

1. שיטות א-פרמטריות

ניתוח נתונים סטטיסטי · הרצאה 1

ד"ר עדי שריד

2026-10-18

על מה נדבר היום

- תזכורת קצרה: מהי בדיקת השערות
- מתי ההנחות של מבחן פרמטרי לא מתקיימות, ולמה זה משנה
- **מבחן הסימן**: השערות על החציון, גם למדגמים מזווגים
- **מבחן הדירוגים המסומנים** (Wilcoxon signed-rank)
- **מבחן סכום הדירוגים** (Wilcoxon rank-sum / Mann–Whitney)
- **טיב התאמה**: מבחן חי-בריבוע ומבחן Kolmogorov–Smirnov
- **מתאם על דירוגים**: Spearman ו-Kendall

תזכורת: בדיקת השערות (Hypothesis testing)

השאלה: האם הזכרים כבדים יותר?

- נתוני Palmer Penguins: מדידות של פינגווינים בתחנת המחקר פאלמר באנטארקטיקה
- נתמקד במין אחד, Gentoo, ובמשתנה אחד: **משקל הגוף** (גרם)
- **השאלה:** האם יש הבדל במשקל הממוצע בין זכרים לנקבות?
- הנתונים מגיעים עם R (מגרסה 4.5), בלי חבילה נוספת

```
gentoo <- penguins |> filter(species == "Gentoo", !is.na(sex))  
gentoo |> summarise(n = n(), mean = mean(body_mass), sd = sd(body_mass), .by = sex)
```

	sex	n	mean	sd
1	female	58	4680	281.6
2	male	61	5485	313.2

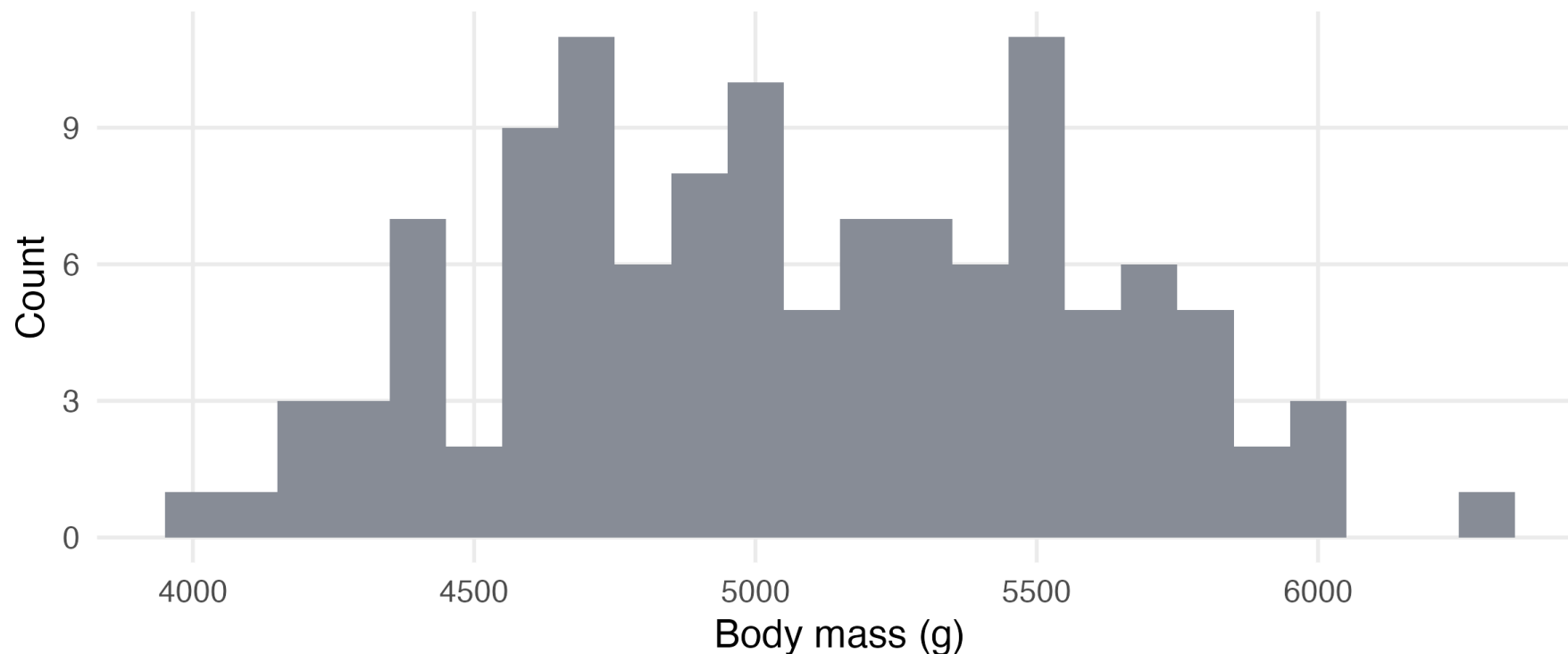
התפלגות המשקל

```
ggplot(gentoo, aes(body_mass)) +  
  geom_histogram(binwidth = 100, fill = nanas_pal[["grey"]]) +  
  labs(title = "משקל הגוף של פינגווינים מסוג Gentoo",  
        subtitle = "פרטים, זכרים ונקבות יחד 119",  
        x = "Body mass (g)", y = "Count")
```

התפלגות המשקל

משקל הגוף של פינגווינים מסוג Gentoo

119 פרטים, זכרים ונקבות יחד



המשקל לפי מין

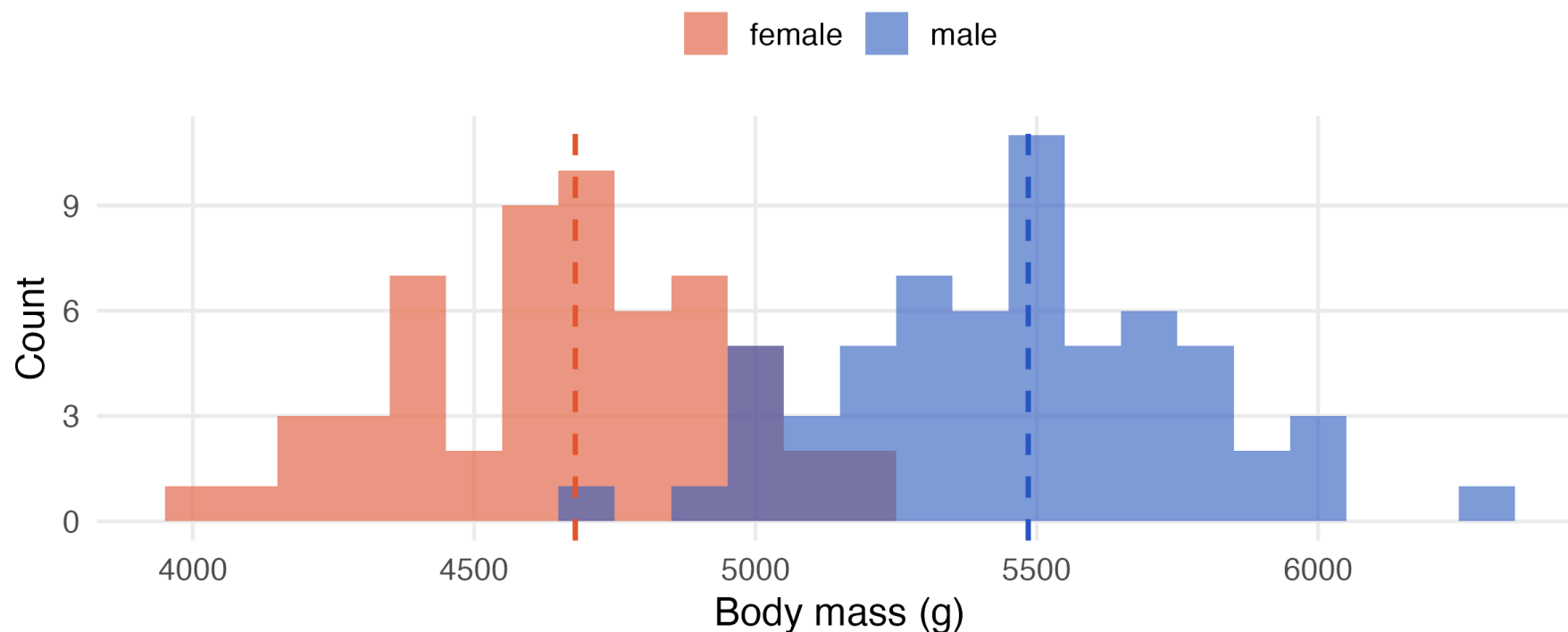
```
sex_means <- gentoo |> summarise(mean = mean(body_mass), .by = sex)

ggplot(gentoo, aes(body_mass, fill = sex)) +
  geom_histogram(binwidth = 100, alpha = 0.6, position = "identity") +
  geom_vline(aes(xintercept = mean, colour = sex), data = sex_means,
             linetype = "dashed", linewidth = 0.9, show.legend = FALSE) +
  scale_fill_manual(values = c(female = nanas_pal[["accent"]], male = nanas_pal[["primary"]])) +
  scale_colour_manual(values = c(female = nanas_pal[["accent"]], male = nanas_pal[["primary"]])) +
  labs(title = "הזכרים כבדים בכ-800 גרם במוצע",
       subtitle = "הקווים המקווקווים: הממוצע בכל קבוצה",
       x = "Body mass (g)", y = "Count", fill = NULL)
```

המשקל לפי מין

הזכרים כבדים בכ-800 גרם בממוצע

הקווים המקווקווים: הממוצע בכל קבוצה



מבחן t לשני מדגמים

$$H_0 : \mu_F = \mu_M \quad H_1 : \mu_F \neq \mu_M$$

$$t = \frac{\bar{x}_F - \bar{x}_M}{\sqrt{\frac{s_F^2}{n_F} + \frac{s_M^2}{n_M}}} = \frac{4679.7 - 5484.8}{\sqrt{\frac{281.6^2}{58} + \frac{313.2^2}{61}}} = \frac{-805.1}{54.5} \approx -14.76$$

```
t.test(body_mass ~ sex, data = gentoo) |>
  print(digits = 6) # a little more precision than the course default
```

Welch Two Sample t-test

data: body_mass by sex

t = -14.76, df = 116.6, p-value <2e-16

alternative hypothesis: true difference in means between group female and group male is not equal to 0

95 percent confidence interval:

-913.113 -697.076

sample estimates:

mean in group female	mean in group male
4679.74	5484.84

איך קוראים את הפלט

- $t = -14.76$: ההפרש בין הממוצעים הוא כ-15 שגיאות תקן. הסימן שלילי כי R מחסר זכרים מנקבות
- $df = 116.6$ Welch לא מניח שוויון שונויות, ולכן דרגות החופש לא שלמות
- $p\text{-value} < 2e-16$: אם H_0 נכונה, הסיכוי לראות הפרש קיצוני כזה כמעט אפס. דוחים את H_0
- **רווח הסמך (95%)**: הנקבות קלות מהזכרים ב-697 עד 913 גרם. הערך 0 לא נמצא ברווח
- **ההנחות**: תצפיות בלתי תלויות, והתפלגות נורמלית בכל קבוצה (או מדגם גדול)

למה צריך שיטות א- פרמטריות? (Non-parametric methods)

הרעיון

- שיטות פרמטריות (מבחן ANOVA, t) מניחות **התפלגות ידועה** של תגובת המערכת, בדרך כלל נורמלית
- שיטות א-פרמטריות מסתפקות ב**מידע חלקי** על ההתפלגות
- הן נשענות פחות על הערכים עצמם ויותר על **הסדר שלהם (הדירוג)**
- אין שיטה אחת לכל דבר: לכל השערה יש מבחן ייעודי

כשהתוחלת בכלל לא קיימת

דוגמה:

$$f(x) = \frac{1}{(1+x)^2}, \quad x \in [0, \infty)$$

$$\mu = \int_0^\infty \frac{x}{(1+x)^2} dx = \infty$$

$$F(a) = 1 - \frac{1}{1+a} = 0.5 \Rightarrow \tilde{\mu} = 1$$

- אין טעם לבדוק השערות על התוחלת, כי היא אינסופית
- **החציון** מוגדר היטב, ואפשר לבדוק השערות עליו

סימולציה: ממוצע מול חציון

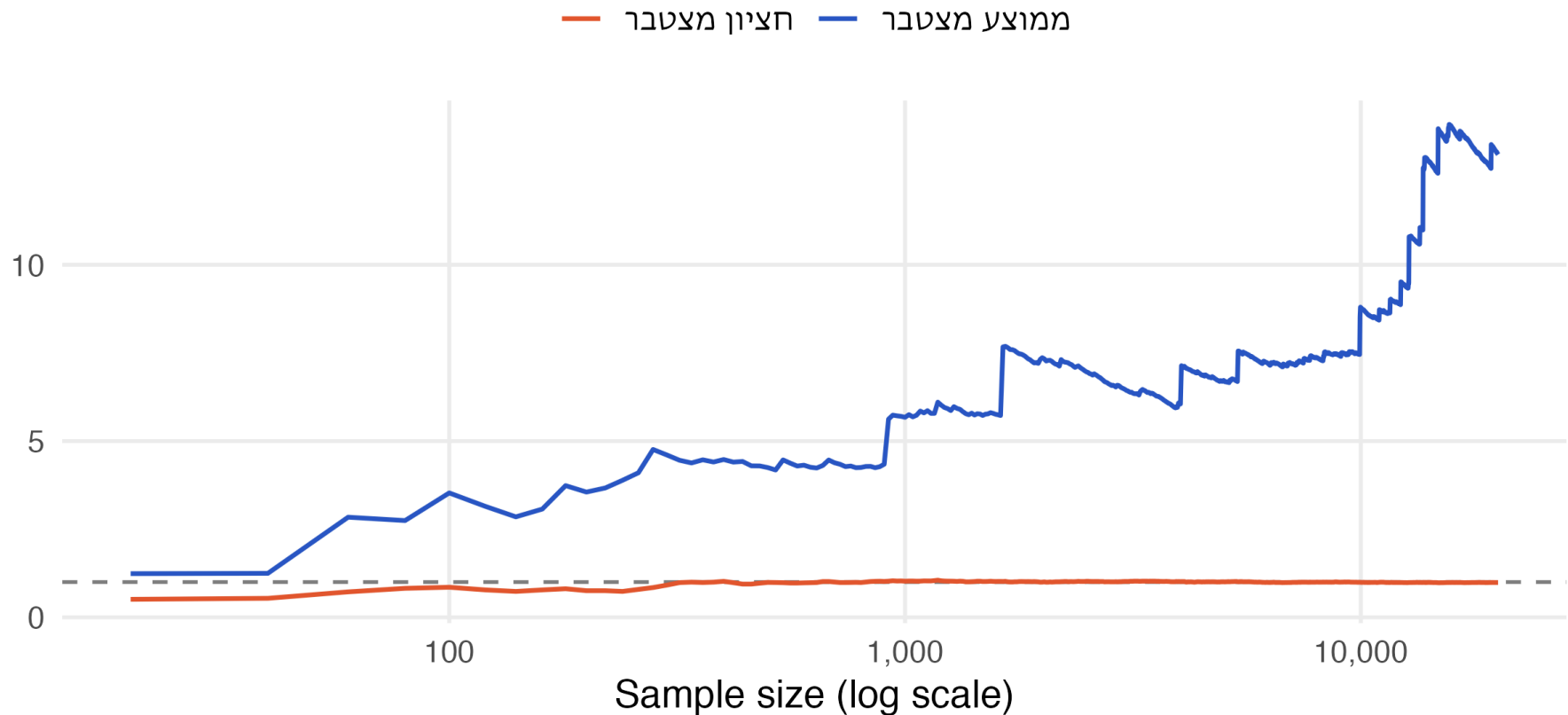
```
#  $F(x) = 1 - 1/(1+x) \Rightarrow x = 1/(1-u) - 1$ 
n <- 20000
heavy <- tibble(
  i = 1:n,
  x = 1 / (1 - runif(n)) - 1
) |>
mutate(
  `ממוצע מצטבר` = cummean(x),
  `חציון מצטבר` = map_dbl(i, \(k) median(x[seq_len(k)]))
) |>
filter(i %% 20 == 0) |>
pivot_longer(c(`חציון מצטבר`, `ממוצע מצטבר`), names_to = "stat")

ggplot(heavy, aes(i, value, colour = stat)) +
  geom_hline(yintercept = 1, linetype = "dashed", colour = "grey50") +
  geom_line(linewidth = 0.8) +
  scale_x_log10(labels = label_comma()) +
  scale_colour_manual(values = c(nanas_pal[["accent"]], nanas_pal[["primary"]])) +
  labs(title = "החציון מתכנס ל-1, הממוצע ממשיך לטפס",
        subtitle = "דגימות מהתפלגות שבה התוחלת אינסופית",
        x = "Sample size (log scale)", y = NULL, colour = NULL)
```

סימולציה: ממוצע מול חציון

החציון מתכנס ל-1, הממוצע ממשיך לטפס

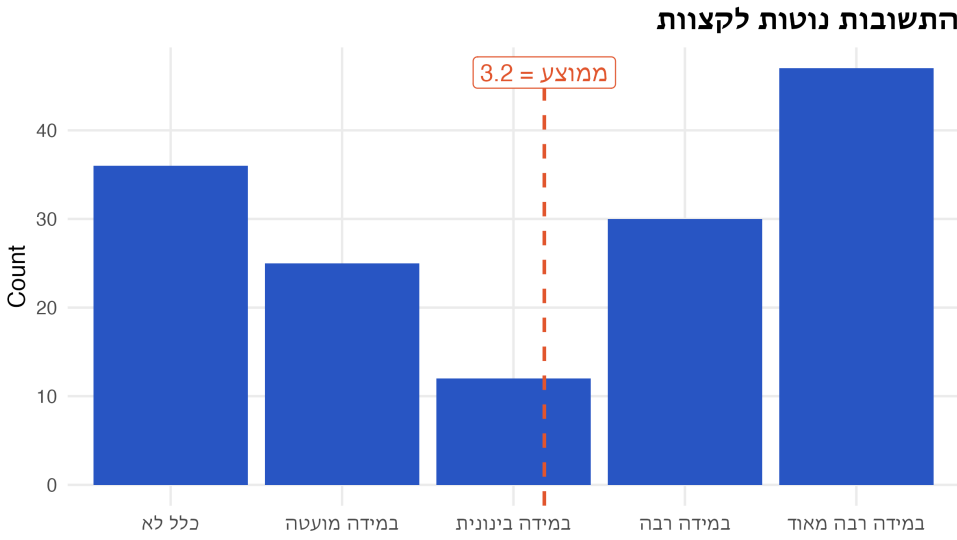
דגימות מהתפלגות שבה התוחלת אינסופית



כשהנתונים סדורים: שאלת סקר בסולם ליקרט

“באיזו מידה את/ה אוהב/ת סטטיסטיקה?” (150 משיבים)

- הסולם **סדור** (ordinal): הסדר ידוע, המרחקים בין התשובות לא
- קידוד 1–5 וחשוב ממוצע מניחים שכל הפערים שווים
- הממוצע (3.2) נופל ליד “במידה בינונית”, התשובה שהכי מעט אנשים בחרו
- ההתפלגות בדידה, חסומה ורחוקה מנורמלית. מבחנים על החציון ועל דירוגים מתאימים כאן יותר



מפת דרכים

השאלה	השיטה הפרמטרית	השיטה הא-פרמטרית
מרכז ההתפלגות	מבחן t / Z על התוחלת	מבחן הסימן (Sign test), מבחן הדירוגים המסומנים (Wilcoxon signed-rank)
שני מדגמים מזווגים	מבחן t מזווג	מבחן הסימן / Wilcoxon signed-rank על ההפרשים
שתי אוכלוסיות בלתי תלויות	מבחן t לשני מדגמים	מבחן סכום הדירוגים (Mann–Whitney)
התאמה להתפלגות נתונה		מבחן חי-בריבוע לטיב התאמה (Goodness-of-fit), מבחן Kolmogorov–Smirnov
השוואה בין שתי התפלגויות		מבחן חי-בריבוע להומוגניות (Chi-square test of homogeneity), Kolmogorov–Smirnov לשני מדגמים
מתאם בין שני משתנים	מתאם פירסון	Spearman's ρ , Kendall's τ

מבחן הסימן (Sign test)

השערות על החציון

$$H_0 : \tilde{\mu} = z \quad H_1 : \tilde{\mu} \neq z$$

- תחת H_0 , כל הפרש $x_i - z$ חיובי או שלילי בהסתברות 0.5
- תצפיות ששוות בדיוק ל- z (הפרש 0) **משמיטים**, ו- n הוא מספר התצפיות שנשארו
- r^+ : מספר ההפרשים החיוביים
- סטטיסטי המבחן: $r = \min(r^+, n - r^+)$
- תחת H_0 נצפה ש- r יהיה קרוב ל- $n/2$, וערך קטן מעיד נגד H_0
- $r^+ \sim \text{Binomial}(n, 0.5) \quad P(r \leq y) = 2 P(r^+ \leq y), \quad y < n/2$

הערך הקריטי

הערך הקריטי R_α^* הוא הערך הגדול ביותר שמקיים

$$P(r \leq R_\alpha^*) = 2 \sum_{y=0}^{R_\alpha^*} \binom{n}{y} \left(\frac{1}{2}\right)^n \leq \alpha$$

כלל ההכרעה: אם $r_0 \leq R_\alpha^*$, דוחים את H_0 .

```
# Critical value of the sign test: largest r with 2 * P(Bin(n, 0.5) <= r) <= alpha
sign_critical <- function(n, alpha) {
  max(which(2 * pbinom(0:n, n, 0.5) <= alpha)) - 1
}
sapply(c(`0.01` = 0.01, `0.05` = 0.05, `0.10` = 0.10), \(a) sign_critical(20, a))
```

0.01	0.05	0.10
3	5	5

דוגמה: זמני תגובה של API

ה-SLA של שירות קובע שהחציון של זמן התגובה הוא **200 ms**. נדגמו 20 בקשות:

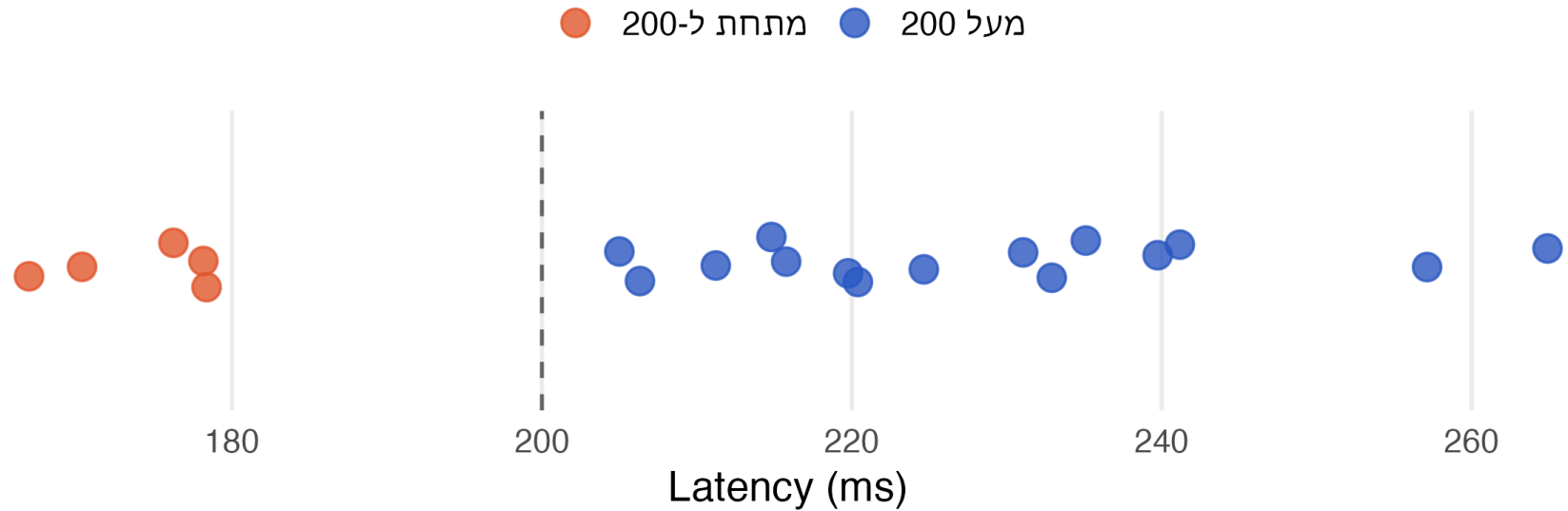
```
api <- read_csv(file.path(data_dir, "api_latency.csv"), show_col_types = FALSE)
api |> mutate(diff = latency_ms - 200, sign = if_else(diff > 0, "+", "-")) |>
  count(sign)
```

```
# A tibble: 2 × 2
  sign      n
<chr> <int>
1 +      15
2 -       5
```

```
ggplot(api, aes(latency_ms, 0, colour = latency_ms > 200)) +
  geom_vline(xintercept = 200, linetype = "dashed", colour = "grey40") +
  geom_point(size = 4, alpha = 0.8, position = position_jitter(height = 0.02, seed = 1)) +
  scale_colour_manual(values = c(`FALSE` = nanas_pal[["accent"]], `TRUE` = nanas_pal[["primary"]]),
    labels = c("מעל 200", "מתחת ל-200")) +
  scale_y_continuous(NULL, breaks = NULL, limits = c(-0.1, 0.1)) +
  labs(title = "15 200 ms מתוך 20 בקשות מעל", x = "Latency (ms)", colour = NULL)
```

דוגמה: זמני תגובה של API

15 מתוך 20 בקשות מעל 200 ms



מבחן הסימן: חישוב

$$r^+ = 15 \Rightarrow r_0 = \min(15, 5) = 5 \bullet$$

$$\bullet \text{ ב-} \alpha = 0.01: R_{0.01}^* = 3, \text{ ו-} 5 > 3, \text{ ולכן לא דוחים את } H_0$$

$$\bullet \text{ ב-} \alpha = 0.05: R_{0.05}^* = 5, \text{ ו-} 5 \leq 5, \text{ ולכן דוחים}$$

```
binom.test(sum(api$latency_ms > 200), n = nrow(api), p = 0.5)
```

Exact binomial test

```
data: sum(api$latency_ms > 200) and nrow(api)
number of successes = 15, number of trials = 20, p-value = 0.04
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.5090 0.9134
sample estimates:
probability of success
      0.75
```

דוגמה: חציון הציונים במועד ב'

מרצה טוען שחציון הציונים במועד ב' שלו הוא 80. ניגשו 11 סטודנטים. האם הטענה נכונה ב- $\alpha = 0.1$?

```
grades <- read_csv(file.path(data_dir, "exam_grades.csv"), show_col_types = FALSE)
grades$grade
```

```
[1] 79 75 85 95 70 60 65 75 74 89 72
```

```
binom.test(sum(grades$grade > 80), n = nrow(grades))$p.value # sign test
```

```
[1] 0.2266
```

• $r_0 = 3 > R_{0.1,11}^* = 2, r^+ = 3$ ולכן טענת המרצה אינה נדחית

• גם מבחן t לתוחלת (בהנחת נורמליות) לא דוחה את H_0

```
t.test(grades$grade, mu = 80) |> print(digits = 6)
```

One Sample t-test

```
data: grades$grade
t = -1.204, df = 10, p-value = 0.256
alternative hypothesis: true mean is not equal to 80
95 percent confidence interval:
 69.3750 83.1704
sample estimates:
mean of x
 76.2727
```

מבחן הסימן למדגמים מזווגים

כשהתצפיות **מזווגות**, מריצים את מבחן הסימן על ההפרשים $d_i = x_i - y_i$ עם $H_0 : \tilde{\mu}_d = 0$.

דוגמה: צוות מוצר שוקל להחליף את Sonnet 5.5 ב-Opus 5.5, היקר יותר, למשימות תיוג. שני המודלים תייגו את אותן 11 משימות. **האם Opus טועה פחות?** בטבלה: שיעור השגיאות (%) בכל משימה. $Opus = d_i$ פחות Sonnet. (נתונים מדומים)

```
llm_err <- read_csv(file.path(data_dir, "llm_labelling_errors.csv"), show_col_types = FALSE)
llm_err |>
  mutate(sign = if_else(opus > sonnet, "+", "-")) |>
  select(task = task_he, sonnet, opus, sign) |>
  kable()
```

מבחן הסימן למדגמים מזווגים

sign	opus	sonnet	task
-	6.9	8.4	ניתוב פניות שירות
-	3.6	4.1	סנטימנט בביקורות
-	2.1	2.3	זיהוי ספאם
-	9.0	11.2	זיהוי כוונה בצ'אט
+	12.4	6.5	זיהוי תוכן פוגעני
-	8.1	9.8	שדות בחשבונות
-	0.6	1.4	זיהוי שפה
-	11.6	14.7	תעדוף פניות רפואיות
-	10.4	13.0	סיווג סעיפים בחוזים
-	6.0	7.2	קטגוריית מוצר
+	22.1	15.3	זיהוי סרקזם

Sonnet מול Opus: התוצאה

```
n_plus <- sum(llm_err$opus > llm_err$sonnet)
binom.test(n_plus, n = nrow(llm_err)) |> print(digits = 6)
```

Exact binomial test

```
data: n_plus and nrow(llm_err)
number of successes = 2, number of trials = 11, p-value = 0.0654
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.0228312 0.5177559
sample estimates:
probability of success
      0.181818
```

- $r^+ = 2$, $p = 0.065 < 0.1$, ולכן דוחים H_0 : חציון ההפרשים אינו 0
- Opus טועה פחות ב-9 מתוך 11 המשימות

מבחן הדירוגים המסומנים (Wilcoxon signed-rank test)

הרחבה של מבחן הסימן

$$H_0 : \tilde{\mu} = z \quad H_1 : \tilde{\mu} \neq z$$

- **הנחה:** ההתפלגות רציפה וסימטרית סביב החציון $\tilde{\mu}$. לכן ההשערות הן על מרכז הסימטריה, שהוא החציון (וגם התוחלת, אם היא קיימת)
- מתבסס לא רק על **הסימן** של הפרשים אלא גם על **גודלם**
- לרוב חזק יותר ממבחן הסימן
- אפשר להפעיל גם על הפרשים של מדגמים מזווגים

שלבי המבחן

1. משמיטים הפרשים ששווים ל-0 ומעדכנים את n . מדרגים את $|x_i - z|$ בסדר עולה. במקרה של שוויון, כל אחד מקבל את ממוצע הדרגות

2. מצמידים לכל דרגה את הסימן של ההפרש שלה

3. R^+ הוא סכום הדרגות החיוביות ו- R^- סכום הדרגות השליליות, וסטטיסטי המבחן הוא $R = \min(R^+, R^-)$

4. דוחים את H_0 אם $R \leq R_{\alpha,n}^*$ (טבלה), או לפי קירוב נורמלי:

$$E(R^+) = \frac{n(n+1)}{4} \qquad V(R^+) = \frac{n(n+1)(2n+1)}{24}$$

דוגמה: $z = 200$

דרגה מסומנת	$ x_i - z $	$x_i - z$	x_i
+2	5	+5	205
-4.5	12	-12	188
נשמט		0	200
+4.5	12	+12	212
-1	4	-4	196
+6	30	+30	230
-3	9	-9	191

$n = 6, \quad R^+ = 12.5, \quad R^- = 8.5$

$R = \min(R^+, R^-) = 8.5$

בדיקה: $R^+ + R^- = \frac{6 \cdot 7}{2} = 21$

מאיפה מגיעות $E(R^+)$ ו- $V(R^+)$?

תחת H_0 ההתפלגות סימטרית סביב z , ולכן לכל הפרש יש סיכוי $\frac{1}{2}$ להיות חיובי, בלי קשר לגודלו. אפשר לחשוב על זה כך: הדרגות $1, 2, \dots, n$ קבועות, ולכל דרגה מטילים מטבע הוגן שקובע את הסימן שלה.

$$R^+ = \sum_{k=1}^n k \cdot I_k, \quad I_k \sim \text{Bernoulli}\left(\frac{1}{2}\right) \text{ בלתי תלויים}$$

• **תוחלת:** סכום כל הדרגות הוא $\frac{n(n+1)}{2}$, ובממוצע חצי ממנו נופל בצד החיובי:

$$E(R^+) = \sum_{k=1}^n k \cdot \frac{1}{2} = \frac{1}{2} \cdot \frac{n(n+1)}{2} = \frac{n(n+1)}{4}$$

• **שונות:** $V(I_k) = \frac{1}{4}$, והמטבעות **בלתי תלויים**, ולכן אין שונויות משותפות:

$$V(R^+) = \sum_{k=1}^n k^2 \cdot \frac{1}{4} = \frac{1}{4} \cdot \frac{n(n+1)(2n+1)}{6} = \frac{n(n+1)(2n+1)}{24}$$

חזרה לזמני התגובה

```
signed <- api |>
  mutate(diff = latency_ms - 200,
         rank = rank(abs(diff)),
         signed_rank = sign(diff) * rank) |>
  arrange(rank)

signed |> summarise(R_plus = sum(rank[diff > 0]),
                  R_minus = sum(rank[diff < 0]))
```

```
# A tibble: 1 × 2
  R_plus R_minus
  <dbl>   <dbl>
1  156.    53.5
```

$$R = \min(156.5, 53.5) = 53.5 < 60 = R_{0.1, 20}^*$$

ולכן דוחים את H_0 ב- $\alpha = 0.1$.

בדיקה ב-R

```
wilcox.test(api$latency_ms, mu = 200, exact = FALSE)
```

Wilcoxon signed rank test with continuity correction

data: api\$latency_ms

V = 156, p-value = 0.06

alternative hypothesis: true location is not equal to 200

• V ב-R הוא R^+ (156.5), לא $\min(R^+, R^-)$

• `exact = FALSE`: יש דרגות זהות, ולכן R משתמש בקירוב הנורמלי

כשהמבחנים לא מסכימים

נחזור ל-Sonnet מול Opus:

```
c(sign      = binom.test(n_plus, nrow(llm_err))$p.value,  
  signed_rank = wilcox.test(llm_err$opus, llm_err$sonnet, paired = TRUE)$p.value)
```

```
sign signed_rank  
0.06543      0.32031
```

- ב- $\alpha = 0.1$ מבחן הסימן דוחה ומבחן הדירוגים המסומנים לא
- אם אפשר להניח סימטריה, שני המבחנים תקפים. מבחן הדירוגים המסומנים מנצל גם את גודל ההפרשים, ולכן בדרך כלל חזק יותר
- אם אי אפשר להניח סימטריה, רק מבחן הסימן בודק את החציון, והוא שקובע

מבחן סכום הדרוגים (Wilcoxon rank-sum test)

שתי אוכלוסיות בלתי תלויות

- n תצפיות מאוכלוסייה א' ו- m תצפיות מאוכלוסייה ב', עם $m \leq n$ ו- $N = n + m$
- $H_0 : F_1(x) = F_2(x)$ מול $H_1 : F_1(x) \neq F_2(x)$
- השיטה (Wilcoxon rank-sum): מאחדים את כל התצפיות לרשימה אחת וממיינים
- סטטיסטי המבחן T : סכום הדרגות של התצפיות מהמדגם הקטן

$$E(T) = \frac{m(n + m + 1)}{2} \quad V(T) = \frac{mn(n + m + 1)}{12}$$

$$\left| \frac{T - E(T)}{\sigma_T} \right| > Z_{1-\alpha/2} \text{ אם } H_0 \text{ דוחים את}$$

מאיפה מגיעה $E(T)$?

נניח שאין תיקו. נסמן ב- $R_1, \dots, R_m$ את הדרגות של תצפיות המדגם הקטן, כך ש- $T = \sum_{i=1}^m R_i$.

תחת H_0 כל התצפיות מגיעות מאותה התפלגות, ולכן כל סידור של N התצפיות סביר באותה מידה. מכאן שכל דרגה R_i מתפלגת אחיד על $\{1, 2, \dots, N\}$:

$$E(R_i) = \frac{1}{N} \sum_{k=1}^N k = \frac{N+1}{2}$$

התוחלת לינארית, ולכן (בלי צורך באי-תלות):

$$E(T) = \sum_{i=1}^m E(R_i) = m \cdot \frac{N+1}{2} = \frac{m(n+m+1)}{2}$$

ובאותה דרך נקבל גם את השונות של דרגה בודדת:

$$V(R_i) = E(R_i^2) - E(R_i)^2 = \frac{(N+1)(2N+1)}{6} - \frac{(N+1)^2}{4} = \frac{N^2-1}{12}$$

מאיפה מגיעה $V(T)$?

הדרגות **תלויות**: שתי תצפיות לא יכולות לקבל אותה דרגה. לכן צריך גם את השונות המשותפת:

$$V(T) = \sum_{i=1}^m V(R_i) + \sum_{i \neq j} \text{Cov}(R_i, R_j) = m \sigma^2 + m(m-1) c$$

כאשר $\sigma^2 = \frac{N^2-1}{12}$ ו- $c = \text{Cov}(R_i, R_j)$ (זהה לכל זוג, מסימטריה).

הטריק: סכום כל N הדרגות הוא קבוע, $1 + 2 + \dots + N$, ולכן השונות שלו אפס. אותו חישוב עם $m = N$ נותן:

$$0 = N\sigma^2 + N(N-1)c \implies c = -\frac{\sigma^2}{N-1} = -\frac{N+1}{12}$$

מציבים בחזרה:

$$V(T) = m\sigma^2 \left(1 - \frac{m-1}{N-1}\right) = m \cdot \frac{(N+1)(N-1)}{12} \cdot \frac{n}{N-1} = \frac{mn(n+m+1)}{12}$$

דוגמה: שיפור בתהליך ייצור

שיפור טכנולוגי בתהליך הייצור לא שינה את צורת ההתפלגות של התפוקה היומית, אבל הזיז אותה כלפי מעלה. האם התפוקה עלתה ב-3 יחידות ביום? ($\alpha = 0.05$)

```
prod <- read_csv(file.path(data_dir, "production_output.csv"), show_col_types = FALSE)
prod |> group_by(period) |> summarise(units = paste(units, collapse = ", "))
```

```
# A tibble: 2 × 2
  period units
  <chr>   <chr>
1 after  7.7, 8.1, 8.3, 8.6, 8.9
2 before 3.9, 4.3, 4.4, 4.6, 4.9, 5.8
```

הטריק: מורידים 3 מהתצפיות שאחרי השיפור. אם התפוקה עלתה ב-3 בדיוק, שני המדגמים מגיעים מאותה התפלגות.

דירוג הרשימה המאוחדת

```
prod_shift <- prod |>
  mutate(adj = if_else(period == "after", units - 3, units),
         rank = rank(adj))

ggplot(prod_shift, aes(adj, 0, colour = period, label = rank)) +
  geom_point(size = 9, alpha = 0.85) +
  geom_text(colour = "white", fontface = "bold", size = 4) +
  scale_colour_manual(values = c(after = nanas_pal[["accent"]], before = nanas_pal[["primary"]]),
                     labels = c(after = "(3 פחות)", before = "לפני")) +
  scale_y_continuous(NULL, breaks = NULL) +
  labs(title = "הדרגות של המדגם המאוחד", x = "Daily output (units)", colour = NULL)
```

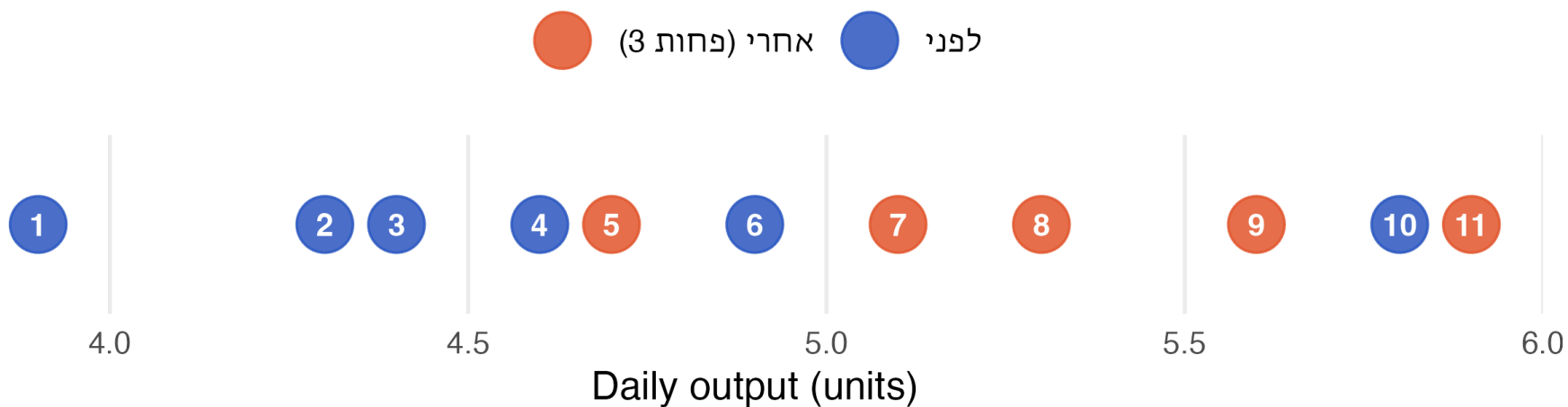
$$T_0 = 5 + 7 + 8 + 9 + 11 = 40, \quad E(T) = \frac{5 \cdot 12}{2} = 30, \quad V(T) = \frac{5 \cdot 6 \cdot 12}{12} = 30$$

$$\left| \frac{40 - 30}{\sqrt{30}} \right| = 1.82 < 1.96$$

ולכן לא דוחים H_0 : הנתונים מתיישבים עם שיפור של 3 יחידות ביום.

דירוג הרשימה המאוחדת

הדרגות של המדגם המאוחד



אותו מבחן ב-R

```
after <- prod$units[prod$period == "after"]  
before <- prod$units[prod$period == "before"]  
wilcox.test(after, before, mu = 3)
```

Wilcoxon rank sum exact test

data: after and before

W = 25, p-value = 0.08

alternative hypothesis: true location shift is not equal to 3

• ב-R הסטטיסטי W הוא $T - \frac{m(m+1)}{2} = 40 - 15 = 25$ (סטטיסטי Mann–Whitney U)

• $\mu = 3$ בודק הזזה של 3 יחידות, בלי לחסר ידנית

• R מחשב p-value מדויק (0.082) ולא את הקירוב הנורמלי, אבל המסקנה זהה

טיב התאמה (Goodness of fit)

תזכורת: מבחן חי-בריבוע לטיב התאמה

H_0 : ההסתברויות של k הקטגוריות הן $p_1, \dots, p_k$. O_i הוא המספר שנצפה, $E_i = np_i$ המספר הצפוי:

$$\chi^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i} \sim \chi_{k-1-m}^2$$

m הוא מספר הפרמטרים שאמדנו מהנתונים. **זה מבחן א-פרמטרי**: ההתפלגות של הסטטיסטי תחת H_0 לא תלויה בהתפלגות של האוכלוסייה, ואותו מבחן עובד לכל F_0 . אבל ההתפלגות χ^2 היא רק **קירוב** למדגם גדול, שנשען על קירוב נורמלי לספירות. לכן צריך $E_i \geq 5$ בכל קטגוריה. ב-KS, לעומת זאת, ההתפלגות של D תחת H_0 מדויקת וזהה **לכל** F_0 רציפה.

דוגמה: האם משקל הגוף של כל 342 הפינגווינים ב-Palmer Penguins (שלושת המינים יחד) מתפלג נורמלית? מחלקים ל-8 קטגוריות ששוות בהסתברותן תחת הנורמלית המותאמת, כך ש- $E_i = 342/8 = 42.75$. אמדנו μ ו- σ , ולכן $m = 2$.

```
mass <- na.omit(penguins$body_mass)
bins <- qnorm(seq(0, 1, by = 1/8), mean(mass), sd(mass)) # 8 equally likely bins under H0
observed <- table(cut(mass, bins))
chisq.test(observed)$statistic # df = 8 - 1 - 2 = 5, not 7
```

X-squared
33.86

$\chi^2 = 33.86 > 11.07 = \chi_{0.95, 5}^2$, ולכן **דוחים** H_0 : המשקל אינו נורמלי. זה תערובת של שלושה מינים.

מבחן Kolmogorov–Smirnov

כשהנתונים רציפים, החלוקה לקטגוריות מאבדת מידע. מבחן KS משווה את פונקציית ההתפלגות המצטברת:

- $F_n(x)$: פונקציית ההתפלגות האמפירית, החלק מהתצפיות שקטנות או שוות ל- x
- $H_0 : F(x) = F_0(x)$ לכל x , כש- F_0 רציפה ונתונה **במלואה**

$$D = \max_x |F_n(x) - F_0(x)|$$

- דוחים את H_0 אם $D > D_{\alpha,n}$ (טבלה). למדגם גדול: $D_{0.05,n} \approx 1.36/\sqrt{n}$
- **בחישוב ידני**: ממיינים את התצפיות, ובודקים בכל $x_{(i)}$ את שני הצדדים של המדרגה:

$$D = \max_i \left\{ \frac{i}{n} - F_0(x_{(i)}), F_0(x_{(i)}) - \frac{i-1}{n} \right\}$$

דוגמה: אותם משקלים, מבחן KS

נחזור למשקלי הפינגווינים, הפעם בלי לחלק לקטגוריות:

```
ks.test(mass, "pnorm", mean(mass), sd(mass)) |> print(digits = 6)
```

Asymptotic one-sample Kolmogorov-Smirnov test

data: mass

D = 0.1041, p-value = 0.00121

alternative hypothesis: two-sided

- $D = 0.104 > 0.074 \approx 1.36/\sqrt{342}$ ולכן דוחים H_0 , בהתאם למבחן חי-בריבוע
- המרחק הגדול ביותר הוא סביב 3,800 גרם: 41% מהפינגווינים קלים מזה, ולפי הנורמלית היו צריכים להיות רק 31%

הפונקציה האמפירית מול הנורמלית

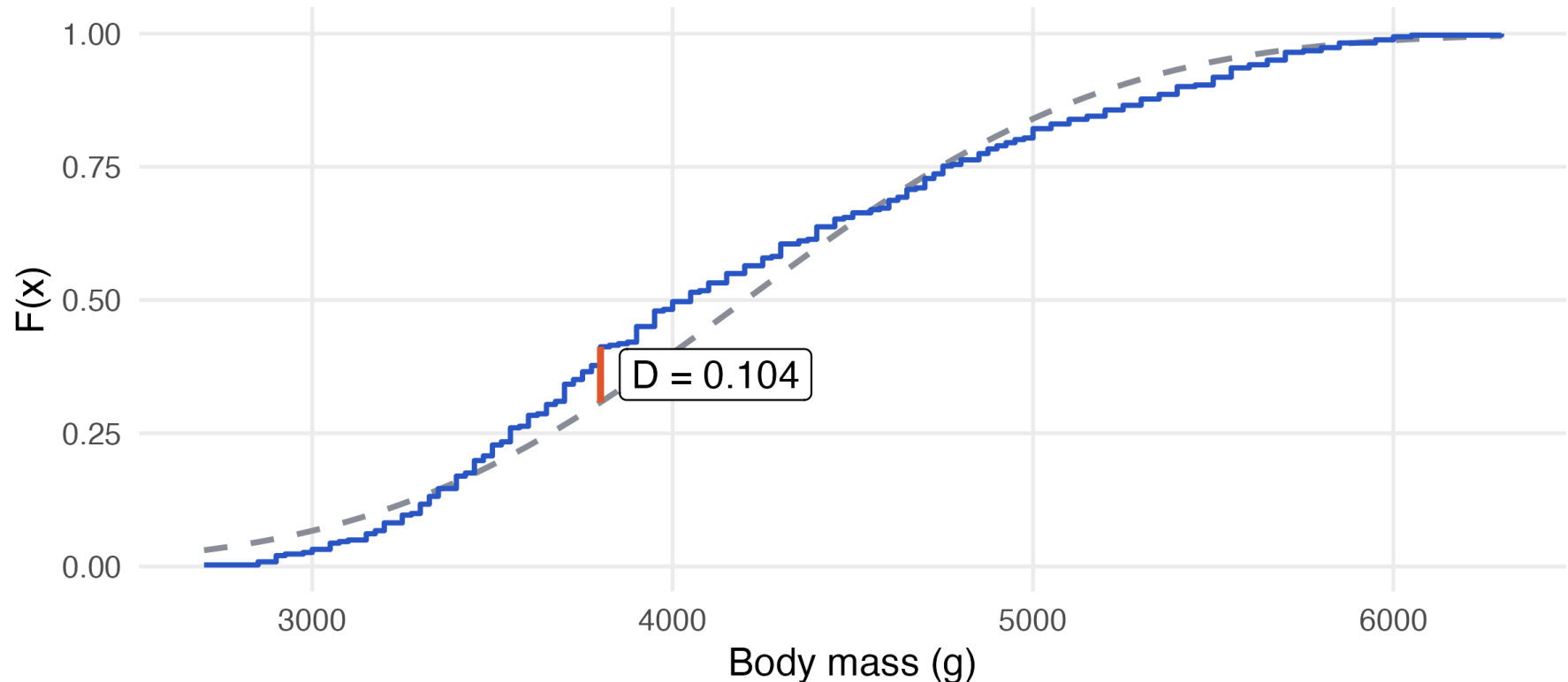
```
ecdf_mass <- tibble(g = sort(unique(mass))) |> # g: each distinct weight
  mutate(Fn = ecdf(mass)(g), F0 = pnorm(g, mean(mass), sd(mass)))
worst <- ecdf_mass |>
  mutate(Fn_left = lag(Fn, default = 0), # the step's value just before g
         Fn_end = if_else(abs(Fn - F0) >= abs(Fn_left - F0), Fn, Fn_left),
         gap = abs(Fn_end - F0)) |>
  slice_max(gap, n = 1, with_ties = FALSE)

ggplot(ecdf_mass, aes(g)) +
  stat_function(fun = \(x) pnorm(x, mean(mass), sd(mass)),
               colour = nanas_pal[["grey"]], linewidth = 1, linetype = "dashed") +
  geom_step(aes(y = Fn), colour = nanas_pal[["primary"]], linewidth = 0.9) +
  geom_segment(aes(x = g, xend = g, y = F0, yend = Fn_end), data = worst,
               colour = nanas_pal[["accent"]], linewidth = 1.2) +
  annotate("label", x = worst$g, y = (worst$F0 + worst$Fn_end) / 2,
           label = paste0("D = ", round(worst$gap, 3)), hjust = -0.1, fill = "white") +
  labs(title = "יותר מדי פינגווינים קלים",
        subtitle = "כחול: התפלגות אמפירית · מקווקו: נורמלית עם אותו ממוצע וסטיות תקן",
        x = "Body mass (g)", y = "F(x)")
```

הפונקציה האמפירית מול הנורמלית

יותר מדי פינגווינים קלים

כחול: התפלגות אמפירית · מקווקו: נורמלית עם אותו ממוצע וסטיית תקן



KS לשני מדגמים

אותו רעיון, בלי F_0 : משווים בין שתי פונקציות התפלגות אמפיריות

$$H_0 : F_1(x) = F_2(x) \quad D = \max_x |F_{1,n}(x) - F_{2,m}(x)|$$

דוגמה: גרסה חדשה של שירות inference עלתה לאוויר. החציון של זמן התגובה כמעט לא השתנה, אבל משתמשים מתלוננים על “קפיצות”. נדגמו 60 בקשות מכל גרסה. (נתונים מדומים)

```
set.seed(1)
latency <- tibble(
  version = rep(c("v1", "v2"), each = 60),
  ms = c(rlnorm(60, log(220), 0.25), rlnorm(60, log(220), 0.65))
)
latency |> summarise(median = median(ms), iqr = IQR(ms), max = max(ms), .by = version)
```

```
# A tibble: 2 × 4
  version median   iqr   max
  <chr>    <dbl> <dbl> <dbl>
1 v1      227.  63.5  361.
2 v2      225. 180. 1048.
```

שתי פונקציות התפלגות מצטברות

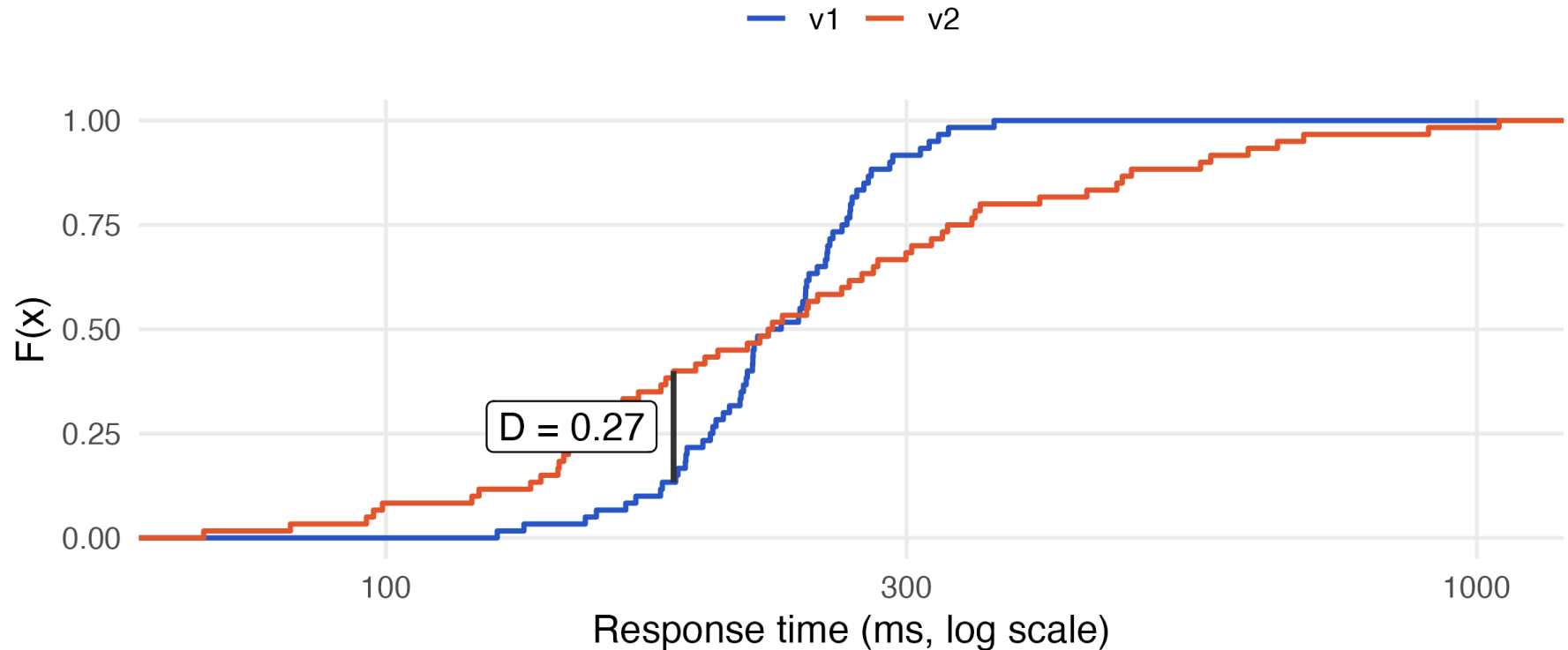
```
grid <- tibble(ms = sort(latency$ms)) |>
  mutate(v1 = ecdf(latency$ms[latency$version == "v1"])(ms),
         v2 = ecdf(latency$ms[latency$version == "v2"])(ms),
         gap = abs(v1 - v2))
max_gap <- slice_max(grid, gap, n = 1, with_ties = FALSE)

ggplot(latency, aes(ms, colour = version)) +
  stat_ecdf(linewidth = 0.9) +
  geom_segment(aes(x = ms, xend = ms, y = v1, yend = v2), data = max_gap,
              colour = "grey20", linewidth = 1, inherit.aes = FALSE) +
  annotate("label", x = max_gap$ms, y = (max_gap$v1 + max_gap$v2) / 2,
          label = paste0("D = ", round(max_gap$gap, 2)), hjust = 1.1,
          fill = "white", label.size = 0) +
  scale_x_log10() +
  scale_colour_manual(values = c(v1 = nanas_pal[["primary"]], v2 = nanas_pal[["accent"]])) +
  labs(title = "אותו חציון, פיזור שונה לגמרי",
       subtitle = "הקו השחור: המרחק הגדול ביותר בין שתי הפונקציות",
       x = "Response time (ms, log scale)", y = "F(x)", colour = NULL)
```

שתי פונקציות התפלגות מצטברות

אותו חציון, פיזור שונה לגמרי

הקו השחור: המרחק הגדול ביותר בין שתי הפונקציות



KS מול סכום הדרוגים

```
v1 <- latency$ms[latency$version == "v1"]  
v2 <- latency$ms[latency$version == "v2"]  
c(rank_sum = wilcox.test(v1, v2)$p.value,  
  ks       = ks.test(v1, v2)$p.value)
```

```
rank_sum      ks  
0.95187 0.02763
```

- **סכום הדרוגים** שואל אם תצפית מגרסה אחת נוטה להיות **גדולה** מתצפית מהגרסה השנייה. ב-v2 יש גם זמנים קצרים מאוד וגם ארוכים מאוד: הדרגות שלו מתפזרות לשני הקצוות, מתקזזות, וסכום הדרגות נשאר קרוב לתוחלת שלו
- **KS** רגיש לכל הבדל בין ההתפלגויות: מיקום, פיזור או צורה. כאן הוא **דוחה** ב- $\alpha = 0.05$
- בשירות אמיתי, הזנב (האחוזון ה-95 או ה-99) הוא מה שהמשתמשים מרגישים. לכן כדאי לבדוק יותר מהחציון

חי-בריבוע להומוגניות (Chi-square test of homogeneity)

גם חי-בריבוע יודע להשוות שתי התפלגויות (או יותר). מחלקים את שני המדגמים לאותן k קטגוריות ובונים טבלה של $2 \times k$:

$$H_0 : F_1 = F_2 \quad E_{ij} = \frac{(i \text{ סכום שורה}) \cdot (j \text{ סכום עמודה})}{N} \quad \chi^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim \chi^2_{(2-1)(k-1)}$$

אותם זמני תגובה, מחולקים ל-5 קטגוריות לפי החמישונים של המדגם המאוחד:

```
breaks <- quantile(latency$ms, seq(0, 1, by = 0.2)) # pooled quintiles
tab <- table(latency$version, cut(latency$ms, breaks, include.lowest = TRUE, dig.lab = 4))
tab
```

	[68.06,165.7]	(165.7,212.3]	(212.3,245.9]	(245.9,300.4]	(300.4,1048]
v1	4	17	17	17	5
v2	20	7	7	7	19

```
chisq.test(tab)
```

Pearson's Chi-squared test

data: tab

X-squared = 31, df = 4, p-value = 3e-06

• בכל קטגוריה 24 בקשות, ולכן $E_{ij} = 12$ בכל תא. ל- v_2 יש עודף בשתי הקטגוריות הקיצוניות: זה הפיזור

ניתוח נתונים סטטיסטי · תשפ"ז

חי-בריבוע מול KS

שני המבחנים א-פרמטריים: ההתפלגות של הסטטיסטי תחת H_0 לא תלויה בהתפלגות של האוכלוסייה. ההבדל ביניהם הוא בסוג הנתונים ובמחיר:

חי-בריבוע		KS
נתונים	מחולקים לקטגוריות. מתאים גם לנתונים קטגוריים ובדידים	רציפים. משתמש בפונקציה האמפירית עצמה
מחיר	החלוקה לקטגוריות מאבדת מידע, והתוצאה תלויה בגבולות	דורש F_0 רציפה ונתונה במלואה (אחרת Lilliefors)
קירוב	אסימפטוטי, צריך $E_i \geq 5$	יש טבלאות מדויקות גם למדגם קטן
שתי התפלגויות	מבחן הומוגניות, <code>chisq.test(table(g, x))</code>	<code>ks.test(x, y)</code> , לשני מדגמים KS

מתאם על דירוגים (Rank correlation)

תזכורת: מתאם פירסון (Pearson correlation)

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}$$

- $-1 \leq r \leq 1$. **הסימן** הוא כיוון הקשר, **והגודל** הוא עד כמה הנקודות קרובות לקו ישר
- $r = 0$ פירושו "אין קשר **לינארי**", לא "אין קשר". ומתאם אינו סיבתיות
- **למה זה חשוב:** זו הדרך המהירה ביותר לכמת קשר בין שני משתנים, והיא הבסיס לרגרסיה (ברגרסיה פשוטה, $R^2 = r^2$)
- **שימוש יומיומי:**
 - האם ציון אוטומטי של מודל שפה (LLM-as-judge) מסכים עם הציונים של בודקים אנושיים?
 - האם זמן הטעינה של אתר קשור לשיעור הנטישה?
 - מטריצת מתאמים כצעד ראשון בכל ניתוח: אילו משתנים "זזים יחד"?

למה לא פירסון?

כפי שראינו, פירסון מודד את חוזק הקשר הלינארי, אבל:

- הוא רגיש מאוד לערכים חריגים
 - הוא לא מתאים למשתנים אורדינליים (סולמות סדר, כמו "מרוצה מאוד" עד "לא מרוצה")
 - הוא לא מתאים כשהשונות לא אחידה
- החלופה:** מדדים שמבוססים על הדרגות, ולכן מודדים קשר מונוטוני, לא בהכרח לינארי.
- **Spearman's ρ** : מתאם פירסון על הדרגות
 - **Kendall's τ** : מבוסס על זוגות בהתאמה מול זוגות בחוסר התאמה

Kendall's τ

זוג תצפיות $(x_i, y_i), (x_j, y_j)$ הוא **בהתאמה** (concordant) אם $x_i > x_j$ וגם $y_i > y_j$ או $x_i < x_j$ וגם $y_i < y_j$.
אחרת הזוג **בחוסר התאמה** (discordant).

$$\tau = \frac{n_c - n_d}{n_0}, \quad n_0 = \binom{n}{2}$$

- n_c : מספר הזוגות בהתאמה, n_d : מספר הזוגות בחוסר התאמה
- $-1 \leq \tau \leq 1$
- יש כמה גרסאות לטיפול בשוויון (ב-R: tau-b)

דוגמה: מתאם ספירמן ביד

שש תשובות של צ'אטבוט קיבלו ציון מבודק אנושי (X) ומשופט אוטומטי (LLM-as-judge, Y). האם השופט מדרג את התשובות באותו סדר כמו הבודק?

תשובה	X	Y	R_X	R_Y	d_i	d_i^2
A	3.1	52	1	2	-1	1
B	4.5	47	2	1	+1	1
C	5.2	61	3	4	-1	1
D	6.8	58	4	3	+1	1
E	7.4	74	5	6	-1	1
F	8.9	70	6	5	+1	1
סה"כ						6

- 1. מדרגים את X לחוד ואת Y לחוד
- 2. לכל תשובה מחשבים את הפער בין שתי הדרגות: $d_i = R_{X,i} - R_{Y,i}$
- 3. מציבים בנוסחה (כשאין שוויונות):

$$\rho_S = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

$$\rho_S = 1 - \frac{6 \cdot 6}{6 \cdot 35} = 0.83$$

אם השופט והבודק מדרגים בדיוק באותו סדר, כל $d_i = 0$ ו- $\rho_S = 1$

דוגמה: קנדל ביד

אותן שש תשובות. עוברים על כל $\binom{6}{2} = 15$ הזוגות. בכל זוג בודקים: האם השופט מסכים עם הבודק **איזו מהשתיים טובה יותר?**

	-	+	F	E	D	C	B	
	1	4	+	+	+	+	-	A
	0	4	+	+	+	+		B
	1	2	+	+	-			C
	0	2	+	+				D
	1	0	-					E
	3	12						סה"כ

למשל, A מול B: הבודק העדיף את B, אבל השופט העדיף את A. הם לא מסכימים, ולכן הזוג **בחוסר התאמה**:

$$X_B = 4.5 > 3.1 = X_A \quad Y_A = 52 > 47 = Y_B$$

- **+**: זוג **בהתאמה**, השופט והבודק מסכימים
- **-**: זוג **בחוסר התאמה**, השופט הפך את הסדר

$$n_c = 12, \quad n_d = 3$$

$$\tau = \frac{n_c - n_d}{n_0} = \frac{12 - 3}{15} = 0.6$$

פרשנות: $\frac{1+\tau}{2} = 0.8$, כלומר השופט והבודק מסכימים על הסדר ב-80% מהזוגות

דוגמה: קשר מונוטוני וערך חריג אחד

```
set.seed(2026) # own seed, so edits elsewhere in the lecture don't change this example
corr_data <- tibble(x = 1:30, y = exp(x / 6) + rnorm(30, sd = 1.5)) |>
  add_row(x = 2, y = 160) # one data-entry error

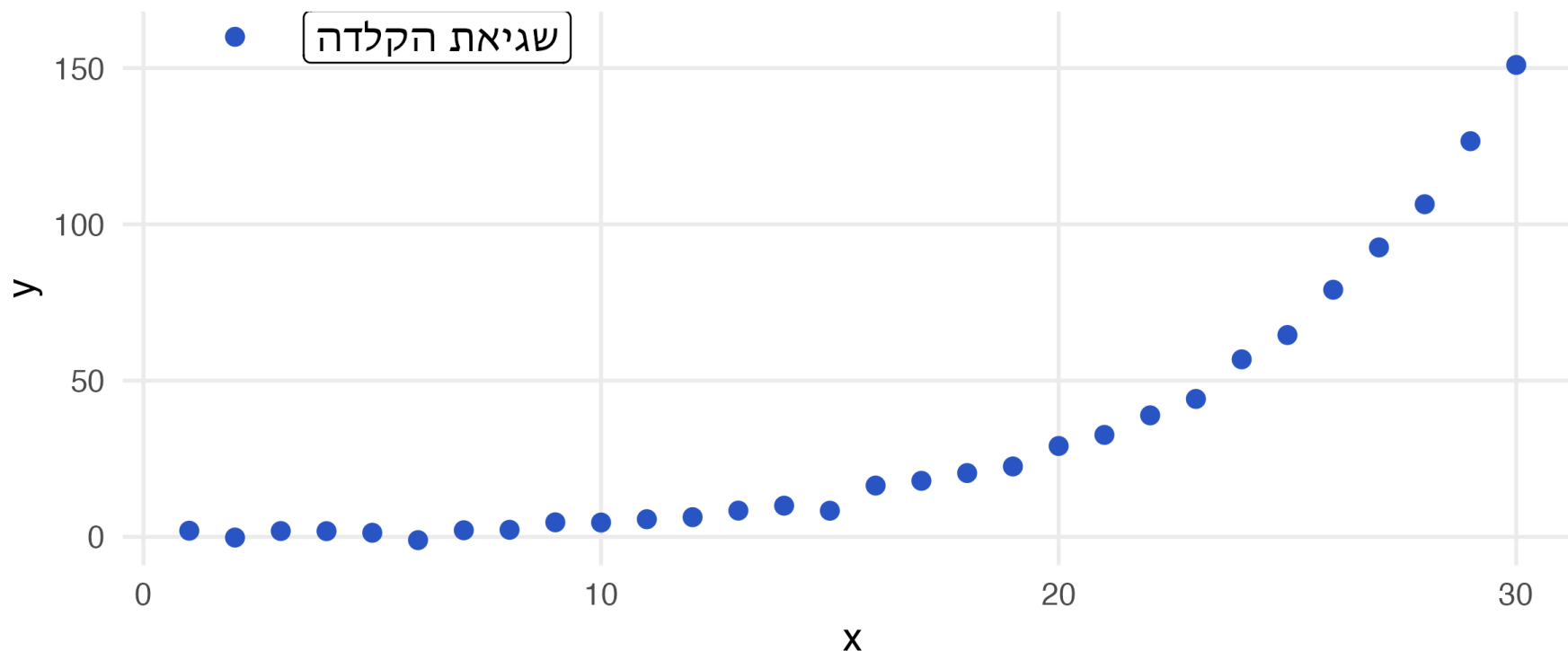
stats <- c(
  pearson = cor(corr_data$x, corr_data$y),
  spearman = cor(corr_data$x, corr_data$y, method = "spearman"),
  kendall = cor(corr_data$x, corr_data$y, method = "kendall")
) # shown in the chart title

ggplot(corr_data, aes(x, y)) +
  geom_point(size = 2.5) +
  annotate("label", x = 3.5, y = 160, label = "שגיאת הקלדה", hjust = 0) +
  labs(title = sprintf("Pearson = %.2f · Spearman = %.2f · Kendall = %.2f",
    stats["pearson"], stats["spearman"], stats["kendall"]),
    subtitle = "קשר מונוטוני ברור, עם תצפית חריגה אחת",
    x = "x", y = "y")
```

דוגמה: קשר מונוטוני וערך חריג אחד

Pearson = 0.59 · Spearman = 0.82 · Kendall = 0.81

קשר מונוטוני ברור, עם תצפית חריגה אחת



דוגמה: סטטיסטיקה מול ספרות

אותם 150 משיבים מסקר הליקרט נשאלו גם “באיזו מידה את/ה אוהב/ת ספרות?”. האם מי שאוהב סטטיסטיקה נוטה לאהוב ספרות פחות?

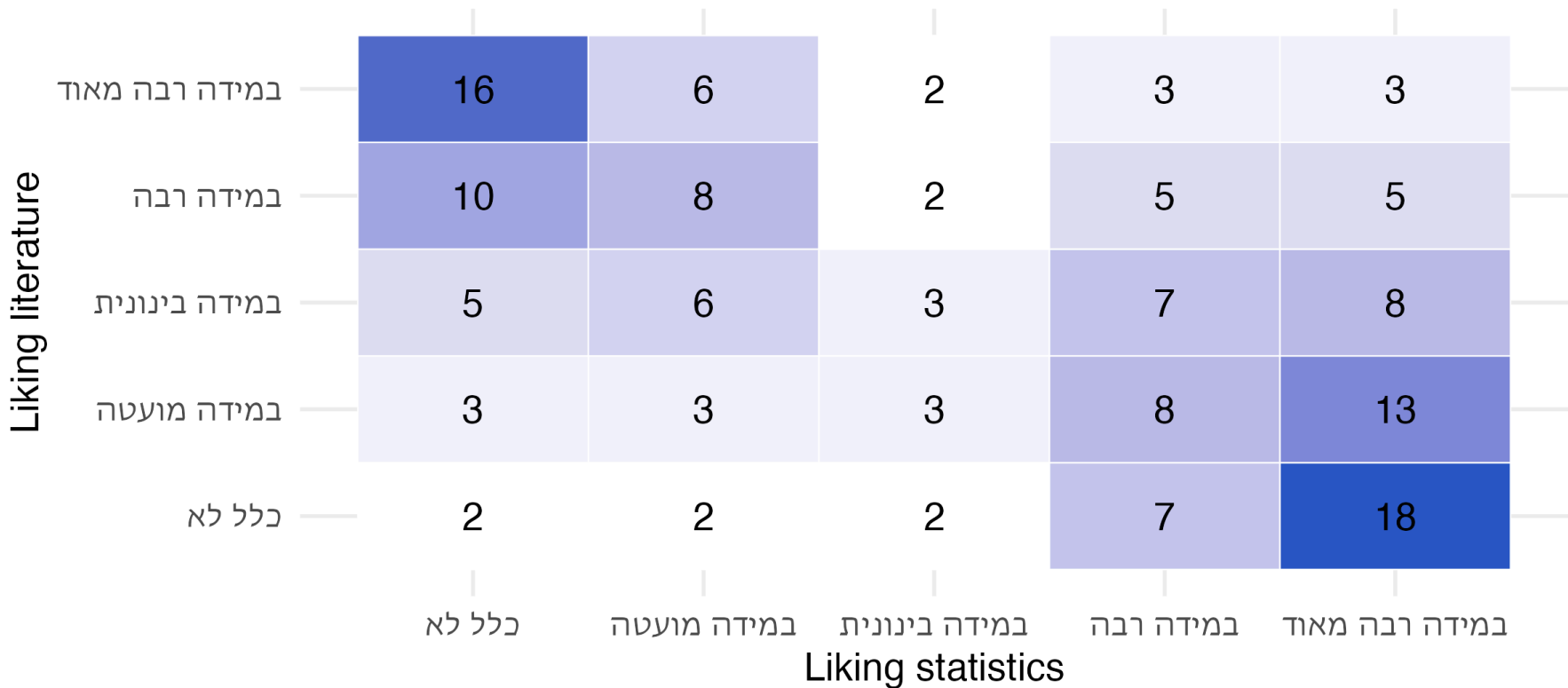
```
# Fixed cross-tab (no RNG). Rows: liking statistics (same margins as the earlier Likert slide)
counts <- matrix(c( 2,  3,  5, 10, 16,
                   2,  3,  6,  8,  6,
                   2,  3,  3,  2,  2,
                   7,  8,  7,  5,  3,
                   18, 13,  8,  5,  3), nrow = 5, byrow = TRUE)
survey <- expand_grid(stats = 1:5, literature = 1:5) |>
  mutate(n = as.vector(t(counts))) |>
  uncount(n)

survey |>
  count(stats, literature) |>
  ggplot(aes(stats, literature, fill = n)) +
  geom_tile(colour = "white") +
  geom_text(aes(label = n)) +
  scale_x_continuous(breaks = 1:5, labels = likert_levels) +
  scale_y_continuous(breaks = 1:5, labels = likert_levels) +
  scale_fill_gradient(low = "white", high = nanas_pal[["primary"]], guide = "none") +
  labs(title = "מי שאוהב סטטיסטיקה נוטה לאהוב ספרות פחות",
       subtitle = "מספר המשיבים בכל צירוף תשובות",
       x = "Liking statistics", y = "Liking literature")
```

דוגמה: סטטיסטיקה מול ספרות

מי שאוהב סטטיסטיקה נוטה לאהוב ספרות פחות

מספר המשיבים בכל צירוף תשובות



סטטיסטיקה מול ספרות: המבחן ב-R

$$H_0 : \tau = 0 \quad H_1 : \tau \neq 0$$

```
cor.test(survey$stats, survey$literature, method = "kendall", exact = FALSE)
```

Kendall's rank correlation tau

```
data: survey$stats and survey$literature
z = -6.2, p-value = 6e-10
alternative hypothesis: true tau is not equal to 0
sample estimates:
    tau
-0.4102
```

- **tau** שלילי: בזוג אקראי של משיבים, מי שאוהב סטטיסטיקה יותר נוטה לאהוב ספרות **פחות**
- **p-value** קטן מאוד, ולכן **דוחים** H_0 : יש קשר מונוטוני שלילי
- **exact = FALSE**: יש הרבה מאוד שוויונות (רק 5 ערכים לכל משתנה), ולכן R משתמש בקירוב הנורמלי עם תיקון לתיקו (**tau-b**)

סיכום

מה לקחת מההרצאה

השאלה	פרמטרי	א-פרמטרי	ב-R
מרכז ההתפלגות	t / Z	סימן, דירוגים מסומנים	<code>binom.test, wilcox.test</code>
מזווג	t מזווג	אותו מבחן על ההפרשים	<code>wilcox.test(paired = TRUE)</code>
שני מדגמים ב"ת	לשני מדגמים t	סכום הדירוגים	<code>wilcox.test(x, y)</code>
התאמה להתפלגות		חי-בריבוע, KS	<code>chisq.test, ks.test</code>
שתי התפלגויות		חי-בריבוע להומוגניות, KS לשני מדגמים	<code>chisq.test(table(g, x)), ks.test(x, y)</code>
מתאם	Pearson	Spearman, Kendall	<code>cor(method =)</code>

- שיטות א-פרמטריות לא "חינמיות": כשהנחות הפרמטריות כן מתקיימות, הן חלשות יותר
- גם להן יש הנחות (רציפות, סימטריה, אותה צורת התפלגות). צריך לבדוק אותן