UNIVERSIDAD DEL NORTE

http://www.uninorte .edu .co

REPASO DE ESTADÍSTICA II

Antalcides Olivo

Julio del 2003

Back

Close

DISTRIBUCIONES MUESTRALES

Aproximación a una distribución normal estándar

Teorema 1.1 (Demoivre-Laplace.) *Supóngase que Y tiene una distribución binomial con n pruebas y con probabilidad de tener*

exito en cualquier prueba denotada por p. Por lo que podemos considerar a Y, como el número de éxitos en n pruebas, como una suma de una muestra formada por ceros y unos. Entonces

$$\bar{X} = \frac{Y}{n}$$

tendrá aproximadamente una distribución normal con

$$E(X_i) = p$$
$$V(X_i) = \frac{p(1-p)}{n}$$

Ahora indicaremos como obtener una buena aproximación.

- **Si p se acerca a 0.5 y $n > 10$ la aproximación es muy buena.**
- **En general si $np > 5$ para$0.5 \geq p$ o cuando $nq > 5$ si $p > 0{,}5$**
- **Si p se acerca al 0 o al 1 la aproximación no es buena**

Distribución ji- cuadrado

Teorema 1.2 ***Sean*** $Z_1, Z_2, Z_3, \cdots, Z_k$ ***variables aleatorias distribuidas normalmente e independientes con media cero y varianza uno, entonces la variable aleatoria***

$$X^2 = \sum_i = 1^k Z_i^2$$

tiene una dsitribución e probabilidad llamada JI-CUADRADO con k grados de libertad.

D

onde $\mu = k$ ***y*** $\sigma^2 = 2k$, ***si***

$$k \longrightarrow \infty \Longrightarrow \chi^2 \sim N\left(\mu, \sigma^2\right)$$

Y su función de densidad es :

$$f_{\chi\alpha}, k \begin{cases} 0 & si \quad x \in (-\infty, 0] \\ \frac{1}{2^{\frac{k}{2}}}\Gamma\left(\frac{k}{2}\right) x^{\frac{k}{2}} - 1 e^{-\frac{x}{2}} & si \quad x \in (0, \infty) \end{cases}$$

Si $k \to \infty \Rightarrow \mathcal{X}^2 \sim N(\mu, \sigma^2)$

Distribución t o Student

Teorema 1.3 ***Sea*** $Z \sim N(0,1)$ ***y*** V ***una variable ji-cuadrada con*** k ***grados de libertad ,si*** Z ***y*** V ***son independientes, entonces la variable aleatoria .***

$$T = \frac{Z}{\sqrt{\frac{V}{k}}}$$

Se dice que tiene una distribución t con k ***grados de libertad y se abrevia*** t_k

La media y la varianza de la distribución t son $\mu = 0$ $\sigma^2 = \frac{k}{k-2}$ **para** $k > 2$**.**

Distribución F

Teorema 1.4 ***Sean*** W ***y*** Y ***variables aleatorias JI-cuadrado independientes con*** u ***y*** v ***grados de libertad respectiva-***

mente . EL cociente

$$F = \frac{W}{u} \Big/ \frac{V}{v}$$

se dice que tiene una distribución f co u grados de libertad en el numerador y v grados de libertad en el denominador y se denota:

$$F_u, v$$

La media y la varianza son:La variable aleatoria F es no negativa y la distribución es asimétrica a la derecha, como podemos observar F se asemeja a la ji-cuadrado, pero F está centrada alrededor de 1 y los paramétros u y v lo cual le proporcionan mayor flexibilidad en cuanto a la forma de su gráfica .

Los puntos porcentuales de la cola inferior $\mathbf{F}_{1-\alpha,u,v} = \frac{1}{F_{\alpha,v,u}}$

Intervalos de confianza

En las secciones anteriores se trató los métodos para estimar un parámetro puntual, usando una estadística adecuada $\widehat{\theta}(x)$ cuyo valor $\widehat{\theta}$ se toma como el valor estimado del valor desconocido θ, pero en muchas situaciones éste método no nos proporciona suficiente información acerca del parámetro de interés, ya que sólo el número puede no tener mucho significado, entonces hay que hacer lo siguiente:

Back

Close




Back
Close

Pruebas de bondad de ajuste

Prueba chi cuadrado de bondad de ajuste

Suponemos ahora un problema que puede ser caracterizado por una variable aleatoria discreta cuyos valores representan K posibles categorias y ocurren con probabilidades $p_k : k = 1, 2, 3, \cdots K$ **en el cual nos interesa la hipótesis** $\mathbf{H_0} : p_k = p_{k_0}$; **Siendo** p_{k_0} **valores fijos, Contra la alternativa lógica** $\mathbf{H_1}$.

Como estadístico de prueba se escoge la llamada χ^2

$$X^2 = \sum_{k=1}^{K} \frac{(N_k - np_{k_0})^2}{np_{k_0}}$$

Se rechaza la hipótesis H_0 **con base en los valores concre-**

tos $n_1, \cdots, n_k$ si y sólo si para el valos x^2 de X^2 vale

$$x^2 \geq c,$$

donde c es algún valos crítico.

Ahora se usa el echo de que, bajo H_0, la estadística $X^2 \rightsquigarrow \chi^2_{K-1}$ en algunos libros se utiliza como notación para el estadístico χ^2

$$X^2 = \sum_{i=1}^{K} \frac{(O_k - E_k)^2}{E_k}$$

donde O_k es la frecuencia observada y E_k es la frecuencia esperada.

En el caso de que la distribución tenga un parámetro la hipótesis es

$$H_0 = p_k = p_k(\theta), k = 1, 2, \cdots, K$$

siendo $\theta = (\theta_1, \cdots, \theta_s)$, $s < K-1$, **un parámeto; Así**

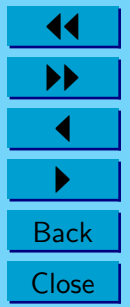
$$X^2 = \sum_{k=1}^{K} \frac{\left(N_k - np_{k_0}\left(\hat{\theta}\right)\right)^2}{np_{k_0}\left(\hat{\theta}\right)} \rightsquigarrow \chi^2_{K-1-s}$$

bajo H_0, **así los grados de libertad se reducen exactamente un número igual al número de párametros, y rechazaria-mos la hipótesis si** $x^2 > \chi^2$

Pruebas de tablas de contingencia

Prueba de independencia **En un estudio estadístico es importante averiguar si dos variables de clasificación , ya sean cualitativas o cuantitativas son independientes.**

En esta prueba los n **elementos de muestra de una población pueden clasificarse de acuerdo con dos criterios diferemtes. Por ello es interesante saber si los dos métodos de calsificación son estadísticamente independientes.**



Supongamos por ejemplo que el primer método tiene r niveles y que el segundo método c niveles . Sea O_{ij} la frecuencia observada para el nivel i y para el nivel j de los dos métodos de clasificación , por lo que los datos aparecerán como una tabla de r renglones y c columnas

Para La prueba de independencia se usan las siguientes hipótesis

H_0 : Las dos variables de clasificación son independientes.

H_1 : Las dos variables de clasificación son dependientes.

Usandose el estadístico de prueba

$$X^2 = \sum_{i,j=1}^{r,c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = \sum_{i,j=1}^{r,c} \frac{O_{ij}^2}{E_{ij}} - n$$

Donde $p_{ij} = u_i v_j$ y

$$\hat{u}_i = \frac{1}{n} \sum_{j=1}^{c} O_{ij}$$

$$\hat{v}_j = \frac{1}{n} \sum_{i=1}^{r} O_{ij}$$

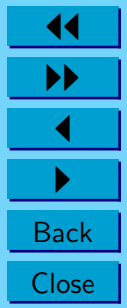
$$E_{ij} = n\hat{u}\hat{v} = \frac{1}{n} \sum_{i=1}^{r} O_{ij} \sum_{j=1}^{c} O_{ij}$$

Así tenemos que $X^2 \rightsquigarrow \chi^2_{(r-1)(c-1)}$ y se rechazará la hipótesis si $X^2 > \chi^2_{(r-1)(c-1)}$

Anova

Anova con un factor

El método de ANOVA es un críterio que requiere del cálculo de dos estimaciones independientes para σ^2, la va-



rianza poblacional común.

estas dos estimaciones las denominaremos S_b^2, S_w^2 ,donde $\mathbf{S}_b^2$ es la estimación de la varianza entre las muestras y $\mathbf{S}_w^2$ es la estimación de la varianza interior de las muestras; con lo que resulta el estadístico

$$F = \frac{S_b^2}{S_w^2}.$$

En este caso tenemos k muestras como se ilustra en la tabla

$$\begin{array}{cccccc} \bar{X}_i & \bar{X}_1 & \bar{X}_1 & \cdots & \bar{X}_k \\ S_i & S_1 & S_2 & \cdots & S_k \\ n_i & n_1 & n_2 & \cdots & n_k \end{array}$$

Para simplificar los cálculos suponemos que

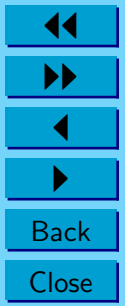
$$n = n_1 = n_2 = \cdots = n_k,$$

entonces hallamos S_w^2 la estimación ponderada para σ^2

$$S_w^2 = \sum_{i=1}^{k} \frac{S_i^2}{k},$$

como esta estimación se basa en k muestras cada una de tamaño n y cada una tiene $(n-1)$ grados de libertad asociados ellas entonces los grados de libertad gl_w asociados a $\mathbf{S}_w^2$ es

$$gl_w = \sum_{i=1}^{k} (n-1) = k\,(n-1)$$



generalizando

$$S_w^2 = \frac{\sum_{i=1}^{k} (n_i - 1)\, S_i^2}{\sum_{i=1}^{k} n_i - k}$$

$$gl_w = \sum_{i=1}^{k} n_i - k$$

$$\bar{X} = \sum_{i=1}^{k} \frac{\bar{X}_i}{k}$$

$$S_b^2 = n \sum_{i=1}^{k} \frac{\left(\bar{X}_i - \bar{X}\right)}{k-1}$$

donde los grados de libertad asociados a S_b^2 son

$$gl_b = k - 1$$

Modelos lineales

Definición 1 ***Sean Y y X variables aleatorias y suponga-mos que la relación existente entre ellas es***

$$Y = \beta_0 + \beta_1 X + \epsilon \tag{1}$$

donde $\epsilon \sim N\left(0, \sigma^2\right)$ ***es un error aleatorio, es decir***

$$E\left(Y|X\right) = \beta_0 + \beta_1 X \tag{2}$$

Notese que el modelo es lineal con relación a los lla-mados coeficientes de regresión β_0, β_1, ***es decir el modelo***

$$Y = \beta_0 + \beta_1 X^n + \epsilon, \tag{3}$$

Donde$n \in IN$ **es también un modelo de regresión lineal.**

En la definición 5.1 se establecieron 4 supuestos los cua-les son

Back
Close

1. **Para cada valor de x_i, las variables aleatorias ϵ_i se distribuyen noermalmente**
2. **Para cada valos x_i, la media o valor esperado de ϵ_i es cero**
3. **Para cada valos x_i, la varianza de ϵ_i es contante**
4. **Los ϵ_i son independientes**

Como consecuencia de los cuatro supuestos se pueden hacer las siguientes observaciones

1. **Los valores de x son fijos**
2. **Los valores de los parámetros $\hat{\beta}_0$ y $\hat{\beta}_1$ son constanes, pero desconocidas, sim embargo pueden estimarse basandonos en los cuatro supuestos y el método de los mínimos cuadrados**
3. **Como el valor de y para un x fijo está determinado por la relación $y = \beta_0 + \beta_1 x + \epsilon$. Por tanto los valores de**

y dependeran de los valores de ϵ. Por tanto y es una variable aleatoria

Para un valor fijo de x, la distribución muestral de y es normal, porque sus valores dependen de ϵ , y los valores de ϵ se distribuyen normalmente. Como muestra la figura.

La distribución muestral de y para un valor fijo de x tiene una media denotada por $\mu_{y|x}$, donde la ecuación $E(y|x) = \beta_0 + \beta_1 x$ se llama ecuación de regresión poblacional.

Ahora determinaremos los LS-Estimadores $\hat{\beta}_0$ y $\hat{\beta}_1$ de β_0 y β_1 respectivamente

Un modelo lineal de acuerdo con la definición 5.1 se llama modelo lineal simple, para este modelo supòngamos que tenemos n pares de observaciones, por ejemplo

$$(y_1, x_1), (y_2, x_2), (y_3, x_3), \cdots, (y_n, x_n).$$

Entonces empleamos estos datos para estimar los parámetros desconocidos β_0 y β_1 por medio del método de los mínimos cuadrados explicado en el capítulo 2.2.3. Por lo que

$$\begin{aligned} W(\beta_0, \beta_1) &= \sum_{i=1}^{n} e_i^2 \\ &= \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \end{aligned}$$

Entonces los LS-Estimadores de β_0 y β_1 deben satisfacer

$$\frac{\partial W}{\partial \beta_0}\Big|_{\hat{\beta}_0, \hat{\beta}_1} = -2 \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right) = 0$$

$$\frac{\partial W}{\partial \beta_1}\Big|_{\hat{\beta}_0, \hat{\beta}_1} = -2 \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right) x_i = 0$$

resolviendo este sistema de ecuaciones se obtienen $\hat{\beta}_0$ y $\hat{\beta}_1$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} y_i x_i - \frac{(\sum_{i=1}^{n} y_i)(\sum_{i=1}^{n} x_i)}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}}$$

Entonces el modelo de regresión lineal simple ajustado es

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

Para simplifiar la notación notaremos

$$S_{xx} = \sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}$$

$$S_{xy} = \sum_{i=1}^{n} y_i x_i - \frac{(\sum_{i=1}^{n} y_i)(\sum_{i=1}^{n} x_i)}{n}$$

Así $\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$.

Estimación puntual analizaremos ahora las propiedades de los estimadores $\hat{\beta}_0$ y $\hat{\beta}_1$

- $E\left(\hat{\beta}_1\right) = E\left(\frac{S_{xy}}{S_{xx}}\right) = \beta_1$
- $V\left(\hat{\beta}_1\right) = V\left(\frac{S_{xy}}{S_{xx}}\right) = \frac{\sigma^2}{S_{xx}}$
- $E\left(\hat{\beta}_0\right) = \beta_0$
- $V\left(\hat{\beta}_0\right) = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)$
- $cov = \left(\hat{\beta}_0, \hat{\beta}_1\right) = -\frac{\sigma^2\bar{x}}{S_{xx}}$

Queda como ejercicio la prueba de estas propiedades .

De acuerdo con esto observamos que $\hat{\beta}_0$ y $\hat{\beta}_1$ son estimadores insesgados .

Ahora si notamos

$$SS_E = \sum_{i=1}^{n} e_i^2 = S_{yy} - \hat{\beta}_1 S_{xy}$$

entonces $\hat{\sigma}^2 = \frac{SS_E}{n-2} \equiv MS_E$ **es un estimador insesgado de** σ^2 **y** $S_{yy} \equiv MS_R$

Prueba de hipótesis

Nos queda presentar los estadísticos que utilizaremos para una prueba de hipótesis

1. **Si tenemos las hipótesis**

$$H_0 : \beta_1 = \beta_{1,0}$$
$$H_1 : \beta_1 \neq \beta_{1,0}$$

entonces se usa el estadístico

$$t = \frac{\hat{\beta}_1 - \beta_{1,0}}{\sqrt{\frac{MS_E}{S_{xx}}}} \rightsquigarrow t_{\alpha/2,n-2}$$

2. Si

$$H_0 : \beta_0 = \beta_{0,0}$$
$$H_1 : \beta_0 \neq \beta_{0,0}$$

Se usa el estadístico

$$t = \frac{\hat{\beta}_0 - \beta_{0,0}}{\sqrt{\frac{MS_E}{S_{xx}}\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}} \rightsquigarrow t_{\alpha/2,n-2}$$

3. Si

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

Se usa el estadístico

$$F = \frac{MS_R}{MS_E} \rightsquigarrow F_{\alpha,1,n-2} \rightsquigarrow t^2_{\alpha/2,n-2}$$

Intervalos de confianza

Es posible encontrar los límites para un intervalo de confianza del $(1-\alpha)\,100\,\%$ para cada caso

1. **Para β_1 es**

$$\hat{\beta} \pm t_{\alpha/2}\sqrt{\frac{MS_E}{S_{xx}}}$$

donde

$$t_{\alpha/2} \rightsquigarrow t_{\alpha/2,n-2}$$

2. **Una vez que se ha encontrado la ecuación de regresión muestral y que se ha determinado el modelo $E(y|x) = \beta_0 + \beta_1 x$, podemos usar la ecuación de regresión $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_i$ para realizar predicciones; al hacerlo queremos estimar el valor promedio para y dado x, es decir $E(y|x)$. Así como predecir el valor aleatorio y para un valor x dado.**

Los límites del intervalo de confianza para $E(y|x_0)$ están dados por la fórmula

$$\hat{y} \pm t_{\alpha/2}\sqrt{SM_E\left(\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right)}$$

donde

$$t_{\alpha/2} \rightsquigarrow t_{\alpha/2,n-2}$$

3. **Los límites del intervalo de confianza para un valor de un solo valor aleatorio y dado un valor particular $x = x_0$ están dados por la expresión**

$$\hat{y} \pm t_{\alpha/2}\sqrt{SM_E\left(1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right)}$$

donde

$$t_{\alpha/2} \rightsquigarrow t_{\alpha/2,n-2}$$

Referencias

[1] Williams Mendenhall. *Estadística matemática con aplicaciones*
Grupo editorial iberoamericano . 1990

[2] Murray R . Spiegel . *Estadística* Mac Graw-Hill .1991

[3] Paul Meyer . *Probabilidad y aplicaciones estadísticas*
Addison Wesley, 1992

[4] Walpole Ronald . *Probabilidad y estadística*
Mac Graw-Hill . 1995

[5] Hines William. *Probabilidad y estadística*
CECSA. 1993