The Art of Statistics – Anna Karenina Principle - 2

BS3033 Data Science for Biologists

Dr Wilson Goh
School of Biological Sciences

Learning Objectives

By the end of this topic, you should be able to:

- Describe non-averaging statistics.

NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

Nonaveraging Statistics

BS3033 Data Science for Biologists

Dr Wilson Goh
School of Biological Sciences

Going Against the Mean!

Many test statistics include a component of centrality (averaging) as a means of summarising data; this includes the mean and median in the t- and U-tests respectively.

These components of centrality are used, assuming that the average (mean, median, mode) makes informative summarisations.

But there are situations when averages are not informative.

The Mean isn't Necessarily Informative

Correlation measures the extent of information in one variable about another, independent of the absolute or average difference between two variables and allows prediction of one variable from another. Averaging is not necessarily related to correlation. Indeed, their real-life data examples suggested that by removing one sample as the holdout, and computing the minimum/ maximum/ mean/ variance over all variables with remaining samples, the maximum and variance often conveys more information on correlation than the mean.

Mitra and Shugan, When and Why Nonaveraging Statistics Work, 2009

In Biology…

It is common in biology for relative changes to be more germane than incremental ones. There are two principal reasons for this. One is that certain biological phenomena can only be properly described and understood through relative changes.

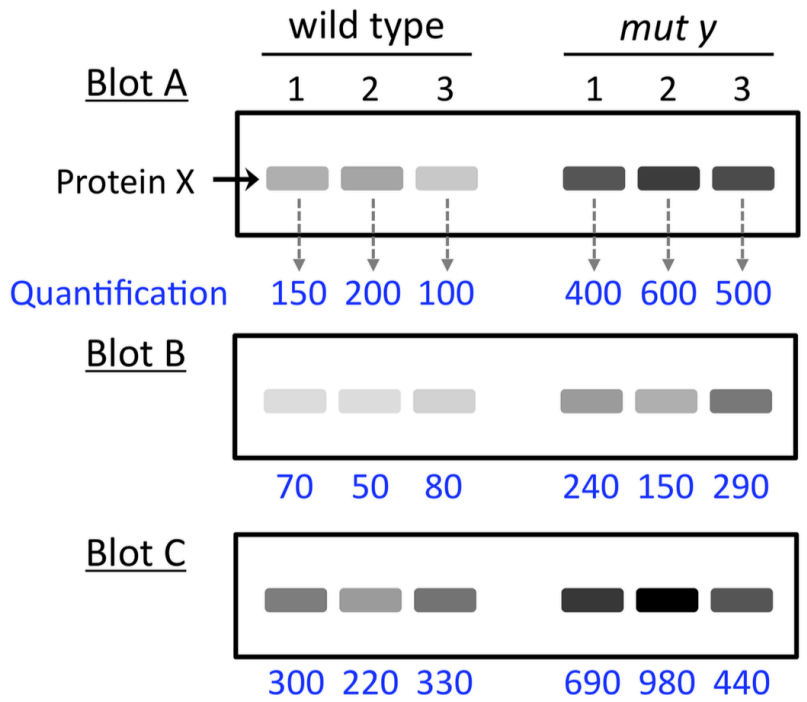
If we were to count the number of bacterial cells in a specified volume of liquid culture every hour, we might derive the following numbers: 1,000, 2,000, 4,000, 8,000, 16,000. The pattern is clear; the cells are doubling every hour.

Conversely, it would be ridiculous to take the mean of the observed changes in cell number and to state that, on average, the cells increase by 3,750 each hour with a 95% CI of −1,174.35 to 8,674.35!

In Biology…

The second reason is due to experimental design. There are many instances where variability between experiments or specimens makes it difficult, if not impossible, to pool mean values from independent repeats in a productive way. Rather, the ratio of experimental and control values within individual experiments or specimens should be our focus.

Western Blots



Fay and Gerow, 2013

This schematic blot shows the outcome of an experiment designed to test the hypothesis that loss of gene *y* activity leads to changes in the expression of protein X in *C. elegans*.

In one scenario, the three blots (A-C) could represent independent biological repeats with lanes 1-3 serving as technical (e.g., loading) repeats.

In another scenario, the three blots could serve as technical repeats with lanes 1-3 representing independent biological repeats.

It seems clear that X and Y are correlated. Does statistics agree?

Western Blots

Only 1 out of 3 biological reps

t-Test 1 (Blot A)	
wild type	*mut y*
150 200 100	400 600 500
$\bar{x}$ = 150	$\bar{x}$ = 500
SD = 50.0	SD = 100
SEM = 28.9	SEM = 57.2

P = 0.013

t-Test 2 (Blot B)	
wild type	*mut y*
70 50 80	240 150 290
$\bar{x}$ = 66.7	$\bar{x}$ = 227
SD = 15.7	SD = 71.0
SEM = 8.82	SEM = 41.0

P = 0.054

t-Test 3 (Blot C)	
wild type	*mut y*
300 220 330	690 980 440
$\bar{x}$ = 283	$\bar{x}$ = 703
SD = 56.9	SD = 270
SEM = 32.8	SEM = 156

P = 0.11

t-Test 4 (Lanes 1)	
wild type	*mut y*
150 70 300	400 240 690
$\bar{x}$ = 173	$\bar{x}$ = 443
SD = 116	SD = 228
SEM = 67.4	SEM = 132

P = 0.17

t-Test 5 (Lanes 2)	
wild type	*mut y*
200 50 220	600 150 980
$\bar{x}$ = 157	$\bar{x}$ = 577
SD = 92.9	SD = 416
SEM = 54.6	SEM = 240

P = 0.22

t-Test 6 (Lanes 3)	
wild type	*mut y*
100 80 330	500 290 440
$\bar{x}$ = 170	$\bar{x}$ = 410
SD = 138	SD = 108
SEM = 80.2	SEM = 62.5

P = 0.082

None of the technical replicates

Western Blots

t-Test 7 (Pooled data)	
wild type	*mut y*
150 70 300	400 240 690
70 50 220	600 150 980
300 80 330	500 290 440
$\bar{x}$ = 167	$\bar{x}$ = 477
SD = 102	SD = 255
SEM = 34.1	SEM = 84.9

P = 0.0065

CI calculations for ratios			
	Ratio (*mut y*/wt)	95% CI	99% CI
Blot 1	3.33	1.87–4.80	1.41–5.26
Blot 2	3.41	1.91–4.90	1.44–5.37
Blot 3	2.48	1.27–3.70	0.882–4.09
Lanes 1	2.56	0.0990–5.02	ND
Lanes 2	3.68	-0.230–7.58	ND
Lanes 3	2.41	0.0682–4.76	ND
All data	2.85	1.34–4.37	0.863–4.85

- Although pooling seems to give significant results, it is not exactly logical. Why?
- All the biological reps now give significant results.
- Why do the technical reps still give non-significant results?

A One-sample t-test of Ratios

Because the null hypothesis is that there is no difference in the expression of protein X between wild-type and *mut y* backgrounds, we would use an expected ratio of 1 for the test.

Thus, the P-value will tell us the probability of obtaining a ratio of 3.07 if the expected ratio is really one. Using the above data points, we do in fact obtain P = 0.02, which would pass our significance cutoff. In fact, this is a perfectly reasonable use of the t-test, even though the test is now being carried out on ratios rather than the unprocessed data.

Note, however, that changing the numbers only slightly to 3.33, 4.51, and 2.48, we would get a mean of 3.44 but with a corresponding *P*-value of 0.054. This again points out the problem with *t*-tests when one has very small sample sizes and moderate variation within samples.

ROM and MOR

Dealing with ratios can produce some rather unexpected behaviours.

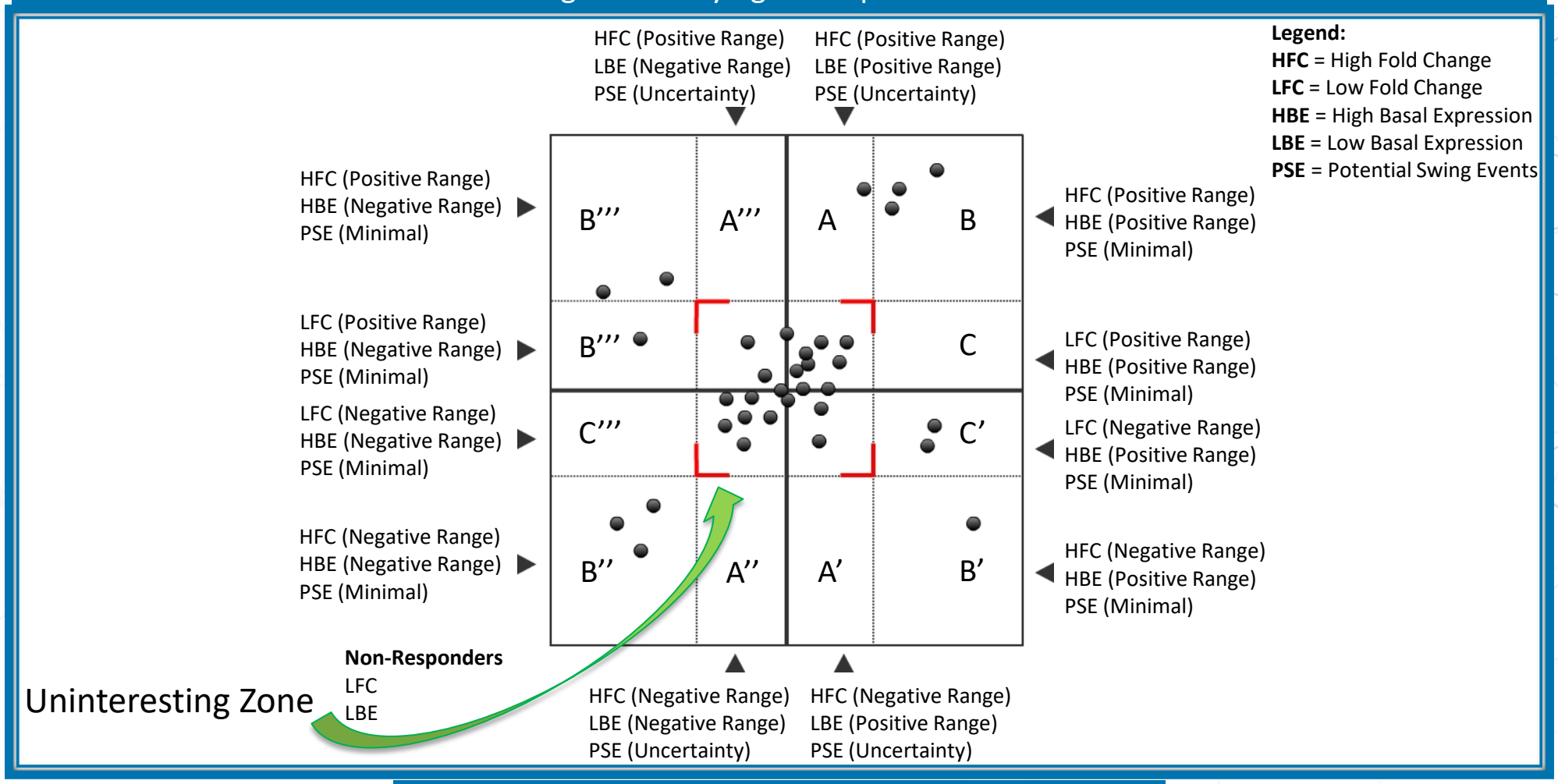
A tinker-toy illustration for increases in house prices in TinyTown (which has only two households).

	Before	After	Relative Increase
	$100,000	$400,000	4
	$300,000	$600,000	2
Means	$200,000	$500,000	MoR ↓
	RoM →	2.5 500k/200k	3 (4+2)/2

MoR tells us about the average effect on individuals, whereas RoM conveys the overall effect on the population as a whole. In the case of the western blot data, 3.07 (i.e., the MoR) is clearly the better indicator, especially given the stated issues with combining data from different blots. It is critical to be aware of the difference between RoM and MoR calculations and to report the statistic that is most relevant to your question of interest.

The Dodeca (12) panels

Dodeca Panel Guide: The panels each are correlated with meaning about the data in those regions. Here is a guide clarifying the implications.



(Non-examinable. But can try and play on own.)

NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

Summary

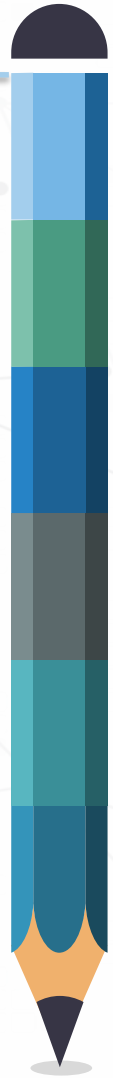
BS3033 Data Science for Biologists

Dr Wilson Goh
School of Biological Sciences

What have we seen here?

1. Do not underestimate the importance of other metrics beyond the mean and the median.

2. Non-averaging statistics convey more information on correlations than measures of centrality.



Avoiding Wrong Conclusions, Getting Deeper Insights

Check for sampling bias

- Are the distributions of the feature of interest in the two samples same as that in the two populations?

Check for exceptions

- Are there large subpopulations for which the test outcome is opposite?
- Are there large subpopulations for which the test outcome becomes much more significant?

Check for validity of the null distribution

- Can you derive it from the null hypothesis?

References

Goh and Wong. Why breast cancer signatures are no better than random signatures explained. Drug Discovery Today, 2018.

Goh and Wong. Turning straw into gold: building robustness into gene signature inference. See readings folder.

Goh and Wong. Dealing with confounders in -omics analysis. Trends in Biotechnology, 36(5):488--498, May 2018.

Goh et al. Why batch effects matter in omics data, and how to avoid them. Trends in Biotechnology, 35(6):498--507, June 2017.

Venet et al. Most Random Gene Expression Signatures Are Significantly Associated with Breast Cancer Outcome, PLOS Comp Bio, 1002240, Oct 2011.

Caniusius et al. A novel independence test for somatic alterations in cancer shows that biology drives mutual exclusivity but chance explains most co-occurrence. Genome Bio, 17:261, Nov 2016.