, y la mediana es el valor que ocupa la posición $\frac{25+1}{2} = 13$ de la muestra ordenada.

0, 0, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, [2], 2, 2, 2, 2, 2, 2, 2, 3, 3, 4

y la mediana es 2 hijos.

Si se trabaja con la tabla de frecuencias, la mediana es el valor más pequeño con una frecuencia acumulada mayor o igual a 13, o con una frecuencia relativa acumulada mayor o igual que 0.5.

x_i	n_i	f_i	N_i	F_i
0	2	0.08	2	0.08
1	6	0.24	8	0.32
2	14	0.56	22	0.88
3	2	0.08	24	0.96
4	1	0.04	25	1
\sum	25	1		

2.6.2.1 Cálculo de la mediana con datos agrupados

Con datos agrupados la mediana se calcula interpolando en el polígono de frecuencias relativas acumuladas para el valor 0.5.

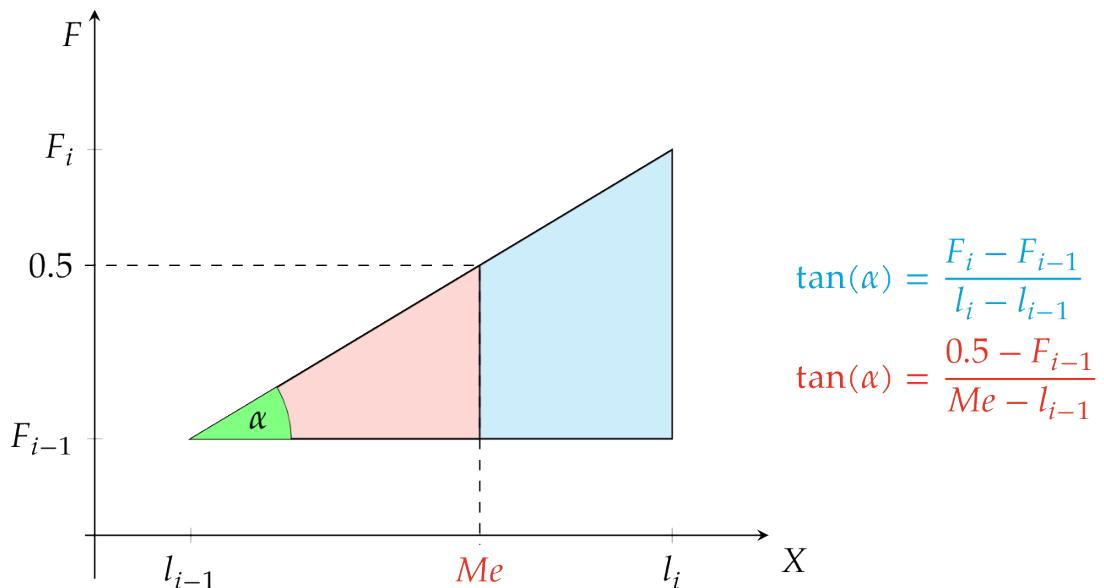


Figura 2.6: Cálculo de la mediana con datos agrupados.

Ambas expresiones son iguales ya que el ángulo α es el mismo, y resolviendo la ecuación se tiene la siguiente fórmula para calcular la mediana

$$Me = l_{i-1} + \frac{0.5 - F_{i-1}}{F_i - F_{i-1}}(l_i - l_{i-1}) = l_{i-1} + \frac{0.5 - F_{i-1}}{f_i}a_i$$

Ejemplo 2.21 (Datos agrupados). Utilizando los datos de la muestra de las estaturas de estudiantes, la mediana cae en la clase (170,180].

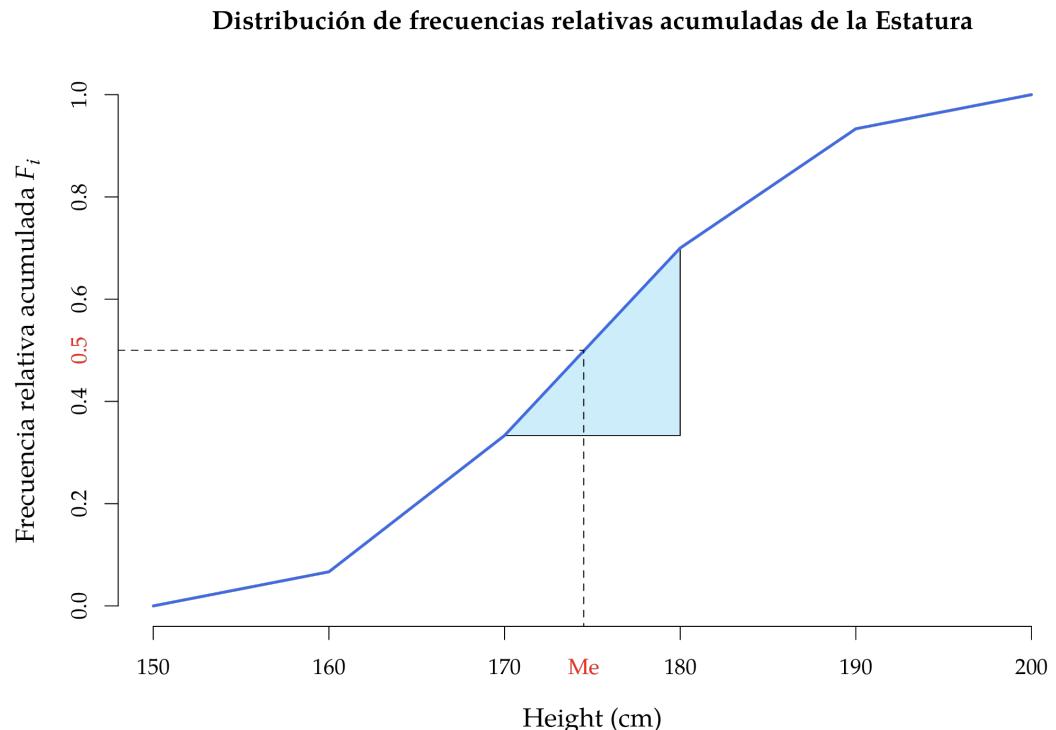


Figura 2.7: Ejemplo de cálculo de la mediana con datos agrupados.

Interpolando en el intervalo (170,180] se tiene

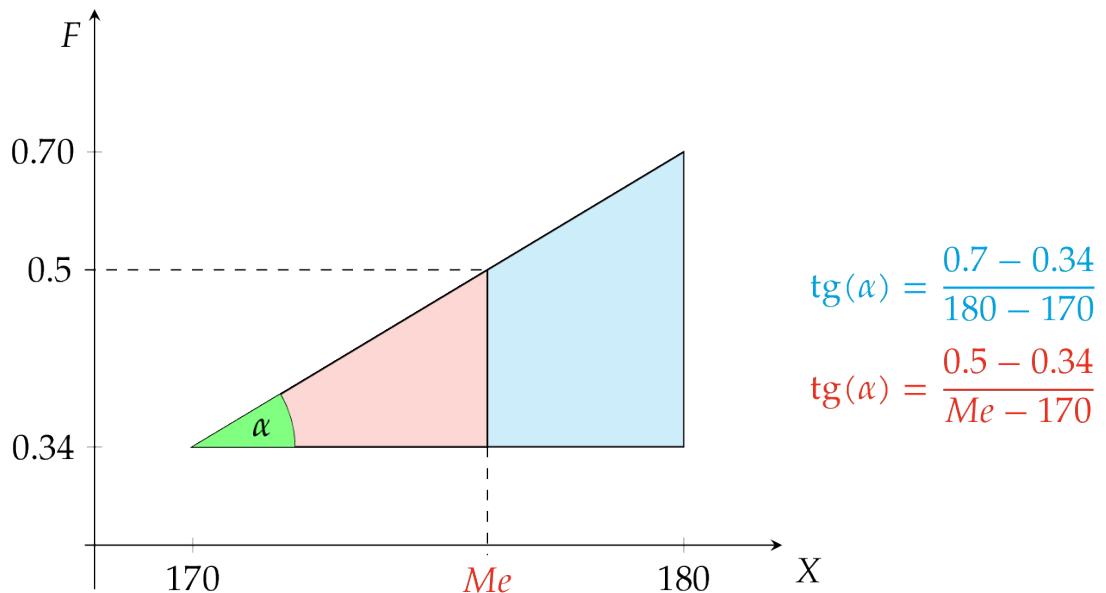


Figura 2.8: Ejemplo de cálculo de la mediana con datos agrupados.

Igualando ambas expresiones y resolviendo la ecuación se obtiene

$$Me = 170 + \frac{0.5 - 0.34}{0.7 - 0.34} (180 - 170) = 170 + \frac{0.16}{0.36} 10 = 174.54 \text{ cm.}$$

Esto significa que la mitad de los estudiantes tienen estaturas menores o iguales que 174.54 cm y la otra mitad mayores o iguales.

2.6.3 Moda

Definición 2.5 (Moda muestral Mo). La *moda muestral* de una variable X es el valor de la variable más frecuente en la muestra.

Con datos agrupados la *clase modal* es la clase con mayor frecuencia en la muestra.

Puede calcularse para todos los tipos de variables (cuantitativas y cualitativas).

Las distribuciones pueden tener más de una moda.



Figura 2.9: Cálculo de la moda.

Ejemplo 2.22. Utilizando los datos de la muestra del número de hijos en las familias, el valor con mayor frecuencia es 2, y por tanto la moda es $Mo = 2$.

x_i	n_i
0	2
1	6
2	14
3	2
4	1

Ejemplo 2.23. Utilizando los datos de la muestra de estaturas de estudiantes, la clase con la mayor frecuencia es $(170, 180]$, que es la clase modal $Mo = (170, 180]$.

X	n_i
$(150, 160]$	2
$(160, 170]$	8
$(170, 180]$	11
$(180, 190]$	7
$(190, 200]$	2

2.6.4 ¿Qué estadístico de tendencia central usar?

En general, siempre que puedan calcularse los estadísticos de tendencia central, es recomendable utilizarlos como valores representativos en el siguiente orden:

1. Media. La media utiliza más información que el resto ya que para calcularla se tiene en cuenta la magnitud de los datos.

2. Mediana. La mediana utiliza menos información que la media, pero más que la moda, ya que para calcularla se tiene en cuenta el orden de los datos.
3. Moda. La moda es la que menos información utiliza ya que para calcularla sólo se tienen en cuenta las frecuencias absolutas.

 **Advertencia**

Hay que tener cuidado con los datos atípicos, ya que la media puede distorsionarse cuando hay datos atípicos. En tal caso es mejor utilizar la mediana como valor más representativo.

Ejemplo 2.24. Si una muestra de número de hijos de 7 familias es

0, 0, 1, 1, 2, 2, 15,

entonces, $\bar{x} = 3$ hijos y $Me = 1$ hijo.

¿Qué medida representa mejor el número de hijos en la muestra?

2.6.5 Medidas de posición no centrales

Las medidas de posición no centrales o *cuantiles* dividen la distribución en partes iguales.

Los más utilizados son:

Cuartiles: Dividen la distribución en 4 partes iguales. Hay 3 cuartiles: C_1 (25% acumulado), C_2 (50% acumulado), C_3 (75% acumulado).

Deciles: Dividen la distribución en 10 partes iguales. Hay 9 deciles: D_1 (10% acumulado), ..., D_9 (90% acumulado).

Percentiles: Dividen la distribución en 100 partes iguales. Hay 99 percentiles: P_1 (1% acumulado), ..., P_{99} (99% acumulado).

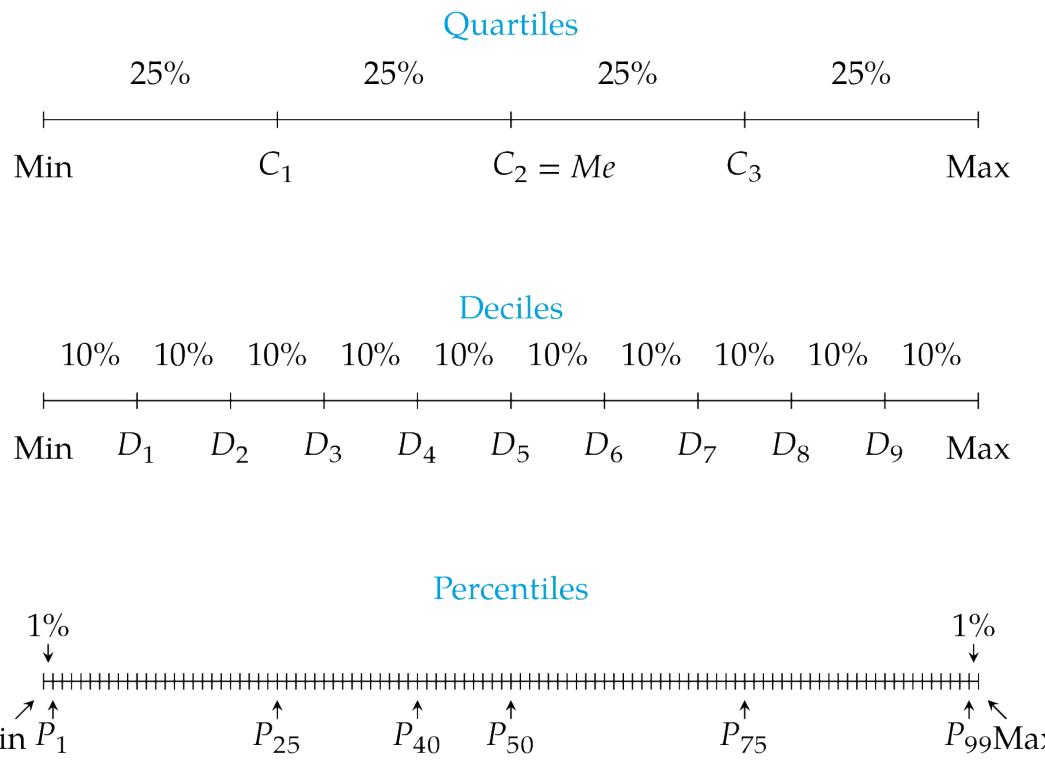


Figura 2.10: Cuartiles, deciles y percentiles.

Obsérvese que hay una correspondencia entre los cuartiles, los deciles y los percentiles. Por ejemplo, el primer cuartil coincide con el percentil 25, y el cuarto decil coincide con el percentil 40.

Los cuantiles se calculan de forma similar a la mediana. La única diferencia es la frecuencia relativa acumulada que corresponde a cada cuantil.

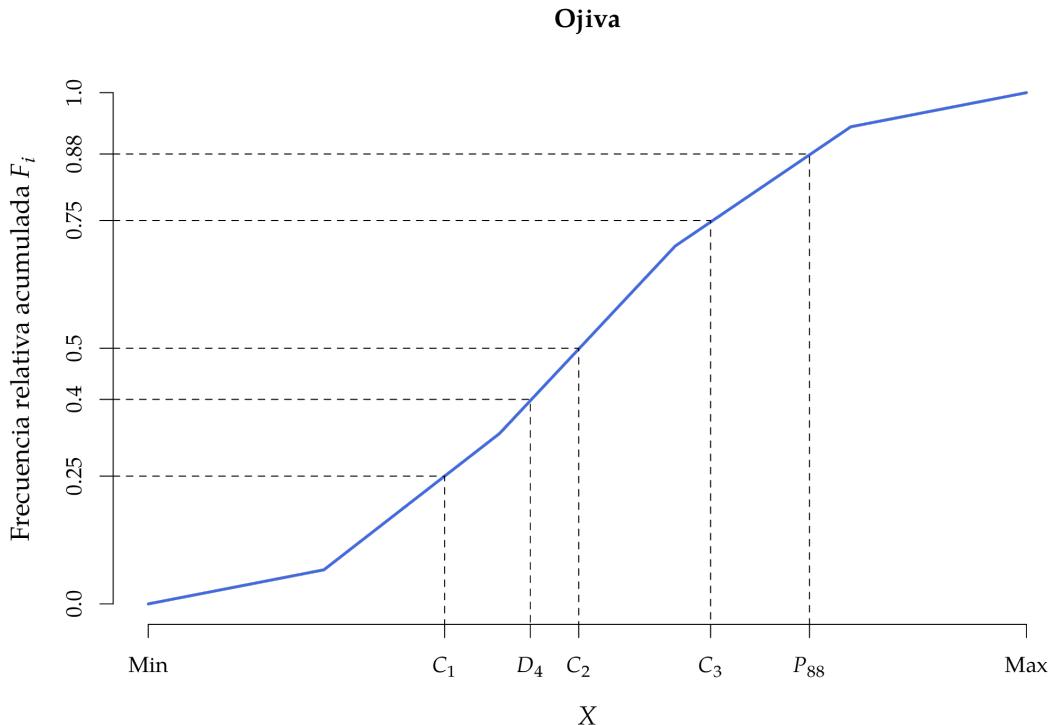


Figura 2.11: Cálculo de cuartiles, deciles y percentiles.

Ejemplo 2.25. Utilizando los datos de la muestra del número de hijos de las familias, la frecuencia relativa acumulada era

x_i	F_i
0	0.08
1	0.32
2	0.88
3	0.96
4	1

$$\begin{aligned}
 F_{C_1} &= 0.25 \Rightarrow Q_1 = 1 \text{ hijos}, \\
 F_{C_2} &= 0.5 \Rightarrow Q_2 = 2 \text{ hijos}, \\
 F_{C_3} &= 0.75 \Rightarrow Q_3 = 2 \text{ hijos}, \\
 F_{D_4} &= 0.4 \Rightarrow D_4 = 2 \text{ hijos}, \\
 F_{P_{92}} &= 0.92 \Rightarrow P_{92} = 3 \text{ hijos}.
 \end{aligned}$$

2.7 Estadísticos de dispersión

La *dispersión* se refiere a la heterogeneidad o variabilidad de los datos. Así pues, los estadísticos de dispersión mide la variabilidad global de los datos, o con respecto a una medida de tendencia central.

Para las variables cuantitativas, las más empleadas son:

- Recorrido.
- Rango Intercuartílico.
- Varianza.
- Desviación Típica.
- Coeficiente de Variación.

2.7.1 Recorrido

Definición 2.6 (Recorrido muestral Re). El *recorrido muestral* o *rango muestral* de una variable X se define como la diferencia entre el máximo y el mínimo de los valores en la muestra.

$$Re = \max_{x_i} - \min_{x_i}$$

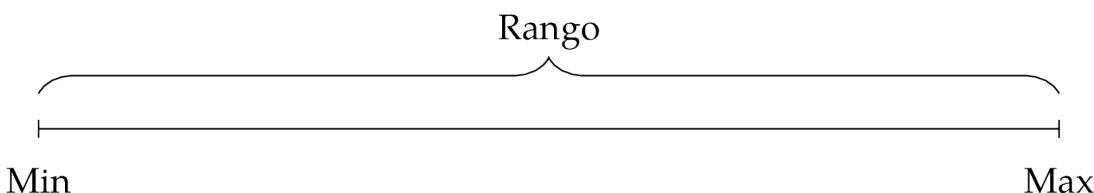


Figura 2.12: Rango muestral.

El recorrido mide la máxima variación que hay entre los datos muestrales. No obstante, es muy sensible a datos atípicos ya que suelen aparecer justo en los extremos de la distribución, por lo que no se suele utilizar mucho.

2.7.2 Rango intercuartílico

Para evitar el problema de los datos atípicos en el recorrido, se puede utilizar el primer y tercer cuartil en lugar del mínimo y el máximo.

Definición 2.7 (Rango intercuartílico muestral RI). El *rango intercuartílico muestral* de una variable X se define como la diferencia entre el tercer y el primer cuartil de la muestra.

$$RI = C_3 - C_1$$

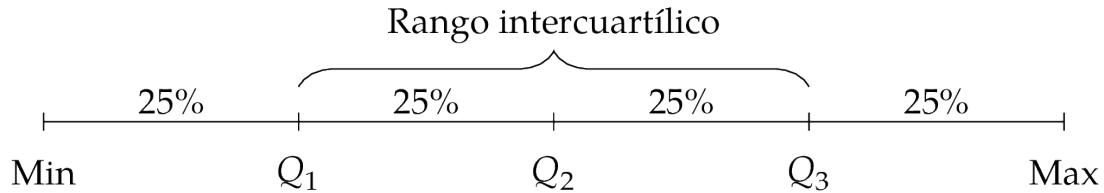


Figura 2.13: Rango intercuartílico.

El rango intercuartílico mide la dispersión del 50% de los datos centrales.

2.7.3 Diagrama de caja y bigotes

La dispersión de una variable suele representarse gráficamente mediante un *diagrama de caja y bigotes*, que representa cinco estadísticos descriptivos (mínimo, cuartiles y máximo) conocidos como los *cinco números*. Consiste en una caja, dibujada desde el primer al tercer cuartil, que representa el rango intercuartílico, y dos segmentos, conocidos como *bigotes* inferior y superior. A menudo la caja se divide en dos por la mediana.

Este diagrama es muy útil y se utiliza para muchos propósitos:

- Sirve para medir la dispersión de los datos ya que representa el rango y el rango intercuartílico.
- Sirve para detectar datos atípicos, que son los valores que quedan fuera del intervalo definido por los bigotes.
- Sirve para medir la simetría de la distribución, comparando la longitud de las cajas y de los bigotes por encima y por debajo de la mediana.

:::{#exm-diagrama-caja} El diagrama siguiente muestra el diagrama de caja y bigotes del peso de una muestra de recién nacidos.

Diagrama de caja y bigotes del peso de recien nacidos

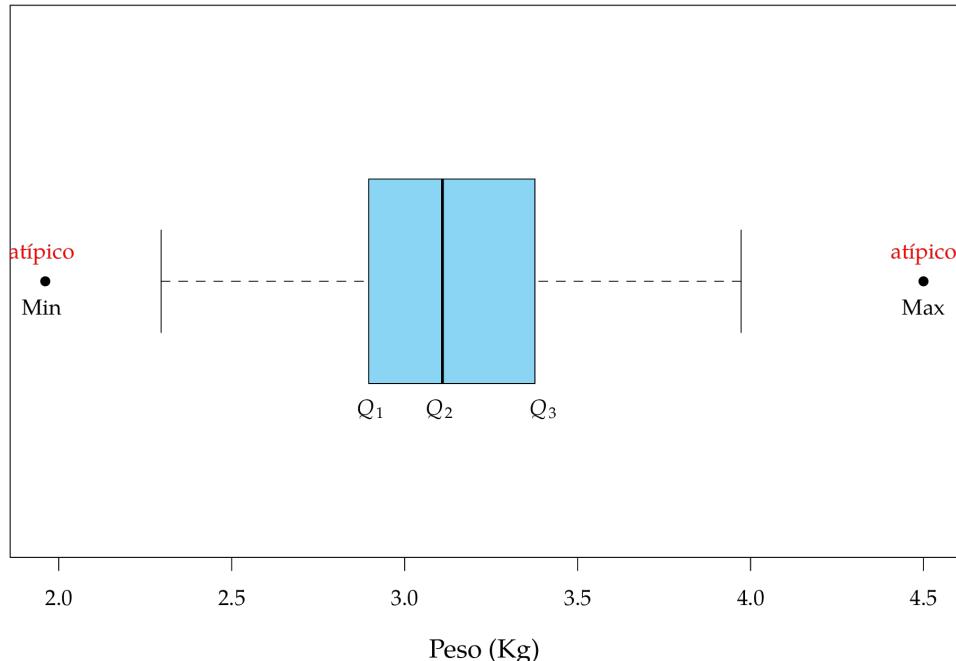


Figura 2.14: Diagrama de caja y bigotes del peso de recién nacidos.

Para construir el diagrama de caja y bigotes hay que seguir los siguientes pasos:

1. Calcular los cuartiles.
2. Dibujar una caja de manera que el extremo inferior caiga sobre el primer cuartil y el extremo superior sobre el tercer cuartil.
3. Dividir la caja con una línea que caiga sobre el segundo cuartil.
4. Para los bigotes inicialmente se calculan dos valores llamados *vallas* v_1 y v_2 . La valla inferior es el primer cuartil menos una vez y media el rango intercuartílico, y la valla superior es el tercer cuartil más una vez y media el rango intercuartílico.

$$v_1 = Q_1 - 1.5 \text{ IQR}$$

$$v_2 = Q_3 + 1.5 \text{ IQR}$$

Las vallas definen el intervalo donde los datos se consideran normales. Cualquier valor fuera de ese intervalo se considera un dato atípico.

El bigote superior se dibuja desde el borde inferior de la caja hasta el menor valor de la muestra que es mayor o igual a la valla inferior, y el bigote superior se dibuja

desde el borde superior de la caja hasta el mayor valor de la muestra que es menor o igual a la valla superior.

⚠️ Advertencia

Los bigotes no son las vallas.

5. Finalmente, si en la muestra hay algún dato atípico, se dibuja un punto para cada uno de ellos.

Ejemplo 2.26. El diagrama de caja y bigotes de la muestra del número de hijos de las familias se muestra a continuación.

Diagrama de caja y bigotes del número de hijos

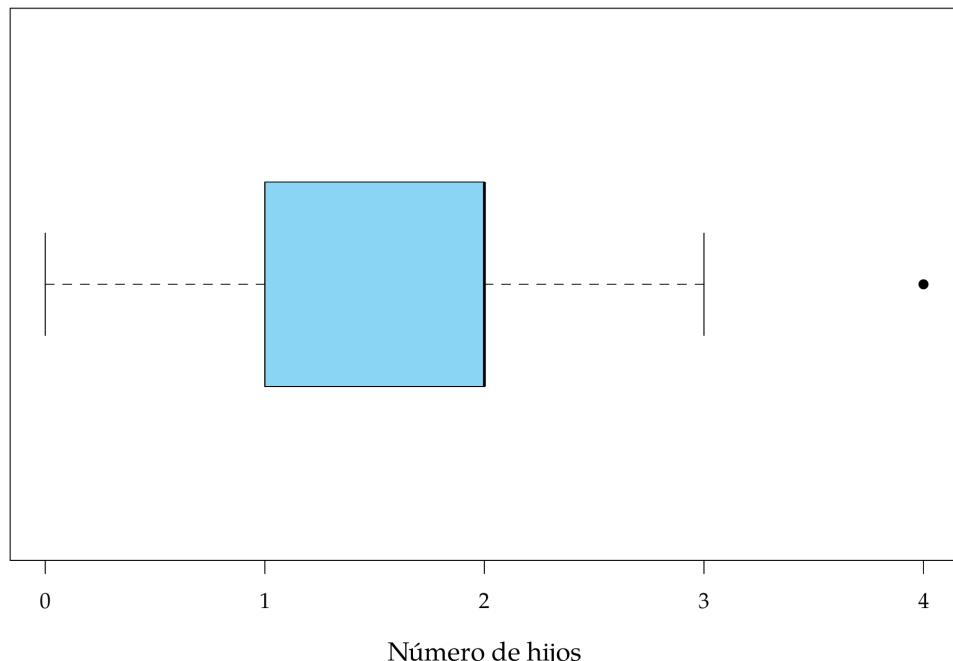


Figura 2.15: Diagrama de caja y bigotes del número de hijos.

2.7.3.1 Desviaciones respecto de la media

Otra forma de medir la variabilidad de una variable es estudiar la concentración de los valores en torno a algún estadístico de tendencia central como por ejemplo la media.

Para ello se suele medir la distancia de cada valor a la media. A ese valor se le llama **desviación de la media**.

desviación - desviación +

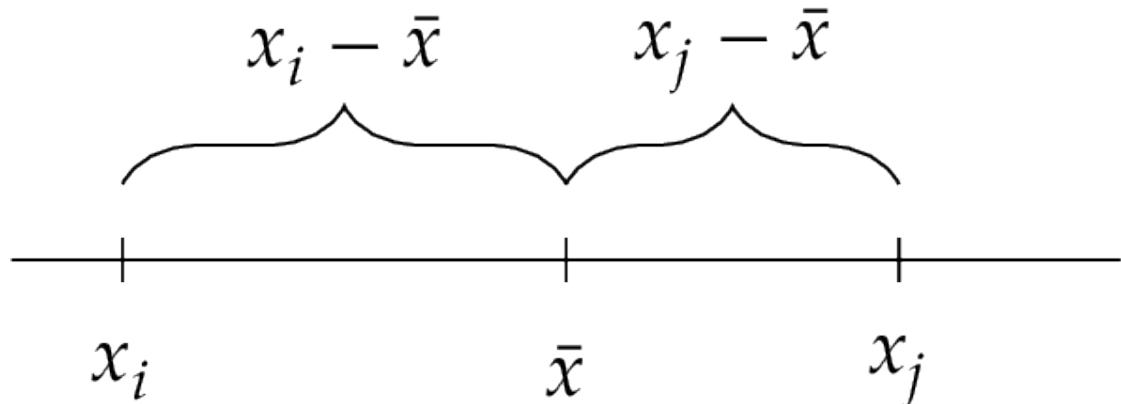


Figura 2.16: Desviaciones con respecto a la media.

Si las desviaciones son grandes la media no será tan representativa como cuando la desviaciones sean pequeñas.

Ejemplo 2.27. La siguiente tabla contiene las notas de 3 estudiantes en un curso con las asignaturas A , B y C .

A	B	C	\bar{x}
0	5	10	5
4	5	6	5
5	5	5	5

Todos los estudiantes tienen la misma media, pero, en qué caso la media representa mejor el rendimiento en el curso?

2.7.4 Varianza y desviación típica

Definición 2.8 (Varianza s^2). La *varianza muestral* de una variable X se define como el promedio del cuadrado de las desviaciones de los valores de la muestra respecto de la media muestral.

$$s^2 = \frac{\sum(x_i - \bar{x})^2 n_i}{n} = \sum(x_i - \bar{x})^2 f_i$$

También puede calcularse de manera más sencilla mediante la fórmula

$$s^2 = \frac{\sum x_i^2 n_i}{n} - \bar{x}^2 = \sum x_i^2 f_i - \bar{x}^2$$

La varianza tiene las unidades de la variable al cuadrado, por lo que para facilitar su interpretación se suele utilizar su raíz cuadrada.

Definición 2.9 (Desviación típica s). La *desviación típica muestral* de una variable X se define como la raíz cuadrada positiva de su varianza muestral.

$$s = +\sqrt{s^2}$$

 Tip

Tanto la varianza como la desviación típica sirven para cuantificar la dispersión de los datos en torno a la media. Cuando la varianza o la desviación típica son pequeñas, los datos de la muestra están concentrados en torno a la media, y la media es una buena medida de representatividad. Por contra, cuando la varianza o la desviación típica son grandes, los datos de la muestra están alejados de la media, y la media ya no representa tan bien.

Desviación típica pequeña	\Rightarrow	Media representativa
Desviación típica grande	\Rightarrow	Media no representativa

Ejemplo 2.28. Las siguientes muestras contienen las notas de dos estudiantes en dos asignaturas.

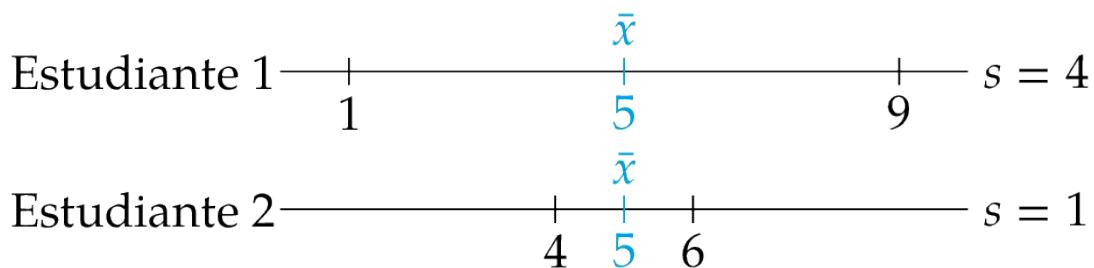


Figura 2.17: Interpretación de la desviación típica.

¿Qué media es más representativa?

Ejemplo 2.29 (Datos no agrupados). Utilizando los datos de la muestra del número de hijos de las familias, con una media $\bar{x} = 1.76$ hijos, y añadiendo una nueva columna a la tabla de frecuencias con los cuadrados de los valores,

x_i	n_i	$x_i^2 n_i$
0	2	0
1	6	6
2	14	56
3	2	18
4	1	16
\sum	25	96

$$s^2 = \frac{\sum x_i^2 n_i}{n} - \bar{x}^2 = \frac{96}{25} - 1.76^2 = 0.7424 \text{ hijos}^2.$$

y la desviación típica es $s = \sqrt{0.7424} = 0.8616$ hijos.

Comparado este valor con el recorrido, que va de 0 a 4 hijos se observa que no es demasiado grande por lo que se puede concluir que no hay mucha dispersión y en consecuencia la media de 1.76 hijos representa bien el número de hijos de las familias de la muestra.

Ejemplo 2.30 (Datos agrupados). Utilizando los datos de la muestra de estaturas de los estudiantes y agrupando las estaturas en clases, se obtenía una media $\bar{x} = 174.67$ cm. El cálculo de la varianza se realiza igual que antes pero tomando como valores de la variable las marcas de clase.

X	x_i	n_i	$x_i^2 n_i$
(150, 160]	155	2	48050
(160, 170]	165	8	217800
(170, 180]	175	11	336875
(180, 190]	185	7	239575
(190, 200]	195	2	76050
\sum		30	918350

$$s^2 = \frac{\sum x_i^2 n_i}{n} - \bar{x}^2 = \frac{918350}{30} - 174.67^2 = 102.06 \text{ cm}^2,$$

y la desviación típica es $s = \sqrt{102.06} = 10.1$ cm.

Este valor es bastante pequeño, comparado con el recorrido de la variable, que va de 150 a 200 cm, por lo que la variable tiene poca dispersión y en consecuencia su media es muy representativa.

2.7.5 Coeficiente de variación

Tanto la varianza como la desviación típica tienen unidades y eso dificulta a veces su interpretación, especialmente cuando se compara la dispersión de variables con diferentes unidades.

Por este motivo, es también común utilizar la siguiente medida de dispersión que no tiene unidades.

Definición 2.10 (Coeficiente de variación muestral cv). El *coeficiente de variación muestral* de una variable X se define como el cociente entre su desviación típica muestral y el valor absoluto de su media muestral.

$$cv = \frac{s}{|\bar{x}|}$$

💡 Tip

El coeficiente de variación muestral mide la dispersión relativa de los valores de la muestra en torno a la media muestral.

Como no tiene unidades, es muy sencillo de interpretar: Cuanto mayor sea, mayor será la dispersión relativa con respecto a la media y menos representativa será la media.

El coeficiente de variación es muy útil para comparar la dispersión de distribuciones de variables diferentes, incluso si las variables tienen unidades diferentes.

Ejemplo 2.31. En la muestra del número de hijos, donde la media era $\bar{x} = 1.76$ hijos y la desviación típica $s = 0.8616$ hijos, el coeficiente de variación vale

$$cv = \frac{s}{|\bar{x}|} = \frac{0.8616}{|1.76|} = 0.49.$$

En la muestra de las estaturas, donde la media era $\bar{x} = 174.67$ cm y la desviación típica $s = 10.1$ cm, el coeficiente de variación vale

$$cv = \frac{s}{|\bar{x}|} = \frac{10.1}{|174.67|} = 0.06.$$

Esto significa que la dispersión relativa en la muestra de estaturas es mucho menor que en la del número de hijos, por lo que la media de las estaturas será más representativa que la media del número de hijos.

2.8 Estadísticos de forma

Son medidas que describen la forma de la distribución.

Los aspectos más relevantes son:

Simetría Mide la simetría de la distribución de frecuencias en torno a la media. El estadístico más utilizado es el *Coeficiente de Asimetría de Fisher*.

Apuntamiento Mide el apuntamiento o el grado de concentración de valores en torno a la media de la distribución de frecuencias. El estadístico más utilizado es el *Coeficiente de Apuntamiento o Curtosis*.

2.8.1 Coeficiente de asimetría

Definición 2.11 (Coeficiente de asimetría muestral g_1). El *coeficiente de asimetría muestral* de una variable X es el promedio de las desviaciones de los valores de la muestra respecto de la media muestral, elevadas al cubo, dividido por la desviación típica al cubo.

$$g_1 = \frac{\sum(x_i - \bar{x})^3 n_i / n}{s^3} = \frac{\sum(x_i - \bar{x})^3 f_i}{s^3}$$

💡 Tip

Mide el grado de simetría de los valores de la muestra con respecto a la media muestra, es decir, cuantos valores de la muestra están por encima o por debajo de la media y cómo de alejados de esta.

- $g_1 = 0$ indica que hay el mismo número de valores por encima y por debajo de la media e igualmente alejados de ella (simétrica).

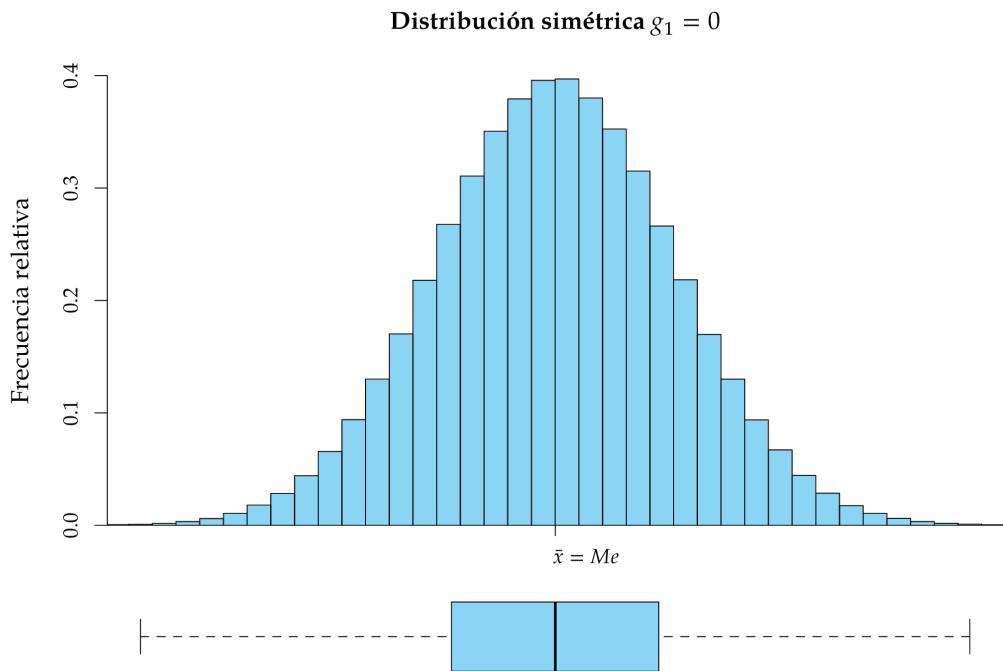


Figura 2.18: Distribución simétrica.

- $g_1 < 0$ indica que la mayoría de los valores son mayores que la media, pero los valores menores están más alejados de ella (asimétrica a la izquierda).

Distribución asimétrica a la izquierda $g_1 < 0$

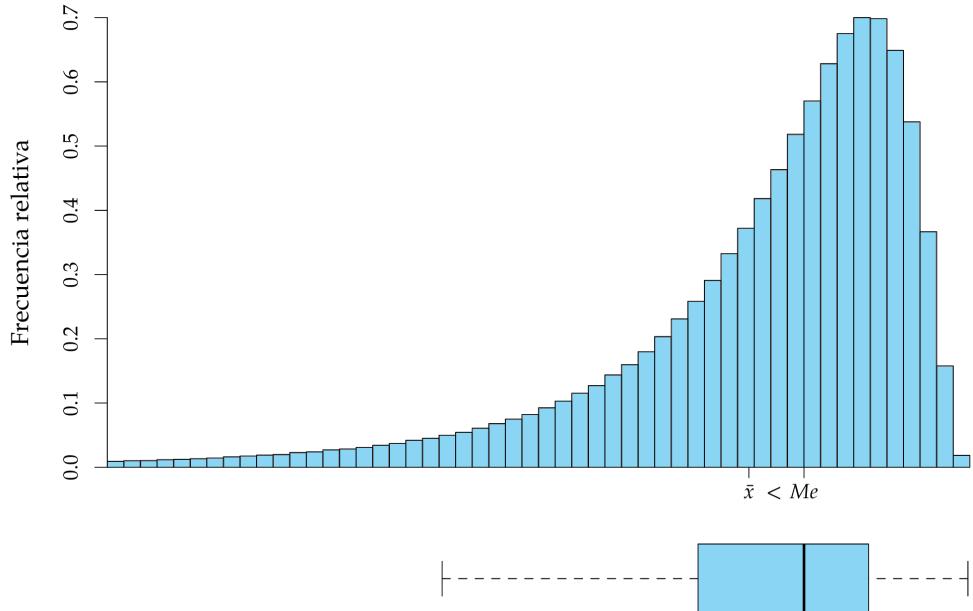


Figura 2.19: Distribución asimétrica hacia la izquierda.

- $g_1 > 0$ indica que la mayoría de los valores son menores que la media, pero los valores mayores están más alejados de ella (asimétrica a la derecha).

Distribución asimétrica a la derecha $g_1 > 0$

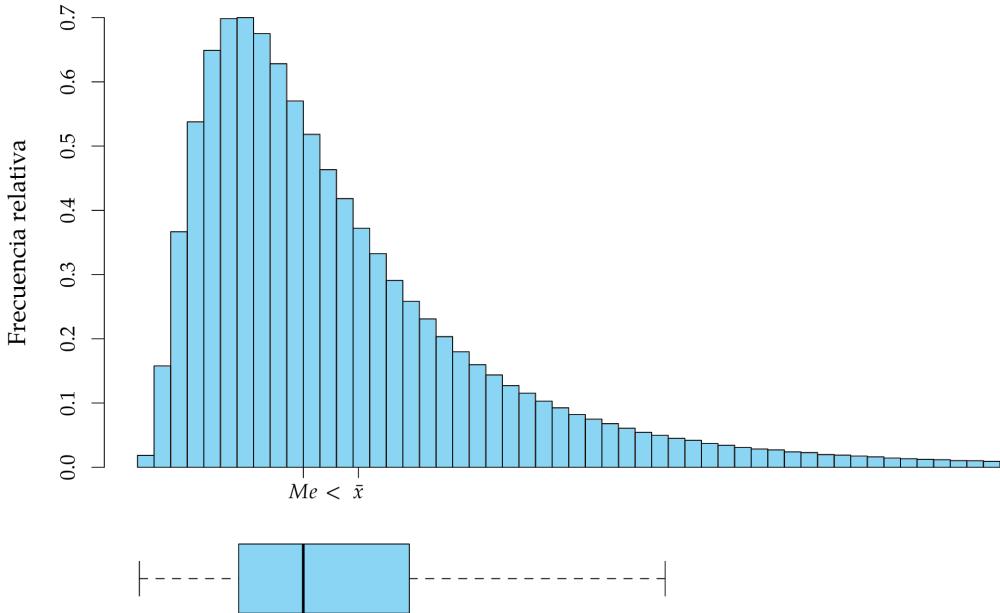


Figura 2.20: Distribución asimétrica hacia la derecha.

Ejemplo 2.32 (Datos agrupados). Utilizando la tabla de frecuencias de la muestra de estaturas y añadiendo una nueva columna con las desviaciones de la media $\bar{x} = 174.67$ cm al cubo, se tiene

X	x_i	n_i	$x_i - \bar{x}$	$(x_i - \bar{x})^3 n_i$
(150, 160]	155	2	-19.67	-15221.00
(160, 170]	165	8	-9.67	-7233.85
(170, 180]	175	11	0.33	0.40
(180, 190]	185	7	10.33	7716.12
(190, 200]	195	2	20.33	16805.14
\sum		30		2066.81

$$g_1 = \frac{\sum (x_i - \bar{x})^3 n_i / n}{s^3} = \frac{2066.81 / 30}{10.1^3} = 0.07.$$

Como está cerca de 0, eso significa que la distribución de las estaturas es casi simétrica.

2.8.2 Coeficiente de apuntamiento o curtosis

Definición 2.12 (Coeficiente de apuntamiento muestral g_2). El *coeficiente de apuntamiento muestral* de una variable X es el promedio de las desviaciones de los valores de la muestra respecto de la media muestral, elevadas a la cuarta, dividido por la desviación típica a la cuarta y al resultado se le resta 3.

$$g_2 = \frac{\sum(x_i - \bar{x})^4 n_i / n}{s^4} - 3 = \frac{\sum(x_i - \bar{x})^4 f_i}{s^4} - 3$$

💡 Tip

El coeficiente de apuntamiento mide la concentración de valores en torno a la media y la longitud de las colas de la distribución. Se toma como referencia la distribución normal (campana de Gauss).

- $g_2 = 0$ indica que la distribución tienen un apuntamiento normal, es decir, la concentración de valores en torno a la media es similar al de una campana de Gauss (*mesocúrtica*).

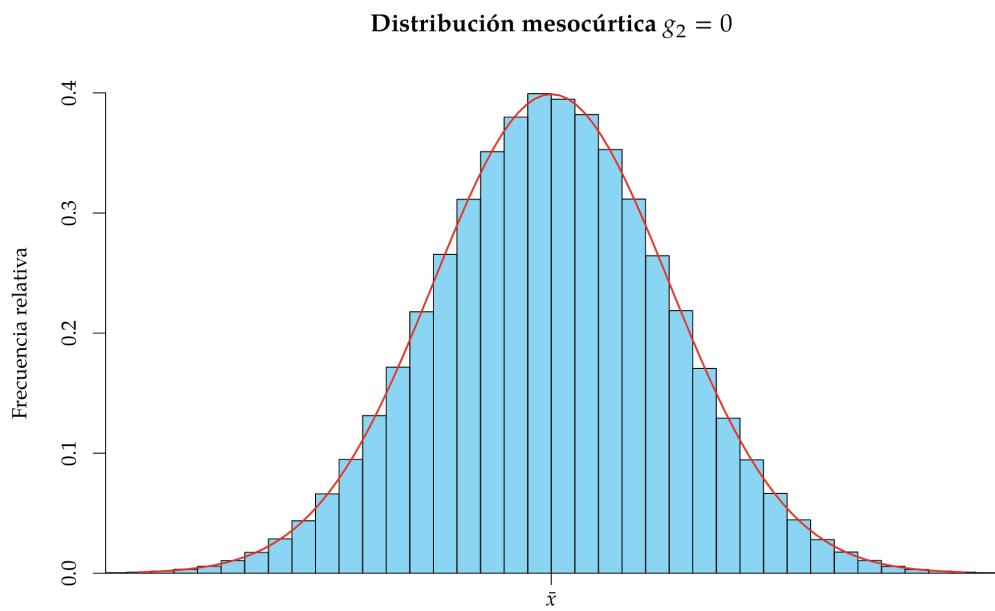


Figura 2.21: Distribución mesocúrtica.

- $g_2 < 0$ indica que la distribución tiene menos apuntamiento de lo normal, es decir, la concentración de valores en torno a la media es menor que en una campana de Gauss (*platicúrtica*).

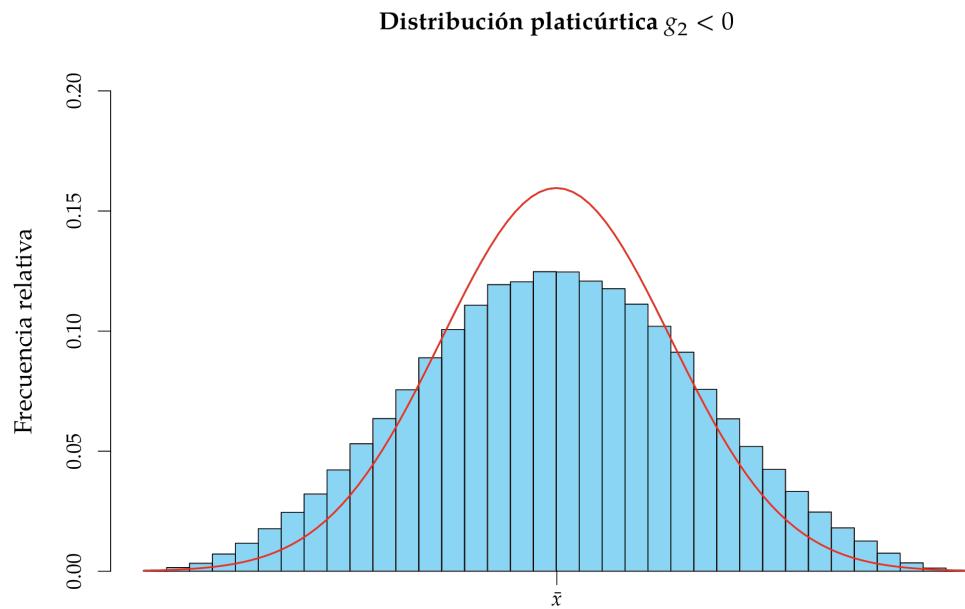


Figura 2.22: Distribución platicúrtica.

- $g_2 > 0$ indica que la distribución tiene más apuntamiento de lo normal, es decir, la concentración de valores en torno a la media es menor que en una campana de Gauss (*leptocúrtica*).

Distribución leptocúrtica $g_2 > 0$

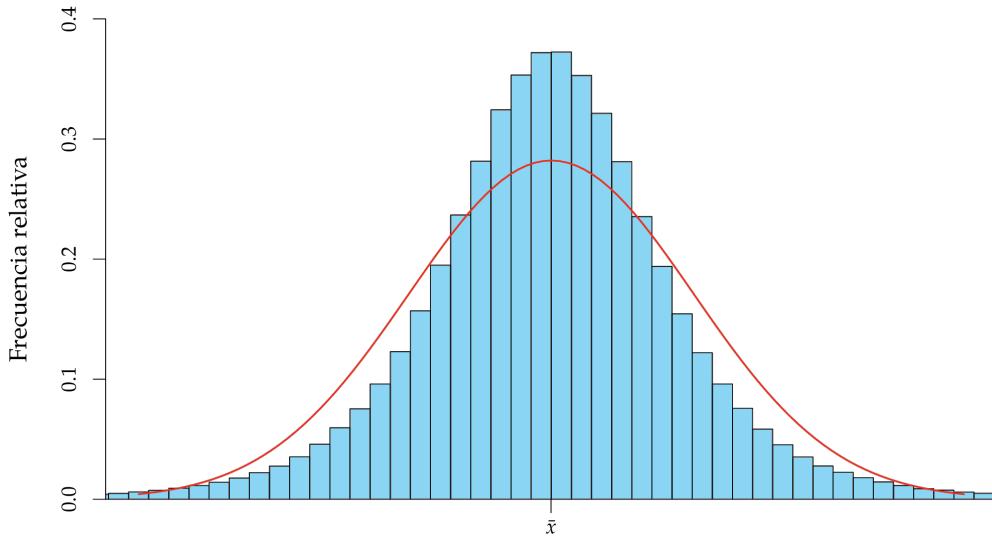


Figura 2.23: Distribución leptocúrtica.

:::{#exm-coeficiente-apuntamiento} ## Datos agrupados Utilizando la tabla de frecuencias de la muestra de estaturas y añadiendo una nueva columna con las desviaciones de la media $\bar{x} = 174.67$ cm a la cuarta, se tiene

X	x_i	n_i	$x_i - \bar{x}$	$(x_i - \bar{x})^4 n_i$
(150, 160]	155	2	-19.67	299396.99
(160, 170]	165	8	-9.67	69951.31
(170, 180]	175	11	0.33	0.13
(180, 190]	185	7	10.33	79707.53
(190, 200]	195	2	20.33	341648.49
\sum		30		790704.45

$$g_2 = \frac{\sum (x_i - \bar{x})^4 n_i / n}{s^4} - 3 = \frac{790704.45 / 30}{10.1^4} - 3 = -0.47.$$

Como se trata de un valor negativo, aunque cercano a 0, podemos decir que la distribución es ligeramente platicúrtica.

Como se verá más adelante en la parte de inferencia, muchas de las pruebas estadísticas solo pueden aplicarse a poblaciones normales.

Las poblaciones normales se caracterizan por ser simétricas y mesocúrticas, de manera que, tanto el coeficiente de asimetría como el de apuntamiento pueden utilizarse para contrastar si los datos de la muestra provienen de una población normal.

💡 Tip

En general, se suele rechazar la hipótesis de normalidad de la población cuando g_1 o g_2 estén fuera del intervalo $[-2, 2]$.

En tal caso, lo habitual es aplicar alguna transformación a la variable para corregir la anormalidad.

2.8.3 Distribuciones no normales

2.8.3.1 Distribución asimétrica a la derecha no normal

Un ejemplo de distribución asimétrica a la derecha es el ingreso de las familias.

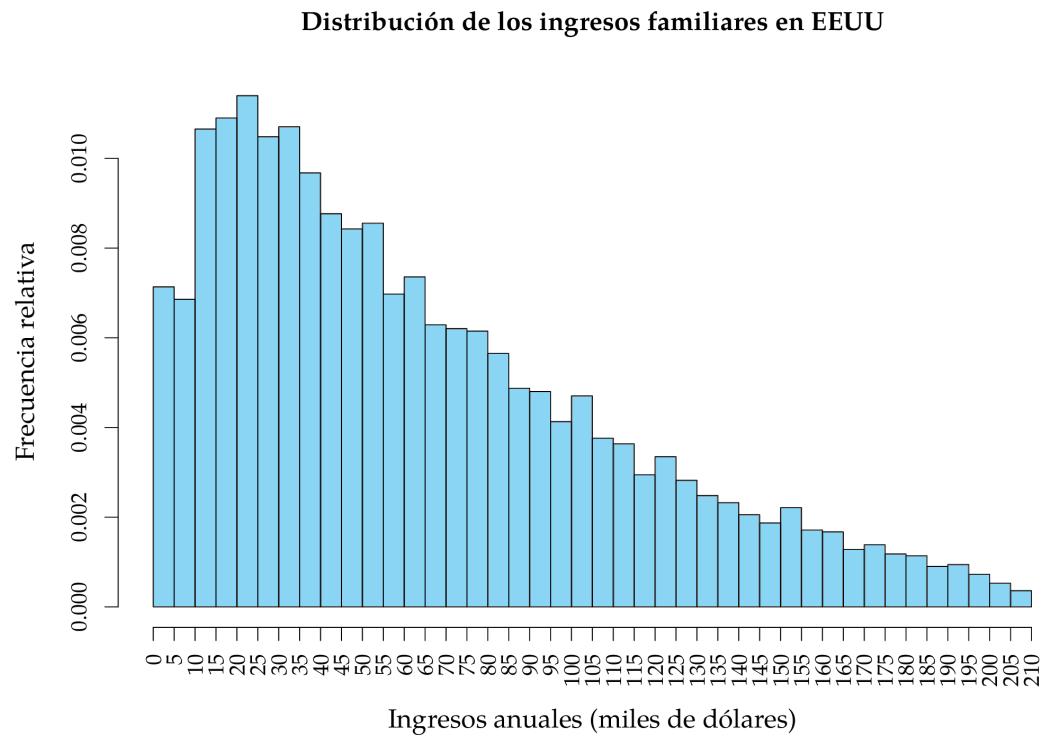


Figura 2.24: Distribucion de los ingresos familiares de EEUU.

2.8.3.2 Distribución asimétrica a la izquierda no normal

Un ejemplo de distribución asimétrica a la izquierda es la edad de fallecimiento.

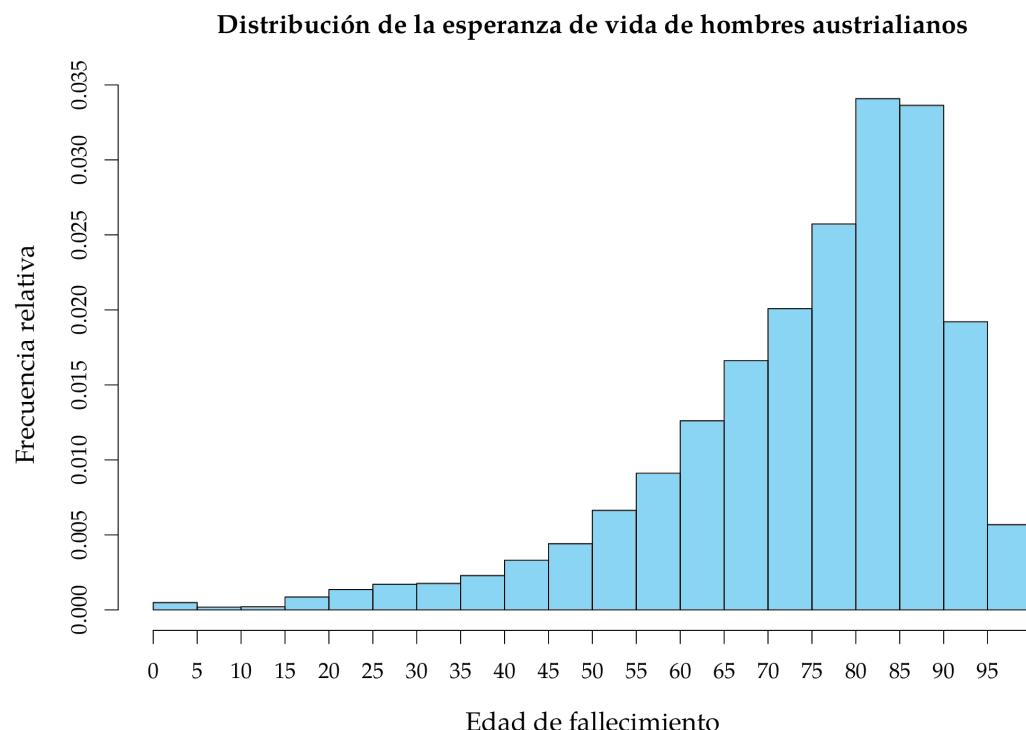


Figura 2.25: Distribucion de la edad de fallecimiento.

2.8.3.3 Distribución bimodal no normal

Un ejemplo de distribución bimodal es la hora de llegada de los clientes de un restaurante.

Distribución de la hora de llegada de los clientes de un restaurante

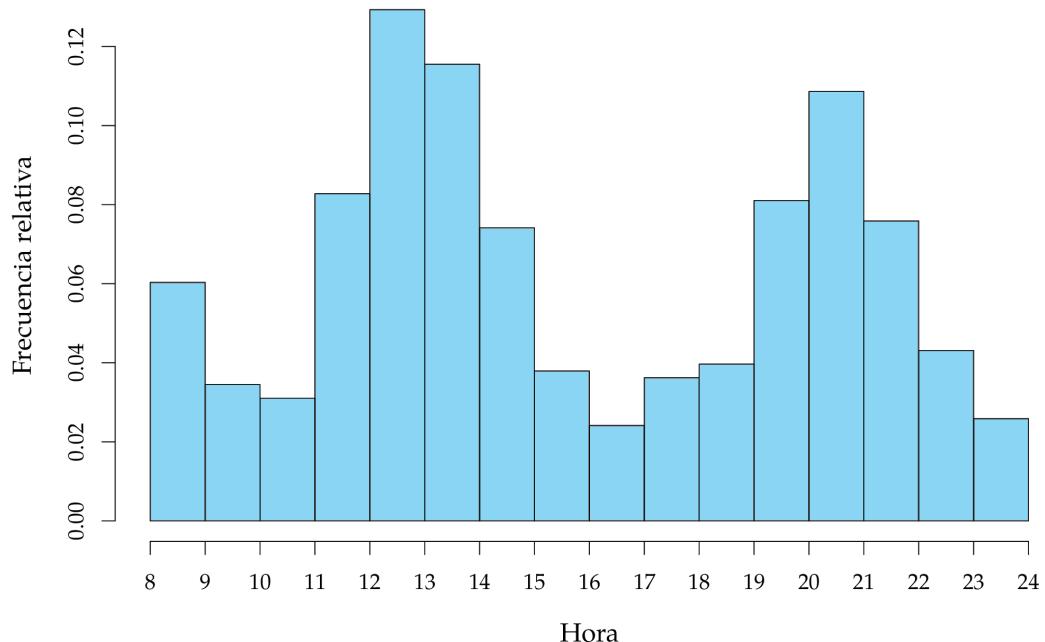


Figura 2.26: Distribucion de la hora de llegada de los clientes de un restaurante.

2.9 Transformaciones de variables

En muchas ocasiones se suelen transformar los datos brutos para corregir alguna anormalidad de la distribución o simplemente para trabajar con unas unidades más cómodas.

Por ejemplo, si estamos trabajando con estaturas medidas en metros y tenemos los siguientes valores:

1.75 m, 1.65 m, 1.80 m,

podemos evitar los decimales multiplicando por 100, es decir, pasando de metros a centímetros:

175 cm, 165 cm, 180 cm,

Y si queremos reducir la magnitud de los datos podemos restarles a todos el menor de ellos, en este caso, 165cm:

10cm, 0cm, 15cm,

Está claro que este conjunto de datos es mucho más sencillo que el original. En el fondo lo que se ha hecho es aplicar a los datos la transformación:

$$Y = 100X - 165$$

2.9.1 Transformaciones lineales

Una de las transformaciones más habituales es la *transformación lineal*:

$$Y = a + bX.$$

Teorema 2.1. *Dada una variable muestral X , si Y es la variable muestral que resulta de aplicar a X la transformación lineal $Y = a + bX$, entonces*

$$\begin{aligned}\bar{y} &= a + b\bar{x}, \\ s_y &= |b|s_x\end{aligned}$$

Además, el coeficiente de curtosis no se altera y el de asimetría sólo cambia de signo si b es negativo.

i Demostración

Se deja como ejercicio.

2.9.2 Transformación de tipificación y puntuaciones típicas

Una de las transformaciones lineales más habituales es la *tipificación*:

Definición 2.13 (Variable tipificada). La *variable tipificada* de una variable estadística X es la variable que resulta de restarle su media y dividir por su desviación típica.

$$Z = \frac{X - \bar{x}}{s_x}$$

Para cada valor x_i de la muestra, la *puntuación típica* es el valor que resulta de aplicarle la transformación de tipificación

$$z_i = \frac{x_i - \bar{x}}{s_x}.$$

 Tip

La puntuación típica es el número de desviaciones típicas que un valor está por encima o por debajo de la media, y es útil para evitar la dependencia de una variable respecto de las unidades de medida empleadas. Esto es útil, por ejemplo, para comparar valores de variables o muestras distintas.

Dada una variable muestral X , si Z es la variable tipificada de X , entonces

$$\bar{z} = 0 \quad s_z = 1.$$

 Demostración

Se deja como ejercicio.

Ejemplo 2.33. Las notas de 5 alumnos en dos asignaturas X e Y son

Alumno:	1	2	3	4	5		
$X :$	2	5	4	8	6	$\bar{x} = 5$	$s_x = 2$
$Y :$	1	9	8	5	2	$\bar{y} = 5$	$s_y = 3.16$

¿Ha tenido el mismo rendimiento el cuarto alumno en la asignatura X que el tercero en la asignatura Y ?

Podría parecer que ambos alumnos han tenido el mismo rendimiento puesto que tienen la misma nota, pero si queremos ver el rendimiento relativo al resto del grupo, tendríamos que tener en cuenta la dispersión de cada muestra y medir sus puntuaciones típicas:

Alumno:	1	2	3	4	5		
$X :$	-1.50	0.00	-0.50	1.50	0.50		
$Y :$	-1.26	1.26	0.95	0.00	-0.95		

Es decir, el alumno que tiene un 8 en X está 1.5 veces la desviación típica por encima de la media de X , mientras que el alumno que tiene un 8 en Y sólo está 0.95 desviaciones típicas por encima de la media de Y . Así pues, el primer alumno tuvo un rendimiento superior al segundo.

Siguiendo con el ejemplo anterior y considerando ambas asignaturas, *¿cuál es el mejor alumno?*

Si simplemente se suman las puntuaciones de cada asignatura se tiene:

Alumno:	1	2	3	4	5
$X :$	2	5	4	8	6
$Y :$	1	9	8	5	2
\sum	3	14	12	13	8

El mejor alumno sería el segundo.

Pero si se considera el rendimiento relativo tomando las puntuaciones típicas se tiene

Alumno:	1	2	3	4	5
$X :$	-1.50	0.00	-0.50	1.50	0.50
$Y :$	-1.26	1.26	0.95	0.00	-0.95
\sum	-2.76	1.26	0.45	1.5	-0.45

Y el mejor alumno sería el cuarto.

2.9.2.1 Transformaciones no lineales

Las transformaciones no lineales son también habituales para corregir la anormalidad de las distribuciones.

La transformación $Y = X^2$ comprime la escala para valores pequeños y la expande para valores altos, de manera que es muy útil para corregir asimetrías hacia la izquierda.

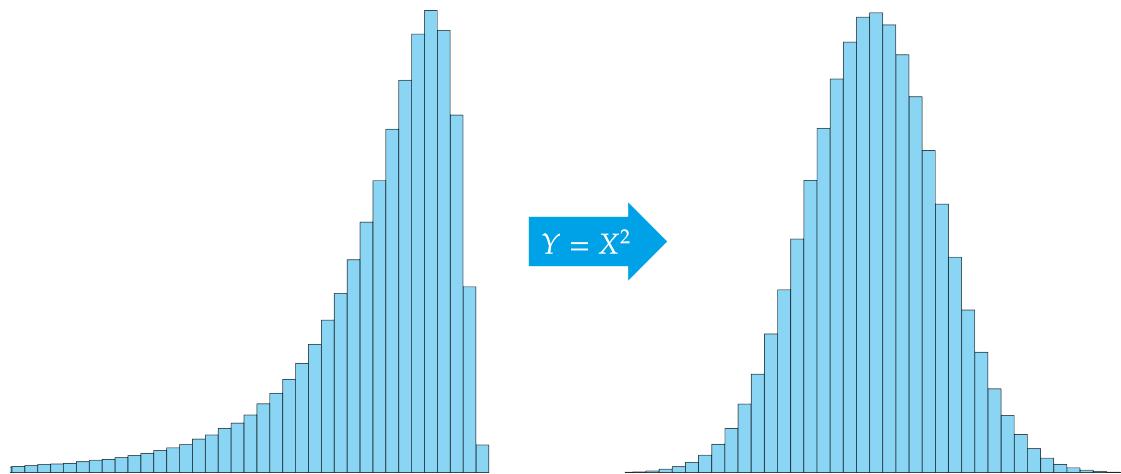


Figura 2.27: Transformación cuadrática.

Las transformaciones $Y = \sqrt{x}$, $Y = \log X$ y $Y = 1/X$ comprimen la escala para valores altos y la expanden para valores pequeños, de manera que son útiles para corregir asimetrías hacia la derecha.

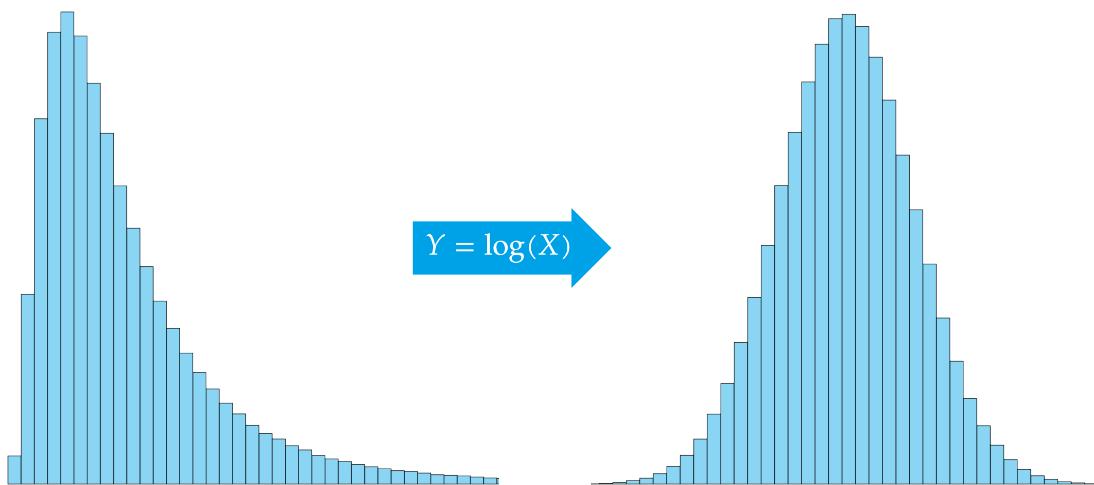


Figura 2.28: Transformación logarítmica.

2.9.3 Variables clasificadoras o factores

En ocasiones interesa describir el comportamiento de una variable, no para toda la muestra, sino para distintos grupos de individuos correspondientes a las categorías de otra variable conocida como **variable clasificadora** o **factor**.

Ejemplo 2.34. Dividiendo la muestra de estaturas según el sexo se obtienen dos submuestras:

Mujeres	173, 158, 174, 166, 162, 177, 165, 154, 166, 182, 169, 172, 170, 168.
Hombres	179, 181, 172, 194, 185, 187, 198, 178, 188, 171, 175, 167, 186, 172, 176, 187.

Habitualmente los factores se usan para comparar la distribución de la variable principal para cada categoría del factor.

Ejemplo 2.35. Los siguientes diagramas permiten comparar la distribución de estaturas según el sexo.

Histograma de estaturas según el sexo

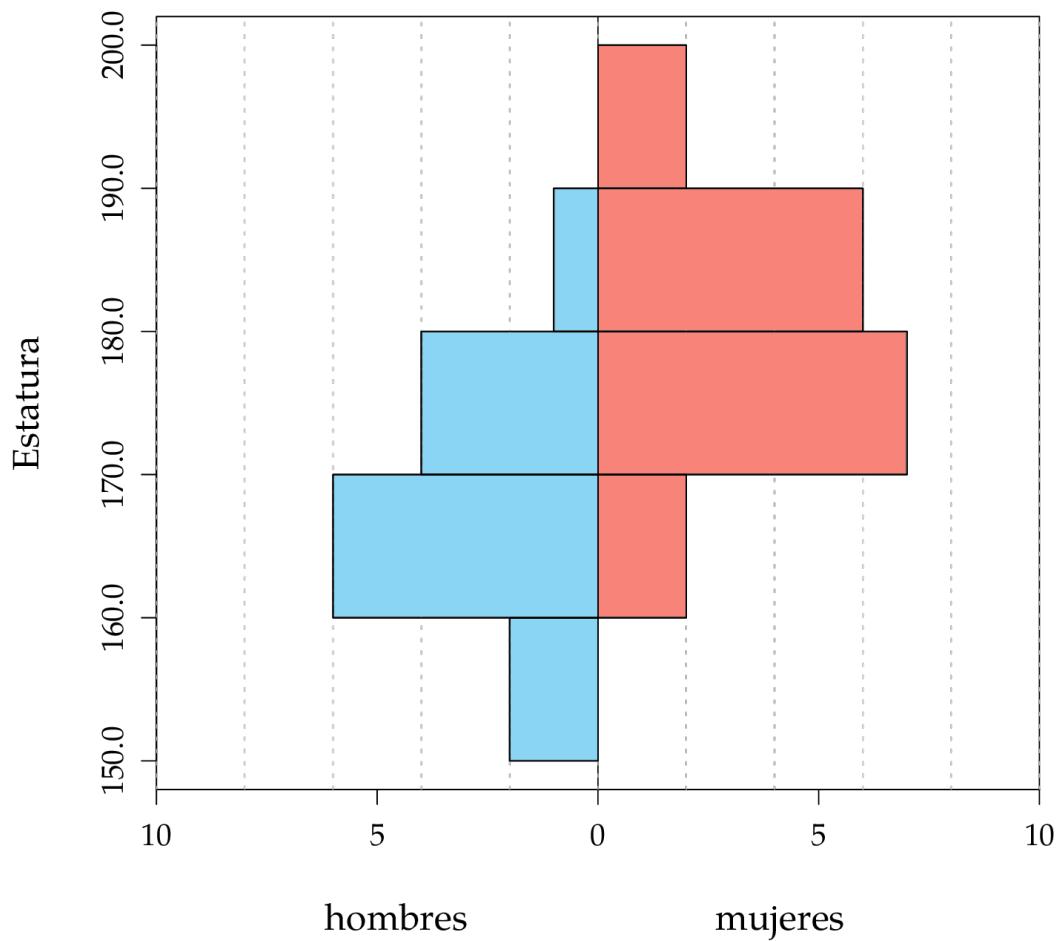


Figura 2.29: Histograma de estaturas por sexo.

Diagrama de cajas de la Estatura según el Sexo

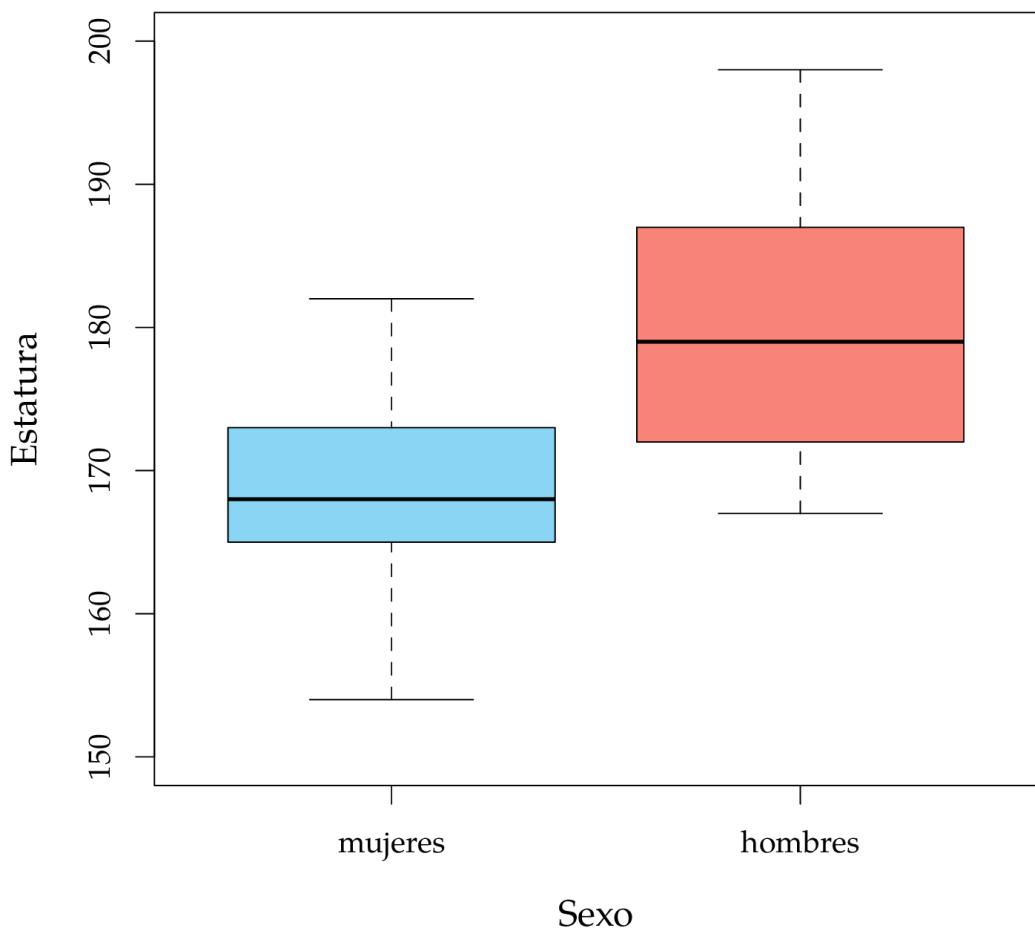


Figura 2.30: Diagramas de cajas de estaturas por sexo.

3 Regresión

Hasta ahora se ha visto como describir el comportamiento de una variable, pero en los fenómenos naturales normalmente aparecen más de una variable que suelen estar relacionadas. Por ejemplo, en un estudio sobre el peso de las personas, deberíamos incluir todas las variables con las que podría tener relación: altura, edad, sexo, dieta, tabaco, ejercicio físico, etc.

Para comprender el fenómeno no basta con estudiar cada variable por separado y es preciso un estudio conjunto de todas las variables para ver cómo interactúan y qué relaciones se dan entre ellas. El objetivo de la estadística en este caso es dar medidas del grado y del tipo de relación entre dichas variables.

Generalmente, en un *estudio de dependencia* se considera una **variable dependiente** Y que se supone relacionada con otras variables X_1, \dots, X_n llamadas **variables independientes**.

El caso más simple es el de una sola variable independiente, y en tal caso se habla de *estudio de dependencia simple*. Para más de una variable independiente se habla de *estudio de dependencia múltiple*.

En este capítulo se verán los estudios de dependencia simple que son más sencillos.

3.1 Distribución de frecuencias conjunta

3.1.1 Frecuencias conjuntas

Al estudiar la dependencia simple entre dos variables X e Y , no se pueden estudiar sus distribuciones por separado, sino que hay que estudiar la distribución conjunta de la **variable bidimensional** (X, Y) , cuyos valores son los pares (x_i, y_j) donde el primer elemento es un valor X y el segundo uno de Y .

Definición 3.1 (Frecuencias muestrales conjuntas). Dada una muestra de tamaño n de una variable bidimensional (X, Y) , para cada valor de la variable (x_i, y_j) observado en la muestra se define

- **Frecuencia absoluta** n_{ij} : Es el número de veces que el par (x_i, y_j) aparece en la muestra.

- **Frecuencia relativa** f_{ij} : Es la proporción de veces que el par (x_i, y_j) aparece en la muestra.

$$f_{ij} = \frac{n_{ij}}{n}$$

 Advertencia

Para las variables bidimensionales no tienen sentido las frecuencias acumuladas.

3.1.2 Distribución de frecuencias bidimensional

Al conjunto de valores de la variable bidimensional y sus respectivas frecuencias muestrales se le denomina **distribución de frecuencias bidimensional**, y se representa mediante una **tabla de frecuencias bidimensional**.

$X \setminus Y$	y_1	\cdots	y_j	\cdots	y_q
x_1	n_{11}	\cdots	n_{1j}	\cdots	n_{1q}
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
x_i	n_{i1}	\cdots	n_{ij}	\cdots	n_{iq}
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
x_p	n_{p1}	\cdots	n_{pj}	\cdots	n_{pq}

Ejemplo 3.1. La estatura (en cm) y el peso (en Kg) de una muestra de 30 estudiantes es:

(179,85), (173,65), (181,71), (170,65), (158,51), (174,66), (172,62), (166,60), (194,90), (185,75), (162,55), (187,78), (198,109), (177,61), (178,70), (165,58), (154,50), (183,93), (166,51), (171,65), (175,70), (182,60), (167,59), (169,62), (172,70), (186,71), (172,54), (176,68), (168,67), (187,80).

La tabla de frecuencias bidimensional es

X/Y	[50, 60)	[60, 70)	[70, 80)	[80, 90)	[90, 100)	[100, 110)
(150, 160]	2	0	0	0	0	0
(160, 170]	4	4	0	0	0	0
(170, 180]	1	6	3	1	0	0
(180, 190]	0	1	4	1	1	0
(190, 200]	0	0	0	0	1	1

3.1.3 Diagrama de dispersión

La distribución de frecuencias conjunta de una variable bidimensional puede representarse gráficamente mediante un **diagrama de dispersión**, donde los datos se representan como una colección de puntos en un plano cartesiano.

Habitualmente la variable independiente se representa en el eje X y la variable dependiente en el eje Y . Por cada par de valores (x_i, y_j) en la muestra se dibuja un punto en el plano con esas coordenadas.

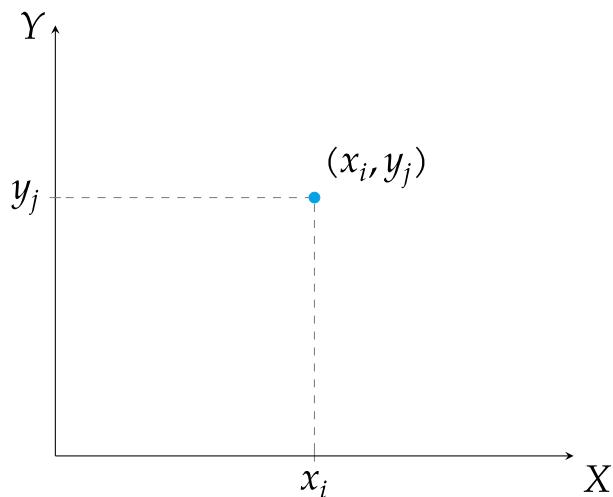


Figura 3.1: Diagrama de dispersión.

El resultado es un conjunto de puntos que se conoce como *nube de puntos*.

Ejemplo 3.2. El siguiente diagrama de dispersión representa la distribución conjunta de estaturas y pesos de la muestra anterior.

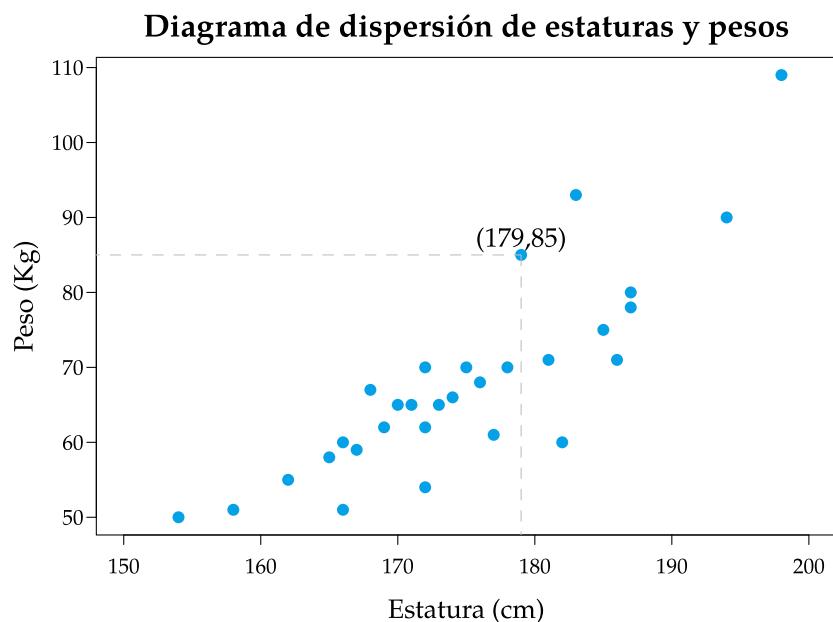


Figura 3.2: Diagrama de dispersión de estaturas y pesos.

i Interpretación

El diagrama de dispersión da información visual sobre el tipo de relación entre las variables.

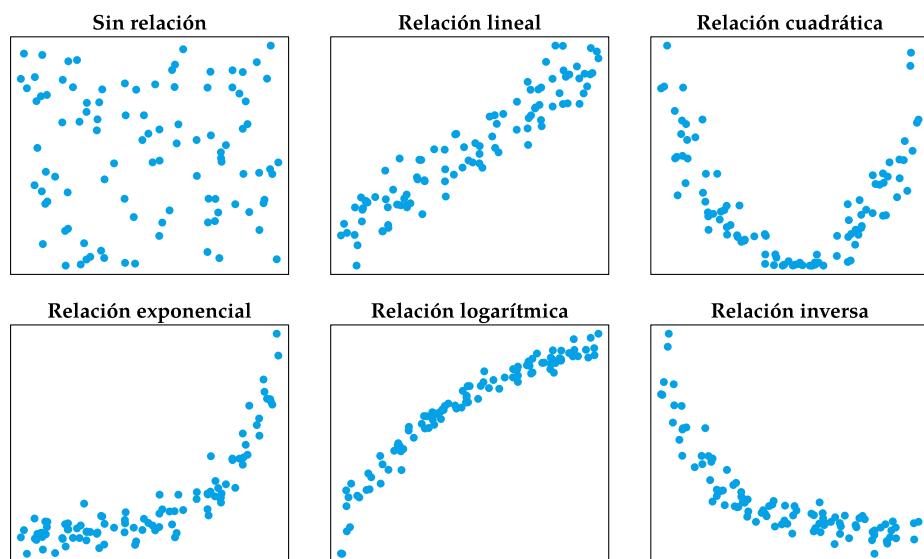


Figura 3.3: Diagramas de dispersión de diferentes tipos de relaciones.

3.1.4 Distribuciones marginales

A cada una de las distribuciones de las variables que conforman la variable bidimensional se les llama .

Las distribuciones marginales se pueden obtener a partir de la tabla de frecuencias bidimensional, sumando las frecuencias por filas y columnas.

$X \setminus Y$	y_1	...	y_j	...	y_q	n_x
x_1	n_{11}	...	n_{1j}	...	n_{1q}	n_{x_1}
\vdots	\vdots	\vdots	$\downarrow +$	\vdots	\vdots	\vdots
x_i	n_{i1}	$\xrightarrow{+}$	n_{ij}	$\xrightarrow{+}$	n_{iq}	n_{x_i}
\vdots	\vdots	\vdots	$\downarrow +$	\vdots	\vdots	\vdots
x_p	n_{p1}	...	n_{pj}	...	n_{pq}	n_{x_p}
n_y	n_{y_1}	...	n_{y_j}	...	n_{y_q}	n

Ejemplo 3.3. En el ejemplo anterior de las estaturas y los pesos, las distribuciones marginales son

X/Y	[50, 60]	[60, 70]	[70, 80]	[80, 90]	[90, 100]	[100, 110]	n_x
(150, 160]	2	0	0	0	0	0	2
(160, 170]	4	4	0	0	0	0	8
(170, 180]	1	6	3	1	0	0	11
(180, 190]	0	1	4	1	1	0	7
(190, 200]	0	0	0	0	1	1	2
n_y	7	11	7	2	2	1	30

y los estadísticos correspondientes son

$$\bar{x} = 174.67 \text{ cm} \quad s_x^2 = 102.06 \text{ cm}^2 \quad s_x = 10.1 \text{ cm}$$

$$\bar{y} = 69.67 \text{ Kg} \quad s_y^2 = 164.42 \text{ Kg}^2 \quad s_y = 12.82 \text{ Kg}$$

3.2 Covarianza

Para analizar la relación entre dos variables cuantitativas es importante hacer un estudio conjunto de las desviaciones respecto de la media de cada variable.

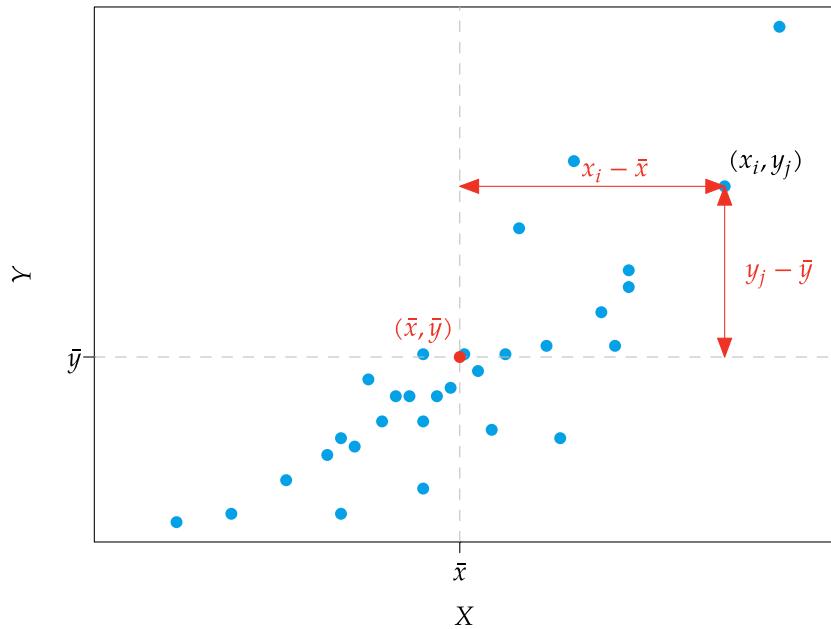


Figura 3.4: Desviaciones de las medias en un diagrama de dispersión.

Si dividimos la nube de puntos del diagrama de dispersión en 4 cuadrantes centrados en el punto de medias (\bar{x}, \bar{y}) , el signo de las desviaciones será:

Cuadrante	$(x_i - \bar{x})$	$(y_j - \bar{y})$	$(x_i - \bar{x})(y_j - \bar{y})$
1	+	+	+
2	-	+	-
3	-	-	+
4	+	-	-

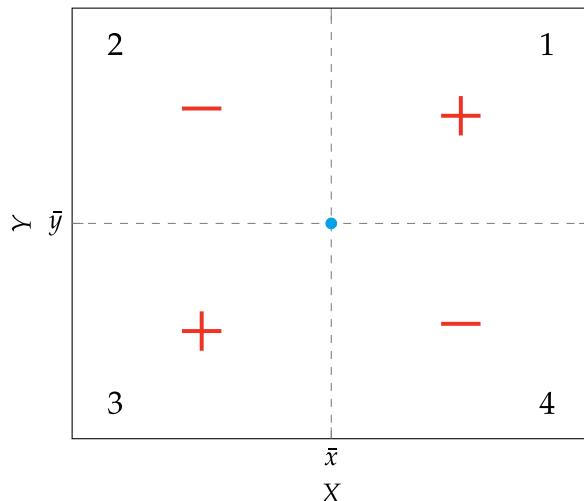


Figura 3.5: Cuadrantes de un diagrama de dispersión.

Si la relación entre las variables es *lineal y creciente*, entonces la mayor parte de los puntos estarán en los cuadrantes 1 y 3 y la suma de los productos de desviaciones será positiva.

$$\sum (x_i - \bar{x})(y_j - \bar{y}) > 0$$

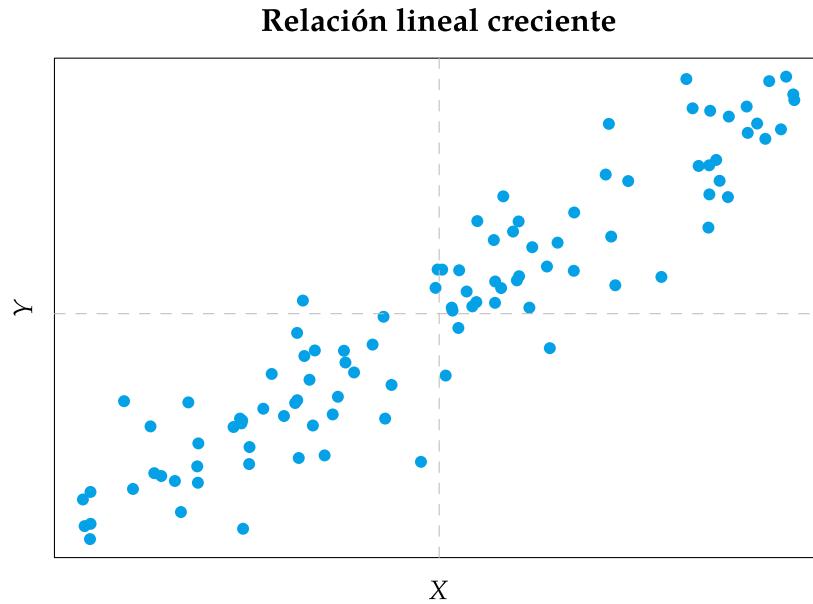


Figura 3.6: Diagrama de dispersión de una relación lineal creciente.

Si la relación entre las variables es *lineal y decreciente*, entonces la mayor parte de los puntos estarán en los cuadrantes 2 y 4 y la suma de los productos de desviaciones será negativa.

$$\sum(x_i - \bar{x})(y_j - \bar{y}) = -$$

Relación lineal decreciente

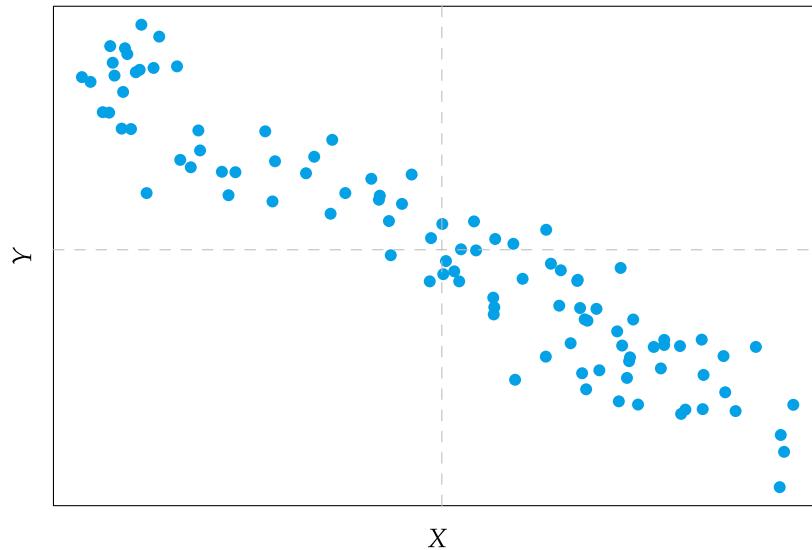


Figura 3.7: Diagrama de dispersión de una relación lineal decreciente.

Usando el producto de las desviaciones respecto de las medias surge el siguiente estadístico.

Definición 3.2 (Covarianza muestral). La *covarianza muestral* de una variable aleatoria bidimensional (X, Y) se define como el promedio de los productos de las respectivas desviaciones respecto de las medias de X e Y .

$$s_{xy} = \frac{\sum(x_i - \bar{x})(y_j - \bar{y})n_{ij}}{n}$$

También puede calcularse de manera más sencilla mediante la fórmula

$$s_{xy} = \frac{\sum x_i y_j n_{ij}}{n} - \bar{x}\bar{y}.$$

i Interpretación

La covarianza sirve para estudiar la relación lineal entre dos variables:

- Si $s_{xy} > 0$ existe una relación lineal creciente.
- Si $s_{xy} < 0$ existe una relación lineal decreciente.
- Si $s_{xy} = 0$ no existe relación lineal.

Ejemplo 3.4. Utilizando la tabla de frecuencias bidimensional de la muestra de estaturas y pesos

X/Y	[50, 60)	[60, 70)	[70, 80)	[80, 90)	[90, 100)	[100, 110)	n_x
(150, 160]	2	0	0	0	0	0	2
(160, 170]	4	4	0	0	0	0	8
(170, 180]	1	6	3	1	0	0	11
(180, 190]	0	1	4	1	1	0	7
(190, 200]	0	0	0	0	1	1	2
n_y	7	11	7	2	2	1	30

$$\bar{x} = 174.67 \text{ cm} \quad \bar{y} = 69.67 \text{ Kg}$$

la covarianza vale

$$s_{xy} = \frac{\sum x_i y_j n_{ij}}{n} - \bar{x}\bar{y} = \frac{155 \cdot 55 \cdot 2 + 165 \cdot 55 \cdot 4 + \dots + 195 \cdot 105 \cdot 1}{30} - 174.67 \cdot 69.67 = \\ = \frac{368200}{30} - 12169.26 = 104.07 \text{ cm} \cdot \text{Kg}.$$

Esto indica que existe una relación lineal creciente entre la estatura y el peso.

3.3 Regresión

En muchos casos el objetivo de un estudio no es solo detectar una relación entre dos variables, sino explicarla mediante alguna función matemática

$$y = f(x)$$

que permita predecir la variable dependiente para cada valor de la independiente.

La **regresión** es la parte de la Estadística encargada de construir esta función, que se conoce como **función de regresión o modelo de regresión**.

3.3.1 Modelos de regresión simple

Dependiendo de la forma de función de regresión, existen muchos tipos de regresión simple. Los más habituales son los que aparecen en la siguiente tabla:

Modelo	Ecuación
Lineal	$y = a + bx$
Cuadrático	$y = a + bx + cx^2$
Cúbico	$y = a + bx + cx^2 + dx^3$
Potencial	$y = a \cdot x^b$
Exponencial	$y = e^{a+bx}$
Logarítmico	$y = a + b \log x$
Inverso	$y = a + \frac{b}{x}$
Sigmoidal	$y = e^{a+\frac{b}{x}}$

La elección de un tipo u otro depende de la forma que tenga la nube de puntos del diagrama de dispersión.

3.3.2 Residuos o errores predictivos

Una vez elegida la familia de curvas que mejor se adapta a la nube de puntos, se determina, dentro de dicha familia, la curva que mejor se ajusta a la distribución, es decir, la función que mejor predice la variable dependiente.

El objetivo es encontrar la función de regresión que haga mínimas las distancias entre los valores de la variable dependiente observados en la muestra, y los predichos por la función de regresión. Estas distancias se conocen como *residuos* o *errores predictivos*.

Definición 3.3 (Residuos o errores predictivos). Dado el modelo de regresión $y = f(x)$ para una variable bidimensional (X, Y) , el *residuo* o *error predictivo* de un valor (x_i, y_j) observado en la muestra, es la diferencia entre el valor observado de la variable dependiente y_j y el predicho por la función de regresión para x_i ,

$$e_{ij} = y_j - f(x_i).$$

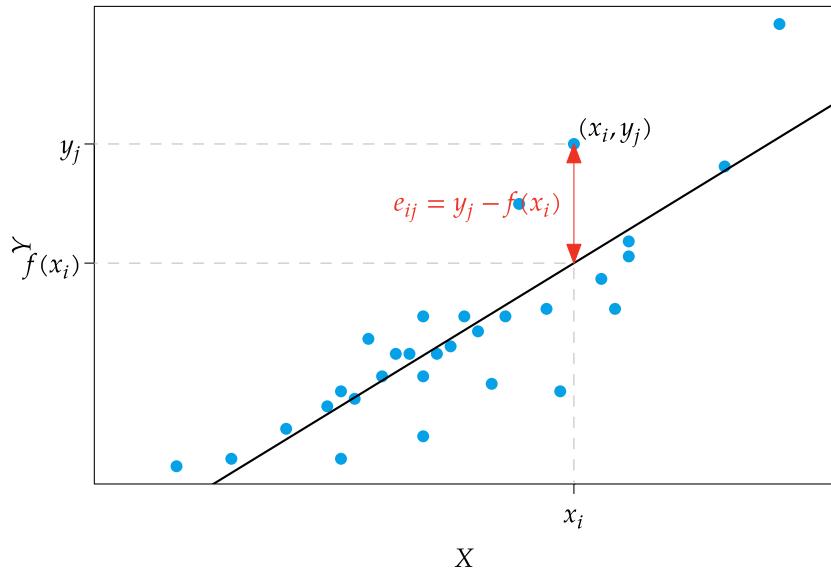


Figura 3.8: Residuos de un modelo de regresión.

3.3.3 Ajuste de mínimos cuadrados

Una forma posible de obtener la función de regresión es mediante el método de *mínimos cuadrados* que consiste en calcular la función que haga mínima la suma de los cuadrados de los residuos

$$\sum e_{ij}^2.$$

En el caso de un modelo de regresión lineal $f(x) = a + bx$, como la recta depende de dos parámetros (el término independiente a y la pendiente b), la suma también dependerá de estos parámetros

$$\theta(a, b) = \sum e_{ij}^2 = \sum (y_j - f(x_i))^2 = \sum (y_j - a - bx_i)^2.$$

Así pues, todo se reduce a buscar los valores a y b que hacen mínima esta suma.

Considerando la suma de los cuadrados de los residuos como una función de dos variables $\theta(a, b)$, se pueden calcular los valores de los parámetros del modelo que hacen mínima esta suma derivando e igualando a 0 las derivadas con respecto a a y b .

$$\frac{\partial \theta(a, b)}{\partial a} = \frac{\partial \sum(y_j - a - bx_i)^2}{\partial a} = 0$$

$$\frac{\partial \theta(a, b)}{\partial b} = \frac{\partial \sum(y_j - a - bx_i)^2}{\partial b} = 0$$

Tras resolver el sistema se obtienen los valores

$$a = \bar{y} - \frac{s_{xy}}{s_x^2} \bar{x} \quad b = \frac{s_{xy}}{s_x^2}$$

Estos valores hacen mínimos los residuos en Y y por tanto dan la recta de regresión óptima.

3.3.4 Coeficiente de determinación

A partir de la varianza residual se puede definir otro estadístico más sencillo de interpretar.

Definición 3.4 (Coeficiente de determinación muestral r^2). Dado un modelo de regresión simple $y = f(x)$ de una variable bidimensional (X, Y) , su *coeficiente de determinación muestral* es

$$r^2 = 1 - \frac{s_{ry}^2}{s_y^2}$$

⚠️ Advertencia

Como la varianza residual puede tomar valores entre 0 y s_y^2 , se tiene que

$$0 \leq r^2 \leq 1$$

ℹ️ Interpretación

Cuanto mayor sea r^2 , mejor explicará el modelo de regresión la relación entre las variables, en particular:

- Si $r^2 = 0$ entonces no existe relación del tipo planteado por el modelo.
- Si $r^2 = 1$ entonces la relación que plantea el modelo es perfecta.

 Advertencia

En el caso de las rectas de regresión, el coeficiente de determinación puede calcularse con esta fórmula

$$r^2 = \frac{s_{xy}^2}{s_x^2 s_y^2}.$$

 Demostración

Prueba. Cuando el modelo ajustado es la recta de regresión la varianza residual vale

$$\begin{aligned} s_{ry}^2 &= \sum e_{ij}^2 f_{ij} = \sum (y_j - f(x_i))^2 f_{ij} = \sum \left(y_j - \bar{y} - \frac{s_{xy}}{s_x^2} (x_i - \bar{x}) \right)^2 f_{ij} = \\ &= \sum \left((y_j - \bar{y})^2 + \frac{s_{xy}^2}{s_x^4} (x_i - \bar{x})^2 - 2 \frac{s_{xy}}{s_x^2} (x_i - \bar{x})(y_j - \bar{y}) \right) f_{ij} = \\ &= \sum (y_j - \bar{y})^2 f_{ij} + \frac{s_{xy}^2}{s_x^4} \sum (x_i - \bar{x})^2 f_{ij} - 2 \frac{s_{xy}}{s_x^2} \sum (x_i - \bar{x})(y_j - \bar{y}) f_{ij} = \\ &= s_y^2 + \frac{s_{xy}^2}{s_x^4} s_x^2 - 2 \frac{s_{xy}}{s_x^2} s_{xy} = s_y^2 - \frac{s_{xy}^2}{s_x^2}. \end{aligned}$$

y, por tanto, el coeficiente de determinación lineal vale

$$r^2 = 1 - \frac{s_{ry}^2}{s_y^2} = 1 - \frac{s_y^2 - \frac{s_{xy}^2}{s_x^2}}{s_y^2} = 1 - 1 + \frac{s_{xy}^2}{s_x^2 s_y^2} = \frac{s_{xy}^2}{s_x^2 s_y^2}.$$

□

Ejemplo 3.5. En el ejemplo de las estaturas y pesos se tenía

$$\begin{array}{ll} \bar{x} = 174.67 \text{ cm} & s_x^2 = 102.06 \text{ cm}^2 \\ \bar{y} = 69.67 \text{ Kg} & s_y^2 = 164.42 \text{ Kg}^2 \\ s_{xy} = 104.07 \text{ cm} \cdot \text{Kg} & \end{array}$$

De modo que el coeficiente de determinación lineal vale

$$r^2 = \frac{s_{xy}^2}{s_x^2 s_y^2} = \frac{(104.07 \text{ cm} \cdot \text{Kg})^2}{102.06 \text{ cm}^2 \cdot 164.42 \text{ Kg}^2} = 0.65.$$

Esto indica que la recta de regresión del peso sobre la estatura explica el 65% de la variabilidad del peso, y de igual modo, la recta de regresión de la estatura sobre el peso explica el 65% de la variabilidad de la estatura.

3.3.5 Coeficiente de correlación lineal

Definición 3.5 (Coeficiente de correlación lineal muestral). Dada una variable bidimensional (X, Y) , el *coeficiente de correlación lineal muestral* es la raíz cuadrada de su coeficiente de determinación lineal, con signo el de la covarianza

$$r = \sqrt{r^2} = \frac{s_{xy}}{s_x s_y}.$$

 Advertencia

Como r^2 toma valores entre 0 y 1, r tomará valores entre -1 y 1,

$$-1 \leq r \leq 1$$

 Interpretación

El coeficiente de correlación lineal no sólo mide el grado de dependencia lineal sino también su dirección (creciente o decreciente):

- Si $r = 0$ entonces no existe relación lineal.
- Si $r = 1$ entonces existe una relación lineal creciente perfecta.
- Si $r = -1$ entonces existe una relación lineal decreciente perfecta.

:::{#exm-coeficiente-correlacion} En el ejemplo de las estaturas y los pesos se tenía

$$\begin{aligned}\bar{x} &= 174.67 \text{ cm} & s_x^2 &= 102.06 \text{ cm}^2 \\ \bar{y} &= 69.67 \text{ Kg} & s_y^2 &= 164.42 \text{ Kg}^2 \\ s_{xy} &= 104.07 \text{ cm} \cdot \text{Kg}\end{aligned}$$

De manera que el coeficiente de correlación lineal es

$$r = \frac{s_{xy}}{s_x s_y} = \frac{104.07 \text{ cm} \cdot \text{Kg}}{10.1 \text{ cm} \cdot 12.82 \text{ Kg}} = +0.8.$$

Esto indica que la relación lineal entre el peso y la estatura es fuerte, y además creciente.

3.3.6 Distintos grados de correlación

Los siguientes diagramas de dispersión muestran modelos de regresión lineales con diferentes grados de correlación.

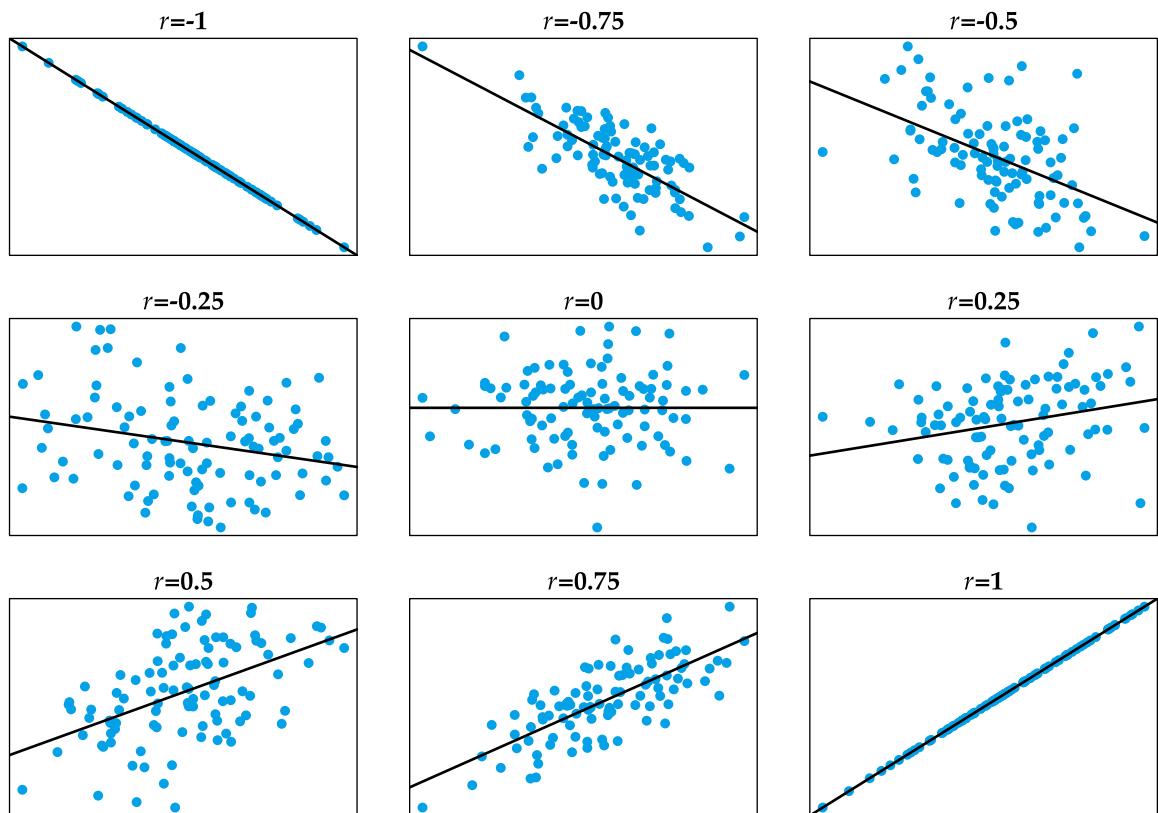


Figura 3.9: Modelos de regresión lineales con diferentes grados de correlación.

3.3.7 Fiabilidad de las predicciones de un modelo de regresión

Aunque el coeficiente de determinación o el de correlación determinan la bondad de ajuste de un modelo de regresión, existen otros factores que influyen en la fiabilidad de las predicciones de un modelo de regresión:

- El coeficiente de determinación: Cuanto mayor sea, menores serán los errores predictivos y mayor la fiabilidad de las predicciones.
- La variabilidad de la población: Cuanto más variable es una población, más difícil es predecir y por tanto menos fiables serán las predicciones.
- El tamaño muestral: Cuanto mayor sea, más información tendremos y, en consecuencia, más fiables serán las predicciones.

Advertencia

Además, hay que tener en cuenta que un modelo de regresión es válido únicamente para el rango de valores observados en la muestra. Fuera de ese rango no hay información del tipo de relación entre las variables, por lo que no deben hacerse predicciones para valores lejos de los observados en la muestra.

3.4 Regresión no lineal

El ajuste de un modelo de regresión no lineal es similar al del modelo lineal y también puede realizarse mediante la técnica de mínimos cuadrados.

No obstante, en determinados casos un ajuste no lineal puede convertirse en un ajuste lineal mediante una sencilla transformación de alguna de las variables del modelo.

3.4.1 Transformación de modelos de regresión no lineales

- **Logarítmico:** Un modelo logarítmico $y = a + b \log x$ se convierte en un modelo lineal haciendo el cambio $t = \log x$:

$$y = a + b \log x = a + bt.$$

- **Exponencial:** Un modelo exponencial $y = ae^{bx}$ se convierte en un modelo lineal haciendo el cambio $z = \log y$:

$$z = \log y = \log(ae^{bx}) = \log a + \log e^{bx} = a' + bx.$$

- **Potencial:** Un modelo potencial $y = ax^b$ se convierte en un modelo lineal haciendo los cambios $t = \log x$ y $z = \log y$:

$$z = \log y = \log(ax^b) = \log a + b \log x = a' + bt.$$

- **Inverso:** Un modelo inverso $y = a + b/x$ se convierte en un modelo lineal haciendo el cambio $t = 1/x$:

$$y = a + b(1/x) = a + bt.$$

- **Sigmoidal:** Un modelo curva S $y = e^{a+b/x}$ se convierte en un modelo lineal haciendo los cambios $t = 1/x$ y $z = \log y$:

$$z = \log y = \log(e^{a+b/x}) = a + b(1/x) = a + bt.$$

3.4.2 Relación exponencial

:::{#exm-regresion-exponencial} El número de bacterias de un cultivo evoluciona con el tiempo según la siguiente tabla:

Horas	Bacterias
0	25
1	28
2	47
3	65
4	86
5	121
6	190
7	290
8	362

El diagrama de dispersión asociado es

Evolución de bacterias

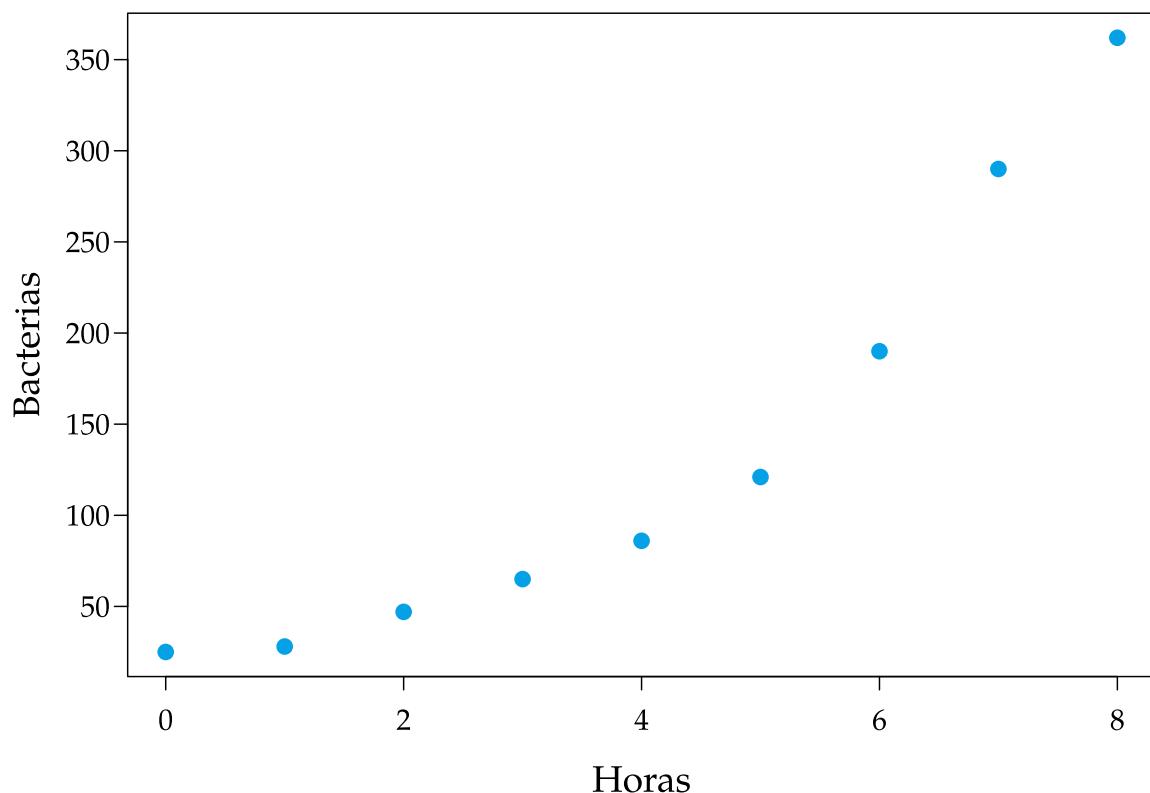


Figura 3.10: Diagrama de dispersión de la evolución de bacterias.

Si realizamos un ajuste lineal, obtenemos la siguiente recta de regresión

$$\text{Bacterias} = -30.18 + 41.27 \text{ Horas}, \text{ with } r^2 = 0.85.$$

Regresión lineal de Bacterias sobre Horas

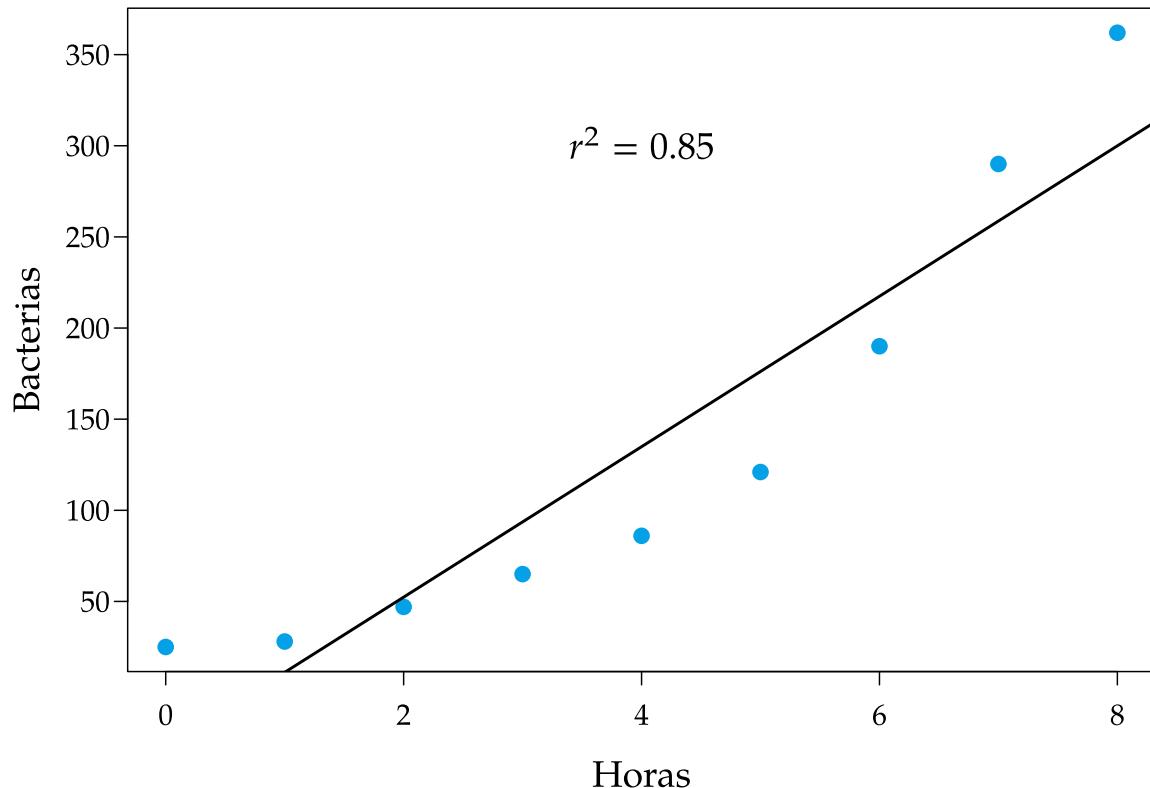


Figura 3.11: Regresión lineal de la evolución de un cultivo de bacterias.

¿Es un buen modelo?

Aunque el modelo lineal no es malo, de acuerdo al diagrama de dispersión es más lógico construir un modelo exponencial o cuadrático.

Para construir el modelo exponencial $y = ae^{bx}$ hay que realizar la transformación $z = \log y$, es decir, aplicar el logaritmo a la variable dependiente.

Horas	Bacterias	$\log(\text{Bacterias})$
0	25	3.22
1	28	3.33
2	47	3.85
3	65	4.17
4	86	4.45
5	121	4.80
6	190	5.25
7	290	5.67
8	362	5.89

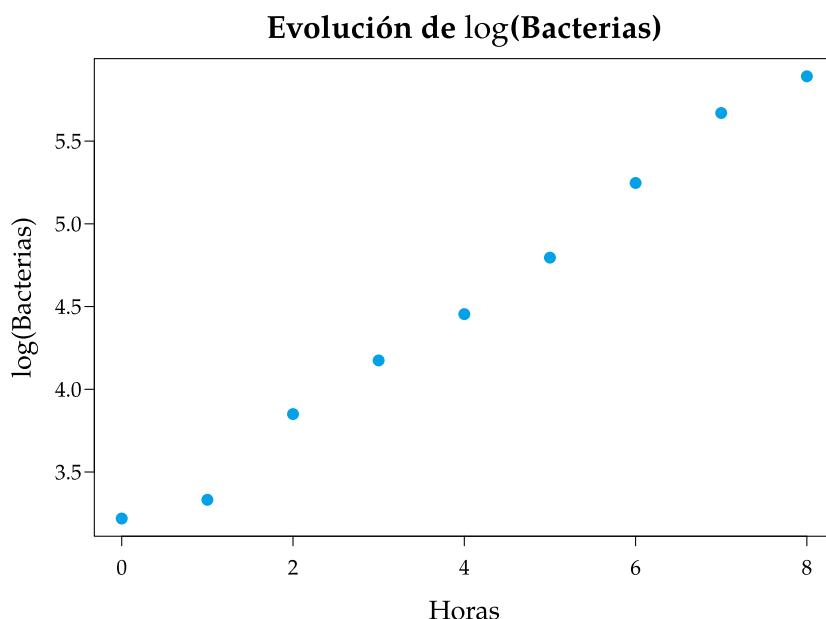


Figura 3.12: Diagrama de dispersión de la evolución del logaritmo de las bacterias de un cultivo.

Ahora sólo queda calcular la recta de regresión del logaritmo de Bacterias sobre Horas

$$\text{Log Bacterias} = 3.107 + 0.352 \text{ Horas.}$$

Y, deshaciendo el cambio de variable, se obtiene el modelo exponencial

$$\text{Bacterias} = e^{3.107+0.352 \text{ Horas}}, \text{ con } r^2 = 0.99.$$

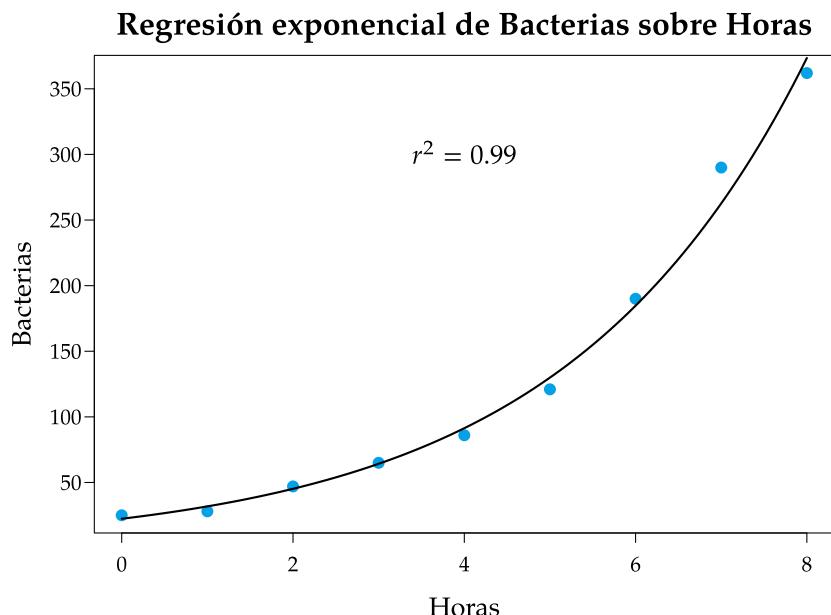


Figura 3.13: Regresión exponencial de la evolución de las bacterias de un cultivo.

Como se puede apreciar, el modelo exponencial se ajusta mucho mejor que el modelo lineal.

3.5 Riesgos de la regresión

3.5.1 La falta de ajuste no significa independencia

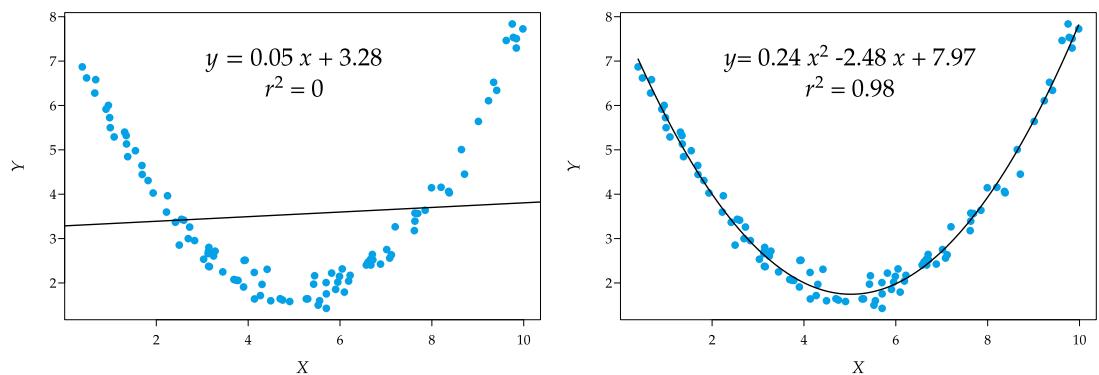
Es importante señalar que cada modelo de regresión tiene su propio coeficiente de determinación.

⚠ Advertencia

Así, un coeficiente de determinación cercano a cero significa que no existe relación entre las variables del tipo planteado por el modelo, pero *eso no quiere decir que las variables sean independientes*, ya que puede existir relación de otro tipo.

3.5.2 Datos atípicos en regresión

Los *datos atípicos* en un estudio de regresión son los puntos que claramente no siguen la tendencia del resto de los puntos en el diagrama de dispersión, incluso si los valores del par no se pueden considerar atípicos para cada variable por separado.



(a) Modelo de regresión lineal en una relación cuadrática.
 (b) Modelo de regresión cuadrático en una relación cuadrática.

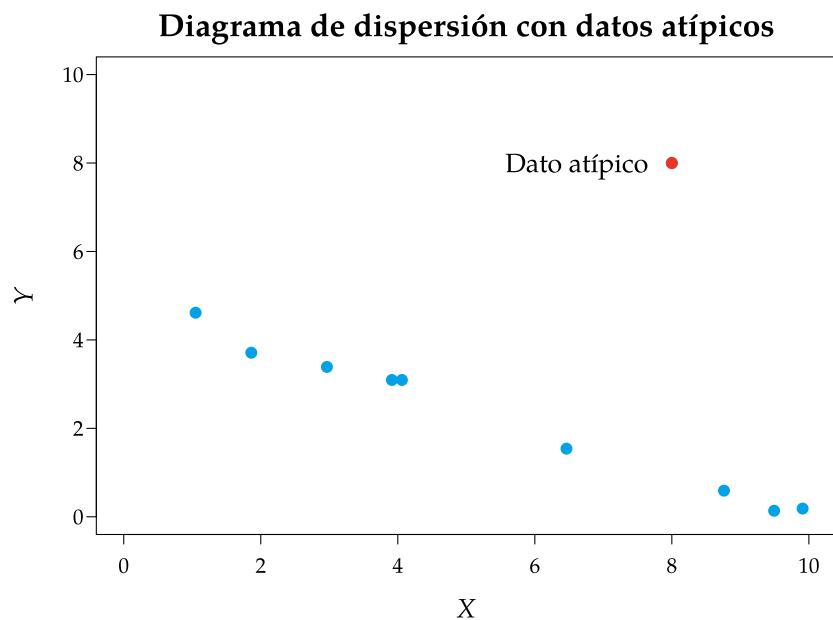
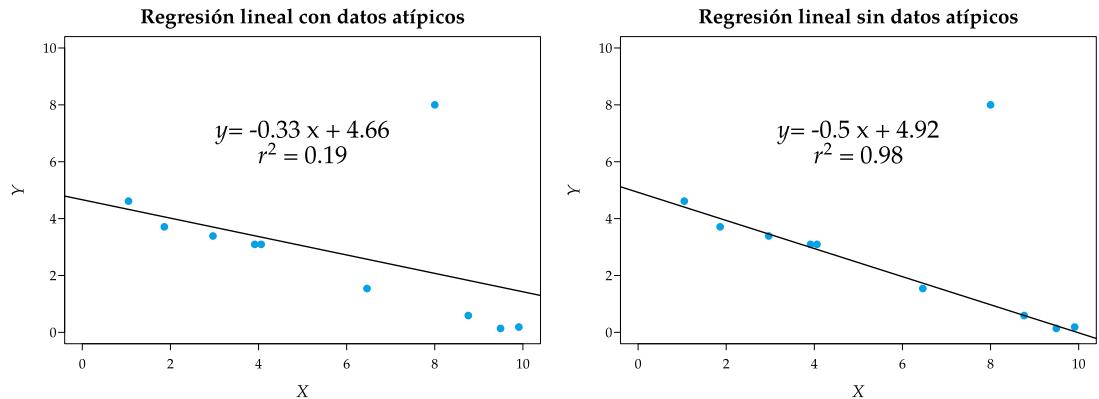


Figura 3.15: Diagrama de dispersión con un dato atípico.

⚠️ Advertencia

Los datos atípicos en regresión suelen provocar cambios drásticos en el ajuste de los modelos de regresión, y por tanto, habrá que tener mucho cuidado con ellos.



(a) Modelo de regresión lineal con datos atípicos. (b) Modelo de regresión lineal sin datos atípicos.

3.5.3 La paradoja de Simpson

A veces, una tendencia desaparece o incluso se revierte cuando se divide la muestra en grupos de acuerdo a una variable cualitativa que está relacionada con la variable dependiente. Esto se conoce como la *paradoja de Simpson*.

:::{#exm-paradoja-simpson} El siguiente diagrama de dispersión muestra una relación inversa entre entre las horas de estudio preparando un examen y la nota del examen.

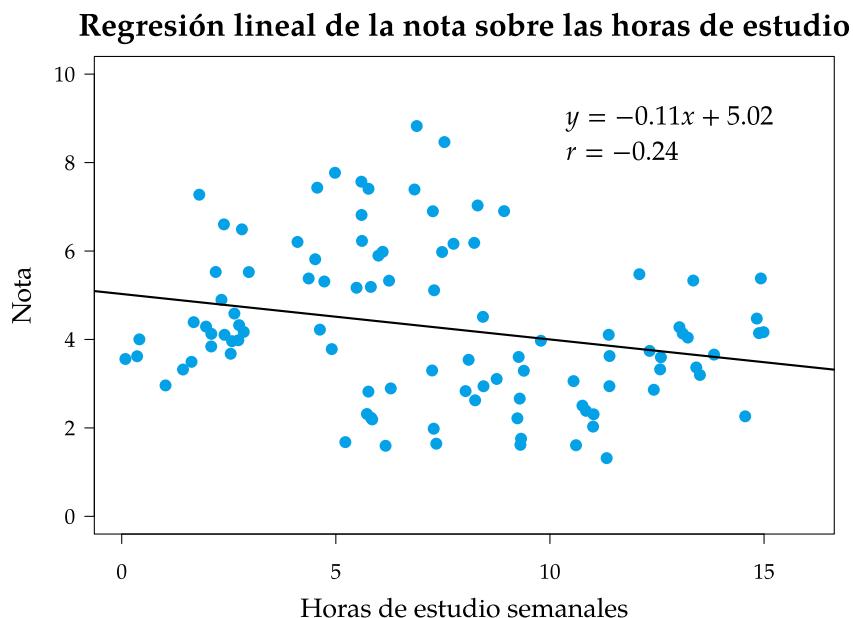


Figura 3.17: Paradoja de Simpson. Relación inversa entre las horas de estudio para un examen y la nota obtenida.

Pero si se divide la muestra en dos grupos (buenos y malos estudiantes) se obtienen diferentes tendencias y ahora la relación es directa, lo que tiene más lógica.

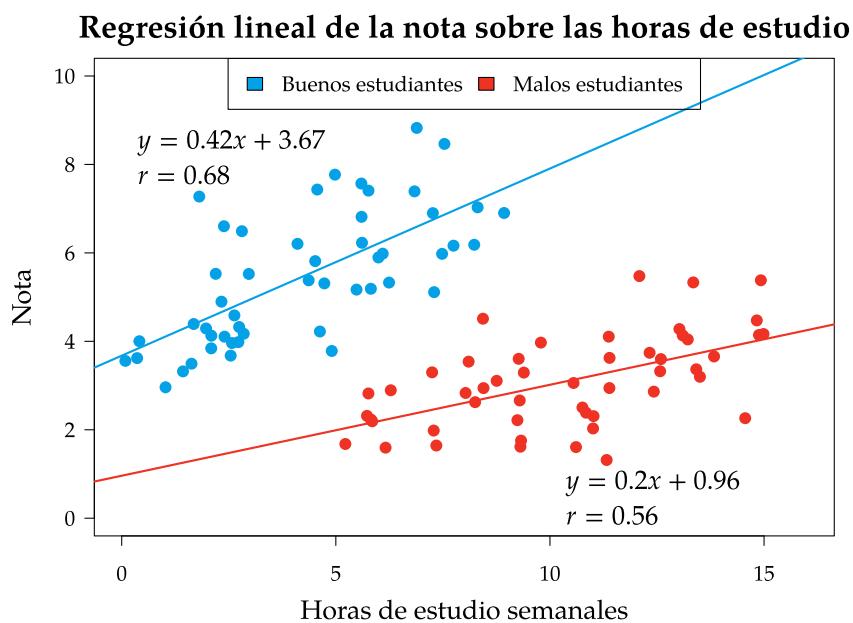


Figura 3.18: Paradoja de Simpson. Relación directa entre las horas de estudio para un examen y la nota obtenida.

4 Relaciones entre variables cualitativas

Los modelos de regresión vistos en el capítulo anterior solamente pueden aplicarse cuando las variables estudiadas son cuantitativas.

Cuando se desea estudiar la relación entre atributos, tanto ordinales como nominales, es necesario recurrir a otro tipo de medidas de relación o de asociación. En este capítulo veremos tres de ellas:

- Coeficiente de correlación de Spearman.
- Coeficiente chi-cuadrado.
- Coeficiente de contingencia.

4.1 Relación entre atributos ordinales

Cuando se quiere estudiar la relación entre dos atributos ordinales, o entre un atributo ordinal y una variable cuantitativa, es importante tener en cuenta el orden de las categorías. En estos casos se puede utilizar el siguiente coeficiente.

4.1.1 Coeficiente de correlación de Spearman

Cuando se tengan atributos ordinales es posible ordenar sus categorías y asignarles valores ordinales, de manera que se puede calcular el coeficiente de correlación lineal entre estos valores ordinales.

Esta medida de relación entre el orden que ocupan las categorías de dos atributos ordinales se conoce como *coeficiente de correlación de Spearman*.

Definición 4.1 (Coeficiente de correlación de Spearman). Dada una muestra de n individuos en los que se han medido dos atributos ordinales X e Y , el coeficiente de correlación de Spearman se define como

$$r_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

donde d_i es la diferencia entre el valor ordinal de X y el valor ordinal de Y del individuo i .

! Importante

Como el coeficiente de correlación de Spearman es en el fondo el coeficiente de correlación lineal aplicado a los órdenes, se tiene que

$$-1 \leq r_s \leq 1,$$

i Interpretación

- Si $r_s = 0$ entonces no existe relación entre los atributos ordinales.
- Si $r_s = 1$ entonces los órdenes de los atributos coinciden y existe una relación directa perfecta.
- Si $r_s = -1$ entonces los órdenes de los atributos están invertidos y existe una relación inversa perfecta.

En general, cuanto más cerca de 1 o -1 esté r_s , mayor será la relación entre los atributos, y cuanto más cerca de 0, menor será la relación.

Ejemplo 4.1. Una muestra de 5 alumnos realizaron dos tareas diferentes X e Y , y se ordenaron de acuerdo a la destreza que manifestaron en cada tarea:

Alumnos	X	Y	d_i	d_i^2
Alumno 1	2	3	-1	1
Alumno 2	5	4	1	1
Alumno 3	1	2	-1	1
Alumno 4	3	1	2	4
Alumno 5	4	5	-1	1
\sum			0	8

El coeficiente de correlación de Spearman para esta muestra es

$$r_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} = 1 - \frac{6 \cdot 8}{5(5^2 - 1)} = 0.6.$$

Esto indica que existe bastante relación directa entre las destrezas manifestadas en ambas tareas.

Ejemplo 4.2 (Empates). Cuando hay empates en el orden de las categorías se atribuye a cada valor empatado la media aritmética de los valores ordinales que hubieran ocupado esos individuos en caso de no haber estado empatados.

Si en el ejemplo anterior los alumnos 4 y 5 se hubiesen comportado igual en la primera tarea y los alumnos 3 y 4 se hubiesen comportado igual en la segunda tarea, entonces se tendría

Alumnos	X	Y	d_i	d_i^2
Alumno 1	2	3	-1	1
Alumno 2	5	4	1	1
Alumno 3	1	1.5	-0.5	0.25
Alumno 4	3.5	1.5	2	4
Alumno 5	3.5	5	-1.5	2.25
\sum			0	8.5

El coeficiente de correlación de Spearman para esta muestra es

$$r_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} = 1 - \frac{6 \cdot 8.5}{5(5^2 - 1)} = 0.58.$$

4.2 Relación entre atributos nominales

Cuando se quiere estudiar la relación entre atributos nominales no tiene sentido calcular el coeficiente de correlación de Spearman ya que las categorías no pueden ordenarse.

Para estudiar la relación entre atributos nominales se utilizan medidas basadas en las frecuencias de la tabla de frecuencias bidimensional, que para atributos se suele llamar *tabla de contingencia*.

Ejemplo 4.3. En un estudio para ver si existe relación entre el sexo y el hábito de fumar se ha tomado una muestra de 100 personas. La tabla de contingencia resultante es

Sexo\Fuma	Si	No	n_i
Mujer	12	28	40
Hombre	26	34	60
n_j	38	62	100

Si el hábito de fumar fuese independiente del sexo, la proporción de fumadores en mujeres y hombres sería la misma.

4.2.1 Frecuencias teóricas o esperadas

En general, dada una tabla de contingencia para dos atributos X e Y ,

$X \setminus Y$	y_1	\cdots	y_j	\cdots	y_q	n_x
x_1	n_{11}	\cdots	n_{1j}	\cdots	n_{1q}	n_{x_1}
\vdots	\vdots	\ddots	\vdots	\ddots	\vdots	\vdots
x_i	n_{i1}	\cdots	n_{ij}	\cdots	n_{iq}	n_{x_i}
\vdots	\vdots	\ddots	\vdots	\ddots	\vdots	\vdots
x_p	n_{p1}	\cdots	n_{pj}	\cdots	n_{pq}	n_{x_p}
n_y	n_{y_1}	\cdots	n_{y_j}	\cdots	n_{y_q}	n

si X e Y fuesen independientes, para cualquier valor y_j se tendría

$$\frac{n_{1j}}{n_{x_1}} = \frac{n_{2j}}{n_{x_2}} = \cdots = \frac{n_{pj}}{n_{x_p}} = \frac{n_{1j} + \cdots + n_{pj}}{n_{x_1} + \cdots + n_{x_p}} = \frac{n_{y_j}}{n},$$

de donde se deduce que

$$n_{ij} = \frac{n_{x_i} n_{y_j}}{n}.$$

A esta última expresión se le llama *frecuencia teórica* o *frecuencia esperada* del par (x_i, y_j) .

4.2.2 Coeficiente chi-cuadrado χ^2

Es posible estudiar la relación entre dos atributos X e Y comparando las frecuencias reales con las esperadas.

Definición 4.2 (Coeficiente Chi-cuadrado χ^2). Dada una muestra de tamaño n en la que se han medido dos atributos X e Y , se define el coeficiente χ^2 como

$$\chi^2 = \sum_{i=1}^p \sum_{j=1}^q \frac{\left(n_{ij} - \frac{n_{x_i} n_{y_j}}{n} \right)^2}{\frac{n_{x_i} n_{y_j}}{n}},$$

donde p es el número de categorías de X y q el número de categorías de Y .

! Importante

Por ser suma de cuadrados, se cumple que

$$\chi^2 \geq 0.$$

i Interpretación

$\chi^2 = 0$ cuando los atributos son independientes, y crece a medida que aumenta la dependencia entre las variables.

Ejemplo 4.4. Siguiendo con el ejemplo anterior, a partir de la tabla de contingencia

Sexo\Fuma	Si	No	n_i
Mujer	12	28	40
Hombre	26	34	60
n_j	38	62	100

se obtienen las siguientes frecuencias esperadas

Sexo\Fuma	Si	No	n_i
Mujer	$\frac{40 \cdot 38}{100} = 15.2$	$\frac{40 \cdot 62}{100} = 24.8$	40
Hombre	$\frac{60 \cdot 38}{100} = 22.8$	$\frac{60 \cdot 62}{100} = 37.2$	60
n_j	38	62	100

y el coeficiente χ^2 vale

$$\chi^2 = \frac{(12 - 15.2)^2}{15.2} + \frac{(28 - 24.8)^2}{24.8} + \frac{(26 - 22.8)^2}{22.8} + \frac{(34 - 37.2)^2}{37.2} = 1.81.$$

Esto indica que no existe gran relación entre el sexo y el hábito de fumar.

4.2.3 Coeficiente de contingencia

El coeficiente χ^2 depende del tamaño muestral, ya que al multiplicar por una constante las frecuencias de todas las casillas, su valor queda multiplicado por dicha constante, lo que podría llevarnos al equívoco de pensar que ha aumentado la relación, incluso cuando las proporciones se mantienen. En consecuencia el valor de χ^2 no está acotado superiormente y resulta difícil de interpretar.

Para evitar estos problemas se suele utilizar el siguiente estadístico.

Definición 4.3 (Coeficiente de contingencia). Dada una muestra de tamaño n en la que se han medido dos atributos X e Y , se define el *coeficiente de contingencia* como

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}}$$

! Importante

De la definición anterior se deduce que

$$0 \leq C \leq 1,$$

i Interpretación

$C = 0$ cuando las variables son independientes, y crece a medida que aumenta la relación.

! Advertencia

Aunque C nunca puede llegar a valer 1, se puede demostrar que para tablas de contingencia con k filas y k columnas, el valor máximo que puede alcanzar C es $\sqrt{(k-1)/k}$.

Ejemplo 4.5. En el ejemplo anterior el coeficiente de contingencia vale

$$C = \sqrt{\frac{1.81}{1.81 + 100}} = 0.13.$$

Como se trata de una tabla de contingencia de 2×2 , el valor máximo que podría tomar el coeficiente de contingencia es $\sqrt{(2-1)/2} = \sqrt{1/2} = 0.707$, y como 0.13 está bastante lejos de este valor, se puede concluir que no existe demasiada relación entre el hábito de fumar y el sexo.

5 Probabilidad

La estadística descriptiva permite describir el comportamiento y las relaciones entre las variables en la muestra, pero no permite sacar conclusiones sobre el resto de la población.

Ha llegado el momento de dar el salto de la muestra a la población y pasar de la estadística descriptiva a la inferencia estadística, y el puente que lo permite es la **Teoría de la Probabilidad**.

Hay que tener en cuenta que el conocimiento que se puede obtener de la población a partir de la muestra es limitado, y que para obtener conclusiones válidas para la población la muestra debe ser representativa de esta. Por esta razón, para garantizar la representatividad de la muestra, esta debe extraerse *aleatoriamente*, es decir, al *azar*.

La teoría de la probabilidad precisamente se encarga de controlar ese azar para saber hasta qué punto son fiables las conclusiones obtenidas a partir de una muestra.

5.1 Experimentos y sucesos aleatorios

El estudio de una característica en una población se realiza a través de experimentos aleatorios.

Definición 5.1 (Experimento aleatorio). Un *experimento aleatorio* es un experimento que cumple dos condiciones:

1. El conjunto de posibles resultados es conocido.
2. No se puede predecir con absoluta certeza el resultado del experimento.

Ejemplo 5.1. Un ejemplo típico de experimentos aleatorios son los juegos de azar. El lanzamiento de un dado, por ejemplo, es un experimento aleatorio ya que:

1. Se conoce el conjunto posibles de resultados $\{1, 2, 3, 4, 5, 6\}$.
2. Antes de lanzar el dado, es imposible predecir con absoluta certeza el valor que saldrá.

Otro ejemplo de experimento aleatorio sería la selección de un individuo de una población al azar y la determinación de su grupo sanguíneo.

En general, la obtención de cualquier muestra mediante procedimientos aleatorios será un experimento aleatorio.

Definición 5.2 (Espacio muestral). Al conjunto Ω de todos los posibles resultados de un experimento aleatorio se le llama *espacio muestral*.

Ejemplo 5.2. Algunos ejemplos de espacios muestrales son:

- Lanzamiento de una moneda: $\Omega = \{c, x\}$.
- Lanzamiento de un dado: $\Omega = \{1, 2, 3, 4, 5, 6\}$.
- Grupo sanguíneo de un individuo seleccionado al azar: $\Omega = \{A, B, AB, 0\}$.
- Estatura de un individuo seleccionado al azar: $\Omega = \mathbb{R}^+$.

En experimentos donde se mide más de una variable, la determinación del espacio muestral puede resultar compleja. En tales casos es recomendable utilizar un para construir el espacio muestral.

En un diagrama de árbol cada variable se representa en un nivel del árbol y cada posible valor de la variable como una rama.

Ejemplo 5.3. El siguiente diagrama de árbol representa el espacio muestral de un experimento aleatorio en el que se mide el sexo y el grupo sanguíneo de un individuo al azar.

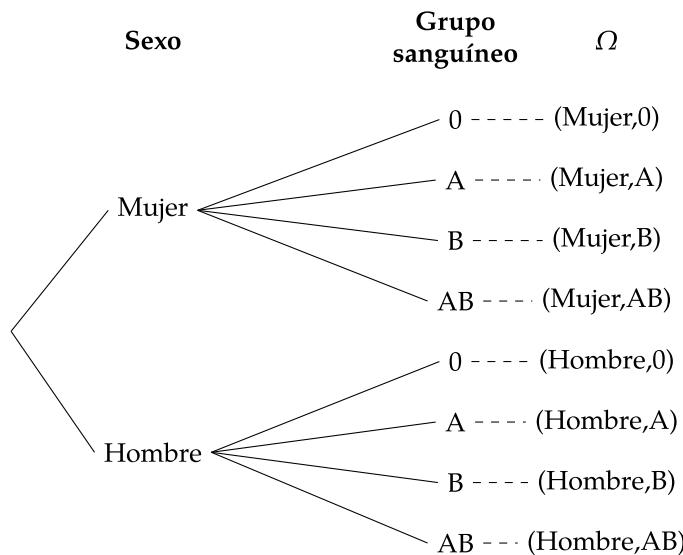


Figura 5.1: Diagrama de árbol del espacio muestral del sexo y el grupo sanguíneo.

Definición 5.3 (Suceso aleatorio). Un *suceso aleatorio* es cualquier subconjunto del espacio muestral Ω de un experimento aleatorio.

Existen distintos tipos de sucesos:

- **Suceso imposible:** Es el suceso vacío \emptyset . Este suceso nunca ocurre.
- **Sucesos elementales:** Son los sucesos formados por un solo elemento.
- **Sucesos compuestos:** Son los sucesos formados por dos o más elementos.
- **Suceso seguro:** Es el suceso que contiene el propio espacio muestral Ω . Este suceso siempre ocurre.

Ejemplo 5.4. En el experimento aleatorio del lanzamiento de un dado, con espacio muestral $\Omega = \{1, 2, 3, 4, 5, 6\}$, el subconjunto $\{2, 4, 6\}$ es un suceso aleatorio que se cumple cuando sale un número par, y el subconjunto $\{1, 2, 3, 4\}$ es un suceso aleatorio que se cumple cuando sale un número menor que 5.

5.1.1 Espacio de sucesos

Definición 5.4 (Espacio de sucesos). Dado un espacio muestral Ω de un experimento aleatorio, el conjunto formado por todos los posibles sucesos de Ω se llama *espacio de sucesos de Ω* y se denota $\mathcal{P}(\Omega)$.

Ejemplo 5.5. Dado el espacio muestral $\Omega = \{a, b, c\}$, su espacio de sucesos es

$$\mathcal{P}(\Omega) = \{\emptyset, \{a\}, \{b\}, \{c\}, \{a, b\}, \{a, c\}, \{b, c\}, \{a, b, c\}\}$$

Puesto que los sucesos son conjuntos, por medio de la teoría de conjuntos se pueden definir las siguientes operaciones entre sucesos:

- Unión.
- Intersección.
- Complementario.
- Diferencia.

5.1.2 Unión de sucesos

Definición 5.5 (Suceso unión). Dados dos sucesos $A, B \subseteq \Omega$, se llama *suceso unión* de A y B , y se denota $A \cup B$, al suceso formado por los elementos de A junto a los elementos de B , es decir,

$$A \cup B = \{x \mid x \in A \text{ o } x \in B\}.$$

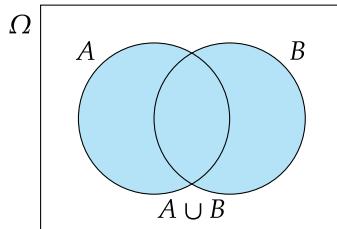


Figura 5.2: Unión de dos sucesos.

El suceso unión $A \cup B$ ocurre siempre que ocurre A o B .

Ejemplo 5.6. Dado el espacio muestral correspondiente al lanzamiento de un dado $\Omega = \{1, 2, 3, 4, 5, 6\}$ y los sucesos $A = \{2, 4, 6\}$ y $B = \{1, 2, 3, 4\}$, la unión de A y B es $A \cup B = \{1, 2, 3, 4, 6\}$.

5.1.3 Intersección de sucesos

Definición 5.6 (Suceso intersección). Dados dos sucesos $A, B \subseteq \Omega$, se llama *suceso intersección* de A y B , y se denota $A \cap B$, al suceso formado por los elementos comunes de A y B , es decir,

$$A \cap B = \{x \mid x \in A \text{ y } x \in B\}.$$

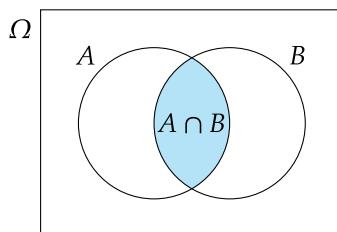


Figura 5.3: Intersección de dos sucesos.

El suceso intersección $A \cap B$ ocurre siempre que ocurren A y B .

Diremos que dos sucesos son **incompatibles** si su intersección es vacía.

Ejemplo 5.7. Dado el espacio muestral correspondiente al lanzamiento de un dado $\Omega = \{1, 2, 3, 4, 5, 6\}$ y los sucesos $A = \{2, 4, 6\}$ y $B = \{1, 2, 3, 4\}$, la intersección de A y B es $A \cap B = \{2, 4\}$, y por tanto, se trata de sucesos compatibles. Sin embargo, el suceso $C = \{1, 3\}$ es incompatible con A ya que $A \cap C = \emptyset$.

5.1.4 Contrario de un suceso

Definición 5.7 (Suceso contrario). Dado suceso $A \subseteq \Omega$, se llama *suceso contrario* o *complementario* de A , y se denota \bar{A} , al suceso formado por los elementos de Ω que no pertenecen a A , es decir,

$$\bar{A} = \{x \mid x \notin A\}.$$

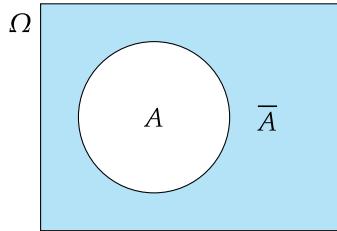


Figura 5.4: Contrario de un suceso.

El suceso contrario \bar{A} ocurre siempre que no ocurre A .

Ejemplo 5.8. Dado el espacio muestral correspondiente al lanzamiento de un dado $\Omega = \{1, 2, 3, 4, 5, 6\}$ y los sucesos $A = \{2, 4, 6\}$ y $B = \{1, 2, 3, 4\}$, el contrario de A es $\bar{A} = \{1, 3, 5\}$.

5.1.5 Diferencia de sucesos

Definición 5.8 (Suceso diferencia). Dados dos sucesos $A, B \subseteq \Omega$, se llama *suceso diferencia* de A y B , y se denota $A - B$, al suceso formado por los elementos de A que no pertenecen a B , es decir,

$$A - B = \{x \mid x \in A \text{ y } x \notin B\} = A \cap \bar{B}.$$

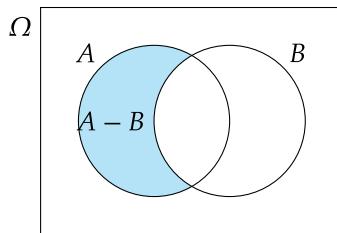


Figura 5.5: Diferencia de sucesos.

El suceso diferencia $A - B$ ocurre siempre que ocurre A pero no ocurre B , y también puede expresarse como $A \cap \bar{B}$.

Ejemplo 5.9. Dado el espacio muestral correspondiente al lanzamiento de un dado $\Omega = \{1, 2, 3, 4, 5, 6\}$ y los sucesos $A = \{2, 4, 6\}$ y $B = \{1, 2, 3, 4\}$, la diferencia de A y B es $A - B = \{6\}$, y la diferencia de B y A es $B - A = \{1, 3\}$.

5.1.6 Álgebra de sucesos

Dados los sucesos $A, B, C \in \mathcal{P}(\Omega)$, se cumplen las siguientes propiedades:

1. $A \cup A = A, A \cap A = A$ (idempotencia).
2. $A \cup B = B \cup A, A \cap B = B \cap A$ (comutativa).
3. $(A \cup B) \cup C = A \cup (B \cup C), (A \cap B) \cap C = A \cap (B \cap C)$ (asociativa).
4. $(A \cup B) \cap C = (A \cap C) \cup (B \cap C), (A \cap B) \cup C = (A \cup C) \cap (B \cup C)$ (distributiva).
5. $A \cup \emptyset = A, A \cap E = A$ (elemento neutro).
6. $A \cup E = E, A \cap \emptyset = \emptyset$ (elemento absorbente).
7. $A \cup \bar{A} = E, A \cap \bar{A} = \emptyset$ (elemento simétrico complementario).
8. $\bar{\bar{A}} = A$ (doble contrario).
9. $\overline{A \cup B} = \bar{A} \cap \bar{B}, \overline{A \cap B} = \bar{A} \cup \bar{B}$ (leyes de Morgan).
10. $A \cap B \subseteq A \cup B$.

5.2 Definición de probabilidad

5.2.1 Definición clásica de probabilidad

Definición 5.9 (Probabilidad - Laplace). Dado un espacio muestral Ω de un experimento aleatorio donde todos los elementos de Ω son equiprobables, la *probabilidad* de un suceso $A \subseteq \Omega$ es el cociente entre el número de elementos de A y el número de elementos de Ω

$$P(A) = \frac{|A|}{|\Omega|} = \frac{\text{nº casos favorables a } A}{\text{nº casos posibles}}$$

Esta definición es ampliamente utilizada, aunque tiene importantes restricciones:

- Es necesario que todos los elementos del espacio muestral tengan la misma probabilidad de ocurrir (*equiprobabilidad*).
- No puede utilizarse con espacios muestrales infinitos, o de los que no se conoce el número de casos posibles.

Precaución

Esto no se cumple en muchos experimentos aleatorios reales.

Ejemplo 5.10. Dado el espacio muestral correspondiente al lanzamiento de un dado $\Omega = \{1, 2, 3, 4, 5, 6\}$ y el suceso $A = \{2, 4, 6\}$, la probabilidad de A es

$$P(A) = \frac{|A|}{|\Omega|} = \frac{3}{6} = 0.5.$$

Sin embargo, si se considera el espacio muestral correspondiente a observar el grupo sanguíneo de un individuo al azar, $\Omega = \{O, A, B, AB\}$, no se puede usar la definición clásica de probabilidad para calcular la probabilidad de que tenga grupo sanguíneo A ,

$$P(A) \neq \frac{|A|}{|\Omega|} = \frac{1}{4} = 0.25,$$

ya que los grupos sanguíneos no son igualmente probables en las poblaciones humanas.

5.2.2 Definición frecuentista de probabilidad

Teorema 5.1 (Ley de los grandes números). *Cuando un experimento aleatorio se repite un gran número de veces, las frecuencias relativas de los sucesos del experimento tienden a estabilizarse en torno a cierto número, que es precisamente su probabilidad.*

De acuerdo al teorema anterior, podemos dar la siguiente definición

Definición 5.10 (Probabilidad frecuentista). Dado un espacio muestral Ω de un experimento aleatorio reproducible, la *probabilidad* de un suceso $A \subseteq \Omega$ es la frecuencia relativa del suceso A en infinitas repeticiones del experimento

$$P(A) = \lim_{n \rightarrow \infty} \frac{n_A}{n}$$

Aunque esta definición es muy útil en experimentos científicos reproducibles, también tiene serios inconvenientes, ya que

- Sólo se calcula una aproximación de la probabilidad real.
- La repetición del experimento debe ser en las mismas condiciones.

Ejemplo 5.11. Dado el espacio muestral correspondiente al lanzamiento de una moneda $\Omega = \{C, X\}$, si después de lanzar la moneda 100 veces obtenemos 54 caras, entonces la probabilidad de C es aproximadamente

$$P(C) = \frac{n_C}{n} = \frac{54}{100} = 0.54.$$

Si se considera el espacio muestral correspondiente a observar el grupo sanguíneo de un individuo al azar, $\Omega = \{O, A, B, AB\}$, si se toma una muestra aleatoria de 1000 personas y se observa que 412 tienen grupo sanguíneo A , entonces la probabilidad del grupo sanguíneo A es aproximadamente

$$P(A) = \frac{n_A}{n} = \frac{412}{1000} = 0.412.$$

5.2.3 Definición axiomática de probabilidad

Definición 5.11 (Probabilidad - Kolmogórov). Dado un espacio muestral Ω de un experimento aleatorio, una función de *probabilidad* es una aplicación que asocia a cada suceso $A \subseteq \Omega$ un número real $P(A)$, conocido como probabilidad de A , que cumple los siguientes axiomas:

1. La probabilidad de un suceso cualquiera es positiva o nula,

$$P(A) \geq 0.$$

2. La probabilidad del suceso seguro es igual a la unidad,

$$P(\Omega) = 1.$$

3. La probabilidad de la unión de dos sucesos incompatibles ($A \cap B = \emptyset$) es igual a la suma de las probabilidades de cada uno de ellos,

$$P(A \cup B) = P(A) + P(B).$$

Teorema 5.2. Si P es una función de probabilidad de un espacio muestral Ω , entonces para cualesquiera sucesos $A, B \in \Omega$, se cumple

1. $P(\bar{A}) = 1 - P(A)$.
2. $P(\emptyset) = 0$.
3. Si $A \subseteq B$ entonces $P(A) \leq P(B)$.
4. $P(A) \leq 1$.
5. $P(A - B) = P(A) - P(A \cap B)$.

6. Si A y B son sucesos compatibles, es decir, su intersección no es vacía, entonces

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

7. Si el suceso A está compuesto por los sucesos elementales e_1, e_2, \dots, e_n , entonces

$$P(A) = \sum_{i=1}^n P(e_i).$$

i Demostración

Prueba.

1. $\bar{A} = \Omega \Rightarrow P(A \cup \bar{A}) = P(\Omega) \Rightarrow P(A) + P(\bar{A}) = 1 \Rightarrow P(\bar{A}) = 1 - P(A).$
2. $\emptyset = \bar{\Omega} \Rightarrow P(\emptyset) = P(\bar{\Omega}) = 1 - P(\Omega) = 1 - 1 = 0.$
3. $B = A \cup (B - A)$. Como A y $B - A$ son incompatibles, $P(B) = P(A \cup (B - A)) = P(A) + P(B - A) \geq P(A).$

Si pensamos en probabilidades como áreas, es fácil de ver gráficamente,

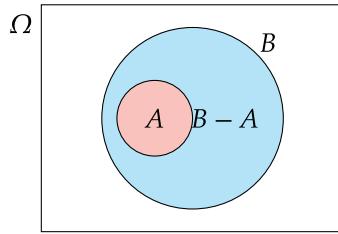


Figura 5.6: Probabilidad de un suceso incluido en otro.

4. $A \subseteq \Omega \Rightarrow P(A) \leq P(\Omega) = 1.$
5. $A = (A - B) \cup (A \cap B)$. Como $A - B$ y $A \cap B$ son incompatibles, $P(A) = P(A - B) + P(A \cap B) \Rightarrow P(A - B) = P(A) - P(A \cap B).$

Si pensamos en probabilidades como áreas, es fácil de ver gráficamente,

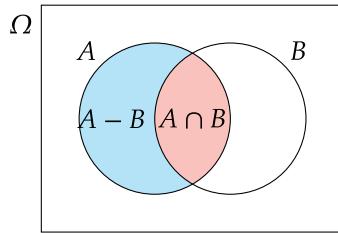


Figura 5.7: Probabilidad de la diferencia de dos sucesos.

6. $A \cup B = (A - B) \cup (B - A) \cup (A \cap B)$. Como $A - B$, $B - A$ y $A \cap B$ son incompatibles, $P(A \cup B) = P(A - B) + P(B - A) + P(A \cap B) = P(A) - P(A \cap B) + P(B) - P(A \cap B) + P(A \cap B) = P(A) + P(B) - P(A \cup B)$.

Si pensamos en probabilidades como áreas, es fácil de ver gráficamente,

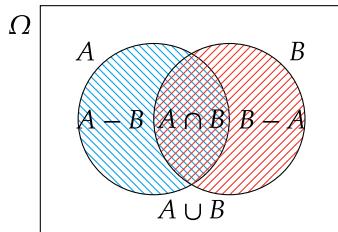


Figura 5.8: Probabilidad de la unión de dos sucesos.

7. $A = \{e_1, \dots, e_n\} = \{e_1\} \cup \dots \cup \{e_n\} \Rightarrow P(A) = P(\{e_1\} \cup \dots \cup \{e_n\}) = P(\{e_1\}) + \dots + P(\{e_n\})$.

□

5.2.4 Interpretación de la probabilidad

Como ha quedado claro en los axiomas anteriores, la probabilidad de un evento A es un número real $P(A)$ que está siempre entre 0 y 1.

En cierto modo, este número expresa la verosimilitud del evento, es decir, la confianza que hay en que ocurra A en el experimento. Por tanto, también nos da una medida de la incertidumbre sobre el suceso.

- La mayor incertidumbre corresponde a $P(A) = 0.5$ (Es tan probable que ocurra A como que no ocurra).
- La menor incertidumbre corresponde a $P(A) = 1$ (A sucederá con absoluta certeza) y $P(A) = 0$ (A no sucederá con absoluta certeza).

Cuando $P(A)$ está más próximo a 0 que a 1, la confianza en que no ocurra A es mayor que la de que ocurra A . Por el contrario, cuando $P(A)$ está más próximo a 1 que a 0, la confianza en que ocurra A es mayor que la de que no ocurra A .

5.3 Probabilidad condicionada

5.3.1 Experimentos condicionados

En algunas ocasiones, es posible que tengamos alguna información sobre el experimento antes de su realización. Habitualmente esa información se da en forma de un suceso B del mismo espacio muestral que sabemos que es cierto antes de realizar el experimento.

En tal caso se dice que el suceso B es un suceso *condicionante*, y la probabilidad de otro suceso A se conoce como y se expresa

$$P(A|B).$$

Esto debe leerse como *probabilidad de A dado B* o *probabilidad de A bajo la condición de B* .

Los condicionantes suelen cambiar el espacio muestral del experimento y por tanto las probabilidades de sus sucesos.

Ejemplo 5.12. Supongamos que tenemos una muestra de 100 hombres y 100 mujeres con las siguientes frecuencias

	No fumadores	Fumadores
Mujeres	80	20
Hombres	60	40

Entonces, usando la definición frecuentista de probabilidad, la probabilidad de que una persona elegida al azar sea fumadora es

$$P(\text{Fumadora}) = \frac{60}{200} = 0.3.$$

Sin embargo, si se sabe que la persona elegida es mujer, entonces la muestra se reduce a la primera fila, y la probabilidad de ser fumadora es

$$P(\text{Fumadora}|\text{Mujer}) = \frac{20}{100} = 0.2.$$

5.3.2 Probabilidad condicionada

Definición 5.12 (Probabilidad condicionada). Dado un espacio muestral Ω de un experimento aleatorio, y dos sucesos $A, B \subseteq \Omega$, la probabilidad de A *condicionada* por B es

$$P(A|B) = \frac{P(A \cap B)}{P(B)},$$

siempre y cuando, $P(B) \neq 0$.

Esta definición permite calcular probabilidades sin tener que alterar el espacio muestral original del experimento.

Ejemplo 5.13. En el ejemplo anterior

$$P(\text{Fumadora}|\text{Mujer}) = \frac{P(\text{Fumadora} \cap \text{Mujer})}{P(\text{Mujer})} = \frac{20/200}{100/200} = \frac{20}{100} = 0.2.$$

5.3.3 Probabilidad del suceso intersección

A partir de la definición de probabilidad condicionada es posible obtener la fórmula para calcular la probabilidad de la intersección de dos sucesos.

$$P(A \cap B) = P(A)P(B|A) = P(B)P(A|B).$$

Ejemplo 5.14. En una población hay un 30% de fumadores y se sabe que el 40% de los fumadores tiene cáncer de pulmón. La probabilidad de que una persona elegida al azar sea fumadora y tenga cáncer de pulmón es

$$P(\text{Fumadora} \cap \text{Cáncer}) = P(\text{Fumadora})P(\text{Cáncer}|\text{Fumadora}) = 0.3 \times 0.4 = 0.12.$$

5.3.4 Independencia de sucesos

En ocasiones, la ocurrencia del suceso condicionante no cambia la probabilidad original del suceso principal.

Definición 5.13 (Sucesos independientes). Dado un espacio muestral Ω de un experimento aleatorio, dos sucesos $A, B \subseteq \Omega$ son *independientes* si la probabilidad de A no se ve alterada al condicionar por B , y viceversa, es decir,

$$P(A|B) = P(A) \quad \text{and} \quad P(B|A) = P(B),$$

si $P(A) \neq 0$ y $P(B) \neq 0$.

Esto significa que la ocurrencia de uno evento no aporta información relevante para cambiar la incertidumbre sobre el otro.

Cuando dos eventos son independientes, la probabilidad de su intersección es igual al producto de sus probabilidades,

$$P(A \cap B) = P(A)P(B).$$

5.4 Espacio probabilístico

Definición 5.14 (Espacio probabilístico). Un *espacio probabilístico* de un experimento aleatorio es una terna (Ω, \mathcal{F}, P) donde

- Ω es el espacio muestral del experimento.
- \mathcal{F} es un conjunto de sucesos del experimento.
- P es una función de probabilidad.

Si conocemos la probabilidad de todos los elementos de Ω , entonces podemos calcular la probabilidad de cualquier suceso en \mathcal{F} y se puede construir fácilmente el espacio probabilístico.

Para determinar la probabilidad de cada suceso elemental se puede utilizar un diagrama de árbol, mediante las siguientes reglas:

1. Para cada nodo del árbol, etiquetar la rama que conduce hasta él con la probabilidad de que la variable en ese nivel tome el valor del nodo, condicionada por los sucesos correspondientes a sus nodos antecesores en el árbol.
2. La probabilidad de cada suceso elemental en las hojas del árbol es el producto de las probabilidades de las ramas que van desde la raíz a la hoja del árbol.

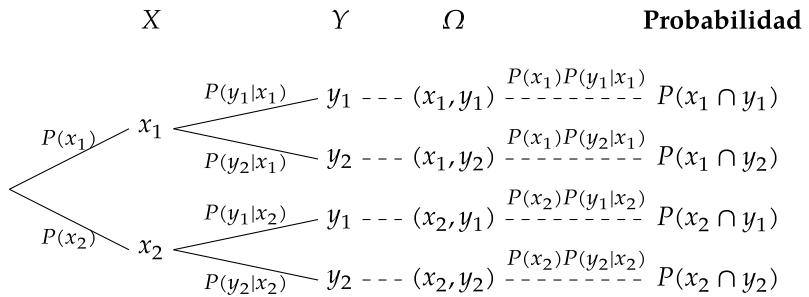


Figura 5.9: Diagrama de árbol de un espacio probabilístico.

5.4.1 Árboles de probabilidad con variables dependientes

Ejemplo 5.15. Sea una población en la que el 30% de las personas fuman, y que la incidencia del cáncer de pulmón en fumadores es del 40% mientras que en los no fumadores es del 10%.

El espacio probabilístico del experimento aleatorio que consiste en elegir una persona al azar y medir las variables Fumar y Cáncer de pulmón se muestra a continuación.

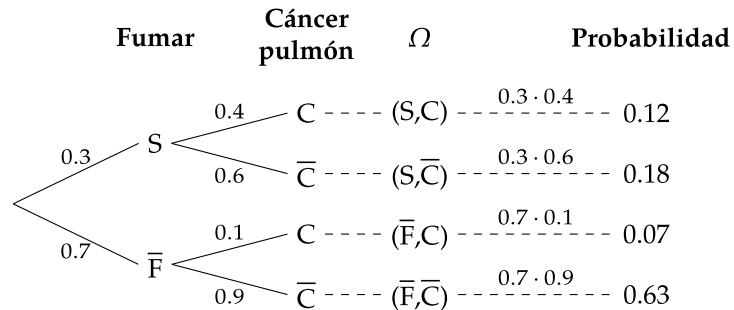


Figura 5.10: Diagrama de árbol del espacio probabilístico de fumar y tener cáncer de pulmón.

5.4.2 Árboles de probabilidad con variables independientes

Ejemplo 5.16. El árbol de probabilidad asociado al experimento aleatorio que consiste en el lanzamiento de dos monedas se muestra a continuación.

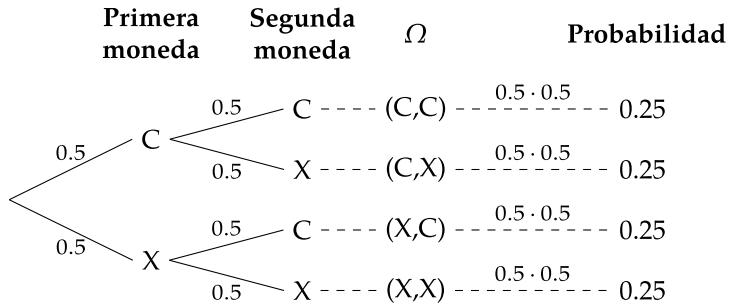


Figura 5.11: Diágrama de árbol del espacio probabilístico del lanzamiento de dos monedas.

Ejemplo 5.17. Dada una población en la que hay un 40% de hombres y un 60% de mujeres, el experimento aleatorio que consiste en tomar una muestra aleatoria de tres personas tiene el árbol de probabilidad que se muestra a continuación.

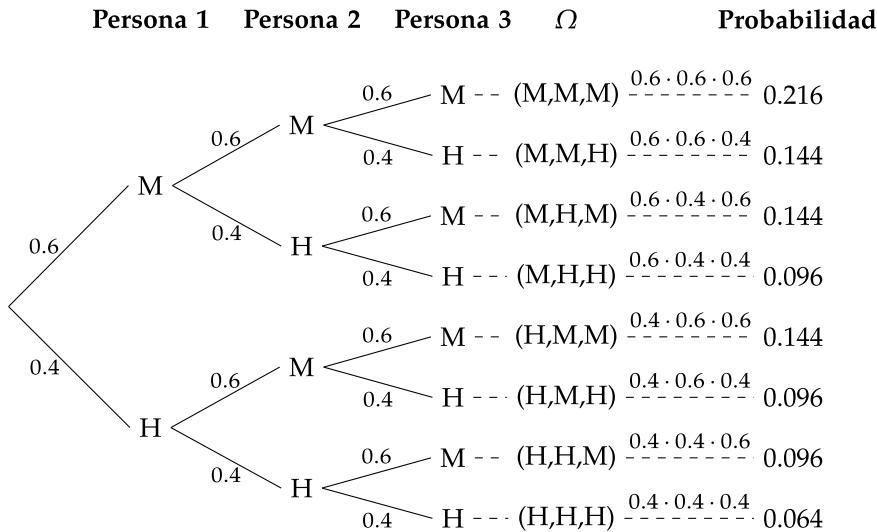


Figura 5.12: Diagrama de árbol del espacio probabilístico del sexo de tres individuos elegidos al azar.

5.5 Teorema de la probabilidad total

Definición 5.15 (Sistema completo de sucesos). Una colección de sucesos A_1, A_2, \dots, A_n de un mismo espacio muestral Ω es un *sistema completo* si cumple las siguientes condiciones:

1. La unión de todos es el espacio muestral: $A_1 \cup \dots \cup A_n = \Omega$.

2. Son incompatibles dos a dos: $A_i \cap A_j = \emptyset \forall i \neq j$.

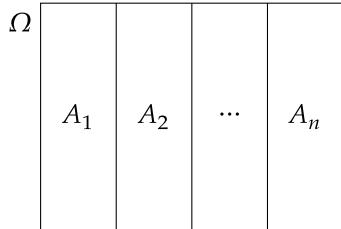


Figura 5.13: Partición del espacio muestral en un sistema completo de sucesos.

En realidad un sistema completo de sucesos es una partición del espacio muestral de acuerdo a algún atributo, como por ejemplo el sexo o el grupo sanguíneo.

5.5.1 Teorema de la probabilidad total

Conocer las probabilidades de un determinado suceso en cada una de las partes de un sistema completo puede ser útil para calcular su probabilidad.

Teorema 5.3 (Probabilidad total). *Dado un sistema completo de sucesos A_1, \dots, A_n y un suceso B de un espacio muestral Ω , la probabilidad de cualquier suceso B del espacio muestral se puede calcular mediante la fórmula*

$$P(B) = \sum_{i=1}^n P(A_i \cap B) = \sum_{i=1}^n P(A_i)P(B|A_i).$$

i Demostración

Prueba. La demostración del teorema es sencilla, ya que al ser A_1, \dots, A_n un sistema completo tenemos

$$B = B \cap E = B \cap (A_1 \cup \dots \cup A_n) = (B \cap A_1) \cup \dots \cup (B \cap A_n)$$

y como estos sucesos son incompatibles entre sí, se tiene

$$\begin{aligned} P(B) &= P((B \cap A_1) \cup \dots \cup (B \cap A_n)) = P(B \cap A_1) + \dots + P(B \cap A_n) = \\ &= P(A_1)P(B/A_1) + \dots + P(A_n)P(B/A_n) = \sum_{i=1}^n P(A_i)P(B/A_i). \end{aligned}$$

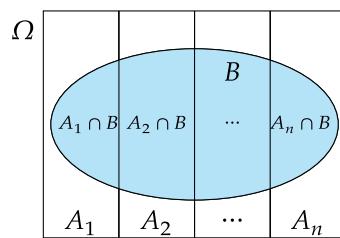


Figura 5.14: Teorema de la probabilidad total.

□

Ejemplo 5.18. Un determinado síntoma S puede ser originado por una enfermedad E pero también lo pueden presentar las personas sin la enfermedad. Sabemos que la prevalencia de la enfermedad E es 0.2. Además, se sabe que el 90% de las personas con la enfermedad presentan el síntoma, mientras que sólo el 40% de las personas sin la enfermedad lo presentan. Si se toma una persona al azar de la población, ¿qué probabilidad hay de que tenga el síntoma?

Para responder a la pregunta se puede aplicar el teorema de la probabilidad total usando el sistema completo $\{E, \bar{E}\}$:

$$P(S) = P(E)P(S|E) + P(\bar{E})P(S|\bar{E}) = 0.2 \cdot 0.9 + 0.8 \cdot 0.4 = 0.5.$$

Es decir, la mitad de la población tendrá el síntoma.

¡En el fondo se trata de una media ponderada de probabilidades!

La respuesta a la pregunta anterior es evidente a la luz del árbol de probabilidad del espacio probabilístico del experimento.

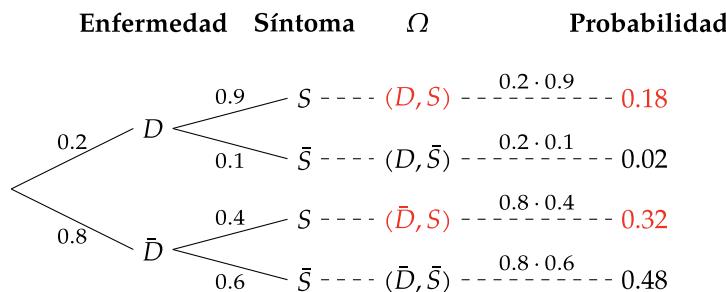


Figura 5.15: Aplicación del teorema de la probabilidad total en un espacio probabilístico.

$$\begin{aligned} P(S) &= P(E, S) + P(\bar{E}, S) = P(E)P(S|E) + P(\bar{E})P(S|\bar{E}) \\ &= 0.2 \cdot 0.9 + 0.8 \cdot 0.4 = 0.18 + 0.32 = 0.5. \end{aligned}$$

5.6 Teorema de Bayes

Los sucesos de un sistema completo de sucesos A_1, \dots, A_n también pueden verse como las distintas hipótesis ante un determinado hecho B .

En estas condiciones resulta útil poder calcular las probabilidades a posteriori $P(A_i|B)$ de cada una de las hipótesis.

Teorema 5.4 (Bayes). *Dado un sistema completo de sucesos A_1, \dots, A_n y un suceso B de un espacio muestral Ω y otro suceso B del mismo espacio muestral, la probabilidad de cada suceso A_i $i = 1, \dots, n$ condicionada por B puede calcularse con la siguiente fórmula*

$$P(A_i|B) = \frac{P(A_i \cap B)}{P(B)} = \frac{P(A_i)P(B|A_i)}{\sum_{i=1}^n P(A_i)P(B|A_i)}.$$

Ejemplo 5.19. En el ejemplo anterior, una pregunta más interesante es qué diagnosticar a una persona que presenta el síntoma.

En este caso se puede interpretar E y \bar{E} como las dos posibles hipótesis para el síntoma S . Las probabilidades a priori para ellas son $P(E) = 0.2$ y $P(\bar{E}) = 0.8$. Esto quiere decir que si no se dispone de información sobre el síntoma, el diagnóstico será que la persona no tiene la enfermedad.

Sin embargo, si al reconocer a la persona se observa que presenta el síntoma, dicha información condiciona a las hipótesis, y para decidir entre ellas es necesario calcular sus probabilidades a posteriori, es decir, $P(E|S)$ y $P(\bar{E}|S)$.

Para calcular las probabilidades a posteriori se puede utilizar el teorema de Bayes:

$$\begin{aligned} P(E|S) &= \frac{P(E)P(S|E)}{P(E)P(S|E) + P(\bar{E})P(S|\bar{E})} = \frac{0.2 \cdot 0.9}{0.2 \cdot 0.9 + 0.8 \cdot 0.4} = \frac{0.18}{0.5} = 0.36, \\ P(\bar{E}|S) &= \frac{P(\bar{E})P(S|\bar{E})}{P(E)P(S|E) + P(\bar{E})P(S|\bar{E})} = \frac{0.8 \cdot 0.4}{0.2 \cdot 0.9 + 0.8 \cdot 0.4} = \frac{0.32}{0.5} = 0.64. \end{aligned}$$

Como se puede ver la probabilidad de tener la enfermedad ha aumentado. No obstante, la probabilidad de no tener la enfermedad sigue siendo mayor que la de tenerla, y por esta razón el diagnóstico seguirá siendo que no tiene la enfermedad.

En este caso se dice que el síntoma *S* no es determinante a la hora de diagnosticar la enfermedad.

5.7 Epidemiología

Una de las ramas de la Medicina que hace un mayor uso de la probabilidad es la , que estudia la distribución y las causas de las enfermedades en las poblaciones, identificando factores de riesgos para las enfermedades de cara a la atención médica preventiva.

En Epidemiología interesa la frecuencia de un *suceso médico E* (típicamente una enfermedad como la gripe, un factor de riesgo como fumar o un factor de protección como vacunarse) que se mide mediante una variable nominal con dos categorías (ocurrencia o no del suceso).

Hay diferentes medidas relativas a la frecuencia de un suceso médico. Las más importantes son:

- Prevalencia
- Incidencia
- Riesgo relativo
- Odds ratio

5.7.1 Prevalencia

Definición 5.16 (Prevalencia). La *prevalencia* de un suceso médico *E* es la proporción de una población que está afectada por el suceso.

$$\text{Prevalencia}(E) = \frac{\text{Nº individuos afectados por } E}{\text{Tamaño poblacional}}$$

A menudo, la prevalencia se estima mediante una muestra como la frecuencia relativa de los individuos afectados por el suceso en la muestra. Es también común expresarla esta frecuencia como un porcentaje.

Ejemplo 5.20. Para estimar la prevalencia de la gripe se estudió una muestra de 1000 personas de las que 150 presentaron gripe. Así, la prevalencia de la gripe es aproximadamente $150/1000 = 0.15$, es decir, un 15%.

5.7.2 Incidencia

La mide la probabilidad de ocurrencia de un suceso médico en una población durante un periodo de tiempo específico. La incidencia puede medirse como una proporción acumulada o como una tasa.

Definición 5.17 (Incidencia acumulada). La *incidencia acumulada* de un suceso médico E es la proporción de individuos que experimentaron el evento en un periodo de tiempo, es decir, el número de nuevos casos afectados por el evento en el periodo de tiempo, dividido por el tamaño de la población inicialmente en riesgo de verse afectada.

$$R(E) = \frac{\text{Nº de nuevos casos con } E}{\text{Tamaño de la población en riesgo}}.$$

Ejemplo 5.21. Una población contenía inicialmente 1000 personas sin gripe y después de dos años se observó que 160 de ellas sufrieron gripe. La incidencia acumulada de la gripe es 160 casos pro 1000 personas por dos años, es decir, 16% en dos años.

5.7.3 Tasa de incidencia o Riesgo absoluto

Definición 5.18 (Riesgo absoluto). La *tasa de incidencia o riesgo absoluto* de un suceso médico E es el número de nuevos casos afectados por el evento dividido por la población en riesgo y por el número de unidades temporales del periodo considerado.

$$R(E) = \frac{\text{Nº nuevos casos con } E}{\text{Tamaño población en riesgo} \times \text{Nº unidades de tiempo}}$$

Ejemplo 5.22. Una población contenía inicialmente 1000 personas sin gripe y después de dos años se observó que 160 de ellas sufrieron gripe. Si se considera el año como intervalo de tiempo, la tasa de incidencia de la gripe es 160 casos dividida por 1000 personas y por dos años, es decir, 80 casos por 1000 personas-año o 8% de personas al año.

5.7.4 Prevalencia vs Incidencia

La prevalencia no debe confundirse con la incidencia. La prevalencia indica cómo de extendido está el suceso médico en una población, sin preocuparse por cuándo los sujetos se han expuesto al riesgo o durante cuánto tiempo, mientras que la incidencia se fija en el riesgo de verse afectado por el suceso en un periodo concreto de tiempo.

Así, la prevalencia se calcula en estudios transversales en un momento temporal puntual, mientras que para medir la incidencia se necesita un estudio longitudinal que permita observar a los individuos durante un periodo de tiempo.

La incidencia es más útil cuando se pretende entender la causalidad del suceso: por ejemplo, si la incidencia de una enfermedad en una población aumenta, seguramente hay un factor de riesgo que lo está promoviendo.

Cuando la tasa de incidencia es aproximadamente constante en la duración del suceso, la prevalencia es aproximadamente el producto de la incidencia por la duración media del suceso, es decir,

$$\text{Prevalencia} = \text{Inciden}\text{cia} \times \text{duración}$$

5.7.5 Comparación de riesgos

Para determinar si un factor o característica está asociada con el suceso médico es necesario comparar el riesgo del suceso en dos poblaciones, una expuesta al factor y la otra no. El grupo expuesto al factor se conoce como *grupo tratamiento* o *grupo experimental T* y el grupo no expuesto como *grupo control C*.

Habitualmente los casos observados para cada grupo se representan en una tabla de 2×2 como la siguiente:

Suceso E

No suceso \bar{E}

Grupo tratamiento T

a

b

Grupo control C

c

d

5.7.6 Riesgo atribuible o diferencia de riesgos RA

Definición 5.19 (Riesgo atribuible). El *riesgo atribuible* o *diferencia de riesgo* de un suceso médico E para los individuos expuestos a un factor es la diferencia entre los riesgos absolutos de los grupos tratamiento y control.

$$RA(E) = R_T(E) - R_C(E) = \frac{a}{a+b} - \frac{c}{c+d}.$$

El riesgo atribuible es el riesgo de un suceso que es debido específicamente al factor de interés.

Obsérvese que el riesgo atribuible puede ser positivo, cuando el riesgo del grupo tratamiento es mayor que el del grupo control, o negativo, de lo contrario.

Ejemplo 5.23. Para determinar la efectividad de una vacuna contra la gripe, una muestra de 1000 personas sin gripe fueron seleccionadas al comienzo del año. La mitad de ellas fueron vacunadas (grupo tratamiento) y la otra mitad recibieron un placebo (grupo control). La tabla siguiente resume los resultados al final del año.

Gripe E

No gripe \bar{E}

Grupo tratamiento (vacunados)

20

480

Grupo control (No vacunados)

80

420

El riesgo atribuible de contraer la gripe cuando se es vacunado es

$$AR(D) = \frac{20}{20 + 480} - \frac{80}{80 + 420} = -0.12.$$

Esto quiere decir que el riesgo de contraer la gripe es un 12% menor en vacunados que en no vacunados.

5.7.7 Riesgo relativo RR

Definición 5.20 (Riesgo relativo). El *riesgo relativo* de un suceso médico E para los individuos expuestos a un factor es el cociente entre las proporciones de individuos afectados por el suceso en un periodo de tiempo de los grupos tratamiento y control. Es decir, el cociente entre las incidencias de grupo tratamiento y el grupo control.

$$RR(D) = \frac{\text{Riesgo grupo tratamiento}}{\text{Riesgo grupo control}} = \frac{R_T(E)}{R_C(E)} = \frac{a/(a+b)}{c/(c+d)}$$

Interpretación

El riesgo relativo compara el riesgo de desarrollar un suceso médico entre el grupo tratamiento y el grupo control.

- $RR = 1 \Rightarrow$ No hay asociación entre el suceso y la exposición al factor.
- $RR < 1 \Rightarrow$ La exposición al factor disminuye el riesgo del suceso.
- $RR > 1 \Rightarrow$ La exposición al factor aumenta el riesgo del suceso.

Cuanto más lejos de 1, más fuerte es la asociación.

Ejemplo 5.24. Para determinar la efectividad de una vacuna contra la gripe, una muestra de 1000 personas sin gripe fueron seleccionadas al comienzo del año. La mitad de ellas fueron vacunadas (grupo tratamiento) y la otra mitad recibieron un placebo (grupo control). La tabla siguiente resume los resultados al final del año.

Gripe E

No gripe \bar{E}

Grupo tratamiento (vacunados)

20

480

Grupo control (No vacunados)

80

420

El riesgo relativo de contraer la gripe cuando se es vacunado es

$$RR(D) = \frac{20/(20 + 480)}{80/(80 + 420)} = 0.25.$$

Así, la probabilidad de contraer la gripe en los individuos vacunados fue la cuarta parte de la de contraerla en el caso de no haberse vacunado, es decir, la vacuna reduce el riesgo de gripe un 75%.

5.7.8 Odds

Una forma alternativa de medir el riesgo de un suceso médico es el *odds*.

Definición 5.21. El *odds* de un suceso médico E en una población es el cociente entre el número de individuos que adquirieron el suceso y los que no en un periodo de tiempo.

$$ODDS(E) = \frac{\text{Nº nuevos casos con } E}{\text{Nº casos sin } E} = \frac{P(E)}{P(\bar{E})}.$$

A diferencia de la incidencia, que es una proporción menor o igual que 1, el odds puede ser mayor que 1. No obstante es posible convertir el odds en una probabilidad con la fórmula

$$P(E) = \frac{ODDS(E)}{ODDS(E) + 1}.$$

Ejemplo 5.25. Una población contenía inicialmente 1000 personas sin gripe. Después de un año 160 de ellas tuvieron gripe. Entonces el odds de la gripe es 160/840.

Obsérvese que la incidencia es 160/1000.

5.7.9 Odds ratio OR

Definición 5.22 (Odds ratio). El *odds ratio* o la *oportunidad relativa* de un suceso médico E para los individuos expuestos a un factor es el cociente entre los odds del sucesos de los grupos tratamiento y control.

$$OR(E) = \frac{\text{Odds en grupo tratamiento}}{\text{Odds en grupo control}} = \frac{a/b}{c/d} = \frac{ad}{bc}.$$

Interpretación

El odds ratio compara los odds de un suceso médico entre el grupo tratamiento y control. La interpretación es similar a la del riesgo relativo:

- $OR = 1 \Rightarrow$ No existe asociación entre el suceso y la exposición al factor.
- $OR < 1 \Rightarrow$ La exposición al factor disminuye el riesgo del suceso.
- $OR > 1 \Rightarrow$ La exposición al factor aumenta el riesgo del suceso.

Cuanto más lejos de 1, más fuerte es la asociación.

Ejemplo 5.26. Para determinar la efectividad de una vacuna contra la gripe, una muestra de 1000 personas sin gripe fueron seleccionadas al comienzo del año. La mitad de ellas fueron vacunadas (grupo tratamiento) y la otra mitad recibieron un placebo (grupo control). La tabla siguiente resume los resultados al final del año.

Gripe E

No gripe \bar{E}

Grupo tratamiento (vacunados)

20

480

Grupo control (No vacunados)

80

420

El odds ratio de sufrir la gripe para los individuos vacunados es

$$OR(D) = \frac{20/480}{80/420} = 0.21875.$$

Esto quiere decir que el odds de sufrir la gripe frente a no sufrirla en los vacunados es casi un quinto del de los no vacunados, es decir, que aproximadamente por cada 22 personas vacunadas con gripe habrá 100 personas no vacunadas con gripe.

5.7.10 Riesgo relativo vs Odds ratio

El riesgo relativo y el odds ratio son dos medidas de asociación pero su interpretación es ligeramente diferente. Mientras que el riesgo relativo expresa una comparación de riesgos entre los grupos tratamiento y control, el odds ratio expresa una comparación de odds, que no es lo mismo que el riesgo. Así, un odds ratio de 2 *no* significa que el grupo tratamiento tiene el doble de riesgo de adquirir el suceso.

La interpretación del odds ratio es un poco más enrevesada porque es *contrafactual*, y nos da cuántas veces es más frecuente el suceso en el grupo tratamiento en comparación con el control, asumiendo que en el grupo control es tan frecuente que ocurra el suceso como que no.

La ventaja del odds ratio es que no depende de la prevalencia o la incidencia del suceso, y debe usarse siempre que el número de individuos que presenta el suceso se selecciona arbitrariamente en ambos grupos, como ocurre en los estudios casos-control.

Ejemplo 5.27. Para determinar la asociación entre el cáncer de pulmón y fumar se tomaron dos muestras (la segunda con el doble de individuos sin cáncer) obteniendo los siguientes resultados:

Muestra 1

Cáncer	
No cáncer	
Fumadores	
60	
80	
No fumadores	
40	
320	

$$RR(D) = \frac{60/(60 + 80)}{40/(40 + 320)} = 3.86.$$

$$OR(D) = \frac{60/80}{40/320} = 6.$$

Muestra 2

Cáncer	
No cáncer	
Fumadores	
60	
160	
No fumadores	
40	
640	

$$RR(D) = \frac{60/(60 + 160)}{40/(40 + 640)} = 4.64.$$

$$OR(D) = \frac{60/160}{40/640} = 6.$$

Así, cuando cambia la incidencia o prevalencia de un suceso (cáncer de pulmón) el riesgo relativo cambia, mientras que el odds ratio no.

La relación entre el riesgo relativo y el odds ratio viene dada por la siguiente fórmula

$$RR = \frac{OR}{1 - R_0 + R_0 \cdot OR} = OR \frac{1 - R_1}{1 - R_0},$$

donde R_C and R_T son la prevalencia o la incidencia en los grupos control y tratamiento respectivamente.

El odds ratio siempre sobreestima el riesgo relativo cuando este es mayor que 1 y lo subestima cuando es menor que 1. No obstante, con sucesos médicos raros (con una prevalencia o incidencia baja) el riesgo relativo y el odds ratio son casi iguales.

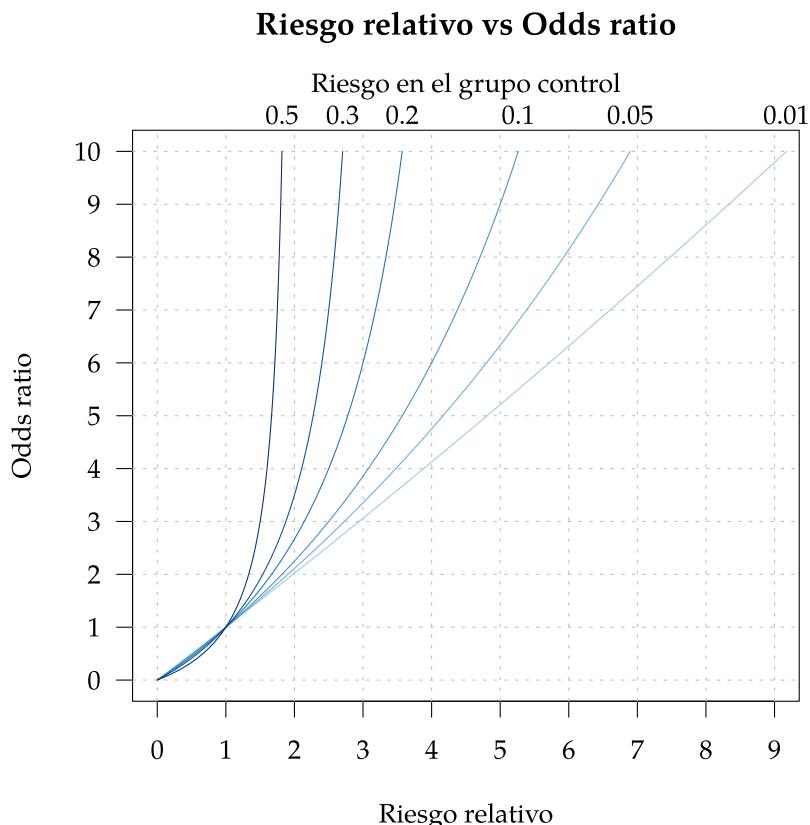


Figura 5.16: Odss ratio versus riesgo relativo.

5.8 Tests diagnósticos

En Epidemiología es común el uso de test para diagnosticar enfermedades.

Generalmente estos test no son totalmente fiables, sino que hay cierta probabilidad de acierto o fallo en el diagnóstico, que suele representarse en la siguiente tabla:

Presencia enfermedad E

Ausencia enfermedad \bar{E}

Test positivo +

Verdadero positivo VP

Falso positivo FP

Test negativo -

Falso negativo FN

Verdadero Negativo VN

5.8.1 Sensibilidad y especificidad de un test diagnóstico

La fiabilidad de un test diagnóstico depende de las siguientes probabilidades.

Definición 5.23 (Sensibilidad). La *sensibilidad* de un test diagnóstico es la proporción de resultados positivos del test en personas con la enfermedad,

$$P(+|E) = \frac{VP}{VP + FN}.$$

Definición 5.24 (Especificidad). La *especificidad* de un test diagnóstico es la proporción de resultados negativos del test en personas sin la enfermedad,

$$P(-|\bar{E}) = \frac{VN}{VN + FP}.$$

Normalmente existe un balance entre la sensibilidad y la especificidad.

Un test con una alta sensibilidad detectará la enfermedad en la mayoría de las personas enfermas, pero también dará más falsos positivos que un test menos sensible. De este modo, un resultado positivo en un test con una gran sensibilidad no es muy útil para confirmar la enfermedad, pero un resultado negativo es útil para descartar la enfermedad, ya que raramente da resultados negativos en personas con la enfermedad.

Por otro lado, un test con una alta especificidad descartará la enfermedad en la mayoría de las personas sin la enfermedad, pero también producirá más falsos negativos que un test menos específico. Así, un resultado negativo en un test con una gran especificidad no es útil para descartar la enfermedad, pero un resultado positivo es muy útil para confirmar la enfermedad, ya que raramente da resultados positivos en personas sin la enfermedad.

Ejemplo 5.28. Un test diagnóstico para la gripe se ha aplicado a una muestra aleatoria de 1000 personas. Los resultados aparecen resumidos en la siguiente tabla.

Presencia de gripe E

Ausencia de gripe \bar{E}

Test +

95

90

Test -

5

810

Según esta muestra, la prevalencia de la gripe puede estimarse como

$$P(E) = \frac{95 + 5}{1000} = 0.1.$$

La sensibilidad del test diagnóstico es

$$P(+|E) = \frac{95}{95 + 5} = 0.95.$$

Y la especificidad es

$$P(-|\bar{E}) = \frac{810}{90 + 810} = 0.9.$$

Así pues, se trata de un buen test tanto para descartar la enfermedad como para confirmarla, pero es un poco mejor para confirmarla que para descartarla porque la especificidad es mayor que la sensibilidad.

Decidir entre un test con una gran sensibilidad o un test con una gran especificidad depende del tipo de enfermedad y el objetivo del test. En general, utilizaremos un test sensible cuando:

- La enfermedad es grave y es importante detectarla.
- La enfermedad es curable.
- Los falsos positivos no provocan traumas serios.

Y utilizaremos un test específico cuando:

- La enfermedad es importante pero difícil o imposible de curar.
- Los falsos positivos pueden provocar traumas serios.
- El tratamiento de los falsos positivos puede tener graves consecuencias.

5.8.2 Valores predictivos de un test diagnóstico

Pero el aspecto más importante de un test diagnóstico es su poder predictivo, que se mide con las siguientes probabilidades a posteriori.

Definición 5.25 (Valor predictivo positivo). El *valor predictivo positivo* de un test diagnóstico es la proporción de personas con la enfermedad entre las personas con resultado positivo en el test,

$$P(E|+) = \frac{VP}{VP + FP}.$$

Definición 5.26 (Valor predictivo negativo). El *valor predictivo negativo* de un test diagnóstico es la proporción de personas sin la enfermedad entre las personas con resultado negativo en el test,

$$P(\bar{E}|-) = \frac{VN}{VN + FN}.$$

i Interpretación

Los valores predictivos positivo y negativo permiten confirmar o descartar la enfermedad, respectivamente, si alcanzan al menos el umbral de 0.5.

$$\begin{aligned} VPP > 0.5 &\Rightarrow \text{Diagnosticar la enfermedad} \\ VPN > 0.5 &\Rightarrow \text{Diagnosticar la no enfermedad} \end{aligned}$$

No obstante, estas probabilidades dependen de la prevalencia de la enfermedad $P(E)$. Pueden calcularse a partir de la sensibilidad y la especificidad del test diagnóstico usando el teorema de Bayes.

$$VPP = P(E|+) = \frac{P(E)P(+|E)}{P(E)P(+|E) + P(\bar{E})P(+|\bar{E})}$$

$$VPN = P(\bar{E}|-) = \frac{P(\bar{E})P(-|\bar{E})}{P(E)P(-|E) + P(\bar{E})P(-|\bar{E})}$$

Así, con enfermedades frecuentes, el valor predictivo positivo aumenta, y con enfermedades raras, el valor predictivo negativo aumenta.

Ejemplo 5.29. Siguiendo con el ejemplo anterior de la gripe, se tiene que el valor predictivo positivo del test es

$$VPP = P(E|+) = \frac{95}{95 + 90} = 0.5135.$$

Como este valor es mayor que 0.5, eso significa que se diagnosticará la gripe si el resultado del test es positivo. No obstante, la confianza en el diagnóstico será baja, ya que el valor es poco mayor que 0.5.

Por otro lado, el valor predictivo negativo es

$$VPN = P(\bar{E}|-) = \frac{810}{5 + 810} = 0.9939.$$

Como este valor es casi 1, eso significa que es casi seguro que no se tiene la gripe cuando el resultado del test es negativo.

Así, se puede concluir que este test es muy potente para descartar la gripe, pero no lo es tanto para confirmarla.

5.8.3 Razón de verosimilitud de un test diagnóstico

Las siguientes medidas también se derivan de la sensibilidad y la especificidad de un test diagnóstico.

Definición 5.27 (Razón de verosimilitud positiva). La *razón de verosimilitud positiva* de un test diagnóstico es el cociente entre la probabilidad de un resultado positivo en personas con la enfermedad y personas sin la enfermedad, respectivamente.

$$RV+ = \frac{P(+|E)}{P(+|\bar{E})} = \frac{\text{Sensibilidad}}{1 - \text{Especificidad}}.$$

Definición 5.28 (Razón de verosimilitud negativa). La *razón de verosimilitud negativa* de un test diagnóstico es el cociente entre la probabilidad de un resultado negativo en personas con la enfermedad y personas sin la enfermedad, respectivamente.

$$RV- = \frac{P(-|E)}{P(-|\bar{E})} = \frac{1 - \text{Sensibilidad}}{\text{Especificidad}}.$$

Interpretación

La razón de verosimilitud positiva puede interpretarse como el número de veces que un resultado positivo es más probable en personas con la enfermedad que en personas sin la enfermedad.

Por otro lado, la razón de verosimilitud negativa puede interpretarse como el número de veces que un resultado negativo es más probable en personas con la enfermedad que en personas sin la enfermedad.

Las probabilidades a posteriori pueden calcularse a partir de las probabilidades a priori usando las razones de verosimilitud

$$P(E|+) = \frac{P(E)P(+|E)}{P(E)P(+|E) + P(\bar{E})P(+|\bar{E})} = \frac{P(E)RV+}{1 - P(E) + P(E)RV+}$$

Así,

- Una razón de verosimilitud positiva mayor que 1 aumenta la probabilidad de la enfermedad.
- Una razón de verosimilitud positiva menor que 1 disminuye la probabilidad de la enfermedad.
- Una razón de verosimilitud 1 no cambia la probabilidad a priori de la de tener la enfermedad.

Relación entre las probabilidades a priori, a posteriori, y la razón de verosimilitud

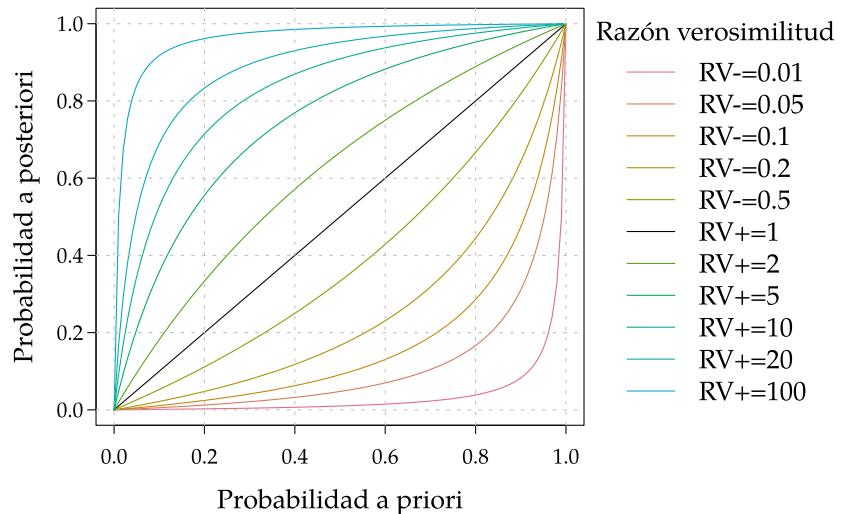


Figura 5.17: Razón de verosimilitud.

6 Estimación de parámetros poblacionales

Los modelos de distribución de probabilidad vistos en el tema anterior explican el comportamiento de las variables aleatorias, pero para ello debemos saber qué modelo de distribución sigue una determinada variable. Este es el primer paso de la etapa de *Inferecia Estadística*.

Para determinar con exactitud el modelo de distribución de una variable hay que conocer la característica estudiada en todos los individuos de la población, lo cual no es posible en la mayoría de los casos (inviabilidad económica, física, temporal, etc.).

Para evitar estos inconvenientes se recurre al estudio de una muestra, a partir de la cual se trata de averiguar, de manera *aproximada*, el modelo de distribución de la variable en la población.

Estudiar un número reducido de individuos de una muestra en lugar de toda la población tiene indudables ventajas:

- Menor coste.
- Mayor rapidez.
- Mayor facilidad.

Pero también presenta algunos inconvenientes:

- Necesidad de conseguir una muestra representativa.
- Posibilidad de cometer errores (*sesgos*).

Afortunadamente, estos errores pueden ser superados: La representatividad de la muestra se consigue eligiendo la modalidad de muestreo más apropiada para el tipo de estudio; en el caso de los errores, aunque no se pueden evitar, se tratará de reducirlos al máximo y acotarlos.

6.1 Distribuciones muestrales

Los valores de una variable X en una muestra de tamaño n de una población pueden verse como el valor de una variable aleatoria n -dimensional.

Definición 6.1 (Variable aleatoria muestral). Una *variable aleatoria muestral* de una variable X estudiada en una población es una colección de n variables aleatorias X_1, \dots, X_n tales que:

- Cada una de las variables X_i sigue la misma distribución de probabilidad que la variable X en la población.
- Todas las variables X_i son mutuamente independientes.

Los valores que puede tomar esta variable n dimensional, serán todas las posibles muestras de tamaño n que pueden extraerse de la población.

Las tres características fundamentales de la variable aleatoria muestral son:

- **Homogeneidad:** Las n variables que componen la variable aleatoria muestral siguen la misma distribución.
- **Independencia:** Las variables son independientes entre sí.
- **Modelo de distribución:** El modelo de distribución que siguen las n variables.

Las dos primeras cuestiones pueden resolverse si se utiliza muestreo aleatorio simple para obtener la muestra. En cuanto a la última, hay que responder, a su vez, a dos cuestiones:

1. ¿Qué modelo de distribución se ajusta mejor a nuestro conjunto de datos? Esto se resolverá, en parte, mediante la utilización de técnicas no paramétricas.
2. Una vez seleccionado el modelo de distribución más apropiado, ¿qué estadístico del modelo nos interesa y cómo determinar su valor? De esto último se encarga la parte de la inferencia estadística conocida como **Estimación de Parámetros**.

En este tema se abordará la segunda cuestión, es decir, suponiendo que se conoce el modelo de distribución de una población, se intentará estimar los principales parámetros que la definen. Por ejemplo, los principales parámetros que definen las distribuciones vistas en el tema anterior son:

Distribución	Parámetro
Binomial	n, p
Poisson	λ
Uniforme	a, b
Normal	μ, σ
Chi-cuadrado	n
T-Student	n
F-Fisher	m, n

La distribución de probabilidad de los valores de la variable muestral depende claramente de la distribución de probabilidad de los valores de la población.

Ejemplo 6.1. Sea una población en la que la cuarta parte de las familias no tienen hijos, la mitad de las familias tiene 1 hijo, y el resto tiene 2 hijos.

Distribución Poblacional

X	P(x)
0	0.25
1	0.50
2	0.25

Muestras de tamaño 2



Distribución muestral

(X_1, X_2)	$P(x_1, x_2)$
(0,0)	0.0625
(0,1)	0.1250
(0,2)	0.0625
(1,0)	0.1250
(1,1)	0.2500
(1,2)	0.1250
(2,0)	0.0625
(2,1)	0.1250
(2,2)	0.0625

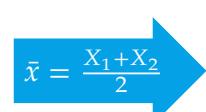
Por ser función de una variable aleatoria, un estadístico en el muestreo es también una variable aleatoria. Por tanto, su distribución de probabilidad también depende de la distribución de la población y de los parámetros que la determinan (μ, σ, p, \dots).

Ejemplo 6.2. Si se toma la media muestral \bar{X} de las muestras de tamaño 2 del ejemplo anterior, su distribución de probabilidad es

Distribución muestral

(X_1, X_2)	$P(x_1, x_2)$
(0,0)	0.0625
(0,1)	0.1250
(0,2)	0.0625
(1,0)	0.1250
(1,1)	0.2500
(1,2)	0.1250
(2,0)	0.0625
(2,1)	0.1250
(2,2)	0.0625

$\bar{x} = \frac{X_1 + X_2}{2}$

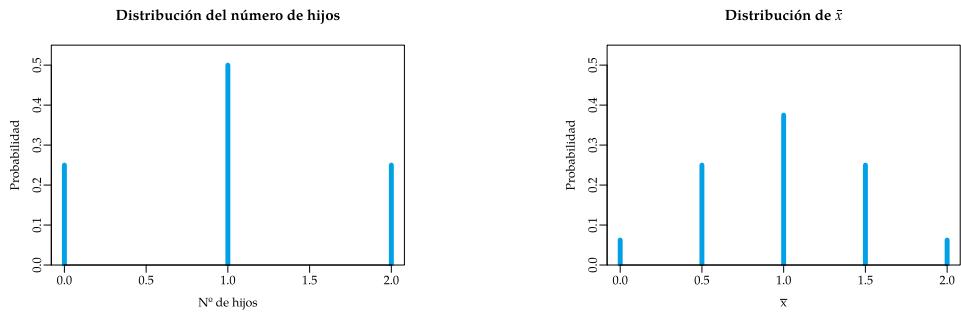


Distribución de \bar{x}

\bar{X}	$P(x)$
0	0.0625
0.5	0.2500
1	0.3750
1.5	0.2500
2	0.0625

¿Cuál es la probabilidad de obtener una media muestral que aproxime la media poblacional con un error máximo de 0.5?

Como hemos visto, para conocer la distribución de un estadístico muestral, es necesario conocer la distribución de la población, lo cual no siempre es posible. Afortunadamente, para muestras grandes es posible aproximar la distribución de algunos estadísticos como la media, gracias al siguiente teorema:



Teorema 6.1 (Teorema central del límite). Si X_1, \dots, X_n son variables aleatorias independientes ($n \geq 30$) con medias y varianzas $\mu_i = E(X_i)$, $\sigma_i^2 = Var(X_i)$, $i = 1, \dots, n$ respectivamente, entonces la variable aleatoria $X = X_1 + \dots + X_n$ sigue una distribución aproximadamente normal de media la suma de las medias y varianza la suma de las varianzas

$$X = X_1 + \dots + X_n \stackrel{n \geq 30}{\sim} N \left(\sum_{i=1}^n \mu_i, \sqrt{\sum_{i=1}^n \sigma_i^2} \right)$$

Este teorema además es la explicación de que la mayoría de las variables biológicas presenten una distribución normal, ya que suelen ser causa de múltiples factores que suman sus efectos de manera independiente.

6.1.1 Distribución de la media muestral para muestras grandes ($n \geq 30$)

La media muestral de una muestra aleatoria de tamaño n es la suma de n variables aleatorias independientes, idénticamente distribuidas:

$$\bar{X} = \frac{X_1 + \dots + X_n}{n} = \frac{X_1}{n} + \dots + \frac{X_n}{n}$$

De acuerdo a las propiedades de las transformaciones lineales, la media y la varianza de cada una de estas variables son

$$E \left(\frac{X_i}{n} \right) = \frac{\mu}{n} \quad \text{y} \quad Var \left(\frac{X_i}{n} \right) = \frac{\sigma^2}{n^2}$$

con μ y σ^2 la media y la varianza de la población de partida.

Entonces, si el tamaño de la muestra es grande ($n \geq 30$), de acuerdo al teorema central del límite, la distribución de la media muestral será normal:

$$\bar{X} \sim N\left(\sum_{i=1}^n \frac{\mu}{n}, \sqrt{\sum_{i=1}^n \frac{\sigma^2}{n^2}}\right) = N\left(\mu, \frac{\sigma}{\sqrt{n}}\right).$$

Ejemplo 6.3 (Ejemplo para muestras grandes ($n \geq 30$)). Supóngase que se desea estimar el número medio de hijos de una población con media $\mu = 2$ hijos y desviación típica $\sigma = 1$ hijo.

¿Qué probabilidad hay de estimar μ a partir de \bar{x} con un error menor de 0.2?

De acuerdo al teorema central del límite se tiene:

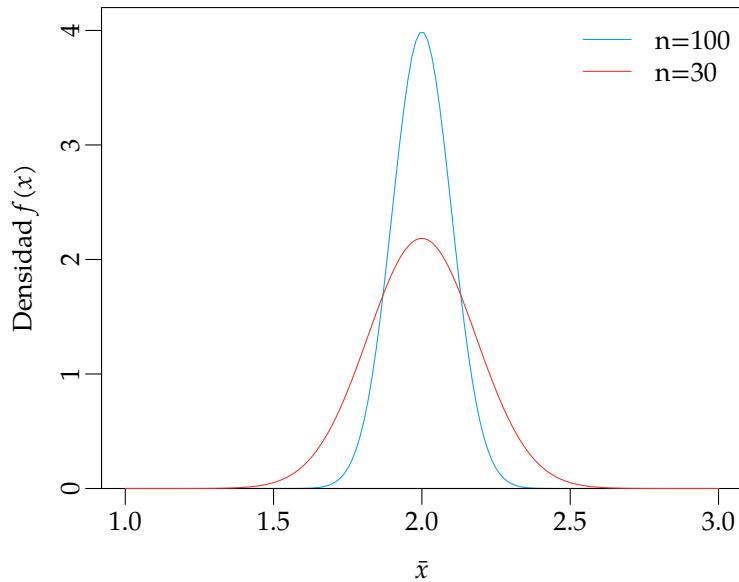
1. Para $n = 30$, $\bar{x} \sim N(2, 1/\sqrt{30})$ y

$$P(1.8 < \bar{x} < 2.2) = 0.7267.$$

1. Para $n = 100$, $\bar{x} \sim N(2, 1/\sqrt{100})$ y

$$P(1.8 < \bar{x} < 2.2) = 0.9545.$$

Distribuciones de la media del nº de hijos



6.1.2 Distribución de una proporción muestral para muestras grandes ($n \geq 30$)

Una proporción p poblacional puede calcularse como la media de una variable dicotómica (0,1). Esta variable se conoce como *variable de Bernouilli* $B(p)$, que es un caso particular de la binomial para $n = 1$. Por tanto, para una muestra aleatoria de tamaño n , una proporción muestral \hat{p} también puede expresarse como la suma de n variables aleatorias independientes, idénticamente distribuidas:

$$\hat{p} = \bar{X} = \frac{X_1 + \dots + X_n}{n} = \frac{X_1}{n} + \dots + \frac{X_n}{n}, \text{ con } X_i \sim B(p)$$

y con media y varianza

$$E\left(\frac{X_i}{n}\right) = \frac{p}{n} \quad \text{y} \quad Var\left(\frac{X_i}{n}\right) = \frac{p(1-p)}{n^2}$$

Entonces, si el tamaño de la muestra es grande ($n \geq 30$), de acuerdo al teorema central del límite, la distribución de la proporción muestral también será normal:

$$\hat{p} \sim N\left(\sum_{i=1}^n \frac{p}{n}, \sqrt{\sum_{i=1}^n \frac{p(1-p)}{n^2}}\right) = N\left(p, \sqrt{\frac{p(1-p)}{n}}\right).$$

6.2 Estimadores

Los estadísticos muestrales pueden utilizarse para aproximar los parámetros de la población, y cuando un estadístico se utiliza con este fin se le llama *estimador del parámetro*.

Definición 6.2 (Estimador y estimación). Un *estimador* es una función de la variable aleatoria muestral

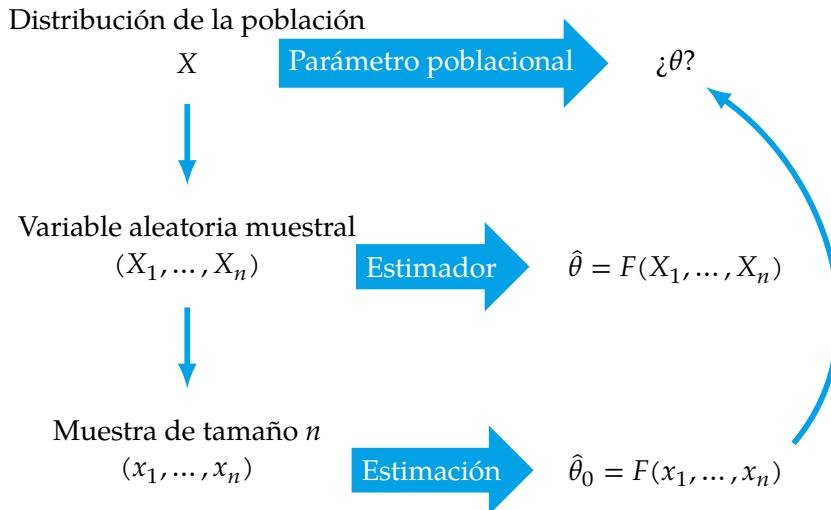
$$\hat{\theta} = F(X_1, \dots, X_n).$$

Dada una muestra concreta (x_1, \dots, x_n) , el valor del estimador aplicado a ella se conoce como *estimación*

$$\hat{\theta}_0 = F(x_1, \dots, x_n).$$

Por ser una función de la variable aleatoria muestral, un estimador es, a su vez, una variable aleatoria cuya distribución depende de la población de partida.

Mientras que el estimador es una función que es única, la estimación no es única, sino que depende de la muestra tomada.



Ejemplo 6.4. Supóngase que se quiere saber la proporción p de fumadores en una ciudad. En ese caso, la variable dicotómica que mide si una persona fuma (1) o no (0), sigue una distribución de Bernoulli $B(p)$.

Si se toma una muestra aleatoria de tamaño 5, $(X_1, X_2, X_3, X_4, X_5)$, de esta población, se puede utilizar la proporción de fumadores en la muestra como estimador para la proporción de fumadores en la población:

$$\hat{p} = \frac{\sum_{i=1}^5 X_i}{5}$$

Este estimador es una variable que se distribuye $\hat{p} \sim \frac{1}{n}B\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$.

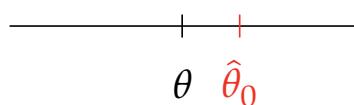
Si se toman distintas muestras, se obtienen diferentes estimaciones:

Muestra	Estimación
(1, 0, 0, 1, 1)	3/5
(1, 0, 0, 0, 0)	1/5
(0, 1, 0, 0, 1)	2/5
...	...

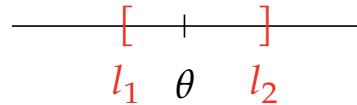
La estimación de parámetros puede realizar de dos formas:

- **Estimación puntual:** Se utiliza un único estimador que proporciona un valor o estimación aproximada del parámetro. El principal inconveniente de este tipo de estimación es que no se especifica la bondad de la estimación.
- **Estimación por intervalos:** Se utilizan dos estimadores que proporcionan los extremos de un intervalo dentro del cual se cree que está el verdadero valor del parámetro con un cierto grado de seguridad. Esta forma de estimar sí permite controlar el error cometido en la estimación.

Estimación puntual



Estimación por intervalo



6.3 Estimación puntual

La estimación puntual utiliza un único estimador para estimar el valor del parámetro desconocido de la población.

En teoría pueden utilizarse distintos estimadores para estimar un mismo parámetro. Por ejemplo, en el caso de estimar la proporción de fumadores en una ciudad, podrían haberse utilizado otros posibles estimadores además de la proporción muestral, como pueden ser:

$$\begin{aligned}\hat{\theta}_1 &= \sqrt[5]{X_1 X_2 X_3 X_4 X_5} \\ \hat{\theta}_2 &= \frac{X_1 + X_5}{2} \\ \hat{\theta}_3 &= X_1 \dots\end{aligned}$$

¿Cuál es el mejor estimador?

La respuesta a esta cuestión depende de las propiedades de cada estimador.

Aunque la estimación puntual no proporciona ninguna medida del grado de bondad de la estimación, existen varias propiedades que garantizan dicha bondad.

Las propiedades más deseables en un estimador son:

- Insesgadez
- Eficiencia
- Consistencia

- Normalidad asintótica
- Suficiencia

Definición 6.3 (Estimador insesgado). Un estimador $\hat{\theta}$ es *insesgado* para un parámetro θ si su esperanza es precisamente θ , es decir,

$$E(\hat{\theta}) = \theta.$$

Distribuciones de estimadores sesgados e insesgados

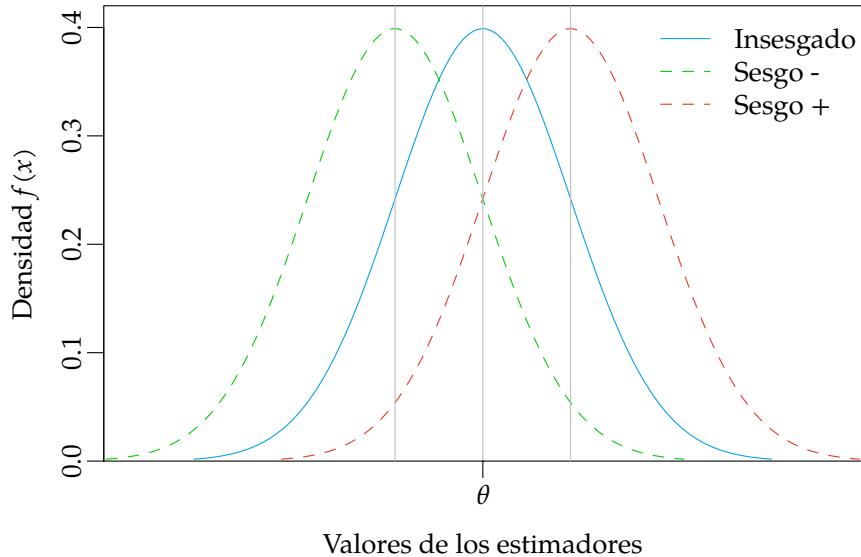


Figura 6.1: Distribución de estimadores sesgados e insesgados.

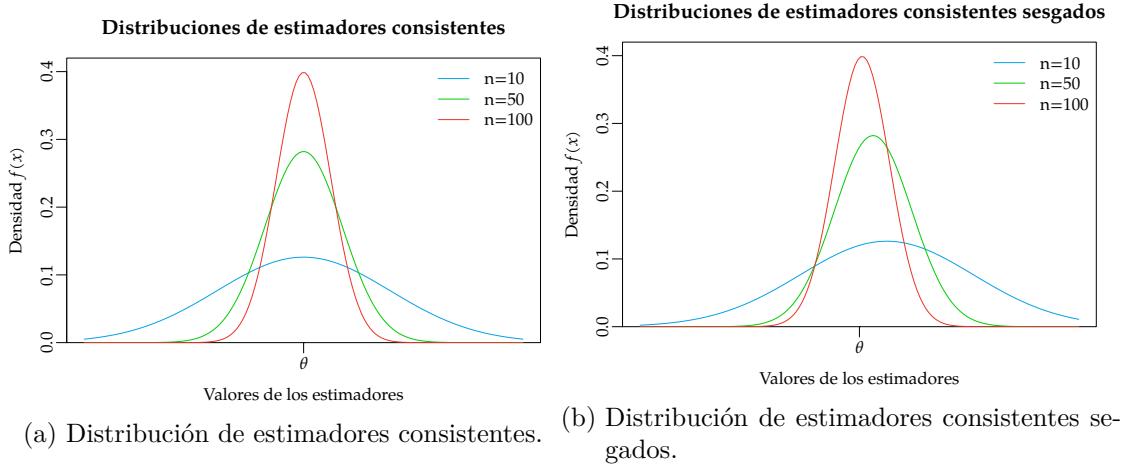
Cuando un estimador no es insesgado, a la diferencia entre su esperanza y el valor del parámetro θ se le llama *sesgo*:

$$Sesgo(\hat{\theta}) = E(\hat{\theta}) - \theta.$$

Cuanto menor sea el sesgo de un estimador, mejor se aproximarán sus estimaciones al verdadero valor del parámetro.

Definición 6.4 (Estimador consistente). Un estimador $\hat{\theta}_n$ para muestras de tamaño n es *consistente* para un parámetro θ si para cualquier valor $\epsilon > 0$ se cumple

$$\lim_{n \rightarrow \infty} P(|\hat{\theta}_n - \theta| < \epsilon) = 1.$$



Las condiciones suficientes para que un estimador sea consistente son:

1. $Sesgo(\hat{\theta}_n) = 0$ o $\lim_{n \rightarrow \infty} Sesgo(\hat{\theta}_n) = 0$.
2. $\lim_{n \rightarrow \infty} Var(\hat{\theta}_n) = 0$.

Así pues, si la varianza y el sesgo disminuyen a medida que aumenta el tamaño de la muestra, el estimador será consistente.

Definición 6.5 (Estimador eficiente). Un estimador $\hat{\theta}$ de un parámetro θ es *eficiente* si tiene el menor error cuadrático medio

$$ECM(\hat{\theta}) = Sesgo(\hat{\theta})^2 + Var(\hat{\theta}).$$

Distribuciones de estimadores insesgado y eficiente sesgado

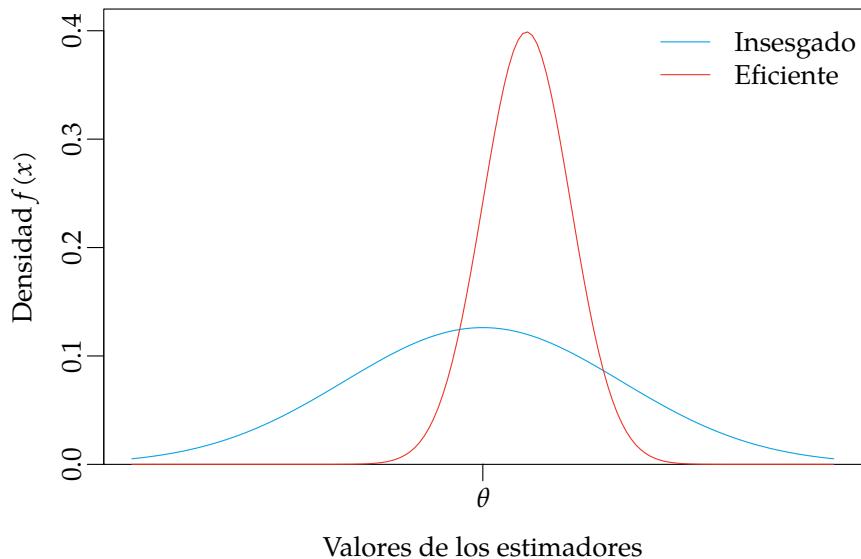


Figura 6.3: Distribución de estimadores insesgados y eficientes sesgados.

Definición 6.6 (Estimador asintóticamente normal). Un estimador $\hat{\theta}$ es *asintóticamente normal* si, independientemente de la distribución de la variable aleatoria muestral, su distribución es normal si el tamaño de la muestra es suficientemente grande.:::

Como veremos más adelante esta propiedad es muy interesante para hacer estimaciones de parámetros mediante intervalos.

Distribuciones de estimadores asintóticamente normales

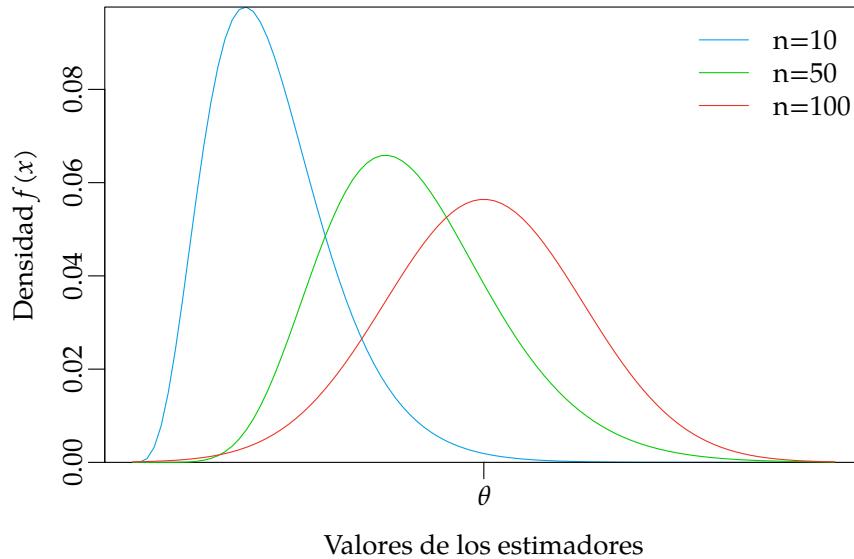


Figura 6.4: Distribución de estimadores asintóticamente normales.

Definición 6.7 (Estimador suficiente). Un estimador $\hat{\theta}$ es *suficiente* para un parámetro θ , si la distribución condicionada de la variable aleatoria muestral, una vez dada la estimación $\hat{\theta} = \hat{\theta}_0$, no depende de θ .

Esto significa que cuando se obtiene una estimación, cualquier otra información es irrelevante para θ .

El estimador que se suele utilizar para estimar la media poblacional es la media muestral.

Para muestras de tamaño n resulta la siguiente variable aleatoria:

$$\bar{X} = \frac{X_1 + \dots + X_n}{n}$$

Si la población de partida tiene media μ y varianza σ^2 se cumple

$$E(\bar{X}) = \mu \quad \text{y} \quad Var(\bar{X}) = \frac{\sigma^2}{n}$$

Así pues, la media muestral es un estimador insesgado, y como su varianza disminuye a medida que aumenta el tamaño muestral, también es consistente y eficiente.

Sin embargo, la varianza muestral

$$S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n}$$

es un estimador sesgado para la varianza poblacional, ya que

$$E(S^2) = \frac{n-1}{n} \sigma^2.$$

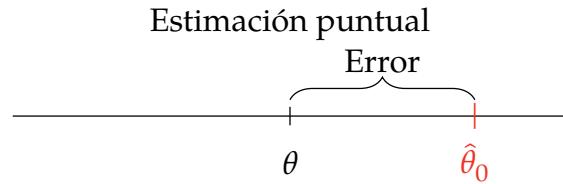
No obstante, resulta sencillo corregir este sesgo para llegar a un estimador insesgado:

Definición 6.8 (Cuasivarianza muestral). Dada una muestra de tamaño n de una variable aleatoria X , se define la *cuasivarianza muestral* como

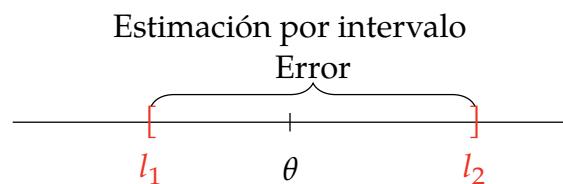
$$\hat{S}^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1} = \frac{n}{n-1} S^2.$$

6.4 Estimación por intervalos

El principal problema de la estimación puntual es que, una vez seleccionada la muestra y hecha la estimación, resulta imposible saber el error cometido.



Para controlar el error de la estimación es mejor utilizar la estimación por intervalos



La estimación por intervalos trata de construir a partir de la muestra un intervalo dentro del cual se supone que se encuentra el parámetro a estimar con un cierto grado de confianza. Para ello se utilizan dos estimadores, uno para el límite inferior del intervalo y otro para el superior.

Definición 6.9 (Intervalo de confianza). Dados dos estimadores $\hat{l}_i(X_1, \dots, X_n)$ y $\hat{l}_s(X_1, \dots, X_n)$, y sus respectivas estimaciones l_1 y l_2 para una muestra concreta, se dice que el intervalo $I = [l_1, l_2]$ es un intervalo de confianza para un parámetro poblacional θ , con un nivel de confianza $1 - \alpha$ (o nivel de significación α), si se cumple

$$P(\hat{l}_i(X_1, \dots, X_n) \leq \theta \leq \hat{l}_s(X_1, \dots, X_n)) = 1 - \alpha.$$

Un intervalo de confianza nunca garantiza con absoluta certeza que el parámetro se encuentra dentro él.

Tampoco se puede decir que la probabilidad de que el parámetro esté dentro del intervalo es $1 - \alpha$, ya que una vez calculado el intervalo, las variables aleatorias que determinan sus extremos han tomado un valor concreto y ya no tiene sentido hablar de probabilidad, es decir, o el parámetro está dentro, o está fuera, pero con absoluta certeza.

Lo que si se deduce de la definición es que el $(1 - \alpha)\%$ de los intervalos correspondientes a las todas las posibles muestras aleatorias, contendrán al parámetro. Es por eso que se habla de *confianza* y no de probabilidad.

Para que un intervalo sea útil su nivel de confianza debe ser alto:

$$1 - \alpha = 0.90 \text{ o } \alpha = 0.10$$

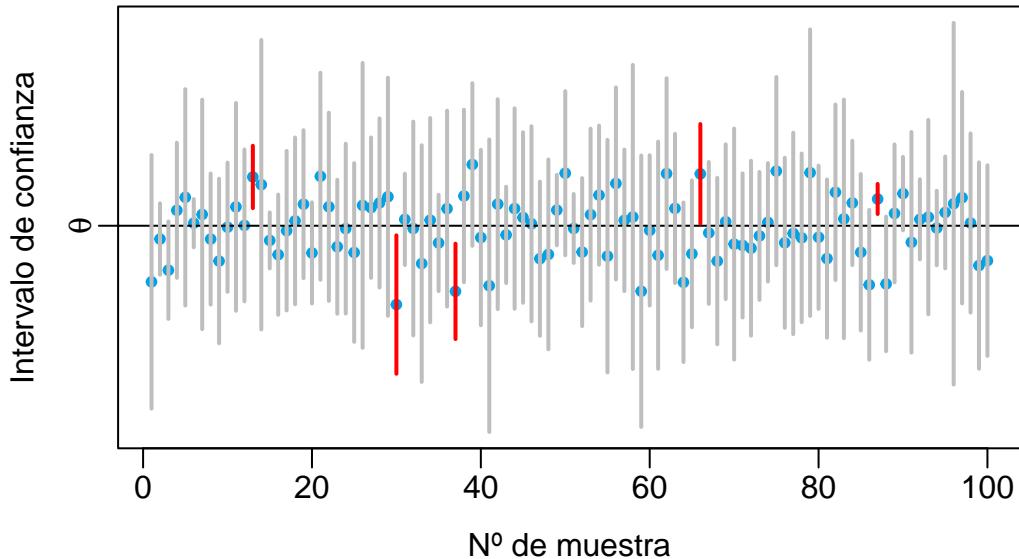
$$1 - \alpha = 0.95 \text{ o } \alpha = 0.05$$

$$1 - \alpha = 0.99 \text{ o } \alpha = 0.01$$

siendo 0.95 el nivel de confianza más habitual y 0.99 en casos críticos.

Teóricamente, de cada 100 intervalos para estimar un parámetro θ con nivel de confianza $1 - \alpha = 0.95$, 95 contendrían a θ y sólo 5 lo dejarían fuera.

50 intervalos de confianza del 95% para θ

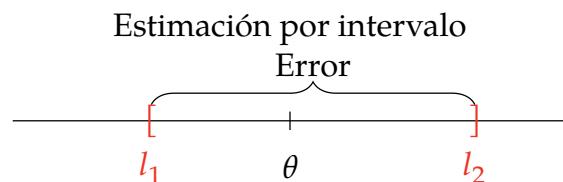


6.4.1 Error de estimación

Otro de los aspectos más importantes de un intervalo de confianza es su error.

Definición 6.10 (Error o imprecisión de un intervalo). El *error* o la *imprecisión* de un intervalo de confianza $[l_i, l_s]$ es su amplitud

$$A = l_s - l_i.$$



Para que un intervalo sea útil no debe ser demasiado impreciso.

En general, la precisión de un intervalo depende de tres factores:

- La dispersión de la población. Cuanto más dispersa sea, menos preciso será el intervalo.
- El nivel de confianza. Cuanto mayor sea el nivel de confianza, menos preciso será el intervalo.

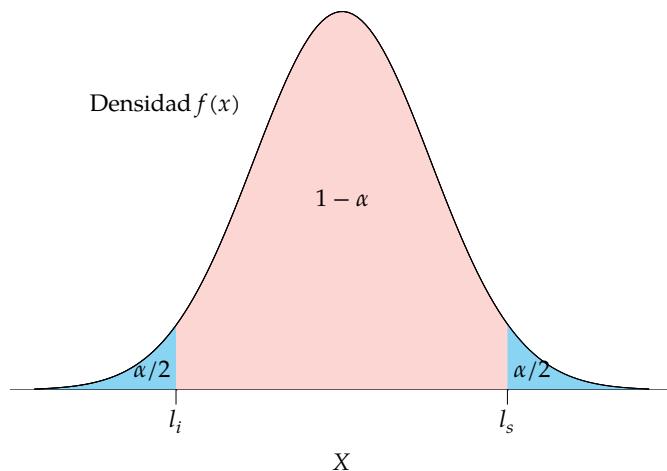
- El tamaño muestral. Cuanto mayor sea el tamaño muestral, más preciso será el intervalo.

Si la confianza y la precisión están reñidas, ¿cómo se puede ganar precisión sin perder confianza?

Habitualmente, para calcular un intervalo de confianza se suele partir de un estimador puntual del que se conoce su distribución muestral.

A partir de este estimador se calculan los extremos del intervalo sobre su distribución, buscando los valores que dejan encerrada una probabilidad $1 - \alpha$. Estos valores suelen tomarse de manera simétrica, de manera que el extremo inferior deje una probabilidad acumulada inferior $\alpha/2$ y el extremo superior deje una probabilidad acumulada superior también de $\alpha/2$.

Distribución del estimador de referencia



6.5 Intervalos de confianza para una población

A continuación se presentan los intervalos de confianza para estimar un parámetro de una población:

- Intervalo para la media de una población normal con varianza conocida.
- Intervalo para la media de una población normal con varianza desconocida.
- Intervalo para la media de una población con varianza desconocida a partir de muestras grandes.
- Intervalo para la varianza de una población normal.
- Intervalo para una proporción de una población.

6.5.1 Intervalo de confianza para la media de una población normal con varianza conocida

Sea X una variable aleatoria que cumple las siguientes hipótesis:

- Su distribución es normal $X \sim N(\mu, \sigma)$.
- La media μ es desconocida, pero su varianza σ^2 es conocida.

Bajo estas hipótesis, la media muestral, para muestras de tamaño n , sigue también una distribución normal

$$\bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

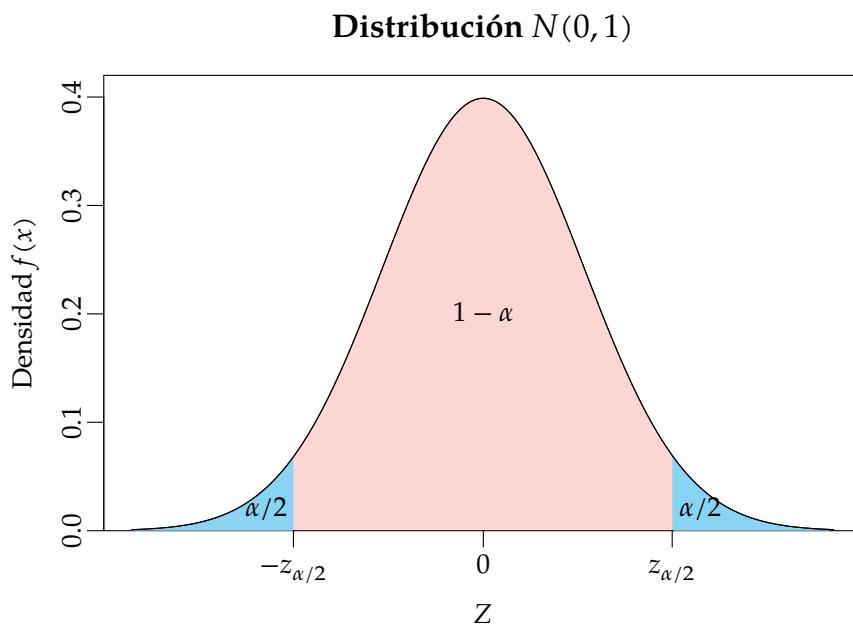
Tipificando la variable se tiene

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

Sobre esta distribución resulta sencillo calcular los valores z_i y z_s de manera que

$$P(z_i \leq Z \leq z_s) = 1 - \alpha.$$

Como la distribución normal estándar es simétrica respecto al 0, lo mejor es tomar valores opuestos $-z_{\alpha/2}$ y $z_{\alpha/2}$ que dejen sendas colas de probabilidad acumulada $\alpha/2$.



A partir de aquí, deshaciendo la tipificación, resulta sencillo llegar a los estimadores que darán los extremos del intervalo de confianza:

$$\begin{aligned}
1 - \alpha &= P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = P\left(-z_{\alpha/2} \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}\right) = \\
&= P\left(-z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \bar{X} - \mu \leq z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\right) = \\
&= P\left(-\bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq -\mu \leq -\bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\right) = \\
&= P\left(\bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\right).
\end{aligned}$$

Así pues, el intervalo de confianza para la media de una población normal con varianza conocida es:

Teorema 6.2 (Intervalo de confianza para la media de una población normal con varianza conocida). *Si $X \sim N(\mu, \sigma)$ con σ conocida, el intervalo de confianza para la media μ con nivel de confianza $1 - \alpha$ es*

$$\left[\bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right]$$

o bien

$$\bar{X} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

De la fórmula del intervalo de confianza

$$\bar{X} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

se deducen varias características:

- a. El intervalo está centrado en la media muestral \bar{X} que era el mejor estimador de la media poblacional.
- b. La amplitud o imprecisión del intervalo es

$$A = 2z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

de manera que depende de:

- σ : cuanto mayor sea la varianza poblacional, mayor será la imprecisión.

- $z_{\alpha/2}$: que a su vez depende del nivel de confianza, y cuanto mayor sea $1 - \alpha$, mayor será la imprecisión.
- n : cuanto mayor sea el tamaño de la muestra, menor será la imprecisión.

Por tanto, la única forma de reducir la imprecisión del intervalo, manteniendo la confianza, es aumentando el tamaño muestral.

6.5.1.1 Cálculo del tamaño muestra para estimar la media de una población normal con varianza conocida

Teniendo en cuenta que la amplitud o imprecisión del intervalo para la media de una población normal con varianza conocida es

$$A = 2z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

se puede calcular fácilmente el tamaño muestral necesario para conseguir un intervalo de amplitud A con confianza $1 - \alpha$:

$$A = 2z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \Leftrightarrow \sqrt{n} = 2z_{\alpha/2} \frac{\sigma}{A},$$

de donde se deduce

$$n = 4z_{\alpha/2}^2 \frac{\sigma^2}{A^2}$$

Ejemplo 6.5. Sea una población de estudiantes en la que la puntuación obtenida en un examen sigue una distribución normal $X \sim N(\mu, \sigma = 1.5)$.

Para estimar la nota media μ , se toma una muestra de 10 estudiantes:

$$4 - 6 - 8 - 7 - 7 - 6 - 5 - 2 - 5 - 3$$

A partir de esta muestra, podemos calcular el intervalo de confianza para μ con un nivel de confianza $1 - \alpha = 0.95$ (nivel de significación $\alpha = 0.05$):

- $\bar{X} = \frac{4+6+8+7+7+6+5+2+5+3}{10} = 5.3$ puntos.
- $z_{\alpha/2} = z_{0.025}$ es el valor de la normal estándar que deja una probabilidad acumulada superior de 0.025, que vale aproximadamente 1.96.

Sustituyendo estos valores en la fórmula del intervalo, se tiene

$$\bar{X} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}} = 5.3 \pm 1.96 \frac{1.5}{\sqrt{10}} = 5.3 \pm 0.93 = [4.37, 6.23].$$

Es decir, μ estaría entre 4.37 y 6.23 puntos con un 95% de confianza.

Ejemplo 6.6. La imprecisión del intervalo anterior es de ± 0.93 puntos.

Si se desea reducir esta imprecisión a ± 0.5 puntos, *¿qué tamaño muestral sería necesario?*

$$n = 4z_{\alpha/2}^2 \frac{\sigma^2}{A^2} = 4 \cdot 1.96^2 \frac{1.5^2}{(2 \cdot 0.5)^2} = 34.57.$$

Por tanto, se necesitaría una muestra de al menos 35 estudiantes para conseguir un intervalo del 95% de confianza y una precisión de ± 0.5 puntos.

6.5.2 Intervalo de confianza para la media de una población normal con varianza desconocida

Sea X una variable aleatoria que cumple las siguientes hipótesis:

- Su distribución es normal $X \sim N(\mu, \sigma)$.
- Tanto su media μ como su varianza σ^2 son desconocidas.

Cuando se desconoce la varianza poblacional se suele estimar mediante la cuasivarianza \hat{S}^2 . Como consecuencia, el estimador de referencia ya no sigue una distribución normal como en el caso de conocer la varianza, sino un T de Student de $n-1$ grados de libertad:

$$\left. \begin{array}{l} \bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right) \\ \frac{(n-1)\hat{S}^2}{\sigma^2} \sim \chi^2(n-1) \end{array} \right\} \Rightarrow \frac{\bar{X} - \mu}{\hat{S}/\sqrt{n}} \sim T(n-1),$$

Como la distribución T de Student, al igual que la normal, también es simétrica respecto al 0, se pueden tomar dos valores opuestos $-t_{\alpha/2}^{n-1}$ y $t_{\alpha/2}^{n-1}$ de manera que

$$\begin{aligned} 1 - \alpha &= P\left(-t_{\alpha/2}^{n-1} \leq \frac{\bar{X} - \mu}{\hat{S}/\sqrt{n}} \leq t_{\alpha/2}^{n-1}\right) \\ &= P\left(-t_{\alpha/2}^{n-1} \frac{\hat{S}}{\sqrt{n}} \leq \bar{X} - \mu \leq t_{\alpha/2}^{n-1} \frac{\hat{S}}{\sqrt{n}}\right) \\ &= P\left(\bar{X} - t_{\alpha/2}^{n-1} \frac{\hat{S}}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{\alpha/2}^{n-1} \frac{\hat{S}}{\sqrt{n}}\right) \end{aligned}$$

Teorema 6.3 (Intervalo de confianza para la media de una población normal con varianza desconocida). *Si $X \sim N(\mu, \sigma)$ con σ desconocida, el intervalo de confianza para la media μ con nivel de confianza $1 - \alpha$ es*

$$\left[\bar{X} - t_{\alpha/2}^{n-1} \frac{\hat{S}}{\sqrt{n}}, \bar{X} + t_{\alpha/2}^{n-1} \frac{\hat{S}}{\sqrt{n}} \right]$$

o bien

$$\bar{X} \pm t_{\alpha/2}^{n-1} \frac{\hat{S}}{\sqrt{n}}$$

6.5.2.1 Calculo del tamaño muestral para estimar la media de una población normal con varianza desconocida

Al igual que antes, teniendo en cuenta que la amplitud o imprecisión del intervalo para la media de una población con varianza desconocida es

$$A = 2t_{\alpha/2}^{n-1} \frac{\hat{S}}{\sqrt{n}}$$

se puede calcular fácilmente el tamaño muestral necesario para conseguir un intervalo de amplitud A con confianza $1 - \alpha$:

$$A = 2t_{\alpha/2}^{n-1} \frac{\hat{S}}{\sqrt{n}} \Leftrightarrow \sqrt{n} = 2t_{\alpha/2}^{n-1} \frac{\hat{S}}{A},$$

de donde se deduce

$$n = 4(t_{\alpha/2}^{n-1})^2 \frac{\hat{S}^2}{A^2}$$

El único problema, a diferencia del caso anterior en que σ era conocida, es que se necesita \hat{S} , por lo que se suele tomar una muestra pequeña previa para calcularla. Por otro lado, el valor de la T de student suele aproximarse asintóticamente por el de la normal estándar $t_{\alpha/2}^{n-1} \approx z_{\alpha/2}$.

Ejemplo 6.7. Supóngase que en el ejemplo anterior no se conoce la varianza poblacional de las puntuaciones.

Trabajando con la misma muestra de las puntuaciones de 10 estudiantes

$$4 - 6 - 8 - 7 - 7 - 6 - 5 - 2 - 5 - 3$$

se puede calcular el intervalo de confianza para μ con un nivel de confianza $1 - \alpha = 0.95$ (nivel de significación $\alpha = 0.05$):

- $\bar{X} = \frac{4+...+3}{10} = \frac{53}{10} = 5.3$ puntos.
- $\hat{S}^2 = \frac{(4-5.3)^2 + \dots + (3-5.3)^2}{9} = 3.5667$ y $\hat{S} = \sqrt{3.5667} = 1.8886$ puntos.
- $t_{\alpha/2}^{n-1} = t_{0.025}^9$ es el valor de la T de Student de 9 grados de libertad, que deja una probabilidad acumulada superior de 0.025, que vale 2.2622.

Sustituyendo estos valores en la fórmula del intervalo, se tiene

$$\bar{X} \pm t_{\alpha/2}^{n-1} \frac{\hat{S}}{\sqrt{n}} = 5.3 \pm 2.2622 \frac{1.8886}{\sqrt{10}} = 5.3 \pm 1.351 = [3.949, 6.651].$$

Ejemplo 6.8. Como se puede apreciar, la imprecisión del intervalo anterior es de ± 1.8886 puntos, que es significativamente mayor que en el caso de conocer la varianza de la población. Esto es lógico pues al tener que estimar la varianza de la población, el error de la estimación se agrega al error del intervalo.

Ahora, el tamaño muestral necesario para reducir la imprecisión a ± 0.5 puntos es

$$n = 4(z_{\alpha/2})^2 \frac{\hat{S}^2}{A^2} = 4 \cdot 1.96^2 \frac{3.5667}{(2 \cdot 0.5)^2} = 54.81.$$

Por tanto, si se desconoce la varianza de la población se necesita una muestra de al menos 55 estudiantes para conseguir un intervalo del 95% de confianza y una precisión de ± 0.5 puntos.

6.5.3 Intervalo de confianza para la media de una población no normal

Sea X una variable aleatoria que cumple las siguientes hipótesis:

- Su distribución no es normal.
- Tanto su media μ como su varianza σ^2 son desconocidas.

Si la población no es normal las distribuciones de los estimadores de referencia cambian, de manera que los intervalos anteriores no son válidos.

No obstante, si la muestra es grande ($n \geq 30$), de acuerdo al teorema central del límite, la distribución de la media muestral se aproximarán a una normal, de modo que sigue siendo cierto

$$\bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

En consecuencia, sigue siendo válido el intervalo anterior.

Teorema 6.4 (Intervalo de confianza para la media de una población no normal con muestras grandes). *Si X es una variable con distribución no normal y $n \geq 30$, el intervalo de confianza para la media μ con nivel de confianza $1 - \alpha$ es*

$$\bar{X} \pm t_{\alpha/2}^{n-1} \frac{\hat{S}}{\sqrt{n}}$$

6.5.4 Intervalo de confianza para la varianza de una población normal

Sea X una variable aleatoria que cumple las siguientes hipótesis:

1. Su distribución es normal $X \sim N(\mu, \sigma)$.
2. Tanto su media μ como su varianza σ^2 son desconocidas.

Para estimar la varianza de una población normal, se parte del estimador de referencia

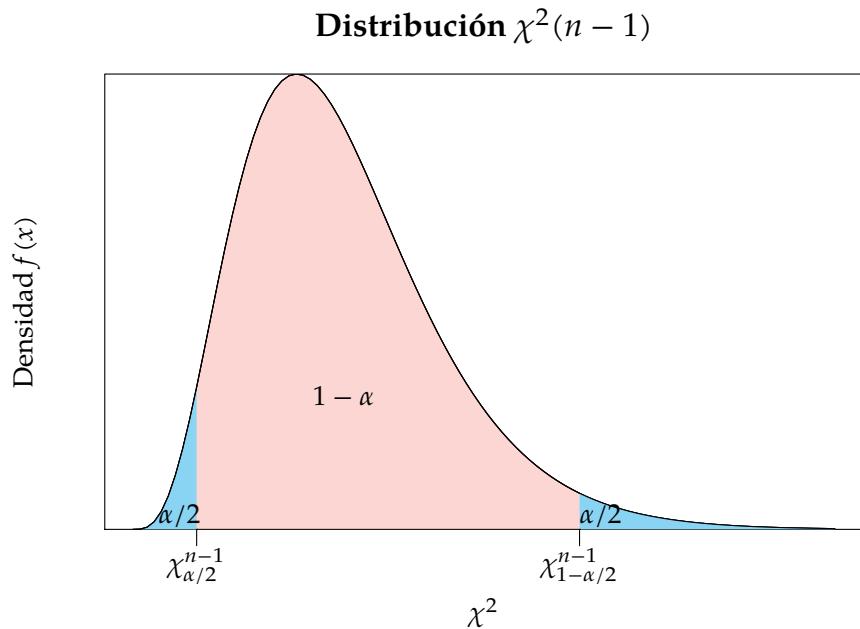
$$\frac{nS^2}{\sigma^2} = \frac{(n-1)\hat{S}^2}{\sigma^2} \sim \chi^2(n-1),$$

que sigue una distribución chi-cuadrado de $n-1$ grados de libertad.

Sobre esta distribución hay que calcular los valores χ_i y χ_s tales que

$$P(\chi_i \leq \chi^2(n-1) \leq \chi_s) = 1 - \alpha.$$

Como la distribución chi-cuadrado no es simétrica respecto al 0, se toman dos valores $\chi_{\alpha/2}^{n-1}$ y $\chi_{1-\alpha/2}^{n-1}$ que dejen sendas colas de probabilidad acumulada inferior de $\alpha/2$ y $1-\alpha/2$ respectivamente.



Así pues, se tiene

$$\begin{aligned}
 1 - \alpha &= P\left(\chi_{\alpha/2}^{n-1} \leq \frac{nS^2}{\sigma^2} \leq \chi_{1-\alpha/2}^{n-1}\right) = P\left(\frac{1}{\chi_{\alpha/2}^{n-1}} \geq \frac{\sigma^2}{nS^2} \geq \frac{1}{\chi_{1-\alpha/2}^{n-1}}\right) = \\
 &= P\left(\frac{1}{\chi_{1-\alpha/2}^{n-1}} \leq \frac{\sigma^2}{nS^2} \leq \frac{1}{\chi_{\alpha/2}^{n-1}}\right) = P\left(\frac{nS^2}{\chi_{1-\alpha/2}^{n-1}} \leq \sigma^2 \leq \frac{nS^2}{\chi_{\alpha/2}^{n-1}}\right).
 \end{aligned}$$

Por tanto, el intervalo de confianza para la varianza de una población normal es:

Teorema 6.5 (Intervalo de confianza para la varianza de una población normal). *Si $X \sim N(\mu, \sigma)$ con σ conocida, el intervalo de confianza para la varianza σ^2 con nivel de confianza $1 - \alpha$ es*

$$\left[\frac{nS^2}{\chi_{1-\alpha/2}^{n-1}}, \frac{nS^2}{\chi_{\alpha/2}^{n-1}} \right]$$

Ejemplo 6.9. Siguiendo con el ejemplo de las puntuaciones en un examen, si se quiere estimar la varianza a partir de la muestra:

$$4 - 6 - 8 - 7 - 7 - 6 - 5 - 2 - 5 - 3$$

para el intervalo de confianza para σ^2 con un nivel de confianza $1 - \alpha = 0.95$ (nivel de significación $\alpha = 0.05$) se tiene:

- $S^2 = \frac{(4-5.3)^2 + \dots + (3-5.3)^2}{10} = 3.21$ puntos².
- $\chi_{\alpha/2}^{n-1} = \chi_{0.025}^9$ es el valor de la chi-cuadrado de 9 grados de libertad, que deja una probabilidad acumulada inferior de 0.025, y vale 2.7.
- $\chi_{1-\alpha/2}^{n-1} = \chi_{0.975}^9$ es el valor de la chi-cuadrado de 9 grados de libertad, que deja una probabilidad acumulada inferior de 0.975, y vale 19.

Sustituyendo estos valores en la fórmula del intervalo, se llega a

$$\left[\frac{nS^2}{\chi_{1-\alpha/2}^{n-1}}, \frac{nS^2}{\chi_{\alpha/2}^{n-1}} \right] = \left[\frac{10 \cdot 3.21}{19}, \frac{10 \cdot 3.21}{2.7} \right] = [1.69, 11.89] \text{ puntos}^2.$$

6.5.5 Intervalo de confianza para una proporción

Para estimar la proporción p de individuos de una población que presentan una determinada característica, se parte de la variable que mide el número de individuos que la presentan en una muestra de tamaño n . Dicha variable sigue una distribución binomial

$$X \sim B(n, p)$$

Como ya se vio, si el tamaño muestral es suficientemente grande (en realidad basta que se cumpla $np \geq 5$ y $n(1-p) \geq 5$), el teorema central de límite asegura que X tendrá una distribución aproximadamente normal

$$X \sim N(np, \sqrt{np(1-p)}).$$

En consecuencia, la proporción muestral \hat{p} también será normal

$$\hat{p} = \frac{X}{n} \sim N \left(p, \sqrt{\frac{p(1-p)}{n}} \right),$$

que es el estimador de referencia.

Trabajando con la distribución del estimador de referencia

$$\hat{p} \sim N \left(p, \sqrt{\frac{p(1-p)}{n}} \right)$$

tras tipificar, se pueden encontrar fácilmente, al igual que hicimos antes, valores $-z_{\alpha/2}$ y $z_{\alpha/2}$ que cumplan

$$P \left(-z_{\alpha/2} \leq \frac{\hat{p} - p}{\sqrt{p(1-p)/n}} \leq z_{\alpha/2} \right) = 1 - \alpha.$$

Así pues, deshaciendo la tipificación y razonando como antes, se tiene

$$\begin{aligned} 1 - \alpha &= P \left(-z_{\alpha/2} \leq \frac{\hat{p} - p}{\sqrt{p(1-p)/n}} \leq z_{\alpha/2} \right) \\ &= P \left(-z_{\alpha/2} \frac{\sqrt{p(1-p)}}{n} \leq \hat{p} - p \leq z_{\alpha/2} \frac{\sqrt{p(1-p)}}{n} \right) \\ &= P \left(\hat{p} - z_{\alpha/2} \frac{\sqrt{p(1-p)}}{n} \leq p \leq \hat{p} + z_{\alpha/2} \frac{\sqrt{p(1-p)}}{n} \right) \end{aligned}$$

Por tanto, el intervalo de confianza para una proporción es

Teorema 6.6 (Intervalo de confianza para una proporción). *Si $X \sim B(n, p)$, y se cumple que $np \geq 5$ y $n(1-p) \geq 5$, entonces el intervalo de confianza para la proporción p con nivel de confianza $1 - \alpha$ es*

$$\left[\hat{p} - z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right]$$

o bien

$$\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

6.5.5.1 Cálculo del tamaño muestra para estimar una proporción

La amplitud o imprecisión del intervalo para la proporción de una población es

$$A = 2z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

así que se puede calcular fácilmente el tamaño muestral necesario para conseguir un intervalo de amplitud A con confianza $1 - \alpha$:

$$A = 2z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \Leftrightarrow A^2 = 4z_{\alpha/2}^2 \frac{\hat{p}(1-\hat{p})}{n},$$

de donde se deduce

$$n = 4z_{\alpha/2}^2 \frac{\hat{p}(1-\hat{p})}{A^2}$$

Para poder hacer el cálculo se necesita una estimación de la proporción \hat{p} , por lo que suele tomarse una muestra previa pequeña para calcularla. En el peor de los casos, si no se dispone de una muestra previa, puede tomarse $\hat{p} = 0.5$.

Ejemplo 6.10. Supóngase que se quiere estimar la proporción de fumadores que hay en una determinada población. Para ello se toma una muestra de 20 personas y se observa si fuman (1) o no (0):

0 – 1 – 1 – 0 – 0 – 0 – 1 – 0 – 0 – 0 – 0 – 1 – 1 – 0 – 1 – 1 – 0 – 0

Entonces:

- $\hat{p} = \frac{8}{20} = 0.4$, por tanto, se cumple $np = 20 \cdot 0.4 = 8 \geq 5$ y $n(1-p) = 20 \cdot 0.6 = 12 \geq 5$.
- $z_{\alpha/2} = z_{0.025}$ es el valor de la normal estándar que deja una probabilidad acumulada superior de 0.025, que vale aproximadamente 1.96.

Sustituyendo estos valores en la fórmula del intervalo, se tiene

$$\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.4 \pm 1.96 \sqrt{\frac{0.4 \cdot 0.6}{10}} = 0.4 \pm 0.3 = [0.1, 0.7].$$

Es decir, p estaría entre 0.1 y 0.7 con un 95% de confianza.

Ejemplo 6.11. Como se puede apreciar la imprecisión del intervalo anterior es ± 0.3 , que es enorme teniendo en cuenta que se trata de un intervalo para una proporción.

Para conseguir intervalos precisos para estimar proporciones se necesitan tamaños muestrales bastante grandes. Si por ejemplo se quiere una precisión de ± 0.05 , el tamaño muestral necesario sería:

$$n = 4z_{\alpha/2}^2 \frac{\hat{p}(1-\hat{p})}{A^2} = 4 \cdot 1.96^2 \frac{0.4 \cdot 0.6}{(2 \cdot 0.05)^2} = 368.79.$$

Es decir, se necesitarían al menos 369 individuos para conseguir un intervalo para la proporción con una confianza del 95%.

6.6 Intervalos de confianza para la comparación dos poblaciones

En muchos estudios el objetivo en sí no es averiguar el valor de un parámetro, sino compararlo con el de otra población. Por ejemplo, comparar si un determinado parámetro vale lo mismo en la población de hombres y en la de mujeres.

En estos casos no interesa realmente estimar los dos parámetros por separado, sino hacer una estimación que permita su comparación.

Se verán tres casos:

- **Comparación de medias:** Se estima la diferencia de medias $\mu_1 - \mu_2$.
- **Comparación de varianzas:** Se estima la razón de varianzas $\frac{\sigma_1^2}{\sigma_2^2}$.
- **Comparación de proporciones:** Se estima la diferencia de proporciones $\hat{p}_1 - \hat{p}_2$.

A continuación se presentan los siguientes intervalos de confianza para la comparación de dos poblaciones:

- Intervalo para la diferencia de medias de dos poblaciones normales con varianzas conocidas.
- Intervalo para la diferencia de medias de dos poblaciones normales con varianzas desconocidas pero iguales.
- Intervalo para la diferencia de medias de dos poblaciones normales con varianzas desconocidas y diferentes.
- Intervalo para el cociente de varianzas de dos poblaciones normales.
- Intervalo para la diferencia de proporciones de dos poblaciones.

6.6.1 Intervalo de confianza para la diferencia de medias de poblaciones normales con varianzas conocidas

Sean X_1 y X_2 dos variables aleatorias que cumplen las siguientes hipótesis:

1. Su distribución es normal $X_1 \sim N(\mu_1, \sigma_1)$ y $X_2 \sim N(\mu_2, \sigma_2)$.
2. Sus medias μ_1 y μ_2 son desconocidas, pero sus varianzas σ_1^2 y σ_2^2 son conocidas.

Bajo estas hipótesis, si se toman dos muestras independientes, una de cada población, de tamaños n_1 y n_2 respectivamente, la diferencia de las medias muestrales sigue una distribución normal

$$\left. \begin{array}{l} \bar{X}_1 \sim N\left(\mu_1, \frac{\sigma_1^2}{\sqrt{n_1}}\right) \\ \bar{X}_2 \sim N\left(\mu_2, \frac{\sigma_2^2}{\sqrt{n_2}}\right) \end{array} \right\} \Rightarrow \bar{X}_1 - \bar{X}_2 \sim N\left(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right).$$

A partir de aquí, tipificando, se pueden buscar los valores de la normal estándar $-z_{\alpha/2}$ y $z_{\alpha/2}$ que cumplen:

$$P \left(-z_{\alpha/2} \leq \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \leq z_{\alpha/2} \right) = 1 - \alpha.$$

Y deshaciendo la tipificación, se tiene

$$\begin{aligned} 1 - \alpha &= P \left(-z_{\alpha/2} \leq \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \leq z_{\alpha/2} \right) \\ &= P \left(-z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \leq (\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2) \leq z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right) \\ &= P \left(\bar{X}_1 - \bar{X}_2 - z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \leq \mu_1 - \mu_2 \leq \bar{X}_1 - \bar{X}_2 + z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right) \end{aligned}$$

Así pues, el intervalo de confianza para la diferencia de medias es

Teorema 6.7 (Intervalo de confianza para la diferencia de medias de poblaciones normales con varianzas conocidas). *Si $X_1 \sim N(\mu_1, \sigma_1)$ y $X_2 \sim N(\mu_2, \sigma_2)$, con σ_1 y σ_2 conocidas, el intervalo de confianza para la diferencia de medias $\mu_1 - \mu_2$ con nivel de confianza $1 - \alpha$ es*

$$\left[\bar{X}_1 - \bar{X}_2 - z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}, \bar{X}_1 - \bar{X}_2 + z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right]$$

o bien

$$\bar{X}_1 - \bar{X}_2 \pm z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

6.6.2 Intervalo de confianza para la diferencia de medias de dos poblaciones normales con varianzas desconocidas e iguales

Sean X_1 y X_2 dos variables aleatorias que cumplen las siguientes hipótesis:

- Su distribución es normal $X_1 \sim N(\mu_1, \sigma_1)$ y $X_2 \sim N(\mu_2, \sigma_2)$.

- Sus medias μ_1 y μ_2 son desconocidas y sus varianzas también, pero son iguales $\sigma_1^2 = \sigma_2^2 = \sigma^2$.

Cuando se desconoce la varianza poblacional se puede estimar a partir de las muestras de tamaños n_1 y n_2 de ambas poblaciones mediante la *cuasivarianza ponderada*:

$$\hat{S}_p^2 = \frac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2 - 2}.$$

El estimador de referencia en este caso sigue una distribución T de Student:

$$\left. \begin{array}{l} \bar{X}_1 - \bar{X}_2 \sim N\left(\mu_1 - \mu_2, \sigma \sqrt{\frac{n_1+n_2}{n_1 n_2}}\right) \\ \frac{n_1 S_1^2 + n_2 S_2^2}{\sigma^2} \sim \chi^2(n_1 + n_2 - 2) \end{array} \right\} \Rightarrow \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\hat{S}_p \sqrt{\frac{n_1+n_2}{n_1 n_2}}} \sim T(n_1 + n_2 - 2).$$

A partir de aquí, se pueden buscar los valores de la T de Student $-t_{\alpha/2}^{n_1+n_2-2}$ y $t_{\alpha/2}^{n_1+n_2-2}$ que cumplen

$$P\left(-t_{\alpha/2}^{n_1+n_2-2} \leq \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\hat{S}_p \sqrt{\frac{n_1+n_2}{n_1 n_2}}} \leq t_{\alpha/2}^{n_1+n_2-2}\right) = 1 - \alpha.$$

Y deshaciendo la transformación se tiene

$$\begin{aligned} 1 - \alpha &= P\left(-t_{\alpha/2}^{n_1+n_2-2} \leq \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\hat{S}_p \sqrt{\frac{n_1+n_2}{n_1 n_2}}} \leq t_{\alpha/2}^{n_1+n_2-2}\right) \\ &= P\left(-t_{\alpha/2}^{n_1+n_2-2} \hat{S}_p \sqrt{\frac{n_1+n_2}{n_1 n_2}} \leq (\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2) \leq t_{\alpha/2}^{n_1+n_2-2} \hat{S}_p \sqrt{\frac{n_1+n_2}{n_1 n_2}}\right) \\ &= P\left(\bar{X}_1 - \bar{X}_2 - t_{\alpha/2}^{n_1+n_2-2} \hat{S}_p \sqrt{\frac{n_1+n_2}{n_1 n_2}} \leq \mu_1 - \mu_2 \leq \bar{X}_1 - \bar{X}_2 + t_{\alpha/2}^{n_1+n_2-2} \hat{S}_p \sqrt{\frac{n_1+n_2}{n_1 n_2}}\right). \end{aligned}$$

Así pues, el intervalo de confianza para la diferencia de medias es

Teorema 6.8 (Intervalo de confianza para la diferencia de medias de poblaciones normales con varianzas desconocidas iguales). *Si $X_1 \sim N(\mu_1, \sigma_1)$ y $X_2 \sim N(\mu_2, \sigma_2)$, con $\sigma_1 = \sigma_2$ desconocidas, el intervalo de confianza para la diferencia de medias $\mu_1 - \mu_2$ con nivel de confianza $1 - \alpha$ es*

$$\left[\bar{X}_1 - \bar{X}_2 - t_{\alpha/2}^{n_1+n_2-2} \hat{S}_p \sqrt{\frac{n_1+n_2}{n_1 n_2}}, \bar{X}_1 - \bar{X}_2 + t_{\alpha/2}^{n_1+n_2-2} \hat{S}_p \sqrt{\frac{n_1+n_2}{n_1 n_2}} \right]$$

o bien

$$\bar{X}_1 - \bar{X}_2 \pm t_{\alpha/2}^{n_1+n_2-2} \hat{S}_p \sqrt{\frac{n_1+n_2}{n_1 n_2}}$$

Si $[l_i, l_s]$ es un intervalo de confianza de nivel $1 - \alpha$ para la diferencia de medias $\mu_1 - \mu_2$, entonces

$$\mu_1 - \mu_2 \in [l_i, l_s]$$

con una confianza del $1 - \alpha\%$.

Por consiguiente, según los valores del intervalo de confianza se tiene:

- Si todos los valores del intervalo son negativos ($l_s < 0$), entonces se puede concluir que $\mu_1 - \mu_2 < 0$ y por tanto $\mu_1 < \mu_2$.
- Si todos los valores del intervalo son positivos ($l_i > 0$), entonces se puede concluir que $\mu_1 - \mu_2 > 0$ y por tanto $\mu_1 > \mu_2$.
- Si el intervalo tiene tanto valores positivos como negativos, y por tanto contiene al 0 ($0 \in [l_i, l_s]$), entonces no se puede afirmar que una media sea mayor que la otra. En este caso se suele asumir la hipótesis de que las medias son iguales $\mu_1 = \mu_2$.

Tanto en el primer como en el segundo caso se dice que entre las medias hay diferencias *estadísticamente significativas*.

Ejemplo 6.12. Supóngase que se quiere comparar el rendimiento académico de dos grupos de alumnos, uno con 10 alumnos y otro con 12, que han seguido metodologías diferentes. Para ello se les realiza un examen y se obtienen las siguientes puntuaciones:

$$\begin{aligned} X_1 : & 4 - 6 - 8 - 7 - 7 - 6 - 5 - 2 - 5 - 3 \\ X_2 : & 8 - 9 - 5 - 3 - 8 - 7 - 8 - 6 - 8 - 7 - 5 - 7 \end{aligned}$$

Si se supone que ambas variables tienen la misma varianza, se tiene

- $\bar{X}_1 = \frac{4+...+3}{10} = 5.3$ y $\bar{X}_2 = \frac{8+...+7}{12} = 6.75$ puntos.
- $S_1^2 = \frac{4^2+...+3^2}{10} - 5.3^2 = 3.21$ y $S_2^2 = \frac{8^2+...+7^2}{12} - 6.75^2 = 2.6875$ puntos².
- $\hat{S}_p^2 = \frac{10 \cdot 3.21 + 12 \cdot 2.6875}{10+12-2} = 3.2175$ puntos², y $\hat{S}_p = 1.7937$.

- $t_{\alpha/2}^{n_1+n_2-2} = t_{0.025}^{20}$ es el valor de la T de Student de 20 grados de libertad que deja una probabilidad acumulada superior de 0.025, y que vale aproximadamente 2.09.

Y sustituyendo en la fórmula del intervalo llegamos a

$$5.3 - 6.75 \pm 2.086 \cdot 1.7937 \sqrt{\frac{10 + 12}{10 \cdot 12}} = -1.45 \pm 1.6021 = [-3.0521, 0.1521] \text{ puntos.}$$

Es decir, la diferencia de puntuaciones medias $\mu_1 - \mu_2$ está entre -3.0521 y 0.1521 puntos con una confianza del 95%.

A la vista del intervalo se puede concluir que, puesto que el intervalo contiene tanto valores positivos como negativos, y por tanto contiene al 0, no puede afirmarse que una de las medias sea mayor que la otra, de modo que se supone que son iguales y no se puede decir que haya diferencias significativas entre los grupos.

6.6.3 Intervalo de confianza para la diferencia de medias de dos poblaciones normales con varianzas desconocidas y distintas

Sean X_1 y X_2 dos variables aleatorias que cumplen las siguientes hipótesis:

- Su distribución es normal $X_1 \sim N(\mu_1, \sigma_1)$ y $X_2 \sim N(\mu_2, \sigma_2)$.
- Sus medias μ_1, μ_2 y varianzas σ_1^2, σ_2^2 , son desconocidas, pero $\sigma_1^2 \neq \sigma_2^2$.

En este caso el estimador de referencia sigue una distribución T de Student

$$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\hat{S}_1^2}{n_1} + \frac{\hat{S}_2^2}{n_2}}} \sim T(g),$$

donde el número de grados de libertad es $g = n_1 + n_2 - 2 - \Delta$, siendo

$$\Delta = \frac{\left(\frac{n_2-1}{n_1} \hat{S}_1^2 - \frac{n_1-1}{n_2} \hat{S}_2^2 \right)^2}{\frac{n_2-1}{n_1^2} \hat{S}_1^4 + \frac{n_1-1}{n_2^2} \hat{S}_2^4}.$$

A partir de aquí, una vez más, se pueden buscar los valores de la T de Student $-t_{\alpha/2}^g$ y $t_{\alpha/2}^g$ que cumplen

$$P \left(-t_{\alpha/2}^g \leq \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\hat{S}_1^2}{n_1} + \frac{\hat{S}_2^2}{n_2}}} \leq t_{\alpha/2}^g \right) = 1 - \alpha.$$

Y deshaciendo la transformación se llega a

$$\begin{aligned}
1 - \alpha &= P \left(-t_{\alpha/2}^g \leq \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\hat{S}_1^2}{n_1} + \frac{\hat{S}_2^2}}}} \leq t_{\alpha/2}^g \right) \\
&= P \left(-t_{\alpha/2}^g \sqrt{\frac{\hat{S}_1^2}{n_1} + \frac{\hat{S}_2^2}} \leq (\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2) \leq t_{\alpha/2}^g \sqrt{\frac{\hat{S}_1^2}{n_1} + \frac{\hat{S}_2^2}} \right) \\
&= P \left(\bar{X}_1 - \bar{X}_2 - t_{\alpha/2}^g \sqrt{\frac{\hat{S}_1^2}{n_1} + \frac{\hat{S}_2^2}} \leq \mu_1 - \mu_2 \leq \bar{X}_1 - \bar{X}_2 + t_{\alpha/2}^g \sqrt{\frac{\hat{S}_1^2}{n_1} + \frac{\hat{S}_2^2}} \right)
\end{aligned}$$

Así pues, el intervalo de confianza para la diferencia de medias es

Teorema 6.9 (Intervalo de confianza para la diferencia de medias de poblaciones normales con varianzas desconocidas distintas). *Si $X_1 \sim N(\mu_1, \sigma_1)$ y $X_2 \sim N(\mu_2, \sigma_2)$, con $\sigma_1 \neq \sigma_2$ desconocidas, el intervalo de confianza para la diferencia de medias $\mu_1 - \mu_2$ con nivel de confianza $1 - \alpha$ es*

$$\left[\bar{X}_1 - \bar{X}_2 - t_{\alpha/2}^g \sqrt{\frac{\hat{S}_1^2}{n_1} + \frac{\hat{S}_2^2}}, \bar{X}_1 - \bar{X}_2 + t_{\alpha/2}^g \sqrt{\frac{\hat{S}_1^2}{n_1} + \frac{\hat{S}_2^2}} \right]$$

o bien

$$\bar{X}_1 - \bar{X}_2 \pm t_{\alpha/2}^g \sqrt{\frac{\hat{S}_1^2}{n_1} + \frac{\hat{S}_2^2}}$$

Como se acaba de ver, existen dos intervalos posibles para estimar la diferencia de medias: uno para cuando las varianzas poblacionales son iguales y otro para cuando no lo son.

Ahora bien, si las varianzas poblacionales son desconocidas,

¿cómo saber qué intervalo utilizar?

La respuesta está en el próximo intervalo que se verá, que permite estimar la razón de varianzas $\frac{\sigma_2^2}{\sigma_1^2}$ y por tanto, su comparación.

Así pues, antes de calcular el intervalo de confianza para la comparación de medias, cuando las varianzas poblacionales sean desconocidas, es necesario calcular el intervalo de confianza para la razón de varianzas y elegir el intervalo para la comparación de medias en función del valor de dicho intervalo.

6.6.4 Intervalo de confianza para el cociente de varianzas

Sean X_1 y X_2 dos variables aleatorias que cumplen las siguientes hipótesis:

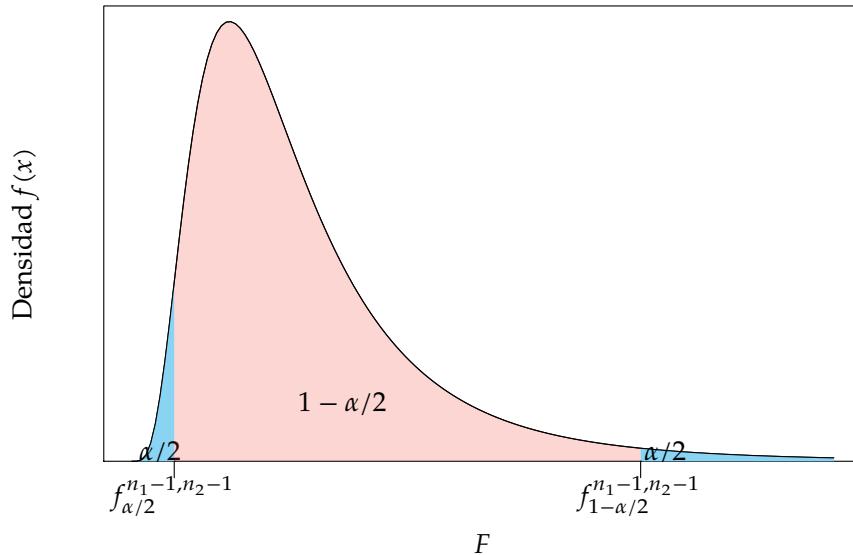
- Su distribución es normal $X_1 \sim N(\mu_1, \sigma_1^2)$ y $X_2 \sim N(\mu_2, \sigma_2^2)$.
- Sus medias μ_1 , μ_2 y varianzas σ_1^2 , σ_2^2 son desconocidas.

En este caso, para muestras de ambas poblaciones de tamaños n_1 y n_2 respectivamente, el estimador de referencia sigue una distribución F de Fisher-Snedecor:

$$\left. \begin{array}{l} \frac{(n_1 - 1)\hat{S}_1^2}{\sigma_1^2} \sim \chi^2(n_1 - 1) \\ \frac{(n_2 - 1)\hat{S}_2^2}{\sigma_2^2} \sim \chi^2(n_2 - 1) \end{array} \right\} \Rightarrow \frac{\frac{(n_2 - 1)\hat{S}_2^2}{\sigma_2^2}}{\frac{(n_1 - 1)\hat{S}_1^2}{\sigma_1^2}} = \frac{\sigma_1^2}{\sigma_2^2} \frac{\hat{S}_2^2}{\hat{S}_1^2} \sim F(n_2 - 1, n_1 - 1).$$

Como la distribución F de Fisher-Snedecor no es simétrica respecto al 0, se toman dos valores $f_{\alpha/2}^{n_2-1, n_1-1}$ y $f_{1-\alpha/2}^{n_2-1, n_1-1}$ que dejen sendas colas de probabilidad acumulada inferior de $\alpha/2$ y $1 - \alpha/2$ respectivamente.

Distribución $F(n_1 - 1, n_2 - 1)$



Así pues, se tiene

$$\begin{aligned} 1 - \alpha &= P \left(f_{\alpha/2}^{n_2-1, n_1-1} \leq \frac{\sigma_1^2}{\sigma_2^2} \frac{\hat{S}_2^2}{\hat{S}_1^2} \leq f_{1-\alpha/2}^{n_2-1, n_1-1} \right) = \\ &= P \left(f_{\alpha/2}^{n_2-1, n_1-1} \frac{\hat{S}_1^2}{\hat{S}_2^2} \leq \frac{\sigma_1^2}{\sigma_2^2} \leq f_{1-\alpha/2}^{n_2-1, n_1-1} \frac{\hat{S}_1^2}{\hat{S}_2^2} \right) \end{aligned}$$

Por tanto, el intervalo de confianza para la comparación de varianzas de dos poblaciones normales es

Teorema 6.10 (Intervalo de confianza para el cociente de varianzas de poblaciones normales). *Si $X_1 \sim N(\mu_1, \sigma_1)$ y $X_2 \sim N(\mu_2, \sigma_2)$, el intervalo de confianza para el cociente de varianzas σ_1^2/σ_2^2 con nivel de confianza $1 - \alpha$ es*

$$\left[f_{\alpha/2}^{n_2-1, n_1-1} \frac{\hat{S}_1^2}{\hat{S}_2^2}, f_{1-\alpha/2}^{n_2-1, n_1-1} \frac{\hat{S}_1^2}{\hat{S}_2^2} \right]$$

Si $[l_i, l_s]$ es un intervalo de confianza de nivel $1 - \alpha$ para la razón de varianzas $\frac{\sigma_1^2}{\sigma_2^2}$, entonces

$$\frac{\sigma_1^2}{\sigma_2^2} \in [l_i, l_s]$$

con una confianza del $1 - \alpha\%$.

Por consiguiente, según los valores del intervalo de confianza se tiene:

- Si todos los valores del intervalo son menores que 1 ($l_s < 1$), entonces se puede concluir que $\frac{\sigma_1^2}{\sigma_2^2} < 1$ y por tanto $\sigma_1^2 < \sigma_2^2$.
- Si todos los valores del intervalo son mayores que 1 ($l_i > 1$), entonces se puede concluir que $\frac{\sigma_1^2}{\sigma_2^2} > 1$ y por tanto $\sigma_1^2 > \sigma_2^2$.
- Si el intervalo tiene tanto valores mayores como menores que 1, y por tanto contiene al 1 ($1 \in [l_i, l_s]$), entonces no se puede afirmar que una varianza sea mayor que la otra. En este caso se suele asumir la hipótesis de que las varianzas son iguales $\sigma_1^2 = \sigma_2^2$.

Ejemplo 6.13. Siguiendo con el ejemplo de las puntuaciones en dos grupos:

$$\begin{aligned} X_1 &: 4 - 6 - 8 - 7 - 7 - 6 - 5 - 2 - 5 - 3 \\ X_2 &: 8 - 9 - 5 - 3 - 8 - 7 - 8 - 6 - 8 - 7 - 5 - 7 \end{aligned}$$

Para calcular el intervalo de confianza para la razón de varianzas con una confianza del 95%, se tiene:

- $\bar{X}_1 = \frac{4+...+3}{10} = 5.3$ puntos y $\bar{X}_2 = \frac{8+...+7}{12} = 6.75$ puntos.
- $\hat{S}_1^2 = \frac{(4-5.3)^2 + \dots + (3-5.3)^2}{9} = 3.5667$ puntos² y $\hat{S}_2^2 = \frac{(8-6.75)^2 + \dots + (7-6.75)^2}{11} = 2.9318$ puntos².
- $f_{\alpha/2}^{n_2-1, n_1-1} = f_{0.025}^{11,9}$ es el valor de la F de Fisher de 11 y 9 grados de libertad que deja una probabilidad acumulada inferior de 0.025, y que vale aproximadamente 0.2787.
- $f_{1-\alpha/2}^{n_2-1, n_1-1} = f_{0.975}^{11,9}$ es el valor de la F de Fisher de 11 y 9 grados de libertad que deja una probabilidad acumulada inferior de 0.975, y que vale aproximadamente 3.9121.

Sustituyendo en la fórmula del intervalo se llega a

$$\left[0.2787 \frac{3.5667}{2.9318}, 3.9121 \frac{3.5667}{2.9318} \right] = [0.3391, 4.7591] \text{ puntos}^2.$$

Es decir, la razón de varianzas $\frac{\sigma_1^2}{\sigma_2^2}$ está entre 0.3391 y 4.7591 con una confianza del 95%.

Como el intervalo tiene tanto valores menores como mayores que 1, no se puede concluir que una varianza sea mayor que la otra, y por tanto se mantiene la hipótesis de que ambas varianzas son iguales.

Si ahora se quisiesen comparar las medias de ambas poblaciones, el intervalo de confianza para la diferencia de medias que habría que tomar es el que parte de la hipótesis de igualdad de varianzas, que precisamente es el que se ha utilizado antes.

6.6.5 Intervalo de confianza para la diferencia de proporciones

Para comparar las proporciones p_1 y p_2 de individuos que presentan una determinada característica en dos poblaciones independientes, se estima su diferencia $p_1 - p_2$.

Si se toma una muestra de cada población, de tamaños n_1 y n_2 respectivamente, las variables que miden el número de individuos que presentan la característica en cada una de ellas siguen distribuciones

$$X_1 \sim B(n_1, p_1) \quad \text{y} \quad X_2 \sim B(n_2, p_2)$$

Cuando los tamaños muestrales son grandes (en realidad basta que se cumpla $n_1 p_1 \geq 5$, $n_1(1-p_1) \geq 5$, $n_2 p_2 \geq 5$ y $n_2(1-p_2) \geq 5$), el teorema central de límite asegura que X_1 y X_2 tendrán distribuciones normales

$$X_1 \sim N(n_1 p_1, \sqrt{n_1 p_1 (1 - p_1)}) \quad \text{y} \quad X_2 \sim N(n_2 p_2, \sqrt{n_2 p_2 (1 - p_2)}),$$

y las proporciones muestrales

$$\hat{p}_1 = \frac{X_1}{n_1} \sim N\left(p_1, \sqrt{\frac{p_1(1-p_1)}{n_1}}\right) \quad \text{y} \quad \hat{p}_2 = \frac{X_2}{n_2} \sim N\left(p_2, \sqrt{\frac{p_2(1-p_2)}{n_2}}\right)$$

A partir de las proporciones muestrales se construye el estimador de referencia

$$\hat{p}_1 - \hat{p}_2 \sim N\left(p_1 - p_2, \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}\right).$$

Tipificando, se buscan valores $-z_{\alpha/2}$ y $z_{\alpha/2}$ que cumplan

$$P\left(-z_{\alpha/2} \leq \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \leq z_{\alpha/2}\right) = 1 - \alpha.$$

Y deshaciendo la tipificación, se llega a

$$\begin{aligned} 1 - \alpha &= P\left(-z_{\alpha/2} \leq \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \leq z_{\alpha/2}\right) \\ &= P\left(-z_{\alpha/2}\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} \leq (\hat{p}_1 - \hat{p}_2) - (p_1 - p_2) \leq z_{\alpha/2}\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}\right) \\ &= P\left(\hat{p}_1 - \hat{p}_2 - z_{\alpha/2}\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} \leq \hat{p}_1 - \hat{p}_2 + p_1 - p_2 \leq z_{\alpha/2}\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}\right) \end{aligned}$$

Así pues, el intervalo de confianza para la diferencia de proporciones es

Teorema 6.11 (Intervalo de confianza para la diferencia de proporciones). *Si $X_1 \sim B(n_1, p_1)$ y $X_2 \sim B(n_2, p_2)$, con $n_1 p_1 \geq 5$, $n_1(1-p_1) \geq 5$, $n_2 p_2 \geq 5$ y $n_2(1-p_2) \geq 5$, el intervalo de confianza para la diferencia de proporciones $p_1 - p_2$ con nivel de confianza $1 - \alpha$ es*

$$\hat{p}_1 - \hat{p}_2 \pm z_{\alpha/2} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

Ejemplo 6.14. Supóngase que se quieren comparar las proporciones o porcentajes de aprobados en dos grupos que han seguido metodologías distintas. En el primer grupo han aprobado 24 alumnos de un total de 40, mientras que en el segundo han aprobado 48 de 60.

Para calcular el intervalo de confianza para la diferencia de proporciones con un nivel de confianza del 95%, se tiene:

- $\hat{p}_1 = 24/40 = 0.6$ y $\hat{p}_2 = 48/60 = 0.8$, de manera que se cumplen las hipótesis $n_1\hat{p}_1 = 40 \cdot 0.6 = 24 \geq 5$, $n_1(1-\hat{p}_1) = 40(1-0.6) = 26 \geq 5$, $n_2\hat{p}_2 = 60 \cdot 0.8 = 48 \geq 5$ y $n_2(1-\hat{p}_2) = 60(1-0.8) = 12 \geq 5$.
- $z_{\alpha/2} = z_{0.025} = 1.96$.

Sustituyendo en la fórmula del intervalo se tiene

$$0.6 - 0.8 \pm 1.96 \sqrt{\frac{0.6(1-0.6)}{40} + \frac{0.8(1-0.8)}{60}} = -0.2 \pm 0.17 = [-0.37, -0.03].$$

Como el intervalo es negativo se tiene $p_1 - p_2 < 0 \Rightarrow p_1 < p_2$, y se puede concluir que hay diferencias significativas en el porcentaje de aprobados.

7 Contrastes de hipótesis paramétricos

7.1 Hipótesis estadística y tipos de contrastes

En muchos estudios estadísticos, el objetivo, más que estimar el valor de un parámetro desconocido en la población, es comprobar la veracidad de una hipótesis formulada sobre la población objeto de estudio.

El investigador, de acuerdo a su experiencia o a estudios previos, suele tener conjeturas sobre la población estudiada que expresa en forma de hipótesis.

Definición 7.1 (Hipótesis estadística). Una *hipótesis estadística* es cualquier afirmación o conjetura que determina, total o parcialmente, la distribución de una o varias variables de la población.

Ejemplo 7.1. Para contrastar el rendimiento académico de un grupo de alumnos en una determinada asignatura, podríamos plantear la hipótesis de si el porcentaje de aprobados es mayor del 50%.

7.1.1 Contraste de hipótesis

En general nunca se sabrá con absoluta certeza si una hipótesis estadística es cierta o falsa, ya que para ello habría que estudiar a todos los individuos de la población.

Para comprobar la veracidad o falsedad de estas hipótesis hay que contrastarlas con los resultados empíricos obtenidos de las muestras. Si los resultados observados en las muestras coinciden, dentro del margen de error admisible debido al azar, con lo que cabría esperar en caso de que la hipótesis fuese cierta, la hipótesis se aceptará como verdadera, mientras que en caso contrario se rechazará como falsa y se buscarán nuevas hipótesis capaces de explicar los datos observados.

Como las muestras se obtienen aleatoriamente, *la decisión de aceptar o rechazar una hipótesis estadística se tomará sobre una base de probabilidad*.

La metodología que se encarga de contrastar la veracidad de las hipótesis estadísticas se conoce como *contraste de hipótesis*.

7.1.2 Tipos de contrastes de hipótesis

- **Contrastes de bondad de ajuste:** El objetivo es comprobar una hipótesis sobre la forma de la distribución de la población.
Por ejemplo, contrastar si las notas de un grupo de alumnos siguen una distribución normal.
- **Contrastes de conformidad:** El objetivo es comprobar una hipótesis sobre alguno de los parámetros de la población.
Por ejemplo, contrastar si la nota media en un grupo de alumnos es igual a 5.
- **Contrastes de homogeneidad :** El objetivo es comparar dos poblaciones con respecto a alguno de sus parámetros.
Por ejemplo, contrastar si el rendimiento de dos grupos de alumnos es el mismo comparando sus notas medias.
- **Contrastes de independencia:** El objetivo es comprobar si existe relación entre dos variables de la población.
Por ejemplo, contrastar si existe relación entre las notas de dos asignaturas diferentes.

Cuando las hipótesis se plantean sobre parámetros de la población, también se habla de **contrastos paramétricos**.

7.1.3 Hipótesis nula e hipótesis alternativa

En la mayoría de los casos un contraste supone tomar una decisión entre dos hipótesis antagonistas:

- **Hipótesis nula:** Es la hipótesis conservadora, ya que se mantendrá mientras que los datos de las muestras no reflejen claramente su falsedad. Se representa como H_0 .
- **Hipótesis alternativa:** Es la negación de la hipótesis nula y generalmente representa la afirmación que se pretende probar. Se representa como H_1 .

Ambas hipótesis se eligen de acuerdo con el principio de simplicidad científica:

“Solamente se debe abandonar un modelo simple por otro más complejo cuando la evidencia a favor del último sea fuerte.” (Navaja de Occam)

Por ejemplo, en el caso de un juicio, en el que el juez debe decidir si el acusado es culpable o inocente, la elección de hipótesis debería ser

$$H_0 : \text{Inocente}$$

$$H_1 : \text{Culpable}$$

ya que la inocencia se asume, mientras que la culpabilidad hay que demostrarla.

Según esto, el juez sólo aceptaría la hipótesis alternativa cuando hubiese pruebas significativas de la culpabilidad del acusado.

El investigador jugaría el papel del fiscal, ya que su objetivo consistiría en intentar rechazar la hipótesis nula, es decir, demostrar culpabilidad del acusado.

 Advertencia

¡Esta metodología siempre favorece a la hipótesis nula!

7.1.4 Contrastes de hipótesis paramétricos

En muchos contrastes, sobre todo en las pruebas de conformidad y de homogeneidad, las hipótesis se formulan sobre parámetros desconocidos de la población como puede ser una media, una varianza o una proporción.

En tal caso, la hipótesis nula siempre asigna al parámetro un valor concreto, mientras que la alternativa suele ser una hipótesis abierta que, aunque opuesta a la hipótesis nula, no fija el valor del parámetro.

Esto da lugar a tres tipos de contrastes:

Bilateral	Unilateral menor	Unilateral mayor
$H_0: \theta = \theta_0$	$H_0: \theta = \theta_0$	$H_0: \theta = \theta_0$
$H_1: \theta \neq \theta_0$	$H_1: \theta < \theta_0$	$H_1: \theta > \theta_0$

Ejemplo 7.2. Supóngase que existen sospechas de que en una población hay menos hombres que mujeres.

¿Qué tipo de contraste debería plantearse para validar o refutar esta sospecha?

1. Las sospechas se refieren al porcentaje o la proporción p de hombres en la población, por lo que se trata de un *contraste paramétrico*.
2. El objetivo es averiguar el valor de p , por lo que se trata de una *prueba de conformidad*. En la hipótesis nula el valor de p se fijará a 0.5 ya que, de acuerdo a las leyes de la genética, en la población debería haber la misma proporción de hombres que de mujeres.

- Finalmente, existen sospechas de que el porcentaje de hombres es menor que el de mujeres, por lo que la hipótesis alternativa será de menor $p < 0.5$.

Así pues, el contraste que debería plantearse es el siguiente:

$$\begin{aligned} H_0 : p &= 0.5 \\ H_1 : p &< 0.5 \end{aligned}$$

7.2 Metodología para realizar un contraste de hipótesis

7.2.1 Estadístico del contraste

La aceptación o rechazo de la hipótesis nula depende, en última instancia, de lo que se observe en la muestra.

La decisión se tomará según el valor que presente algún estadístico de la muestra relacionado con el parámetro o característica que se esté contrastando, y cuya distribución de probabilidad debe ser conocida suponiendo cierta la hipótesis nula y una vez fijado el tamaño de la muestra. Este estadístico recibe el nombre de **estadístico del contraste**.

Para cada muestra, el estadístico dará una estimación a partir de la cual se tomará la decisión: *si la estimación difiere demasiado del valor esperado bajo la hipótesis H_0 , entonces se rechazará, y en caso contrario se aceptará.*

La lógica que guía la decisión es la de mantener la hipótesis nula a no ser que en la muestra haya pruebas contundentes de su falsedad. Siguiendo con el símil del juicio, se trataría de mantener la inocencia mientras no haya pruebas claras de culpabilidad.

Ejemplo 7.3. Volviendo al ejemplo del contraste sobre la proporción de hombres de una población

$$\begin{aligned} H_0 : p &= 0.5 \\ H_1 : p &< 0.5 \end{aligned}$$

Si para resolver el contraste se toma una muestra aleatoria de 10 personas, podría tomarse como estadístico del contraste X el número de hombres en la muestra.

Suponiendo cierta la hipótesis nula, el estadístico del contraste seguiría una distribución binomial $X \sim B(10, 0.5)$, de manera que el número esperado de hombres en la muestra sería 5.

Así pues, es lógico aceptar la hipótesis nula si en la muestra se obtiene un número de hombres próximo a 5 y rechazarla cuando el número de hombres sea muy inferior a 5. Pero, *¿dónde poner el límite entre los valores X que lleven a la aceptación y los que lleven al rechazo?*

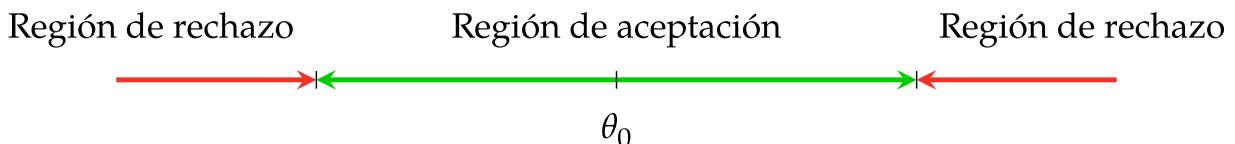
7.2.2 Regiones de aceptación y de rechazo

Una vez elegido el estadístico del contraste, lo siguiente es decidir para qué valores de este estadístico se decidirá aceptar la hipótesis nula y para qué valores se rechazará. Esto divide del conjunto de valores posibles del estadístico en dos regiones:

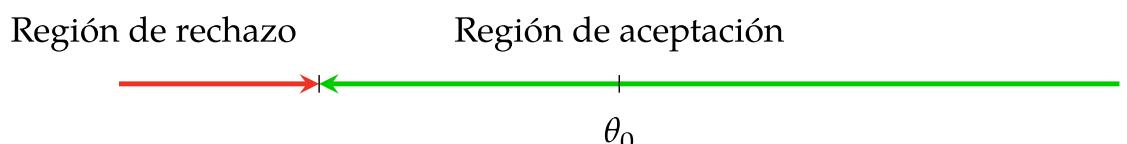
- **Región de aceptación:** Es el conjunto de valores del estadístico del contraste a partir de los cuales se decidirá aceptar la hipótesis nula.
 - **Región de rechazo:** Es el conjunto de valores del estadístico del contraste a partir de los cuales se decidirá rechazar la hipótesis nula, y por tanto, aceptar la hipótesis alternativa.

Dependiendo de la dirección del contraste, la región de rechazo quedará a un lado u otro del valor esperado del estadístico del contraste según la hipótesis nula:

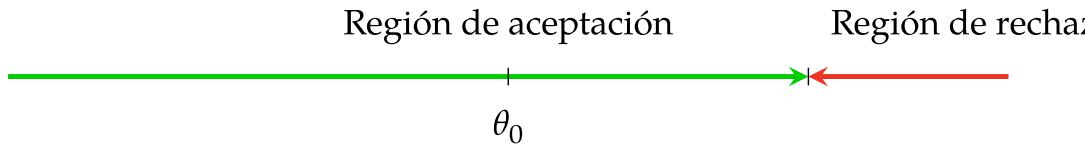
- Contraste bilateral $H_0 : \theta = \theta_0$ $H_1 : \theta \neq \theta_0$.



- Contraste unilateral de menor $H_0 : \theta = \theta_0$ & $H_1 : \theta < \theta_0$.



- Contraste unilateral de mayor $H_0 : \theta = \theta_0$, $H_1 : \theta > \theta_0$.

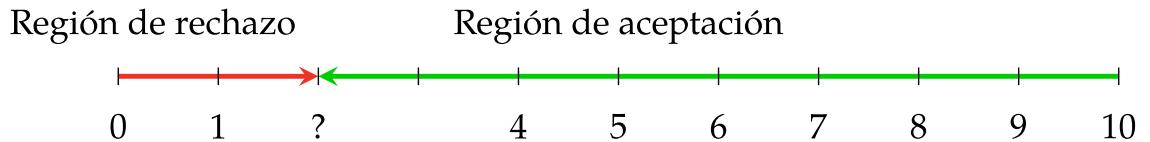


Ejemplo 7.4. Siguiendo con el ejemplo del contraste sobre la proporción de hombres de una población

$$H_0 : p = 0.5$$

$$H_1 : p < 0.5$$

Como el estadístico del contraste tenía una distribución binomial $X \sim B(10, 0.5)$ suponiendo cierta la hipótesis nula, su recorrido será de 0 a 10 y su valor esperado 5, por lo que, al tratarse de un contraste unilateral de menor, la región de rechazo quedará por debajo del 5. Pero, *¿dónde poner el límite entre las regiones de aceptación y de rechazo?*



7.2.3 Errores en un contraste de hipótesis

Hemos visto que un contraste de hipótesis se realiza mediante una regla de decisión que permite aceptar o rechazar la hipótesis nula dependiendo del valor que tome el estadístico del contraste.

Al final el contraste se resuelve tomando una decisión de acuerdo a esta regla. El problema es que nunca se conocerá con absoluta certeza la veracidad o falsedad de una hipótesis, de modo que al aceptarla o rechazarla es posible que se esté tomando una decisión equivocada.

Los errores que se pueden cometer en un contraste de hipótesis son de dos tipos:

- **Error de tipo I:** Se comete cuando se rechaza la hipótesis nula siendo esta verdadera.

- **Error de tipo II:** Se comete cuando se acepta la hipótesis nula siendo esta falsa.

Decisión	H_0 cierta	H_1 cierta
Aceptar H_0	Decisión correcta	Error de tipo II
Rechazar H_0	Error de tipo I	Decisión correcta

7.2.4 Riesgos de los errores de un contraste de hipótesis

Los riesgos de cometer cada tipo de error se cuantifican mediante probabilidades:

Definición 7.2 (Riesgos α y β). En un contraste de hipótesis, se define el *riesgo α* como la máxima probabilidad de cometer un error de tipo I, es decir,

$$P(\text{Rechazar } H_0 | H_0) \leq \alpha,$$

y se define el *riesgo β* como la máxima probabilidad de cometer un error de tipo II, es decir,

$$P(\text{Aceptar } H_0 | H_1) \leq \beta.$$

 **Advertencia**

En principio, puesto que esta metodología favorece a la hipótesis nula, el error del tipo I suele ser más grave que el error del tipo II, y por tanto, el riesgo α suele fijarse a niveles bajos de 0.1, 0.05 o 0.01, siendo 0.05 lo más habitual.

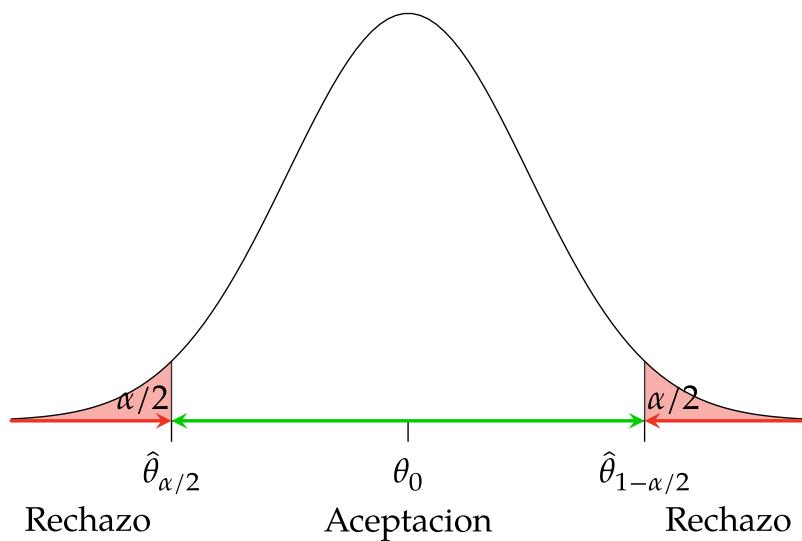
Debe tenerse cuidado al interpretar el riesgo α ya que se trata de una probabilidad condicionada a que la hipótesis nula sea cierta. Por tanto, cuando se rechace la hipótesis nula con un riesgo $\alpha = 0.05$, es erróneo decir 5 de cada 100 veces nos equivocaremos, ya que esto sería cierto sólo si la hipótesis nula fuese siempre verdadera.

Tampoco tiene sentido hablar de la probabilidad de haberse equivocado una vez tomada una decisión a partir de una muestra concreta, pues en tal caso, si se ha tomado la decisión acertada, la probabilidad de error es 0 y si se ha tomado la decisión equivocada, la probabilidad de error es 1.

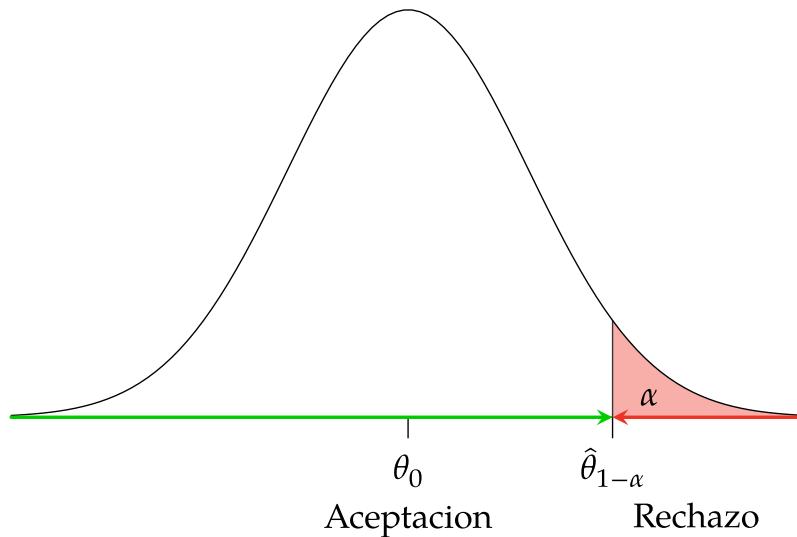
7.2.5 Determinación de las regiones de aceptación y de rechazo en función del riesgo α

Una vez fijado el riesgo α que se está dispuesto a tolerar, es posible delimitar las regiones de aceptación y de rechazo para el estadístico del contraste de manera que la probabilidad acumulada en la región de rechazo sea α , suponiendo cierta la hipótesis nula.

Regiones de un contraste bilateral



Regiones de un contraste unilateral de mayor



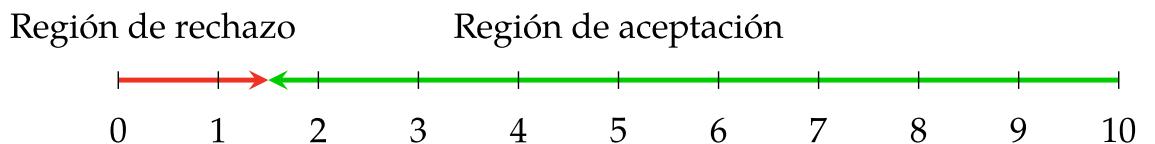
Ejemplo 7.5. Siguiendo con el contraste sobre la proporción de hombres de una población, como el estadístico del contraste sigue una distribución binomial $X \sim B(10, 0.5)$, si se decide rechazar la hipótesis nula cuando en la muestra haya 2 o menos hombres, la probabilidad de cometer un error de tipo I será

$$P(X \leq 2) = f(0) + f(1) + f(2) = 0.0010 + 0.0098 + 0.0439 = 0.0547.$$

Si riesgo máximo de error de tipo I que se está dispuesto a tolerar es $\alpha = 0.05$, ¿qué valores del estadístico permitirán rechazar la hipótesis nula?

$$P(X \leq 1) = f(0) + f(1) = 0.0010 + 0.0098 = 0.0107.$$

Es decir, sólo se podría rechazar la hipótesis nula con 0 o 1 hombres en la muestra.



7.2.6 Riesgo β y tamaño del efecto

Aunque el error de tipo II pueda parecer menos grave, también interesa que el riesgo β sea bajo, ya que de lo contrario será difícil rechazar la hipótesis nula (que es lo que se persigue la mayoría de las veces), aunque haya pruebas muy claras de su falsedad.

El problema, en el caso de contrastes paramétricos, es que la hipótesis alternativa es una hipótesis abierta en la que no se fija el valor del parámetro a contrastar, de modo que, para poder calcular el riesgo β es necesario fijar dicho valor.

Lo normal es fijar el valor del parámetro del contraste a la mínima cantidad para admitir diferencias significativas desde un punto de vista práctico o clínico. Esa mínima diferencia que se considera clínicamente significativa se conoce como **tamaño del efecto** y se representa por δ .

7.2.7 Potencia de un contraste

Puesto que el objetivo del investigador suele ser rechazar la hipótesis nula, a menudo, lo más interesante de un contraste es su capacidad para detectar la falsedad de la hipótesis nula cuando realmente hay diferencias mayores que δ entre el verdadero valor del parámetro y el que establece la hipótesis nula.

Definición 7.3 (Potencia de un contraste). La *potencia* de un contraste de hipótesis se define como

$$\text{Potencia} = P(\text{Rechazar } H_0|H_1) = 1 - P(\text{Aceptar } H_0|H_1) = 1 - \beta.$$

Así pues, al reducir el riesgo β se aumentará la potencia del contraste.

Un contraste poco potente no suele ser interesante ya que no permitirá rechazar la hipótesis nula aunque haya evidencias en su contra.

7.2.8 Cálculo del riesgo β y de la potencia $1 - \beta$

:::{#exm-** Supóngase que en el contraste sobre la proporción de hombres no se considera importante una diferencia de menos de un 10% con respecto al valor que establece la hipótesis nula, es decir, $\delta = 0.1$.

Esto permite fijar la hipótesis alternativa

$$H_1 : p = 0.5 - 0.1 = 0.4.$$

Suponiendo cierta esta hipótesis el estadístico del contraste seguiría una distribución binomial $X \sim B(10, 0.4)$.

En tal caso, el riesgo β para las regiones de aceptación y rechazo fijadas antes será

$$\beta = P(\text{Aceptar } H_0 | H_1) = P(X \geq 2) = 1 - P(X < 2) = 1 - 0.0464 = 0.9536.$$

Como puede apreciarse, se trata de un riesgo β muy alto, por lo que la potencia del contraste sería sólo de

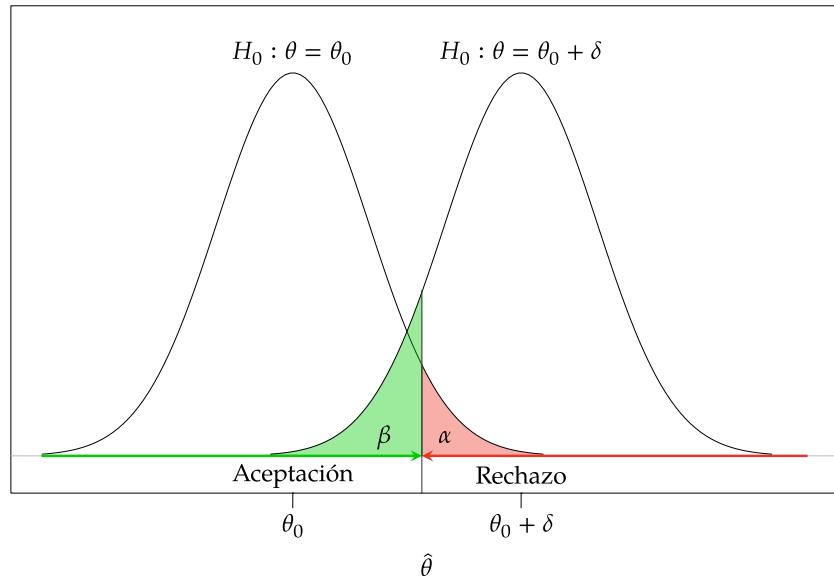
$$1 - \beta = 1 - 0.9536 = 0.0464,$$

lo que indica que no se trataría de un buen contraste para detectar diferencias de un 10% en el valor del parámetro.

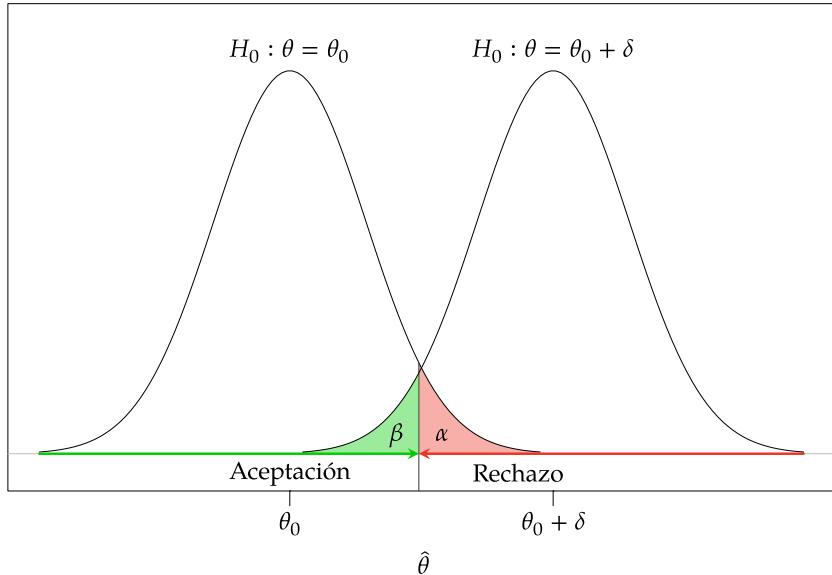
7.2.9 Relación del riesgo β y el tamaño del efecto δ

El riesgo β depende directamente de la mínima diferencia δ que se desea detectar con respecto al valor del parámetro que establece la hipótesis nula.

Relación entre el riesgo β y el tamaño del efecto δ



Relación entre el riesgo β y la mínima diferencia importante δ



Ejemplo 7.6. Si en el contraste sobre la proporción de hombres se desease detectar una diferencia de al menos un 20% con respecto al valor que establece la hipótesis nula, es decir, $\delta = 0.2$, entonces la hipótesis alternativa se fijaría a

$$H_1 : p = 0.5 - 0.2 = 0.3,$$

y bajo esta hipótesis el estadístico del contraste seguiría una distribución binomial $X \sim B(10, 0.3)$.

En tal caso, el riesgo β para las regiones de aceptación y rechazo fijadas antes sería

$$\beta = P(\text{Aceptar } H_0 | H_1) = P(X \geq 2) = 1 - P(X < 2) = 1 - 0.1493 = 0.8507,$$

por lo que el riesgo riesgo β disminuiría y la potencia del contraste aumentaría

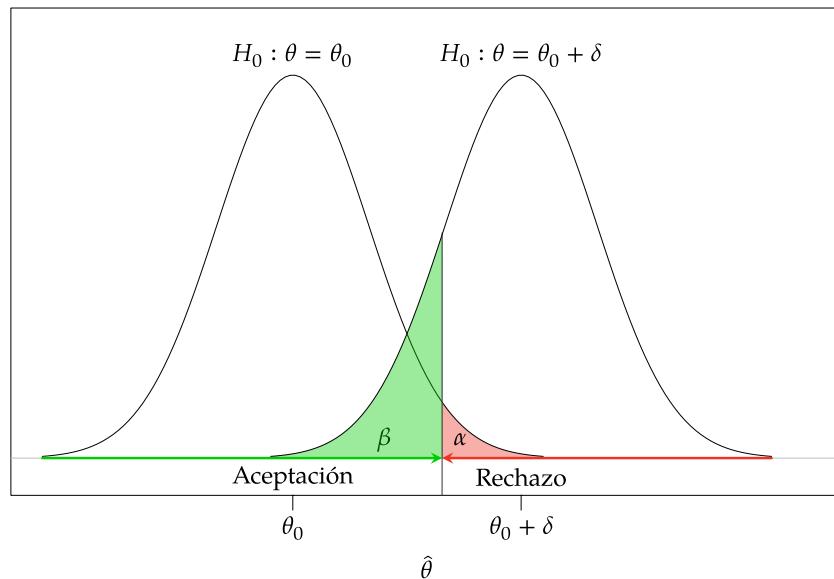
$$1 - \beta = 1 - 0.8507 = 0.1493,$$

aunque seguiría siendo un contraste poco potente.

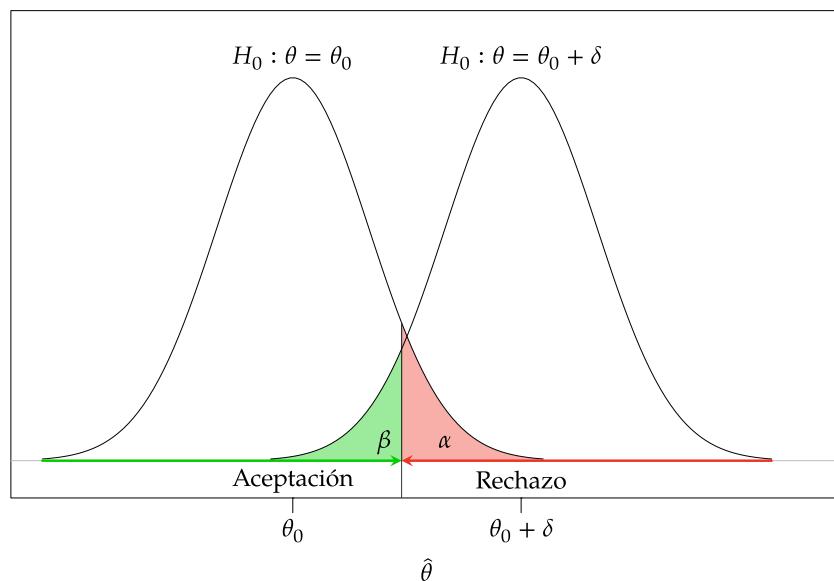
7.2.10 Relación entre los riesgos α y β

Los riesgos α y β están enfrentados, es decir, cuando uno aumenta el otro disminuye y viceversa.

Relación entre los riesgos α y β



Relación entre los riesgos α y β



Ejemplo 7.7. Si en el contraste sobre la proporción de hombres toma como riesgo $\alpha = 0.1$, entonces la región de rechazo sería $X \leq 2$ ya que, suponiendo cierta la hipótesis nula, $X \sim B(10, 0.5)$, y

$$P(X \leq 2) = 0.0547 \leq 0.1 = \alpha.$$

Entonces, para una diferencia mínima $\delta = 0.1$ y suponiendo cierta la hipótesis alternativa, $X \sim B(10, 0.4)$, el riesgo β será

$$\beta = P(\text{Aceptar } H_0 | H_1) = P(X \geq 3) = 1 - P(X < 3) = 1 - 0.1673 = 0.8327,$$

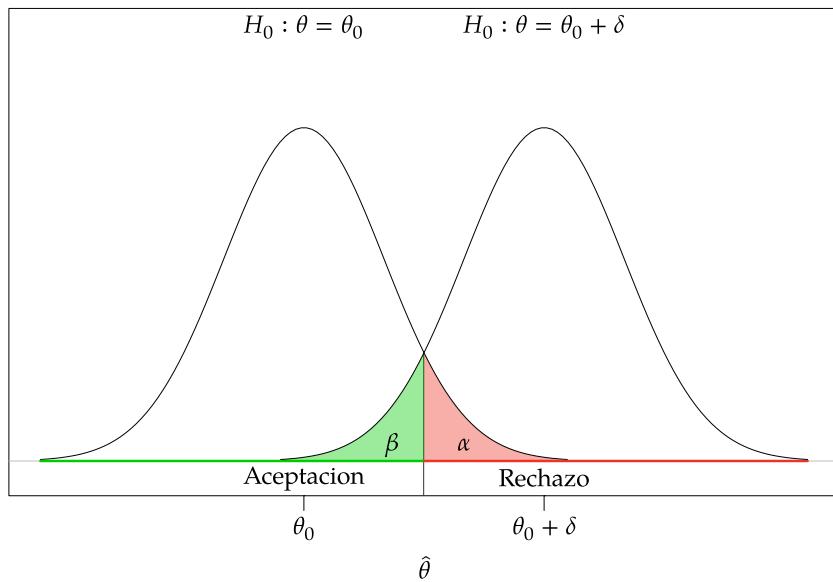
y ahora la potencia ha subido hasta

$$1 - \beta = 1 - 0.8327 = 0.1673.$$

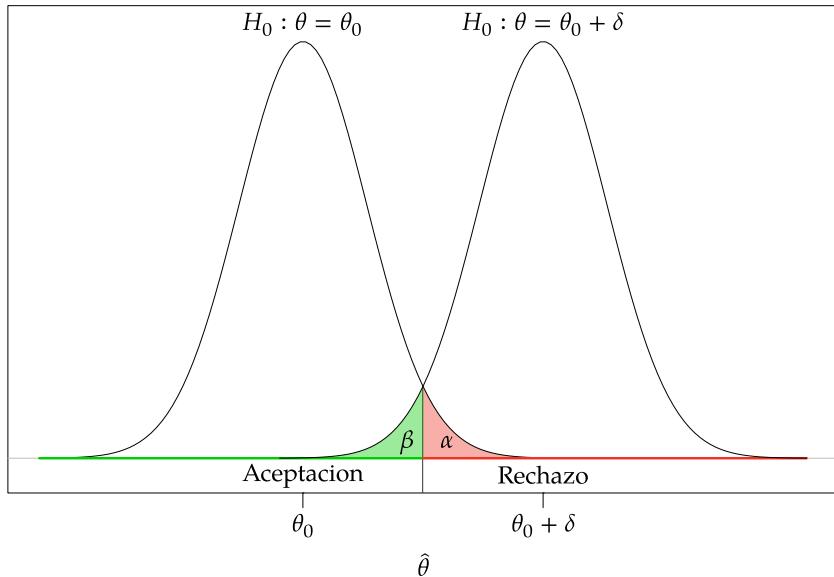
7.2.11 Relación de los riesgos de error y el tamaño muestral

Los riesgos de error también dependen el tamaño de la muestra, ya que al aumentar el tamaño de la muestra, la dispersión del estadístico del contraste disminuye y con ello también lo hacen los riesgos de error.

Riesgos de error para muestras pequeñas



Riesgos de error para muestras grandes



Ejemplo 7.8. Si para realizar el contraste sobre la proporción de hombres se hubiese tomado una muestra de tamaño 100, en lugar de 10, entonces, bajo la suposición de certeza de la hipótesis nula, el estadístico del contraste seguiría una distribución binomial $B(100, 0.5)$, y ahora la región de rechazo sería $X \leq 41$, ya que

$$P(X \leq 41) = 0.0443 \leq 0.05 = \alpha.$$

Entonces, para $\delta = 0.1$ y suponiendo cierta la hipótesis alternativa, $X \sim B(100, 0.4)$, el riesgo β sería

$$\beta = P(\text{Aceptar } H_0 | H_1) = P(X \geq 42) = 0.3775,$$

y ahora la potencia habría aumentado considerablemente

$$1 - \beta = 1 - 0.3775 = 0.6225.$$

Este contraste sería mucho más útil para detectar una diferencia de al menos un 10% con respecto al valor del parámetro que establece la hipótesis nula.

7.3 Curva de potencia

La potencia de un contraste depende del valor del parámetro que establezca la hipótesis alternativa y, por tanto, es una función de este

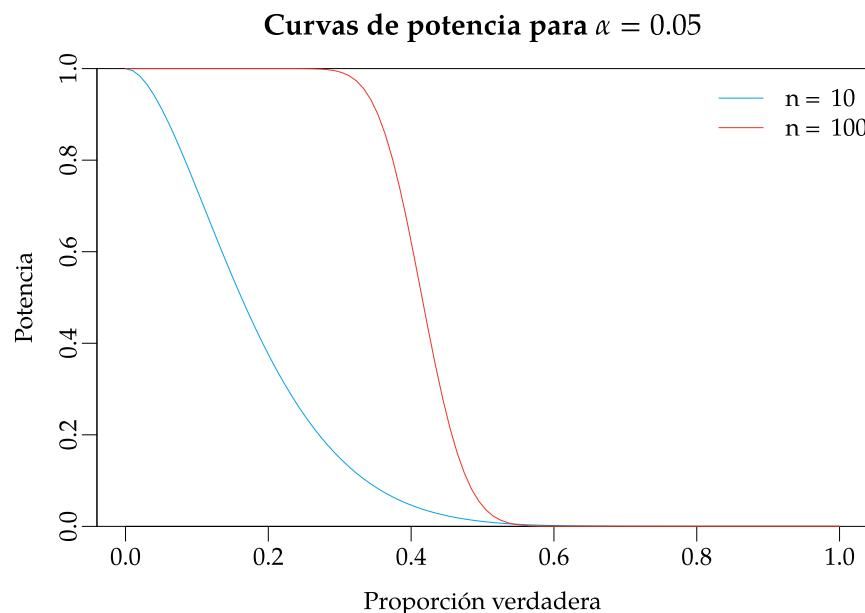
$$\text{Potencia}(x) = P(\text{Rechazar } H_0 | \theta = x).$$

Esta función da la probabilidad de rechazar la hipótesis nula para cada valor del parámetro y se conoce como **curva de potencia**.

Cuando no se puede fijar el valor concreto del parámetro en la hipótesis alternativa, resulta útil representar esta curva para ver la bondad del contraste cuando no se rechaza la hipótesis nula. También es útil cuando sólo se dispone de un número determinado de individuos en la muestra, para ver si merece la pena hacer el estudio.

Un contraste será mejor cuanto mayor sea el área encerrada por debajo de la curva de potencia.

Ejemplo 7.9. La curva de potencia correspondiente al contraste sobre la proporción de hombres en la población es la siguiente



7.3.1 *p*-valor de un contraste de hipótesis

En general, siempre que la estimación del estadístico caiga dentro de la región de rechazo, rechazaremos la hipótesis nula, pero evidentemente, si dicha estimación se aleja bastante de la región de aceptación tendremos más confianza en el rechazo que si la estimación está cerca del límite entre las regiones de aceptación y rechazo.

Por este motivo, al realizar un contraste, también se calcula la probabilidad de obtener una discrepancia mayor o igual a la observada entre la estimación del estadístico del contraste y su valor esperado según la hipótesis nula.

Definición 7.4 (*p*-valor). En un contraste de hipótesis, para cada estimación x_0 del estadístico del contraste X , dependiendo del tipo de contraste, se define el *p*-valor del contraste como

$$\begin{aligned}\text{Contraste bilateral :} & \quad 2P(X \geq x_0 | H_0) \\ \text{Contraste unilateral de menor :} & \quad P(X \leq x_0 | H_0) \\ \text{Contraste unilateral de mayor :} & \quad P(X \geq x_0 | H_0)\end{aligned}$$

En cierto modo, el *p*-valor expresa la confianza que se tiene al tomar la decisión de rechazar la hipótesis nula. Cuanto más próximo esté el *p*-valor a 1, mayor confianza existe al aceptar la hipótesis nula, y cuanto más próximo esté a 0, mayor confianza hay al rechazarla.

7.3.2 Regla de decisión de un contraste

Una vez fijado el riesgo α , la regla de decisión para realizar un contraste también puede expresarse de la siguiente manera:

i Interpretación

Regla de decisión

$$\begin{aligned}\text{Si } p\text{-valor} \leq \alpha & \rightarrow \text{ Rechazar } H_0 \\ \text{Si } p\text{-valor} > \alpha & \rightarrow \text{ Aceptar } H_0.\end{aligned}$$

De este modo, el *p*-valor nos da información de para qué niveles de significación puede rechazarse la hipótesis nula y para cuales no.

Ejemplo 7.10. Si el contraste sobre la proporción de hombres se toma una muestra de tamaño 10 y se observa 1 hombre, entonces el *p*-valor, bajo a supuesta certeza de la hipótesis nula, $X \sim B(10, 0.5)$, será

$$p = P(X \leq 1) = 0.0107,$$

mientras que si en la muestra se observan 0 hombres, entonces el p -valor será

$$p = P(X \leq 0) = 0.001.$$

En el primer caso se rechazaría la hipótesis nula para un riesgo $\alpha = 0.05$, pero no podría rechazarse para un riesgo $\alpha = 0.01$, mientras que en el segundo caso también se rechazaría para $\alpha = 0.01$. Es evidente que en el segundo la decisión de rechazar la hipótesis nula se tomaría con mayor confianza.

7.3.3 Pasos para la realización de un contraste de hipótesis

1. Formular la hipótesis nula H_0 y la alternativa H_1 .
2. Fijar los riesgos α y β deseados.
3. Seleccionar el estadístico del contraste.
4. Fijar la mínima diferencia clínicamente significativa (tamaño del efecto) δ .
5. Calcular el tamaño muestral necesario n .
6. Delimitar las regiones de aceptación y rechazo.
7. Tomar una muestra de tamaño n .
8. Calcular el estadístico del contraste en la muestra.
9. Rechazar la hipótesis nula si la estimación cae en la región de rechazo o bien si el p -valor es menor que el riesgo α y aceptarla en caso contrario.

7.4 Contrastes paramétricos más importantes

Pruebas de conformidad:

- Contraste para la media de una población normal con varianza conocida.
- Contraste para la media de una población normal con varianza desconocida.
- Contraste para la media de una población con varianza desconocida a partir de muestras grandes.
- Contraste para la varianza de una población normal.
- Contraste para una proporción de una población.

Pruebas de homogeneidad:

- Contraste de comparación de medias de dos poblaciones normales con varianzas conocidas.
- Contraste de comparación de medias de dos poblaciones normales con varianzas desconocidas pero iguales.

- Contraste de comparación de medias de dos poblaciones normales con varianzas desconocidas y diferentes.
- Contraste de comparación de varianzas de dos poblaciones normales.
- Contraste de comparación de proporciones de dos poblaciones.

7.5 Contraste para la media de una población normal con varianza conocida

Sea X una variable aleatoria que cumple las siguientes condiciones:

- Su distribución es normal $X \sim N(\mu, \sigma)$.
- La media μ es desconocida, pero su varianza σ^2 es conocida.

Contraste:

$$\begin{aligned} H_0 &: \mu = \mu_0 \\ H_1 &: \mu \neq \mu_0 \end{aligned}$$

Estadístico del contraste:

$$\bar{x} \sim N\left(\mu_0, \frac{\sigma}{\sqrt{n}}\right) \Rightarrow Z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \sim N(0, 1).$$

Región de aceptación: $z_{\alpha/2} < Z < z_{1-\alpha/2}$.

Región de rechazo: $Z \leq z_{\alpha/2}$ y $Z \geq z_{1-\alpha/2}$.

7.6 Contraste para la media de una población normal con varianza desconocida

Sea X una variable aleatoria que cumple las siguientes condiciones:

- Su distribución es normal $X \sim N(\mu, \sigma)$.
- Tanto su media μ como su varianza σ^2 son desconocidas.

Contraste:

$$\begin{aligned} H_0 &: \mu = \mu_0 \\ H_1 &: \mu \neq \mu_0 \end{aligned}$$

Estadístico del contraste: Utilizando la cuasivarianza como estimador de la varianza poblacional se tiene

$$\bar{x} \sim N\left(\mu_0, \frac{\sigma}{\sqrt{n}}\right) \Rightarrow T = \frac{\bar{x} - \mu_0}{\hat{s}/\sqrt{n}} \sim T(n-1).$$

Región de aceptación: $t_{\alpha/2}^{n-1} < T < t_{1-\alpha/2}^{n-1}$.

Región de rechazo: $T \leq t_{\alpha/2}^{n-1}$ y $T \geq t_{1-\alpha/2}^{n-1}$.

Ejemplo 7.11. En un grupo de alumnos se quiere contrastar si la nota media de estadística es mayor que 5 puntos. Para ello se toma la siguiente muestra:

$$6.3, 5.4, 4.1, 5.0, 8.2, 7.6, 6.4, 5.6, 4.3, 5.2$$

El contraste que se plantea es

$$H_0 : \mu = 5 \quad H_1 : \mu > 5$$

Para realizar el contraste se tiene:

- $\bar{x} = \frac{6.3 + \dots + 5.2}{10} = \frac{58.1}{10} = 5.81$ puntos.
- $\hat{s}^2 = \frac{(6.3 - 5.81)^2 + \dots + (5.2 - 5.81)^2}{9} = \frac{15.949}{9} = 1.7721$ puntos², y $\hat{s} = 1.3312$ puntos.

Y el estadístico del contraste vale

$$T = \frac{\bar{x} - \mu_0}{\hat{s}/\sqrt{n}} = \frac{5.81 - 5}{1.3312/\sqrt{10}} = 1.9246.$$

El p -valor del contraste es $P(T(9) \geq 1.9246) = 0.04323$, lo que indica que se rechazaría la hipótesis nula para $\alpha = 0.05$.

La región de rechazo es

$$T = \frac{\bar{x} - 5}{1.3312/\sqrt{10}} \geq t_{0.95}^9 = 1.8331 \Leftrightarrow \bar{x} \geq 5 + 1.8331 \frac{1.3312}{\sqrt{10}} = 5.7717,$$

de modo que se rechazará la hipótesis nula siempre que la media de la muestra sea mayor que 5.7717 y se aceptará en caso contrario.

Suponiendo que en la práctica la mínima diferencia importante en la nota media fuese de un punto $\delta = 1$, entonces bajo la hipótesis alternativa $H_1 : \mu = 6$, si se decidiese rechazar la hipótesis nula, el riesgo β sería

$$\beta = P\left(T(9) \leq \frac{5.7717 - 6}{1.3312\sqrt{10}}\right) = P(T(9) \leq -0.5424) = 0.3004,$$

de manera que la potencia del contraste para detectar una diferencia de $\delta = 1$ punto sería $1 - \beta = 1 - 0.3004 = 0.6996$.

7.6.1 Determinación del tamaño muestral en un contraste para la media

Se ha visto que para un riesgo α la región de rechazo era

$$T = \frac{\bar{x} - \mu_0}{\hat{s}/\sqrt{n}} \geq t_{1-\alpha}^{n-1} \approx z_{1-\alpha} \quad \text{para } n \geq 30.$$

o lo que es equivalente

$$\bar{x} \geq \mu_0 + z_{1-\alpha} \frac{\hat{s}}{\sqrt{n}}.$$

Si el tamaño del efecto es δ , para una hipótesis alternativa $H_1 : \mu = \mu_0 + \delta$, el riesgo β es

$$\beta = P\left(Z < \frac{\mu_0 + z_{1-\alpha} \frac{\hat{s}}{\sqrt{n}} - (\mu_0 + \delta)}{\frac{\hat{s}}{\sqrt{n}}}\right) = P\left(Z < \frac{z_{1-\alpha} \frac{\hat{s}}{\sqrt{n}} - \delta}{\frac{\hat{s}}{\sqrt{n}}}\right).$$

de modo que

$$z_\beta = \frac{z_{1-\alpha} \frac{\hat{s}}{\sqrt{n}} - \delta}{\frac{\hat{s}}{\sqrt{n}}} \Leftrightarrow \delta = (z_{1-\alpha} - z_\beta) \frac{\hat{s}}{\sqrt{n}} \Leftrightarrow n = (z_{1-\alpha} - z_\beta)^2 \frac{\hat{s}^2}{\delta^2} = (z_\alpha + z_\beta)^2 \frac{\hat{s}^2}{\delta^2}.$$

Ejemplo 7.12. Se ha visto en el ejemplo anterior que la potencia del contraste para detectar una diferencia en la nota media de 1 punto era del 69.96%. Para aumentar la potencia del test hasta un 90%, ¿cuántos alumnos habría que tomar en la muestra?

Como se desea una potencia $1 - \beta = 0.9$, el riesgo $\beta = 0.1$ y mirando en la tabla de la normal estándar se puede comprobar que $z_\beta = z_{0.1} = 1.2816$.

Aplicando la fórmula anterior para determinar el tamaño muestral necesario, se tiene

$$n = (z_\alpha + z_\beta)^2 \frac{\hat{s}^2}{\delta^2} = (1.6449 + 1.2816)^2 \frac{1.7721}{1^2} = 15.18,$$

de manera que habría que haber tomado al menos 16 alumnos.

7.7 Contraste para la media de una población con varianza desconocida y muestras grandes

Sea X una variable aleatoria que cumple las siguientes condiciones:

- Su distribución puede ser de cualquier tipo.
- Tanto su media μ como su varianza σ^2 son desconocidas.

Contraste:

$$\begin{aligned} H_0 &: \mu = \mu_0 \\ H_1 &: \mu \neq \mu_0 \end{aligned}$$

Estadístico del contraste: Utilizando la cuasivarianza como estimador de la varianza poblacional y gracias al teorema central del límite por tratarse de muestras grandes ($n \geq 30$) se tiene

$$\bar{x} \sim N\left(\mu_0, \frac{\sigma}{\sqrt{n}}\right) \Rightarrow Z = \frac{\bar{x} - \mu_0}{\hat{s}/\sqrt{n}} \sim N(0, 1).$$

Región de aceptación: $-z_{\alpha/2} < Z < z_{\alpha/2}$.

Región de rechazo: $Z \leq -z_{\alpha/2}$ y $Z \geq z_{\alpha/2}$.

7.8 Contraste para la varianza de una población normal

Sea X una variable aleatoria que cumple las siguientes hipótesis:

- Su distribución es normal $X \sim N(\mu, \sigma)$.
- Tanto su media μ como su varianza σ^2 son desconocidas.

Contraste:

$$\begin{aligned} H_0 &: \sigma = \sigma_0 \\ H_1 &: \sigma \neq \sigma_0 \end{aligned}$$

Estadístico del contraste: Partiendo de la cuasivarianza muestral como estimador de la varianza poblacional, se tiene

$$J = \frac{nS^2}{\sigma_0^2} = \frac{(n-1)\hat{S}^2}{\sigma_0^2} \sim \chi^2(n-1),$$

que sigue una distribución chi-cuadrado de $n - 1$ grados de libertad.

Región de aceptación: $\chi_{\alpha/2}^{n-1} < J < \chi_{1-\alpha/2}^{n-1}$.

Región de rechazo: $J \leq \chi_{\alpha/2}^{n-1}$ y $J \geq \chi_{1-\alpha/2}^{n-1}$.

Ejemplo 7.13. En un grupo de alumnos se quiere contrastar si la desviación típica de la nota es mayor de 1 punto. Para ello se toma la siguiente muestra:

$$6.3, 5.4, 4.1, 5.0, 8.2, 7.6, 6.4, 5.6, 4.3, 5.2$$

El contraste que se plantea es

$$H_0 : \sigma = 1 \quad H_1 : \sigma > 1$$

Para realizar el contraste se tiene:

- $\bar{x} = \frac{6.3+\dots+5.2}{10} = \frac{58.1}{10} = 5.81$ puntos.
- $\hat{s}^2 = \frac{(6.3-5.81)^2+\dots+(5.2-5.81)^2}{9} = \frac{15.949}{9} = 1.7721$ puntos².

El estadístico del contraste vale

$$J = \frac{(n-1)\hat{S}^2}{\sigma_0^2} = \frac{9 \cdot 1.7721}{1^2} = 15.949,$$

y el p -valor del contraste es $P(\chi(9) \geq 15.949) = 0.068$, por lo que no se puede rechazar la hipótesis nula para $\alpha = 0.05$.

7.9 Contraste para proporción de una población

Sea p la proporción de individuos de una población que tienen una determinada característica.

Contraste:

$$\begin{aligned} H_0 &: p = p_0 \\ H_1 &: p \neq p_0 \end{aligned}$$

Estadístico del contraste: La variable que mide el número de individuos con la característica en una muestra aleatoria de tamaño n sigue una distribución binomial $X \sim B(n, p_0)$. De acuerdo al teorema central del límite, para muestras grandes ($np \geq 5$ y $n(1-p) \geq 5$), $X \sim N(np_0, \sqrt{np_0(1-p_0)})$, y se cumple

$$\hat{p} = \frac{X}{n} \sim N\left(p_0, \sqrt{\frac{p_0(1-p_0)}{n}}\right) \Rightarrow Z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} \sim N(0, 1).$$

Región de aceptación: $z_{\alpha/2} < Z < z_{1-\alpha/2}$.

Región de rechazo: $Z \leq z_{\alpha/2}$ y $Z \geq z_{1-\alpha/2}$.

Ejemplo 7.14. En un grupo de alumnos se desea estimar si el porcentaje de aprobados es mayor del 50%. Para ello se toma una muestra de 80 alumnos entre los que hay 50 aprobados.

El contraste que se plantea es

$$\begin{aligned} H_0 : p &= 0.5 \\ H_1 : p &> 0.5 \end{aligned}$$

Para realizar el contraste se tiene que $\hat{p} = 50/80 = 0.625$ y como se cumple $n\hat{p} = 80 \cdot 0.625 = 50 \geq 5$ y $n(1-\hat{p}) = 80(1-0.625) = 30 \geq 5$, el estadístico del contraste vale

$$Z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} = \frac{0.625 - 0.5}{\sqrt{0.5(1-0.5)/80}} = 2.2361.$$

y el p -valor del contraste es $P(Z \geq 2.2361) = 0.0127$, por lo que se rechaza la hipótesis nula para $\alpha = 0.05$ y se concluye que el porcentaje de aprobados es mayor de la mitad.

7.10 Contraste de comparación de medias de dos poblaciones normales con varianzas conocidas

Sean X_1 y X_2 dos variables aleatorias que cumplen las siguientes condiciones:

- Su distribución es normal $X_1 \sim N(\mu_1, \sigma_1^2)$ $X_2 \sim N(\mu_2, \sigma_2^2)$.
- Sus medias μ_1 y μ_2 son desconocidas, pero sus varianzas σ_1^2 y σ_2^2 son conocidas.

Contraste:

$$\begin{aligned} H_0 &: \mu_1 = \mu_2 \\ H_1 &: \mu_1 \neq \mu_2 \end{aligned}$$

Estadístico del contraste:

$$\begin{aligned} \bar{X}_1 &\sim N\left(\mu_1, \frac{\sigma_1}{\sqrt{n_1}}\right) \\ \bar{X}_2 &\sim N\left(\mu_2, \frac{\sigma_2}{\sqrt{n_2}}\right) \end{aligned} \Rightarrow$$

$$\Rightarrow \bar{X}_1 - \bar{X}_2 \sim N\left(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right) \Rightarrow Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0, 1).$$

Región de aceptación: $-z_{\alpha/2} < Z < z_{\alpha/2}$.

Región de rechazo: $Z \leq -z_{\alpha/2}$ y $Z \geq z_{\alpha/2}$.

7.11 Contraste de comparación de medias de dos poblaciones normales con varianzas desconocidas e iguales

Sean X_1 y X_2 dos variables aleatorias que cumplen las siguientes condiciones:

- Su distribución es normal $X_1 \sim N(\mu_1, \sigma_1)$ y $X_2 \sim N(\mu_2, \sigma_2)$.
- Sus medias μ_1 y μ_2 son desconocidas y sus varianzas también, pero son iguales $\sigma_1^2 = \sigma_2^2 = \sigma^2$.

Contraste:

$$\begin{aligned} H_0 &: \mu_1 = \mu_2 \\ H_1 &: \mu_1 \neq \mu_2 \end{aligned}$$

Estadístico del contraste:

$$\begin{aligned} \bar{X}_1 - \bar{X}_2 &\sim N\left(\mu_1 - \mu_2, \sigma \sqrt{\frac{n_1 + n_2}{n_1 n_2}}\right) \\ \frac{n_1 S_1^2 + n_2 S_2^2}{\sigma^2} &\sim \chi^2(n_1 + n_2 - 2) \end{aligned} \Rightarrow T = \frac{\bar{X}_1 - \bar{X}_2}{\hat{S}_p \sqrt{\frac{n_1 + n_2}{n_1 n_2}}} \sim T(n_1 + n_2 - 2).$$

Región de aceptación: $-t_{\alpha/2}^{n_1+n_2-2} < T < t_{\alpha/2}^{n_1+n_2-2}$.

Región de rechazo: $T \leq -t_{\alpha/2}^{n_1+n_2-2}$ y $T \geq t_{\alpha/2}^{n_1+n_2-2}$.

Ejemplo 7.15. Se quiere comparar el rendimiento académico de dos grupos de alumnos, uno con 10 alumnos y otro con 12, que han seguido metodologías diferentes. Para ello se les realiza un examen y se obtienen las siguientes puntuaciones:

$$X_1 : 4 - 6 - 8 - 7 - 7 - 6 - 5 - 2 - 5 - 3 \\ X_2 : 8 - 9 - 5 - 3 - 8 - 7 - 8 - 6 - 8 - 7 - 5 - 7$$

El contraste que se plantea es

$$H_0 : \mu_1 = \mu_2 \quad H_1 : \mu_1 \neq \mu_2$$

Para realizar el contraste, se tiene

- $\bar{X}_1 = \frac{4+...+3}{10} = 5.3$ puntos y $\bar{X}_2 = \frac{8+...+7}{12} = 6.75$ puntos.
- $S_1^2 = \frac{4^2+...+3^2}{10} - 5.3^2 = 3.21$ puntos² y $S_2^2 = \frac{8^2+...+7^2}{12} - 6.75^2 = 2.69$ puntos².
- $\hat{S}_p^2 = \frac{10 \cdot 3.21 + 12 \cdot 2.69}{10+12-2} = 3.2175$ puntos², y $\hat{S}_p = 1.7937$.

Si se suponen varianzas iguales, el estadístico del contraste vale

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\hat{S}_p \sqrt{\frac{n_1+n_2}{n_1 n_2}}} = \frac{5.3 - 6.75}{1.7937 \sqrt{\frac{10+12}{10 \cdot 12}}} = -1.8879,$$

y el p -valor del contraste es $2P(T(20) \leq -1.8879) = 0.0736$, de modo que no se puede rechazar la hipótesis nula y se concluye que no hay diferencias significativas entre las notas medias de los grupos.

7.12 Contraste de comparación de medias de dos poblaciones normales con varianzas desconocidas

Sean X_1 y X_2 dos variables aleatorias que cumplen las siguientes condiciones:

- Su distribución es normal $X_1 \sim N(\mu_1, \sigma_1)$ y $X_2 \sim N(\mu_2, \sigma_2)$.
- Sus medias μ_1 , μ_2 y varianzas σ_1^2 , σ_2^2 , son desconocidas, pero $\sigma_1^2 \neq \sigma_2^2$.

Contraste:

$$H_0 : \mu_1 = \mu_2 \\ H_1 : \mu_1 \neq \mu_2$$

Estadístico del contraste:

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\hat{S}_1^2}{n_1} + \frac{\hat{S}_2^2}{n_2}}} \sim T(g),$$

con $g = n_1 + n_2 - 2 - \Delta$ y

$$\Delta = \frac{\left(\frac{n_2-1}{n_1}\hat{S}_1^2 - \frac{n_1-1}{n_2}\hat{S}_2^2\right)^2}{\frac{n_2-1}{n_1^2}\hat{S}_1^4 + \frac{n_1-1}{n_2^2}\hat{S}_2^4}.$$

Región de aceptación: $-t_{\alpha/2}^g < T < t_{\alpha/2}^g$.

Región de rechazo: $T \leq -t_{\alpha/2}^g$ y $T \geq t_{\alpha/2}^g$.

7.13 Contraste de comparación de varianzas de dos poblaciones normales

Sean X_1 y X_2 dos variables aleatorias que cumplen las siguientes condiciones:

- Su distribución es normal $X_1 \sim N(\mu_1, \sigma_1)$ y $X_2 \sim N(\mu_2, \sigma_2)$.
- Sus medias μ_1, μ_2 y varianzas σ_1^2, σ_2^2 son desconocidas.

Contraste:

$$\begin{aligned} H_0 &: \sigma_1 = \sigma_2 \\ H_1 &: \sigma_1 \neq \sigma_2 \end{aligned}$$

Estadístico del contraste:

$$\left. \begin{aligned} \frac{(n_1-1)\hat{S}_1^2}{\sigma_1^2} &\sim \chi^2(n_1-1) \\ \frac{(n_2-1)\hat{S}_2^2}{\sigma_2^2} &\sim \chi^2(n_2-1) \end{aligned} \right\} \Rightarrow F = \frac{\frac{(n_1-1)\hat{S}_1^2}{\sigma_1^2}}{\frac{(n_2-1)\hat{S}_2^2}{\sigma_2^2}} = \frac{\sigma_2^2}{\sigma_1^2} \frac{\hat{S}_1^2}{\hat{S}_2^2} \sim F(n_1-1, n_2-1).$$

Región de aceptación: $F_{\alpha/2}^{n_1-1, n_2-1} < F < F_{1-\alpha/2}^{n_1-1, n_2-1}$.

Región de rechazo: $F \leq F_{\alpha/2}^{n_1-1, n_2-1}$ y $F \geq F_{1-\alpha/2}^{n_1-1, n_2-1}$.

Ejemplo 7.16. Siguiendo con el ejemplo de las puntuaciones en dos grupos:

$$\begin{aligned} X_1 &: 4 - 6 - 8 - 7 - 7 - 6 - 5 - 2 - 5 - 3 \\ X_2 &: 8 - 9 - 5 - 3 - 8 - 7 - 8 - 6 - 8 - 7 - 5 - 7 \end{aligned}$$

Si se desea comparar las varianzas, el contraste que se plantea es

$$H_0 : \sigma_1 = \sigma_2 \quad H_1 : \sigma_1 \neq \sigma_2$$

Para realizar el contraste, se tiene

- $\bar{X}_1 = \frac{4+\dots+3}{10} = 5.3$ puntos y $\bar{X}_2 = \frac{8+\dots+7}{12} = 6.75$ puntos.
- $\hat{S}_1^2 = \frac{(4-5.3)^2 + \dots + (3-5.3)^2}{9} = 3.5667$ y $\hat{S}_2^2 = \frac{(8-6.75)^2 + \dots + (3-6.75)^2}{11} = 2.9318$ puntos².

El estadístico del contraste vale

$$F = \frac{\hat{S}_1^2}{\hat{S}_2^2} = \frac{3.5667}{2.9318} = 1.2165,$$

y el p -valor del contraste es $2P(F(9, 11) \leq 1.2165) = 0.7468$, por lo que se mantiene la hipótesis de igualdad de varianzas.

7.14 Contraste de comparación de proporciones de dos poblaciones

Sean p_1 y p_2 las respectivas proporciones de individuos que presentan una determinada característica en dos poblaciones.

Contraste:

$$H_0 : p_1 = p_2 \quad H_1 : p_1 \neq p_2$$

Estadístico del contraste: Las variables que miden el número de individuos con la característica en dos muestras aleatorias de tamaños n_1 y n_2 respectivamente, siguen distribuciones binomiales $X_1 \sim B(n_1, p_1)$ y $X_2 \sim B(n_2, p_2)$. Si las muestras son grandes ($n_i p_i \geq 5$ y $n_i(1 - p_i) \geq 5$), de acuerdo al teorema central del límite, $X_1 \sim N(np_1, \sqrt{np_1(1 - p_1)})$ y $X_2 \sim N(np_2, \sqrt{np_2(1 - p_2)})$, y se cumple

$$\left. \begin{array}{l} \hat{p}_1 = \frac{X_1}{n_1} \sim N\left(p_1, \sqrt{\frac{p_1(1-p_1)}{n_1}}\right) \\ \hat{p}_2 = \frac{X_2}{n_2} \sim N\left(p_2, \sqrt{\frac{p_2(1-p_2)}{n_2}}\right) \end{array} \right\} \Rightarrow Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \sim N(0, 1)$$

Región de aceptación: $z_{\alpha/2} < Z < z_{1-\alpha/2}$.

Región de rechazo: $z \leq z_{\alpha/2}$ y $z \geq z_{1-\alpha/2}$.

Ejemplo 7.17. Se quiere comparar los porcentajes de aprobados en dos grupos que han seguido metodologías distintas. En el primer grupo han aprobado 24 alumnos de un total de 40, mientras que en el segundo han aprobado 48 de 60.

El contraste que se plantea es

$$H_0 : p_1 = p_2 \quad H_1 : p_1 \neq p_2$$

Para realizar el contraste, se tiene $\hat{p}_1 = 24/40 = 0.6$ y $\hat{p}_2 = 48/60 = 0.8$, de manera que se cumplen las condiciones $n_1\hat{p}_1 = 40 \cdot 0.6 = 24 \geq 5$, $n_1(1-\hat{p}_1) = 40(1-0.6) = 26 \geq 5$, $n_2\hat{p}_2 = 60 \cdot 0.8 = 48 \geq 5$ y $n_2(1-\hat{p}_2) = 60(1-0.8) = 12 \geq 5$, y el estadístico del contraste vale

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} = \frac{0.6 - 0.8}{\sqrt{\frac{0.6(1-0.6)}{40} + \frac{0.8(1-0.8)}{60}}} = -2.1483,$$

y el p -valor del contraste es $2P(Z \leq -2.1483) = 0.0317$, de manera que se rechaza la hipótesis nula para $\alpha = 0.05$ y se concluye que hay diferencias.

7.15 Realización de contrastes mediante intervalos de confianza

Una interesante alternativa a la realización de un contraste

$$H_0 : \theta = \theta_0 \quad H_1 : \theta \neq \theta_0$$

con un riesgo α , es calcular el intervalo de confianza para θ con un nivel de confianza $1 - \alpha$, ya que este intervalo se puede interpretar como el conjunto aceptable de hipótesis para θ , de manera que si θ_0 está fuera del intervalo, la hipótesis nula es poco creíble y puede rechazarse, mientras que si está dentro la hipótesis es creíble y se acepta.

Cuando el contraste sea unilateral de menor, el contraste se realizaría comparando θ_0 con el límite superior del intervalo de confianza para θ con un nivel de confianza $1 - 2\alpha$,

mientras que si el contraste es unilateral de mayor, se comparará con el límite inferior del intervalo.

Contraste	Intervalo de confianza	Decisión
Bilateral	$[l_i, l_s]$ con nivel de confianza $1 - \alpha$	Rechazar H_0 si $\theta_0 \notin [l_i, l_s]$
Unilateral menor	$[-\infty, l_s]$ con nivel de confianza $1 - 2\alpha$	Rechazar H_0 si $\theta_0 \geq l_s$
Unilateral mayor	$[l_i, \infty]$ con nivel de confianza $1 - 2\alpha$	Rechazar H_0 si $\theta_0 \leq l_i$

Ejemplo 7.18. Volviendo al contraste para comparar el rendimiento académico de dos grupos de alumnos que han obtenido las siguientes puntuaciones:

$$X_1 : 4 - 6 - 8 - 7 - 7 - 6 - 5 - 2 - 5 - 3$$

$$X_2 : 8 - 9 - 5 - 3 - 8 - 7 - 8 - 6 - 8 - 7 - 5 - 7$$

El contraste que se planteaba era

$$H_0 : \mu_1 = \mu_2 \quad H_1 : \mu_1 \neq \mu_2$$

Como se trata de un contraste bilateral, el intervalo de confianza para la diferencia de medias $\mu_1 - \mu_2$ con nivel de confianza $1 - \alpha = 0.95$, suponiendo varianzas iguales, vale $[-3.0521, 0.1521]$ puntos. Y como según la hipótesis nula $\mu_1 - \mu_2 = 0$, y el 0 cae dentro del intervalo, se acepta la hipótesis nula.

La ventaja del intervalo es que, además de permitirnos realizar el contraste, nos da una idea de la magnitud de la diferencia entre las medias de los grupos.

8 Análisis de la Varianza

8.1 Análisis de la varianza de 1 factor

El *Análisis de la Varianza con un Factor* (ANOVA por sus siglas en inglés), es una técnica estadística de contraste de hipótesis, que sirve para comparar las medias una variable cuantitativa, que suele llamarse *variable dependiente* o *respuesta*, en distintos grupos o muestras definidas por una variable cualitativa, llamada *variable independiente* o *factor*. Las distintas categorías del factor que definen los grupos a comparar se conocen como *niveles* o *tratamientos* del factor.

Se trata, por tanto, de una generalización de la *prueba T para la comparación de medias de dos muestras independientes*, para diseños experimentales con más de dos muestras. Y se diferencia de un análisis de regresión simple, donde tanto la variable dependiente como la independiente eran cuantitativas, en que en el análisis de la varianza de un factor, la variable independiente o factor es una variable cualitativa, aunque como veremos más adelante en los contrastes de regresión, se puede plantear un contraste de ANOVA como si fuese un contraste de regresión lineal.

Un ejemplo de aplicación de esta técnica podría ser la comparación del nivel de colesterol medio según el grupo sanguíneo. En este caso, la dependiente o factor es el grupo sanguíneo, con cuatro niveles (A, B, O, AB), mientras que la variable respuesta es el nivel de colesterol.

Para comparar las medias de la variable respuesta según los diferentes niveles del factor, se plantea un contraste de hipótesis en el que la hipótesis nula, H_0 , es que la variable respuesta tiene igual media en todos los niveles, mientras que la hipótesis alternativa, H_1 , es que hay diferencias estadísticamente significativas entre al menos dos de las medias. Dicho contraste se realiza mediante la descomposición de la varianza total de la variable respuesta; de ahí procede el nombre de esta técnica.

8.1.1 El contraste de ANOVA

La notación habitual en ANOVA es la siguiente:

- k : es el número de niveles del factor.
- n_i : es el tamaño de la muestra aleatoria correspondiente al nivel i -ésimo del factor.
- $n = \sum_{i=1}^k n_i$: es el número total de observaciones.

- X_{ij} ($i = 1, \dots, k$; $j = 1, \dots, n_i$): es una variable aleatoria que indica la respuesta del j -ésimo individuo al i -ésimo nivel del factor.
- x_{ij} : es el valor concreto, en una muestra dada, de la variable X_{ij} .

Niveles del factor			
1	2	...	k
X_{11}	X_{21}	...	X_{k1}
X_{12}	X_{22}	...	X_{k2}
\vdots	\vdots	\vdots	\vdots
X_{1n_1}	X_{2n_2}	...	X_{kn_k}

- μ_i : es la media de la población del nivel i .
- $X_i = \sum_{j=1}^{n_i} X_{ij}/n_i$: es la variable media muestral del nivel i , y estimador de μ_i .
- $\bar{x}_i = \sum_{j=1}^{n_i} x_{ij}/n_i$: es la estimación concreta para una muestra dada de la variable media muestral del nivel i .
- μ : es la media global de la población (incluidos todos los niveles).
- $\bar{X} = \sum_{i=1}^k \sum_{j=1}^{n_i} X_{ij}/n$: es la variable media muestral de todas las respuestas, y estimador de μ .
- $\bar{x} = \sum_{i=1}^k \sum_{j=1}^{n_i} x_{ij}/n$: es la estimación concreta para una muestra dada de la variable media muestral.

Con esta notación podemos expresar la variable respuesta mediante un modelo matemático que la descompone en componentes atribuibles a distintas causas:

$$X_{ij} = \mu + (\mu_i - \mu) + (X_{ij} - \mu_i),$$

es decir, la respuesta j -ésima en el nivel i -ésimo puede descomponerse como resultado de una media global, más la desviación con respecto a la media global debida al hecho de que recibe el tratamiento i -ésimo, más una nueva desviación con respecto a la media del nivel debida a influencias aleatorias.

Sobre este modelo se plantea la hipótesis nula: las medias correspondientes a todos los niveles son iguales; y su correspondiente alternativa: al menos hay dos medias de nivel que son diferentes.

$$\begin{aligned} H_0 : & \mu_1 = \mu_2 = \cdots = \mu_k \\ H_1 : & \mu_i \neq \mu_j \text{ para algún } i \neq j. \end{aligned}$$

Para poder realizar el contraste con este modelo es necesario plantear ciertas hipótesis estructurales (supuestos del modelo):

- ndependencia: Las k muestras, correspondientes a los k niveles del factor, representan muestras aleatorias independientes de k poblaciones con medias $\mu_1 = \mu_2 = \dots = \mu_k$ desconocidas.
- Normalidad: Cada una de las k poblaciones es normal.
- Homocedasticidad: Cada una de las k poblaciones tiene la misma varianza σ^2 .

Teniendo en cuenta la hipótesis nula y los supuestos del modelo, si se sustituye en el modelo las medias poblacionales por sus correspondientes estimadores muestrales, se tiene

$$X_{ij} = \bar{X} + (\bar{X}_i - \bar{X}) + (X_{ij} - \bar{X}_i),$$

o lo que es lo mismo,

$$X_{ij} - \bar{X} = (\bar{X}_i - \bar{X}) + (X_{ij} - \bar{X}_i).$$

Elevando al cuadrado y teniendo en cuenta las propiedades de los sumatorios, se llega a la ecuación que recibe el nombre de *identidad de la suma de cuadrados*:

$$\sum_{i=1}^k \sum_{j=1}^{n_i} (X_{ij} - \bar{X})^2 = \sum_{i=1}^k n_i (\bar{X}_i - \bar{X})^2 + \sum_{i=1}^k \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2,$$

donde:

- $\sum_{i=1}^k \sum_{j=1}^{n_i} (X_{ij} - \bar{X})^2$: recibe el nombre de *suma total de cuadrados*, (*SCT*), y es la suma de cuadrados de las desviaciones con respecto a la media global; por lo tanto, una medida de la variabilidad total de los datos.
- $\sum_{j=1}^k n_i (\bar{X}_i - \bar{X})^2$: recibe el nombre de *suma de cuadrados de los tratamientos o suma de cuadrados intergrupos*, (*SCI*), y es la suma ponderada de cuadrados de las desviaciones de la media de cada nivel con respecto a la media global; por lo tanto, una medida de la variabilidad atribuida al hecho de que se utilizan diferentes niveles o tratamientos.
- $\sum_{i=1}^k \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2$: recibe el nombre de *suma de cuadrados residual o suma de cuadrados intragrupo*, (*SCIntra*), y es la suma de cuadrados de las desviaciones de las observaciones con respecto a las medias de sus respectivos niveles o tratamientos; por lo tanto, una medida de la variabilidad en los datos atribuida a las fluctuaciones aleatorias dentro del mismo nivel.

Con esta notación la identidad de suma de cuadrados se expresa:

$$SCT = SCI + SCIntra$$

Y un último paso para llegar al estadístico que permitirá contrastar H_0 , es la definición de los *Cuadrados Medios*, que se obtienen al dividir cada una de las sumas de cuadrados

por sus correspondientes grados de libertad. Para SCT el número de grados de libertad es $n - 1$; para $SCInter$ es $k - 1$; y para $SCIIntra$ es $n - k$.

Por lo tanto,

$$\begin{aligned} CMT &= \frac{SCT}{n - 1} \\ CMInter &= \frac{SCInter}{k - 1} \\ CMIntra &= \frac{SCIIntra}{n - k} \end{aligned}$$

Y se podría demostrar que, en el supuesto de ser cierta la hipótesis nula y los supuestos del modelo, el cociente

$$\frac{CMInter}{CMIntra}$$

sigue una distribución F de Fisher con $k - 1$ y $n - k$ grados de libertad.

De esta forma, si H_0 es cierta, el valor del cociente para un conjunto de muestras dado, estará próximo a 0 (aún siendo siempre mayor que 0); pero si no se cumple H_0 crece la variabilidad intergrupos y la estimación del estadístico crece. En definitiva, realizaremos un contraste de hipótesis unilateral con cola a la derecha de igualdad de varianzas, y para ello calcularemos el p -valor de la estimación de F obtenida y aceptaremos o rechazaremos en función del nivel de significación fijado.

8.1.1.1 Tabla de ANOVA

Todos los estadísticos planteados en el apartado anterior se recogen en una tabla denominada Tabla de ANOVA, en la que se ponen los resultados de las estimaciones de dichos estadísticos en las muestras concretas objeto de estudio. Esas tablas también son las que aportan como resultado de cualquier ANOVA los programas estadísticos, que suelen añadir al final de la tabla el p -valor del estadístico F calculado, y que permite aceptar o rechazar la hipótesis nula de que las medias correspondientes a todos los niveles del factor son iguales.

	Suma de cuadrados	Grados de libertad	Cuadrados medios	Estadístico F	p-valor
Intergrupos	$SCInter$	$k - 1$	$CMInter = \frac{SCInter}{\frac{k-1}{n-k}}$	$f = \frac{CMInter}{CMIntra}$	$P(F > f)$
Intragrupos	$SCIIntra$	$n - k$	$CMIntra = \frac{SCIIntra}{n-k}$		

	Suma de cuadrados	Grados de libertad	Cuadrados medios	Estadístico F	p-valor
Total	SCT	$n - 1$			

8.1.2 Test de comparaciones múltiples y por parejas

Una vez realizado el ANOVA de un factor para comparar las k medias correspondientes a los k niveles o tratamientos del factor, se puede concluir aceptando la hipótesis nula, en cuyo caso se da por concluido el análisis de los datos en cuanto a detección de diferencias entre los niveles, o rechazándola, en cuyo caso es natural continuar con el análisis para tratar de localizar con precisión dónde está la diferencia, cuáles son los niveles cuyas respuestas son estadísticamente diferentes.

En el segundo caso, hay varios métodos que permiten detectar las diferencias entre las medias de los diferentes niveles, y que reciben el nombre de *test de comparaciones múltiples*. A su vez este tipo de test se suele clasificar en:

- Test de comparaciones por parejas: Su objetivo es la comparación una a una de todas las posibles parejas de medias que se pueden tomar al considerar los diferentes niveles. Su resultado es una tabla en la que se reflejan las diferencias entre todas las posibles parejas y los intervalos de confianza para dichas diferencias, con la indicación de si hay o no diferencias significativas entre las mismas. Hay que aclarar que los intervalos obtenidos no son los mismos que resultarían si se considera cada pareja de medias por separado, ya que el rechazo de H_0 en el contraste general de ANOVA implica la aceptación de una hipótesis alternativa en la que están involucrados varios contrastes individuales a su vez; y si queremos mantener un nivel de significación α en el general, en los individuales debemos utilizar un α' considerablemente más pequeño.
- Test de rango múltiple: Su objetivo es la identificación de subconjuntos homogéneos de medias que no se diferencian entre sí.

Para los primeros se puede utilizar el test de Bonferroni; para los segundos, el test de Duncan; y para ambas categorías a la vez los test HSD de Tukey y Scheffé.

8.2 ANOVA de dos o más factores

En muchos problemas aparece no ya un único factor que permite clasificar los individuos de la muestra en k diferentes niveles, sino que pueden presentarse dos o más factores que permiten clasificar a los individuos de la muestra en múltiples grupos según diferentes criterios, que se pueden analizar para ver si hay o no diferencias significativas entre las medias de la variable respuesta.

Para tratar con este tipo de problemas surge el *ANOVA de Dos o Más Factores* (o también *ANOVA de Dos o Más Vías*) como una generalización del proceso de un factor, que además de permitir el análisis de la influencia de cada uno de los factores por separado también hace posible el estudio de la *interacción* entre ellos.

Por otra parte, también son frecuentes los problemas en los que se toma más de una medida de una variable cuantitativa (respuesta) en cada sujeto de la muestra, y se procede al análisis de las diferencias entre las diferentes medidas. Si sólo se toman dos, el procedimiento adecuado es la T de Student de datos pareados, o su correspondiente no paramétrico, el test de Wilcoxon; pero si se han tomado tres o más medidas, el test paramétrico correspondiente a la T de Student de datos pareados es el *ANOVA de Medidas Repetidas*.

Incluso también se puede dar el caso de un problema en el que se analice una misma variable cuantitativa medida en varias ocasiones en cada sujeto de la muestra pero teniendo en cuenta a la vez la influencia de uno, dos o más factores que permiten clasificar a los individuos en varios subgrupos diferentes. En definitiva, pueden aparecer problemas donde a la par que un ANOVA de medidas repetidas se requiera realizar un ANOVA de dos o más vías.

Por último, la situación más compleja que se puede plantear en el análisis de una respuesta cuantitativa se presenta cuando, añadida a medidas repetidas y dos o más vías o factores de clasificación, se tienen una o más variables cuantitativas, llamadas *covariables*, que se piensa que pueden influir en la variable respuesta. Se procede entonces a realizar un *ANCOVA* o *Análisis de Covarianza*, con el que se pretende analizar la influencia de los factores y también ver si hay diferencias entre las medidas repetidas pero habiendo eliminado previamente la influencia (variabilidad) debida a la presencia de las covariables que se pretenden controlar.

8.2.1 ANOVA de dos factores con dos niveles cada factor

Para entender qué es un ANOVA de dos o más factores, conviene partir de un caso sencillo con dos factores y dos niveles en cada factor. Por ejemplo, se puede plantear un experimento con individuos que siguen o no una dieta (primer factor: dieta, con dos niveles: sí y no), y que a su vez toman o no un determinado fármaco (segundo factor: fármaco, con dos niveles: sí y no) para reducir su peso corporal (variable respuesta numérica: reducción del peso corporal expresada en Kg). En esta situación, se generan cuatro grupos diferentes: los que no hacen dieta ni toman fármaco (No-No), los que no hacen dieta pero sí toman fármaco (No-Sí), los que hacen dieta y no toman fármaco (Sí-No), y los que hacen dieta y toman fármaco (Sí-Sí). Y se pueden plantear tres efectos diferentes:

- El de la dieta: viendo si hay o no diferencias significativas en los Kg perdidos entre los individuos que la han seguido y los que no.

- El del fármaco: viendo si hay o no diferencias significativas en los Kg perdidos entre los individuos que lo han tomado y los que no.
- El de la interacción: viendo si el efecto combinado de dieta y fármaco es diferente del que tendrían sumando sus efectos por separado, y entonces se diría que sí que hay interacción; o si por el contrario el efecto de la combinación de dieta y fármaco es el mismo que la suma de los efectos por separado, y entonces se diría que no hay interacción.

A su vez, si hay interacción se puede dar en dos sentidos: si la combinación de dieta y fármaco ha hecho perder más kilos a los pacientes de los que cabría esperar con la suma de dieta y fármaco por separado, entonces la interacción de ambos factores ha actuado en sinergia con los mismos, mientras que si la combinación ha hecho perder menos kilos de los que cabría esperar con dieta y fármaco por separado, entonces la interacción ha actuado en antagonismo con ambos.

Siguiendo con el ejemplo, supongamos que la tabla que aparece a continuación refleja la media de Kg perdidos dentro de cada uno de los grupos comentados. Por simplificar el ejemplo, no se reflejan los Kg en cada individuo con la consiguiente variabilidad de los mismos, pero el ANOVA de dos vías sí que tendría en cuenta esa variabilidad para poder hacer inferencia estadística, plantear contrastes de hipótesis y calcular sus correspondientes p-valores.

	Fármaco No	Fármaco Sí
Dieta No	0	5
Dieta Sí	3	8

Si los resultados obtenidos fuesen los de la tabla anterior, se diría que no hay interacción entre fármaco y dieta, ya que el efecto del fármaco en el grupo de los que no hacen dieta ha hecho perder 5 Kg en media a los individuos, el efecto de la dieta en el grupo de los que no toman fármaco les ha hecho perder 3 Kg en media, y el efecto combinado de dieta y fármaco ha hecho perder 8 Kg con respecto a los que no hacen dieta y tampoco toman fármaco. Estos 8 Kg son iguales a la suma de 3 y 5, es decir iguales a la suma de los efectos de los factores por separado, sin ningún tipo de interacción (de término añadido) que cambie el resultado de la suma.

Con las medias de los cuatro grupos que se generan en el cruce de los dos factores, cada uno con dos niveles (2x2), se representan los gráficos de medias que aparecen más adelante. En estos gráficos, cuando no hay interacción las rectas que unen las medias correspondientes a un mismo nivel de uno de los factores son paralelas dentro de cierto margen de variabilidad.

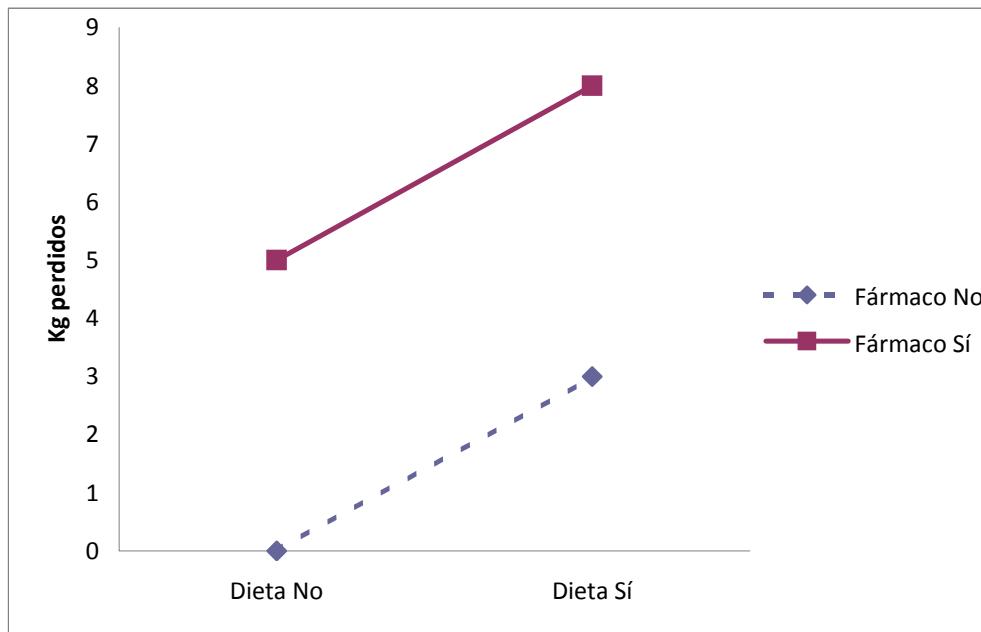


Figura 8.1: Gráfico de medias de dos factores sin interacción

Por el contrario, también podría obtenerse una tabla en la que la suma de los efectos por separado fuese menor que el efecto combinado de dieta y fármaco:

	Fármaco No	Fármaco Sí
Dieta No	0	5
Dieta Sí	3	12

En este caso, dejando al margen las variabilidad dentro de cada uno de los grupos y suponiendo que la misma es lo suficientemente pequeña como para que las diferencias sean significativas, los 8 Kg en media que se perderían al sumar los efectos por separado de dieta y fármaco son menores que los 12 que, en media, han perdido los individuos que han tomado el fármaco y han seguido la dieta a la vez. Por lo tanto, se ha producido una interacción de los dos factores que, al unirlos, ha servido para potenciar sus efectos por separado. Dicho de otra forma, para explicar el resultado final de los individuos que han tomado el fármaco y también han seguido la dieta habría que introducir un nuevo término en la suma, el término de interacción, que contribuiría con 4 Kg de pérdida añadidos a los 8 Kg que se perderían considerando simplemente la suma de dieta y fármaco. Como este nuevo término contribuye a aumentar la pérdida que se obtendría al

sumar los efectos por separado de ambos factores, se trataría de un caso de interacción en sinergia con los dos factores de partida.

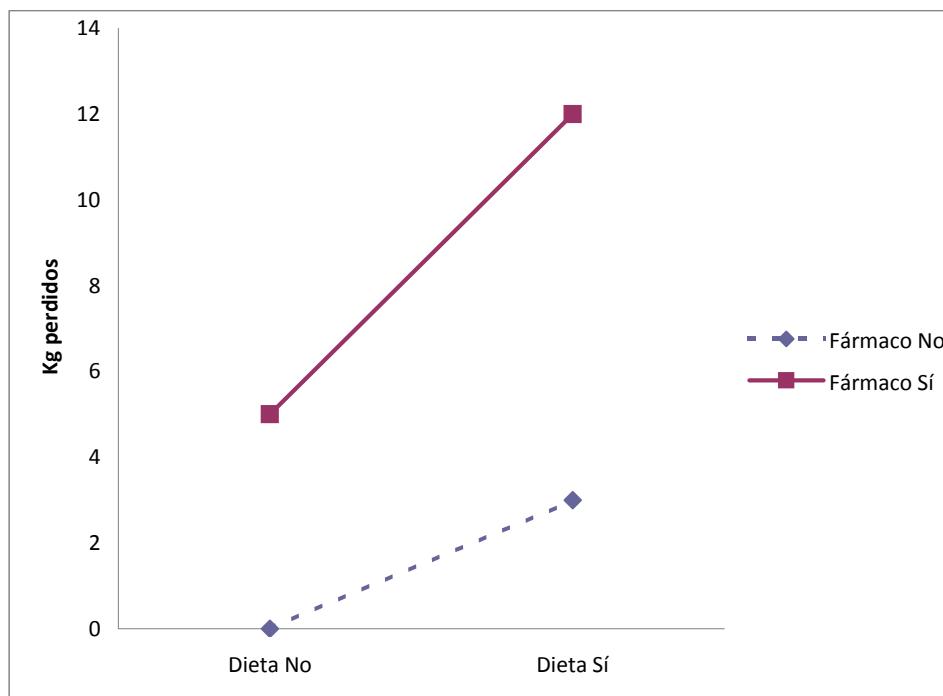


Figura 8.2: Gráfico de medias de dos factores con interacción sinérgica

Por último, también se podría obtener una tabla en la que la suma de los efectos por separado fuese mayor que el efecto combinado de los dos factores:

	Fármaco No	Fármaco Sí
Dieta No	0	5
Dieta Sí	3	4

Igualmente, en este nuevo ejemplo los 8 Kg en media que se perderían al sumar los efectos por separado de los dos factores son mayores que los 4 que en realidad pierden, en media, los individuos que han seguido la dieta y utilizado el fármaco. Por lo tanto, para explicar el resultado obtenido en el grupo de los que toman el fármaco y siguen la dieta habría que introducir un término añadido a la suma de efectos sin más, que se restaría a los 8 Kg hasta dejarlos en 4 Kg. Se trataría de un caso de interacción en antagonismo con los dos factores de partida.

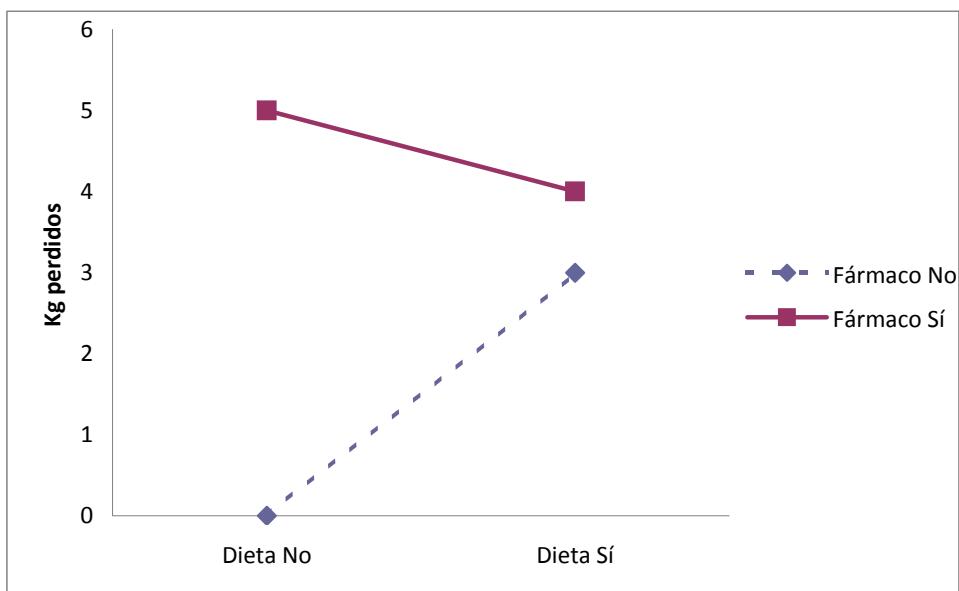


Figura 8.3: Gráfico de medias de dos factores con interacción antagónica

En realidad, la interacción también puede producirse en sinergia con uno de los factores y en antagonismo con el otro, ya que a veces los dos factores pueden producir un efecto con signo contrario. Por ejemplo, al hablar del factor dieta, se tiende a pensar que se trata de una dieta que sirve para bajar el peso, pero también cabe plantearse un experimento con personas que siguen una dieta de alto contenido calórico que en principio debería hacerles subir peso y ver qué evolución siguen cuando a la vez toman un fármaco para bajarlo.

Como puede deducirse fácilmente de las tablas y gráficas anteriores, la presencia de interacción implica que la diferencia entre las medias de los dos grupos dentro de un mismo nivel de uno de los factores no es la misma que para el otro nivel. Por ejemplo, en la segunda tabla, la diferencia entre las medias de Kg perdidos entre los que sí que toman el fármaco y los que no lo toman vale: $5-0=5$ Kg en los que no hacen dieta, y $12-3=9$ Kg en los que sí que hacen dieta. Lo cual gráficamente se traduce en que la pendiente de la recta que une las medias dentro del grupo de los que sí que toman el fármaco es diferente de la pendiente que une las medias dentro del grupo de los que no lo toman. En las ideas anteriores se basará el planteamiento del contraste de hipótesis para ver si la interacción ha resultado o no significativa.

Como ya se ha comentado, en cualquiera de las tablas anteriores se podrían analizar tres efectos diferentes: el de la dieta, el del fármaco y el de la interacción de dieta con fármaco; lo cual, en términos matemáticos, se traduce en tres contrastes de hipótesis diferentes:

- Efecto de la dieta sobre la cantidad de peso perdido:

$$H_0 : \mu_{\text{con dieta}} = \mu_{\text{sin dieta}}$$

$$H_1 : \mu_{\text{con dieta}} \neq \mu_{\text{sin dieta}}$$

- Efecto del fármaco sobre la cantidad de peso perdido:

$$H_0 : \mu_{\text{con fármaco}} = \mu_{\text{sin fármaco}}$$

$$H_1 : \mu_{\text{con fármaco}} \neq \mu_{\text{sin fármaco}}$$

- Efecto de la interacción entre dieta y fármaco, que a su vez se puede plantear de dos formas equivalentes:

- Viendo si dentro dentro de los grupos definidos en función de la dieta la diferencia de Kg perdidos entre los que toman fármaco y los que no lo toman es la misma:

$$H_0 : (\mu_{\text{con fármaco}} - \mu_{\text{sin fármaco}})_{\text{sin dieta}} = (\mu_{\text{con fármaco}} - \mu_{\text{sin fármaco}})_{\text{con dieta}}$$

$$H_1 : (\mu_{\text{con fármaco}} - \mu_{\text{sin fármaco}})_{\text{sin dieta}} \neq (\mu_{\text{con fármaco}} - \mu_{\text{sin fármaco}})_{\text{con dieta}}$$

- Viendo si dentro de los grupos definidos en función del fármaco la diferencia de Kg perdidos entre los que hacen dieta y los que no la hacen es la misma:

$$H_0 : (\mu_{\text{con dieta}} - \mu_{\text{sin dieta}})_{\text{sin fármaco}} = (\mu_{\text{con dieta}} - \mu_{\text{sin dieta}})_{\text{con fármaco}}$$

$$H_1 : (\mu_{\text{con dieta}} - \mu_{\text{sin dieta}})_{\text{sin fármaco}} \neq (\mu_{\text{con dieta}} - \mu_{\text{sin dieta}})_{\text{con fármaco}}$$

Aunque los detalles matemáticos más precisos sobre cómo el ANOVA de dos o más vías da respuesta a los contrastes expuestos quedan fuera del nivel de esta práctica, la idea general es sencilla y muy parecida a la explicada con más detalle en la práctica de ANOVA de una vía. En el ANOVA de una vía, la variabilidad total de los datos, expresada como suma de distancias al cuadrado con respecto a la media global (llamada Suma de Cuadrados Total), se descompone en dos diferentes fuentes de variabilidad: las distancias al cuadrado de los datos de cada grupo con respecto a la media del grupo, *Suma de Cuadrados Intra*, más las distancias al cuadrado entre las diferentes medias de los grupos y la media general, *Suma de Cuadrados Inter*. La suma de cuadrados intra-grupos es también llamada *Variabilidad Residual* o *Suma de Cuadrados Residual*, ya que su cuantía es una medida de la dispersión residual, remanente incluso después de haber dividido los datos en grupos. Estas sumas de cuadrados, una vez divididas por sus correspondientes grados de libertad, generan varianzas llamadas *Cuadrados Medios*, y el cociente de cuadrados medios (cuadrado medio inter dividido entre cuadrado medio intra) bajo la hipótesis nula de igualdad de medias en todos los grupos sigue una distribución *F* de Fisher que se puede utilizar para calcular un *p*-valor del contraste de igualdad de medias. En el ANOVA de dos factores, en lugar de dos fuentes de variabilidad tenemos cuatro: una por el primer factor, otra por el segundo, otra por la interacción y otra más que contempla la variabilidad residual o variabilidad intragrupo. En el ejemplo anterior, las cuatro fuentes de variabilidad son:

- La debida al primer factor: la dieta.

2. La debida al segundo factor: el fármaco.
3. La debida a la interacción entre ambos.
4. La residual.

Las tres primeras fuentes de variabilidad llevan asociadas sus correspondientes sumas de cuadrados, similares a la suma de cuadrados inter del ANOVA de una vía, mientras que la variabilidad residual lleva asociada su suma de cuadrados residual, similar a la suma de cuadrados intra del ANOVA de una vía. Dividiendo las sumas de cuadrados entre sus respectivos grados de libertad se obtienen varianzas, que divididas entre la varianza residual generan, bajo la hipótesis nula de igualdad de medias, valores f de la distribución F de Fisher que pueden utilizarse para calcular el p -valor del correspondiente contraste.

Lo anterior se resume en forma de tabla de un ANOVA de dos vías, considerando un primer factor con k_1 niveles, un segundo factor con k_2 niveles y un total de datos n . Si se denomina F1 al primer factor, F2 al segundo, I a la interacción y R al residual, la tabla de un ANOVA de dos vías tiene la siguiente forma:

Fuente	Suma Cuad	Grad Lib	Cuad Medios	f	p -valor
F1	$SF1$	$GF1 = k_1 - 1$	$CF1 = \frac{SF1}{GF1}$	$f1 = \frac{CF1}{CR}$	$P(F > f1)$
F2	$SF2$	$GF2 = k_2 - 1$	$CF2 = \frac{SF2}{GF2}$	$f2 = \frac{CF2}{CR}$	$P(F > f2)$
I	SI	$GI = GF1 \cdot GF2$	$CI = \frac{SI}{GI}$	$fI = \frac{CI}{CR}$	$P(F > fI)$
R	SR	$GR = n - 1 - GF1 - GF2 - GI$	$CR = \frac{SR}{GR}$		
Total	ST	$GT = n - 1$			

Una vez obtenida la tabla, habitualmente mediante un programa de estadística para evitar realizar la gran cantidad de cálculos que conlleva (los distintos programas pueden proporcionar tablas ligeramente diferentes a la expuesta en esta práctica, en las que pueden aparecer filas añadidas cuya interpretación dependerá del programa utilizado), el siguiente paso es la interpretación de los p -valores obtenidos en cada uno de los factores y en la interacción. Para ello, resulta clave el p -valor de la interacción porque condicionará completamente el análisis:

- Si la interacción no ha resultado significativa (p -valor de la interacción mayor que el nivel de significación, habitualmente 0.05), se puede considerar por separado la actuación de los dos factores y ver si hay o no diferencias significativas en sus niveles atendiendo al p -valor que aparece en la tabla para cada uno de ellos. Por ejemplo, en la primera de las tablas del análisis de Kg perdidos en función de la dieta y el fármaco, se obtendría que la interacción no es significativa, lo cual implicaría que habría que analizar el efecto de los factores por separado. Para ello, se acudiría al p -valor del factor dieta y si es menor que el nivel de significación

fijado, entonces el factor dieta habría resultado significativo, lo cual quiere decir que habría diferencias significativas (más allá de las asumibles por azar) entre los Kg perdidos por los individuos que hacen dieta y los que no; y todo ello, independientemente de si los individuos están tomando o no el fármaco, ya que no hay una interacción significativa que ligue los resultados de la dieta con el fármaco. Igualmente, con el factor fármaco, se acudiría a su *p*-valor y se vería si hay o no diferencias significativas entre los Kg perdidos por los que toman el fármaco y los que no lo hacen, independientemente de si siguen o no la dieta.

- Si la interacción ha resultado significativa (*p*-valor de la interacción menor que el nivel de significación, habitualmente 0.05), no se puede considerar por separado la actuación de los dos factores, la presencia de uno de los factores condiciona lo que sucede en el otro y el análisis de diferencias debidas al segundo factor debe realizarse por separado dentro de cada uno de los niveles del primero; y a la inversa, el análisis de diferencias debidas al primero debe realizarse por separado dentro de cada uno de los niveles del segundo. Por ejemplo, en la segunda de las tablas del análisis de Kg perdidos en función de la dieta y el fármaco, muy probablemente se obtendría que la interacción sí que es significativa, con lo cual no habría un único efecto del fármaco: en el grupo de los que no toman el fármaco, la diferencia de Kg perdidos entre los que sí que hacen dieta y los que no la hacen no sería la misma que en el grupo de los que sí que toman el fármaco. E igualmente, tampoco habría un único efecto de la dieta: en el grupo de los que no hacen dieta, la diferencia de Kg perdidos entre los que sí que toman el fármaco y los que no lo hacen no sería la misma que en el grupo de los que sí que hacen dieta.

Una aclaración final importante es que en ningún caso un ANOVA de dos factores con dos niveles en cada vía equivale a hacer por separado una T de Student de datos independientes en cada uno de los factores. Ni siquiera en el caso de que no haya interacción el *p*-valor que se obtiene en cada uno de los dos factores coincide con el que se obtendría en la comparación de los niveles mediante la T de Student. El ANOVA de dos factores es una técnica multivariante que cuantifica la influencia de cada una de las variables independientes en la variable dependiente después de haber eliminado la parte de la variabilidad que se debe a las otras variables independientes que forman parte del modelo. En el ejemplo de los Kg perdidos, no sería lo mismo analizar la influencia de la variable dieta después de eliminar la variabilidad explicada mediante la variable fármaco e incluso la interacción entre dieta y fármaco, que es lo que haría el ANOVA de dos factores, que analizar simplemente la influencia de la variable dieta sin más, o fármaco sin más, que es lo que podríamos hacer mediante una T de Student de datos independientes. Tampoco el análisis de la interacción en el ANOVA de dos factores equivale a realizar un ANOVA de una vía considerando una nueva variable independiente con cuatro categorías diferentes (1:Sí-Sí, 2:Sí-No, 3:No-Sí, 4:No-No), por el mismo motivo: las conclusiones del ANOVA de dos vías hay que entenderlas en el contexto de una técnica multivariante en que la importancia de cada variable independiente se obtiene después de eliminar de los datos la variabilidad debida a las demás.

8.2.2 ANOVA de dos factores con tres o más niveles en algún factor

El planteamiento y resolución de un ANOVA de dos factores con tres o más niveles en algún factor es muy parecido al ya expuesto de dos niveles en cada factor. Únicamente cambian ligeramente las hipótesis nulas planteadas en los factores en las que habría que incluir la igualdad de tantas medias como niveles tenga el factor analizado, y las alternativas en las que se supone que alguna de las medias es diferente. En cuanto a las interacciones, también se contemplarían diferencias de medias pero teniendo en cuenta que hay más diferencias posibles al tener más niveles dentro de cada factor.

En cuanto a la interpretación final de los resultados de la tabla del ANOVA, si no hay interacción y sin embargo hay diferencias significativas en cualquiera de los factores con 3 o más niveles, el siguiente paso sería ver entre qué medias se dan esas diferencias. Por ejemplo, si no hay interacción y se ha rechazado la hipótesis nula de igualdad de medias entre los tres niveles del factor 1, habría que ver si esas diferencias aparecen entre los niveles 1 y 2, o entre el 1 y 3, e incluso entre el 2 y el 3, independientemente del factor 2; e igualmente con el factor 2. Para poder ver entre qué niveles hay diferencias, habría que realizar *Test de Comparaciones Múltiples y por Parejas*; por ejemplo un test de Bonferroni o cualquier otro de los vistos en la práctica de ANOVA de una vía. Si la interacción saliese significativa, habría que hacer lo mismo pero considerando las posibles diferencias entre los 3 niveles del factor 1 dentro de cada nivel del factor 2 y viceversa.

Como ya se ha comentado para el ANOVA de dos factores con dos niveles en cada factor y la T de Student de datos independientes, igualmente el ANOVA de dos factores con tres o más niveles en algún factor no equivale a dos ANOVAS de una vía. El *p*-valor que se obtiene en el de dos factores no es el mismo que que se obtendría en los ANOVAS de una vía realizados teniendo en cuenta cada uno de los factores por separado, incluso si la interacción no es significativa.

8.2.3 ANOVA de tres o más factores

Aunque los fundamentos del ANOVA de tres o más factores son muy parecidos a los de dos y la tabla obtenida es muy similar, la complejidad en la interpretación sube un escalón. Por ejemplo, en un ANOVA de tres factores la tabla presentaría los tres efectos de cada uno de los factores por separado, las tres interacciones dobles (1 con 2, 1 con 3 y 2 con 3), e incluso también podría mostrar la interacción triple (los programas de estadística permiten considerar o no las interacciones de cualquier orden). Si la interacción triple fuese significativa, entonces no se podría hablar del efecto general del factor 1, sino que habría que analizar el efecto del factor 1 dentro de cada nivel del 2 y a su vez dentro de cada nivel del 3, y así sucesivamente. Si la interacción triple no fuese significativa pero sí que lo fuese la del factor 1 con el 2, entonces habría que analizar el efecto del factor 1 dentro de cada uno de los niveles del 2 pero independientemente del factor 3. Y así hasta completar un conjunto muy grande de análisis posibles y de *Test de Comparaciones Múltiples* aplicados. No obstante, es el propio experimentador el

que debe limitar el conjunto de análisis a realizar con un planteamiento muy claro del experimento, reduciendo en la medida de lo posible el número de factores considerados y teniendo claro que no merece la pena considerar interacciones triples, o de órdenes superiores, si no hay forma clara de interpretar su resultado.

En ningún caso un ANOVA de tres o más factores equivale a tres ANOVAS de una vía realizados teniendo en cuenta los factores considerados por separado.

8.2.4 Factores fijos y Factores aleatorios

A la hora de realizar un ANOVA de varios factores, el tratamiento de la variabilidad debida a cada uno de ellos y también las conclusiones que se pueden obtener después de realizarlo, son diferentes dependiendo de que los factores sean fijos o aleatorios.

Se entiende como *Factor Fijo o Factor de Efectos Fijos* aquel cuyos niveles los establece, los fija de antemano, el investigador (por ejemplo, cantidades concretas de fármaco o de tiempo transcurrido), o vienen dados por la propia naturaleza del factor (por ejemplo, el sexo o la dieta). Su variabilidad es más fácil de controlar y también resulta más sencillo su tratamiento en los cálculos que hay que hacer para llegar a la tabla final del ANOVA, pero tienen el problema de que los niveles concretos que toma el factor constituyen la población de niveles sobre los que se hace inferencia. Es decir, no se pueden sacar conclusiones poblacionales que no se refieran a esos niveles fijos con los que se ha trabajado.

Por contra, un *Factor Aleatorio o Factor de Efectos Aleatorios* es aquel cuyos niveles son seleccionados de forma aleatoria entre todos los posibles niveles del factor (por ejemplo, cantidad de fármaco, con niveles 23 mg, 132 mg y 245 mg, obtenidos al escoger 3 niveles de forma aleatoria entre 0 y 250 mg). Su tratamiento es más complicado, pero al constituir una muestra aleatoria de niveles, se pretende sacar conclusiones extrapolables a todos los niveles posibles.

8.2.4.1 Supuestos del modelo de ANOVA de dos o más vías

Como ya sucedía con el ANOVA de una vía, el de dos o más vías es un test paramétrico que supone que:

- Los datos deben seguir distribuciones normales dentro de cada categoría, entendiendo por categorías todas las que se forman del cruce de todos los niveles de todos los factores. Por ejemplo, en un ANOVA de 2 factores con 3 niveles en cada factor, se tienen 3^2 categorías diferentes.
- Todas las distribuciones normales deben tener igualdad de varianzas (homocedasticidad).

Cuando no se cumplen las condiciones anteriores y además las muestras son pequeñas, no se debería aplicar el ANOVA de dos o más vías, con el problema añadido de que no hay un test no paramétrico que lo sustituya. Mediante test no paramétricos (sobre todo mediante el test de Kruskall-Wallis) se podría controlar la influencia de cada uno de los factores por separado en los datos, pero nunca el importantísimo papel de la interacción.

8.3 ANOVA de medidas repetidas

En muchos problemas se cuantifica el valor de una variable dependiente en varias ocasiones en el mismo sujeto (por ejemplo: en un grupo de individuos que están siguiendo una misma dieta, se puede anotar el peso perdido al cabo de un mes, al cabo de dos y al cabo de tres), y se intenta comparar la media de esa variable en las diferentes ocasiones en que se ha medido, es decir, ver si ha habido una evolución de la variable a lo largo de las diferentes medidas (en el ejemplo anterior, una evolución del peso perdido). Conceptualmente es una situación análoga a la estudiada al comparar dos medias con datos emparejados mediante una T de Student de datos emparejados, o su correspondiente no paramétrico, el test de Wilcoxon, pero ahora hay más de dos medidas emparejadas, realizadas en el mismo individuo. En estas situaciones se utiliza el ANOVA de medidas repetidas.

El ANOVA de medidas repetidas, como también sucede con cualquier otro test que utilice datos emparejados, tiene la ventaja de que las comparaciones que se realizan están basadas en lo que sucede dentro de cada sujeto (intra-sujetos), lo cual reduce el ruido o variabilidad que se produce en comparaciones entre diferentes grupos de sujetos. Por ejemplo, en el estudio sobre la evolución del peso perdido con personas que siguen la misma dieta, se podría haber cuantificado la variable al cabo de uno, dos y tres meses, pero en tres grupos diferentes que hubiesen seguido la misma dieta, pero con este diseño del estudio no se controlan otras variables que pueden influir en el resultado final, por ejemplo el sexo, la edad, o la cantidad de ejercicio que se hace al día. Dicho de otra forma, en el diseño con grupos independientes es posible que alguno de los grupos tenga una media de edad superior, o no haya igual número de hombres que de mujeres, y todo ello tener su reflejo en el número de Kg perdidos. Mientras que, con el diseño de datos emparejados, la segunda medida se compara con la primera que también se ha realizado en la misma persona, y por lo tanto es igual su sexo, su edad y la cantidad de deporte que realiza; y así con todas las demás medidas que se comparan entre sí pero dentro del mismo individuo. Eso permite controlar la variabilidad y detectar pequeñas diferencias que de otra forma serían indetectables.

8.3.0.1 ANOVA de medidas repetidas como ANOVA de dos vías sin interacción

El ANOVA de medidas repetidas puede realizarse como un ANOVA de dos vías sin interacción sin más que realizar los cálculos oportunos introduciendo adecuadamente los datos en un programa estadístico.

En la situación de partida, si suponemos que tenemos k medidas emparejadas de una variable dependiente numérica y n individuos en los que hemos tomado las medidas, los datos se pueden organizar como aparecen en la tabla siguientes:

2-5	VarDep 1	VarDep 2	...	VarDep k
Individuo 1	$x_{1,1}$	$x_{1,2}$...	$x_{1,k}$
Individuo 2	$x_{2,1}$	$x_{2,2}$...	$x_{2,k}$
...
Individuo n	$x_{n,1}$	$x_{n,2}$...	$x_{n,k}$

Pero esos mismos datos también se pueden ordenar en un formato de tabla mucho más conveniente para poderles aplicar un ANOVA de dos vías:

2-4	Var Dep	Individuo	Medida
Fila 1	$x_{1,1}$	1	1
Fila 2	$x_{2,1}$	2	1
...
Fila n	$x_{n,1}$	n	1
Fila $n+1$	$x_{1,2}$	1	2
Fila $n+2$	$x_{2,2}$	2	2
...
Fila $2n$	$x_{n,2}$	n	2
...
Fila $(k-1)n+1$	$x_{1,k}$	1	k
Fila $(k-1)n+2$	$x_{2,k}$	2	k
...
Fila kn	$x_{n,k}$	n	k

Con ello, tanto Individuo como Medida son variables categóricas que dividen la muestra total ($n \cdot k$ datos de la variable dependiente) en grupos: n grupos en la variable Individuo y k grupos en la variable Medida. Además, considerando el cruce de ambas variables (Medida x Individuo) se forman $n \cdot k$ grupos con un único dato de la variable dependiente en cada grupo.

Para explicar la variabilidad de los datos de la variable dependiente cuantitativa se pueden considerar tres fuentes: la debida a la variable Medida, la debida a la variable

Individuo, y la residual. Ahora no cabe hablar de la variabilidad debida a la interacción entre Medida e Individuo ya que los grupos que surgen del cruce de los dos factores sólo tienen un dato y no es viable calcular medias y dispersiones dentro de un grupo con un único dato. Y el análisis de la influencia de cada uno de los factores se realiza mediante un ANOVA de dos factores sin interacción, que genera la siguiente tabla:

Fuente	Suma Cuad	Grad Lib	Cuad Med	F	p-valor
F1=Medida	SF_1	$GF_1 = k - 1$	$CF_1 = \frac{SF_1}{GF_1}$	$f_1 = \frac{CF_1}{CR}$	$P(F > f_1)$
F2=Individuo	SF_2	$GF_2 = n - 1$	$CF_2 = \frac{SF_2}{GF_2}$	$f_2 = \frac{CF_2}{CR}$	$P(F > f_2)$
Residual	SR	$GR = (n \cdot k) - 1 - GF_1 - GF_2$	$CR = \frac{SR}{GR}$		
Total	ST	$GT = (n \cdot k) - 1$			

Y permite dar respuesta a los siguientes contrastes:

1. En la variable Medida:

$$H_0 : \mu_{\text{Medida } 1} = \mu_{\text{Medida } 2} = \dots = \mu_{\text{Medida } k}$$

$$H_1: \text{Alguna de las medias es diferente.}$$

Si el p -valor obtenido es menor que el nivel de significación fijado querrá decir que alguna de las medias es significativamente diferente del resto. Este es el contraste más importante del ANOVA de medidas repetidas y supone que la variabilidad dentro de cada individuo (intra-sujeto) es lo suficientemente grande como para que se descarte el azar como su causa. Por lo tanto la variable Medida ha tenido un efecto significativo.

2. En la variable Individuo:

$$H_0 : \mu_{\text{Individuo } 1} = \mu_{\text{Individuo } 2} = \dots = \mu_{\text{Individuo } n}$$

$$H_1: \text{Alguna de las medias es diferente.}$$

Si el p -valor obtenido es menor que el nivel de significación fijado querrá decir que alguna de las medias es significativamente diferente del resto, y por lo tanto alguno de los individuos analizados ha tenido un comportamiento en la variable dependiente diferente del resto. En realidad no es un contraste importante en el ANOVA de medidas repetidas ya que supone un análisis de la variabilidad entre individuos (inter-sujetos), pero es muy difícil que en un experimento dado esta variabilidad no esté presente.

Si la conclusión del ANOVA es que hay que rechazar alguna de las dos hipótesis nulas, ya sea la de igualdad de medias en los grupos formados por la variable Medida o la de igualdad de medias en los grupos formados por la variable Individuo, entonces en el siguiente paso se podría aplicar un Test de Comparaciones Múltiples y por Parejas, por

ejemplo un test de Bonferroni, para ver qué medias son diferentes, especialmente para ver entre qué niveles del la variable Medida se dan las diferencias.

8.3.0.2 Supuestos del ANOVA de medidas repetidas

Como en cualquier otro ANOVA, en el de medidas repetidas se exige que:

- Los datos de la variable dependiente deben seguir distribuciones normales dentro de cada grupo, ya sea formado por la variable Medida o por la variable Individuo. Como el contraste más importante se realiza en la variable Medida, resultará especialmente importante que sean normales las distribuciones de todas las Medidas.
- Todas las distribuciones normales deben tener igualdad de varianzas (homocedasticidad), especialmente las de las diferentes Medidas.

Cuando en un ANOVA de medidas repetidas se cumple la normalidad y la homocedasticidad de todas las distribuciones se dice que se cumple la *Esfericidad* de los datos, y hay tests estadísticos especialmente diseñados para contrastar la esfericidad como la *prueba de Mauchly*.

Cuando no se cumplen las condiciones anteriores y además las muestras son pequeñas, no se debería aplicar el ANOVA de medidas repetidas, pero al menos sí que hay una prueba no paramétrica que permite realizar el contraste de si hay o no diferencias significativas entre los distintos niveles de la variable Medida, que es el *test de Friedman*.

8.3.1 ANOVA de medidas repetidas + ANOVA de una o más vías

No son pocos los problemas en los que, además de analizar el efecto intra-sujetos en una variable dependiente cuantitativa medida varias veces en los mismos individuos para el que cabría plantear un ANOVA de medidas repetidas, también aparecen variables cualitativas que se piensa que pueden estar relacionadas con la variable dependiente. Estas últimas variables introducen un efecto que aunque habitualmente es catalogado como inter-sujetos más bien se trataría de un efecto inter-grupos, ya que permiten definir grupos entre los que se podría plantear un ANOVA de una o más vías. Por ejemplo, se podría analizar la pérdida de peso en una muestra de individuos al cabo de uno, dos y tres meses de tratamiento (ANOVA de medidas repetidas), pero teniendo en cuenta que los individuos de la muestra han sido divididos en seis grupos que se forman por el cruce de dos factores, Dieta y Ejercicio, con tres dietas diferentes: a, b y c, y dos niveles de ejercicio físico diferentes: bajo y alto. Para analizar la influencia de estos dos factores inter-sujetos, habría que plantear un ANOVA de dos vías con interacción. Para un ejemplo como el comentado, aunque los datos podrían disponerse de una forma similar a la que permite realizar el ANOVA de medidas repetidas como un ANOVA de dos factores (variables Medida e Individuo), y añadirle dos factores más (Dieta y Ejercicio), no resulta cómodo tener que introducir en la matriz de datos varias filas para

un mismo individuo (tantas como medidas repetidas diferentes se hayan realizado). Por ello, determinados programas de estadística, como PASW, permiten realizar ANOVAs de medidas repetidas introduciendo los datos en el formato clásico, una fila para cada individuo y una variable para cada una de las medidas repetidas, definiendo factores intra-sujeto que en realidad estarían compuestos por todas las variables que forman parte de las medidas repetidas. Además, a los factores intra-sujeto permiten añadirle nuevos factores inter-sujeto (categorías) que pueden influir en las variables respuesta (las diferentes medidas), e incluso comprobar si hay o no interacción entre los factores inter-sujeto entre sí y con los factores intra-sujeto. Por lo tanto, son procedimientos que realizan a la vez un ANOVA de medidas repetidas y un ANOVA de una o más vías, con la ventaja de que se pueden introducir los datos en la forma clásica: una fila para cada individuo.

El resultado de la aplicación de estos procedimientos es muy parecido a los comentados en apartados previos: se generan tablas de ANOVA en las que se calcula un *p*-valor para cada uno de los factores, ya sean intra-sujeto (medidas repetidas) o inter-sujeto (categorías), y también para la interacción, ya sea de los factores inter-sujeto entre sí o de factores inter-sujeto con los intra-sujeto.

8.4 Análisis de la covarianza: ANCOVA

El análisis de la covarianza, ANCOVA, es una extensión del ANOVA (ya sea de una o varias vías y de medidas repetidas), que permite analizar la influencia que sobre la variable dependiente cuantitativa tienen todas las variables independientes categóricas (factores) y las medidas repetidas contempladas en el ANOVA, pero, además, eliminando el efecto que otra u otras variables independientes cuantitativas podrían tener sobre la variable respuesta. Las variables independientes cuyo efecto se pretende eliminar (controlar o ajustar) son llamadas *Covariables* o *Covariantes* porque se espera que covarién, es decir, que estén correlacionadas con la variable dependiente.

Aunque la explicación detallada de cómo se realiza el ANCOVA va más allá del nivel de lo expuesto en esta práctica, la idea es sencilla: se puede plantear un análisis de regresión de la variable dependiente en función de la covariable (o de las covariables si hay más de una), y eliminar la parte de la variabilidad de la dependiente que se puede explicar gracias a la covariable sin más que trabajar con los residuos del modelo de regresión en lugar de con los datos originales. Posteriormente, se procede a realizar una ANOVA, de uno o varios factores e incluso de medidas repetidas, aplicado a los residuos.

El resultado final de la aplicación del ANCOVA es una tabla muy parecida a la del ANOVA, pero con una línea añadida por para cada una de las covariables. En esas líneas se recoge la cantidad de variabilidad explicada por cada una de las covariables y su correspondiente *p*-valor, que da respuesta al contraste de si la covariable es o no prescindible para explicar lo que sucede en la variable dependiente (en términos más

técnicos, el contraste sería si la pendiente del modelo de regresión de la variable independiente en función de la covariante puede o no ser igual a 0). En la tabla del ANCOVA no hay ninguna línea añadida que contemple la posible interacción entre la covariante y los distintos factores inter-sujetos, simplemente porque si hubiese interacción no debería aplicarse un modelo de ANCOVA ya que el efecto del factor no podría estimarse porque dependería del valor concreto considerado en la covariante, que, por ser continua, tiene infinitos valores, luego habría infinitos diferentes efectos del factor y no se le podría asignar un *p*-valor concreto. Pero sí que la tabla añade una línea para la interacción de cada uno de los factores intra-sujetos con cada una de las covariantes, ya que cada factor intra-sujetos internamente está compuesto por varias variables cuantitativas que pueden presentar diferentes pendientes en la regresión en función de la covariante.

Si la representación gráfica habitual para ver si una serie de factores influyen o no en una variable respuesta cuantitativa (ANOVA) es el denominado gráfico de medias, en el ANCOVA el efecto de la covariante en la variable respuesta se puede ver mediante la nube puntos de la variable respuesta en función de la covariante, que presentará un aspecto más o menos rectilíneo dependiendo del nivel de correlación lineal entre ambas. Además, también se puede intuir si un determinado factor influye en la variable respuesta una vez eliminada la influencia de la covariante:

- Si la nube de puntos puede ajustarse adecuadamente mediante una única recta de pendiente nula, independientemente de los niveles del factor, entonces quiere decir que ni la covariante ni el factor son significativos para explicar la variable respuesta.
- Si la nube de puntos se ajusta adecuadamente mediante una única recta de pendiente no nula, independientemente de los niveles del factor, entonces quiere decir que la covariante sí que es significativa pero no así el factor, ya que, una vez eliminada la influencia de la covariante (es decir, tomando como variable dependiente los residuos del ajuste inicial) no habría diferencias entre los distintos niveles (los puntos de las diferentes categorías quedarían a la misma altura).

Por ejemplo, en la siguiente figura aparece el resultado de un experimento en el que se han anotado los Kg perdidos por personas que han seguido dos tipos diferentes de dieta, pero teniendo en cuenta como covariante el índice de masa corporal, que se piensa que también puede influir en el número de Kg perdidos pero que, sin embargo, no se ha controlado a la hora de elaborar los grupos y claramente han quedado desequilibrados en la covariante (los que han tomado la dieta 2 tienen en media mayor índice de masa corporal que los que han tomado la dieta 1). Según la figura, cabría esperar que haya diferencias significativas en los Kg perdidos según la dieta (parece que la dieta 2 hace perder más Kg que la dieta 1), pero en realidad todo se debe a la covariante, y eliminando su efecto (si la pendiente de la recta fuese 0) los dos grupos habrían perdido cantidades muy similares de peso. En definitiva, en similares condiciones de índice de masa corporal, la dieta 2 no haría perder más Kg.

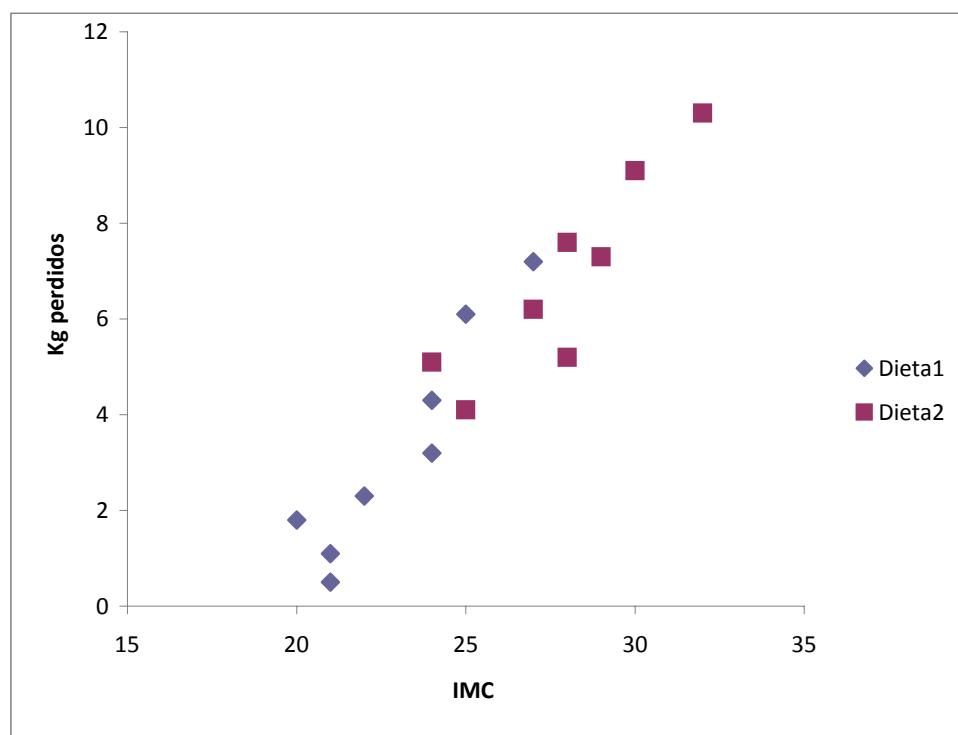


Figura 8.4: Nube de puntos con covariable significativa y factor no significativo

- Si la nube de puntos se ajusta adecuadamente mediante varias rectas de pendiente nula, una por cada nivel del factor, entonces la covariable no es significativa, pero sí el factor.
- Si la nube de puntos se ajusta adecuadamente con rectas, una por cada nivel del factor, con igual pendiente no nula, y al menos una de las rectas es diferente de todas las demás (al menos uno de los niveles aparece desplazado), entonces tanto la covariable como el factor serían significativos a la hora de explicar la variable dependiente.

Por ejemplo, en la siguiente figura aparece el resultado de un experimento similar al ya comentado: Kg perdidos en función de la dieta y de la covariable índice de masa corporal que no se ha controlado adecuadamente a la hora de hacer los grupos (el grupo de los que toman la dieta 2 tiene mayor IMC de partida). A la vista de la gráfica incluso parece que hay diferencias significativas en el número de Kg perdidos de tal forma que los de la dieta 2 haría perder más, pero todo es consecuencia de la covariable; eliminado su efecto, el número de Kg perdidos por

los individuos que toman la dieta 1 es mayor (eliminada la pendiente de la recta, los puntos de la dieta 1 quedarían por arriba).

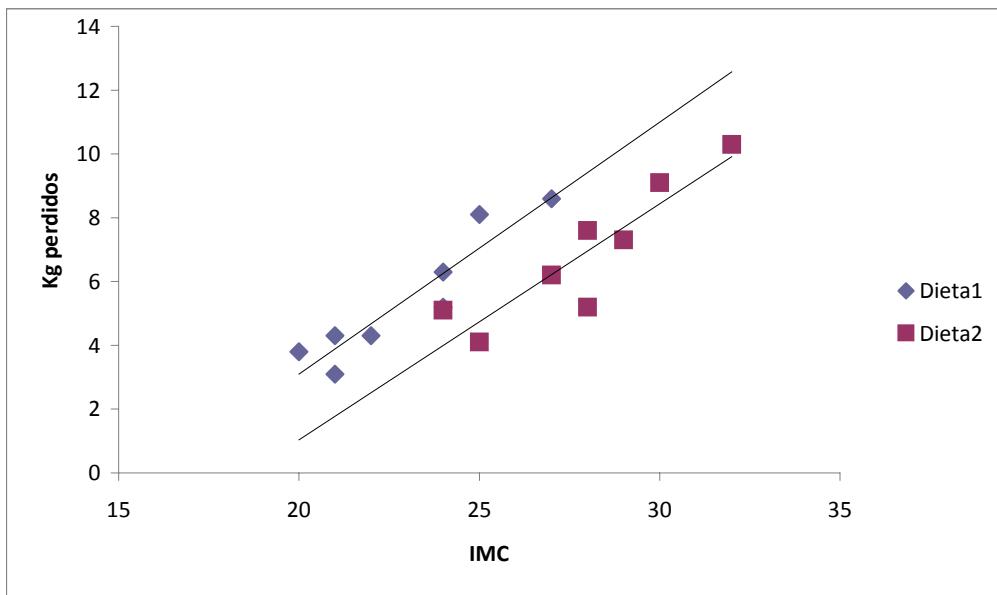


Figura 8.5: Nube de puntos con covariante significativa y factor también significativo

- Si la nube de puntos se ajusta adecuadamente con diferentes rectas, una por cada nivel del factor, con pendientes no nulas pero diferentes, entonces quiere decir que habría interacción entre covariante y factor y no debería plantearse un modelo de ANCOVA.