The Two Most Critical Preliminary Data Analysis Tasks:

1) Strong Establishment of the Sample Size, and
2) Distributional Understanding of All Variables

Michael J. Schell
Moffitt Cancer Center
June 7, 2023

Strong Establishment of the Sample Size

Statistics is a branch of scientific analysis methods that uses inference

Data are obtained from the sample, which is a subset of the population for which the inference is to be derived

Once this subset is defined properly, the resulting **sample size** is the critical first quantitative result. It is most valuable to etch this number in one's memory so that all future analyses are obtained with reference to it, knowing that sometimes the analyzed sample size may be smaller due to missing data or further exclusion of some data.

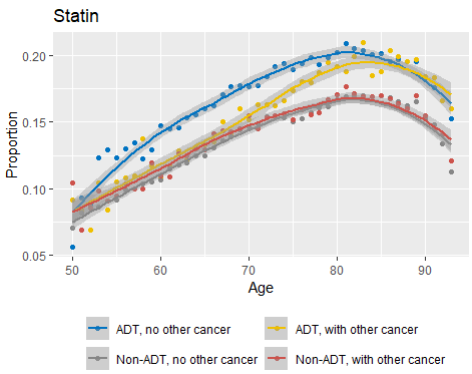
Strong Establishment of the Sample Size – 2

Consequently, a critical double check for all analyses conducted should be with regard to this number. In particular, it should never be larger than the sample size. However, this can too easily happen if one does not check the number. This common error occurs when one does not apply all the exclusions from some raw data set.

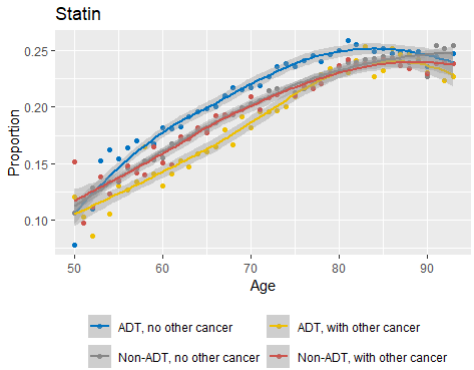
This is why analysts should:

Always check the sample size for a given analysis compared to **the sample size**.

Incorrect Plot on 3.6M People, with One Exclusion Left Out of the Program



Plot for the Established Sample Size, 2.5M



Distributional Understanding of All Variables

The distributional properties of every variable that is used in an analysis should be examined. This should be viewed as a critical, not an optional step, in a high-quality analysis.

There are two main kinds of examination:

1) Visual
2) Quantitative

Of the two, a ***quantitative examination is the critical one***, especially if the analyst is processing hundreds or thousands of variables

Visualization

Three common types of visualization for distributions are:

Histograms, boxplots, and violin plots.

Often scaled automatically by the software.

Sometimes too much space is consumed by 1+ outlier values. Consequently, interpretability is impeded, rendering the default visualization tool insufficient.

Mean and SD, or 5-Number Summary?

Of course, one can (and should) choose to do both. However, if given the draconian choice, the 5-number summary is easily the best.

The mean and SD values are NOT robust

Why?

The mean and SD can be radically altered by even a single outlier.

In the language of **robust statistics**, both have a **breakdown point** of zero, which means that a single infinite data point takes both the mean and the SD off to infinity as well.

The 5-Number Summary

The 5-number summary is robust

The median has the maximum possible breakdown point – at 50%.

The **minimum** and **maximum** can be mainly non-informative.

For many adult illnesses, the extremes may often be 21- ~95.

They could be VERY informative if the values are so implausible that they suggest a **gross error***

***which we define as an outlier value that is thought to be an error, not a valid, but extreme value.

The Inner Core of the 5-Number Summary

While the extremes MIGHT be very useful,

the central 3 values are ALWAYS valuable.

The Q1, median, and Q3 values represent the 25^{th}, 50^{th}, and 75^{th} percentiles of the distribution.

Thus, they map out the middle half of the distribution – by construction.

This **inner core of the distribution** provides a robust basis for the identification of outliers, whereas the SD may fail.

The Four Moments … and The Two Ratios

While mathematically speaking, moments don't have to exist by being infinite, practically speaking, I've never encountered such a distribution. If all moments exist, they essentially define a distribution. Numbered from 1 on, they become increasingly less important. Thus, 4 moments should suffice to obtain a strong sense of the distribution under examination. The first four moments are: **mean, SD, skewness**, **and kurtosis**. There are two very important ratios of these moments as well. The first is the **coefficient of variation** (CV), which is 100 times the ratio of the SD to the mean. The second is the **distributional slope**, which is the ratio of the square of the skewness to the kurtosis.

The 6 Parametric Distributional Tools

These six numbers, **the six parametric distributional tools**, give us very powerful quantitative tools, if we know how to use them.

Obtained by **arithmetical calculation** quickly by computer algorithms

Thus **can be scaled up** when working with hundreds or thousands of variables.

In this way, this 6-parameter approach succeeds where visualization can easily fail, and where the 5-number summary is much less deployable.

Coefficient of Variation

The coefficient of variation describes the relationship of the SD to the mean.

For non-negative normally-distributed data, the CV should be ≤ 50, because that would correspond to having 0 be 2 SDs below the mean.

Highly right-skewed data: CV values ≥ 50

(e.g. CV=100 for the exponential)

For biological variables, lower CV implies greater physiological control.

Example of Physiological Control

Variable	Range	Mean	SD	CV
Hemoglobin (Males)	13.2-16.6	14.9	0.85	5.7
Hemoglobin (Females)	11.6-15	13.3	0.85	6.4
Hematocrit (Males)	38.3-48.6%	43.4%	2.58%	5.9
Hematocrit (Females)	35.5-44.9%	40.2%	2.35%	5.8
Platelet count (Males)	135-317	226	45.5	20.1
Platelet count (Females)	157-371	264	53.5	20.3

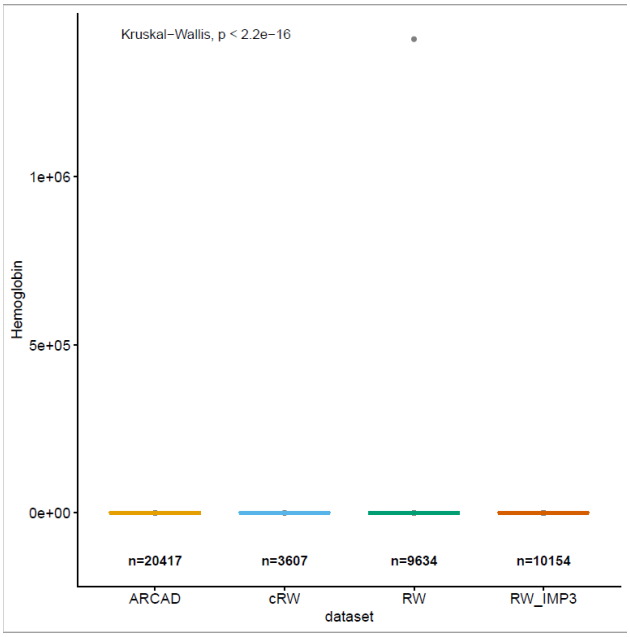
SD estimated by dividing the range by 4

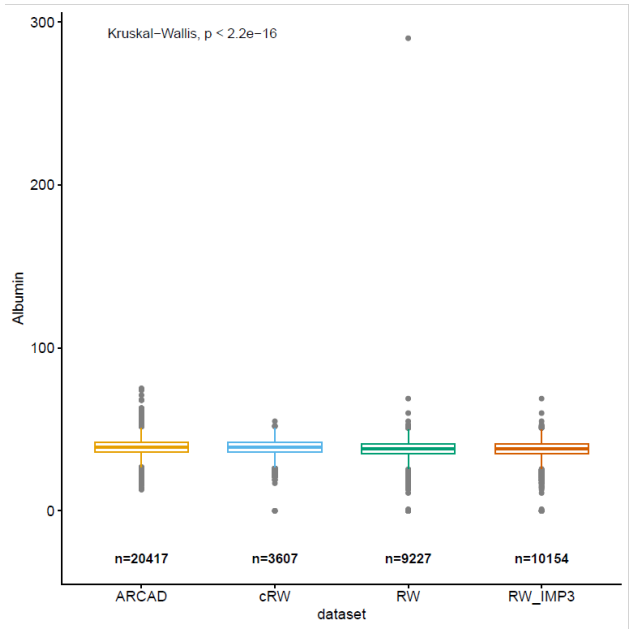
https://www.mayoclinic.org/tests-procedures/complete-blood-count/about/pac-20384919

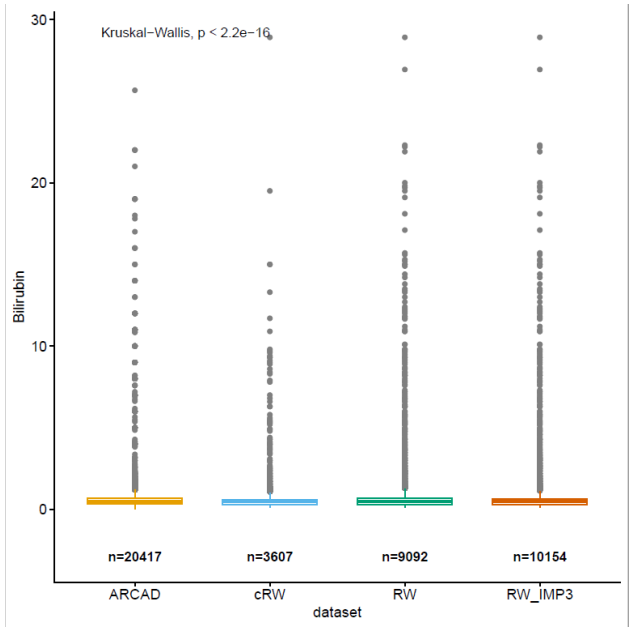
Example of Physiological Control – 2

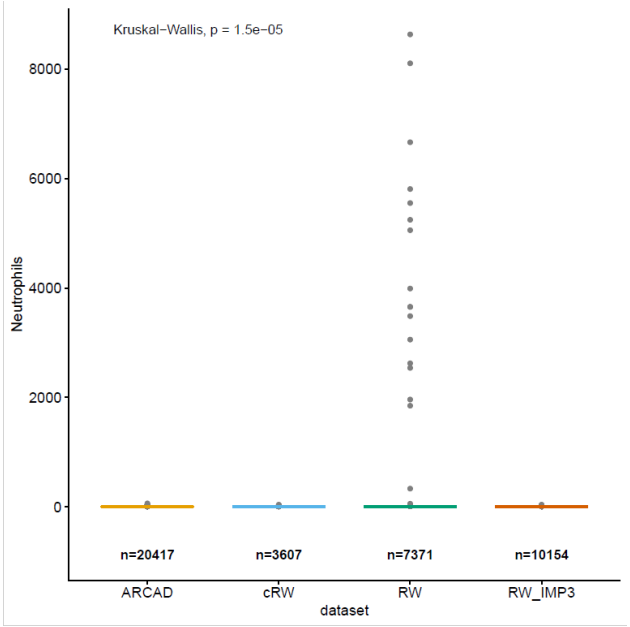
Variable	Range	Mean	SD	CV
ALT	7-55	31	12	38.7
ALP	40-129	84.5	22.2	26.3
Albumin	3.5-5.0	4.25	.375	8.8
Total protein	6.3-7.9	7.1	.4	5.6
Bilirubin	0.1-1.2	0.65	.275	42.3
PT	9.4-12.5	11.0	.775	7.0

https://www.mayoclinic.org/tests-procedures/liver-function-tests/about/pac-20394595#:~:text=6.3%20to%207.9%20g%2FdL




Kruskal–Wallis, p < 2.2e−16
Albumin
300
200
100
0
n=20417
n=3607
n=9227
n=10154
ARCAD
cRW
RW
RW_IMP3
dataset


Kruskal–Wallis, p < 2.2e−16
Bilirubin
30
20
10
0
n=20417
n=3607
n=9092
n=10154
ARCAD
cRW
RW
RW_IMP3
dataset



Distributional slope

Distribution	Skewness	Kurtosis	Ex. Kurtosis	Dist. slope
Uniform	0	1.8	-1.2	0
Normal	0	3	0	0
Poisson (λ=1)	1	3+1	1	1.00
Poisson (λ=9)	1/3	3+1/9	1/9	1.00
Exponential	2	9	6	.667
Gamma (df=9)	2/3	3+ 6/9	6/9	.667
Lognormal (σ=1)	6.18	113.9	110.9	.345
Lognormal (σ=0.5)	1.75	8.90	5.90	.519

Average Data from 12 Patients

Which of the 6 is Best For Distributional Understanding?

Day	Mean	SD	Skew	Kurt	Slope	CV
Pre-exp	2.76	9.79	16	423	.61	355
Post-exp	1.97	9.69	111	20298	.61	492
007	1.12	0.76	52	6276	.43	68
014	1.44	6.37	81	8723	.75	442
028	1.38	3.14	57	5578	.58	228
063	1.91	15.41	108	13882	.84	969
091	1.59	4.47	58	6733	.50	285
180	1.57	2.91	25	1192	.52	185
360	1.70	6.91	159	37211	.68	406

Log_{10} Data from Patient 10

Which of the 6 is Best For Distributional Understanding?

Day	Mean	SD	Skew	Kurt	Max	Slope	CV
Pre-exp	.021	.08	3.64	14.9	0.48	.89	378
Post-exp	.054	.17	4.07	23.8	1.79	.70	309
007	.032	.11	3.86	21.3	1.28	.70	332
014	.055	.17	4.40	29.2	2.54	.66	312
028	.047	.15	4.09	24.6	1.80	.68	316
063	.072	.20	3.60	18.9	2.12	.69	274
091	.122	.27	2.90	12.8	2.42	.65	226
180	.070	.19	3.33	16.3	1.79	.68	266
360	.114	.25	2.81	12.3	2.12	.65	222

QUESTIONS?