Visualizing distributions

HUGEN 2073

Genomic Data Visualization and Integration

Slides borrowed/adapted from H. J. Park with permission

Topics

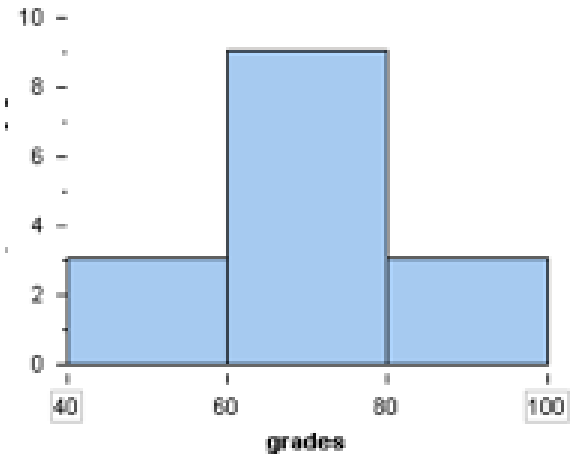
Visualizing

- Distributions of single variable
- Distributions of multiple variables
- Statistical properties underlying a distribution

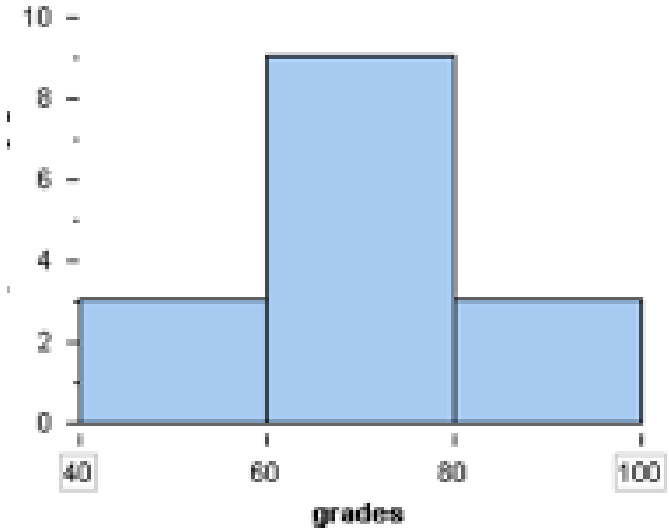
Histograms for the distribution of 1 variable

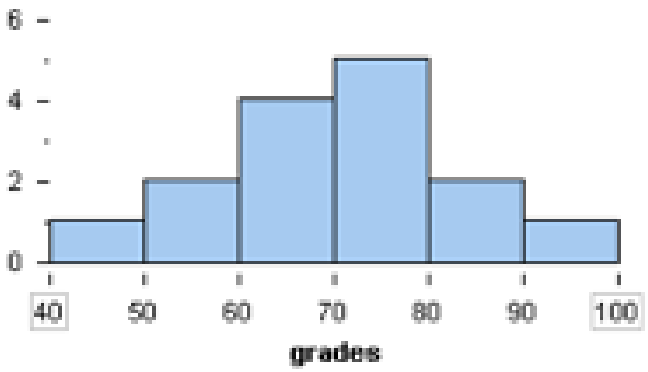
Name	Grade
Adams	10
Baker	2
Clark	34
Davis	24
Evans	45
Frank	6
Ghosh	53
Hills	81
Irwin	21
Jones	52
Klein	36
Lopez	15
….	…

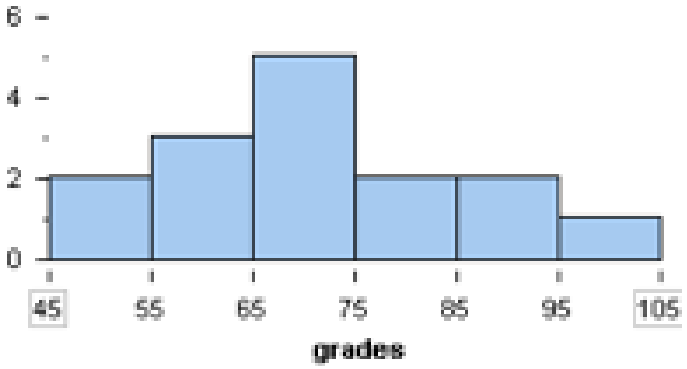
- X-axis: bins (of a continuous variable)
- Y-axis: # of items in the bin



Bin width is the most important parameter



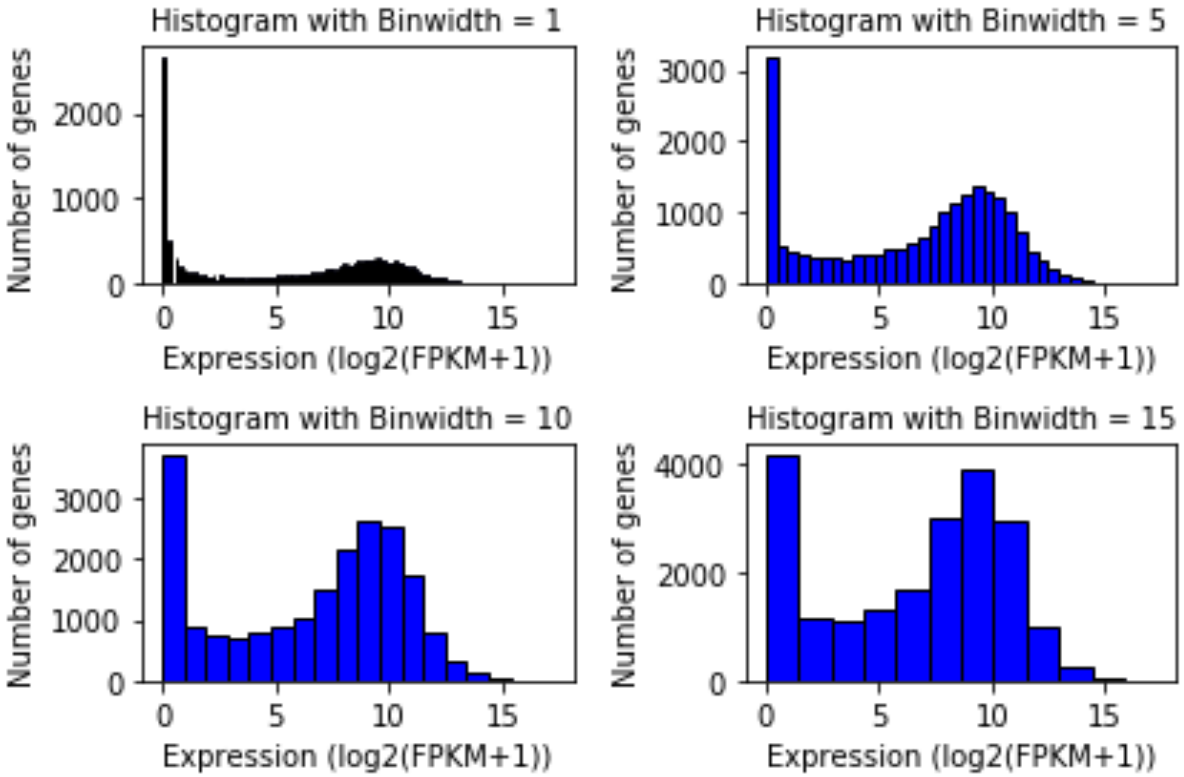




What bin width (and position) would be best here?

A “good” bin width may give a more direct interpretation

Bin width is the most important parameter



Informative bin width

- Gives direct interpretation
- May conform to context-specific prior knowledge (e.g. the ratio of lowly expressed genes to other genes)
- Default is not necessarily best!

Distribution of a single variable in a table

E.g., expression (FPKM) for 20,530 genes in the breast tumor sample TCGA-BH-A0C7-01

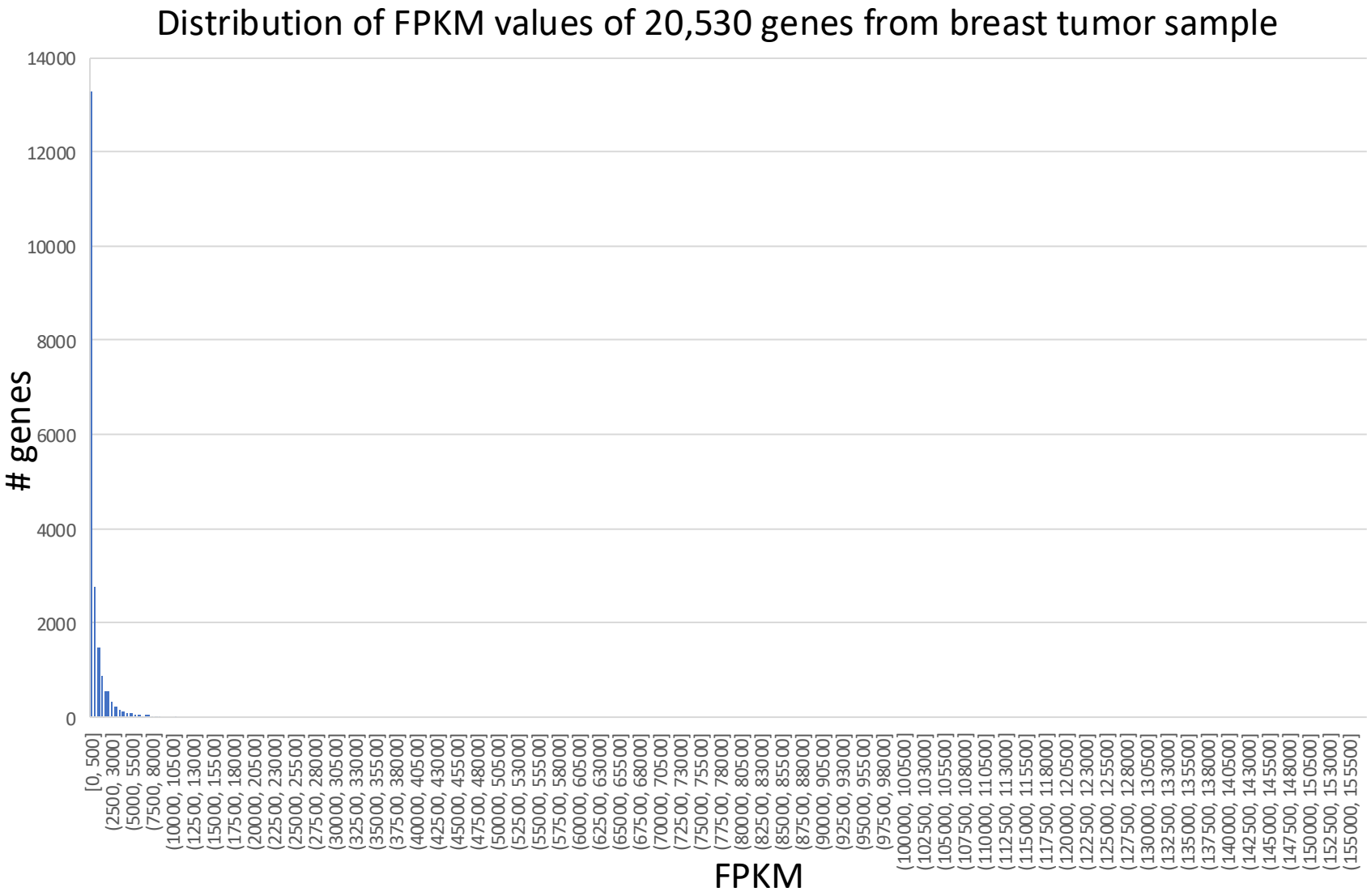
Gene/sample	TCGA-BH-A0C7-01
ARHGEF10L	10.0931
HIF3A	3.609
RNF17	0
RNF10	11.382
RNF11	10.7663
RNF13	10.3366
GTF2IP1	12.724
REM1	2.2163
MTVR2	0.334
RTN4RL2	5.091
C16orf13	8.6721
C16orf11	0
FGFR1OP2	8.6967
TSKS	0.605
ATRX	10.7342
PMM2	9.7939
LOC100272146	5.2176
ASS1	8.1664
NCBP1	10.8287
ZNF709	7.5449
ZNF708	7.1908
RBM14	9.9606
NCBP2	10.9786
DISC1	6.3806
CAMK1	8.0451
RPL37	12.619
……	

×20,530

Histogram of the same expression data

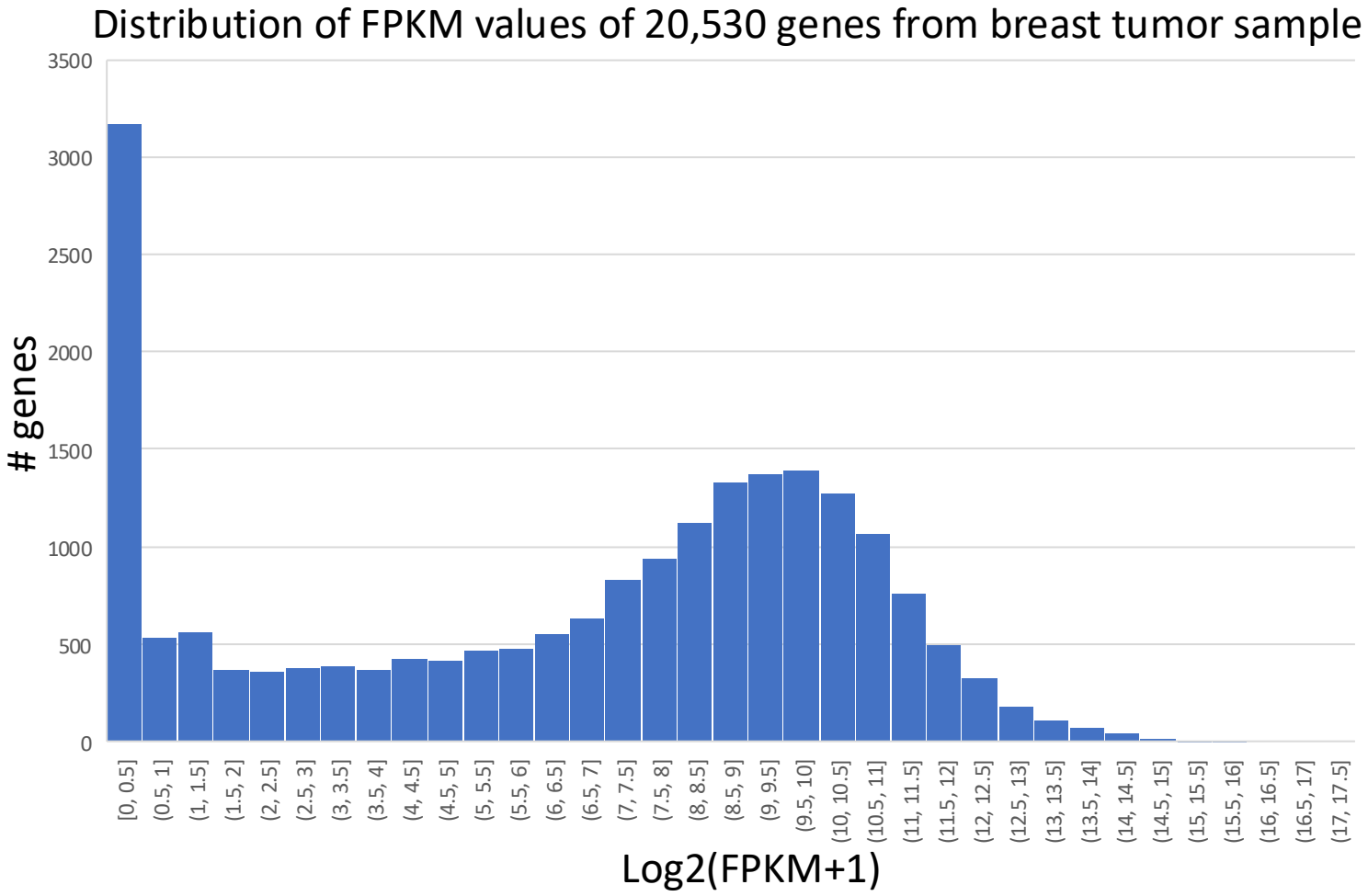
Gene/sample	TCGA-BH-A0C7-01
ARHGEF10L	10.0931
HIF3A	3.609
RNF17	0
RNF10	11.382
RNF11	10.7663
RNF13	10.3366
GTF2IP1	12.724
REM1	2.2163
MTVR2	0.334
RTN4RL2	5.091
C16orf13	8.6721
C16orf11	0
FGFR1OP2	8.6967
TSKS	0.605
ATRX	10.7342
PMM2	9.7939
LOC100272146	5.2176
ASS1	8.1664
NCBP1	10.8287
ZNF709	7.5449
ZNF708	7.1908
RBM14	9.9606
NCBP2	10.9786
DISC1	6.3806
CAMK1	8.0451
RPL37	12.619
……	

×20,530



Histogram on logarithmic scale to visualize non-Gaussian distribution

- Nonlinear x-axis of log2(FPKM+1)
- Why +1?



Making a histogram

In ggplot

geom_histogram has several ways you can control the bins

- binwidth = width of bins (can get fancy here)
- bins = number of bins
- breaks = a vector giving the bin boundaries
- position (how to arrange the bars of multiple overlaid histograms)

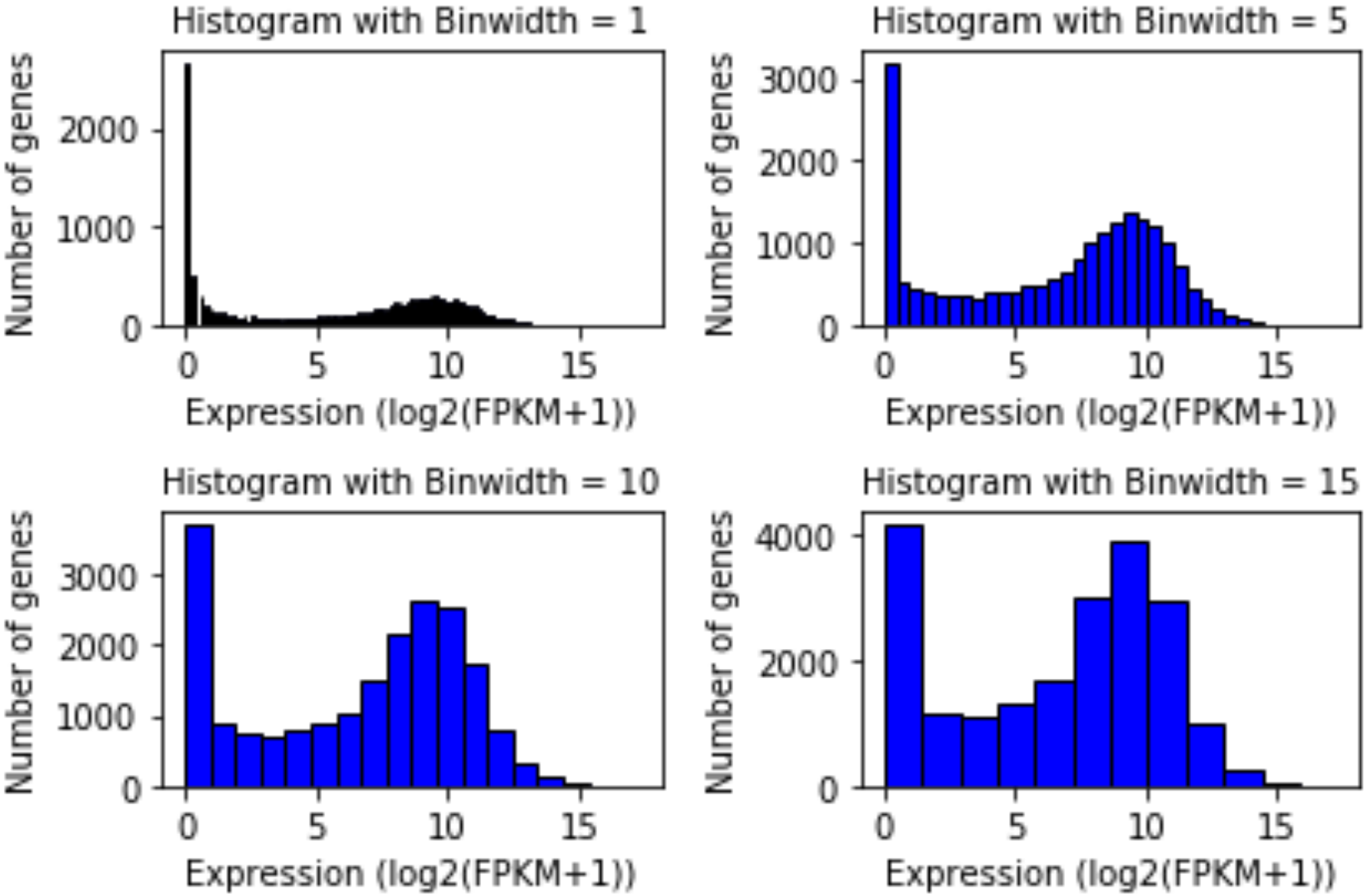
aesthetics are like those for geom_bar, including

- x or y
- fill
- color

Choosing a bin width

With wider bins…

- broader patterns
- diluted noise
- lose patterns in small intervals



Bin width *can* be determined based on formulas

Scott's normal reference rule

$$h = \frac{3.5\hat{\sigma}}{\sqrt[3]{n}}$$

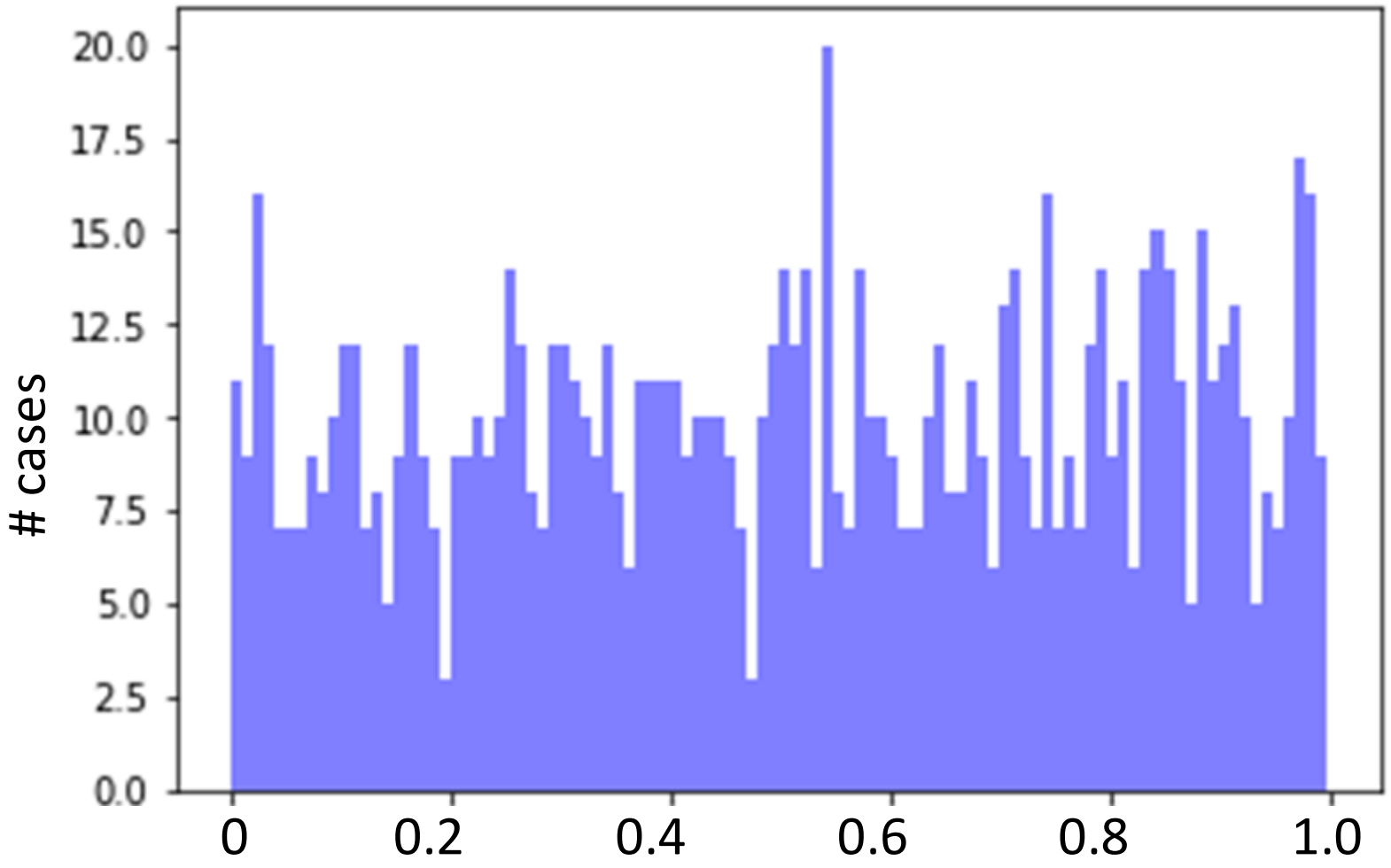
Freedman-Diaconis' choice

$$h = 2\frac{\mathrm{IQR}(x)}{\sqrt[3]{n}}$$

- We want to use too wide a bin width (i.e., less sensitive a bin width)
- You don't have to remember these, because they are provided (e. g., n.bins(x, rule=c("freedman.diaconis", "sturges", "scott", "simple") in R)
- I prefer to test and see by myself to make it informative

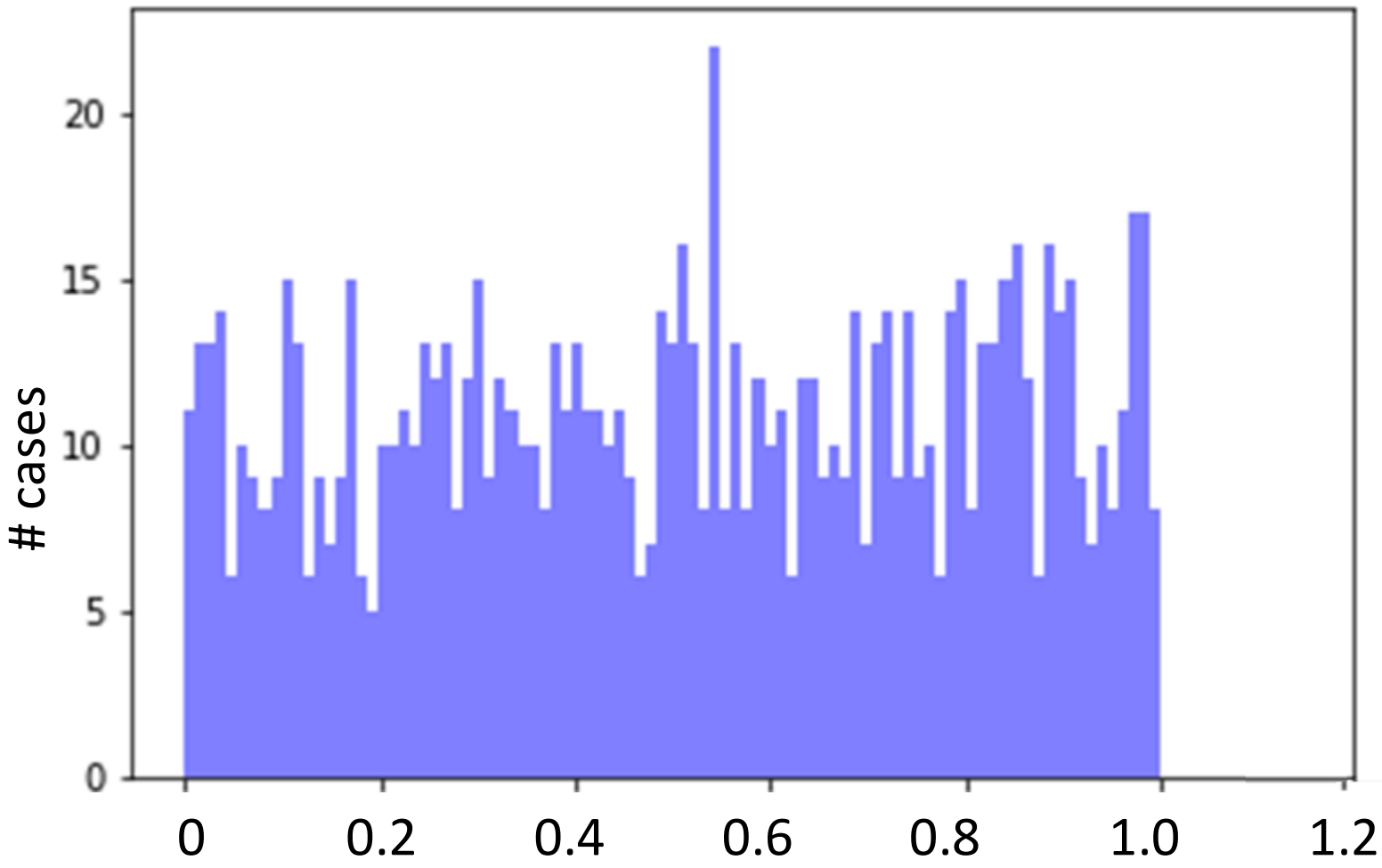
Axes should be appropriate

- E. g., this system produces values between 0 and 1



Axes should conform to the range of the variable

- This system apparently produces (could produce?) values between 0 and 1.2
- So are the values between 1 and 1.2 missing?



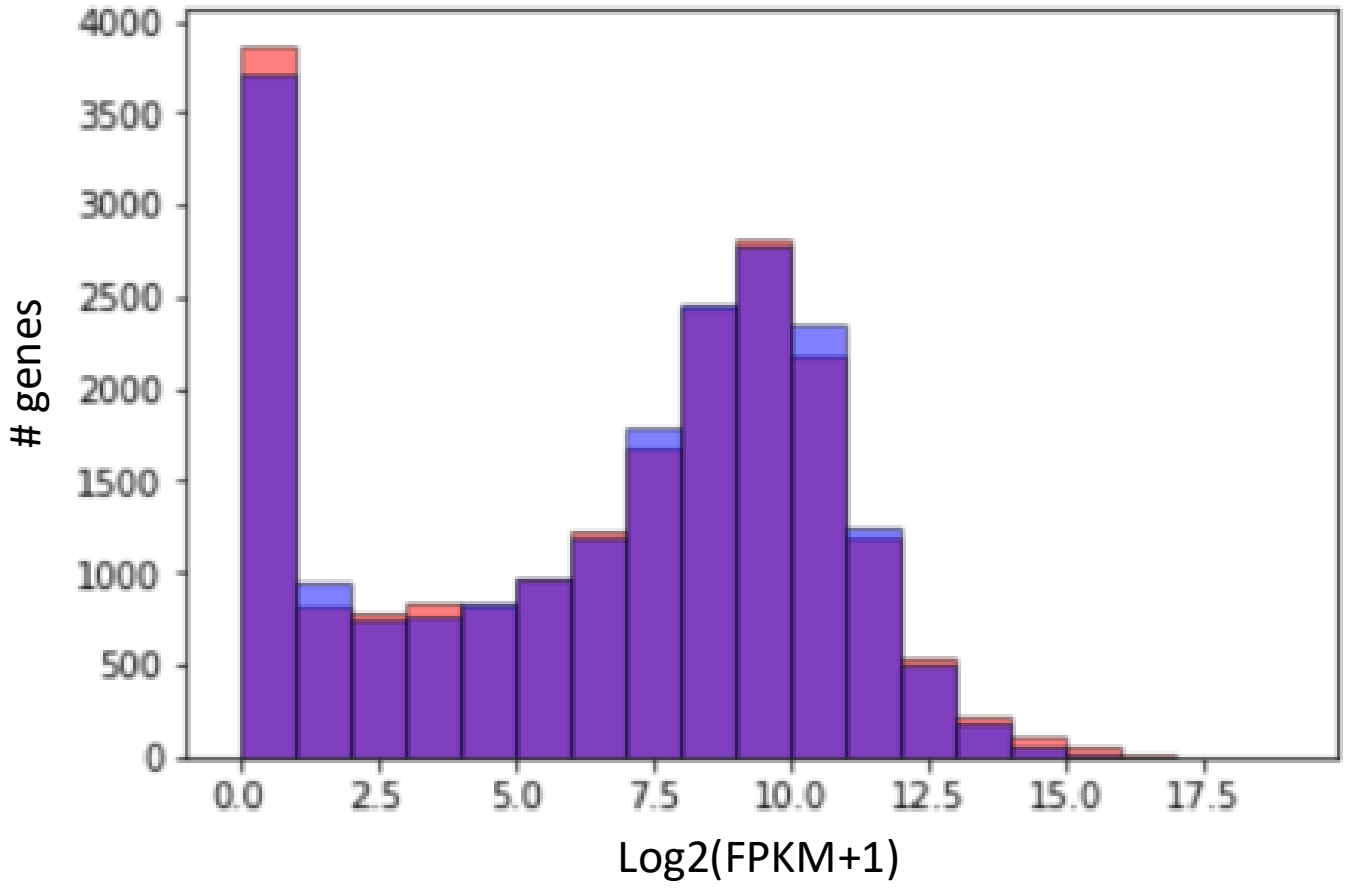
The distribution of 2+ variables

Expression (FPKM) for 20,530 expressed genes in two breast tumor samples (TCGA-BH-A0C7-01, TCGA-AR-A5QQ-01, …)

sample	TCGA-AR-A5QQ-01	TCGA-BH-A0C7-01	……
ARHGEF10L	9.5074	10.0931	
HIF3A	1.5787	3.609	
RNF17	0	0	
RNF10	11.3676	11.382	
RNF11	11.1292	10.7663	
RNF13	9.9722	10.3366	
GTF2IP1	11.5966	12.724	
REM1	3.2396	2.2163	
MTVR2	0	0.334	
RTN4RL2	3.5764	5.091	
C16orf13	9.794	8.6721	
C16orf11	0.5819	0	
FGFR1OP2	8.844	8.6967	
TSKS	1.993	0.605	
ATRX	9.4263	10.7342	
PMM2	10.3735	9.7939	
LOC100272146	3.0792	5.2176	
ASS1	9.5279	8.1664	
NCBP1	10.3001	10.8287	
ZNF709	5.2763	7.5449	
ZNF708	6.3037	7.1908	
RBM14	10.3336	9.9606	
NCBP2	10.8729	10.9786	
DISC1	6.0783	6.3806	
CAMK1	8.6979	8.0451	
	……		

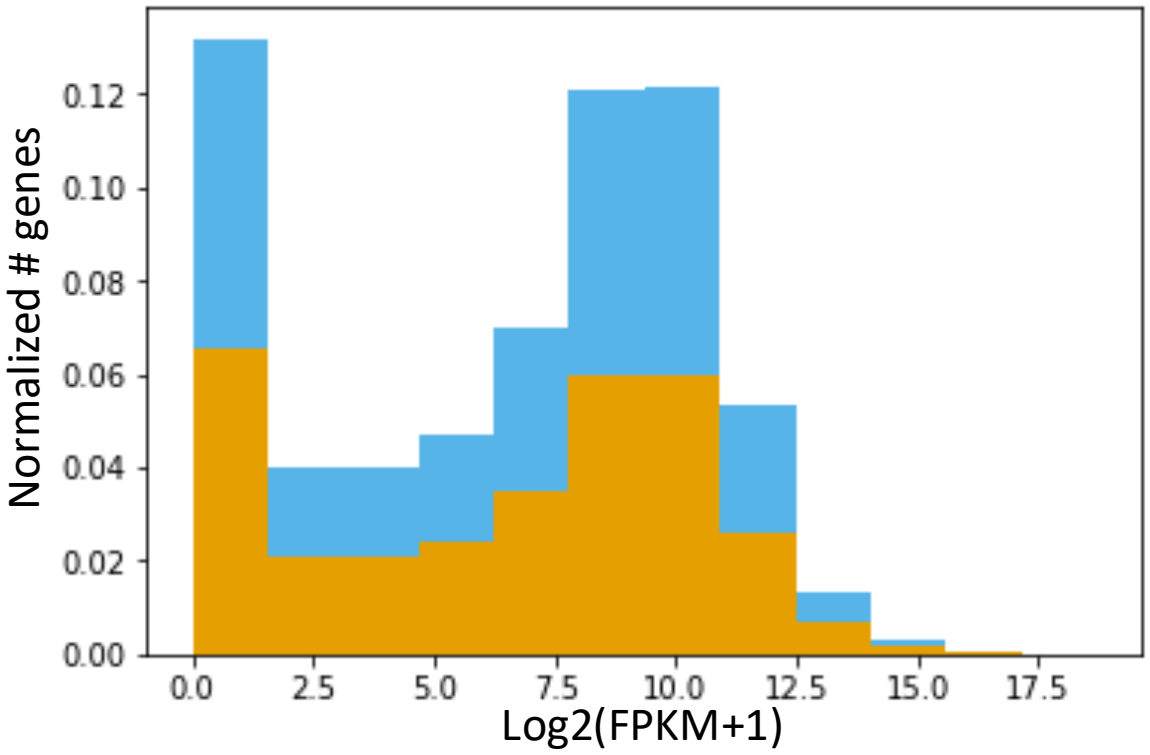
Overlapped histogram for two variables?

- Need to set
  - the same bin width starting at the same x-axis points
  - transparency
- Problems
  - Comparisons difficult?
  - 3 colors now
  - Looks stacked



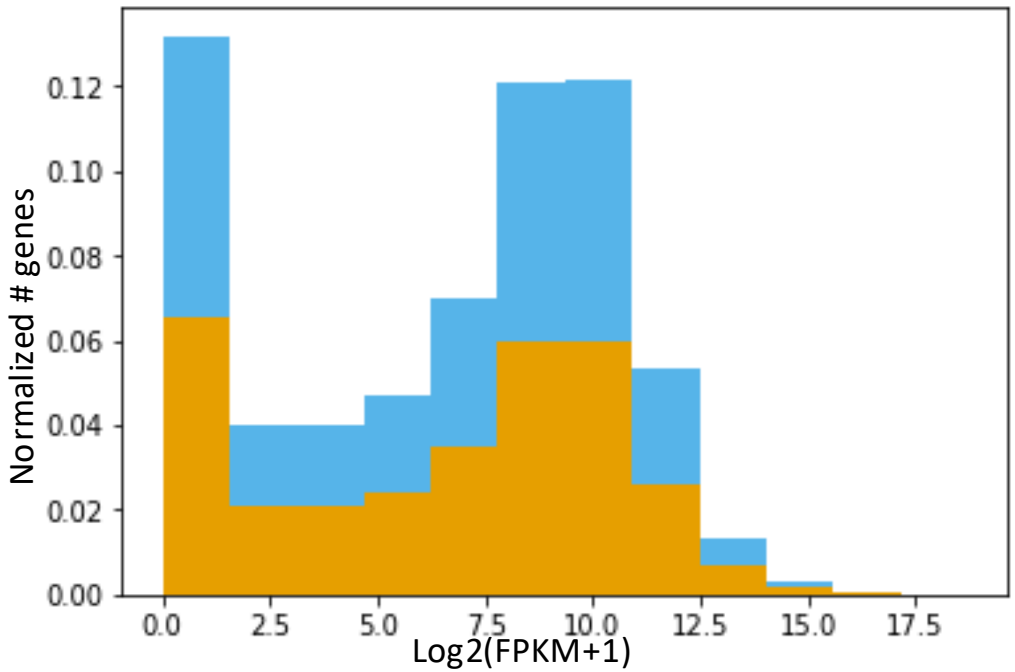
Stacked histogram for two variables?

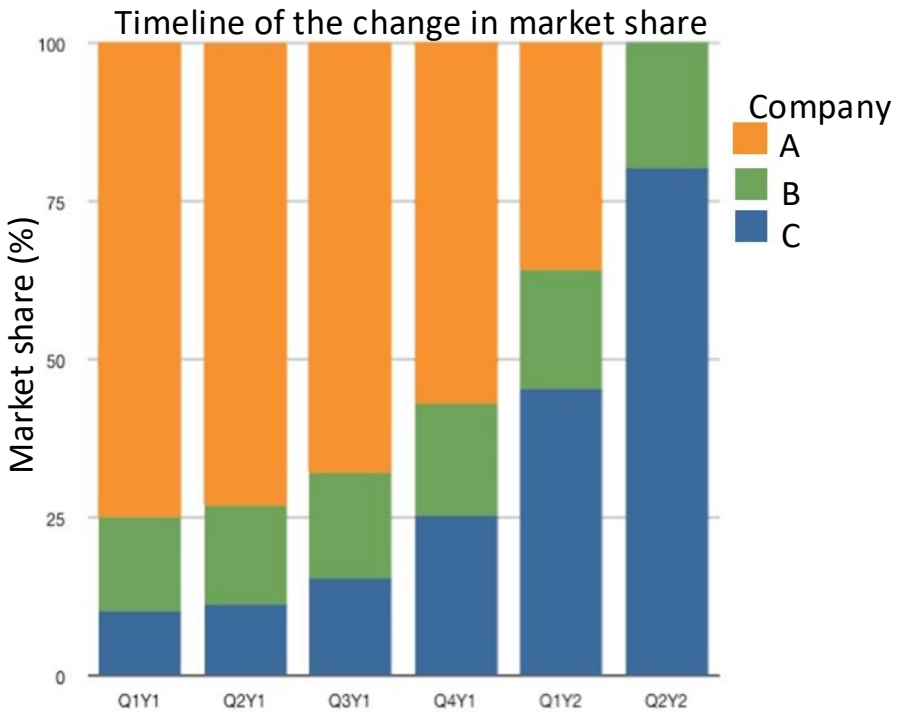
- As a *proportion* of the whole for each bin → counts are not shown!
- Still not easy to compare between colors
- Confusing – not clear whether these are *stacked* or *overlapped*!



Stacked histogram?

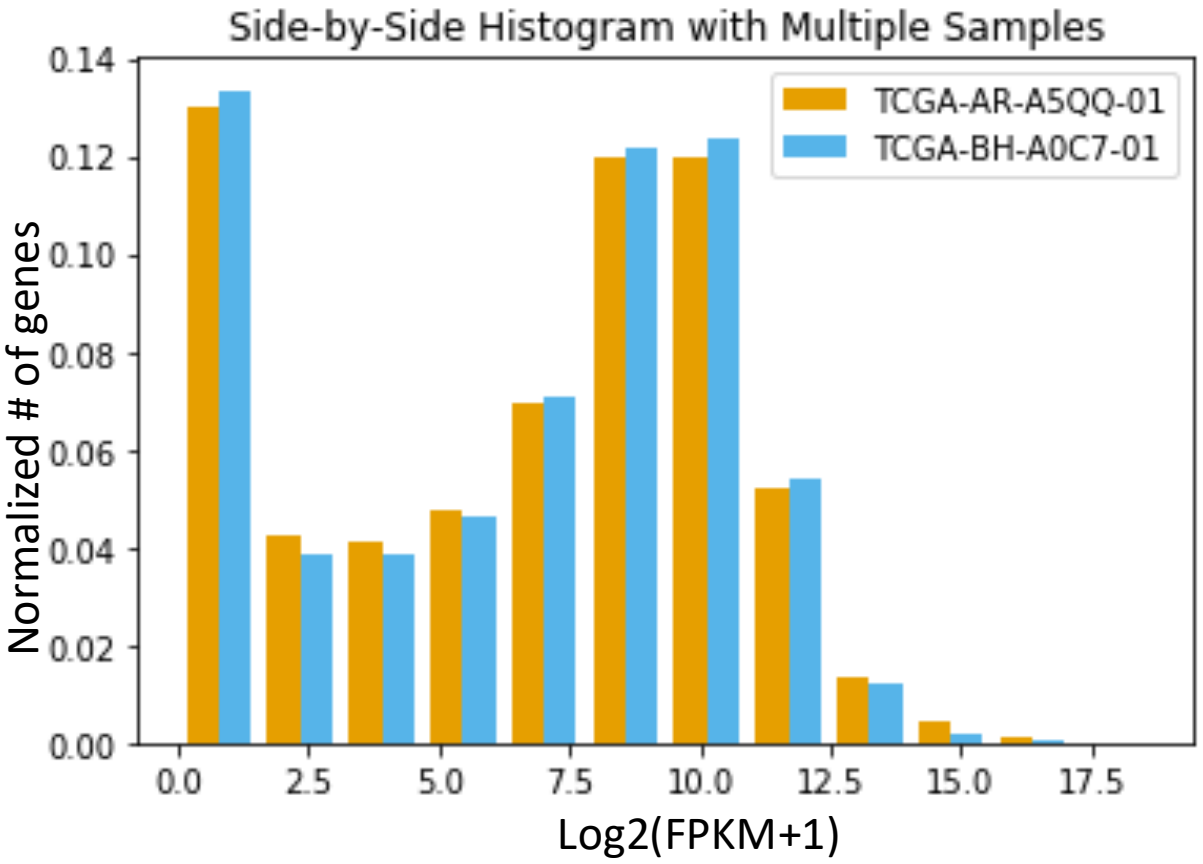
- Sample as a proportion of the whole for each bin for comparison
- <u>Maybe</u> OK when the total amount is fixed
- Still has the problems of a stacked bar plot/area plot





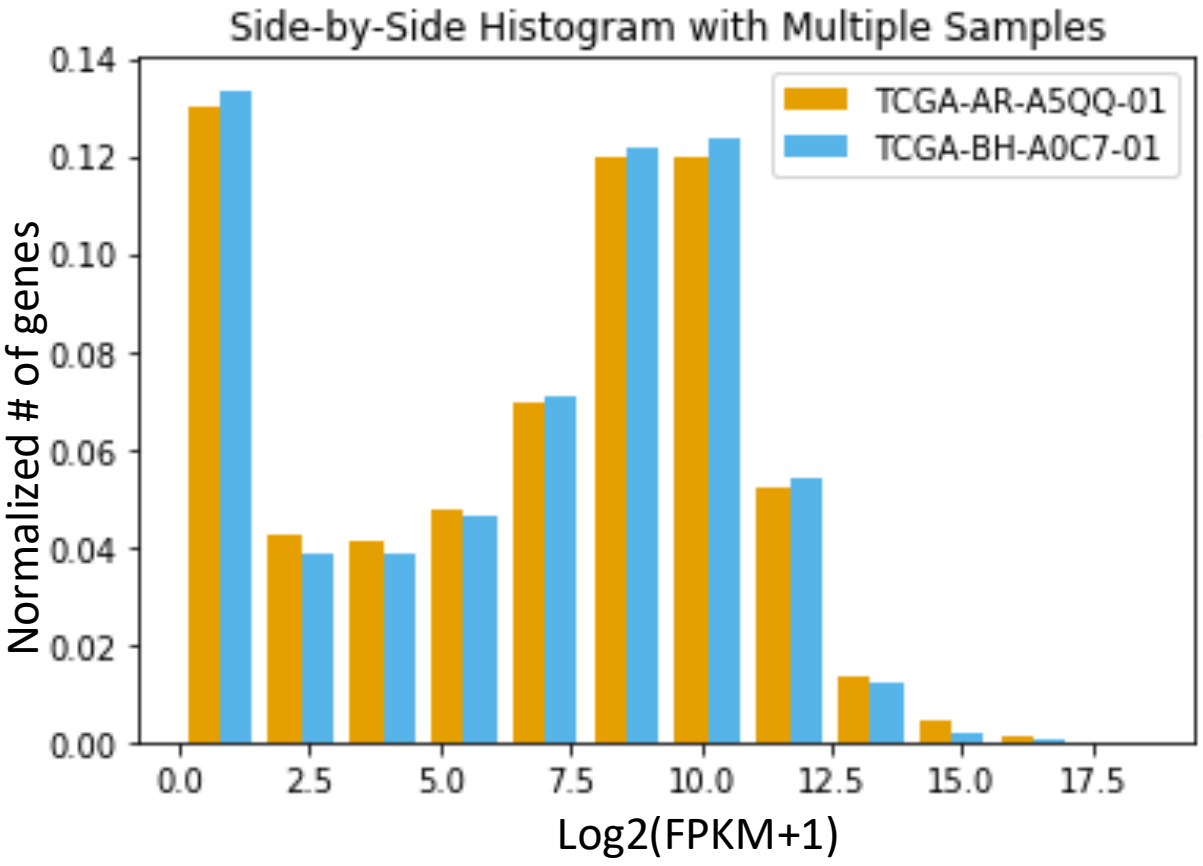
http://www.leancrew.com/all-this/2011/11/i-hate-stacked-area-charts/

Side-by-side histograms!

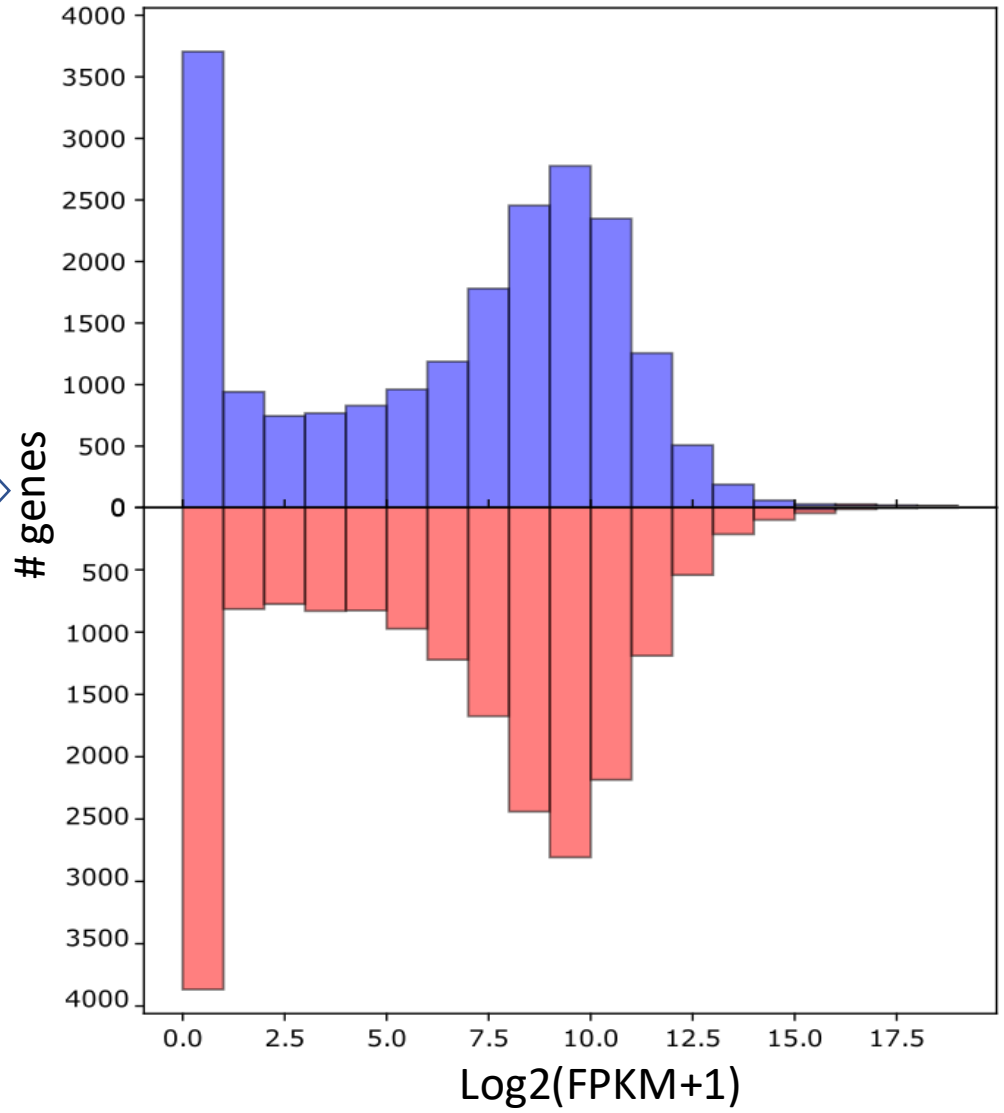


- In ggplot2, use position="dodge"

Another type of histogram for two distributions?

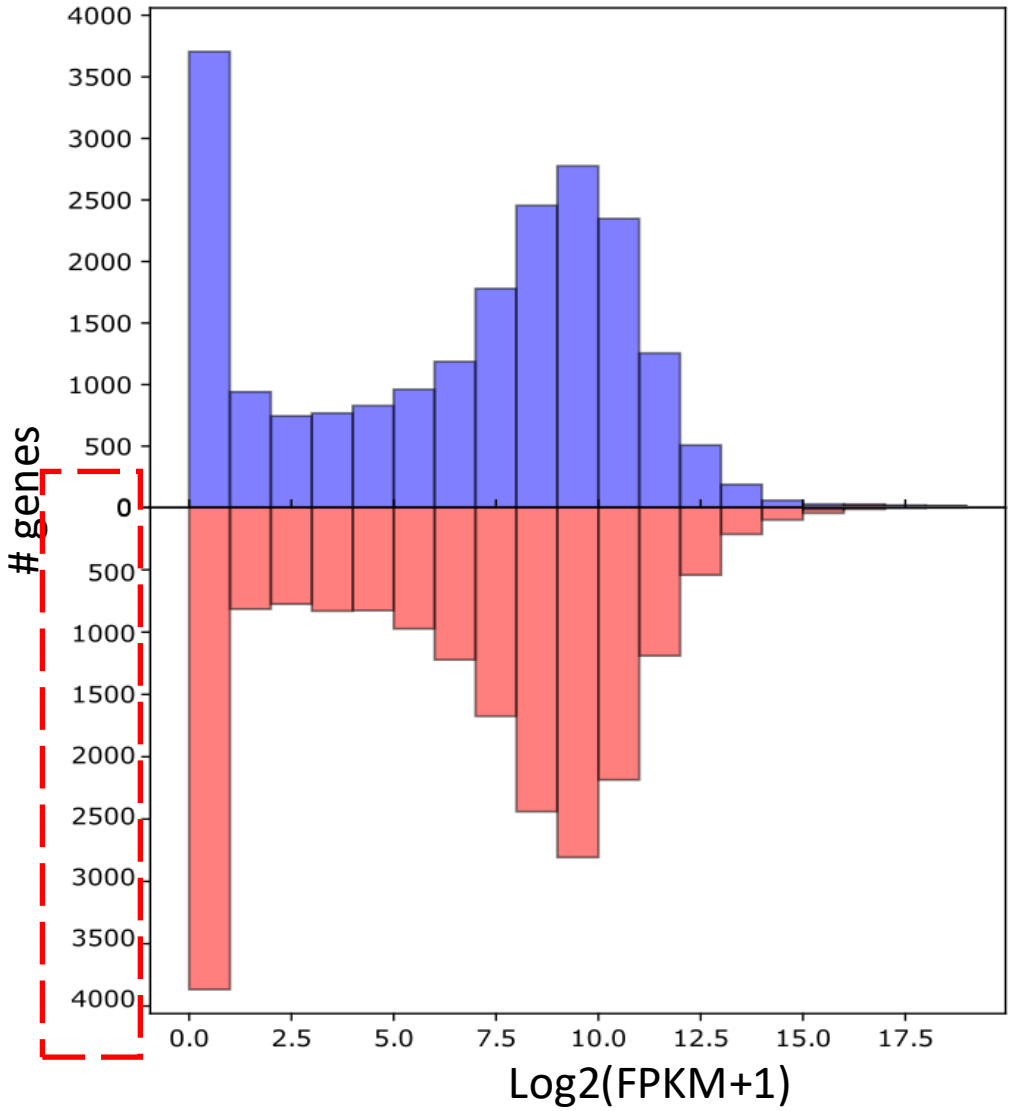


?

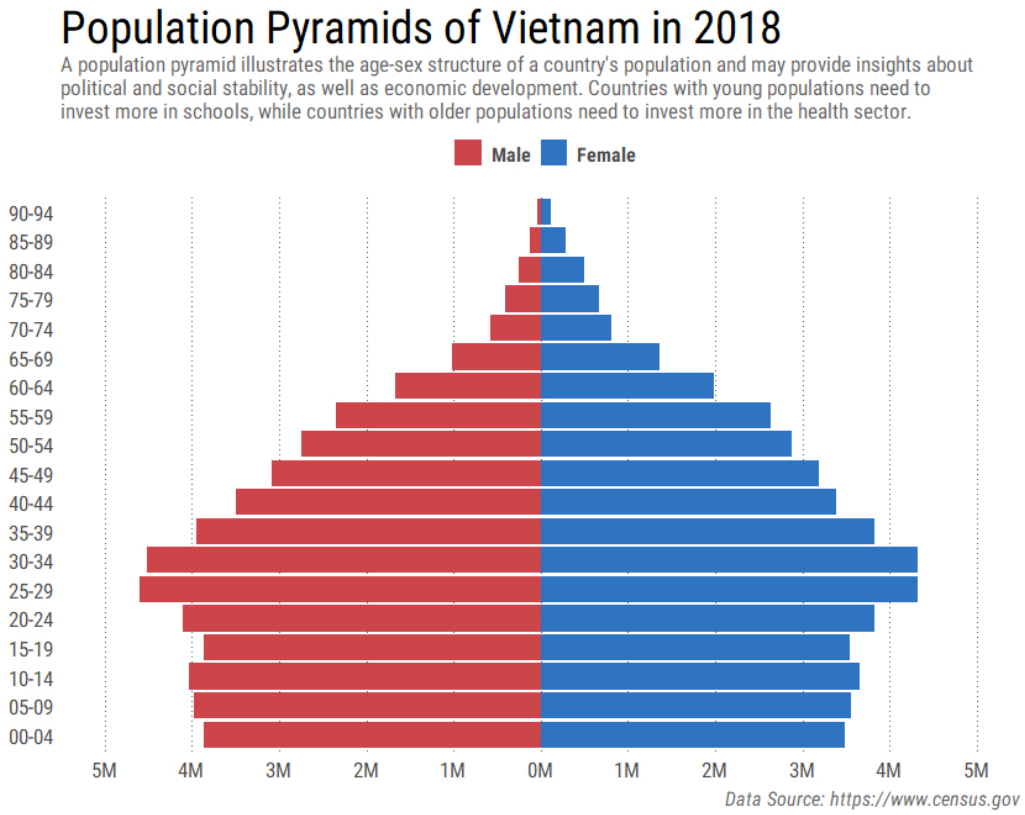


A new type of histogram to visualize distribution of multiple variables

- Y-axis advances both directions
- Can catch and compare the trend
- Difficult to compare within each bin
- Need to be careful to set the negative sign of y-axis right



Similar: age pyramid plots



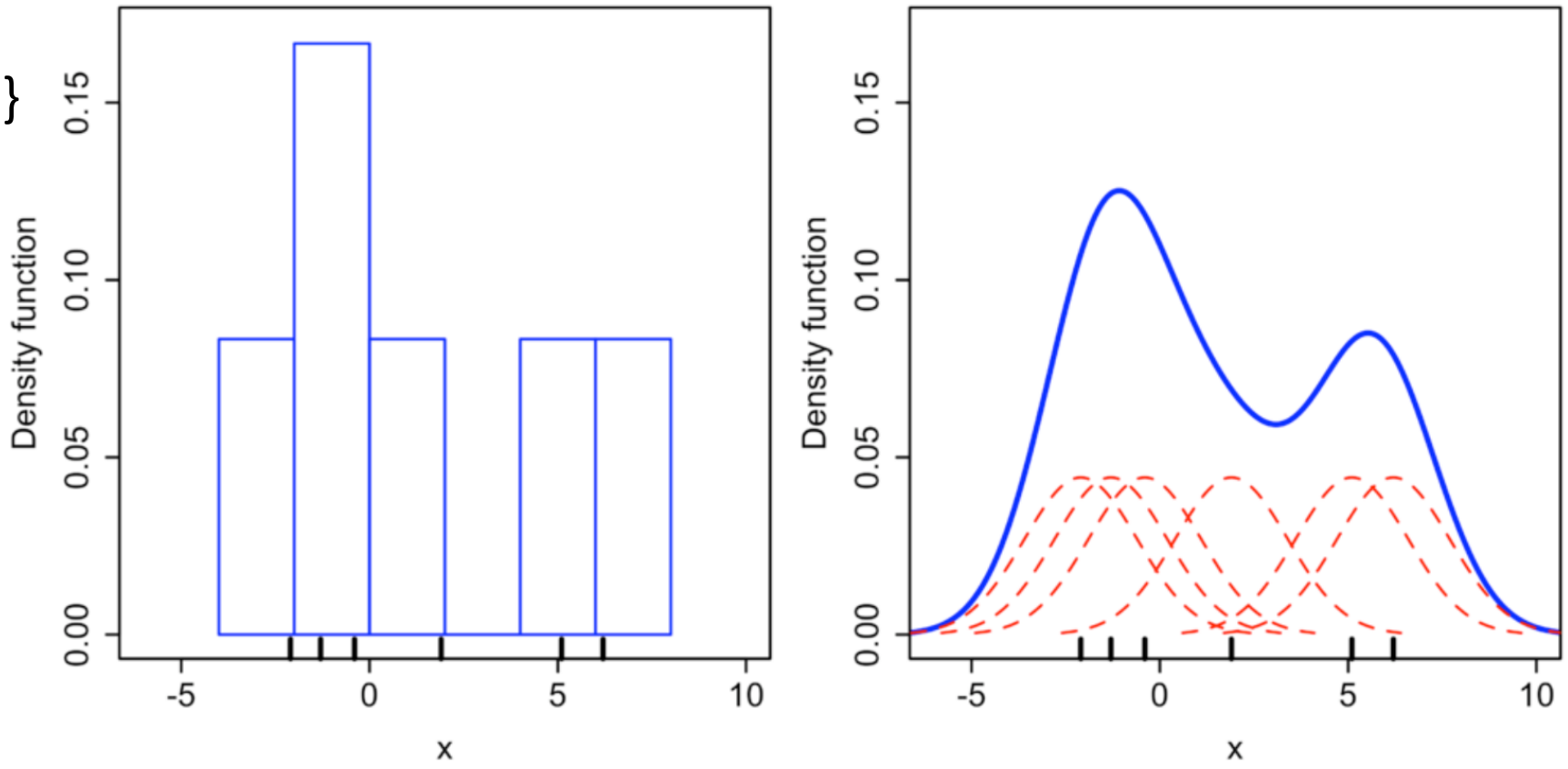
https://rpubs.com/chidungkt/505486

Density plots for visualizing probability distributions

E. g., X-axis values=
{-2.05, -1.5, -0.1, 1.93, 5.2, 6.02}
Y-axis

- The count in histogram
- The probability density function different from a probability

Histogram vs. Density plot

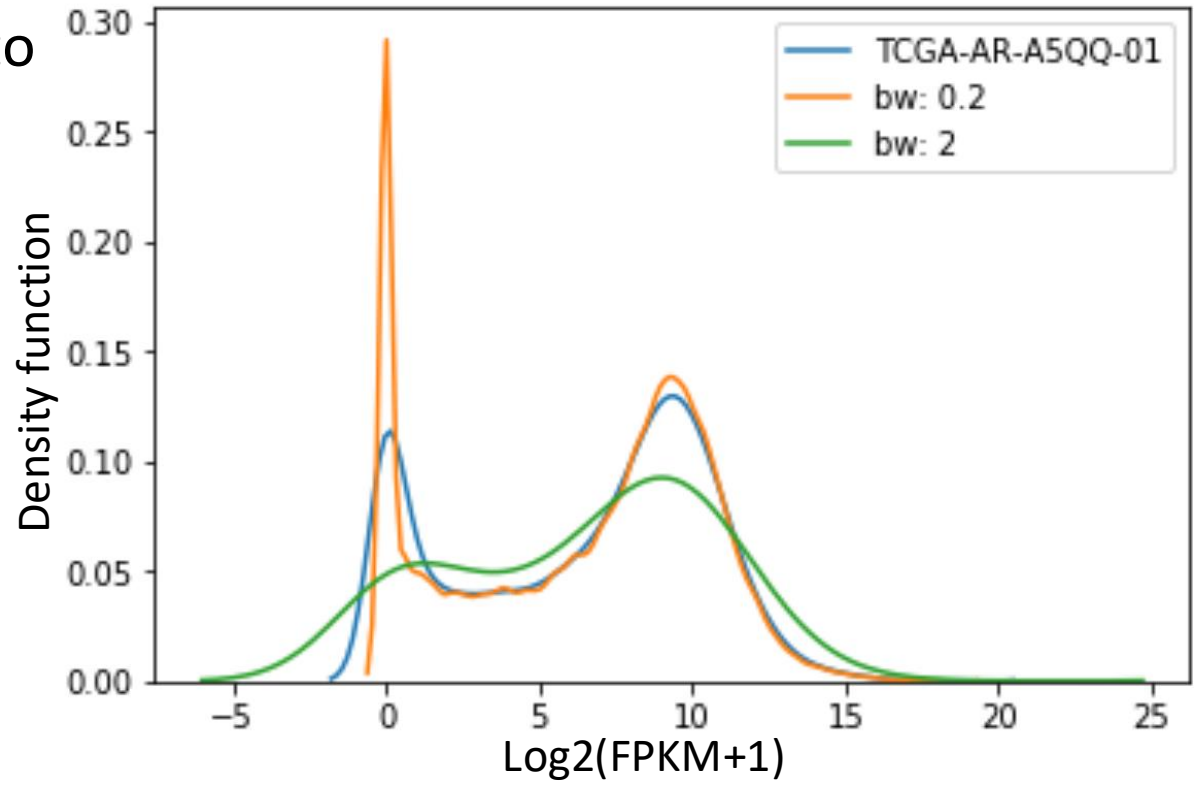


Bandwidth in density plot

Equivalent to bin width in a histogram, we need to find a good <u>bandwidth</u> in a density plot

ggplot2::geom_density

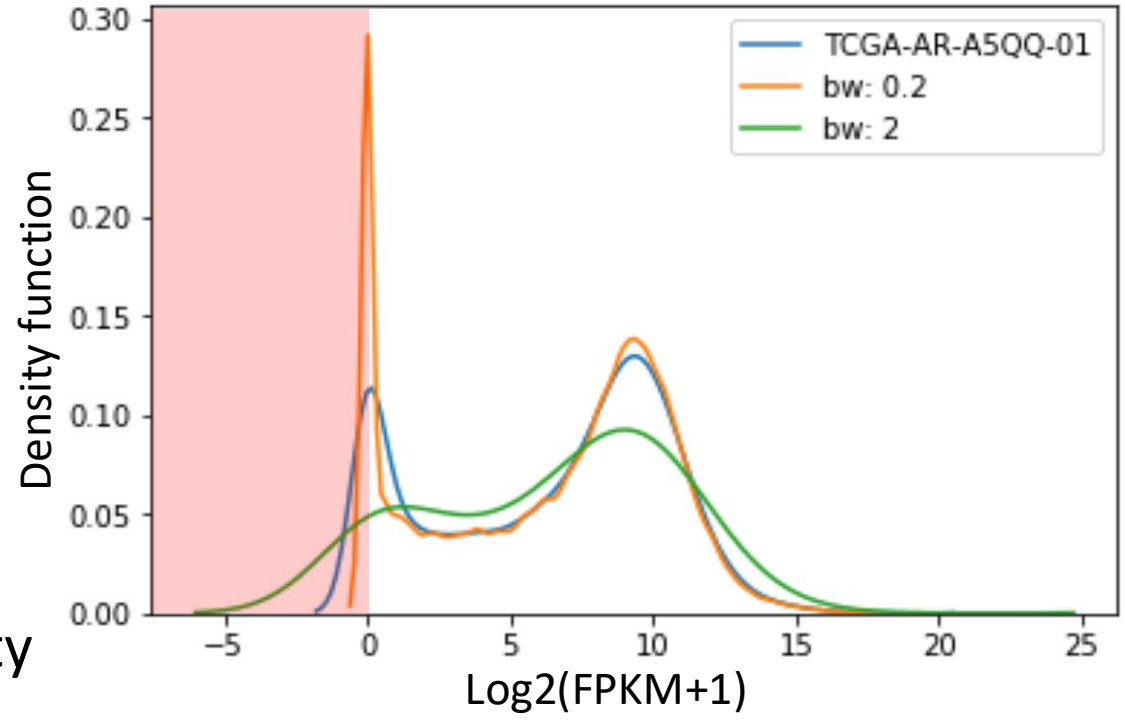
- Similar aesthetics to geom_histogram
- Essential parameters
  - kernel (default = gaussian)
  - bw = bandwidth = sd of kernel
  - bounds = upper/lower bounds for the estimated data



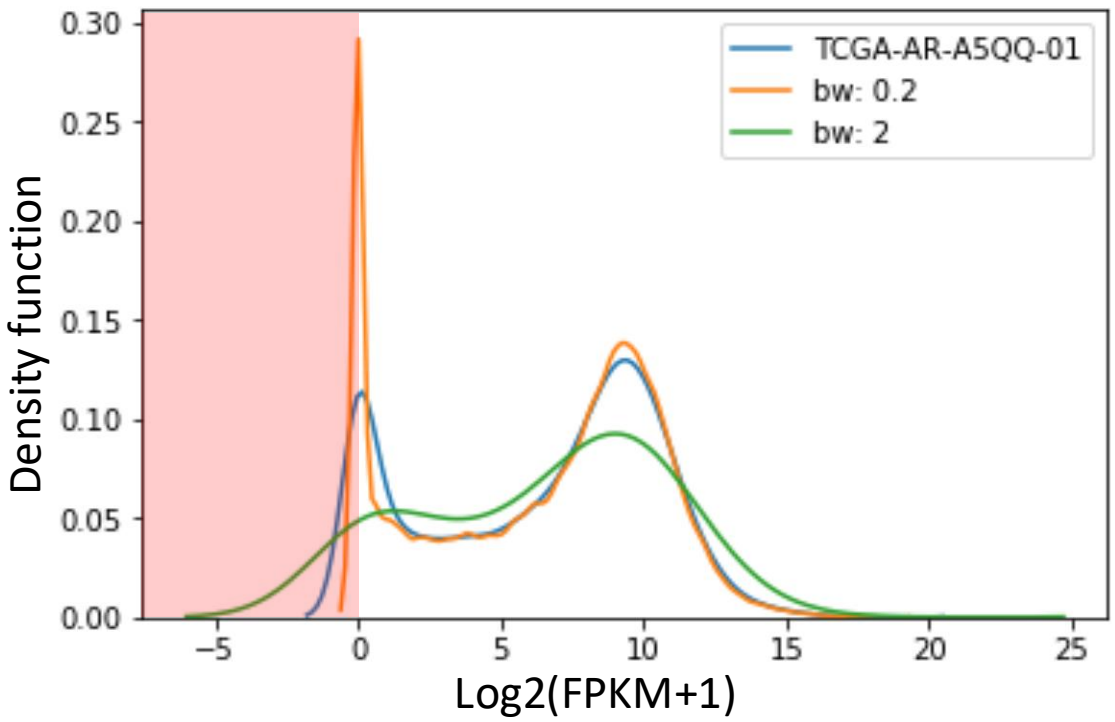
Density plots may yield unreasonable tails

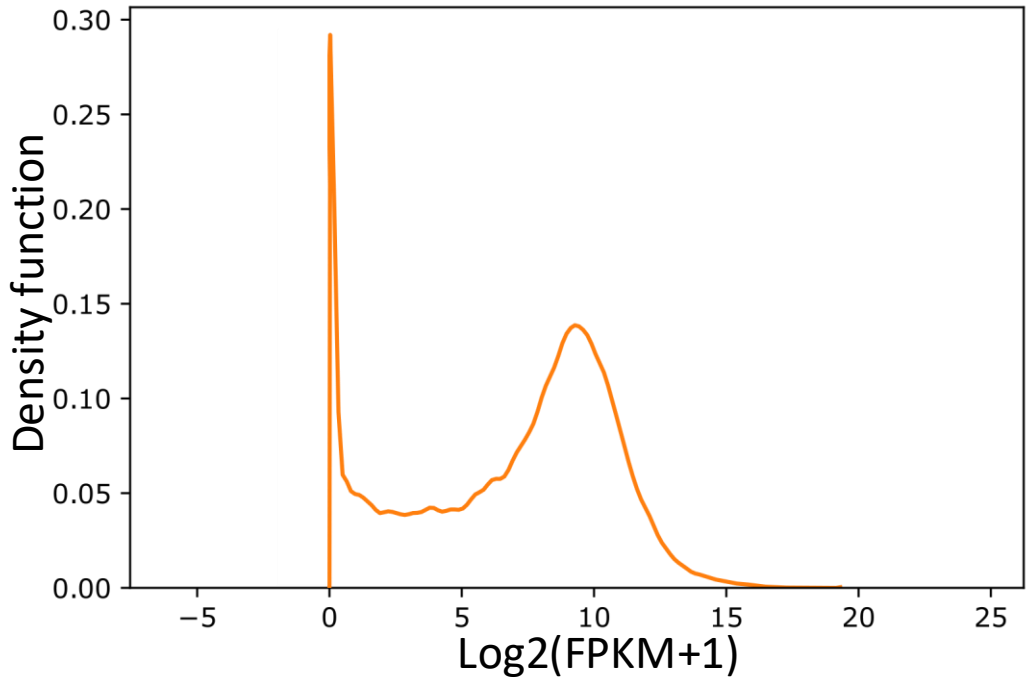
Equivalent to bin width in a histogram, we need to

- Find a good bandwidth in a density plot
- Be careful of a tendency to produce the appearance of data when none exists
  - especially in the tails
  - can lead to figures that make nonsensical statements
  - use the “bounds” argument of geom_density



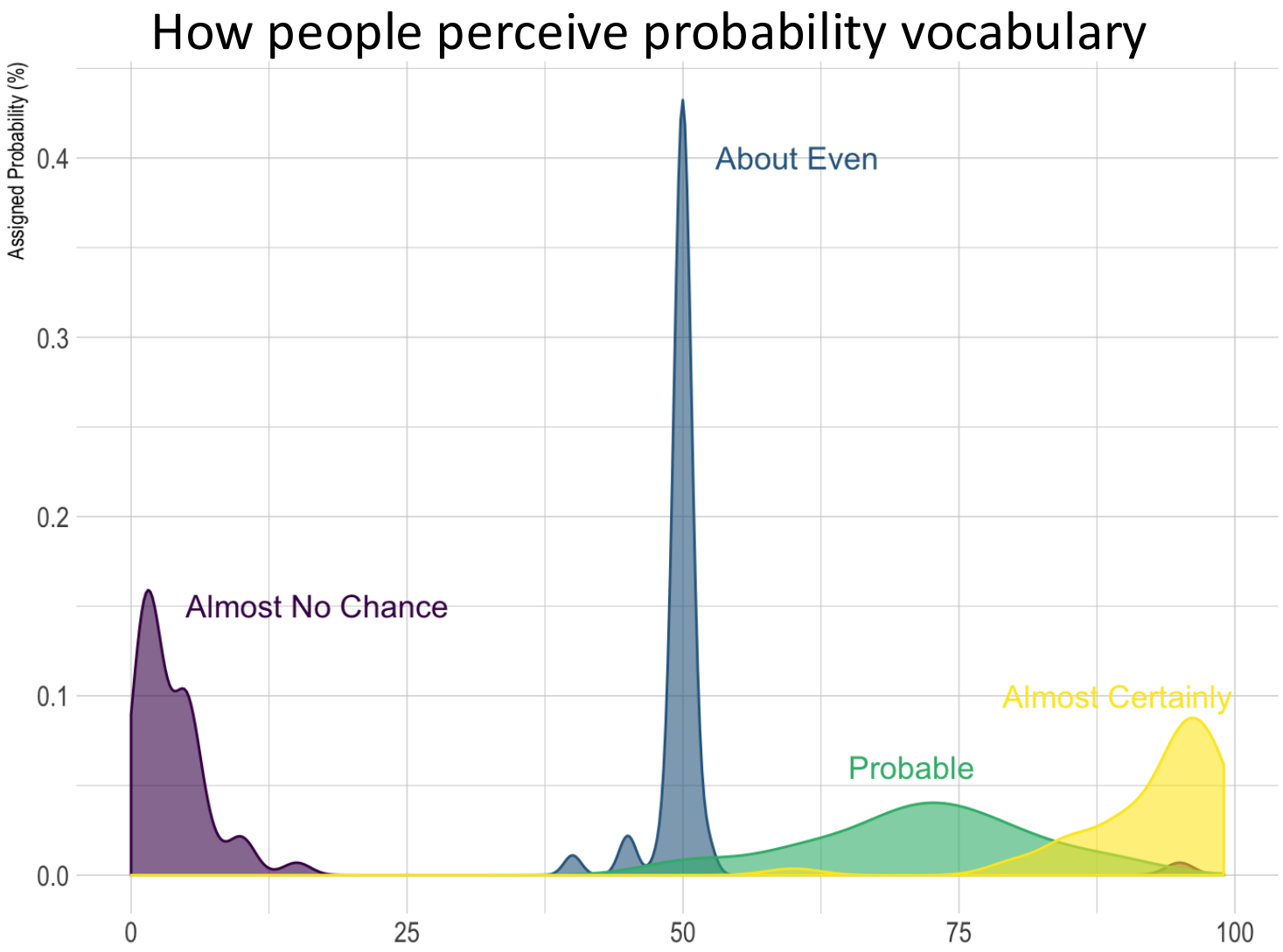
Density plots may yield unreasonable tails





Don’t just use an editor to remove the unreasonable tail.

Density plot *without* unreasonable tails



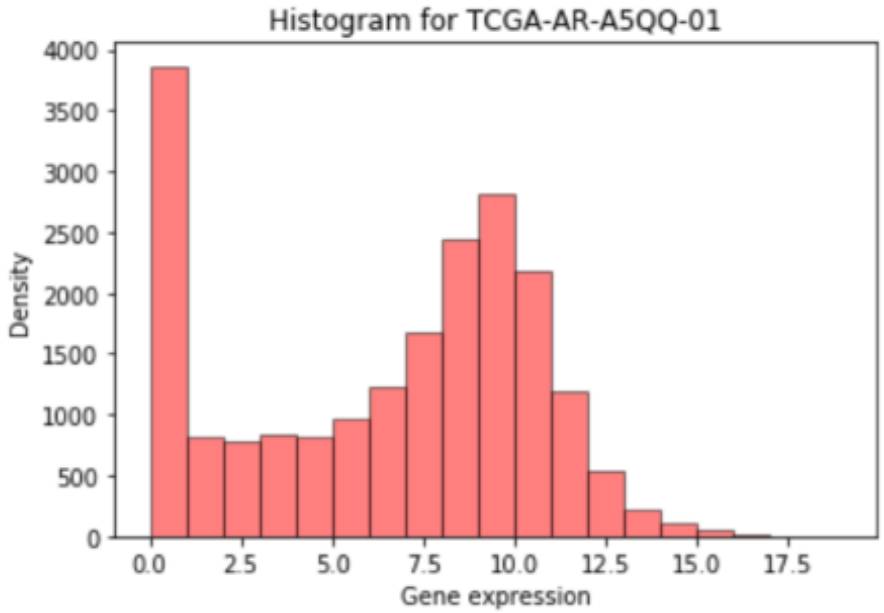
Fundamentals of Data Visualization, Wilke, O'Reilly, 1st Ed.

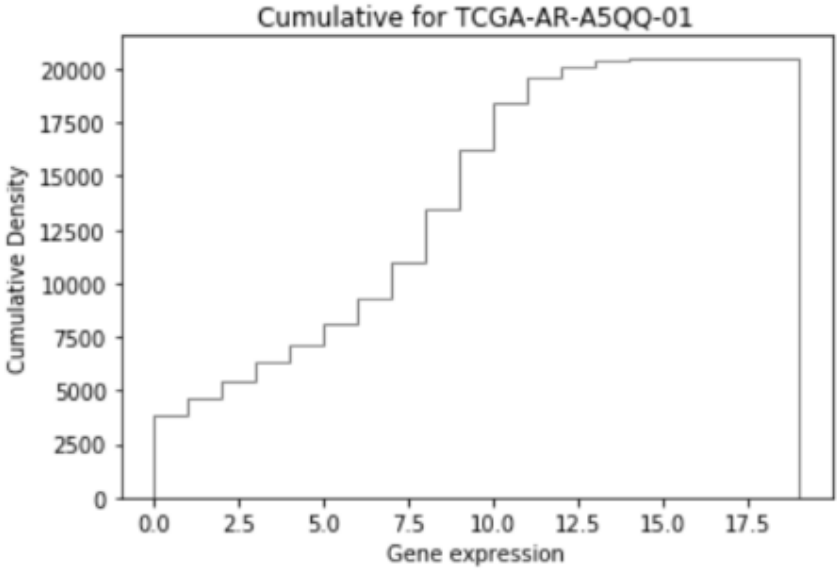
Visualizing *cumulative* distributions

The CDF is the function that says how much of the observed data (or pdf) is less than each given value.

F(x) = P[X <= x] for each possible value x.
(This is basically the integral of the histogram or pdf.)

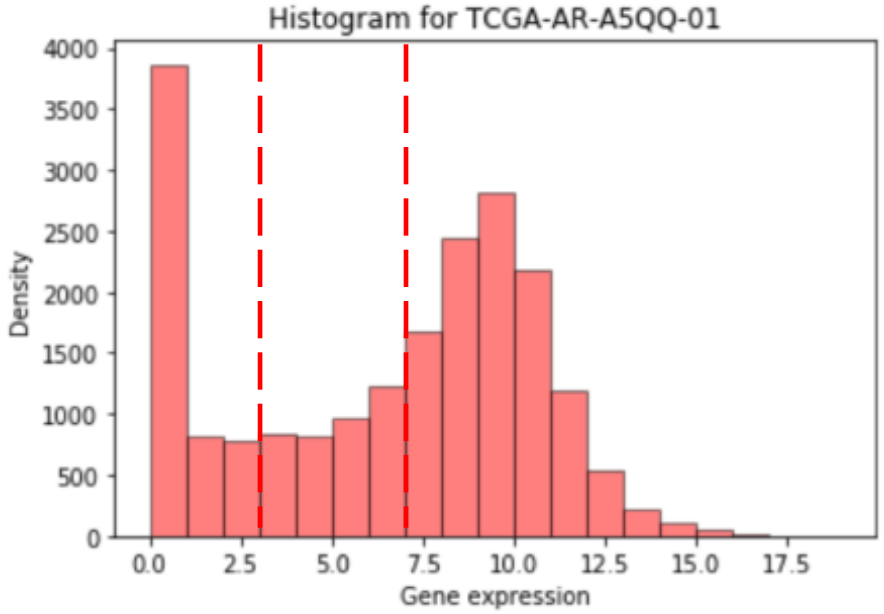
In R, use the ecdf() function.

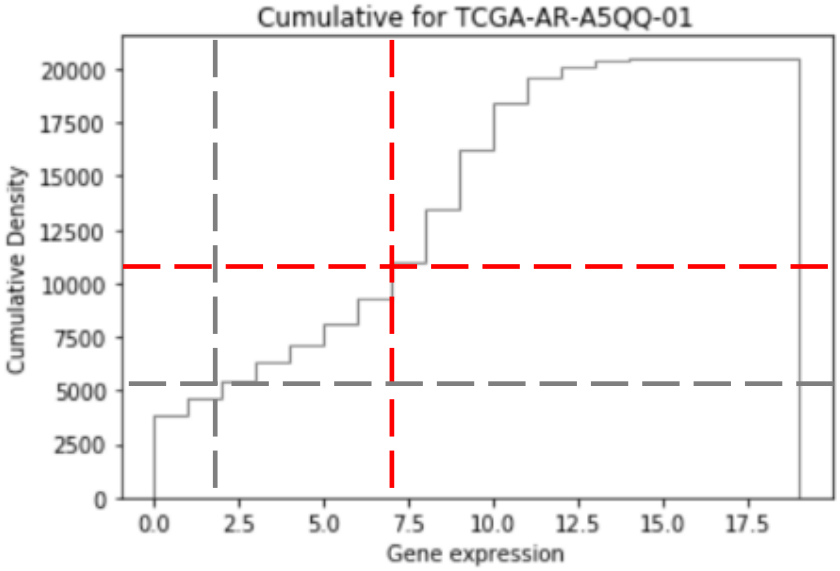




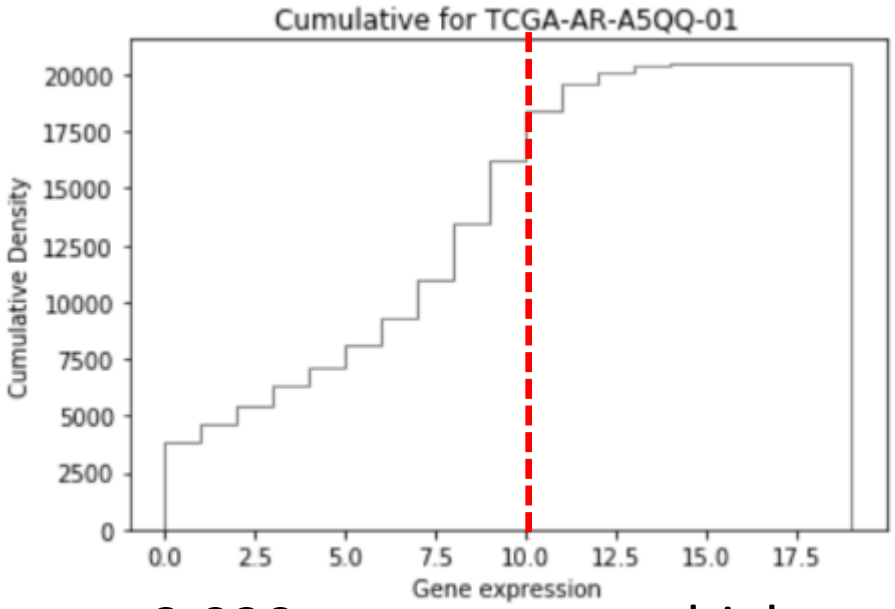
Visualizing cumulative distribution

- Use CDF when you want to discuss how many observations fall in some interval
- E.g., how many genes above ~2^2? Above ~2^7?
- How many below 2^2?
- To answer this you can either
  - Add up the bins of the histogram
  - or subtract the two values of the CDF

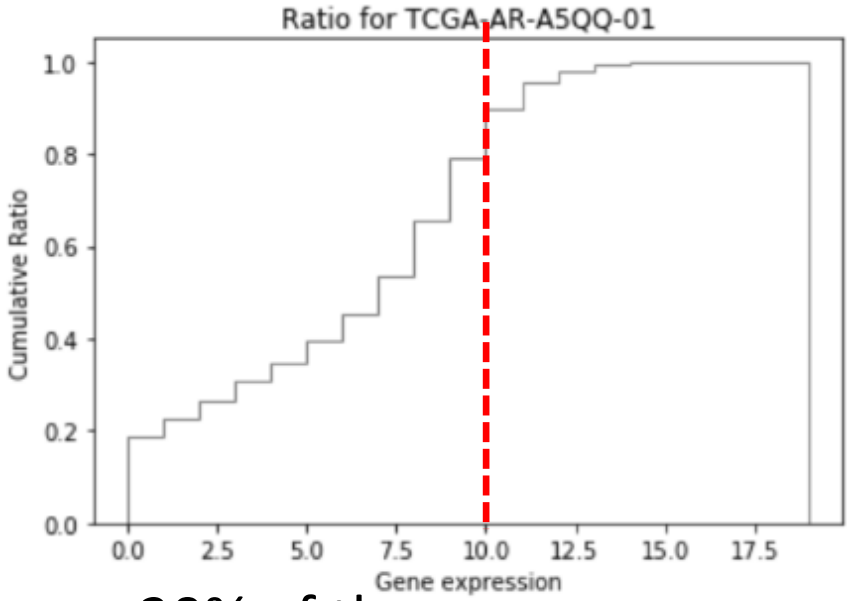




Cumulative distribution to visualize percentiles



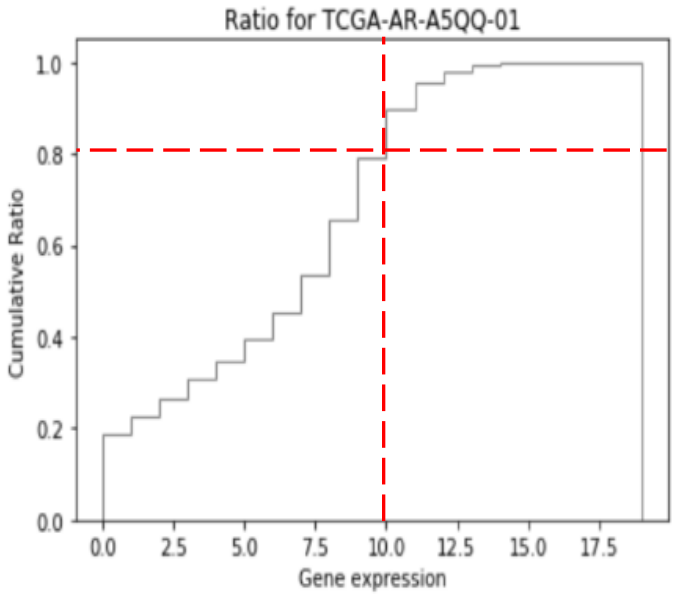
3,030 genes express higher than 10 (log2 FPKM)

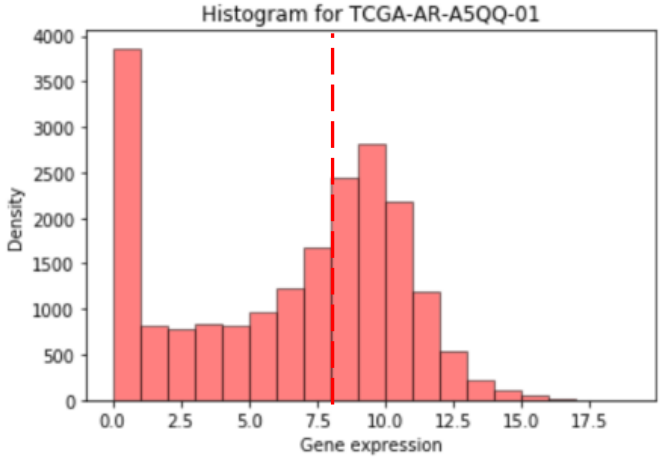


20% of the genes express higher than 10 (log2 FPKM)

Inference from cumulative dist. is “indirect”

Q: Which figure is better for backing up the following statement?



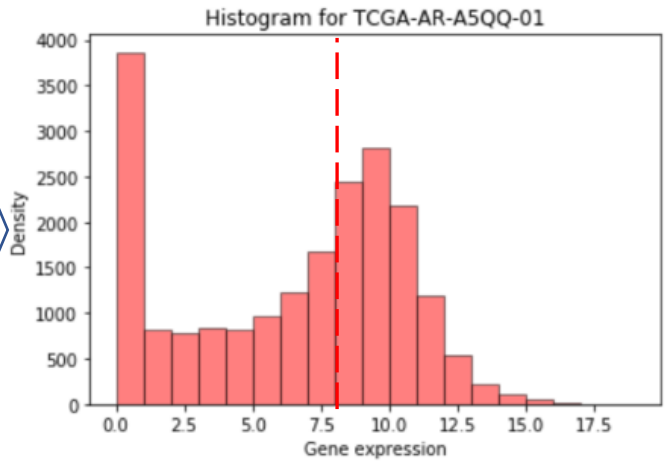


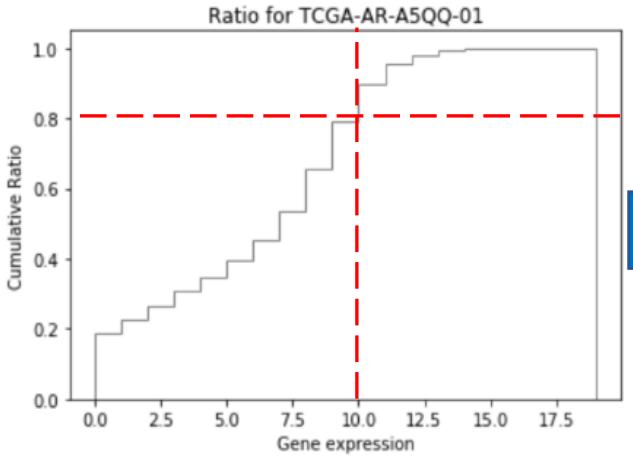
?

20% of the genes express higher than 10 (log2 FPKM)

Cumulative dist'n. can be "indirect"

Gene/sample	TCGA-BH-A0C7-01
ARHGEF10L	10.0931
HIF3A	3.609
RNF17	0
RNF10	11.382
RNF11	10.7663
RNF13	10.3366
GTF2IP1	12.724
REM1	2.2163
MTVR2	0.334
RTN4RL2	5.091
...	

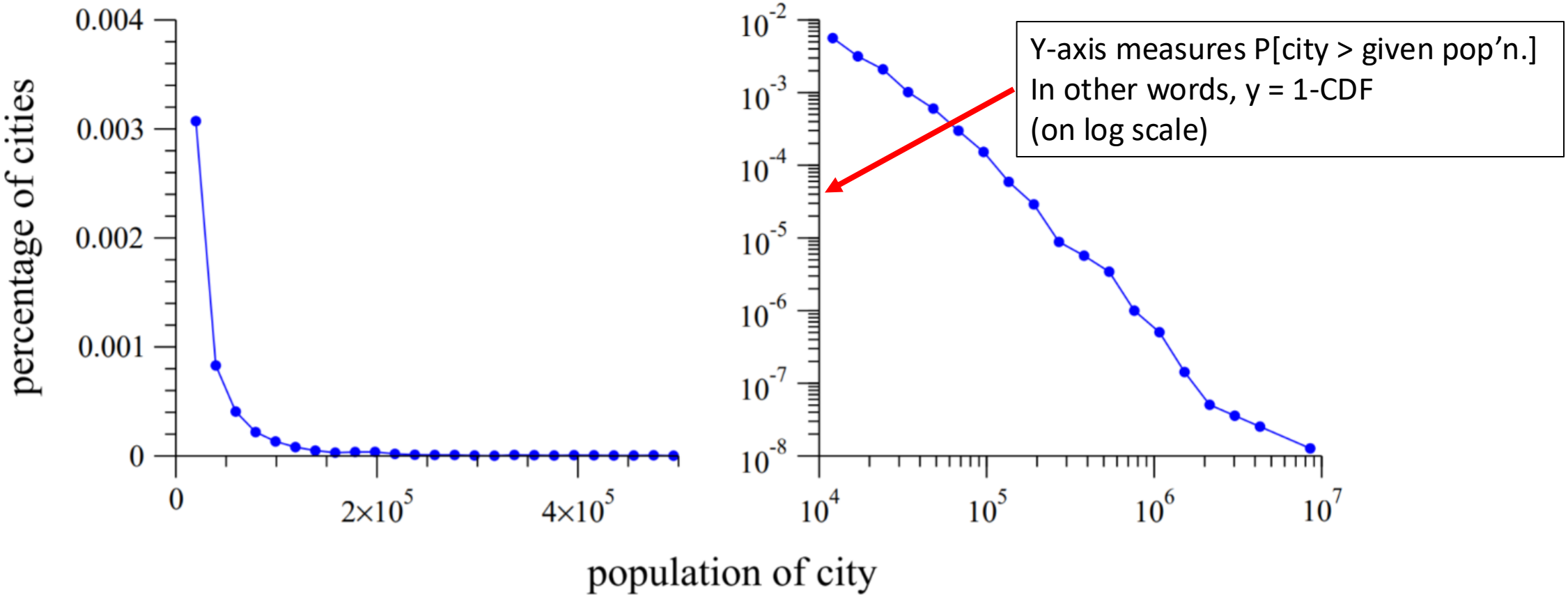




20% of the genes express higher than 10 (log2 FPKM)

Visualization to reveal **underlying statistical property**

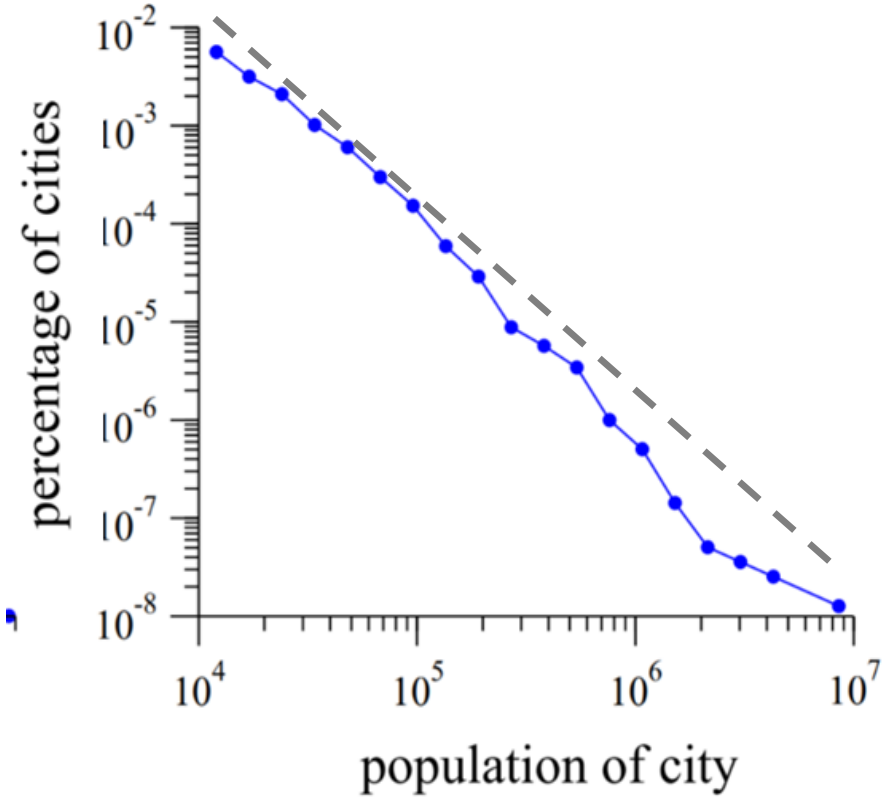
The populations of all US cities with population of 10,000 or more



http://www-personal.umich.edu/~mejn/courses/2006/cmplxsys899/powerlaws.pdf

Distribution visualization to reveal underlying statistical property

The populations of all US cities



Suppose the CDF obeys an equation like this:

$\ln[1- F(x)] = -a \ln x + b$

(as in the picture at the left)

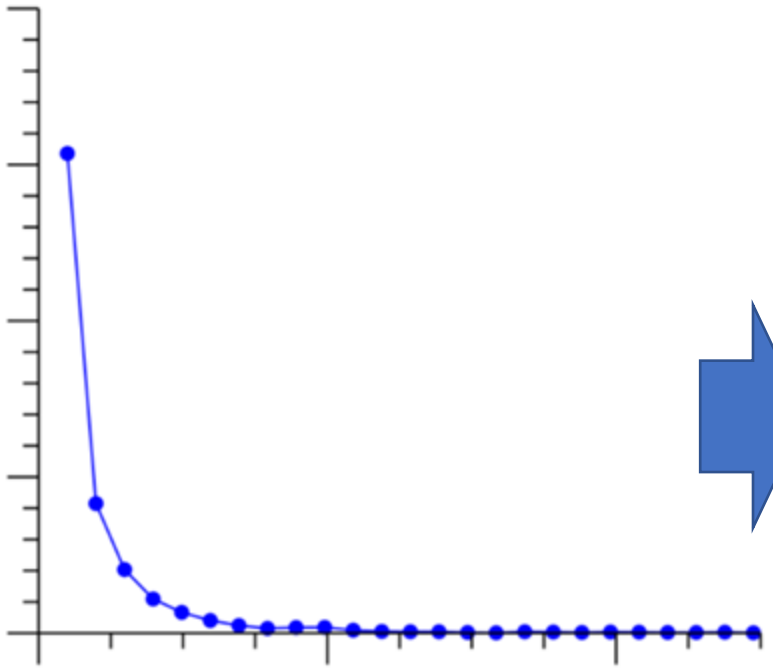
Then the PDF obeys an equation like this:

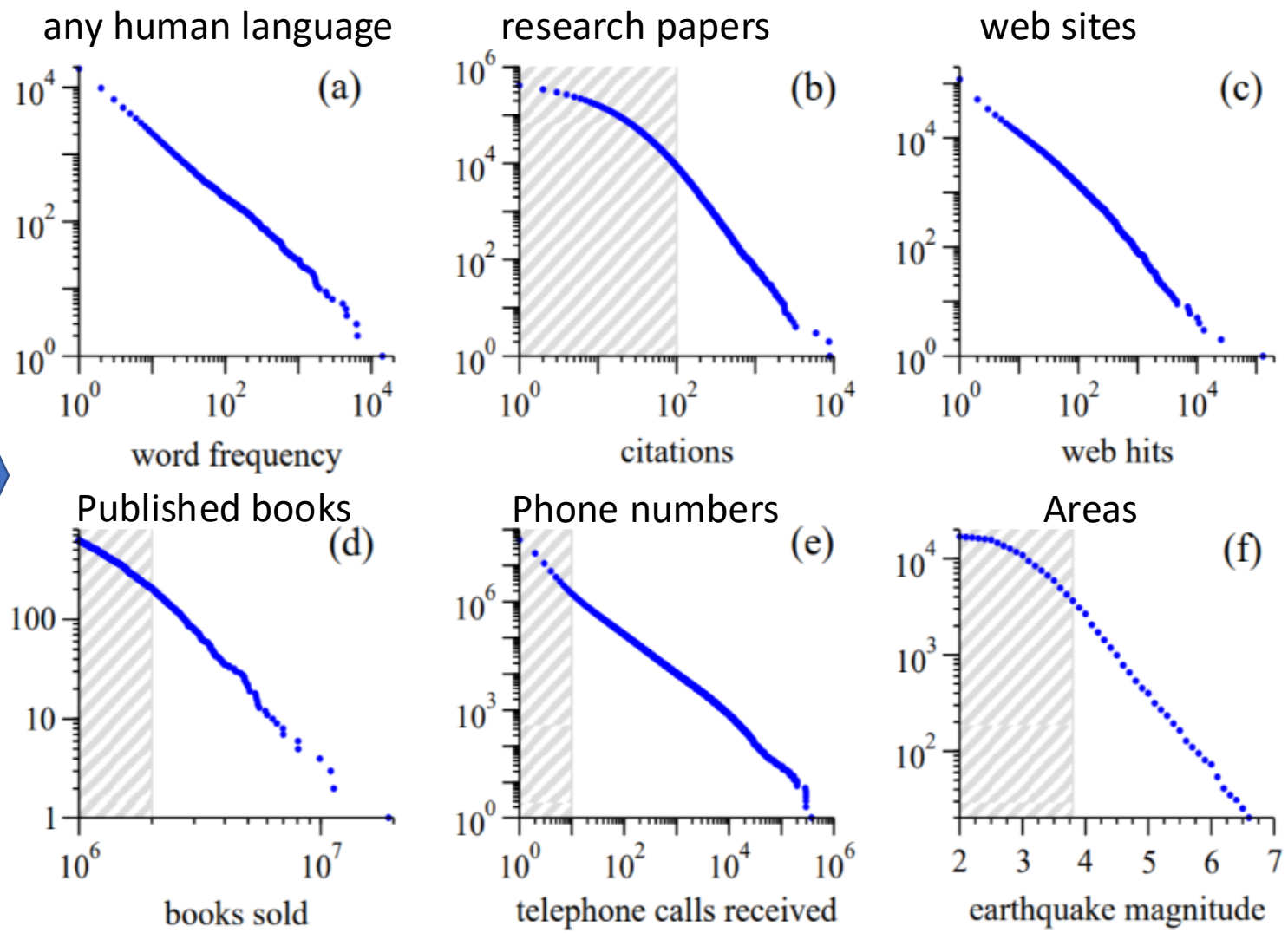
$f(x) = Cx^{-\alpha}$

(the power law)

http://www-personal.umich.edu/~mejn/courses/2006/cmplxsys899/powerlaws.pdf

Some observed power laws

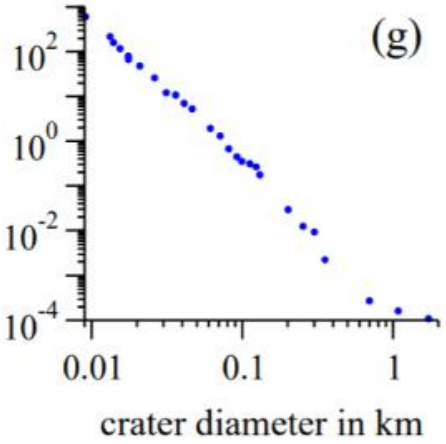




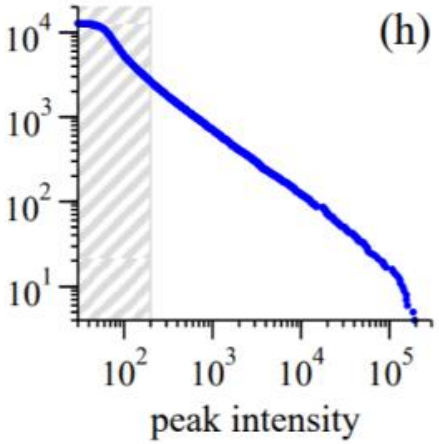
http://www-personal.umich.edu/~mejn/courses/2006/cmplxsys899/powerlaws.pdf

Some more observed power laws

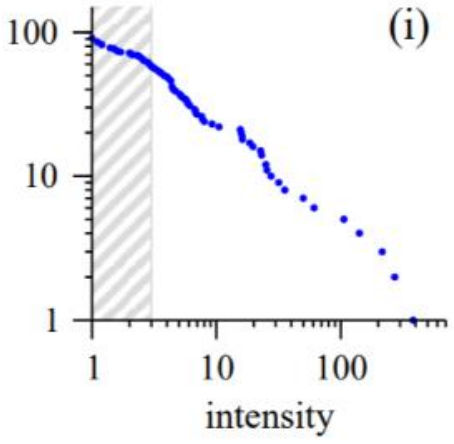
Moon craters



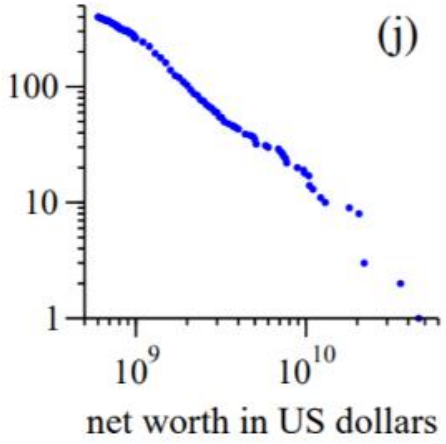
Gamma ray in solar flares



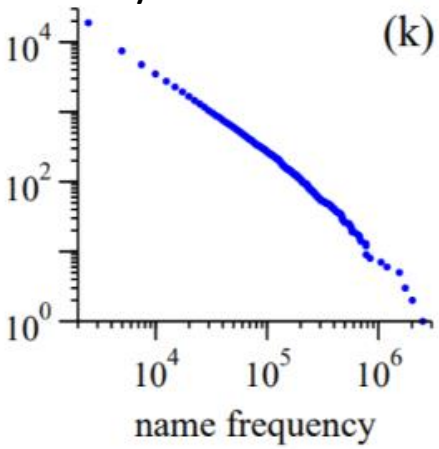
Of wars b/w 1816 & 1980

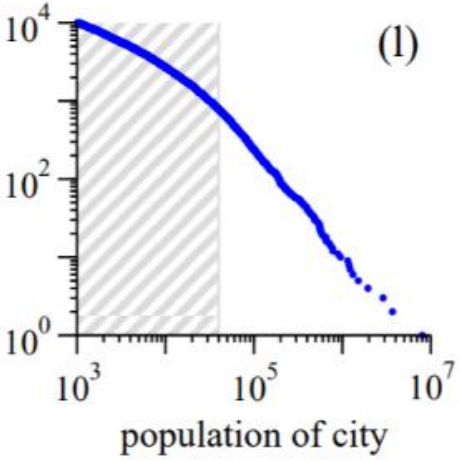


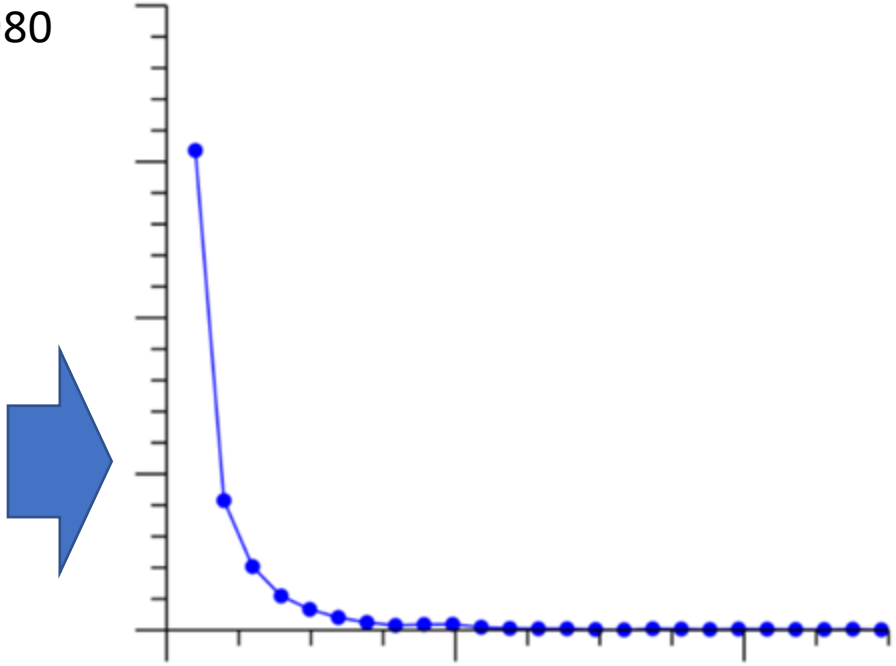
Annual income



Family name in US





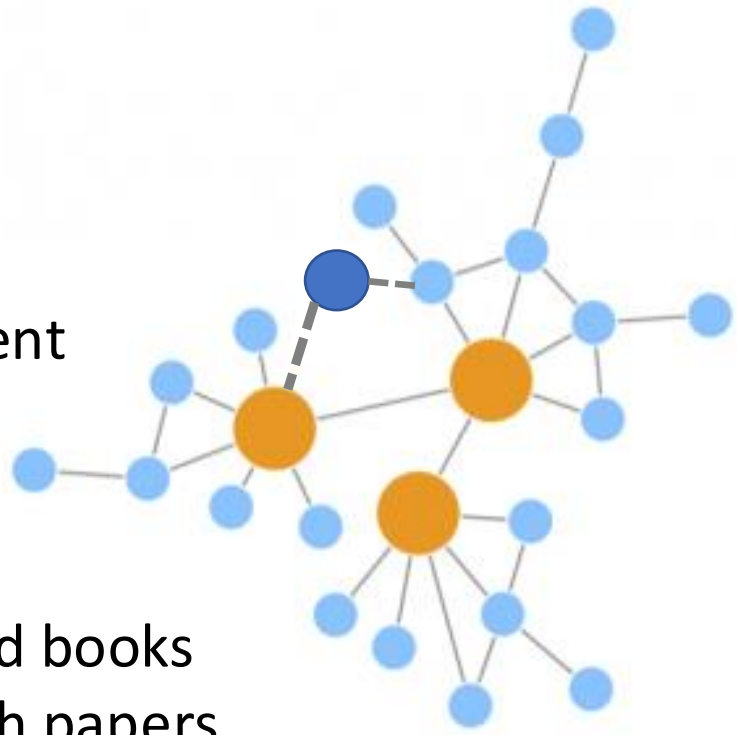


In terms of network nodes:

- “rich-get-richer process”
- Preferential attachment

http://www-personal.umich.edu/~mejn/courses/2006/cmplxsys899/powerlaws.pdf

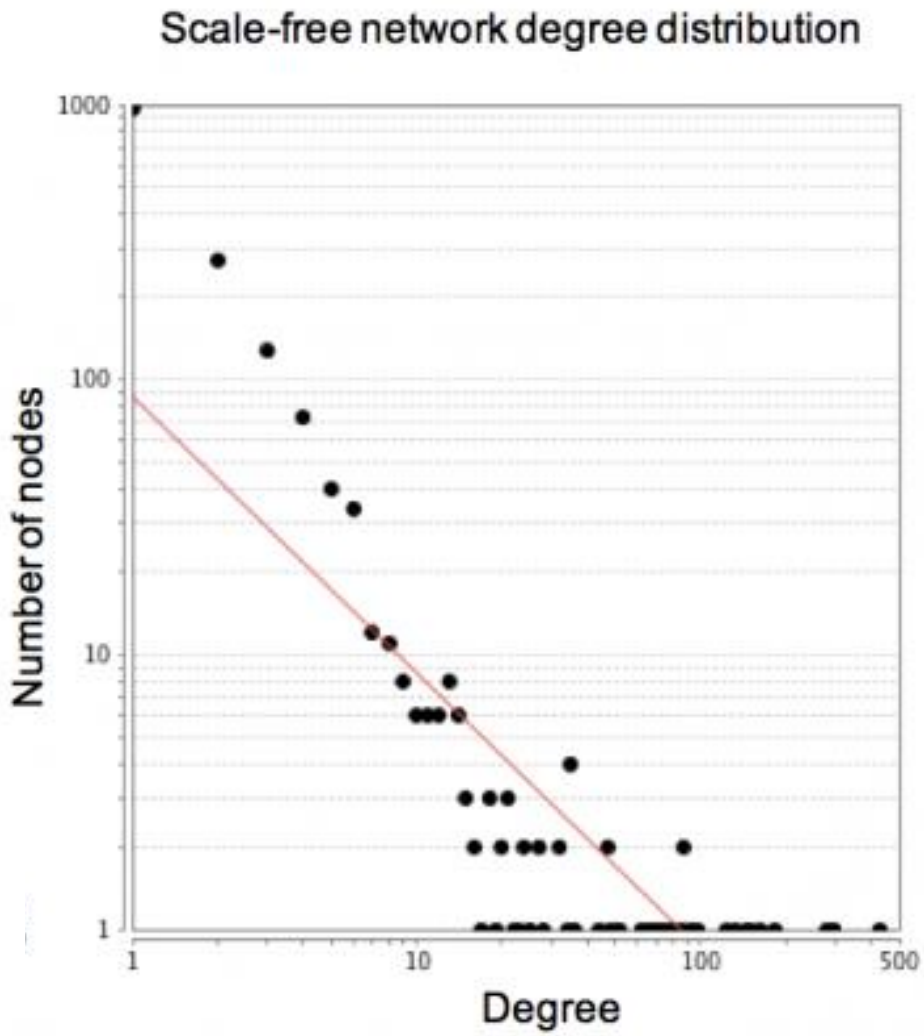
Power laws appear in a ‘scale-free’ network that underwent preferential attachment



Preferential attachment

E. g.,

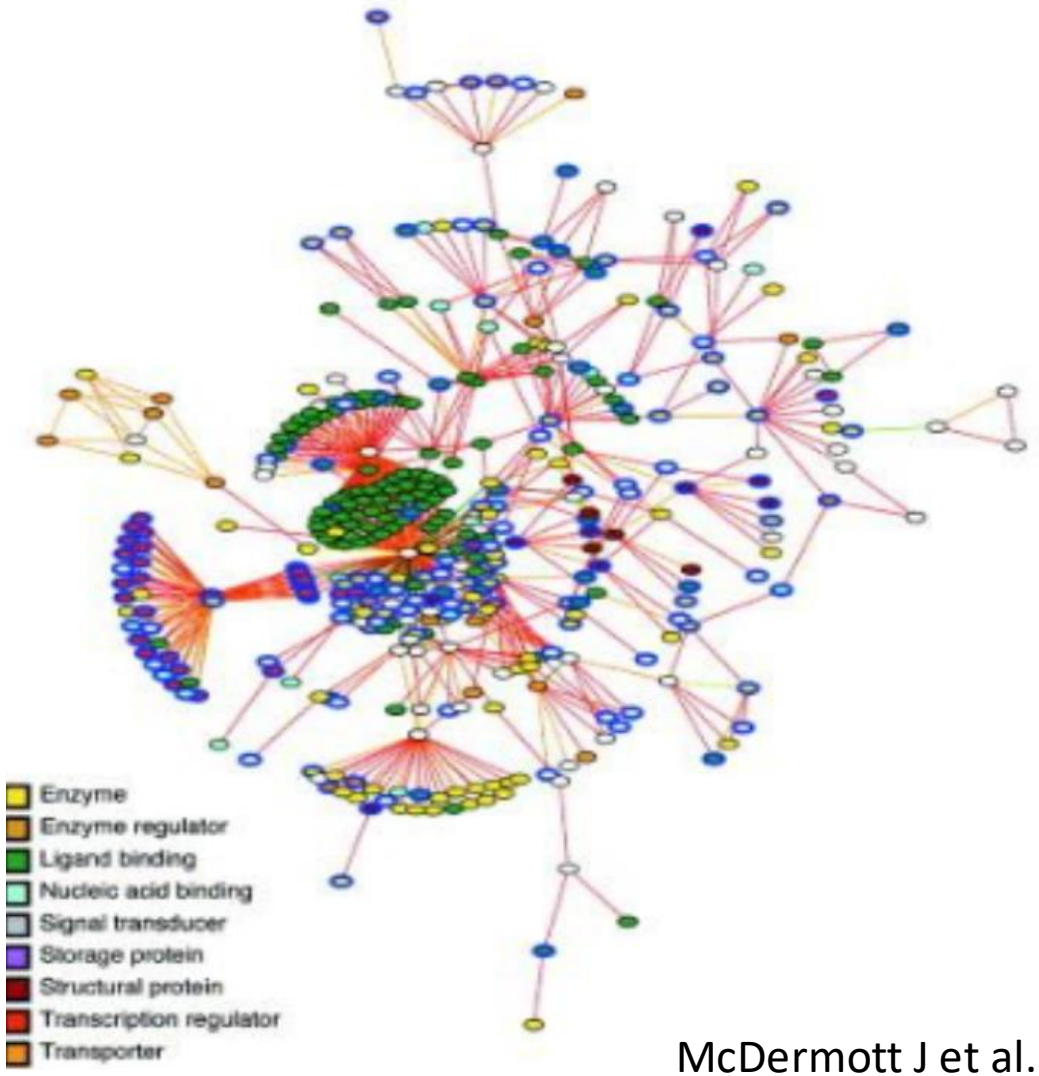
- Calling frequency
- Hits of web sites
- Selling of published books
- Citation of research papers
- Use frequency of words in a language
- Earthquake frequency across areas



https://medium.com/@pmontee/ventures-role-a-power-law-economy-and-society-f128f6691dc5

Typical look of scale-free networks

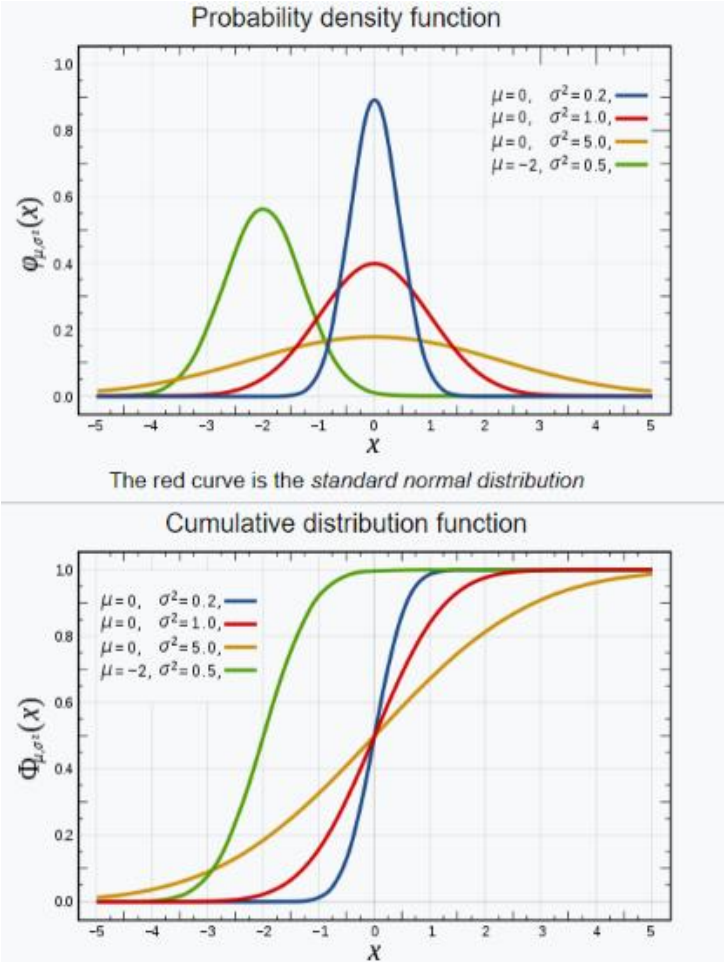
Predicted protein networks for E.coli.



McDermott J et al. Trends Biotechnol. (2004)

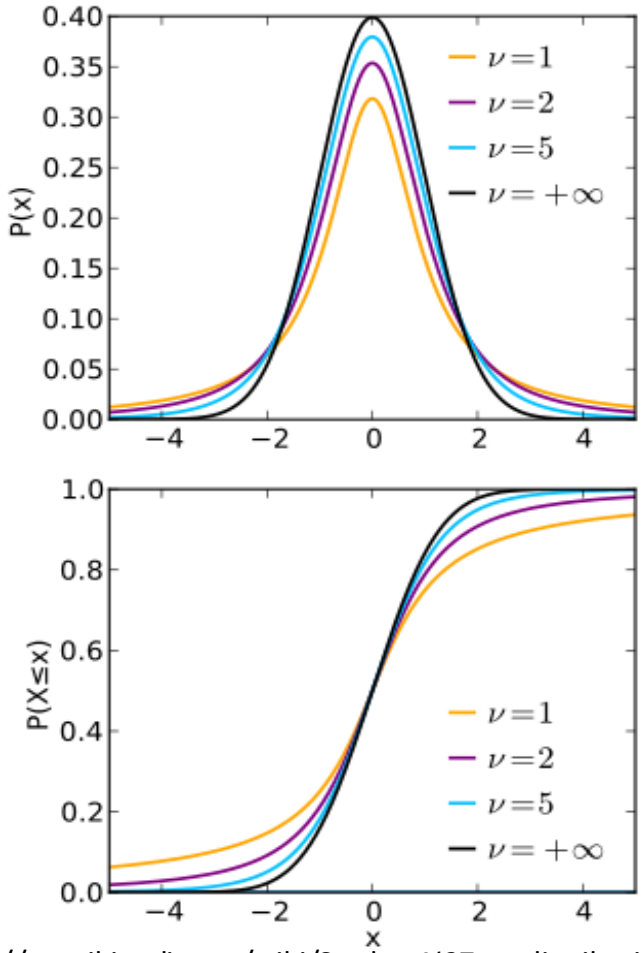
Some common probability distributions

Normal distribution



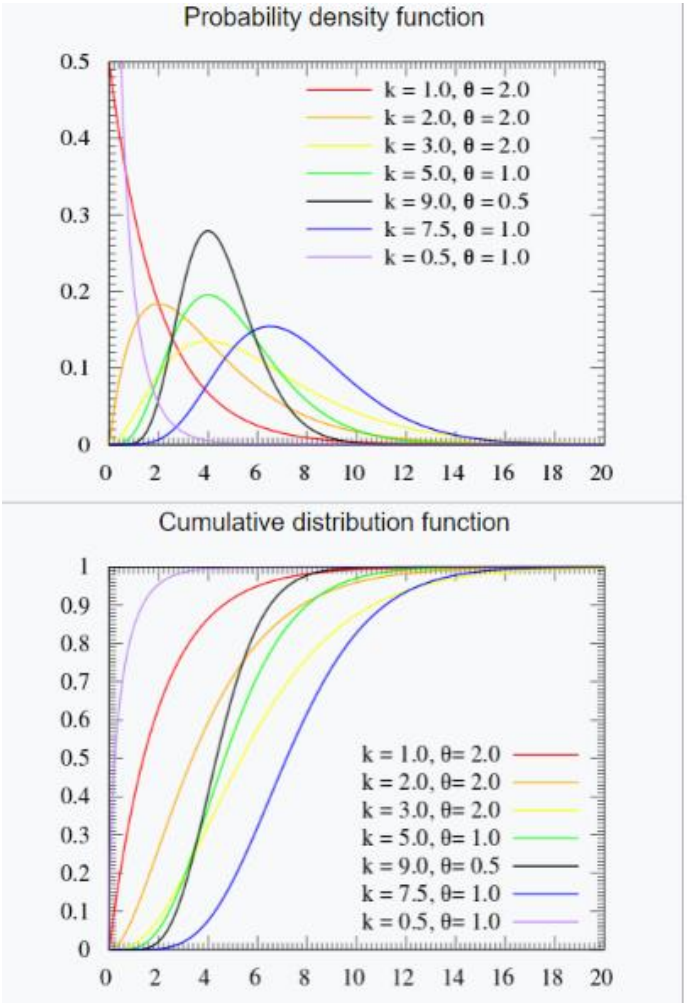
https://en.wikipedia.org/wiki/Normal_distribution

Student's t distribution



https://en.wikipedia.org/wiki/Student%27s_t-distribution

Gamma distribution

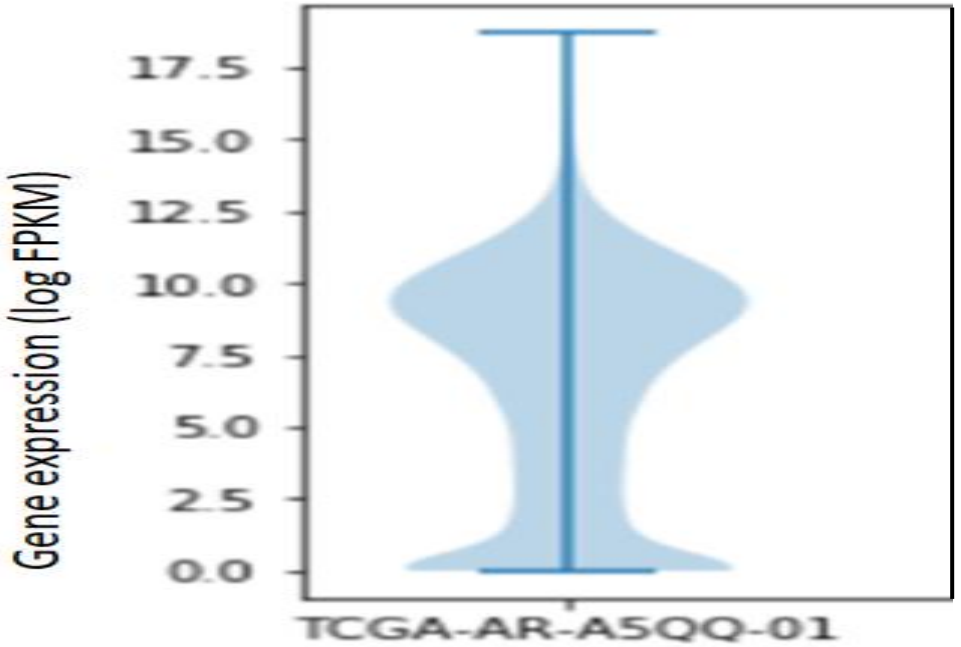


https://en.wikipedia.org/wiki/Gamma_distribution

How to identify the underlying distribution?

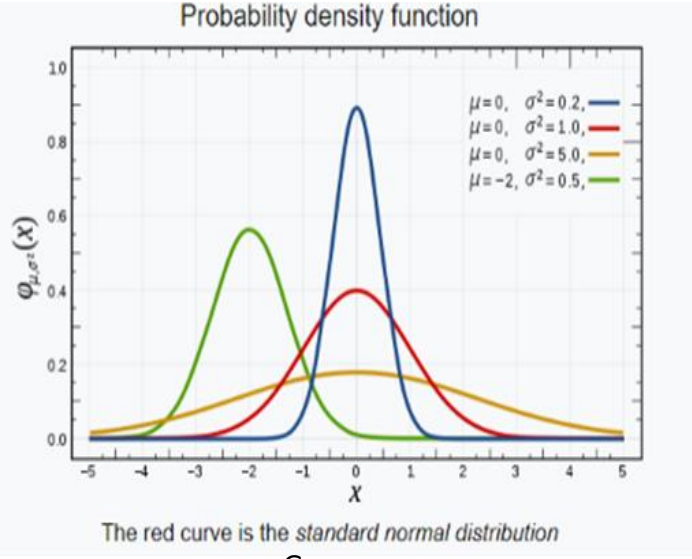
Questions: to what extent does each follow

- Normal distribution?
- T-distribution
- Gamma distribution?
- Bimodal?

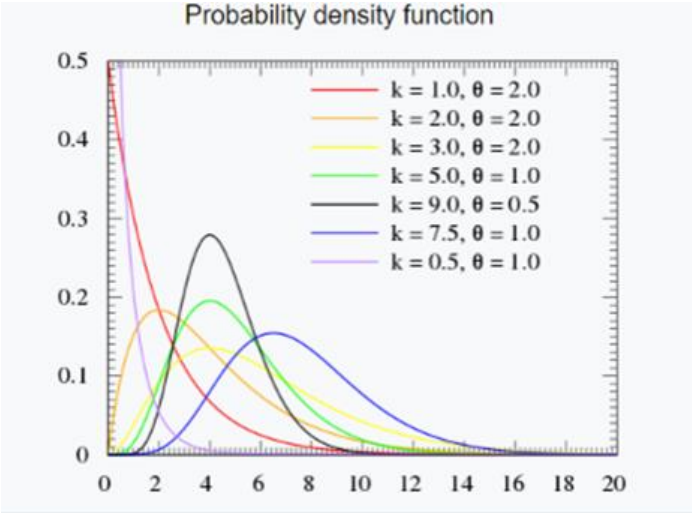


Some distributions may look quite alike

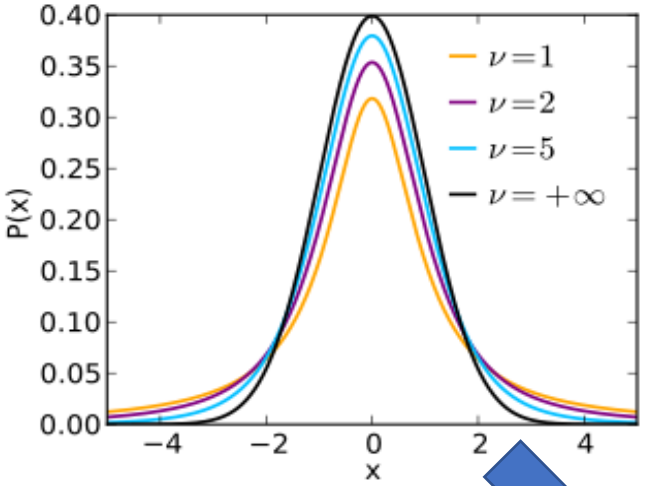
Normal distribution

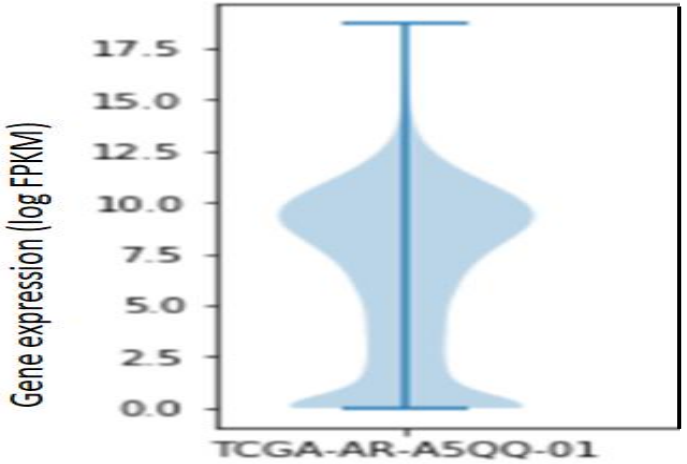


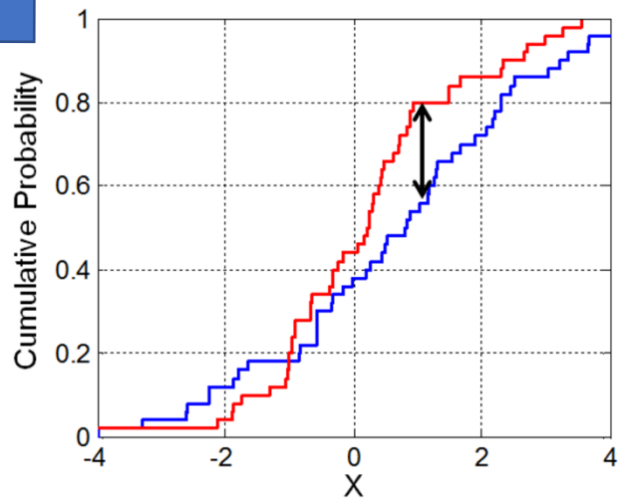
Gamma



Student's t-distribution
Probably density function





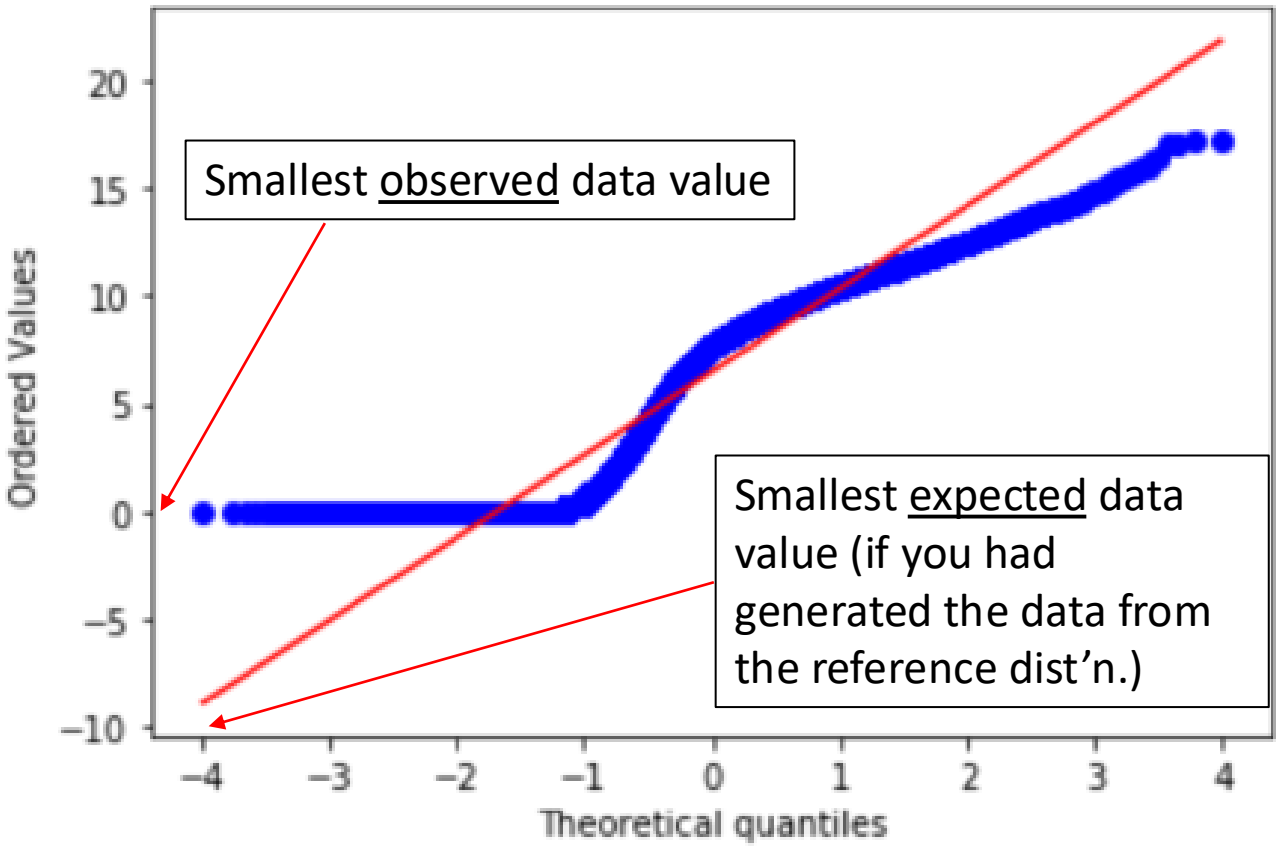


https://en.wikipedia.org/wiki/Kolmogorov%E2%80%93Smirnov_test

Quantile-quantile (Q-Q) plots compare your data's distribution with a reference distribution (of your choice)

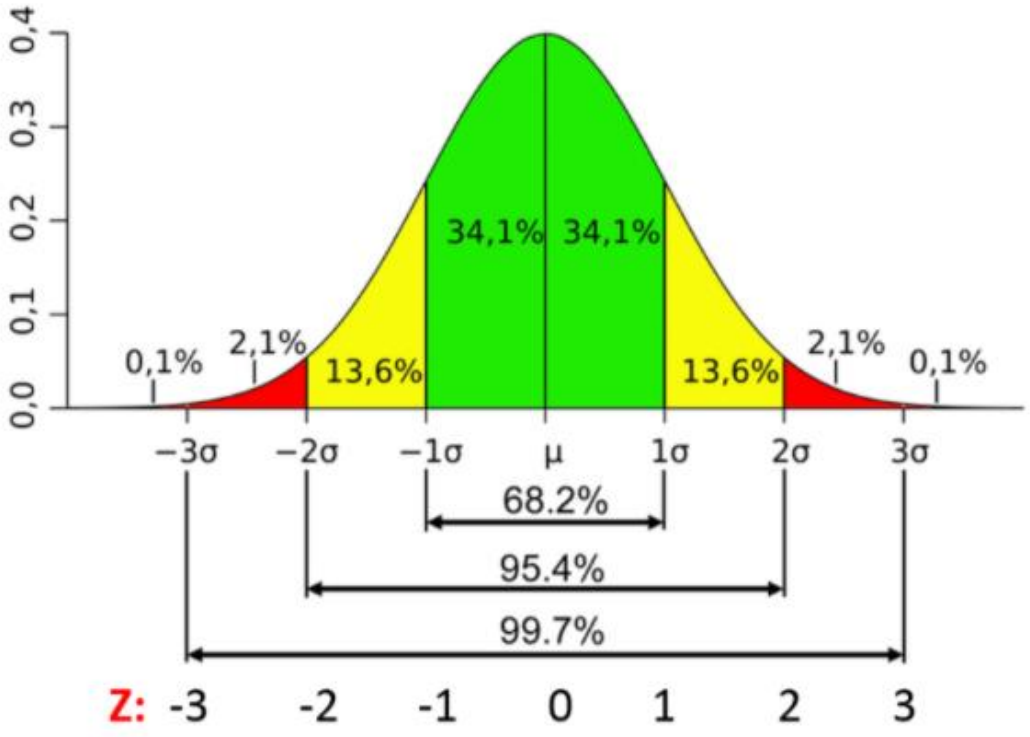
- For a given reference distribution, you can always calculate quantiles
  - What would the smallest data point probably be?
  - The second smallest?
  - …
  - The biggest?
- Plot the observed vs. theoretical quantiles
- Points: x-coordinate represents the position in a reference distribution and y-coordinate represents the actual value
- Diagonal line: where a given data point should fall if the data follows a specific reference distribution
- TCGA-BH-A0C7-01 seems not to follow Gaussian

TCGA-BH-A0C7-01 for normal distribution



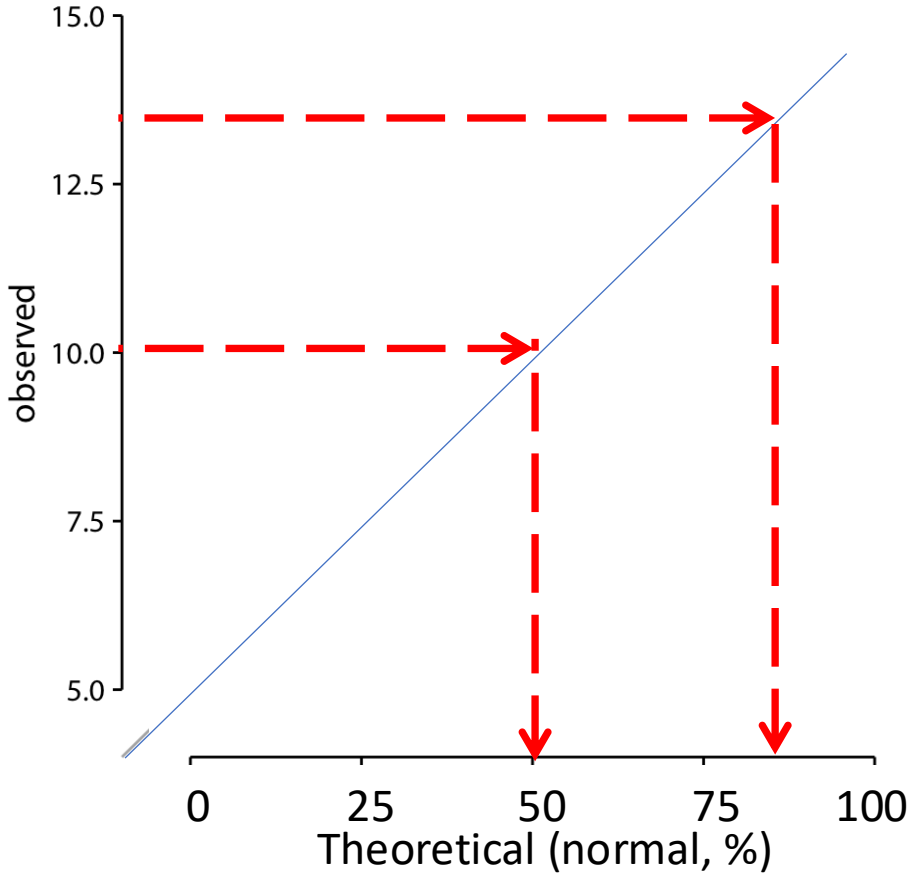
Quantile-quantile plot compares two dist'ns.

- E.g. data ~ N(mean=10, sd=3)



https://sphweb.bumc.bu.edu/otlt/MPH-Modules/PH717-QuantCore/PH717-Module6-RandomError/PH717-Module6-RandomError5.html

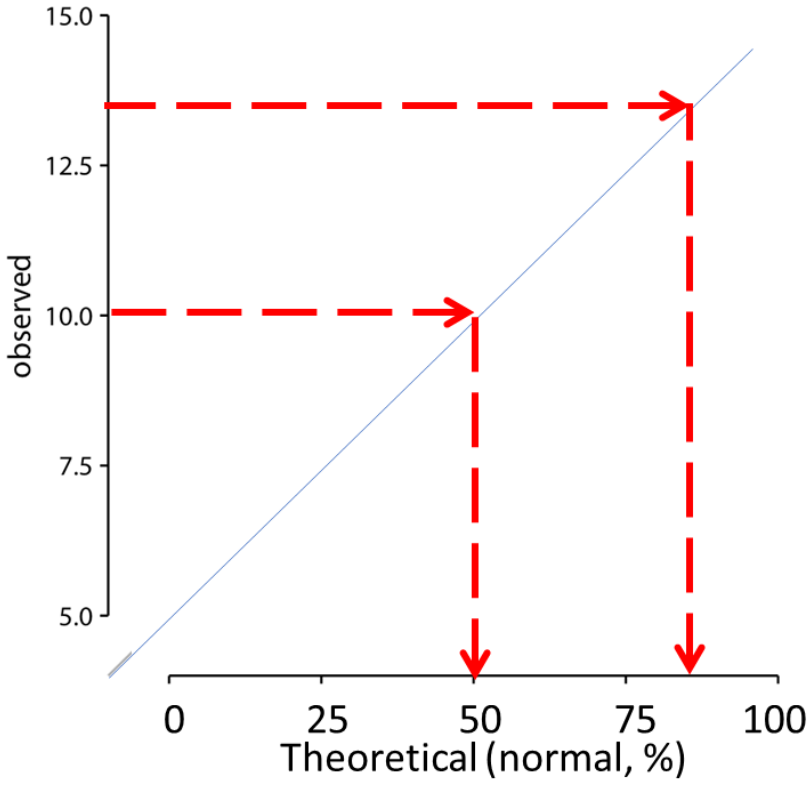
- E.g. 10 should be at the 50th percentile and 13 should be the 84.1th percentile (50+34.1)



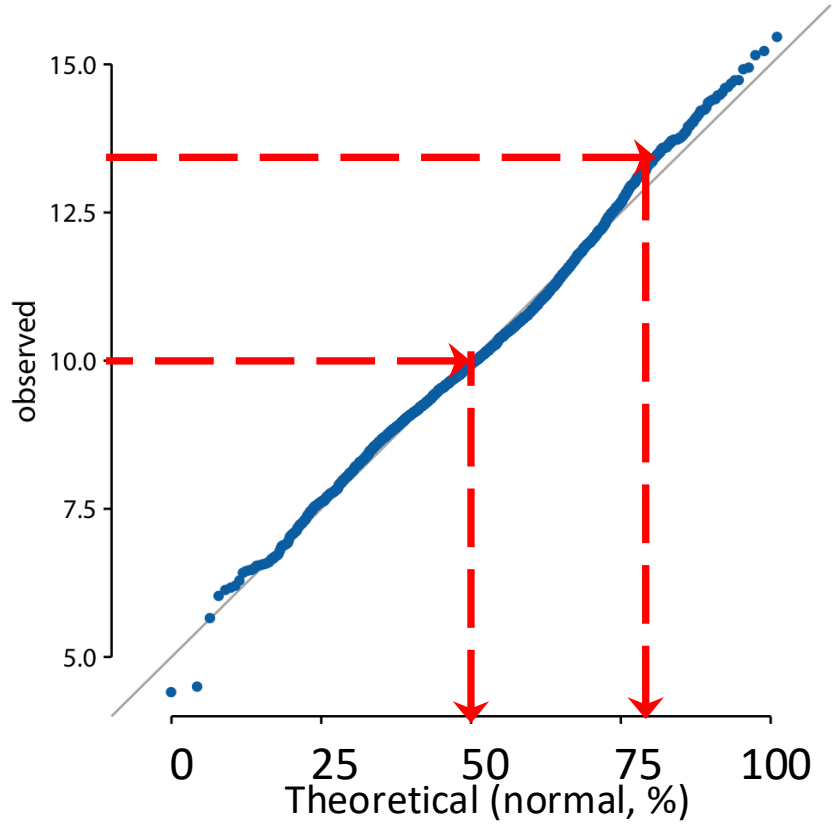
Adapted from Fundamentals of Data Visualization, Wilke, O'Reilly, 1st Ed.

Quantile-quantile plot checks the association of the observed variable to a given distribution

- E.g. 10 should be at the 50th percentile and 13 should be the 84.1th percentile (50+34.1)



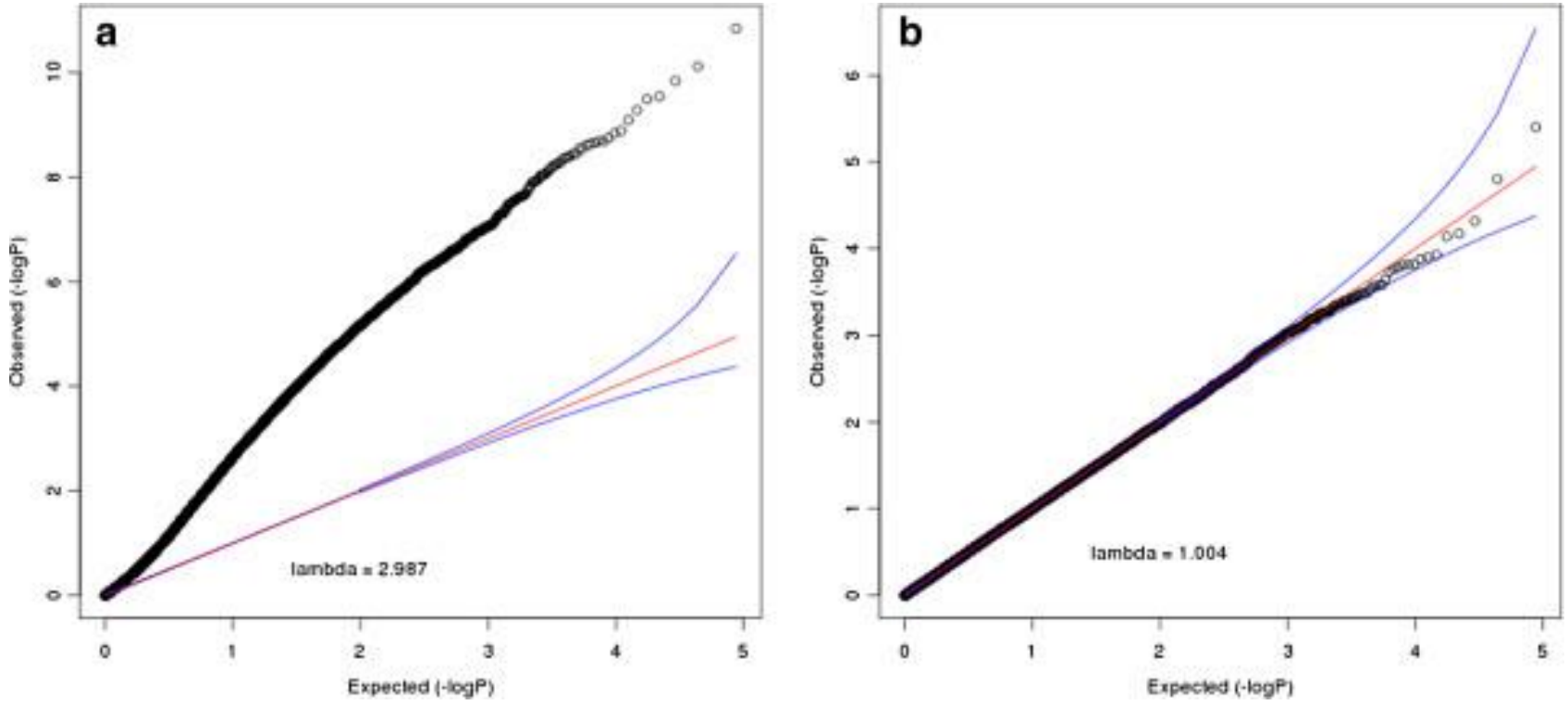
- E.g. In actual data, 10 is at the 50th percentile but 13 is stacked up lower than the 84.1th percentile due to a local value inflation



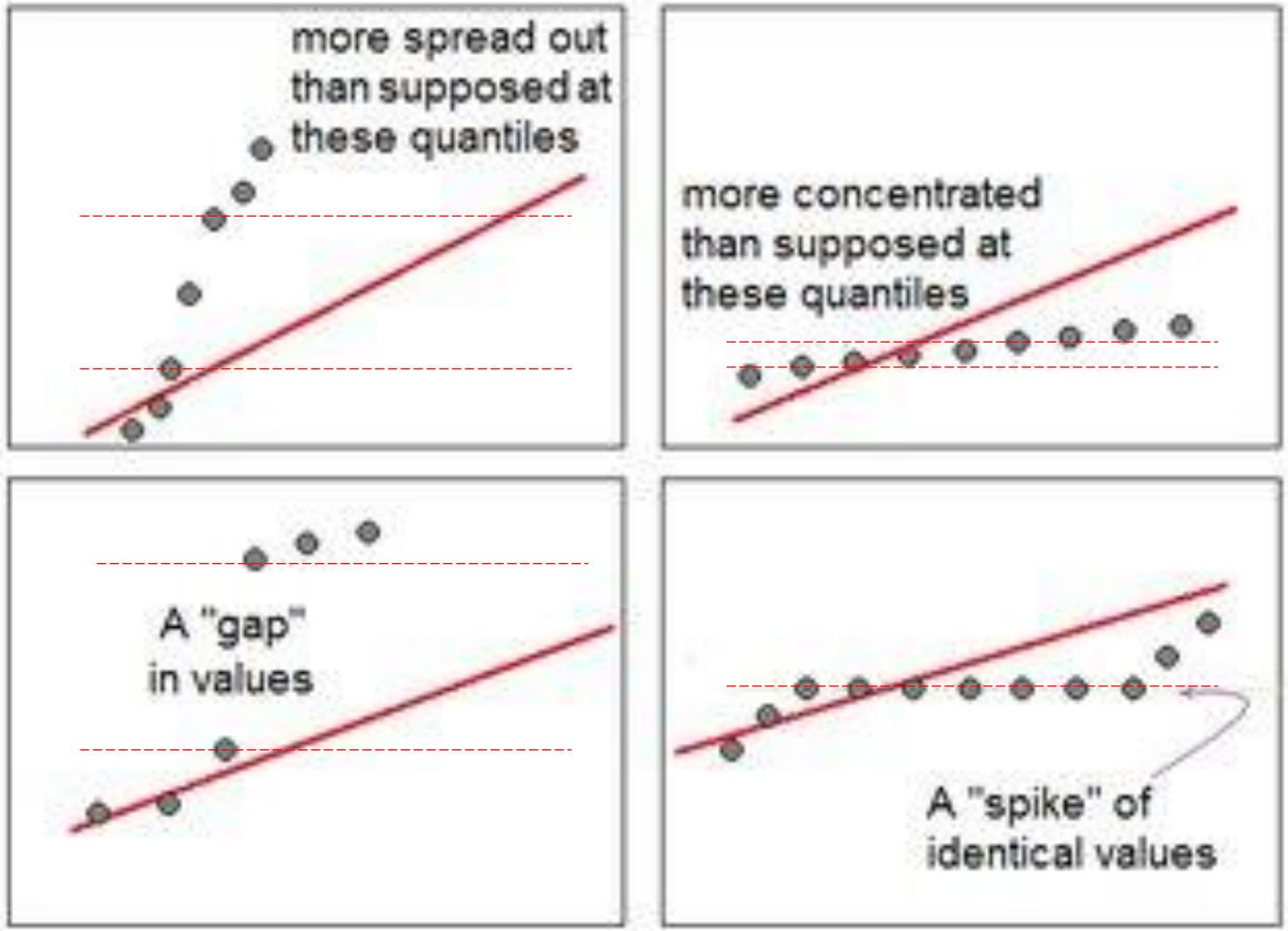
Adapted from Fundamentals of Data Visualization, Wilke, O'Reilly, 1st Ed.

Quantile-quantile plot checks the association of the observed variable to a given distribution

- Lambda quantifies how much it deviates from our expectation. Often want it to be roughly 1.

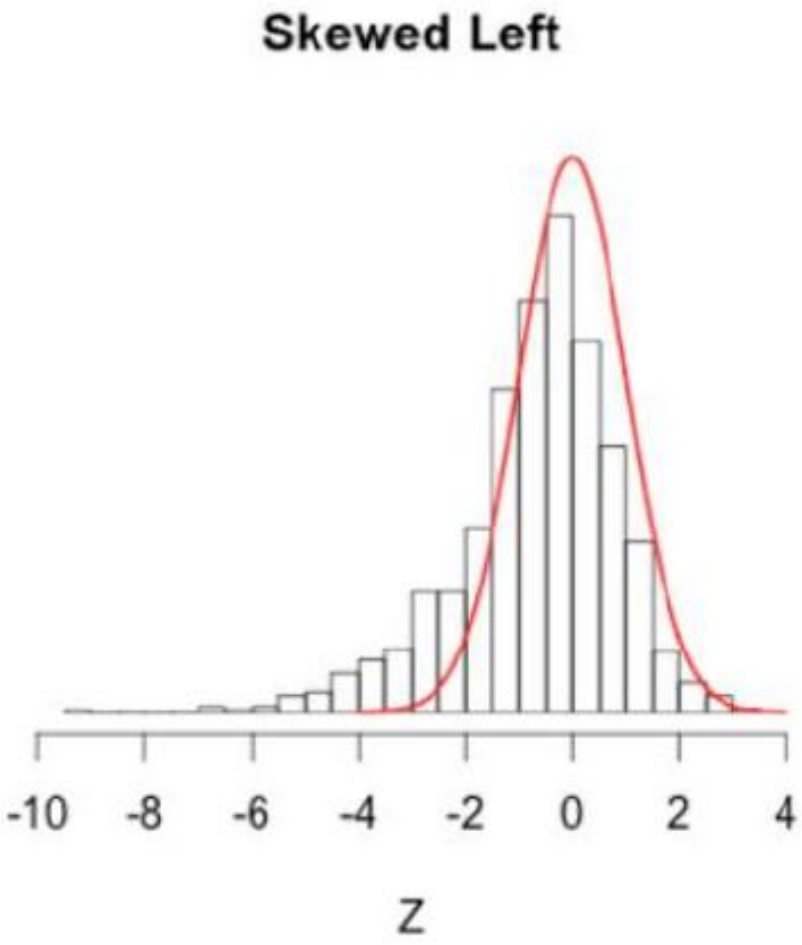


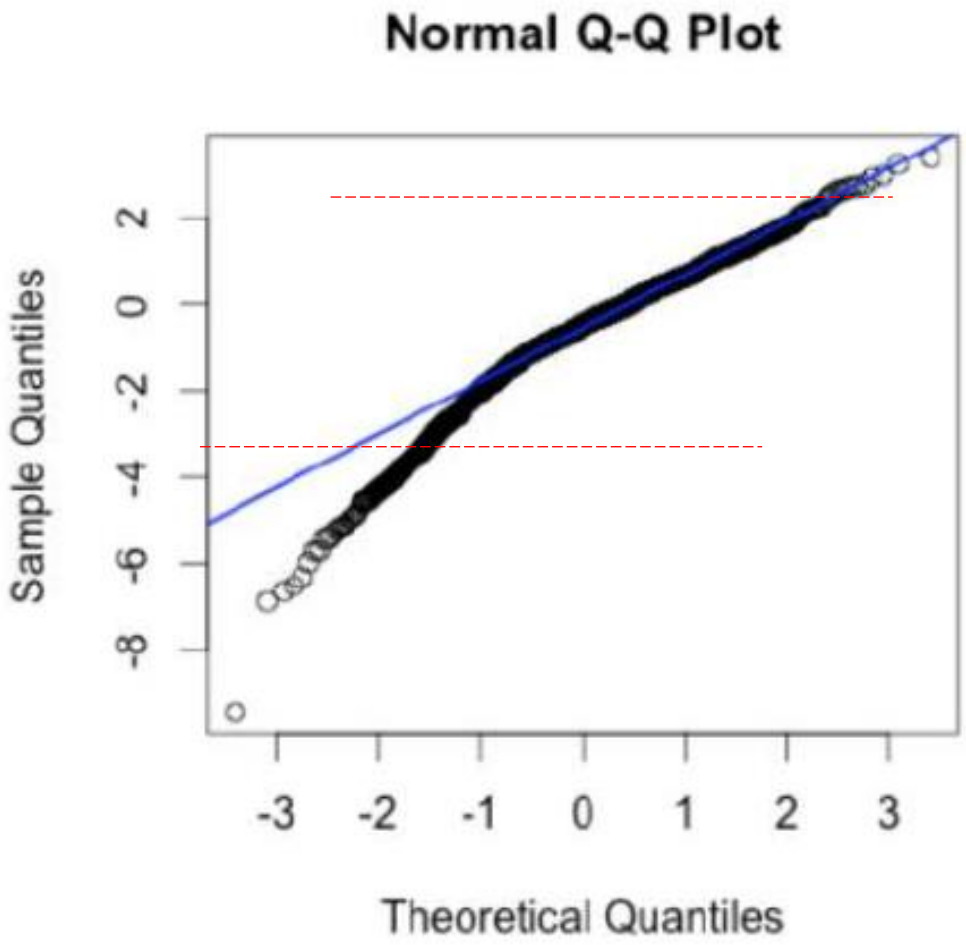
Local behavior in a quantile-quantile plot



https://stats.stackexchange.com/questions/101274/how-to-interpret-a-qq-plot

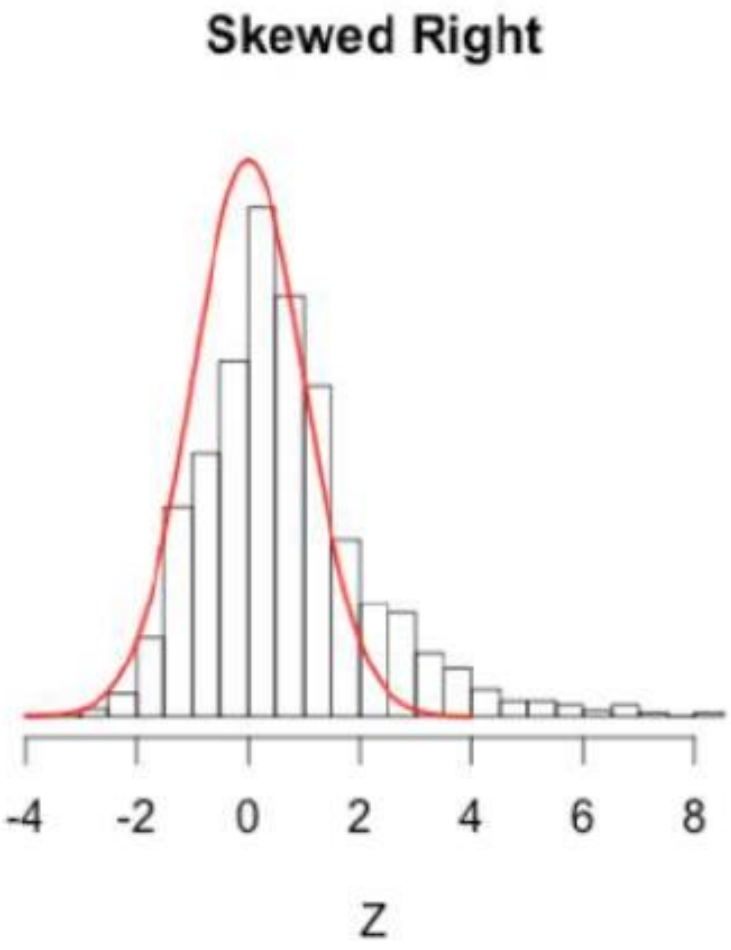
Skewness in normal quantile-quantile plot

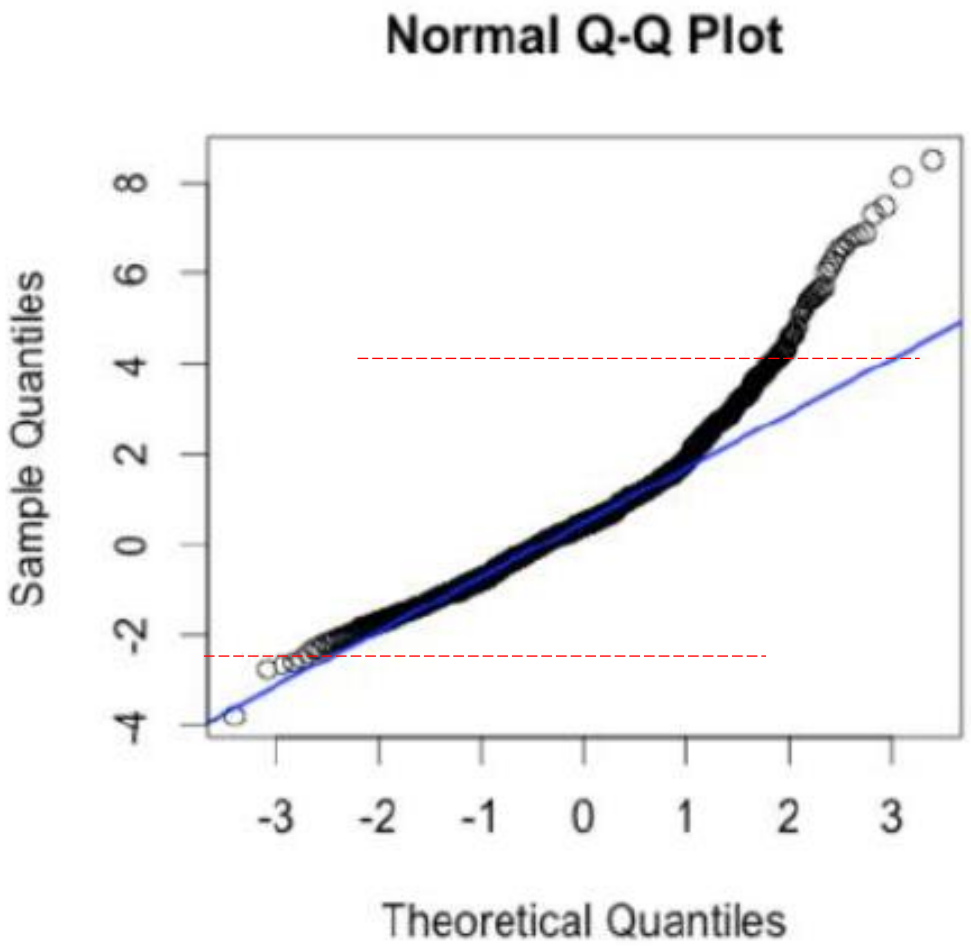




https://towardsdatascience.com/q-q-plots-explained-5aa8495426c0

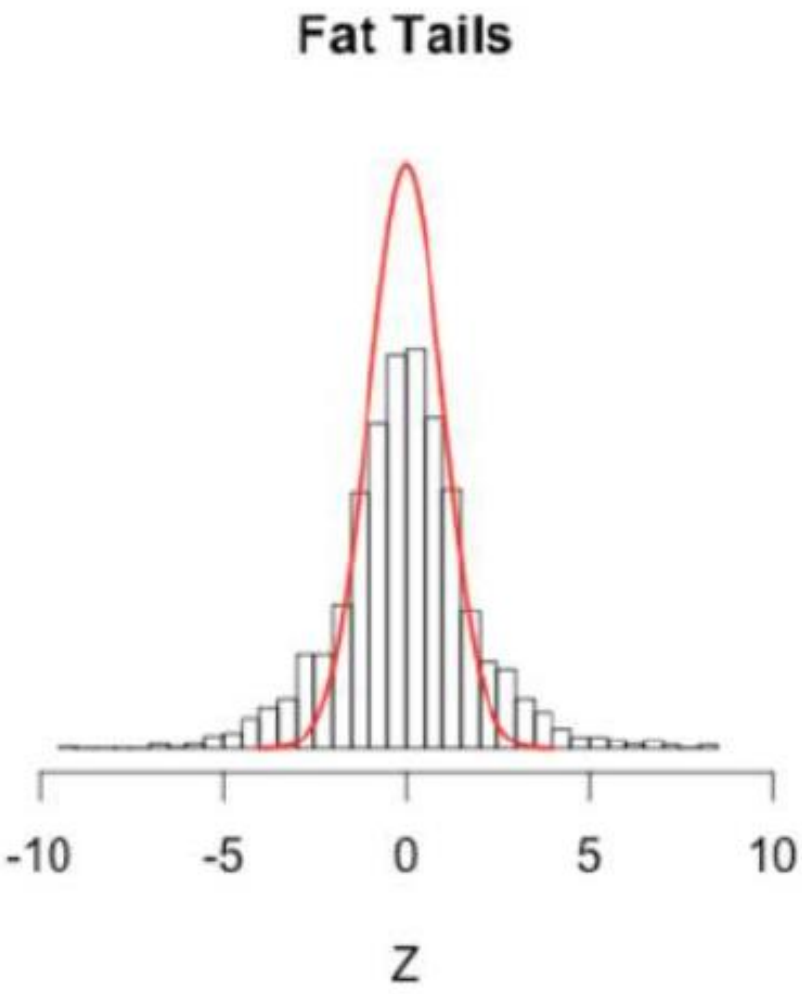
Skewness in normal quantile-quantile plot

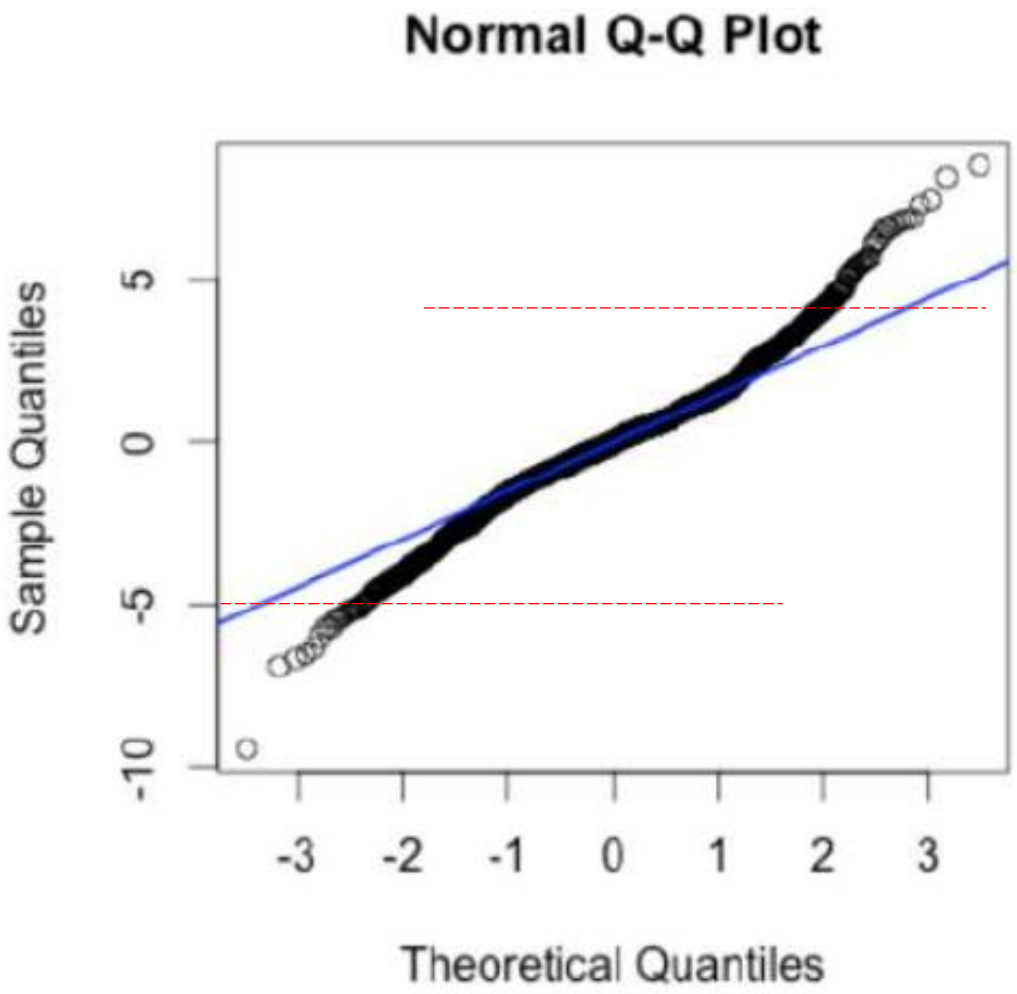




https://towardsdatascience.com/q-q-plots-explained-5aa8495426c0

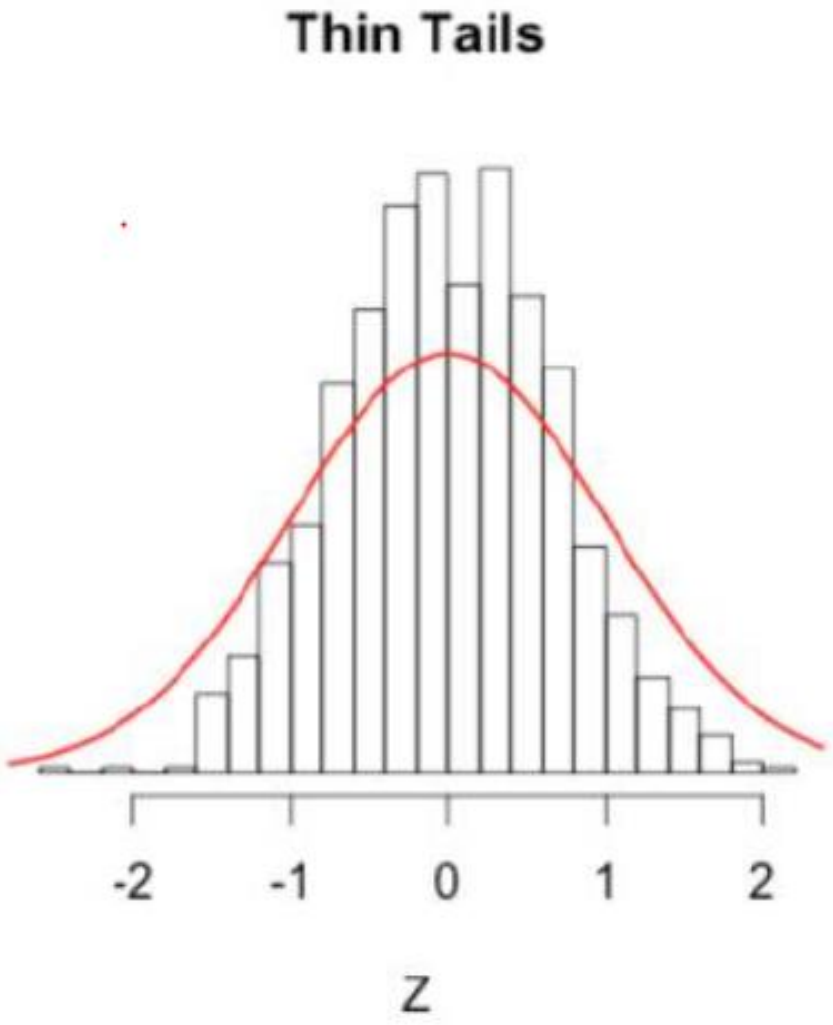
Fat tail in normal quantile-quantile plot

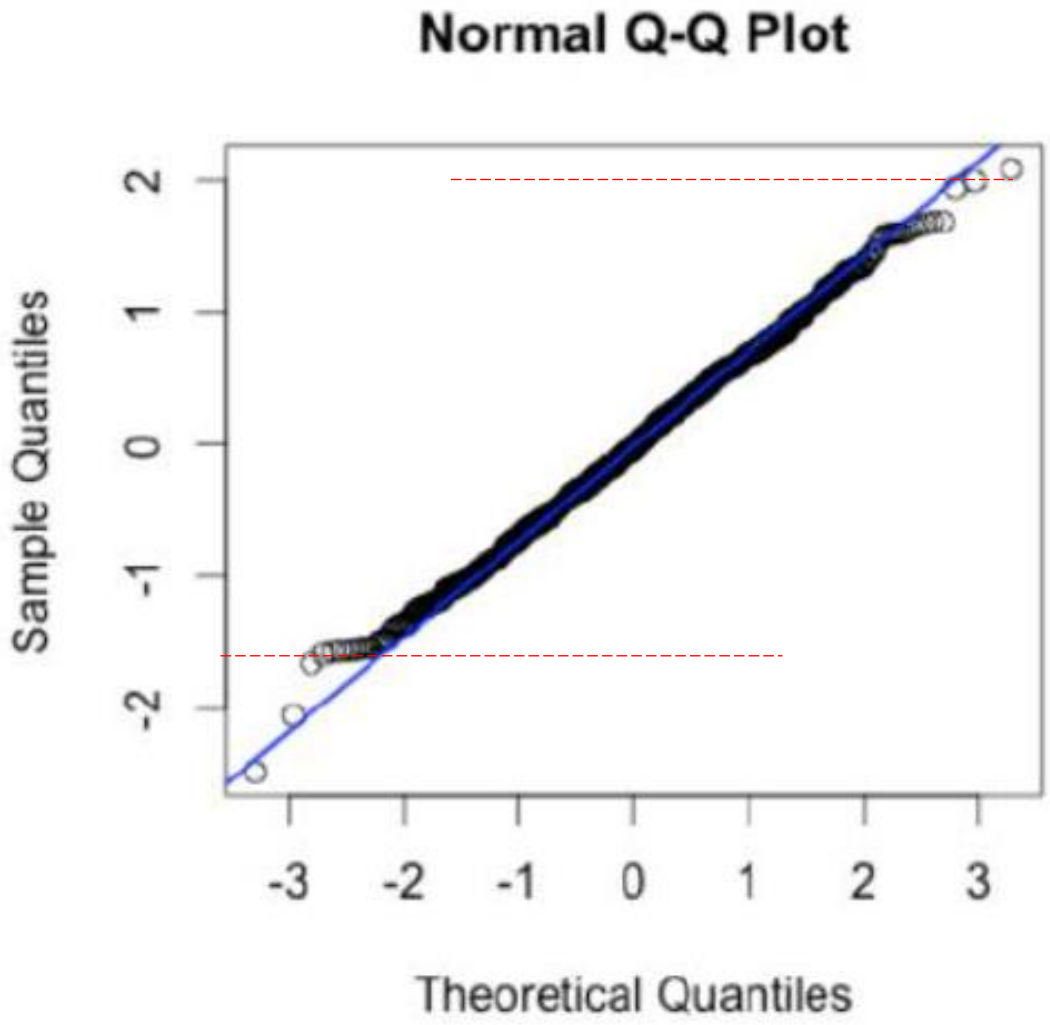




https://towardsdatascience.com/q-q-plots-explained-5aa8495426c0

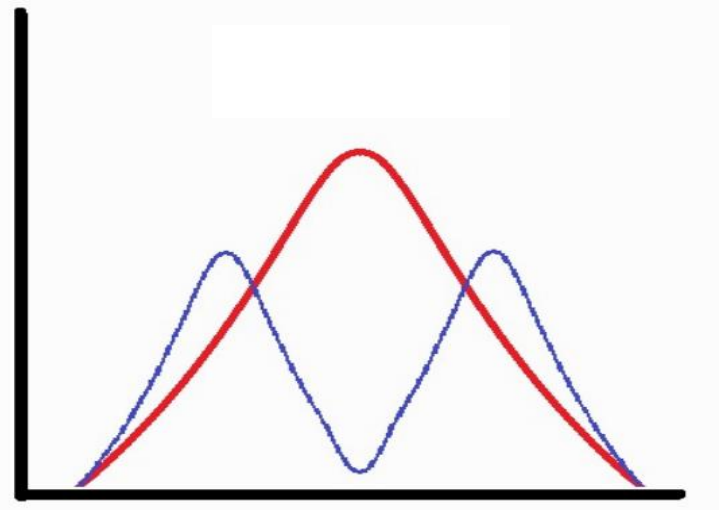
Light tail in normal quantile-quantile plot

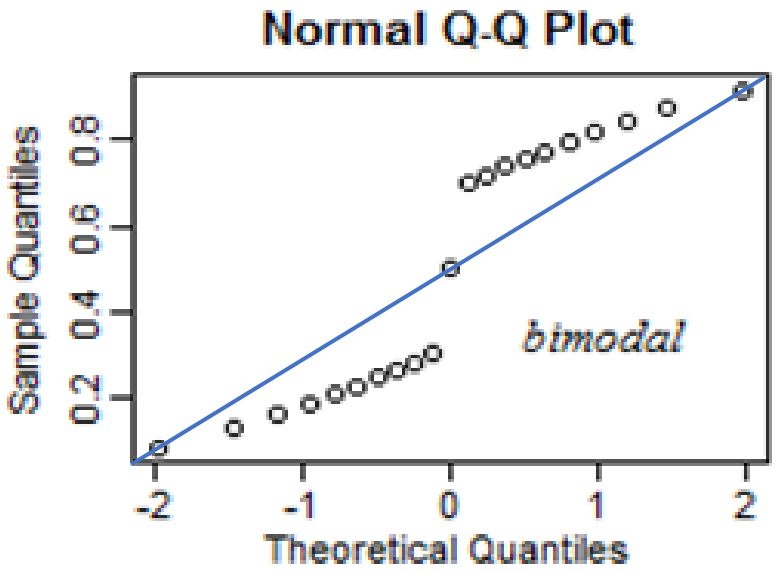




https://towardsdatascience.com/q-q-plots-explained-5aa8495426c0

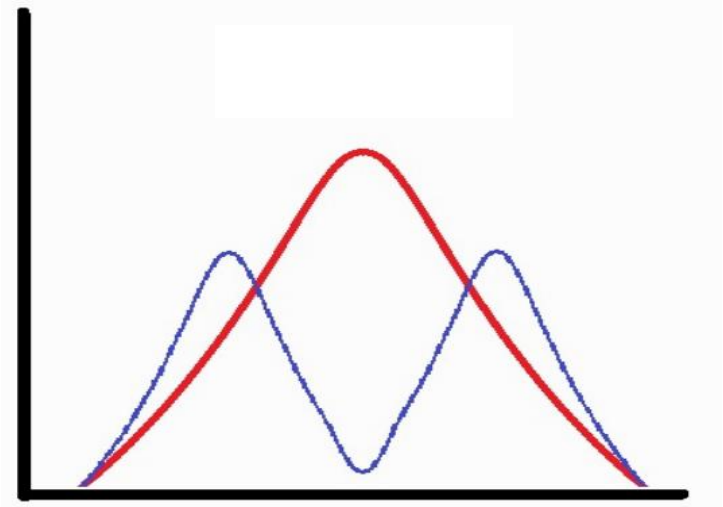
Bimodality in normal quantile-quantile plot

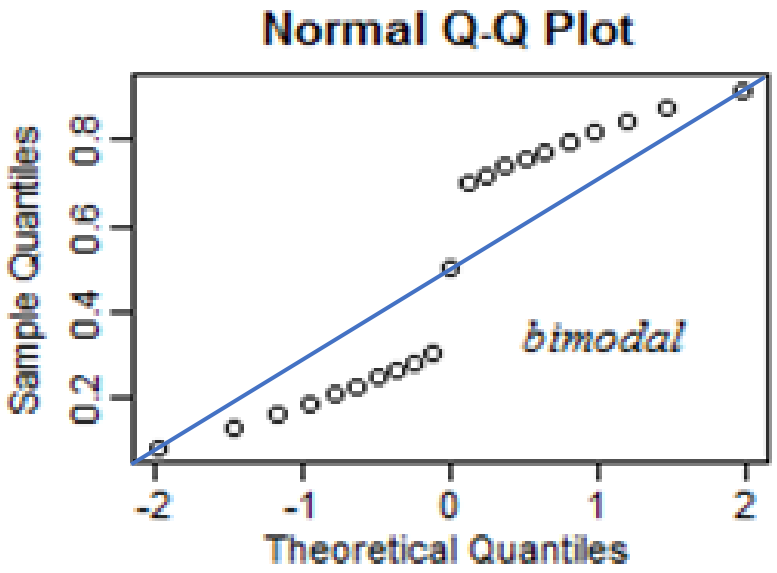


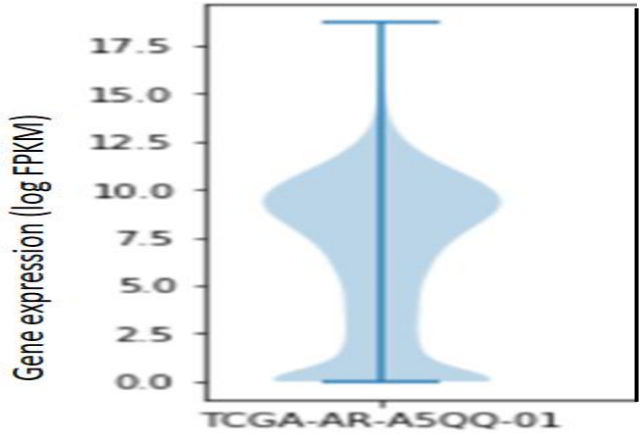


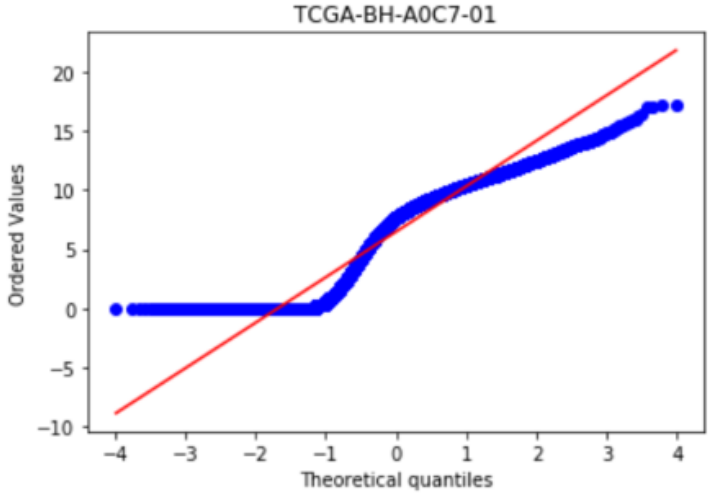
Adapted from https://stats.stackexchange.com/questions/101274/how-to-interpret-a-qq-plot & https://helenowskiirene.wordpress.com/tag/bimodal-distribution/

Bimodality in normal quantile-quantile plot





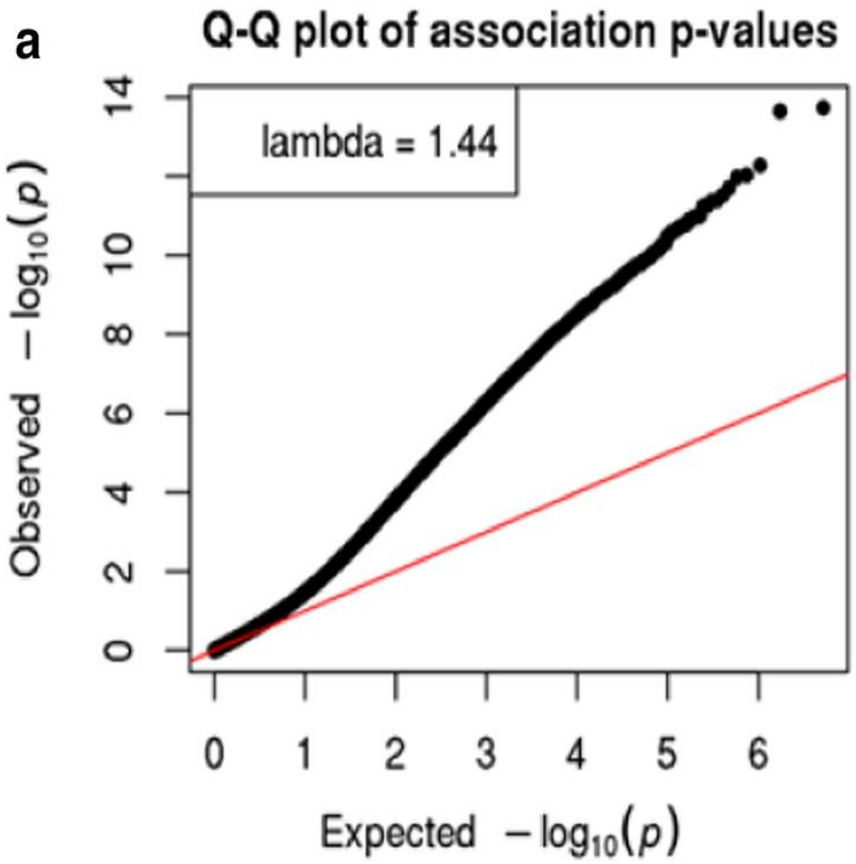


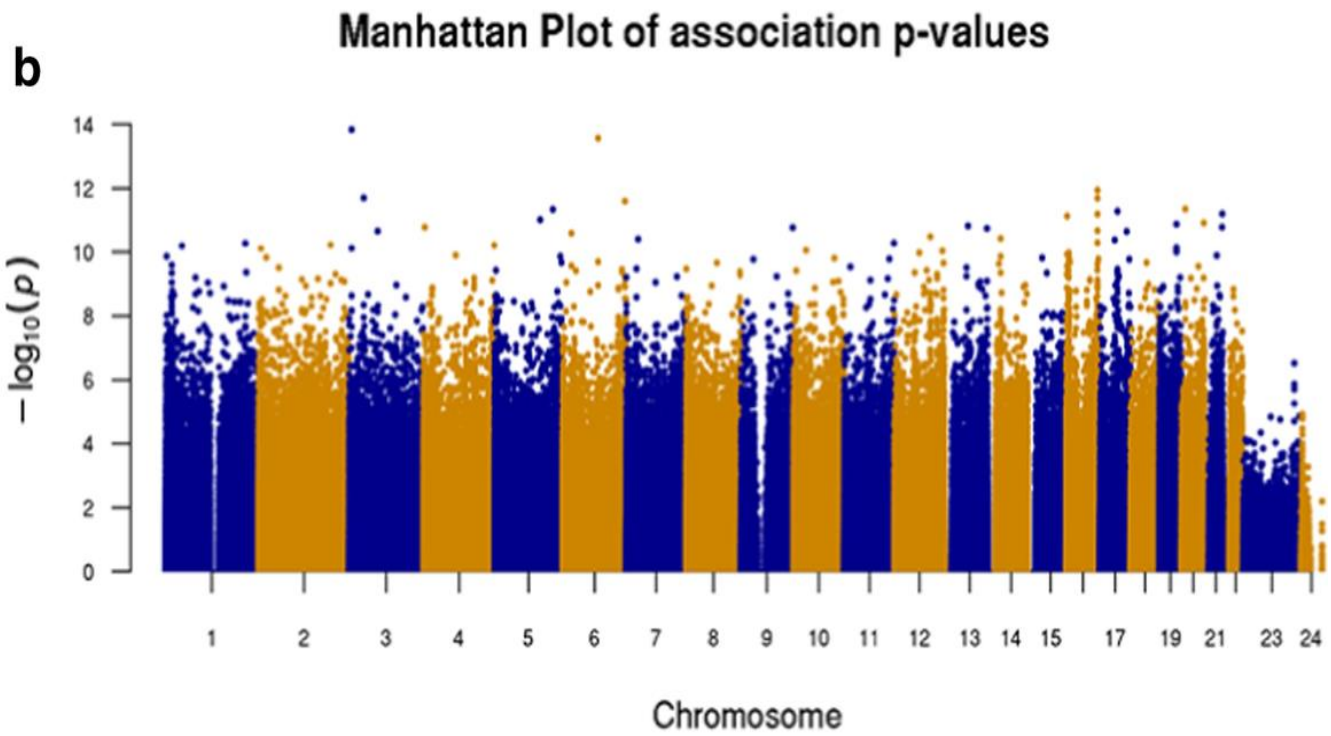


Adapted from https://stats.stackexchange.com/questions/101274/how-to-interpret-a-qq-plot & https://helenowskiirene.wordpress.com/tag/bimodal-distribution/

Quantile-quantile plot can be drawn for any distribution

Under the null hypothesis, p-values are uniformly distributed between 0 and 1

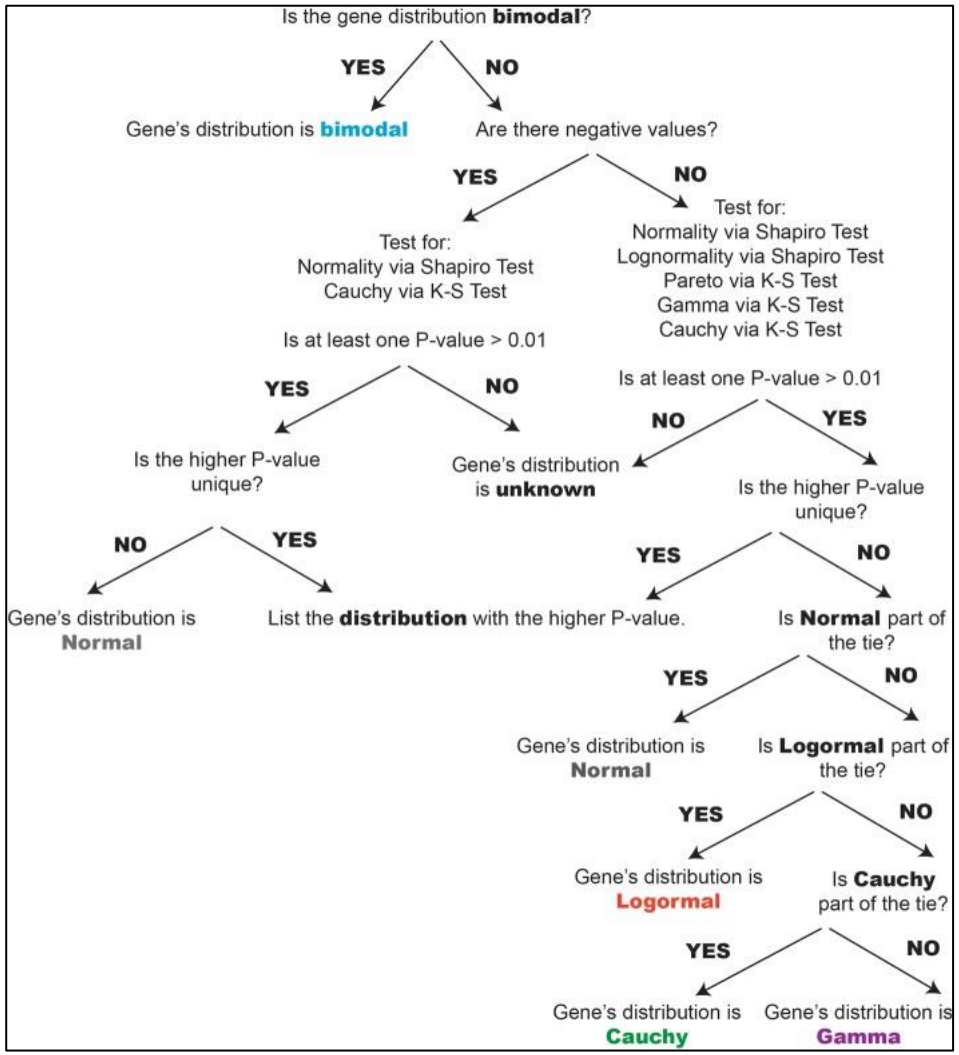




Cao et al. Clinical Epigenetics, 12, Article number: 192 (2020)

How to tell the underlying distribution?

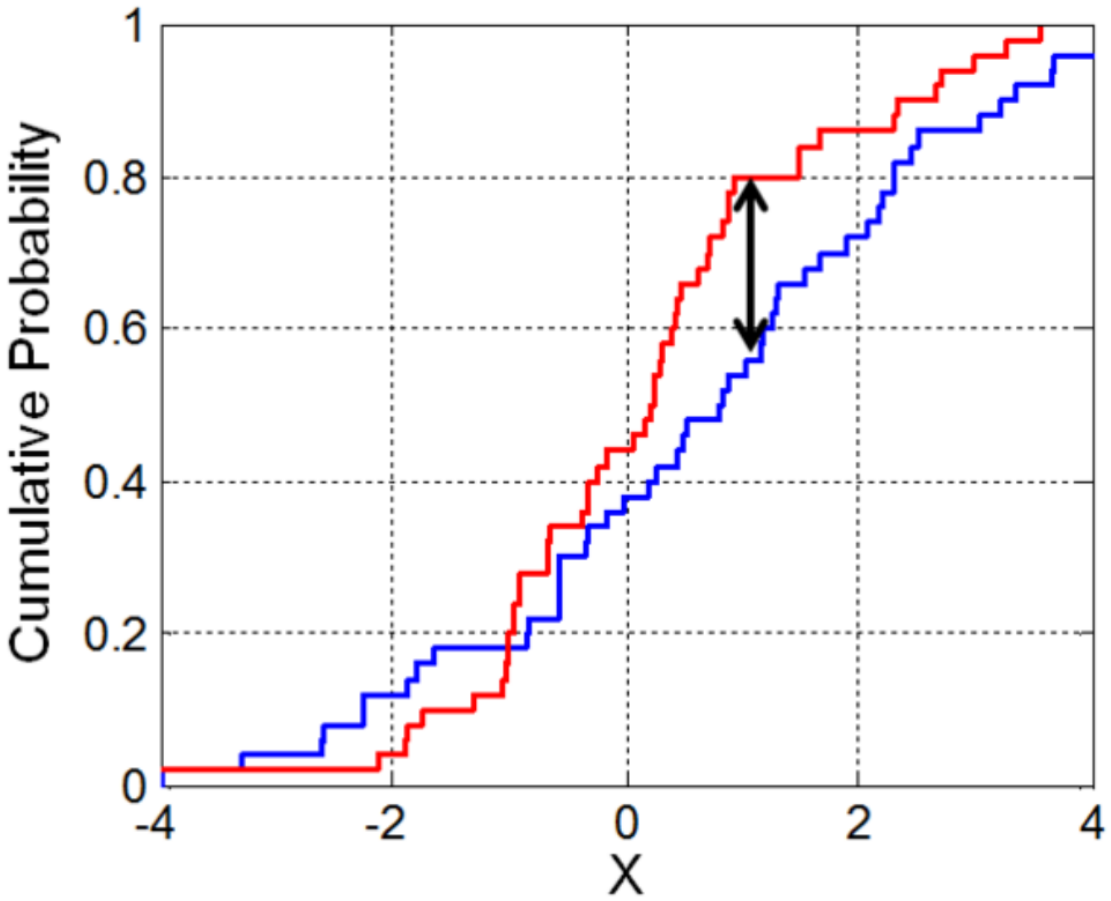
A quantitative way to describe the underlying distribution <u>by looking at data</u>?



Torrenté et al. BMC Bioinformatics, 2020

K-S test identifies a distribution in reference to the shape of theoretical distributions

- Kolmogorov-Smirnov test
  - Test H_0 that data came from a specified null distribution
  - Uses distance between data's eCDF the null CDF
  - $P < 0.05$ → reject H
  - Can be handy for "identifying" a distribution (i.e., a possibly good model)
- Shapiro-Wilk test is specifically for testing normality



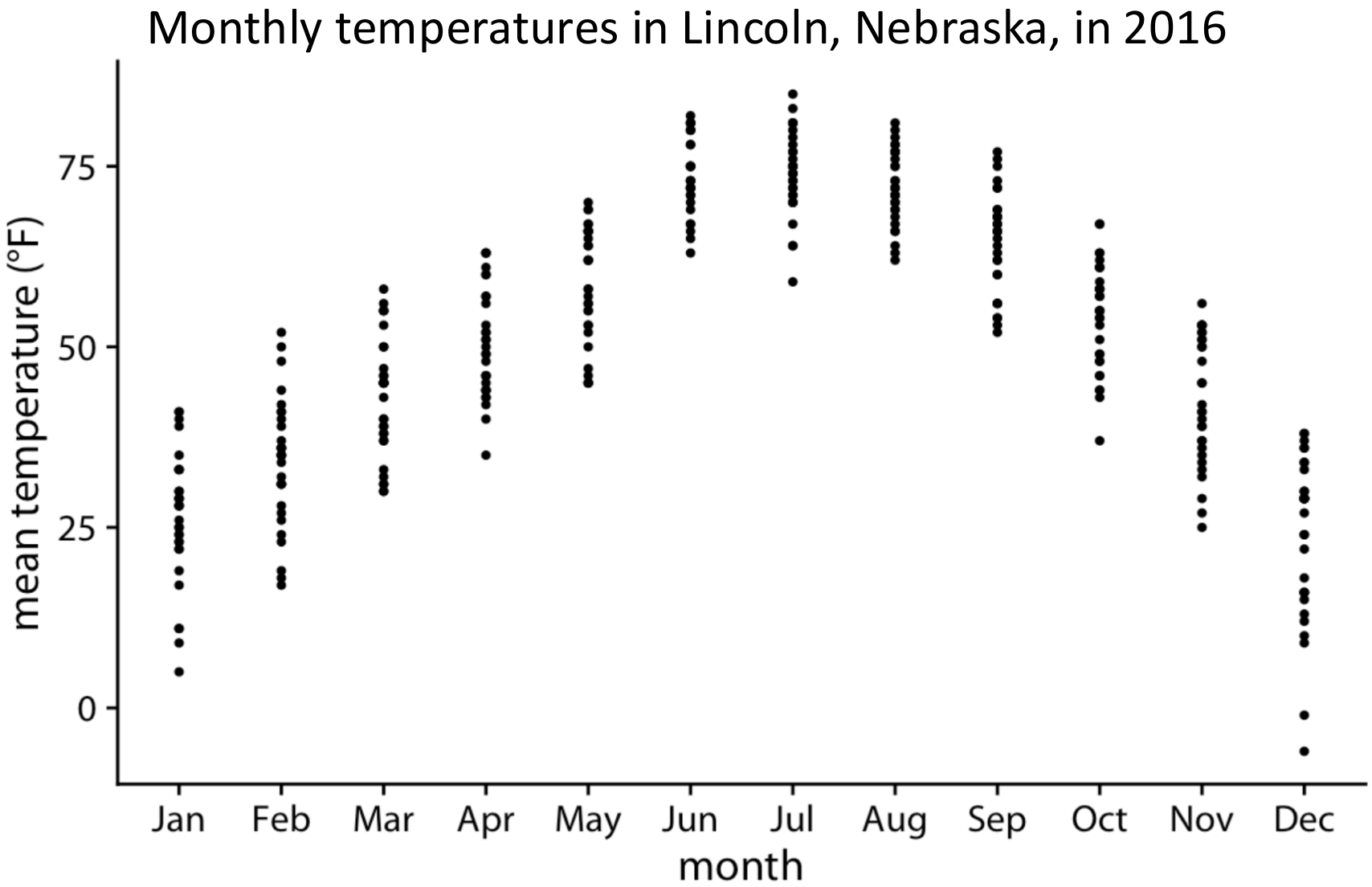
https://en.wikipedia.org/wiki/Kolmogorov%E2%80%93Smirnov_test

More plots for multiple distributions

- Boxplots
- Violin plots
- Strip chart
- Jittered strip charts
- Beeswarm plots
- Ridgeline plots

Many distributions in Strip chart

- Pros: actual data instead of estimates
- Cons: too many points lead to overplotting

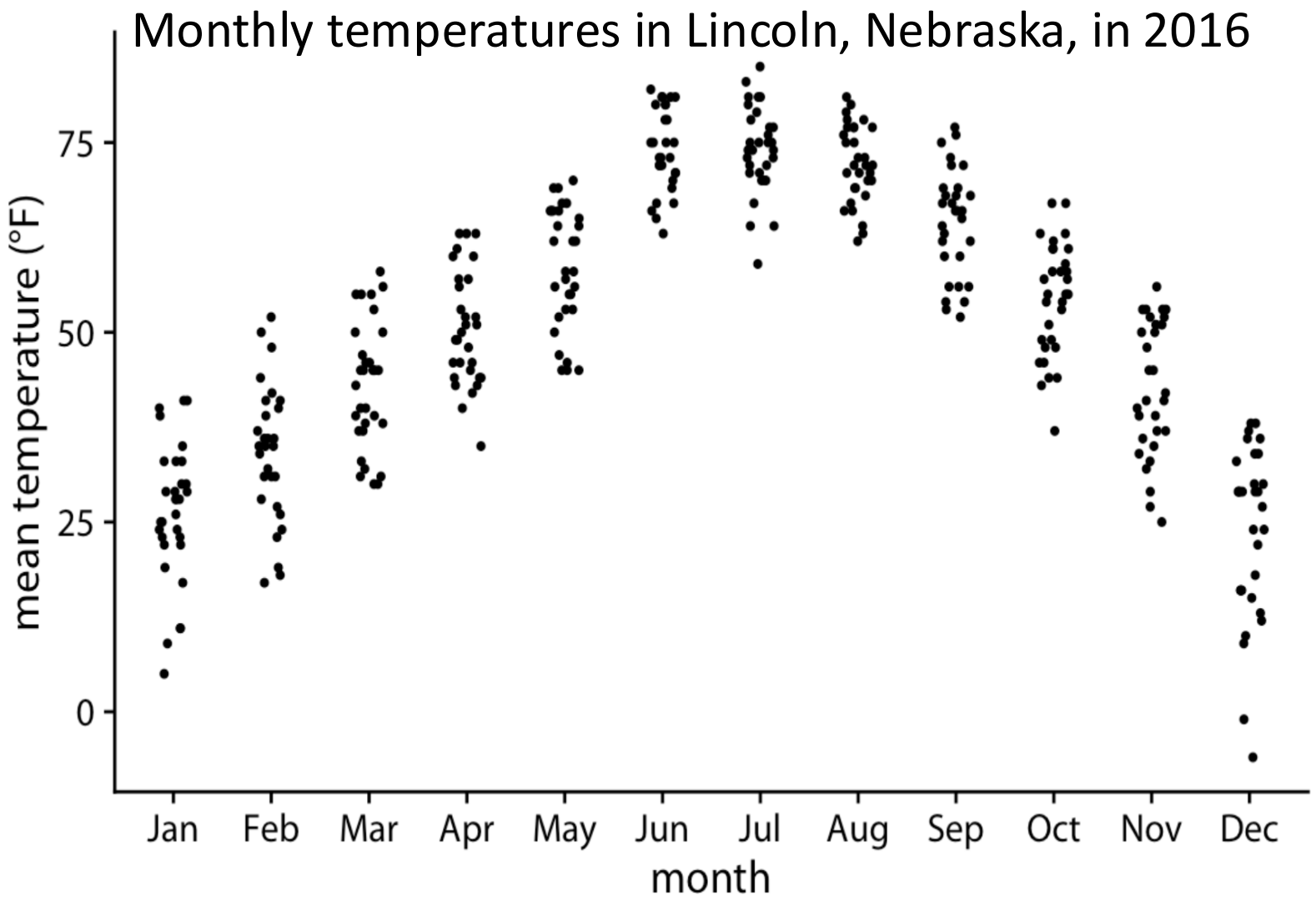


Fundamentals of Data Visualization, Wilke, O'Reilly, 1st Ed.

Many distributions in jittered strip chart

Jittering: slight irregular variation (< a constant)
Pros: solving the overplotting issue
Cons:

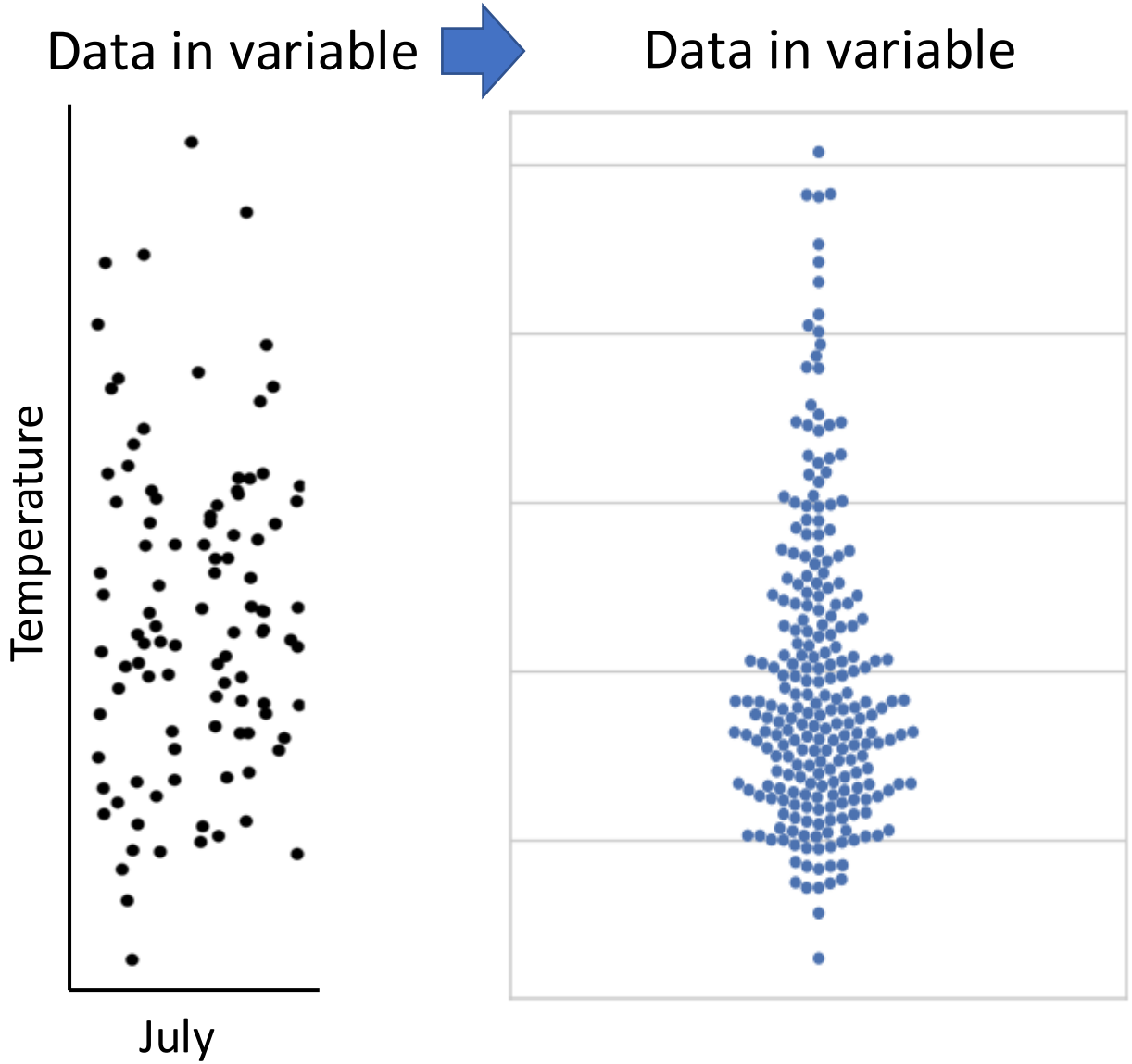
- Changes the data
- If too much, it will send the dots too far
- Not easy to read multiple data points
- Takes up space
- Hard to see "density"



Fundamentals of Data Visualization, Wilke, O'Reilly, 1st Ed.

Many distributions in (bee)swarm plot

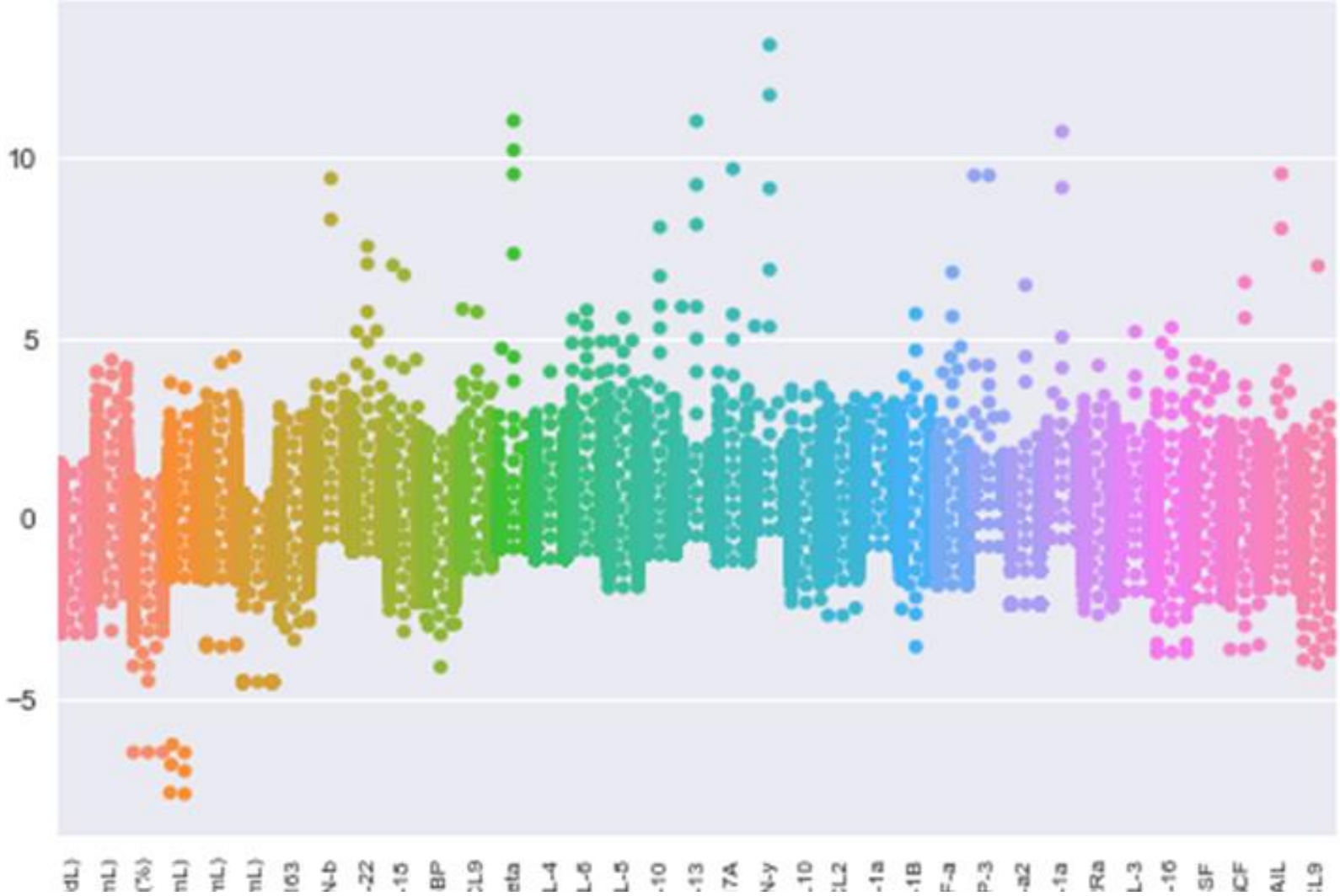
- Pros: helps you to see actual distribution over y-axis
- Pros: more efficient than jittered strip chart
- Cons: still non-data visualization along x-axis



https://seaborn.pydata.org/generated/seaborn.swarmplot.html

Not ideal for too many variables

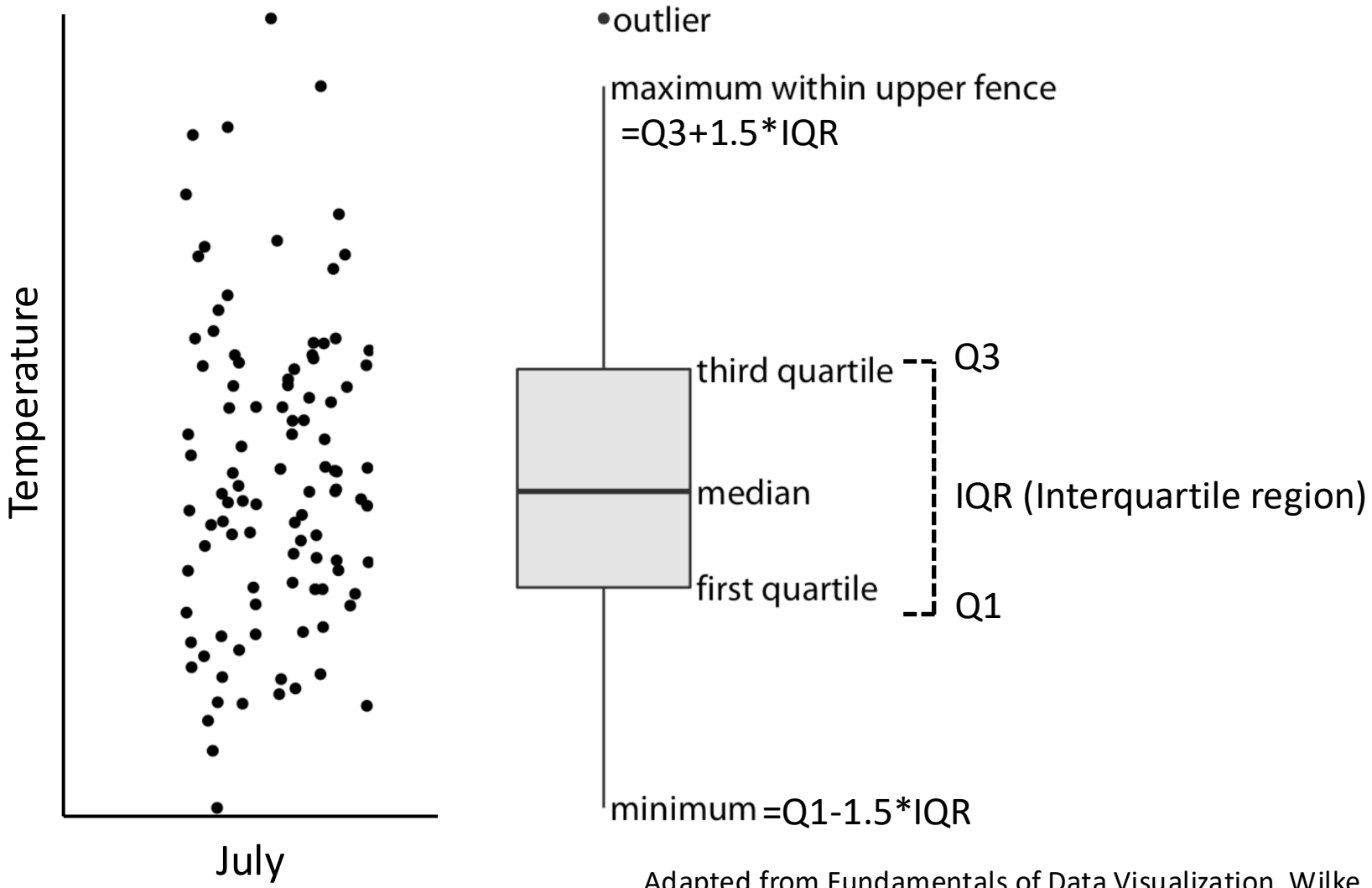
- 404 data points for 36 clinical features
- Same with the plots showing actual data (e.g. jittered strip chart)
- Time to use plots showing statistical properties, e.g. boxplot and violinplot



Unpublished

Distribution in boxplot

Data in variable → Descriptive statistics of data

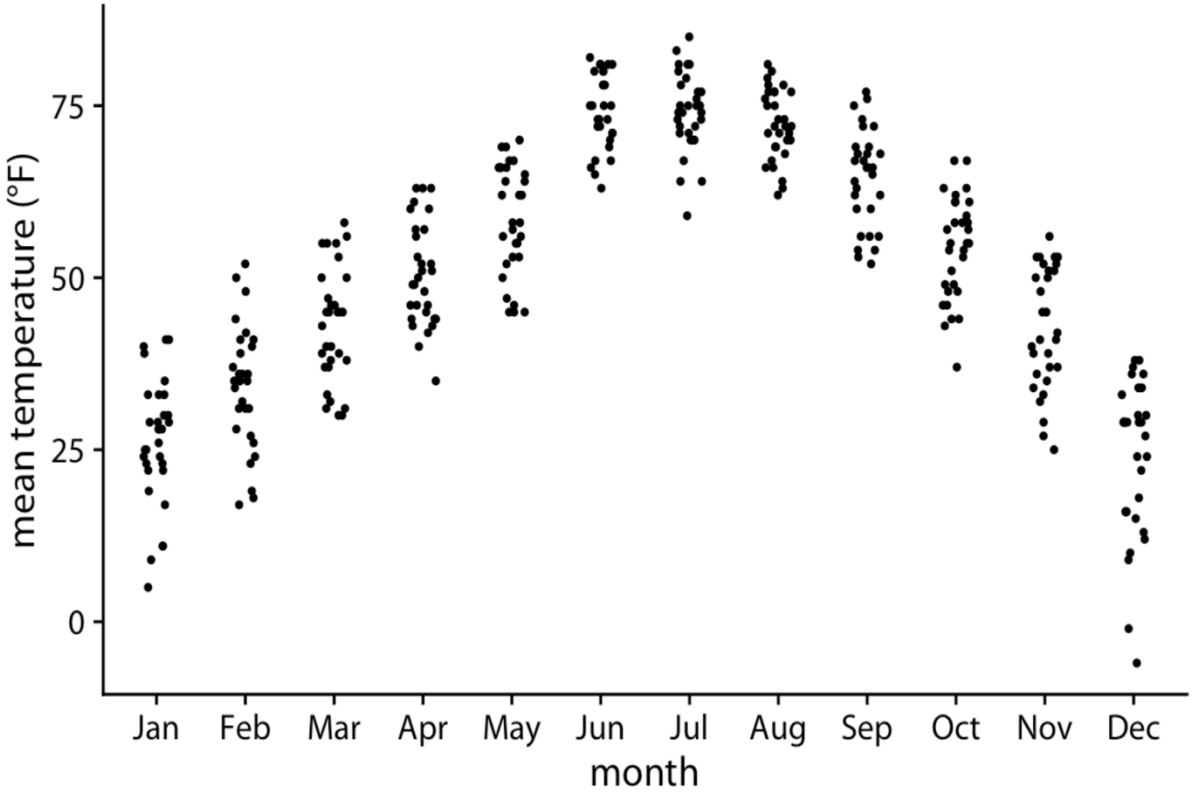


Adapted from Fundamentals of Data Visualization, Wilke, O'Reilly, 1st Ed.

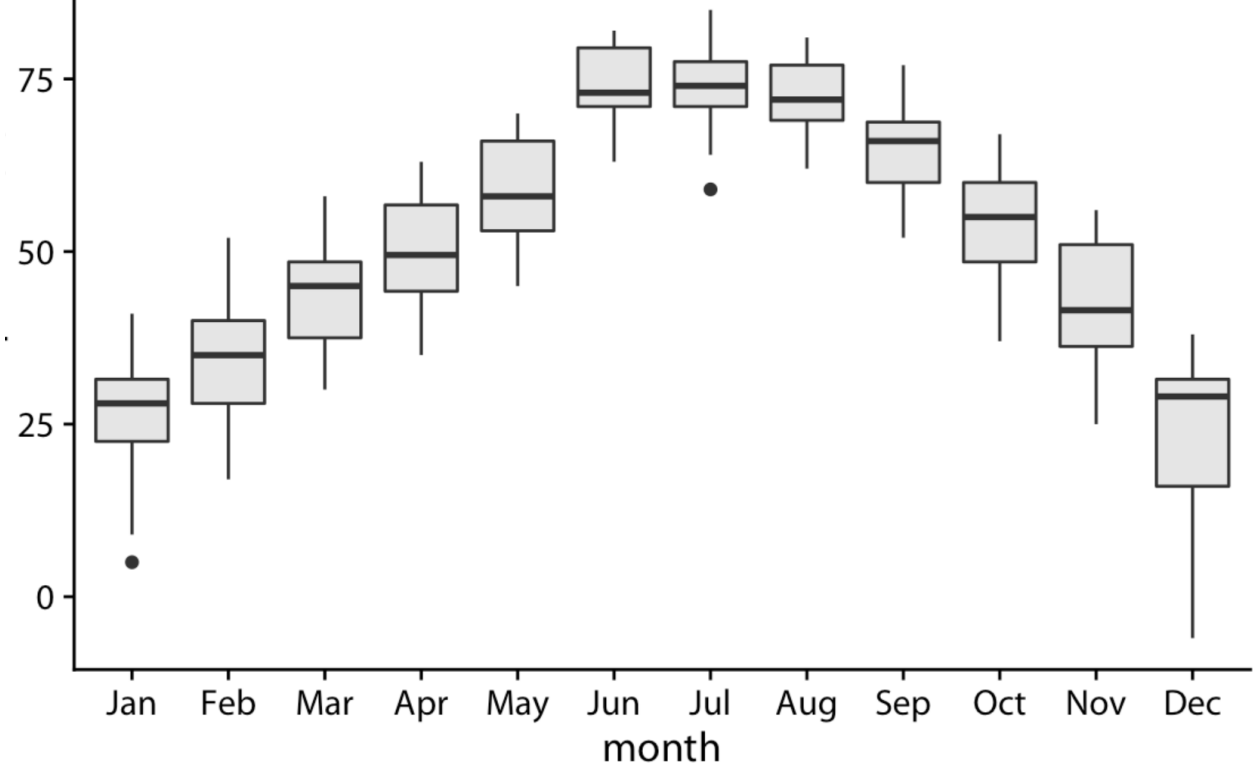
Many distributions in Boxplot

With median, IQR, and whiskers, can easily read patterns in the multiple distributions. e.g.

- Skewness in December
- No skewness in July



Monthly temperatures in Lincoln, Nebraska, in 2016



Fundamentals of Data Visualization, Wilke, O'Reilly, 1st Ed.

Boxplot helps us see multiple aspects of data

Values of usual interest

- Median

e.g. monotonical increase

- IQR

e.g. words with large/small IQR

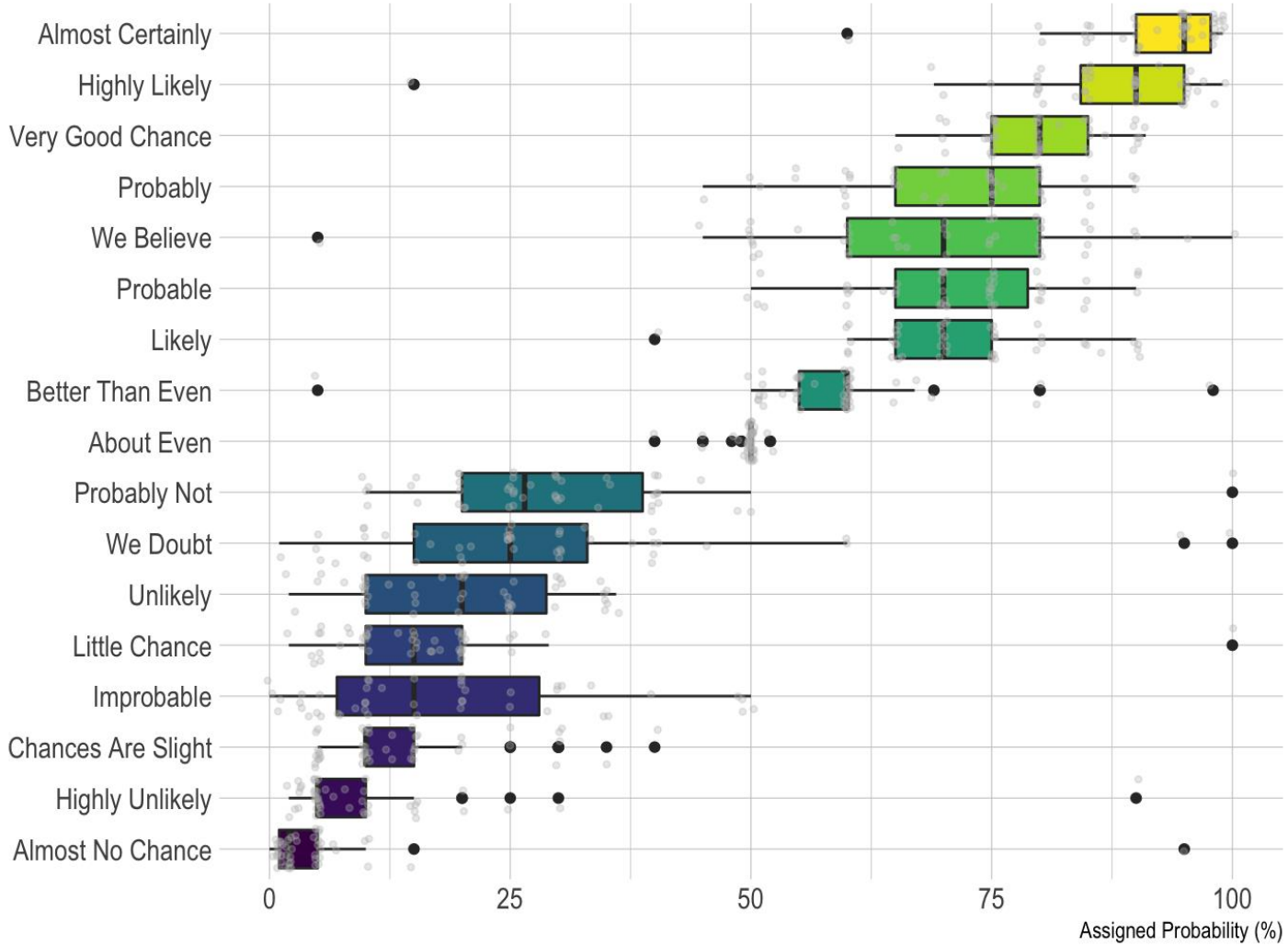
- Skewness

e.g. Better Than Even

- Outliers

e.g. words with many/few outliers

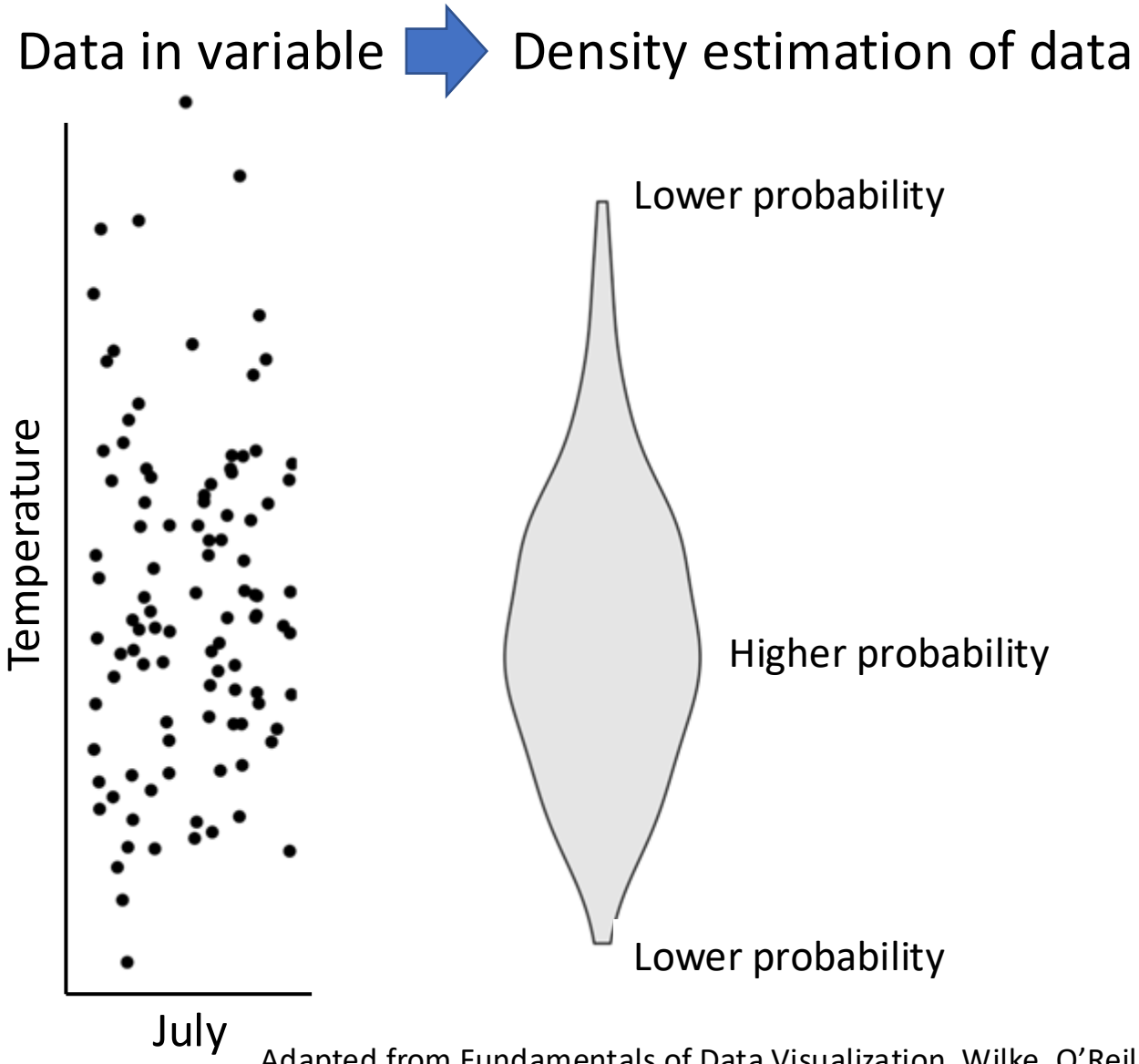
Raw data can be overlaid!



Fundamentals of Data Visualization, Wilke, O'Reilly, 1st Ed.

Violin plots

- Boxplots replaced by violin plots
- Pros
  - More nuanced picture
- Cons
  - Estimation, not data

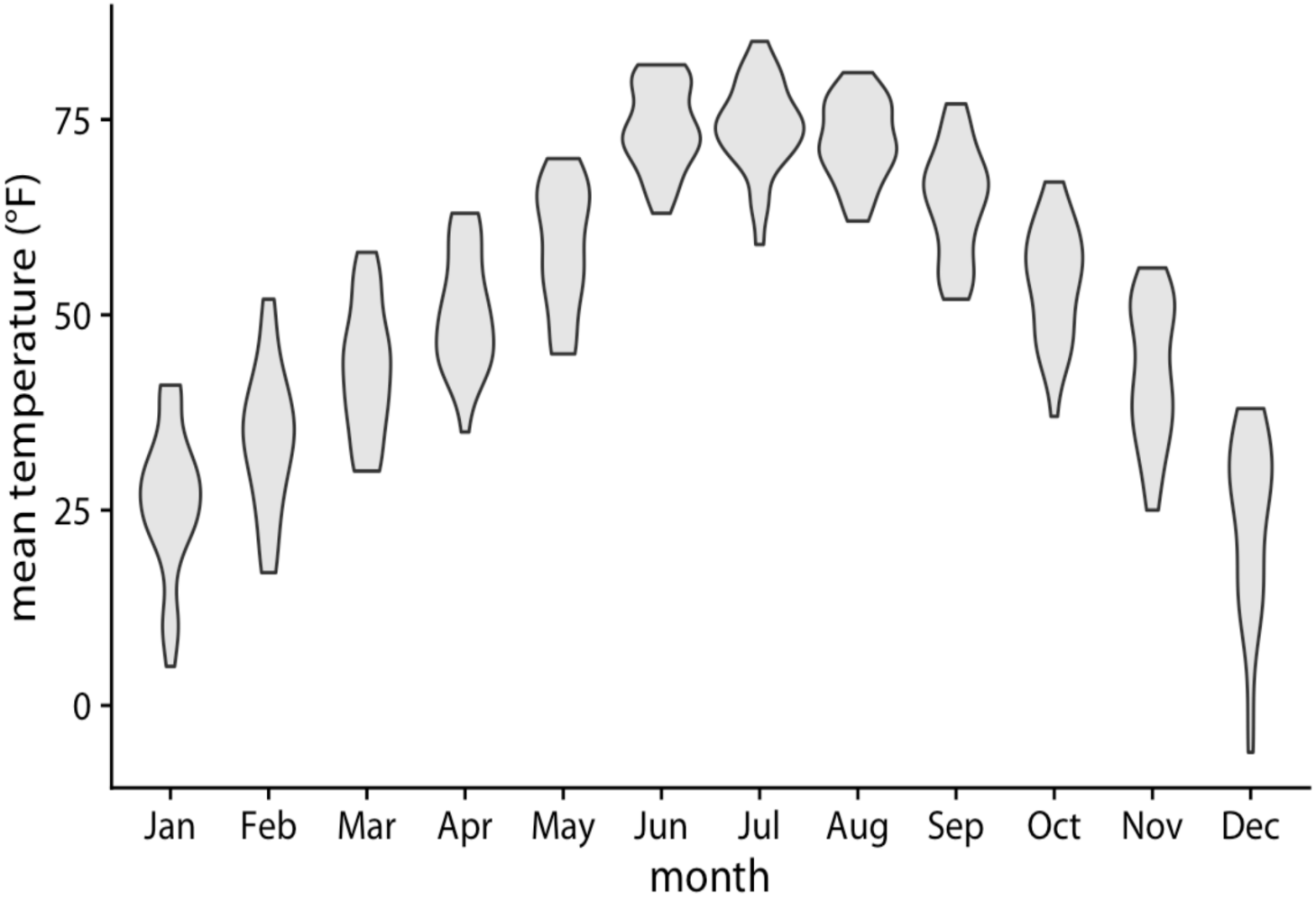


Adapted from Fundamentals of Data Visualization, Wilke, O'Reilly, 1st Ed.

Many distributions in Violin plot

- Pros: show distribution shape
e. g., unimodal vs. bimodal
- Cons: does not show summary statistics, difficult to compare

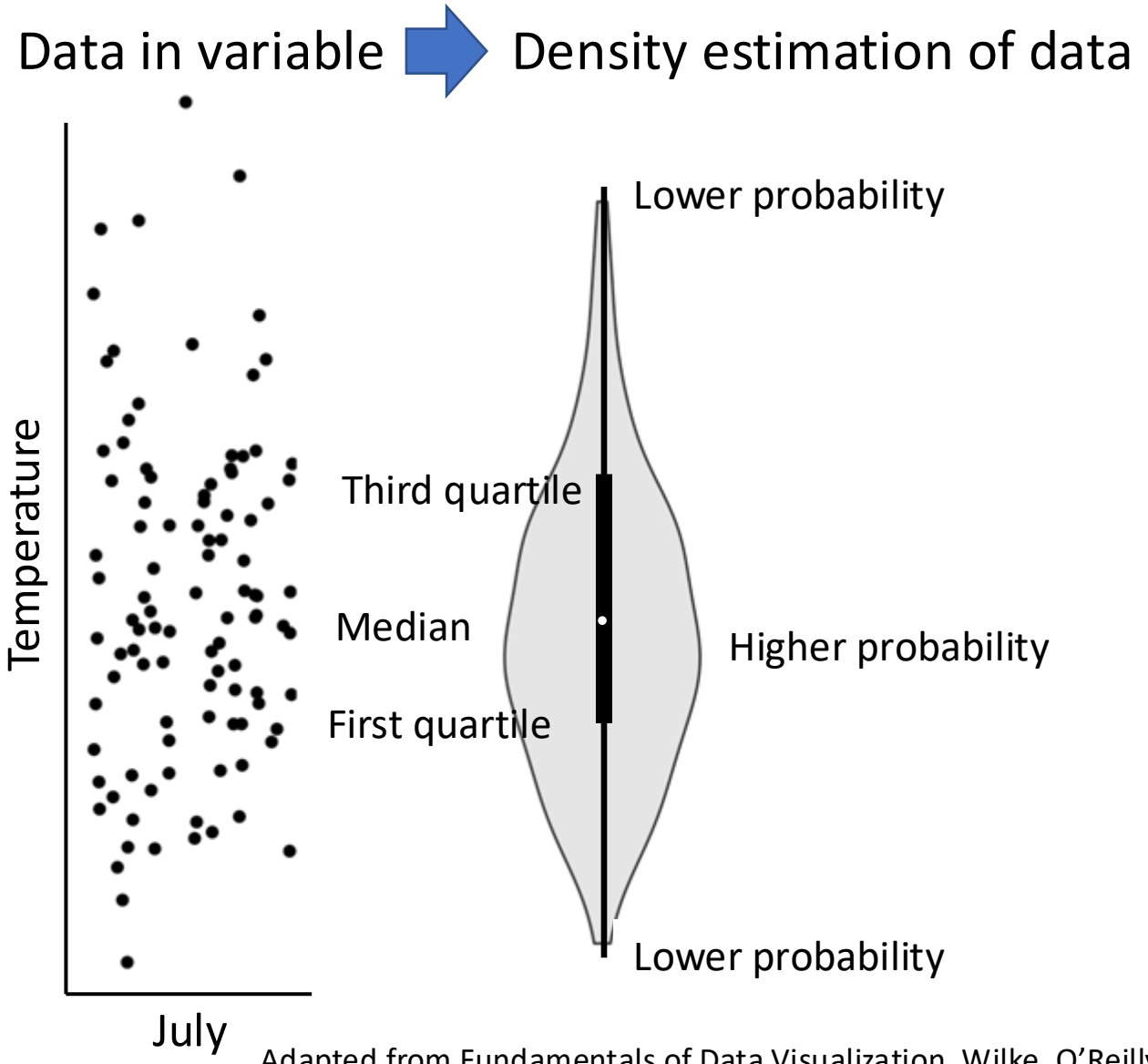
Monthly temperatures in Lincoln, Nebraska, in 2016



Fundamentals of Data Visualization, Wilke, O'Reilly, 1st Ed.

Distribution in violin plot

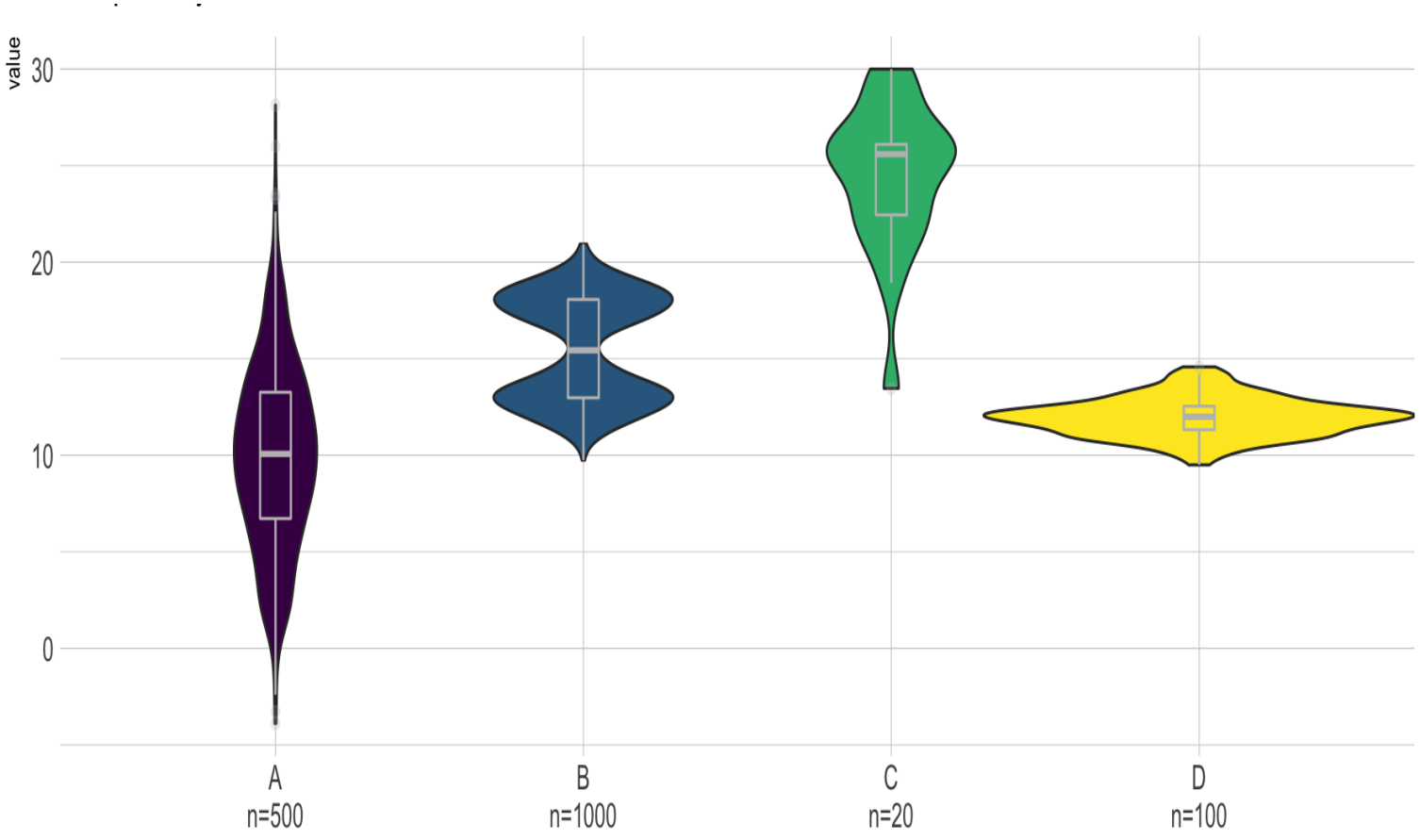
- Boxplots replaced by violin plots
- Pros
  - More nuanced picture
- Cons
  - Largely estimation, not data



Adapted from Fundamentals of Data Visualization, Wilke, O'Reilly, 1st Ed.

Distribution with different sample size has different statistical power

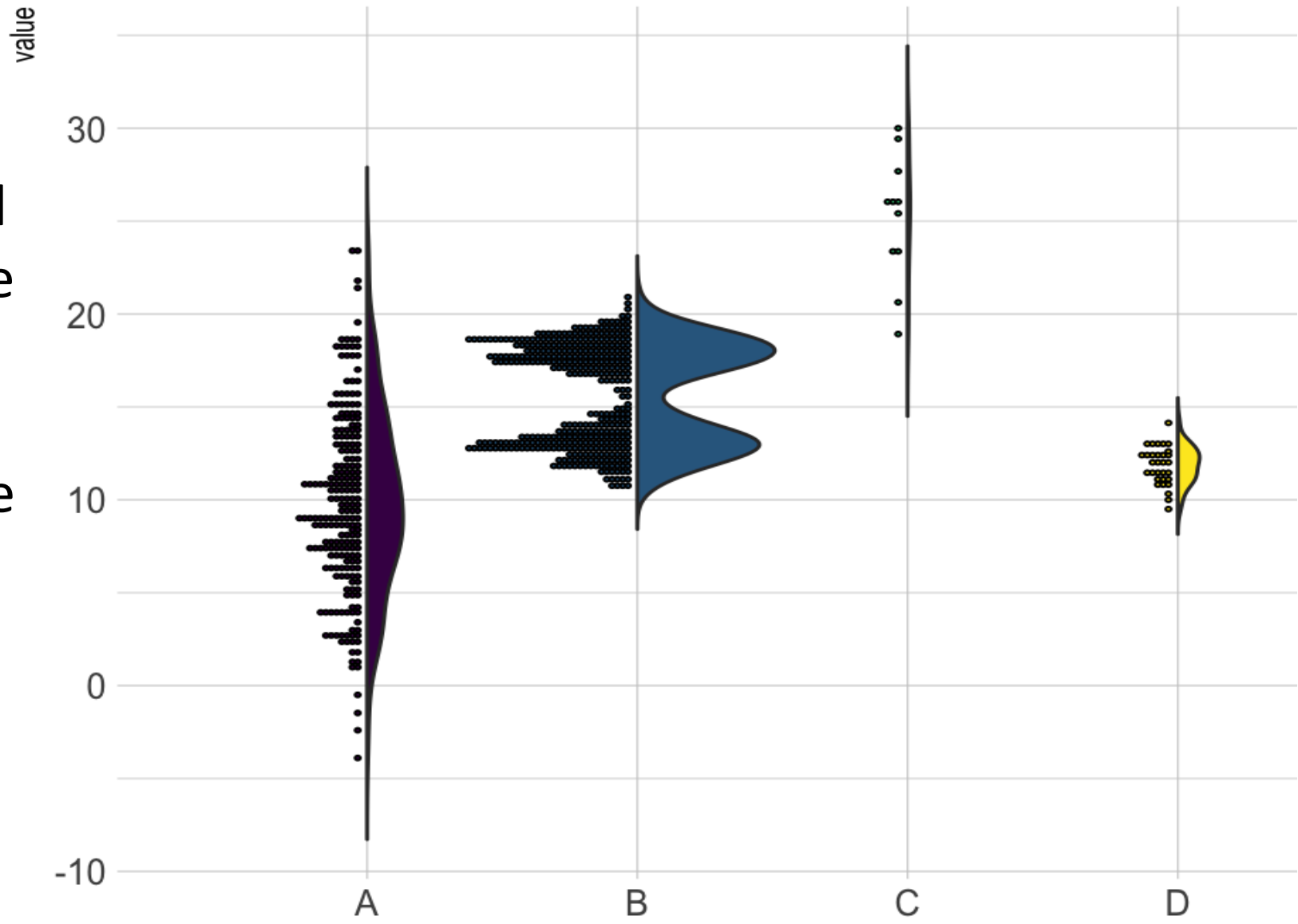
- Between the shape of B and C, which would you be more confident about?
- To estimate the statistical power, annotate sample size whenever possible



https://www.data-to-viz.com/caveat/boxplot.html

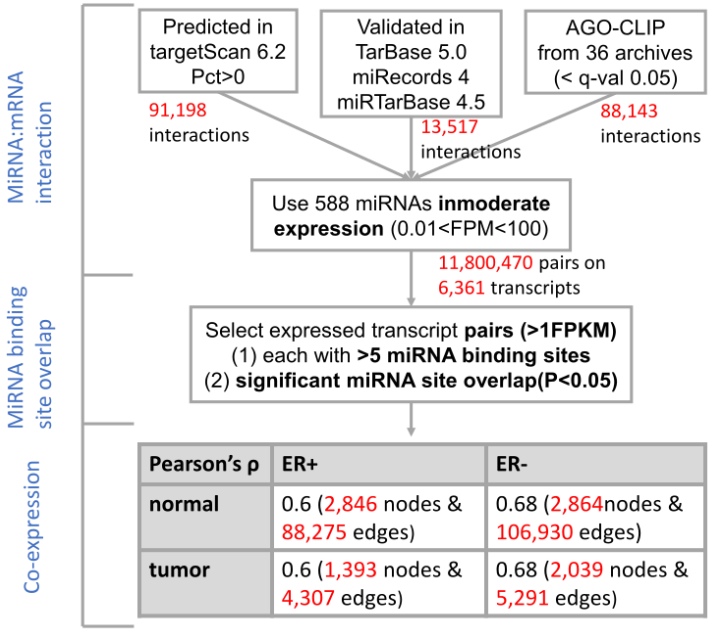
Violinplot with sample size

- Between the shape of B and C, which would you be more confident about?
- To estimate the statistical power, annotate sample size whenever possible

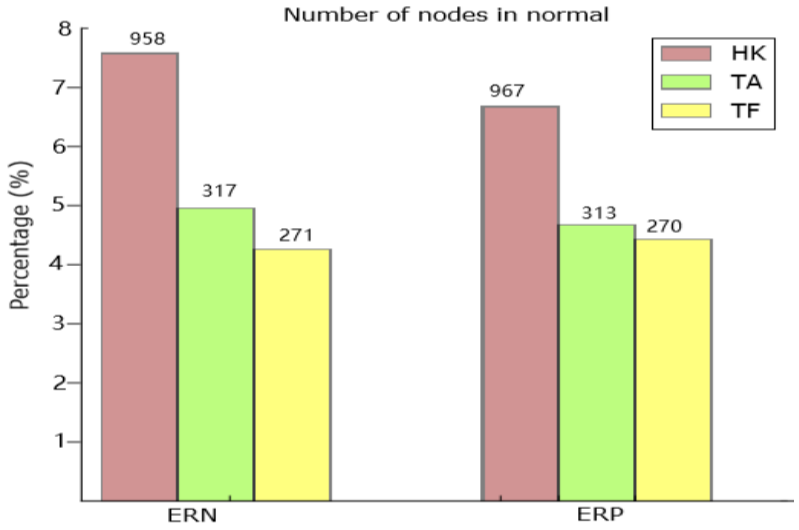


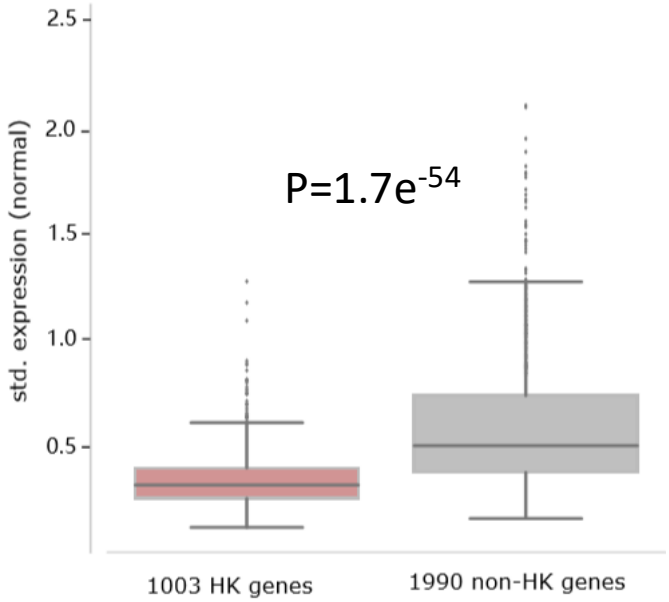
https://www.data-to-viz.com/caveat/boxplot.html

Data visualization with sample size



Pearson's ρ	ER+	ER-
normal	0.6 (2,846 nodes & 88,275 edges)	0.68 (2,864nodes & 106,930 edges)
tumor	0.6 (1,393 nodes & 4,307 edges)	0.68 (2,039 nodes & 5,291 edges)



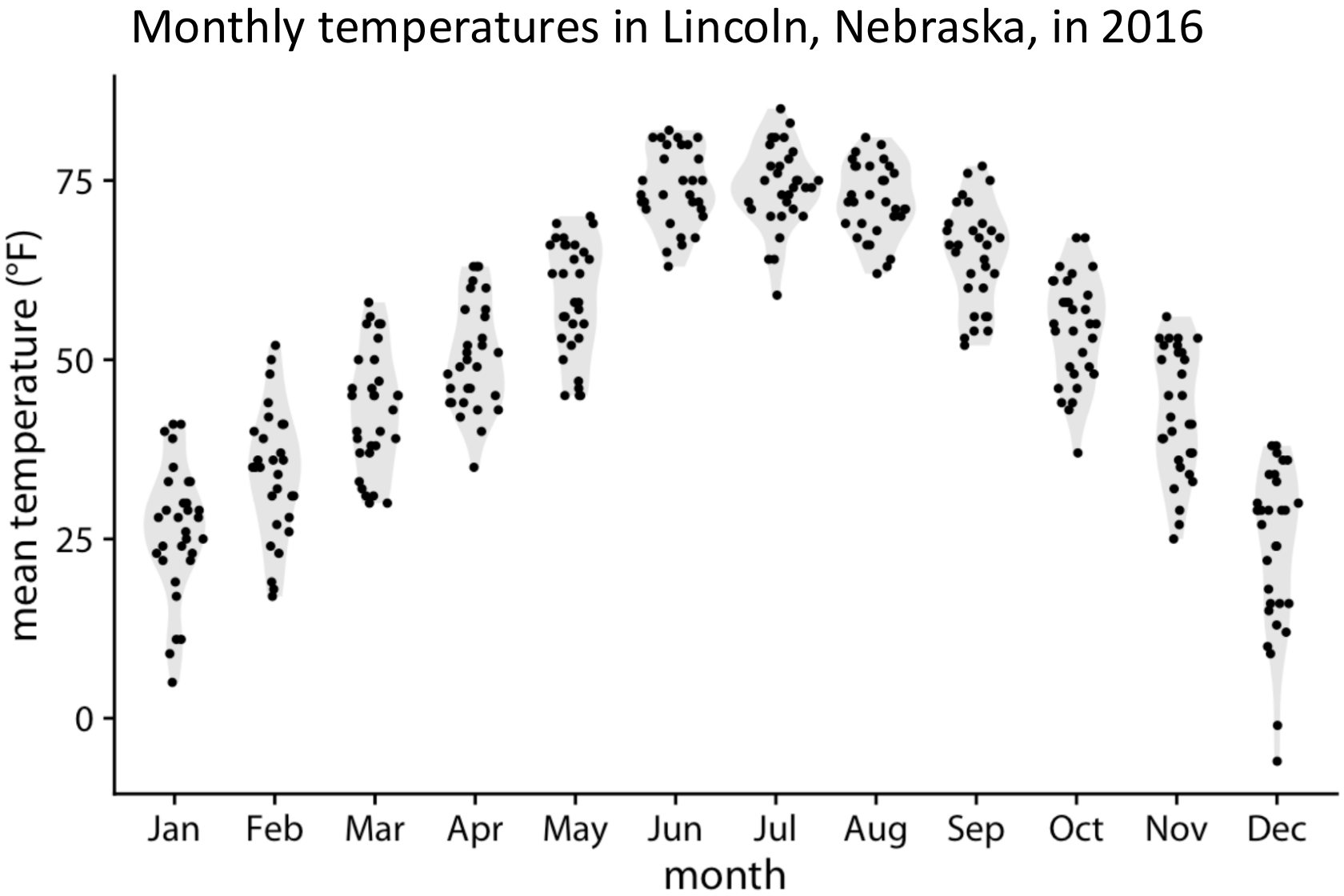


Zhenjiang et al. In revision

Many distributions in jittered strip chart and violin plot

Hybrid of

- violin and jittered strip (a.k.a. sina plot)
- violin and box plot

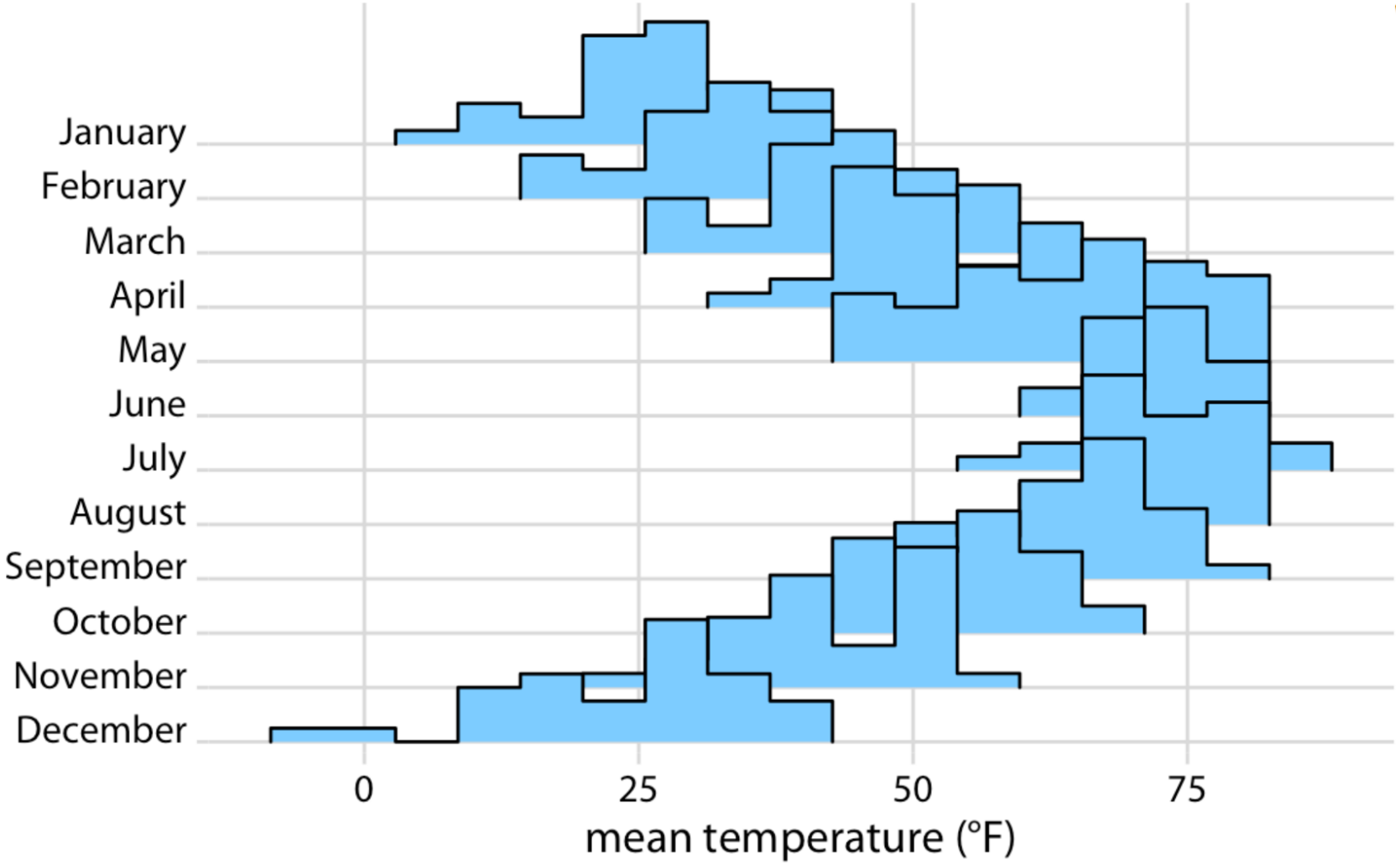


Fundamentals of Data Visualization, Wilke, O'Reilly, 1st Ed.

Histogram is not ideal for distributions over many variables

- Cons: the vertical lines running together
- Solution:
  - Spread the variables
  - A visualization that doesn't require vertical lines to be drawn

Monthly temperatures in Lincoln, Nebraska, in 2016

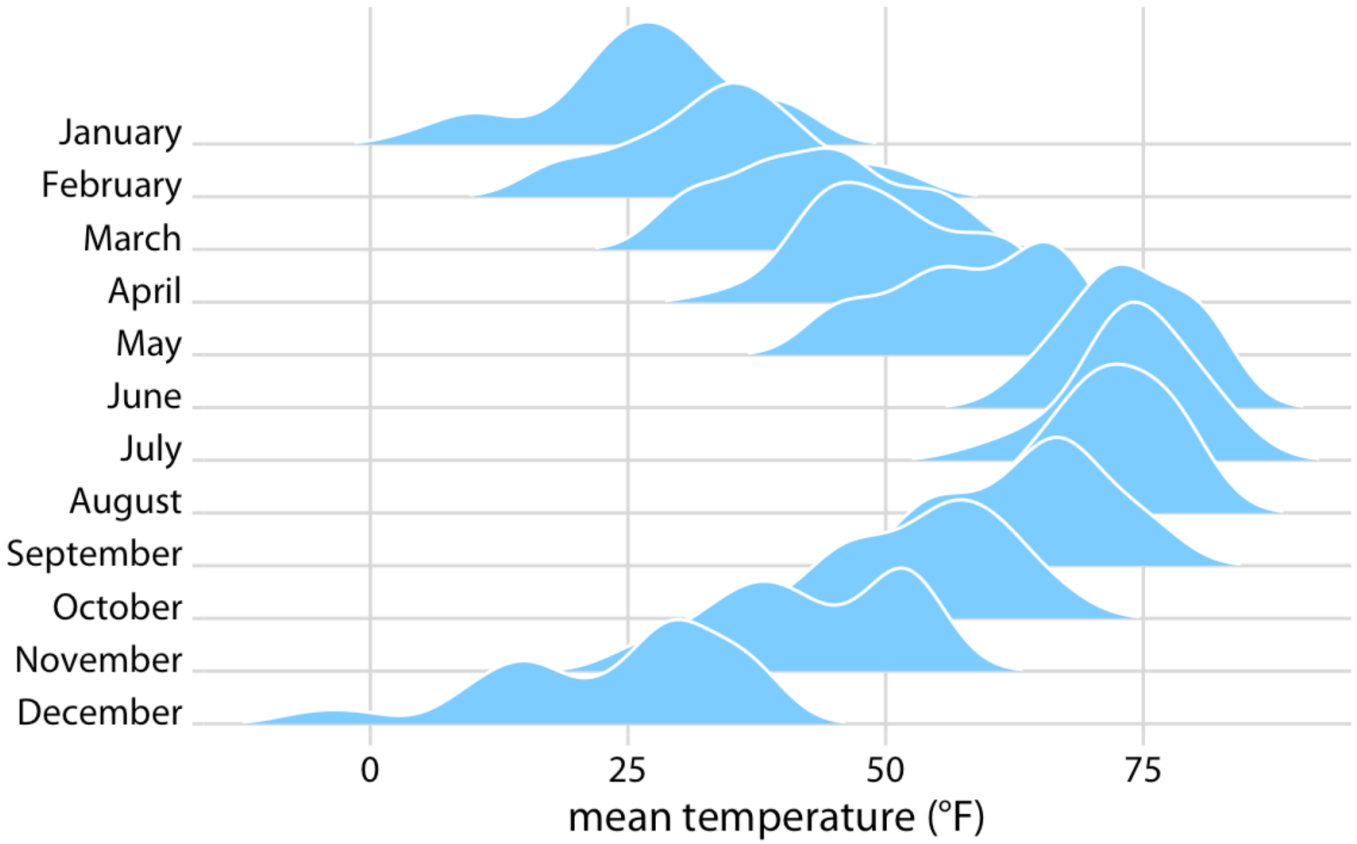


Fundamentals of Data Visualization, Wilke, O'Reilly, 1st Ed.

Many distributions in ridgeline plot

- Density estimates like violin plot with x-axis showing the data values and the y-axis showing the variables
- Pros: work well to show trends in distributions across variables (e. g., over time)
- Overlap usually discouraged to use, but maybe OK with estimate visualization

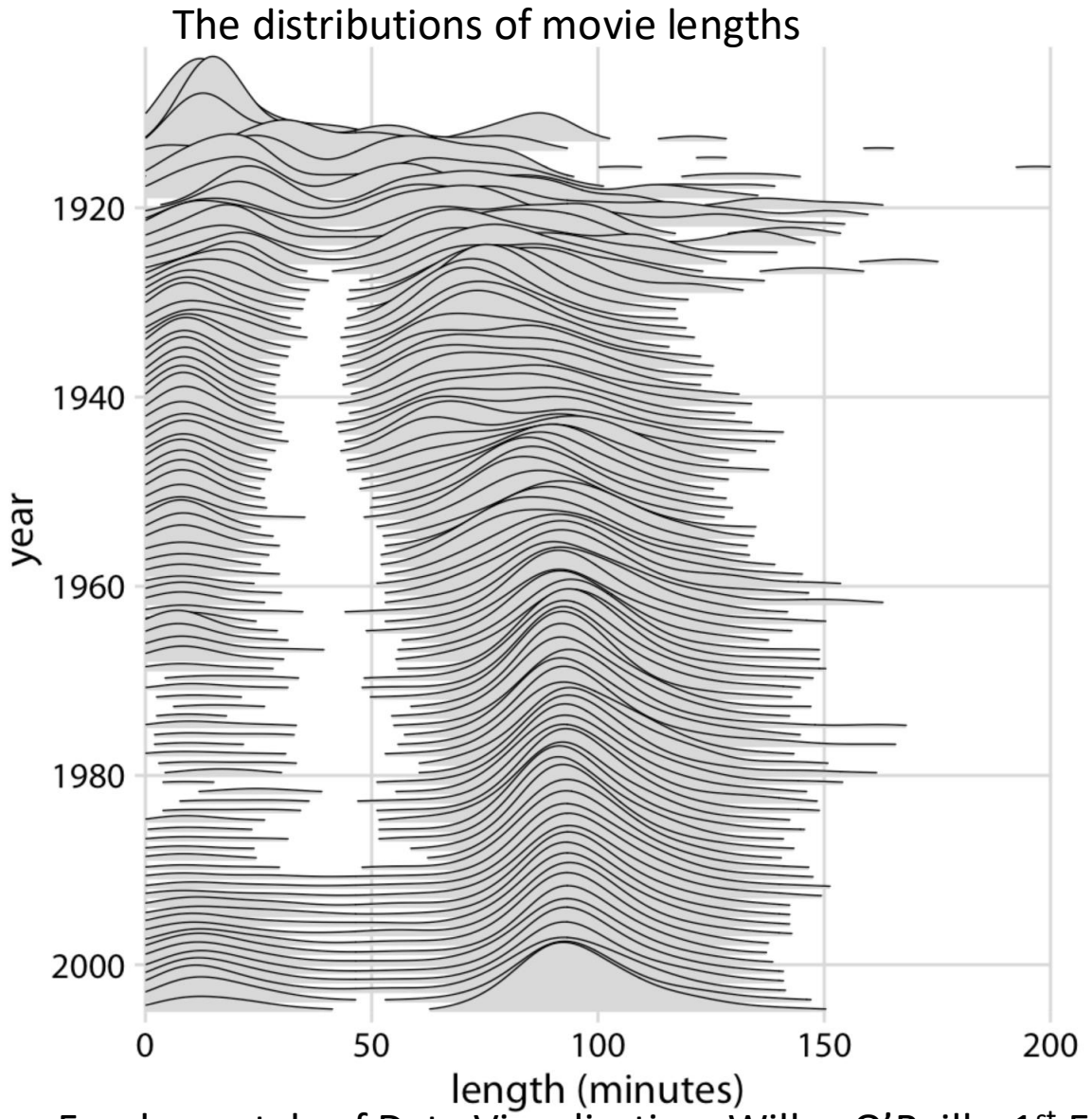
Monthly temperatures in Lincoln, Nebraska, in 2016



Fundamentals of Data Visualization, Wilke, O'Reilly, 1st Ed.

Ridgeline plot for very large numbers of distributions

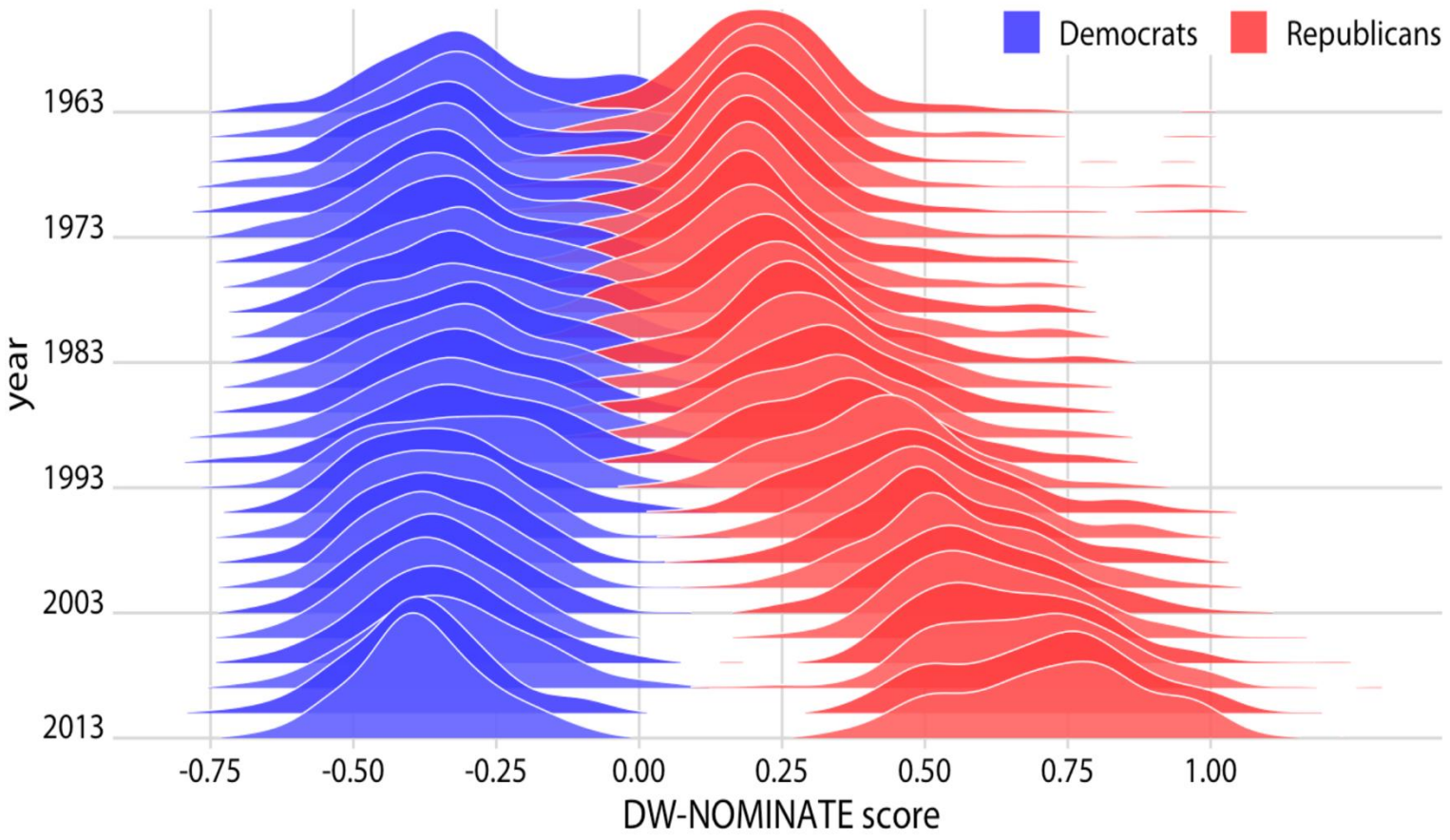
- Pros: scale to very large numbers of distributions, because it saves space compared to boxplot assuming gradual changes between neighboring variables



Fundamentals of Data Visualization, Wilke, O'Reilly, 1st Ed.

Ridgeline plot to compare many distributions

- Can compare two trends over times (color being another dimension)
- DW-NOMINATE scores shows voting patterns of the party congressmen.
- Expansion of ideological divides in Congress



Fundamentals of Data Visualization, Wilke, O'Reilly, 1st Ed.