Tests Estadísticos para Comparar Recomendaciones

IIC 3633 - Sistemas Recomendadores

Denis Parra
Profesor Asistente, DCC, PUC CHile

TOC

En esta clase

1. Significancia Estadistica de los Resultados
    - T-test
    - Signed test
    - Wilcoxon

Comparando Métricas de Performance entre Recomendadores

- Hipótesis nula (H0): No existe diferencia entre la media métrica de performance (RMSE, MAP, nDCG, etc.) del recomendador R_1 versus el recomendador R_2.

$$H_0 : \bar{metrica}_{R_1} = \bar{metrica}_{R_2}$$

- Hipótesis alternativa (H1): Si existe diferencia

$$H_1 : \bar{metrica}_{R_1} \neq \bar{metrica}_{R_2}$$

- Opciones de Test para chequear si $rechazamos$ o $fallamos\ en\ rechazar$ la hipótesis nula H_0
  - T-test (paired y not paired): test paramétrico, válido bajo ciertos supuestos
  - Signed y Wilcoxon: No paramétrico, no requiere los supuestos del T-test pero tiene menos poder (en el sentido estadistico)
- Debemos definir un nivel de significacion α, por lo general se rechaza la hipotesis nula con $p\text{-}value < 0,05$.

Supuestos del T-test

- Variable Bivariada independiente (grupos A, B)
- Variable dependiente continua (MAP, precision, recall, etc.)
- Cada observación de la variable es independiente de las otras observaciones:
  - El MAP de un usuario es independiente del MAP de otro usuario
  - En el t-test pareado, requerimos sólo las diferencias de pares ($A_i - B_i$) que sean independientes
- La variable dependiente tiene una distribución normal, con la misma varianza σ^2 en cada grupo (como si la distribución del grupo A y del grupo B fueran la misma, pero una desplazada respecto de la otra, sin cambiar de forma)

** REF: http://www.csic.cornell.edu/Elrod/t-test/t-test-assumptions.html

Ejemplo 1: T-Test

```
# Datasets de prueba
# lista de MAP para recomendador 1, con 30 usuarios, media de 0.2 y desv. st. de 0.1
rec1_map <- rnorm(30, mean = 0.2, sd = 0.1)

# lista de MAP para recomendador 1, con 30 usuarios, media de 0.2 y desv. st. de 0.1
rec2_map <- rnorm(30, mean = 0.4, sd = 0.15)

summary(rec1_map)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0672  0.1330  0.1970  0.2040  0.2810  0.3930
```

```
summary(rec2_map)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.113   0.277   0.386   0.399   0.553   0.693
```

5/11

Grafico de las distribuciones

```
# Graficos
plot(density(rec1_map), col=2)
lines(density(rec2_map), col=3)
```

T-test de Muestras Independientes

- Revisamos si el p-value es menor de 0.05 (nuestro α $level$)

```
# Independent samples T-test
t.test(rec1_map,rec2_map)
```

```
##
##  Welch Two Sample t-test
##
## data:  rec1_map and rec2_map
## t = -5.56, df = 45.2, p-value = 1.38e-06
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.2665 -0.1248
## sample estimates:
## mean of x mean of y
##    0.2037    0.3994
```

T-test de Pares

- Tiene mayor poder en términos estadísticos: La probablidad de encontrar un efecto, dado que existe, es mayor que en un t-test de muestras independientes.

```
# Paired samples T-test
t.test(rec1_map,rec2_map,paired=TRUE )
```

```
##
##  Paired t-test
##
## data:  rec1_map and rec2_map
## t = -4.984, df = 29, p-value = 2.654e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.2759 -0.1154
## sample estimates:
## mean of the differences
##                 -0.1956
```

8/11

Tests alternativos no-paramétricos

Cuando no se cumplen los supuestos (normalidad) y no se puede hacer alguna corrección o relajo de ellos, debemos usar alternativas (que usualmente tienen menos poder estadístico)

- Wilcoxon rank sum test (no es el mismo que signed rank test)
- Wilcoxon Signed Rank Test: Para datos pareados

Wilcoxon Rank Sum Test

- También llamado Mann-Whitney U, Wilcoxin-Mann-Whitney test, o Wilcoxin rank sum test.
- Consiste en calcular la métrica U basada en rankear las observaciones luego de mezclar ambas muestras.

```
wilcox.test(rec1_map,rec2_map)
```

```
##
##  Wilcoxon rank sum test
##
## data:  rec1_map and rec2_map
## W = 147, p-value = 2.446e-06
## alternative hypothesis: true location shift is not equal to 0
```

Wilcoxon Signed-Rank test

- Se basa en calcular diferencias entre pares
- La estadística de test corresponde al número de diferencias positivos o negativas
- H_0: la mediana de las diferencias entre pares es igual a zero

```
wilcox.test(rec1_map,rec2_map, paired=TRUE)
```

```
##
##  Wilcoxon signed rank test
##
## data:  rec1_map and rec2_map
## V = 45, p-value = 3.05e-05
## alternative hypothesis: true location shift is not equal to 0
```