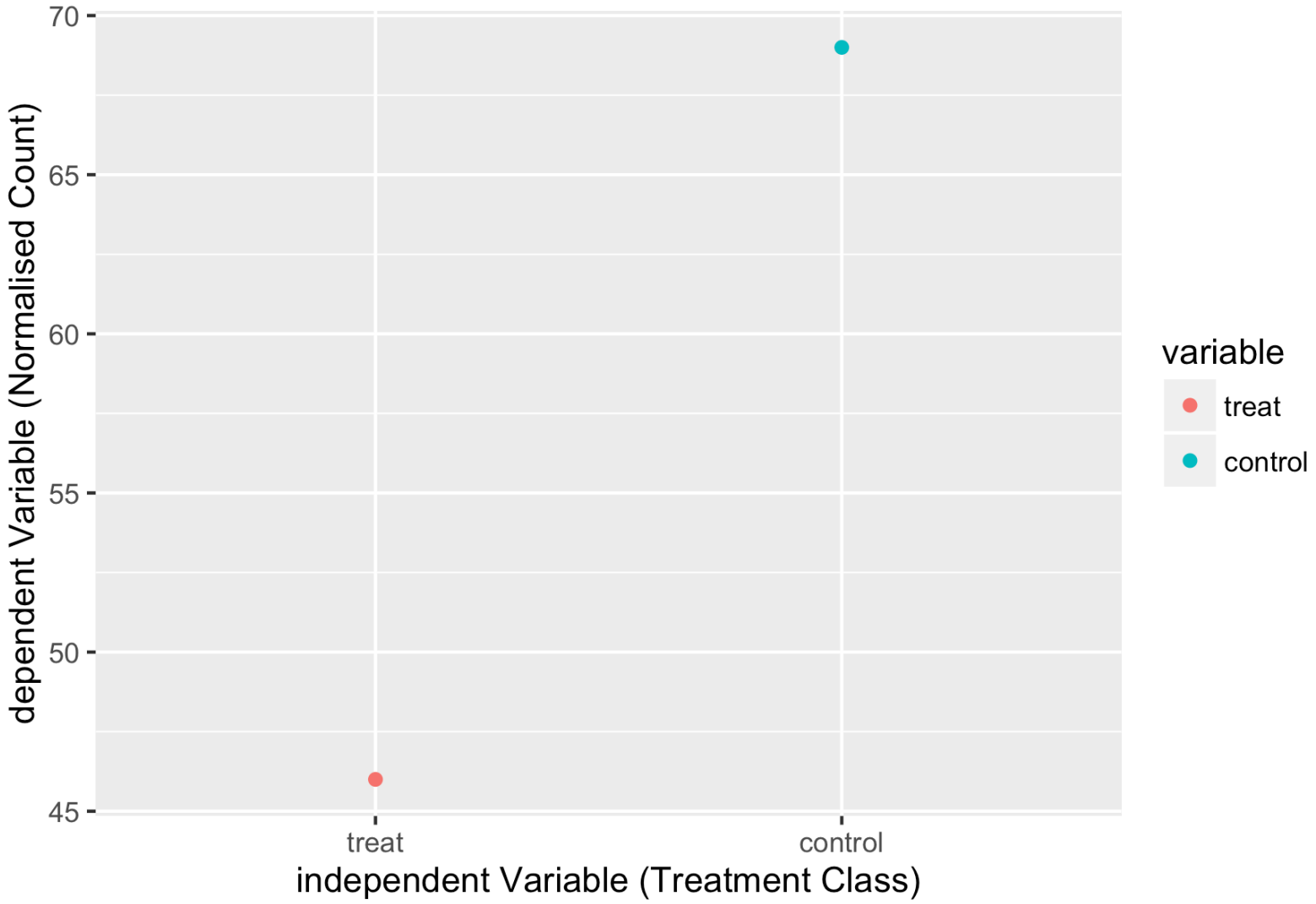
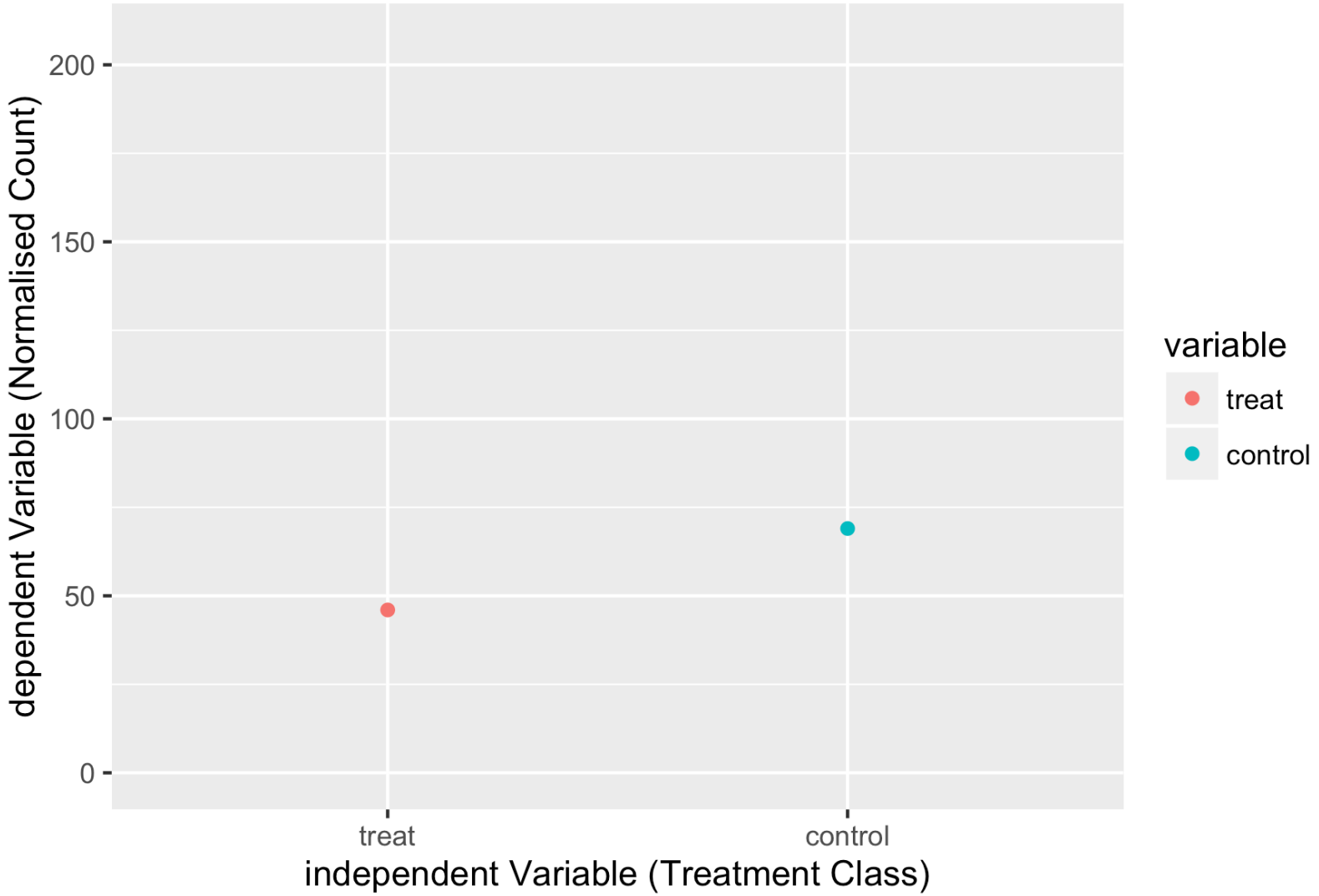
Statistics Refresher

"Statistics is, or should be, about scientific investigation and how to do it better, but many statisticians believe it is a branch of mathematics. Now I agree that the physicist, the chemist, the engineer, and the statistician can never know too much mathematics, but their objectives should be better physics, better chemistry, better engineering, and in the case of statistics, better scientific investigation. Whether in any given study this implies more or less mathematics is incidental."

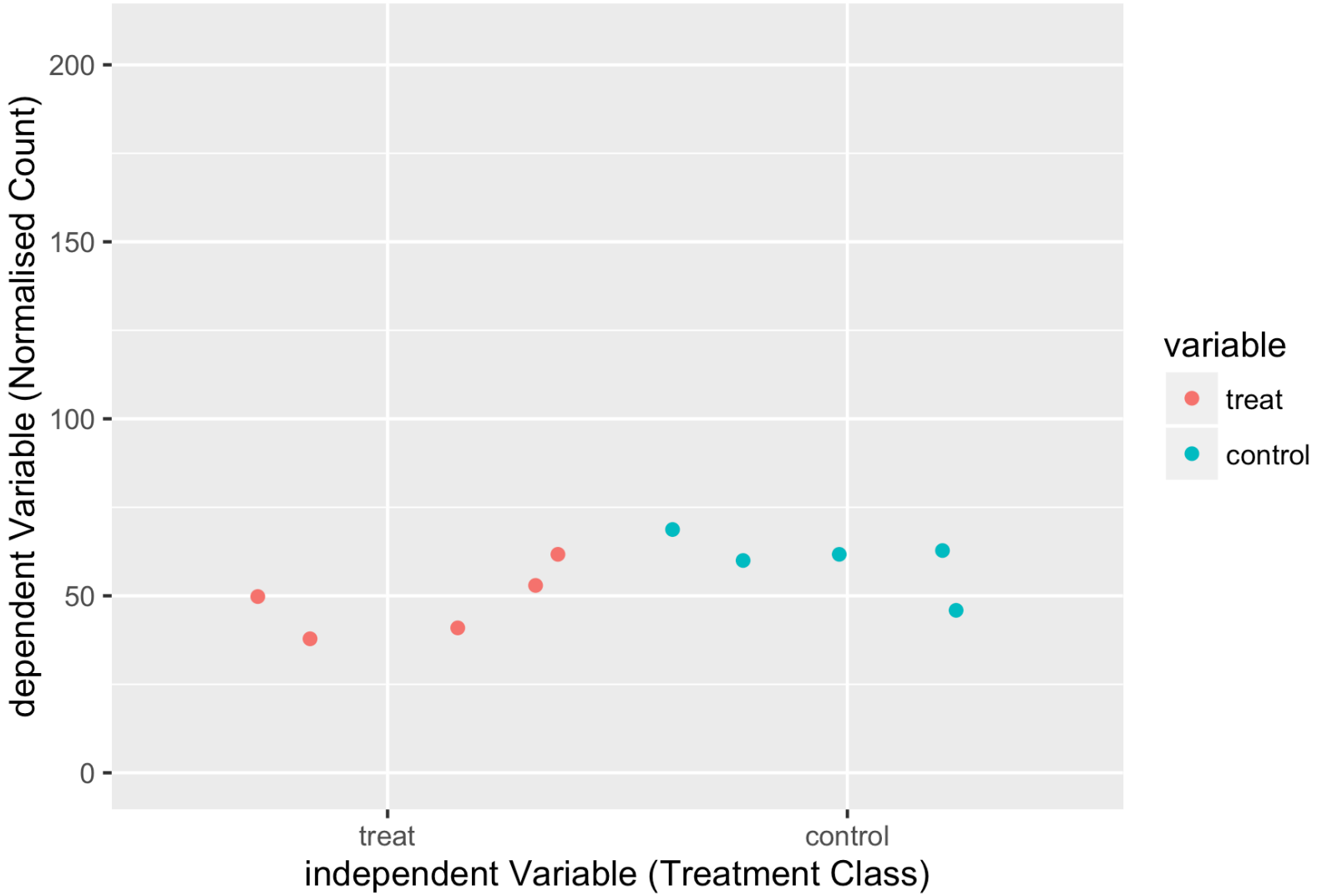
George E. P. Box



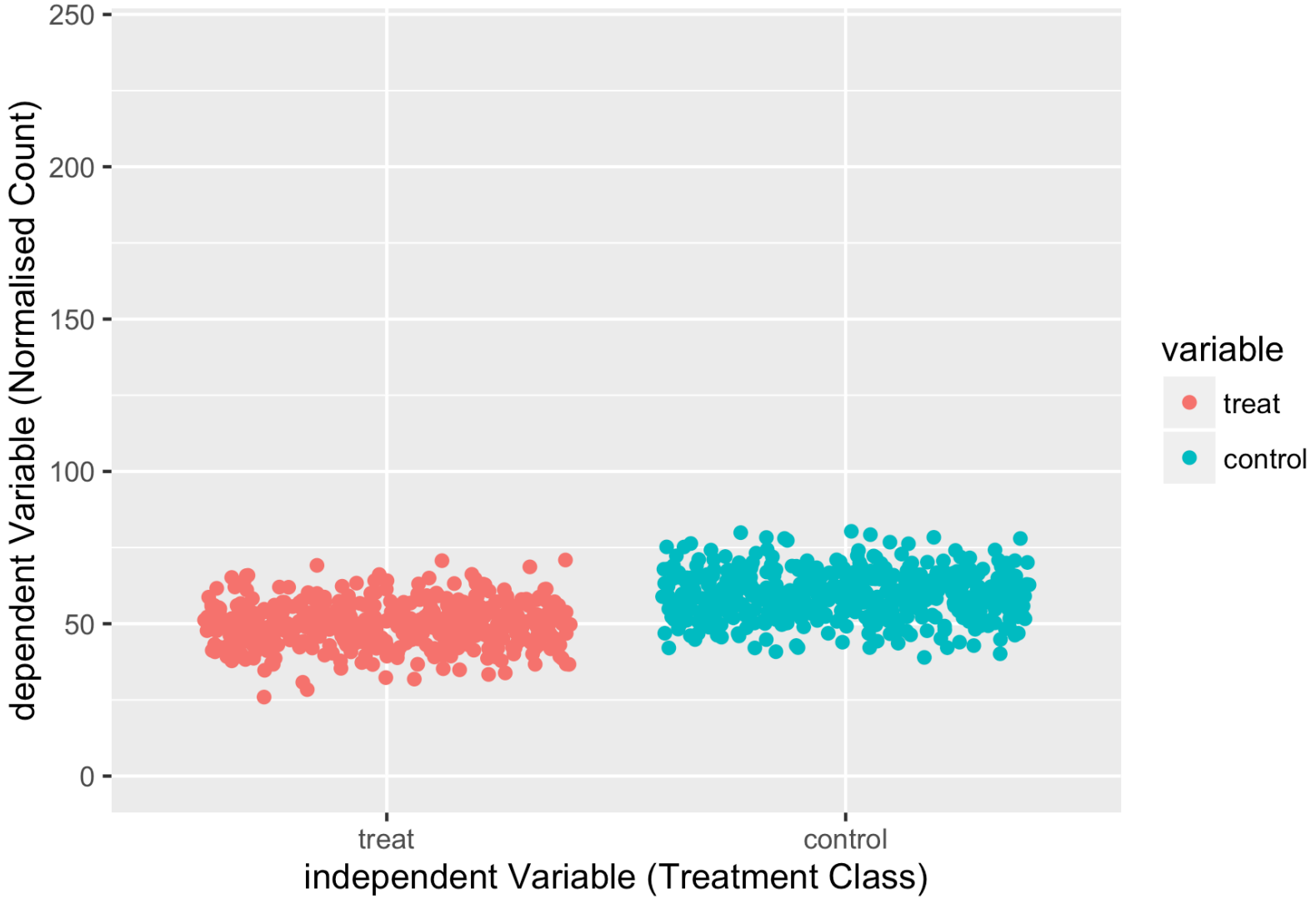
Two conditions is there a treatment effect?



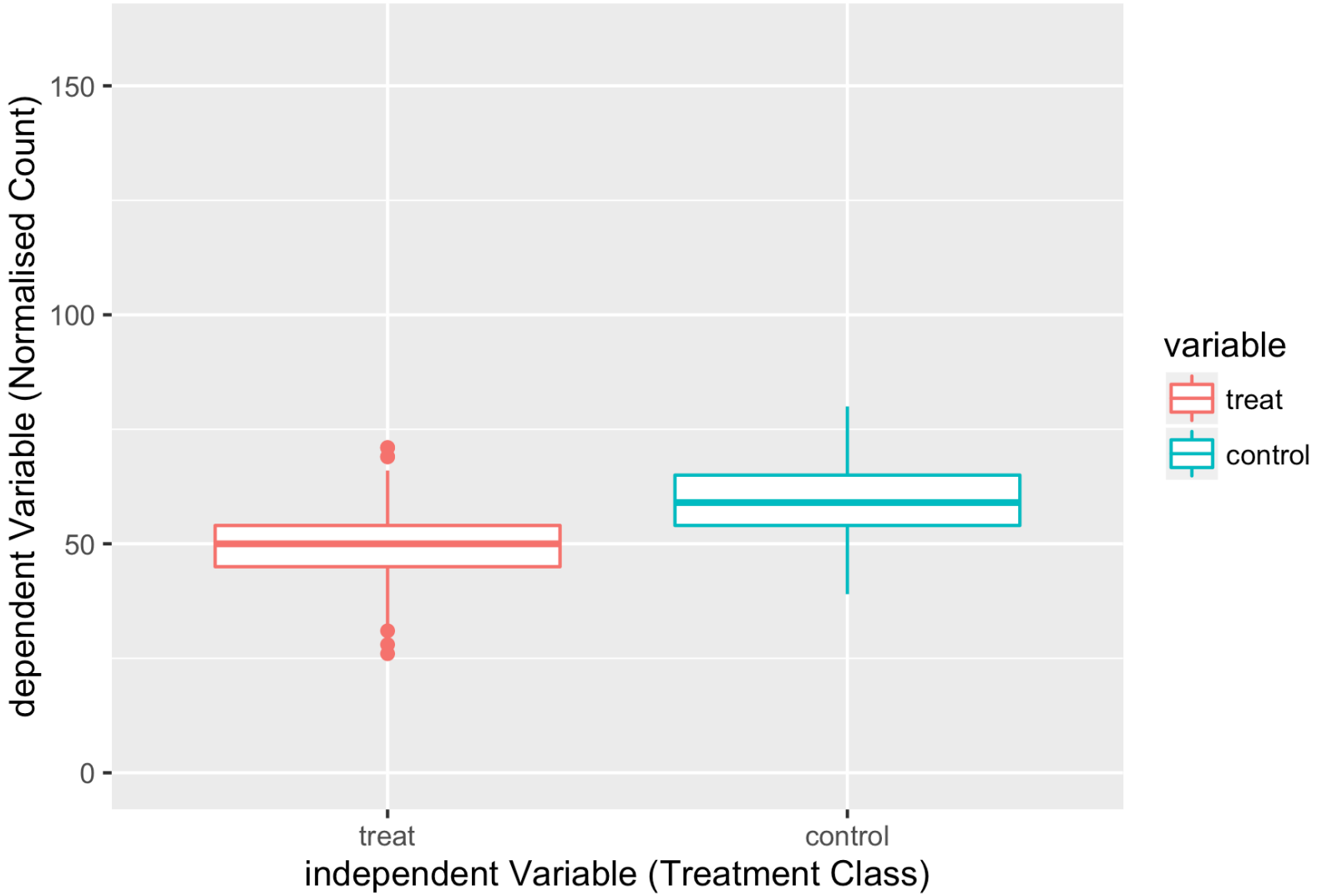
If consider possible range of values



If more samples are considered



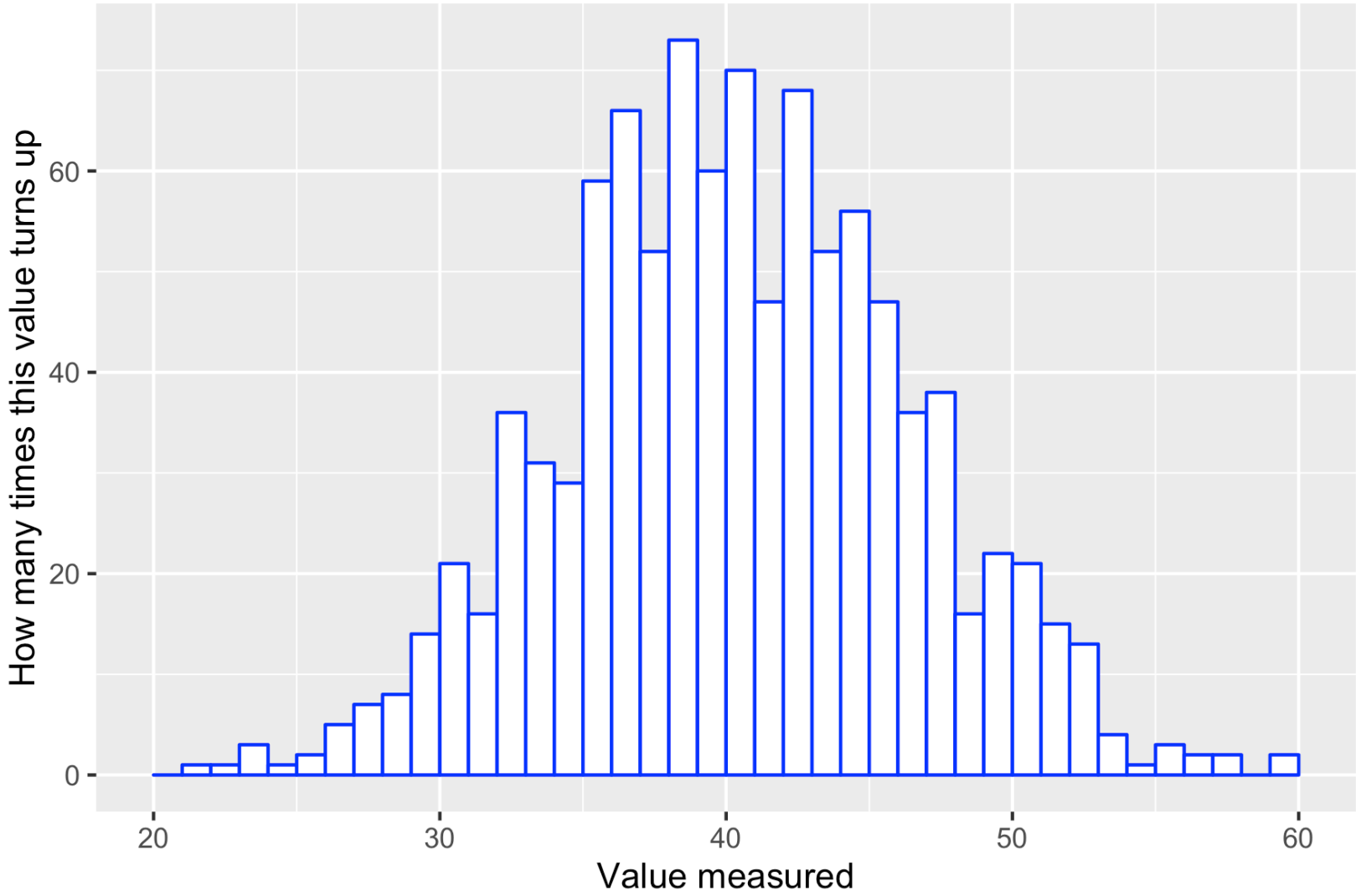
If lots of samples are considered



Use statistics to help “see the woods for the trees”

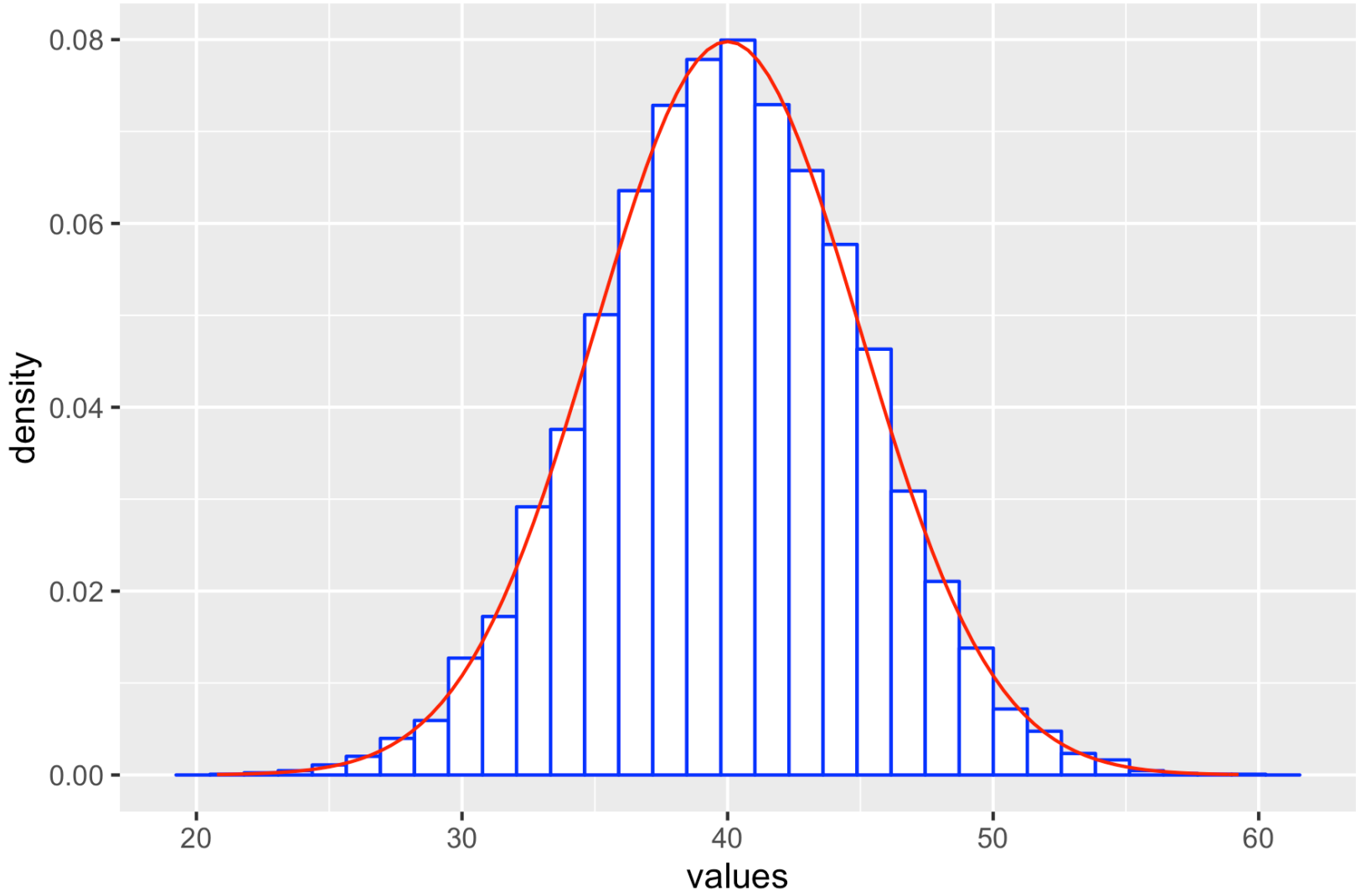
Hypothesis Testing

- null hypothesis
  - there is no difference between the values for the two groups
- alternative hypothesis
  - there is a difference
- p-value
  - what is the probability of seeing this data if the null hypothesis (there is no difference) is true
- the smaller the p-value the less likely the two groups are the same.

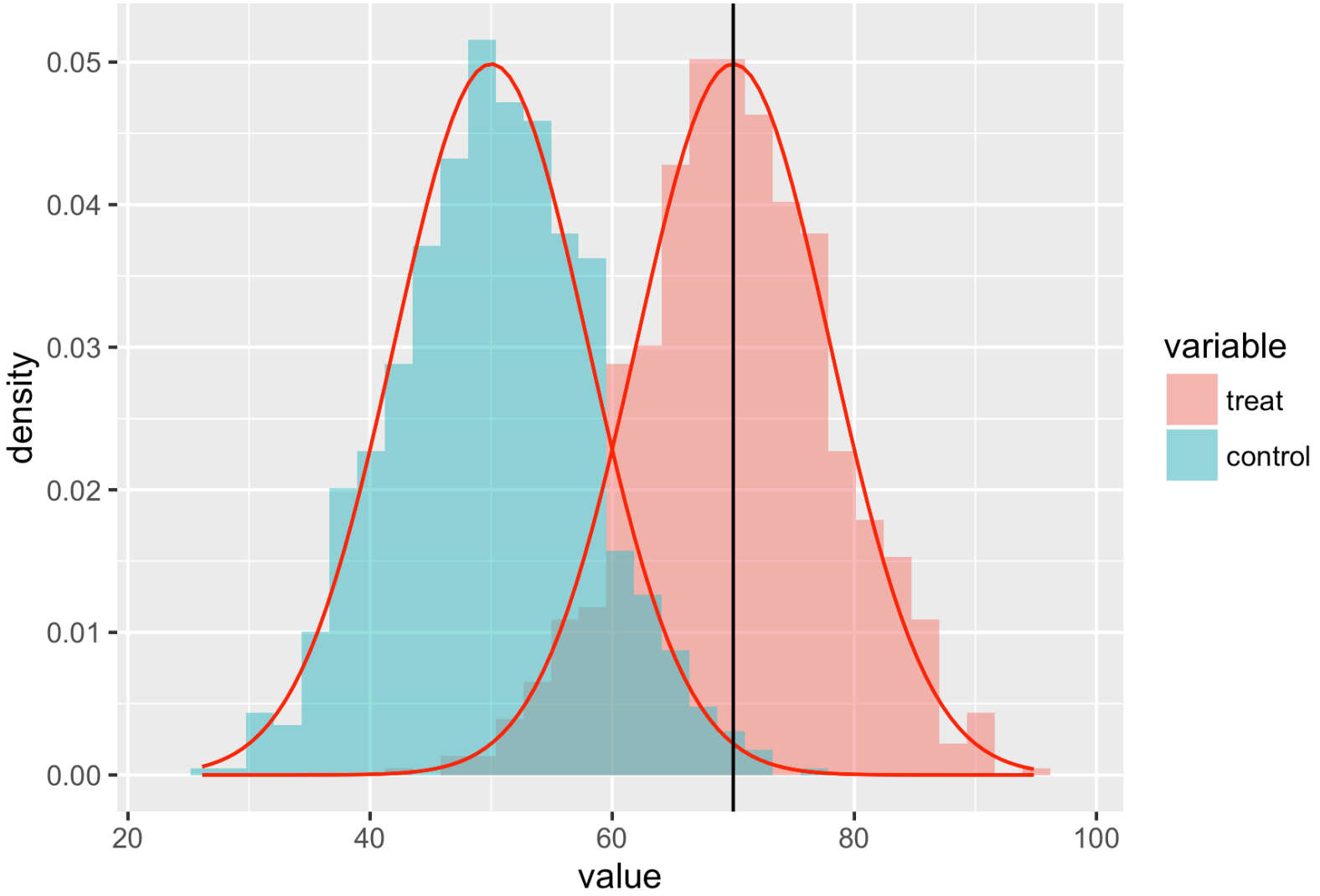


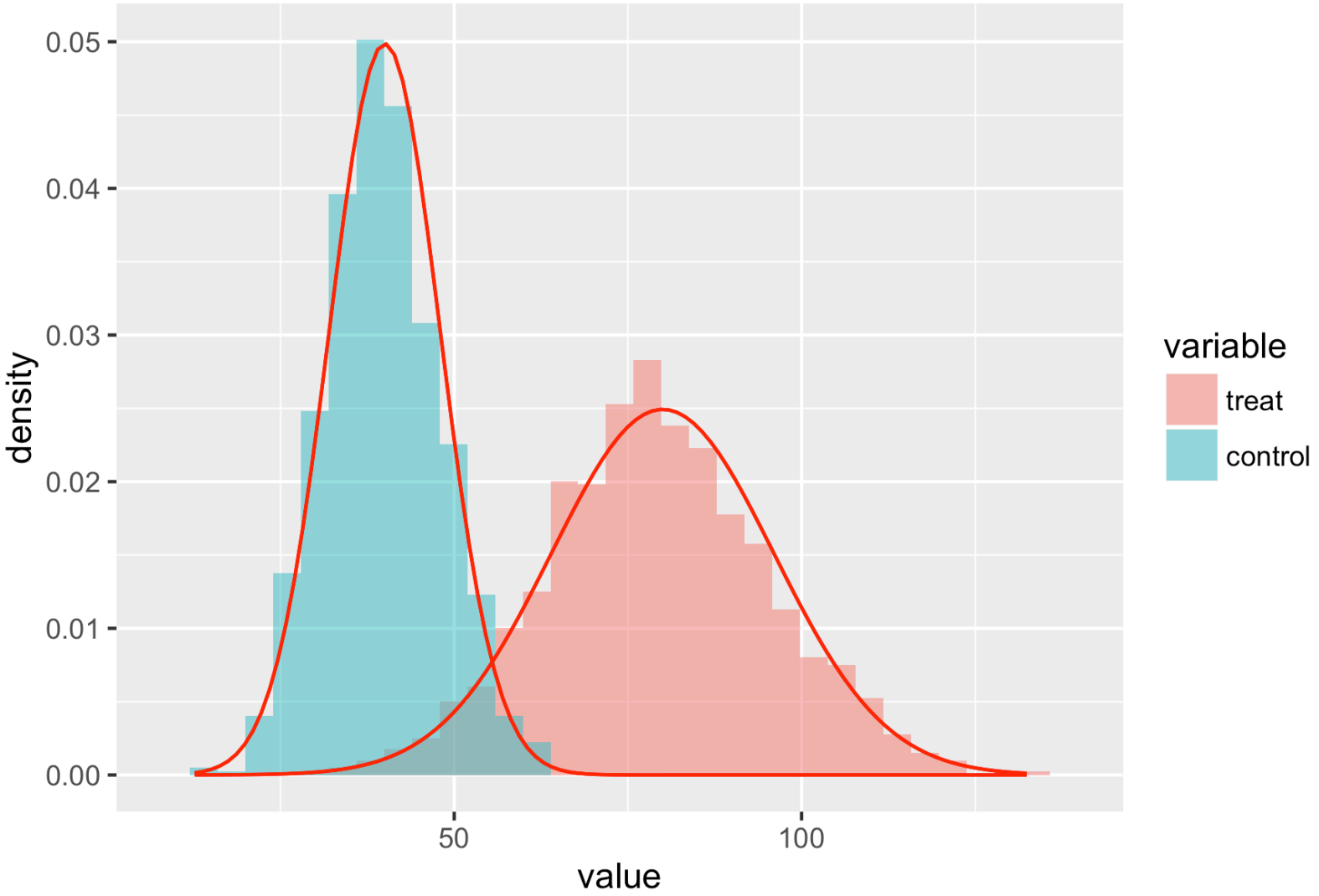
No only the location (central tendency) but the scale (spread) of the data

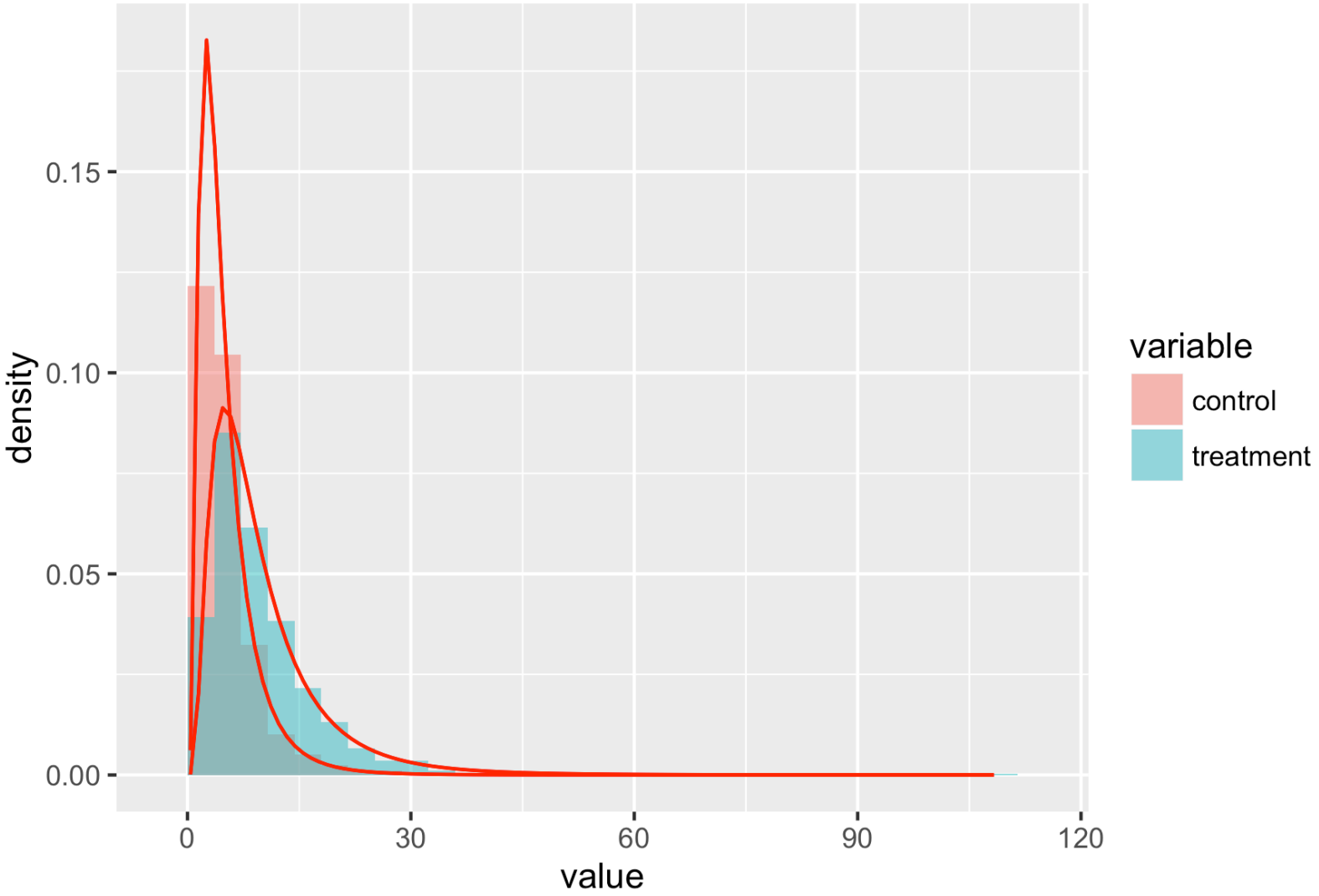
- Parametric Statistics
  - make an assumption about a mathematical model for the shape of the data
- Non-parametric Statistic
  - no assumption or mathematical model for the shape of the data



Gaussian (or normal) model


0.05
0.04
0.03
0.02
0.01
0.00
density
20
40
60
80
100
value
variable
treat
control

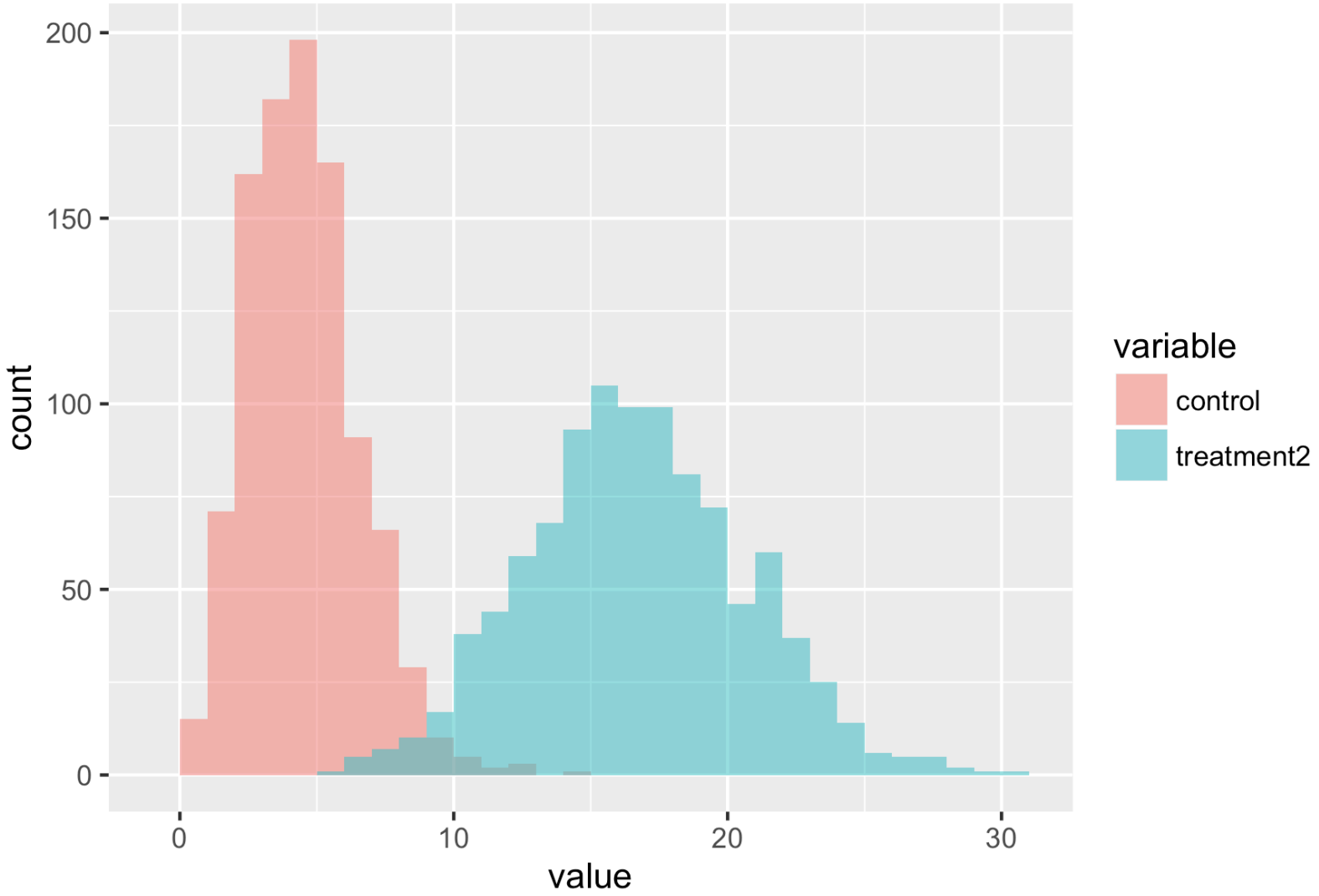

0.05
0.04
0.03
0.02
0.01
0.00
density
50
100
value
variable
treat
control

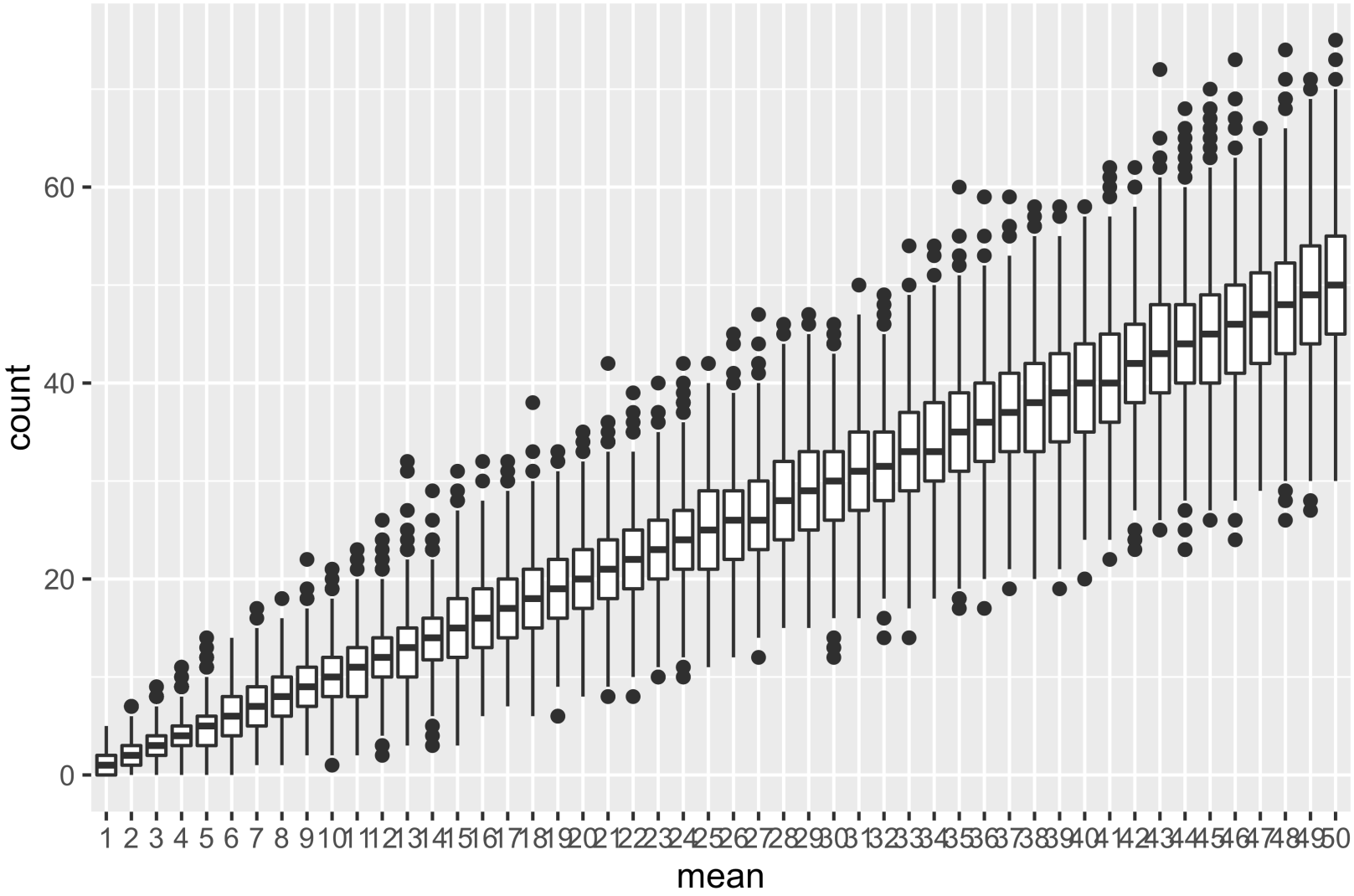

density
0.15
0.10
0.05
0.00
0
30
60
90
120
value
variable
control
treatment

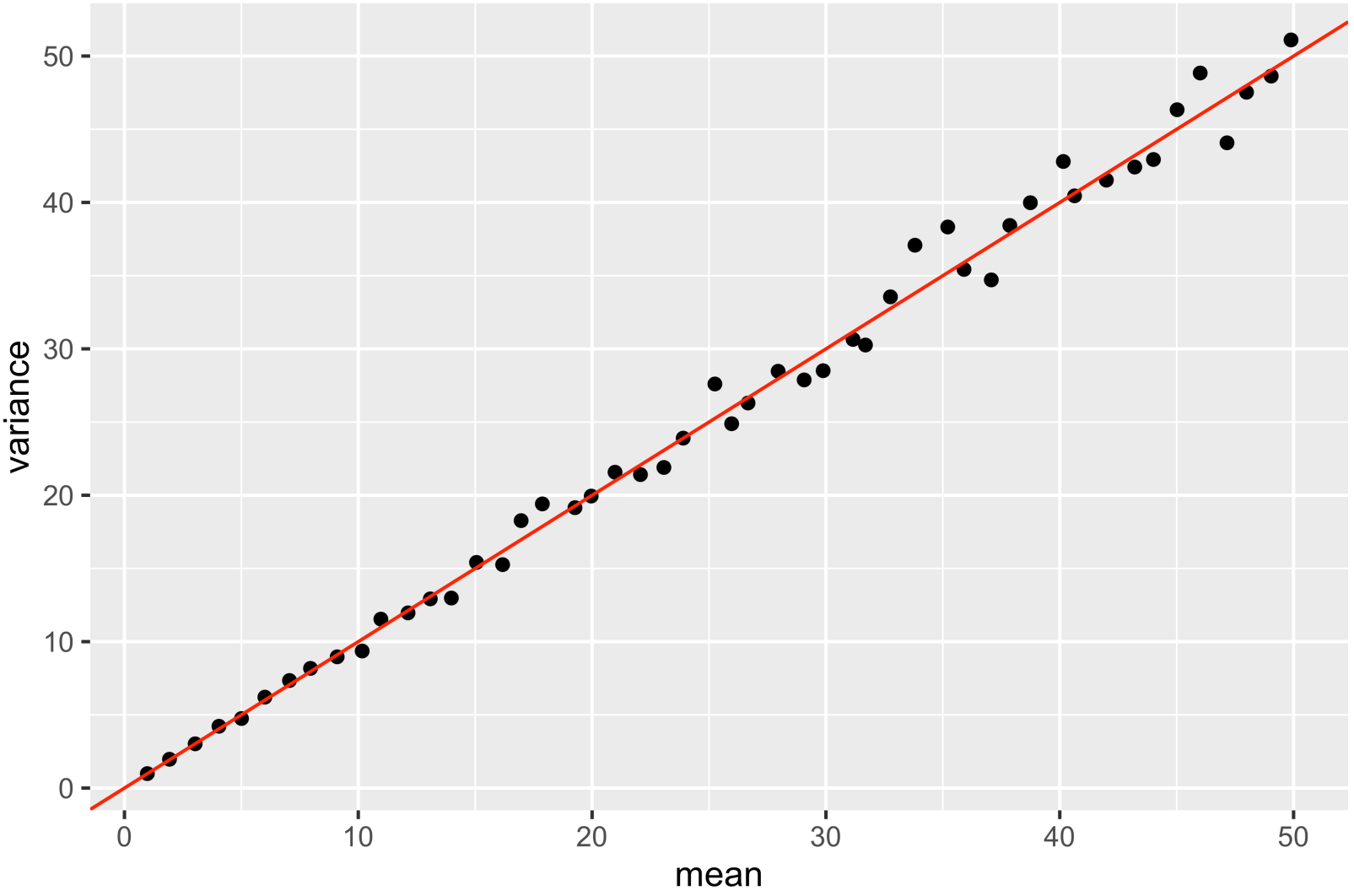
The Shape of RNA-Seq Data

"The sciences do not try to explain, they hardly even try to interpret, they mainly make models. By a model is meant a mathematical construct which, with the addition of certain verbal interpretations, describes observed phenomena. The justification of such a mathematical construct is solely and precisely that it is expected to work"

John Von Neumann


200
150
100
50
0
count
0
10
20
30
value
variable
control
treatment2


count
0
20
40
60
mean
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50


variance
mean
0
10
20
30
40
50

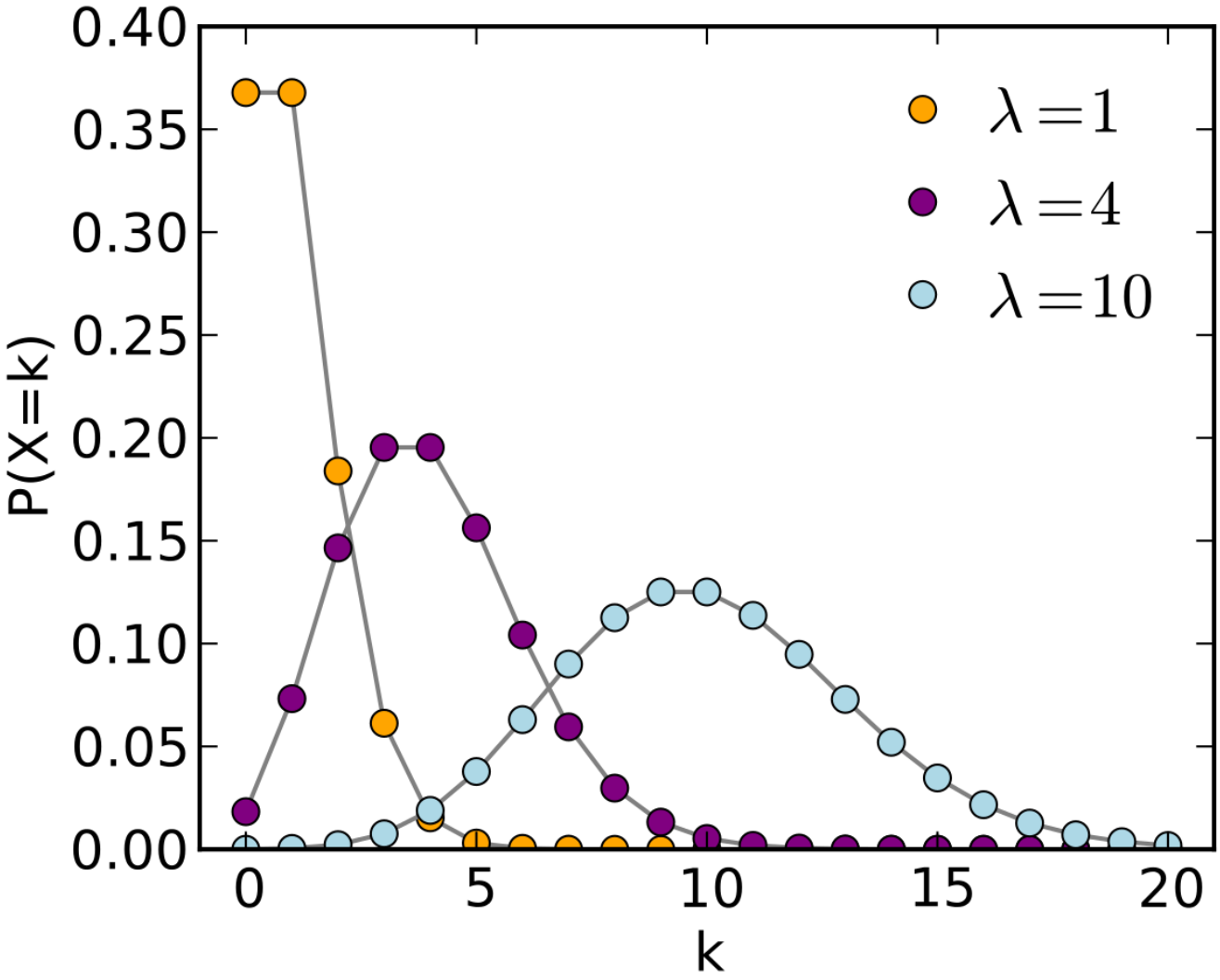
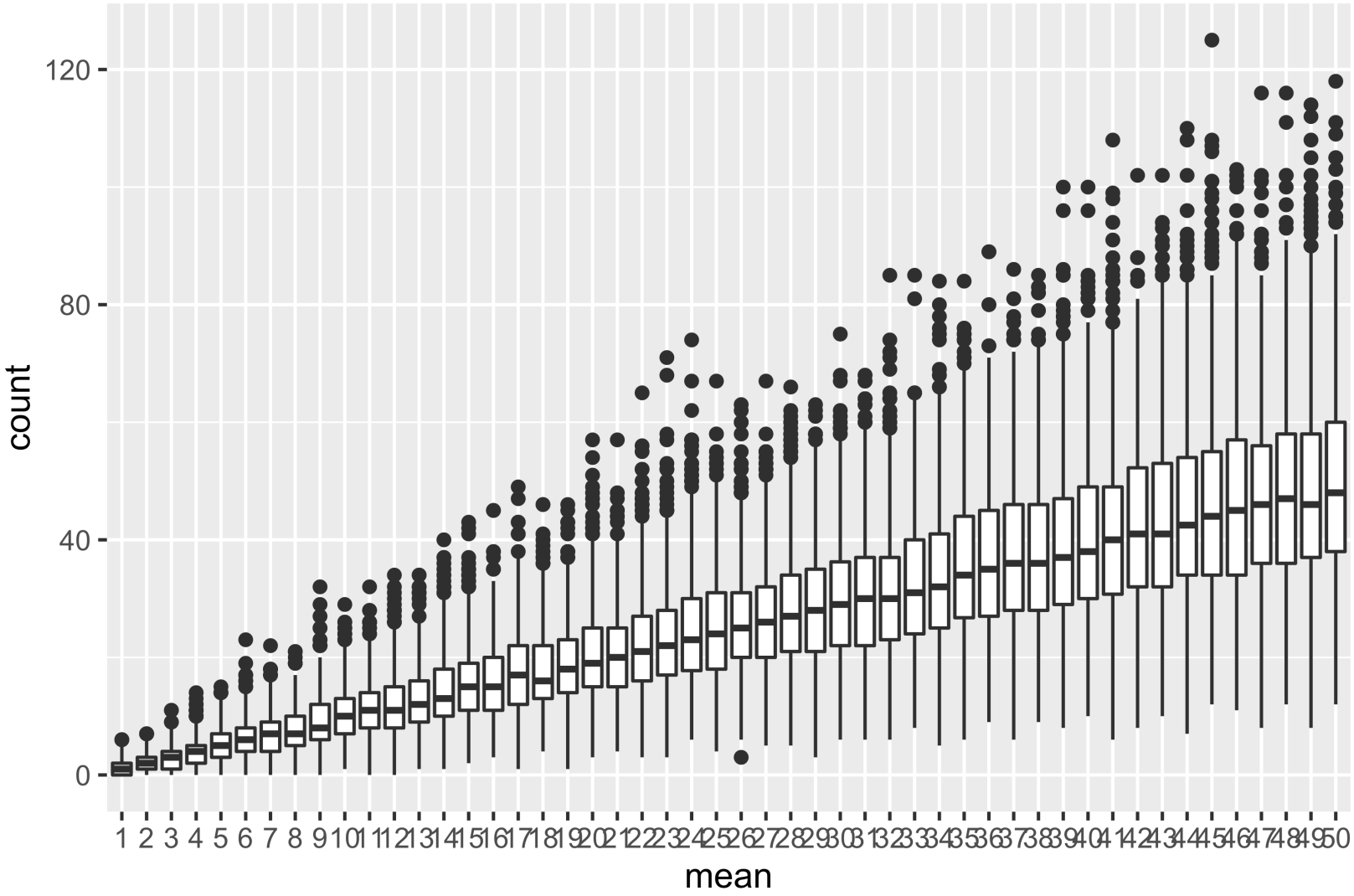
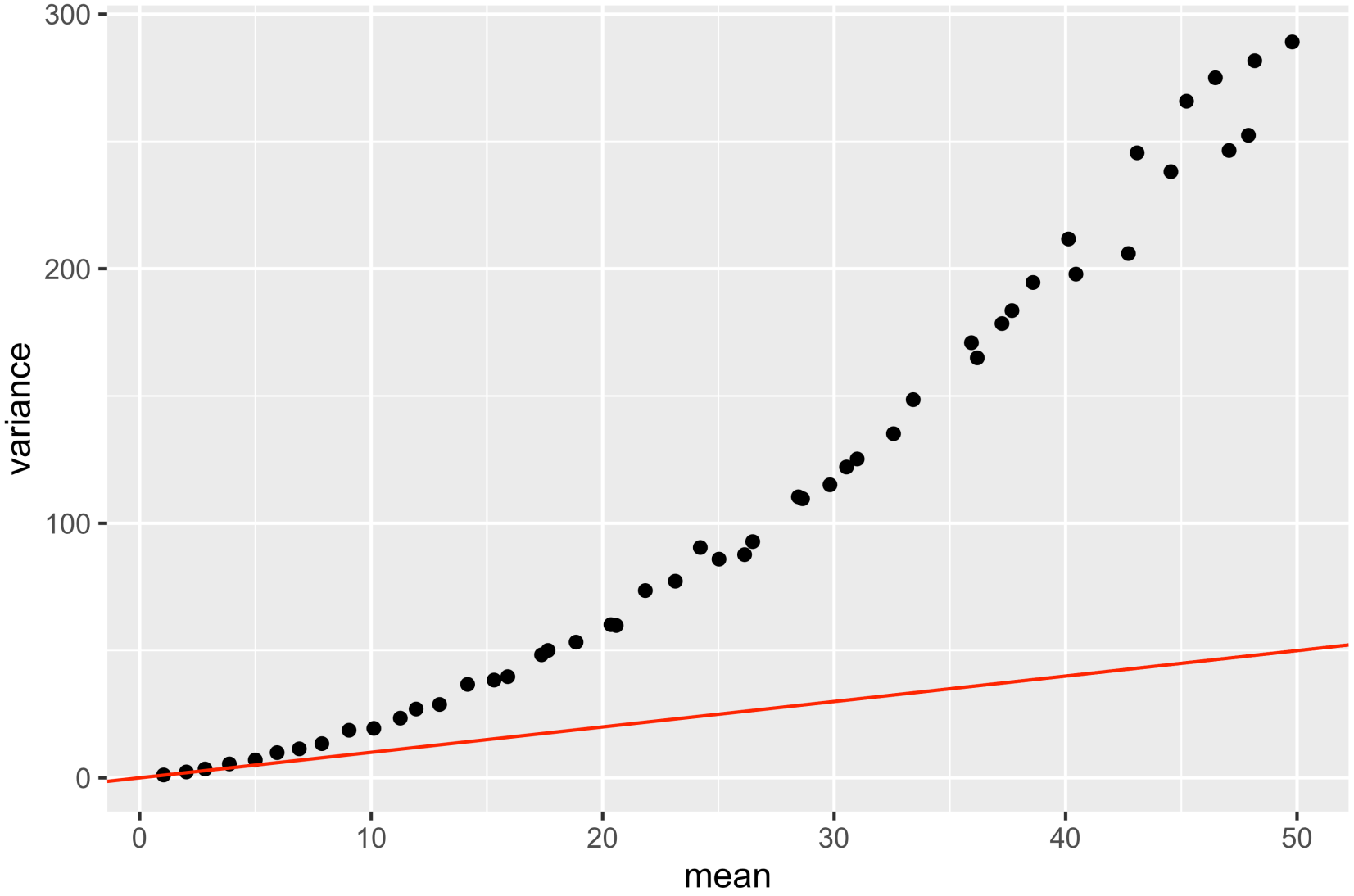


Image Wikipedia


120
80
40
0
count
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50
mean


300
200
100
0
variance
0
10
20
30
40
50
mean

- RNA-seq Data
  - is count data
    - count data is often near poissonian
    - technical replicates tend to display poisson distribution
    - biological replicates tend to be overdisperse ie the variance increase faster than the mean
  - alternative models
    - poisson mixture
    - negative bionomial
    - beta binomial
    - quasai-likelihood methods
  - transform the data

RNA-seq data is generally low replicate data and therefore statistical power is low

- use data points from all genes to calculate or model the variance (variance shrinkage)
  - use emperical bayes methods sample the data for to estimate variance
  - fit models of mean variance relationships to all data to estimate variances for specific individual genes
- use robust statistical methods to lessen the impact of outliers

Normalisation

"You can't compare an apple to an orange. It will cause a lot of self-esteem issues."

Craig Sheffer

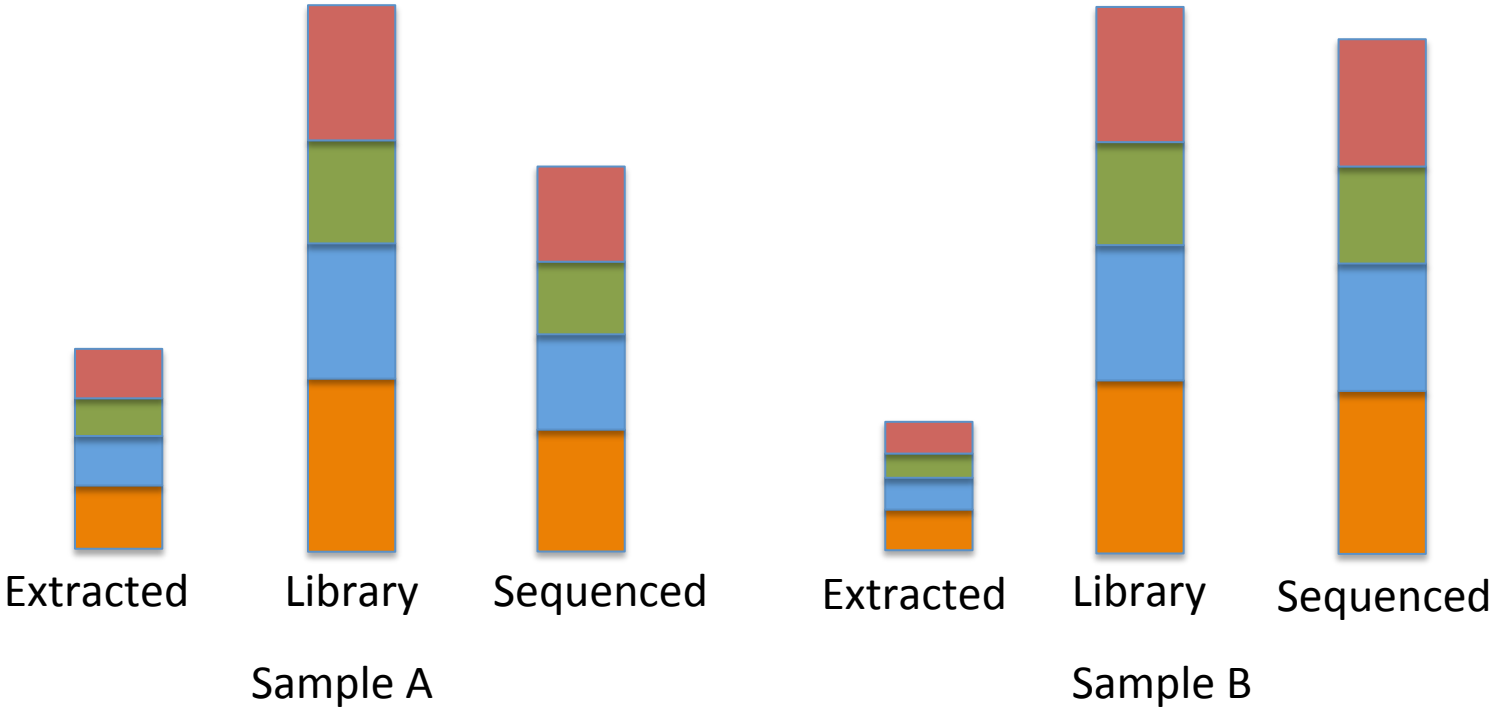
Normalisation does not have an exact definition but broadly

- it attempts to ensure comparison of like-with-like.
- post hoc reduction of unwanted variation (not related to the experimental conditions)

It is an essential step, and choice of the factors to normalise over can make a big difference to the result.

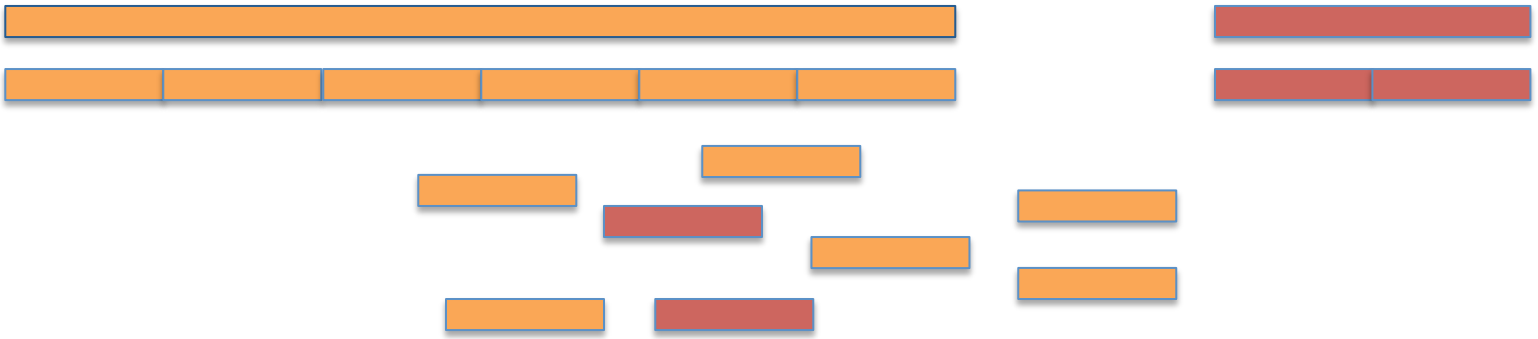
Library size variation

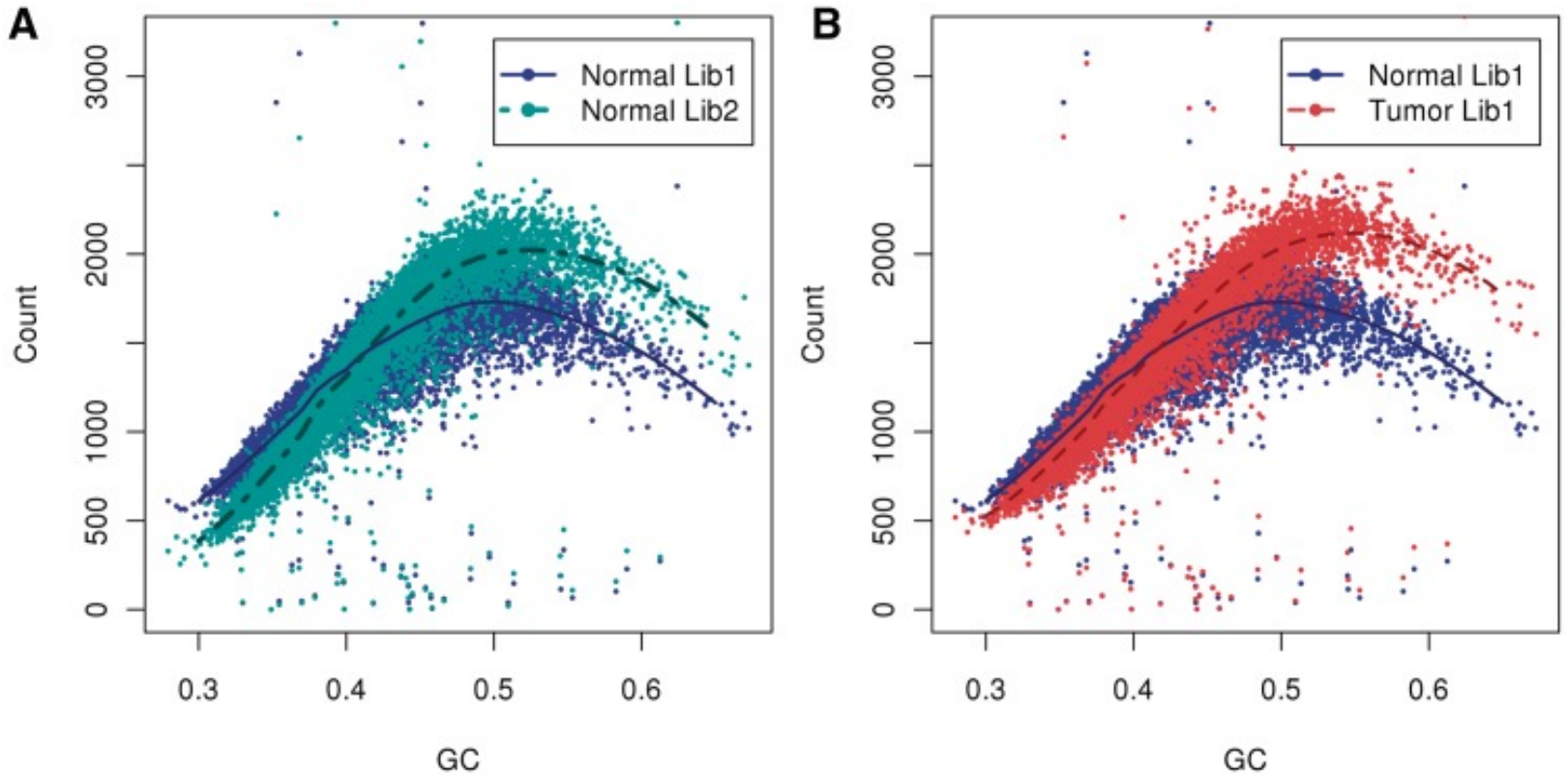
- total number sequence reads will vary between samples for non-biological reasons



Gene Length

- longer gene can be broken into more different fragments and therefore it is more likely fragments from long genes will be sequenced
  - complicated if genes change isoform between treatments
  - smaller counts makes significant differntial expression harder to detect

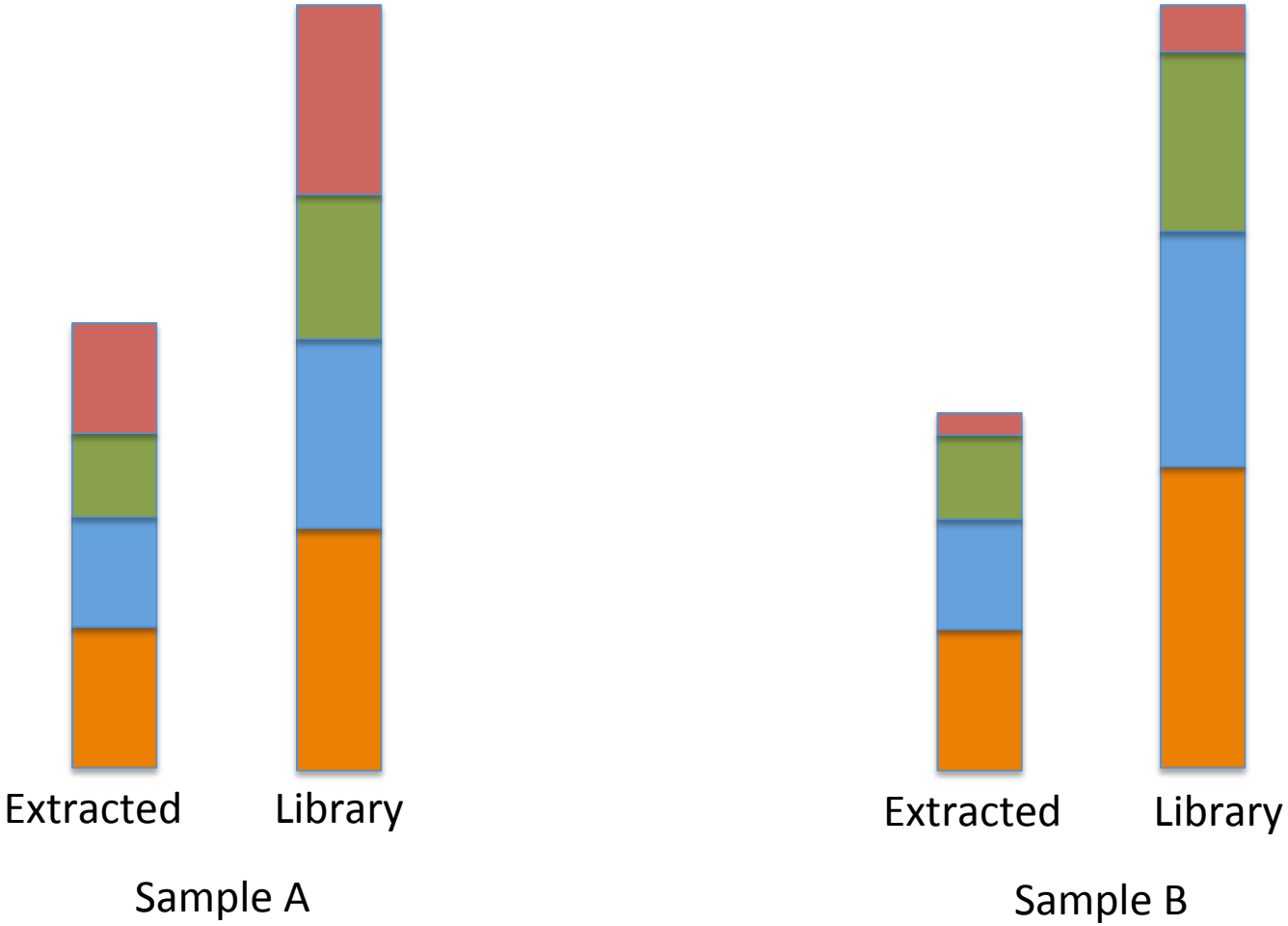




A bias associated with it's GC content has been observed in the number of counts sequenced from a fragment

image Benjamini & Speed 2012 Nucleic Acid Res.

Total RNA composition



Normalisation approaches differ between Tools

- How normalisation is applied depends on method
  - Pseudocount methods (transformed counts) scale the counts by normalisation factors and fit the model to the adjusted counts
  - Count base methods tend to introduce extra components (offsets) into the model and fit the adjusted model of the raw counts
- Which factors to normalise by and what correction method chosen make a bigger difference.

Models

"All models are wrong but some are useful”

George Box 1978

"Models should be as simple as possible, but not more so."

Attributed to Einstein

- There are a bewildering number of different proposed models for RNA-seq data
- Even more bewildering number of implementations as software packages
- Each paper claims their method “out performs” the others
  - most of these claims based on simulated data
  - simulated data require generating model assumption

Pairwise Exact Test

- tests the null hypothesis that the the proportion of counts for some gene x amongst two conditions is the same as that of the remaining genes

	Condition1	Condition2	Marginal Totals
Gene X	n_{11}	n_{12}	$n_{11} + n_{12}$
All other Genes	n_{21}	n_{22}	$n_{21} + n_{22}$
Marginal Totals	$n_{11} + n_{21}$	$n_{12} + n_{22}$	

$$\frac{n_{11}}{n_{12}} = \frac{n_{21}}{n_{22}}$$

- calculates a probability of getting a set of values (n11, n12, n21, n22)
  - assuming a binomial distribution

$$p(n_{11}, n_{12}, n_{21}, n_{22}) = \frac{\binom{n_{11}+n_{12}}{n_{11}}\binom{n_{21}+n_{22}}{n_{21}}}{\binom{N}{n_{11}+n_{21}}}$$

$$\ni \binom{k}{l} = \frac{k!}{(k-l)!l!}$$

$$\ni k! = 1 \times 2 \times 3 \times \ldots \times k$$

Fisher's Exact Test

$$p(r \geq n_{11}) = \sum_{k=n_{11}}^{n_{11}+n_{12}} \frac{\binom{k+n_{12}}{k}\binom{n_{21}+n_{22}}{n_{21}}}{\binom{N}{k+n_{21}}}$$

Generalised Linear Models

- are an group of models which include
  - general linear models
  - multiple regression
  - simple linear models
  - ANOVA

  as special cases of GLMs with limiting assumptions
- expand the available distribution model for the dependent variable to other exponential family distributions than just the normal
- use a link function to transform the linear predictor to the

	WT-1	WT-2	WT-3	Mut-1	Mut-2	Mut-3
gene1	$r_{1,1}$	$r_{1,2}$	$r_{1,3}$	$r_{1,4}$	$r_{1,5}$	$r_{1,6}$
gene2	$r_{2,1}$	$r_{2,2}$	$r_{2,3}$	$r_{2,4}$	$r_{2,5}$	$r_{2,6}$
gene3	$r_{3,1}$	$r_{3,2}$	$r_{3,3}$	$r_{3,4}$	$r_{3,5}$	$r_{3,6}$

where $r_{i,j}$ is the read count for gene i in the jth sample

Consider a single gene

	WT-1 wildtype	WT-2 wildtype	WT-3 wildtype	Mut-1 mutant	Mut-2 mutant	Mut-3 mutant
gene1	r_1	r_2	r_3	r_4	r_5	r_6

$$r_i = \alpha + \begin{bmatrix} \beta_{WT} & \beta_{Mut} \end{bmatrix} \begin{bmatrix} c_{WTi} \\ c_{Muti} \end{bmatrix} + \varepsilon_i$$

$$c_{WTi} = \begin{cases} 1 & \text{wildtype} \\ 0 & \text{mutant} \end{cases} \qquad c_{Muti} = \begin{cases} 0 & \text{wildtype} \\ 1 & \text{mutant} \end{cases}$$

Have n equations in this case 6

- an assumption about the distribution of the error term ε,
- can calculate the best fit values for α, β_{WT} and β_{Mut}
- can calculate statistic that $\beta_{WT} \neq \beta_{Mut}$
  - or **foldchange** $\beta_{Mut}/\beta_{WT} \neq 1$, $\log(\beta_{Mut}/\beta_{WT}) \neq 0$
  - it is important to have an assumption about the shape of ε.

Tools differ in

- how complicated your design can be
  - extra factors in the model other than just treatment/ condition
- their assumptions about ε
  - poisson, negative binomial etc.
- the method they use to fit the parameters
- the statistical approach they use to test for $\log(\beta_{Mut}/\beta_{WT}) \neq 0$
- generally log foldchange agree well between tools but significance values can vary considerably

Multiple Test Correction

"Littlewood's law of miracles states that in the course of any normal person's life, miracles happen at the rate of roughly one per month."

Freeman Dyson (no not the vacuum cleaner man!)

Littlewood's Law.

- we are awake eight hours each day
- we see and hear things happening at a rate of one per second.
- total number of events that happen to us is about 30,000 per day, or about a million per month.
- If a miracle is and event that happens about one per million events.
- Therefore we should expect about one miracle to happen, on the average, every month.

The multiple comparison problem

- as the number of genes compared between two conditions increases, it becomes more likely that the condition being compared will appear to differ in terms of at least one gene just by chance.
- if I set 9 H or T out off 10 tosses as indicating a coin is biased, the more coins I toss the greater the chance of getting 9 out of 10 for one of them even though it is not biased

- There are various strategies for reducing this ie. for controling the false positives in a set of multiple comparisons
- Which strategy to choose can depends on the purpose of the study.
  - If preventing getting false positives is important choose a conservative correction
  - if preventing getting false negitives is important choose a more permissive correction
- If you are on a fishing expidition looking for candidate genes you should control for multiple comparisons

"This model will be a simplification and an idealization, and consequently a falsification. It is hoped that the features retained for discussion are those of greatest importance in the present state of knowledge"

Alan Turing 1952