Workshop 2 11:30-12:30
Using GRIM, SPRITE, RIVETS

In groups of 2-5 work through the following activities

~ GRIM Activities (20-mins) ~

Exercise 1 - Identify GRIM Inconsistencies by hand

Fifty-two university students were asked “How likely are to use ChatGPT (or an equivalent form of Generative AI) 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).

1a) Is the mean GRIM Consistent? What would happen if the scale was 1 to 10?

Twenty-one participants in their 30s rated how youthful, physi01/07/2024cally fit, and relevant they felt using Likert-type scales (0 = not at all, 7 = extremely). Each participant's ratings for these three items were averaged to create a composite "30 is the new 20" score (M = 3.77, SD = 0.63).

1b) Is the mean GRIM Consistent?

Exercise 2 - Identify GRIM Inconsistent Means using the App

Highlight any GRIM Inconsistent means in the below tables using the web app – link here.

2a) Example from Festinger & Carlsmith (1959)[1]
Let’s start with this paper, one of the most influential and well-known experiments that provided strong empirical support for the theory of Cognitive Dissonance, now cited just under 6,000 times.

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

Note down any inconsistencies you find here _____

[1]Thanks go to Dr Nick Brown for pointing out this example to me back in 2019, and for blog by Matti TJ Heino on which it is based - link here

Drs Samuel Westwood, Nick Brown

2b) Siğirci & Wansink (2015)[2]

Now let's do the same as in Exercise 2a for the below table, again using the GRIM app. These data are from the retracted Siğirci & Wansink (2015) paper, which demonstrated that the price paid for a meal influences perceptions of fullness and feelings of guilt and regret about how much food was eaten. Ignore the $4 Condition, focus on the $8 Condition.

Table 2 How price paid influence overeating

	$4 (Discounted-price)			$8 (Full-price)		
	One piece (N = 18)	Two pieces (N = 18)	Three pieces (N = 7)	One piece (N = 17)	Two pieces (N = 19)	Three pieces (N = 10)
I ate more pizza than I should have	2.63 (2.06)	4.82 (2.55)	6.00 (2.00)	1.76 (1.82)	3.53 (2.39)	4.40 (3.24)
I feel guilty about how much I ate	2.39 (1.94)	3.44 (2.47)	3.71 (1.49)	2.26 (1.79)	1.68 (1.42)	2.90 (2.08)
I am physically uncomfortable	2.17 (1.88)	2.94 (2.12)	2.43 (1.51)	1.97 (1.68)	1.45 (0.94)	2.25 (1.81)
I overate	2.11 (1.81)	3.89 (2.59)	3.71 (1.79)	1.67 (1.28)	1.67 (1.24)	3.50 (2.74)
I ate more than I should have	2.50 (2.20)	4.28 (2.44)	4.57 (2.22)	2.00 (1.45)	2.14 (1.77)	3.92 (2.81)

It's important to note that all inconsistencies in the table were explained by incorrect sample sizes or rounding errors. Significant issues in Siğirci & Wansink (2015) were not detected because even randomly generated numbers can pass the GRIM test. Thus, while GRIM can identify some errors, it is not comprehensive. Detected GRIM inconsistencies should be considered a minimum estimate, as many more errors may be undetected.

Exercise 3 - Identify odd sample sizes – example from Mazar & Zhong (2010)[3]

As well as means, GRIM can also be used to identify irregularities in sample sizes in subgroups when these are not reported in the paper (as long as the means have been correctly reported).

Mazar & Zhong (2010) asked 59 undergraduate students to complete a survey. They were randomly assigned to rate either a person who purchases organic foods and environmentally friendly products or a person who purchases conventional foods and products. They used a 7-point scale (1 = not at all, 7 = very) to indicate how cooperative, altruistic, and ethical they thought such a person was. The mean and standard deviation of their ratings are provided in the below Table 1.

Note that at no point do the authors specify the sample size for each group, just the total sample of 59. No matter, we can often infer those using GRIM!

[2]This exercise is taken from van der Zee, T., Anaya, J. & Brown, N.J.L. Statistical heartburn: An attempt to digest four pizza publications from the Cornell Food and Brand Lab. BMC Nutr 3, 54 (2017). https://doi.org/10.1186/s40795-017-0167-x. You may find the original article in question here: Siğirci, Ö., Wansink, B. RETRACTED ARTICLE: Low prices and high regret: how pricing influences regret at all-you-can-eat buffets. BMC Nutr 1, 36 (2015). https://doi.org/10.1186/s40795-015-0030-x

[3]This exercise was based heavily off the following 2018 blog post by James Heathers – link here

Drs Samuel Westwood, Nick Brown

Table 1. Mean (SD) of ratings for purchases of organic foods or environmentally friendly "Green" products versus conventional foods or products

Attribute	"Green" Group (N=?)	"Conventional" Group (N=?)
Cooperative	4.75 (1.37)	3.62 (1.76)
Altruistic	5.07 (1.01)	3.36 (1.23)
Ethical	5.55 (1.44)	3.36 (1.70)

To do this, we need to determine a common denominator (i.e., sample size) that can accommodate the eventual means of 4.75, 5.07, and 5.55 for one group, and 3.62 and 3.36 for the other group. Additionally, the total sample size for both groups should equal *n*=59.

Let's start by identifying GRIM Consistent sample sizes for the "Conventional" Group for means 3.62 and 3.36. I've tested sample sizes of 1 to 38 without success – i.e., none were consistent - so we need to check sizes 39 to 59.

I recommend one person tests each sample size using the GRIM app while another records "Y" for consistent or "N" for inconsistent in the table below.

	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59
3.36																					
3.62																					

3a) What are the calculated sample sizes for the for the "Conventional" Group?

List the GRIM Consistent sample sizes consistent for both means here:

_____

Now let's do the same for the "Green" Group – i.e., identify GRIM Consistent sample sizes for means 4.75, 5.07, and 5.55.

	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59
4.75																					
5.07																					
5.55																					

3b) What are the calculated sample sizes for the for the "Green" Group?

List the GRIM Consistent sample sizes consistent for all three means here:

_____

3c) Notice anything strange about the sample sizes?
Examine the sample for each group, what strikes you as being odd?

Drs Samuel Westwood, Nick Brown

SPRITE Activities (20-mins)

Exercise 4 – Can you get children to eat more carrots?[4]

Wansink et al. (2012) demonstrated that giving carrots more appealing names significantly boosts their consumption at school dinners in elementary school. The paper was eventually retracted thanks to several oddities, which we will explore below.

Study 1: elementary students consumed more carrots when attractively named

	Named as "X-ray Vision Carrots" (n=32) Mean (SD)	Named as "Food of the Day" (n=38) Mean (SD)	Unnamed (Control) (n=45) Mean (SD)
Number taken	17.1 (17.6)	14.6 (14.5)	19.4 (19.9)
Number eaten	11.3 (16.3)	4.7 (6.7)	6.8 (8.7)
Number uneaten	6.7 (9.6)	10.3 (12.5)	13.2 (16.9)

4a) Notice anything odd?

The table above displays the number of carrots taken by children and, of these, how many were eaten and how many were uneaten. Examine the values reported in each column.

Do you notice anything unusual or unexpected about these figures? Consider how the values in each column relate to each other.

Do you notice anything odd about the mean and standard deviation of the number of carrots taken in the control group?

4b) Use the SPRITE shiny app

Using the SPRITE shiny app, inspect the possible distributions of the values you examined in the control group in Exercise 1 (i.e., the number of carrots eaten, M=19.4, SD=19.9, n=45).

To run the app, SPRITE needs the reported mean, standard deviation, sample size and range of values. I have provided them below.

a) **Minimum Scale Value** = 0.
b) **Maximum Scale Value** = 50[5].
c) **Sample Size** = 45
d) **Target Mean** = 19.4
e) **Target SD** = 19.9
f) **Number of Results** = 4. This parameter determines how many plotted results you want to generate. Feel free to adjust this number later to explore different outcomes. Note that if you check the box labelled "Use fixed seed" you will always get the same results, which can sometimes be useful.
g) **Decimal Places** = 1 (because your input mean and SD have 1dp of precision)

You can ignore **Fixed Value**, **Fixed Count**, and **Use fixed seed** for this exercise.

[4]This was based off another 2018 blog post by James Heathers – link here

[5]The minimum and maximum number of carrots eaten were not reported, so we will assume a range of possible carrots that could be consumed by a child at lunchtime. Normally, in the absence of reported data, we would assume theoretically consistent or plausible values. Clearly, 50 is wildly improbable (approximately 300grams of baby cut carrots!), but the reason for setting the maximum so high will become clear as you proceed with the exercise.

Drs Samuel Westwood, Nick Brown

4c) Examine the four plots.
Do you notice anything unexpected about the values? Remember that the plots are simulations of datasets that could potentially fit the summary statistics[6].

Exercise 5 – "Lady in red, is dancing with me…"[7]

In a now retracted paper, Nicolas Guéguen (2012)[8] asked participants to rate the attractiveness and sexual intent of a woman in a photograph wearing a red, blue, green, or white t-shirt. Men significantly attributed higher sexual intent to the woman in the red clothing condition, which was not influenced by the attractiveness rating.

Using SPRITE, examine the distribution of ages reported by Guéguen (2012). You will need to download the paper for this exercise – link here.

RIVETS Activities (20-mins)

Exercise 6 – Is this RIVETing stuff?

Read the following extract from an article in the *Journal of Riveting Research*.

Panels constructed with steel rivets (*N*=78) performed better than panels constructed with aluminium rivets (*N*=75) on all three outcome measures: flexibility (*M*=33.95, *SD*=4.75 versus *M*=28.55, *SD*=5.01; *t*(151)=6.84), durability (*M*=78.80, *SD*=7.54 versus *M*=67.91, *SD*=7.27, *t*(151)=9.09; and production cost (*M*=19.24, *SD*=3.71 versus *M*=15.66, *SD*=3.80, *t*(151)=5.90).

6a) RIVETS Calculations
For each of the results reported above, calculate the smallest, largest, and nominal *t* statistic corresponding to the reported rounded means and SDs. You may use the table provided, and the GraphPad app to calculate the t-tests.

Flexibility

N1	M1	SD1	N2	M2	SD2	t	Type of t
							Nominal
							Smallest
							Largest

Durability

N1	M1	SD1	N2	M2	SD2	t	Type of t
							Nominal
							Smallest
							Largest

Production Costs

6You can also download the simulated data in a CSV file to explore them with other software, even if that's just to verify that the means and SDs exactly match what you asked for.
[7]This exercise is based of a very illuminating thread on the history of this paper and its ultimate retraction in Nov 2022 – link here.
[8]Guéguen, N. (2012). RETRACTED ARTICLE: Color and Women Attractiveness: When Red Clothed Women Are Perceived to Have More Intense Sexual Intent. *The Journal of social psychology*, *152*(3), 261-265.

Drs Samuel Westwood, Nick Brown

N1	M1	SD1	N2	M2	SD2	t	Type of t
							Nominal
							Smallest
							Largest

6b) Evaluate your RIVETs

Having completed Exercise 6b, how likely do you think it is that this study was faked, with the test statistics having been calculated “by hand” from fabricated summary statistics?

Drs Samuel Westwood, Nick Brown

ANSWERS

1a) Firstly, no, the mean is GRIM *Inconsistent*. The first step is to multiply the mean by the sample size (6.28 x 52 = 326.56). Next, round this product to the nearest integer, resulting in 327. Then, divide this integer by the sample size and round the result to two decimal places: (327 / 52 = 6.29). Finally, compare this result with the original mean. If they are identical, the mean is consistent with the sample size and integer data; if they differ, as in this case (6.28 versus 6.29), the mean is inconsistent.

Secondly, nothing would happen if the scale was, say, 0 to 10 rather than 1 to 7. The GRIM technique is independent of the possible values of the measured variable, as long as they are integers or (in the case of composite items) can be represented as integers divided by a sufficiently small number. Thus, the number of response options for a Likert-type item is irrelevant.

1b) No, it is GRIM Inconsistent. The possible values for the mean score can be 1.00, 1.33, 1.66, 2.00, 2.33, 2.66, 3.00, etc. This level of granularity is finer compared to reporting integer scores, as it reflects the mean of the means of the three components rather than the total scores. Despite the small sample size, a GRIM test can be performed by multiplying the sample size by the number of items averaged for the composite measure (i.e., three).

Here's the process:

1. Multiply the sample size (21) by the number of items (3) to get 63.
2. Multiply 63 by the mean score (3.77) to get 237.51.
3. Round 237.51 to the nearest whole number, resulting in 238.
4. Divide 238 by 63 to get approximately 3.777, which rounds to 3.78.

This calculated mean (3.78) is inconsistent with the reported mean (3.77) given the sample size, indicating a potential discrepancy.

Drs Nick Brown and James Heathers have provided an Excel spreadsheet that automates these calculations, which is available at https://osf.io/3fcbr.

2a,b)

TABLE 1

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

Table 2 How price paid influence overeating

	$4 (Discounted-price)			$8 (Full-price)		
	One piece (n = 18)	Two pieces (n = 18)	Three pieces (n = 7)	One piece (n = 17)	Two pieces (n = 19)	Three pieces (n = 10)
I ate more pizza than I should have	2.63 (2.06)	4.82 (2.55)	6.00 (2.00)	1.76 (1.82)	3.53 (2.39)	4.40 (3.24)
I feel guilty about how much I ate	2.39 (1.94)	3.44 (2.47)	3.71 (1.49)	2.26 (1.79)	1.68 (1.42)	2.90 (2.08)
I am physically uncomfortable	2.17 (1.88)	2.94 (2.12)	2.43 (1.51)	1.97 (1.68)	1.45 (0.94)	2.25 (1.81)
I overate	2.11 (1.81)	3.89 (2.59)	3.71 (1.79)	1.67 (1.28)	1.67 (1.24)	3.50 (2.74)
I ate more than I should have	2.50 (2.20)	4.28 (2.44)	4.57 (2.22)	2.00 (1.45)	2.14 (1.77)	3.92 (2.81)

Drs Samuel Westwood, Nick Brown

3a) GRIM Consistent Ns for the "Conventional" Group = 39, 42, 45, 47, 50, 53, 55, 58
3b) GRIM Consistent Ns for the "Green" Group = 44, 55, 56
3c) Two things should strike you as odd. First, you would expect a more balanced set of sample sizes across each group, ideally around 30, give or take one or two or even three. Second, even if randomization led to a sizeable imbalance in sample sizes between the two groups, the total sample size must be 59. Yet, this total is impossible to achieve with any combination of sample sizes that are GRIM Consistent.

4a) Two things should strike you as odd. First, the number of carrots taken is not the sum of the number eaten and uneaten. If you start with a certain number of items [A], and then consume some of them [B], leaving the rest uneaten [C], the total number of items [A] should equal the sum of those eaten [B] and those uneaten [C]. If this equation does not hold true, an error has been made. For example, consider the first column, 17.1 is not the sum of 11.3 and 6.7, but it should be. Second, number of carrots taken in the control group (i.e., Mean=19.4, SD=19.9, n=45) implies that there were negative carrots[9].

4b) What should immediately catch your attention is that in most of the simulated distributions, an implausibly high number of carrots were apparently taken by some children, which in some instances reached as high as 50. This means that among the plausible datasets generating a mean of 19.4 and an SD of 19.9 for a sample of 45 children, many children could be consuming nearly 300 grams of carrots…This is particularly striking given that these data were based on the control group – i.e., where the carrots were *not* renamed, so as to establish a baseline for what children aged 8 and 11 normally eat. What is more, these were the number of carrots taken, meaning that a someone was behind the counter dishing out exorbitant amounts of carrots!

Note, SPRITE is not calculating a probability for a specific distribution. Instead, you're using the M/SD and limits to show a range of possible distributions, so caution is needed. With GRIM, after verifying multiple-item measures, you can directly report inconsistencies to an editor since they must be corrected. SPRITE acts like a smell test, prompting further investigation for objective evidence. In cases with smaller samples and extreme distributions, SPRITE can be very accurate.

5) Notice that there is a high number of both 18 and 21-year-olds, with 18-year-olds bing the majority in most generated age distributions of the participants for a given mean of 19.2, SD of 1.35, and age range of 18 to 21.

6a)

[9]For anyone interested, Wansink made a comment on this in an interview for Retraction Watch – link here – where he tried to excuse this discrepancy by saying, "For anybody who can remember school lunches as an 8 year old – some amount of it ends up on the floor or in pockets." Thanks to Dr Brown for pointing this out, and for noting that '"eaten" wasn't determined by a camera in each child's gullet; it was *defined* (and should have been calculated) as "taken minus uneaten", regardless of where the food physically ended up.

Drs Samuel Westwood, Nick Brown

Flexibility

N1	M1	SD1	N2	M2	SD2	t	Type of t
78	33.95	4.75	75	28.55	5.01	6.84	Nominal
78	33.945001	4.754999	75	28.554999	5.014999	6.82	Smallest
78	33.954999	4.745001	75	28.545001	5.005001	6.86	Largest

Durability

e	M1	SD1	N2	M2	SD2	t	Type of t
78	78.80	7.54	75	67.91	7.27	9.09	Nominal
78	78.795001	7.544999	75	67.914999	7.274999	9.07	Smallest
78	78.804999	7.535001	75	67.905001	7.265001	9.10	Largest

Production Costs

N1	M1	SD1	N2	M2	SD2	t	Type of t
78	19.24	3.71	75	15.66	3.80	5.90	Nominal
78	19.235001	3.714999	75	15.664999	3.804999	5.87	Smallest
78	19.244999	3.705001	75	15.655001	3.795001	5.92	Largest

The smallest *t* value is obtained by making the larger mean as small as possible (consistent with its rounded value), the smaller mean as large as possible, and both SDs as large as possible (because the SDs appear — converted to a pooled standard error — in the denominator of the calculation of the *t* statistic). Conversely, the largest *t* value is obtained by making the larger mean larger, the smaller mean smaller, and both SDs as small as possible.

6b) There is no correct answer here! Naively we might say that there were five possible rounded values for flexibility, four for durability, and six for production cost, and the authors just happened to report the nominal value in each case, so that was one chance in (5x4x6)=120. But things are actually more complicated than that, as the nominal values are somewhat more likely to occur than the extreme ones. Perhaps the actual chance of this result is one in 30, which is a low threshold to start thinking about fraud, let alone accusing anyone of it. If we had 10 such "perfect" results instead of three, though, we would definitely be justified in asking the authors for their dataset!

Drs Samuel Westwood, Nick Brown