1

Granularity consistency checks

(GRIM, GRIMMER)

Granularity consistency test for Means

I have 2 participants.

I calculate their mean age.

I round their mean age to 1 decimal place to report in a manuscript.

Without knowing either of their ages,

we know the rounded mean age of 2 participants must end in .0 or .5

Possible: Mean age = 23.5

Impossible: Mean age = 23.1

Granularity consistency test for Means

What if I have 3 participants?

What if I have 7 participants?

Granularity consistency test for Means

Technically, GRIM test is consistent when:

reported_mean * n_participants * n_items = total_sum

round(total_sum, 0) / (N_participants * N_items) = recalculated_mean

round(recalculated_mean, n_digits) == reported_mean

But there are simple calculators

- http://nickbrown.fr/GRIM

Exercise 1

Fifty-two university students were asked “How likely are to use ChatGPT in your coursework?”, with responses given via a 1–7 Likert-type scale (1=very unlikely; 7=very likely). Results showed a high likelihood (M = 5.22, SD = 1.33).

- Is the mean GRIM Consistent?

Exercise 1

Fifty-two university students were asked “How likely are to use ChatGPT in your coursework?”, with responses given via a 1–7 Likert-type scale (1=very unlikely; 7=very likely). Results showed a high likelihood (M = 5.22, SD = 1.33).

- Is the mean GRIM Consistent?
- Without using the calculator:
    - What would happen if the scale was 1 to 10?

Exercise 2

Festinger & Carlsmith (1959) Cognitive Dissonance Theory

Cited roughly 6000 times

- Are the means GRIM Consistent?

AVERAGE RATINGS ON INTERVIEW QUESTIONS FOR EACH CONDITION

Question on Interview	Experimental Condition		
	Control (N = 20)	One Dollar (N = 20)	Twenty Dollars (N = 20)
How enjoyable tasks were (rated from −5 to +5)	−.45	+1.35	−.05
How much they learned (rated from 0 to 10)	3.08	2.80	3.15
Scientific importance (rated from 0 to 10)	5.60	6.45	5.18
Participate in similar exp. (rated from −5 to +5)	−.62	+1.20	−.25

Exercise 2

Festinger & Carlsmith (1959) Cognitive Dissonance Theory

Cited roughly 6000 times

- Are the means GRIM Consistent?

AVERAGE RATINGS ON INTERVIEW QUESTIONS FOR EACH CONDITION

Question on Interview	Experimental Condition		
	Control (N = 20)	One Dollar (N = 20)	Twenty Dollars (N = 20)
How enjoyable tasks were (rated from −5 to +5)	−.45	+1.35	−.05
How much they learned (rated from 0 to 10)	3.08	2.80	3.15
Scientific importance (rated from 0 to 10)	5.60	6.45	5.18
Participate in similar exp. (rated from −5 to +5)	−.62	+1.20	−.25

Granularity consistency test for SDs

Just like GRIM, but for Standard Deviations (SDs)

Math is more complex but principle is the same: only some values are possible

Interpretation of GRIM inconsistency

- Maybe you made an error
  - Extracted the wrong number, made a typo etc
  - Double check
- Maybe the authors made an error
  - Reported the wrong number, made a typo, rounded it inappropriately
- Maybe the authors changed some of the real values inappropriately
  - Started with the real results, changed some values to make them more favourable (fabrication type 1)
- Maybe there was no real data at all
  - Maybe the numbers are completely invented (fabrication type 2)

51%

Of psychology articles contain inconsistencies

Brown & Heathers (2017)

Remember

- Understand how to use “N items”

- Understand when the GRIM test will always consistent, even when the reported values are errors or fabrications:
  - When the mean is reported to 1 decimal places (eg M = 4.4), GRIM stops working at N_participants * N_items = 10
  - When the mean is reported to two decimal places (eg M = 4.37), GRIM stops working at N_participants * N_items = 100
  - Etc.
  - This makes GRIM useless above this these values!

Exercise 3

- Examine the articles you've been assigned for GRIM/MER inconsistencies

- Simplest
  - http://nickbrown.fr/GRIM

- Fancier
  - https://errors.shinyapps.io/scrutiny/

2

Range consistency checks

(TIDES)

Range consistency checks

- Example 1: intervals (Haller et al., 2022)
  - “Divorced: Beta = −0.05, 95% CI [−0.50, −0.11]”

- Example 2: means
  - “Sum scores on the five-item scale (response options 1-7) … Mean = 38.1”

- Example 3: SDs
  - “1-7 Likert scale … Mean = 2.81, SD = 3.10”

Relies on information that is repeated or overlooked

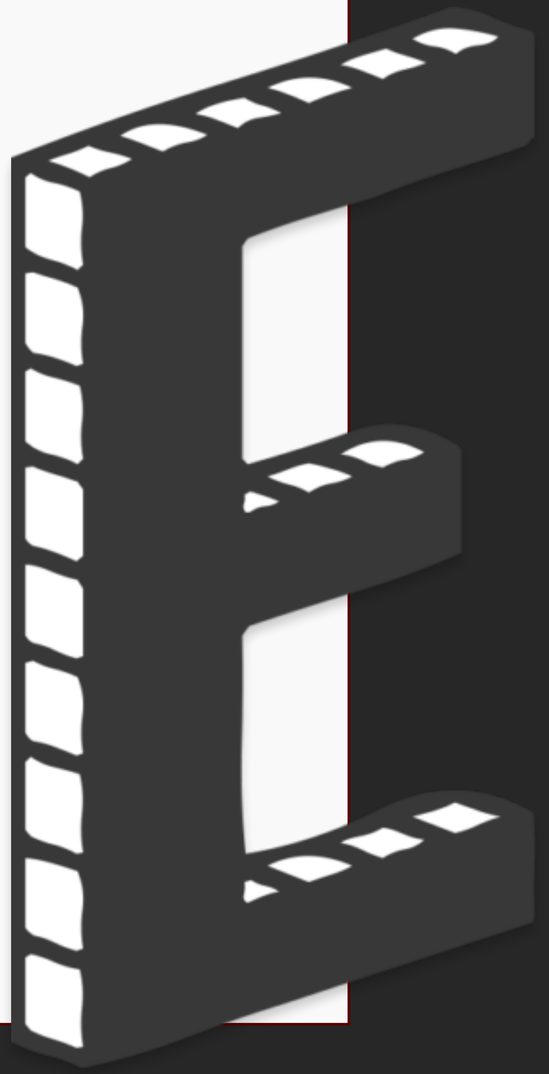
01.

Range consistency checks

Expertise problem

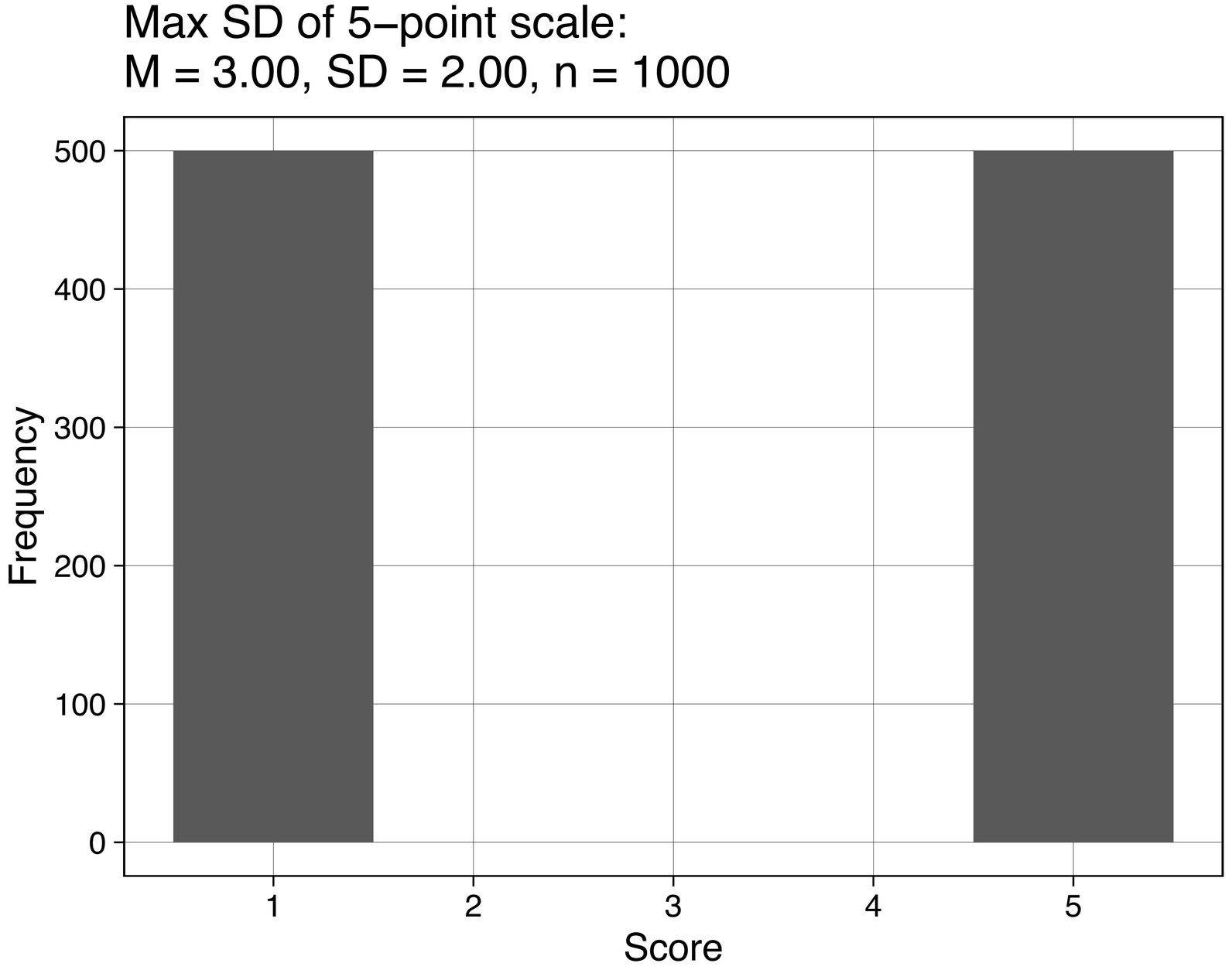
Beck Depression Inventory (BDI-II)

$N = 23, M = 20.70, SD = 3.40$

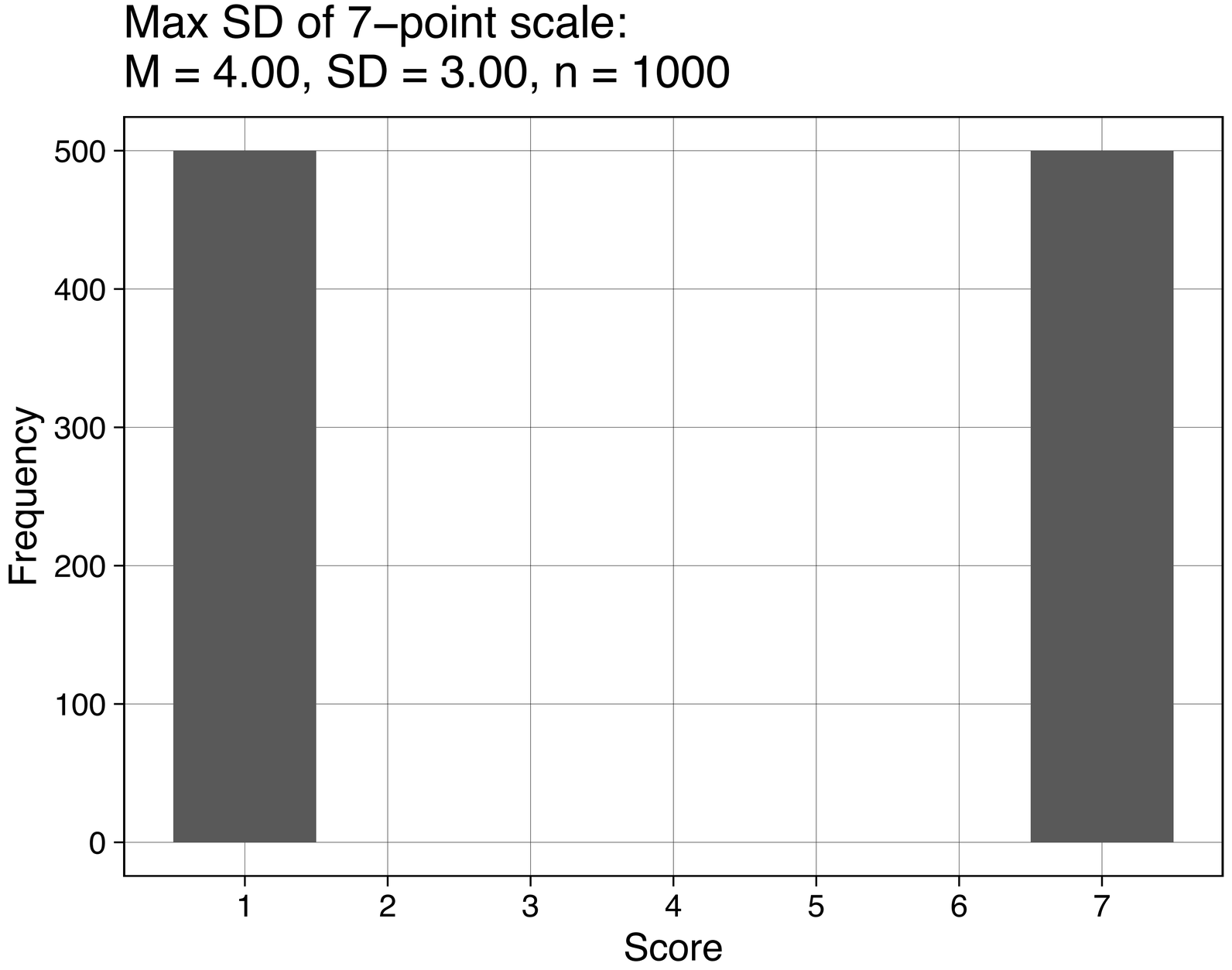


Range consistency checks for SDs

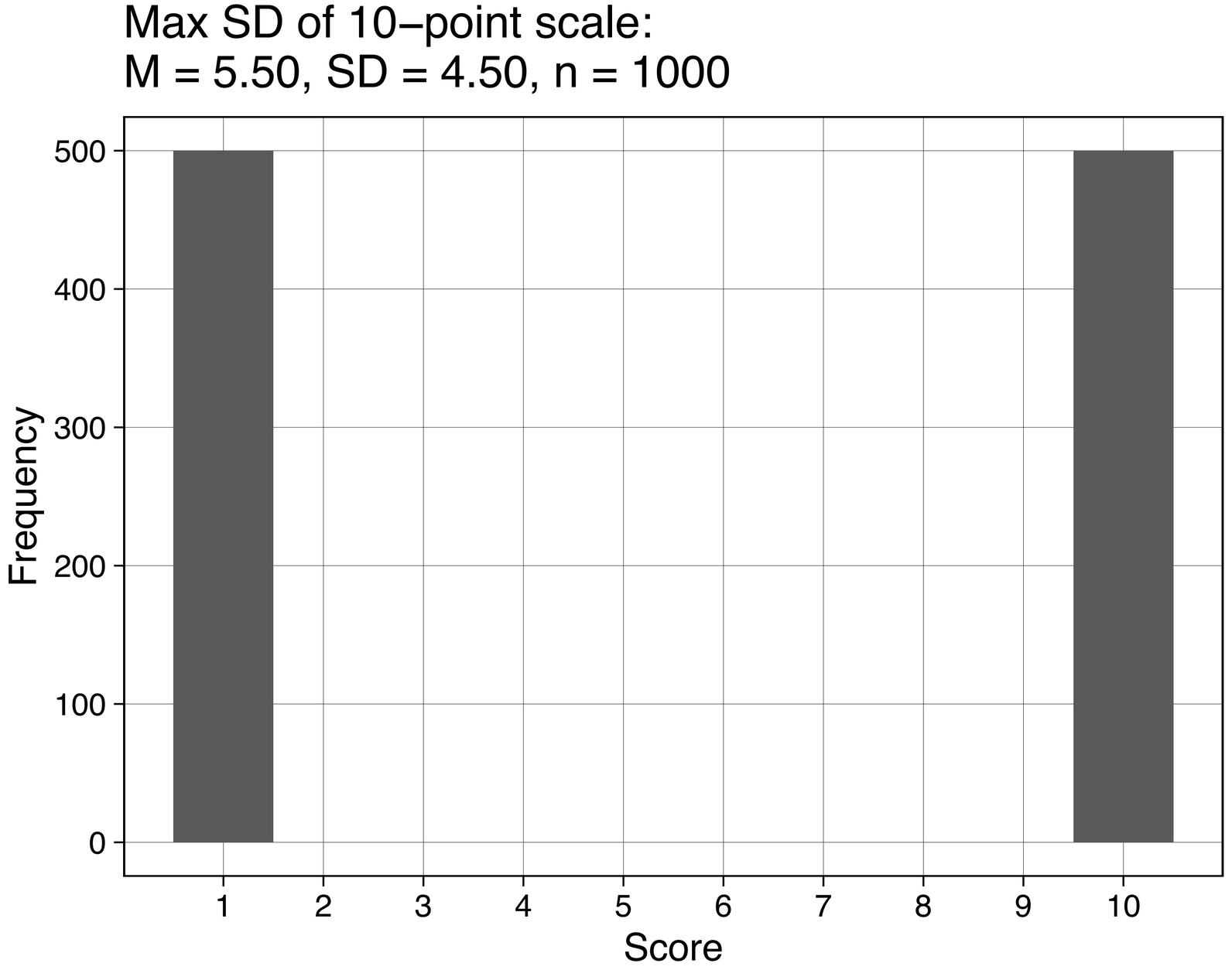
Max SD for whole scale



Max SD for whole scale



Max SD for whole scale



Max SD for whole scale

Response options	Max SD (in large N)
5	2
7	3
10	4.5
General Rule	(response_options - 1) / 2

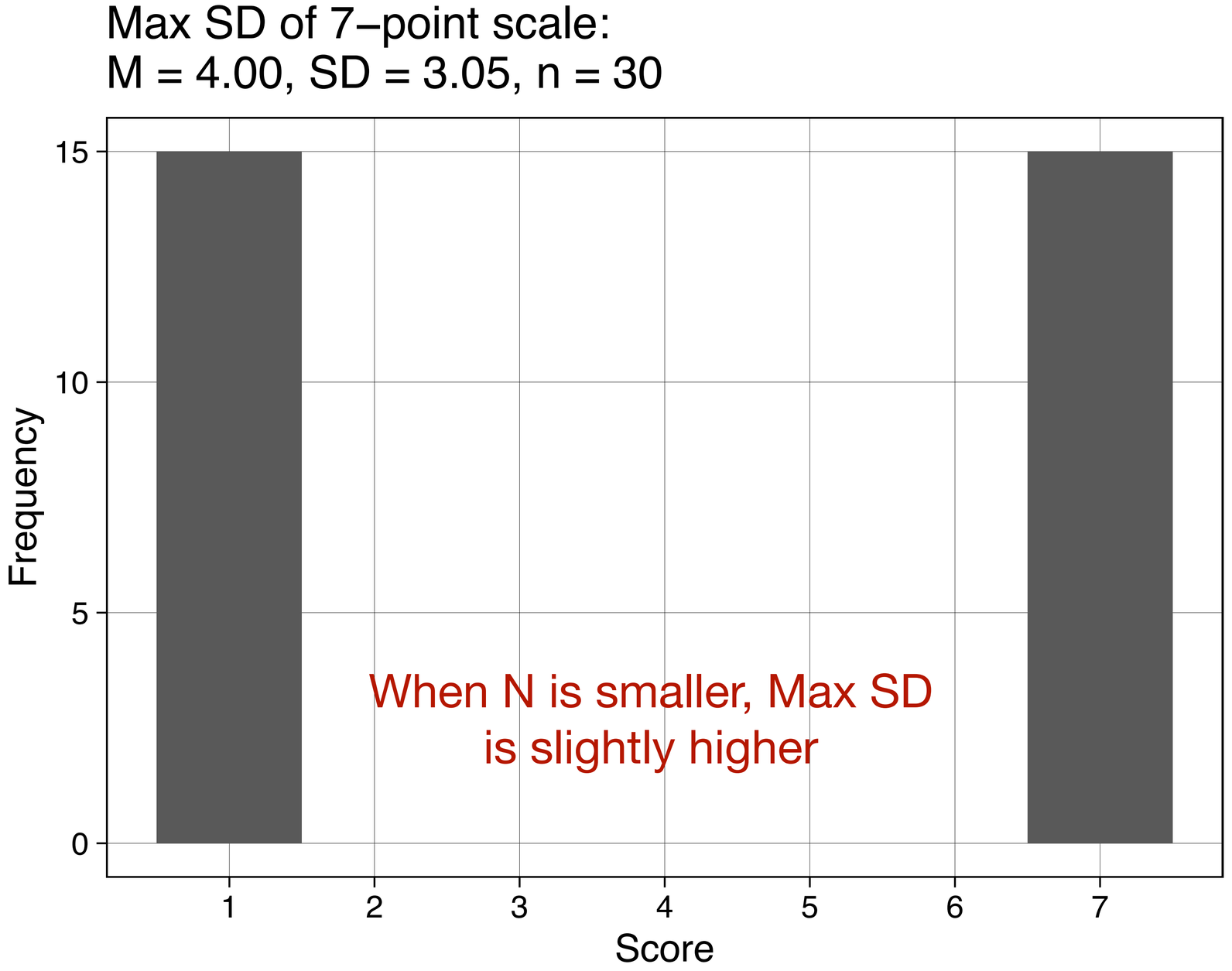
Max SD for whole scale

Example	Scale	[min, max]	SD
1	BDI	[0, 63]	13.49
2	BDI	[0, 63]	31.51
3	CES-D	[0, 60]	44.63
4	PHQ-9	[0, 27]	95.00
5	4-item Likert scale 1-7 response options mean scored	[1, 7]	3.11

Max SD for whole scale

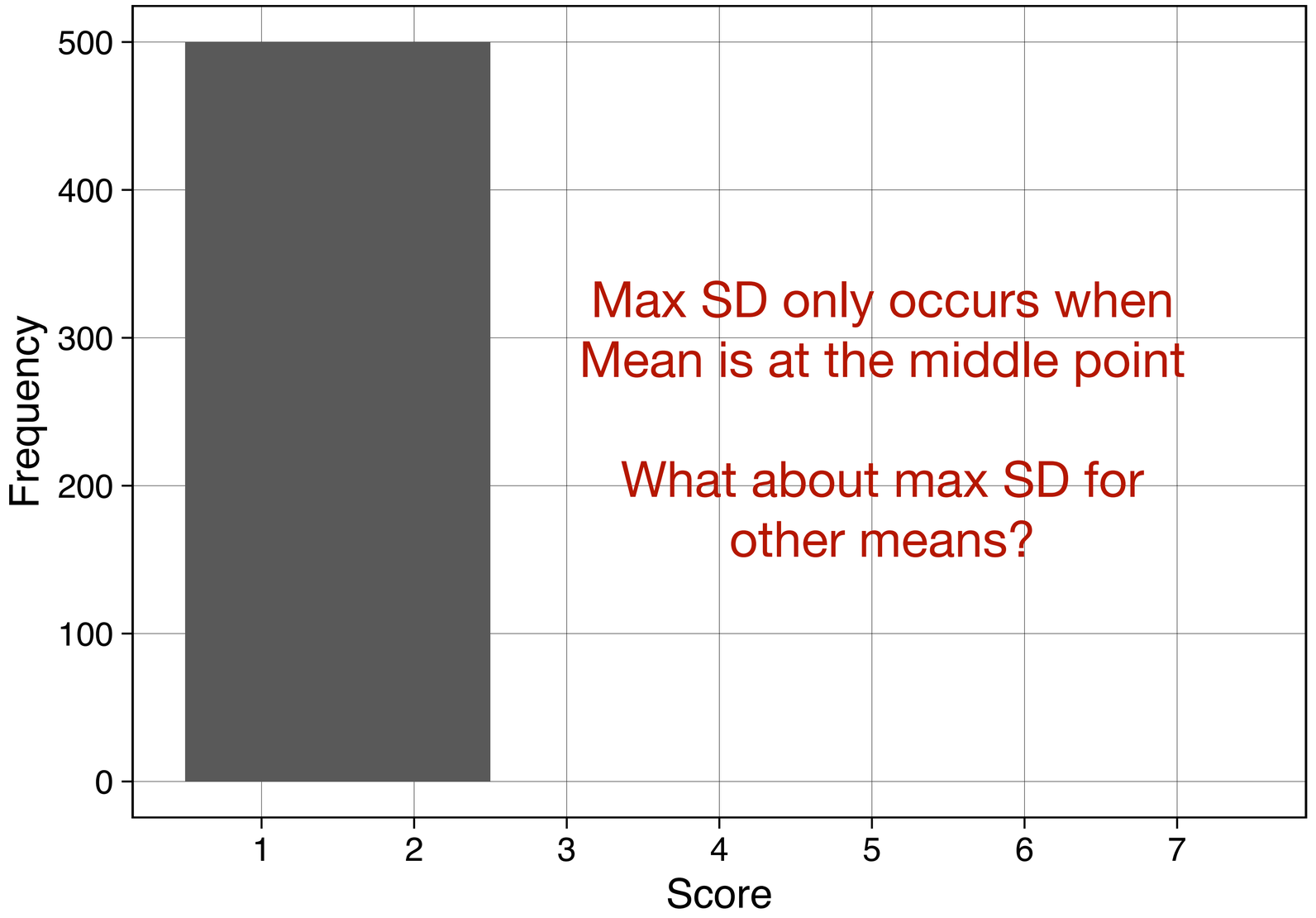
Example	Scale	[min, max]	SD	
1	BDI	[0, 63]	13.49	
2	BDI	[0, 63]	31.51	Zemestani et al. 2020
3	CES-D	[0, 60]	44.63	Ede et al. 2020
4	PHQ-9	[0, 27]	95.00	Wright et al. 2022
5	4-item Likert scale 1-7 response options mean scored	[1, 7]	3.11	

Max SD for whole scale

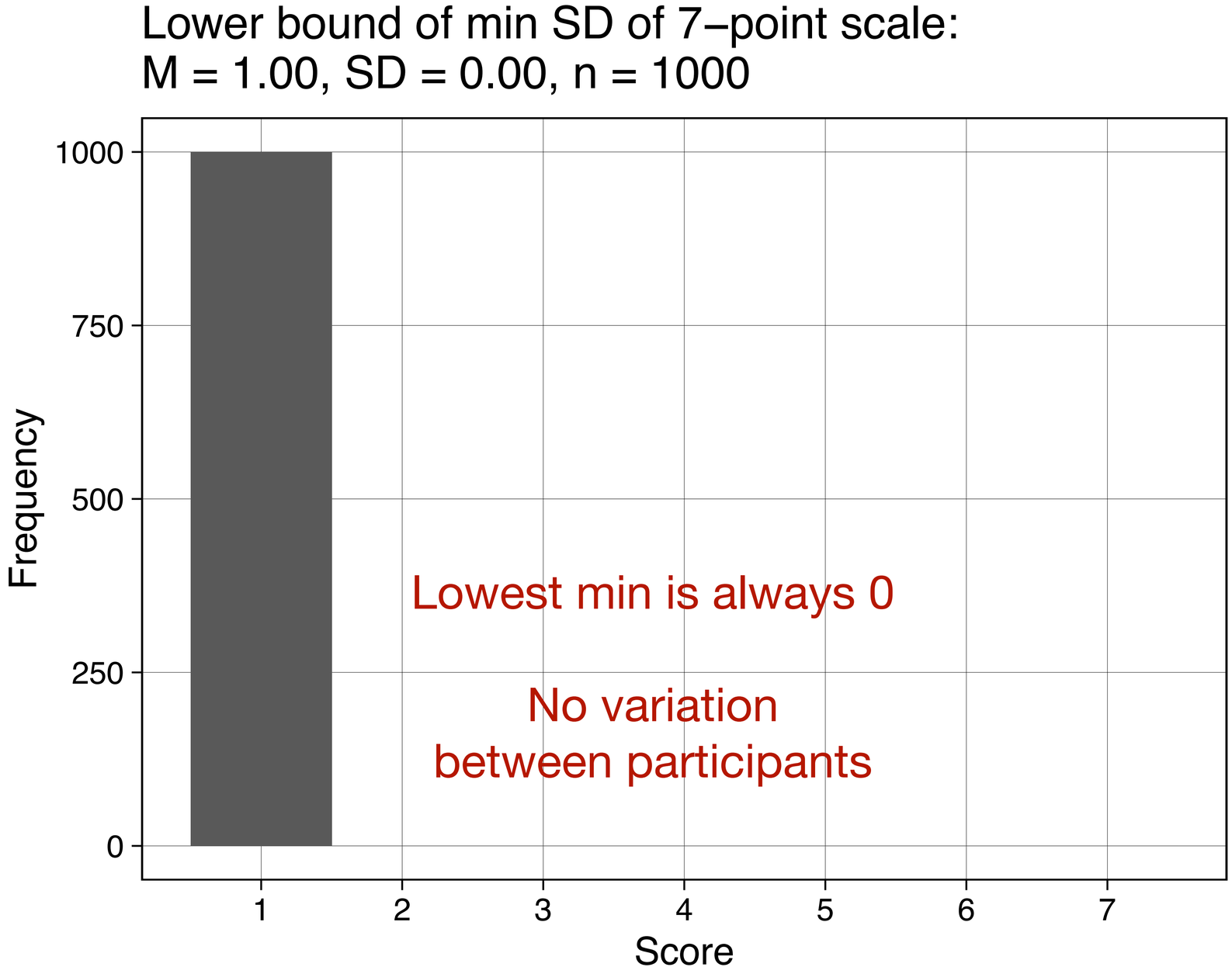


Max SD for a specific mean

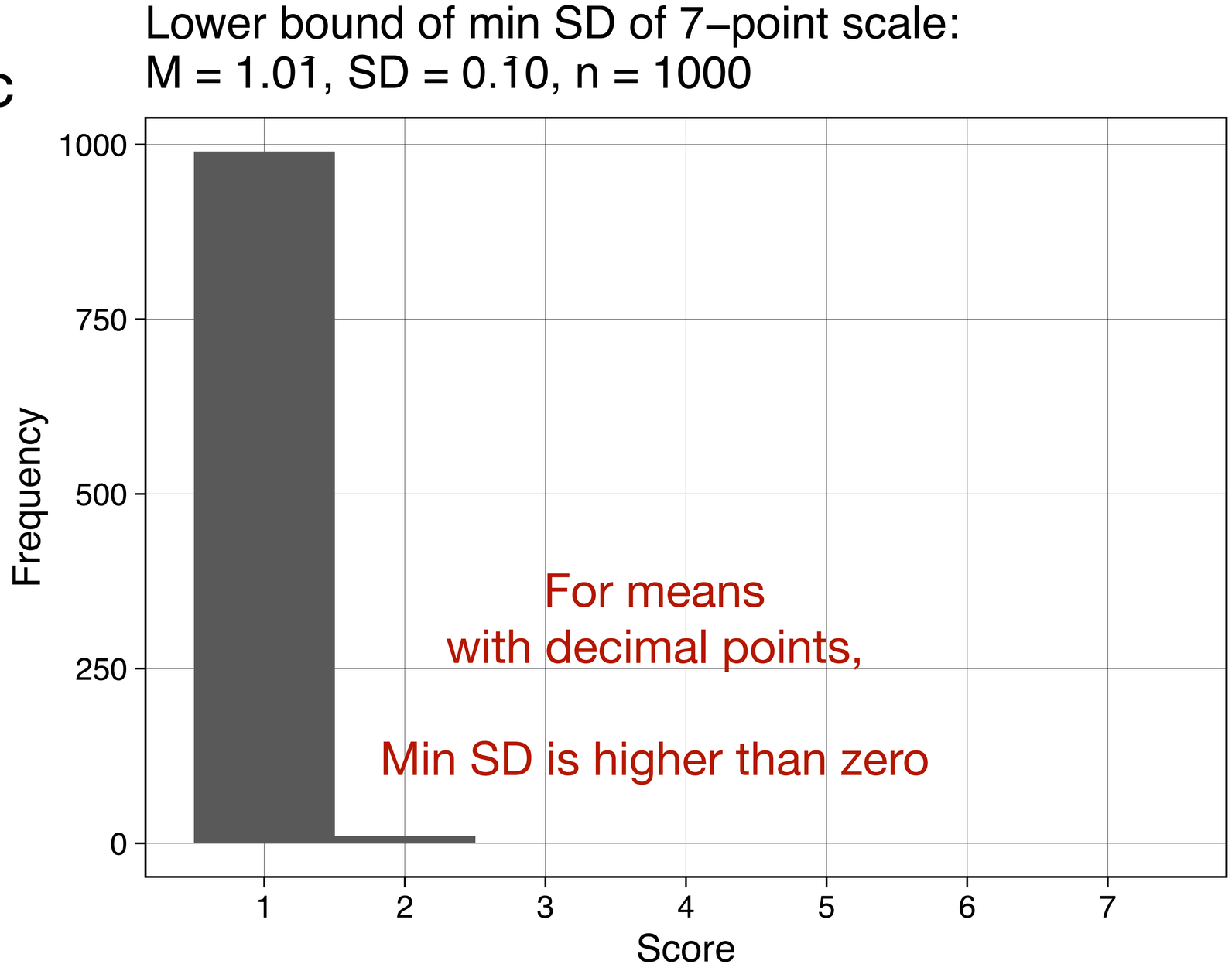
Upper bound of min SD of 7–point scale:
M = 1.50, SD = 0.50, n = 1000



Min SD for whole scale



Min SD for a specific mean



3

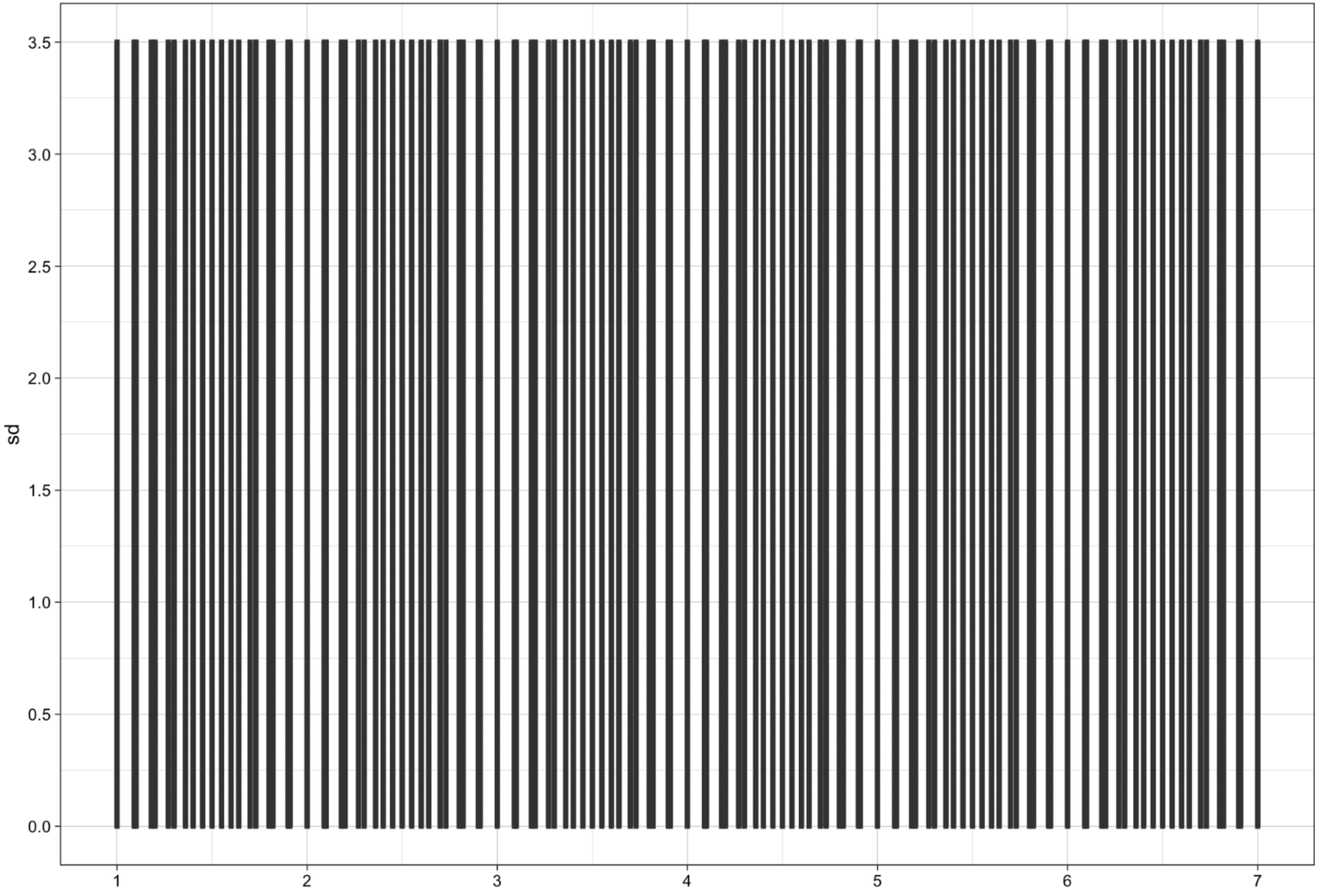
Granularity + range consistency checks

(GRIM/MER + TIDES)

Granularity

GRIM

Only some means are possible

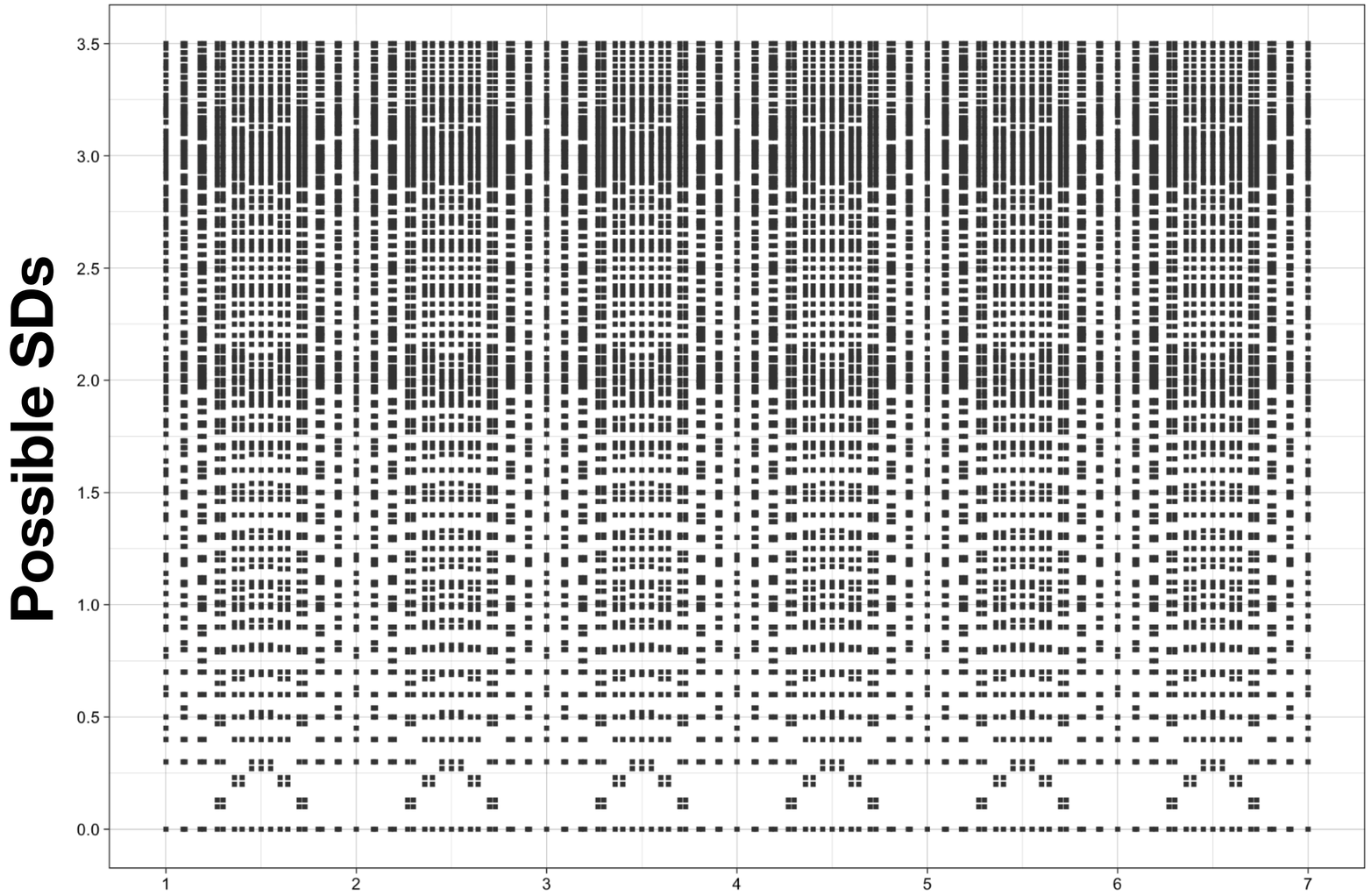


when N = 14 +
1-7 Likert scale

Granularity

GRIM + GRIMMER

Only some means, some SDs are possible

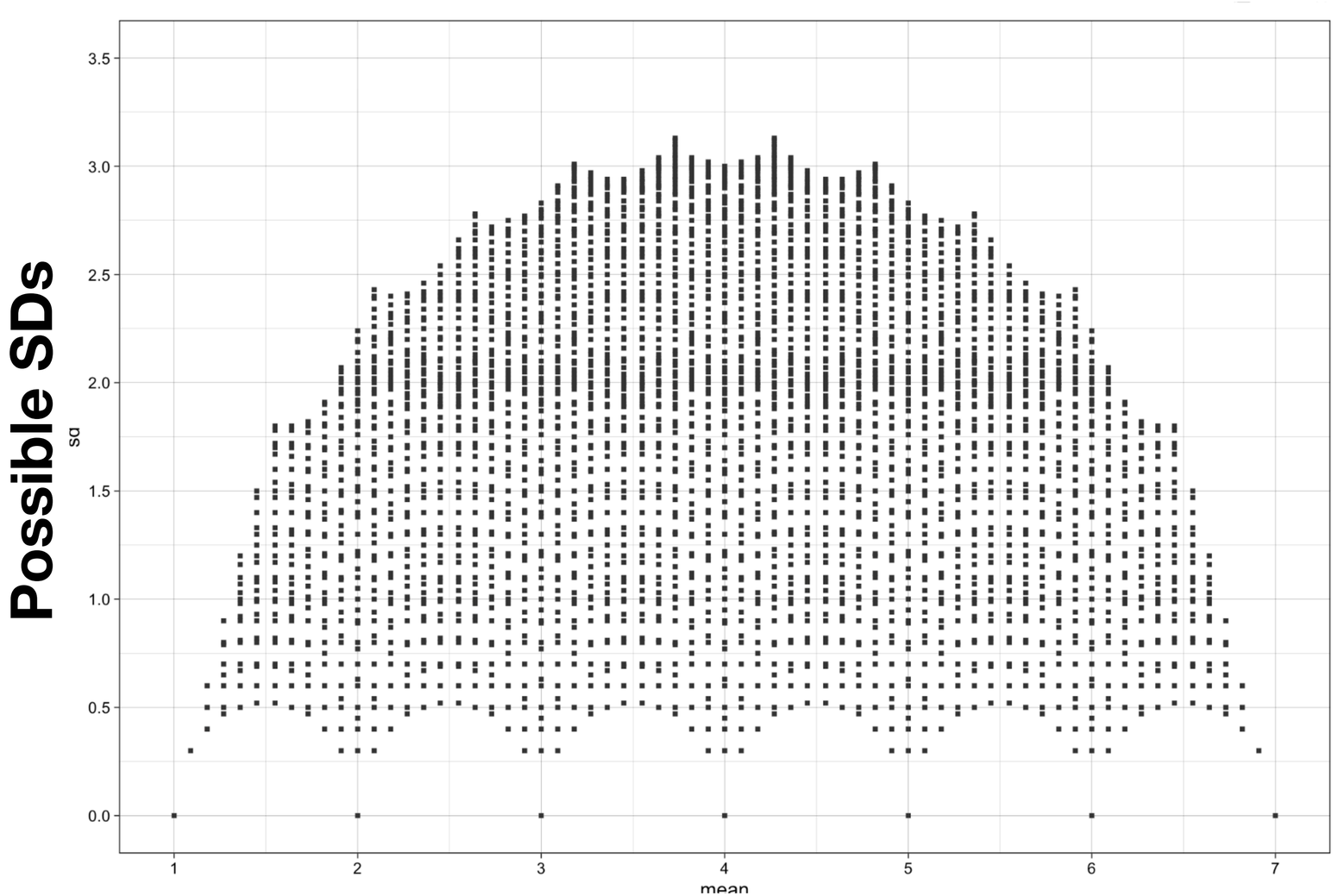


when N = 14 +
1-7 Likert scale

Granularity + range

GRIM + GRIMMER + TIDES

Only some means + SDs are possible



"Umbrella plot"
Heathers (2018)

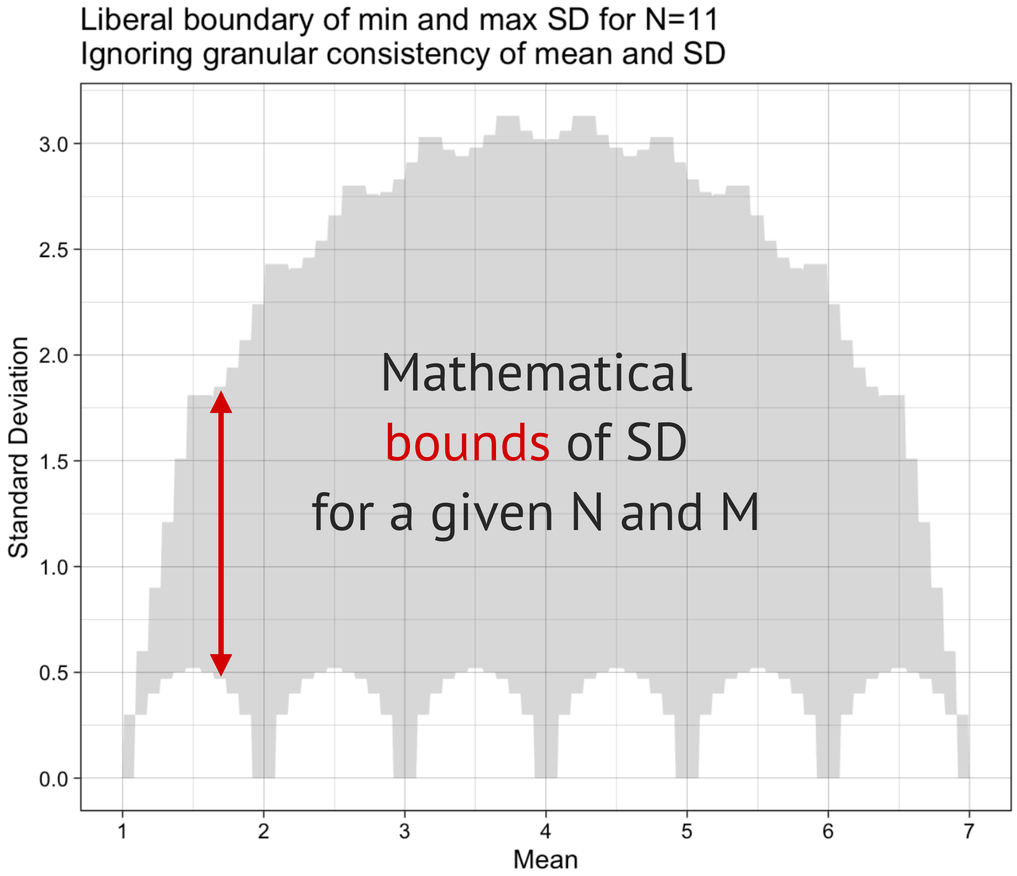
when N = 14 +
1-7 Likert scale

TIDES

Only some means + SDs are possible

Even when N > 100 and GRIM/MER stop working

Possible SDs



Even when N is too high for GRIM/MER

For any N + 1-7 Likert scale

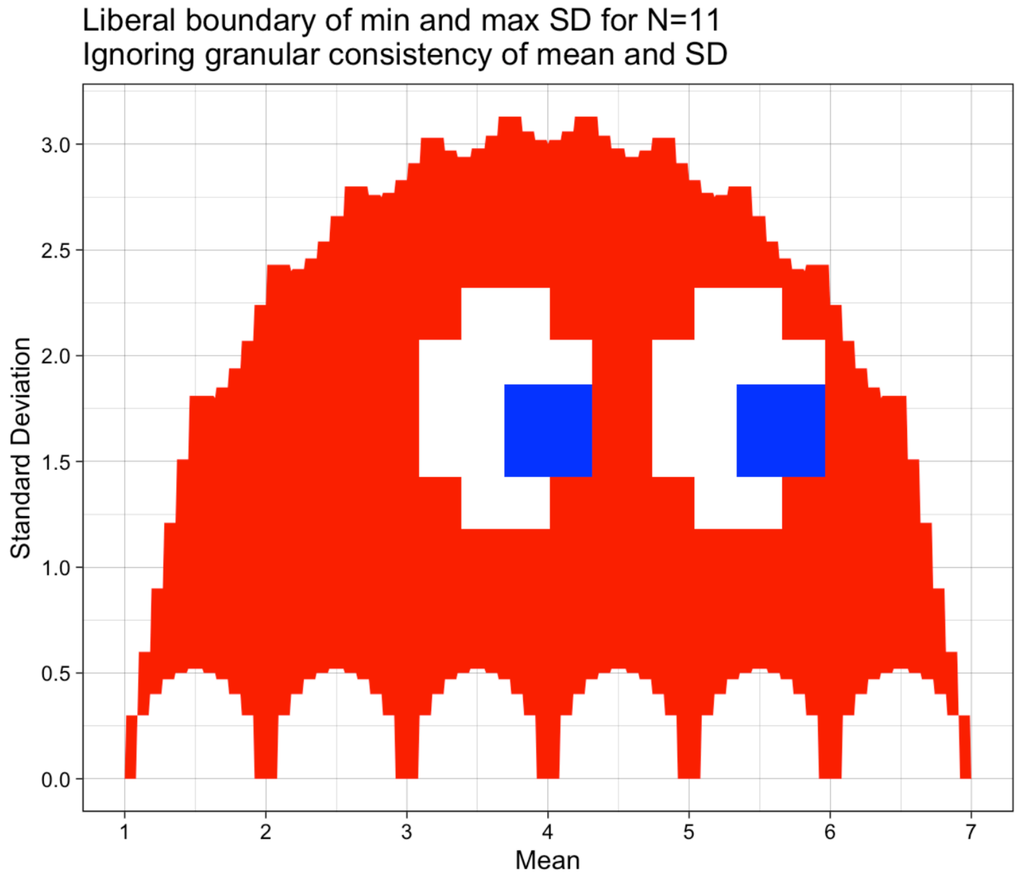
Possible Means

TIDES

Only some means + SDs are possible

Even when N > 100 and GRIM/MER stop working

Possible SDs



Even when N is too high for GRIM/MER

For any N + 1-7 Likert scale

Possible Means

TIDES

errors.shinyapps.io/TIDES

- What happens when:
  - N is very small?
  - N is very large?
  - Scale range is very small?
  - Scale range is very large?

Summary of interpretation of GRIM/MER & TIDES

Result fails GRIM/ER/TIDES

Meaning

Reported result is not possible

Causes

Typos
Errors
… *There was no dataset*

Result passes GRIM/ER/TIDES

Meaning

Reported result is possible

Causes

There was a dataset and summary stats were reported accurately

Summary of interpretation of GRIM/MER & TIDES

Could this be bad?

How?

Result passes GRIM/ER/TIDES

Meaning

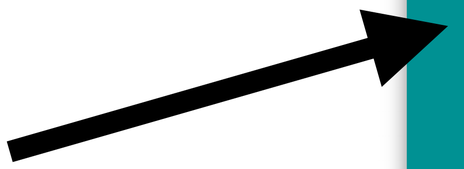
Reported result is possible

Causes

There was a dataset and summary stats were reported accurately

Summary of interpretation of GRIM/MER & TIDES

The underlying data could be fake, erroneous, or implausible



Result passes GRIM/ER/TIDES

Meaning

Reported result is possible

Causes

There was a dataset and summary stats were reported accurately

Break