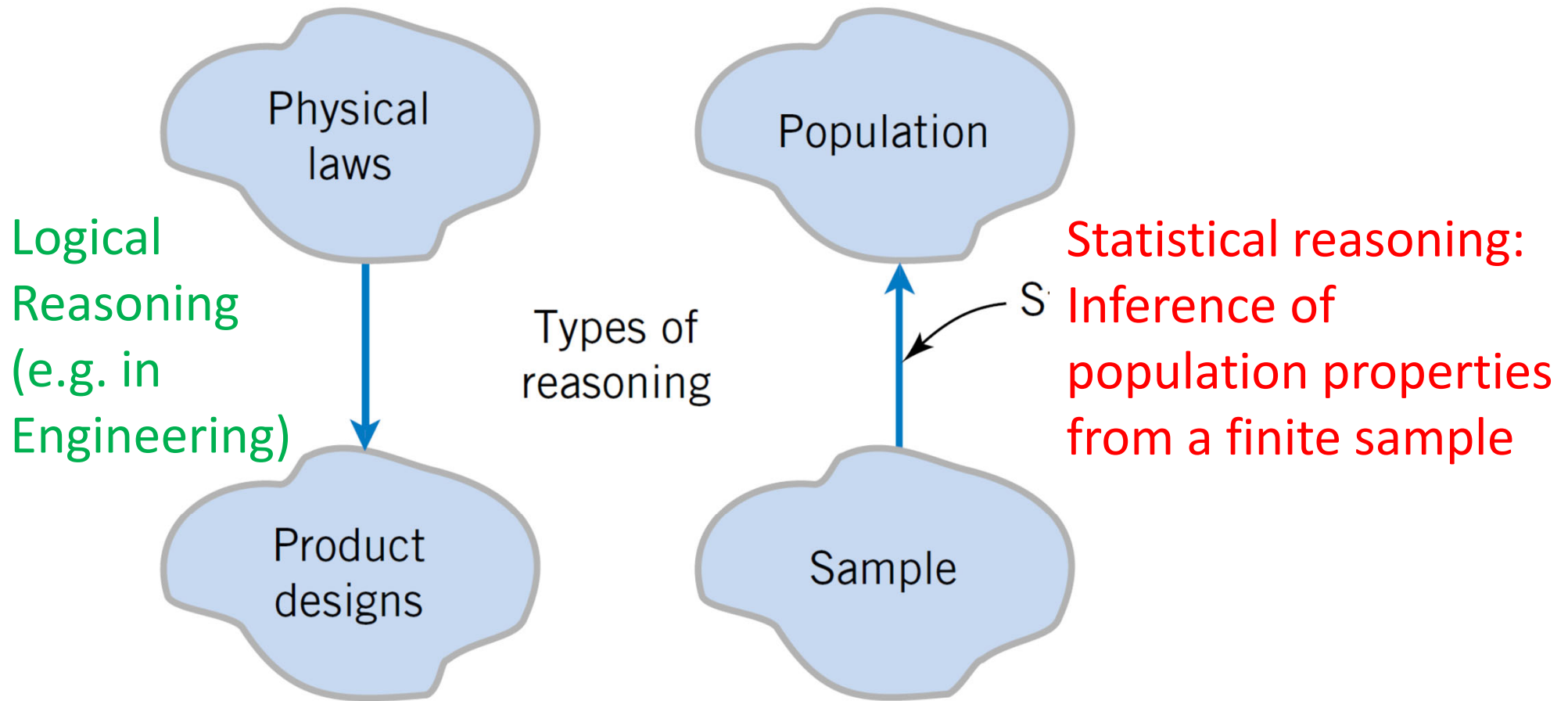


Descriptive statistics:
Populations, Samples
Histograms, Quartiles
Sample mean and
variance

Two types of reasoning



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**.

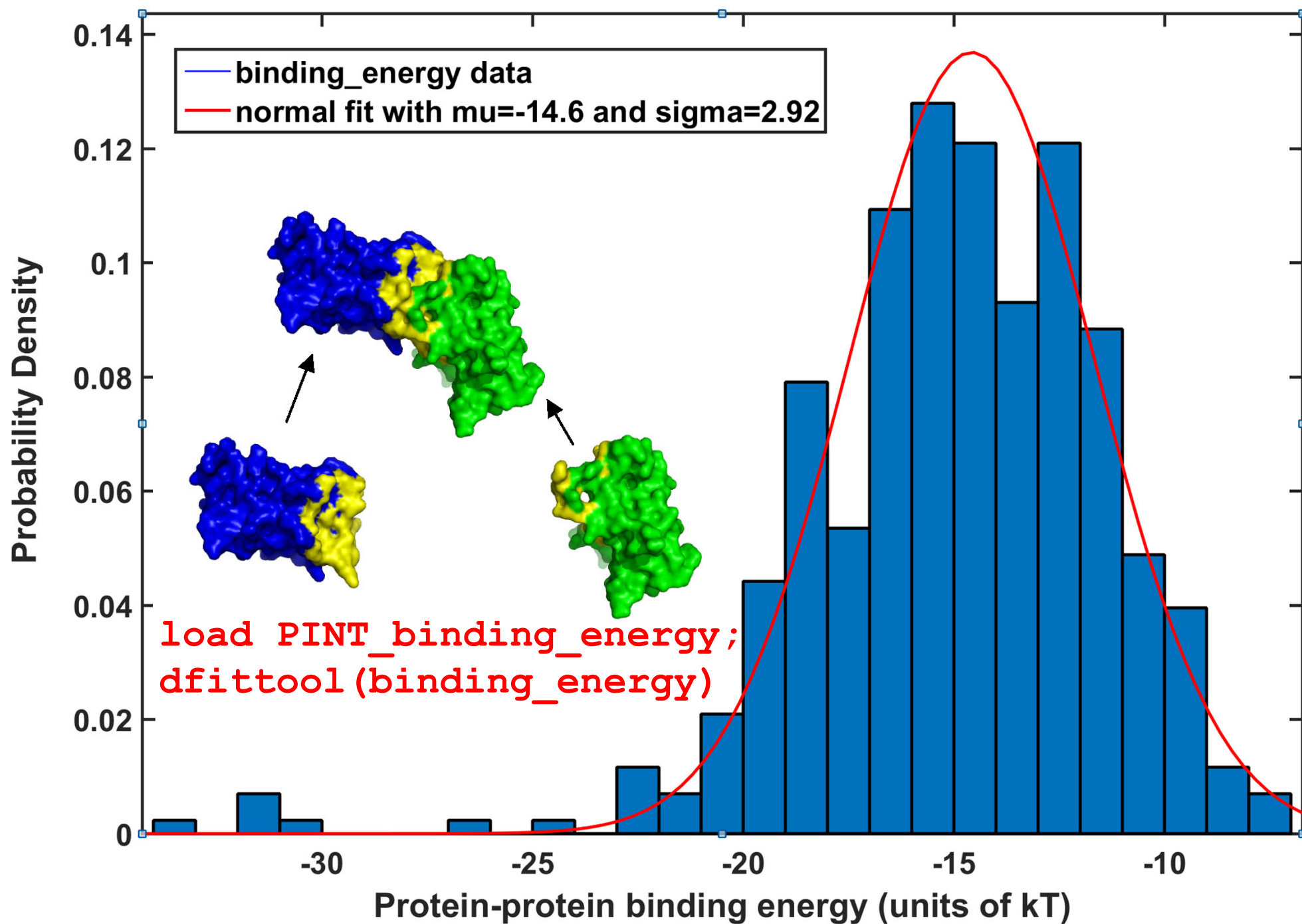
Some Definitions

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

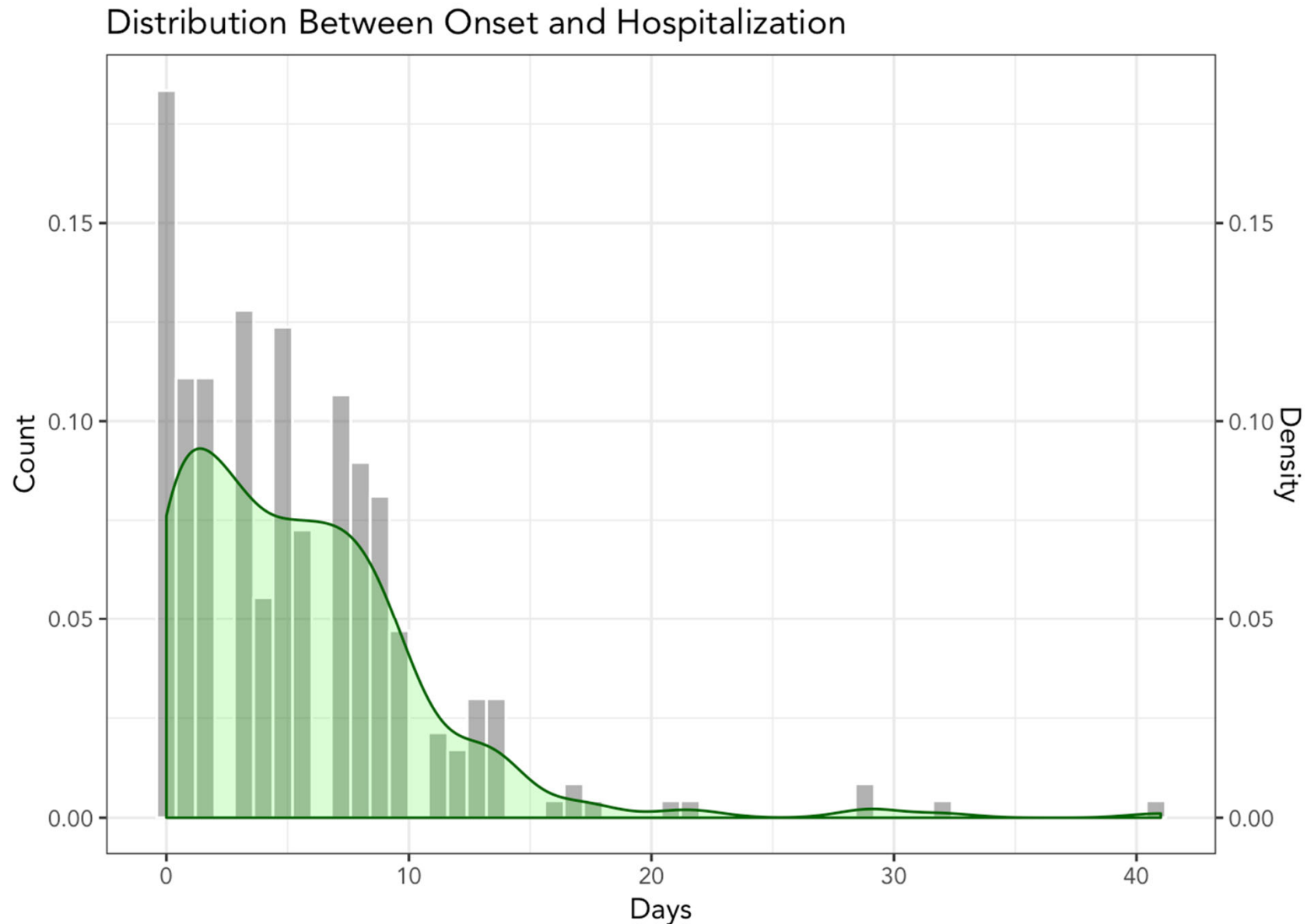
Ways to describe a sample:

Histogram

approximates PDF
(or PMF)



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

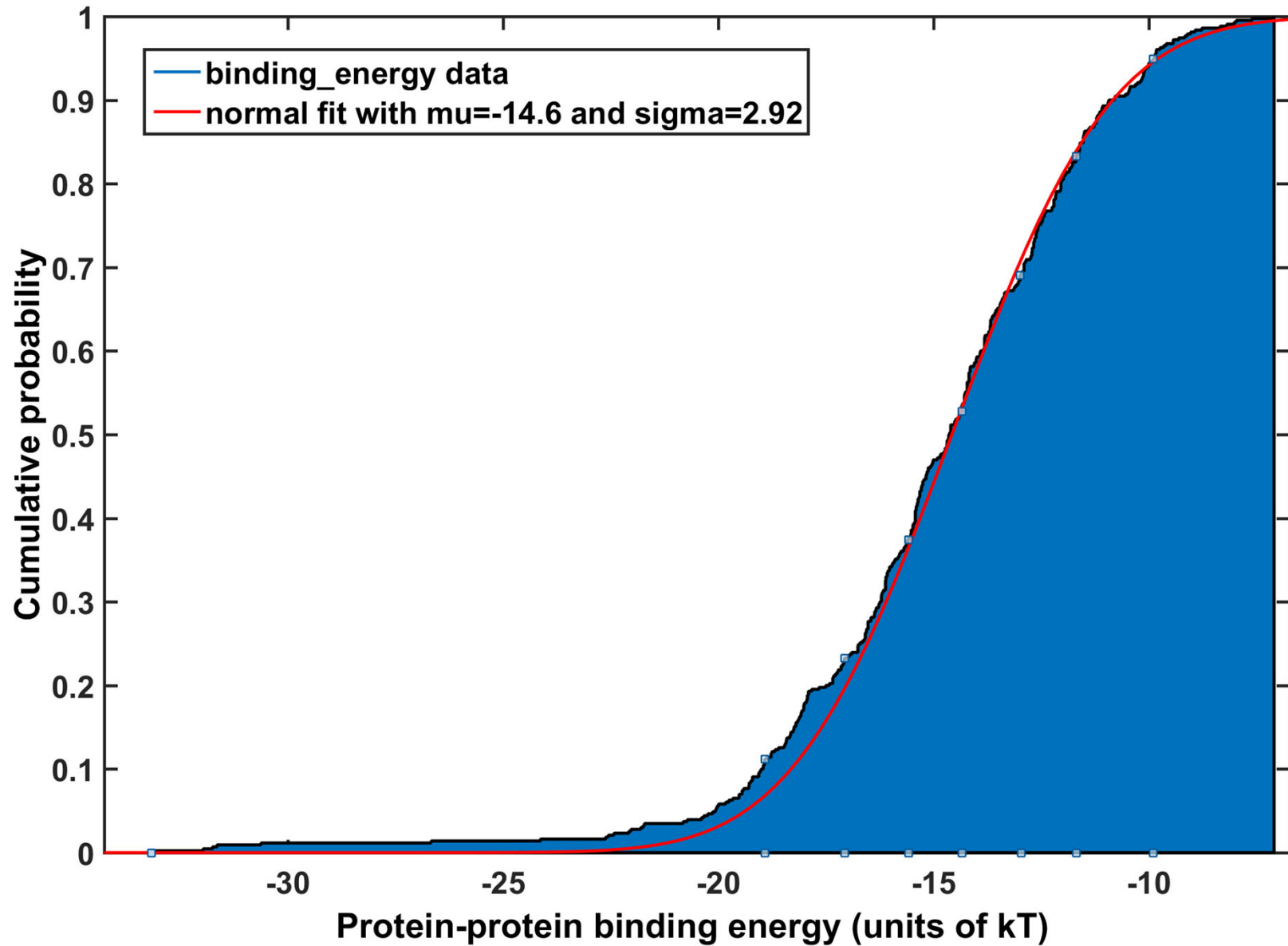


Histograms with Unequal Bin Widths

- 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

Matlab exercise #1

- Find the median and lower & upper quartiles of a $n=1000$ sample drawn from a standard normal distribution
- Do not use built-in Matlab functions for this exercise!
- Hint: use `[a,b]=sort(r1);` to rank order your sample. The variable `a` returns `r1` sorted in the increasing order.
- How to find the median and both quartiles from `a`?

How to find the median & quartiles

- % Example: find median and lower quartile of
- % a sample with $n=100$ drawn from uniform
- `r1=randn(1000,1);`
- `[a,b]=sort(r1);`
- `med=(a(500)+a(501))./2`
- `sum(r1<med) % verify`
- `q1=(a(250)+a(251))./2`
- `sum(r1<q1) % verify`
- `q3=(a(750)+a(751))./2`
- `sum(r1<q3) % verify`

Box-and-Whisker Plot

- A box plot is a graphical display showing **S**pread, **O**utliers, **C**enter, and **S**hape (**SOCS**).
- It displays the **5-number summary**: *min*, q_1 , *median*, q_3 , and *max*.

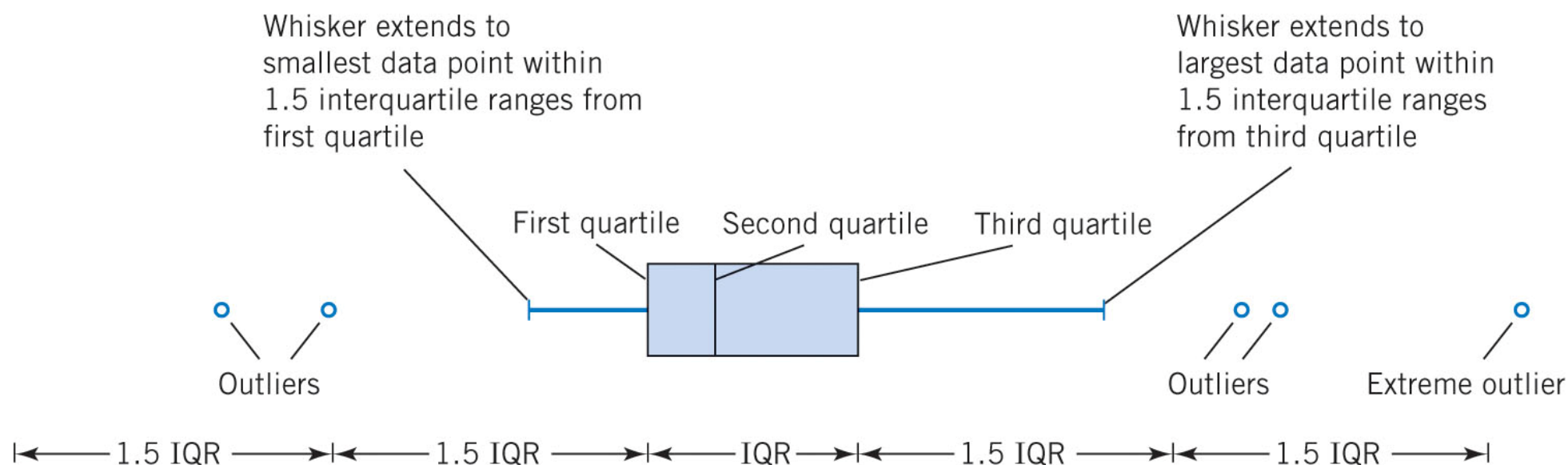
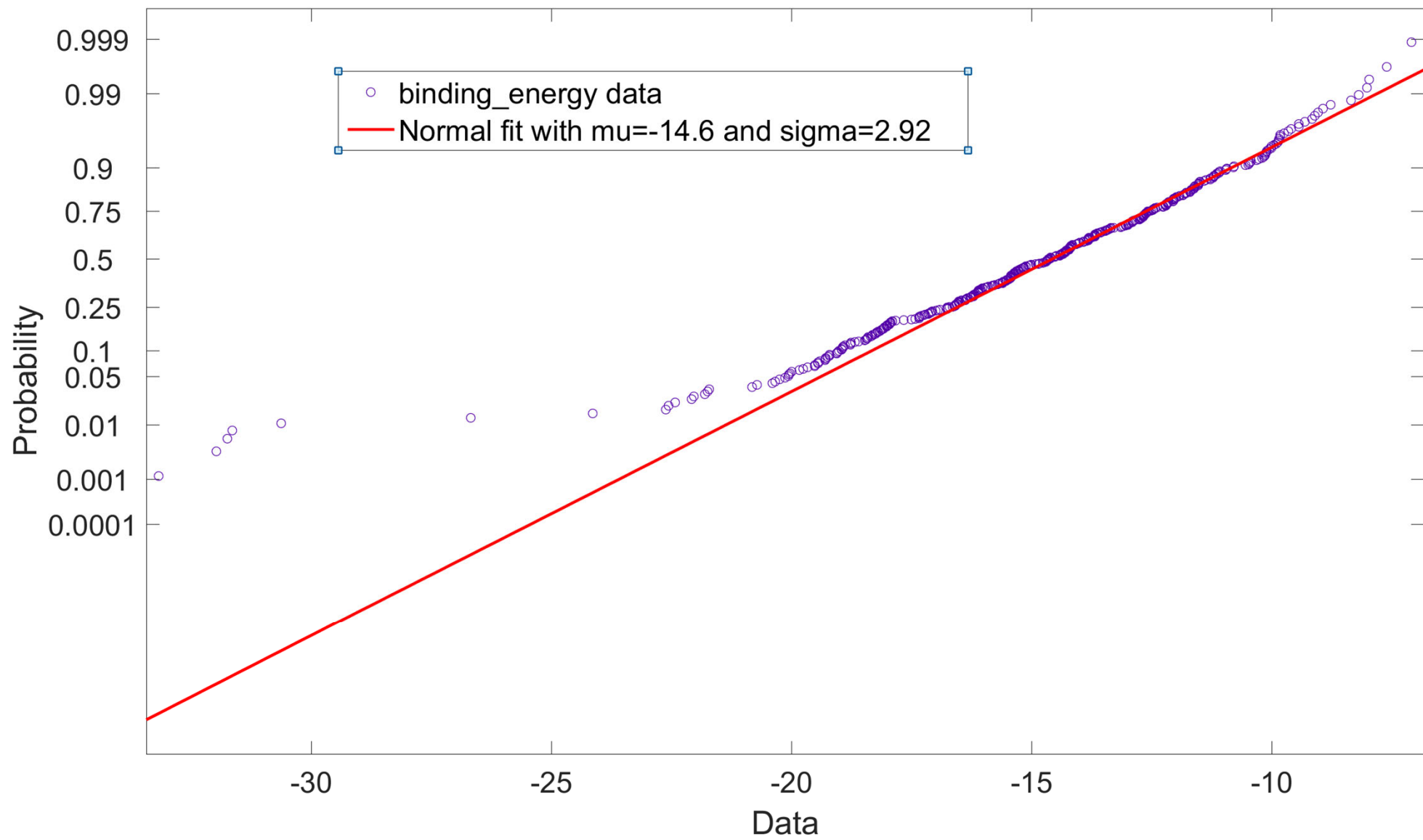


Figure 6-13 Description of a box plot.

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.



How To Build a Probability Plot

- Sort the data observations in ascending order:
 $X_{(1)}, X_{(2)}, \dots, X_{(n)}$.
- Empirically determined cumulative frequency
 $\text{Prob}(x \leq x_{(j)}) = j/n$. To correct for discreteness
of $x_{(j)}$ better use $\text{Prob}(x \leq x_{(j)}) = (j-0.5)/n$
- If you believe that $\text{CDF}(x)$ 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

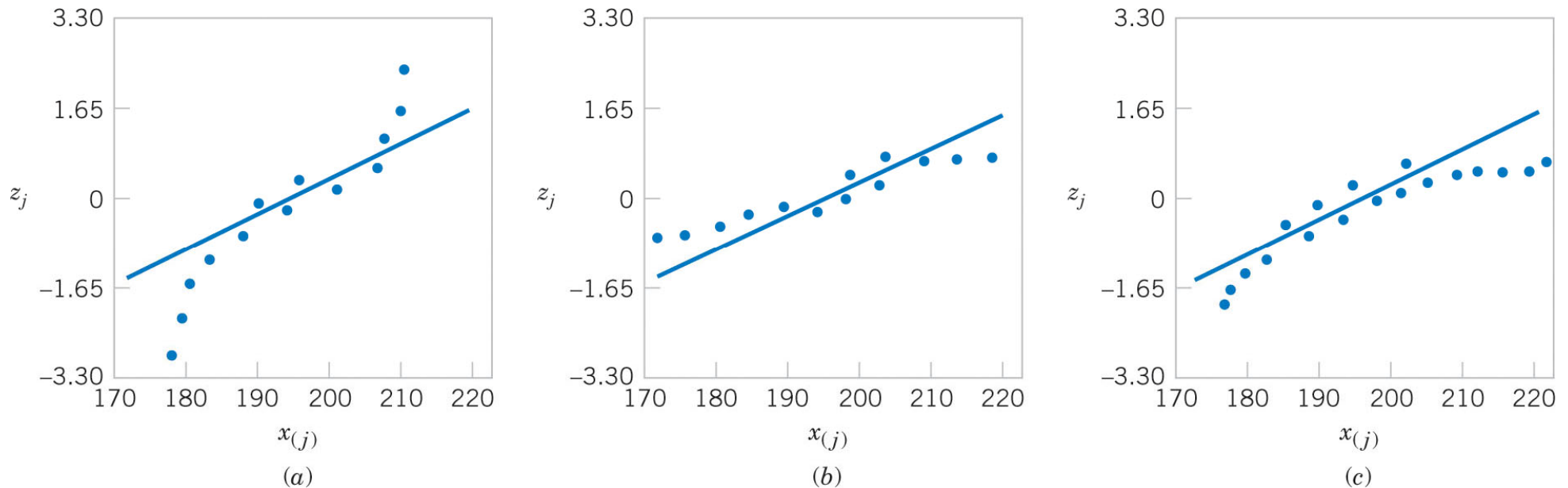
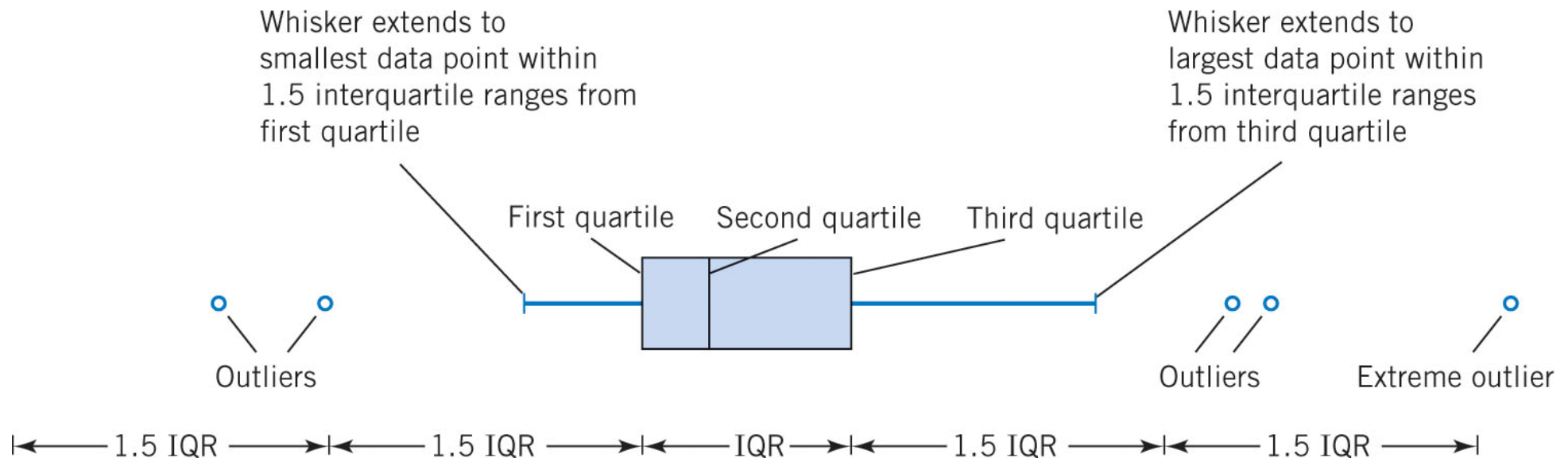


Figure 6-21 Normal probability plots indicating a non-normal distribution.

- (a) Light tailed distribution (squeezed together)
- (b) Heavy tailed distribution (stretched out)
- (c) Right skewed distribution (left end squeezed, right end stretched)

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 - Interquartile Range**
- **IQR=2.*norminv(0.75) % 1.3490**
- **% Calculate $0.5 \cdot \text{IQR} + 1.5 \cdot \text{IQR}$ - 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

WHY ARE THERE SLAVES IN THE BIBLE

WHY DO TWINS HAVE DIFFERENT FINGERPRINTS
WHY ARE AMERICANS AFRAID OF DRAGONS

WHY IS HTTPS CROSSED OUT IN RED
WHY IS THERE A LINE THROUGH HTTPS
WHY IS THERE A RED LINE THROUGH HTTPS ON FACEBOOK

WHY IS HTTPS IMPORTANT

WHY AREN'T MY
ARMS GROWING



WHY ARE THERE WEEKS
WHY DO I FEEL DIZZY

QUESTIONS

FOUND IN GOOGLE AUTOCOMPLETE

WHY AREN'T ECONOMISTS RICH

WHY DO AMERICANS CALL IT SOCCER

WHY ARE MY EARS RINGING

WHY ARE THERE SO MANY AVENGERS

WHY ARE THE AVENGERS FIGHTING THE X MEN

WHY IS WOLVERINE NOT IN THE AVENGERS

WHY ARE THERE SO MANY CROWS IN ROCHESTER, MN

WHY IS PSYCHIC WEAK TO BUG

WHY DO CHILDREN GET CANCER

WHY IS POSEIDON ANGRY WITH ODYSSEUS

WHY IS THERE ICE IN SPACE

WHY ARE THERE ANTS IN MY LAPTOP

WHY IS EARTH TILTED

WHY IS SPACE BLACK

WHY IS OUTER SPACE SO COLD

WHY ARE THERE PYRAMIDS ON THE MOON

WHY IS NASA SHUTTING DOWN

WHY ARE THERE MALE AND FEMALE BIKES

WHY ARE THERE TINY SPIDERS IN MY HOUSE

WHY DO SPIDERS COME INSIDE

WHY ARE THERE HUGE SPIDERS IN MY HOUSE

WHY ARE THERE LOTS OF SPIDERS IN MY HOUSE

WHY ARE THERE SPIDERS IN MY ROOM

WHY ARE THERE SO MANY SPIDERS IN MY ROOM

WHY DO SPIDER BITES ITCH

WHY IS DYING SO SCARY

WHY IS THERE NO GPS IN LAPTOPS

WHY DO KNEES CLICK

WHY AREN'T THERE E GRADES

WHY IS ISOLATION BAD

WHY DO BOYS LIKE ME

WHY DON'T BOYS LIKE ME

WHY IS THERE ALWAYS A JAVA UPDATE

WHY ARE THERE RED DOTS ON MY THIGHS

WHY IS LYING GOOD

WHY IS GPS FREE

WHY IS SEX SO IMPORTANT

WHY ARE THERE SWARMS OF GNATS

WHY IS THERE PHLEGM

WHY ARE THERE
GHOSTS



WHY IS THERE AN OWL IN MY BACKYARD

WHY IS THERE AN OWL OUTSIDE MY WINDOW

WHY IS THERE AN OWL ON THE DOLLAR BILL

WHY DO OWLS ATTACK PEOPLE

WHY ARE AK 47s SO EXPENSIVE

WHY ARE THERE HELICOPTERS CIRCLING MY HOUSE

WHY ARE THERE GODS

WHY ARE THERE TWO SPOCKS

WHY IS MT VESUVIUS THERE

WHY DO THEY SAY T MINUS

WHY ARE THERE OBELISKS

WHY ARE WRESTLERS ALWAYS WET

WHY ARE OCEANS BECOMING MORE ACIDIC

WHY IS ARWEN DYING

WHY AREN'T MY QUAIL LAYING EGGS

WHY AREN'T MY QUAIL EGGS HATCHING

WHY AREN'T THERE ANY FOREIGN MILITARY BASES IN AMERICA

WHY ARE CIGARETTES LEGAL
WHY ARE THERE DUCKS IN MY POOL
WHY IS JESUS WHITE
WHY IS THERE LIQUID IN MY EAR
WHY DO Q TIPS FEEL GOOD
WHY DO GOOD PEOPLE DIE

WHY AREN'T
THERE GUNS IN
HARRY POTTER



WHY ARE ULTRASOUNDS IMPORTANT
WHY ARE ULTRASOUND MACHINES EXPENSIVE
WHY IS STEALING WRONG

WHY DO WHALES JUMP
WHY ARE WITCHES GREEN
WHY ARE THERE MIRRORS ABOVE BEDS

WHY DO I SAY UH
WHY IS SEA SALT BETTER
WHY ARE THERE TREES IN THE MIDDLE OF FIELDS

WHY IS THERE NOT A POKEMON MMO
WHY IS THERE LAUGHING IN TV SHOWS
WHY ARE THERE DOORS ON THE FREEWAY

WHY ARE THERE SO MANY SVCHOST.EXE RUNNING
WHY AREN'T THERE ANY COUNTRIES IN ANTARCTICA
WHY ARE THERE SCARY SOUNDS IN MINECRAFT

WHY IS THERE KICKING IN MY STOMACH
WHY ARE THERE TWO SLASHES AFTER HTTP
WHY ARE THERE CELEBRITIES

WHY DO SNAKES EXIST
WHY DO OYSTERS HAVE PEARLS
WHY ARE DUCKS CALLED DUCKS

WHY DO THEY CALL IT THE CLAP
WHY ARE KYLE AND CARTMAN FRIENDS
WHY IS THERE AN ARROW ON AANG'S HEAD

WHY ARE TEXT MESSAGES BLUE
WHY ARE THERE MUSTACHES ON CLOTHES
WHY ARE THERE MUSTACHES ON CARS

WHY ARE THERE MUSTACHES EVERYWHERE
WHY ARE THERE SO MANY BIRDS IN OHIO
WHY IS THERE SO MUCH RAIN IN OHIO

WHY IS OHIO WEATHER SO WEIRD
WHY ARE THERE MALE AND FEMALE BIKES
WHY ARE THERE BRIDESMAIDS

WHY DO DYING PEOPLE REACH UP
WHY AREN'T THERE VARICOSE ARTERIES
WHY ARE OLD KINGDOMS DIFFERENT

WHY ARE THERE SQUIRRELS
WHY IS PROGRAMMING SO HARD
WHY IS THERE A 0 OHM RESISTOR

WHY DO AMERICANS HATE SOCCER
WHY DO RHYMES SOUND GOOD
WHY DO TREES DIE

WHY IS THERE NO SOUND ON CNN
WHY AREN'T POKEMON REAL
WHY AREN'T BULLETS SHARP
WHY DO DREAMS SEEM SO REAL



WHY IS SEX
SO IMPORTANT



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 $X_1, X_2, \dots, X_p$ and constants $c_1, c_2, \dots, c_p$
$$Y = c_1X_1 + c_2X_2 + \dots + c_pX_p \quad (5-24)$$
is a linear combination of $X_1, X_2, \dots, X_p$.

Mean & Variance of a Linear Function

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

$$E(Y) = c_1E(X_1) + c_2E(X_2) + \dots + c_pE(X_p) \quad (5-25)$$

$$V(Y) = c_1^2V(X_1) + c_2^2V(X_2) + \dots + c_p^2V(X_p) + 2 \sum_{i < j} \sum c_i c_j \text{cov}(X_i X_j) \quad (5-26)$$

If $X_1, X_2, \dots, X_p$ are **independent**, then $\text{cov}(X_i X_j) = 0$,

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

Descriptive statistics:
Sample mean and
variance

Some Definitions

- The random variables $X_1, X_2, \dots, X_n$ 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 $X_1, X_2, \dots, X_n$ 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

$$C_i = \frac{1}{n}$$

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

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

$$\bar{x} = \sum_{i=1}^n x_i / n = \sum_{i=1}^n \frac{1}{n} x_i$$

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 $\bar{X} = \frac{(X_1 + X_2 + \dots + X_n)}{n}$ and $E(X_i) = \mu$

$$\text{Then } E(\bar{X}) = \frac{n \cdot \mu}{n} = \mu \quad (5-28a)$$

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

$$\text{Then } V(\bar{X}) = \frac{n \cdot \sigma^2}{n^2} = \frac{\sigma^2}{n} \quad (5-28b)$$

Central Limit Theorem

If $X_1, X_2, \dots, X_n$ 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

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

for **large n** , is the **standard normal distribution**.

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