6. השוואות מרובות (המשך) וניתוח שונות דו-כיווני

ניתוח נתונים סטטיסטי · הרצאה 6

ד״ר עדי שריד

2026-11-22

על מה נדבר היום

- **חזרה קצרה:** איפה עצרנו בשבוע שעבר
- **השלמה מהרצאה 5:**
  - שיטת Duncan, מההתחלה ועד הסוף, כולל תרגיל
  - שיטת Tukey (HSD)
  - Kruskal–Wallis וניגודים עם חציונים
- **ניתוח שונות דו-כיווני** (two-way ANOVA):
  - אפקטים ראשיים ואינטראקציה
  - המודל, פירוק השונות וטבלת ANOVA
  - בלוקים: אפקט היום במוקד
  - אפקטים אקראיים ומודל היררכי (nested)

חזרה: איפה עצרנו

המודל החד-כיווני ופירוק השונות

המודל: a רמות, n תצפיות בכל רמה, $N = an$

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij}, \qquad \sum_{i=1}^{a} \tau_i = 0, \qquad \varepsilon_{ij} \sim N(0, \sigma^2)$$ בלתי תלויים

$H_0 : \tau_1 = \cdots = \tau_a = 0$ מול H_1: לפחות $\tau_i \neq 0$ אחד

פירוק סכום הריבועים:

$$\underbrace{\sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{\cdot\cdot})^2}_{SS_T\ (SS_y),\ N-1} = \underbrace{n\sum_{i=1}^{a}(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2}_{SS_{Tr},\ a-1} + \underbrace{\sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{i\cdot})^2}_{SS_E,\ N-a}$$

- SS_T (או SS_y): השונות הכוללת של התגובה סביב הממוצע הכללי
- SS_{Tr}: כמה ממוצעי הרמות רחוקים מהממוצע הכללי (**בין** הקבוצות)
- SS_E: הפיזור של התצפיות סביב ממוצע הרמה שלהן (**בתוך** הקבוצות)
- $F_0 = \dfrac{MS_{Tr}}{MS_E} = \dfrac{SS_{Tr}/(a-1)}{SS_E/(N-a)} \sim F_{a-1,\,N-a}$ תחת H_0, ו-MS_E הוא האומד של σ^2

מוקד השירות: חמישה מודלי שפה

5 מודלים (M1, M2 קטנים, M3 בינוני, M4, M5 גדולים), 9 ימים לכל מודל. התגובה: משך הטיפול הממוצע בפנייה (דקות).

```
summary(fit_ai)
```

```
            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
```

- $F_0 = 5.28 > F_{0.95;\,4,\,40} = 2.61$: **יש הבדל בין המודלים**. $MS_E = 0.854$
- מבחן ה-F לא אומר **בין מי למי**. שתי דרכים לענות:

1. **ניגודים:** מעט שאלות שניסחנו **מראש**. למשל "קטנים מול גדולים": $SS_C = 15.42$, $F_0 = 18.06$, כ-86% מ-SS_{Tr}
2. **השוואות מרובות:** כל הזוגות, עם תיקון לכך שבודקים הרבה

למה צריך תיקון?

עם 5 רמות יש $\binom{5}{2} = 10$ זוגות. 10 מבחני t ב-$\alpha = 0.05$, כשאין שום הבדל:

שיעור הטעות לכל השוואה	$\alpha = 0.05$
שיעור הטעות למשפחה (FWER): P(לפחות טעות אחת)	עד $1 - 0.95^{10} = 0.40$. בסימולציה: 27%

- כדי לשמור על ה-FWER, משווים כל **הפרש בין ממוצעים** לסף גבוה יותר מזה של מבחן t
- שתי השיטות שנלמד משתמשות בשגיאת התקן של ממוצע:

$$s_{\bar{y}} = \sqrt{\frac{MS_E}{n}} = \sqrt{\frac{0.854}{9}} = 0.308$$

השיטה	**הסף**	**FWER**
Duncan	$R_p = r_\alpha(p, f)\, s_{\bar{y}}$, תלוי במרחק p בדירוג	גבוה מ-α
Tukey HSD	$q_\alpha(a, f)\, s_{\bar{y}}$, סף אחד לכל הזוגות	בדיוק α

שיטת Duncan

הרעיון: הסף תלוי במרחק בדירוג

Duncan's multiple range test משווה את כל הזוגות, אחרי שמבחן ה-F דחה. החידוש: **לא כל הזוגות מקבלים אותו סף**.

ניסוי מחשבתי: אין שום הבדל בין 5 המודלים. ממיינים את 5 הממוצעים:

- שני ממוצעים **שכנים** בדירוג: הפער ביניהם הוא פער רגיל בין שני ממוצעים
- הממוצע **הגבוה** והממוצע **הנמוך** מבין החמישה: זה הפער **הגדול ביותר** מתוך 10 פערים. הוא גדול כמעט תמיד, גם בלי שום אפקט
- לכן זוג שנמצא בקצוות של הדירוג צריך לעבור סף **גבוה** יותר, וזוג של שכנים סף **נמוך** יותר

כשאין שום אפקט: הפער בין הקצוות גדול בהרבה מהפער בין שכנים

5 ממוצעים מאותה התפלגות, 4,000 חזרות · הפער ביחידות של שגיאת התקן



המרחק p

p = כמה ממוצעים נמצאים בטווח בין שני הממוצעים שמשווים, כולל שניהם.

דירוג	1	2	3	4	5
מודל	M2	M1	M3	M5	M4
ממוצע	5.33	5.00	3.99	3.96	3.76

הזוג	M2, M1	M2, M3	M1, M4	M2, M4
הממוצעים בטווח	M2, M1	M2, M1, M3	M1, M3, M5, M4	כל החמישה
p	2 (שכנים)	3	4	5 (הקצוות)

- עם a רמות, p נע בין 2 ל-a
- לכל p יש סף משלו: **הטווח המובהק הקטן ביותר** (least significant range)

$$R_p = r_\alpha(p, f) \cdot s_{\bar{y}} \qquad s_{\bar{y}} = \sqrt{\frac{MS_E}{n}} \qquad f = N - a$$

טבלת המקדמים $r_{0.05}(p,\, f)$

בבחינה מקבלים טבלה. **שורה** לפי f, דרגות החופש של MS_E. **עמודה** לפי p:

f	p = 2	p = 3	p = 4	p = 5	p = 6
10	3.15	3.29	3.38	3.43	3.47
20	2.95	3.10	3.19	3.25	3.30
25	2.91	3.06	3.15	3.22	3.27
30	2.89	3.04	3.13	3.20	3.25
40	2.86	3.01	3.10	3.17	3.22
60	2.83	2.98	3.07	3.14	3.20

- המקדם **גדל עם** p: ככל שהטווח רחב יותר, הסף גבוה יותר
- המקדם **קטן עם** f: יותר דרגות חופש, אומד מדויק יותר ל-σ^2, סף נמוך יותר
- למודלים, $f = 40$: $R_p = (2.86,\ 3.01,\ 3.10,\ 3.17) \times 0.308 = (0.88,\ 0.93,\ 0.96,\ 0.98)$

כמה סיכון מרשים בכל טווח?

נקודת המוצא (מהניסוי המחשבתי): גם בלי שום אפקט, ככל שיש יותר ממוצעים בטווח, הפער בין הקצוות גדול יותר. לכן הסף **חייב** לגדול עם p.

השאלה הפתוחה: כמה סיכון לאזעקת שווא מרשים בכל טווח? יש שתי תשובות אפשריות:

גישה מחמירה: 5% לכל ההשוואות יחד

- כל ההשוואות הן משפחה אחת
- רוצים סיכוי של 95% שלא תהיה אף אזעקת שווא
- לכן **כל** הזוגות נמדדים בסף של הטווח הרחב ביותר $(p = a)$
- זו שיטת Tukey, שנלמד מיד אחרי Duncan

Duncan: 5% לכל צעד

- טווח של p ממוצעים מכיל $p - 1$ צעדים בין שכנים
- כל צעד מקבל סיכון של 5%
- הסיכונים מצטברים, ולכן טווח רחב יותר מקבל **יותר** סיכון

p	2	3	4	5
הסיכון ב-Duncan	5%	10%	14%	19%
הסף ב-Duncan	2.86	3.01	3.10	3.17
הסף בגישה המחמירה	4.04	4.04	4.04	4.04

הסיכון: $1 - 0.95^{p-1}$. הספים: מקדמים עבור $f = 40$, כמו במוקד השירות.

בשורה אחת: Duncan נותן 5% לכל צעד, והצעדים מצטברים. לכן הסף שלו נמוך יותר מזה של הגישה המחמירה, והוא מוצא יותר הבדלים.

מאיפה מגיעים המקדמים?

ההתפלגות: studentized range. הטווח של p ממוצעים, מחולק בשגיאת התקן, כש-H_0 נכונה:

$$q_{p,\,f} = \frac{\bar{y}_{max} - \bar{y}_{min}}{s_{\bar{y}}}, \qquad s_{\bar{y}} = \sqrt{\frac{MS_E}{n}}$$

- **איפה f?** בתוך $s_{\bar{y}}$. את σ^2 אנחנו אומדים עם MS_E, שיש לו $f = N - a$ דרגות חופש. לכן ההתפלגות של q תלויה ב-p **וב-f**, בדיוק כמו ש-t תלוי בדרגות החופש (ו-$f \to \infty$ הוא המקרה ש-σ ידוע)

רמת המובהקות: לטווח של p ממוצעים, Duncan משתמש ב-α_p, ו-$r_\alpha(p, f)$ הוא האחוזון המתאים של $q_{p,\,f}$:

$$\alpha_p = 1 - (1-\alpha)^{p-1} \qquad \Longrightarrow \qquad r_\alpha(p, f) = q_{p,\,f}\left[(1-\alpha)^{p-1}\right]$$

- α_p היא הסיכון המצטבר של $p-1$ הצעדים מהשקף הקודם
- ב-R אין צורך בטבלה: `qtukey(prob, nmeans = p, df = f)`

```
p <- 2:5
r_duncan <- qtukey((1 - 0.05)^(p - 1), nmeans = p, df = 40)   # Duncan's r_alpha(p, 40)
rbind(p = p, alpha_p = 1 - 0.95^(p - 1), r = r_duncan, R_p = r_duncan * sqrt(mse / 9)) |> round(3)
```

```
         [,1]  [,2]  [,3]  [,4]
p       2.000 3.000 4.000 5.000
alpha_p 0.050 0.098 0.143 0.185
r       2.858 3.005 3.102 3.171
R_p     0.880 0.926 0.955 0.977
```

האלגוריתם

1. **מיינים** את ממוצעי הרמות בסדר יורד
2. **מחשבים** R_p לכל $p = 2, \ldots, a$
3. **בודקים מבחוץ פנימה:** קודם הזוג של הקצוות ($p = a$), אחר כך כל הזוגות עם $p = a - 1$, וכן הלאה עד השכנים ($p = 2$)
4. **ההפרש מובהק** אם $\bar{y}_{i\cdot} - \bar{y}_{k\cdot} > R_p$
5. **כלל ההגנה:** אם ההפרש בין שני ממוצעים **לא** מובהק, כל הממוצעים שביניהם נחשבים שווים, ו**לא בודקים** אף זוג בתוך הטווח הזה
6. **מסכמים באותיות:** רמות שחולקות אות לא שונות זו מזו באופן מובהק

המודלים: שלב אחר שלב

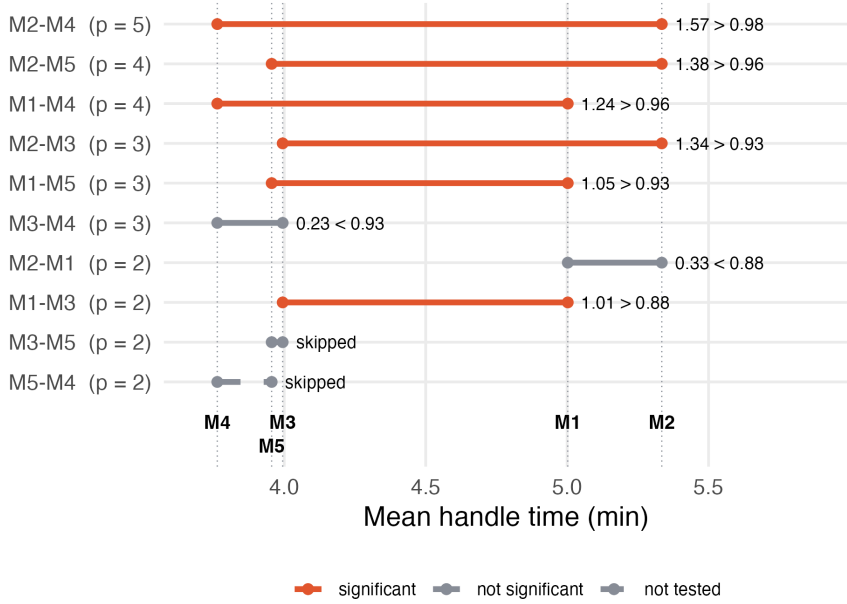
- $R_2 = 0.88$, $R_3 = 0.93$, $R_4 = 0.96$, $R_5 = 0.98$
- הממוצעים: M2 5.33 · M1 5.00 · M3 3.99 · M5 3.96 · M4 3.76

שלב	הזוג	p	ההפרש	R_p	החלטה
1	M2–M4	5	1.57	0.98	**מובהק**
2	M2–M5	4	1.38	0.96	**מובהק**
	M1–M4	4	1.24	0.96	**מובהק**
3	M2–M3	3	1.34	0.93	**מובהק**
	M1–M5	3	1.05	0.93	**מובהק**
	M3–M4	3	0.23	0.93	לא מובהק
4	M2–M1	2	0.33	0.88	לא מובהק
	M1–M3	2	1.01	0.88	**מובהק**
	M3–M5	2			**לא נבדק***
	M5–M4	2			**לא נבדק***

* כלל ההגנה: שניהם בתוך הטווח M3–M4, שלא היה מובהק

בודקים מלמעלה למטה: מהקצוות פנימה

כל קו הוא טווח בין שני ממוצעים · מובהק אם ההפרש גדול מ-Rp



התוצאה

	M2	M1	M3	M5	M4
ממוצע	5.33	5.00	3.99	3.96	3.76
קבוצה	**א**	**א**	**ב**	**ב**	**ב**

- **M2 ו-M1** לא שונים זה מזה ($0.33 < 0.88$), ושניהם שונים מכל השאר
- **M3, M5 ו-M4** לא שונים זה מזה: הטווח של שלושתם הוא $R_3 = 0.93 > 0.23$
- רמות עם אותה **אות** לא שונות זו מזו באופן מובהק. לפעמים מסמנים את אותו הדבר בקו מתחת לממוצעים: M2 M1 M3 M5 M4

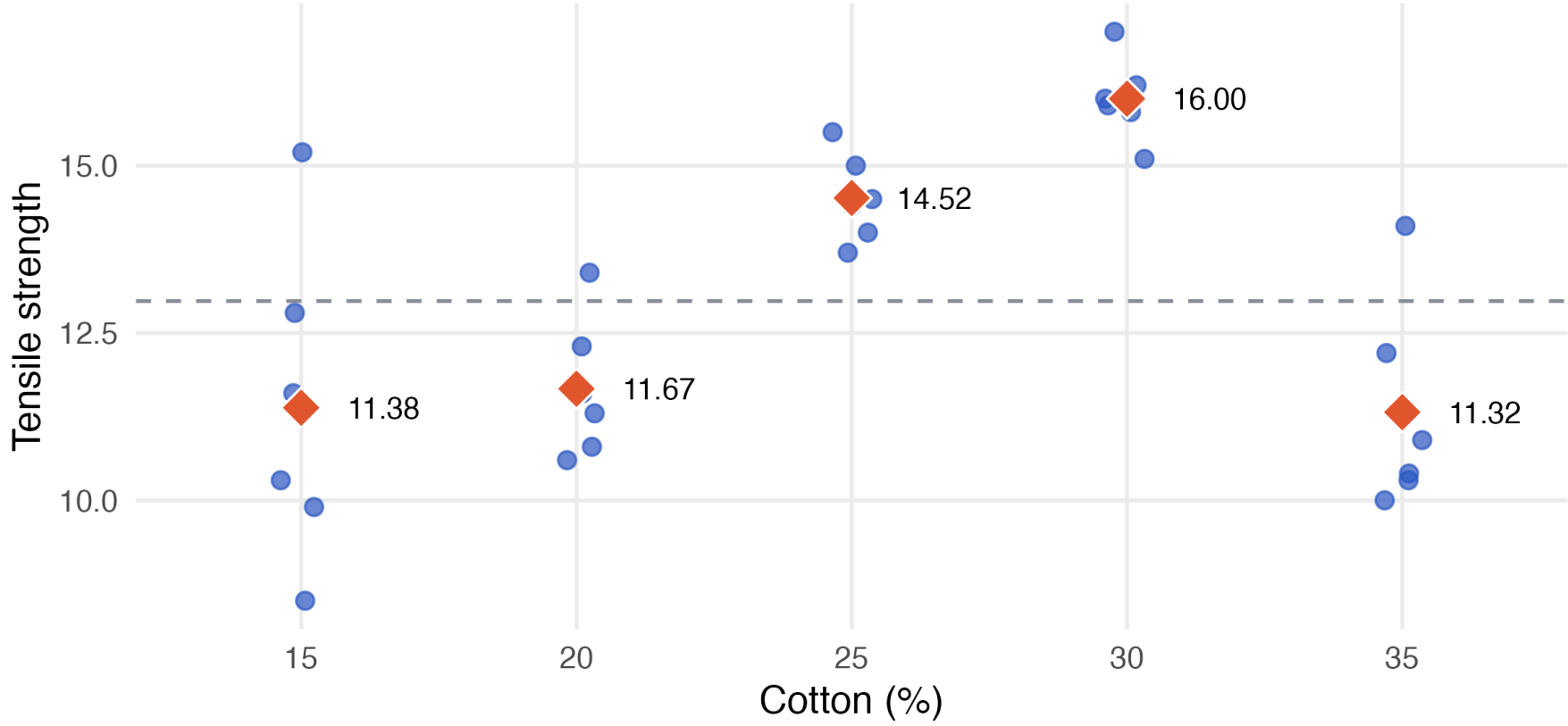
ההמלצה: להפעיל את אחד מ-M3, M4 או M5. משך הטיפול לא מבדיל ביניהם, ולכן הבחירה תלויה בשיקולים אחרים, למשל עלות: M3 הוא המודל הבינוני, M4 ו-M5 הגדולים והיקרים.

חוזק הבד: הנתונים

מפעל טקסטיל בודק איזה אחוז כותנה בבד נותן את החוזק הגבוה ביותר. 5 רמות (15%–35%), 6 דגימות בכל רמה. התגובה: חוזק מתיחה.

חוזק הבד לפי אחוז הכותנה

נקודות: 6 דגימות בכל רמה · מעוין: ממוצע הרמה · קו מקווקו: הממוצע הכללי



תרגיל: Duncan על חוזק הבד

מפעל טקסטיל בודק 5 רמות של אחוז כותנה בבד, 6 דגימות בכל רמה. התגובה: חוזק מתיחה. ה-F מובהק.
אילו רמות שונות זו מזו?

```r
fit_cot <- aov(strength ~ cotton_pct, data = cotton)
summary(fit_cot)
```

```
            Df Sum Sq Mean Sq F value  Pr(>F)
cotton_pct   4  111.1    27.8    13.9 4.3e-06 ***
Residuals   25   50.1     2.0
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
sort(tapply(cotton$strength, cotton$cotton_pct, mean), decreasing = TRUE)
```

```
   30    25    20    15    35
16.00 14.52 11.67 11.38 11.32
```

השלימו: $s_{\bar{y}}$, $R_2, \ldots, R_5$, כל השלבים, והחלוקה לקבוצות.

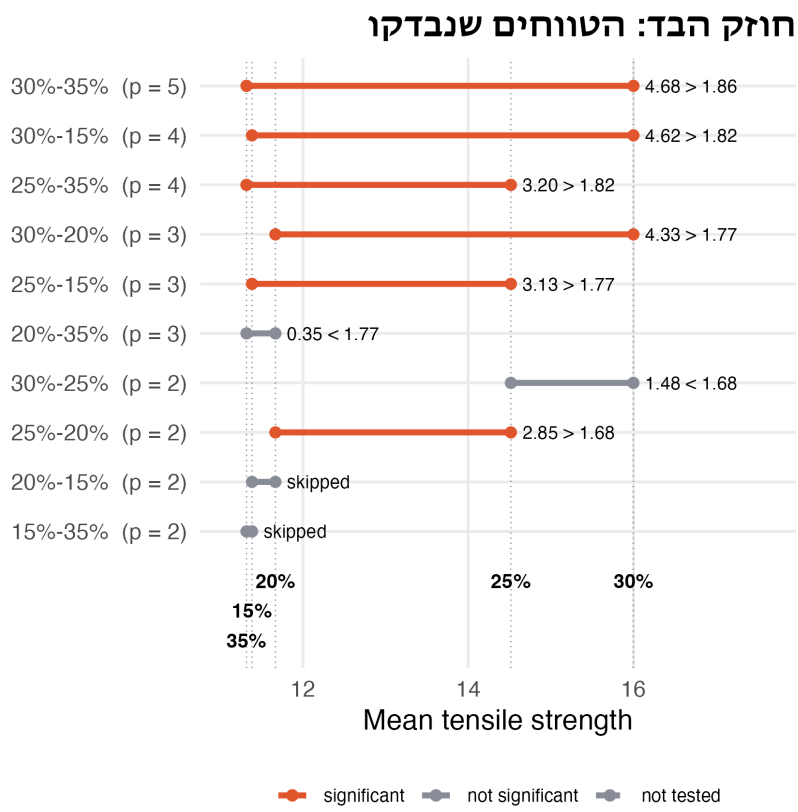
רמז: לא צריך את הנתונים הגולמיים. כל מה שצריך נמצא בפלט: MS_E ו-f בשורת `Residuals`, $n = 6$ מהתיאור, והממוצעים בשורה האחרונה.

פתרון, שלב 1: מה לוקחים מהפלט

- **MS_E:** עמודת `Mean Sq` בשורת `Residuals`: $MS_E = 2$. זה האומד של σ^2, ולכן לא צריך את התצפיות עצמן
- **f:** עמודת `Df` באותה שורה: $f = N - a = 30 - 5 = 25$
- **שגיאת התקן של ממוצע:** $s_{\bar{y}} = \sqrt{MS_E/n} = \sqrt{2/6} = 0.578$
- **המקדמים:** בטבלה, שורה $f = 25$: $r_{0.05}(2, \ldots, 5) = 2.91,\ 3.06,\ 3.15,\ 3.22$
- **הספים:** $R_p = r_p \cdot s_{\bar{y}}$, כלומר $R_2, \ldots, R_5 = 1.68,\ 1.77,\ 1.82,\ 1.86$
- **מיון הממוצעים** מהגדול לקטן (המרחק p נקבע לפי הסדר הזה):

הרמה	30%	25%	20%	15%	35%
הממוצע	16.00	14.52	11.67	11.38	11.32
המיקום	1	2	3	4	5

פתרון, שלב 2: מהקצוות פנימה



- $p = 5$ (הטווח כולו), סף $R_5 = 1.86$:
  - 30–35: $4.68 > 1.86$ **מובהק**
- $p = 4$, סף $R_4 = 1.82$:
  - 30–15: $4.62 > 1.82$ **מובהק**
  - 25–35: $3.20 > 1.82$ **מובהק**
- $p = 3$, סף $R_3 = 1.77$:
  - 30–20: $4.33 > 1.77$ **מובהק**
  - 25–15: $3.13 > 1.77$ **מובהק**
  - 20–35: $0.35 < 1.77$ **לא מובהק**. לכן 20%, 15% ו-35% נחשבים שווים
- $p = 2$, סף $R_2 = 1.68$:
  - 25–30: $1.48 < 1.68$ **לא מובהק**
  - 20–25: $2.85 > 1.68$ **מובהק**
  - 15–20, 15–35: **לא נבדקים** (בתוך 20–35, כלל ההגנה)

פתרון, שלב 3: הקבוצות

	30%	25%	20%	15%	35%
הממוצע	16.00	14.52	11.67	11.38	11.32
הקבוצה	**א**	**א**	**ב**	**ב**	**ב**

- **30% ו-25%** לא שונים זה מזה, ושניהם חזקים יותר מכל השאר
- **20%, 15% ו-35%** לא שונים זה מזה
- **המסקנה המעשית:** בד חזק דורש 25%–30% כותנה. הנתונים לא מבדילים בין שתי הרמות האלה, ולכן אפשר לבחור ביניהן לפי שיקול אחר (למשל עלות)
- **מה עלול להשתבש בבחינה:**
  - למיין לפי אחוז הכותנה במקום לפי הממוצעים, ולקבל p שגוי
  - לקחת את המקדם מהשורה $f = N - 1$ במקום $f = N - a$
  - לבדוק זוגות פנימיים אחרי שהטווח שמכיל אותם לא היה מובהק

ביקורת על Duncan

- α_p גדלה עם מספר הממוצעים: עם 5 רמות, $\alpha_5 = 1 - 0.95^4 = 0.185$
- כלומר, ה-FWER של השיטה **גבוה מ-5%**, וגדל עם מספר הרמות. עם 10 רמות, $\alpha_{10} = 0.37$
- השיטה "מוצאת" יותר הבדלים, וחלקם שגויים
- בעבודה מעשית מקובל להשתמש ב-Tukey

שיטת Tukey (HSD)

הרעיון

Tukey's Honest Significant Difference: סף **אחד** לכל הזוגות, שמבטיח FWER $= \alpha$ בדיוק (בניסוי מאוזן).

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

- $q_{\alpha}(a, f)$: הערך הקריטי של התפלגות ה-studentized range של **כל** a הממוצעים, עם $f = N - a$ דרגות חופש
- **ההיגיון:** אם הזוג **הרחוק ביותר** לא חוצה את הסף תחת H_0 בהסתברות $1 - \alpha$, אף זוג אחר לא חוצה אותו. לכן ההסתברות לטעות אחת לפחות היא α
- **הקשר ל-Duncan:** זו "הגישה המחמירה" שראינו קודם. אותה התפלגות, אבל תמיד עם $p = a$ ועם α (לא α_p). לכן גם בקצוות הסף של Tukey (4.04) גבוה מזה של Duncan (3.17), ובזוגות הפנימיים הפער גדול עוד יותר

רווחי סמך משותפים לכל ההפרשים: $(\bar{y}_{i\cdot} - \bar{y}_{k\cdot}) \pm HSD$. **בניסוי לא מאוזן** (Tukey–Kramer):
$q_{\alpha}(a, f)\sqrt{\frac{MS_E}{2}\left(\frac{1}{n_i} + \frac{1}{n_k}\right)}$.

המודלים: חישוב ידני

$$HSD = q_{0.05}(5, 40) \cdot 0.308 = 4.04 \cdot 0.308 = 1.244$$

```
c(q = qtukey(0.95, nmeans = 5, df = 40), HSD = qtukey(0.95, 5, 40) * sqrt(mse / 9))
```

```
    q   HSD
4.039 1.244
```

```
duncan |> mutate(tukey = diff > 1.244) |> select(high, low, diff, duncan = significant, tukey)
```

```
# A tibble: 10 × 5
   high  low     diff duncan tukey
   <chr> <chr>  <dbl> <lgl>  <lgl>
 1 M2    M4    1.57   TRUE   TRUE
 2 M2    M5    1.38   TRUE   TRUE
 3 M1    M4    1.24   TRUE   FALSE
 4 M2    M3    1.34   TRUE   TRUE
 5 M1    M5    1.05   TRUE   FALSE
 6 M3    M4    0.231  FALSE  FALSE
 7 M2    M1    0.332  FALSE  FALSE
 8 M1    M3    1.01   TRUE   FALSE
 9 M3    M5    0.0389 FALSE  FALSE
10 M5    M4    0.192  FALSE  FALSE
```

Tukey ב-R

```
TukeyHSD(fit_ai)
```

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = aht ~ model, data = calls)

$model
          diff     lwr       upr  p adj
M2-M1  0.33222 -0.9117  1.576159 0.9397
M3-M1 -1.00778 -2.2517  0.236159 0.1617
M4-M1 -1.23889 -2.4828  0.005048 0.0514
M5-M1 -1.04667 -2.2906  0.197270 0.1354
M3-M2 -1.34000 -2.5839 -0.096063 0.0292
M4-M2 -1.57111 -2.8150 -0.327174 0.0072
M5-M2 -1.37889 -2.6228 -0.134952 0.0233
M4-M3 -0.23111 -1.4750  1.012826 0.9836
M5-M3 -0.03889 -1.2828  1.205048 1.0000
M5-M4  0.19222 -1.0517  1.436159 0.9918
```

איך קוראים את הפלט

- **`diff`**: ההפרש בין הממוצעים (השורה − העמודה). `M2-M1 = 0.33`: עם M2 הטיפול ארוך מאשר עם M1 ב-0.33 דקות
- **`lwr`, `upr`**: רווח סמך **משותף** של 95%: $\text{diff} \pm 1.244$. ב-95% מהניסויים **כל** עשרת הרווחים יכילו את ההפרש האמיתי יחד
- **`p adj`**: p-value מתוקן להשוואות מרובות. מובהק אם קטן מ-0.05, כלומר אם הרווח לא כולל את 0
- **שימו לב ל-`M4-M1`:** ההפרש 1.239, וה-HSD הוא 1.244. `p adj = 0.051`. **לא** מובהק, בהפרש של חמש אלפיות דקה

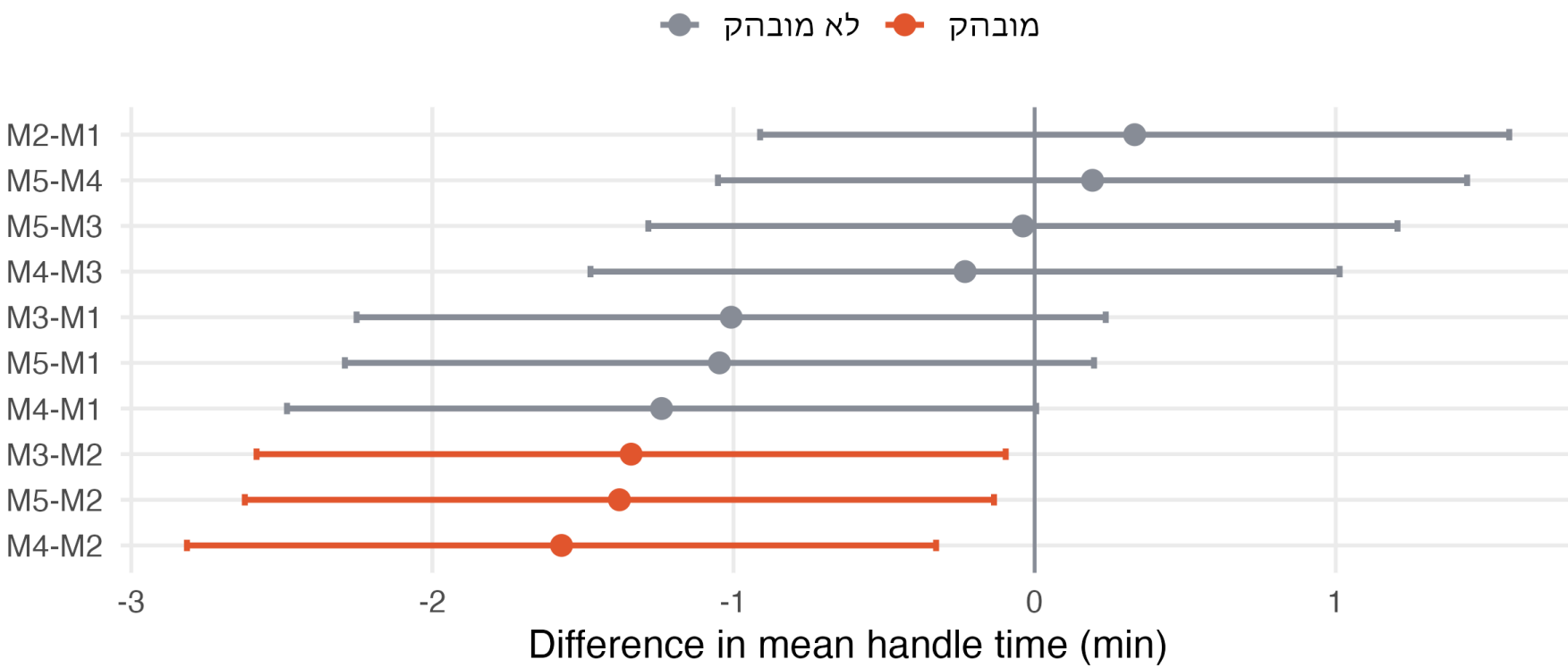
רווחי הסמך בתרשים

```
TukeyHSD(fit_ai)$model |>
  as_tibble(rownames = "pair") |>
  mutate(pair = fct_reorder(pair, diff), significant = `p adj` < 0.05) |>
  ggplot(aes(diff, pair, colour = significant)) +
  geom_vline(xintercept = 0, colour = nanas_pal[["grey"]]) +
  geom_errorbar(aes(xmin = lwr, xmax = upr), width = 0.25, linewidth = 1, orientation = "y") +
  geom_point(size = 3) +
  scale_colour_manual(values = c(`TRUE` = nanas_pal[["accent"]], `FALSE` = nanas_pal[["grey"]]),
                      labels = c(`TRUE` = "מובהק", `FALSE` = "לא מובהק")) +
  labs(title = "שלושה זוגות מובהקים לפי Tukey: M2 מול M3, M4 ו־M5",
       subtitle = "רווחי סמך משותפים 95% להפרשים · רווח שלא חוצה את 0 הוא מובהק",
       x = "Difference in mean handle time (min)", y = NULL, colour = NULL)
```

רווחי הסמך בתרשים

שלושה זוגות מובהקים לפי Tukey: M2 מול M3, M4 ו-M5

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



Duncan מול Tukey

זוג	הפרש	Duncan (R_p)	Tukey (1.244)
M2–M4	1.57	✓ (0.98)	✓
M2–M5	1.38	✓ (0.96)	✓
M2–M3	1.34	✓ (0.93)	✓
M1–M4	1.24	✓ (0.96)	✗
M1–M5	1.05	✓ (0.93)	✗
M1–M3	1.01	✓ (0.88)	✗
M2–M1, M3–M4, M3–M5, M5–M4	≤ 0.33	✗	✗

- Duncan מוצא 6 זוגות. Tukey מוצא 3. ב-Tukey, M1 "תלוי באמצע": לא שונה מ-M2, וגם לא מ-M3, M4, M5
- **אין כאן "נכון" ו"לא נכון".** זו בחירה בין שתי טעויות: Tukey שומר על הסיכוי לתגלית שגויה **אחת** מתחת ל-5%, במחיר של עוצמה

איזו שיטה מתי?

המצב	השיטה	ב-R
מעט שאלות, מתוכננות מראש	ניגודים, α רגילה	ידני, `pf()`
כל הזוגות, אחרי F מובהק	Tukey HSD	`TukeyHSD()`
כל הזוגות, כשהבחינה דורשת	Duncan	`qtukey()` + טבלה

הכלל החשוב ביותר: מחליטים על השיטה **לפני** שמסתכלים על התוצאות. אם בוחרים את השיטה שנותנת את התוצאה ה"נכונה", כל התיקון לא שווה כלום.

ניתוח שונות א-פרמטרי: Kruskal–Wallis

מתי צריך מבחן א-פרמטרי?

הדוגמה: זמן התגובה של API בארבעה אזורי ענן (regions), 12 בקשות מכל אזור. זמני תגובה מוטים ימינה, עם ערכים חריגים.

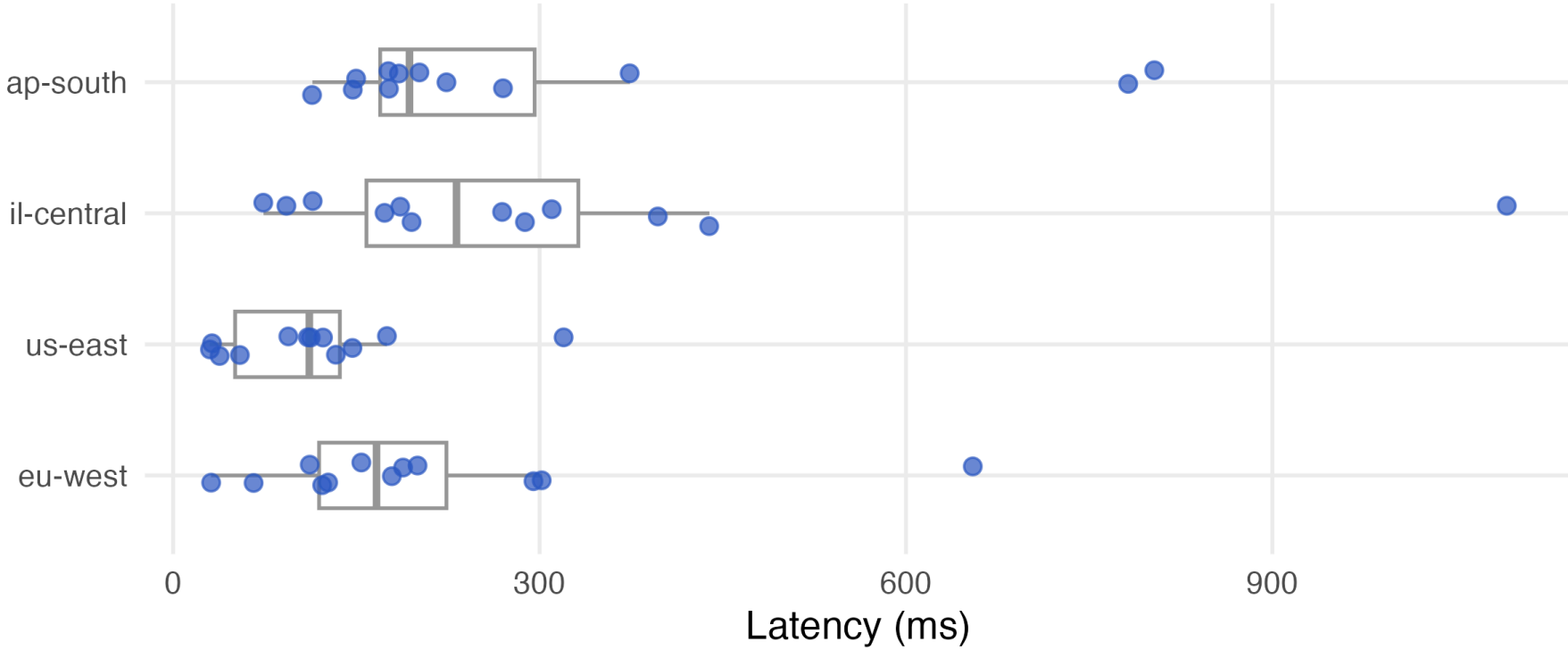
```
set.seed(162)
latency <- tibble(region = factor(rep(c("eu-west", "us-east", "il-central", "ap-south"), each = 12),
                                  levels = c("eu-west", "us-east", "il-central", "ap-south")),
                  ms = round(rlnorm(48, log(rep(c(150, 155, 210, 220), each = 12)), 0.8)))

ggplot(latency, aes(ms, region)) +
  geom_boxplot(width = 0.5, outlier.shape = NA, colour = "grey60") +
  geom_point(colour = nanas_pal[["primary"]], size = 2.5, alpha = 0.7,
             position = position_jitter(height = 0.1, seed = 3)) +
  labs(title = "זמני תגובה מוטים, עם כמה בקשות איטיות מאוד",
       subtitle = "12 בקשות מכל אזור",
       x = "Latency (ms)", y = NULL)
```

מתי צריך מבחן א-פרמטרי?

זמני תגובה מוטים, עם כמה בקשות איטיות מאוד

12 בקשות מכל אזור



ANOVA על נתונים מוטים

```r
fit_lat <- aov(ms ~ region, data = latency)
summary(fit_lat)
```

```
            Df  Sum Sq Mean Sq F value Pr(>F)
region       3  292464   97488    2.35  0.085 .
Residuals   44 1823008   41432
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
ks.test(rstandard(fit_lat), "pnorm")              # standardised residuals vs N(0, 1)
```

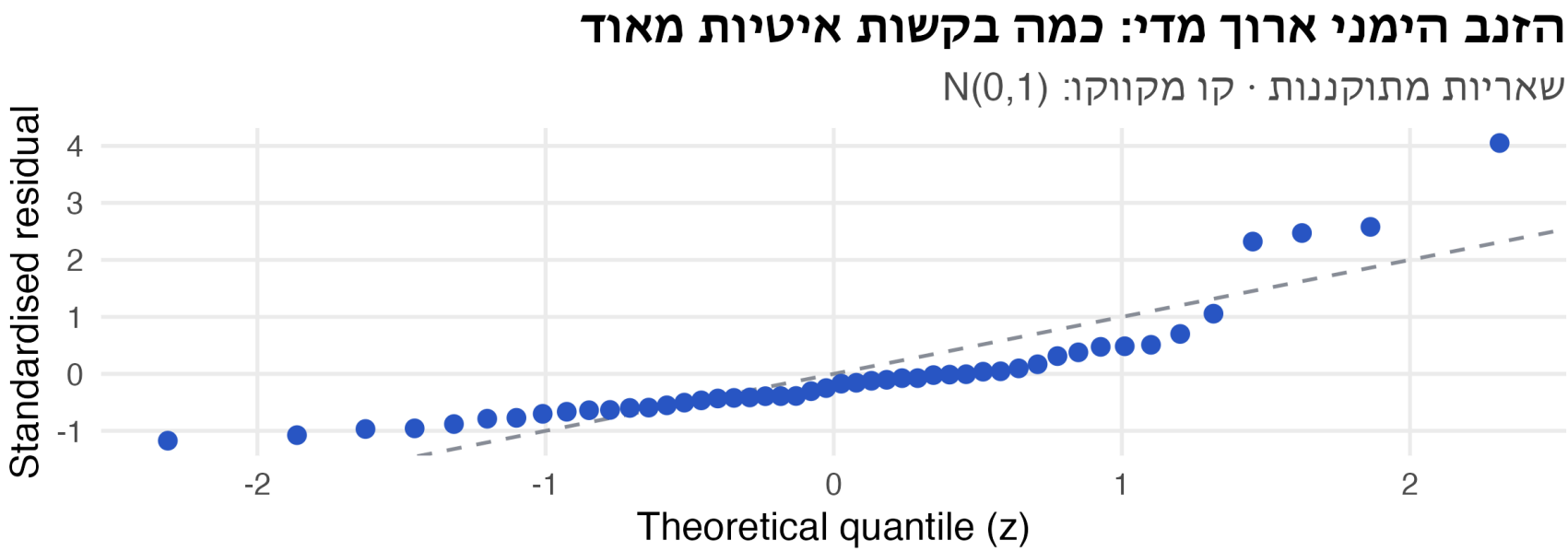
```
    Exact one-sample Kolmogorov-Smirnov test

data:  rstandard(fit_lat)
D = 0.21, p-value = 0.02
alternative hypothesis: two-sided
```

- ה-ANOVA **לא דוחה** ($p = 0.085$)
- **אבל השאריות לא נורמליות.** מבחן Kolmogorov–Smirnov על השאריות המתוקננות, כמו בהרצאה 3: $D = 0.212 > 1.36/\sqrt{48} = 0.196$, ולכן **דוחים** נורמליות

השאריות בגרף Q-Q

```
tibble(r = sort(rstandard(fit_lat))) |>
  mutate(z = qnorm((row_number() - 0.5) / n())) |>
  ggplot(aes(z, r)) +
  geom_abline(intercept = 0, slope = 1, colour = nanas_pal["grey"], linetype = "dashed") +
  geom_point(colour = nanas_pal["primary"], size = 2.5) +
  labs(title = "הזנב הימני ארוך מדי: כמה בקשות איטיות מאוד",
       subtitle = "שאריות מתוקננות · קו מקווקו: N(0,1)",
       x = "Theoretical quantile (z)", y = "Standardised residual")
```



- הבקשות האיטיות מנפחות את MS_E ו"מסתירות" את ההבדלים בין האזורים
- **שתי אפשרויות:** טרנספורמציה $(\log)$, או מבחן שלא מניח נורמליות

Kruskal–Wallis: הרעיון

הכללה של מבחן סכום הדירוגים (הרצאה 1) ליותר משני מדגמים.

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij}$$

- **לא** מניחים $\varepsilon_{ij} \sim N(0, \sigma^2)$. מניחים רק שכל ה-$\varepsilon_{ij}$ מגיעים מאותה התפלגות **רציפה**
- $H_0 : \tau_1 = \cdots = \tau_a = 0$ (כל הרמות מאותה התפלגות) מול H_1: לפחות רמה אחת מוזזת
- **השיטה:** מדרגים את כל N התצפיות יחד, מ-1 (הקטנה) עד N (הגדולה), ומריצים "ANOVA על הדרגות"

תחת H_0 הדרגות מתפזרות באקראי בין הרמות. כמו במבחן סכום הדירוגים, כל דרגה מתפלגת אחיד על $\{1, \ldots, N\}$:

$$E(R_{ij}) = \frac{N+1}{2} \qquad V(R_{ij}) = \frac{N^2 - 1}{12}$$

הסטטיסטי

ב-ANOVA (בהנחת נורמליות):

$$\frac{SS_{Tr}}{\sigma^2} = \frac{\sum_{i=1}^{a} n_i(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2}{\sigma^2} \sim \chi^2_{a-1}$$

ב-Kruskal–Wallis: אותו ביטוי, על הדרגות. $\bar{R}_{i\cdot}$ הוא ממוצע הדרגות ברמה i, ו-σ^2 מוחלף בשונות הידועה של דרגה, $\frac{N(N+1)}{12}$ (בקירוב):

$$K = \frac{12}{N(N+1)} \sum_{i=1}^{a} n_i \left(\bar{R}_{i\cdot} - \frac{N+1}{2} \right)^2 = \frac{12}{N(N+1)} \sum_{i=1}^{a} \frac{R_{i\cdot}^2}{n_i} - 3(N+1)$$

- $R_{i\cdot}$ הוא **סכום** הדרגות ברמה i
- תחת H_0, $K \sim \chi^2_{a-1}$ בקירוב (טוב כש-$n_i \geq 5$)
- דוחים את H_0 אם $K > \chi^2_{1-\alpha,\, a-1}$
- **שימו לב:** אין כאן σ^2 לאמוד. השונות של הדרגות ידועה מראש

דוגמה לחישוב ידני: חוזק הבד

אותם נתונים מהתרגיל על Duncan. מדרגים את כל 30 הדגימות יחד:

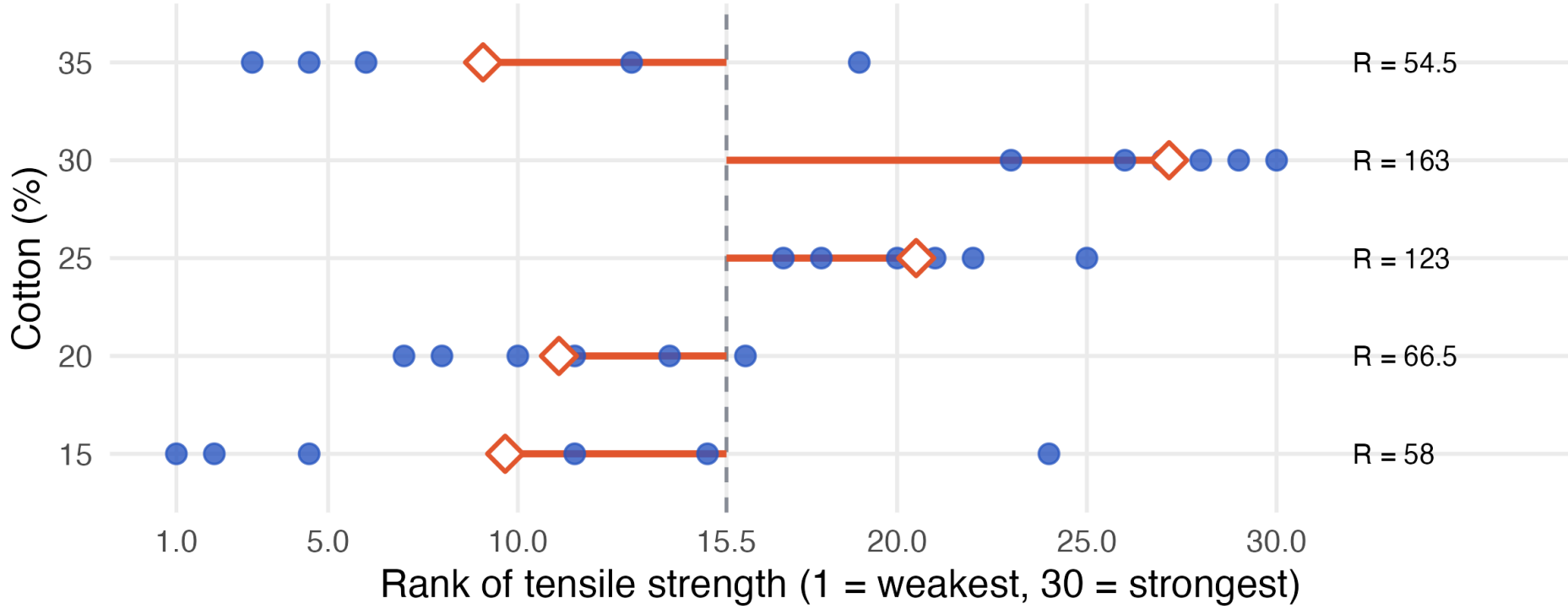
```
cotton |>
  group_by(cotton_pct) |>
  summarise(n = n(), strength = paste(sort(strength), collapse = ", "),
            R = sum(rank), R_bar = mean(rank))
```

```
# A tibble: 5 × 5
  cotton_pct     n strength                                R R_bar
  <fct>      <int> <chr>                                <dbl> <dbl>
1 15             6 8.5, 9.9, 10.3, 11.6, 12.8, 15.2    58    9.67
2 20             6 10.6, 10.8, 11.3, 11.6, 12.3, 13.4  66.5 11.1
3 25             6 13.7, 14, 14.4, 14.5, 15, 15.5     123   20.5
4 30             6 15.1, 15.8, 15.9, 16, 16.2, 17     163   27.2
5 35             6 10, 10.3, 10.4, 10.9, 12.2, 14.1    54.5  9.08
```

הדרגות בתרשים

ב-25% וב-30% הדרגות גבוהות, ב-15%, 20% ו-35% נמוכות

נקודה: דרגה של דגימה · מעוין: ממוצע הדרגות · קו מקווקו: 15.5, הממוצע הצפוי תחת H0



- K מסכם את **אורכי הקווים הכתומים** (בריבוע, משוקללים ב-n_i): $K = \frac{12}{N(N+1)} \sum n_i(\bar{R}_{i\cdot} - 15.5)^2$. תחת H_0
  כל המעוינים קרובים לקו המקווקו

חישוב K

$N = 30$, $a = 5$, $n_i = 6$. סכומי הדרגות: $R_{i\cdot} = 58,\ 66.5,\ 123,\ 163,\ 54.5$.

$$K = \frac{12}{30 \cdot 31} \cdot \frac{58^2 + 66.5^2 + 123^2 + 163^2 + 54.5^2}{6} - 3 \cdot 31 = 19.81$$

$$K = 19.81 > \chi^2_{0.95,\,4} = 9.49$$

דוחים את H_0: אחוז הכותנה משפיע על חוזק הבד.

```
kruskal.test(strength ~ cotton_pct, data = cotton)
```

```
    Kruskal-Wallis rank sum test

data:  strength by cotton_pct
Kruskal-Wallis chi-squared = 20, df = 4, p-value = 5e-04
```

בחזרה לזמני התגובה

תזכורת: זמן התגובה של API בארבעה אזורי ענן, 12 בקשות מכל אזור. הנתונים מוטים ימינה, עם כמה בקשות איטיות מאוד. ANOVA על הנתונים הגולמיים **לא דחה** ($p = 0.085$), והשאריות לא נורמליות. עכשיו ננסה את Kruskal–Wallis:

```
kruskal.test(ms ~ region, data = latency)
```

```
    Kruskal-Wallis rank sum test

data:  ms by region
Kruskal-Wallis chi-squared = 12, df = 3, p-value = 0.009
```

```
fit_lat_log <- aov(log(ms) ~ region, data = latency)
summary(fit_lat_log)
```

```
             Df Sum Sq Mean Sq F value Pr(>F)
region        3   7.42   2.473    4.75 0.0059 **
Residuals    44  22.91   0.521
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- **Kruskal–Wallis דוחה** ($p = 0.0085$): יש הבדל בין האזורים
- גם ANOVA על $\log(\text{ms})$ דוחה. הלוגריתם "מכווץ" את הזנב הימני ומקרב לנורמליות (השקף הבא)
- **הדרגות לא רגישות לערכים חריגים:** בקשה של 1,000 מילישניות היא פשוט "הדרגה הגבוהה ביותר", בדיוק כמו בקשה של 400

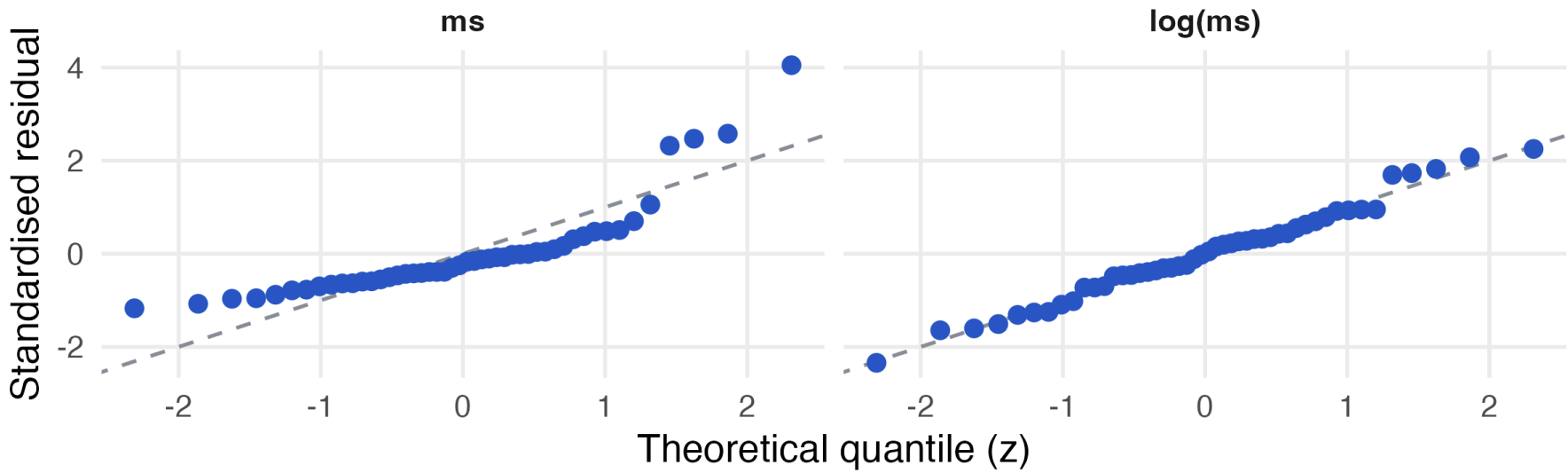
מה הלוגריתם עושה לשאריות?

```
bind_rows(tibble(scale = "ms", r = rstandard(fit_lat)),
          tibble(scale = "log(ms)", r = rstandard(fit_lat_log))) |>
  mutate(scale = fct_inorder(scale)) |>
  group_by(scale) |>
  arrange(r, .by_group = TRUE) |>
  mutate(z = qnorm((row_number() - 0.5) / n())) |>
  ggplot(aes(z, r)) +
  geom_abline(intercept = 0, slope = 1, colour = nanas_pal[["grey"]], linetype = "dashed") +
  geom_point(colour = nanas_pal[["primary"]], size = 2.5) +
  facet_wrap(~ scale) +
  labs(title = "אחרי הטרנספורמציה, השאריות יושבות על הקו",
       subtitle = "שאריות מתוקננות · קו מקווקו: N(0,1)",
       x = "Theoretical quantile (z)", y = "Standardised residual")
```

מה הלוגריתם עושה לשאריות?

אחרי הטרנספורמציה, השאריות יושבות על הקו

שאריות מתוקננות · קו מקווקו: N(0,1)




```
ks_log <- ks.test(rstandard(fit_lat_log), "pnorm")
c(D = unname(ks_log$statistic), p_value = ks_log$p.value)      # KS on log scale
```

```
      D p_value
0.06632 0.97494
```

אילו אזורים שונים? השוואות מרובות על דרגות

אחרי Kruskal–Wallis מובהק, משווים זוגות במבחן סכום הדירוגים, ומתקנים להשוואות מרובות:

```
pairwise.wilcox.test(latency$ms, latency$region, p.adjust.method = "bonferroni")
```

```
    Pairwise comparisons using Wilcoxon rank sum exact test

data:  latency$ms and latency$region

           eu-west us-east il-central
us-east    0.468   -       -
il-central 1.000   0.060   -
ap-south   1.000   0.003   1.000

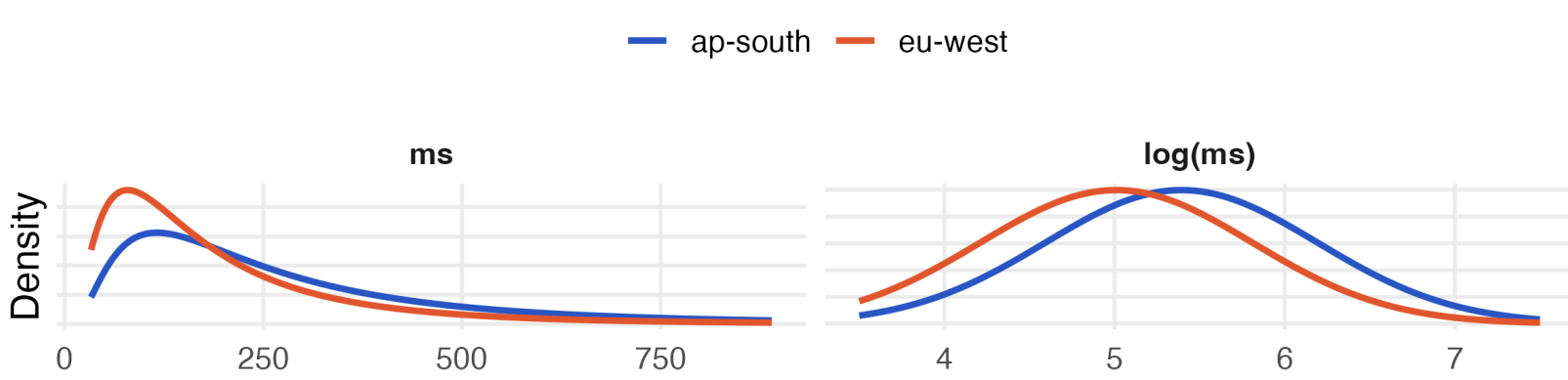
P value adjustment method: bonferroni
```

- כל תא הוא p-value של מבחן סכום הדירוגים בין שני אזורים, **מתוקן** בשיטת Bonferroni: מוכפל במספר ההשוואות (וחסום ב-1). דוחים אם הוא קטן מ-0.05
- **אבל האם זה מבחן על חציונים?** תלוי בהנחת ההזזה (השקף הבא)

מה בודק מבחן סכום הדירוגים? הנחת ההזזה

- **H_0 של המבחן:** לשתי הרמות **אותה התפלגות**. הוא דוחה כשתצפיות מרמה אחת נוטות להיות גדולות מתצפיות מהרמה השנייה
- **הנחת ההזזה** (shift): שתי ההתפלגויות זהות בצורה ובפיזור, ואחת פשוט **מוזזת** ימינה: $F_2(x) = F_1(x - \delta)$. רק תחתיה דחייה פירושה "החציונים שונים"
- **בלי ההנחה,** המבחן יכול לדחות גם כשהחציונים שווים, למשל כשלאחת הרמות פיזור גדול יותר

בסקאלה הלוגריתמית: אותה צורה, רק מוזזת



- **בדוגמה:** הנתונים סומלצו מהתפלגות **לוג-נורמלית**. ב-ms האזור האיטי גם **מפוזר** יותר, ולכן זו לא הזזה. ב-$\log(\text{ms})$ זו הזזה נקייה
- **והדרגות לא מושפעות מ-**log: הלוגריתם שומר על הסדר, ולכן המבחן על ms והמבחן על $\log(\text{ms})$ **זהים**. ההנחה מתקיימת בסקאלה הלוגריתמית, וחציון של log הוא log של החציון. לכן המסקנה על החציונים נכונה גם ב-ms

ניגודים עם חציונים

אחרי Kruskal–Wallis, איך אומדים ניגוד $\Gamma = \sum c_i \tau_i$ בלי ממוצעים?

- **הניסיון הטבעי:** $\hat{\tau}_i = \text{median}_j\{y_{ij}\} - \text{median}_{\text{all}}$
- **בממוצעים** $\sum \tau_i = 0$ היא רק מוסכמה (תמיד אפשר להעביר קבוע בין μ ל-τ_i), והאומדים מקיימים אותה מעצמם: הממוצע הכללי הוא ממוצע ממוצעי הקבוצות
- **בחציונים** החציון הכללי **אינו** ממוצע של חציוני הקבוצות, ולכן האומדים הטבעיים **לא** מקיימים את המוסכמה

דוגמה קטנה: $y_{1j} = \{3, 7, 8\}$ ו-$y_{2j} = \{0, 1, 2\}$. החציון הכללי הוא 2.5, וחציוני הקבוצות הם 7 ו-1:

$$\hat{\tau}_1 = 7 - 2.5 = 4.5 \qquad \hat{\tau}_2 = 1 - 2.5 = -1.5 \qquad \hat{\tau}_1 + \hat{\tau}_2 = 3 \neq 0$$

הפתרון: חציונים של הפרשים

הרעיון: במקום להשוות כל רמה ל"מרכז" משותף, אומדים **בכמה כל רמה גבוהה מכל רמה אחרת**, ורק אז ממצעים.

1. **לכל זוג רמות (i, k):** מחשבים את **כל** ההפרשים בין תצפית מרמה i לתצפית מרמה k ($n_i \cdot n_k$ הפרשים), ולוקחים את החציון שלהם:

$$Z_{ik} = \underset{j,\, m}{\text{median}} \{ y_{ij} - y_{km} \}$$

Z_{ik} עונה על השאלה: **"בכמה, בדרך כלל, תצפית מרמה i גבוהה מתצפית מרמה k?"**

2. **מסדרים בטבלה** $a \times a$. על האלכסון $Z_{ii} = 0$, ו-$Z_{ki} = -Z_{ik}$: אם i גבוהה מ-k ב-6, אז k נמוכה מ-i ב-6

3. **האפקט של רמה** הוא ממוצע השורה שלה, "בכמה היא גבוהה מהרמות, בממוצע":

$$\hat{\tau}_i = \frac{1}{a} \sum_{k=1}^{a} Z_{ik}$$

4. **הסכום אפס:** כל זוג $Z_{ik} + Z_{ki} = 0$ מתבטל, ולכן $\sum_i \hat{\tau}_i = 0$. הבעיה מהשקף הקודם נפתרה

   ובונוס: ההפרשים Z_{ik} עצמם לא מצטברים. בכותנה, $Z_{15,20} + Z_{20,25} = -0.65 + (-3) = -3.65$, אבל $Z_{15,25} = -3.55$. ההפרשים $\hat{\tau}_i - \hat{\tau}_k$ תמיד מצטברים

5. **הניגוד:** $\hat{\Gamma} = \sum c_i \hat{\tau}_i$

בדוגמה הקטנה

$y_{1j} = \{3, 7, 8\}$ ו-$y_{2m} = \{0, 1, 2\}$. כל 9 ההפרשים $y_{1j} - y_{2m}$:

	$y_{2m} = 0$	$y_{2m} = 1$	$y_{2m} = 2$
$y_{1j} = 3$	3	2	1
$y_{1j} = 7$	7	6	5
$y_{1j} = 8$	8	7	6

ממוינים: $1, 2, 3, 5, \mathbf{6}, 6, 7, 7, 8$. החציון: $Z_{12} = 6$, ולכן $Z_{21} = -6$.

	רמה 1	**רמה 2**	**$\hat{\tau}_i$ (ממוצע השורה)**
רמה 1	0	6	$(0 + 6)/2 = 3$
רמה 2	–6	0	$(-6 + 0)/2 = -3$

$\hat{\tau}_1 + \hat{\tau}_2 = 0$, וההפרש ביניהם, 6, הוא ההזזה האמודה בין הרמות.

ניגוד בדוגמת הכותנה

השאלה: האם הקצוות (15% ו-35%) חלשים מ-30%? $\Gamma = \frac{1}{2}(\tau_1 + \tau_5) - \tau_4$.

```
y_by_level <- split(cotton$strength, cotton$cotton_pct)
Z <- outer(seq_along(y_by_level), seq_along(y_by_level),
           Vectorize(\(i, k) median(outer(y_by_level[[i]], y_by_level[[k]], "-"))))
dimnames(Z) <- list(names(y_by_level), names(y_by_level))
round(Z, 2)                                    # Z[i, k] = median of y_i - y_k
```

```
     15    20    25   30    35
15 0.00 -0.65 -3.55 -5.0 -0.10
20 0.65  0.00 -3.00 -4.5  0.55
25 3.55  3.00  0.00 -1.5  3.60
30 5.00  4.50  1.50  0.0  5.10
35 0.10 -0.55 -3.60 -5.1  0.00
```

```
tau_hat <- rowMeans(Z)                         # row means; they sum to 0
round(c(tau_hat, sum = sum(tau_hat), Gamma = 0.5 * (tau_hat[[1]] + tau_hat[[5]]) - tau_hat[[4]]), 2)
```

```
   15    20    25    30    35   sum Gamma
-1.86 -1.26  1.73  3.22 -1.83  0.00 -5.06
```

$\hat{\Gamma} = \frac{1}{2}\big((-1.86) + (-1.83)\big) - 3.22 = -5.06$: בד עם 15% או 35% כותנה חלש בכ-5 יחידות מבד עם 30% כותנה.

ניגוד עם חציונים: איך בודקים?

$\hat{\Gamma} = -5.06$ הוא **אומדן**, לא מבחן. את המבחן עושים על אותו ניגוד ($c = (\frac{1}{2}, 0, 0, -1, \frac{1}{2})$), אבל עם **ממוצעי הדרגות** (כמו ב-Kruskal–Wallis), ומשווים להתפלגות **נורמלית**:

35%	30%	25%	20%	15%	
9.08	27.17	20.50	11.08	9.67	$\bar{R}_i$
0.5	−1	0	0	0.5	c_i

$$Z = \frac{\sum c_i \bar{R}_i}{\sqrt{\frac{N(N+1)}{12}\sum \frac{c_i^2}{n_i}}} = \frac{-17.79}{4.4} = -4.04$$

$|Z| > 1.96$, ולכן **דוחים** $H_0 : \Gamma = 0$ ($p = 5.3 \times 10^{-5}$). **בשורה התחתונה:** הקצוות חלשים מ-30%, בכ-5.1 יחידות חוזק.

מה יכול להשתבש בהשוואות מרובות

1. **לבחור את ההשוואה אחרי שרואים את הנתונים**, ולבדוק אותה כאילו תוכננה מראש. ניגוד מתוכנן חייב להיות מתוכנן
2. **לבחור את שיטת התיקון לפי התוצאה:** "ב-Tukey זה לא יצא מובהק, אז ננסה Duncan". זו הגרסה המתוחכמת של p-hacking
3. **לא לתקן בכלל:** להריץ מבחן t על כל הזוגות, ולהכריז על הזוג שיצא מובהק
4. **לפרש Kruskal–Wallis כמבחן על חציונים** בלי הנחת הזזה. אם לרמות יש צורות שונות (למשל פיזור שונה), המבחן בודק משהו אחר
5. **להתעלם מבלוקים:** כשאותן יחידות (ימים, משימות, משתמשים) מופיעות בכל הרמות. על זה עכשיו

ניתוח שונות דו-כיווני (Two-way ANOVA)

השאלה: RAG, ואיזה מודל?

צוות ה-AI במוקד בונה בוט שעונה על שאלות של לקוחות. שתי החלטות פתוחות:

- **גודל המודל:** קטן, בינוני או גדול. מודל גדול מדויק יותר, אבל יקר פי כמה בכל פנייה
- **RAG** (retrieval-augmented generation): לשלוף מסמכי מדיניות רלוונטיים ולהכניס אותם לפרומפט. עולה בזמן תגובה ובתשתית

השאלות העסקיות:

1. האם RAG משפר את הדיוק?
2. האם השיפור **תלוי בגודל המודל**?
3. האם מודל קטן עם RAG יכול להחליף מודל גדול, ולחסוך בעלויות?

הניסוי: כל אחד מ-$3 \times 2 = 6$ השילובים רץ 4 פעמים על סט הערכה (eval set) של 200 שאלות, כל פעם על מדגם שאלות אחר. התגובה: **אחוז התשובות הנכונות**.

למה לא שני ניסויים נפרדים?

הגישה הנאיבית (one factor at a time): ניסוי אחד שמשווה גדלים (בלי RAG), וניסוי שני שבודק RAG (עם המודל הבינוני).

הבעיות:

1. **מפספסים אינטראקציה:** אם RAG עוזר רק למודלים קטנים, ניסוי על המודל הבינוני ייתן תשובה חלקית, ולגבי המודל הגדול שגויה
2. **בזבוז:** בניסוי דו-כיווני **כל תצפית משמשת פעמיים**. כל אחת מ-24 הריצות תורמת גם להשוואת הגדלים (8 ריצות לכל גודל) וגם להשוואת RAG (12 ריצות לכל צד)

ניסוי עם כל השילובים של הרמות נקרא **ניסוי עצרתי** (factorial design), או **מודל מצטלב** (crossed): כל רמה של פקטור אחד מופיעה עם כל רמה של הפקטור השני.

אפקט ראשי ואינטראקציה: שתי דוגמאות

דיוק ממוצע (%) בשני עולמות דמיוניים:

עולם 1: בלי אינטראקציה

	כבוי RAG	פעיל RAG
מודל קטן	60	70
מודל גדול	75	85

RAG מוסיף 10 נקודות **לכל** מודל.

עולם 2: עם אינטראקציה

	כבוי RAG	פעיל RAG
מודל קטן	60	75
מודל גדול	80	80

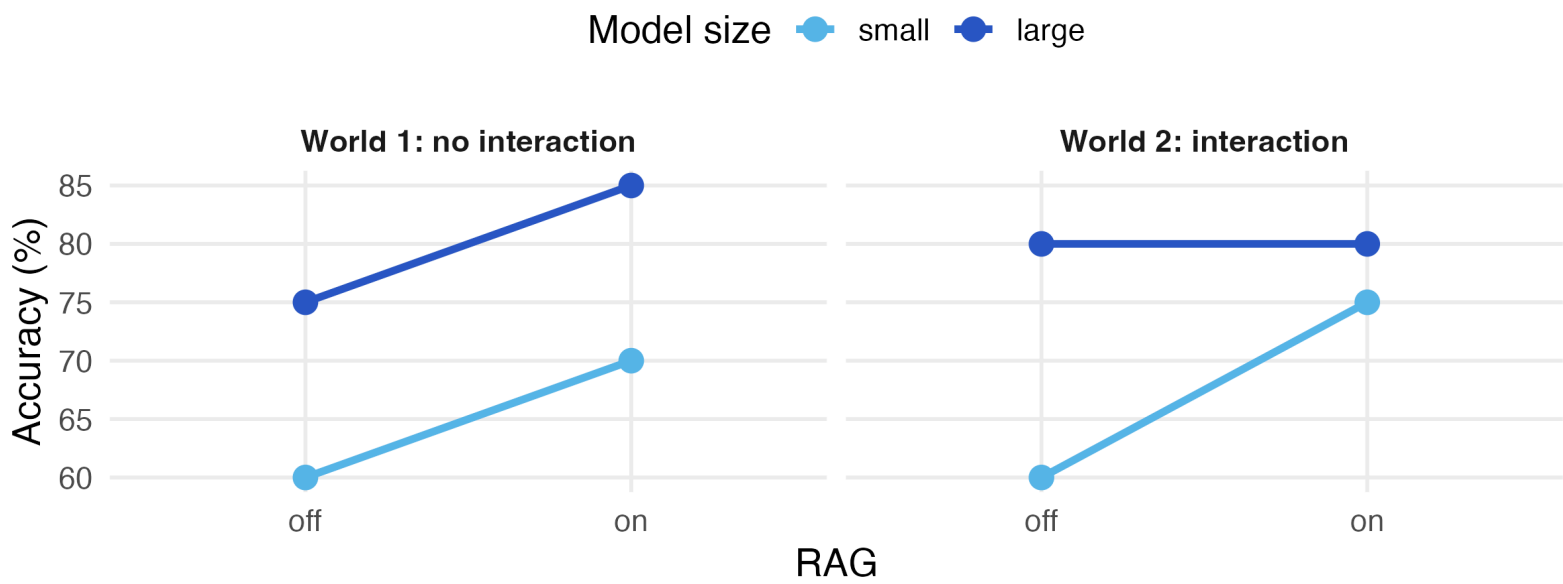
RAG מוסיף 15 נקודות לקטן, ו-0 לגדול.

- **אפקט ראשי** (main effect) של RAG: ההפרש בין הממוצעים, כשממצעים על פני רמות המודל. עולם 1: $\frac{70+85}{2} - \frac{60+75}{2} = 10$. עולם 2: $\frac{75+80}{2} - \frac{60+80}{2} = 7.5$
- **אינטראקציה** (interaction): ההשפעה של פקטור אחד **תלויה ברמה** של הפקטור השני. בעולם 2, "RAG מוסיף 7.5 נקודות" לא נכון לאף מודל

גרף אינטראקציה

קווים מקבילים = אין אינטראקציה

כל קו הוא מודל · השיפוע הוא האפקט של RAG על המודל הזה



- **אין אינטראקציה:** הקווים מקבילים. האפקט של RAG זהה לכל מודל, ואפשר לדבר על "האפקט של RAG"
- **יש אינטראקציה:** הקווים לא מקבילים (מתקרבים, מתרחקים או נחתכים). חייבים לענות "**תלוי** במודל"

הנתונים

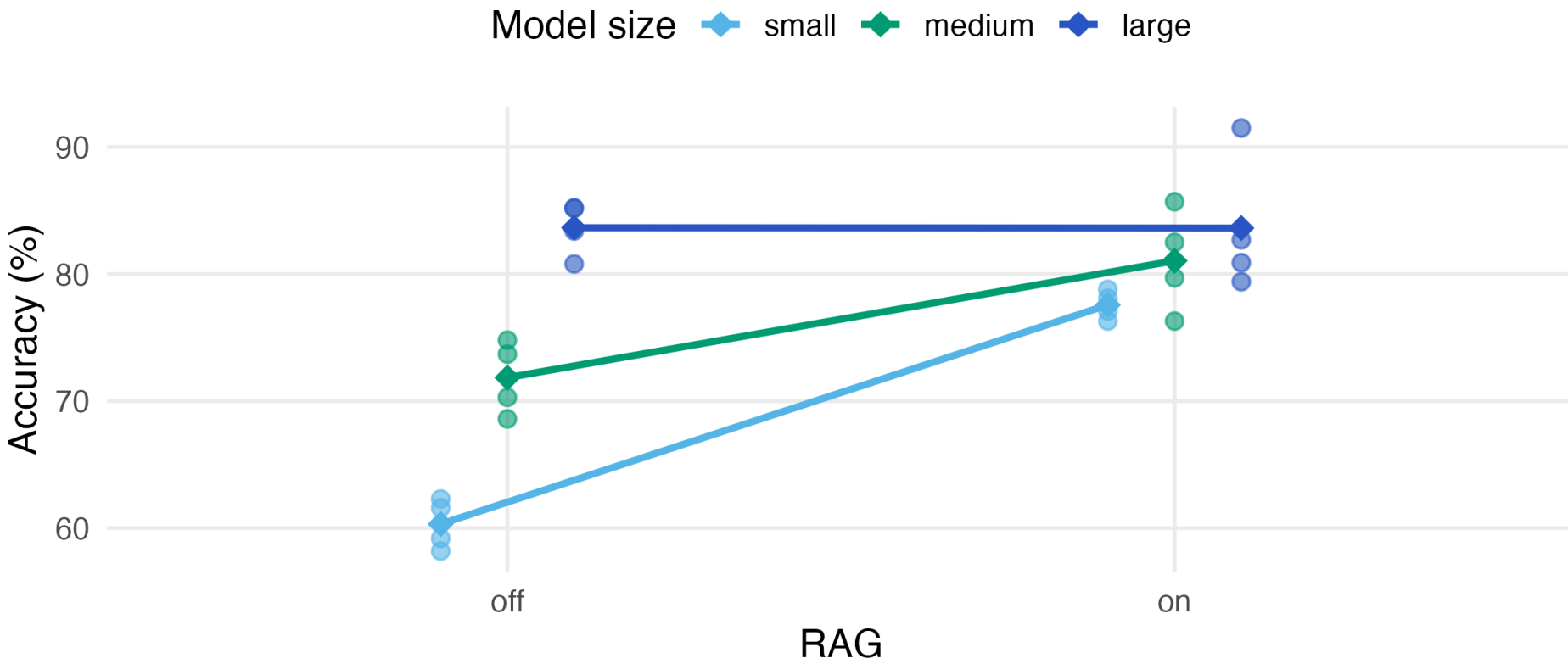
```
set.seed(606)
rag <- tribble(~size, ~retrieval, ~m,
  "small", "off", 62, "small", "on", 78, "medium", "off", 71,
  "medium", "on", 82, "large", "off", 80, "large", "on", 84) |>
  expand_grid(run = 1:4) |>
  mutate(acc = round(m + rnorm(n(), 0, 3), 1),
         size = factor(size, levels = c("small",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



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:

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

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

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

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

- **האינטראקציה:** כמה התא סוטה ממה שהמודל האדיטיבי צופה, $\bar{y}_{i..} + \bar{y}_{.j.} - \bar{y}_{...}$
- **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 $F_0 > F_{1-\alpha;\,(a-1)(b-1),\,ab(n-1)}$
- bn ב-$E(MS_A)$: כל ממוצע $\bar{y}_{i\cdot\cdot}$ מבוסס על bn תצפיות

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} הוא סכום הריבועים בין כל ab התאים, כלומר ה-SS_{Tr} של ניתוח חד-כיווני על 6 התאים. מתוכו, מה שלא מוסבר על ידי A ו-B לבד הוא האינטראקציה

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
```

- $\widehat{(\tau\beta)}_{small,\,on} > 0$, $\widehat{(\tau\beta)}_{large,\,on} < 0$: עם RAG, המודל הקטן מרוויח **יותר** ממה שהמודל האדיטיבי צופה, והגדול **פחות**

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**. Tukey על כל 6 התאים, עם ה-MS_E של המודל הדו-כיווני:

```
TukeyHSD(fit_rag, "size:retrieval")$`size:retrieval`[c("small:on-large:off", "large:on-small:on"), ]
```

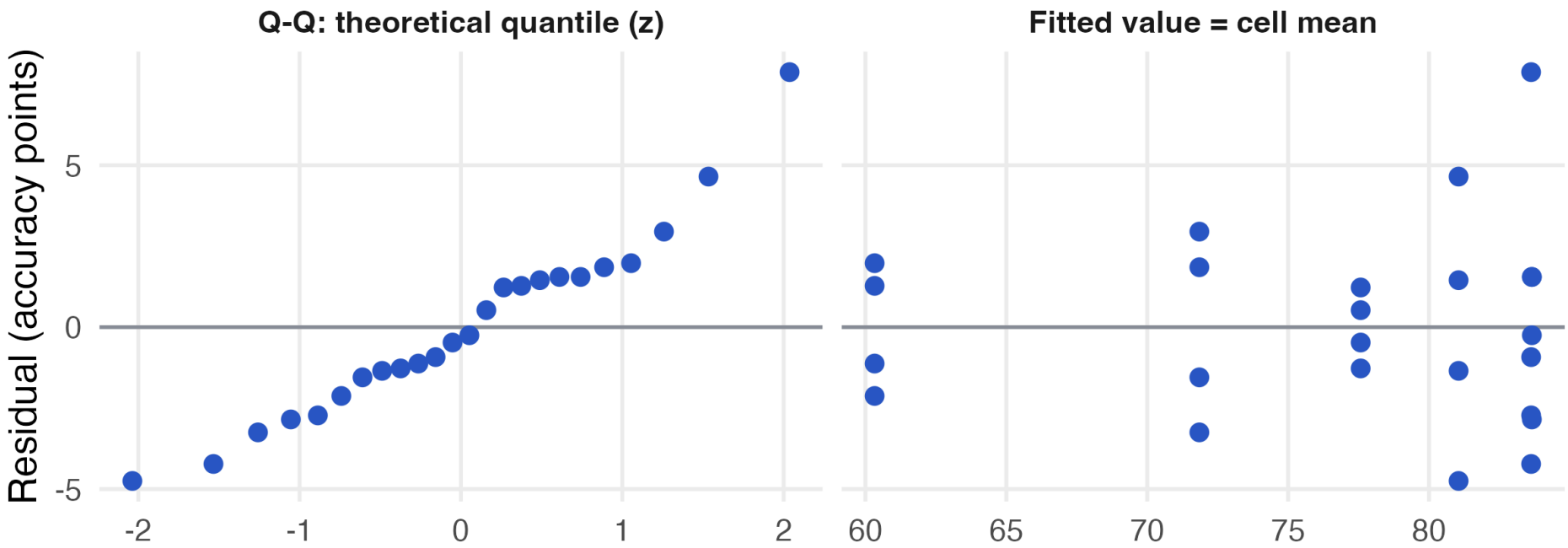
```
                     diff     lwr    upr  p adj
small:on-large:off -6.075 -13.362  1.212 0.1355
large:on-small:on   6.050  -1.237 13.337 0.1382
```

- `small:on-large:off`: הקטן עם RAG מפגר אחרי הגדול בלי RAG בכ-6 נקודות, ו**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 יותר מבדיקות

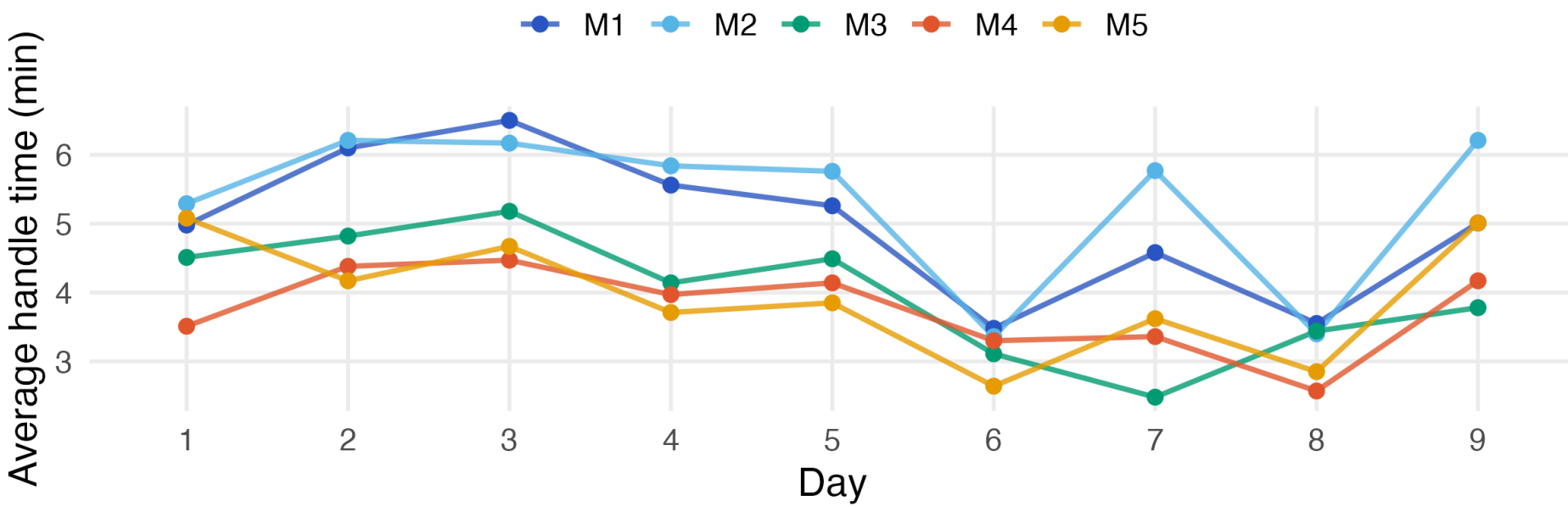
בלוקים: ניתוח דו-כיווני בלי חזרות

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:** 8 דרגות חופש של השגיאה (40 → 32). כשהבלוק באמת משנה, 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?

- **מתייגים:** במאגר יש כ-40. **מגרילים 4**
- **סוגי משימות:** בקטלוג יש עשרות סוגים (סיווג כוונה, זיהוי ישויות, דירוג תשובות…). **מגרילים 3**
- כל מתייג מתייג **2 פריטים** מכל סוג משימה. התגובה: שניות לפריט

```
set.seed(707)
annot <- expand_grid(annotator = paste0("A", 1:4), task = paste0("T", 1:3)) |>
  mutate(u_a = rep(rnorm(4, 0, 3), each = 3),      # annotator effect, sd 3 s
         u_t = rep(rnorm(3, 0, 8), times = 4),     # task-type effect, sd 8 s
         u_at = rnorm(12, 0, 2)) |>                 # interaction, sd 2 s
  expand_grid(item = 1:2) |>
  mutate(sec = round(40 + u_a + u_t + u_at + rnorm(n(), 0, 2.5), 1)) |>
  select(annotator, task, item, sec)
```

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. קודם נתקן ידנית (כמו בבחינה), ובשקף הבא נראה איך אומרים ל-R:

```r
ms <- summary(fit_ann)[[1]][["Mean Sq"]]; dfs <- summary(fit_ann)[[1]][["Df"]]   # A, B, AB, E
F_A <- ms[1] / ms[3]; F_B <- ms[2] / ms[3]
rbind(annotator = c(F = F_A, p = pf(F_A, dfs[1], dfs[3], lower.tail = FALSE)),
      task      = c(F = F_B, p = pf(F_B, dfs[2], dfs[3], lower.tail = FALSE)))
```

```
                F         p
annotator    5.58 3.596e-02
task       163.83 5.815e-06
```

```r
vc <- c(annotator = (ms[1] - ms[3]) / (3 * 2), task = (ms[2] - ms[3]) / (4 * 2),
        interaction = (ms[3] - ms[4]) / 2, error = ms[4])
round(cbind(variance = vc, share = vc / sum(vc)), 3)
```

```
            variance share
annotator      7.725 0.035
task         205.950 0.932
interaction    2.905 0.013
error          4.309 0.020
```

איך קוראים את התוצאה

- **מתייגים:** $F_A = 5.58$ עם $(3, 6)$ דרגות חופש, $p = 0.036$. עם המכנה של R (MS_E): $F = 13.1$, $p = 4.3 \times 10^{-4}$. ה-p-value הנכון גדול פי 84, כי MS_{AB} גדול פי 2.3 מ-MS_E, ויש רק 6 דרגות חופש במכנה
- **סוגי משימות:** כ-93% מהשונות. **ההחלטה:** לתמחר תיוג **לפי סוג משימה**, לא מחיר אחיד לפריט
- **מתייגים:** כ-3% מהשונות. יש מתייגים מהירים ואיטיים, אבל זה פחות חשוב מסוג המשימה
- **זהירות:** את σ^2_β אומדים מ-3 סוגי משימות בלבד. האומד מאוד לא מדויק. כדי לאמוד אותו טוב, צריך **יותר סוגי משימות**, לא יותר פריטים

ב-R: הדרך הנכונה, עם `Error()`

את החישוב הידני אפשר להחליף בשורה אחת. `Error()` אומר ל-R **מול איזה איבר אקראי** לבדוק את האפקטים:

```r
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
- $F_A = 5.58$ עם $(3, 6)$ דרגות חופש: **בדיוק** החישוב הידני. בשכבה `Within`, השורה `Residuals` היא MS_E
- את רכיבי השונות עדיין מחשבים מתוחלות ה-MS, כמו קודם

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:** כשמספר התצפיות בתאים שונה, הפירוק $SS_A + SS_B + SS_{AB} + SS_E$ כבר לא יחיד, ו-`aov()` נותן תוצאה **שתלויה בסדר** הפקטורים בנוסחה

סיכום

מה לקחת מההרצאה

השאלה	הכלי	ב-R
כל הזוגות (גרסת הבחינה)	Duncan, $R_p = r_\alpha(p, f)\, s_{\bar{y}}$, מהקצוות פנימה	`qtukey((1-α)^(p-1), p, f)`
כל הזוגות, FWER $= \alpha$	Tukey HSD, $q_\alpha(a, f)\, s_{\bar{y}}$	`TukeyHSD()`
נתונים מוטים, חריגים	Kruskal–Wallis, $K \sim \chi^2_{a-1}$	`kruskal.test()`
ניגוד עמיד לחריגים	חציונים של הפרשים, Z_{ik}	
שני פקטורים, עם חזרות	שלושה מבחני F, $SS_A + SS_B + SS_{AB} + SS_E$	`aov(y ~ A * B)`
האם האפקט של A תלוי ב-B?	אינטראקציה, גרף אינטראקציה, אפקטים פשוטים	`emmeans(fit, ~ A, by = "B")`
מקור רעש ידוע (יום, משמרת)	בלוק, $n = 1$, מודל אדיטיבי	`aov(y ~ A + block)`
הרמות הוגרלו	רכיבי שונות, $F_A = MS_A / MS_{AB}$	`aov(y ~ A + B + Error(A:B))`
A קיים רק בתוך B	מודל היררכי, $F_A = MS_A / MS_{B(A)}$	`aov(y ~ A + Error(A:B))`

בהרצאה הבאה: השלמת ניתוח שונות דו-כיווני, תרגיל מסכם מקצה לקצה, ותחילת רגרסיה לוגיסטית.