Preliminares - Estadística

Los siguientes serán conceptos de Matemática 3 utilizados a lo largo del desarrollo teórico del método de *Regresión Lineal*, por lo que es importante que se tomen el tiempo de revisarlos.

1 Variables Aleatorias

En cualquier experimento existen múltiples aspectos o características que pueden ser observadas o medidas, pero en la mayoría de los casos un experimentador se enfoca en uno o algunos aspectos específicos de una muestra, como será en nuestro caso.

Por ejemplo, un investigador prueba una muestra de componentes electrónicos para los cuales podía anotar sólo el número de los que han fallado dentro de las 1000 horas, o, en cambio, podría tomar nota de los tiempos de falla de cada uno de ellos.

En general, cada resultado de un experimento puede ser asociado con un número especificando una regla de asociación o función. Tal regla se llama **variable aleatoria**, *variable* porque diferentes valores numéricos son posibles y *aleatoria* porque el valor observado depende de cuál de los posibles resultados experimentales resulte.

Definición 1.1. *Para un espacio muestral $\mathcal{S}$ de algún experimento, una* ***variable aleatoria*** *(v.a) es una función que asigna a cada elemento de $\mathcal{S}$ un número real. Es decir, si X es una v.a de $\mathcal{S}$, $X: \mathcal{S} \to \mathbb{R}$*

Por ejemplo, dado que se tira una moneda dos veces y sea X la v.a. tal que X: "Número de secas obtenidas luego de los dos tiros". El espacio muestral del experimento será:

$$\mathcal{S} = \{(c,c),(c,s),(s,c),(s,s)\}$$

entonces

$$X(c,c) = 0 \qquad X(c,s) = X(s,c) = 1 \qquad X(s,s) = 2$$

Luego, el rango (o imagen) de X es $R_X = \{0, 1, 2\}$

Las variables aleatorias se clasifican según su rango (o imagen). Sea X una v.a. con rango R_X, si R_X es un conjunto *finito* o *infinito numerable*, entonces se dice que X es una **variable aleatoria discreta**. En cambio, si R_X es un conjunto *infinito no numerable* entonces X es una **variable aleatoria continua**

2 Esperanza o valor esperado

La **esperanza**, **valor esperado** o **valor medio** de una variable aleatoria X se puede definir informalmente como el promedio ponderado de los valores del rango de la variable, donde los "+pesos" de cada valor x_i es la probabilidad $P(X = x_i)$, es decir, la probabilidad de que X tome el valor x_i. El valor esperado de X se anota $E(X)$

Para calcular la esperanza de una variable aleatoria en necesario conocer su distribución de probabilidad, lo que indica cómo está distribuida (asignada) la probabilidad total de 1 entre todos los valores posibles de la variable. Para las variables discretas se define una *función de masa de probabilidad* y para las continuas una *función de densidad de probabilidad*. Sean X una variable aleatoria discreta e Y una variable continua, se define a su valor esperado como:

$$E(X) = \sum_{x_i \in R_X} x_i P(X = x_i) \qquad E(Y) = \int_{-\infty}^{\infty} y_i P(Y = y_i)\, dx$$

Indistintamente de que la variable aleatoria sea discreta o continua, la esperanza resulta ser una *función lineal*, es decir, dados $a_0, a_1 \in \mathbb{R}$ y una v.a X, se cumple:

$$E(a_0 + a_1 X) = a_0 + a_1 E(X)$$

3 Varianza

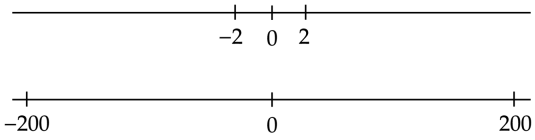
Mientras que la esperanza de una v.a. determina dónde está centrada la distribución de probabilidad, la **varianza** mide cuán dispersos o "alejados" están los valores de la variable con respecto a su esperanza.

Veamos que dadas dos variables aleatorias X e Y discretas, con las siguientes distribuciones de probabilidad

x	-2	2
$p(x)$	0.5	0.5

y	-200	200
$p(y)$	0.5	0.5

se puede verificar fácilmente, por lo visto en la sección anterior, que $E(X) = E(Y) = 0$, pero notemos que los valores que toma la variable Y están mucho mas "alejados" de su valor medio que los valores de X:



El concepto de varianza refleja esta situación, asignando valores más grandes a aquellas variables aleatorias cuyos valores se encuentran más lejos de sus respectivas esperanzas.

Dada una variable aleatoria X cuya esperanza es $E(X) = \mu$ se define a la varianza de X de la siguiente manera:

$$V(X) = E[(X - \mu)^2] = \sigma^2$$

La cantidad $(X-\mu)^2$ es el cuadrado de la desviación de X desde su valor medio, y la varianza resulta la esperanza de dicha desviación al cuadrado. Cuanto más cerca se encuentren los valores de X del valor esperado μ, menor será la varianza. De forma contraria, si hay valores alejados de μ que tengan alta probabilidad, entonces σ^2 será grande.

Notemos que la varianza de una v.a. nunca es negativa

Tanto para variables discretas como continuas valen las siguientes propiedades:

Dados X una variable aleatoria, $\mu = E(X)$ y a y b constantes

$$V(X) = E(X^2) - \mu^2$$

$$V(aX + b) = a^2 V(X) \qquad y \qquad \sigma_{aX+b} = |a|\sigma_X$$

4 Estadísticos y Estimación puntual

El objetivo de la inferencia estadística casi siempre es obtener algún tipo de conclusión sobre uno o más parámetros (características) de las observaciones realizadas.

A menudo, en los problemas de inferencia estadística es poco práctico o imposible analizar la totalidad de las observaciones, es decir, toda la **población**. En ese caso el investigador debe tomar una parte o subconjunto de la población denominada **muestra**

Para que las inferencias sean válidas, la muestra debe ser representativa de la población. Se selecciona una **muestra aleatoria** como el resultado de un mecanismo aleatorio. En consecuencia, la selección de una muestra es un *experimento aleatorio*, y cada observación de la muestra es el valor observado de una *variable aleatoria.* Las observaciones en la población determinan la distribución de probabilidad de la variable aleatoria.

Definición 4.1. *Se dice que las variables aleatorias $X_1, X_2, \ldots, X_n$ forman una* ***muestra aleatoria*** *simple de tamaño n si*

1. *Las X_i son variables aleatorias independientes.*
2. *Cada X_i tiene la misma distribución de probabilidad.*

Esta variación en los valores observados implica, a su vez, que el valor de cualquier función de las observaciones muestrales, tal como la media muestral o la desviación estándar muestral también varía de una muestra a otra. Dicha función o cantidad dependiente de los datos muestrales observados recibe el nombre de **estadístico**.

Definición 4.2. *Un* ***estadístico*** *es cualquier cantidad cuyo valor puede ser calculado a partir de datos muestrales (función). Antes de obtener los datos, existe incertidumbre sobre el valor del estadístico particular. En consecuencia, un estadístico es una variable aleatoria.*

El objetivo de la estimación puntual es seleccionar un solo número (*puntual*), con base en los datos muestrales, que represente un valor sensible o aproximado (estimación) de un parámetro

θ de la población observada.

Definición 4.3. *Una* ***estimación puntual*** *de un parámetro θ es un número único que puede ser considerado como un valor sensible de θ. Se obtiene una estimación puntual seleccionando un estadístico apropiado y calculando su valor con los datos muestrales dados. El estadístico seleccionado se llama* ***estimador puntual*** *de θ y se lo suele anotar $\hat{\theta}$.*

A continuación veremos estadisticos usualmente utilizados como estimadores puntuales ...

Sean $X_1, X_2, \ldots, X_n$ una muestra aleatoria de la v.a X donde

$$E(X) = \mu \qquad \text{y} \qquad V(X) = \sigma^2$$

En primer lugar, si suponemos que no conocemos el valor de μ (*media poblacional*), se utiliza como aproximación de éste al estadístico conocido como **media o promedio muestal**:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Es decir, la media muestral es un *estimador puntual* del parámetro μ. Simbólicamente $\hat{\mu} = \overline{X}$.

Por otro lado si desconocemos el valor de σ^2, un estadístico utilizado como su estimador puantual es la **varianza muestral** ($\hat{\sigma}^2 = S^2$):

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

Finalmente, para estimar el valor de la *desviación estándar poblacional* el estadístico que se utiliza es la **desviación estándar muestral** ($\hat{\sigma} = S$):

$$S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

Es claro que, en el mejor de todos los mundos, sería ideal hallar un estimador $\hat{\theta}$ de un parámetro θ con el cual $\hat{\theta} = \theta$ siempre, es decir, que el estimador siempre de como resultado el valor verdadero del parámtero. Sin embargo, $\hat{\theta}$ es una función de las X_i muestrales, por lo que es una

variable aleatoria. Con algunas muestras, $\hat{\theta}$ dará un valor más grande que θ, mientras que con otras muestras $\hat{\theta}$ subestimará θ.

Si pensamos

$$\hat{\theta} = \theta + \text{error de estimación}$$

resulta evidente que un estimador preciso (e ideal) sería uno que genere errores de estimación pequeños, consiguiendo así que los valores estimados se acerquen al valor verdadero.

Ahora pensemos a un estimador puntual como un instrumento de medición que puede: estar calibrado con precisión o dar sistemáticamente valores menores al valor verdadero que se está midiendo. Si el instrumento está calibrado, las mediciones producidas se distribuirán en torno al valor verdadero de tal modo que en promedio este instrumento mide lo que se propone medir, por lo que se conoce como instrumento *insesgado*. En cambio, si siempre resulta dar valores más pequeños, se dice que el instrumento que tiene un *sesgo* sistemático.

Definición 4.4. *Un estimador puntual $\hat{\theta}$ es un* ***estimador insesgado*** *de θ si $E(\hat{\theta}) = \theta$ con todo valor posible de θ. Si $\hat{\theta}$ no es insesgado, la diferencia $E(\hat{\theta}) - \theta$ se conoce como el* ***sesgo*** *de $\hat{\theta}$.*

Es decir, $\hat{\theta}$ es insesgado si su distribución de probabilidad siempre está "centrada" en el valor verdadero del parámetro.

Notar que los estadísticos mencionados anteriormente resultan ser estimadores puntuales insesgados de la media, varianza y desviación poblacional, respectivamente.

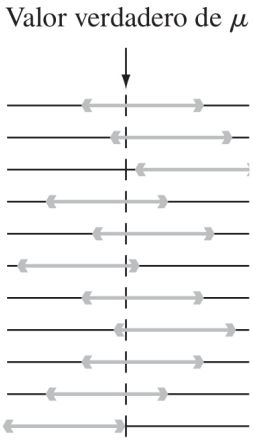
5 Intervalos de confianza

Como vimos en la sección anterior, la estimación puntual de un parámetro da como resultado un solo número. Éste no proporciona información sobre la precisión y confiabilidad de esta estimación. Consideremos, por ejemplo, utilizar el estadístico $\overline{X}$ para calcular una estimación puntual de la media poblacional (μ) en un experimento. Debido a la variabilidad del muestreo, casi nunca es el caso de que el valor calculado $\overline{x}$ cumpla $\overline{x} = \mu$. La estimación puntual no dice nada sobre qué tan *cerca* pudiera estar a μ.

Una variante a dar un solo número puntual es calcular un intervalo completo de los valores

posibles del parámetro estudiado, un **intervalo de confianza** (IC). Para construir el intervalo de confianza primero debe determinarse un **nivel de confianza**, es decir, el grado de confiabilidad de que el verdadero valor del parámetro se encuentre en dicho intervalo ¿Pero qué indica este "nivel de confianza" en verdad?

Pensemos que deseamos hallar un intervalo de confianza del 95% para la media μ de una población normalmente distribuída cuando se conoce el valor de σ. Entonces si tomáramos muestras aleatorias una y otra vez, y calculáramos el intervalo de confianza, a la larga, el 95% de los intervalos resultantes contendrán a μ.



La variación en los intervalos es originada por la participación del estadístico $\overline{X}$ en su cálculo. En particular, los intervalos son de la forma:

$$\left(\overline{X} - 1,96\ \frac{\sigma}{\sqrt{n}} \quad ; \quad \overline{X} + 1.96\ \frac{\sigma}{\sqrt{n}}\right)$$

El intervalo es aleatorio porque los dos puntos extremos del intervalo implican a la variable aleatoria $\overline{X}$. Esto significa que, en cada una de las muestras aleatorias mencionadas, el valor de $\overline{X}$ va cambiando y por ende el de los extremos del intervalo.

Por otro lado el **ancho** del intervalo da información sobre la precisión de la estimación de intervalo. Si el nivel de confianza es alto y el intervalo resultante es angosto, el conocimiento del valor del parámetro es razonablemente preciso. Un muy amplio intervalo de confianza, sin embargo, indica que existe gran cantidad de incertidumbre sobre el valor verdadero de lo que se está estimando.

Entonces, si podemos elegir el nivel de confianza, ¿por qué no elegimos un nivel mayor al 95%, como 99% o 100%? Esto es debido a que cuanto más alto es el grado de confianza, más ancho

es el intervalo resultante. En particular, el único intervalo de 100% para μ es $(-\infty, \infty)$, que no nos aporta mucha información sobre el valor verdadero de este parámetro. Por lo tanto notemos que al elegir un alto nivel de confianza estamos "sacrificando" precisión en la estimación.

6 Tests de Hipótesis

Como hemos visto, el valor de un parámetro puede ser estimado a partir de datos muestrales mediante estimadores puntuales o intervalos de confianza. Con frecuencia, sin embargo, el objetivo de una investigación no es estimar un parámetro sino decidir cuál de dos presunciones contradictorias sobre el parámetro es la correcta. Los métodos para conseguir esto forman parte de la inferencia estadística llamada *prueba de hipótesis*.

Una **hipótesis estadística** es una aseveración sobre el valor de uno o varios parámetros de una población. Un ejemplo de una hipótesis es la pretensión de que $\mu = 1000$ es el número de horas promedio de funcionamiento de las componentes electrónicas fabricadas en una empresa.

En cualquier problema de prueba de hipótesis siempre existen dos hipótesis contradictorias que se consideran. Una podría ser la presunción $\mu = 1000$ y la otra $\mu \neq 1000$. O bien, si se toma como aseveración que la proporción de componentes electrónicas defectuosas es $p < 0.1$, la hipótesis contraria es que $p \geq 0.1$. El objetivo de estas pruebas es determinar, en base a los datos muestrales, qué afirmación es la correcta.

Así como dice la frase conocida "serás inocente hasta que se demuestre lo contrario", en el caso de las pruebas de hipótesis, siempre una de las ellas será inicialmente favorecida, es decir, se tomará como cierta. Luego, esta pretensión inicial no será rechazada a menos que se presente evidencia muestral suficiente que la contradiga y apoye fuertemente la hipótesis alternativa.

Definición 6.1. *La **hipótesis nula** denotada por H_0, es la pretensión que inicialmente se supone cierta. La **hipótesis alternativa** denotada por H_a, es la aseveración contradictoria a H_0.*

La hipótesis nula será rechazada en favor de la hipótesis alternativa sólo si la evidencia muestral sugiere que H_0 es falsa. Si la muestra no contradice fuertemente a H_0, se continuará creyendo en la verdad de la hipótesis nula. Las dos posibles conclusiones derivadas de un análisis de prueba de hipótesis son entonces *rechazar* H_0 o *no rechazar* H_0.

Una **prueba de hipótesis** es el método de utilizar datos muestrales para decidir si la hipótesis nula debe ser rechazada. Por consiguiente se podría probar $H_0 : \mu = 1000$ contra $H_a : \mu \neq 1000$. Sólo si los datos muestrales sugieren fuertemente que μ es diferente de 1000 deberá ser rechazada la hipótesis nula. En el tratamiento de la prueba de hipótesis, H_0 siempre será formulada como una afirmación de igualdad. Si θ denota el parámetro de interés, la hipótesis nula tendrá la forma $H_0 : \theta = \theta_0$ donde θ_0 es un número específico llamado **valor nulo** del parámetro.

Una vez recordados estos conceptos, pasemos a especificar cómo es el proceso de prueba de una hipótesis. Como mencionamos, deseamos determinar si existe información en los datos muestrales que funcionen como contradicción a nuestra hipótesis nula. De esta forma, el procedimiento de prueba consta de:

1. Un **estadístico de prueba**, una función de los datos muestrales en la cual se basará la decisión.

2. Una **región de rechazo**, el conjunto de todos los valores del estadístico de prueba para los cuales H_0 será rechazada.

Por lo tanto, la hipótesis nula será rechazada si y sólo si el valor estadístico de prueba calculado queda en la región de rechazo. La base de elección del rango de valores de rechazo radica en la consideración de los errores que podrían cometerse al sacar una conclusión. En una prueba de hipótesis siempre existen dos tipos de errores:

> Un **error de tipo I** consiste en rechazar la hipótesis nula cuando ésta es verdadera.
> Un **error de tipo II** consiste en no rechazar la hipótesis nula cuando ésta es falsa.

La dificultad con la utilización de un procedimiento basado en datos muestrales es que, debido a la *variabilidad* del muestreo, el resultado podría ser una muestra no representativa, resultando en errores a la hora de tomar una decisión.

Como no podemos considerar un procedimiento que no tengan errores (son aquellos en los que se prueba con toda la población), habrá que buscar procedimientos con los cuales sea *improbable* que ocurra cualquier tipo de error. Es decir, un buen procedimiento es uno con el cual la probabilidad de cometer cualquier tipo de error es pequeña. La selección de un valor de corte

en una región de rechazo particular fija las probabilidades de errores de tipo I y tipo II. Estas probabilidades de error son usualmente denotadas por α y β respectivamente.

Con respecto a estas dos probabilidades de error debemos aclarar lo siguiente: Suponiendo que un tenemos un experimento y un tamaño de muestra fijos, dado un estadístico de prueba, entonces si se reduce el tamaño de la región de rechazo para obtener un valor más pequeño de α se obtiene un valor más grande de β. De esta forma, no existe una región de rechazo que haga que al mismo tiempo α y β sean pequeños. Se debe seleccionar una región para establecer un "compromiso" entre α y β. El método más utilizado para esto es especificar el valor más grande de α que pueda ser tolerado y encontrar una región de rechazo que tenga valor de α. Esto hace a β tan pequeño como sea posible dependiendo del límite en α.

Al valor resultante de α se lo conoce como **nivel de significación** de la prueba y el procedimiento de prueba correspondiente se llama **prueba de nivel** α. Usualmente el nivel de significación será pequeño, del orden de 0,1, 0,01 o 0,05, como veremos en las pruebas de hipótesis sobre parámteros de regresión lineal.

Finalmente, cabe notar que la probabilidad de estos dos errores α y β esta atada a la distribución de probabilidad del estadístico que se elija para la prueba.

7 Aclaraciones

Este apunte preliminar de conceptos de estadística se desarrolló en su mayoría sin ejemplos ya que, por un lado, fueron vistos previamente en Matemática 3, y por otro, el uso que les daremos a los conceptos será con un enfoque hacia el método de regresión lineal.

A pesar de esto, no dejamos de destacar la importancia de la probabilidad y la estadística para el desarrollo de modelos matemáticos e informáticos, así como para el análisis de datos.

Aquellas personas que deseen profundizar, tanto en ejemplos como en conceptos, pueden dirigirse a los apuntes provistos por la profesora María Beatriz Pintarelli en el sitio del departamento de matemática de la Facultad de Ciencias Exactas: Apuntes Matemática 3