7. ניתוח שונות דו-כיווני (המשך) ותרגיל מסכם

ניתוח נתונים סטטיסטי · הרצאה 7

ד״ר עדי שריד

2026-11-22

על מה נדבר היום

- **השלמת ניתוח שונות דו-כיווני** (two-way ANOVA):
  - המודל, פירוק השונות, טבלת ANOVA וחישוב ידני
  - אפקטים פשוטים ובדיקת ההנחות
  - בלוקים, אפקטים אקראיים ומודל היררכי (nested)
- **תרגיל מסכם מקצה לקצה:** איך לתמחר ספרי שמע לפי סוגה. כל מה שלמדנו עד עכשיו, בשאלה אחת בסגנון הבחינה

חזרה: איפה עצרנו

ניתוח דו-כיווני: השאלה

צוות ה-AI במוקד בונה בוט שעונה ללקוחות. שתי החלטות פתוחות:

- **גודל המודל** (פקטור A): קטן, בינוני או גדול. גדול מדויק יותר, אבל יקר
- **RAG** (פקטור B): לשלוף מסמכי מדיניות ולהכניס אותם לפרומפט, או לא

הניסוי: כל אחד מ-$3 \times 2 = 6$ השילובים רץ 4 פעמים על סט הערכה. התגובה: אחוז התשובות הנכונות. ניסוי עם כל השילובים נקרא **ניסוי עצרתי** (factorial design).

- **אפקט ראשי** (main effect): ההשפעה של פקטור אחד, בממוצע על פני הרמות של השני
- **אינטראקציה** (interaction): ההשפעה של פקטור אחד **תלויה ברמה** של השני
- **בגרף אינטראקציה:** קווים מקבילים = אין אינטראקציה

הנתונים

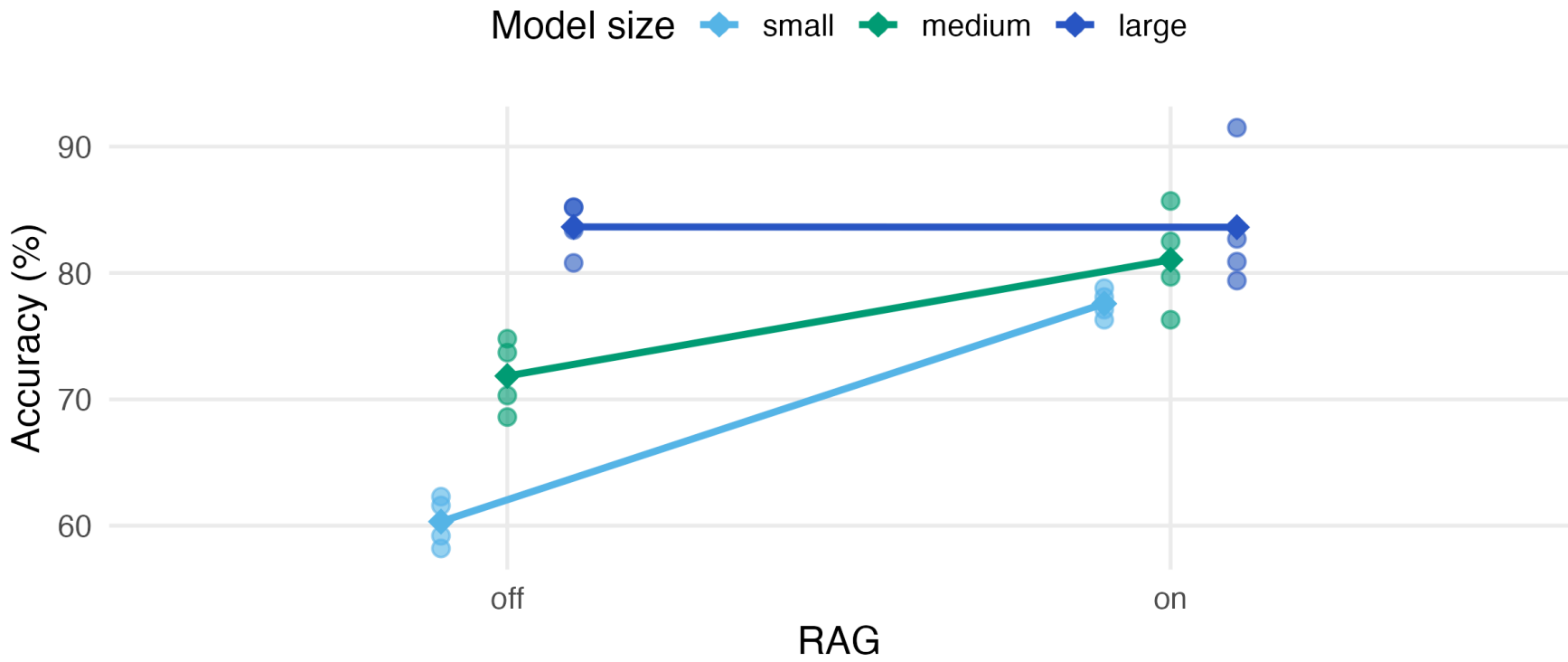
```
set.seed(606)
rag <- tribble(~size, ~retrieval, ~m,
  "small", "off", 62, "small", "on", 78, "medium", "off", 71,
  "medium", "on", 82, "large", "off", 80, "large", "on", 84) |>
  expand_grid(run = 1:4) |>
  mutate(acc = round(m + rnorm(n(), 0, 3), 1),
         size = factor(size, levels = c("small", "medium", "large")),
         retrieval = factor(retrieval, levels = c("off", "on"))) |>
  select(size, retrieval, run, acc)
rag |> pivot_wider(names_from = run, values_from = acc, names_prefix = "run ")
```

```
# A tibble: 6 × 6
  size   retrieval `run 1` `run 2` `run 3` `run 4`
  <fct>  <fct>       <dbl>   <dbl>   <dbl>   <dbl>
1 small  off          62.3    59.2    58.2    61.6
2 small  on           76.3    77.1    78.1    78.8
3 medium off          73.7    74.8    68.6    70.3
4 medium on           76.3    82.5    85.7    79.7
5 large  off          85.2    85.2    83.4    80.8
6 large  on           82.7    79.4    91.5    80.9
```

הנתונים בתרשים

הוספת RAG עוזרת מאוד למודל הקטן, ולא עוזרת לגדול

נקודה: ריצה אחת (200 שאלות) · מעוין וקו: ממוצע התא



- הקווים לא מקבילים: נראה שיש אינטראקציה. עכשיו נבנה את המודל ונבדוק אותה בצורה פורמלית

ניתוח שונות דו-כיווני: המודל והמבחנים

סימון

פקטור A עם a רמות (גודל המודל, $a=3$), פקטור B עם b רמות (RAG, $b=2$), n **חזרות בכל תא** ($n=4$).
$N = abn = 24$.

	B_1	$\cdots$	B_b	**סכום**
A_1	$y_{111}, \ldots, y_{11n}$	$\cdots$	$y_{1b1}, \ldots, y_{1bn}$	$y_{1\cdot\cdot}$
$\vdots$		$y_{ij\cdot}$		
A_a	$y_{a11}, \ldots, y_{a1n}$	$\cdots$	$y_{ab1}, \ldots, y_{abn}$	$y_{a\cdot\cdot}$
סכום	$y_{\cdot 1\cdot}$	$\cdots$	$y_{\cdot b\cdot}$	$y_{\cdots}$

y_{ijk}: התצפית ה-k ברמה i של A וברמה j של B. **נקודה במקום אינדקס = סכום על האינדקס הזה**:

$$y_{ij\cdot} = \sum_{k=1}^{n} y_{ijk} \qquad y_{i\cdot\cdot} = \sum_{j=1}^{b}\sum_{k=1}^{n} y_{ijk} \qquad y_{\cdot j\cdot} = \sum_{i=1}^{a}\sum_{k=1}^{n} y_{ijk} \qquad y_{\cdots} = \sum_{i,j,k} y_{ijk}$$

והממוצעים: $\bar{y}_{ij\cdot} = y_{ij\cdot}/n$, $\bar{y}_{i\cdot\cdot} = y_{i\cdot\cdot}/(bn)$, $\bar{y}_{\cdot j\cdot} = y_{\cdot j\cdot}/(an)$, $\bar{y}_{\cdots} = y_{\cdots}/(abn)$.

המודל

$$y_{ijk} = \mu + \tau_i + \beta_j + (\tau\beta)_{ij} + \varepsilon_{ijk} \qquad \varepsilon_{ijk} \overset{iid}{\sim} N(0, \sigma^2)$$

האיבר	המשמעות	אילוץ
μ	התוחלת הכללית	
τ_i	האפקט של רמה i של A (גודל המודל)	$\sum_i \tau_i = 0$
β_j	האפקט של רמה j של B (RAG)	$\sum_j \beta_j = 0$
$(\tau\beta)_{ij}$	**האינטראקציה:** התוספת של השילוב (i,j), מעבר לסכום האפקטים	$\sum_i (\tau\beta)_{ij} = 0$ לכל j, $\sum_j (\tau\beta)_{ij} = 0$ לכל i

- התוחלת בתא: $\mu_{ij} = \mu + \tau_i + \beta_j + (\tau\beta)_{ij}$
- **בלי אינטראקציה** ($(\tau\beta)_{ij} = 0$): $\mu_{ij} = \mu + \tau_i + \beta_j$. המודל **אדיטיבי**, וזה בדיוק "קווים מקבילים"

שלוש מערכות השערות

A:	$H_0 : \tau_1 = \cdots = \tau_a = 0$	$H_1 : \tau_i \neq 0$ עבור i כלשהו
B:	$H_0 : \beta_1 = \cdots = \beta_b = 0$	$H_1 : \beta_j \neq 0$ עבור j כלשהו
AB:	$H_0 : (\tau\beta)_{ij} = 0$ לכל i, j	$H_1 : (\tau\beta)_{ij} \neq 0$ עבור i, j כלשהם

סדר הקריאה: קודם האינטראקציה.

- **אם האינטראקציה מובהקת:** האפקטים הראשיים הם ממוצע של אפקטים שונים, ולכן הם לא מספרים את כל הסיפור. בודקים את ההשפעה של A **בכל רמה** של B בנפרד
- **אם היא לא מובהקת:** אפשר לפרש את האפקטים הראשיים ישירות

פירוק השונות

כמו בחד-כיווני, מפרקים כל סטייה מהממוצע הכללי לחלקים:

$$y_{ijk} - \bar{y}_{\cdots} = \underbrace{(\bar{y}_{i\cdot\cdot} - \bar{y}_{\cdots})}_{\hat{\tau}_i} + \underbrace{(\bar{y}_{\cdot j\cdot} - \bar{y}_{\cdots})}_{\hat{\beta}_j} + \underbrace{(\bar{y}_{ij\cdot} - \bar{y}_{i\cdot\cdot} - \bar{y}_{\cdot j\cdot} + \bar{y}_{\cdots})}_{\widehat{(\tau\beta)}_{ij}} + \underbrace{(y_{ijk} - \bar{y}_{ij\cdot})}_{e_{ijk}}$$

מעלים בריבוע וסוכמים. כל המכפלות המעורבות מתאפסות (בניסוי מאוזן), ונשאר:

$$SS_T = \underbrace{bn\sum_i(\bar{y}_{i\cdot\cdot} - \bar{y}_{\cdots})^2}_{SS_A} + \underbrace{an\sum_j(\bar{y}_{\cdot j\cdot} - \bar{y}_{\cdots})^2}_{SS_B} + \underbrace{n\sum_{i,j}(\bar{y}_{ij\cdot} - \bar{y}_{i\cdot\cdot} - \bar{y}_{\cdot j\cdot} + \bar{y}_{\cdots})^2}_{SS_{AB}} + \underbrace{\sum_{i,j,k}(y_{ijk} - \bar{y}_{ij\cdot})^2}_{SS_E}$$

דרגות חופש: $abn - 1 = (a-1) + (b-1) + (a-1)(b-1) + ab(n-1)$

- **האינטראקציה:** כמה התא סוטה ממה שהמודל האדיטיבי צופה, $\bar{y}_{i\cdot\cdot} + \bar{y}_{\cdot j\cdot} - \bar{y}_{\cdots}$
- **השגיאה:** הפיזור **בתוך** כל תא. לכן צריך $n \geq 2$ כדי לאמוד אותה

טבלת ANOVA

מקור	סכום ריבועים	דרגות חופש	ממוצע ריבועים	$E(MS)$	F_0
A	SS_A	$a-1$	MS_A	$\sigma^2 + \frac{bn\sum \tau_i^2}{a-1}$	MS_A/MS_E
B	SS_B	$b-1$	MS_B	$\sigma^2 + \frac{an\sum \beta_j^2}{b-1}$	MS_B/MS_E
AB	SS_{AB}	$(a-1)(b-1)$	MS_{AB}	$\sigma^2 + \frac{n\sum\sum (\tau\beta)_{ij}^2}{(a-1)(b-1)}$	MS_{AB}/MS_E
שגיאה	SS_E	$ab(n-1)$	MS_E	σ^2	
סך הכול	SS_T	$abn-1$			

- אותו היגיון כמו בחד-כיווני: תחת H_0 המונה והמכנה אומדים את σ^2, ו-$F_0 \sim F$ עם דרגות החופש שלהם
- למשל, דוחים H_0 לאינטראקציה אם $F_0 > F_{1-\alpha;\,(a-1)(b-1),\,ab(n-1)}$

נוסחאות לחישוב ידני

עובדים עם **סכומים**. מגדירים את **גורם התיקון** $C = \frac{y_{...}^2}{abn}$:

$$SS_T = \sum_{i,j,k} y_{ijk}^2 - C$$

$$SS_A = \frac{1}{bn}\sum_{i} y_{i..}^2 - C \qquad SS_B = \frac{1}{an}\sum_{j} y_{.j.}^2 - C$$

$$SS_{\text{cells}} = \frac{1}{n}\sum_{i,j} y_{ij.}^2 - C$$

$$SS_{AB} = SS_{\text{cells}} - SS_A - SS_B$$

$$SS_E = SS_T - SS_{\text{cells}}$$

- **הכלל:** סכום בריבוע, חלקי מספר התצפיות שבסכום, פחות C
- SS_{cells}: סכום הריבועים **בין התאים**. מאיפה הוא בא, ולמה מחסירים ממנו את SS_A ואת SS_B כדי לקבל את האינטראקציה: בשני השקפים הבאים

מאיפה בא SS_{cells}?

צעד 1: שוכחים לרגע שיש שני פקטורים. כל צירוף של גודל ו-RAG הוא "טיפול" נפרד: 6 קבוצות של 4 ריצות. זה ניתוח חד-כיווני רגיל, עם $a \cdot b = 6$ רמות:

$$SS_T = \underbrace{n \sum_{i,j} (\bar{y}_{ij\cdot} - \bar{y}_{\cdots})^2}_{SS_{\text{cells}} = SS_{Tr}} + \underbrace{\sum_{i,j,k} (y_{ijk} - \bar{y}_{ij\cdot})^2}_{SS_E}$$

SS_{cells} הוא ה-SS_{Tr} של הניתוח הזה: **כל** מה שמבדיל בין התאים. מה שנשאר הוא הפיזור בתוך התאים, ולכן $SS_E = SS_T - SS_{\text{cells}}$.

צעד 2: למה התאים שונים זה מזה? יש שלוש סיבות אפשריות: הגודל (A), ה-RAG (B), או הצירוף הספציפי של שניהם (AB). זה אותו פירוק כמו בשקף "פירוק השונות", בלי השגיאה:

$$\bar{y}_{ij\cdot} - \bar{y}_{\cdots} = \hat{\tau}_i + \hat{\beta}_j + \widehat{(\tau\beta)}_{ij} \quad \Longrightarrow \quad SS_{\text{cells}} = SS_A + SS_B + SS_{AB}$$

צעד 3: האינטראקציה היא מה שנשאר. אם אין אינטראקציה, כל ממוצע תא הוא בדיוק "ממוצע כללי + אפקט שורה + אפקט עמודה", והקווים בתרשים מקבילים. SS_{AB} מודד כמה התאים **רחוקים מהתחזית האדיטיבית** הזו.

האינטראקציה במספרים

Gap (interaction)	Additive prediction	Actual mean	RAG	Size
-4.2	64.5	60.3	off	small
4.2	73.4	77.6	on	small
-0.2	72.0	71.8	off	medium
0.2	80.9	81.0	on	medium
4.4	79.2	83.7	off	large
-4.4	88.0	83.6	on	large

- **התחזית האדיטיבית** היא $\bar{y}_{i\cdot\cdot} + \bar{y}_{\cdot j\cdot} - \bar{y}_{\cdots}$. זה מה שהיינו מצפים לו אם RAG היה מוסיף אותו דבר לכל גודל
- **הפער** הוא $\widehat{(\tau\beta)}_{ij}$. למודל הקטן RAG עוזר **יותר** מהממוצע (פער חיובי עם RAG), ולגדול **פחות** (פער שלילי)
- $SS_{\text{cells}} - SS_A - SS_B$ בחישוב הידני $SS_{AB} = n\sum_{i,j}(\text{gap}_{ij})^2 = 4 \cdot 74.72 = 298.9$. אותו מספר שיוצא מ-

חישוב ידני: טבלת הסכומים

```
cell_tot <- rag |> group_by(size, retrieval) |> summarise(y_ij = sum(acc), .groups = "drop")
cell_tot |>
  pivot_wider(names_from = retrieval, values_from = y_ij) |>
  mutate(y_i = off + on)
```

```
# A tibble: 3 × 4
  size     off    on   y_i
  <fct>  <dbl> <dbl> <dbl>
1 small   241.  310.  552.
2 medium  287.  324.  612.
3 large   335.  334.  669.
```

```
c(y_off = sum(rag$acc[rag$retrieval == "off"]), y_on = sum(rag$acc[rag$retrieval == "on"]),
  y_total = sum(rag$acc), sum_y2 = sum(rag$acc^2))
```

```
   y_off     y_on  y_total   sum_y2
   863.3    969.0   1832.3 141705.2
```

חישוב ידני: סכומי הריבועים

$$C = \frac{1,832.3^2}{24} = 139,888.5 \qquad SS_T = 141,705.17 - C = 1,816.7$$

$$SS_A = \frac{551.6^2 + 611.6^2 + 669.1^2}{2 \cdot 4} - C = 863.0 \qquad SS_B = \frac{863.3^2 + 969.0^2}{3 \cdot 4} - C = 465.5$$

$$SS_{\text{cells}} = \frac{241.3^2 + 310.3^2 + 287.4^2 + 324.2^2 + 334.6^2 + 334.5^2}{4} - C = 1,627.4$$

$$SS_{AB} = 1,627.4 - 863.0 - 465.5 = 298.9 \qquad SS_E = 1,816.7 - 1,627.4 = 189.3$$

מקור	SS	df	MS	F_0	$F_{0.95}$
גודל (A)	863.0	2	431.5	41.04	3.55
RAG (B)	465.5	1	465.5	44.27	4.41
AB	298.9	2	149.4	14.21	3.55
שגיאה	189.3	18	10.52		

אותו ניתוח ב-R

```r
fit_rag <- aov(acc ~ size * retrieval, data = rag)   # size * retrieval = size + retrieval + size:retrieval
summary(fit_rag)
```

```
              Df Sum Sq Mean Sq F value Pr(>F)
size           2    863     432    41.0  2e-07 ***
retrieval      1    466     466    44.3  3e-06 ***
size:retrieval 2    299     149    14.2  2e-04 ***
Residuals     18    189      11
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- **`size:retrieval`** היא שורת האינטראקציה. $F_0 = 14.21$, $p < 0.001$: **יש אינטראקציה**
- גם שני האפקטים הראשיים מובהקים, אבל בגלל האינטראקציה הם לא מספרים את כל הסיפור
- $MS_E = 10.52$, ולכן $\hat{\sigma} = 3.24$ נקודות אחוז: הרעש בין ריצות על מדגמי שאלות שונים

אומדים לאפקטים

$$\hat{\mu} = \bar{y}_{\cdots} \qquad \hat{\tau}_i = \bar{y}_{i\cdot\cdot} - \bar{y}_{\cdots} \qquad \hat{\beta}_j = \bar{y}_{\cdot j\cdot} - \bar{y}_{\cdots} \qquad \widehat{(\tau\beta)}_{ij} = \bar{y}_{ij\cdot} - \bar{y}_{i\cdot\cdot} - \bar{y}_{\cdot j\cdot} + \bar{y}_{\cdots} \qquad \hat{\sigma}^2 = MS_E$$

```
mu_hat <- mean(rag$acc)
eff <- rag |>
  group_by(size, retrieval) |> summarise(cell = mean(acc), .groups = "drop") |>
  group_by(size) |> mutate(row_mean = mean(cell)) |>
  group_by(retrieval) |> mutate(col_mean = mean(cell)) |> ungroup() |>
  mutate(tau = row_mean - mu_hat, beta = col_mean - mu_hat,
         tau_beta = cell - row_mean - col_mean + mu_hat)
eff |> select(size, retrieval, cell, tau, beta, tau_beta) |>
  mutate(across(where(is.numeric), \(x) round(x, 2)))
```

```
# A tibble: 6 × 6
  size   retrieval  cell   tau  beta tau_beta
  <fct>  <fct>     <dbl> <dbl> <dbl>    <dbl>
1 small  off        60.3 -7.4   -4.4    -4.22
2 small  on         77.6 -7.4    4.4     4.22
3 medium off        71.8  0.1   -4.4    -0.2
4 medium on         81.0  0.1    4.4     0.2
5 large  off        83.6  7.29  -4.4     4.42
6 large  on         83.6  7.29   4.4    -4.42
```

- עם RAG, המודל הקטן מרוויח **יותר** ממה שהמודל האדיטיבי צופה, והגדול **פחות**. לכן האינטראקציה של קטן עם RAG חיובית, $\widehat{(\tau\beta)}_{small,\,on} > 0$, ושל גדול עם RAG שלילית, $\widehat{(\tau\beta)}_{large,\,on} < 0$

כשהאינטראקציה מובהקת: אפקטים פשוטים

אפקט פשוט (simple effect): האפקט של RAG **בכל גודל מודל בנפרד**. כל השוואה משתמשת ב-MS_E של המודל המלא:

$$t_0 = \frac{\bar{y}_{i2\cdot} - \bar{y}_{i1\cdot}}{\sqrt{2MS_E/n}} \sim t_{ab(n-1)}$$

```
emm_rag <- emmeans::emmeans(fit_rag, ~ retrieval | size)
emmeans::contrast(emm_rag, "revpairwise")
```

```
size = small:
 contrast estimate   SE df t.ratio p.value
 on - off   17.250 2.29 18   7.523 <0.0001

size = medium:
 contrast estimate   SE df t.ratio p.value
 on - off    9.200 2.29 18   4.012  0.0008

size = large:
 contrast estimate   SE df t.ratio p.value
 on - off   -0.025 2.29 18  -0.011  0.9914
```

- **קטן:** RAG מוסיף כ-17 נקודות. **בינוני:** כ-9. **גדול:** 0
- SE זהה בשלושתם: $\sqrt{2 \cdot 10.52/4} = 2.29$

ומה עם מודל קטן עם RAG מול מודל גדול?

השאלה העסקית: האם מודל קטן עם RAG, הזול יותר, יכול להחליף את המודל הגדול? זו השוואה בין **תאים**, לא בין רמות של פקטור. לכן מתייחסים ל-6 התאים כאילו היו 6 קבוצות בניתוח חד-כיווני, ומריצים עליהם Tukey. הסף אחיד לכל 15 הזוגות:

$$HSD = q_{\alpha}(ab,\ f_E) \cdot \sqrt{\frac{MS_E}{n}}$$

רכיב	ערך	למה
מספר הממוצעים	$ab = 6$	כל התאים, לא $a = 3$
דרגות חופש לשגיאה	$f_E = ab(n-1) = 18$	מה-MS_E של המודל הדו-כיווני
n	4	ריצות בכל תא
$q_{0.05}(6, 18)$	4.49	טבלת studentized range, או `qtukey()`
$s_{\bar{y}} = \sqrt{MS_E/n}$	$\sqrt{10.52/4} = 1.62$	

נקודות אחוז $HSD = 4.49 \cdot 1.62 = 7.29$

- **כל** הפרש בין שני תאים שגדול מ-7.29 מובהק. קטן ממנו: לא מובהק

הסף האחיד: חישוב ב-R

```r
mse_rag <- deviance(fit_rag) / df.residual(fit_rag)          # MS_E = SS_E / 18
q_rag   <- qtukey(0.95, nmeans = 6, df = df.residual(fit_rag))
hsd_rag <- q_rag * sqrt(mse_rag / 4)
c(MSE = mse_rag, q = q_rag, s_ybar = sqrt(mse_rag / 4), HSD = hsd_rag)
```

```
   MSE      q s_ybar    HSD
10.515  4.494  1.621  7.287
```

```r
cells <- rag |> group_by(size, retrieval) |> summarise(mean = mean(acc), .groups = "drop") |>
  mutate(cell = paste(size, retrieval, sep = ":")) |> arrange(mean)
cells |> select(cell, mean)
```

```
# A tibble: 6 × 2
  cell        mean
  <chr>      <dbl>
1 small:off   60.3
2 medium:off  71.8
3 small:on    77.6
4 medium:on   81.0
5 large:on    83.6
6 large:off   83.6
```

- התאים מסודרים מהנמוך לגבוה. כל זוג שהמרחק ביניהם גדול מ-7.29 נבדל

Tukey על 6 התאים: הפלט המלא

```r
tk_rag <- TukeyHSD(fit_rag, "size:retrieval")$`size:retrieval`
round(tk_rag, 3)
```

```
                        diff     lwr    upr p adj
medium:off-small:off  11.525   4.238 18.812 0.001
large:off-small:off   23.325  16.038 30.612 0.000
small:on-small:off    17.250   9.963 24.537 0.000
medium:on-small:off   20.725  13.438 28.012 0.000
large:on-small:off    23.300  16.013 30.587 0.000
large:off-medium:off  11.800   4.513 19.087 0.001
small:on-medium:off    5.725  -1.562 13.012 0.177
medium:on-medium:off   9.200   1.913 16.487 0.009
large:on-medium:off   11.775   4.488 19.062 0.001
small:on-large:off    -6.075 -13.362  1.212 0.136
medium:on-large:off   -2.600  -9.887  4.687 0.861
large:on-large:off    -0.025  -7.312  7.262 1.000
medium:on-small:on     3.475  -3.812 10.762 0.659
large:on-small:on      6.050  -1.237 13.337 0.138
large:on-medium:on     2.575  -4.712  9.862 0.866
```

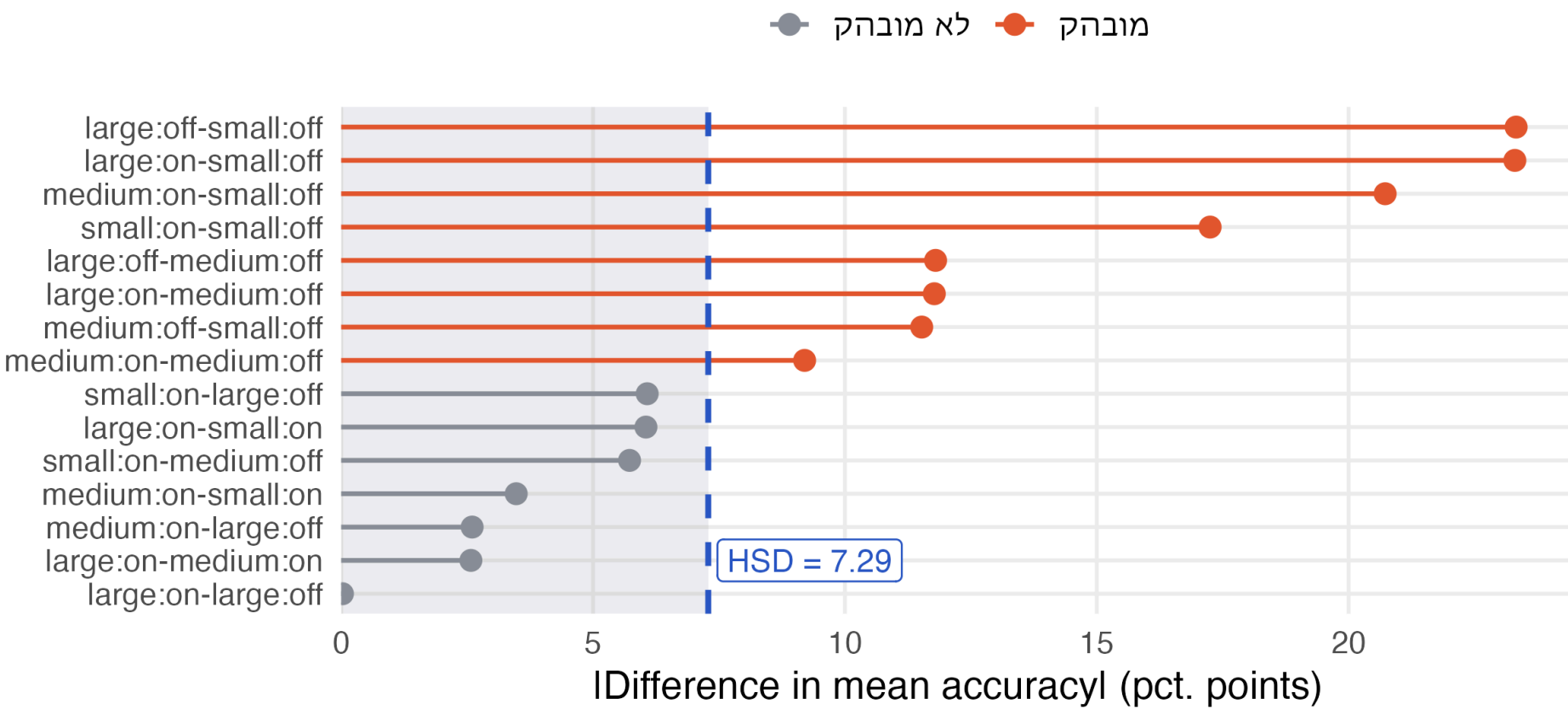
איך קוראים את הפלט

- **רוחב אחיד:** בכל שורה, $\text{upr} - \text{diff} = \text{diff} - \text{lwr} = 7.287$. זה ה-HSD. הניסוי מאוזן, ולכן לכל 15 הזוגות יש אותו רווח
- **שלוש דרכים לומר שהפרש מובהק:** $|\text{diff}| > HSD$, או שהרווח לא כולל את 0, או `p adj < 0.05`
- **10 מתוך 15 זוגות מובהקים.** 5 לא מובהקים, וכולם בין התאים "הטובים": קטן+RAG, בינוני+RAG, גדול עם RAG ובלי
- **קבוצות:**
  - **A:** קטן בלי RAG (60.3), נמוך מכולם
  - **B:** בינוני בלי RAG (71.8) וקטן+RAG (77.6). ההפרש ביניהם, 5.72, קטן מה-HSD
  - **C:** קטן+RAG, בינוני+RAG, גדול+RAG, גדול בלי RAG. כל ההפרשים ביניהם עד 6.08
- קטן+RAG נמצא גם ב-B וגם ב-C: הוא לא שונה מובהקות מאף אחד מהשכנים שלו

15 ההפרשים מול הסף האחיד

סרגל אחד לכל הזוגות: 10 מתוך 15 עוברים את ה-HSD

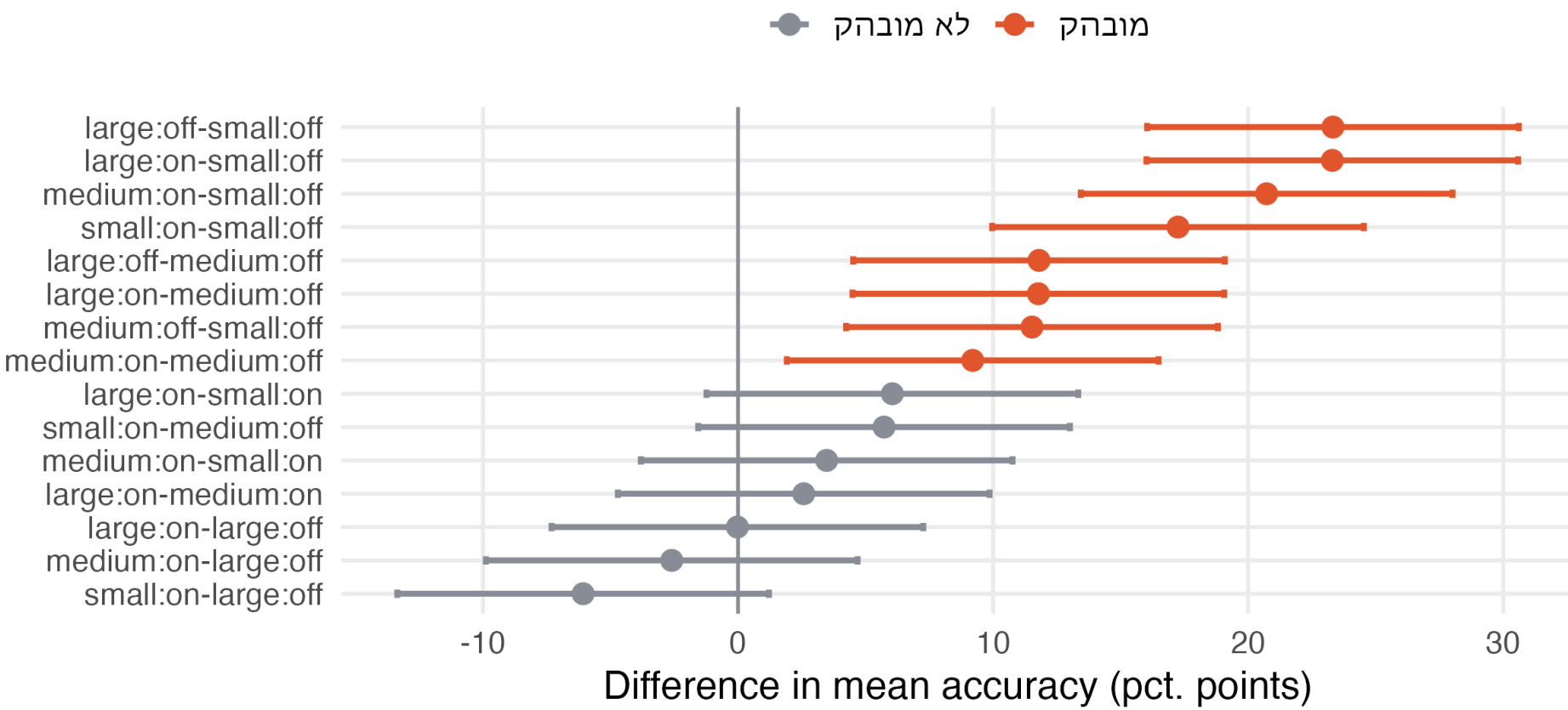
האזור האפור: הפרשים קטנים מה-HSD, כלומר לא מובהקים



רווחי הסמך המשותפים בתרשים

כל הרווחים באותו רוחב: ההפרש ± HSD

רווחי סמך משותפים 95% · רווח שלא חוצה את 0 הוא מובהק



בחזרה לשאלה: קטן+RAG מול גדול

```
round(tk_rag[c("small:on-large:off", "large:on-small:on"), ], 3)
```

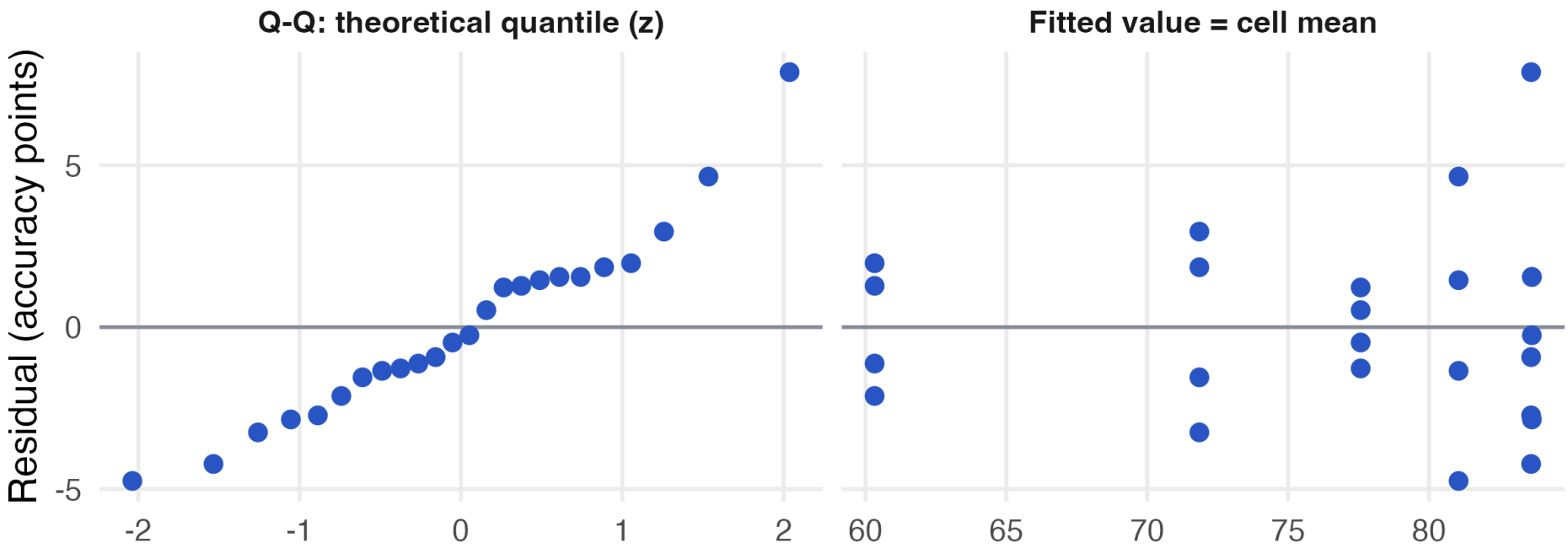
```
                     diff     lwr    upr p adj
small:on-large:off -6.075 -13.362  1.212 0.136
large:on-small:on   6.050  -1.237 13.337 0.138
```

- `small:on-large:off`: הקטן עם RAG מפגר אחרי הגדול בלי RAG בכ-6 נקודות. $HSD = 7.29 < 6.08$, ולכן **ההפרש לא מובהק** ($p = 0.14$)
- **אבל "לא מובהק" אינו "שווה".** רווח הסמך המשותף: מ-13 נקודות לטובת הגדול ועד נקודה אחת לטובת הקטן. ייתכן שהפער אפסי, וייתכן שהוא 13 נקודות
- **ההמלצה:** RAG לכל מודל קטן או בינוני. לפני שמחליפים את הגדול בקטן עם RAG, צריך ניסוי ממוקד עם יותר ריצות. עכשיו אפשר לחשב כמה (ניתוח עוצמה, הרצאה 4)

בדיקת ההנחות

אותן הנחות כמו בחד-כיווני, ואותם כלים. השאריות: $e_{ijk} = y_{ijk} - \bar{y}_{ij\cdot}$

השאריות קרובות לנורמליות, והפיזור דומה בתאים



- הנקודות ב-Q-Q קרובות לקו. הפיזור דומה ברוב התאים. Bartlett על 6 התאים: $p = 0.18$
- עם 4 תצפיות לתא, אי אפשר לראות הרבה. זה המצב הרגיל בניסויים, ולכן תכנון טוב (הגרלה, ריצות עצמאיות) חשוב מאוד

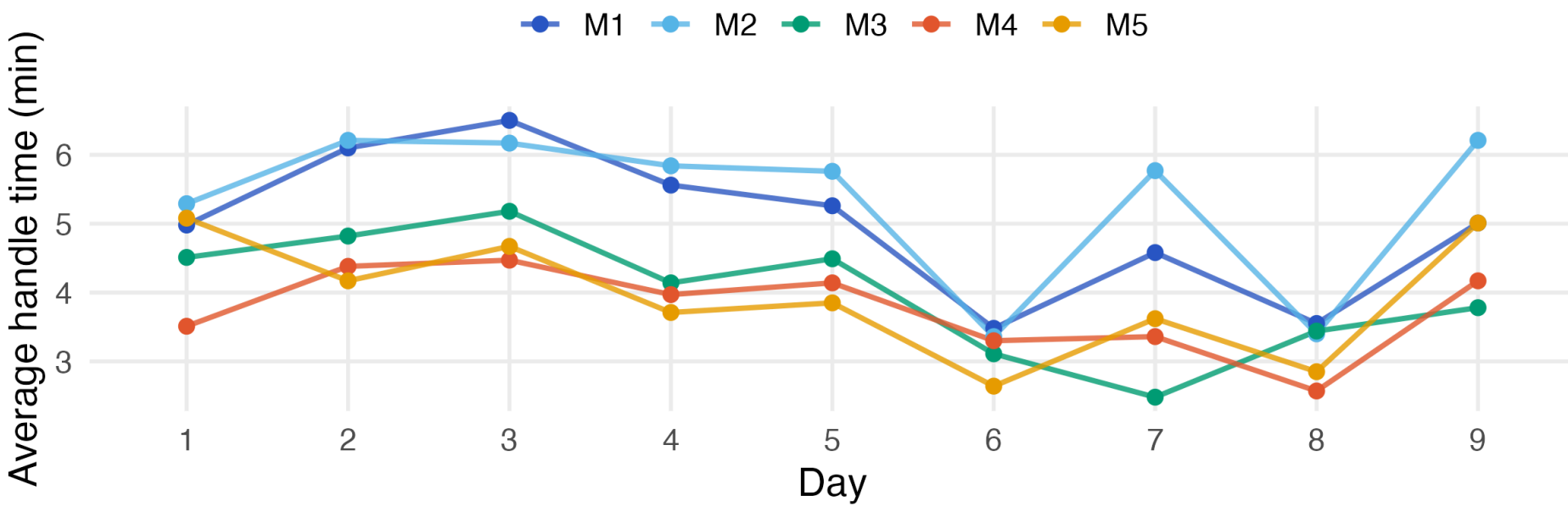
בלוקים: ניתוח דו-כיווני בלי חזרות

חזרה למוקד: אפקט היום

בהרצאה 4 ראינו שבימים 6 ו-8 הטיפול היה קצר **בכל המודלים**. כל יום הפעלנו את כל חמשת המודלים, על פניות אחרות.

הקווים עולים ויורדים יחד: יש ימים קלים וימים קשים

משך הטיפול הממוצע, לכל מודל בכל יום



- **היום הוא פקטור שני.** הוא לא מעניין אותנו כשלעצמו, אבל הוא מוסיף רעש. פקטור כזה נקרא **בלוק** (block)
- **בכל תא יש תצפית אחת** ($n = 1$): מודל אחד ביום אחד

המודל עם בלוק

$$y_{ij} = \mu + \tau_i + \beta_j + \varepsilon_{ij} \qquad i = 1, \ldots, a \text{ (מודל)}, \quad j = 1, \ldots, b \text{ (יום)}$$

- **אין איבר אינטראקציה.** עם $n = 1$ אין פיזור בתוך תא, ולכן אי אפשר להפריד אינטראקציה משגיאה. מניחים שהמודל **אדיטיבי**: יום עמוס מאריך את הטיפול בכל המודלים באותה מידה
- מה שהיה SS_{AB} משמש כשגיאה:

$$SS_E = SS_T - SS_A - SS_B \qquad df_E = (a-1)(b-1)$$

מקור	SS	df	F_0
מודל (A)	SS_A	$a-1$	MS_A/MS_E
יום (בלוק)	SS_B	$b-1$	
שגיאה	SS_E	$(a-1)(b-1)$	
סך הכול	SS_T	$ab-1$	

התוצאה: בלי בלוק ועם בלוק

```r
summary(aov(aht ~ model, data = calls_d))
```

```
            Df Sum Sq Mean Sq F value Pr(>F)
model        4   18.0    4.51    5.28 0.0016 **
Residuals   40   34.1    0.85
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
summary(aov(aht ~ model + day, data = calls_d))
```

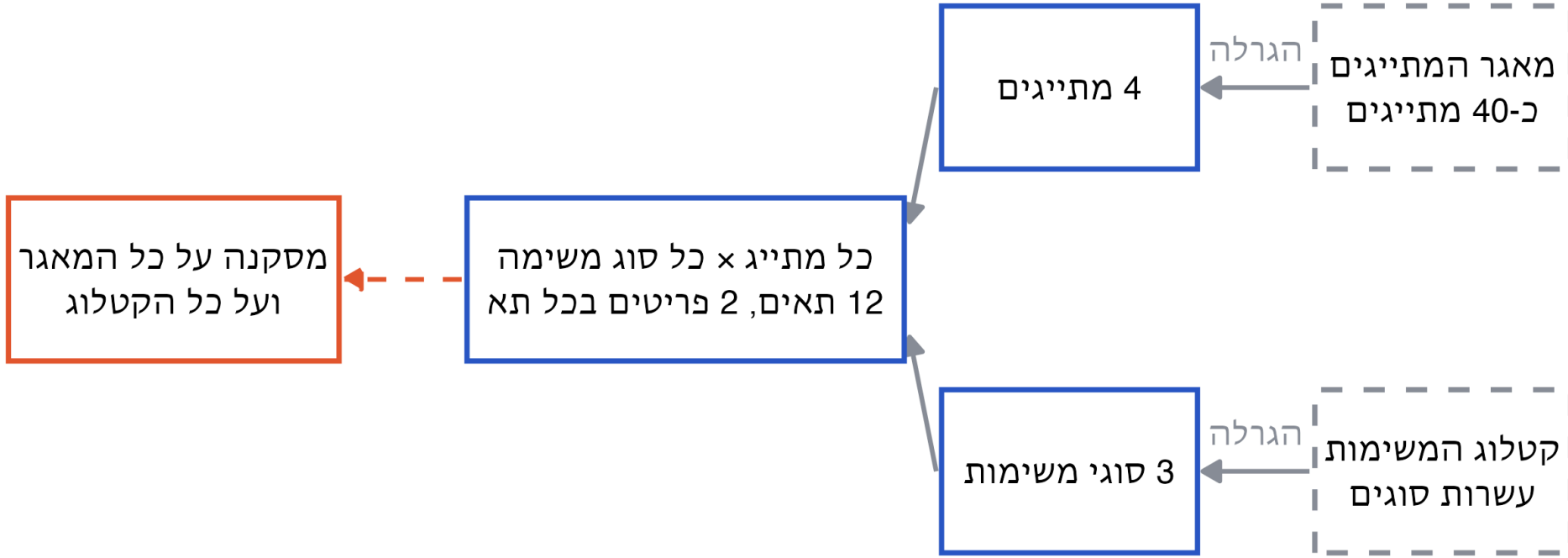
```
            Df Sum Sq Mean Sq F value  Pr(>F)
model        4   18.0    4.51    17.4 1.1e-07 ***
day          8   25.9    3.23    12.5 6.6e-08 ***
Residuals   32    8.3    0.26
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- $SS_{model} = 18.02$ **לא השתנה**. מה שהשתנה: 25.9 מתוך 34.1 של השגיאה הם בעצם הבדלים בין ימים
- MS_E ירד מ-0.854 ל-0.259, ו-F של המודלים קפץ מ-5.3 ל-17.4
- **המחיר:** הבלוק לוקח 8 דרגות חופש מהשגיאה (40 → 32). $MS_E = SS_E/df_E$, ולכן אם הבלוק לא היה מסביר כלום, היינו מחלקים כמעט אותו SS_E במספר קטן יותר, ו-MS_E היה **עולה**. כאן הבלוק הוריד את SS_E פי 4, ולכן המחיר זניח

אפקטים אקראיים בניתוח דו-כיווני

השאלה: מתייגים וסוגי משימות

צוות ה-AI משלם למתייגים (annotators) לתייג נתוני אימון. כמה זמן לוקח לתייג פריט, ומאיפה מגיעה השונות?



השאלה לא "האם מתייג A3 איטי מ-A1", אלא: **כמה מהשונות בזמן התיוג נובעת ממתייגים, כמה מסוגי משימות, וכמה מהשילוב?**

המודל

$$y_{ijk} = \mu + \tau_i + \beta_j + (\tau\beta)_{ij} + \varepsilon_{ijk}$$

$$\tau_i \sim N(0, \sigma_\tau^2) \qquad \beta_j \sim N(0, \sigma_\beta^2) \qquad (\tau\beta)_{ij} \sim N(0, \sigma_{\tau\beta}^2) \qquad \varepsilon_{ijk} \sim N(0, \sigma^2)$$

כולם בלתי תלויים. **ההשערות:** $H_0 : \sigma_\tau^2 = 0$, $H_0 : \sigma_\beta^2 = 0$, $H_0 : \sigma_{\tau\beta}^2 = 0$.

סכומי הריבועים ודרגות החופש זהים למודל הקבוע. מה שמשתנה הן התוחלות:

מקור	$E(MS)$, **אפקטים קבועים**	$E(MS)$, **אפקטים אקראיים**
A	$\sigma^2 + \frac{bn\sum \tau_i^2}{a-1}$	$\sigma^2 + n\sigma_{\tau\beta}^2 + bn\sigma_\tau^2$
B	$\sigma^2 + \frac{an\sum \beta_j^2}{b-1}$	$\sigma^2 + n\sigma_{\tau\beta}^2 + an\sigma_\beta^2$
AB	$\sigma^2 + \frac{n\sum\sum (\tau\beta)_{ij}^2}{(a-1)(b-1)}$	$\sigma^2 + n\sigma_{\tau\beta}^2$
שגיאה	σ^2	σ^2

מה מחלקים במה?

הכלל: המכנה של F הוא ה-MS שהתוחלת שלו **שווה לתוחלת המונה כש-H_0 נכונה**.

- **A:** תחת $H_0 : \sigma_\tau^2 = 0$, $E(MS_A) = \sigma^2 + n\sigma_{\tau\beta}^2 = E(MS_{AB})$. לכן:

$$F_A = \frac{MS_A}{MS_{AB}} \sim F_{a-1,\,(a-1)(b-1)} \qquad F_B = \frac{MS_B}{MS_{AB}} \sim F_{b-1,\,(a-1)(b-1)} \qquad F_{AB} = \frac{MS_{AB}}{MS_E}$$

- **האינטואיציה:** אם המתייגים שהוגרלו מגיבים אחרת לכל סוג משימה, הממוצעים שלהם "קופצים" גם בלי הבדל אמיתי ביניהם. הרעש הזה נכנס ל-MS_A, ולכן צריך להשוות אותו לרעש של האינטראקציה ולא רק לרעש בתוך תא

אומדים לרכיבי השונות (פותרים את משוואות התוחלת):

$$\hat{\sigma}^2 = MS_E \qquad \hat{\sigma}_{\tau\beta}^2 = \frac{MS_{AB} - MS_E}{n} \qquad \hat{\sigma}_\tau^2 = \frac{MS_A - MS_{AB}}{bn} \qquad \hat{\sigma}_\beta^2 = \frac{MS_B - MS_{AB}}{an}$$

ב-R: למה צריך להגיד לו מה אקראי

```r
fit_ann <- aov(sec ~ annotator * task, data = annot)
summary(fit_ann)
```

```
               Df Sum Sq Mean Sq F value  Pr(>F)
annotator       3    169      56   13.11 0.00043 ***
task            2   3315    1658  384.73 1.3e-11 ***
annotator:task  6     61      10    2.35 0.09812 .
Residuals      12     52       4
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

R לא יודע שהפקטורים אקראיים, אלא אם אומרים לו. הוא מחלק הכול ב-MS_E, ולכן ה-F של המתייגים ושל סוגי המשימות בפלט הזה **שגוי**. בשקף הבא נראה איך אומרים ל-R.

ב-R: הדרך הנכונה, עם `Error()`

`Error()` אומר ל-R **מול איזה איבר אקראי** לבדוק את האפקטים:

```
summary(aov(sec ~ annotator + task + Error(annotator:task), data = annot))
```

```
Error: annotator:task
          Df Sum Sq Mean Sq F value  Pr(>F)
annotator  3    169      56    5.58   0.036 *
task       2   3315    1658  163.83 5.8e-06 ***
Residuals  6     61      10
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Error: Within
          Df Sum Sq Mean Sq F value Pr(>F)
Residuals 12   51.7    4.31
```

- הפלט מחולק ל**שכבות** (strata). בשכבה `annotator:task`, השורה `Residuals` היא MS_{AB}, ושני האפקטים נבדקים מולה
- $F_A = MS_A / MS_{AB} = 5.58$ עם $(3, 6)$ דרגות חופש. זה המכנה הנכון. בשכבה `Within`, השורה `Residuals` היא MS_E
- **למה `annotator` ו-`task` לא בתוך `Error()`?** כי `Error()` לא מצהיר מי אקראי. הוא קובע **מול מה** בודקים את מה שכתוב מחוצה לו. אם נכתוב `Error(annotator * task)`, כל פקטור יקבל שכבה משלו, בלי מבחן F בכלל
- את רכיבי השונות מחשבים מתוחלות ה-MS, לפי הנוסחאות בשקף על האומדים

חישוב רכיבי השונות

הרעיון: כל $E(MS)$ הוא סכום של רכיבי שונות. מחסירים שני MS שההפרש ביניהם הוא **רכיב אחד בלבד**, ומחלקים במקדם שלו. למשל:

$$E(MS_A) - E(MS_{AB}) = \left(\sigma^2 + n\sigma^2_{\tau\beta} + bn\sigma^2_{\tau}\right) - \left(\sigma^2 + n\sigma^2_{\tau\beta}\right) = bn\,\sigma^2_{\tau}$$

מהפלט:

מקור	MS
מתייגים (MS_A)	56.47
סוגי משימות (MS_B)	1,657.7
אינטראקציה (MS_{AB})	10.12
שגיאה (MS_E)	4.31

$$a = 4, \quad b = 3, \quad n = 2$$

$$\hat{\sigma}^2 = MS_E = 4.31$$

$$\hat{\sigma}^2_{\tau\beta} = \frac{MS_{AB} - MS_E}{n} = \frac{10.12 - 4.31}{2} = 2.91$$

$$\hat{\sigma}^2_{\tau} = \frac{MS_A - MS_{AB}}{bn} = \frac{56.47 - 10.12}{3 \cdot 2} = 7.72$$

$$\hat{\sigma}^2_{\beta} = \frac{MS_B - MS_{AB}}{an} = \frac{1,657.7 - 10.12}{4 \cdot 2} = 205.95$$

- **סך הכול:** $4.31 + 2.91 + 7.72 + 205.95 = 220.9$. החלק של כל רכיב הוא הרכיב חלקי הסכום, למשל סוגי משימות: $205.95/220.9 = 93\%$
- **אומד שלילי?** אם $MS_A < MS_{AB}$ יוצא אומד שלילי. שונות לא יכולה להיות שלילית, ולכן מדווחים 0

איך קוראים את התוצאה

בלי `Error()`: מול MS_E (שגוי)

מקור	F	p
מתייגים	13.1	0.00043
סוגי משימות	384.7	< 0.0001

עם `Error()`: מול MS_{AB} (נכון)

מקור	F	p
מתייגים	5.6	0.036
סוגי משימות	163.8	< 0.0001

רכיבי שונות (מהשקף הקודם)

רכיב	שונות	חלק
סוגי משימות	206.0	93%
מתייגים	7.7	3%
אינטראקציה	2.9	1%
שגיאה	4.3	2%

- **מתייגים:** עם המכנה הנכון ה-p-value גדול פי 84, כי MS_{AB} גדול פי 2.3 מ-MS_E, ויש רק 6 דרגות חופש במכנה. כאן זה עדיין מובהק, אבל עם פחות נתונים המסקנה הייתה יכולה להתהפך
- **סוגי משימות:** כ-93% מהשונות. **ההחלטה:** לתמחר תיוג **לפי סוג משימה**, לא מחיר אחיד לפריט. מתייגים: רק כ-3%
- **זהירות:** את σ^2_β אומדים מ-3 סוגי משימות בלבד, ולכן האומד מאוד לא מדויק. כדי לאמוד אותו טוב צריך **יותר סוגי משימות**, לא יותר פריטים

מודל היררכי (Nested)

מצטלב מול היררכי

מצטלב (crossed): כל רמה של A מופיעה עם **כל** רמה של B. כל מתייג תייג כל סוג משימה.

היררכי (nested): כל רמה של B קיימת **רק בתוך רמה אחת** של A. במקרה כזה אומרים ש-B מקונן ב-A, ומסמנים $B(A)$.

הדוגמה: שלושה מרכזי נתונים (ישראל, אירופה, ארה״ב). בכל אחד מגרילים **4 שרתים**, ושולחים לכל שרת 3 בקשות.

מרכז נתונים	ישראל				אירופה				ארה״ב			
שרת	s1	s2	s3	s4	s5	s6	s7	s8	s9	s10	s11	s12

- שרת s1 קיים רק בישראל. אין משמעות ל”שרת 1 באירופה”, ולכן **אין אינטראקציה** מרכז × שרת
- גם אם נמספר את השרתים 1–4 בכל מרכז, “שרת 1” בישראל ו”שרת 1” באירופה הם שרתים **שונים**

המודל והטבלה

$$y_{ijk} = \mu + \tau_i + \beta_{j(i)} + \varepsilon_{ijk} \qquad i = 1..a \text{ (מרכז)},\ j = 1..b \text{ (שרת בתוך מרכז)},\ k = 1..n$$

$$SS_T = \underbrace{bn\sum_i (\bar{y}_{i\cdot\cdot} - \bar{y}_{\cdots})^2}_{SS_A} + \underbrace{n\sum_{i,j} (\bar{y}_{ij\cdot} - \bar{y}_{i\cdot\cdot})^2}_{SS_{B(A)}} + \underbrace{\sum_{i,j,k} (y_{ijk} - \bar{y}_{ij\cdot})^2}_{SS_E}$$

- $SS_{B(A)}$: הפיזור של השרתים **סביב הממוצע של המרכז שלהם**. דרגות חופש: $a(b-1)$
- $SS_{B(A)} = SS_B + SS_{AB}$ של המודל המצטלב

מקור	df	$E(MS)$, **השרתים אקראיים**	F_0
A (מרכז)	$a-1$	$\sigma^2 + n\sigma^2_\beta + \frac{bn\sum \tau_i^2}{a-1}$	$MS_A / MS_{B(A)}$
B(A) (שרת)	$a(b-1)$	$\sigma^2 + n\sigma^2_\beta$	$MS_{B(A)} / MS_E$
שגיאה	$ab(n-1)$	σ^2	

השרתים: ב-R ובחישוב נכון

```r
set.seed(808)
dc_servers <- tibble(dc = rep(c("il", "eu", "us"), each = 4), server = paste0("s", 1:12),
                     dc_mean = rep(c(120, 110, 135), each = 4), u = rnorm(12, 0, 8)) |>
  expand_grid(rep = 1:3) |>
  mutate(ms = round(dc_mean + u + rnorm(n(), 0, 6), 1), dc = factor(dc, levels = c("il", "eu", "us")))
fit_nest <- aov(ms ~ dc / server, data = dc_servers)   # dc / server = dc + dc:server
summary(fit_nest)
```

```
            Df Sum Sq Mean Sq F value  Pr(>F)
dc           2   5964    2982   100.5 2.2e-12 ***
dc:server    9   4502     500    16.9 2.4e-08 ***
Residuals   24    712      30
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
tab_n <- summary(fit_nest)[[1]]
F_dc <- tab_n[1, "Mean Sq"] / tab_n[2, "Mean Sq"]
c(F_dc = F_dc, p = pf(F_dc, 2, 9, lower.tail = FALSE), F_crit = qf(0.95, 2, 9))
```

```
   F_dc       p  F_crit
5.96038 0.02246 4.25649
```

- R בודק את המרכזים מול MS_E: $F = 100.5$. **זה שגוי.** השרתים הוגרלו, ולכן המכנה הוא $MS_{B(A)}$: $F = 5.96$ עם $(2, 9)$ דרגות חופש
- המסקנה נשארת (דוחים), אבל ה-p-value שונה בסדרי גודל. **יש לנו 4 שרתים לכל מרכז, לא 12 בקשות**

השרתים: הדרך הנכונה, עם `Error()`

השרתים הוגרלו, ולכן אומרים ל-R שהם האיבר האקראי:

```
summary(aov(ms ~ dc + Error(server), data = dc_servers))
```

```
Error: server
          Df Sum Sq Mean Sq F value Pr(>F)
dc         2   5964    2982    5.96  0.022 *
Residuals  9   4502     500
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Error: Within
          Df Sum Sq Mean Sq F value Pr(>F)
Residuals 24    712    29.7
```

- בשכבה `server`, השורה `Residuals` היא $MS_{B(A)}$ עם 9 דרגות חופש, והמרכזים נבדקים מולה: $F = 5.96$, כמו בחישוב הידני
- שמות השרתים כאן ייחודיים (`s1`–`s12`), ולכן מספיק `Error(server)`. אם השרתים היו ממוספרים מחדש בכל מרכז (`s1`–`s4` בכל אחד), היינו כותבים `Error(dc:server)`

מה יכול להשתבש

חמש מלכודות בניתוח דו-כיווני

1. **לפרש אפקט ראשי כשיש אינטראקציה:** "RAG מוסיף 9 נקודות בממוצע" נכון בממוצע, ולא נכון לאף מודל. מסתכלים קודם על האינטראקציה ועל הגרף
2. **מכנה לא נכון:** עם אפקטים אקראיים או מודל היררכי, `aov()` רגיל מחלק הכול ב-MS_E. צריך לבחור את המכנה לפי תוחלות ה-MS, או להגדיר אותו ב-`Error()`
3. **מצטלב במקום היררכי:** אם מריצים `dc * server` על שרתים שמקוננים במרכזים, מקבלים שורת "אינטראקציה" חסרת משמעות
4. **להתעלם מבלוק:** אפקט היום נשאר בשגיאה, MS_E מנופח, והמבחן מאבד עוצמה
5. **ניסוי לא מאוזן:** נניח שרוב הריצות של המודל הגדול היו עם RAG, ורוב הריצות של הקטן בלי. עכשיו גודל ו-RAG **מתואמים**, וחלק מהשונות אפשר להסביר בכל אחד מהם. `aov()` מוסיף את הפקטורים **לפי הסדר בנוסחה**: הראשון מקבל את כל מה שהוא מסביר, כולל החלק המשותף, והשני רק את מה שהוא מוסיף מעבר לזה. בנתוני ה-RAG, אחרי שמוחקים כמה ריצות: $SS_{RAG} = 248.7$ כשהוא שני, ו-814.7 כשהוא ראשון. ההפרש, 566.1, הוא החלק המשותף. **לכן מתכננים ניסוי מאוזן**

תרגיל מסכם: לתמחר ספרי שמע לפי סוגה

הסיפור

ליסנה (שם בדוי) היא סטארט-אפ שמשיק חנות ספרי שמע (audiobooks). לפני ההשקה, ההנהלה רוצה לקבוע **מחיר לפי סוגה** (genre): מחיר גבוה לסוגות שיש עליהן ביקוש גבוה, ומחיר נמוך יותר כדי למשוך קהל לסוגות עם ביקוש נמוך.

לליסנה עוד אין נתוני מכירות משלה. לכן אנליסטית בחברה אוספת מדגם מהקטלוג הציבורי של Audible:

- **5 סוגות:** פנטזיה, מותחן, רומנטיקה, עזרה עצמית והיסטוריה
- **סוג הקריינות:** קריין אנושי, או קול סינתטי (AI narration). ההפקה עם קול סינתטי זולה בהרבה
- מכל צירוף של סוגה וקריינות היא מגרילה **12 ספרים** שיצאו ב-2025. סך הכול $N = 120$
- **מדד הביקוש:** מספר הדירוגים שהספר קיבל בשנה הראשונה. Audible לא מפרסמת מכירות, ומספר הדירוגים הוא מדד עקיף נפוץ

השאלה העסקית: אילו סוגות זוכות לביקוש גבוה, האם הקריינות משנה, ומה זה אומר על המחיר?

הנתונים

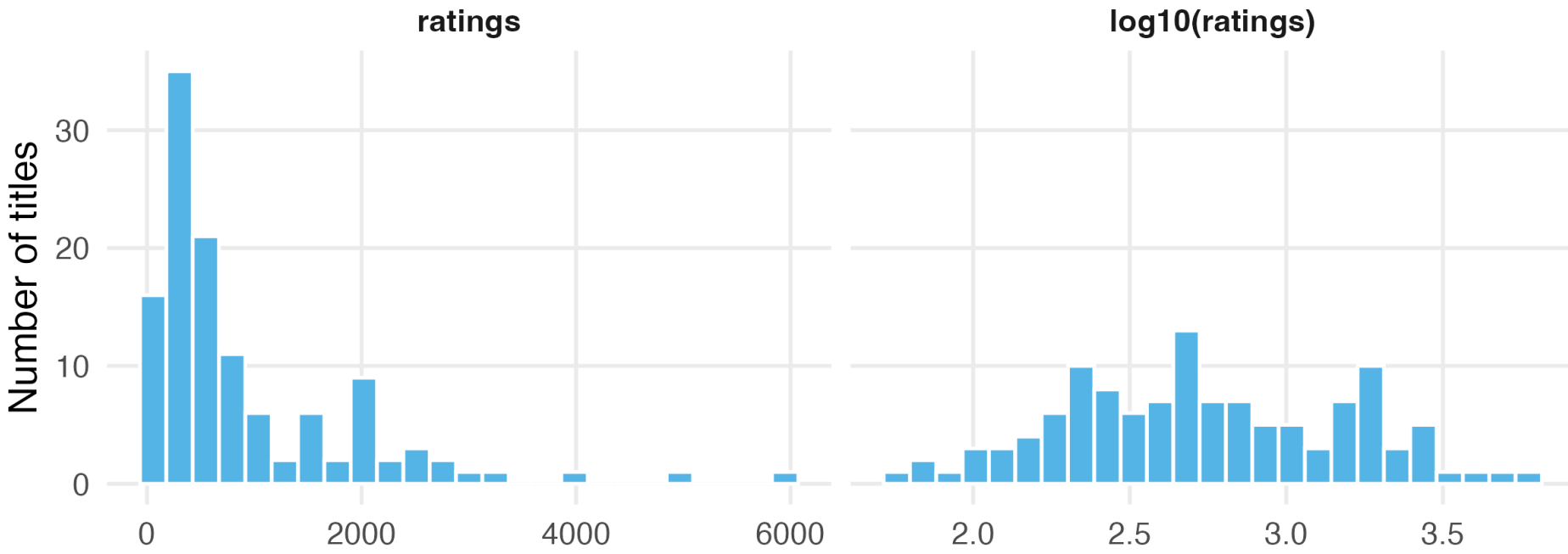
```
books |> slice_sample(n = 6)
```

```
# A tibble: 6 × 5
  genre     narration hours price ratings
  <fct>     <fct>     <dbl> <dbl>   <dbl>
1 history   ai         15.8  24.0     217
2 self_help human       6.4  11.0     147
3 fantasy   ai         20.3  30.0    5950
4 thriller  human      16.7  30.0    1962
5 history   human      13.5  27.0     444
6 romance   ai         10.9  25.0     159
```

- `hours`: משך ההקלטה בשעות · `price`: מחיר מחירון בדולרים · `ratings`: מספר הדירוגים בשנה הראשונה

שאלה 1: מה מודדים?

מספר הדירוגים: זנב ימני ארוך. אחרי לוגריתם: סימטרי



השאלה: האנליסטית מנתחת את $\log_{10}(\text{ratings})$ ולא את מספר הדירוגים עצמו. למה? ואיך מפרשים הפרש של 0.3 בסולם הלוגריתמי?

תשובה 1

- **ההתפלגות מוטה ימינה:** רוב הספרים מקבלים כמה מאות דירוגים, ומעטים מקבלים אלפים. ממוצע של סוגה נגרר אחרי ספר מצליח אחד
- **הפיזור גדל עם הממוצע:** בסוגה פופולרית ההבדלים בין ספרים גדולים יותר. זה מפר את הנחת שוויון השונויות
- **ההשפעות כפליות:** קריינות טובה מכפילה ביקוש, היא לא "מוסיפה 200 דירוגים". הלוגריתם הופך השפעות כפליות לחיבוריות, וזה מה ש-ANOVA ורגרסיה מניחות

הפירוש: הפרש d בסולם $\log_{10}$ הוא **יחס** של 10^d בסולם המקורי.

הפרש ב-$\log_{10}$	**0.1**	**0.3**	**0.5**	**1**
יחס בין מספרי הדירוגים	×1.26	×2	×3.16	×10

```
books <- books |> mutate(log_r = log10(ratings))
```

שאלה 2: השלימו את טבלת ה-ANOVA

ניתוח שונות חד-כיווני של $\log_{10}(\text{ratings})$ לפי סוגה (5 סוגות, 24 ספרים בכל אחת). חלק מהפלט נמחק:

מקור	SS	df	MS	F_0
סוגה	5.00	(א)	(ב)	(ה)
שגיאה	(ג)	(ד)	0.162	
סך הכול	23.63	119		

1. השלימו את הטבלה
2. נסחו את ההשערות, ובדקו אותן ב-$\alpha = 0.05$ ($F_{0.95;\,4,\,115} = 2.45$)
3. כמה אחוזים מהשונות בביקוש מוסברים על ידי הסוגה?

תשובה 2

- (א) $a - 1 = 4$ · (ב) $MS = 5.00/4 = 1.251$ · (ג) $SS_E = 23.63 - 5.00 = 18.63$ · (ד) $N - a = 115$
- (ה) $F_0 = 1.251/0.162 = 7.72 > 2.45$: **דוחים** את $H_0 : \tau_1 = \cdots = \tau_5 = 0$. יש הבדל בביקוש בין הסוגות
- $\eta^2 = SS_{Tr}/SS_T = 5.00/23.63 = 0.21$: הסוגה מסבירה כ-21% מהשונות. **מובהק, אבל רוב השונות היא בין ספרים בתוך אותה סוגה**

```
summary(fit_g)
```

```
             Df Sum Sq Mean Sq F value  Pr(>F)
genre         4    5.0   1.251    7.72 1.5e-05 ***
Residuals   115   18.6   0.162
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

שאלה 3: האם אפשר לסמוך על המבחן?

```
bartlett.test(ratings ~ genre, data = books)$p.value     # raw counts
```

```
[1] 2.457e-14
```

```
bartlett.test(log_r ~ genre, data = books)$p.value       # log scale
```

```
[1] 0.1738
```

```
kruskal.test(ratings ~ genre, data = books)
```

```

    Kruskal-Wallis rank sum test

data:  ratings by genre
Kruskal-Wallis chi-squared = 25, df = 4, p-value = 6e-05
```

השאלה: מה אומרות שתי בדיקות Bartlett? למה האנליסטית הריצה גם Kruskal–Wallis, ומה היה יוצא אם הייתה מריצה אותו על `log_r`?

תשובה 3

- **על המספרים הגולמיים:** p זעיר. השונויות שונות מאוד בין הסוגות, כצפוי כשהפיזור גדל עם הממוצע
- **על הלוגריתם:** $p = 0.17$. אין עדות לשונויות שונות. **הלוגריתם תיקן את הבעיה**
- **Kruskal–Wallis** לא מניח נורמליות. הוא מגיע לאותה מסקנה, וזה מחזק את התוצאה
- **על `log_r` הוא היה נותן בדיוק אותה תוצאה.** הלוגריתם הוא פונקציה עולה, ולכן הוא לא משנה את **הדירוגים** (ranks) של התצפיות. מבחנים א-פרמטריים על דרגות לא רגישים לטרנספורמציה עולה

שאלה 4: אילו סוגות שונות זו מזו?

```r
tk <- TukeyHSD(fit_g)$genre
round(tk[order(tk[, "p adj"]), ], 3)
```

```
                   diff    lwr    upr p adj
history-fantasy    -0.622 -0.944 -0.300 0.000
history-romance    -0.454 -0.776 -0.132 0.001
history-self_help  -0.401 -0.723 -0.079 0.007
history-thriller   -0.372 -0.694 -0.050 0.015
thriller-fantasy   -0.250 -0.572  0.072 0.205
self_help-fantasy  -0.221 -0.543  0.101 0.323
romance-fantasy    -0.168 -0.490  0.154 0.601
romance-thriller    0.083 -0.239  0.405 0.954
self_help-romance  -0.053 -0.375  0.269 0.991
self_help-thriller  0.030 -0.293  0.352 0.999
```

1. חשבו את ה-HSD ידנית ($q_{0.95}(5, 115) = 3.92$), ובדקו שהוא מתאים לרוחב רווחי הסמך בפלט
2. חלקו את הסוגות לקבוצות. מה המשמעות לתמחור?

תשובה 4

$$HSD = q_{0.95}(5, 115)\sqrt{\frac{\overline{MS_E}}{n}} = 3.92\sqrt{\frac{0.162}{24}} = 0.322$$

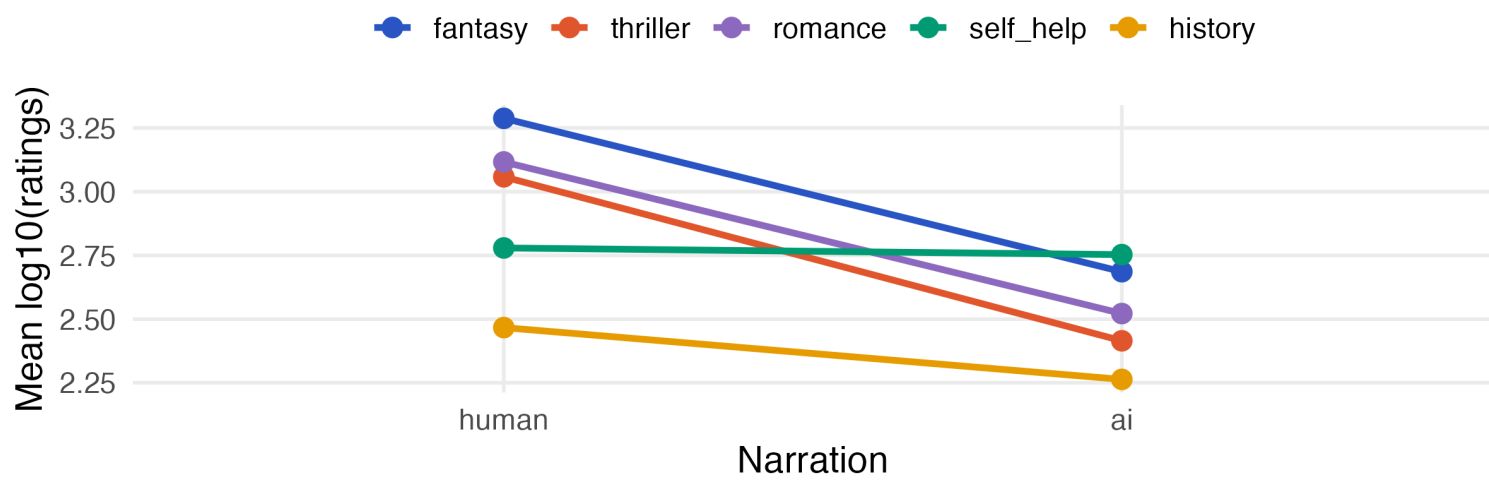
זה בדיוק חצי הרוחב של כל רווח סמך בפלט: $\text{diff} \pm HSD$.

- **היסטוריה** נמוכה מכל ארבע הסוגות האחרות. ההפרש מפנטזיה, -0.62 בסולם $\log_{10}$, הוא יחס של $10^{-0.62} = 0.24$: **כרבע** מהדירוגים
- **בין ארבע האחרות אין הבדל מובהק.** שתי קבוצות: {פנטזיה, מותחן, רומנטיקה, עזרה עצמית} ו-{היסטוריה}
- **לתמחור:** לפי הניתוח הזה, שתי מדרגות מחיר. אבל רגע: עוד לא הסתכלנו על הקריינות

שאלה 5: סוגה × קריינות

קול סינתטי מוריד את הביקוש לסיפורת, ולא לעיון

ממוצע log10(ratings) בכל תא, 12 ספרים לתא




```r
fit_gn <- aov(log_r ~ genre * narration, data = books)
summary(fit_gn)
```

```
                Df Sum Sq Mean Sq F value  Pr(>F)
genre            4   5.00    1.25   11.87 4.7e-08 ***
narration        1   5.14    5.14   48.79 2.3e-10 ***
genre:narration  4   1.89    0.47    4.49  0.0021 **
Residuals      110  11.59    0.11
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

השאלה: מה סדר הקריאה של הטבלה? מה המסקנה? בדקו ידנית את F_0 של האינטראקציה.

תשובה 5

- **קודם האינטראקציה:** $F_0 = \frac{1.89/4}{11.59/110} = \frac{0.474}{0.105} = 4.49 > F_{0.95;\,4,\,110} = 2.45$. **ההשפעה של הקריינות תלויה בסוגה**
- לכן "קול סינתטי מוריד את הביקוש" (האפקט הראשי) הוא ממוצע שלא נכון לאף סוגה. צריך לבדוק **אפקטים פשוטים**: את השפעת הקריינות בכל סוגה בנפרד
- MS_E ירד מ-0.162 ל-0.105: הקריינות הסבירה חלק ממה שבניתוח החד-כיווני נחשב לרעש

שאלה 5, המשך: פנטזיה מול עזרה עצמית

השאלה: בעזרת התרשים והטבלה מהשאלה הקודמת (12 ספרים בכל תא):

1. האם לקול סינתטי יש השפעה מובהקת על הביקוש **בפנטזיה**? ובעזרה עצמית?
2. בנו רווח סמך של 95% להפרש בעזרה עצמית, ופרשו אותו כיחס בין מספרי הדירוגים

הפרש בין שני תאים, עם ה-MS_E של המודל הדו-כיווני:

$$SE = \sqrt{\frac{2\,MS_E}{n}} = \sqrt{\frac{2 \cdot 0.105}{12}} = 0.133 \qquad t_{0.975;\,110} = 1.98$$

- **פנטזיה:** $2.69 - 3.29 = -0.60$, ולכן $t = -0.60/0.133 = -4.5$. **מובהק.** בערך רבע מהדירוגים: $10^{-0.60} = 0.25$
- **עזרה עצמית:** $2.75 - 2.78 = -0.03$, ולכן $t = -0.2$. **לא מובהק**
- **רווח הסמך:** $-0.03 \pm 1.98 \cdot 0.133 = [-0.29,\ 0.24]$, וכיחס: $[0.51,\ 1.72]$
- **"לא מובהק" זה לא "אין השפעה".** הרווח כולל גם ירידה של כמחצית בדירוגים וגם עלייה של יותר מחצי. עם 12 ספרים לתא אי אפשר לדעת

שאלה 5, המשך: כמה השוואות עשינו?

השאלה: בשאלה הקודמת בדקנו כל סוגה ברמת מובהקות של 5%. אבל את פנטזיה ואת עזרה עצמית **בחרנו אחרי שהסתכלנו על התרשים**, ובסוף נרצה לדווח על כל 5 הסוגות. מה הבעיה? איך מתקנים? אילו מסקנות שורדות?

- **הבעיה:** 5 מבחנים, כל אחד ב-$\alpha = 0.05$. הסיכוי לפחות לדחייה שגויה אחת (FWER) מגיע עד $1 - 0.95^5 = 0.23$. ובחירת הסוגות "המעניינות" מהתרשים היא בעצמה השוואה: בפועל השווינו את כל 5
- **תיקון Bonferroni:** כל מבחן ברמה $0.05/5 = 0.01$, ולכן הסף הוא $t_{1-0.01/2;\,110} = 2.62$ במקום 1.98

סוגה	fantasy	thriller	romance	self_help	history
הפרש (log10)	-0.60	-0.64	-0.60	-0.03	-0.20
t	-4.5	-4.9	-4.5	-0.2	-1.5

- **מה שורד:** בשלוש סוגות הסיפורת $|t| > 2.62$, גם אחרי התיקון. ההיסטוריה והעזרה העצמית לא מובהקות בכל מקרה. כאן המסקנה לא השתנתה, אבל עם אפקטים קטנים יותר היא הייתה יכולה להשתנות

שאלה 6: ומה עם המחיר?

האנליסטית מוסיפה את המחיר ואת המשך, ומריצה רגרסיה:

```
fit_lm <- lm(log_r ~ price + hours + genre + narration, data = books)
round(summary(fit_lm)$coefficients, 4)
```

```
               Estimate Std. Error t value Pr(>|t|)
(Intercept)      2.9841     0.2139  13.950   0.0000
price            0.0218     0.0091   2.388   0.0186
hours           -0.0220     0.0173  -1.272   0.2058
genrethriller   -0.2566     0.1111  -2.311   0.0227
genreromance    -0.1642     0.1304  -1.259   0.2105
genreself_help  -0.2183     0.1409  -1.550   0.1240
genrehistory    -0.5995     0.0993  -6.039   0.0000
narrationai     -0.3379     0.0694  -4.871   0.0000
```

1. פרשו את המקדם של `narrationai`, ובנו לו רווח סמך של 95% ($t_{0.975;\,112} = 1.98$)
2. מנהל המוצר אומר: "המקדם של המחיר **חיובי** ומובהק. נעלה מחירים, והביקוש יעלה." מה דעתכם?
3. למה האנליסטית משתמשת ב-`log_r` כמשתנה התלוי ברגרסיה? האם אפשר לבנות את מודל הרגרסיה על המשתנה המקורי, מספר הדירוגים (`ratings`)?

תשובה 6

- **narrationai:** $-0.338 \pm 1.98 \cdot 0.0694 = [-0.475, \ -0.200]$. ביחס: $10^{-0.34} = 0.46$, כלומר ספר עם קול סינתטי מקבל **בערך חצי ממספר הדירוגים** של ספר דומה עם קריין אנושי (פי 0.33 עד פי 0.63), **כשהסוגה, המחיר והמשך קבועים**
- **אבל** זה מודל בלי אינטראקציה. בשאלה 5 ראינו שהאפקט שונה מאוד בין סיפורת לעיון, ולכן "פי חצי" הוא ממוצע שלא מתאר אף סוגה (מלכודת 1 של הדו-כיווני)
- **המחיר:** 0.0218 לכל דולר, יחס של $10^{0.0218} = 1.052$. **אסור להסיק מזה שהעלאת מחיר מעלה ביקוש:**
  - אלה **נתוני תצפית**, לא ניסוי. אף אחד לא הגריל מחירים
  - מו"לים קובעים מחיר גבוה לספרים **שהם מצפים שיצליחו** (מחבר מוכר, שיווק). ה"באזז" הזה מעלה גם את המחיר וגם את הביקוש, והוא לא במודל
  - המקדם החיובי מתאר את **ההחלטות של המו"לים**, לא את התגובה של הקונים למחיר

תשובה 6 (המשך): למה `log_r`?

טכנית, אפשר. `lm()` ירוץ גם על `ratings`. השאלה היא אם ההנחות של המודל מתקיימות. נשווה בין שני המודלים:

```
fit_raw <- lm(ratings ~ price + hours + genre + narration, data = books)
shapiro.test(resid(fit_raw))$p.value   # normality of residuals, raw scale
```

```
[1] 3.454e-09
```

```
shapiro.test(resid(fit_lm))$p.value    # normality of residuals, log scale
```

```
[1] 0.8789
```

```
sum(fitted(fit_raw) < 0)               # books with a negative predicted count
```

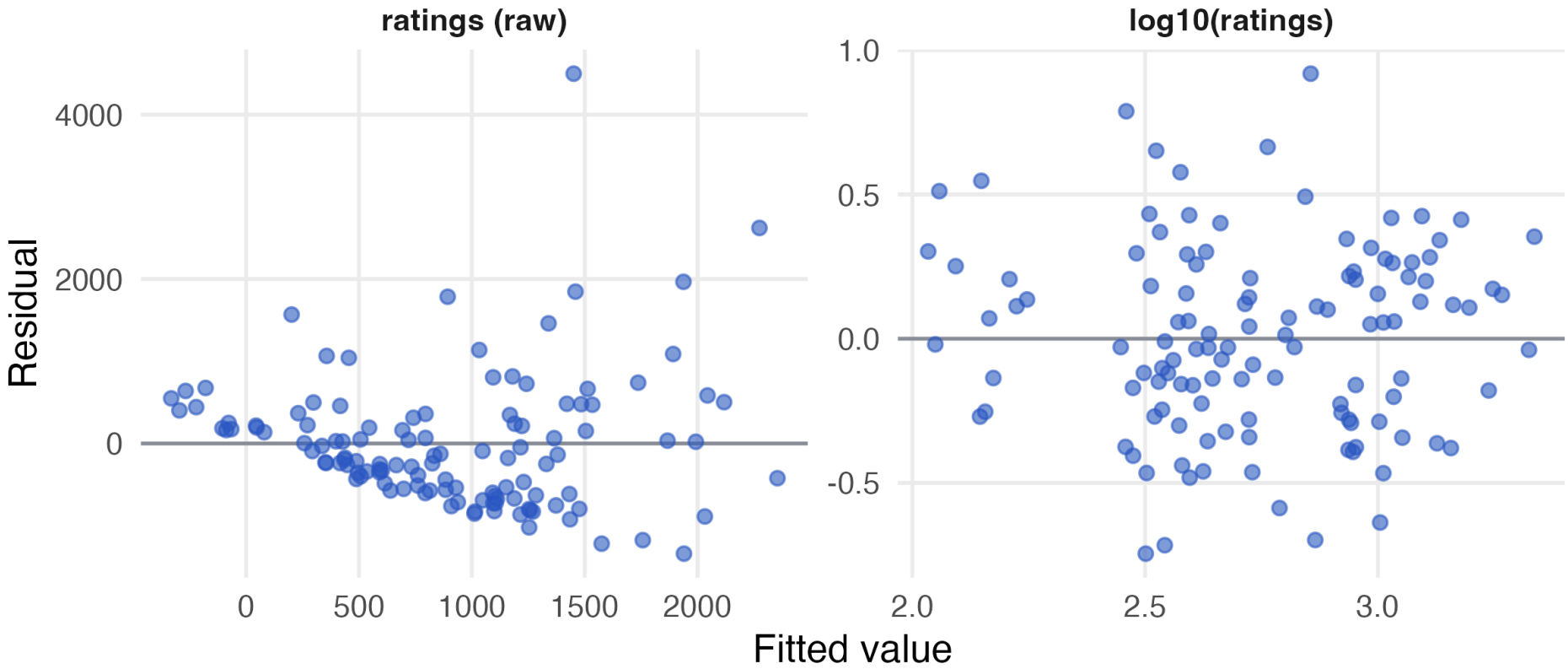
```
[1] 9
```

- **השאריות לא נורמליות** במודל על `ratings` ($p < 0.001$ ב-Shapiro–Wilk), ונורמליות על `log_r` ($p = 0.88$)
- **השונות לא קבועה:** סטיית התקן של השאריות עולה מכ-443 בתחזיות הנמוכות לכ-1,168 בגבוהות. זו אותה בעיה ש-Bartlett מצא בשאלה 3
- **תחזיות בלתי אפשריות:** המודל חוזה מספר דירוגים **שלילי** ל-9 ספרים

למה `log_r`? השאריות של שני המודלים

על הסולם המקורי: משפך. על הסולם הלוגריתמי: ענן אחיד

שאריות מול ערכים חזויים · אותם משתנים מסבירים בשני המודלים



למה `log_r`? הפירוש

- **ביקוש פועל באחוזים, לא ביחידות.** במודל על `ratings`, המקדם של `narrationai` הוא -601 דירוגים, **אותו מספר בכל סוגה**. אבל לספר היסטוריה ממוצע יש רק 294 דירוגים, ולספר פנטזיה 1,594. הורדה של 601 דירוגים גדולה מכל הביקוש הממוצע בהיסטוריה, ובפנטזיה היא רק כ-38% ממנו
- **על `log_r`, המקדם הוא יחס:** $10^{-0.34} = 0.46$, כלומר "פי חצי" בכל סוגה. זה פירוש שאפשר לקחת להנהלה
- **המסקנה:** אפשר לבנות את המודל על `ratings`, אבל הוא מפר את הנחות הנורמליות ושוויון השונויות, חוזה ערכים שליליים ומניח שהאפקטים חיבוריים. כשהתגובה היא ספירה מוטה ימינה, שמשתנה בסדרי גודל, מנתחים את הלוגריתם שלה

שאלה 7: ההמלצה

מה ממליצים להנהלה של ליסנה?

1. **סוגה:** ביקוש גבוה לסיפורת ולעזרה עצמית, נמוך להיסטוריה (כרבע). אפשר להתחיל בשתי מדרגות מחיר
2. **קריינות:** בסיפורת קול סינתטי מוריד את הביקוש לכרבע. **בעזרה עצמית אין לו מחיר נראה לעין.** לכן: להפיק עזרה עצמית בקול סינתטי (זול) ולתמחר אותה נמוך, להשקיע בקריינים לסיפורת, ולבדוק היסטוריה בניסוי קטן לפני שמחליטים
3. **מחיר:** הנתונים האלה **לא יכולים** לענות כמה הביקוש רגיש למחיר. ההמלצה: ניסוי מחיר מוגרל בהשקה, למשל שני מחירים לכל סוגה
4. **הסתייגויות:** מספר דירוגים ≠ מכירות. Audible ≠ השוק של ליסנה. ומדגם של 12 ספרים לתא מספיק לאפקטים גדולים בלבד

מה ראינו בתרגיל

שלב	הכלי	מה למדנו
להחליט מה מודדים	התפלגות, טרנספורמציה	$\log$ כשההשפעות כפליות והפיזור גדל עם הממוצע
האם יש הבדל?	חד-כיווני ANOVA	יש, אבל הסוגה מסבירה רק כ-21%
האם אפשר לסמוך?	Bartlett, Kruskal–Wallis	הלוגריתם תיקן את השונויות, והמסקנה עמידה
בין מי למי?	Tukey HSD	היסטוריה לבד, השאר יחד
האם ההשפעה תלויה במשהו?	דו-כיווני, אפקטים פשוטים ANOVA	קול סינתטי פוגע בסיפורת בלבד
ומה עם משתנים רציפים?	רגרסיה מרובה	מקדם מובהק ≠ סיבתיות

סיכום

מה לקחת מההרצאה

השאלה	הכלי	ב-R
שני פקטורים, עם חזרות	$SS_A + SS_B + SS_{AB} + SS_E$, קודם בודקים את האינטראקציה	`aov(y ~ A * B)`
יש אינטראקציה	אפקטים פשוטים, השוואות בין תאים	`emmeans(fit, ~ A, by = "B")`
מקור רעש ידוע (יום, משמרת)	בלוק, $n = 1$, מודל אדיטיבי	`aov(y ~ A + block)`
הרמות הוגרלו	$F_A = MS_A / MS_{AB}$, רכיבי שונות	`Error(A:B)`
A קיים רק בתוך B	מודל היררכי, $F_A = MS_A / MS_{B(A)}$	`Error(A:B)`
תגובה מוטה, השפעות כפליות	$\log$, ופירוש של הפרשים כיחסים	`log10(y)`
מקדם ברגרסיה על נתוני תצפית	מתאם, לא סיבתיות. לסיבתיות צריך ניסוי	

בהרצאה הבאה

- **רגרסיה לוגיסטית** (logistic regression): כשהתגובה היא כן או לא. האם משתמש בשבוע ניסיון של ליסנה יירשם למנוי?
- סיכוי (odds), logit, יחס סיכויים (odds ratio)
- אמידה בנראות מקסימלית, `glm()` ומבחן deviance
- כמה משתנים מסבירים, מבחן חלקי וסיווג