

# Tests Estadísticos para Comparar Recomendaciones

IIC 3633 - Sistemas Recomendadores

Denis Parra



# TOC

## En esta clase

1. Significancia Estadística de los Resultados
  - T-test -- Tails -- Comparación Múltiple: correcciones
  - Signed test
  - Wilcoxon
2. Tests a grupos
  - ANOVA
  - Kruskal–Wallis
3. ¿Cómo reproducir resultados de papers?
4. Demostraciones interactivas

# Antes de empezar

## 1. Métricas de predicción vistas la clase anterior

- RMSE, MSE, MAE
- Precision, Recall, F-1
- MRR
- AP, MAP
- nDCG
- [Pending] Se mencionó Kendall-Tau y Spearman Rank Correlation

## 2. Otras métricas [Pending]

- Diversity (Ziegler)
- Lathia's Diversity (over time)
- MPR (for implicit feedback)

# Comparando Métricas de Performance entre Recomendadores

- Hipótesis nula ( $H_0$ ): No existe diferencia entre la media métrica de performance (RMSE, MAP, nDCG, etc.) del recomendador  $R_1$  versus el recomendador  $R_2$ .

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

- Hipótesis alternativa ( $H_1$ ): Si existe diferencia

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

- 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 estadístico )
- Debemos definir un nivel de significancia estadística  $\alpha$ , por lo general se rechaza la hipótesis 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  $\sigma^2$  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 2, con 30 usuarios, media de 0.4 y desv. st. de 0.15
rec2_map <- rnorm(30, mean = 0.4, sd = 0.15)

summary(rec1_map)
```

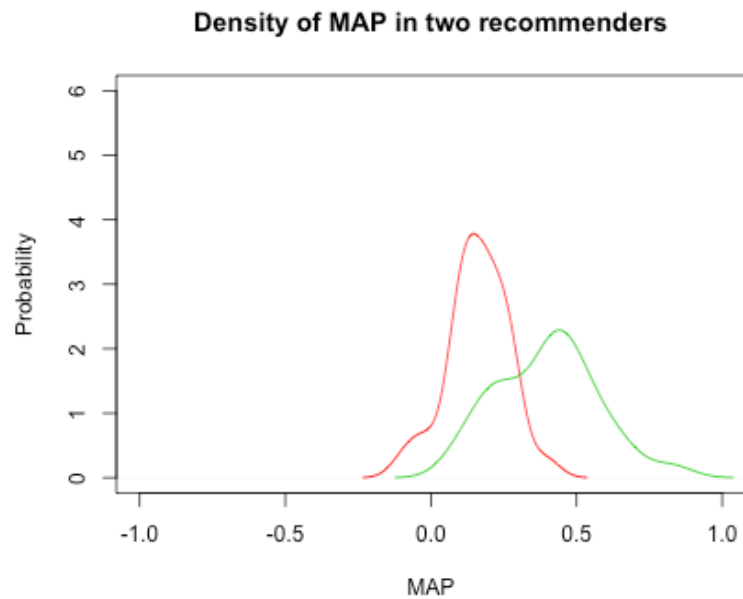
```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -0.09290  0.09766  0.16842  0.16218  0.23658  0.39594
```

```
summary(rec2_map)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.08946  0.26800  0.39900  0.39948  0.47719  0.82629
```

# Gráfico de las distribuciones

```
# Graficos
plot(density(rec1_map), col=2, main="Density of MAP in two recommenders",
     xlab="MAP", ylab="Probability",
     xlim=c(-1, 1), ylim=c(0, 6))
lines(density(rec2_map), col=3)
```



# T-test de Muestras Independientes

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

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

```
##  
## Paired t-test  
##  
## data:  rec1_map and rec2_map  
## t = -5.5585, df = 29, p-value = 5.38e-06  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -0.3246170 -0.1499887  
## sample estimates:  
## mean of the differences  
## -0.2373029
```



# T-test de Múltiples Pares

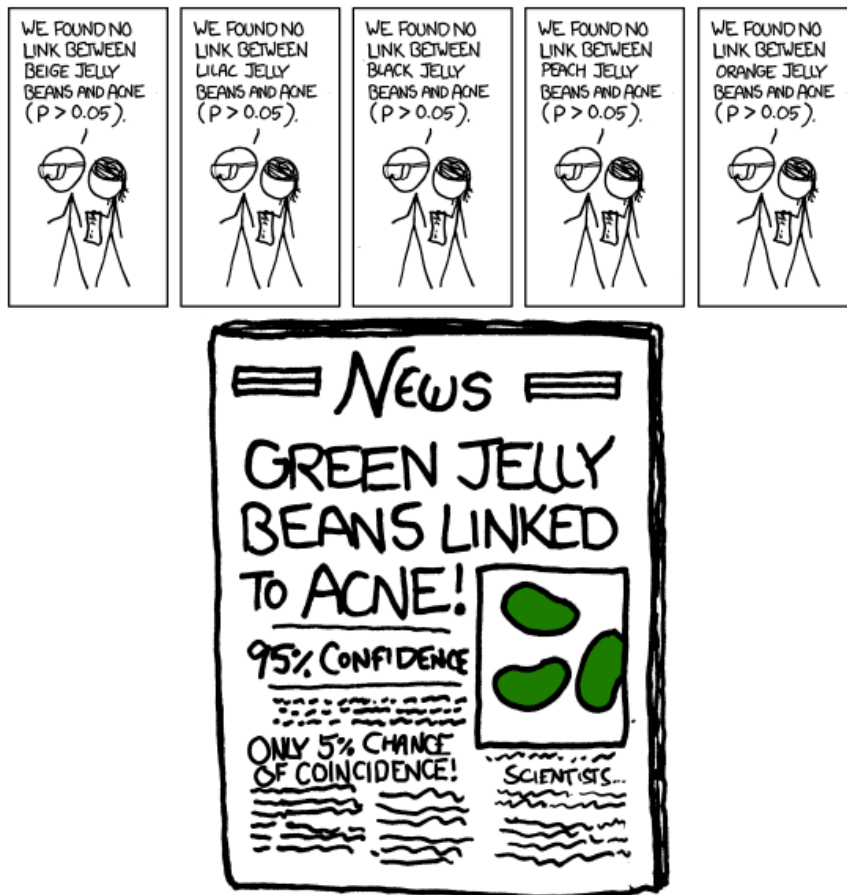
- En este caso es importante hacer alguna corrección, ya que al momento de

```
# multiple pair-wise t-test  
t.test(rec1_map,rec2_map,paired=TRUE )
```

```
##  
## Paired t-test  
##  
## data:  rec1_map and rec2_map  
## t = -5.5585, df = 29, p-value = 5.38e-06  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
##  -0.3246170 -0.1499887  
## sample estimates:  
## mean of the differences  
##          -0.2373029
```

# T-test

- xkcd: [Significant](#)



# T-test de Múltiples Pares

- En este caso es importante hacer alguna corrección, ya que al momento de

```
rec3_map <- rnorm(30, mean = 0.45, sd = 0.1)
library(reshape2)
df1 <- melt(cbind(rec1_map, rec2_map, rec3_map))
df1$Var2 <- factor(df1$Var2)
# Paired samples T-test
pairwise.t.test(df1$value, df1$Var2, p.adj="bonf")
```

```
##
## Pairwise comparisons using t tests with pooled SD
##
## data: df1$value and df1$Var2
##
##      rec1_map rec2_map
## rec2_map 1.7e-09  -
## rec3_map 2.2e-12  0.47
##
## P value adjustment method: bonferroni
```

# 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 = 114, p-value = 1.018e-07  
## 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 = 29, p-value = 3.239e-06  
## alternative hypothesis: true location shift is not equal to 0
```

# ANOVA

- Técnica de análisis estadístico que permite identificar si en un grupo ( $n_{\text{grupos}} > 2$ ) hay diferencias significativas.
- ANOVA no indica quién produce las diferencias, para eso necesitamos comparaciones post-hoc (multiple t-tests, por ejemplo)
- Hay distintas versiones de ANOVA según el diseño del experimento (within o between subjects) y según las combinaciones de grupos a comparar (1-way, 2-way, etc.)
- Kruskal-Wallis es la alternativa no-paramétrica de ANOVA

# Referencias

- Ziegler, C. N., McNee, S. M., Konstan, J. A., & Lausen, G. (2005, May). Improving recommendation lists through topic diversification. In Proceedings of the 14th international conference on World Wide Web (pp. 22-32). ACM.
- Lathia, N., Hailes, S., Capra, L., & Amatriain, X. (2010, July). Temporal diversity in recommender systems. In Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval (pp. 210-217). ACM.
- Hu, Y., Koren, Y., & Volinsky, C. (2008, December). Collaborative filtering for implicit feedback datasets. In Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on (pp. 263-272). IEEE.