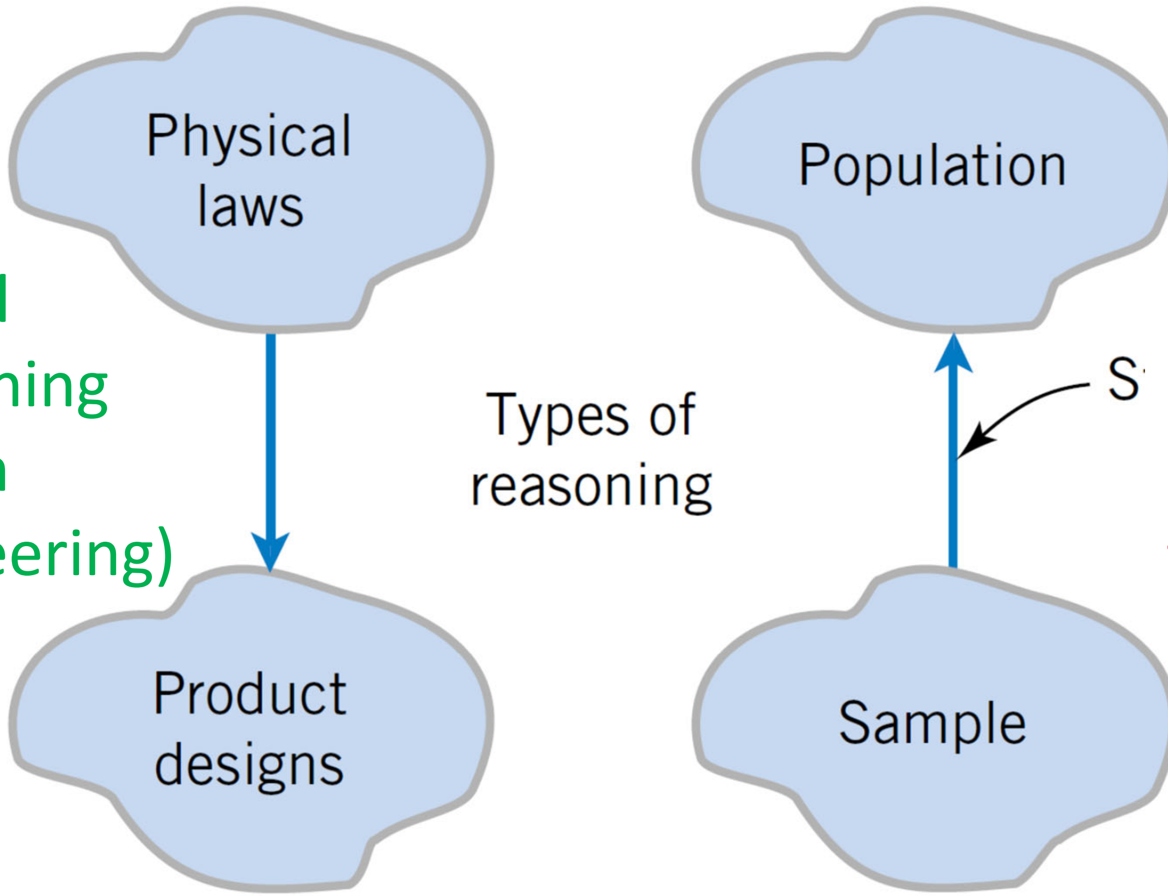
This Thursday 4/3/2025
we will have a remedial Matlab lab.
Bring your laptops

Same time (9:30am-10:50am)
and place (here)
as our usual lecture

Descriptive statistics:
Populations, Samples
Histograms, Quartiles
Sample mean and
variance

Two types of reasoning

Logical Reasoning (e.g. in Engineering)



Statistical reasoning: Inference of population properties from a finite sample

Numerical Summaries of Data

- Data are the numerical observations of a phenomenon of interest.
- The totality of all observations is a population.
  - Population can be infinite (e.g. abstract random variables)
  - It can be very large (e.g. 7 billion humans or all patients who have cancer of a given type)
- A (usually small) portion of the population collected for analysis is a random sample.
- We want to use sample to infer facts about populations
- The inference is not perfect but gets better and better as sample size increases.

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.

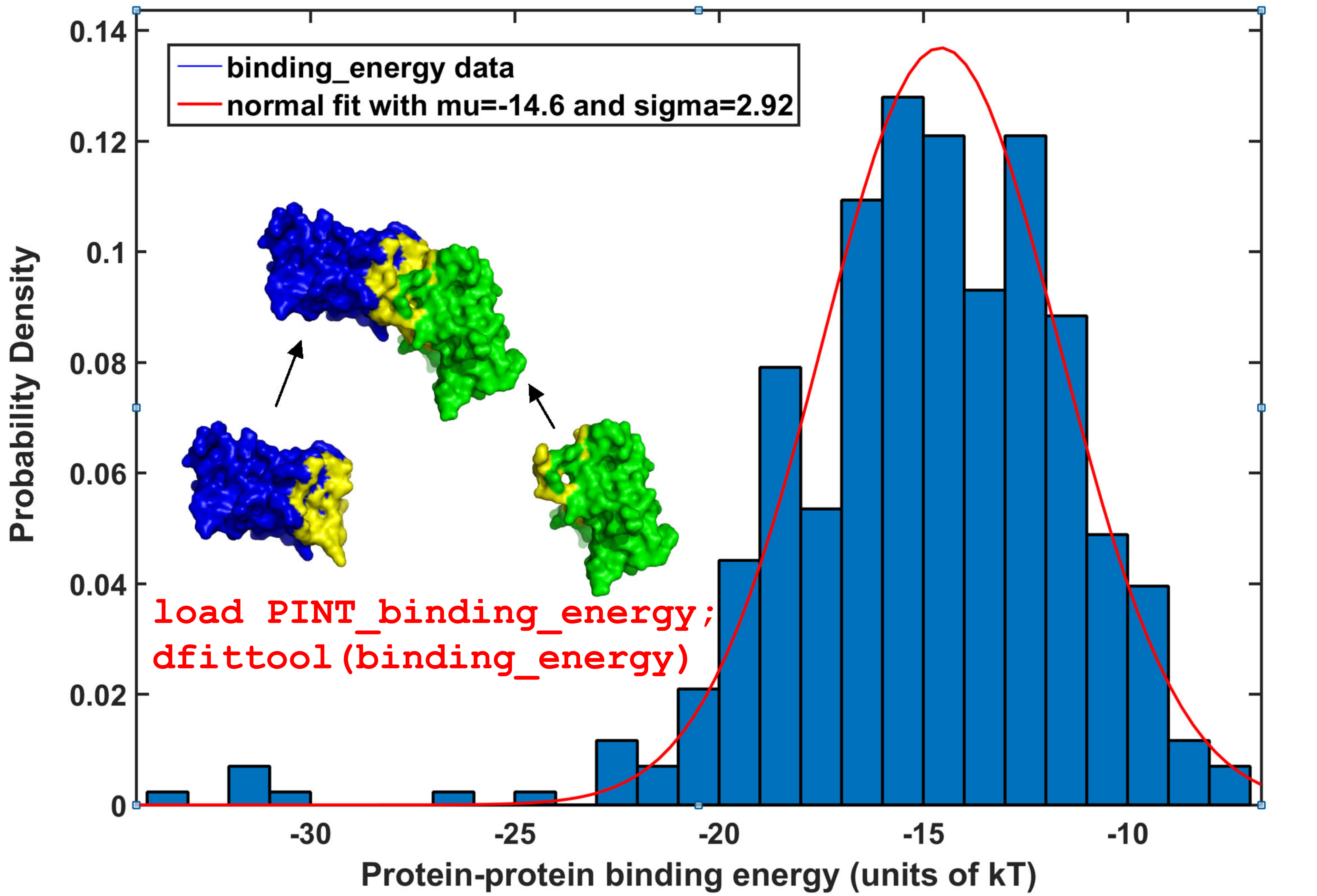
- Such $X_1, X_2, \ldots, X_n$ are also called independent and identically distributed (or i. i. d.) random variables

Ways to describe a sample:

Histogram

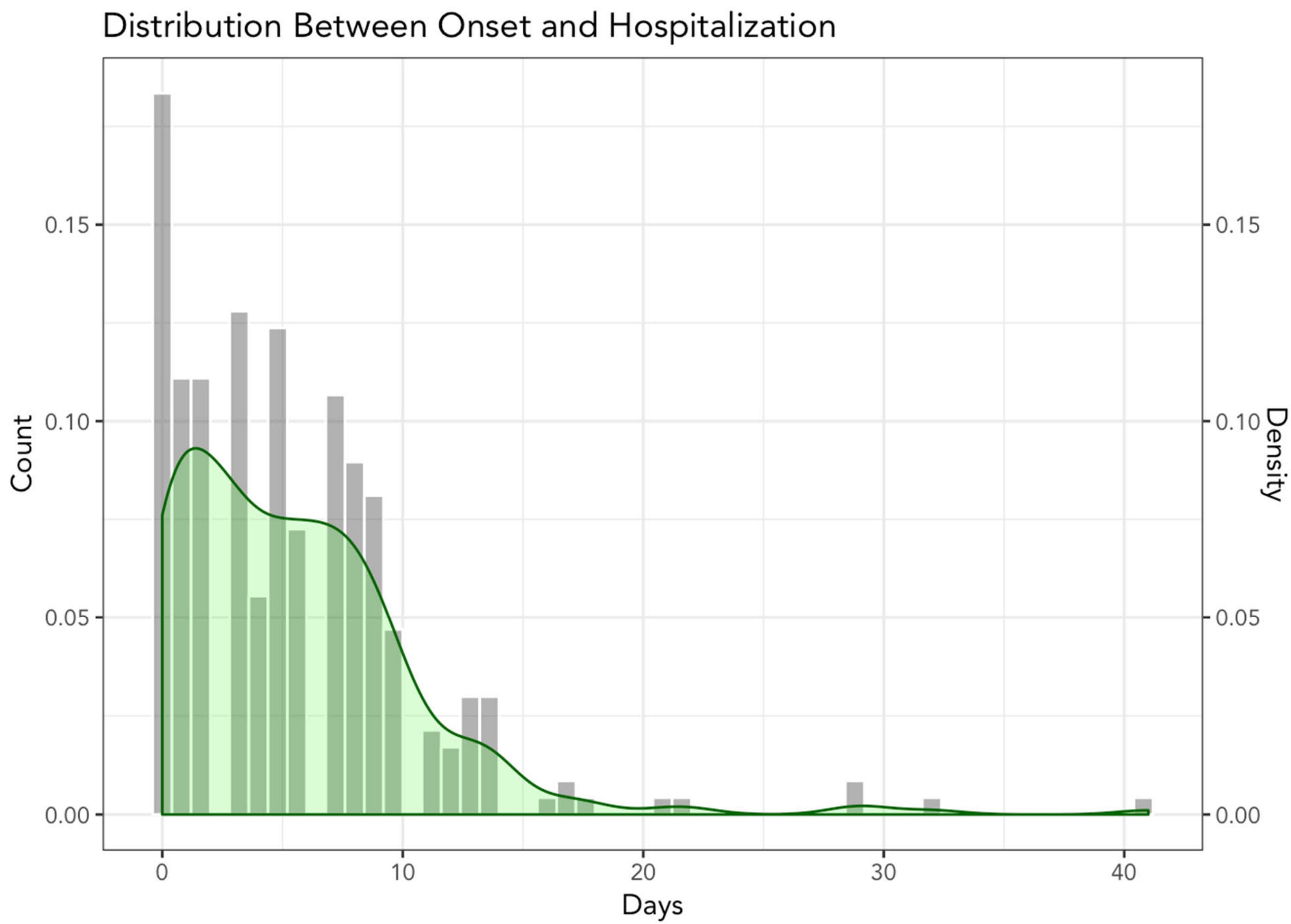
approximates PDF

(or PMF)



binding_energy data
normal fit with mu=-14.6 and sigma=2.92
load PINT_binding_energy;
dfittool(binding_energy)
Probability Density
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
-30
-25
-20
-15
-10
Protein-protein binding energy (units of kT)

PDF of time between COVID-19 symptoms onset and hospitalization in IL, April 2020

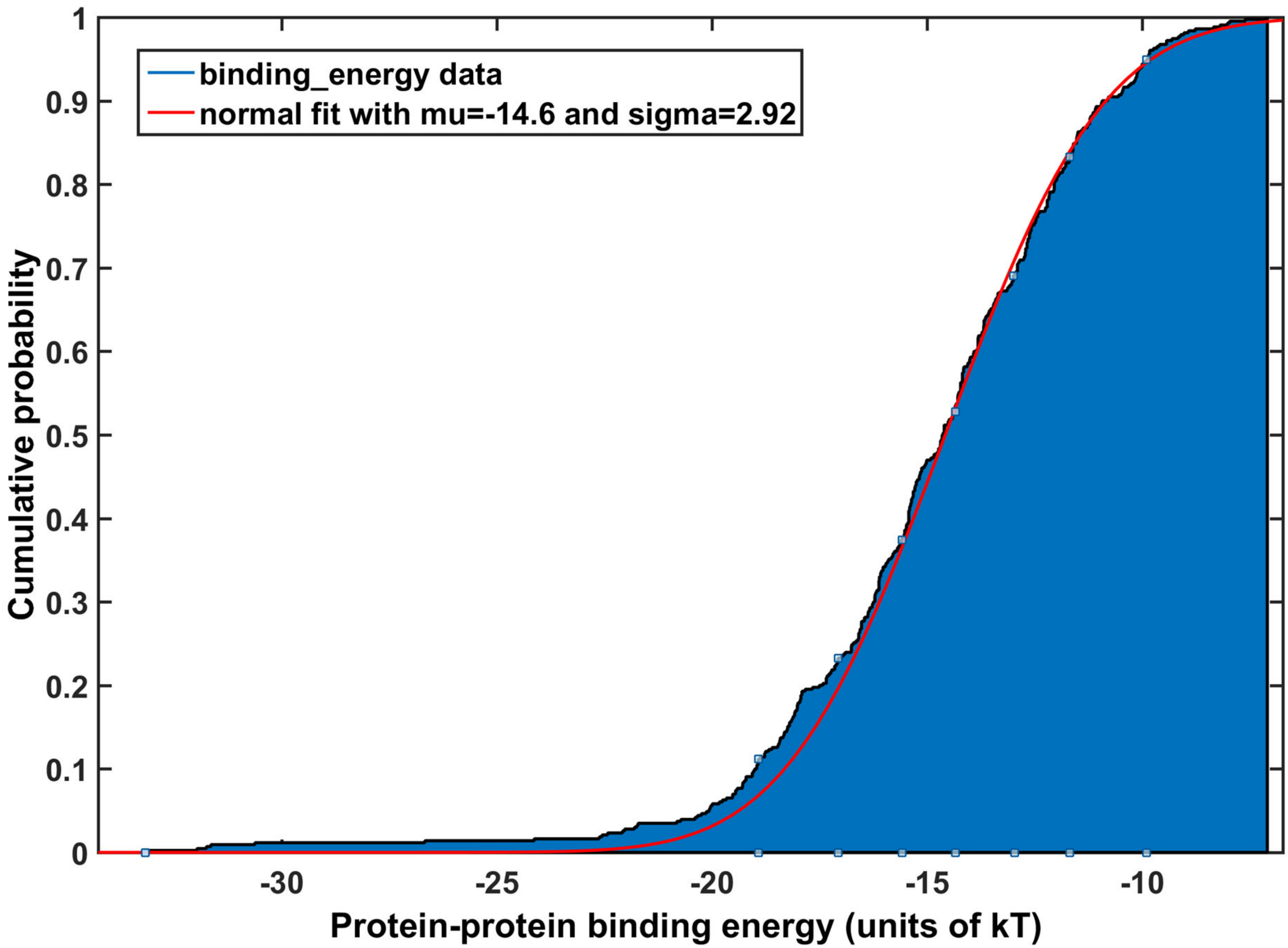


Histograms with Unequal Bin Widths

- If the data is tightly clustered in some regions and scattered in others, it is visually helpful to use narrow bin widths in the clustered region and wide bin widths in the scattered areas.
- To approximate the PDF, the rectangle area, not the height, must be proportional to the bin relative frequency.

$$\text{Rectangle height} = \frac{\text{bin relative frequency}}{\text{bin width}}$$

Cumulative Frequency Plot



Median, Quartiles, Percentiles

- The median q_2 divides the sample into two equal parts: 50% (n/2) of sample points below q_2 and 50% (n/2) points above q_2
- The three quartiles partition the data into four equally sized counts or segments.
  - 25% of the data is less than q_1.
  - 50% of the data is less than q_2, the median.
  - 75% of the data is less than q_3.
- There are 100 percentiles. n-th percentile p_n is defined so that n% of the data is less than p_n

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*.

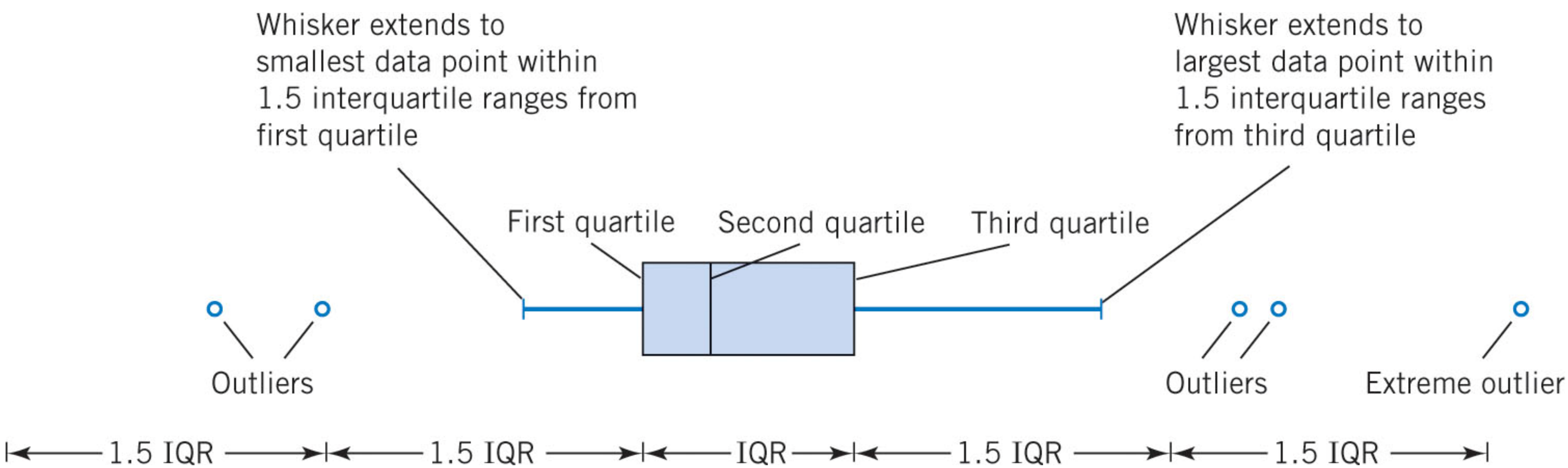
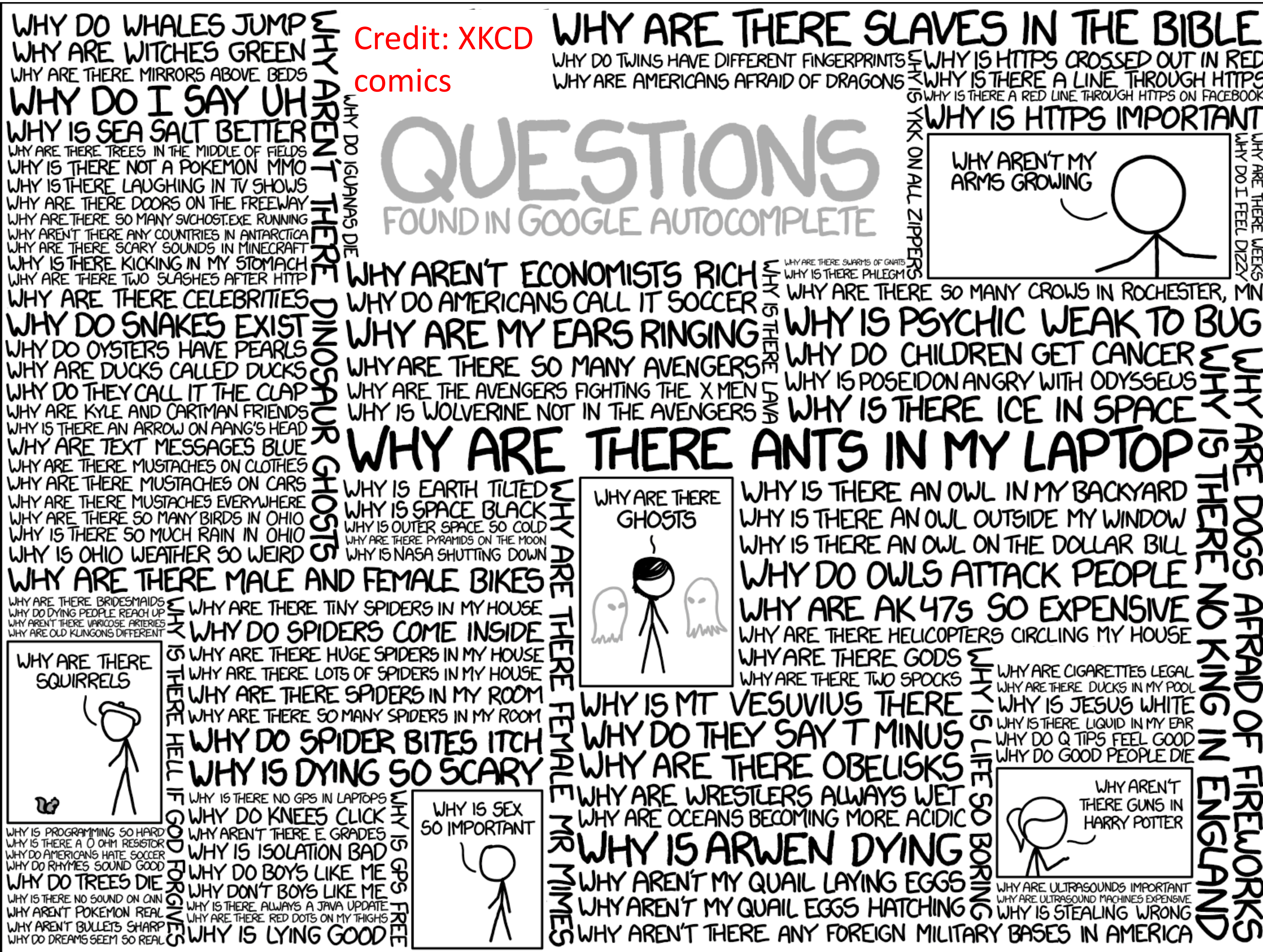


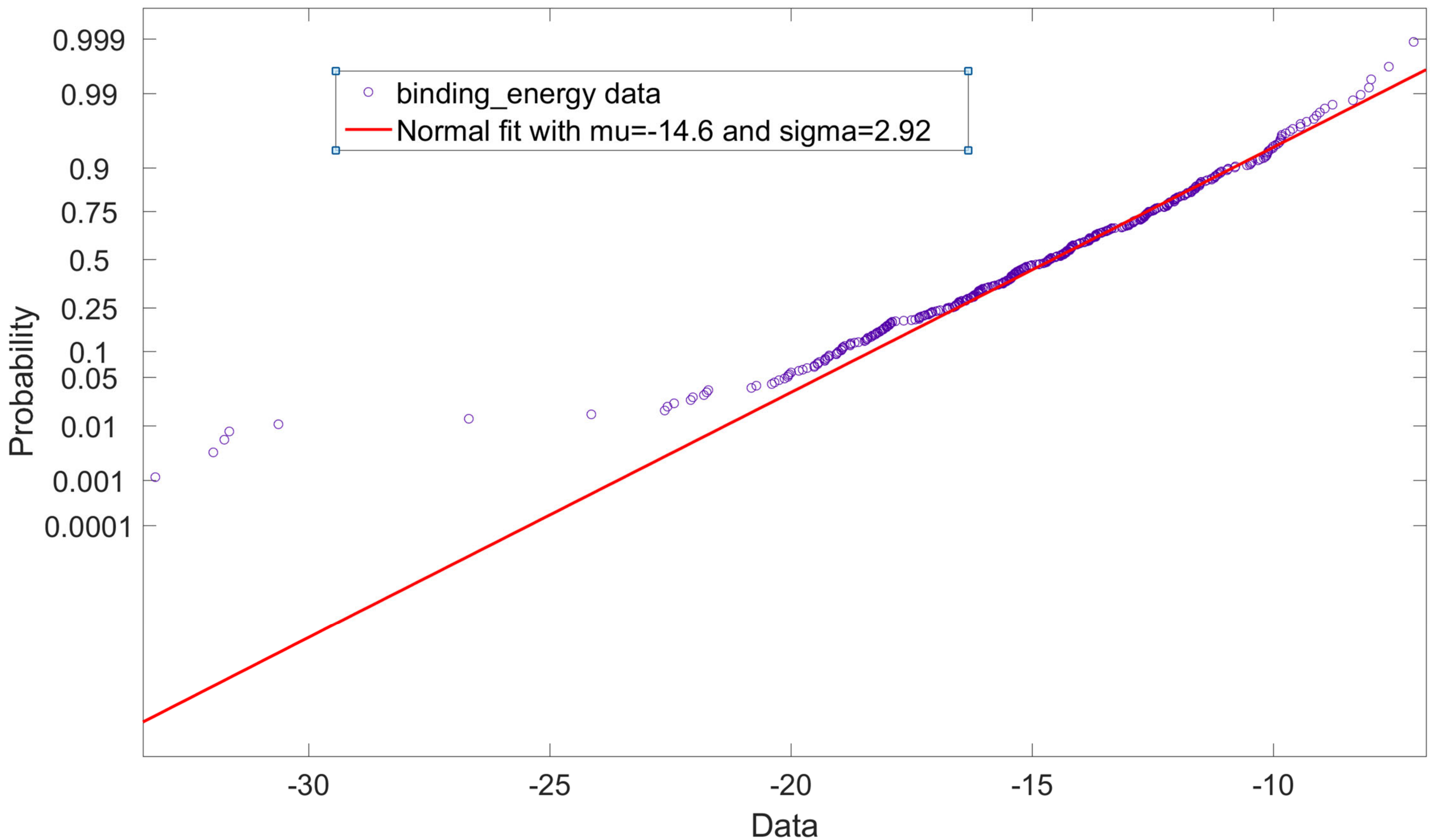
Figure 6-13 Description of a box plot.



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: $x_{(1)}, x_{(2)},\dots, x_{(n)}$.
- Empirically determined cumulative frequency Prob($x \le x_{(j)}$) =j/n. To correct for discreteness of $x_{(j)}$ better use 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 CDF($x_{(j)}$)
- Probability plot is $x_{(j)} \cdot$ [(j-0.5)/n]/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

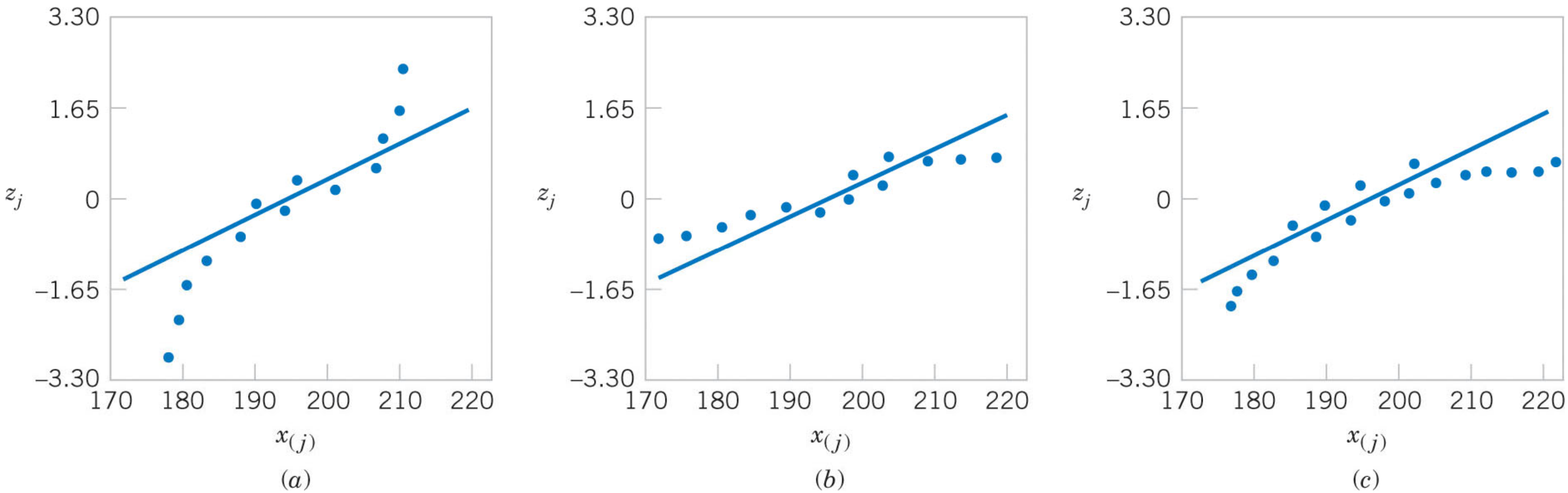
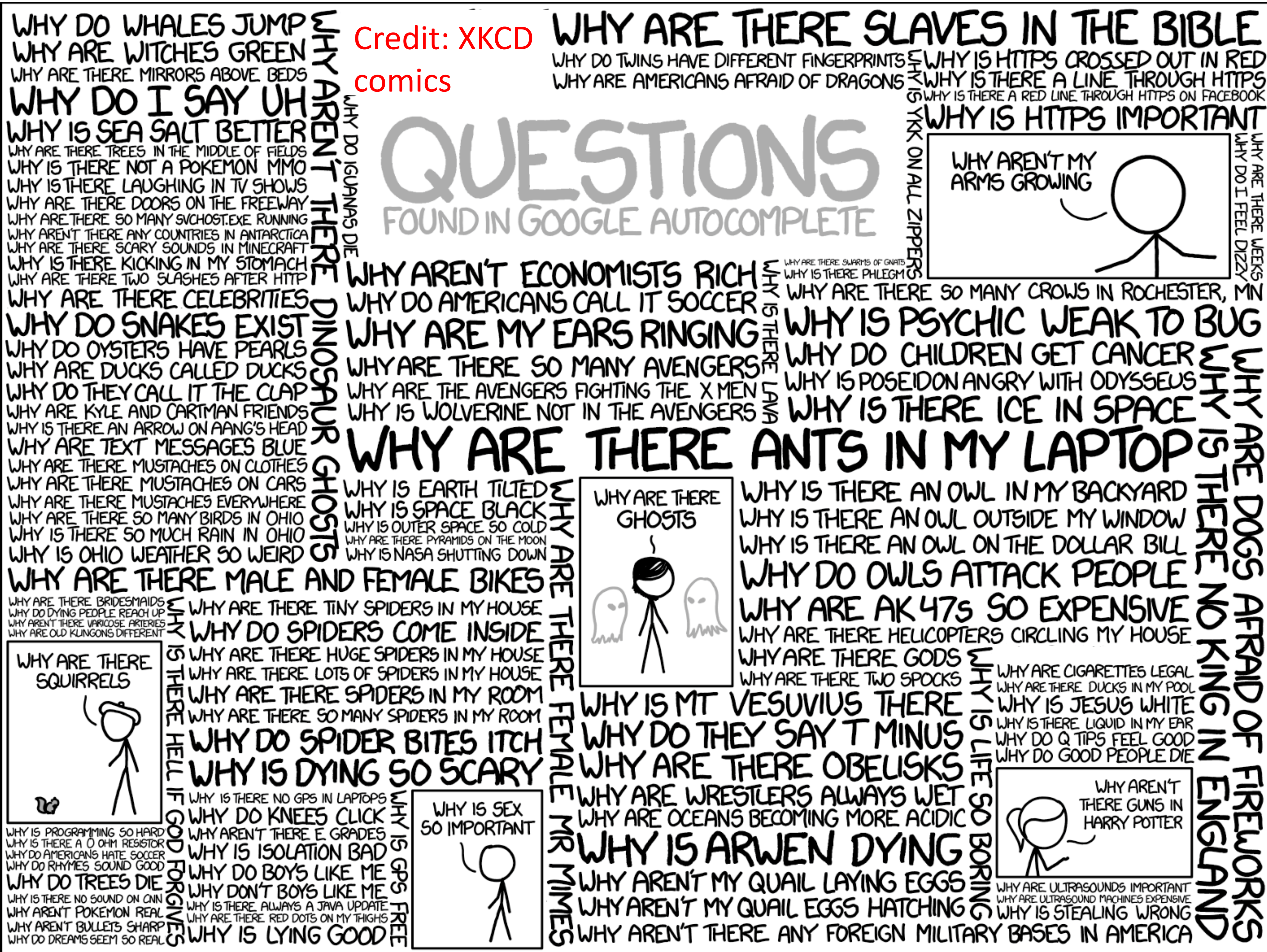


Figure 6-21 Normal probability plots indicating a non-normal distribution.
(a) Light tailed distribution (squeezed together)
(b) Heavy tailed distribution (stretched out)
(c) Right skewed distribution (left end squeezed, right end stretched)



Credit: XKCD comics

Linear Functions of Random Variables

- A function of multiple random variables is itself a random variable.
- A function of random variables can be formed by either linear or nonlinear relationships. We will only work with linear functions.
- Given random variables $X_1, X_2, \ldots, X_p$ and constants $c_1, c_2, \ldots, c_p$

  $Y = c_1X_1 + c_2X_2 + \ldots + c_pX_p$ (5-24)

  is a linear combination of $X_1, X_2, \ldots, X_p$.

Mean & Variance of a Linear Function

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

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

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

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

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

Descriptive statistics:

Sample mean and variance

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

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

IMPORTANT:

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

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

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

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

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

Mean & Variance of the Sample Average

If $\bar{X} = \dfrac{(X_1 + X_2 + \ldots + X_n)}{n}$ and $E(X_i) = \mu$

Then $E(\bar{X}) = \dfrac{n \cdot \mu}{n} = \mu$ (5−28a)

If the X_i are independent with $V(X_i) = \sigma^2$

Then $V(\bar{X}) = \dfrac{n \cdot \sigma^2}{n^2} = \dfrac{\sigma^2}{n}$ (5−28b)

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

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

for large n, is the standard normal distribution.

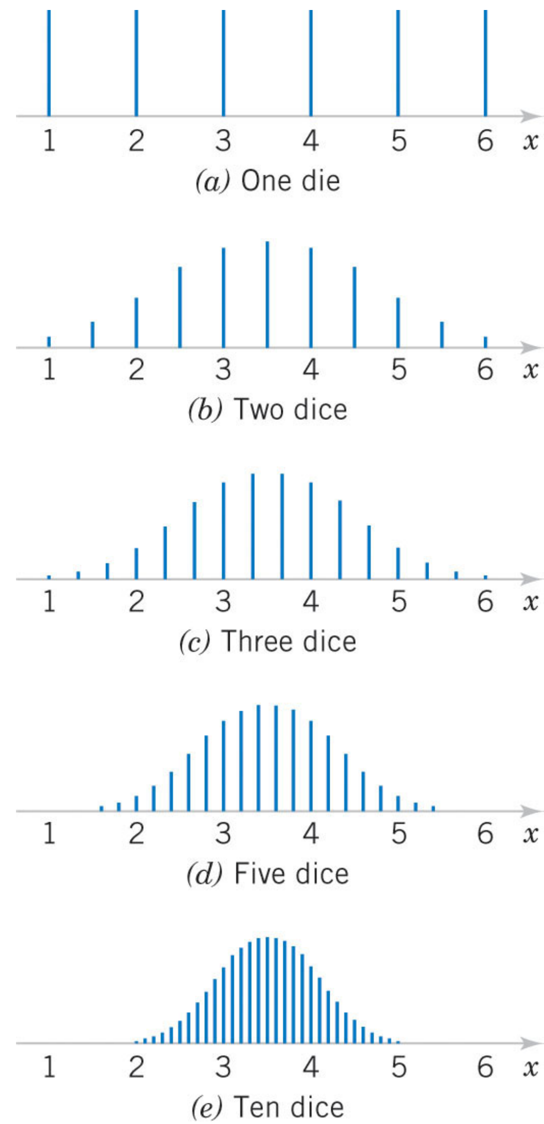
If $X_1, X_2, \ldots, X_n$ are themselves normally distributed – for any n

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")

Sampling Distributions of Sample Means

Figure 7-1 Distributions of average scores from throwing dice.
Mean = (6+1)/2=3.5
Sigma^2 = [(6-1+1)^2-1]/12=2.92
Sigma=1.71



Formulas

$$\mu = \frac{b+a}{2}$$

$$\sigma_X^2 = \frac{(b-a+1)^2 - 1}{12}$$

$$\sigma_{\bar{X}}^2 = \frac{\sigma_X^2}{n}$$

show
Matlab