Sesion 3: Inferencia Estadística

Carlos Ignacio Patiño (cpatinof@gmail.com)

Julio 31, 2015

Agenda

- Propuestas proyecto final (presentaciones grupales)
- Caso de Estudio: Inferencia Estadística
- Break
- Probabilidad y Variables Aleatorias
- Inferencia Estadística
- Tutorial R

Proyectos finales (Propuestas)

- 15 minutos por grupo
- Título
- Motivación
- Preguntas a resolver
- Hipótesis central
- Datos a emplear
- Habilidades requeridas por el grupo

Caso de Estudio

(En grupos) Seleccionar uno de los artículos escogidos por los miembros del grupo y con base en su lectura desarrollar la siguiente guía.

- Defina claramente la hipótesis planteada por el (los) autor(es)
- Identifique los supuestos planteados en el artículo
- Identifique y explique el método empleado en el artículo para validar la hipótesis
- ¿Qué información emplea el artículo?
- ¿Cuál es la conclusión?
- ¿Está de acuerdo con la conclusión y la forma en la que se llegó a ésta?

Break

45 minutos

Probabilidad y Variables Aleatorias

¿Cuál es la función de la probabilidad en un curso de métodos cuantitativos?

- Probabilidad como medición de la incertidumbre
- Reglas básicas para encontrar probabilidades
- Probabilidad como medida de confiabilidad de una inferencia

Definiciones

- **Experimiento:** Proceso de observación que conlleva a un resultado que no puede ser predicho con plena certeza (lanzamiento de un dado, moneda)
- **Punto muestral:** El resultado más básico de un experimento
- Ejemplo: Se lanzan dos monedas y se registra la cara en la que caen. Enumere todos los *puntos muestrales* de este *experimento*
- **Espacio muestral:** Conjunto de todos los puntos muestrales

Probabilidad

La probabilidad de un punto muestral es un número entre 0 y 1 que mide la verosimilitud con que el resultado va a ocurrir en el momento de realizar el experimento.

- $0 \leq p_i \leq 1$
- $\sum(p_i) = 1$

¿Cómo asignar probabilidades a cada punto muestral?

- Ejemplo de una moneda
- Ejemplo de un dado
- Ejemplo de un accidente de tránsito en Hato Corozal, Casanare (2013)

Una definición más

- **Evento** es un conjunto específico de puntos muestrales. Por ejemplo, observar un número par al lanzar un dado es un evento compuesto por tres posibles puntos muestrales (2, 4, y 6)

Práctica en clase

Experimento: Se lanzan dos monedas desbalanceadas (i.e. resultado no es equiprobable). La probabilidad asociada a cada punto muestral se reporta en la siguiente tabla.

Punto	Probabilidad
CC	$^4/_9$
CS	$^2/_9$
SC	$^2/_9$
SS	$^1/_9$

(Verifique que las propiedades para asignar probabilidades a puntos muestrales se cumplen)

Considere dos eventos: a) Observar exactamente una cara y b) observar al menos una cara. Calcule la probabilidad de a y b.

Pasos para calcular probabilidades a eventos

1. Definir experimento
2. Listar puntos muestrales
3. Asignar probabilidades a esos puntos
4. Determinar el conjunto de puntos que contiene el evento de interés
5. **SUMAR** las probabilidades de cada punto para obtener la probabilidad del **evento**

Unión e Intersección

La **unión** o **intersección** de dos o más eventos genera **eventos compuestos**

- **Unión:** $A \cup B$ consiste en todos los puntos muestrales que pertenecen a A, B o a ambos eventos
- **Intersección:** $A \cap B$ consiste en todos los puntos muestrales que pertenecen a A y a B

Recuerde que la probabilidad de un evento es igual a la suma de las probabilidades de los puntos muestrales que lo componen.

ENUT: Estudiantes que reportan estudiar en casa fuera de la jornada escolar

	Fin de semana	Semana
Preescolar	1.3	4.1
Primaria	9.6	29.8
Secundaria o Media	9.4	28.6
Técnico	0.7	1.8
Tecnológico	0.5	1.4
Universitario	3.1	8.7
Especialización	0.2	0.5
Maestría	0.1	0.3
Doctorado	0	0.1

Definamos dos eventos

A: [El individuo estudia en el hogar durante los fines de semana]

B: [El individuo que estudia en el hogar está cursando un programa de posgrado]

¿Cuál es $P(A)$?

¿Cuál es $P(B)$?

¿Cuál es $P(A \cup B)$?

¿Cuál es $P(A \cap B)$?

Soluciones

$P(A)$: Suma de probabilidades en la primera columna (24.8)

$P(B)$: Suma de probabilidades de puntos muestrales para las últimas tres filas y las dos columnas (1.2)

$P(A \cup B)$: Todos los puntos en A o B (o ambos) (25.6)

$P(A \cap B)$: Todos los puntos en ambos eventos A y B (0.3)

Reglas de probabilidad I

Regla Aditiva:

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

Probabilidad Condicional

Hasta ahora, hemos analizado probabilidades no-condicionales (*unconditional probabilities*), es decir, aquellas que no asumen una condición inicial, aparte de las definidas por el experimento.

A menudo, contamos con información adicional, que condiciona la probabilidad de un resultado en un experimento dado.

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

Ejemplo de probabilidad condicional

Definamos como **experimento** el lanzamiento de un dado.

Dos eventos:

A: $\{$Cae 1$\}$ B: $\{$Cae impar$\}$ o $\{1, 3, 5\}$

¿Cambia la probabilidad P(A), al saber que el evento B ocurrió?

$P(A|B) = ??$

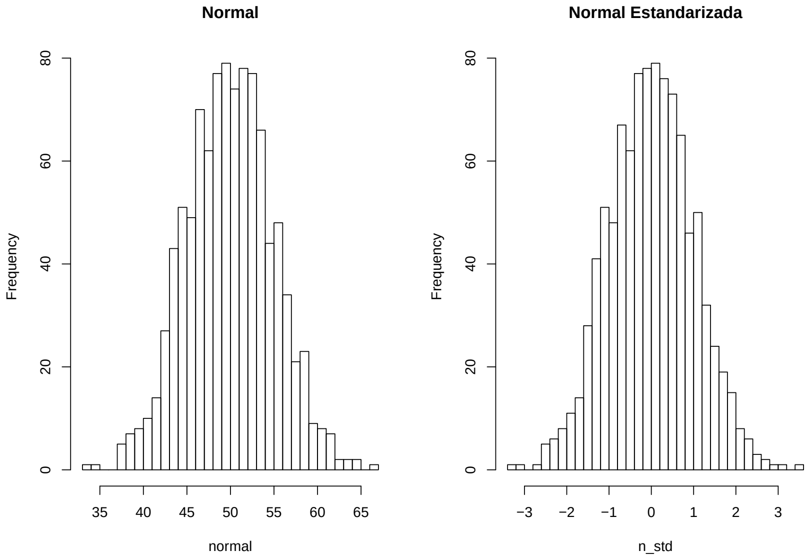
Regla Multiplicativa:

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

La distribución Normal

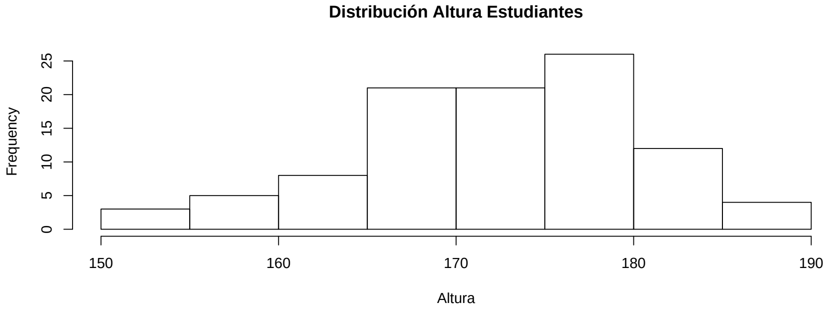
- Una de las distribuciones más comunmente observadas
- Muchos fenómenos sociales y económicos generan variables aleatorias que siguen distribuciones de probabilidad que se pueden aproximar por una distribución normal
- Perfectamente simétrica alrededor de su media (μ)
- Su dispersión está determinada por su desviación estándar (σ)

Usando la tabla de la distribución normal estandarizada



¿$P(-z_0 < z < z_0)$?

Supongamos que la altura de la clase se distribuye normalmente con media 172cm y desviación estándar 8.3cm



Si seleccionamos un estudiante al azar, ¿cual es la probabilidad de que su estatura esté entre 165cm y 179cm?

Pasos para emplear la tabla normal estandarizada

1. Calcular el valor z_0. En este caso, es igual a 0.875 (o -0.875 para el caso del límite inferior)
2. Ubicar el (los) valor(es) z_0 en el gráfico de la normal estándar
3. Definir el area que se busca
4. Ir a la tabla y buscar la probabilidad (area)
5. La tabla nos da el area bajo el segmento desde 0 hasta el valor del z_0
6. En este caso, el valor del area en la tabla corresponde a un z de 0.87 por lo que debemos promediar el area de 0 a 0.87 y de 0 a 0.88, lo que nos da 0.3092
7. Dada la simetría de la distribución, el área bajo la curva en el rango -0.875 a 0.875 es igual a 2x0.3092 o 0.6184 (62% de probabilidad)

Ejemplo para trabajar en clase

Usted es el director de operaciones de un emprendimiento social que ofrece sus servicios a través de un portal web. Según estudios de tráfico (con datos de los últimos 2 años), en promedio, la página web recibe **10 visitas diarias**, con una desviación estándar de 2. Actualmente, usted cuenta con un equipo comercial capaz de gestionar y procesar hasta **14 solicitudes** de servicio al día. Recientemente, un proveedor externo se acercó a usted para ofrecerle un servicio que permitiría incrementar esta capacidad. Suponga que no existen planes de crecimiento inmediatos en la compañía. ¿Cuál sería su sugerencia a la dirección de la compañía?

Chequeando normalidad

- Histograma
- Intervalos $\bar{x} \pm s$, $\bar{x} \pm 2s$ y $\bar{x} \pm 3s$: determine el porcentaje de datos que caen en cada intervalo (68%, 95%, 100% para normal)
- Calcule IQR y la desviación estándar. Si los datos son normales, IQR/s $= 1.3$

Intervalos de confianza

- Estimación puntual de un parámetro sobre la población
- Intervalo estimado de ese parámetro

Definido como:

$$\bar{X} \pm z \frac{\sigma}{\sqrt{n}}$$

Requiere: 1) Muestra aleatoria y 2) n grande (n>=30)

Muestras pequeñas

- La desviación estándar de la población es usualmente desconocida
- Si bien ésta se puede aproximar como $\sigma_{\bar{x}=\sigma/\sqrt{n}}$, la desviación estándar de la muestra (s) puede ser una pobre aproximación a σ cuando la muestra es pequeña (menos de 30 observaciones)
- Para ajustar por este hecho, empleamos el estadístico t en lugar del z

Intervalos de confianza para la proporción de una población

1. La media de la distribución muestral de $\hat{p}$ es p
2. La desviación estándar de a distribución muestral es $\sqrt{pq/n}$, donde $q = 1 - p$
3. Para muestras grandes, la distribución es aproximandamente normal.

El intervalo de confianza para la proporción poblacional se define como:

$$\hat{p} \pm z\sqrt{\frac{\hat{p}\hat{q}}{n}}$$

Prueba de hipótesis

Una **hipótesis estadística** es una declaración sobre el valor numérico de un *parámetro poblacional*

Hipótesis Nula: Representa la hipótesis a ser aceptada a no ser que los datos provean evidencia sobre su falsedad. Usualmente representa el *status quo*

Hipótesis Alternativa: Representa la hipótesis a ser aceptada ante la presencia de evidencia convincente. Usualmente representa el valor del parámetro poblacional para el cual el investigador recolecta evidencia

Estadístico de prueba: Se calcula usando información de la muestra y es empleado para validar o rechazar la hipótesis nula

Zona de Rechazo: Valores numéricos del estadistico de la prueba para los cuales se puede rechazar la hipótesis nula

Supuestos, **Conclusiones**

Pasos para seleccionar el tipo de prueba

1. Seleccionar la hipótesis alternativa como aquella para la que se define el experimento de muestreo (OJO: nunca debe incluir esta el signo "$=$")
2. Seleccionar la hipótesis nula como el *status quo*. Se presume que esta hipótesis será verdadera a no ser que se presente evidencia en favor de la alternativa

Las hipótesis pueden ser de una cola o dos colas. **Una cola** para pruebas del estilo *mayor a* o *menor a*. **Dos colas** para pruebas donde lo que se bsuca es probar que el valor poblacional es *diferente de* algún valor definido como *status quo*

Hipótesis sobre la media poblacional (muestras grandes)

- Dada una muestra grande, podemos asegurar que la distribución muestral de la media es aproximadamente normal, por lo que es correcto emplear un estadístico z
- Igualmente, dada una muestra grande, la desviación estándar muestral es una buena aproximación a la verdadera desviación estándar poblacional

El estadístico z puede ser entonces aproximado como :

$$z = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

Donde μ_0 representa el valor de la media poblacional especificado en la hipótesis nula.

Zona de rechazo para prueba de cola inferior

$$H_o : \mu \geq \mu_0$$

$$H_a : \mu < \mu_0$$

La hipótesis nula se rechaza en caso de que $z \leq -z_\alpha$, donde z_α corresponde al $100(1-\alpha)$ percentil de la distribución normal estandarizada.

Zona de rechazo para prueba de cola superior

$$H_o : \mu \leq \mu_0$$

$$H_a : \mu > \mu_0$$

La hipótesis nula se rechaza en caso de que $z \geq z_\alpha$, donde z_α corresponde al $100(1-\alpha)$ percentil de la distribución normal estandarizada.

Zona de rechazo para prueba de dos colas

$$H_o : \mu = \mu_0$$

$$H_a : \mu \neq \mu_0$$

Rechazamos la hipótesis nula cuando $z \leq -z_{\alpha/2}$ o cuando $z \geq z_{\alpha/2}$, donde $z_{\alpha/2}$ corresponde al $100(1-\alpha/2)$ percentil de la distribución normal estandarizada.

Muestras pequeñas

Para muestras pequeñas, la igual que en el caso de los intervalos de confianza, empleamos el estadístico t

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

Hipótesis sobre la proporción poblacional

Cola inferior

$$H_o : p \geq p_0$$
$$H_a : p < p_0$$

Donde p_0 es el límite inferior hipotético de la proporción poblacional verdadera p.

El estadístico z se define en términos de la proporción muestral y el tamaño de la muestra:

$$z = \frac{\bar{p} - p_0}{\sqrt{p_0(1 - p_0)/n}}$$

La hipótesis nula se rechaza en caso de que $z \leq -z_\alpha$, donde z_α corresponde al $100(1-\alpha)$ percentil de la distribución normal estandarizada.

Cola superior

$$H_o : p \leq p_0$$

$$H_o : p > p_0$$

La hipótesis nula se rechaza en caso de que $z \geq z_\alpha$, donde z_α corresponde al $100(1-\alpha)$ percentil de la distribución normal estandarizada.

Muestras independientes

El intervalo de confianza de la diferencia de dos medias independietes está definido como:

$$(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2}\sqrt{\frac{{s_1}^2}{n_1} + \frac{{s_2}^2}{n_2}}$$

Muestras independientes

Miremos el caso general:

$$H_o : (\mu_1 - \mu_2) = D_0$$
$$H_a : (\mu_1 - \mu_2) \neq D_0$$

donde D_0 es la diferencia hipotética, a menudo 0 (con el fin de verificar si ambas medias son iguales o no).

El estadístico z se calcula como en casos anteriores: $\frac{(\bar{x_1}-\bar{x_2})-D_0}{\sigma_{(\bar{x_1}-\bar{x_2})}}$, donde:

$$\sigma_{(\bar{x_1}-\bar{x_2})} = \sqrt{\frac{{s_1}^2}{n_1} + \frac{{s_2}^2}{n_2}}$$

Rechazamos H_0 si $|z| \geq z_{\alpha/2}$

Hipótesis sobre la diferencia de medias

Muestras emparejadas (paired)

Miremos el caso general:

$$H_o : (\mu_1 - \mu_2) = 0$$

$$H_a : (\mu_1 - \mu_2) \neq 0$$

El estadístico z se computa como en casos anteriores: $\frac{(\bar{x_1} - \bar{x_2})}{s_{(\bar{x_1} - \bar{x_2})}/\sqrt{n}}$

Rechazamos H_0 si $|z| \geq z_{\alpha/2}$

Hipótesis sobre la diferencia de proporciones

Miremos el caso general:

$$H_o : (p_1 - p_2) = 0$$

$$H_a : (p_1 - p_2) \neq 0$$

El estadístico z se computa como en casos anteriores: $\frac{(\hat{p}_1 - \hat{p}_2)}{\sigma_{(\hat{p}_1 - \hat{p}_2)}}$, donde:

$$\sigma_{(\hat{p}_1 - \hat{p}_2)} = \sqrt{\hat{p}\hat{q}(\frac{1}{n_1} + \frac{1}{n_2})}$$

Rechazamos H_0 si $|z| \geq z_{\alpha/2}$

Pasemos a la práctica

Tutorial R (1 hora para lectura, 1 hora para realización caso de estudio)