Introducción al diseño y análisis de encuestas

Aplicaciones estadísticas a la selección de muestras y al análisis de cuestionarios

Ángel A. Juan y Alicia Vila

PID_00161062

Universitat Oberta de Catalunya
www.uoc.edu

Índice

Introducción 5

Objetivos 6

1. Diseño de cuestionarios 7
1.1. Elaboración de las preguntas de un cuestionario 7
1.2. Uso de escalas en preguntas estructuradas 10

2. Diseño y selección de la muestra 14
2.1. Muestreo aleatorio simple 15
2.2. Muestreo sistemático 17
2.3. Muestreo aleatorio estratificado (grupos homogéneos) 17
2.4. Muestreo por conglomerados (*clusters* o grupos heterogéneos) 20

3. Análisis de cuestionarios: estudio parcial de un caso 25
3.1. Ejemplo de uso de estadísticos descriptivos e intervalos de confianza 25
3.2. Ejemplo de uso de contrastes de hipótesis para comparar dos grupos 27
3.3. Ejemplo de uso de ANOVA para comparar más de dos grupos 29
3.4. Ejemplo de uso de correlación y regresión lineal 30

Resumen 32

Ejercicios de autoevaluación 33

Solucionario 35

Introducción

Las encuestas y cuestionarios se han convertido en una herramienta de investigación de uso cotidiano en la llamada "sociedad de la información". La idea de usar datos provenientes de una muestra –compuesta por un número relativamente pequeño de elementos– para obtener información sobre toda una población es utilizada a diario por los medios de comunicación, ya sea prensa escrita, televisión, radio o incluso Internet.

En efecto, las encuestas y los cuestionarios se usan para sondear el estado de opinión de los potenciales votantes de unas elecciones, para conocer el potencial interés de nuevos bienes o servicios en el mercado, para predecir la aceptación que tendrán determinadas decisiones gubernamentales o estratégicas, para conocer mejor a los miembros de una comunidad, para detectar demandas potenciales de los consumidores que no están siendo satisfechas, etc. En investigación, además, las técnicas basadas en el uso de encuestas y cuestionarios representan probablemente la herramienta de investigación social más común en artículos y publicaciones científicas.

Sin embargo, el paso de datos muestrales a información sobre la población no es trivial, ya que requiere de todo un proceso metódico que incluye el diseño de las preguntas (para evitar introducir sesgos innecesarios en las mismas), el diseño de la muestra (para minimizar en lo posible el error muestral), la realización de la encuesta y el análisis de los resultados. En muchas ocasiones este proceso se hace demasiado a la ligera y de forma poco rigurosa, con lo que los resultados que se obtienen son poco fiables y nada creíbles desde un punto de vista científico. En este módulo se presentan y discuten los conceptos básicos de estas técnicas, desde las claves de un buen cuestionario y de un buen diseño muestral hasta ejemplos de cómo pueden aplicarse las técnicas estadísticas trabajadas durante el curso para representar numérica y gráficamente la información obtenida sobre la población.

Objetivos

Los objetivos docentes que se pretenden alcanzar con este módulo son los siguientes:

1. Entender la importancia de las encuestas y los cuestionarios en la sociedad de la información.

2. Conocer los aspectos clave a considerar cuando se elaboran las preguntas de un cuestionario.

3. Conocer los tipos de escalas más habituales en los cuestionarios, así como el tipo de datos que produce cada una de ellas.

4. Introducirse en los tipos de muestreo más habituales en los estudios de encuestas, en particular: el muestreo aleatorio simple, el muestreo sistemático, el muestreo por estratos y el muestreo por conglomerados.

5. Saber calcular estimaciones puntuales y por intervalos para diversos parámetros poblacionales según el tipo de muestreo usado.

6. Aprender a usar las técnicas estadísticas trabajadas durante el curso para analizar cuestionarios.

7. Aprender a usar programas estadísticos o de análisis de datos como instrumento básico en la aplicación práctica de los conceptos y técnicas estadísticas.

1. Diseño de cuestionarios

Las técnicas de investigación basadas en el uso de encuestas se aplican a multitud de ámbitos diferentes: en los negocios, en la administración pública, en las ciencias sociales y del comportamiento, en las ciencias de la información y la comunicación, en las ciencias de la salud, en las ciencias políticas, y en cualquier otro ámbito en el que los datos que puedan aportar los usuarios de un servicio o los consumidores de un producto jueguen un papel fundamental. En la Sociedad de la Información, las organizaciones e instituciones hacen un uso intensivo de los datos que explican cómo se comportan los individuos, cuáles son sus gustos y sus necesidades, qué opinión tienen sobre determinados temas, etc. En este contexto, las técnicas de investigación basadas en el uso de encuestas permiten obtener unos datos que, tras su posterior análisis estadístico, proporcionan una valiosa información tanto a los investigadores teóricos de una determinada disciplina como a los responsables de tomar decisiones sobre el funcionamiento de las organizaciones.

En general, se pueden distinguir seis fases secuenciales en el desarrollo de cualquier estudio basado en el uso de encuestas (figura 1): (a) identificación de los temas concretos sobre los que se desea obtener información así como de la población a encuestar, (b) diseño del cuestionario como instrumento para obtener los datos que se necesitan, (c) diseño y selección de una muestra representativa de la población, (d) obtención de los datos mediante el envío del cuestionario a los individuos que componen la muestra, (e) análisis estadístico de las observaciones muestrales a fin de inferir información sobre la población, y (f) elaboración de informes y conclusiones.

Figura 1. Fases en el desarrollo de una encuesta

En este apartado se hará especial énfasis en la fase de diseño del cuestionario, dejando para apartados posteriores otras fases clave en las que las técnicas estadísticas tienen una aportación decisiva, es decir, la fase de diseño y selección de la muestra y la fase de análisis de las observaciones muestrales.

1.1. Elaboración de las preguntas de un cuestionario

Las preguntas que se formulan en un cuestionario constituyen el aspecto más relevante de cualquier encuesta. Para que éstas cumplan su papel de forma efi-

ciente, las preguntas de un cuestionario deben centrarse en aquellos aspectos esenciales sobre los que se desea obtener información. Asimismo, dichas preguntas deben ser lo más breves y claras posibles a fin de facilitar la tarea de las personas encuestadas y maximizar la fiabilidad y validez del cuestionario. Se trata de evitar posibles problemas tales como: interpretaciones erróneas de las preguntas, agotamiento del encuestado o, incluso, rechazo a contestar una parte o la totalidad del cuestionario por la longitud del mismo o el esfuerzo necesario para entender las preguntas y contestarlas. Estas problemáticas podrían introducir sesgos y errores muestrales en los datos, lo que mermaría la fiabilidad y validez de la encuesta y de sus resultados.

Es importante ser cuidadoso en la elaboración de las preguntas a fin de evitar introducir en el cuestionario problemas de **error muestral** –debido al uso de una muestra para estimar parámetros poblacionales– o de **sesgo** (cualquier otro tipo de error en el cuestionario diferente del error muestral): si en la propia formulación de las pregunta se está induciendo al encuestado a responder en un sentido concreto, entonces se está introduciendo un sesgo en el cuestionario; si la formulación de las preguntas es ambigua y da pie a diferentes interpretaciones, entonces se está favoreciendo una excesiva dispersión de las respuestas, lo que incrementa el error muestral. Por tanto, la manera en cómo las preguntas se formulan en un cuestionario es determinante a la hora de evitar introducir patrones de sesgo y error muestral en el mismo. Así, se pueden establecer las siguientes recomendaciones generales a tener presentes cuando se elaboran las preguntas de un cuestionario:

- Criterios de interpretación y respuesta claros: los criterios en los que el encuestado debe basarse para interpretar y contestar a una pregunta deben estar claramente especificados en el cuestionario.

- Preguntas apropiadas al conjunto de individuos que configuran la muestra: todos los encuestados deben poder responder a las preguntas sobre la base de su experiencia o condición personal.

- Uso adecuado de expresiones, ejemplos o alternativas de respuesta: debe evitarse incluir en la pregunta expresiones que inciten a una determinada respuesta, así como ejemplos de posibles respuestas, ya que ello podría inducir a los encuestados a responder de una determinada manera y de este modo introducir un factor de sesgo en las respuestas.

- Nivel de actualidad de las preguntas: no se debería presuponer que el encuestado será capaz de recordar con precisión cuál fue su comportamiento en el pasado o su opinión sobre un tema acontecido hace ya bastante tiempo.

- Preguntas con un nivel de generalización o concreción adecuado: se debería evitar formular preguntas demasiado genéricas o ambiguas que se pue-

dan interpretar de formas muy distintas y cuya respuesta no aporte demasiada información, así como preguntas demasiado específicas que el encuestado no sea capaz de contestar con el nivel de detalle requerido.

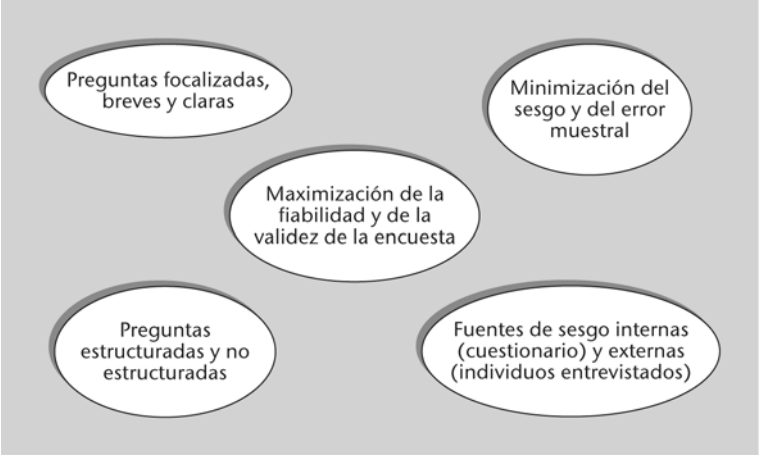
Además de estas fuentes internas de sesgo causadas por el propio instrumento de la encuesta, existen también otras potenciales fuentes de sesgo que no se originan por cómo se han elaborado las preguntas, sino por las condiciones en las que se ha respondido al cuestionario. Conviene conocer y tener presentes estas otras fuentes potenciales de sesgo para evitarlas en lo posible con una correcta elección de las condiciones de la encuesta y, en particular, de la muestra. Así, algunas de estas fuentes externas de sesgo son las siguientes: respuestas que buscan estar en coherencia con lo que es "socialmente deseable" o con lo que el entrevistador espera obtener, respuestas orientadas a dar una buena imagen del encuestado, respuestas con excesiva tendencia a la dicotomía (sí o no, positivo o negativo, etc.) o hacia las opciones extremas, respuestas hostiles excesivamente condicionadas por experiencias negativas recientes, etc.

Existen dos formatos básicos para elaborar preguntas de un cuestionario: las **preguntas abiertas** o no estructuradas son aquellas que permiten al encuestado responder libremente sin estar condicionado por un conjunto de posibles alternativas de respuesta. Por el contrario, las **preguntas estructuradas** o cerradas son aquellas que contienen en la propia pregunta un conjunto de posibles respuestas o categorías a elegir por el encuestado. La preguntas estructuradas son las que habitualmente más se usan en los cuestionarios, ya que además de acotar más claramente el contexto de la información que se espera obtener, suelen ser más fáciles y rápidas de contestar, permiten comparar mejor diferentes grupos de encuestados y, sobre todo, facilitan enormemente el procesado y análisis posterior de los datos.

Cuando se usan preguntas estructuradas es importante elegir bien las categorías o posibles respuestas alternativas de manera que éstas constituyan una lista completa de opciones (incluyendo opciones como "otros" o "no sabe o no contesta" cuando sea necesario) y sean mutuamente excluyentes (a menos que sean de opción múltiple). Por lo que respecta al número de categorías o respuestas alternativas, lo recomendable es que se sitúe entre un mínimo de dos para preguntas dicotómicas y un máximo de seis. Añadir más categorías suele dificultar en exceso la tarea del encuestado. Hay que tener presente, sin embargo, que en caso de duda sobre el nivel de detalle que se quiera ofrecer en las categorías, suele ser preferible optar por la opción con más categorías, puesto que siempre es posible combinar o agregar categorías a posteriori –durante la fase de análisis–, mientras que la operación de desagregar respuestas ya obtenidas en nuevas categorías no suele ser posible sin la consiguiente pérdida de precisión e información.

La figura 2 sintetiza los conceptos clave que se deben tener en cuenta en la elaboración de las preguntas de cualquier cuestionario.

Figura 2. Conceptos clave en la elaboración de las preguntas de un cuestionario



1.2. Uso de escalas en preguntas estructuradas

Las respuestas a preguntas estructuradas consisten, por lo general, en elegir una opción concreta en una lista de categorías posibles. Estas categorías siguen una escala o graduación que puede ser simplemente nominal o bien puede implicar algún tipo de relación ordinal o numérica entre las distintas categorías implicadas:

- **Escalas nominales:** son aquellas en las que las categorías no están asociadas a una relación de orden o de magnitud. Un ejemplo sería una escala en la que las categorías fuesen distintos códigos postales, prefijos telefónicos o identificadores del sexo ("hombre", "mujer"). Este tipo de escala proporciona datos de tipo nominal que simplemente identifican categorías, por lo que es el más limitado desde el punto de vista de las técnicas estadísticas que se pueden aplicar a las observaciones obtenidas.

- **Escalas ordinales:** son aquellas cuyas categorías siguen una relación de orden o preferencia, aunque no de magnitud, que permite clasificarlas. Un ejemplo sería una escala de tareas secuenciales a realizar en un proceso, en el que la pregunta podría ser elegir aquella tarea que se considere más crítica. Este tipo de escalas posibilita el uso de las llamadas técnicas estadísticas no paramétricas para analizar los datos obtenidos.

- **Escalas de intervalos equidistantes:** son las que asocian una magnitud a cada categoría y en las que el cero no significa ausencia de magnitud. Un ejemplo sería una escala graduada del 1 al 7 para representar niveles de importancia. Esta escala permite el uso de técnicas de inferencia estadística, por lo que resulta altamente recomendable.

- **Escalas de ratio:** son las que asocian una magnitud a cada categoría y en las que el cero representa ausencia de magnitud. Un ejemplo sería una escala graduada del 0 al 50 para indicar la distancia en kilómetros recorrida por el encuestado para acudir a su lugar de trabajo. Al igual que ocurría con las escalas de intervalos equidistantes, las de ratio también permiten el uso de técnicas de inferencia estadística.

A continuación, se describen algunos ejemplos de escalas particulares que se usan habitualmente en los cuestionarios:

Nota

Los ejemplos sólo cubren algunas de las tipologías de escalas más usadas. En Internet es fácil encontrar ejemplos de cuestionarios completos y otros tipos de escalas sin más que buscar por términos como *survey examples*, *questionnaire examples*, etc.

- La **escala de Likert**: esta escala suele usarse para obtener el grado de acuerdo o desacuerdo del encuestado con una determinada afirmación (figura 3). Puesto que todas las categorías en una escala de Likert suelen estar etiquetadas (y las etiquetas o identificadores de cada categoría no tienen por qué representar magnitudes equidistantes), hay cierta discrepancia entre los expertos sobre si esta escala debe considerarse simplemente como una escala ordinal o bien puede incluso considerarse como una escala de intervalos. Una posible solución a este problema sería mantener únicamente los identificadores o etiquetas de los extremos (p. ej.: "(1) Muy en desacuerdo" y "(5) muy de acuerdo"), dejando el resto de ítems numerados pero sin etiquetar, de modo que los números definan intervalos equidistantes. En todo caso, es éste un tema bastante discutible sobre el que no parece haber un total consenso. Obviamente, resulta muy ventajoso poder considerar una escala de Likert como de intervalos para poder así aplicar técnicas de inferencia estadística de forma lícita.

Figura 3. Ejemplo de preguntas usando una escala de Likert

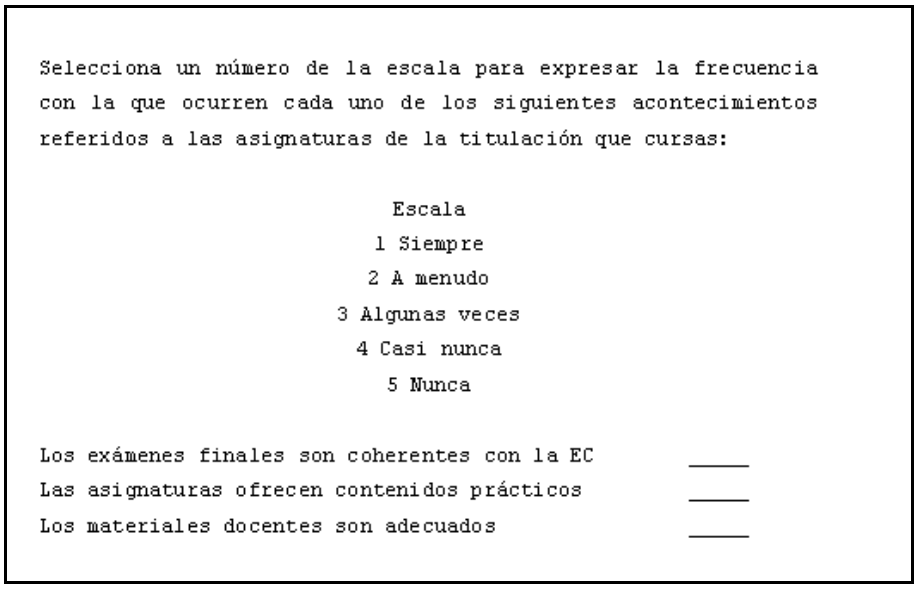
Selecciona un número de la escala para expresar en qué medida estás en acuerdo o en desacuerdo con cada una de las afirmaciones siguientes referidas a la asignatura Estadística:

Escala
1 Totalmente de acuerdo
2 De acuerdo
3 Neutral
4 En desacuerdo
5 Totalmente en desacuerdo

Los exámenes finales son coherentes con la EC _____
La asignatura ofrece contenidos prácticos _____
Los materiales docentes son adecuados _____

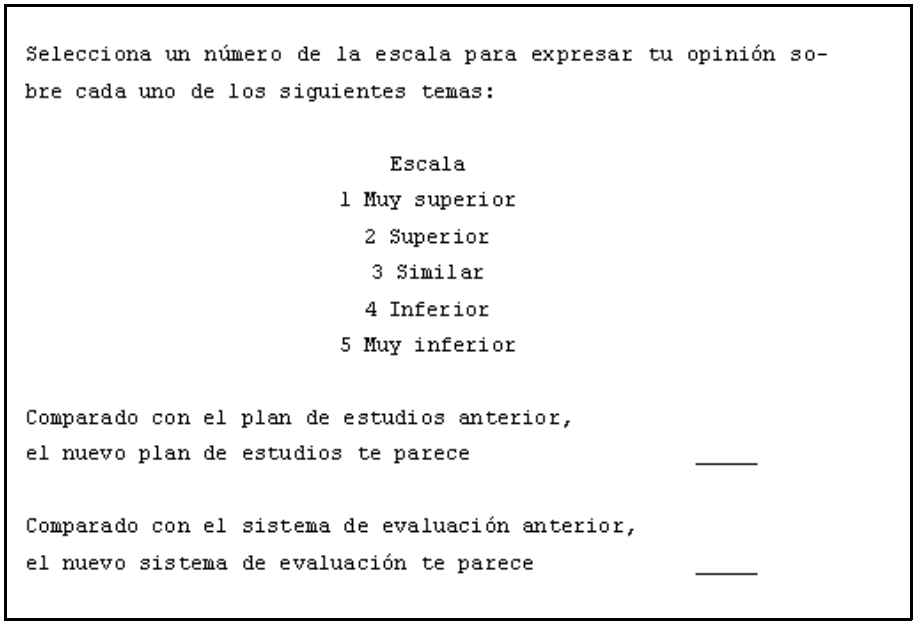
- La **escala de frecuencia verbal**: esta escala es muy similar a la de Likert, con la diferencia de que los ítems de la escala indican con qué frecuencia se ha llevado a cabo una determinada acción (figura 4).

Figura 4. Ejemplo de preguntas usando una escala de frecuencia verbal



- La **escala comparativa**: a diferencia de las anteriores, los ítems de esta escala indican cómo se comparan dos elementos entre sí a criterio del encuestado (figura 5). Esta escala se considera como una escala de intervalos, por lo que es lícito aplicar las técnicas de inferencia a los datos obtenidos con ella.

Figura 5. Ejemplo de uso de una escala comparativa



- La **escala linealnumérica**: esta escala también es similar a la de Likert, aunque los ítems extremos suelen hacer referencia al grado de importancia que asigna el encuestado a un tema y los ítems intermedios no suelen estar etiquetados (figura 6). Por esto último, se considera una escala de intervalos.

Figura 6. Ejemplo de uso de una escala lineal-numérica

Selecciona un número de la escala para expresar tu opinión sobre el nivel de relevancia de cada uno de los siguientes temas referidos a las asignaturas que cursas:

	Escala						
	Máxima relevancia	1	2	3	4	5	6 Mínima relevancia
El uso de recursos de Internet							_____
El uso de materiales actualizados							_____
El uso de los foros y debates							_____

- La **escala de diferencias semánticas**: esta escala consiste en definir dos extremos caracterizados por adjetivos contrapuestos y, posteriormente, definir una graduación de ítems no etiquetados entre ambos (figura 7). También se considera como una escala de intervalos.

Figura 7. Ejemplo de uso de una escala de diferencias semánticas

En relación a la formación que recibes en esta universidad, selecciona un valor numérico según lo próxima que esté con respecto a cada adjetivo:

Teórica	1	2	3	4	5	6	7	Práctica
Económica	1	2	3	4	5	6	7	Cara
Actualizada	1	2	3	4	5	6	7	Desfasada

2. Diseño y selección de la muestra

Como ya se ha comentado en el apartado anterior, en toda encuesta hay dos tipos de errores que conviene tener presentes: (a) el error muestral, que es la diferencia entre el estimador obtenido a partir de las observaciones (p. ej., la media muestral $\bar{x}$) y el verdadero valor del parámetro poblacional (p. ej., la media poblacional μ), y (b) el sesgo o error no muestral, que engloba todos los restantes tipos de errores que pueden ocurrir durante el desarrollo y análisis de una encuesta, es decir, errores en el diseño de las preguntas, errores causados por la "no-respuesta" (*missing data*), errores en la selección de los individuos a encuestar, errores en el registro y procesado de los datos, etc.

Ejemplo

Recordar que el término *estadístico* hace referencia a una muestra mientras que el término *parámetro* hace referencia a toda la población. Así, por ejemplo, el estadístico media muestral es un estimador del parámetro media poblacional.

Las encuestas pueden clasificarse en función del método de muestreo usado. Así, se habla de **muestreo probabilístico** cuando cada uno de los individuos que componen el marco del muestreo (elementos de la población susceptibles de ser elegidos) tiene una probabilidad conocida de ser seleccionado. Por el contrario, se habla de **muestreo no probabilístico** cuando no es posible saber cuál es la probabilidad de cada elemento de ser seleccionado. Los muestreos no probabilísticos pueden ser de gran utilidad como herramienta exploratoria, pero no permiten conocer la precisión de las estimaciones que se obtienen para los parámetros poblacionales, es decir, no dan información sobre el error muestral que se está cometiendo. Ejemplos de muestreos no probabilísticos serían los siguientes:

- A fin de conocer la opinión de los estudiantes de una universidad presencial sobre su nuevo Campus Virtual, se encuesta a los matriculados de una asignatura concreta.

- A fin de conocer la opinión de los clientes de un nuevo centro comercial, se piden voluntarios para responder a un cuestionario.

- A fin de conocer la opinión de los usuarios de una base de datos documental, un directivo selecciona una muestra de usuarios que, según su criterio, son representativos del conjunto de usuarios.

Los muestreos probabilísticos, por su parte, sí permiten calcular intervalos de confianza para los parámetros poblacionales a partir de los las observaciones de la muestra. Esto es, los muestreos probabilísticos permiten conocer la magnitud del error muestral que se está cometiendo. En este apartado se describirán cuatro de los métodos probabilísticos más populares: el muestreo aleatorio simple, el muestreo sistemático, el muestreo estratificado, y el muestreo por conglomerados.

2.1. Muestreo aleatorio simple

En un **muestreo aleatorio simple**, todos los elementos del marco muestral (elementos de la población que son candidatos a ser seleccionados) tienen la misma probabilidad de ser elegidos. Para seleccionar, mediante muestreo aleatorio simple, n elementos de entre los N que componen la lista de candidatos a ser elegidos, se suele asignar un número natural (1, 2, 3, …, N) a cada uno de los elementos de la lista y, a continuación, se generan al azar n números aleatorios distintos, que identificaran a los elementos seleccionados.

De acuerdo con la teoría de la estadística inferencial, si se selecciona una muestra aleatoria suficientemente grande (en la práctica $n \geq 30$ suele ser suficiente), el teorema central del límite permite obtener intervalos de confianza para la media poblacional μ. En particular:

Para un nivel de confianza del 95%, un intervalo de confianza para la media poblacional, μ, viene dado por:

$$\bar{x} \pm 1{,}96 \cdot \sqrt{\frac{N-n}{N}}\left(\frac{s}{\sqrt{n}}\right)$$

donde s representa la desviación estándar de las observaciones muestrales.

Ejemplo: un periódico de economía tiene actualmente $N = 8.000$ lectores suscritos. Una muestra aleatoria simple de $n = 484$ lectores es elegida para realizar una encuesta. Tras analizar los datos de dicha encuesta, se sabe que la media de los ingresos mensuales de los lectores seleccionados en la muestra es de $\bar{x} = 30.500$ euros y que la correspondiente desviación estándar es de $s = 7.040$ euros.

La media muestral, $\bar{x}$, es un buen estimador de la media poblacional, μ. Además, un intervalo de confianza del 95% para dicha media poblacional será: $30.500 \pm 1{,}96 \cdot \sqrt{\frac{8.000-484}{8.000}}\left(\frac{7.040}{\sqrt{484}}\right)$ = (29.892,07, 31.107,93). En otras palabras, para un nivel de confianza del 95%, los ingresos medios del conjunto de los 8.000 lectores suscritos al periódico oscilarán entre 29.892 y 31.108 euros.

De forma similar, es posible calcular intervalos de confianza para otros parámetros de la población, como el total acumulado de una población, por ejemplo, la demanda total de la población, la riqueza total de una población, etc.

El estadístico $N \cdot \bar{x}$ es un buen estimador del total acumulado de una población, $N \cdot \mu$. Además, si (a, b) es un intervalo de confianza del 95% para la media poblacional, μ, un intervalo de confianza del 95% para $N \cdot \mu$, viene dado por $(N \cdot a, N \cdot b)$.

Ejemplo: se desea estimar el número total de visitas anuales que reciben los portales web de las universidades pertenecientes a una clasificación que incluye las quinientas mejores del mundo. Para ello, se ha seleccionado una muestra aleatoria de cincuenta universidades pertenecientes a esa clasificación y se han obtenido los siguientes estadísticos muestrales: el número medio de visitas anuales es de veintidós mil, siendo la desviación estándar de cuatro mil.

En primer lugar, cabe destacar que $N \cdot \bar{x}$ = 11.000.000 será un buen estimador para el número total de visitas anuales que reciben los portales de las quinientas mejores universidades. Un intervalo de confianza del 95% para el número total de visitas anuales será: $500 \cdot 22.000 \pm 1,96 \cdot 500 \cdot \sqrt{\frac{500-50}{500}} \left(\frac{4.000}{\sqrt{50}} \right)$ = (10.474.077, 11.525.923). En otras palabras, para un nivel de confianza del 95%, el número total de visitas anuales que recibirán los quinientos portales web estará entre 10,47 millones y 11,53 millones.

Indicación

Al realizar los cálculos, se recomienda usar al menos cuatro decimales para no perder demasiada precisión en el redondeo, especialmente cuando *N* es un número muy grande.

Finalmente, también es posible obtener intervalos de confianza para la proporción de elementos de una población que satisfacen unas determinadas condiciones, por ejemplo, proporción de individuos que usan un servicio, proporción de individuos con estudios superiores, etc.

Para un nivel de confianza del 95%, un intervalo de confianza para la proporción p de elementos de una población que cumple una determinada condición viene dado por:

$$p' \pm 1,96 \cdot \sqrt{\left(\frac{N-n}{N} \right) \cdot \left(\frac{p'(1-p')}{n-1} \right)}$$

donde p' es la proporción de elementos de la muestra que la cumplen.

Ejemplo: siguiendo con el ejemplo anterior de los portales web de las universidades pertenecientes a la clasificación de las 500 mejores, se desea estimar el porcentaje de portales que disponen de un programa institucional –al estilo del MIT OpenCourseWare– para ofrecer contenidos formativos en abierto. De las cincuenta universidades que constituyen la muestra, un total de treinta y cinco disponen de dicho programa.

La proporción muestral, p' = 35/50 = 0,70 = 70%, es un buen estimador del porcentaje de universidades en las quinientas mejores que tendrán un programa así. Además, es posible obtener un intervalo de confianza del 95% para dicha proporción poblacional: $0,7 \pm 1,96 \cdot \sqrt{\left(\frac{500-50}{500} \right) \cdot \left(\frac{0,7(1-0,7)}{50-1} \right)}$ = (0,5783, 0,8217). En otras palabras, con un nivel de confianza del 95% se puede afirmar que entre el 58% y el 82% de universidades entre las quinientas mejores disponen de un programa de contenidos formativos en abierto. Observar que, en

este caso, el intervalo de confianza es poco preciso (hay unos veinticuatro puntos porcentuales de diferencia entre los extremos del intervalo), lo cual se debe a que el tamaño de la muestra es relativamente pequeño.

2.2. Muestreo sistemático

El **muestreo sistemático** consiste en usar una regla para seleccionar de forma sistemática los elementos de una muestra. Este muestreo se suele usar en poblaciones grandes y homogéneas como alternativa al muestreo aleatorio simple, especialmente en aquellas situaciones en las que el proceso de asignar un número entero a cada elemento de una larga lista puede resultar complicado o costoso en tiempo (p. ej.: asignar un número entero a cada uno de los números de una guía telefónica, asignar un número entero a cada uno de los clientes que accede a un centro comercial en un día determinado, etc.). Así, por ejemplo, si se desea seleccionar una muestra de treinta teléfonos de la guía telefónica de una ciudad, una forma sistemática de hacerlo sería escoger al azar el primero y, posteriormente, escoger un teléfono cualquiera de cada una de las veintinueve páginas siguientes. Otro ejemplo: si se desea entrevistar a cuarenta clientes de un gran centro comercial, una forma sistemática de seleccionar la muestra sería empezar por uno al azar y, a continuación, escoger cada cinco minutos al nuevo cliente que acceda al centro en ese preciso instante. A menudo, este tipo de muestreo se puede considerar como equivalente a un muestreo aleatorio simple, especialmente cuando el listado o marco muestral sigue un orden aleatorio, es decir, realizar una selección sistemática de elementos en una lista que sigue un orden aleatorio es técnicamente equivalente a realizar directamente una selección aleatoria de elementos que no sigan un orden aleatorio.

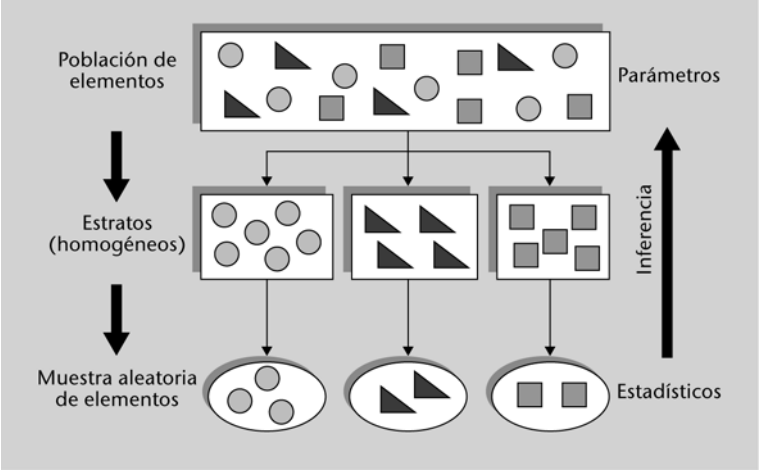
2.3. Muestreo aleatorio estratificado (grupos homogéneos)

El **muestreo aleatorio estratificado** se suele usar en los casos en que resulta fácil agrupar los elementos de la población considerada en subgrupos de composición homogénea llamados **estratos**. Por ejemplo: trabajadores de una organización agrupados por departamento, estudiantes de una universidad agrupados por titulación, habitantes de un país agrupados por nivel de renta o edad, revistas científicas agrupadas por ámbito temático, etc. Cuando la variabilidad dentro de cada estrato es menor que la variabilidad entre estratos, este tipo de muestreo tiende a proporcionar más precisión que un muestreo aleatorio simple a la hora de estimar los parámetros poblacionales.

Así, el muestreo aleatorio por estratos consiste en: (a) clasificar los N elementos de una población en H grupos o estratos (de manera que los elementos de cada estrato sean similares entre ellos), y (b) seleccionar a continuación una muestra aleatoria simple para cada uno de los estratos (figura 8). Los estadísticos obtenidos para cada estrato se combinan posteriormente para obtener es-

timaciones de algunos parámetros como la media, el total acumulado o la proporción de la población.

Figura 8. Muestreo aleatorio estratificado



En un muestreo por estratos, es posible obtener un buen estimador de la media poblacional haciendo un promedio ponderado de las medias muestrales obtenidas en cada estrato. En concreto, $\bar{x}_E = \frac{1}{N}\sum_{i=1}^{H} N_i \cdot \bar{x}_i$ es un buen estimador de μ, donde N_i representa el número total de elementos del estrato *i*-ésimo y $\bar{x}_i$ representa la media de la muestra asociada a dicho estrato.

Para un nivel de confianza del 95%, un intervalo de confianza para la media poblacional, μ, viene dado por:

$$\bar{x}_E \pm 1{,}96 \cdot \sqrt{\frac{1}{N^2}\sum_{i=1}^{H} N_i (N_i - n_i)\frac{s_i^2}{n_i}}$$

donde n_i y s_i representan, respectivamente, el tamaño y la desviación estándar de la muestra asociada al estrato *i*-ésimo.

Por otro lado, el estadístico $N \cdot \bar{x}_E$ es un buen estimador del total acumulado de una población, $N \cdot \mu$. Además, si (a, b) es un intervalo de confianza del 95% para la media poblacional, μ, un intervalo de confianza del 95% para $N \cdot \mu$, viene dado por $(N \cdot a, N \cdot b)$.

Finalmente, un intervalo de confianza para la proporción p de elementos de una población que cumple una determinada condición viene dado por:

$$p'_E \pm 1{,}96 \cdot \sqrt{\frac{1}{N^2}\sum_{i=1}^{H} N_i (N_i - n_i) \cdot \left(\frac{p'_i (1 - p'_i)}{n_i - 1}\right)}$$

donde $p'_E = \frac{1}{N}\sum_{i=1}^{H} N_i \cdot p'_i$ es un promedio ponderado de las proporciones p'_i de elementos de la muestra que la cumplen para el estrato *i*-ésimo.

Ejemplo: hace dos años se graduaron en una universidad un total de 1.500 estudiantes. Para conocer el salario medio de dichos estudiantes, tanto a nivel global como por titulación, se agruparon los estudiantes por titulaciones (estratos) y se encuestó a un total de ciento ochenta exestudiantes. La tabla 1 incluye, por orden de columnas, el número de graduados en cada estrato, el tamaño de cada muestra, la media muestral, la desviación estándar muestral y la proporción de estudiantes con un sueldo superior a los 36.000 euros anuales.

Tabla 1. Estadísticos obtenidos para cada estrato

Titulación (estrato)	N_i	n_i	$\bar{x}_i$	s_i	p'_i
Administración y Dirección de Empresas	500	45	30.000	2.000	4/45
Información y Documentación	350	40	28.500	1.700	2/40
Ingeniería Informática	200	30	31.500	2.300	7/30
Psicología	300	35	27.000	1.600	1/35
Ingeniería de Telecomunicaciones	150	30	31.000	2.250	6/30
Total	1.500	180			

Un buen estimador del salario medio para el conjunto de mil quinientos graduados viene dado por el promedio ponderado de las distintas medias muestrales:

$$\bar{x}_E = \frac{1}{1.500}\left(500 \cdot 30.000 + 350 \cdot 28.500 + 200 \cdot 31.500 + 300 \cdot 27.000 + 150 \cdot 31.000\right)$$

$= 29.350$ euros.

Además, se puede obtener el correspondiente intervalo de confianza, para un nivel de confianza del 95%, para la media poblacional:

$$29.350 \pm 1,96 \cdot \sqrt{\frac{1}{1.500^2}\left(500 \cdot (500-45)\frac{2.000^2}{45} + \ldots + 150 \cdot (150-30)\frac{2.250^2}{30}\right)} =$$

(29.079,33, 29.620,67), es decir, se puede afirmar, con un nivel de confianza del 95%, que el salario medio del total de mil quinientos graduados de esta universidad está entre 29.079 y 29.621 euros por año. Para hacer este tipo de cálculos es conveniente usar una hoja de cálculo (figura 9).

Figura 9. Uso de Excel para realizar cálculos en muestreo estratificado

	A	B	C	D	E	F	G	H
1	**Titulación (estrato)**	**N(i)**	**n(i)**	**x-bar(i)**	**s(i)**	**p(i)**	**N(i) * x-bar(i)**	**N(i) * (N(i) - n(i)) * (s(i)^2 / n(i))**
2	Dirección de Empresas	500	45	30.000	2.000	16.528	15.000.000	20.222.222.222
3	Información y Documentación	350	40	28.500	1.700	14.642	9.975.000	7.839.125.000
4	Ing. Informática	200	30	31.500	2.300	40.024	6.300.000	5.995.333.333
5	Psicología	300	35	27.000	1.600	12.785	8.100.000	5.814.857.143
6	Ing. Telecomunicación	150	30	31.000	2.250	39.994	4.650.000	3.037.500.000
7	*Totales*	*1.500*	*180*				*44.025.000*	*42.909.037.698*
8								
9			z =	1,96			$\sum_{i=1}^{5} N_i \cdot \bar{x}_i$	$\sum_{i=1}^{5} N_i (N_i - n_i) \frac{s_i^2}{n_i}$
10			x(E) =	29.350				
11			s(E) =	138,10				
12			x(E) - z*s(E) =	29.079,3306				
13			x(E) + z*s(E) =	29.620,6694				
14								

En segundo lugar, se pueden estimar los ingresos anuales totales del conjunto de los mil quinientos graduados, $N \cdot \mu$, para saber cuál será su potencial impacto sobre la economía local. En este caso, puesto que el estimador de μ era $\bar{x}_E$ = 29.350 euros, el estimador puntual de $N \cdot \mu$ será 1.500 $\bar{x}_E$ = 44.025.000 y un intervalo de confianza al 95% vendrá dado por: (1.500 · 29.079,3306, 1.500 · 29.620,6694) = (43.618.995,86, 44.431.004,14). En otras palabras, se puede afirmar con un nivel de confianza del 95% que serán necesarios entre 43,6 y 44,4 millones de euros para cubrir los salarios anuales de los mil quinientos graduados.

En tercer lugar, un buen estimador del porcentaje de estudiantes de la población cuyos ingresos superan los 36.000 euros vendrá dado por el promedio ponderado de los porcentajes en cada estrato: $p'_E = \frac{1}{1.500}\left(500\frac{4}{45} + \ldots + 150\frac{6}{30}\right) = 0{,}0981$, es decir, aproximadamente, sólo un 9,8% de los salarios de los mil quinientos graduados será superior a los 36.000 euros anuales. Finalmente, se puede obtener un intervalo de confianza del 95% para el porcentaje poblacional anterior:

$$0{,}0981 \pm 1{,}96 \cdot \sqrt{\frac{1}{1.500^2}\left(500(500-45)\frac{(4/45)(41/45)}{45-1} + \ldots + 150(150-30)\frac{(6/30)(24/30)}{30-1}\right)}$$

= (0,0584, 0,1379), es decir, se puede afirmar con un 95% de confianza que el porcentaje de graduados en la promoción de hace dos años cuyos ingresos superan los 36.000 euros anuales oscila entre un 5,8% y un 13,8%.

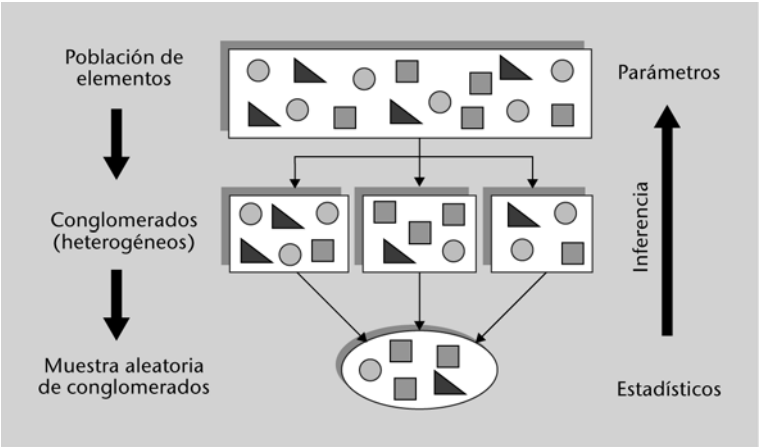
2.4. Muestreo por conglomerados (*clusters* o grupos heterogéneos)

El **muestreo por conglomerados** se suele usar en los casos en que resulta fácil agrupar los elementos de la población considerada en subgrupos de

composición heterogénea llamados conglomerados, cada uno de los cuales viene a ser una representación a pequeña escala de la población total (es decir, se presupone una gran variabilidad entre los elementos de un mismo conglomerado). Por ejemplo: los habitantes de una gran ciudad pueden agruparse por barrios, los usuarios de un servicio web pueden agruparse por países de procedencia, las revistas científicas pueden agruparse por editorial, etc. De hecho, una de las principales aplicaciones del muestreo por conglomerados está relacionada con el muestreo por áreas o regiones geográficas, donde los conglomerados suelen ser países, regiones, ciudades o barrios. El muestreo por conglomerados permite reducir los costes de desplazamientos entre zonas geográficamente dispersas y, a la vez, evita tener que generar listados exhaustivos de toda la población, puesto que sólo son necesarios listados exhaustivos de cada conglomerado seleccionado.

Así, el muestreo por conglomerados consiste en: (a) clasificar los N elementos de una población en H grupos o conglomerados (de manera que los elementos de cada conglomerado presenten mucha variabilidad entre ellos), (b) seleccionar a continuación una muestra aleatoria simple de h conglomerados, y (c) para cada conglomerado de la muestra seleccionada, o bien encuestar a cada uno de los elementos que lo componen –muestreo por conglomerados en una etapa– o bien seleccionar una nueva muestra aleatoria de elementos para encuestar –muestreo en dos etapas– (figura 10). Si bien tanto en un caso como en otro es posible obtener estimadores puntuales y por intervalos para varios parámetros poblacionales, se tratará sólo el muestreo por conglomerados en una etapa (es decir, se supondrá que, una vez seleccionada la muestra de conglomerados, se encuesta a todos los elementos de cada conglomerado seleccionado).

Figura 10. Muestreo por conglomerados



En un muestreo por conglomerados es posible obtener un buen estimador de la media poblacional μ mediante la expresión $\bar{x}_C = \frac{\sum_{i=1}^{h} y_i}{\sum_{i=1}^{h} N_i}$, donde N_i represen-

ta el número total de elementos del conglomerado *i*-ésimo e y_i representa el valor total de las observaciones de dicho conglomerado.

Para un nivel de confianza del 95%, un intervalo de confianza para la media poblacional, μ, viene dado por:

$$\bar{x}_C \pm 1{,}96 \cdot \sqrt{\frac{H-h}{H \cdot h\left(\frac{N}{H}\right)^2}\left(\frac{\sum_{i=1}^{h}\left(y_i - \bar{x}_C \cdot N_i\right)^2}{h-1}\right)}$$

Por otro lado, el estadístico $N \cdot \bar{x}_C$ es un buen estimador del total acumulado de una población, $N \cdot \mu$. Además, si (a, b) es un intervalo de confianza del 95% para la media poblacional μ, un intervalo de confianza del 95% para $N \cdot \mu$ viene dado por $(N \cdot a, N \cdot b)$

Finalmente, un intervalo de confianza para la proporción *p* de elementos de una población que cumple una determinada condición viene dado por:

$$p'_C \pm 1{,}96 \cdot \sqrt{\frac{H-h}{H \cdot h\left(\frac{N}{H}\right)^2}\left(\frac{\sum_{i=1}^{h}\left(m_i - p'_C \cdot N_i\right)^2}{h-1}\right)}$$

donde m_i es el número de elementos del conglomerado *i*-ésimo que cumple una determinada característica y $p'_C = \frac{\sum_{i=1}^{h} m_i}{\sum_{i=1}^{h} N_i}$ es buen estimador del promedio de elementos de la población que cumplen dicha característica.

Ejemplo: el sistema sanitario de atención primaria de un país está compuesto por un total de doce mil médicos distribuidos en mil centros de atención primaria (conglomerados). Con el fin de obtener cierta información sobre la población de médicos considerada, y ante la dificultad de realizar encuestas a médicos de todos los centros, se lleva a cabo un muestreo por conglomerados en el que se seleccionan de forma aleatoria un total de diez centros de atención primaria. A continuación, se pasa una encuesta a los médicos de cada uno de los centros escogidos. La tabla 2 incluye, por orden de columnas, el identificador del centro, el número de médicos que en él trabajan, el número total de visitas asociadas con una cierta enfermedad que recibe el centro en una semana normal y el número de médicos que son mujeres.

Tabla 2. Estadísticos obtenidos para cada conglomerado de la muestra

Centro (conglomerado)	Número de médicos N_i	Total de visitas y_i	Número de mujeres m_i
CAP-01	8	320	2
CAP-02	25	1.125	8
CAP-03	4	115	0
CAP-04	17	714	6
CAP-05	7	247	1
CAP-06	3	94	2
CAP-07	15	634	2
CAP-08	4	147	0
CAP-09	12	481	5
CAP-10	33	1.567	9
Totales	**128**	**5.444**	**35**

En primer lugar, un buen estimador para el número medio de visitas semanales que recibe cada médico viene dado por: $\bar{x}_C = \frac{5.444}{128} = 42{,}5313$, es decir, en promedio cada médico del sistema sanitario recibirá unas cuarenta y tres visitas semanales.

Es posible obtener un intervalo de confianza del 95% para dicha media poblacional:

$$42{,}5313 \pm 1{,}96 \cdot \sqrt{\frac{1.000-10}{1.000 \cdot 10\left(\frac{12.000}{1.000}\right)^2}\left(\frac{(320-42{,}5313\cdot 8)^2+\ldots+(1.567-42{,}5313\cdot 33)^2}{10-1}\right)}$$

$= 42{,}5313 \pm 1{,}96 \cdot 1{,}7299 = (39{,}14,\ 45{,}92)$. En otras palabras, se puede afirmar con un nivel de confianza del 95% que el promedio de visitas semanales por médico en el sistema sanitario del país está entre 39 y 46 (figura 11).

Figura 11. Uso de Excel para realizar cálculos en muestreo por conglomerados

	A	B	C	D	E
1	Centro	Número de médicos	Total de visitas	Número de mujeres	
2	(conglomerado)	N_i	Y_i	m_i	[y(i) - x(C)*N(i)]^2
3	CAP-01	8	320	2	410,06
4	CAP-02	25	1125	8	3809,20
5	CAP-03	4	115	0	3038,77
6	CAP-04	17	714	6	81,56
7	CAP-05	7	247	1	2572,39
8	CAP-06	3	94	2	1128,54
9	CAP-07	15	634	2	15,75
10	CAP-08	4	147	0	534,77
11	CAP-09	12	481	5	862,89
12	CAP-10	33	1567	9	26722,03
13	*Totales*	*128*	*5444*	*35*	*39175,97*
14					
15		z =	1,96		
16		x(C) =	42,53		
17		s(C) =	1,73		
18		x(C) - z*s(C) =	39,14		
19		x(C) + z*s(C) =	45,92		

$$\sum_{i=1}^{10}\left(y_i - \bar{x}_C \cdot N_i\right)^2$$

En segundo lugar, se pueden estimar las visitas semanales totales del conjunto de los doce mil médicos, $N \cdot \mu$, para saber cuál será su potencial impacto sobre el sistema sanitario. En este caso, puesto que el estimador de μ era $\bar{x}_C = 42{,}5313$, el estimador puntual de $N \cdot \mu$ será $12.000\,\bar{x}_C = 510.375$ y un intervalo de confianza del 95% vendrá dado por: (12.000 · 39,1.406, 12.000 · 45,9219) = (469.687,38, 551.062,62). En otras palabras, se puede afirmar con un nivel de confianza del 95% que el sistema de atención primaria del país recibirá entre 469.687 y 551.063 visitas en una semana normal.

En tercer lugar, un buen estimador del porcentaje de médicos que son mujeres vendrá dado por: $p'_C = \frac{35}{128} = 0{,}2734$, es decir, aproximadamente el 27,3% de los médicos del sistema de atención primaria son mujeres. Finalmente, se puede obtener un intervalo de confianza del 95% para el porcentaje poblacional anterior:

$$0{,}2734 \pm 1{,}96 \cdot \sqrt{\frac{1.000-10}{1.000 \cdot 10\left(\frac{12.000}{1.000}\right)^2}\left(\frac{(2-0{,}2734\cdot 8)^2 + \ldots + (9-0{,}2734\cdot 33)^2}{10-1}\right)}$$

= (0,2066, 0,3402), es decir, se puede afirmar con un 95% de confianza que el porcentaje de mujeres en la población de médicos de asistencia primaria oscila entre un 20,7% y un 34,0%.

3. Análisis de cuestionarios: estudio parcial de un caso

En este apartado se presenta un caso de estudio en el que se muestran ejemplos del uso de técnicas estadísticas para analizar diferentes tipos de preguntas pertenecientes a una encuesta. El objetivo de la encuesta era obtener información concreta sobre la visión (y la actitud) de las grandes empresas de una determinada comunidad autónoma respecto al fenómeno de la externalización de los servicios, sistemas y tecnologías de la información. Para ello, se diseñó una encuesta formada por varias preguntas, algunas de ellas basadas en escalas nominales y otras en escalas de intervalos equidistantes. La población objetivo de la encuesta eran los directivos de servicios, sistemas y tecnologías de la información de las empresas, con sede social en dicha comunidad autónoma, cuyo volumen de facturación o de empleados superaban unas determinadas cantidades establecidas a priori por los investigadores. Del listado completo de empresas que cumplían dichos requisitos, se seleccionó una muestra aleatoria de cien empresas y se mandó el cuestionario a los correspondientes directivos, tras lo que se obtuvo una tasa de respuesta superior al 80%. La aleatoriedad de la muestra y la alta tasa de respuesta obtenida son dos factores imprescindibles a la hora de generalizar, con ciertas garantías, los resultados de la encuesta al conjunto de la población de empresas que satisfacen las características anteriormente descritas.

Aclaración

El objetivo último de esta sección no es explicar con detalle un caso completo de análisis de una encuesta (ya que ello requeriría de un módulo entero), sino proporcionar ejemplos concretos de cómo se pueden utilizar muchos de los conceptos y técnicas vistas en módulos anteriores para analizar encuestas. Así pues, esta sección muestra cómo se pueden combinar muchas de las técnicas estadísticas anteriormente vistas para extraer información a partir de los datos de una encuesta.

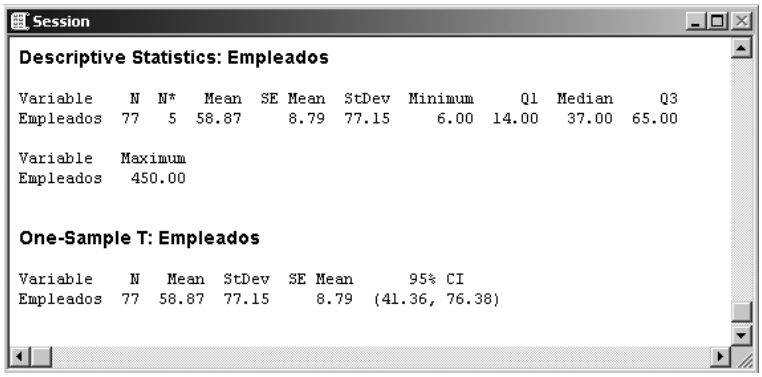
3.1. Ejemplo de uso de estadísticos descriptivos e intervalos de confianza

Una de las preguntas de la encuesta pedía especificar el número de trabajadores de plantilla del departamento de tecnologías de la información y la comunicación (TIC). Dicha pregunta está asociada a una variable aleatoria discreta, por lo que se pueden considerar los estadísticos descriptivos de la misma como muestra la figura 12.

Nota

Tanto los *outputs* como los gráficos de esta sección han sido generados con Minitab, usando los menús y opciones ya explicadas en módulos anteriores.

Figura 12. Estadísticos descriptivos de la variable "N.º de empleados"



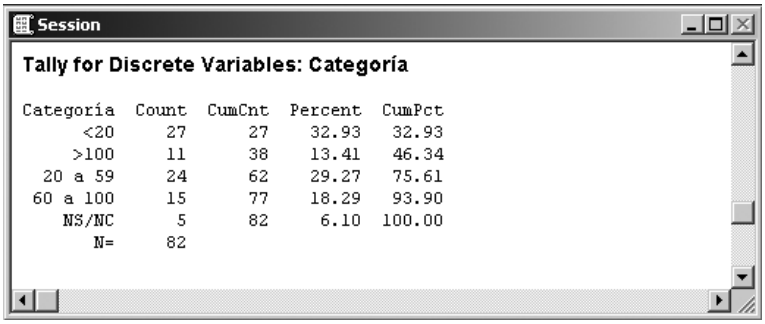
Recordatorio Minitab

Para obtener los estadísticos descriptivos, usar ***Stat*** > ***Basic Statistics*** > ***Display Descriptive Statistics***. Para obtener el Intervalo de Confianza usar ***Stat*** > ***Basic Statistics*** > ***1-Sample t***.

Esta pregunta fue contestada correctamente por un total de setenta y siete de los ochenta y dos directivos que respondieron la encuesta (cinco directivos dejaron

sin contestar esta pregunta). El promedio de trabajadores del departamento TIC es de cincuenta y nueve para las empresas que contestaron a la pregunta. Se observa también que el número de trabajadores en dicho departamento es muy variable, oscilando entre un mínimo de seis trabajadores y un máximo de cuatrocientos cincuenta, lo que hace pensar en diferentes niveles de externalización de los servicios y sistemas TIC. Puesto que la muestra es aleatoria, se ha podido obtener un intervalo de confianza para el promedio de trabajadores en departamentos TIC de todas las empresas de la población considerada. En este caso, usando un nivel de confianza del 95% se ha obtenido el intervalo (41,36, 76,38), es decir: con un 95% de confianza se puede afirmar que en promedio estos departamentos tienen entre 41 y 77 empleados. Asimismo, resulta posible agrupar los valores obtenidos para la variable anterior en categorías de empresas según el número de empleados en el departamento TIC, lo que permite obtener tablas y gráficos circulares para representar las frecuencias asociadas a cada tipo de empresa participante en la encuesta (figuras 13 y 14).

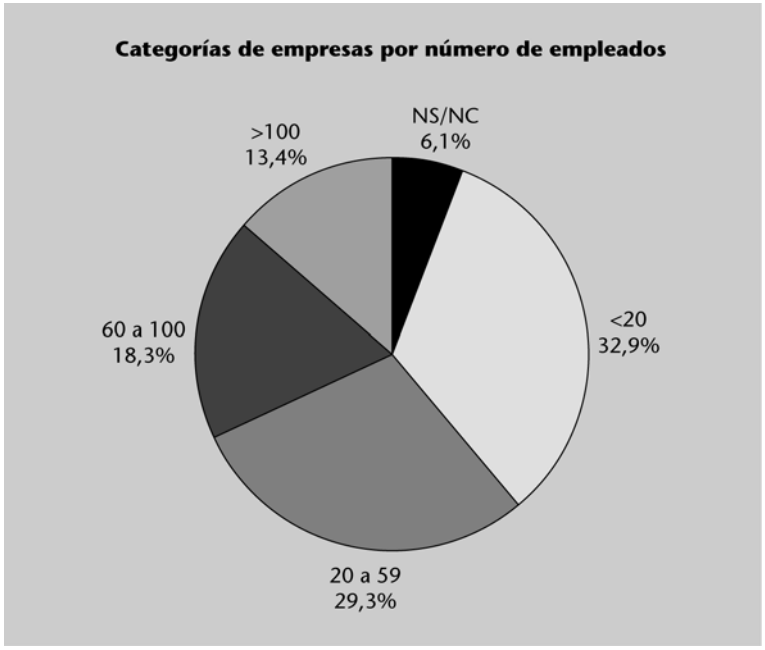
Figura 13. Tabla de frecuencias para cada categoría



Recordatorio Minitab

Para obtener una tabla de frecuencias, usar ***Stat*** > ***Tables*** > ***Tally Individual Variables***.

Figura 14. Gráfico circular que representa los porcentajes de cada categoría



Recordatorio Minitab

Para obtener un diagrama circular, usar ***Graph*** > ***Pie Chart***.

En este caso se aprecia que aproximadamente un tercio (32,9%) de las empresas participantes tienen departamentos TIC relativamente pequeños (menos de 20 empleados), lo que induce a pensar que tendrán bastantes servicios y sistemas de información externalizados.

Otra de las preguntas de la encuesta pedía seleccionar, de entre una lista de factores, aquellos (uno o más) que se tenían en cuenta a la hora de valorar el nivel de éxito de un proyecto TIC finalizado. Puesto que se trata de una pregunta con respuesta múltiple (se pueden seleccionar varios factores a la vez), en este caso se puede emplear un diagrama de barras, como se muestra en la figura 15, para representar el porcentaje de citas de cada factor y caracterizar así aquellos factores más frecuentemente citados.

Figura 15. Gráfico de barras con frecuencia de citas de factores de éxito

Recordatorio Minitab

Para obtener un diagrama de barras, usar ***Graph*** > ***Bar Chart***. Notar que es posible personalizar los gráficos (p. ej., mostrando porcentajes, haciendo que las barras sean horizontales) mediante los botones ***Chart Options***, ***Scale***, etc. (la ayuda contextual de Minitab incluye explicaciones detalladas de todas estas opciones).

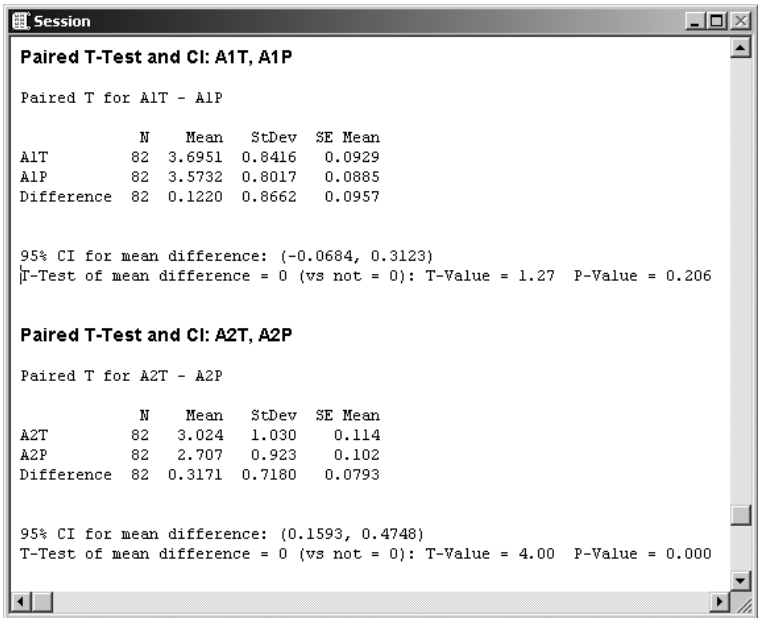
En este caso, queda claro que a la hora de valorar el éxito de un proyecto hay tres factores que se usan casi siempre ("resultados y funcionalidad", "coste final acorde a presupuesto" y "cumplimiento de los plazos previstos"). Cabe observar que el factor "otros" ha sido seleccionado en un 14,3% de las respuestas, lo que indica que tal vez exista un factor no considerado entre los anteriores que también tenga su importancia relativa.

3.2. Ejemplo de uso de contrastes de hipótesis para comparar dos grupos

En otra de las preguntas del cuestionario se le proponían al encuestado una lista de cinco ítems o motivos por los cuales una empresa podía optar por la externalización de sus servicios y sistemas TIC (p. ej.: "superar las limitaciones de las calificaciones profesionales y técnicas del equipo interno", "promover cambios organizativos, estructurales o culturales internos", "conseguir mejores niveles de calidad del servicio o sistema final", "reducir los costes totales", etc.). A continuación se le pedía valorar, usando una escala linealnumérica, la importancia de cada uno de dichos ítems o motivos de externalización, tanto desde un punto de vista teórico como desde un punto de vista práctico (es decir, el encuestado debía emitir dos evaluaciones para cada ítem: por un lado la

correspondiente a la importancia teórica o hipotética del motivo de externalización y, por otro, la correspondiente a la importancia real manifestada en la práctica cotidiana). La escala linealnumérica oscilaba entre 1 (muy poco importante) y 5 (muy importante). Uno de los objetivos de esta pregunta era determinar si para cada uno de los ítems existían diferencias significativas entre su importancia hipotética o teórica y su importancia real en la práctica del día a día (tales diferencias pondrían de manifiesto la existencia de otros factores asociados con la práctica diaria que alteraban significativamente el nivel de importancia teórico de cada motivo). En este caso se optó por realizar un contraste de hipótesis para comparar las dos medias que se obtenían para cada uno de los ítems (es decir, para cada motivo se realiza un contraste de hipótesis sobre la igualdad de la puntuación media teórica y la puntuación media práctica). La figura 16 muestra el *output* de Minitab para los dos primeros tests correspondientes a los dos primeros motivos de la lista (ítems A1 y A2). Se observa que en el caso del primer motivo de externalización considerado, no parece haber diferencias significativas, para un nivel de significación $\alpha = 0{,}05$, entre las medias respectivas de las puntuaciones teóricas (A1T) y las prácticas (A1P). Por el contrario, en el caso del segundo motivo, el *p*-valor obtenido es muy bajo (*p*-valor = 0,000), lo que evidencia la existencia de diferencias significativas entre la importancia hipotética del motivo y su importancia en la práctica.

Figura 16. Test de hipótesis para comparar medias de motivos



```
Session

Paired T-Test and CI: A1T, A1P

Paired T for A1T - A1P

             N    Mean   StDev  SE Mean
A1T         82  3.6951  0.8416   0.0929
A1P         82  3.5732  0.8017   0.0885
Difference  82  0.1220  0.8662   0.0957


95% CI for mean difference: (-0.0684, 0.3123)
T-Test of mean difference = 0 (vs not = 0): T-Value = 1.27  P-Value = 0.206


Paired T-Test and CI: A2T, A2P

Paired T for A2T - A2P

             N    Mean   StDev  SE Mean
A2T         82   3.024   1.030    0.114
A2P         82   2.707   0.923    0.102
Difference  82  0.3171  0.7180   0.0793


95% CI for mean difference: (0.1593, 0.4748)
T-Test of mean difference = 0 (vs not = 0): T-Value = 4.00  P-Value = 0.000
```

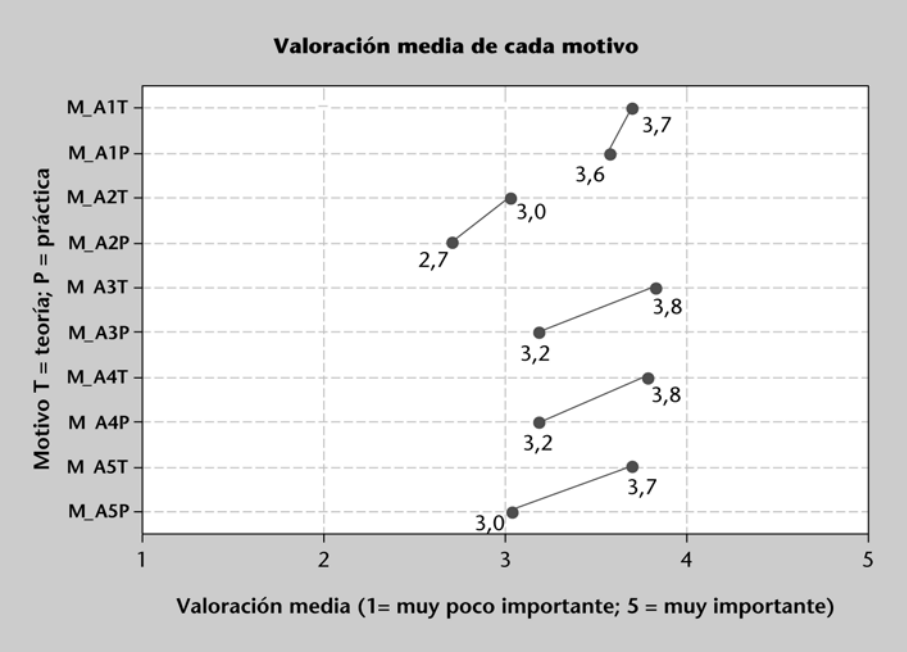
Recordatorio Minitab

Para realizar un contraste de hipótesis para dos muestras dependientes, usar ***Stat*** > ***Basic Statistics*** > ***Paired t***.

La figura 17 muestra el valor de importancia medio obtenido para cada uno de los cinco motivos de externalización considerados, tanto desde un punto de vista teórico como desde un punto de vista práctico. Se observa que, para todos los pares teoría-práctica, el valor teórico siempre es superior al valor práctico. Esto hace sospechar que si bien algunos motivos de externalización deberían ser considerados como muy importantes, en la práctica ello no siempre es posible debido a la influencia de otros factores (condi-

ciones laborales, recursos disponibles, etc.). Precisamente los contrastes de hipótesis permiten detectar aquellos casos en los que las diferencias entre teoría y práctica son significativas. Se observa también en esta figura cuál es la importancia relativa de cada motivo a la hora de decidir sobre externalizar o no los servicios y sistemas TIC.

Figura 17. Comparación visual de la importancia relativa de los ítems



Recordatorio Minitab

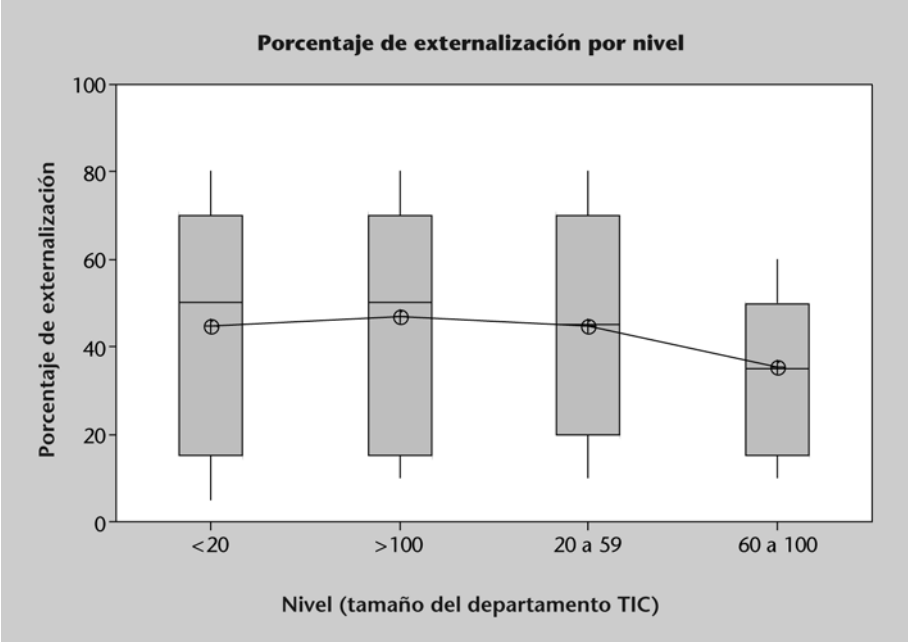
La figura muestra una nube de puntos obtenida con ***Graph*** > ***Scatterplot***. Las líneas de unión entre puntos se han generado con las opciones de dibujo de Minitab con el fin de visualizar mejor las diferencias.

3.3. Ejemplo de uso de ANOVA para comparar más de dos grupos

A fin de disponer de información sobre el porcentaje de servicios y sistemas TIC que las empresas externalizaban, en una de las preguntas se le pidió al encuestado estimar ese valor porcentual. En particular, se pretendía analizar si este porcentaje era el mismo para todas las empresas con independencia del tamaño de su departamento TIC o si, por el contrario, este porcentaje dependía de forma significativa del número de trabajadores en nómina que tuviera dicho departamento.

Puesto que se habían predefinido cuatro categorías o niveles distintos de empresas según la dimensión del departamento TIC (véase la figura 14), resulta necesario aplicar un test ANOVA para dar respuesta a la duda formulada. La figura 18 muestra una comparativa de los distintos diagramas de cajas y bigotes (*boxplots*) por categoría o nivel. Visualmente no se observan grandes diferencias entre los diferentes grupos, salvo quizá una cierta diferencia con el grupo de empresas con departamentos entre sesenta y cien empleados, cuyos porcentajes de externalización parecen algo inferiores al resto (incluso a las de mayor tamaño). En todo caso, estas posibles diferencias visuales no parecen demasiado claras.

Figura 18. *Boxplots* de porcentaje de externalización por nivel

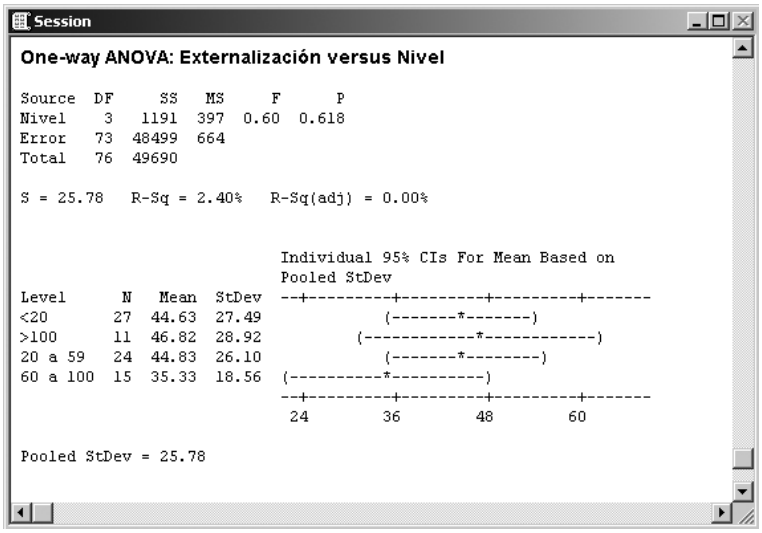


Recordatorio Minitab

Para obtener un *boxplot* múltiple, se ha de usar la opción ***Graph*** > ***Boxplot***. Las líneas de unión entre los distintos *boxplots* se generan mediante las opciones del botón ***Data View***.

La figura 19 muestra el *output* ANOVA, que ayuda a despejar las dudas: un *p*-valor de 0,618 indica que no se han hallado indicios suficientes como para rechazar la hipótesis nula de que el porcentaje medio de externalización es el mismo para todos los grupos, es decir, no parece que el tamaño del departamento TIC tenga una influencia decisiva en el porcentaje de servicios y sistemas TIC que acaban externalizándose.

Figura 19. Contraste ANOVA para comparar las medias de porcentajes

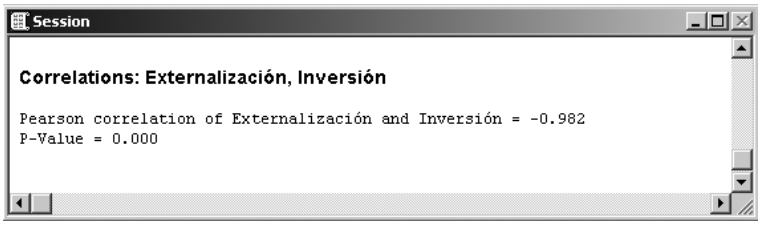


3.4. Ejemplo de uso de correlación y regresión lineal

En una de las últimas preguntas del cuestionario se pedía a los encuestados estimar las cantidades (en euros) que tenían previsto invertir durante el próximo año en adquisición de programas y nuevos sistemas informáticos. Parece lógico pen-

sar que estas cantidades pueden estar inversamente relacionadas con los porcentajes de externalización de cada empresa, esto es, cabría esperar que a mayor porcentaje de externalización de servicios y sistemas TIC, menor inversión prevista en adquisición de programas y nuevos sistemas informáticos. Para tratar de corroborar esta impresión y detectar una posible correlación lineal entre ambas variables se calculó el coeficiente de correlación lineal entre ambas. La figura 20 muestra que, en efecto, existe una fuerte correlación lineal negativa entre ambas variables, ya que el coeficiente de correlación es de –0,982.

Figura 20. Coeficiente de correlación lineal entre externalización e inversión

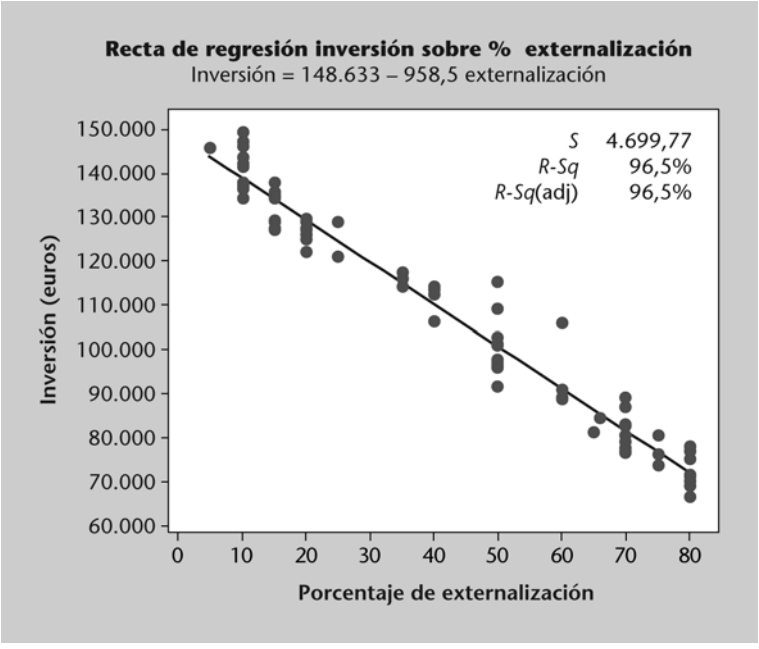


Recordatorio Minitab

Para calcular el coeficiente de correlación, usar la opción ***Stat*** > ***Basic Statistics*** > ***Correlation***.

Tiene sentido, pues, representar la recta de regresión de la inversión sobre el nivel de externalización. Esta recta se muestra en la figura 21. Puesto que el coeficiente de determinación asociado es muy alto (*R–sq* = 96,5%), se puede incluso usar al ecuación de dicha recta para hacer estimaciones sobre la inversión futura de las empresas en nuevos equipos informáticos a partir de su nivel de externalización de servicios y sistemas TIC.

Figura 21. Recta de regresión de la inversión sobre el nivel de externalización



Recordatorio Minitab

Para representar la recta de regresión, usar la opción ***Stat*** > ***Regression*** > ***Regression*** (alternativamente, también se puede usar ***Stat*** > ***Regression*** > ***Fitted Line Plot***).

Resumen

Las técnicas de investigación social basadas en el uso de encuestas y cuestionarios están cada vez más extendidas en todos los ámbitos. Sin embargo, diseñar un buen cuestionario no es una tarea fácil y conviene tener presentes aspectos clave como la brevedad y claridad de las preguntas, el tipo de escala usada o el análisis posterior que se pretende aplicar a los datos de la muestra.

En el diseño del cuestionario y del muestreo hay que tratar de minimizar tanto el error muestral como el error no muestral o sesgo. Para ello resulta necesario conocer bien las diferentes técnicas básicas de muestreo que se usan en cada caso (muestreo aleatorio simple, muestreo sistemático, muestreo estratificado y muestreo por conglomerados).

Finalmente, una vez obtenidos los datos de la encuesta, conviene saber qué técnicas estadísticas se pueden aplicar en cada caso y qué tipo de información pueden proporcionar, tanto de forma numérica como gráfica. Precisamente, el análisis de los resultados obtenidos mediante el uso de estas técnicas comporta a menudo un proceso de reflexión importante, es decir, el programa estadístico siempre será capaz de calcular números y generar resultados, pero no siempre estos resultados tendrán sentido ni serán válidos. Es tarea del investigador comprobar si se satisfacen las hipótesis necesarias para aplicar cada técnica estadística, e interpretar y validar, si procede, los resultados generados por los ordenadores.

Ejercicios de autoevaluación

1) Seleccionar un tema y diseñar un cuestionario para obtener información sobre el mismo. El cuestionario debe contener una pregunta por cada tipo de escala (nominal, ordinal, de intervalos equidistantes y de ratio). Argumentar la validez del cuestionario y especificar qué tipo de técnicas estadísticas se pueden hacer servir para analizar cada pregunta.

2) Entre los investigadores de una universidad se ha realizado un estudio para conocer sus hábitos de trabajo. Entre otras cosas, el estudio pretendía obtener información sobre el número medio de artículos que un investigador lee anualmente, así como sobre qué porcentaje de los mismos están en inglés. Dado que la universidad tiene tres grandes ámbitos de investigación (*E-learning*, Computación y Sociedad de la Información), se diseñó un muestreo por estratos en el que se clasificó a cada investigador en el estrato correspondiente a su ámbito de investigación. La tabla siguiente resume los datos de la encuesta:

Ámbito de investigación (estrato)	N_i	n_i	x_i	s_i	p'_i
E-learning	200	20	138	30	0,50
Computación	250	30	103	25	0,78
Sociedad de la Información	100	25	210	50	0,21

Con la ayuda de un modelo de hoja de cálculo (MS Excel u Open Office Calc), se pide:

a) Obtener un intervalo de confianza del 95% para el promedio de artículos leídos anualmente por la población de investigadores de la universidad.

b) Obtener un intervalo de confianza del 95% para el total de artículos leídos anualmente por el conjunto de investigadores de la universidad.

c) Obtener un intervalo de confianza del 95% para el porcentaje de artículos leídos que están en inglés.

3) Las veinticinco bibliotecas universitarias de un país emplean un total de trescientos profesionales en su servicio de obtención de documentos (SOD). A fin de obtener información sobre el número medio de documentos "difíciles de obtener" que se solicitan anualmente, se selecciona una muestra aleatoria de cuatro bibliotecas universitarias y se encuesta a cada uno de los profesionales del SOD respectivo. También se quiere obtener información sobre el número de expertos en Tecnologías de la Información y Comunicación de cada servicio SOD analizado. La tabla inferior muestra la información obtenida:

Biblioteca (conglomerado)	Número de profesionales N_i	Total de documentos "difíciles" y_i	Número de expertos en TIC m_i
SOD-01	7	95	1
SOD-02	18	325	6
SOD-03	15	190	6
SOD-04	10	140	2

Con la ayuda de un modelo de hoja de cálculo (MS Excel o Open Office Calc), se pide:

a) Obtener un intervalo de confianza del 95% para el promedio de documentos "difíciles de obtener" procesados anualmente por un SOD.

b) Obtener un intervalo de confianza del 95% para el total de documentos "difíciles de obtener" que son procesados anualmente por el global de los SOD.

c) Obtener un intervalo de confianza del 95% para el porcentaje de especialistas en TIC que trabajan en los SOD del sistema universitario.

4) En un estudio se entrevistó a ocho individuos elegidos al azar para evaluar el potencial de venta de un producto antes y después de lanzar una fuerte campaña publicitaria por televisión. El

interés por comprar el producto fue determinado por cada individuo, antes y después de la campaña, usando una escala entre 0 y 10, donde los valores más grandes representaban un interés mayor en adquirir el producto. La tabla siguiente muestra los resultados obtenidos:

Individuo	Después	Antes
1	6	5
2	6	4
3	7	7
4	4	3
5	3	5
6	9	8
7	7	5
8	6	6

Contrastad la hipótesis nula de que, en promedio, el interés por adquirir el producto no ha variado tras la campaña. Usad un nivel de confianza del 95%.

5) En un estudio se visitaron cinco ciudades de una provincia para preguntar a los residentes sobre sus hábitos a la hora de hacer la compra. Una de las preguntas versaba sobre el número de días por mes que realizaban la compra fuera de su provincia. Un total de treinta personas participaron en la encuesta y proporcionaron las observaciones que se incluyen en la tabla siguiente:

Ciudad 1	Ciudad 2	Ciudad 3	Ciudad 4	Ciudad 5
1	3	1	2	5
3	3	6	5	3
2	4	2	7	2
1	3	5	4	9
1	9	6	8	8
0	7	3	1	6

Se pide:

a) Determinar si existen o no diferencias significativas entre los hábitos de compra de los residentes en función de su ciudad (usar un nivel de confianza del 99%).

b) Obtener el coeficiente de correlación entre la ciudad de residencia y el número de veces por mes que se compra fuera de la provincia.

Solucionario

1) Pregunta abierta, consultad el primer apartado de este material para comprobar la validez del cuestionario propuesto.

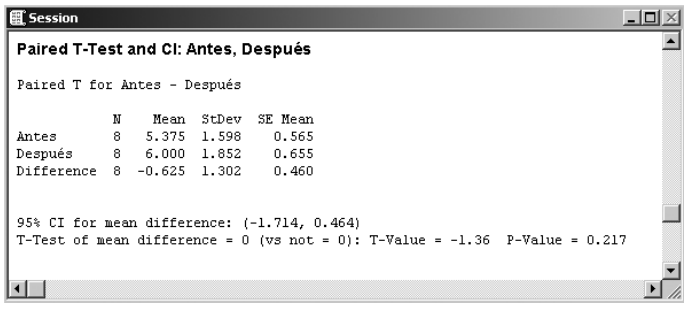
2) La figura siguiente muestra los resultados obtenidos con el modelo Excel. Así, con un nivel de confianza del 95% se puede afirmar que:
a) El número medio de artículos leídos por año e investigador oscila entre 129 y 142.
b) El total de artículos leídos por año oscila entre 70.675 y 78.025.
c) El porcentaje de los artículos leídos que está en inglés oscila entre el 47% y el 68%.

	A	B	C	D	E	F	G	H	I	J
1	**investigación (estrato)**	**N(i)**	**n(i)**	**x-bar(i)**	**s(i)**	**p'(i)**	**N(i) * x-bar(i)**	**N(i) * (N(i) - n(i)) * (s(i)^2 / n(i))**	**N(i) * p'(i)**	**N(i) * (N(i) - n(i)) * [p'(i) * (1 - p'(i)) / (n(i) - 1)]**
2	E-learning	200	20	138	30	0,50	27.600	1.620.000	100	473,68
3	Computación	250	30	103	25	0,78	25.750	1.145.833	195	325,45
4	Sociedad de la Información	100	25	210	50	0,21	21.000	750.000	21	51,84
5	*Totales*	*550*	*75*				*74.350*	*3.515.833*	*316*	*850,98*
6										
7			z =	1,96						
8			x(E) =	135,18		p'(E) =	0,57			
9			s(E) =	3,41		sp(E) =	0,05			
10		x(E) - z*s(E) =		**128,50**	p'(E) - z*sp(E) =		**0,47**			
11		x(E) + z*s(E) =		**141,86**	p'(E) + z*sp(E) =		**0,68**			
12			N * a =	**70.674,89**						
13			N * b =	**78.025,11**						
14										

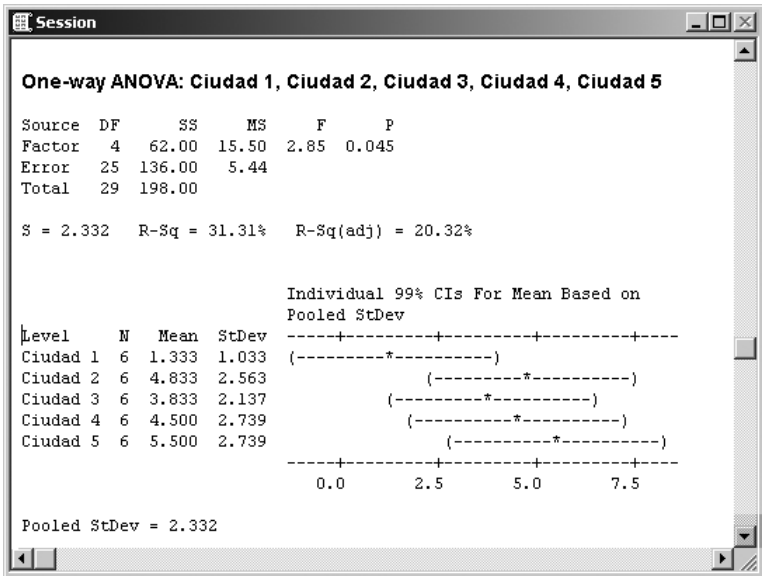
3) La figura siguiente muestra los resultados obtenidos con el modelo Excel. Así, con un nivel de confianza del 95% se puede afirmar que:
a) El número medio de documentos "difíciles" solicitados por año en cada SOD oscila entre 129 y 142.
b) El total de documentos "difíciles" solicitados por año en el conjunto de los SOD oscila entre 3.635 y 5.365.
c) El porcentaje de especialistas en TIC de entre los empleados en el conjunto de los SOD oscila entre 0,21 y 0,39.

	A	B	C	D	E	F
1	**SOD**	**Número de profesionales**	**Total de documentos "dificiles"**	**NNúmero de especialistas TIC**		
2	**(conglomerado)**	**N(i)**	**y(i)**	**m(i)**	**[y(i) - x(C)*N(i)]^2**	**[m(i) - p'(c)*N(i)]^2**
3	SOD-01	7	95	1	100,00	1,21
4	SOD-02	18	325	6	3025,00	0,36
5	SOD-03	15	190	6	1225,00	2,25
6	SOD-04	10	140	2	100,00	1,00
7	*Totales*	*50*	*750*	*15*	*4450,00*	*4,82*
8						
9		z =	1,96			
10		x(C) =	15,00	p'(C) =	0,3	
11		s(C) =	1,47	sp(C) =	0,05	
12		x(C) - z*s(C) =	**12,12**	p'(C) - z*sp(C) =	**0,21**	
13		x(C) + z*s(C) =	**17,88**	p'(C) + z*sp(C) =	**0,39**	
14		N*a =	**3635,18**			
15		N*b =	**5364,82**			

4) En este caso, se requiere usar un contraste de hipótesis para dos poblaciones dependientes (ya que son los mismos individuos los que contestan al test en dos momentos distintos). El *output* de Minitab muestra un p-valor = 0,217 > 0,05 = α, es decir, no se puede rechazar la hipótesis nula de que ambas medias son iguales. En otras palabras, no se han encontrado evidencias suficientes como para afirmar, a un nivel de confianza del 95%, que la campaña publicitaria ha tenido efecto en la intención de compra del producto por parte de los consumidores.



5) En el *output* siguiente de Minitab se muestra un estadístico $F = 2{,}85$ que tiene un p-valor asociado $p = 0{,}045 > 0{,}01 = \alpha$ (puesto que en este caso el nivel de confianza era del 99%). Así pues, no hay evidencias suficientes como para rechazar la hipótesis nula de que todas las medias son iguales, es decir, no parece haber diferencias significativas entre los hábitos de compra de los residentes de las distintas ciudades. Se observa que, en efecto, los intervalos de confianza se solapan unos sobre otros.



El *output* siguiente muestra que la correlación entre la variable Ciudad y la variable Días (ambas generadas a partir de los datos iniciales) es de 0,440, valor que no parece corresponder con una correlación fuerte. En efecto, el p-valor de 0,015 hace pensar que, para un nivel de confianza del 99%, ambas variables no están fuertemente correlacionadas. Observad que esta conclusión es bastante coherente con la obtenida anteriormente para el test ANOVA.