Probabilidad y Estadística I
Semana 8
Teoremas límite y estadística descriptiva

Profesor: Nicolás López MSc

Universidad del Rosario

Universidad del Rosario

Contenido

Universidad del Rosario

Contenido

Universidad del Rosario

Distribuciónes conjuntas

Múltiples variables independientes. Ejemplo 1

Recuerde el caso que $X_i \sim Ber(p)$ con $i = 1, ..., n$. Cada realización es independiente entre sí, por lo cual se tiene que $X = \sum_i X_i \sim Bin(n, p)$ (explicación intututitiva, libro guía pg 93. Demostración: función generadora de momentos). Por la linealidad del valor esperado

$$E(X) = \sum_i E(X_i) = np$$

Y por la independecia de las X_i

$$V(X) = \sum_i V(X_i) = np(1 - p)$$

Distribuciónes conjuntas

Múltiples variables independientes. Ejemplo 2

Piense en X_i con $i = 1, ..., n$ v.a. independientes indicando la aprobación del presidente. En este caso $X_i \sim Ber(p)$ con p desconocido, pero fijo. Puede ser de interés la v.a.

$$\bar{X} = \frac{\sum X_i}{n}$$

Por la linealidad del valor esperado, nuevamente

$$E(\bar{X}) = \frac{\sum_i E(X_i)}{n} = \frac{np}{n} = p$$

Universidad del Rosario

Distribuciónes conjuntas

Múltiples variables independientes. Ejemplo 2

Asumiendo que cada realización es independiente entre sí:

$$V(\bar{X}) = \frac{1}{n \times n} \sum_i V(X_i) = \frac{p(1-p)}{n}$$

Y así, se espera que la media $\bar{X}$ en promedio sea igual a p. A medida que n es mayor, se tiene mayor precisión en la estimación. **Este resultado se mantiene para v.a. no Bernoulli, siempre y cuando sean independientes con misma media y varianza.**

Distribuciónes conjuntas

Múltiples variables independientes

Suponga entonces que tiene X_1,... v.a. independientes con la misma distribución de media μ y varianza σ^2, sea S_n la suma de las n primeras v.a:

$$S_n = X_1 + ... + X_n$$

Por linealidad

$$E(S_n) = n\mu$$

Y por independencia

$$V(S_n) = n\sigma^2$$

Esta última crece con n, sin embargo al tener

$$M_n = \frac{S_n}{n}$$

Y tiene sentido en pensar en el comportamiento asintótico de M_n

Universidad del Rosario

Distribuciónes conjuntas

Múltiples variables independientes

Recuerde que se tiene:

$$E(M_n) = \mu \qquad V(M_n) = \frac{\sigma^2}{n}$$

Por lo cual, sin necesidad de supuestos distribucionales

$$Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}}$$

Cumple que

$$E(Z_n) = 0 \qquad V(Z_n) = 1$$

El TLC indica que, independiente de la distribución de los X_i con $i = 1, \ldots,$ Z_n se distribuye normal estándar.

Universidad del Rosario

Desigualdades de Markov y Chebyshev

Markov

Si una v.a. X toma solamente valores no negativos:

$$P(X > a) \leq E(X)/a$$

Para todo $a > 0$. Es decir, si una v.a. no negativa tiene una media pequeña, es poco probable que tome valores grandes.

Universidad del Rosario

Desigualdades de Markov y Chebyshev

Markov

Sea $X \sim U[0,4]$, entonces $E(X) = 2$ y

$$P(X > 3) \leq 2/3 = 0{,}67$$

En efecto

$$P(X > 3) = 0{,}25 \leq 0{,}67$$

Universidad del Rosario

Desigualdades de Markov y Chebyshev

Chebyshev

Para una v.a. X con media μ y varianza σ^2

$$P(|X - \mu| > c) \leq \sigma^2/c^2$$

Para $c > 0$. Es decir, si una v.a. tiene una varianza baja, entonces la probabilidad que se aleje de su media también lo es.

Universidad del Rosario

Note que la desigualdad de Chebyshev se obtiene mediante la desigualdad de Markov con la v.a. no negativa $(X-\mu)^2$ y $a = c^2$. Esta desigualdad puede ser usada en la detección de atipicidades.

Teorema del límite central

Suponga que tiene X_1,... v.a. independientes con la misma distribución de media μ y varianza σ^2 (familia de variables **iid**) Sea S_n la suma de las n primeras v.a:

$$S_n = X_1 + ... + X_n$$

El TLC indica que

$$Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}}$$

Cumple

$$E(Z_n) = 0 \qquad V(Z_n) = 1$$

Y Z_n converge a una distribución $N(0,1)$ en el sentido que

$$lim_{n\longrightarrow\infty}(Z_n \leq z) = P(Z \leq z)$$

Para cada z con $Z \sim N(0,1)$.

Universidad del Rosario

Teorema del límite central

Este resultado es de gran importancia en la estadística:

- Conceptualmente: *the sum of a large number of independent random variables is approximately normal*.
- Prácticamente: *eliminates the need for detailed probabilistic models, and for tedious manipulations of PMFs and PDFs*.

Universidad del Rosario

Contenido

Universidad del Rosario

Teorema del límite central

Aproximación de Moivre-Laplace

Como se destacó al inicio de la clase, $X_i \sim Ber(p)$ para $i = 1, ..., n$ con $E(X_i) = p$ y $V(X_i) = p(1-p)$ entonces, como cada realización es independiente entre sí, se tiene que $S_n = \sum_i X_i \sim Bin(n, p)$. Por el TLC

$$\frac{S_n - np}{\sqrt{np(1-p)}} \sim_A N(0, 1)$$

Con lo cual, la aproximación de Moivre-Laplace a la distribución binomial no es más que el uso del TLC sobre la suma de X_i-s independientes con la misma distribución bernoulli.

Teorema del límite central

Aproximación de Moivre-Laplace

Sea $S_n \sim Bin(36, 0{,}5)$, entonces:

$$P(S_n \leq 21) = 0{,}8785$$

Esta es la probabilidad **exacta** del evento. De manera **aproximada**, mediante el TLC, se tiene:

$$P(S_n \leq 21) \approx P\left(Z \leq \frac{21-8}{3}\right) = P(Z \leq 1) = \Phi(1) = 0{,}8413$$

Contenido

Universidad del Rosario

Se lanza un dado corriente $n = 1000$ veces. Calcule la probabilidad exacta y aproximada de que el número 4 aparezca por lo menos 150 veces. Considere X_i como la v.a. obtener el número 4 en el i ésimo lanzamiento, con $i = 1, ..., 1000$.

Teorema del límite central

Ejercicios

De manera exacta se tiene que

$$X = \sum_i X_i \sim Ber(n = 1000, p = 1/6)$$

Asumiendo independencia entre los 1000 lanzamientos. De manera exacta se pregunta la siguiente probabilidad:

$$P(X \geq 150) = \texttt{sum(dbinom(150:1000,size=1000,p=1/6))} = 0{,}92894$$

Se tiene que $X_i \sim Bin(p = 1/6)$, entonces $E(X_i) = 1/6$ y $V(X_i) = (1/6)(5/6)$ y de manera aproximada se tiene entonces que

$$X \sim_A N(\mu = 1000 \times (1/6), \sigma^2 = 1000 \times (1/6)(5/6) = 1000 \times 5/36)$$

Asumiendo independencia entre los 1000 lanzamientos. Se pregunta

$$P(X \geq 150) = P\left(\frac{X - 1000/6}{\sqrt{5000/36}} \geq \frac{150 - 1000/6}{\sqrt{5000/36}}\right)$$

Mediante el TLC, con $Z \sim N(0, 1)$, se puede expresar como

$$P(X \geq 150) \approx_{TLC} P\left(Z \geq -1{,}41\right) = 1 - \Phi(-1{,}41) = 1 - \texttt{pnorm(-1.41)}$$

Que es igual a 0.92135.

Teorema del límite central

Ejercicios

Un elevador de carga puede transportar 1.000kg. Una carga de 45 cajas debe transportarse en el elevador. El peso de una caja es una v.a. con promedio de 200kg y desviación estándar de 55kg. Calcule la probabilidad de que las 45 cajas puedan transportarse de manera simultánea en el elevador.

En este caso se tiene

$$X_i = \text{Peso de la } i\text{-ésima caja en kg}$$

Con $E(X_i) = 200kg$ y $\sqrt{V(X_i)} = 55kg$ para $i = 1, ..., 45$. Suponiendo independencia entre las variables el TLC indica que, con $Z \sim N(0, 1)$, la probabilidad preguntada se puede expresar como:

$$P(\sum X_i \leq 1000) \approx_{TLC} P\left(Z \leq \frac{1000 - 45 \times 200}{55 \times \sqrt{45}}\right) = \Phi(-21{,}68) \approx 0$$

Teorema del límite central

Ejercicios

Una compañia tabacalera afirma que la cantidad de nicotina en sus cigarrillos es una v.a. con media 2.2mg y desviación estándar de 3mg. En una **muestra aleatoria** de 100 cigarrillos, la media muestral observada fue de 3.1mg. Si en efecto los valores de la tabacalera son ciertos, ¿cuál es la probabilidad de haber encontrado una media muestral de 3.1mg o mayor?

Universidad del Rosario

Si en efecto los valores de la tabacalera son ciertos, se tiene X_i v.a. que indica la cantidad de nicotina en el cigarrillo i con $i = 1, ..., 100$ con $E(X_i) = 2{,}2mg$ y $\sqrt{V(X_i)} = 3mg$. Al ser muestra aleatoria, se tiene que X_i son **iid**, con lo cual podemos usar el TLC sobre la suma, pero también **sobre el promedio**.

El TLC indica que

$$Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}} = \frac{S_n/n - n\mu/n}{\sigma\sqrt{n}/n} = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

Cumple

$$E(Z_n) = 0 \qquad V(Z_n) = 1$$

Con lo cual podemos hacer uso del TLC sobre totales y también sobre promedios.

Se pregunta entonces

$$P(\bar{X} \geq 3{,}1) = P\left(\frac{\bar{X} - 2{,}2}{3/\sqrt{100}} \geq \frac{3{,}1 - 2{,}2}{3/\sqrt{100}}\right) \approx_{TLC} P(Z \geq 3)$$

Con $Z \sim N(0,1)$, se tiene entonces

$$P(\bar{X} \geq 3{,}1) \approx_{TLC} 1 - \Phi(3) = 0{,}0013$$

¿Que concluye del ejercicio?

Se muestra a continuación la diferencia en peso (en libras) de 15 individuos después de una dieta particular. Diferencias positivas implican aumento de peso, por otra parte, diferencias negativas indican una disminución de peso

```
-1.51 1.73 -2.34 7.38 2.32 -2.28 2.95 3.95

3.30 -0.22 7.05 2.56 -1.48 -7.86 5.50
```

Asuma que las diferencias provienen de una distribución normal con media 0 libras y desviación estándar de a) 16 libras, b) 9 libras y c) 3 libras, independientes. Calcule la media observada y responda ¿cuál es la probabilidad de encontrar un promedio mayor o igual al encontrado en la muestra?

Se tiene que $X_1, ..., X_{15} \sim N(0, \sigma^2)$ con $\sigma = 16, 9, 3$ libras dependiendo del escenario. **Sin necesidad del TLC**, se sabe que

$$\bar{X} \sim N(0, \sigma^2/15)$$

Es decir que $E(\bar{X}) = 0$ y $\sqrt{V(\bar{X})} = \sigma/\sqrt{15}$, con lo cual

```
set.seed(1)

x = round(rnorm(15,mean=1,sd=4),2)

1 - pnorm(mean(x),0,sd=c(16,9,3)/sqrt(15))
```

Igual a 0.3670, 0.2729, y 0.035. ¿Que concluye del ejercicio?

Contenido

Universidad del Rosario

Definiciones iniciales

- Población y muestra.
- Unidad experimental.
- Diferencia entre **variable** y **dato** $\rightarrow$ medición

Repaso definiciones iniciales

Medición

Es un proceso sistemático de asignación de un valor a las características de interés de las UE. Esta medición puede ser bastante concreta (como la estatura en cms de una persona) ó bastante abstracta (como inteligencia, ó algún otro *constructo*).

Contenido

Universidad del Rosario

Clasificación de las variables

Tipos de datos según su **escala de medida**

Al observar una variable, podemos encontrar una diferenciación por la "forma" en la cual la medimos. Esta clasificación induce un mayor nivel de **sofisticación** en la medición:

1. Nominal. Ej: Sexo al nacer
   Relación de igualdad - desigualdad
2. Ordinal. Ej: Escolaridad
   Relación de igualdad - desigualdad + Orden
3. Intervalo. Ej: Puntaje IQ, temperatura (°C, °F)
   Relación de igualdad - desigualdad + Orden + Cero **relativo**
4. Razón. Ej: Ingresos mensuales, temperatura (°Kelvin)
   Relación de igualdad - desigualdad + Orden + Cero **absoluto**

Clasificación de las variables

Tipos de datos según su **escala de medida**. Ejemplo detallado 1

Para la temperatura en grados centígrados:

- Se puede determinar si dos temperaturas son iguales ó no: 10°C es igual a 10°C, pero diferente a 20°C.
- Se puede determinar si dos temperaturas son mayores o menores entre si: 10°C mayor a 0°C, pero menor a 20°C .

Universidad del Rosario

Clasificación de las variables

Tipos de datos según su **escala de medida**. Ejemplo detallado 1

Adicionalmente, iguales intervalos implican iguales diferencias, por ejemplo, una diferencia de 10°C tiene el mismo significado en cualquier punto de la escala:

+ La diferencia entre 0°C y 10°C es igual a la diferencia entre 10°C y 20°C.
- Sin embargo, 0°C, no implica ausencia de temperatura.

Con lo cual, la temperatura en °C cumple con las características de una escala de intervalo, pero no de razón.

Universidad del Rosario

Clasificación de las variables

Tipos de datos según su **escala de medida**. Ejemplo detallado 2

Figura 1: Tomado de *The Fundamental Process of Measurement*

Athlete	True Time	Watch V (Nominal)	Watch W (Ordinal)	Watch X (Ordinal)	Watch Y (Interval)	Watch Z (Ratio)
A	10	23	11	2	21	20
B	11	12	14	3	23	22
C	13	20	15	4	27	26
D	20	19	18	5	41	40
E	13	20	15	4	27	26
S	0	26	9	1	1	0

A group of athletes are trying out for the track team at the university, and they are all being timed in the 100 meter dash. Several coaches record times for each athlete. However, they use some very unusual stopwatches. The stopwatches are at several different levels of measurement

Universidad del Rosario

Clasificación de las variables

Tipos de datos según su **escala de medida**. Ejemplo detallado 2

Notas

- Las escalas de medida tienen una estructura jerárquica, es decir, están ordenadas: cada una posee sus propiedades junto a las propiedades de las anteriores escalas.
- ¿Qué pasará si hago el mismo análisis de la tabla, intercambiando la primera columna por la última?

Universidad del Rosario

Clasificación de las variables

Tipos de datos según su **escala de medida**. Ejemplo detallado 2

Da lo mismo!!!... es decir que cualquiera de estas dos columnas puede considerarse como aquella denotada como "true times".

- Puede que la primera columna tenga más sentido que la última, pues da el **tiempo en segundos** para cada atleta, una **unidad de medida** que hemos usado toda la vida. Sin embargo, ambas columnas dan la misma información.
- Si los inventores del reloj hubieran definido un segundo como la mitad de tiempo que dura actualmente, la última columna sería **más familiar** para nosotros.
- Parece arbitrario porque lo es. De la misma manera que medir nuestra estatura, ya sea en centímetros o en pulgadas, ambas son **equivalentes**.

Contenido

Universidad del Rosario

Clasificación de las variables

Tipos de datos según su **dimensión**

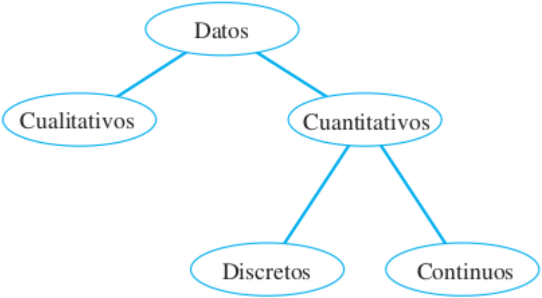
Resultan datos univariados cuando se mide una sola variable en una sola unidad experimental. Resultan datos bivariados cuando se miden dos variables en una sola unidad experimental. Resultan datos multivariados cuando se miden más de dos variables.

Clasificación de las variables

Tipos de datos según su **clase**

Al observar una variable, podemos clasificar esta observación en las siguientes clases:

Figura 2: Tomado de Introduccion a la Probabilidad y Estadística



Contenido

Universidad del Rosario

Distribución de los datos

Una vez recolectados y dispuestos en una base, los datos pueden resumirse en forma de tablas estadísticas y gráficas, las cuales pueden usarse para mostrar la **distribución** de los datos:

- ¿Qué valores de la variable han sido medidos?
- ¿Con qué frecuencia se presenta cada uno de los valores?

Lo que se busca es un **resumen más compacto de los datos** que la simple lista de observaciones.

Universidad del Rosario

Distribución de los datos

Se **sacrifica** la información individual para ganar **interpretabilidad** en los datos

Distribución de los datos

Ejemplo escolaridad

Se cuenta con el nivel educativo de una muestra de tamaño $n = 3000$ de la cual se muestran a continuación 30 observaciones:

```
[1] 1. < HS Grad 4. College Grad 3. Some College 4. College Grad
[5] 2. HS Grad 4. College Grad 3. Some College 3. Some College
[9] 3. Some College 2. HS Grad 3. Some College 2. HS Grad
[13] 2. HS Grad 4. College Grad 2. HS Grad 3. Some College
[17] 4. College Grad 1. < HS Grad 4. College Grad 3. Some College
[21] 4. College Grad 2. HS Grad 4. College Grad 4. College Grad
[25] 2. HS Grad 1. < HS Grad 2. HS Grad 2. HS Grad
[29] 2. HS Grad 2. HS Grad
```

¿Qué comentarios surgen al ver la lista de observaciones para la variable *Escolaridad*? (respecto a su distribución)

Universidad del Rosario

Distribución de los datos

Ejemplo escolaridad

Al revisar con cuidado, se observa que la escolaridad de estas 30 personas toma diferentes valores:

- `<HS Grad`. Menor de bachillerato.
- `HS Grad`. Bachillerato.
- `Some College`. Menor de universitario.
- `College Grad`. Universitario.

Esta es una variable **cualitativa ordinal**.

Universidad del Rosario

Distribución de los datos

Ejemplo escolaridad

Preguntas

- ¿Representarán estas 4 categorías todos los posibles valores de la variable *Escolaridad*?
- ¿Cuál es le valor más frecuente de la variable *Escolaridad*? ¿cuál es el menos frecuente? ¿cómo se **distribuye** la variable?

Distribución de los datos

Ejemplo escolaridad

Respuestas

- Hay 3000 observaciones, puede haber personas con menor, o mayor nivel educativo. Requiero un análisis completo de la base de datos.
- Aún con **tan sólo 30 registros**, la lista no es clara: no se puede determinar cuál valor es más o menos frecuente.
- No se pueden determinar las características de los 3000 datos sin una descripción univariada de la variable.

Distribución de los datos

Ejemplo escolaridad

Exiten múltiples herramientas **gráficas** y **numéricas** para describir un conjunto de datos. Entre las gráficas se encuentran:

- Boxplot.
- Diagrama de barras.
- Estimación kernel de la densidad.
- Histograma con polígono de frecuencias.
- *Tree map*.
- *Dot chart*.

Y las numéricas

- Tendencia central: media, mediana, moda.
- Dispersión: coeficiente de variación, rango intercuartíl.
- Localización: cuartíles 1 y 3. Mínimo y máximo.
- Asimetría: tercer momento alrededor de la media.