Instituto Valenciano de Investigaciones Agrarias

Seminario

MÉTODOS ESTADÍSTICOS PARA LA INVESTIGACIÓN AGRONÓMICA

Sesión 3

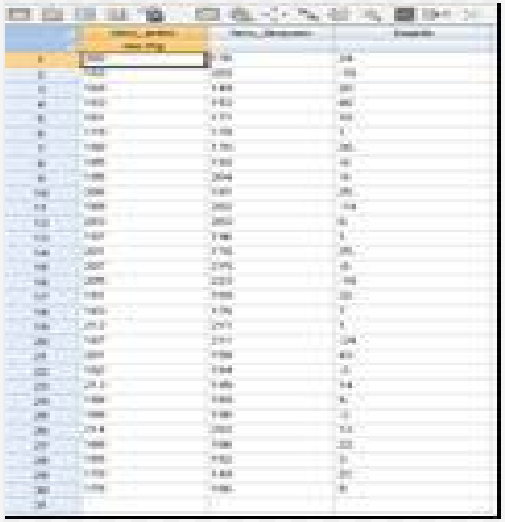
INFERENCIA ESTADÍSTICA BÁSICA SOBRE UNA POBLACIÓN NORMAL

Inferencia Estadística Básica sobre una población normal

- Ejemplo
- Cuestión previa: “¿son los datos normales?”
- Inferencia sobre la media poblacional
  - Contraste de la hipótesis m = 0.
  - La distribución t de Student
  - Intervalo de confianza para m
- Determinación del tamaño muestral
- Intervalo de confianza para σ
- Intervalo de confianza para una proporción p
- Ejercicios

Inferencia sobre una población normal: Ejemplo

- Se desea estudiar si el consumo regular de cierta infusión reduce la tensión en personas hipertensas.
- En 30 personas hipertensas se tomó la tensión (en condiciones estandarizadas) antes y después de estar dos semanas tomando regularmente dicha infusión.
- Los resultados obtenidos son:



- ¿Qué puede afirmarse sobre la efectividad de la infusión para reducir la tensión?

Planteamiento estadístico del ejemplo

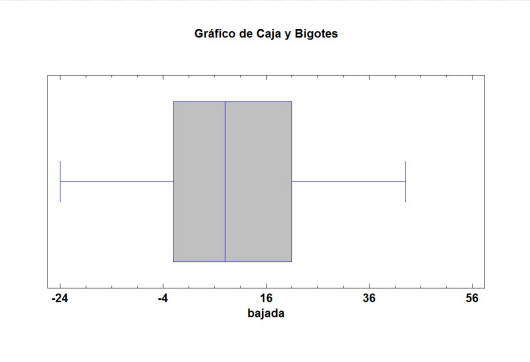
- La población objeto del estudio la constituyen todas las personas hipertensas (habría que precisarla más: ¿ambos sexos?, ¿adultos y niños? ¿qué es hipertenso?,…)
- La variable estudiada es la bajada de tensión tras estar dos semanas tomando la infusión. Esta variable tiene en la población un valor medio desconocido **m**
- Si **m** es igual a 0 la infusión no influye sobre la tensión. Si **m** es mayor que 0 la infusión es efectiva, dependiendo su grado de efectividad del valor concreto de **m**
- Las 30 personas estudiadas se considera que pueden ser consideradas, *"a los efectos del estudio"*, como extraídas al azar de la población estudiada (discutir esto)
- Se desea utilizar la información contenida en la muestra para sacar conclusiones sobre **m**

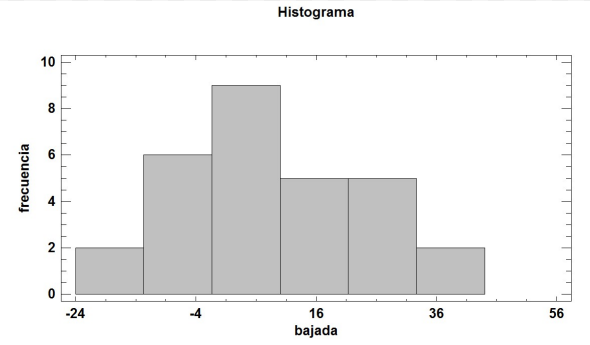
Cuestión previa: ¿son los datos normales?

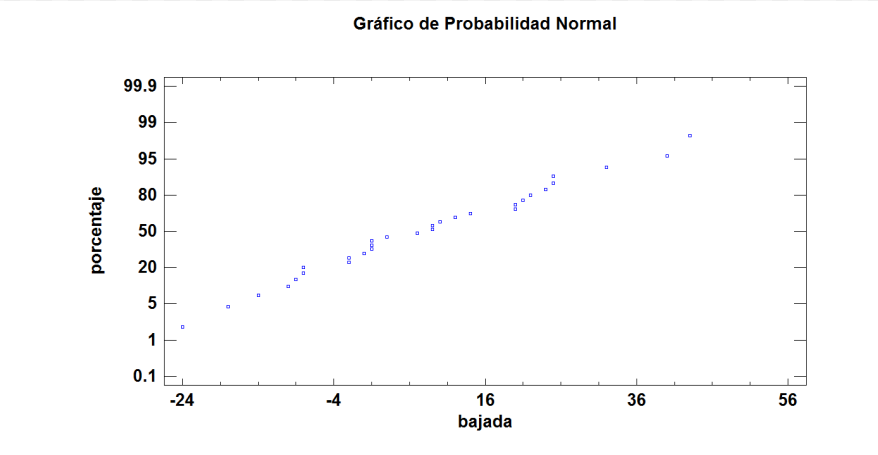
- La mayor parte de las técnicas de inferencia estadística clásicas sobre variables aleatorias continuas, asumen que las variables muestreadas se distribuyen normalmente.
- Antes, por tanto, de aplicar estas técnicas es aconsejable analizar hasta qué punto el modelo Normal es adecuado, a la vista de los datos obtenidos, para analizar la pauta de variabilidad de la variable estudiada
- Análisis de la normalidad de los datos:
  - **Tests formales** de ajuste a distribuciones: test Gi_2, Kolmogorv_Smirnov, Saphiro_Wilks,… **Dudosa utilidad práctica,** constituyendo "la respuesta correcta a una pregunta equivocada".
  - Cálculo de parámetros descriptivos de la no normalidad: coeficientes de asimetría y curtosis
  - Diagrama Box-Whisker e Histograma (si hay al menos unos 40 datos)
  - **Gráfico en papel probabilístico normal**: es el procedimiento más aconsejable para detectar datos anómalos, asimetrís, mezcla de poblaciones, …
- La conclusión no debe ser “mis datos se distribuyen normalmente” (carece de sentido epistemológico), sino más bien “el modelo teórico de la distribución normal es adecuado para analizar mis datos”

Normalidad de los datos del ejemplo

Número de datos	30
Promedio	8.13333
Desviación Estándar	16.7285
Mínimo	-24.0
Máximo	43.0
Sesgo Estandarizado	**0.44530**
Curtosis Estandarizada	**-0.52119**







Conclusión: la distribución Normal es un modelo adecuado para analizar estos datos

Contraste de la Hipótesis m = 0

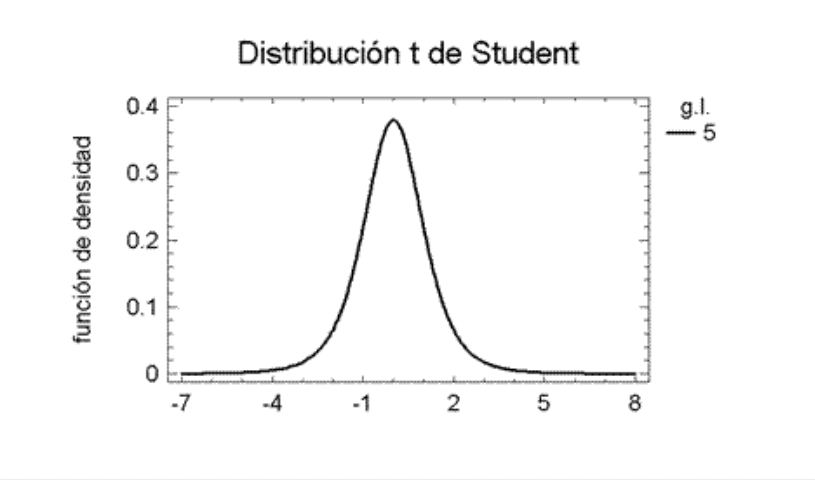
- ¿Hay evidencia en los datos de que la infusión tiene un efecto positivo reduciendo la tensión en hipertensos?
- Hipótesis nula: **H_0: m = 0** o sea : la infusión no tiene efecto (prudente escepticismo)
- Medida natural de la discrepancia entre la muestra y H_0: valor absoluto de (x_{bar} – 0)
- Medida estadística de la discrepancia: como (x_{bar} – 0) puede ser grande si x_{bar} tiene mucha variabilidad, se debe dividir por la desviación típica estimada de la media muestral, obteniéndose el estadístico **t**

$$t = \frac{x - 0}{s/\sqrt{N}}$$

- (justificar el divisor que aparece en la expresión anterior)
- Se demuestra que si H_0 es cierta el estadístico t sigue una distribución denominada **t de Student** con N-1 grados de libertad
- Por el contrario, si m $\neq$ 0, **t** tiende a tomar valores mayores (<u>en valor absoluto</u>) de los que cabe esperar para una t de Student

La distribución t de Student

- Matemáticamente una distribución t de Student con n grados de libertad aparece cuando se divide una variable N(0,1) por la raíz cuadrada de una Gi-2 con n grado de libertad e independiente de la N(0,1)
- La distribución t_n se parece a la N(0,1), siendo simétrica respecto a 0 y con σ cercana a 1, pero tiene más curtosis que una normal



- Las probabilidades asociadas a una t_n vienen en tablas o las calculan directamente los softwares estadísticos

Contraste de la Hipótesis m = 0 (continuación)

- H_0 se rechazará si **t** es "demasiado grande" (en valor absoluto) para ser una t de Student, o sea si $P(|t_{N-1}| > |\mathbf{t}|) < \alpha$ (riesgo de 1ª especie)
- A la probabilidad $P(|t_{N-1}| > |\mathbf{t}|)$ se le denomina **p-value**
- En el ejemplo resultan: x_{bar}=8.13 s=16.7 N=30 y, por tanto

$$t = \frac{\bar{x} - 0}{s/\sqrt{N}} = \frac{8.13 - 0}{16.7/\sqrt{30}} = 2.66$$

- En tablas (o con el software) se obtiene:
  - $P(|t_{N-1}| > |2.66|) = 0.012$ (p-value)
- Al ser el p-value < 0.05 se rechaza H_0 y se concluye que **m** es significativamente mayor (por ser **t** positivo) que 0 (para α=0.05)
- **Conclusión**: la infusión tiene un efecto favorable estadística-mente significativo, reduciendo en promedio la tensión en personas hipertensas

Intervalo de confianza para m

- Una forma alternativa, y más informativa, de inferir conclusiones sobre la media poblacional **m** es obtener un **intervalo de confianza** para la misma
- Intervalo de confianza para **m**: intervalo construido a partir de la muestra (y por tanto **aleatorio**) de forma que tenga a priori una probabilidad elevada (1-α) de contener el valor real de **m**. A (1-α) se le denomina **nivel de confianza** del intervalo
- Fórmula del intervalo de confianza para **m**:

$$\left[\bar{x} - t_{N-1}(\alpha/2)\frac{s}{\sqrt{N}}\,;\,\bar{x} + t_{N-1}(\alpha/2)\frac{s}{\sqrt{N}}\right]$$

- donde $t_{n-1}(\alpha/2)$ cumple $P(-t_{n-1}(\alpha/2) < t_{n-1} < +t_{n-1}(\alpha/2)) = 1-\alpha$
- En el ejemplo $t_{30-1}(0.05/2)=2.042$ con lo que el intervalo es:

$$\left[8.13 - 2.042\frac{16.7}{\sqrt{30}}\,;\,8.13 + 2.042\frac{16.7}{\sqrt{30}}\right] = [1.9;14.4]$$

- Puede afirmarse, con un 95% de confianza, que la infusión reduce **en promedio** la tensión entre 1.9 y 14.4 mmHg

Intervalo de confianza para m (observaciones)

- El intervalo [1.9 ; 14.4] no contiene a 0, por lo tanto la H_0 m=0 se rechaza
- Cuanto más elevado sea el nivel de confianza deseado para el intervalo, más amplio resultará éste (justificarlo)
- El intervalo de confianza es para la media **m** y **ino para los valores individuales de la variable!**
- **Una costumbre errónea**: frecuentemente se ve en las revistas al exponer unos resultados como los anteriores: "*la media obtenida ha sido 8.13 (±16.7)".* **La expresión x_{bar}(±s) no tiene ningún sentido**.
- Una forma correcta de exponer este tipo de resultados sería, por ejemplo: *"en la tabla se recogen los valores medios obtenidos en cada muestra y, entre paréntesis, la desviación típica de las mismas (en la tabla aparecería 8.13(16.7) sin el ±)*

Consideraciones sobre el tamaño de la muestra

- En el estudio anterior podríamos habernos preguntado: ¿Qué número mínimo de personas deberíamos haber incluido en el estudio?
- ¿Mínimo … para qué?
- Para responder a la pregunta se necesita conocer tres cosas:
  - El valor mínimo δ que se desearía detectar como significativo, operando con un riesgo de 1ª especie α, si **m** fuera realmente $\geq \delta$
  - El riesgo β que se está dispuesto a asumir de no detectar como significativo un valor de **m** que sea realmente $\geq \delta$
  - Conocer, o tener una estimación previa aproximada σ^*, del valor de σ en la población muestreada
- Conociendo estos datos, el valor mínimo de N viene dado por la expresión aproximada

$$N = \left(\frac{\left(z_{\alpha/2} + z_{1-\beta} \right) \sigma^*}{\delta} \right)^2$$

- donde $z_{\alpha/2}$ y $z_{1-\beta}$ son percentiles de la distribución normal tipificada que verifican. $P(N(0,1) > z_{\alpha/2} = \alpha/2$ y $P(N(0,1) > z_{1-\beta}) = 1-\beta$

Ejercicios

- Un meteorólogo afirma que en la Comunidad Valenciana más de la mitad de las nubes que generan granizo dañino tienen una altura sobre el nivel del mal superior a 9.000 mts. ¿Qué pude decirse sobre esta afirmación tras analizar los datos del archivo *granizo.xls*?
  - Nota 1: se desea que la probabilidad de equivocarse afirmando que la opinión del meteorólogo es errónea sea inferior al 1%
  - Nota 2: tener en cuenta el carácter marcadamente asimétrico de la variable analizada

- Una bajada de tensión producida por un fármaco se considera relevante en la práctica si, como media, es al menos de 5 mmHg. Se conoce que la bajada de tensión al tomar un fármaco fluctúa de unas personas a otras (incluyendo el efecto de la imprecisión de la medida) con una σ que vale aproximadamente 15 mmHg.
  - ¿Qué tamaño mínimo de muestra es necesario si se desea que la probabilidad de detectar como significativo (trabajando con α=0.05) una bajada de tensión que sea realmente relevante, sea superior al90%?
  - ¿Cuál es el tamaño mínimo necesario si se considera suficiente detectar como significativos efectos que sean superiores a 20 mmHg?

Intervalo de confianza para σ

- Sea s^2 la varianza muestral obtenida en una muestra de tamaño N extraída de una población Normal de varianza σ^2
- Se demuestra que $(N-1)s^2/\sigma^2$ sigue una distribución Gi-2 con N-1 grados de libertad.
- De lo anterior se deduce fácilmente la siguiente expresión del intervalo de confianza para σ

$$\left[\sqrt{(N-1)\frac{s^2}{g_2}}\ ;\sqrt{(N-1)\frac{s^2}{g_1}}\right]$$

- Donde g_1 y g_2 se obtienen de la tabla de la gi_2 de forma que verifiquen

$$P\left(g_1 < \chi^2_{N-1} < g_2\right) = 1-\alpha$$

- **Ejercicio**: En el ejemplo analizado sobre la bajada de tensión, ¿es admisible la hipótesis de que la bajada de tensión provocada por la infusión fluctúa de unos individuos a otros con una σ igual a 20 mmHg?

Intervalo de confianza para una proporción p

- ¿Cómo obtener un Intervalo de Confianza para la proporción p de individuos de una población que verifican cierto suceso, a partir de la frecuencia X observada en una muestra de tamaño N?
- Se sabe que el valor observado de la muestra de la proporción (p* = X/N) fluctúa de forma aproximadamente normal con media p y varianza p(1-p)/N
- Un intervalo de confianza aproximado para p viene dado por

$$\left(P^* - t_{n-1}^{\alpha/2}\sqrt{\frac{P^*(1-P^*)}{N}} \;;\; P^* + t_{n-1}^{\alpha/2}\sqrt{\frac{P^*(1-P^*)}{N}} \right)$$

- Nota: existe una forma exacta de obtener el intervalo de confianza, aunque es más laboriosa computacionalmente
- **Ejercicio**: se desea hacer una encuesta para estimar el % de huertos citrícolas afectadospor cierta virosis en la Comu nidad Valenciana. ¿Qué tamaño mínimo N debe tener la muestra si se desea que la probabilidad de cometer un error superior a 5 puntos porcentuales en la estimación de dicha proporción P sea inferior al 1%?