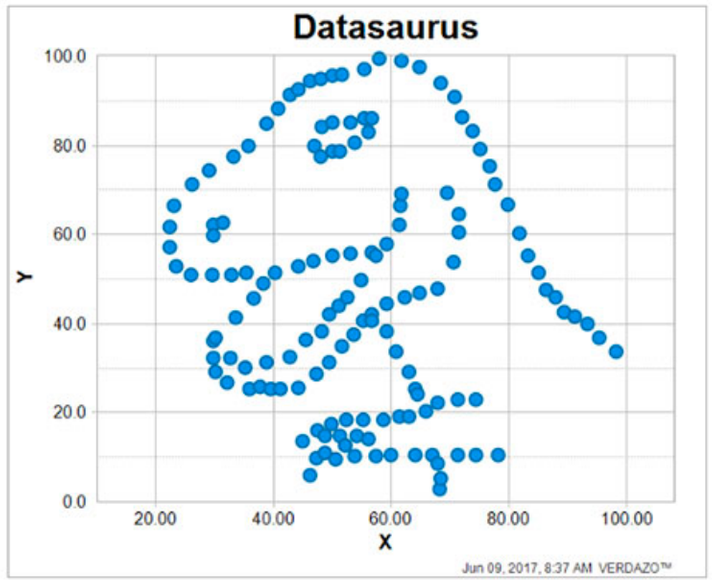
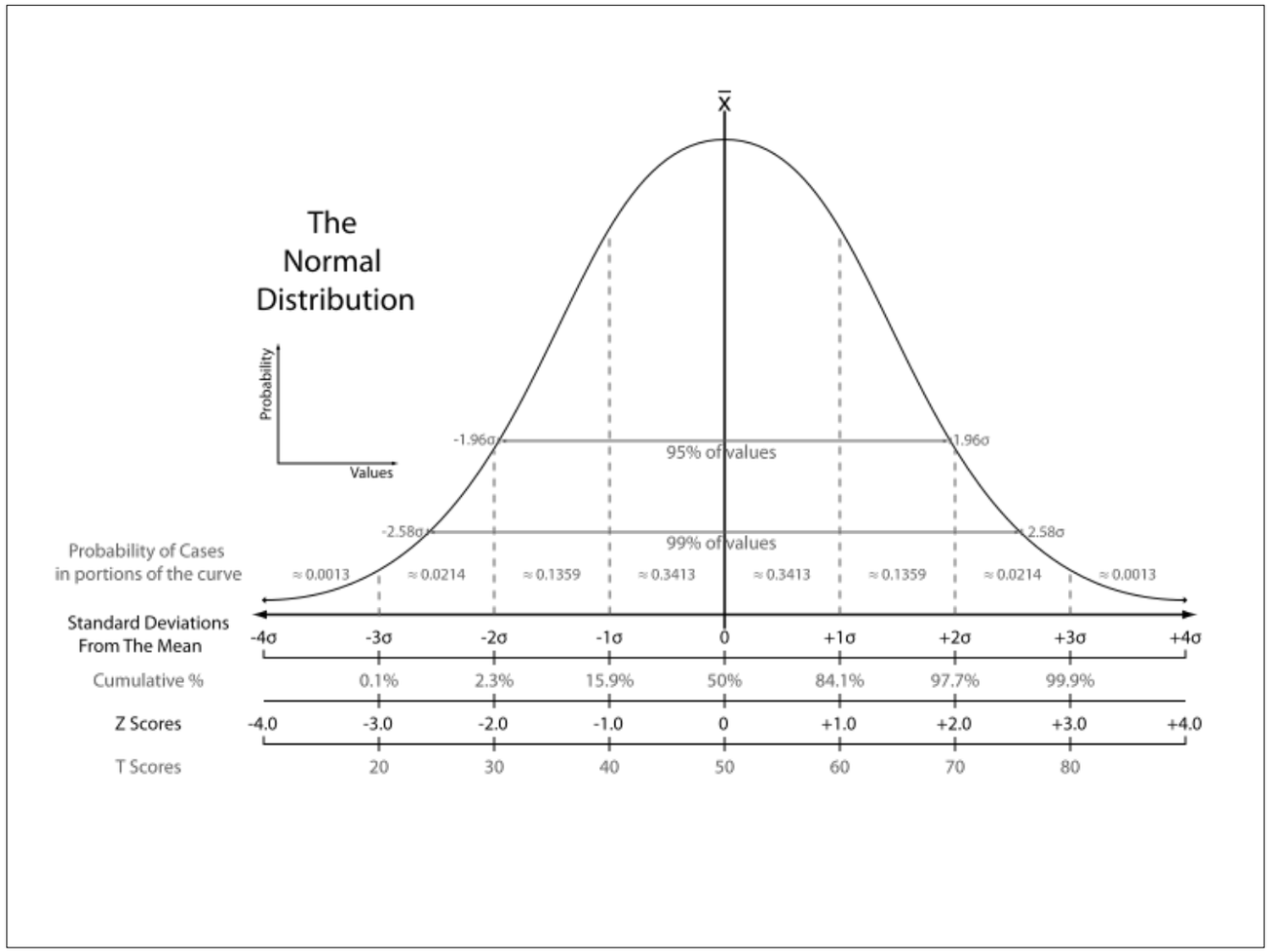
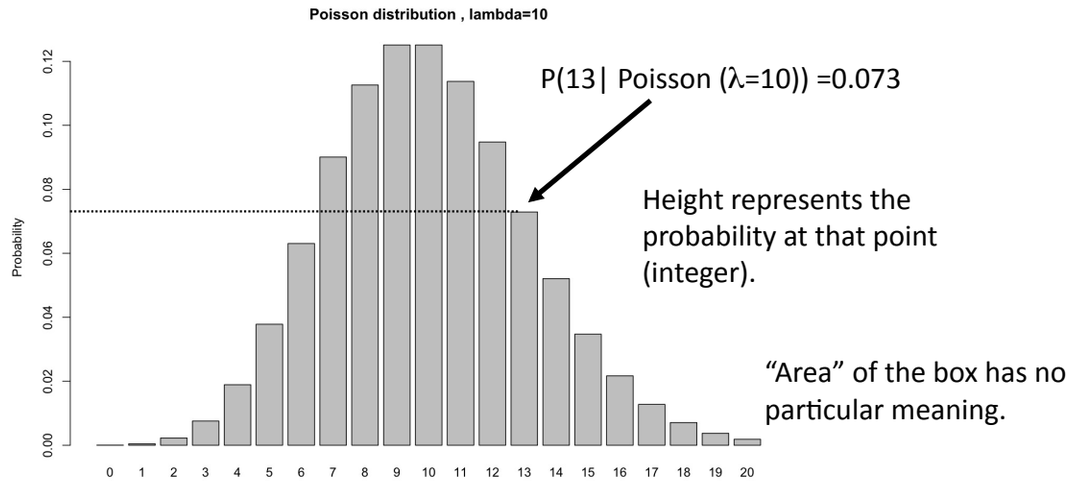


Statistics (or something)

Amanda Charbonneau


The Normal Distribution
x̄
Probability
Values
-1.96σ
1.96σ
95% of values
-2.58σ
2.58σ
99% of values
Probability of Cases in portions of the curve
≈ 0.0013
≈ 0.0214
≈ 0.1359
≈ 0.3413
≈ 0.3413
≈ 0.1359
≈ 0.0214
≈ 0.0013
Standard Deviations From The Mean
-4σ
-3σ
-2σ
-1σ
0
+1σ
+2σ
+3σ
+4σ
Cumulative %
0.1%
2.3%
15.9%
50%
84.1%
97.7%
99.9%
Z Scores
-4.0
-3.0
-2.0
-1.0
0
+1.0
+2.0
+3.0
+4.0
T Scores
20
30
40
50
60
70
80

Probability Mass function (For discrete distributions, like read counts)



P(13| Poisson (λ=10)) =0.073

Height represents the probability at that point (integer).

"Area" of the box has no particular meaning.

P(integer) ≥ 0
P(non-integers) = 0.

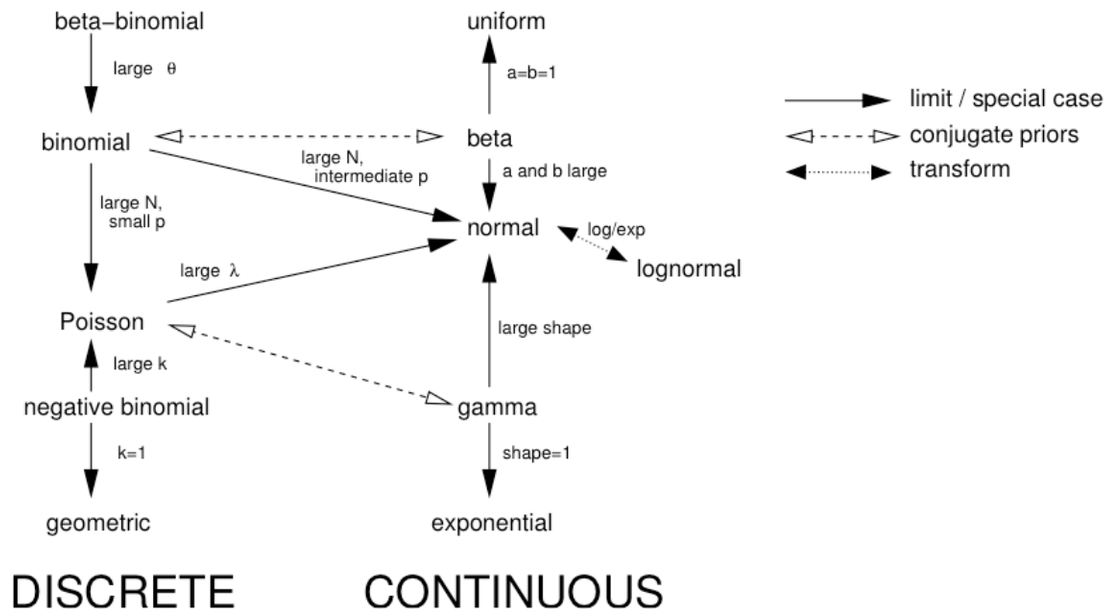


Figure 4.17 Relationships among probability distributions.

Bolker 2007 CH4 page 137

Negative binomial

$$\text{Negative Binomial Distribution} = \frac{\Gamma(k+x)}{\Gamma(k)x!}\left(\frac{k}{k+\mu}\right)^{k}\left(\frac{\mu}{k+\mu}\right)^{x}$$

Expected number of counts =μ
Over-dispersion parameter = k

For our purposes all we care about is that

$$\text{var}(x) = \mu + k\mu^2$$

“To consult the statistician after an experiment is finished is often merely to ask him(her) to conduct a post mortem examination. He(she) can perhaps say what the experiment died of.”

–Ronald Fisher

I have an idea…

I have an idea…

- 150 individuals

I have an idea…

- 150 individuals
- 50 of each treatment

I have an idea…

- 150 individuals
- 50 of each treatment
- Treatment lasts 1 week

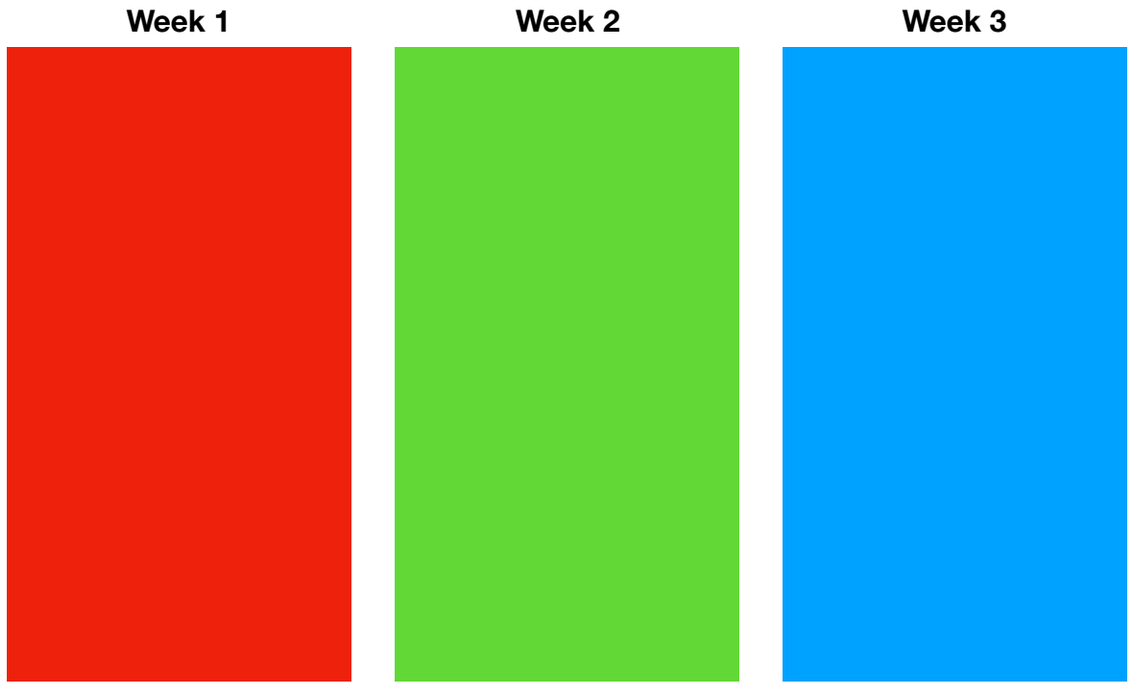
I have an idea…

- 150 individuals
- 50 of each treatment
- Treatment lasts 1 week
- We have 3 incubators/ greenhouses/tanks/cages which each hold 50 individuals

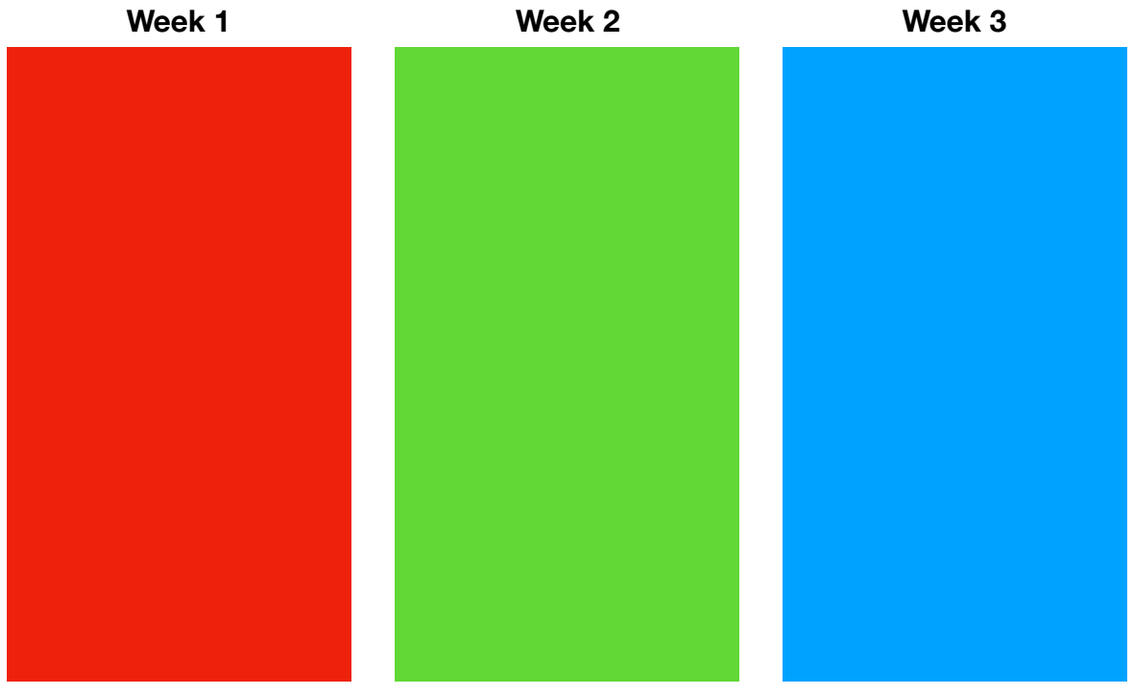
I have an idea…

- 150 individuals
- 50 of each treatment
- Treatment lasts 1 week
- We have 3 incubators/ greenhouses/tanks/cages which each hold 50 individuals
- Let’s do the blue treatment in week 1, green treatment in week 2 and red treatment in week 3

Experimental Design

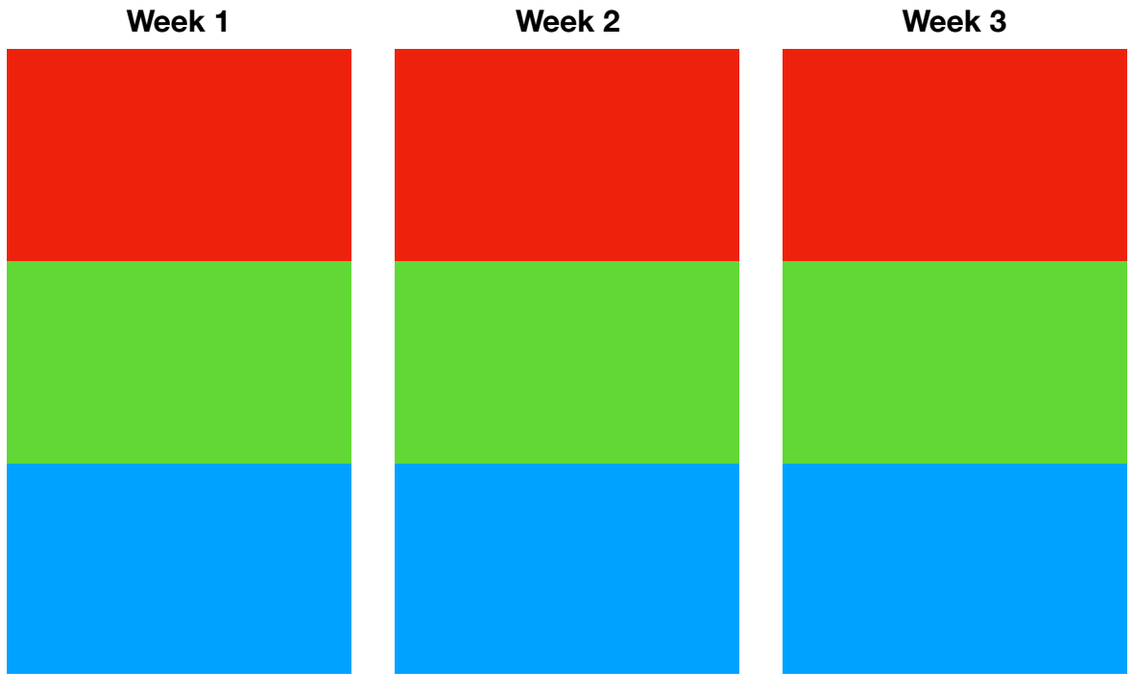


Experimental Design

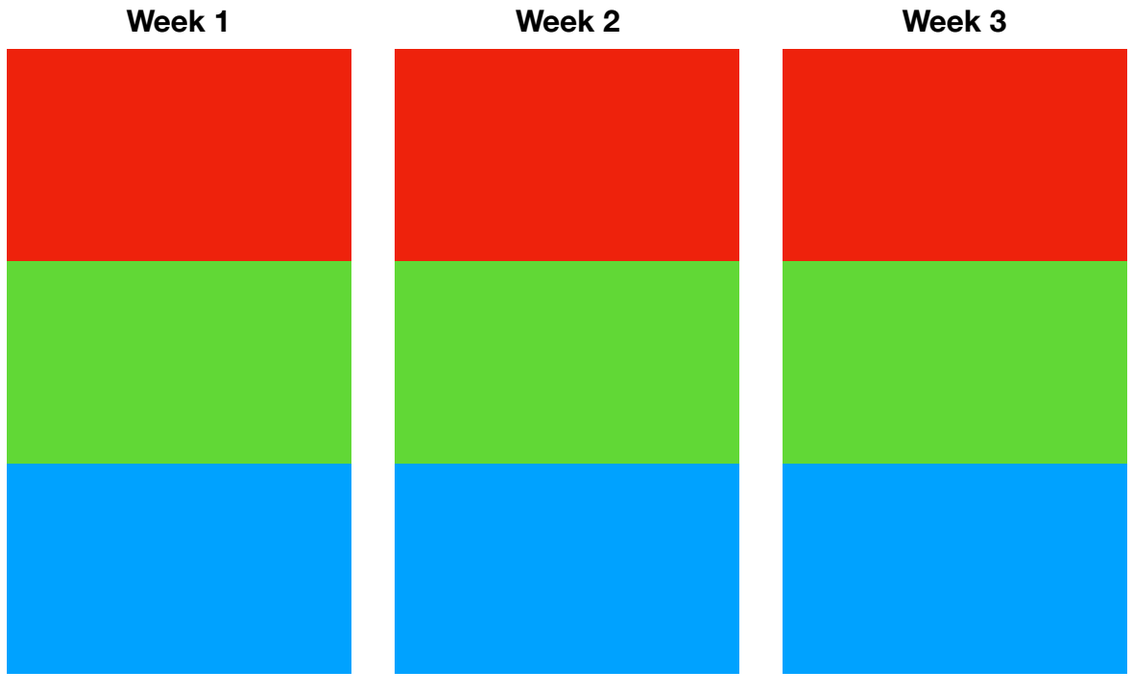


How's this?

Experimental Design

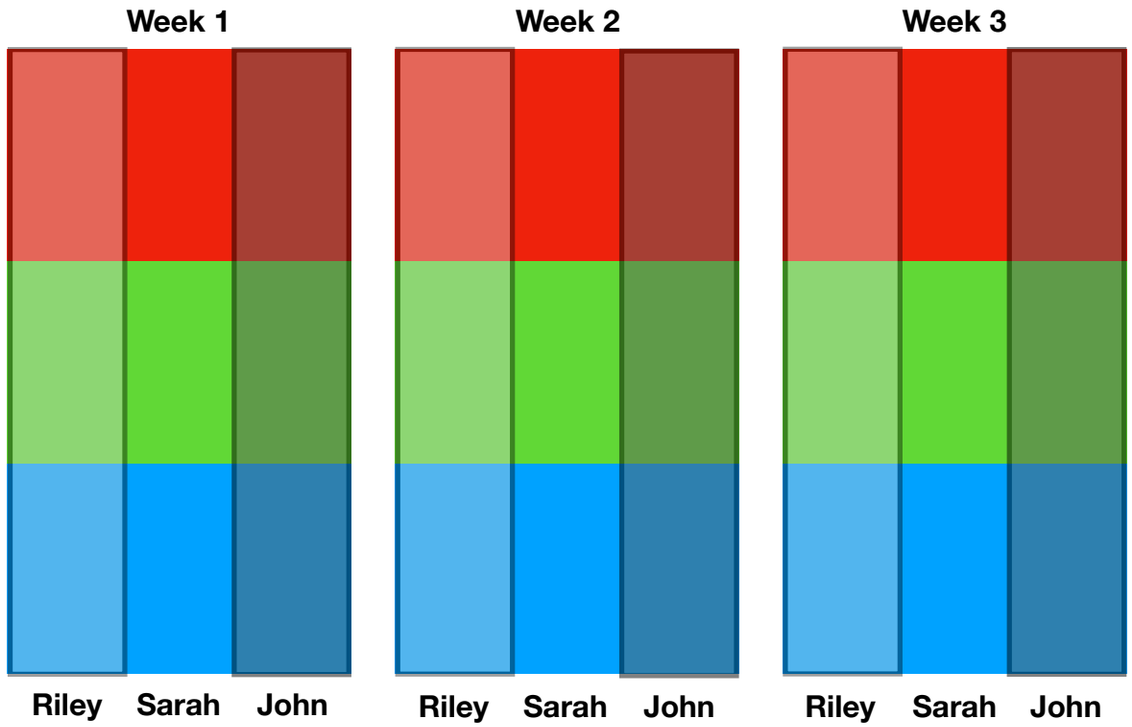


Experimental Design



You have 3 undergrads. How should they split the data collection work?

Experimental Design



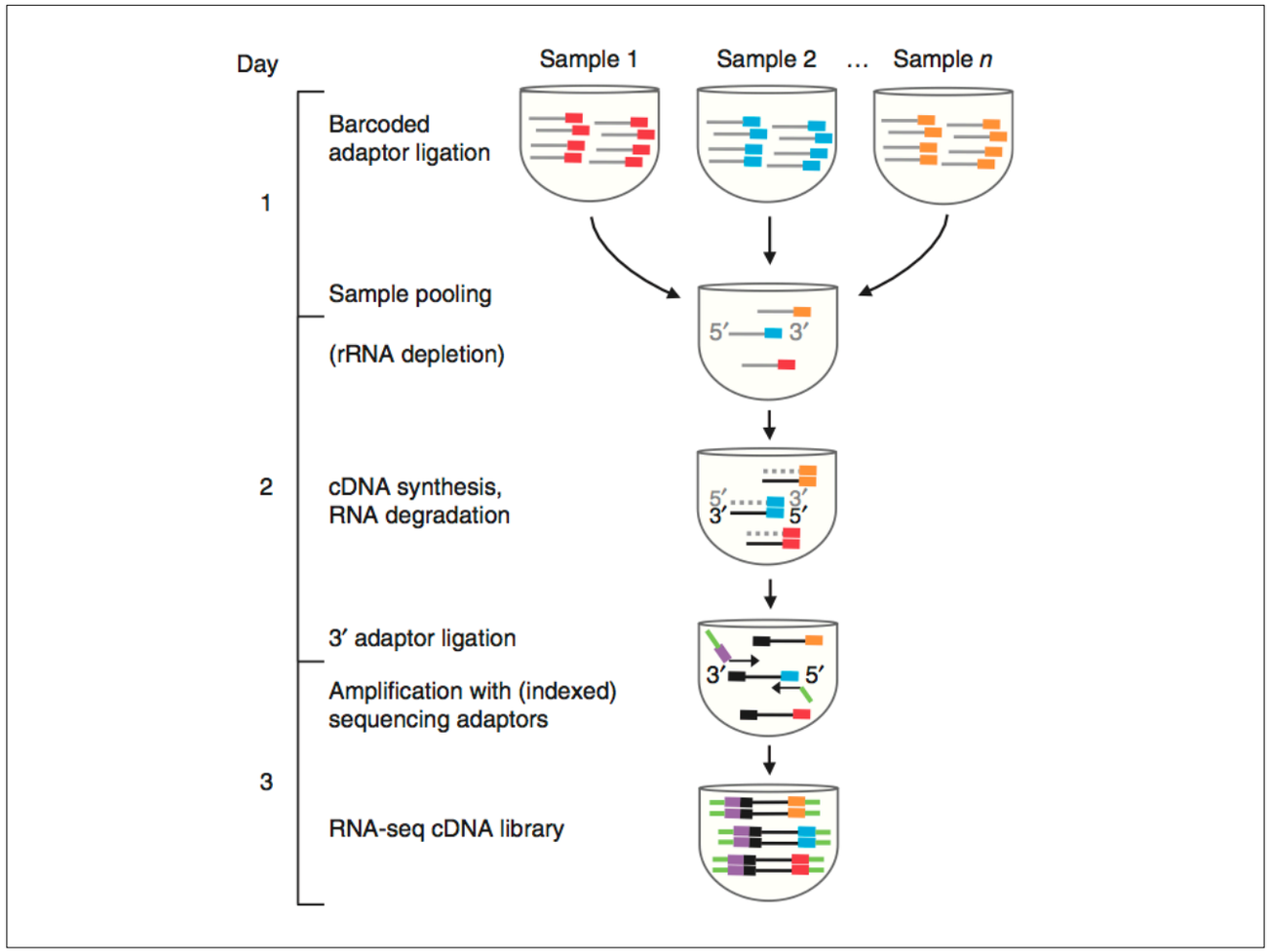
Experimental Design

Sample	Treatment	Week	Student	Measurement
1		1	Riley	92
2		1	Sarah	56
3		1	John	21
4		2	John	77
5		2	Riley	35
6		2	Sarah	26
7		3	Sarah	68
8		3	John	41
9		3	Riley	42

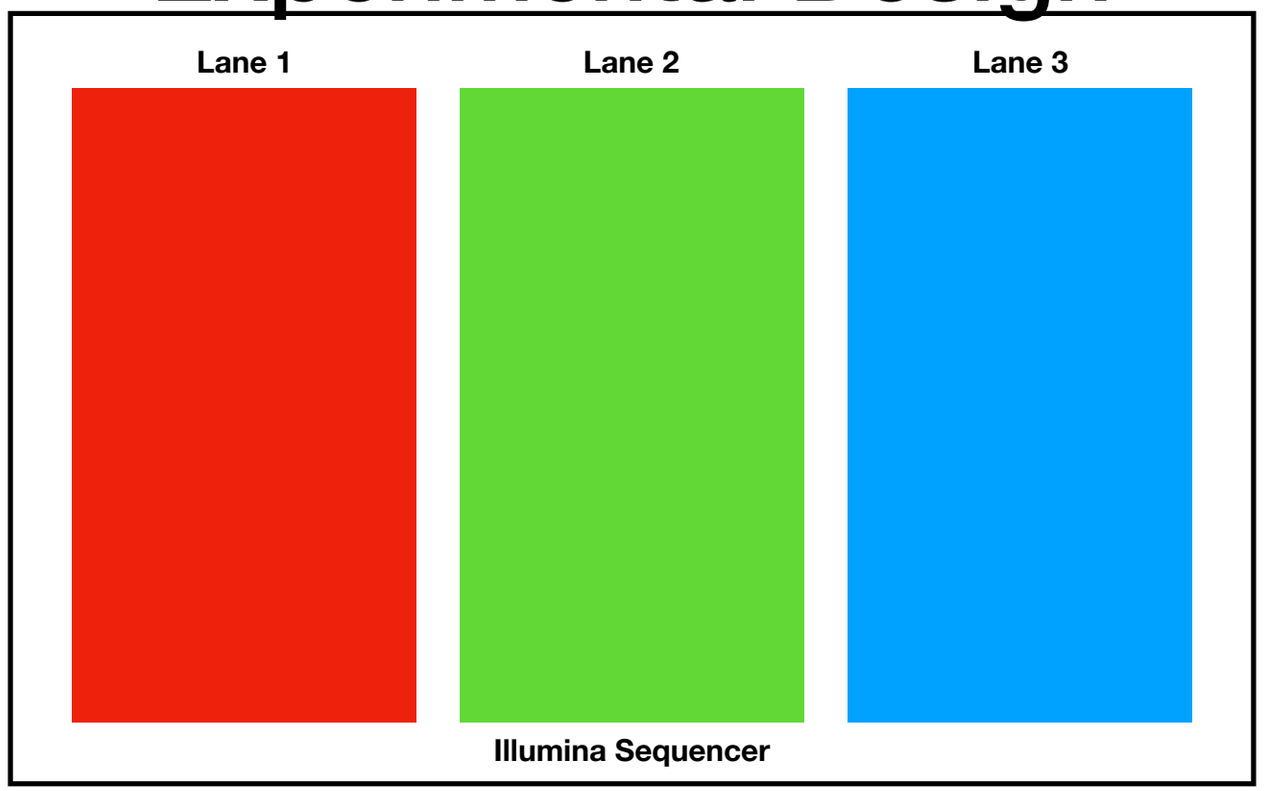
If this is going to be an RNAseq experiment, what variables should we try to account for?

- Time collected RNA
- Extraction method
- Different lanes
- Time in storage
- PCR
- When/where samples collected

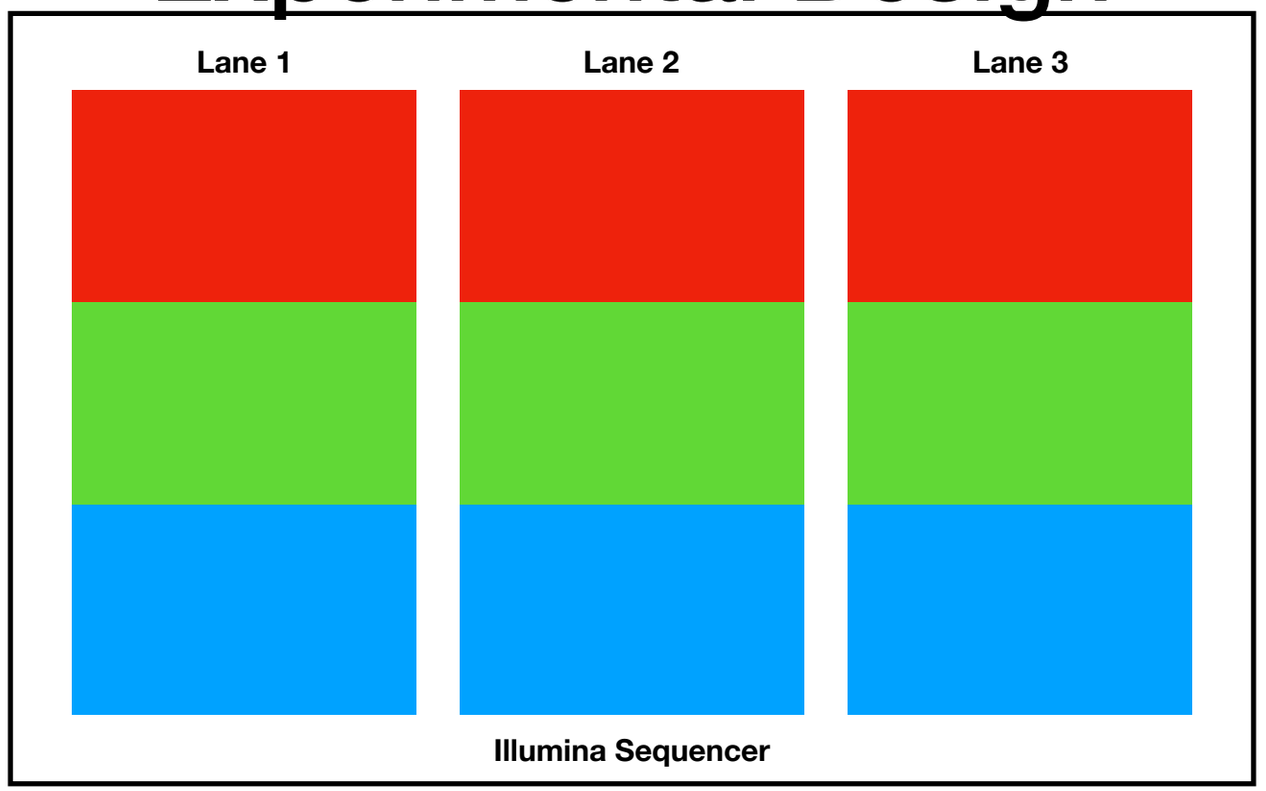
lane, barcode, prep kit, prep date, person prepping, run date,


Day
Sample 1
Sample 2
…
Sample n
1
Barcoded adaptor ligation
Sample pooling
(rRNA depletion)
2
cDNA synthesis, RNA degradation
3′ adaptor ligation
Amplification with (indexed) sequencing adaptors
3
RNA-seq cDNA library
5′
3′

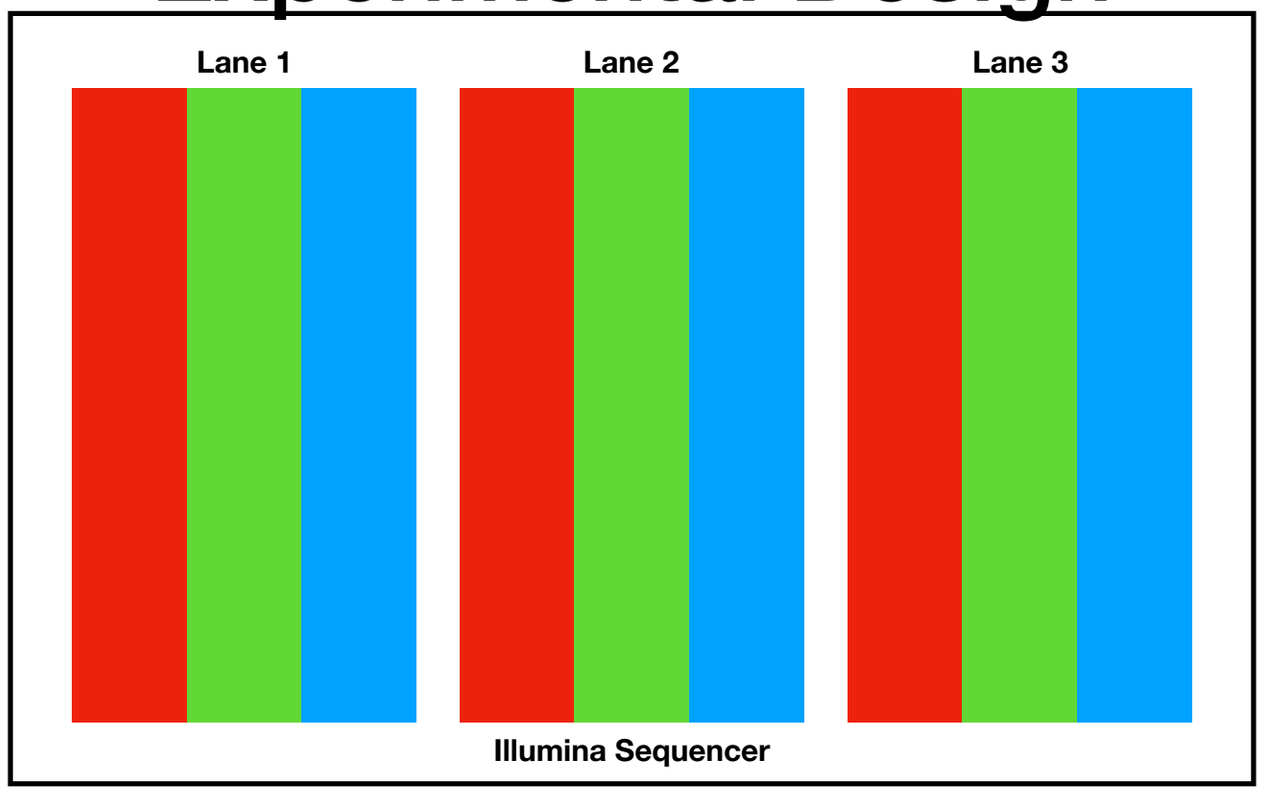
Experimental Design



Experimental Design

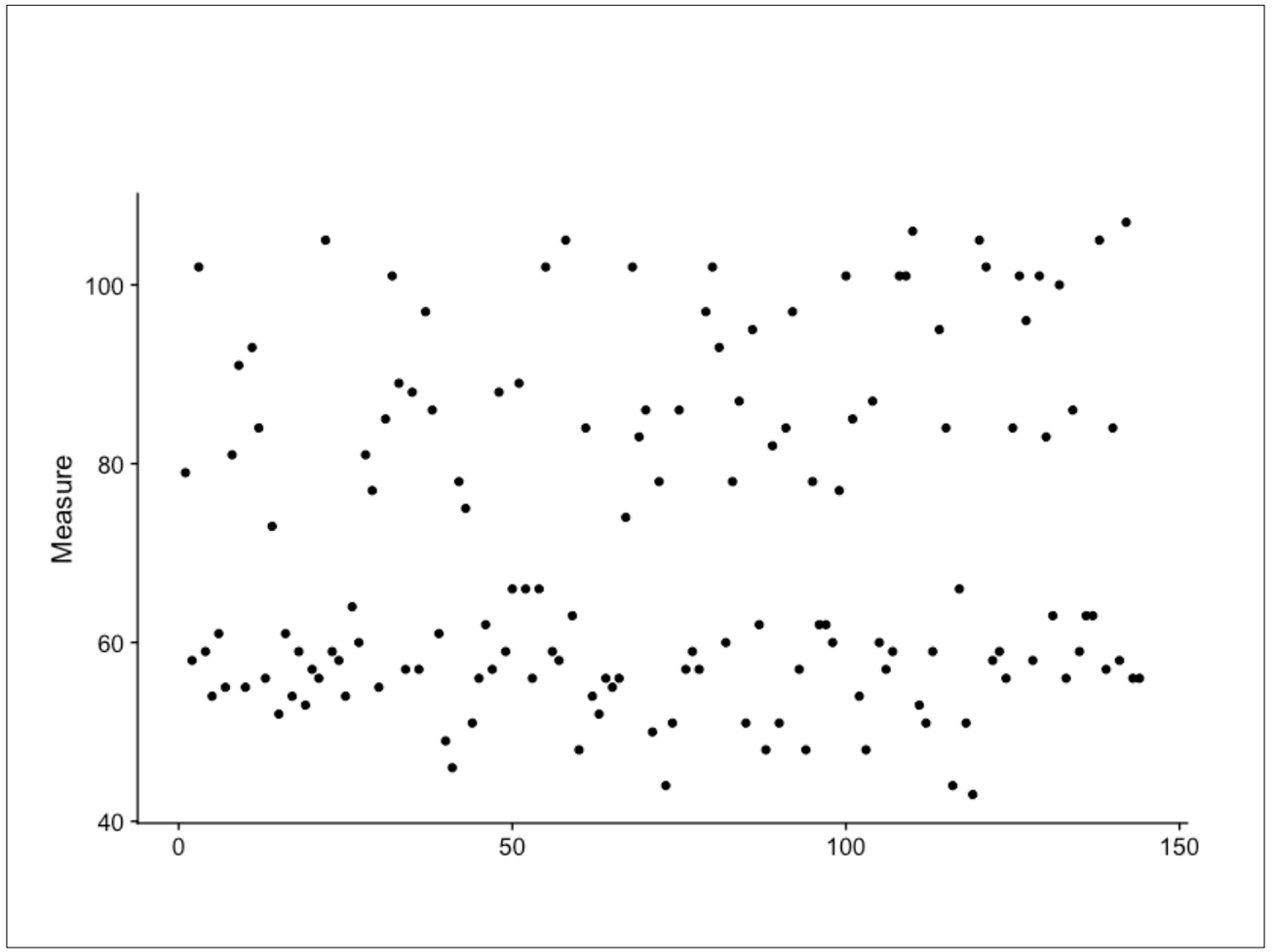


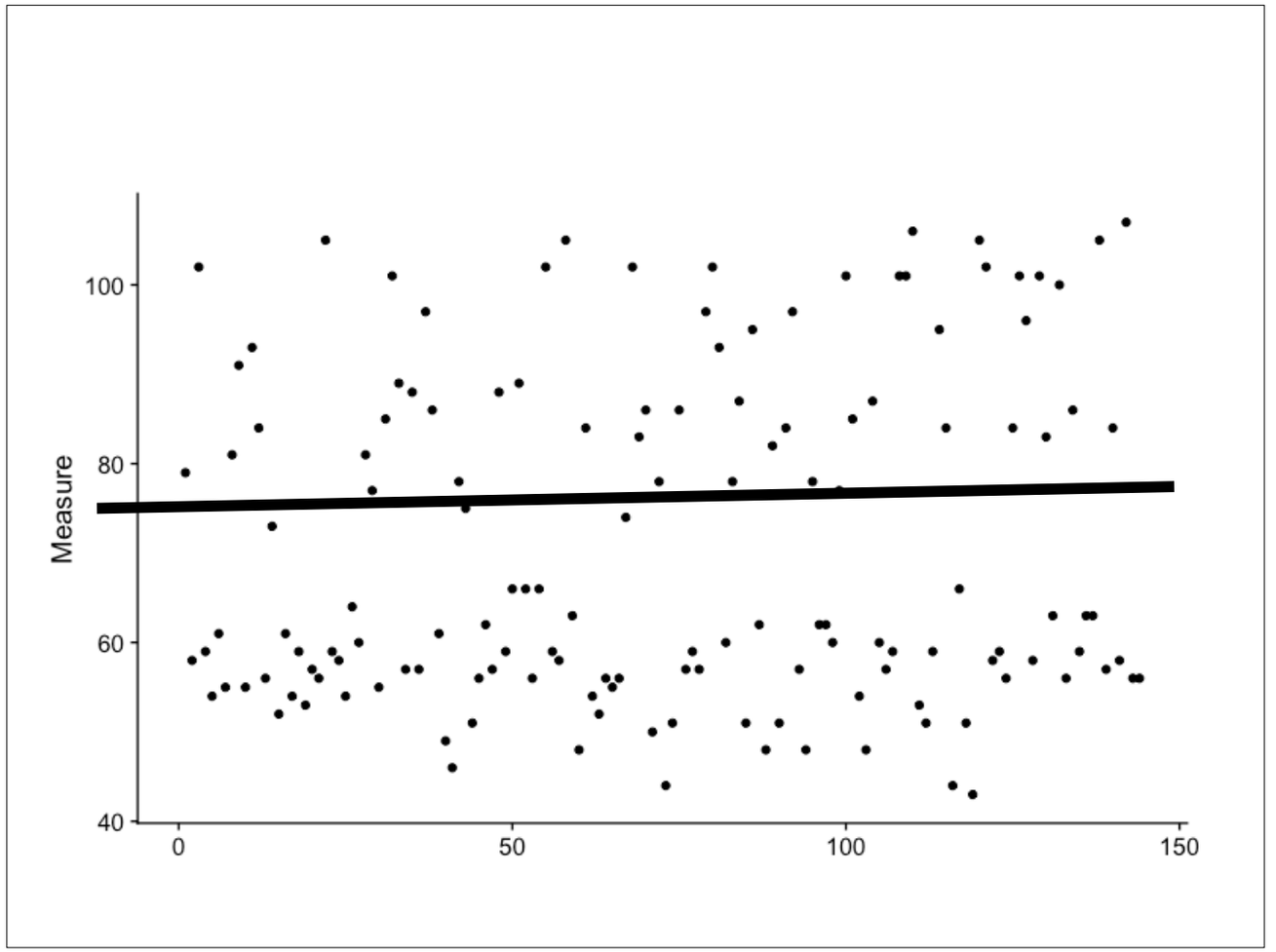
Experimental Design

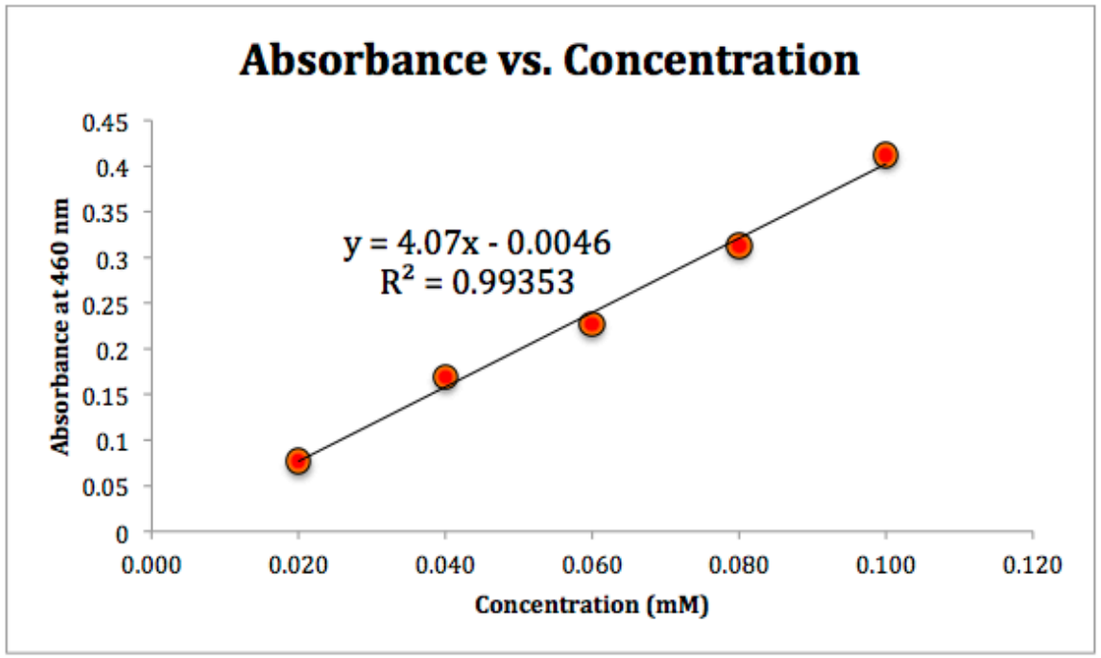


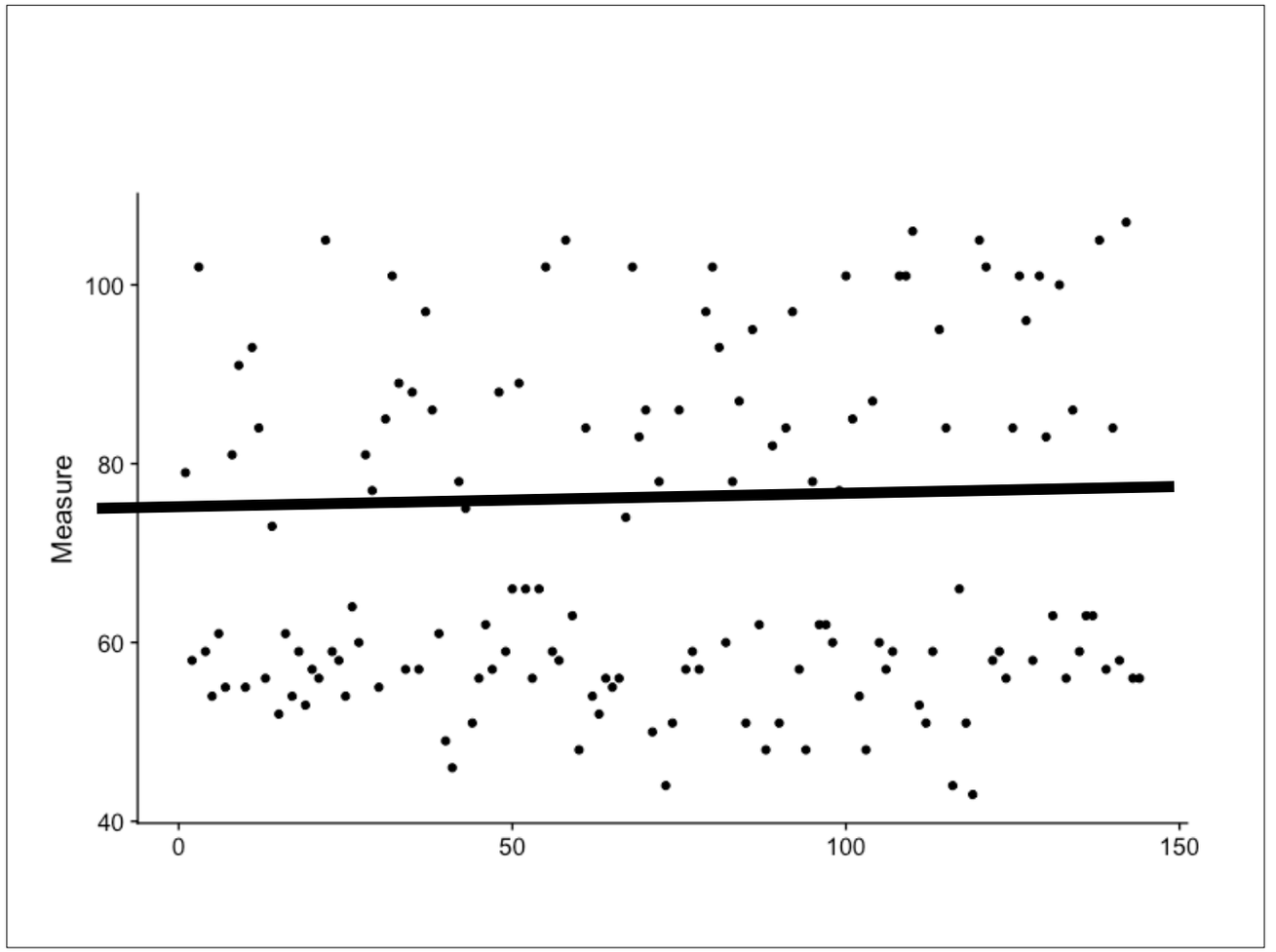
Experimental Design

Sample	Treatment	Week	Student	ReadCount	Lane	Library Prep Date
1		1	Riley	92	1	5/23/18
2		1	Sarah	56	2	6/30/18
3		1	John	21	1	5/23/18
4		2	John	77	2	6/30/18
5		2	Riley	35	1	5/23/18
6		2	Sarah	26	2	6/30/18
7		3	Sarah	68	1	5/23/18
8		3	John	41	2	6/30/18
9		3	Riley	42	1	5/23/18


Measure
100
80
60
40
0
50
100
150

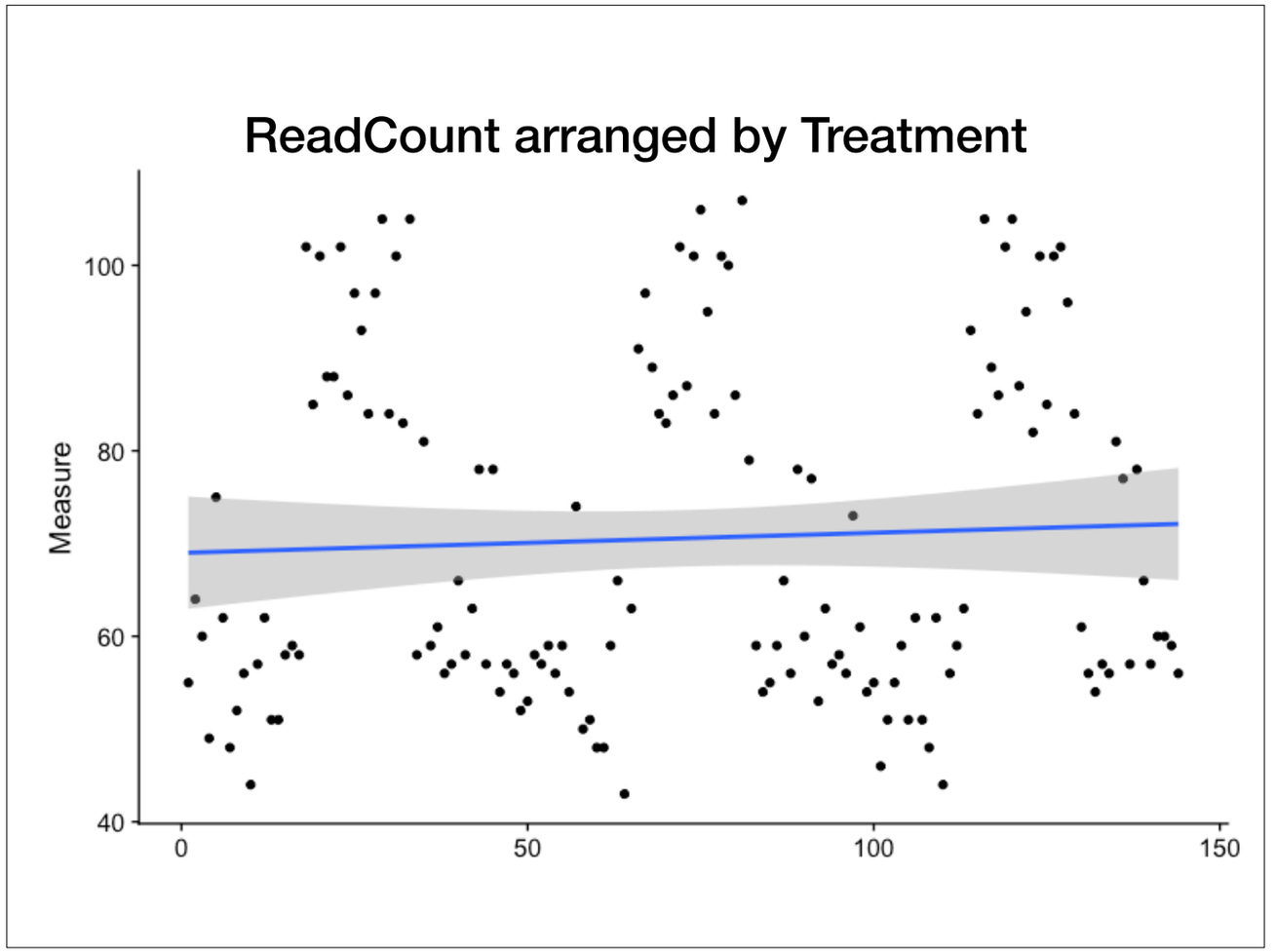

Measure
100
80
60
40
0
50
100
150

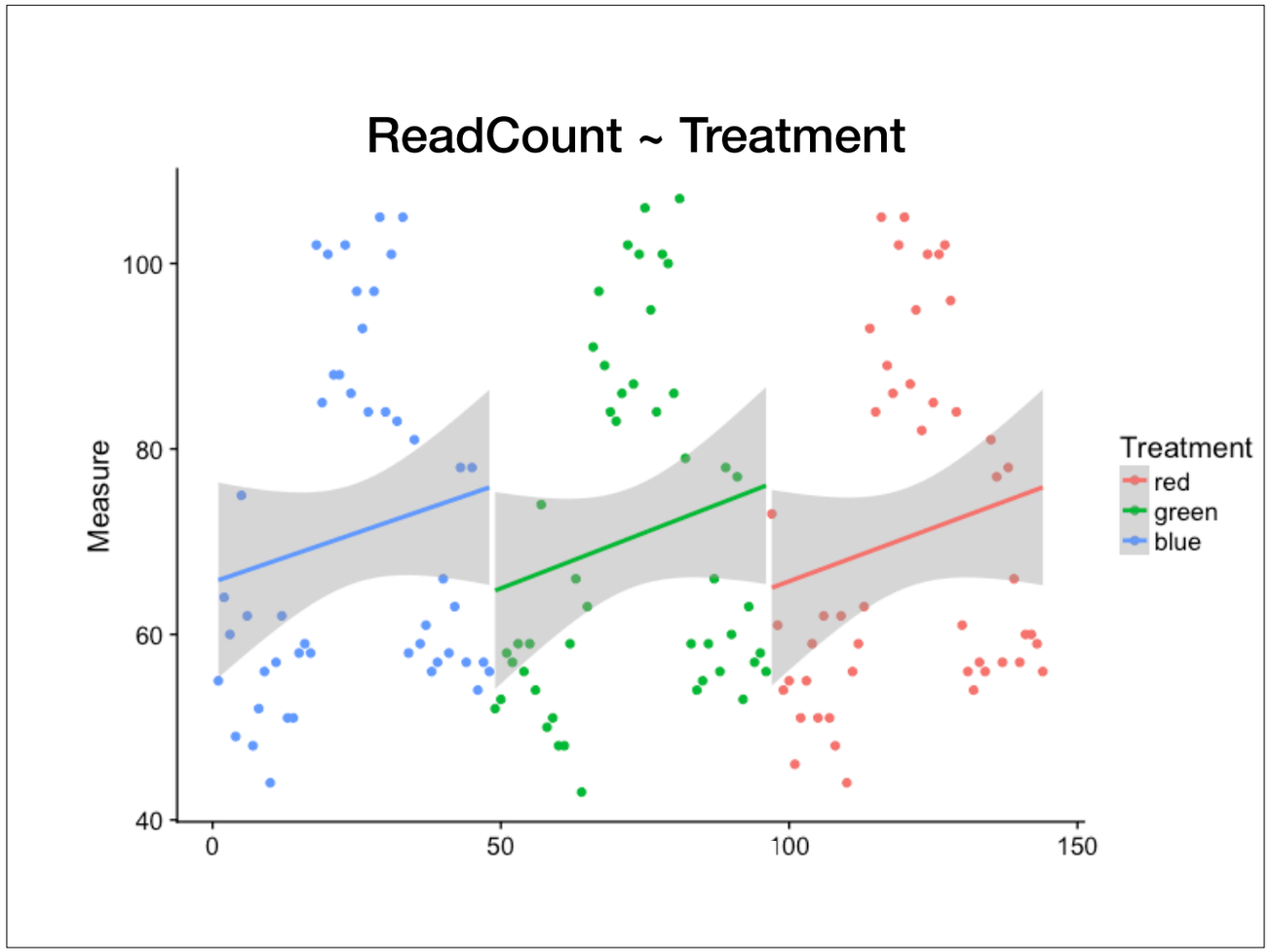

Absorbance vs. Concentration
y = 4.07x - 0.0046
R² = 0.99353
Absorbance at 460 nm
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0.000
0.020
0.040
0.060
0.080
0.100
0.120
Concentration (mM)


Measure
100
80
60
40
0
50
100
150

ReadCount ~ Treatment

Sample	Treatment	Week	Student	ReadCount	Lane	Library Prep Date
1		1	Riley	92	1	5/23/18
2		1	Sarah	56	2	6/30/18
3		1	John	21	1	5/23/18
4		2	John	77	2	6/30/18
5		2	Riley	35	1	5/23/18
6		2	Sarah	26	2	6/30/18
7		3	Sarah	68	1	5/23/18
8		3	John	41	2	6/30/18
9		3	Riley	42	1	5/23/18

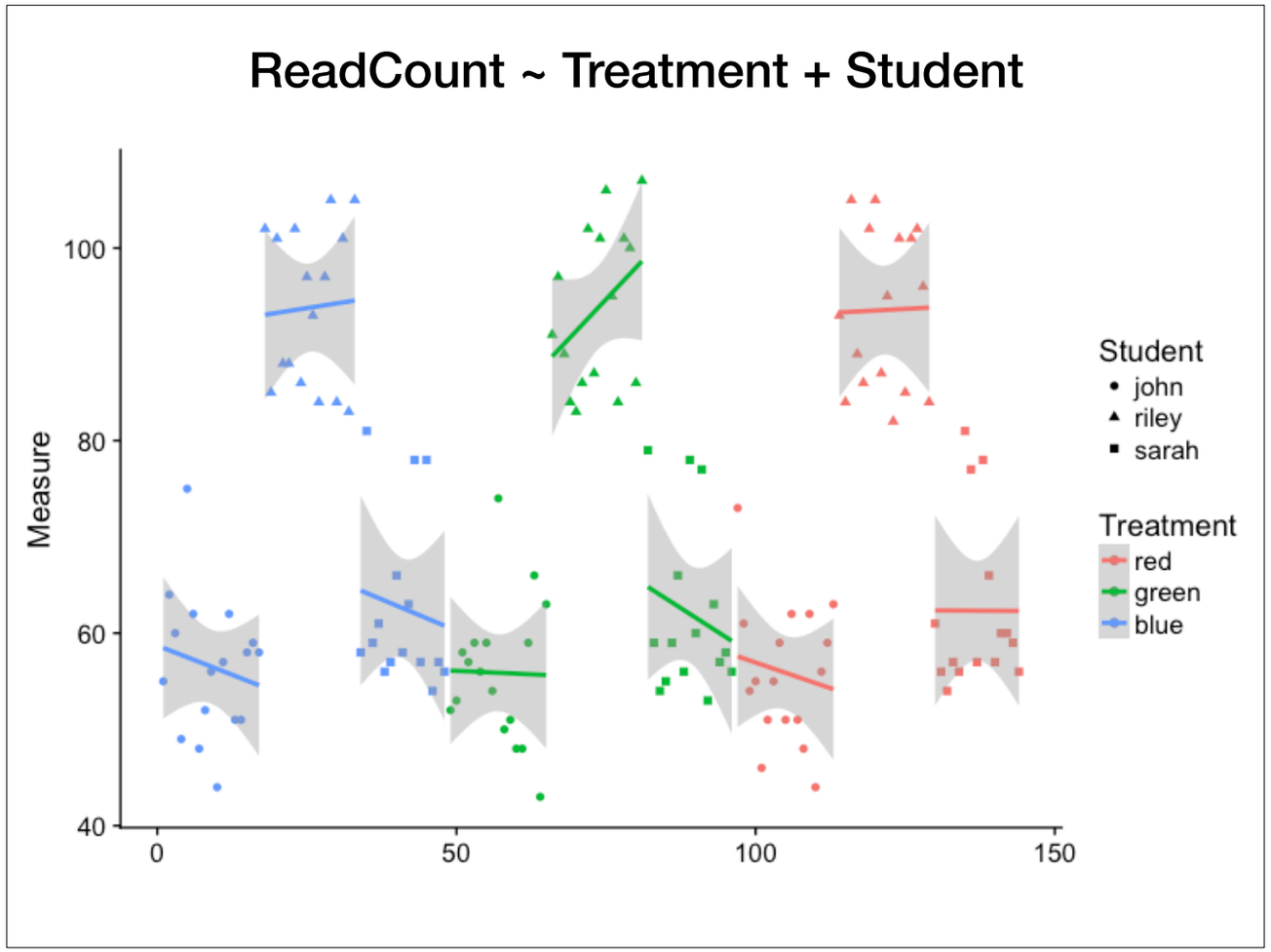

ReadCount arranged by Treatment
Measure
100
80
60
40
0
50
100
150

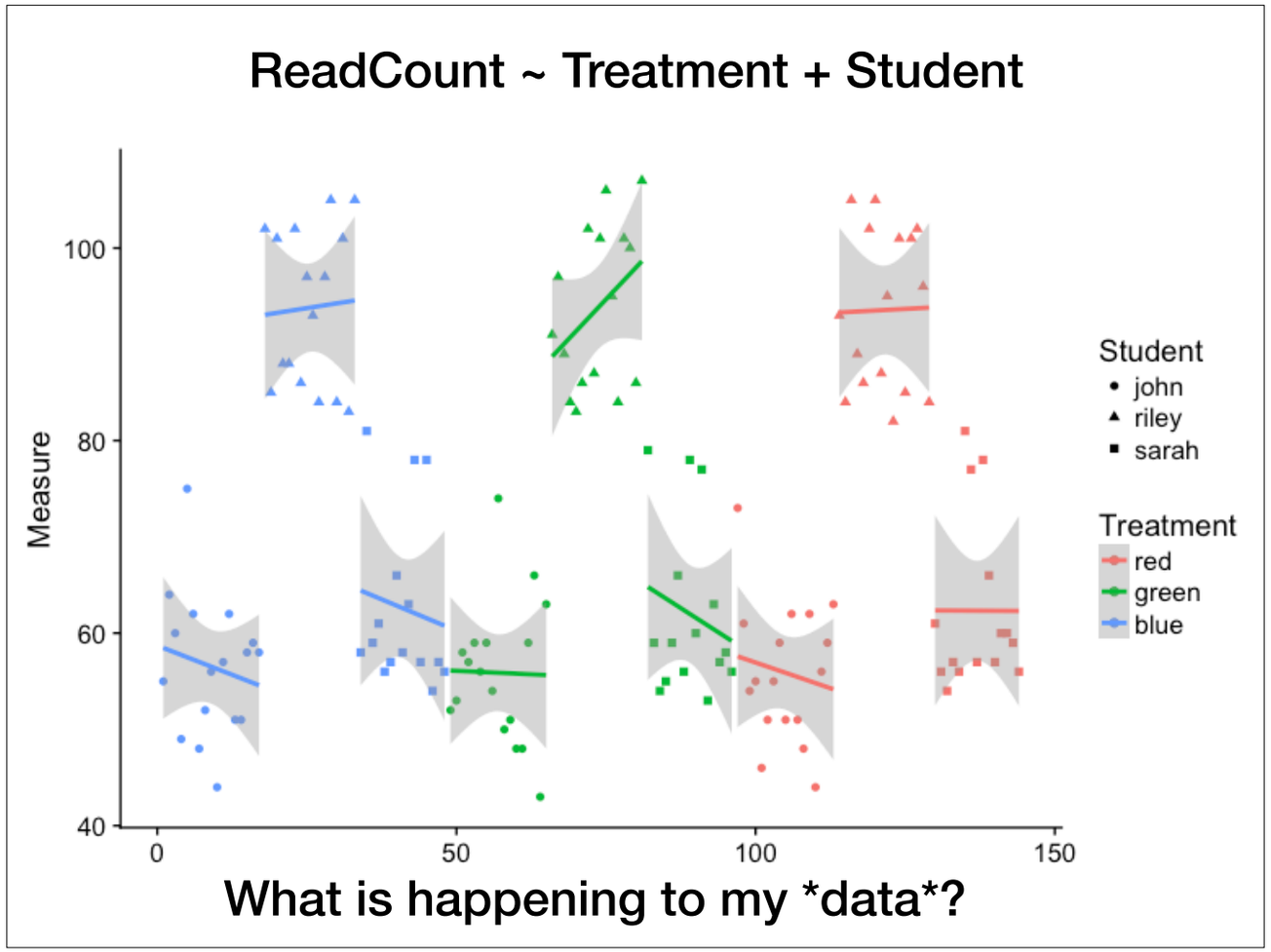

ReadCount ~ Treatment
Measure
100
80
60
40
0
50
100
150
Treatment
red
green
blue

Sample	Treatment	Week	Student	ReadCount	Lane	Library Prep Date
1		1	Riley	92	1	5/23/18
2		1	Sarah	56	2	6/30/18
3		1	John	21	1	5/23/18
4		2	John	77	2	6/30/18
5		2	Riley	35	1	5/23/18
6		2	Sarah	26	2	6/30/18
7		3	Sarah	68	1	5/23/18
8		3	John	41	2	6/30/18
9		3	Riley	42	1	5/23/18

ReadCount ~ Treatment + Student

Sample	Treatment	Week	Student	ReadCount	Lane	Library Prep Date
1		1	Riley	92	1	5/23/18
2		1	Sarah	56	2	6/30/18
3		1	John	21	1	5/23/18
4		2	John	77	2	6/30/18
5		2	Riley	35	1	5/23/18
6		2	Sarah	26	2	6/30/18
7		3	Sarah	68	1	5/23/18
8		3	John	41	2	6/30/18
9		3	Riley	42	1	5/23/18

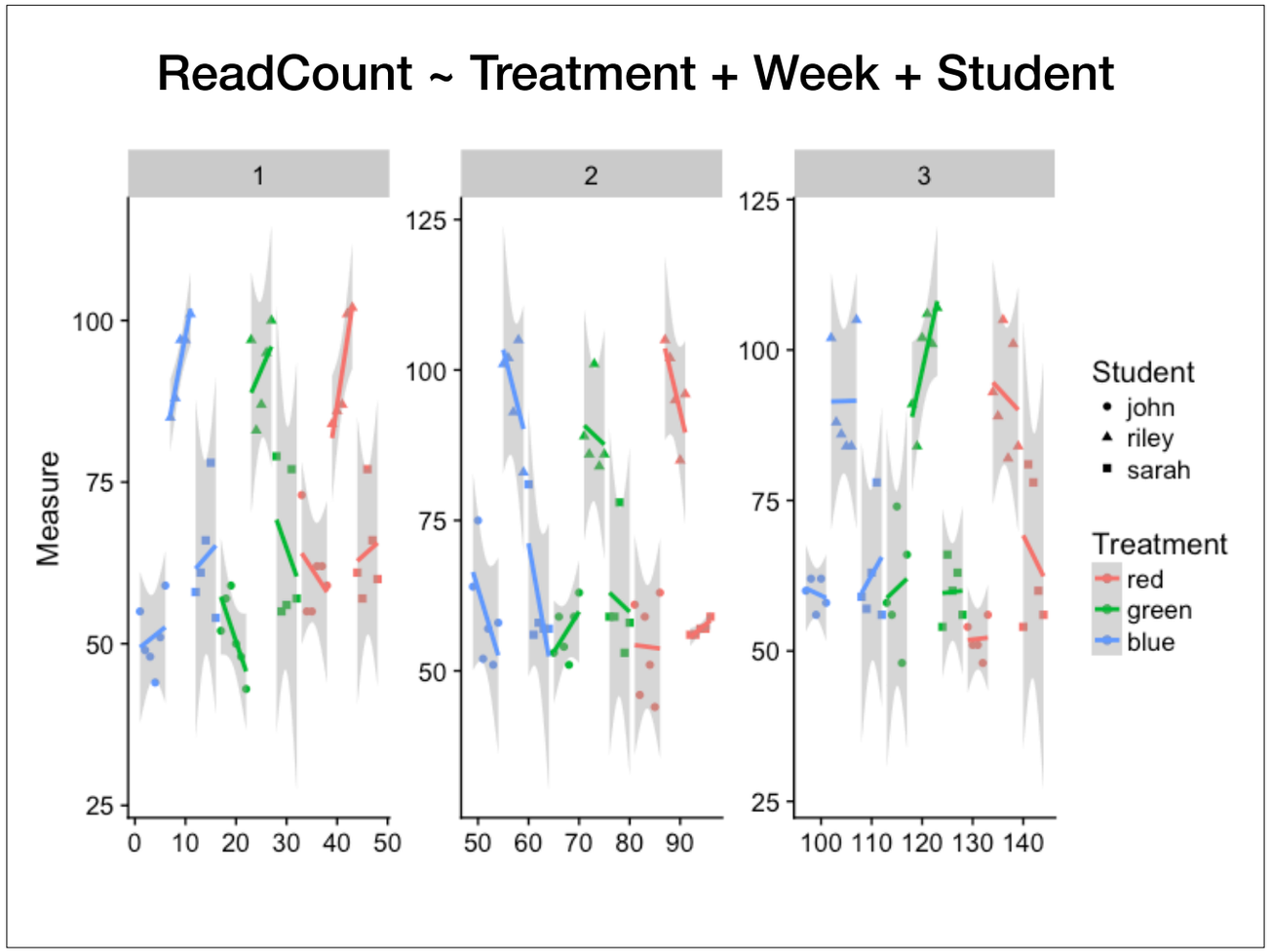

ReadCount ~ Treatment + Student
Measure
100
80
60
40
0
50
100
150
Student
john
riley
sarah
Treatment
red
green
blue


ReadCount ~ Treatment + Student
Measure
100
80
60
40
0
50
100
150
Student
john
riley
sarah
Treatment
red
green
blue
What is happening to my *data*?

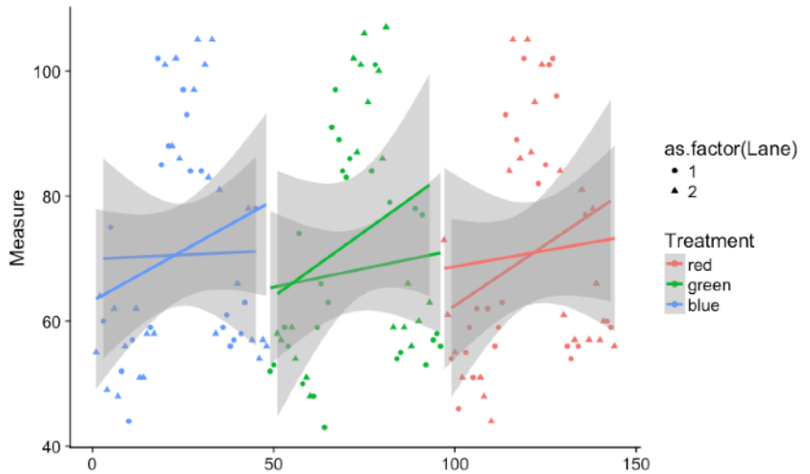
Sample	Treatment	Week	Student	ReadCount	Lane	Library Prep Date
1		1	Riley	92	1	5/23/18
2		1	Sarah	56	2	6/30/18
3		1	John	21	1	5/23/18
4		2	John	77	2	6/30/18
5		2	Riley	35	1	5/23/18
6		2	Sarah	26	2	6/30/18
7		3	Sarah	68	1	5/23/18
8		3	John	41	2	6/30/18
9		3	Riley	42	1	5/23/18

ReadCount ~ Treatment + Week + Student

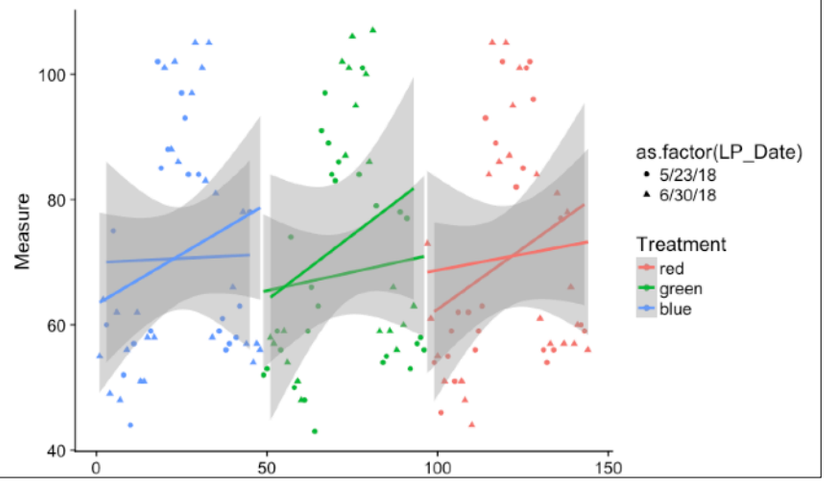
Sample	Treatment	Week	Student	ReadCount	Lane	Library Prep Date
1		1	Riley	92	1	5/23/18
2		1	Sarah	56	2	6/30/18
3		1	John	21	1	5/23/18
4		2	John	77	2	6/30/18
5		2	Riley	35	1	5/23/18
6		2	Sarah	26	2	6/30/18
7		3	Sarah	68	1	5/23/18
8		3	John	41	2	6/30/18
9		3	Riley	42	1	5/23/18


ReadCount ~ Treatment + Week + Student
1
2
3
Measure
125
100
75
50
25
0 10 20 30 40 50
50 60 70 80 90
100 110 120 130 140
Student
john
riley
sarah
Treatment
red
green
blue

ReadCount ~ Treatment + Lane



ReadCount ~ Treatment + LP_Date

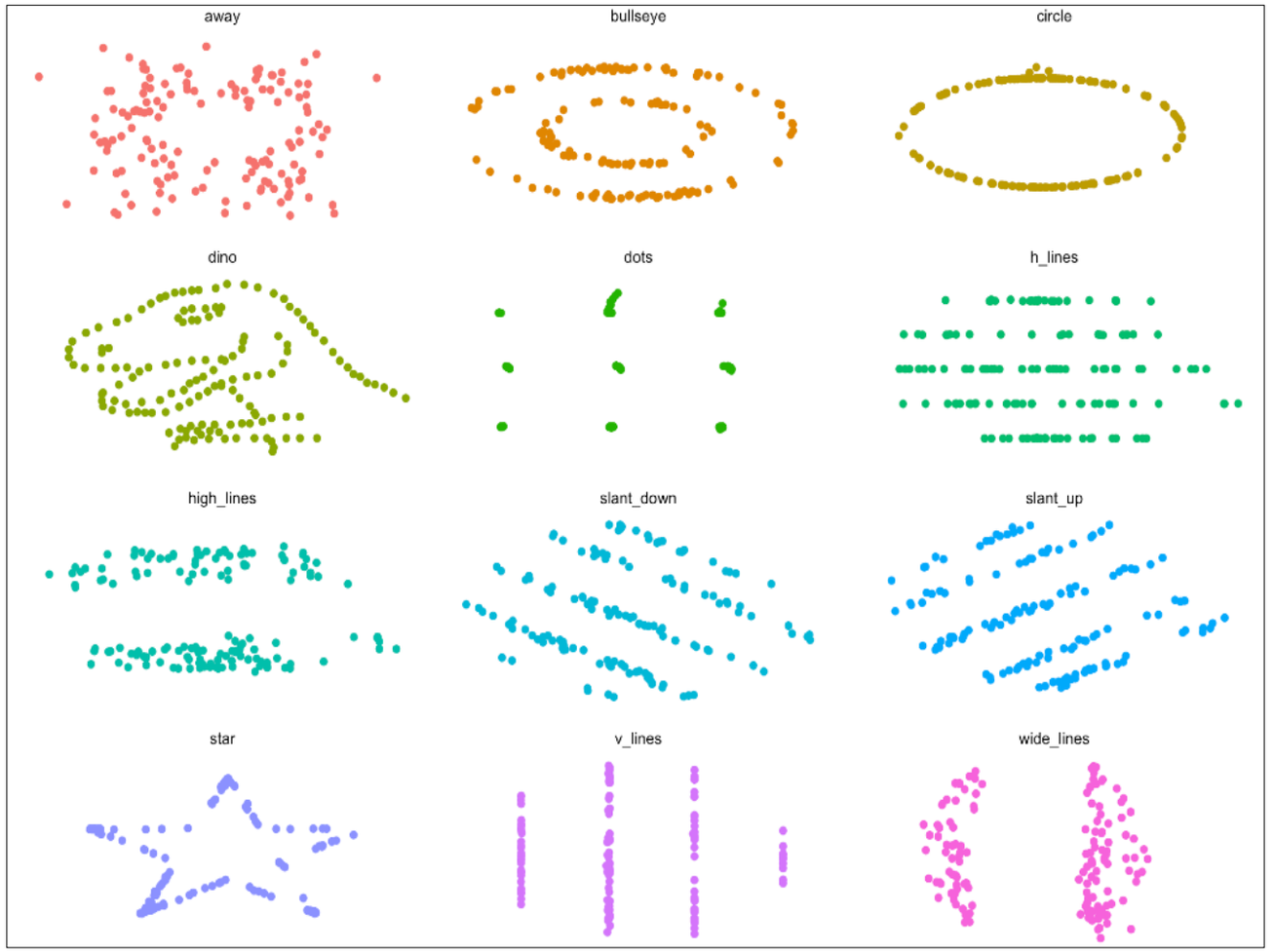


Sample	Treatment	Week	Student	ReadCount	Lane	Library Prep Date
1		1	Riley	92	1	5/23/18
2		1	Sarah	56	2	6/30/18
3		1	John	21	1	5/23/18
4		2	John	77	2	6/30/18
5		2	Riley	35	1	5/23/18
6		2	Sarah	26	2	6/30/18
7		3	Sarah	68	1	5/23/18
8		3	John	41	2	6/30/18
9		3	Riley	42	1	5/23/18

Lane and LP_Date are perfectly confounded!

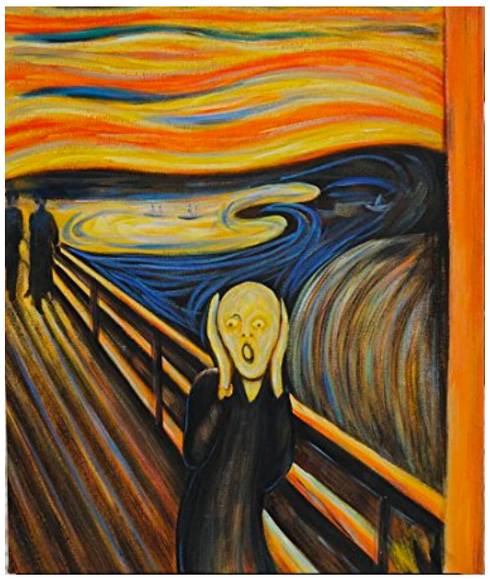
Sample	Treatment	Week	Student	ReadCount	Lane	Library Prep Date
1		1	Riley	92	1	5/23/18
2		1	Sarah	56	2	6/30/18
3		1	John	21	1	5/23/18
4		2	John	77	2	6/30/18
5		2	Riley	35	1	5/23/18
6		2	Sarah	26	2	6/30/18
7		3	Sarah	68	1	5/23/18
8		3	John	41	2	6/30/18
9		3	Riley	42	1	5/23/18

All that stuff at the beginning
is actually important

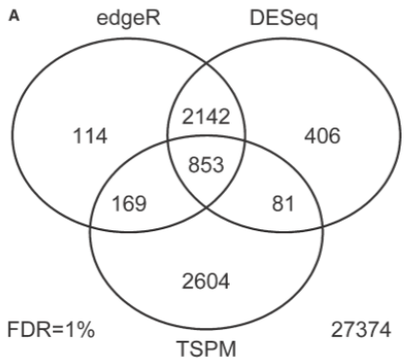

away
bullseye
circle
dino
dots
h_lines
high_lines
slant_down
slant_up
star
v_lines
wide_lines

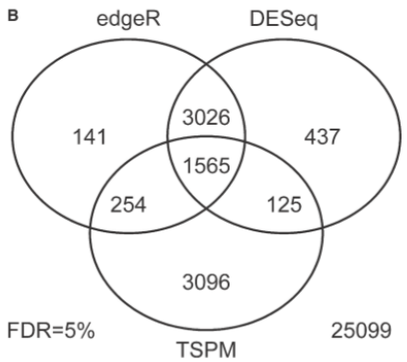
RNAseq Math

RNAseq Math



How do the methods compare for real data?

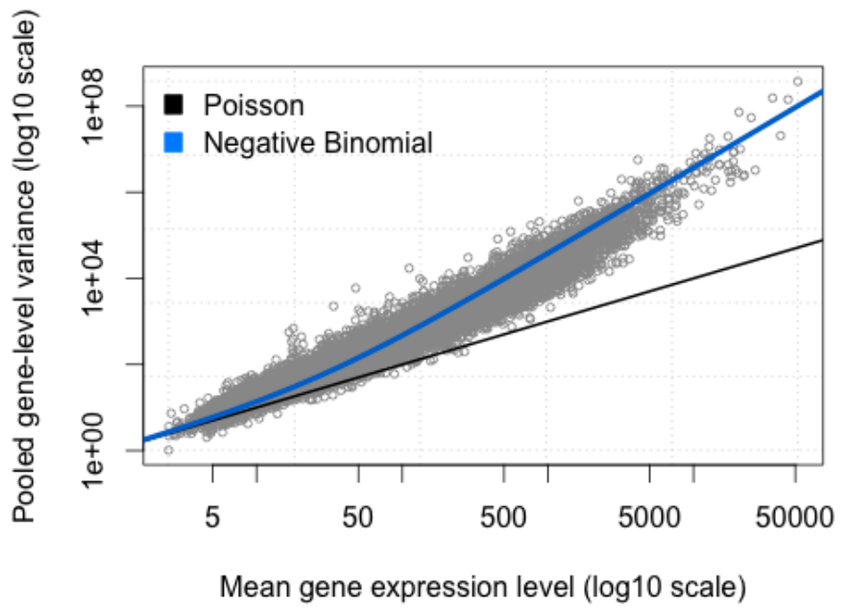




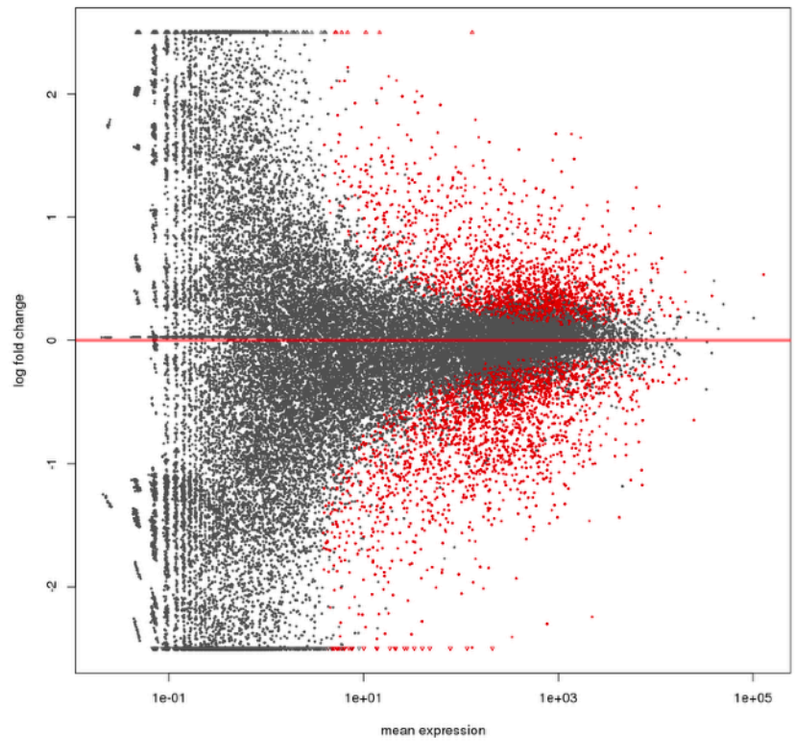
Kvam et al. 2012

Dispersion

Dispersion


Poisson
Negative Binomial
Pooled gene-level variance (log10 scale)
1e+00
1e+04
1e+08
5
50
500
5000
50000
Mean gene expression level (log10 scale)

Numbers are hard



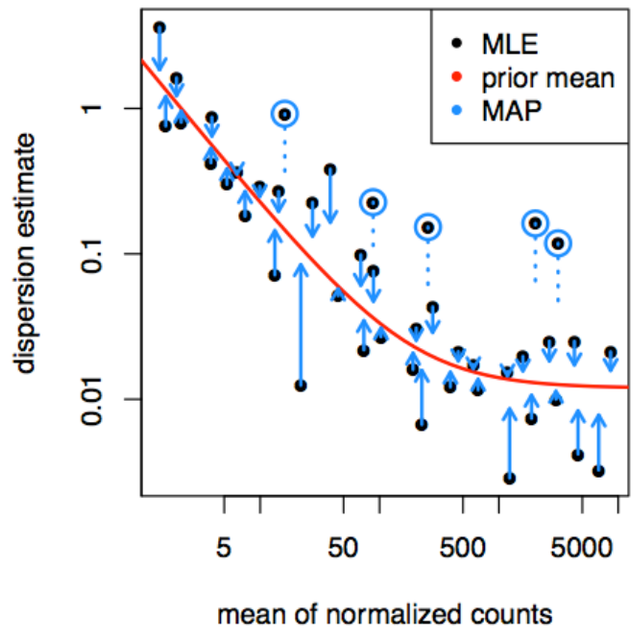
	Gene 1 Count	Gene 2 Count
blue	5	10
blue	5	10
blue	5	10
blue	5	10
blue	5	10
blue	5	10
blue	5	10
blue	5	10
blue	5	10
blue	5	10
blue	5	10
blue	5	10
yellow	5	1
yellow	5	1
yellow	5	1
yellow	5	1
yellow	5	1
yellow	5	1
yellow	5	1
yellow	5	1
yellow	5	1

	Gene 1 Count	Gene 2 Count	Gene 3 Count
blue	5	10	-
blue	5	10	-
blue	5	10	4
blue	5	10	-
blue	5	10	-
blue	5	10	-
blue	5	10	9
blue	5	10	-
blue	5	10	-
blue	5	10	-
blue	5	10	30
yellow	5	1	45
yellow	5	1	50
yellow	5	1	50
yellow	5	1	43
yellow	5	1	45
yellow	5	1	44
yellow	5	1	41
yellow	5	1	43

	Gene 1 Count	Gene 2 Count	Gene 3 Count	Gene 4 Count	Gene 5 Count
blue	5	10	-	4	5
blue	5	10	-	90	6
blue	5	10	4	51	9
blue	5	10	-	76	6
blue	5	10	-	97	10
blue	5	10	-	33	1
blue	5	10	9	88	7
blue	5	10	-	96	10
blue	5	10	-	26	2
blue	5	10	-	68	4
blue	5	10	30	48	2
yellow	5	1	45	60	65
yellow	5	1	50	65	60
yellow	5	1	50	75	63
yellow	5	1	43	35	53
yellow	5	1	45	20	65
yellow	5	1	44	84	63
yellow	5	1	41	89	52
yellow	5	1	43	48	60

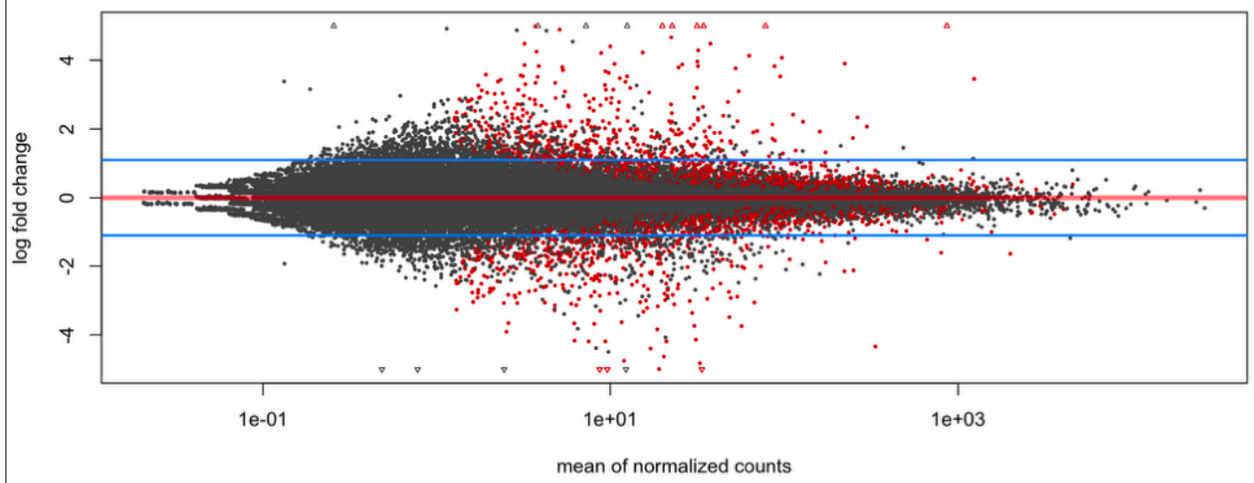
Model Inception

Dispersion



Love, Huber & Anders 2014 BioRXiV doi: 10.1101/002832

RNAseq without over-dispersion correction



RNAseq with over-dispersion correction