Estadistica II
Tema 0. Repaso de conceptos básicos

Curso 2009/10

Tema 0. Repaso de conceptos básicos

Contenidos

- Variables aleatorias y distribuciones de probabilidad
- La distribución normal
- Muestras aleatorias, estadísticos y distribuciones en el muestreo
- La distribución de la media muestral
- Teorema central del límite

Tema 0. Repaso de conceptos básicos

Referencias en la bibliografía

- Meyer, P. "Probabilidad y aplicaciones estadísticas" (1992)
  - Capítulos 4, 9, 12 y 13
- Newbold, P. "Estadística para los negocios y la economía" (1997)
  - Capítulos 4, 5 y 6

Variables aleatorias

- Experimento aleatorio: proceso con distintos resultados posibles, siendo incierto el correspondiente a una realización concreta
- Variable aleatoria: variable que toma valores (numéricos) asociados a los resultados de un experimento aleatorio
  - Variable aleatoria discreta: Si puede tomar un número finito o numerable de valores
  - Variable aleatoria continua: Si puede tomar un número infinito no numerable de valores (por ejemplo, valores en un intervalo de la recta real)
- Notación: representaremos las variables aleatorias con letras mayúsculas, X, y sus valores con letras minúsculas, x_1.

Ejemplos:

- El número de parados en un periodo de tiempo en una cierta región
- Los beneficios de una empresa en un periodo de tiempo
- La variación en la cotización en Bolsa de una empresa en una semana

Distribuciones de probabilidad

- Asociamos probabilidades a los valores de una variable aleatoria
- Definimos funciones (funciones de probabilidad, distribución, densidad) correspondientes a las probabilidades para valores o conjuntos de valores de la variable
- Clasificamos las variables aleatorias por la forma de sus funciones de probabilidad, densidad

Variables discretas

- Para una variable aleatoria discreta X con posibles valores $\{x_1, x_2, \ldots\}$, definimos su función de probabilidad o función de masa como

$$p_i = P[X = x_i], \text{ para } i = 1, 2, \ldots$$

- La función de probabilidad acumulada se define como

$$F(x_0) = P[X \leq x_0] = \sum_{i:x_i \leq x_0} p_i.$$

Distribuciones de probabilidad

Propiedades

- $0 \leq P[X = x_i] \leq 1$
- $F(\infty) = \sum_i P[X = x_i] = 1$
- $F(y) \leq F(x), \quad \forall y \leq x$
- $P[X > x] = 1 - P[X \leq x] = 1 - F(x)$
- $\mathrm{E}[X] = \sum_i x_i p_i$
- $\mathrm{Var}[X] = \sum_i (x_i - \mathrm{E}[X])^2 p_i = \sum_i x_i^2 p_i - \mathrm{E}[X]^2$

Distribuciones de probabilidad

Ejercicio 0.1

Una variable aleatoria discreta X toma valores con las probabilidades que se indican en la tabla siguiente:

Valor	0	1	2	3
Probabilidad	0.1	0.3	0.2	0.4

- Calcula su valor esperado y su varianza.
- Calcula el valor esperado de la variable definida como $Y = \text{máx}(2, X)$

Resultados

$$\begin{aligned} E[X] &= 0 \times 0{,}1 + 1 \times 0{,}3 + 2 \times 0{,}2 + 3 \times 0{,}4 = 1{,}9 \\ \text{Var}[X] &= \textstyle\sum_i (x_i - E[X])^2 p_i = 1{,}09 \\ E[Y] &= 2 \times (0{,}1 + 0{,}3 + 0{,}2) + 3 \times 0{,}4 = 2{,}4 \end{aligned}$$

Distribuciones de probabilidad

Ejercicio 0.1

Una variable aleatoria discreta X toma valores con las probabilidades que se indican en la tabla siguiente:

Valor	0	1	2	3
Probabilidad	0.1	0.3	0.2	0.4

- Calcula su valor esperado y su varianza.
- Calcula el valor esperado de la variable definida como $Y = \text{máx}(2, X)$

Resultados

$$\begin{aligned} \mathrm{E}[X] &= 0 \times 0{,}1 + 1 \times 0{,}3 + 2 \times 0{,}2 + 3 \times 0{,}4 = 1{,}9 \\ \mathrm{Var}[X] &= \textstyle\sum_i (x_i - E[X])^2 p_i = 1{,}09 \\ \mathrm{E}[Y] &= 2 \times (0{,}1 + 0{,}3 + 0{,}2) + 3 \times 0{,}4 = 2{,}4 \end{aligned}$$

Distribuciones de probabilidad

Variables continuas

- Para una variable aleatoria continua X, su función de distribución se define como
$$F(x) = P[X \leq x]$$
- Para una variable aleatoria continua (sin masa en ningún punto) se tiene $P(X = x) = 0$
- En lugar de la función de probabilidad, definimos la función de densidad como
$$f(x) = \frac{dF(x)}{dx} = F'(x)$$

Distribuciones de probabilidad

Propiedades

- $F(-\infty) = 0$
- $F(\infty) = 1$
- $F(y) \leq F(x), \quad \forall y \leq x$
- $f(x) \geq 0 \quad \forall x \in \mathbb{R}$
- $P(a \leq X \leq b) = \int_a^b f(x)dx \quad \forall a, b \in \mathbb{R}$
- $F(x) = P(X \leq x) = \int_{-\infty}^{x} f(u)du$
- $\int_{-\infty}^{\infty} f(x)dx = 1$
- $\mathrm{E}[X] = \int x f(x)dx$
- $\mathrm{Var}[X] = \int x^2 f(x)dx - \mathrm{E}[X]^2$

La distribución normal

Descripción

- Distribución continua más conocida y utilizada
- Relacionada con otras distribuciones (sumas, media muestral)

Definición

- Una variable continua X sigue una distribución normal con parámetros μ y σ, $X \sim \mathcal{N}(\mu, \sigma)$, si su densidad tiene la forma

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

- En este caso, $\mathrm{E}[X] = \mu$ y $\mathrm{Var}[X] = \sigma^2$

La distribución normal

Transformaciones lineales

- La combinación lineal de un número finito de variables aleatorias normales independientes tiene distribución normal
- $\mathrm{E}[\sum_i a_i X_i] = \sum_i a_i \mathrm{E}[X_i]$ y $\mathrm{Var}[\sum_i a_i X_i] = \sum_i a_i^2 \mathrm{Var}[X_i]$
- Si $X \sim \mathcal{N}(\mu, \sigma)$, entonces $Y = aX + b \sim \mathcal{N}(a\mu + b, a\sigma)$

Estandarización

- Si $X \sim \mathcal{N}(\mu, \sigma)$, se cumple que

$$Z = \frac{X - \mu}{\sigma} \sim \mathcal{N}(0, 1)$$

- La distribución $\mathcal{N}(0, 1)$ se conoce como distribución normal estándar
  - Basta con disponer de (una tabla de) valores para la distribución normal estándar

La distribución normal

Ejercicio 0.2

- Para una distribución normal con media 2,5 y varianza 4, calcula la probabilidad de que tome un valor mayor que 4.
- Considera dos variables normales independientes X_1 y X_2 con los parámetros anteriores. Calcula la probabilidad de que $X_1 - X_2$ sea mayor que 1.

Resultados

$$\begin{aligned} P(X > 4) &= P\left(\frac{X - 2{,}5}{\sqrt{4}} > \frac{4 - 2{,}5}{\sqrt{4}}\right) = P(Z > 0{,}75) = 0{,}227 \\ \mathrm{Var}[X_1 - X_2] &= 2\mathrm{Var}[X_1] = 8, \quad \sim N(0, \sqrt{8}) \\ P(X_1 - X_2 > 1) &= P\left(\frac{X_1 - X_2}{\sqrt{8}} > \frac{1}{\sqrt{8}}\right) = P(Z > 0{,}354) = 0{,}362 \end{aligned}$$

La distribución normal

Ejercicio 0.2

- Para una distribución normal con media 2,5 y varianza 4, calcula la probabilidad de que tome un valor mayor que 4.
- Considera dos variables normales independientes X_1 y X_2 con los parámetros anteriores. Calcula la probabilidad de que $X_1 - X_2$ sea mayor que 1.

Resultados

$$
\begin{aligned}
P(X > 4) &= P\left(\frac{X - 2{,}5}{\sqrt{4}} > \frac{4 - 2{,}5}{\sqrt{4}}\right) = P(Z > 0{,}75) = 0{,}227 \\
\mathrm{Var}[X_1 - X_2] &= 2\mathrm{Var}[X_1] = 8, \quad \sim N(0, \sqrt{8}) \\
P(X_1 - X_2 > 1) &= P\left(\frac{X_1 - X_2}{\sqrt{8}} > \frac{1}{\sqrt{8}}\right) = P(Z > 0{,}354) = 0{,}362
\end{aligned}
$$

Muestras aleatorias

Definiciones

- Población: conjunto completo de información sobre el valor de interés
- Muestra: subconjunto de valores de la población
- Inferencia: proceso de obtención de información sobre valores desconocidos de la población a partir de valores muestrales

Motivación

- Obtener información fiable sobre el conjunto de la población estudiando un subconjunto de la población (muestra) con coste reducido

Muestra aleatoria simple

- Cada miembro de la población tiene la misma probabilidad de pertenecer a una muestra aleatoria simple
- La selección se realiza de manera independiente
  - La selección de un individuo no afecta a la probabilidad de seleccionar cualquier otro

Distribuciones en el muestreo

Definiciones

- Estadístico: una función de la información en la muestra (su media, varianza, etc.)
  - Un estadístico es una variable aleatoria, su valor depende de la muestra escogida
- Distribución muestral: distribución de probabilidad de un estadístico sobre todas las muestras con el mismo tamaño. Cambia con el tamaño de la muestra

La media de la población

- La media poblacional es un parámetro muy relevante en muchas situaciones prácticas
- Estadístico: para una muestra aleatoria simple $X_1, \ldots, X_n$, haremos inferencia sobre la media poblacional partiendo de la media muestral,

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

La distribución de la media muestral

Propiedades

- El valor esperado de la media de la muestra es la media de la población

$$\mathrm{E}[\bar{X}] = \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \mathrm{E}[X]$$

- La varianza de la media muestral vale

$$\mathrm{Var}[\bar{X}] = \mathrm{Var}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\mathrm{Var}[X]$$

  - El valor de la varianza decrece si n aumenta
  - Podemos reducir el error aumentando el tamaño de la muestra

La distribución de la media muestral

Distribución

- Deseamos conocer la distribución de la media muestral
  - En muchos casos se necesita información adicional a la media y varianza (probabilidades)
- Si la variable X sigue una distribución normal $\mathcal{N}(\mu, \sigma)$, para una muestra aleatoria simple de tamaño n,

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

- Si X no sigue una distribución normal, resultados aproximados basados en el teorema central del límite
  - Si las variables X_i tienen media μ y desviación típica σ (ambas finitas), y si n es grande, se tiene (aproximadamente) que

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

El teorema central del límite

Ejercicio 0.3

Se tiene una muestra aleatoria simple de 100 valores de una distribución con media 25 y desviación típica 20
Calcula la probabilidad de que la media de la muestra esté entre 22 y 28

Resultados

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

$$\begin{aligned} P(22 \leq \bar{X} \leq 28) &= P\left(\frac{22-25}{20/\sqrt{100}} \leq \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \leq \frac{28-25}{20/\sqrt{100}}\right) \\ &= P(-1{,}5 \leq Z \leq 1{,}5) = 0{,}866 \end{aligned}$$

El teorema central del límite

Ejercicio 0.3

Se tiene una muestra aleatoria simple de 100 valores de una distribución con media 25 y desviación típica 20
Calcula la probabilidad de que la media de la muestra esté entre 22 y 28

Resultados

$$\begin{aligned} \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} &\sim N(0,1) \\ P(22 \leq \bar{X} \leq 28) &= P\left(\frac{22-25}{20/\sqrt{100}} \leq \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \leq \frac{28-25}{20/\sqrt{100}}\right) \\ &= P(-1{,}5 \leq Z \leq 1{,}5) = 0{,}866 \end{aligned}$$

El teorema central del límite

Aproximaciones para proporciones

- Queremos estudiar las proporciones de cumplimiento de una propiedad en la población
  - Si X_i representa si se cumple o no la propiedad de interés en un miembro de una muestra aleatoria simple de tamaño n, y la probabilidad de cumplimiento es p, entonces
    - X_i sigue una distribución Bernoulli
    - el número de casos en la muestra $X = \sum_i X_i$ sigue una distribución binomial con parámetros n y p
  - Si n es elevado, se puede aplicar el teorema central del límite para aproximar la distribución de $\hat{p} = X/n$, la proporción en la muestra (y un estimador para p)
- En este caso, las variables X_i tienen media p y desviación típica $p(1-p)$ y por el teorema central del límite se tiene (aproximadamente) que

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

El teorema central del límite

Ejercicio 0.4

Un candidato en unas elecciones locales ha encargado un sondeo sobre una muestra de 36 personas. Si la proporción de personas dispuestas a votarle en la población fuese del 36 %,

- Calcula la probabilidad de que la proporción observada en la encuesta sea superior al 38 %
- ¿Como cambia el resultado si la muestra aumenta a 100 personas?

Resultados

$$\frac{\hat{p}-p}{\sqrt{p(1-p)/n}} \sim N(0,1)$$

$$\begin{aligned} P(\hat{p} \geq 0{,}38) &= P\left(\frac{\hat{p}-p}{\sqrt{p(1-p)/n}} \geq \frac{0{,}38-0{,}36}{\sqrt{0{,}36(1-0{,}36)/36}}\right) \\ &= P(Z \geq 0{,}25) = 0{,}401 \end{aligned}$$

$$P(\hat{p} \geq 0{,}38) = P\left(Z \geq \frac{0{,}38-0{,}36}{\sqrt{0{,}36(1-0{,}36)/100}}\right) = 0{,}338$$

El teorema central del límite

Ejercicio 0.4

Un candidato en unas elecciones locales ha encargado un sondeo sobre una muestra de 36 personas. Si la proporción de personas dispuestas a votarle en la población fuese del 36 %,

- Calcula la probabilidad de que la proporción observada en la encuesta sea superior al 38 %
- ¿Como cambia el resultado si la muestra aumenta a 100 personas?

Resultados

$$\frac{\hat{p}-p}{\sqrt{p(1-p)/n}} \sim N(0,1)$$

$$\begin{aligned} P(\hat{p} \geq 0{,}38) &= P\left(\frac{\hat{p}-p}{\sqrt{p(1-p)/n}} \geq \frac{0{,}38-0{,}36}{\sqrt{0{,}36(1-0{,}36)/36}}\right) \\ &= P(Z \geq 0{,}25) = 0{,}401 \end{aligned}$$

$$P(\hat{p} \geq 0{,}38) = P\left(Z \geq \frac{0{,}38-0{,}36}{\sqrt{0{,}36(1-0{,}36)/100}}\right) = 0{,}338$$