Box-and-Whisker Plot

- A box plot is a graphical display showing Spread, Outliers, Center, and Shape (SOCS).
- It displays the 5-number summary: *min*, q_1, *median*, q_3, and *max*.

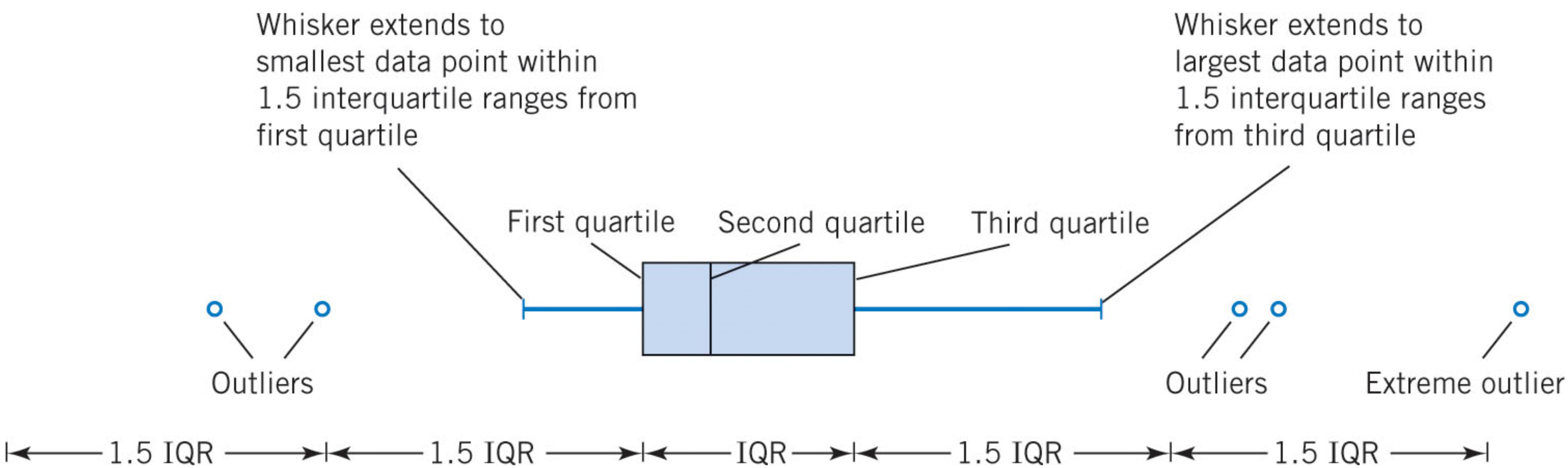
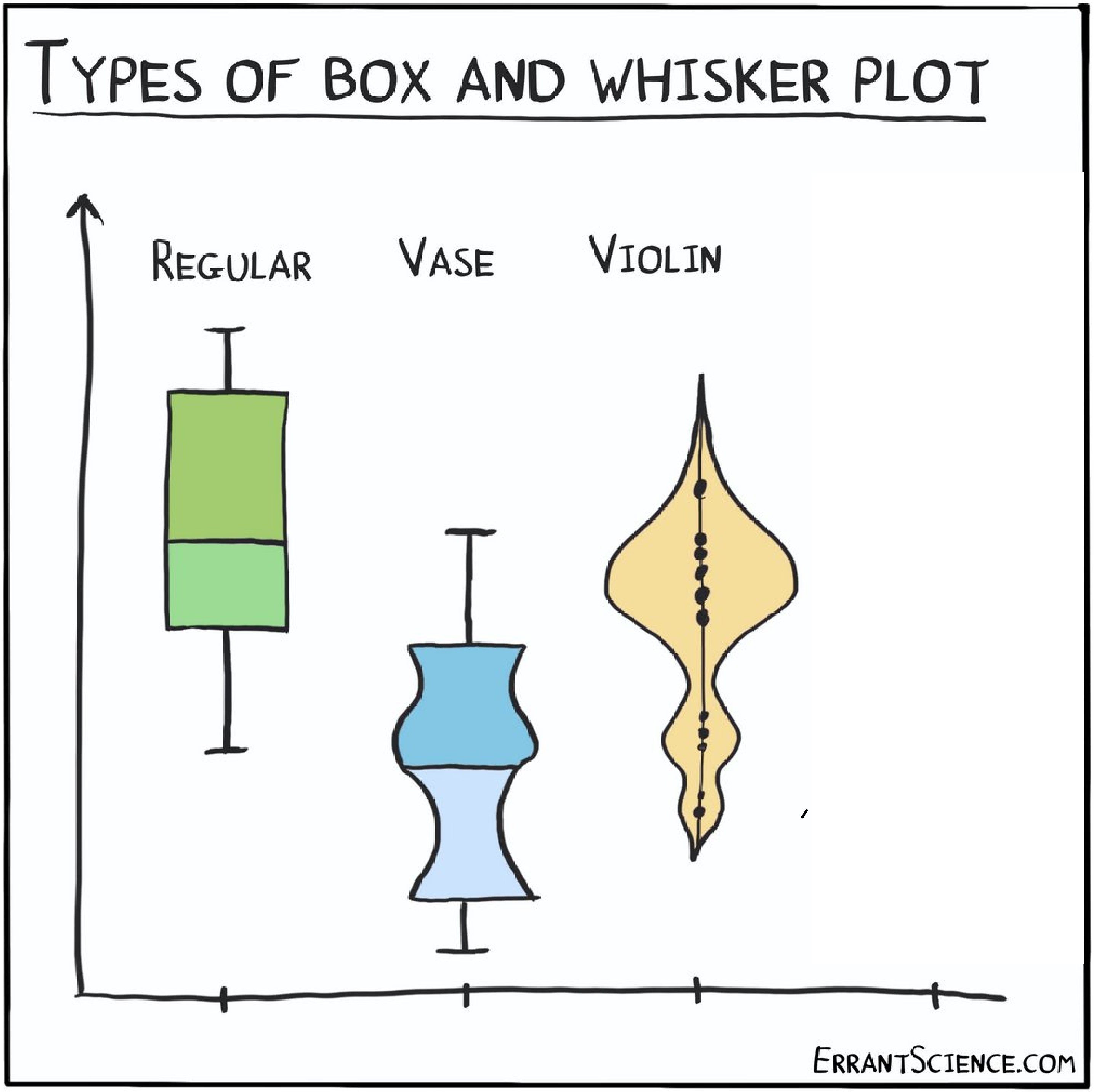


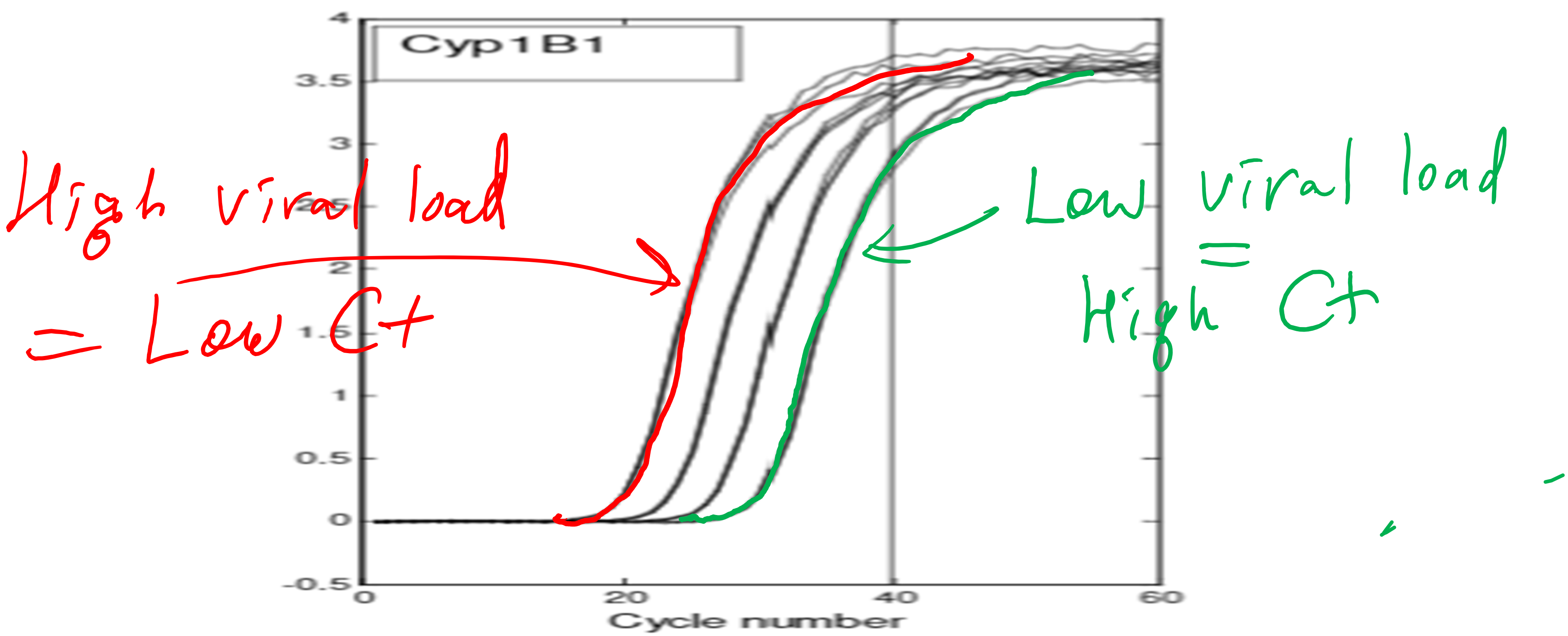
Figure 6-13 Description of a box plot.


TYPES OF BOX AND WHISKER PLOT
REGULAR
VASE
VIOLIN
ERRANTSCIENCE.COM

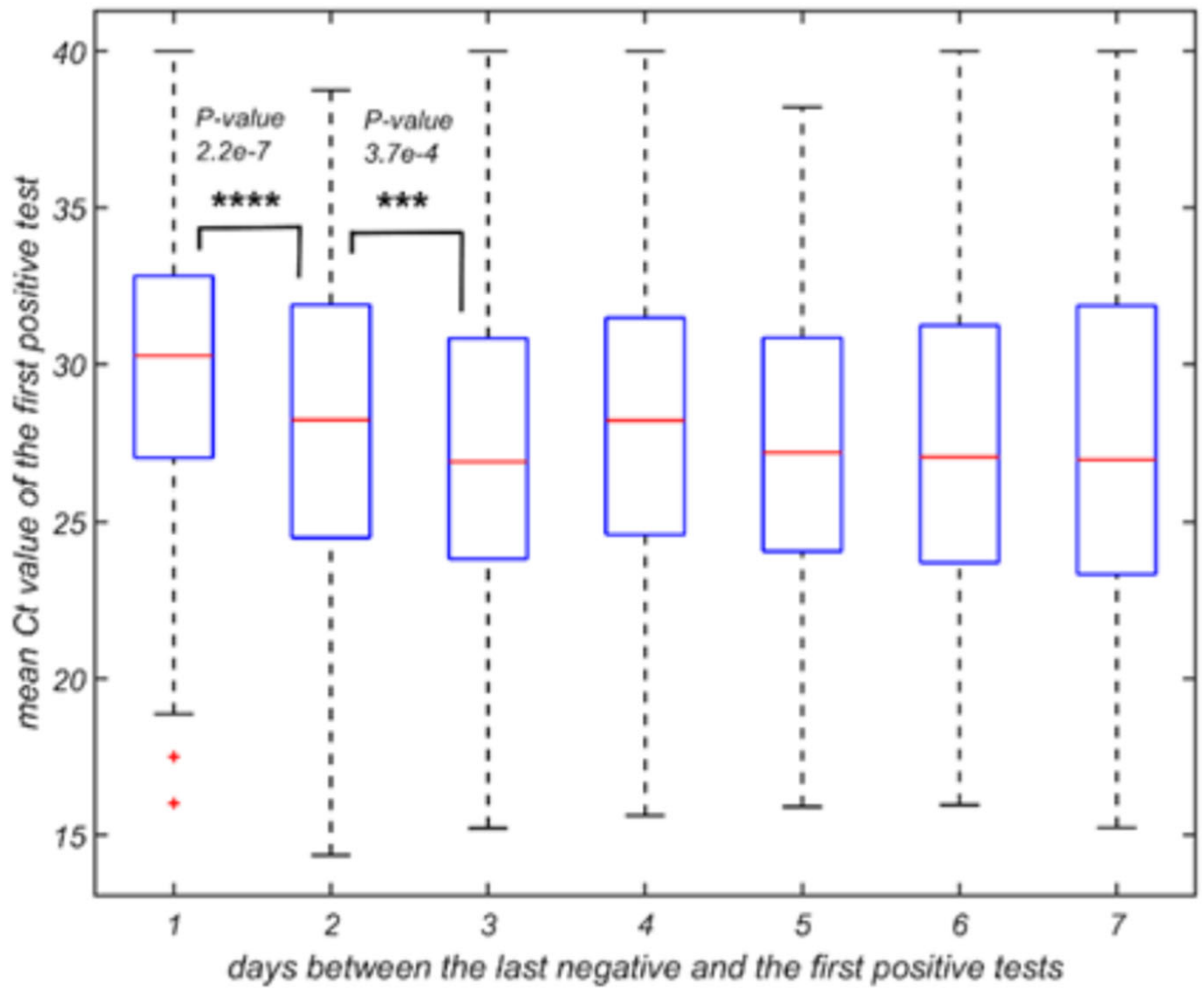
Reminder

What is the Cycle threshold (Ct) value of a PCR test?

Ct = const – log2(viral DNA concentration)



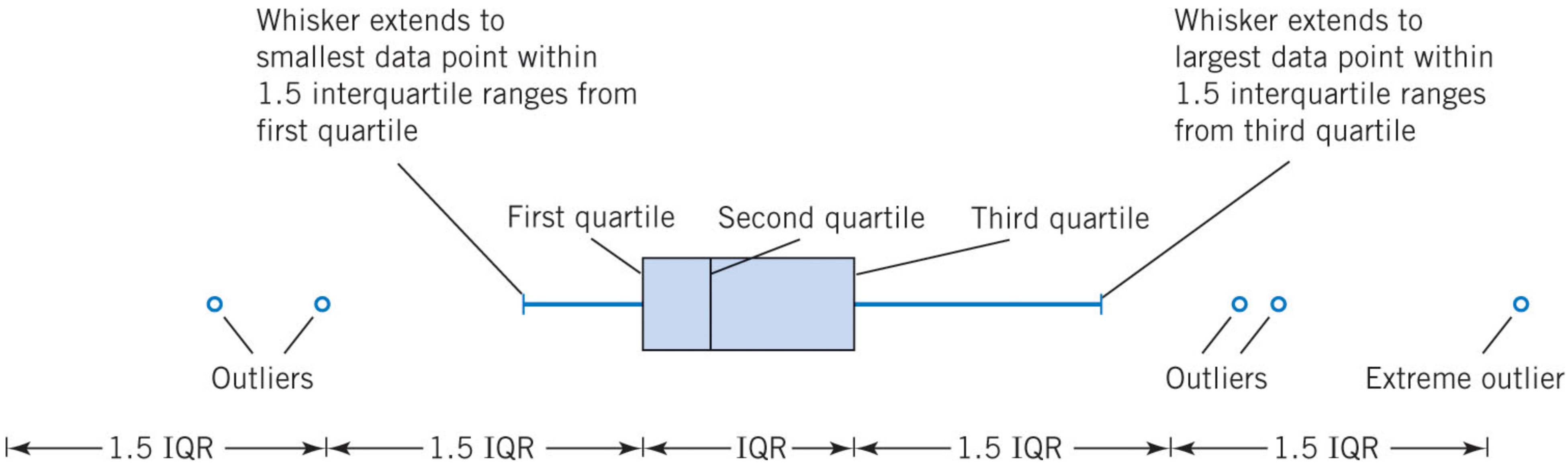
Bar plot based on COVID-19 tests at UIUC



Ranoa, D. R. E. et al. Mitigation of SARS-CoV-2 transmission at a large public university. Nat Commun 13, 3207 (2022)

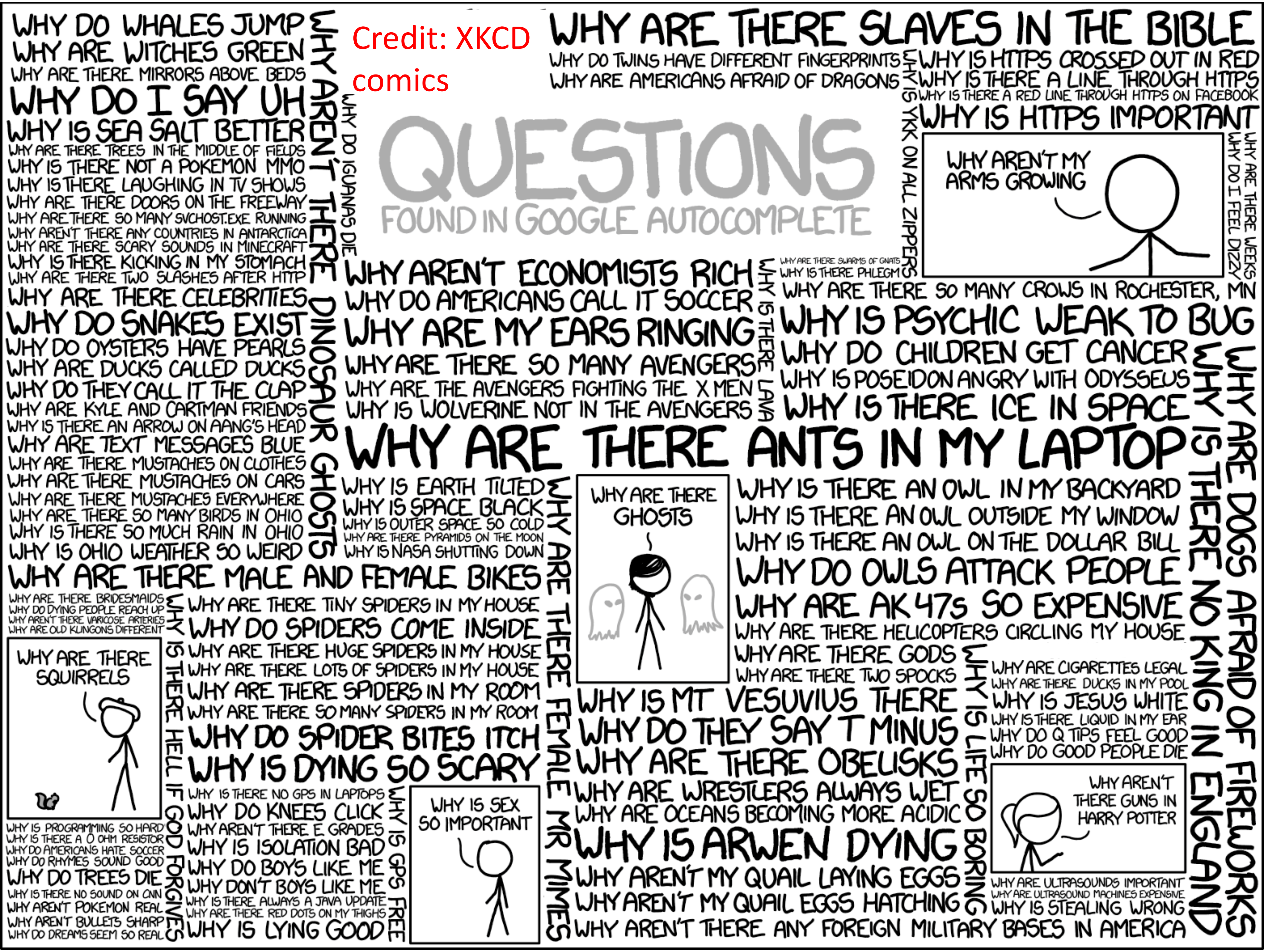
Matlab exercise #2:

- Generate a sample with n= 1000 following standard normal distribution
- Calculate median, first, and third quartiles
- Calculate IQR and find ranges shown below
- Find and count left and right outliers
- Do not use built-in Matlab functions for this!
- Make box and whisker plot: use boxplot



How many right outliers one expects in a sample of n=1000 following normal distribution?

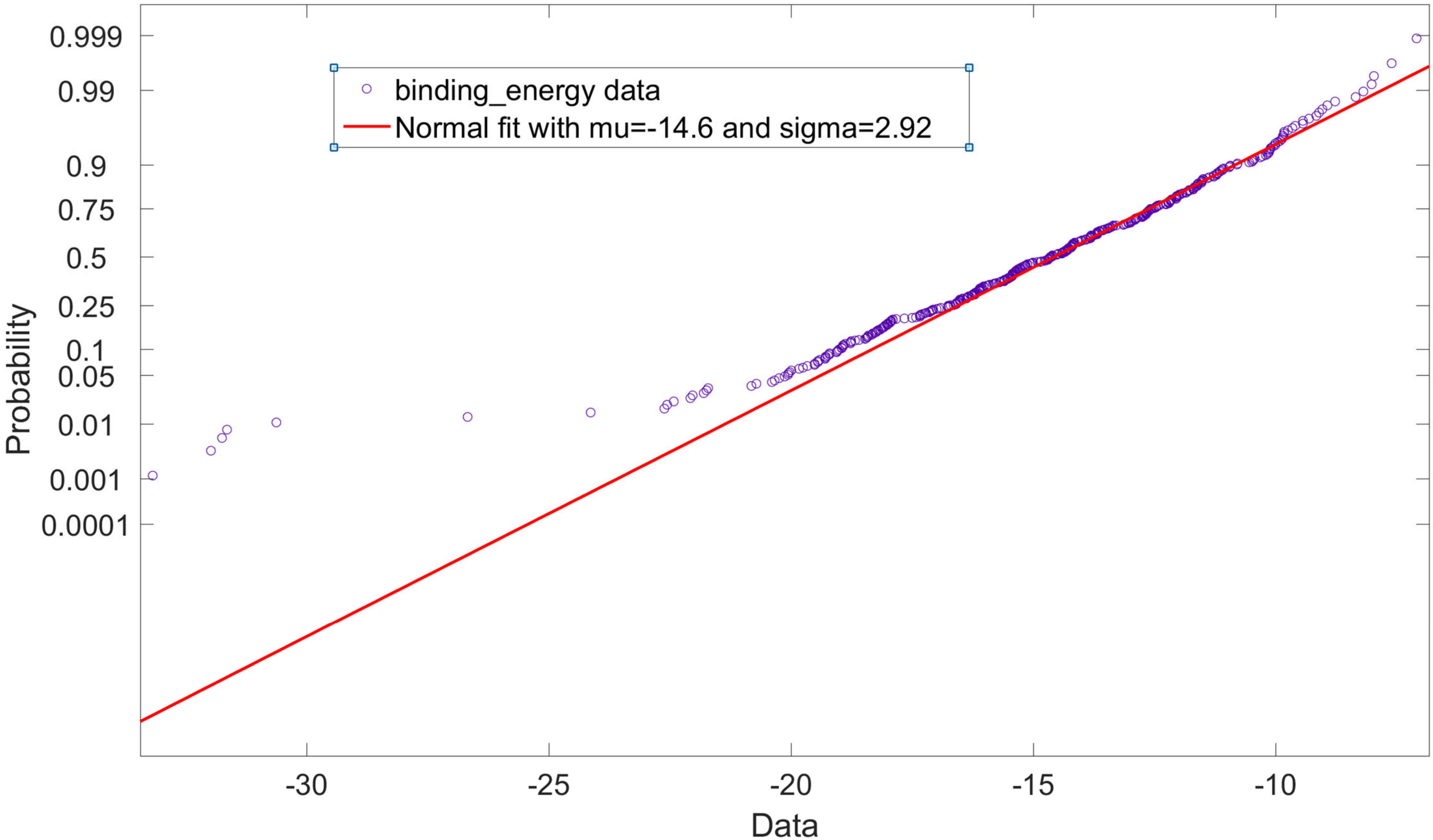
- % find the third quartile of a standard distribution
- norminv(0.75) %ans = 0.6745
- % Calculate IQR - Inter Quartale Range
- IQR=2.*norminv(0.75) % 1.3490
- % Calculate 0.5*IGR+1.5*IQR - the right whisker position
- whisker=0.5.*IQR+1.5*IQR %ans = 2.6980
- % Find the probability to be above the right whisker
- 1-normcdf(whisker) %ans = 0.00349
- % Find number of right outliers in a sample of 1000 points
- 1000.*(1-normcdf(whisker)) %ans = 3.49



Credit: XKCD comics

Probability Plots

- How do we know if a particular probability distribution is a reasonable model for a data set?
- A histogram of a large data set reveals the shape of a distribution. The histogram of a small data set does not provide a clear picture.
- A probability plot is helpful for all data set size. How good is the model based on a particular probability distribution can be verified using a subjective visual examination.


binding_energy data
Normal fit with mu=-14.6 and sigma=2.92
Probability
0.999
0.99
0.9
0.75
0.5
0.25
0.1
0.05
0.01
0.001
0.0001
-30
-25
-20
-15
-10
Data

How To Build a Probability Plot

- Sort the data observations in ascending order: $x_{(1)}, x_{(2)},\ldots, x_{(n)}$.
- Empirically determined cumulative frequency $\text{Prob}(x \le x_{(j)}) = j/n$. To correct for discreteness of $x_{(j)}$ better use $\text{Prob}(x \le x_{(j)}) = (j-0.5)/n$
- If you believe that CDF(x) describes your random variable $(j-0.5)/n$ should be close to $\text{CDF}(x_{(j)})$
- Probability plot is $x_{(j)} \cdot [(j-0.5)/n]/\text{CDF}(x_{(j)})$ plotted versus the observed value $x_{(j)}$.
- If the fit is good - one gets a straight line
- Deviations can be seen especially at tails.

Probability Plot Variations

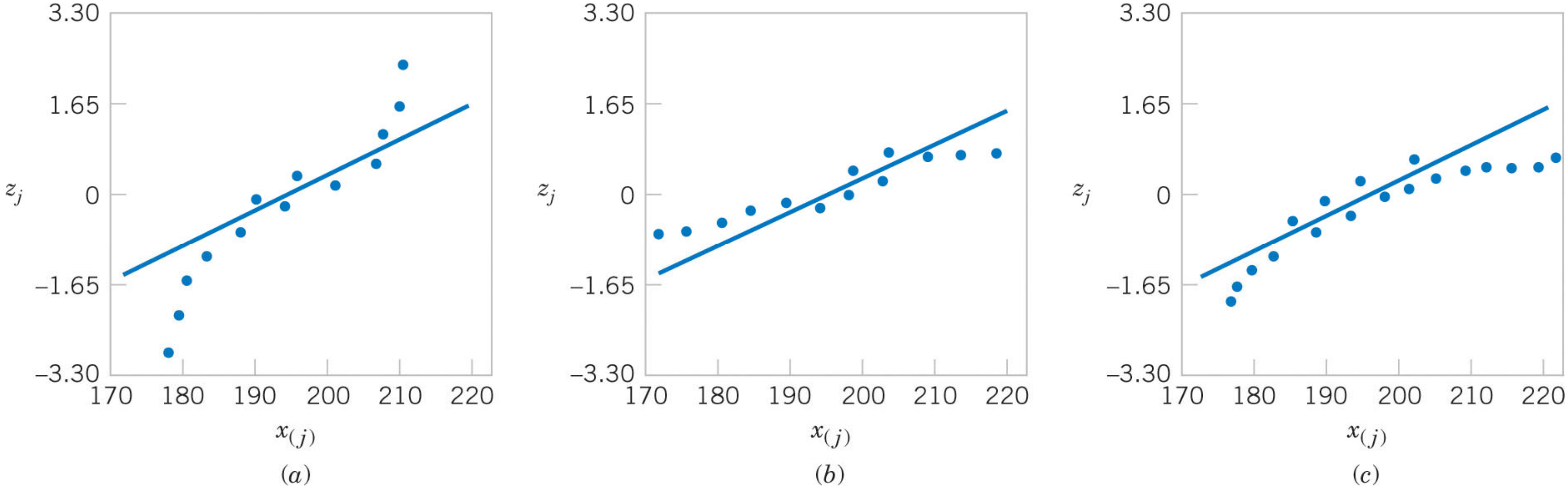
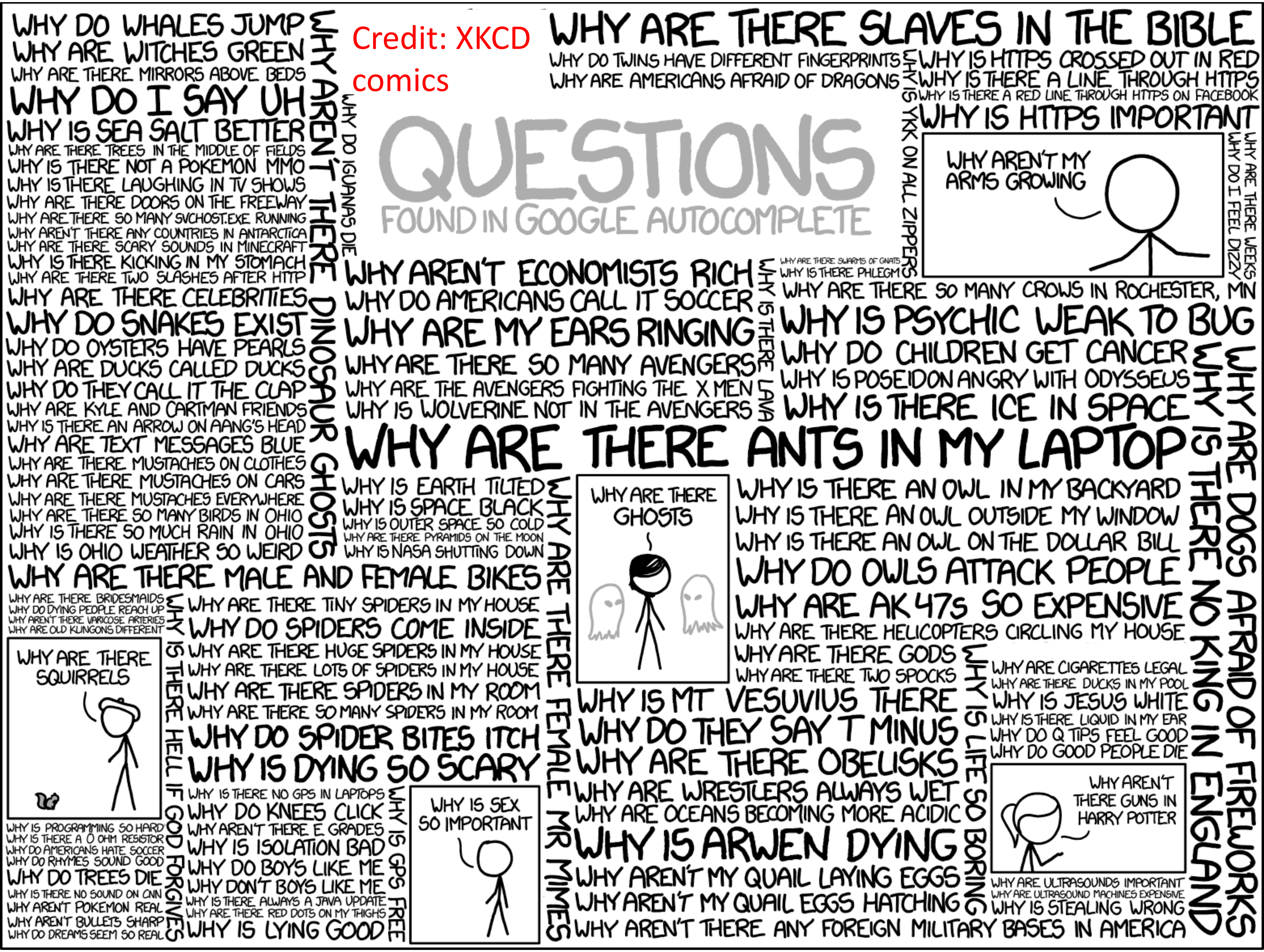


Figure 6-21 Normal probability plots indicating a non-normal distribution.
(a) Light tailed distribution (squeezed together)
(b) Heavy tailed distribution (stretched out)
(c) Right skewed distribution (left end squeezed, right end stretched)



Credit: XKCD comics

Descriptive statistics:

Sample mean and its variance

Standard error vs Standard deviation

Some Definitions

- The random variables $X_1, X_2, \ldots, X_n$ are a random sample of size n if:
  a) The X_i are independent random variables.
  b) Every X_i has the same probability distribution.

  Such $X_1, X_2, \ldots, X_n$ are also called independent and identically distributed (or i. i. d.) random variables
- A **statistic** is any function of the observations in a random sample.
- The probability distribution of a statistic is called a **sampling distribution**.

Statistic #1: Sample Mean

If the values of n observations in a random sample are denoted by $x_1, x_2, \ldots, x_n,$ the sample mean is

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n} \quad (6-1)$$

New random variable $\bar{X}$ is a linear combination of *n* independent identically distributed variables $X_1, X_2, \ldots, X_n$

$$\bar{X} = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

Mean & Variance of a Linear Function

$Y = c_1X_1 + c_2X_2 + \ldots + c_pX_p$

$$E(Y) = c_1E(X_1) + c_2E(X_2) + \ldots + c_pE(X_p) \quad (5\text{-}25)$$

$$\mathrm{V}(Y) = c_1^2V(X_1) + c_2^2V(X_2) + \ldots + c_p^2V(X_p) + 2\sum_{i<j}\sum c_ic_j \operatorname{cov}(X_iX_j) \quad (5\text{-}26)$$

If $X_1, X_2, \ldots, X_p$ are independent, then $\operatorname{cov}(X_iX_j) = 0,$

$$V(Y) = c_1^2V(X_1) + c_2^2V(X_2) + \ldots + c_p^2V(X_p) \quad (5\text{-}27)$$

IMPORTANT:

Sample mean $\bar{x}$ is drawn from a random variable

$$\bar{X} = \frac{X_1 + X_2 + \dots + X_n}{n}$$

$$E(\bar{X}) = \frac{n \cdot E(X_i)}{n} = \frac{n \cdot \mu}{n} = \mu$$

$$V(\bar{X}) = \frac{n \cdot V(X_i)}{n^2} = \frac{n \cdot \sigma^2}{n^2} = \frac{\sigma^2}{n}$$

$$\text{Stand. dev.}(\bar{X}) = \frac{\sigma}{\sqrt{n}}$$

Central Limit Theorem

If $X_1, X_2, \ldots, X_n$ is a random sample of size n is taken from a population with mean μ and finite variance σ^2, and **any distribution**. If $\bar{X}$ is the sample mean, then the limiting form of the distribution of

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \qquad (7-1)$$

for large n, is the standard normal distribution.

If $X_1, X_2, \ldots, X_n$ are themselves normally distributed – for any n

Test CLT for your own random variable

- Go to: https://onlinestatbook.com/stat_sim/sampling_dist/
- Select "Custom" at the top and use mouse to sketch the PMF of your own random variable
- Select "mean" and n=5 in the third panel
- Choose "Animated" in the second panel and use number_of_experiments=5 to see one sample being generated
- Repeat with number_of _experiments =10,000
- Now select "mean" and n=25 in the fourth panel
- Skewness and Curtosis are measures of how good is the normal (Gaussian) fit (choose "fit normal")