Inferencia: Estimación por intervalos

Grado en Ingeniería Biomédica

Clase 2

Jorge Calero Sanz

Departamento de Teoría de la Señal y Comunicaciones
Universidad Rey Juan Carlos

7 de marzo de 2023

Intervalos de confianza

Intervalo de confianza de la media en una normal

- Hasta ahora hemos visto métodos para estimar el valor de un parámetro. A veces, podemos estar interesados simplemente en encontrar un rango de valores posibles para ese parámetro.

Intervalos de confianza

Teorema Central de Límite

- Recordemos: el Teorema Central del Límite nos dice que cuando el tamaño de la muestra es grande ($n > 30$ ya empieza a funcionar), independientemente de como se distribuya la población X, la distribución de las medias muestrales $\overline{X}$ se distribuye como una normal $N(\mu, \frac{\sigma}{\sqrt{n}})$.
- Dicho de otra manera, si $X_i \sim X$ e independientes con media y varianza μ y σ^2 (finitas):

$$\sum_{i=1}^{n} X_i = X_1 + ... + X_n \sim N(n\mu, n\sigma^2)$$

Intervalos de confianza

Teorema Central de Límite

- Y si *estandarizamos* el estimador $\overline{X}$, obtenemos que:

$$Z = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

donde Z es una distribución normal $N(0, 1)$ (normal estándar).

Intervalos de confianza

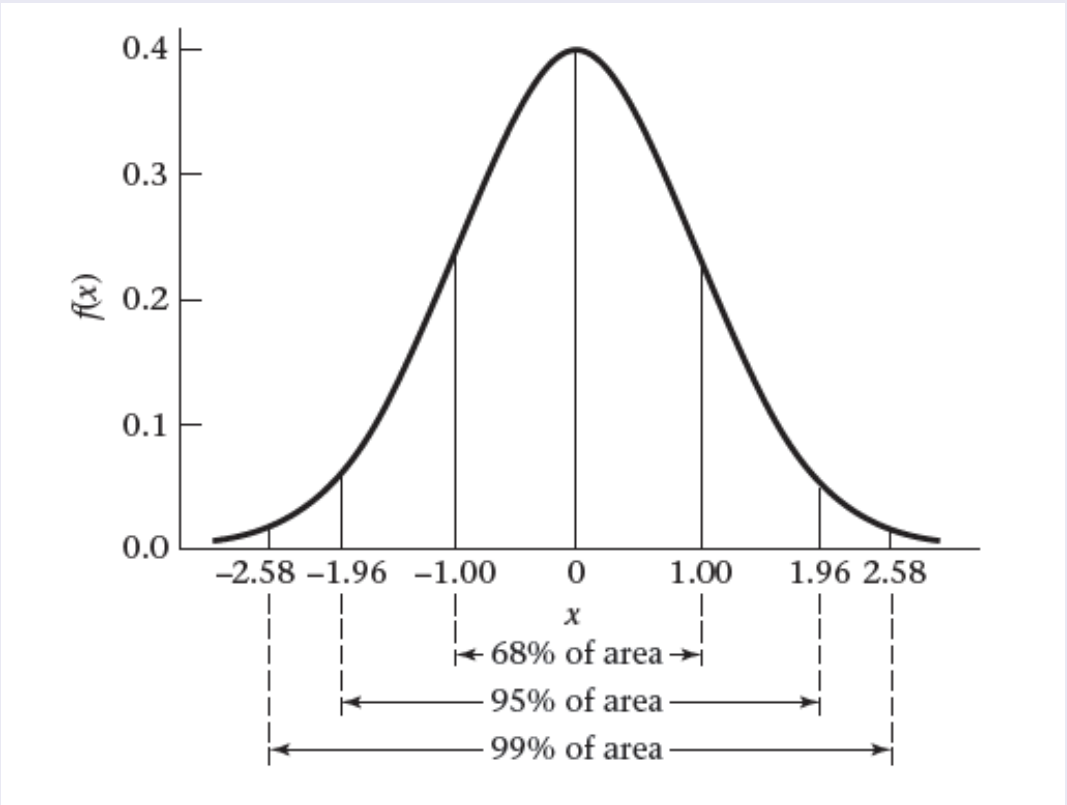
Estimación de la media μ

- Volvamos al caso de tener una población normal: $X \sim N(\mu, \sigma^2)$. En general, esto funciona para cualquier X con $E[X] = \mu$ y $Var(X) = \sigma^2$.
- Distinguiremos 2 casos:
  - la varianza σ^2 es conocida,
  - o la varianza σ^2 es desconocida.
- Gracias al TCL, tenemos que $Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$, que tenemos perfectamente cuantificada.

Intervalos de confianza

Estimación de la media μ con σ conocida

- Por ejemplo, sabemos que $P(-1.96 < Z < 1.96) = 0.95$, o dicho de otra forma: el 95 % del área de la pdf de Z queda comprendida en el intervalo $[-1.96, 1.96]$.



Intervalos de confianza

Estimación de la media μ con σ conocida

- Tenemos $P(-1.96 < Z < 1.96) = 0.95$.
- Podemos deshacer el cambio $\overline{X} = Z \cdot \frac{\sigma}{\sqrt{n}} + \mu$, obteniendo:

$$P(\mu - 1.96 \cdot \sigma/\sqrt{n} < \overline{X} < \mu + 1.96 \cdot \sigma/\sqrt{n}) = 0.95$$

- Esto nos habla del estimador $\overline{X}$ y nos dice que *tenemos un* 95 % *de certeza* de que $\overline{X}$ caerá en el intervalo $[\mu - 1.96 \cdot \sigma/\sqrt{n}, \mu + 1.96 \cdot \sigma/\sqrt{n}]$.

Intervalos de confianza

Estimación de la media μ con σ conocida

- Pero queremos encontrar un rango de valores para el parámetro μ.
- Operando, llegamos a:

$$P(\overline{X} - 1.96 \cdot \sigma/\sqrt{n} < \mu < \overline{X} + 1.96 \cdot \sigma/\sqrt{n}) = 0.95$$

Intervalos de confianza

Estimación de la media μ con σ conocida

- Retomando la normal estándar $P(-1.96 < Z < 1.96) = 0.95$, podemos sustituir el valor 95 % por $(1 - \alpha)$ %, y de este modo, los valores -1.96 y 1.96 se convierten en los percentiles que dejen encerrada un área igual a $1 - \alpha$.
- Estos percentiles, para la normal estándar, los denotamos por $\pm z_{1-\frac{\alpha}{2}}$ (recordando que $N(0, 1)$ es simétrica respecto del 0, y, por tanto, $z_\alpha = -z_{1-\alpha}$).

Intervalos de confianza

Estimación de la media μ con σ conocida

- $1 - \alpha$ es lo que llamamos **confianza** y α es el **nivel de significación**.
- Antes de la infórmatica (cuenta la leyenda que el ser humano estuvo un tiempo sin ordenadores...) estas estimaciones se hacían con tablas, y normalmente se tomaban como valores de confianza 95 %, 99 % y 99, 5 %, que incluso tienen nombre:
  - 95 %: casi significativa
  - 99 %: significativa
  - 99, 5 %: muy significativa

Intervalos de confianza

Estimación de la media μ con σ conocida

- Podemos reescribir el intervalo de confianza $1-\alpha$ como:

$$P(\overline{X} - z_{1-\frac{\alpha}{2}} \cdot \sigma/\sqrt{n} < \mu < \overline{X} + z_{1-\frac{\alpha}{2}} \cdot \sigma/\sqrt{n}) = 1-\alpha$$

- Y en notación más corta:

$$\overline{x} \pm z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$$

Estimación de intervalos

Interpretación

- Otra manera de expresarlo es decir que tenemos un 95 % de certeza que μ caerá en el intervalo $[\overline{X} - 1.96 \cdot \sigma/\sqrt{n}, \overline{X} + 1.96 \cdot \sigma/\sqrt{n}]$.
- Esto significa que si tomáramos 100 muestras de tamaño n, obtendríamos 100 valores de $\overline{X}$. De todos esos intervalos, aproximadamente en el 95 caería nuestro parámetro μ.

Intervalos de confianza

Estimación de la media μ con σ desconocida

- Sin embargo, es posible (suele ser lo general) que no conozcamos el valor de σ.
- Para arreglar esto, podemos estimar σ a través del estimador S_* (la ráiz de la cuasivarianza).
- El problema de la expresión:

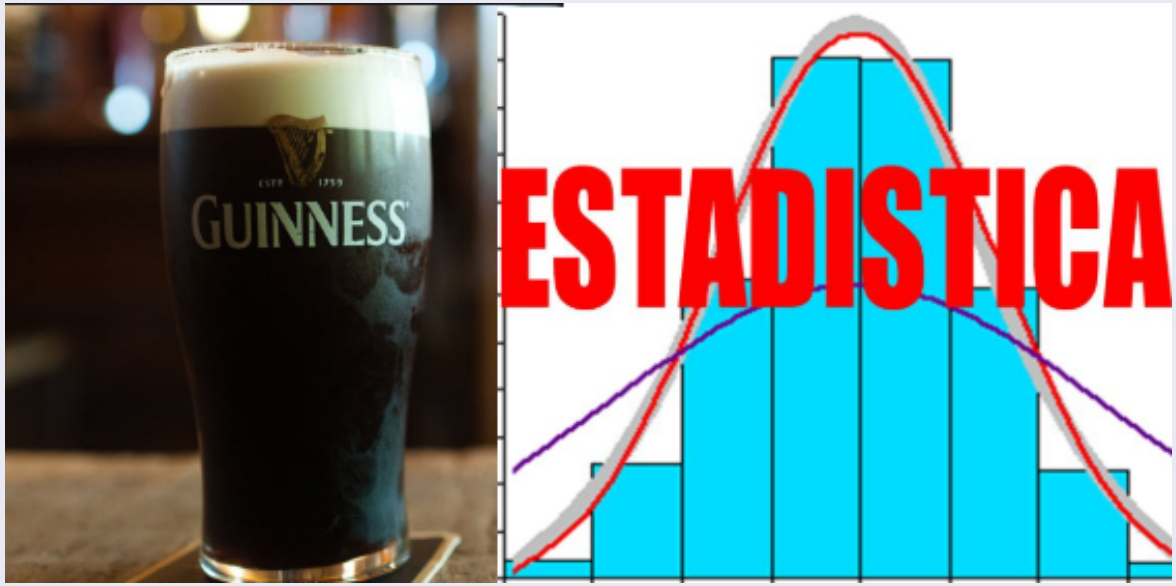
$$\frac{\overline{X} - \mu}{S_*/\sqrt{n}}$$

es que no se distribuye como una normal, pues el cambio de σ por S_* hace que la anterior expresión ya no sea la estandarización de la normal.

Intervalos de confianza

Estimación de la media μ con σ desconocida

¿Qué tienen en común la Estadística y la cerveza Guinness?



Intervalos de confianza

Estimación de la media μ con σ desconocida

En 1908, un trabajador de la cervercería Guinness, WIlliam Gossett, bajo el seudónimo de Student, encontró la distribución de la fórmula

$$\frac{\overline{X} - \mu}{S_* / \sqrt{n}}$$

donde $X_i \sim N(\mu, \sigma)$ son independientes, para $i = 1, ..., n$.
La forma de esta distribución solo depende del tamaño n y se denota como **t de Student** con $n - 1$ grados de libertad (t_{n-1}).

Intervalos de confianza

Estimación de la media μ con σ desconocida

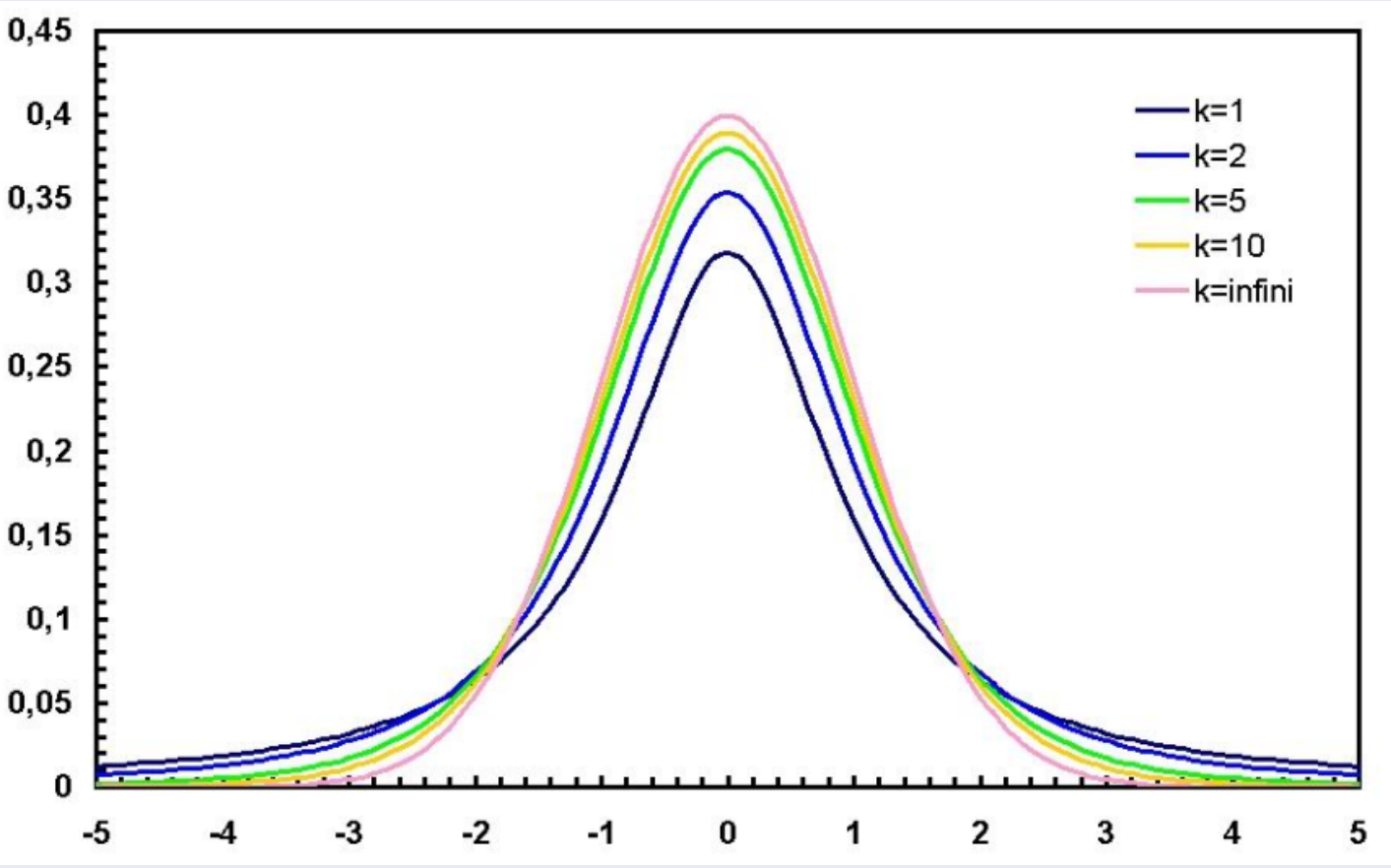
La función de densidad de la t_n de Student es:

$$f(t) = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\Gamma(\frac{n}{2})}(1 + \frac{t^2}{n})^{-\frac{n+1}{2}}$$

De modo que $E[t] = 0$ y $Var(t) = \frac{n}{n-2}$

Intervalos de confianza

Estimación de la media μ con σ desconocida



Intervalos de confianza

Estimación de la media μ con σ desconocida

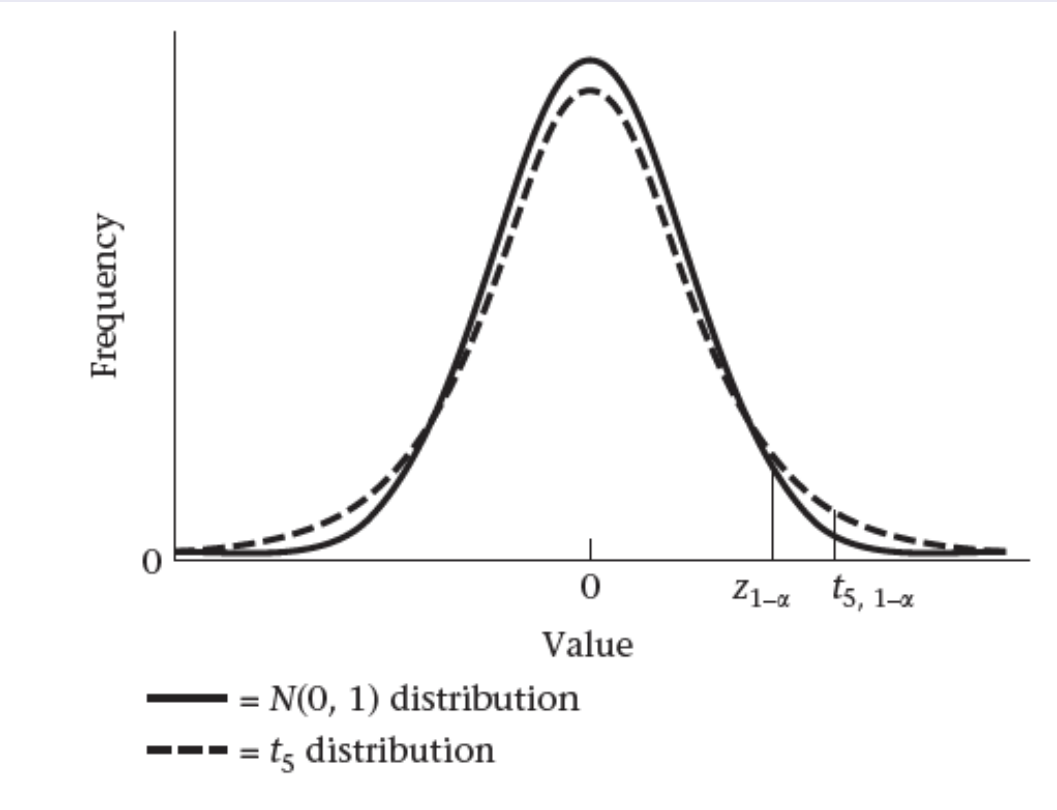
La t_d de Student está cuantificada también en tablas. Los cálculos usuales que se hacen es a través del u-ésimo percentil, que se denota por $t_{d,u}$, donde d son los grados de libertad. Entonces tenemos:

$$P(t_d < t_{d,u}) \equiv u$$

Por ejemplo, $t_{20,95}$ es aquel valor que deja el 95 % de la probabilidad a la izquierda de t_{20} (20 grados de libertad).

Intervalos de confianza

Estimación de la media μ con σ desconocida



A medida que vamos tomando más grados de libertad, la t de Student converge a $N(0, 1)$.

En particular, la t de Student es simétrica respecto al 0.

Intervalos de confianza

Estimación de la media μ con σ desconocida

En resumen:

- Si σ es desconocido, lo reemplazamos por S_*, y obtenemos:

$$t = \frac{\overline{X} - \mu}{S_*/\sqrt{n}}$$

que es una t de Student con $n-1$ grados de libertad.

Intervalos de confianza

Estimación de la media μ con σ desconocida

- Tenemos un 95 % de certeza de que t caerá entre los percentiles 2.5 y 97.5, es decir,

$$P(t_{n-1,0.025} < t < t_{n-1,0.975}) = 0.95$$

- Sustituyendo 95 % por $(1 - \alpha)$ %:

$$P(t_{n-1,\alpha/2} < t < t_{n-1,1-\alpha/2}) = 1 - \alpha$$

Intervalos de confianza

Estimación de la media μ con σ desconocida

La primera desigualdad es:

$$t_{n-1,\alpha/2} < t = \frac{\overline{X} - \mu}{S_*/\sqrt{n}}$$

Operando, podemos dejarlo como:

$$\mu < \overline{X} - t_{n-1,\alpha/2} S_*/\sqrt{n}$$

Podemos hacer lo mismo para la otra desigualdad, y usando que $t_{n-1,\alpha/2} = -t_{n-1,1-\alpha/2}$ (usando que t es simétrica), llegamos a:

$$P(\overline{X} - t_{n-1,\alpha/2} S_*/\sqrt{n} < \mu < \overline{X} + t_{n-1,1-\alpha/2} S_*/\sqrt{n}) = 1 - \alpha$$

Intervalos de confianza

Estimación de la media μ con σ desconocida

Decimos entonces que el intervalo:

$$\left[\overline{x} - t_{n-1,1-\alpha/2}\frac{s_*}{\sqrt{n}}, \overline{x} + t_{n-1,1-\alpha/2}\frac{s_*}{\sqrt{n}}\right]$$

o en notación abreviada:

$$\overline{x} \pm t_{n-1,1-\alpha/2}\frac{s_*}{\sqrt{n}}$$

que es el intervalo de confianza $(1-\alpha)$ para la media μ de una distribución normal con d.típica desconocida.

Intervalos de confianza

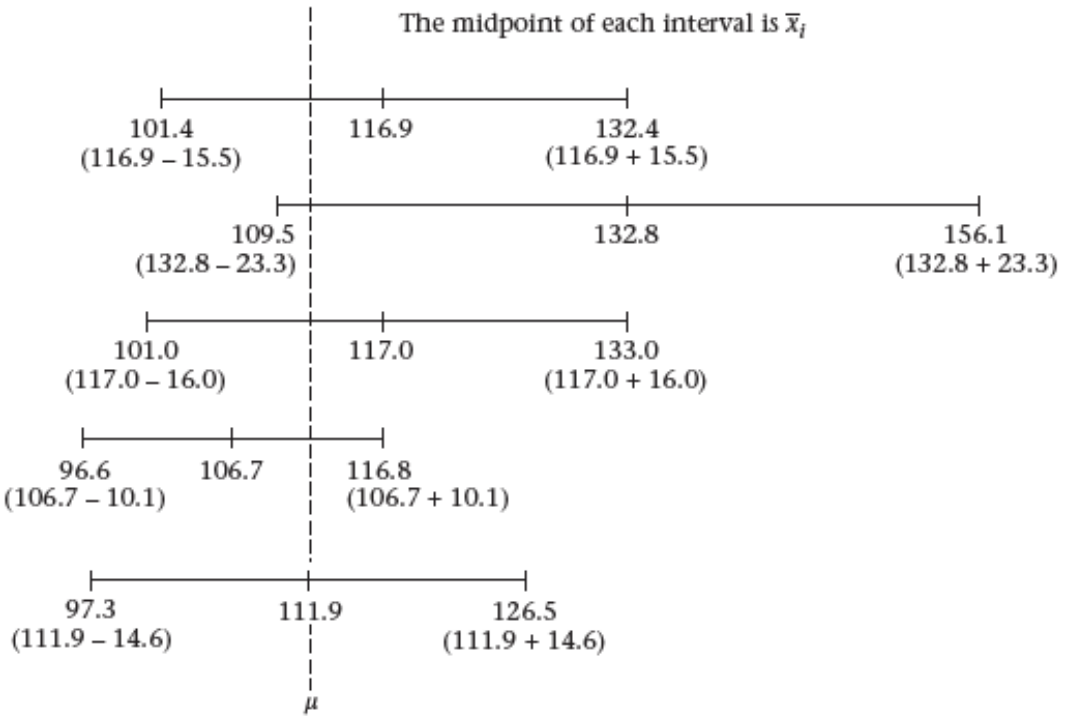
Estimación de la media μ con σ desconocida

Es importante enteder el significado de los intervalos de confianza:

- **No** significa que para una muestra particular, haya un 95 % de probabilidad de que el parámetro caiga dentro de dicho intervalo.
- Lo que **si** significa es que de una colección de intervalos de confianza 95 %, el 95 % de estos intervalos contendrán el parámetro.

Estimación de intervalos

Figure 6.7 A collection of 95% CIs for the mean μ as computed from repeated samples of size 10 (see Table 6.3) from the population of birthweights given in Table 6.2



Estimación de intervalos

Estimación de intervalos

Si la muestra que tomamos es $n > 200$, la t de Student se aproxima muy bien a $N(0, 1)$, y podemos tomar el intervalo:

$$\overline{x} \pm z_{1-\alpha/2} \frac{s_*}{\sqrt{n}}$$

Estimación de intervalos

Factores que alteran los intervalos de confianza

- Como hemos visto, hay 3 variables que pueden cambiar el tamaño del intervalo de confianza:
  - n: Cuanto más grande sea la muestra, mejor será la estimación, y por tanto, la longitud decrece.
  - s_*: Cuanto mayor sea la cuasiv de la muestra, la longitud del intervalo aumenta.
  - α Cuanto menor sea el nivel de significación, la confianza del intervalo será mayor. Pero para que estemos más seguros de que el parámetro está contenido en ese intervalo, tendremos que tomarlo de mayor longitud.

Estimación por intervalos de confianza

Intervalo de confianza para la varianza σ^2 de una normal

- Ya hemos visto como estimar la media, tanto si σ es conocido como si no, pero ... ¿Cómo estimamos σ?
- Ya sabemos que la cuasivarianza muestral S_*^2 es un estimador insesgado para la varianza σ^2.
- Para dar un intervalo de confianza de este estimador, debemos definir una nueva familia de distribuciones.

Estimación por intervalos de confianza

Intervalo de confianza para la varianza σ^2 de una normal

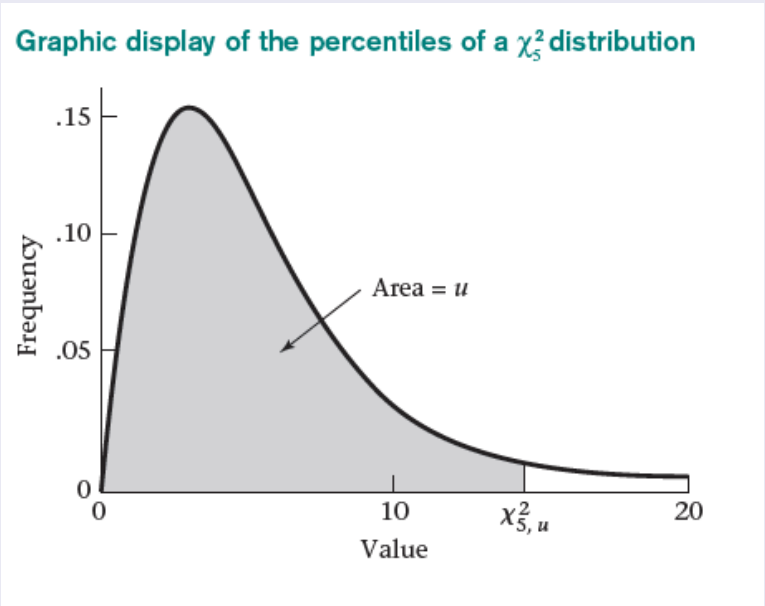
- Si $X_1, ..., X_n \sim N(0,1)$ son independientes, entonces $G = \sum_{i=1}^{n} X_i^2$, donde decimos que G sigue una distribución chi-cuadrado con n grados de libertad, y la llamamos $G = \chi_n^2$.
- Para esta distribución, tenemos los parámetros:
  - La esperanza es $E[\chi_n^2] = n$.
  - La varianza es $Var(\chi_n^2) = 2n$.

Estimación por intervalos de confianza

Intervalo de confianza para la varianza σ^2 de una normal

Estimación por intervalos de confianza

Intervalo de confianza para la varianza σ^2 de una normal



Percentil de la chi-cuadrado

Definimos el u-percentil de χ^2_n como $\chi^2_{n,u}$ al valor que verifica $P(\chi^2_n < \chi^2_{n,u}) = u$,

Estimación por intervalos de confianza

Intervalo de confianza para la varianza σ^2 de una normal

En este caso, se puede demostrar que:

$$\chi^2_{n-1} \sim \frac{(n-1)S^2_*}{\sigma^2}$$

Y por tanto, tenemos la siguiente expresión de probabilidad:

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

Estimación por intervalos de confianza

Intervalo de confianza para la varianza σ^2 de una normal

A través de unas pocas cuentas, finalmente tendremos el siguiente intervalo de confianza para σ^2:

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

Estimación por intervalos de confianza

Efecto del tabaco en la BMD en mujeres de mediana edad

- Se tomaron 41 parejas de gemelas con el fin de estudiar la relación entre fumar tabaco y la densidad de mineral en los huesos (BMD) o densidad ósea.
- Se realizaron las medidas pertinentes a cada pareja y se obtuvo la diferencia de BMD entre la gemela menos fumadora y la gemela más fumadora.
- Con estos datos, tenemos una muestra de $n = 41$, de la que se obtuvo la media $\overline{x} = -0.036$, y el error estándar $s/\sqrt{n} = 0.014$ (g/cm^2).
- Calcular el intervalo de confianza del 95 % para la media poblacional. ¿Qué conclusiones podemos sacar?

Estimación por intervalos de confianza

Efecto del tabaco en la BMD en mujeres de mediana edad

- El intervalo sería
  $-0.036 \pm t_{40,0.975}(s/\sqrt{41}) = -0.036 \pm 2.021 \cdot (0.014) =$
  $-0.036 \pm 0.028 = (-0.064, -0.008)$.
- Tenemos entonces que incluso el valor más alejado de la media dentro del intervalo -0.008 , es negativo, es decir, podemos estar bastante seguros de que la media será un valor negativo.
- Pero hemos calculado la media entre la diferencia del BMD de la persona fumadora y la menos fumadora, es decir, la BMD de la persona fumadora será mayor que la de la fumadora.
- En términos estadísticos, podemos asegurar que hay una significancia entre fumar y el BMD.

Estimación por intervalos de confianza

Estimación puntual para la binomial

- Tenemos una población X, donde la prevalencia de una característica sigue un parámetro p.
- Tomamos una muestra de tamaño n: $X_1, X_2, ..., X_n$, donde $X_i \sim Bernouilli(p)$.
- El estimador $\hat{p} = \frac{1}{n} \sum X_i$, que es la proporción muestral, es un estimador insesgado para el parámetro p.

Estimación por intervalos de confianza

Estimación puntual para la binomial

- El estimador $\hat{p} = \frac{1}{n}\sum X_i$ es insesgado:

$$E[\hat{p}] = \frac{1}{n}E[\sum X_i] = \frac{1}{n}\sum E[X_i] = p$$

- La varianza del estimador es:

$$Var(\hat{p}) = \frac{1}{n^2}Var(\sum X_i) = \frac{1}{n^2}\sum Var(X_i) = \frac{npq}{n^2} = \frac{pq}{n}$$

Estimación por intervalos de confianza

Intervalo de confianza para la binomial

- Para dar una estimación por intervalos del parámetro p de una binomial, tenemos dos opciones:
  - Si podemos aproximar la binomial por una normal.
  - O si no podemos hacerlo

Estimación por intervalos de confianza

Aproximación a la normal

- Podremos aproximar la binomial a una normal si $\hat{p}\hat{q}n \geq 5$, donde $\hat{p}, \hat{q}$ son los valores muestrales.
- En ese caso, el intervalo de confianza vendrá dado por:

$$\hat{p} \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}\hat{q}}{n}}$$

donde $z_{1-\alpha/2}$ es el percentil de la normal estándar.

Estimación por intervalos de confianza

Intervalo de confianza para la binomial

- Si no es posible aproximar la binomial a una normal (cuando tengamos $\hat{p}\hat{q}n \leq 5$, el intervalo de confianza será (p_1, p_2), donde los valores se obtienen de:

$$P(X \geq x | p = p_1) = \frac{\alpha}{2}$$
$$P(X \leq x | p = p_2) = \frac{\alpha}{2}$$

donde $x =$ *número de valores observados*

Estimación por intervalos de confianza

Intervalo de confianza para la Poisson

- Recordemos que la distribución Poisson se usaba para modelizar la probabilidad de que ocurran cierto número de sucesos en un intervalo fijo de tiempo ,cuando los eventos ocurren de manera independiente y con una media constante.
- Para estudios longitudinales (se estudia una población a lo largo del tiempo), definimos X como el número de eventos ocurridos en el tiempo estudiado.
- $X \sim Poisson(\mu = T\lambda)$, donde
  - T es el intervalo de tiempo.
  - λ es el número de eventos por unidad de tiempo.
- Tendremos $E[X] = \mu = T\lambda$.

Estimación por intervalos de confianza

Intervalo de confianza para la Poisson

- Persona-años es unidad de tiempo definida como 1 persona estudiada en 1 año).
- Ejemplo: Durante 10 años se observan a 12.000 menores de edad para detectar si poseen una enfermedad.
- Si hay 12000 elementos en la muestra observados en 10 años, habrá $T = 120.000$ personas-año.

Estimación por intervalos de confianza

Intervalo de confianza para la Poisson

- Un estimador para λ está dado por $\hat{\lambda} = X/T$, donde X es el número observado de eventos en T personas-año.
- El estimador es insesgado pues

$$E[\hat{\lambda}] = E[X/T] = \frac{E[X]}{T} = \frac{\lambda T}{T} = \lambda$$

Estimación por intervalos de confianza

Intervalo de confianza para la Poisson

- Si en esos 10 años se detectaron en total 12 casos, tenemos que la estimación del parámetro λ es $\hat{\lambda} = X/T = 12/120000 = 0.0001$.
- Multiplicando por 10^5, tenemos que ratio de incidencia es $0.0001 \cdot 10^5 = 10$ por $10^5 = 100000$ personas-año.

Estimación por intervalos de confianza

Intervalo de confianza para la Poisson

- El intervalo de confianza $(1-\alpha)$ para el parámetro λ viene dado por $(\mu_1/T, \mu_2/T)$ donde :

$$P(X \geq x|\mu = \mu_1) = \frac{\alpha}{2}$$
$$P(X \leq x|\mu = \mu_2) = \frac{\alpha}{2}$$

donde $x =$ número de eventos observados y $T =$ número de personas-año observadas

Estimación por intervalos de confianza

Intervalo de confianza para la Poisson

Exact Method for Obtaining a CI for the Poisson Parameter λ

An exact $100\% \times (1-\alpha)$ CI for the Poisson parameter λ is given by $(\mu_1/T, \mu_2/T)$, where μ_1, μ_2 satisfy the equations

$$Pr(X \geq x \mid \mu = \mu_1) = \frac{\alpha}{2} = \sum_{k=x}^{\infty} e^{-\mu_1} \mu_1^k / k!$$

$$= 1 - \sum_{k=0}^{x-1} e^{-\mu_1} \mu_1^k / k!$$

$$Pr(X \leq x \mid \mu = \mu_2) = \frac{\alpha}{2} = \sum_{k=0}^{x} e^{-\mu_2} \mu_2^k / k!$$

and x = observed number of events, T = number of person-years of follow-up.