Estadística Aplicada

Dr. Alfonso Alba Cadena
fac@fc.uaslp.mx

Facultad de Ciencias
UASLP

Contenido

1. Conceptos básicos de estadística
2. Medidas descriptivas
3. Estimación
4. Simulación
5. Inferencia estadística
6. Inferencia basada en dos muestras
7. Análisis de varianza
8. Regresión lineal simple

Bibliografía sugerida

- **Probabilidad & Estadística.**
  Walpole et al., Pearson Prentice Hall.

- **Probabilidad y Estadística para Ingeniería y Ciencias.**
  Jay L. Devore., Thompson Learning.

Unidad I

Conceptos básicos de estadística

Introducción

- A grandes rasgos, la estadística tiene dos objetivos:

  1. Describir y entender el comportamiento de un conjunto de datos

  2. Asistir en la toma de decisiones cuando existe incertidumbre

Probabilidad versus Estadística

- La probabilidad
  - Estudia modelos que describen eventos aleatorios, que posiblemente aún no ocurren, o podrían nunca ocurrir.
  - Es estrictamente formal y matemática.

- La estadística
  - Estudia modelos que describan, de manera plausible pero aproximada, datos que provienen de eventos conocidos.
  - Utiliza herramientas matemáticas (en particular, la probabilidad), pero también se basa mucho en interpretaciones subjetivas.

Tipos de estadística

- De manera arbitraria, las técnicas que se estudian en estadística suelen dividirse en dos categorías:

  1. **Estadística descriptiva:** Tiene por objetivo describir las características de los datos de manera resumida.

  2. **Inferencia estadística:** Tiene por objetivo extraer conclusiones o tomar decisiones a partir de los datos.

- En la práctica existe un gran traslape y retroalimentación entre ambas categorías.

Estudios de bioestadística

- El primer paso en cualquier estudio estadístico consiste en establecer objetivos. Por ejemplo:
  - Determinar si un factor específico en un estudio tiene un impacto significativo en el resultado.
  - Comparar dos poblaciones con respecto a alguna característica de sus miembros.
  - Validar nuevas metodologías.

Tipos de estudios

- Encuestas
- Estudios clínicos
- Análisis de datos experimentales
- Estudios de campo

Tipos de datos

- **Nominales:** Representan categorías que no son comparables. Puede ser necesario agrupar los datos de acuerdo a los criterios nominales y posteriormente analizar las posibles diferencias entre categorías.

- **Ordinales:** Representan datos numéricos que pueden ordenarse, pero no existe una manera exacta de calcular la distancia entre ellos. Por ejemplo, opciones en encuestas como: Pésimo, Malo, Regular, Bueno, Excelente.

- **Métricos:** Representan cantidades enteras o reales (usualmente físicas) que pueden compararse en términos de alguna distancia.

Agrupación de datos

- **Casos de prueba:** Representan la fracción de los datos que tiene el factor o la característica que se desea estudiar, o sobre la cual se desea tomar decisiones. Por ejemplo: pacientes a los que se les aplica un nuevo tratamiento o medicamento, fumadores en un estudio de los efectos del tabaco, etc.

- **Grupos de control:** Representan aquellos datos que se sabe que no tienen el factor o la característica bajo estudio. Por ejemplo: pacientes a los que se les aplica el tratamiento estandard (en lugar del nuevo), o no fumadores en el estudio de los efectos del tabaco.

Poblaciones y muestras

- Por lo general, uno quisiera describir o extraer conclusiones sobre un conjunto relativamente numeroso de sujetos. A este conjunto se le llama *población* o *universo*. Por ejemplo: de todas las personas que fuman, qué porcentaje llegan a desarrollar cáncer pulmonar? O cuál es la estatura promedio de los mexicanos?

- En la práctica, suele ser muy costoso o incluso imposible evaluar a todos los elementos de la población, por lo que se considera únicamente un subconjunto "suficientemente grande" de ella, llamado *muestra*.

Poblaciones y muestras

- Si uno pudiera entrevistar o evaluar a toda la población, la descripción o conclusiones resultantes estarían libres de error y serín irrefutables.

- Sin embargo, dado que solamente se evalua una muestra, el análisis realizado siempre tendrá un cierto grado de incertidumbre. Una de las tareas de la estadística es cuantificar esta incertidumbre.

Selección y tamaño de la muestra

- Dos de los principales factores que influyen en el grado de incertidumbre son:
  - La manera de seleccionar la muestra
  - La cantidad de elementos o *tamaño* de la muestra

Selección de la muestra

- Comúnmente se utiliza un muestreo uniforme e independiente. Es decir, cualquier elemento de la población tiene la misma probabilidad de ser elegido, y no influye en la elección de otros elementos.

- En algunos casos puede justificarse el uso de otros esquemas de muestreo; por ejemplo, cuando algún segmento de la población podría verse subrepresentado en una muestra uniforme.

- En caso de no poder muestrear uniformemente la población, debe limitarse explícitamente el estudio al segmento del que proviene la muestra y evitar generalizar las conclusiones a la población total.

Tamaño de muestras

- El número de elementos en la muestra por lo general influye de manera directa en el grado de confiabilidad del estudio (o de manera inversa en el grado de incertidumbre).

- En muchos estudios, es posible ir incrementando el tamaño de muestra conforme se van adquiriendo nuevos datos.

- En un estudio prospectivo, donde se planea obtener nuevos datos conforme transcurre el tiempo, se debe considerar la posibilidad de que a algunos sujetos no se les pueda dar seguimiento. En general, es deseable iniciar el estudio con el mayor número de sujetos posible.

Muestreo con y sin reemplazo

- En la mayoría de los casos, no se permite seleccionar a un elemento de la población mas de una vez en la muestra. A esto se le llama *muestreo sin reemplazo*.

- Desde el punto de vista teórico, sin embargo, la mayoría de las fórmulas usadas en estadística suponen un muestreo con reemplazo, donde el mismo elemento podría aparecer dos o más veces en la muestra.

- Si el tamaño de la muestra es relativamente pequeño comparado con el tamaño de la población, entonces se considera válido aplicar las fórmulas estadísticas a muestras sin reemplazo.

Manejo de datos

- En la mayoría de las aplicaciones modernas de la estadística, suele adquirirse una gran cantidad de datos.

- Muchas veces, estos datos se almacenan en forma de tablas: para cada objeto de estudio (sujetos, pacientes, etc) se capturan varios *campos* de datos. Los campos se organizan en columnas y los sujetos forman los renglones de una tabla o matriz.

- En otras ocasiones, la información obtenida de cada objeto de estudio es mucho mas compleja (e.g., señales de EEG, imágenes de resonancia magnética, etc). En estos casos, la información suele organizarse de diversas maneras; aunque existen varios formatos estandarizados (EDF, DICOM, etc).

Formatos de archivo para tablas

- La manera mas sencilla de manejar una tabla de datos en una computadora es utilizar un formato de archivo de texto.

- En este caso, cada renglón de la tabla se traduce a un renglón en el archivo de texto, y los elementos de cada renglón se separan mediante uno o mas *caracteres separadores* (usualmente espacios, tabulaciones, o comas).

- Uno de los formatos mas comunes (y que muchos programas pueden leer y escribir) es el formato CSV (comma separated values), donde los elementos de los renglones se separan únicamente por comas. Sin embargo, los datos pueden ser numéricos o alfanuméricos.

Lectura y escritura de tablas en Octave

- Lectura de matrices numéricas:
  - Datos separados por espacios o tabuladores:
    `m = load("archivo.txt");`
  - Datos separados por comas:
    `m = csvread("archivo.csv");`
- Escritura de matrices numéricas:
  - Datos separados por espacios o tabuladores:
    `save -ascii "archivo.txt" m`
  - Datos separados por comas:
    `csvwrite("archivo.csv", m)`

Lectura de tablas en C/C++

- Lectura de un archivo de texto separado por espacios o tabuladores en C++:

```
#include <fstream.h>
...
ifstream ifs("archivo.txt");
for (i = 0; i < n; i++) {
    ifs >> registro[i].propiedad_1;
    ...
    ifs >> registro[i].propiedad_m;
}
ifs.close();
```

Escritura de tablas en C/C++

- Escritura de una tabla separada por espacios o tabuladores en C++:

```
#include <fstream.h>
...
ofstream ofs("archivo.txt");
for (i = 0; i < n; i++) {
    ofs << registro[i].propiedad_1 << "\t";
    ...
    ofs << registro[i].propiedad_m << endl;
}
ofs.close();
```

- En caso de ser necesario, se puede cambiar el tabulador por un espacio o coma (por ejemplo, para generar un archivo CSV).

Técnicas de conteo

- En un estudio moderno, es común contar con muestras de gran tamaño (cientos, miles, o millones).

- Una de las principales maneras de resumir la información, es contando cuántas veces se observa un dato particular en la muestra (frecuencia).

- Para realizar el conteo, es importante distinguir entre dos tipos de datos:
  - **Discretos:** Nominales, ordinales, y enteros.
  - **Contínuos:** Reales y enteros en un rango es suficientemente grande.

Tablas de frecuencias

- Para elaborar una tabla de frecuencias, es necesario determinar primero el *rango* de los datos; es decir, los valores mínimo y máximo que pueden tomar.

- El rango se divide (en caso de ser necesario) en un número adecuado de intervalos (disjuntos y cuya unión sea el rango completo), llamados *categorías* o *bins*. Usualmente, los bins son equiespaciados, pero en algunas aplicaciones tiene mas sentido usar intervalos con longitudes distintas.

- Para cada bin se encuentra el número de datos que caen dentro del intervalo correspondiente (para datos contínuos), o que son iguales al valor que representa el bin (para datos discretos).

Frecuencias relativas

- En algunos casos puede ser más util presentar las frecuencias como proporciones o porcentajes con respecto al número total de datos.

- Esto es particularmente útil cuando se desean comparar dos o mas muestras de distintos tamaños.

Histogramas

- Un histograma es una representación gráfica de una tabla de frecuencias, donde comúnmente el eje horizontal corresponde a los valores de los datos, y el vertical a las frecuencias. La frecuencia de cada bin se grafica mediante barras, líneas verticales, o como una función contínua.

- Un aspecto importante es el tamaño de los bins (con respecto al número de datos): si éste es demasiado grande, los datos se concentrarán en pocos bins y no será posible apreciar los detalles de la distribución; si es demasiado pequeõ, se apreciarán "agujeros" en las regiones donde hay menor densidad de datos.

Histogramas multi-dimensionales

- Un histograma multi-dimensional es una tabla de frecuencias que cuenta el número de veces que ocurre cada posible combinación de resultados cuando se observa más de una variable (por ejemplo, peso y altura de una persona, o el hecho de ser fumador y que uno de los padres también lo sea).

- El histograma multidimensional permite apreciar y cuantificar correlaciones entre variables.

- Qué pasa cuando hay demasiadas variables? $\Longrightarrow$ La maldición de la dimensionalidad.

Unidad II

Medidas descriptivas

Introducción

- La estadística descriptiva tiene por objetivo presentar de una manera resumida ciertas características (numéricas) de una población.

- Típicamente, estas características se estiman a partir de una muestra, y representan una aproximación de las verdaderas características de la población.

- Para distinguir entre ambas cosas, llamaremos *parámetros* a las características (posiblemente desconocidas) de la población, y *estadísticos* a las que podemos estimar a partir de la muestra.

Medidas descriptivas

- Las principales medidas descriptivas que caracterizan a una población o una muestra se pueden clasificar como sigue:
  - **Medidas de tendencia central o localización**
  - **Medidas de variabilidad o dispersión**
  - **Medidas relacionales**

Medidas de tendencia central

- Estas características describen el valor central de una muestra o población.

- Las medidas mas comunes son: **media**, **mediana**, y **moda**.

Media

- La *media*, también llamada *media aritmética* o *promedio* es la medida central mas utilizada en estadística.

- Se define simplemente como la suma de todas las observaciones en la muestra, dividida entre el número de observaciones.

- Si los datos de una muestra de tamaño n son $x_1, x_2, \ldots, x_n$, la media de la población (posiblemente desconocida) se representa típicamente como μ_x, mientras que la media muestral se representa con $\bar{x}$:

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{1}{n} \sum_{j=1}^{n} x_n.$$

Propiedades de la media

- El promedio de las desviaciones alrededor de la media es cero:

$$\frac{1}{n}\sum_{j=1}^{n}(x_j - \bar{x}) = 0.$$

- El promedio de las desviaciones alrededor de cualquier valor distinto de la media es distinto de cero:

$$\frac{1}{n}\sum_{j=1}^{n}(x_j - z) \neq 0, \quad \text{si } z \neq \bar{x}.$$

- El promedio de los cuadrados de las desviaciones alrededor de la media es menor o igual que el promedio de los cuadrados de las desviaciones alrededor de cualquier otro número:

$$\frac{1}{n}\sum_{j=1}^{n}(x_j - \bar{x})^2 \leq \frac{1}{n}\sum_{j=1}^{n}(x_j - z)^2.$$

- Un incremento o decremento aplicado de forma idéntica a todos los datos produce el mismo cambio en la media:

$$y_j = x_j + r,\ j = 1, \ldots, n \quad \Longrightarrow \quad \bar{y} = \bar{x} + r.$$

- Un cambio en la escala de los datos produce el mismo cambio de escala en la media:

$$y_j = sx_j,\ j = 1, \ldots, n,\ s \neq 0 \quad \Longrightarrow \quad \bar{y} = s\bar{x}.$$

Mediana

- Para una muestra de tamaño n, dada por los datos $x_1, x_2, \ldots, x_n$, la mediana se define como aquél valor m tal que exactamente $n/2$ datos son menores a m y otros $n/2$ datos son mayores a m. Si n es impar, m es uno de los datos de la muestra.

- Para encontrar la mediana, se ordenan primero los datos de menor a mayor (o viceversa). Para n impar, la mediana es el dato que queda en la $(n+1)/2$-ésima posición.

- Si n es par, entonces la mediana se calcula como el promedio de los datos que quedan en las posiciones $n/2$ y $n/2+1$.

Propiedades de la mediana

- El promedio de las desviaciones absolutas de los datos con respecto a la mediana es menor que el promedio de las desviaciones absolutas con respecto a cualquier otro número:

$$\frac{1}{n}\sum_{j=1}^{n}|x_j - m| \leq \frac{1}{n}\sum_{j=1}^{n}|x_j - z|.$$

- La mediana suele ser un mejor descriptor del centro de una muestra cuando existen valores muy atípicos (outliers). Los valores atípicos tienden a sesgar la media de manera importante.

Moda

- La moda representa aquél valor que ocurre con mayor frecuencia en una población o una muestra.

- Cuando los datos son discretos, la moda puede encontrarse simplemente haciendo un histograma de los datos y buscando el dato que maximiza la frecuencia.

- Para datos contínuos, la búsqueda de la moda se puede realizar interpolando el histograma y encontrando el máximo de la función de interpolación.

Medidas de variabilidad

- Otro parámetro importante acerca de una población es el grado de variabilidad o dispersión de los datos.

- En general, uno espera que las medidas de tendencia central representen de manera fiel a la población o a la muestra; sin embargo, es posible que existan muchos datos alejados del centro. En estos casos, se esperaría observar una alta variabilidad. Si en cambio los datos se concentran cerca del centro, la variabilidad será menor.

- Las medidas de variabilidad mas utilizadas son la *varianza* y la *desviación estándard* (las cuales están fuertemente relacionadas).

- Otras medidas de variabilidad son: el *rango*, el *rango intercuartil*, y la *entropía*.

Desviaciones

- Consideremos una muestra $x_1, \ldots, x_n$. Uno puede obtener una medida de la variabilidad de los datos considerando la desviación promedio con respecto a alguna medida de localización, por ejemplo, la media $\overline{x}$.

- Sin embargo, si se mide la desviación tomando únicamente las diferencias $x_j - \overline{x}$, sabemos que el promedio de las desviaciones será cero.

- Para evitar cancelaciones, podemos considerar mejor el valor absoluto o el cuadrado de las desviaciones. En muchos casos, se prefiere utilizar el cuadrado, debido a que tiene propiedades analíticas que el valor absoluto no tiene.

Varianza

- En base a lo anterior, uno se vería tentado a proponer como medida de variabilidad el promedio de los cuadrados de las desviaciones con respecto a la media. Esta medida se conoce como *varianza poblacional* (σ^2) o *varianza muestral sesgada* (s^2):

$$s_n^2 = \frac{1}{n}\sum_{j=1}^{n}(x_j - \bar{x})^2.$$

- Cuando se utiliza la fórmula anterior para calcular la varianza de una muestra y considerarla como una aproximación a la varianza poblacional, la estimación resultante tiene un sesgo que ocasiona que la varianza poblacional sea *subestimada*.

- Por la razón anterior, en muchas ocasiones se prefiere utilizar el siguiente estimador no-sesgado:

$$s^2 = \frac{1}{n-1}\sum_{j=1}^{n}(x_j - \bar{x})^2.$$

- Se puede demostrar que $E[s^2] = \sigma^2$ y $E[s_n^2] = \frac{n-1}{n}\sigma^2$.

Desviación estándar

- Dado que la varianza es el promedio de los cuadrados de las desviaciones con respecto a la media, sus unidades son las unidades de los datos al cuadrado. Por ejemplo, si los datos están dados en metros, las unidades de la varianza serán metros cuadrados.

- En muchas ocasiones es útil contar con una medida de variabilidad en las mismas unidades que los datos (y que las unidades de localización).

- Esto puede lograrse simplemente tomando la raíz cuadrada (positiva) de la varianza. A esta medida se le conoce como la *desviación estándar*.

- La desviación estándar poblacional se representa con σ, mientras que la desviación estándar muestral se denota con s.

Propiedades de la varianza y la desviación estándar

- Ni la varianza, ni la desviación estándar se ven afectadas si se agrega un valor constante a cada observación. Es decir:

$$y_j = x_j + r,\ j = 1, \ldots, n \quad \Longrightarrow \quad s_y = s_x.$$

- Si se cambia la escala de los datos por un factor a (distinto de cero), la desviación estándar se verá afectada de la misma manera, mientras que la varianza se verá escalada por un factor a^2:

$$y_j = ax_j,\ j = 1, \ldots, n,\ a \neq 0 \ \Longrightarrow \ s_y = as_x,\ s_y^2 = a^2 s_x^2.$$

Rango

- El *rango* se define como la diferencia entre el mayor y el menor valor observados en una población o unamuestra.

- Aunque el rango puede utilizarse como una medida rudimentaria de variabilidad, es también una medida altamente sensible a valores atípicos.

- Por otra parte, el rango tiende a incrementarse con el tamaño de la muestra. A diferencia de la varianza, no hay forma que el rango disminuya a partir de nuevas observaciones.

Cuartiles

- Considere una muestra de tamaño n dada por $x_1, x_2, \ldots, x_n$, y sin pérdida de generalidad supongamos que los datos están ordenados de forma ascendente; es decir $x_1 \leq x_2 \leq \ldots \leq x_n$.

- Los *cuartiles* q_1, q_2, q_3 se definen como aquellos valores que dividen el número de observaciones en cuatro partes iguales, de manera que un 25% de las observaciones son menores que q_1, un 50% son menores que $q_2 = m$ (la mediana), y un 75% son menores que q_3.

- Una medida de variabilidad menos sensible a valores atípicos que el rango es el *rango inter-cuartil* (IQR por sus siglas en inglés), el cual se define como IQR $= q_3 - q_1$.

Entropía

- Formalmente, la *entropía* (según la definición utilizada en la teoría de la información - también llamada *entropía de Shannon*) es una medida de la incertidumbre de una variable aleatoria.

- Dada una muestra $x_1, \ldots, x_n$, se puede estimar la entropía calculando primero un histograma normalizado p_k, $k = 1, \ldots, q$ de los datos, donde q es el número de bins y p_k representa la proporción de datos que caen en el k-ésimo bin.

- De esta manera, se define la entropía H_x de la muestra como

$$H_x = -\sum_{k=1}^{q} p_k \log_b(p_k),$$

  donde b es la base del logaritmo (típicamente se utilizan las bases 2, e y 10).

- Una propiedad interesante de la entropía es que no se ve afectada por cambios en la escala o en el centro de los datos.

Medidas relacionales

- Las medidas relacionales tienen por objetivo establecer el grado de relación o interacción entre parejas de variables o mediciones de una misma población o muestra. Por ejemplo, pueden usarse para cuantificar la relación entre el nivel de glucosa y el índice de masa corporal en pacientes con diabetes.

- Las principales medidas relacionales son: *covarianza*, *correlación*, e *información mutua*.

- Estas medidas no pueden utilizarse para comparar dos poblaciones o muestras distintas.

Covarianza

- La covarianza mide qué tanto cambian dos variables *juntas* con respecto a sus medias.

- Considere una muestra de tamaño n con dos rasgos, dados por $x_1, x_2, \ldots, x_n$ y $y_1, y_2, \ldots, y_n$. La covarianza $\text{cov}_{x,y}$ entre el rasgo x y el rasgo y se define como

$$\text{cov}_{x,y} = \frac{1}{n} \sum_{j=1}^{n} (x_j - \bar{x})(y_j - \bar{y}).$$

- La definición anterior se refiere a la *covarianza poblacional*. Para estimar la *covarianza muestral* se divide entre $n - 1$ en lugar de n.

- Si la covarianza entre dos variables es igual a cero, decimos que las variables son *independientes*.

Propiedades de la covarianza

- La covarianza entre dos variables es positiva si al incrementarse una de ellas por lo general se incrementa también la otra.

- La covarianza entre dos variables es negativa si al incrementarse una de las variables por lo general la otra disminuye.

- La covarianza de una variable consigo misma es igual a su varianza: $cov_{x,x} = s_x^2$.

- La covarianza es simétrica: $\text{cov}_{x,y} = \text{cov}y, x$.

Correlación

- La magnitud de la covarianza depende de la escala de los datos. Por ejemplo, si los valores de una de las variables se multiplican por 2, entonces la covarianza resultante será también el doble.

- Una medida relacional cuya magnitud no depende de la escala de los datos se puede obtener al dividir la covarianza entre las desviaciones estándar de las variables involucradas.

- A esta medida se le conoce como *coeficiente de correlación* y se define como

$$\text{cor}_{x,y} = \frac{\text{cov}_{x,y}}{s_x s_y}.$$

- El coeficiente de correlación siempre toma valores entre -1 y 1. En particular, $\text{cor}_{x,x} = 1$.

Información Mutua

- Otra manera de medir la interdependencia entre dos variables es mediante la *información mutua*, la cual se define como

$$I_{x,y} = H_x + H_y - H_{x,y},$$

donde H_x y H_y representan la entropía de x y y, respectivamente, y $H_{x,y}$ es la *entropía conjunta* de ambas variables.

- La entropía conjunta se obtiene calculando primero el histograma conjunto $p_{x,y}$; es decir, el histograma bidimensional que se obtiene al dividir los rangos de x y de y en regiones rectangulares, y contando cuántos datos caen en cada región. Posteriormente se calcula

$$H_{x,y} = -\sum_{k=1}^{q}\sum_{l=1}^{r} p_{x,y}\log_b(p_{x,y}),$$

donde q y r representan el número de bins en los histograma de x y de y, respectivamente.

Unidad III

Estimación

Introducción

- Uno de los objetivos de la inferencia estadística es realizar *generalizaciones* acerca de la población en base a los datos de una muestra.

- En la estadística clásica, la inferencia se realiza de dos maneras: mediante *estimación* y mediante *pruebas de hipótesis*.

- La estimación tiene por objetivo encontrar estadísticos (también llamados estimadores) que representen fielmente las características de la población que se desean estudiar.

La distribución normal

- En muchos casos, la distribución de los datos puede aproximarse mediante la distribución normal, la cual tiene la forma siguiente (para el caso univariado):

$$P_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\},$$

  donde μ y σ^2 representan, respectivamente, la media y la varianza de la distribución normal (y no necesariamente de los datos).

- La notación que se utiliza generalmente para representar esta distribución es $\mathcal{N}(\mu, \sigma^2)$.

Teorema del Límite Central

- El Teorema del Límite Central establece que la media de un número suficientemente grande de variables aleatorias independientes e igualmente distribuídas (i.i.d.) tiene una distribución aproximadamente normal.

- Específicamente, si se tienen n variables $x_1, x_2, \ldots, x_n$ provenientes de una distribución con media μ y varianza σ^2, entonces, para un valor suficientemente grande de n, la media

$$\bar{x}_n = (x_1 + x_2 + \ldots + x_n)/n$$

  tiene una distribución aproximadamente normal con media μ y varianza σ^2/n.

- Equivalentemente, uno puede decir que $\sqrt{n}(\bar{x}_n - \mu)/\sigma \sim \mathcal{N}(0, 1)$.

Estimación paramétrica

- Considere una cierta población de la cual se desea estimar alguna característica, la cual corresponde a un parámetro θ de la distribución P_θ que representa a la población.

- Para lograr esto, se obtiene una muestra de tamaño n, a partir de la cual se calcula un estadístico $\hat{\Theta}$.

- El valor de este estadístico depende de la muestra que se elija, por lo tanto $\hat{\Theta}$ puede ser considerado una variable aleatoria (razón por la cual se denota con mayúscula).

- Una cuestión importante es cómo elegir un estadístico que aproxime bien al parámetro que se desea estimar.

Propiedades de un estimador

- Las siguientes propiedades pueden ayudar a determinar si un estimador es mejor que otro:
  - Ser *insesgado*: Un estimador es insesgado si su valor esperado es igual al parámetro que se desea estimar. Es decir, $E[\hat{\Theta}] = \theta$.
  - *Varianza de un estimador*: se define como $var(\hat{\Theta}) = E[(\hat{\Theta} - \theta)^2]$. Lo deseable es tener un estimador con la menor varianza posible.
  - Ser *consistente*: Un estimador es consistente si converge al verdadero valor conforme el tamaño de la muestra crece a infinito.

Estimadores puntuales

- Un *estimador puntual* $\hat{\Theta}$ de un parámetro θ consiste en un solo número (una aproximación de θ), el cual no proporciona información adicional acerca de qué tan buena o mala es la estimación (es decir, qué tan cerca está $\hat{\Theta}$ de θ).

- Por ejemplo, la media muestral $\bar{X}$ es un estimador puntual del promedio de la población μ.

Intervalos de confianza

- En muchos casos, es preferible estimar un intervalo $(\hat{\Theta}_L, \hat{\Theta}_R)$ dentro del cual se puede esperar razonablemente que se encuentre el parámetro que se desea estimar.

- Formalmente, se desea encontrar valores $\hat{\Theta}_L$ y $\hat{\Theta}_R$ tales que

$$P(\hat{\Theta}_L < \theta < \hat{\Theta}_R) = 1 - \alpha,$$

  para $0 < \alpha < 1$.

- Al intervalo $(\hat{\Theta}_L, \hat{\Theta}_R)$ se le llama *intervalo de confianza* de (1-α)100%, y al valor (1-α) se le llama *grado de confianza* o *índice de confianza*.

Intervalos de confianza

- Es fácil calcular un intervalo de confianza si se conoce la distribución $P(\hat{\Theta})$ del estimador. Sin embargo, esta distribución rara vez se conoce por completo: posiblemente solo se conozcan algunas de sus características.

- Por ejemplo, si se desea estimar la media μ de una población, el estimador que se suele utilizar es la media muestral, la cual sabemos que tiene una distribución aproximadamente normal (suponiendo un tamaño de muestra suficientemente grande) centrada en μ, pero con varianza σ^2 desconocida.

Intervalo de confianza para la media

- Consideremos el caso en que se desea estimar la media poblacional μ a partir de la media muestral $\bar{x}$ obtenida de una muestra de tamaño n (suficientemente grande), y supongamos que conocemos la varianza σ^2 de los datos.

- Por lo que hemos visto anteriormente, sabemos que

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1),$$

  por lo que podemos encontrar un valor $z_{\alpha/2}$ tal que

$$P\left(-z_{\alpha/2} < \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} < z_{\alpha/2}\right) = 1 - \alpha,$$

  donde $z_{\alpha/2}$ es el cuantil de la distribución normal $\mathcal{N}(0, 1)$ al que le corresponde un área de $1 - \alpha/2$.

- Manipulando la desigualdad se puede obtener el intervalo deseado:

$$P\left(\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha.$$

Intervalo de confianza para la media

- Dada una muestra $x_1, x_2, \ldots, x_n$ (con n suficientemente grande) donde cada dato proviene de una distribución con media μ desconocida y varianza σ^2 conocida, entonces se cumplen las siguientes afirmaciones:
  - Si se utiliza $\bar{x}$ como estimador de μ, se puede tener una confianza de (1-α)100% de que el error no excederá $z_{\alpha/2}\sigma/\sqrt{n}$.
  - Si se utiliza $\bar{x}$ como estimador de μ, se puede tener una confianza de (1-α)100% de que el error no excederá un umbral e cuando el tamaño de la muestra sea

$$n = \left(\frac{z_{\alpha/2}\sigma}{e}\right)^2.$$

Intervalos de confianza para σ desconocida

- Cuando la varianza de los datos es desconocida y los datos tienen una distribución aproximadamente normal, se puede calcular la varianza insesgada S^2 de la muestra y utilizar el estadístico

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}},$$

el cual tiene una distribución t de Student con $n-1$ grados de libertad.

- Por lo tanto, el intervalo de confianza de $(1-\alpha)100\%$ se obtiene encontrando el cuantil $t_{\alpha/2}$ tal que

$$P\left(-t_{\alpha/2} < \frac{\bar{X} - \mu}{S/\sqrt{n}} < t_{\alpha/2}\right) = 1 - \alpha.$$

Intervalos de confianza para σ desconocida

- Nuevamente, podemos manipular la desigualdad para llegar a:

$$P\left(\bar{X} - t_{\alpha/2}\frac{S}{\sqrt{n}} < \mu < \bar{X} + t_{\alpha/2}\frac{S}{\sqrt{n}}\right) = 1 - \alpha.$$

- De esta forma, si $\bar{x}$ y s son la media y desviación estándar de una muestra aleatoria de una población con varianza σ^2 desconocida, se puede asegurar con un $(1-\alpha)100\%$ de confianza que μ se encuentra en el rango

$$\left(\bar{x} - t_{\alpha/2}\frac{s}{\sqrt{n}},\ \bar{x} - t_{\alpha/2}\frac{s}{\sqrt{n}}\right),$$

donde $t_{\alpha/2}$ representa el $(1-\alpha/2)$-cuantil de una distribución t de Student con $v = n - 1$ grados de libertad.

Intervalos de confianza para muestras grandes

- Dado que el estimador s^2 de la varianza es consistente (es decir, converge a la varianza poblacional conforme el tamaño de la muestra crece), para muestras grandes ($n \geq 30$) se puede suponer que s es lo suficientemente aproximado a σ.

- En este caso, una buena aproximación para el intervalo de confianza se obtiene tomando los extremos

$$\overline{x} \pm z_{\alpha/2}\frac{s}{\sqrt{n}}.$$

Estimación de la diferencia entre dos medias

- Considere el caso donde se desea estimar la diferencia entre las medias de dos poblaciones (e.g., la diferencia promedio de estatura entre hombres y mujeres de una cierta población), donde las medias poblacionales son, respectivamente μ_1 y μ_2, y sus varianzas son σ_1^2 y σ_2^2.

- Suponer que de cada población se cuenta con una muestra, de tamaños n_1 y n_2, respectivamente. Dado que la combinación lineal de variables normales tiene también una distribución normal, entonces la diferencia de las medias muestrales $\bar{X}_1 - \bar{X}_2$ tiene una distribución aproximadamente normal con media $\mu_1 - \mu_2$ y varianza $\sigma_{12} = \sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}$.

Caso para varianzas conocidas

- Por lo tanto, si las varianzas poblacionales σ_1^2 y σ_2^2 son conocidas, entonces el estadístico

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$$

  tiene una distribución aproximadamente normal con media 0 y varianza 1.

- El intervalo de confianza de $(1-\alpha)100\%$ para la diferencia de medias $\mu_1 - \mu_2$ queda entonces dado por $(\bar{x}_1 - \bar{x}_2) \pm e$, donde el error e está dado por

$$e = z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}.$$

Caso para varianzas desconocidas e iguales

- En caso de que las varianzas σ_1^2 y σ_2^2 sean desconocidas, pero iguales ($\sigma_1^2 = \sigma_2^2$), el estadístico que se debe utilizar es

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{S_p\sqrt{1/n_1 + 1/n_2}},$$

donde

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}.$$

- T tiene una distribución t de Student con $n_1 + n_2 - 2$ grados de libertad.

Caso para varianzas desconocidas e iguales

- El intervalo de confianza de $(1-\alpha)100\%$ para la diferencia de medias $\mu_1 - \mu_2$ cuando las varianzas poblacionales son desconocidas e iguales está dado por

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}},$$

donde

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}.$$

Caso para varianzas desconocidas distintas

- Cuando las varianzas poblacionales son desconocidas y no es razonable suponer que sean iguales, el estadístico que suele utilizarse es:

$$T' = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{S_1^2/n_1 + S_2^2/n_2}},$$

el cual tiene aproximadamente una distribución t con v grados de libertad, donde

$$v = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{\left[(s_1^2/n_1)^2/(n_1 - 1)\right] + \left[(s_2^2/n_2)^2/(n_2 - 1)\right]}.$$

- Es importante notar que v depende de variables aleatorias, por lo cual representa una *estimación* de los grados de libertad, además de que será necesario redondear v al entero mas cercano.

Estimación de una proporción

- En muchos casos se tienen datos que representan variables binomiales (por ejemplo, la preferencia de un medicamento con respecto a otro, o el hecho de que una persona sea fumador o no), y el interés radica en estimar la proporción de la población que tiene una cierta característica.

- Esta proporción se calcula simplemente como $\hat{P} = X/n$, donde X representa el número de elementos que tienen la propiedad de interés en una muestra de tamaño n.

- Es importante notar que X se puede obtener como una suma de n variables aleatorias tipo Bernoulli cuyo valor es 1 si el elemento correspondiente tiene la propiedad de interés, y 0 si no la tiene.

Estimación de una proporción

- Ya que X es la suma de variables i.i.d., para n suficientemente grande podemos decir que X está normalmente distribuida con media p y varianza σ^2, donde p es la proporción de la población que se desea estimar y $\sigma^2 = p(1-p)/n$.

- Por lo tanto, si se espera que p no sea muy cercana a cero o a uno, podemos estimar un intervalo de confianza de $(1-\alpha)100\%$ considerando el estadístico

$$Z = \frac{\hat{P} - p}{\sqrt{p(1-p)/n}},$$

y encontrando el cuantil $z_{\alpha/2}$ tal que

$$P\left(-z_{\alpha/2} < Z < z_{\alpha/2}\right) = 1 - \alpha.$$

Intervalo de confianza para una proporción

- Manipulando la ecuación anterior para aislar p en el centro de la desigualdad se llega a

$$P\left(\widehat{P} - z_{\alpha/2}\sqrt{p(1-p)/n} < p < \widehat{P} + z_{\alpha/2}\sqrt{p(1-p)/n}\right) = 1-\alpha,$$

  sin embargo sigue apareciendo p (la cual es desconocida) en los extremos de la desigualdad.

- Una alternativa es que, para n grande, el error que se obtiene al reemplazar p por $\widehat{p}$ es muy pequeño, por lo que podemos escribir

$$P\left(\widehat{P} - z_{\alpha/2}\sqrt{\widehat{p}(1-\widehat{p})/n} < p < \widehat{P} + z_{\alpha/2}\sqrt{\widehat{p}(1-\widehat{p})/n}\right) \approx 1-\alpha.$$

- De manera que el intervalo de confianza para p está dado por

$$\widehat{p} \pm z_{\alpha/2}\sqrt{\widehat{p}(1-\widehat{p})/n}.$$

Selección del tamaño de muestra

- Si se utiliza $\widehat{p}$ como estimación de p, entonces podemos asegurar con un $(1-\alpha)100\%$ de confianza que el error de estimación no excederá de

$$z_{\alpha/2}\sqrt{\widehat{p}(1-\widehat{p})/n}.$$

- Por lo tanto, si se desea tener un error máximo e, el tamaño de la muestra debe ser aproximadamente

$$n \geq \frac{z^2_{\alpha/2}\widehat{p}(1-\widehat{p})}{e^2}.$$

- Note que se requiere conocer $\widehat{p}$ para calcular n, pero $\widehat{p}$ se obtiene a partir de una muestra. Típicamente se utiliza una muestra preliminar con tamaño $n \geq 30$ para estimar $\widehat{p}$, y posteriormente se estima el tamaño de muestra para obtener el error deseado.

Estimación de la diferencia entre dos proporciones

- Consideremos ahora el problema de estimar la diferencia entre dos proporciones, por ejemplo $p_1 - p_2$ donde p_1 es la proporción de fumadores con cáncer pulmonar y p_2 es la proporción de no-fumadores con cáncer pulmonar.

- El estimador que suele utilizarse es

$$Z = \frac{(\hat{P}_1 - \hat{P}_2) - (p_1 - p_2)}{\sqrt{p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2}} \sim \mathcal{N}(0,1).$$

- Nuevamente será necesario reemplazar p_1 y p_2 por sus estimaciones $\widehat{p}_1$ y $\widehat{p}_2$ en la estimación del intervalo de confianza para $(p_1 - p_2)$, el cual queda como

$$(\widehat{p}_1 - \widehat{p}_2) \pm z_{\alpha/2}\sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n_1} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{n_2}}.$$

Estimación de la varianza

- Dada una muestra de tamaño n, la varianza muestral S^2 puede utilizarse como un estimador de la varianza poblacional σ^2.

- Si las población tiene una distribución normal, entonces el estadístico

$$X^2 = \frac{(n-1)S^2}{\sigma^2}$$

tiene una distribución chi cuadrada con $n-1$ grados de libertad.

- Por lo tanto, podemos encontrar los cuantiles $\chi^2_{1-\alpha/2}$ y $\chi^2_{\alpha/2}$ de la distribución chi cuadrada con $(n-1)$ grados de libertad que dejan áreas de $1-\alpha/2$ y $\alpha/2$, respectivamente, a la derecha, de manera que

$$P\left(\chi^2_{1-\alpha/2} < X^2 < \chi^2_{\alpha/2}\right) = 1-\alpha.$$

Intervalo de confianza para la varianza

- Dividiendo la desigualdad entre $(n-1)S^2$ y tomando el recíproco de cada término (cambiando así el sentido de las desigualdades) llegamos a

$$P\left(\frac{(n-1)S^2}{\chi^2_{\alpha/2}} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{1-\alpha/2}}\right) = 1-\alpha.$$

- Por lo tanto, para una población con distribución normal, el intervalo de confianza de $(1-\alpha)100\%$ para la varianza poblacional σ^2 está dado por

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}},$$

donde s^2 es la varianza muestral de una muestra de tamaño n, y $\chi^2_{\alpha/2}$ y $\chi^2_{1-\alpha/2}$ son cuantiles de la distribución chi cuadrada con $n-1$ grados de libertad.

Estimación de la razón de dos varianzas

- Cuando se desean comparar las varianzas de dos poblaciones, se suele considerar la razón de las varianzas en lugar de la diferencia.

- De esta manera, el parámetro de interés es σ_1^2/σ_2^2, y el estimador utilizado es s_1^2/s_2^2.

- Si ambas poblaciones son normales y dadas muestras de tamaños n_1 y n_2, respectivamente, de las poblaciones 1 y 2, entonces el estadístico

$$F = \frac{\sigma_2^2 S_1^2}{\sigma_1^2 S_2^2}$$

tiene una distribución F con $v_1 = n_1 - 1$ y $v_2 = n_2 - 1$ grados de libertad.

Intervalo de confianza para la razón de varianzas

- Con base en lo anterior, podemos encontrar los cuantiles $f_{1-\alpha/2}(v_1, v_2)$ y $f_{\alpha/2}(v_1, v_2)$ tales que

$$P\left(f_{1-\alpha/2}(v_1, v_2) < \frac{\sigma_2^2 S_1^2}{\sigma_1^2 S_2^2} < f_{\alpha/2}(v_1, v_2)\right) = 1 - \alpha.$$

- Aprovechando el hecho de que $f_{1-\alpha/2}(v_1, v_2) = 1/f_{\alpha/2}(v_2, v_1)$, podemos manipular la desigualdad anterior para encontrar el intervalo de confianza de $(1-\alpha)100\%$, el cual está dado por

$$\frac{s_1^2}{s_2^2}\frac{1}{f_{\alpha/2}(v_1, v_2)} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{s_1^2}{s_2^2} f_{\alpha/2}(v_2, v_1).$$

Unidad IV

Pruebas de hipótesis

Introducción

- Esta unidad se enfoca en el problema de formular un procedimiento de decisión en base a los datos que se tienen. Por ejemplo, considere el problema de decidir si un tratamiento nuevo para cierta enfermedad es mas efectivo que el tratamiento común.

- En estos casos, el analista/ingeniero/científico suelen *postular una conjetura o hipótesis* acerca de la o las poblaciones en las que está interesado. Por ejemplo, la conjetura podría ser que los dos tratamientos antes mencionados son igual de efectivos.

Hipótesis estadística

- Formalmente, la conjetura que se postula se escribe en forma de una *hipótesis estadística*, la cual es una aseveración con respecto a una o más poblaciones.

- Por ejemplo, si μ_1 y μ_2 representan los tiempos medios de recuperación con dos tratamientos distintos, uno podría postular la hipótesis $\mu_1 - \mu_2 = 0$.

- Dado que se desconocen μ_1 y μ_2, el problema consiste en verificar la validez de la hipótesis para una muestra dada.

Noción de prueba de hipótesis

- Para verificar una hipótesis estadística uno debe preguntarse si existe una probabilidad razonable de observar una muestra tan extrema como la que se tiene bajo la suposición de que la hipótesis es cierta.

- Retomando el ejemplo anterior, suponga que $\bar{x}_1$ y $\bar{x}_2$ representan las medias muestrales del tiempo de recuperación de los dos tratamientos que se desea comparar. Uno debe entonces hacerse la siguiente pregunta: si suponemos que $\mu_1 = \mu_2$, cuál es la probabilidad de obtener una muestra (de cada población) cuya diferencia de medias muestrales $\bar{X}_1 - \bar{X}_2$ sea al menos tan extremo como $\bar{x}_1 - \bar{x}_2$.

- Solamente cuando dicha probabilidad es muy pequeña se puede *rechazar* la hipótesis; sin embargo, en ningún momento podremos aceptar la hipótesis.

Hipótesis nula e hipótesis alternativa

- La *hipótesis nula* (denotada por H_0) representa aquella conjetura que se desea probar. Esta hipótesis debe considerarse como plausible (aunque no necesariamente cierta) hasta que no se tenga evidencia suficiente para rechazarla. La hipótesis nula se rechaza cuando la probabilidad de observar un resultado igual o más extremo que el que se obtiene de la muestra es muy pequeña.

- Al rechazar H_0 se debe aceptar una *hipótesis alternativa* (H_1), la cual por lo general está asociada con la pregunta que se desea responder.

- La hipótesis nula H_0 siempre se opone a H_1, y en muchos casos una es el complemento lógico de la otra.

Resultado de una prueba de hipótesis

- Por lo tanto, una prueba de hipótesis solo puede dar como resultado uno de los dos siguientes:
  - Se rechaza H_0 a favor de H_1 dada la suficiente evidencia en los datos
  - No se rechaza H_0 (H_0 sigue siendo plausible, pero no necesariamente aceptada)
- Solo se llega a una conclusión firme cuando H_0 es rechazada.

Tipos de errores

- Cuando se realiza una prueba de hipótesis, uno puede incurrir en dos tipos de errores:
  - **Error Tipo I:** Rechazar la hipótesis nula cuando ésta es verdadera.
  - **Error Tipo II:** No rechazar la hipótesis nula cuando en realidad es falsa.

- La probabilidad de cometer un error tipo I se denota por α, y también se le llama *nivel de significancia*. Típicamente el analista establece el nivel de significancia deseado.

- La probabilidad de cometer un error tipo II se denota por β.

- La *potencia* de una prueba es $1 - \beta$; es decir, la probabilidad de rechazar H_0 cuando H_1 es verdadera. La potencia se utiliza para comparar dos pruebas estadísticas con el mismo nivel de significancia.

Estadísticos de prueba

- El *estadístico de prueba* X es la variable aleatoria sobre la cual se basa la hipótesis, y cuyo valor se conoce para la muestra con la que se cuenta.

- En el ejemplo anterior, el estadístico de prueba es $X = \bar{X}_1 - \bar{X}_2$, cuyo valor para la muestra que se tiene es $\bar{x}_1 - \bar{x}_2$.

- Para llevar a cabo la prueba es necesario conocer la distribución del estadístico de prueba bajo la suposición de que la hipótesis nula es cierta. A esta distribución se le llama la *distribución nula* y se denota por $P_{H_0} = P(X \mid H_0)$.

- Retomando el ejemplo anterior, uno rechazaría la hipótesis nula (de que $\mu_1 - \mu_2 = 0$) si

$$P_{H_0}\left(\bar{X}_1 - \bar{X}_2 < \bar{x}_1 - \bar{x}_2\right) < \alpha,$$

  suponiendo que $H_1 : \ \mu_1 - \mu_2 < 0$.

Región crítica

- Por lo general, el analista determina el nivel de significancia (probabilidad de error Tipo I) deseado. Por lo tanto, si se conoce la distribución nula P_{H_0}, es posible calcular un cuantil x_α tal que la probabilidad de observar un valor mas extremo que x_α es justamente α. En nuestro ejemplo, tendrímos que

$$P(X < x_\alpha) = \alpha, \quad \text{donde } X = \bar{X}_1 - \bar{X}_2.$$

- Note que cualquier valor de $\bar{x}_1 - \bar{x}_2$ que sea menor que x_α ocasionará que la hipótesis nula sea rechazada. Por lo tanto, x_α puede verse como un umbral de decisión (también llamado *valor crítico*).

- El conjunto de valores del estadístico de prueba que ocasionan que se rechace H_0 constituyen la *región crítica*.

- En nuestro ejemplo, H_0 será rechazada (en favor de H_1) si $\bar{x}_1 - \bar{x}_2 < x_\alpha$, por lo que la región crítica es $\{x \in \mathbb{R} \ : \ x < x_\alpha\}$.

Pruebas de una y dos colas

- Consideremos la hipótesis nula $H_0: \ \mu_1 - \mu_2 = 0$. Dependiendo de la pregunta que uno desea responder, es posible establecer distintas hipótesis alternativas.

- Cuando la hipótesis alternativa es **unilateral**, por ejemplo $H_1: \mu_1 - \mu_2 < 0$, o bien $H_1: \ \mu_1 - \mu_2 > 0$ se dice que se trata de una *prueba de una sola cola*. En este caso, la región crítica contempla solamente una cola de la distribución nula, por lo que el umbral de decisión es x_α.

- Cuando la hipótesis alternativa es **bilateral**, por ejemplo $H_1: \mu_1 - \mu_2 \neq 0$, entonces se trata de una *prueba de dos colas*, y para definir la región crítica se requerirán dos cuantiles o valores críticos $x_\alpha/2$ y $x_{1-\alpha/2}$ tales que

$$P(X < x_\alpha/2) = \alpha/2, \quad \text{y} \quad P(X < x_{1-\alpha/2}) = \alpha/2.$$

P-valores

- Una alternativa que no requiere el cálculo de un valor crítico (o incluso la especificación del nivel de significancia) consiste en calcular lo que se conoce como el p-valor (o valor p) de la prueba.

- El p-valor se define como la probabilidad de obtener un valor del estadístico de prueba al menos tan extremo que el que se observa a partir de la muestra, bajo la suposición de que la hipótesis nula es cierta.

- Por ejemplo, es para una prueba de una sola cola donde H_0 : $\mu = 0$, el p-valor se calcula como

$$P(X < x), \quad \text{si} \quad H_1 : \mu < 0,$$

  o bien,

$$P(X > x), \quad \text{si} \quad H_1 : \mu > 0.$$

- En una prueba estadística, un p-valor suficientemente pequeño (en particular, menor que α) ocasionará el rechazo de la hipótesis nula.

P-valores vs. valores críticos

- Ventajas de utilizar valores y regiones críticas
  - Si la distribución nula es conocida, puede proponerse un valor de α para el que el valor crítico x_α es fácil de encontrar (e.g., en tablas).
  - Si se tienen que realizar múltiples pruebas con la misma distribución nula y el mismo nivel de significancia, solo es necesario calcular una vez el umbral de decisión / valor crítico. (esto, sin embargo, acarrea otros problemas)

- Ventajas de utilizar p-valores
  - Por lo general es mas simple estimar probabilidades cumulativas que encontrar cuantiles, aún cuando se desconozca la forma de la distribución nula (e.g., histogramas).
  - No se requiere establecer de antemano un nivel de significancia. Puede calcularse primero el p-valor y posteriormente (en caso de ser necesario) proponer un valor para α.

Pruebas de hipótesis sobre la media

- Caso para cuando la varianza poblacional σ^2 es conocida:
  - Hipótesis nula: $H_0: \ \mu = \mu_0$ (para un valor dado de μ_0).
  - Estadístico de prueba: $Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \ \sim \mathcal{N}(0, 1)$.
  - Región crítica:
    - $|z| > z_{\alpha/2}$ cuando $H_1: \ \mu \neq \mu_0$.
    - $z > z_\alpha$ cuando $H_1: \ \mu > \mu_0$.
    - $z < -z_\alpha$ cuando $H_1: \ \mu < \mu_0$.
  - P-Valor:
    - $1 - P(Z < z)$ cuando $z > 0$.
    - $P(Z < z)$ cuando $z < 0$.

    donde P representa la distribución $\mathcal{N}(0, 1)$.

Pruebas de hipótesis sobre la media

- Caso para cuando la varianza poblacional σ^2 es desconocida:
  - Hipótesis nula: $H_0: \ \mu = \mu_0$ (para un valor dado de μ_0).
  - Estadístico de prueba: $T = \frac{\bar{X}-\mu_0}{S/\sqrt{n}} \sim t(n-1)$, $S^2 = \frac{\sum_{j=1}^{n}(X_j-\bar{X})^2}{n-1}$.
  - Región crítica:
    - $|t| > t_{\alpha/2}$ cuando $H_1: \ \mu \neq \mu_0$.
    - $t > t_\alpha$ cuando $H_1: \ \mu > \mu_0$.
    - $t < -t_\alpha$ cuando $H_1: \ \mu < \mu_0$.
  - P-Valor:
    - $1 - P(T < t)$ cuando $t > 0$.
    - $P(T < t)$ cuando $t < 0$.

    donde P representa la distribución t de Student con $n-1$ grados de libertad.

Pruebas de hipótesis sobre una proporción

- Hipótesis nula: $H_0: \ p = p_0$ (para un valor dado de p_0).
- Estadístico de prueba:

$$Z = \frac{\bar{X} - p_0}{\sqrt{p_0(1-p_0)/n}} \ \sim \mathcal{N}(0,1)$$

  para n suficientemente grande.

- Región crítica:
  - $|z| > z_{\alpha/2}$ cuando $H_1: \ p \neq p_0$.
  - $z > z_\alpha$ cuando $H_1: \ p > p_0$.
  - $z < -z_\alpha$ cuando $H_1: \ p < p_0$.
- P-Valor:
  - $1 - P(Z < z)$ cuando $z > 0$.
  - $P(Z < z)$ cuando $z < 0$.

  donde P representa la distribución $\mathcal{N}(0,1)$.

Pruebas de hipótesis sobre la varianza

- Hipótesis nula: $H_0: \ \sigma^2 = \sigma_0^2$ (para un valor dado de σ_0^2).
- Estadístico de prueba:

$$X^2 = \frac{(n-1)S^2}{\sigma_0^2} \ \sim \chi^2(n-1).$$

  para una población normalmente distribuida.

- Región crítica:
  - $\chi^2 < \chi^2_{1-\alpha/2}$ ó $\chi^2 > \chi^2_{\alpha/2}$ cuando $H_1: \ \sigma^2 \neq \sigma_0^2$.
  - $\chi^2 > \chi^2_{\alpha}$ cuando $H_1: \ \sigma^2 > \sigma_0^2$.
  - $\chi^2 < \chi^2_{1-\alpha}$ cuando $H_1: \ \sigma^2 < \sigma_0^2$.
- P-Valor:
  - $1 - P(X^2 < \chi^2)$ cuando $\chi^2 > n-1$.
  - $P(X^2 < \chi^2)$ cuando $\chi^2 < n-1$.

  donde P representa la distribución chi cuadrada con $n-1$ grados de libertad.

Pruebas de hipótesis sobre las medias de dos poblaciones

- Caso para cuando las varianzas poblacionales σ_1^2 y σ_2^2 son conocidas:
  - Hipótesis nula: $H_0: \ \mu_1 = \mu_2$.
  - Estadístico de prueba: $Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim \mathcal{N}(0,1)$.
  - Región crítica:
    - $|z| > z_{\alpha/2}$ cuando $H_1: \ \mu_1 \neq \mu_2$.
    - $z > z_\alpha$ cuando $H_1: \ \mu_1 > \mu_2$.
    - $z < -z_\alpha$ cuando $H_1: \ \mu_1 < \mu_2$.
  - P-Valor:
    - $1 - P(Z < z)$ cuando $z > 0$.
    - $P(Z < z)$ cuando $z < 0$.

    donde P representa la distribución $\mathcal{N}(0,1)$.

Pruebas de hipótesis sobre las medias de dos poblaciones

- Caso para cuando las varianzas poblacionales σ_1^2 y σ_2^2 son iguales y desconocidas:
  - Hipótesis nula: $H_0: \ \mu_1 = \mu_2$.
  - Estadístico de prueba: $T = \frac{\bar{X}_1 - \bar{X}_2}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \ \sim t(n_1 + n_2 - 2)$, donde $S_p^2 = \frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}$.
  - Región crítica:
    - $|t| > t_{\alpha/2}$ cuando $H_1: \ \mu_1 \neq \mu_2$.
    - $t > t_\alpha$ cuando $H_1: \ \mu_1 > \mu_2$.
    - $t < -t_\alpha$ cuando $H_1: \ \mu_1 < \mu_2$.
  - P-Valor:
    - $1 - P(T < t)$ cuando $t > 0$.
    - $P(T < t)$ cuando $t < 0$.

    donde P representa la distribución t de Student con $n_1 + n_2 - 2$ grados de libertad.

Pruebas de hipótesis sobre las varianzas de dos poblaciones

- Hipótesis nula: $H_0: \ \sigma_1 = \sigma_2$.
- Estadístico de prueba: $S = \frac{S_1^2}{S_2^2} \ \sim F(n_1 - 1, n_2 - 1)$, cuando las poblaciones están normalmente distribuidas.
- Región crítica:
  - $f < f_{1-\alpha/2}$ ó $f > f_{\alpha/2}$ cuando $H_1: \ \sigma_1 \neq \sigma_2$.
  - $f > f_{\alpha}$ cuando $H_1: \ \sigma_1 > \sigma_2$.
  - $f < f_{1-\alpha}$ cuando $H_1: \ \sigma_1 < \sigma_2$.
- P-Valor:
  - $1 - P(F < f)$ cuando $f > 1$.
  - $P(F < f)$ cuando $f < 1$.

  donde P representa la distribución F con $v_1 = n_1 - 1$ y $v_2 = n_2 - 1$ grados de libertad.

Unidad V

Inferencia no paramétrica

Motivación

- La inferencia estadística (estimación y pruebas de hipótesis) clásica se fundamenta en la obtención de estadísticos cuya distribución, bajo ciertas condiciones, tiene una forma paramétrica conocida; por ejemplo, una distribución tipo t de Student con un cierto número de grados de libertad.

- Sin embargo, habrá ocasiones en las que la característica que se desea estimar o probar no tiene una distribución conocida.

- Por ejemplo, en el análisis de señales biomédicas es común realizar pruebas para determinar si ciertas características de la señal como la potencia o la entropía, o bien, la correlación entre dos señales distintas, es significativamente mayor o menor que un cierto valor (típicamente conocido como la *línea de base*). Otro ejemplo ocurre cuando se desea hacer inferencia sobre la varianza de una población que no necesariamente está normalmente distribuida.

- En algunos de estos casos, es posible obtener la distribución del estadístico de forma aproximada mediante modelos no paramétricos.

Estimación de densidad no paramétrica

- Considere el siguiente problema: se tiene una muestra $z_1, z_2, \ldots, z_n$ y se desea modelar la función de densidad $p_Z(z)$ de estos datos, de manera que se puedan estimar la probabilidad cumulativas

$$P_Z(z) = p_Z(Z < z^*) = \int_{-\infty}^{z^*} p_Z(z)dz$$

y cuantiles

$$z_\alpha, \quad \text{tal que } P_Z(z_\alpha) = \alpha.$$

Estimación de densidad a través del histograma

- Si los datos $z_1, \ldots, z_n$ corresponden a una muestra uniforme e independiente de la distribución p_Z que se desea estimar, y si el número de datos n es suficientemente grande, entonces se puede estimar la distribución a través de un histograma con N bins, donde N es suficientemente grande.

- Sean $a = \min_j\{z_j\}$, $b = \max_j\{z_j\}$, y $h = (b-a)/N$, entonces

$$p_Z(a + kh < Z < a + (k+1)h) \approx \frac{p_k}{n},$$

  donde p_k corresponde al k-ésimo bin del histograma (indexados a partir de $k = 0$).

- Por lo tanto, una estimación burda de la probabilidad cumulativa $P_Z(z^*)$ se puede obtener sumando las frecuencias desde el primer bin hasta el bin que contiene a z^*. Específicamente, sea $k^* = \lfloor (z^* - a)/h \rfloor$; entonces,

$$P_Z(z^*) \approx \frac{1}{n} \sum_{k=0}^{k^*} p_k.$$

Estimación de cuantiles a través del histograma

- El cuantil z_α puede estimarse como la posición donde inicia el primer bin tal que la suma de todos los bins anteriores representa una proporción igual a α. En otras palabras, $z_\alpha \approx a + k_\alpha h$, donde

$$k_\alpha = \arg\min_k \left\{ \sum_{j=0}^{k} p_j > \alpha n \right\}.$$

- En la práctica, k_α se encuentra sumando las frecuencias de los bins a partir del primer bin. Cada vez que se agrega un bin, se compara la suma hasta el momento con αn, y en caso de ser mayor la suma, el índice del último bin agregado corresponde a k_α.

Estimación de densidad basada en kernels

- En un histograma normalizado, cada dato contribuye de igual manera al área total que le corresponde a cada bin. El área con la que contribuye cada dato es igual a $1/n$ (donde n es el número de datos), de manera que el área total del histograma normalizado es igual a uno.

- Una alternativa para estimar la función de densidad de una muestra consiste en considerar que cada dato defina de cierta forma su propio bin, el cual estará centrado en el valor del dato, y tendrá una anchura definida por h. Al igual que en el caso de los histogramas normalizados, cada dato debe contribuir con un área de $1/n$ al área total de la función de densidad.

Estimación de densidad basada en kernels

- Consideremos una muestra $z_1, \ldots, z_n$. Con base en lo anterior, podemos estimar la función de densidad de probabilidad p_Z como

$$p_Z(z) \approx \frac{1}{n} \sum_{j=1}^{n} K_h(z - z_j) = \frac{1}{nh} \sum_{j=1}^{n} K\left(\frac{z - z_j}{h}\right),$$

  donde K es una función integrable simétrica no-negativa, llamada *kernel* cuya integral es igual a uno. Por otra parte, K_h se conoce como el *kernel escalado* y por lo general se define como $K_h(x) = K(x/h)/h$. A h se le conoce como el *ancho de banda* del kernel.

- Los kernels mas comúnmente utilizados son: uniforme (rectangular), triangular, Gaussiano (normal) y el de Epanechnikov.

- Dado que el kernel es una función integrable, la probabilidad cumulativa $P_Z(z)$ por lo general puede encontrarse de manera analítica.

Selección del ancho de banda

- El ancho de banda h tiene una gran influencia en la forma de la distribución estimada. La elección correcta de h dependerá de los datos. Si se utiliza un valor muy pequeño de h, la distribución estimada presentará muchos artefactos que se observan como variaciones abruptas en la función de densidad. En cambio, un valor demasiado grande de h resultará en estimaciones de distribuciones muy suaves que tienden a la uniformidad.

- Para el caso del kernel Gaussiano, una regla práctica para encontrar un ancho de banda óptimo es la regla de Silverman:

$$h = \left(\frac{4s^5}{3n}\right)^{1/5} \approx 1.06sn^{-1/5},$$

  donde s es la desviación estándar de la muestra.

Histogramas vs Kernels

- La estimación de densidad basada en histogramas tiene como ventaja su facilidad de implementación y su eficiencia computacional (cuando el número de bins es considerablemente menor que el número de datos). Sus desventajas son la resolución limitada (por el número de bins), y que por lo general requiere un gran número de datos para evitar regiones submuestreadas del histograma.

- La estimación de densidad basada en kernels tiene ventajas analíticas y puede producir buenos resultados aún cuando el número de datos es pequeño. Sin embargo, los resultados pueden ser muy sensibles a la elección del ancho de banda, y cuando se tiene un gran número de datos, la estimación tiene un costo computacional considerable.

Unidad VI

Regresión lineal simple

Motivación

- En ocasiones, uno requiere realizar pruebas estadísticas para determinar si dos o más variables tienen alguna relación entre sí. Por ejemplo, determinar si existe una relación cuantificable entre el peso y la presión sanguínea de una cierta población.

- La existencia de tales relaciones permite *predecir* el comportamiento de una de las variables cuando se conocen las demás.

Gráficas de dispersión

- Suponga que estamos interesados en la posible relación entre dos variables X y Y de una cierta población, y que contamos con una muestra (x_1, y_1), (x_2, y_2), $ldots$, (x_n, y_n).

- Un primer paso para determinar si existe una relación entre dos variables, consiste en graficar los datos de la muestra como puntos en un plano cartesiano donde el eje de las abscisas representa la variable X y el de las ordenadas representa a Y. Esto se conoce como una *gráfica de dispersión* o *scatterplot*.

- Si los datos quedan mas o menos a lo largo de una línea recta o curva, entonces existe una relación fuerte entre ambas variables. Si los datos forman mas bien una nube, la relación entre ellos no es clara o es inexistente.

Modelos de regresión

- Para cuantificar la relación entre dos variables o mas variables se propone un *modelo paramétrico* que expresa matemáticamente la forma general de la curva que describen los datos cuando son graficados. La forma exacta dependerá de ciertos *parámetros* los cuales deben ser estimados a partir de una muestra.

- Básicamente, el modelo de regresión describe el comportamiento de una de las variables (la *variable dependiente* o *respuesta* del modelo) con respecto a las demás (las *variables independientes* o *regresores*).

- Los modelos de regresión pueden clasificarse de acuerdo al número de variables independientes. Cuando solo se tiene una variable independiente, se trata de un modelo de regresión *simple*, de lo contrario, se denomina modelo de regresión *múltiple*.

- Los modelos de regresión también pueden clasificarse de acuerdo a la forma de la curva que describen: lineal, polinomial, exponencial, etc.

Regresión lineal simple

- El modelo de regresión lineal simple considera que la respuesta Y depende linealmente de un solo regresor X, y por lo tanto está dado por:

$$Y = \alpha + \beta X + \epsilon,$$

donde α y β son los parámetros del modelo, y ϵ es una variable aleatoria con $E(\epsilon) = 0$ y $\text{var}(\epsilon) = \sigma^2$. Esta variable aleatoria (también conocida como *residuo*) representa las posibles desviaciones de cada dato con respecto al modelo, dadas por la variabilidad inherente a la población.

- Los objetivos del análisis de regresión suelen ser

1. Obtener una estimación de los parámetros α y β.
2. Determinar la variabilidad de los parámetros
3. Con base en lo anterior, determinar si existe una relación estadísticamente significativa entre X y Y.

Estimación de parámetros por mínimos cuadrados

- Si se tiene un modelo que se ajuste bien a los datos, entonces uno puede esperar que los residuos sean relativamente pequeños (comparados, por ejemplo, con un modelo que no se ajusta bien).

- Por lo tanto, una manera de encontrar los parámetros del modelo consiste en minimizar el *error cuadrático total*, el cual puede definirse como la suma de los residuos al cuadrado. Por ejemplo, si modelamos cada dato como

$$y_j = \alpha + \beta x_j + e_j,$$

entonces el error cuadrático total está dado por

$$E = \sum_{j=1}^{n} e_j^2 = \sum_{j=1}^{n} (y_j - \alpha - \beta x_j)^2.$$

Estimación de parámetros por mínimos cuadrados

- Para minimizar el error, se calculan las derivadas parciales del error con respecto a los parámetros:

$$\frac{\partial E}{\partial \alpha} = -2\sum_{j=1}^{n}(y_i - a - bx_j),\ \frac{\partial E}{\partial \beta} = -2\sum_{j=1}^{n}(y_i - a - bx_j)x_j.$$

- Igualando a cero ambas derivadas y resolviendo el sistema de ecuaciones se llega a la siguiente solución:

$$b = \frac{\sum_{j=1}^{n}(x_j - \bar{x})(y_j - \bar{y})}{\sum_{j=1}^{n}(x_j - \bar{x})^2} = \frac{\mathrm{cov}_{x,y}}{\mathrm{var}(x)},$$

$$a = \bar{y} - b\bar{x},$$

para una muestra dada (x_j, y_j), $j = 1, \ldots, n$.

Propiedades de los estimadores de α y β

- Para una muestra de tamaño n, llamemos A y B a los estimadores de α y β. Estos estimadores son variables aleatorias ya que dependen de la elección de la muestra, mientras que a y b representarán la estimación para una muestra específica.

- Bajo la suposición de que el error ϵ tiene media cero y varianza σ^2, y toma valores independientes en cada instancia, entonces la respuesta Y, para un valor dado de la variable independiente x tiene media $E[Y|X = x] = \alpha + \beta x$ y varianza var$(Y|X = x) = \sigma^2$.

Propiedades de los estimadores de α y β

- Se puede demostrar que los estimadores B y A son insesgados; es decir, que $E[B] = \beta$ y $E[A] = \alpha$. Así mismo, se puede verificar que la varianza de los estimadores está dada por

$$\text{var}(B) = \frac{\sigma^2}{\sum_{j=1}^{n}(x_j - \bar{x})^2},$$

$$\text{var}(A) = \frac{\sigma^2 \sum_{j=1}^{n} x_j^2}{n \sum_{j=1}^{n}(x_j - \bar{x})^2}.$$

Estimación de la varianza del error

- Para estimar la varianza de los estimadores A y B se requiere conocer la varianza σ^2 de los errores.

- Un estimador insesgado de σ^2 es

$$s^2 = \frac{E}{n-2} = \frac{S_{yy} - bS_{xy}}{n-2},$$

donde $S_{xy} = \sum_j (x_j - \bar{x})(y_j - \bar{y})$ y $S_{yy} = \sum_j (y_j - \bar{y})^2$.

- La división entre $n-2$ se debe a que para estimar σ^2 a través de s^2 se requiere estimar dos parámetros, que son α y β.

Inferencia sobre α y β

- Podemos realizar inferencia sobre la pendiente del modelo lineal a partir del siguiente estadístico:

$$T = \frac{B - \beta}{S/\sqrt{S_{xx}}} \sim t(n-2),$$

donde S representa un estimador de σ (ver diapositiva anterior) y $S_{xx} = \sum_j (x_j - \bar{x})^2$.

- Mediante este estimador es posible construir intervalos de confianza o realizar pruebas de hipótesis para β.