Test estadístico de hipótesis sobre la MEDIA

Antonio Sala Piqueras

Notas sobre control de sistemas multivariables/complejos

Dept. Ing. Sistemas y Automatica (DISA)
Universitat Politècnica de València (UPV)

Video-presentación disponible en:

personales.upv.es/asala/videos/**thmd**.html

Presentación

Motivación:

En muchas aplicaciones se desea comprobar si la media de una variable aleatoria tiene un valor “normal” o si supera un cierto “umbral”. Si las medidas son ruidosas, hay que promediar varias muestras.

Objetivos:

Comprender cómo establecer umbrales sobre la media muestral para desmentir la hipótesis anterior (con una cierta probabilidad de equivocarse –falsa alarma–).

Contenidos:

Tipos de hipótesis sobre la media. Caso varianza conocida: test bilateral/unilateral. Caso varianza desconocida. Caso no estacionario (Filtro de Kalman). Conclusiones.

Hipótesis sobre la **media**

Suponemos: (*a*) se miden n muestras de y al azar[1] $\{y_1, \ldots, y_n\}$, (*b*) Estas muestras tienen distribución (aprox.) normal, (*c*) desviación típica de y es conocida, igual a σ.

Hipótesis de partida $\mathcal{H}_0$:

1. la media de y es μ. [Test bilateral]
2. la media de y es $\geq \mu$. [Test unilateral]
3. la media de y es $\leq \mu$. [Test unilateral]

Si $y \sim N(\mu, \sigma^2)$, la media muestral es $N(\mu, \left(\frac{\sigma}{\sqrt{n}}\right)^2)$. Normalizando:

$$Z := \frac{\overbrace{\left(\frac{1}{n}\sum_{i=1}^{n} y_i\right)}^{\text{media muestral}} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

[1]Si son medidas tomadas a lo largo del tiempo (serie temporal), que las características estadísticas no varíen en el tiempo y sea "ruido blanco" (muestras independientes, idénticamente distribuidas).

Hipótesis sobre la **media**, con dt conocida

- **Prueba bilateral (two-tailed):** Si la media de y es μ, Z debería ser (aprox.) una distribución normal de media cero y desv. típica unidad. Dado un nivel de confianza α, la tabla de distr. normal dará el intervalo de Z fuera del cual se rechaza la hipótesis.

  Conf. 95%: el 95% de las veces en que $\mathcal{H}_0$ sea cierta, $Z \in [-1.96, +1.96]$

  Matlab: El $(1-\alpha)$% de las veces que $\mathcal{H}_0$ sea cierta, $Z \in [-\xi, +\xi]$, $\xi =$`sqrt(2)*erfcinv(`α`)`. También se obtendrían los límites inferior y superior con `norminv(`$\alpha/2$`)`, `norminv(1-`$\alpha/2$`)`.

- **Prueba unilateral (one-tailed):** Si la media de y es $\leq \mu$, eso se desmentirá con Z grande; el peor caso es igualdad. Se desmiente si $Z \geq \xi$, $\xi =$`sqrt(2)*erfcinv(2*`α`)`, o, con el mismo resultado `norminv(1-`α`)`.

  Por ejemplo, 95% resulta $\xi = 1.645$, 99% resulta $\xi = 2.32$.

¿Y si "y" no es un "ruido blanco"?

Ejemplo 1: Proceso no estacionario $m_{k+1} = m_k + w_k$, $y_k = m_k + v_k$ donde la media m "varía" con un paseo aleatorio (w_k es ruido de proceso "pequeño" comparado con v_k).

Ejemplo 2: Además, con autocorrelación temporal:

$$m_{k+1} = m_k + w_k$$
$$x_{k+1} = Ax_k + B\xi_k$$
$$y_k = m_k + Cx_k + v_k$$

La solución*, si los ruidos v, ξ, v tienen distribución normal y sus varianzas son conocidas, se basa en el **Filtro de Kalman no estacionario**: obtiene la distribución **a posteriori** de m, y se calcula la probabilidad de que la media sea inferior (o superior) a μ con fórmulas de la distribución normal (`erf`).

*Las fórmulas en transp. anteriores son un caso particular, con $w_k \equiv 0$, $E(v_k^2) = \sigma^2$.

Varianza Desconocida

Definimos (resto de suposiciones igual):

- Media muestral: $\bar{y} := \frac{1}{n}\sum_{i=1}^{n} y_i$
- Desv. típica muestral: $\hat{\sigma} := \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2}$

Definimos:

$$Z = \frac{(\bar{y} - \mu)}{\hat{\sigma}/\sqrt{n}}$$

Entonces, si la media "real" fuera μ, Z tendría sería una variable aleatoria con **distribución "t" de $(n-1)$ grados de libertad**. Su intervalo de confianza (unilateral/bilateral) a nivel $1-\alpha$ daría el criterio de rechazo.

Ejemplo: Con 20 muestras, 5% de nivel de significación estadístico (confianza 95%), Matlab obtiene:

```
x = tinv(0.95,19), ans=1.729
```

con lo que $Z \geq 1.729$ **rechazaría** que la media fuera menor que μ. Con infinitas muestras, sale el valor igual a σ conocido (1.645).

Conclusiones

- Con muestras de un "ruido blanco" de varianza conocida puede determinarse si su media es [igual/por encima/por debajo] de un cierto valor μ (distrib. normal).
- El nivel de confianza (1-probabilidad de falsa alarma) alto aumenta el número de muestras necesario para detectarlo...
- También con varianza desconocida se puede resolver (distribución T), aunque al ser la varianza un "estimado muestral" el test es más conservativo (rechazamos menos veces con los mismos datos).
- Si el proceso no es estacionario pero las varianzas/correlaciones temporales (modelo) que lo excitan son conocidas, el desarrollo se convierte en un Filtro de Kalman.
- Con varianzas desconocidas y proceso no estacionario, uff... se complica mucho.