UNIVERSIDAD DE SANTIAGO DE CHILE

DEPARTAMENTO DE INGENIERÍA INFORMÁTICA

Redondeemos los intervalos de confianza: otras distribuciones

Extra análisis estadístico

- Deben recordar la distribución χ^2
  - Que es la distribución la **suma de los cuadrados** de *n* variables aleatorias independientes con distribución normal estándar[1]

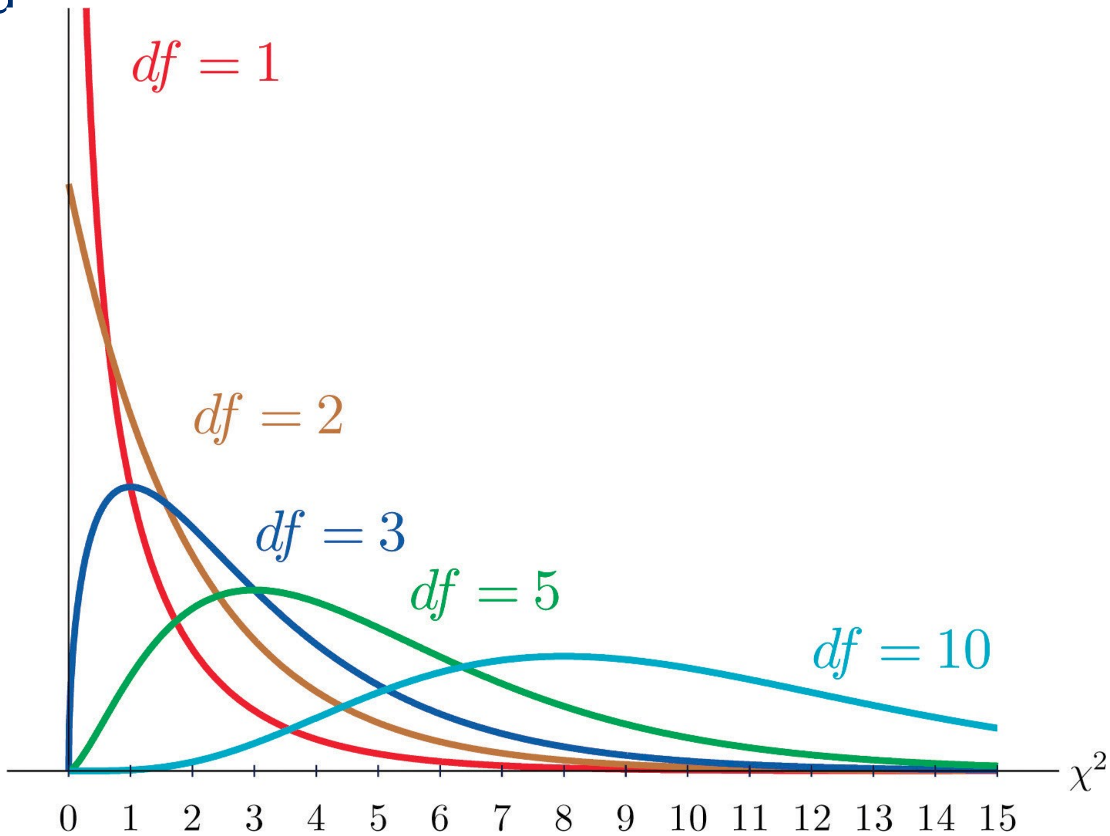
$$\chi^2_{(\nu=n)} = \sum_{i=1}^{n} {z_i}^2$$

  - Esta distribución tiene **un solo parámetro**
    - **grados de libertad (*ν*)**
    - que corresponde al número de valores *z* en la suma de los cuadrados
    - luego *ν* siempre es un valor entero

Extra análisis estadístico

- Deben recordar la distribución χ^2
  - Respecto a cada función de distribución de probabilidad[2]:
    - son positivas (solo con dominio $x \geq 0$)
    - con asimetría positiva (una larga cola superior)
    - aunque se mueven a la derecha y se vuelven más simétricas a medida que se ν incrementa

(Figura tomada de [3])



Extra análisis estadístico

- Deben recordar la distribución χ^2
  - Describe la distribución del **estimador de la varianza**[1]

$$\chi^2 = \sum {z_i}^2 = \sum \left(\frac{x - \bar{x}}{\sigma} \right)^2 = \frac{\sum (x - \bar{x})^2}{\sigma^2} = \frac{SS}{\sigma^2}$$

  - Recordando:

$$s^2 = \frac{\sum (x - \bar{x})^2}{n - 1} = \frac{SS}{n - 1}$$

  - Se obtiene:

$$\chi^2_{(n-1)} = \frac{(n-1)s^2}{\sigma^2}$$

Extra análisis estadístico

- Deben recordar la distribución χ^2
  - Luego[2]:

$$P\left(\chi^2_{(1-\alpha/2,n-1)} < \frac{(n-1)s^2}{\sigma^2} < \chi^2_{(\alpha/2,n-1)}\right) = 1-\alpha$$

y

$$\left[\frac{(n-1)s^2}{\chi^2_{(\alpha/2,n-1)}}, \frac{(n-1)s^2}{\chi^2_{(1-\alpha/2,n-1)}}\right]$$

es un **intervalo de confianza para σ^2** con nivel de confianza aproximadamente de 100·(1 - α)%

Extra análisis estadístico

- La distribución χ^2 también ayuda con las medias[1]
  - Recordando que para la **distribución muestral de la media** se cumple:

$$Z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

  - Como ***σ* es desconocida** en la práctica, se reemplaza por su **estimador *s***

$$\frac{\bar{x} - \mu}{s / \sqrt{n}}$$

  - Vimos que cuando n es suficientemente grande, este valor se aproxima a Z
  - Pero este **no es el caso general**

Extra análisis estadístico

- La distribución χ^2 también ayuda con las medias[1]
  - William S. Gosset, bajo el pseudónimo "Student", derivó la distribución de este estadístico, la que llamó **distribución t**

$$\frac{\bar{x}-\mu}{s/\sqrt{n}} \sim t_\nu = \frac{Z}{\sqrt{\frac{\chi^2_\nu}{\nu}}}$$

con $\nu = n - 1$ grados de libertad

  - La curva de *t* es muy similar a curva *z*, pero más esparcida
  - La dispersión disminuye a medida que ν se incrementa, y **la curva z a menudo se llama curva *t* con $\nu = \infty$**

Extra análisis estadístico

- Como resultado[2]:

Si $\bar{x}$ es la media y *s* la desviación estándar muestrales calculadas con una muestra aleatoria de tamaño *n* tomada de una población normal con media *μ*. Entonces

$$P\left(-t_{(\alpha/2,n-1)} < \frac{\bar{x}-\mu}{s/\sqrt{n}} < t_{(\alpha/2,n-1)}\right) \approx 1-\alpha$$

y

$$\bar{x} \pm t_{(\alpha/2,n-1)} \cdot \frac{s}{\sqrt{n}}$$

es un **intervalo de confianza para *μ*** con nivel de confianza aproximadamente de 100·(1 - α)%.

Extra-extra análisis estadístico

- Con mucha más controversia es el caso del **parámetro *p***: proporciones de éxitos

- Uno suele encontrar algo como:

  Sea $\hat{p}$ la **fracción muestral de éxitos** en una muestra de tamaño n. Si $\boldsymbol{n \cdot p \geq 5}$ (o 10) y $n \cdot (1 - p) \geq 5$ (o 10), $\hat{p}$ tiene **aproximadamente una distribución normal** con media p y varianza $p \cdot (1 - p)/n$.

  - Pero esto omite un sin número de detalles
  - Y la literatura está llena de estudios **a favor y en contra** de estas simplificaciones
  - Pero hoy en día, con el poder de cómputo moderno, la idea de usar mejores alternativas va ganado fuerza

Extra-extra análisis estadístico

- Al minuto (año 2017), parece que el método **más robusto**[2, 4-6] es el propuesto por Edwin Wilson en 1927[7] (y algunas variaciones):

Calculando: $p' = \dfrac{\hat{p} + \frac{z^2_{\alpha/2}}{2n}}{1 + \frac{z^2_{\alpha/2}}{n}}$ y $s' = \dfrac{\sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{z^2_{\alpha/2}}{4n^2}}}{1 + \frac{z^2_{\alpha/2}}{n}}$, entonces

$$p' \pm z_{\alpha/2} \cdot s'$$

es un **intervalo de confianza para *p*** con nivel de confianza aproximadamente de 100·(1 - α)%

Referencias

[1] Rudolf J. Freund, William J. Wilson, Donna L. Mohr (2010). Statistical Methods; 3rd Edition, Academic Press.

[2] Jay L. Devore (2011). Probability and Statistics for Engineering and the Sciences; 8th Edition, Duxbury Press.

[3] Douglas S. Shafer, Zhiyi Zhang (2012). Beginning Statistics - Creative Commons version (v. 1.0).Available at http://2012books.lardbucket.org.

[4] Alan Agresti, Brent A. Coull (1998). Approximate is better than “exact” for interval estimation of binomial proportions. The American Statistician, 52: 119-126.

[5] Lawrence D. Brown, T. Tony Cai, Anirban DasGupta (2001). Interval Estimation for a Binomial Proportion. Statistical Science, 16(2): 101-133.

[6] Sean A. Wallis (2013). Binomial confidence intervals and contingency tests: mathematical fundamentals and the evaluation of alternative methods. Journal of Quantitative Linguistics, 20(3): 178-208.

[7] Edwin B. Wilson (1927). Probable inference, the law of succession, and statistical inference. Journal of the American Statistical Association, 22: 209-212.