AMS 7: Statistical Methods For the Biological, Environmental and Health Sciences

1: Introduction and Descriptive Methods

David Draper

Department of Applied Mathematics and Statistics
University of California, Santa Cruz

draper@ams.ucsc.edu

http://www.ams.ucsc.edu/~draper

L-5

Outline

- **Introduction**: **populations** and **samples**; **parameters** and **statistics (estimates)**

- **Data types**: qualitative and quantitative variables; nominal and ordinal; discrete and continuous; interval and ratio; dichotomous

- **Descriptive methods** for a single variable
  - **Graphical**: histograms, bar charts
  - **Numerical**: measures of **center** (mean, median, mode) and **spread** (standard deviation, variance)

- Using the **normal distribution** descriptively

1.1 Introduction

Statistics is the **study of uncertainty**: how to **measure it**, and what to **do about it**.

Uncertainty is a state of **incomplete** or **imperfect information** about something of interest to you, for example

theta = p

the **percentage** θ of the deer who live on the UCSC campus as of ~~31 December 2006~~ who have **chronic wasting disease**. 31 Mar 2018

I notice that I **don't know** the value of θ **exactly**; I have the impression that θ is **rather small**, since the deer on campus seem **relatively healthy**, but I have **substantial uncertainty** about its **precise value**.

I can **reduce my uncertainty** by **gathering data** on the disease status of campus deer; **how** should this data-gathering be done?

The **set**

$\mathcal{P}$ = {**the deer who live on the UCSC campus as of** ~~**31 December 2006**~~} 31 Mar 2018

is an example of a **population**: a collection of **subjects** or **elements** (in this case, deer) of interest to me.

There is an **aspect** of each of these population subjects that I'm curious about: if I encountered one of these deer, the **question** I would ask is **"Does this deer have chronic wasting disease or not?"**

Things that can be **measured** on population subjects are called **variables**; in this case the variable of interest takes on only two values, {**yes, no**} (such variables are called **dichotomous** or **binary**).

Populations and Samples; Parameters and Statistics (Estimates)

We'll see soon that a **handy** way to work with dichotomous variables is to assign **1** and **0** to their two possible values (hence the term **binary**); for example, with the variable **(chronic wasting disease or not)** the coding **(1 = yes, 0 = no)** is particularly useful.

A **numerical summary** of a population is called a **parameter**; θ is an example of one possible parameter of interest about the population $\mathcal{P}$ above (others might include the **average weight** of the deer who are **more than three months old**).

If I had enough **time** and **money** (and a way of ensuring that I could **find** all the deer and mark them **uniquely**, so that I didn't **double-count** any individual), in principle I could perform a **complete census** of the entire population, obtaining the **disease status** for each individual, and at the end of this census **I would no longer have any uncertainty** about the parameter θ.

In practice people **rarely** have enough time and money to perform a **complete census** of a population $\mathcal{P}$; instead it's natural to choose a **subset** $\mathcal{S}$ of $\mathcal{P}$ and evaluate the variable(s) of interest **only on the population subjects in the subset**. &

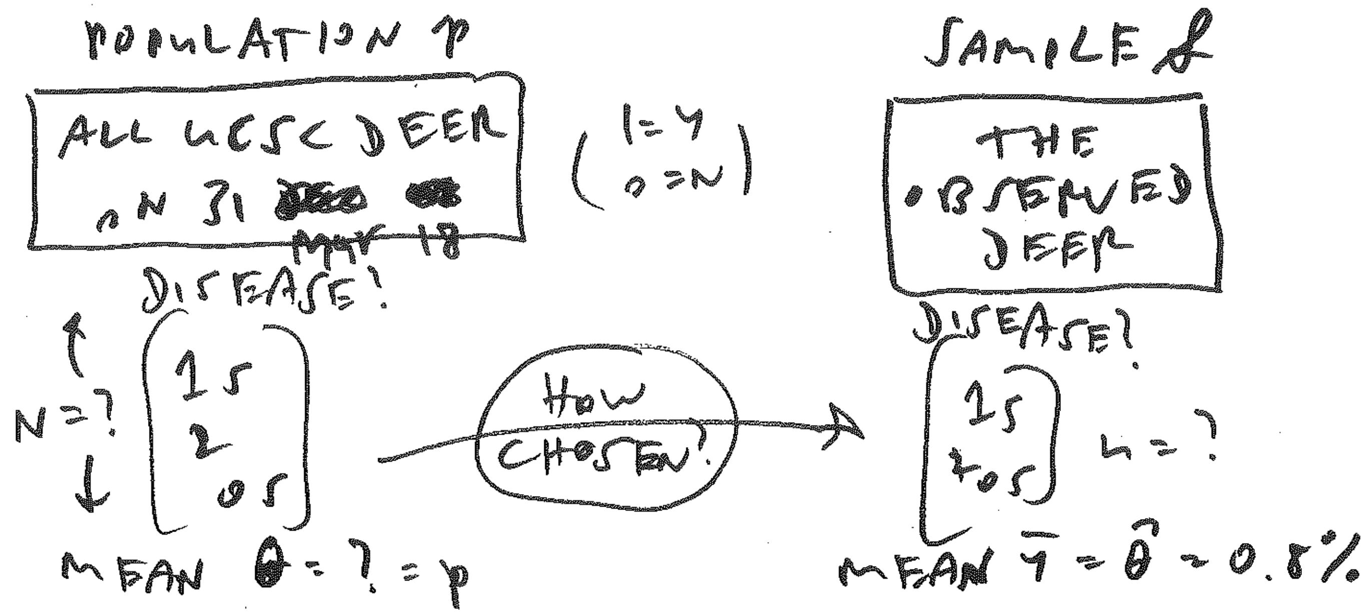
Such a subset is called a **sample** from the population $\mathcal{P}$ — if the sample is chosen well, it seems like a good idea to use the data in the sample to make an **estimate** of (an educated guess at) the population parameter θ of interest.

An estimate $\hat{\theta}$ of a population parameter θ is also sometimes called a **statistic**.

Populations and Samples

Let's let N stand for the **number of subjects** in the **population** $\mathcal{P}$, and n denote the **number of subjects** in the **sample** $\mathcal{S}$.

Then both the **population** and the **sample** can be thought of as <u>data sets</u>, which can be further visualized as **rectangular tables*** with **one row** for each **subject** and **one column** for each **variable**, as in the following **diagram**:



In this class we'll be looking a lot at diagrams like this one: such diagrams are the basis of both **probability models** and **statistical models**, both of which are crucial to the process of **quantifying uncertainty**.

To fill out a diagram like this I need to specify the following **ingredients**:

- In the **box** above the population data set $\mathcal{P}$ I describe the subjects in $\mathcal{P}$ by saying to myself **"There's one row in the population for each ____"** and filling in the blank; for example, here there's **one row for each deer living on the Santa Cruz campus on 31 December 2006**.

*__Math note:__ the official name for such a rectangular table is a **matrix**.

Random Sampling

• Above each **column** in the population data set I write the name of the **variable** summarized by that column (in this part of the class we'll typically work with only **one variable at a time**); here the variable of interest is the answer to the question **(chronic wasting disease?)**.

• I identify the **number** N of subjects in the population if I know it (here I'm **not sure** how many deer there are on the UCSC campus in December 2006, so I just put a **question mark**).

• Then I do the same three things for the **sample** data set: in the box above S I describe the **subjects** in the **sample** (here I might just say **"the observed deer"**); above each **column** in the sample data set I write the name of the **variable** summarized by that column (this will be the **same** as in the population); and I identify the number n of **subjects** in the **sample** if I know it (here we haven't yet talked about **how large the sample should be**, so again I just put a **question mark**).

There's one crucial thing about this concept of **using the sample data to estimate a parameter** of interest in the population: I said above that this is a good idea **"if the sample is chosen well,"** and we need to figure out what this means.

Evidently, if the sample is to serve as a **good stand-in** for the rest of the population, the basic principle we want to follow is to try to make the **sample** and the unsample (the part of the population not chosen in the sample) **as similar as possible in all relevant ways**.

The simplest way to achieve this goal turns out to be to **draw the sample at random from the population** (so that all subsets have an **equal chance** of being chosen).

SRS and IID Sampling

So the last (and perhaps **most crucial**) step in filling out the **diagram** above is as follows:

- Finally, in the **circle** above the **arrow** from the population to the sample I describe the **sampling method** (in this case, random).

To **literally take a random sample** of size n of deer from $\mathcal{P}$, you'd have to

(a) make a **list** of all the **population subjects** (deer), with **unique identifying tags**;

(b) choose n of these tags at random (using, for example, **pseudo-random numbers** generated by computer) **without replacement** (the sampling method **at random without replacement** is called **simple random sampling** (SRS), as opposed to **at random with replacement**, which is called **independent identically distributed** (IID*) sampling); and

(c) measure the **variables** of interest on the sampled deer by finding the ones with the **chosen tags**.

In practice people would often instead use a **simpler** method that's not literally SRS (for example, if the deer were well distributed spatially, you could **partition** the UCSC campus into n non-overlapping and exhaustive spatial subsets and have n people each get data on the **first deer they encounter** in their subset on a given day) and then argue that their simpler method was **like** what you would get with SRS.

*__Textbook note:__ IID is a term not mentioned in Triola & Triola.

1.2 Data Types

It's useful to have a **classification** of the various **types of data** that variables can keep track of (because some methods of analysis are definitely **not appropriate** for some data types).

Example 1: **Genetic phenotype**. **Eye color** in an animal you're studying may take on only **two values** (brown, blue) that have **no unique place on the number line** (earlier we called such variables **dichotomous** or **binary**); similarly, **hair color** might take on **four values** (predominately brown, black, red or white).

Variables like this are said to occur on a **nominal** scale of measurement (so dichotomous variables with values like {yes, no} are **special cases** of nominal variables).

Example 2: **Success in running a maze** might be recorded

1 (**very slow**), 2 (**slow**), 3 (**moderate**), 4 (**fast**), 5 (**very fast**)

There are still **no unique places on the number line** for such values, but (unlike example 1) there's a **natural ordering** to these values.

Variables like this are said to occur on an **ordinal** scale.

Some other names for **nominal** and **ordinal** variables are **qualitative** and **categorical**.

Data Types (continued)

$\boxed{\text{Example 3:}}$ **Size of a plant**. Two measures of the **size** of a plant (which, in turn, is a measure of its **competitiveness**) would include its **height** (in centimeters (cm)) and the **number of leaves** it has.

Unlike the situations in Examples 1 and 2, the values taken on by these variables **do have unique places on the number line**, and in fact there are two important characteristics of the numerical values of these variables:

(a) there is a **constant size interval** between any adjacent units on the measurement scale, so that the concept of **1 unit** means the same thing anywhere on the scale (for example, plants A, B, C and D are (respectively) 14, 15, 62, and 63 cm high; the **amount** by which B is **taller** than A is the same as the **amount** by which D is **taller** than C); and

(b) there is a **true zero** on the measurement scale with a **direct physical meaning** — this allows us to make meaningful statements about **ratios** (for example, plant C is $\frac{62\text{cm}}{15\text{cm}} \doteq 4.1$ times taller than plant B).

Variables like this are said to occur on a **ratio** scale.

$\boxed{\text{Example 4:}}$ **Growing temperature** at which a plant produces the most buds. Temperature (measured either in $^{\circ}C$ or $^{\circ}F$) does have a **constant size interval** but **lacks a true zero**, so (contrary to statements you see in the newspaper or on TV) when it's $80^{\circ}F$ outside you can't correctly say that it's **twice as hot** as when it's $40^{\circ}F$.

Variables like this are said to occur on an **interval** scale.

Data Types (continued)

Some other names for **ratio** and **interval** variables are **quantitative** and **numerical**.

One last distinction: plant **height** and **number of leaves** are different in that with plant height, **conceptually** (with finer and finer measuring instruments) there are **no possible gaps between the possible values**, whereas with number of leaves, **distinct structural gaps** exist (it doesn't make sense to talk about $4\frac{1}{2}$ leaves).

Quantitative variables with **gaps** between the possible values are called **discrete**; quantitative variables with **no conceptual gaps** between the possible values are called **continuous**.

Why these distinctions matter. Suppose I choose to **code** the **age** of some animals I'm observing in the following way, when **storing** the values of this variable in a **computer**:

less than 1 year old = 0, **between 1 and 2 years old** = 1, **between 2 and 3 years old** = 2, **between 3 and 4 years old** = 3, ...

Suppose further that I choose to **code** the hair color of these animals in the following way:

brown = 0, **black** = 1, **red** = 2, **white** = 3

Here's the data set I get (written, to save space, in a **transposed** fashion in relation to the convention on page 5 above: here the **rows** are the **variables** and the **columns** are the **subjects** (animals)):

							Mean
Animal Identifier	45	333	167	2	⋯	501	243.9
Age	2	0	1	1	⋯	3	1.7
Hair Color	0	2	3	2	⋯	1	1.2

1.3 Descriptive Methods

As we'll soon discuss, it's sometimes both **useful** and **meaningful** to **summarize** a variable by taking its **mean** (just add 'em up and divide by how many there are); the computer has done this for us in the table above in the **final column**.

The problem is, of course, that the **mean** is **meaningful** only for the **age** variable (because it's **quantitative** [ratio, discrete]; the other two variables are **qualitative** [nominal]).

The point: The **right way** to **analyze** a variable often depends on the **scale** on which it's measured.

1.3.1 Graphical descriptive methods. **Example: butterfly wing lengths**. Zar (1999) gives data from a sample of $n = 24$ immature monarch butterflies, in which the **variable** of interest (we might call it y; T&T would call it x) is **wing length** (in cm):

4.4 3.6 4.1 3.3 3.5 3.8 4.5 4.3 4.3 4.0 4.1 3.6
4.0 4.0 3.8 3.8 3.9 4.2 4.2 4.1 3.7 3.9 4.0 3.9

(This is just **shorthand** for a **data set** with $n = 24$ **rows** (**subjects** = **butterflies**) and 1 **column** (**variable** = **wing length**), written in this manner to save space.)

How might we **summarize** this variable in a way that would allow us to **see patterns** (**graphical** summaries) and to capture **most of the information it contains** in **fewer than 24 numbers** (**numerical** summaries)?

Raw Frequency Distribution

As long as the **order** in which the data values were listed above is **not relevant**, the first step would be to **sort** the data from **smallest** to **largest**:

3.3 3.5 3.6 3.6 3.7 3.8 3.8 3.8 3.9 3.9 3.9 4.0
4.0 4.0 4.0 4.1 4.1 4.1 4.2 4.2 4.3 4.3 4.4 4.5

Now we can see that there are a number of **duplicate values** (caused by **rounding** the wing length measurement to the nearest cm).

This suggests a **further summary** in which we keep track of the **values** of the variable and the **raw frequencies** (the numbers of times those values are attained):

Value	Frequency
3.3	1
3.4	0
3.5	1
3.6	2
3.7	1
3.8	3
3.9	3
4.0	4
4.1	3
4.2	2
4.3	2
4.4	1
4.5	1
Total	$n = 24$

This is called a **raw frequency distribution** (or **frequency table**) for the variable y (sometimes people just refer to the **distribution** of y, or ask "How is y **distributed**?").

Raw Frequency Histogram

The **table** on the previous page is not as easy to **absorb** as it would be if we could display it **graphically**.

Since it has **two columns** or **dimensions**, it's natural to make a plot in which one dimension (**horizontal**, say) is the **values** the variable takes on and the other (**vertical**, say) is the **raw frequencies** — the resulting graph is a (raw frequency) **histogram** of the variable y:

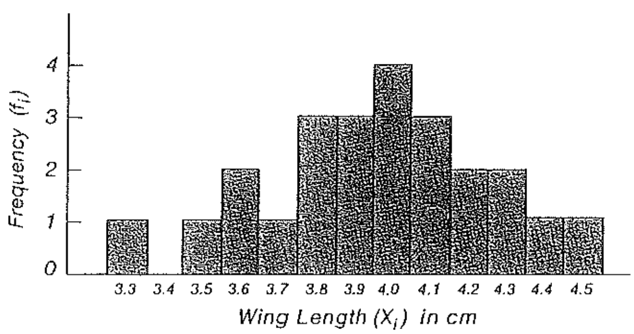


Figure 3.1 A histogram of the data in Example 3.2. The mean (3.96 cm) is the center of gravity of the histogram, and the median (3.975 cm) divides the histogram into two equal areas.

A **histogram** is a special case of a **bar graph**: a plot with **locations** identified along the **horizontal** axis corresponding to **values** a variable takes on and **bars** over those locations with **heights** give by the (raw) **frequencies** of those values.

A **bar graph** can be drawn as a summary of **any qualitative (nominal or ordinal) variable**; there is no unique place called "yes" or "red" on the number line, but you can just **invent arbitrary horizontal locations** and make a useful plot anyway.

Strictly speaking, what makes a **histogram** a histogram is that the variable in question is **quantitative** (so that the values do have unique locations on the number line) — histograms can be made for either **discrete** or **continuous** variables.

More Graphical Examples

EXAMPLE 1.1 The location of sparrow nests. A frequency table of nominal data.

Nest site	*Number of nests observed*
A. Vines	56
B. Building eaves	60
C. Low tree branches	46
D. Tree and building cavities	49

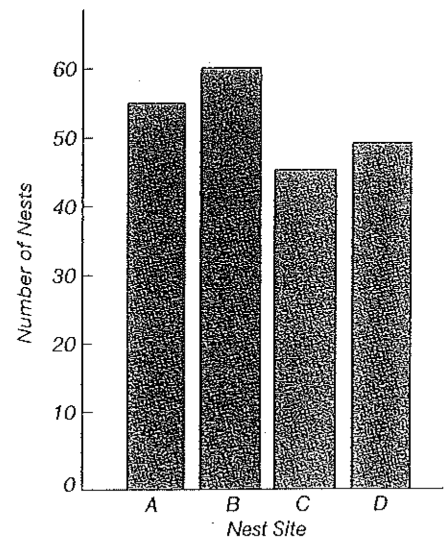


Figure 1.1 A bar graph of the sparrow nest data of Example 1.1. An example of a bar graph for nominal data.

EXAMPLE 1.2 Numbers of sunfish, tabulated according to amount of black pigmentation. A frequency table of ordinal data.

Pigmentation class	*Amount of pigmentation*	*Number of fish*
0	No black pigmentation	13
1	Faintly speckled	68
2	Moderately speckled	44
3	Heavily speckled	21
4	Solid black pigmentation	8

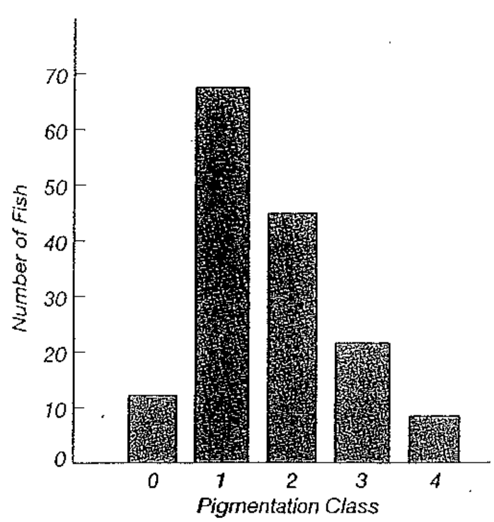


Figure 1.3 A bar graph of the sunfish pigmentation data of Example 1.2. An example of a bar graph for ordinal data.

L- 18

Graphical Examples (continued)

EXAMPLE 1.3 **Frequency of occurrence of various litter sizes in foxes. A frequency table of discrete, ratio-scale data.**

Litter size	*Frequency*
3	10
4	27
5	22
6	4
7	1

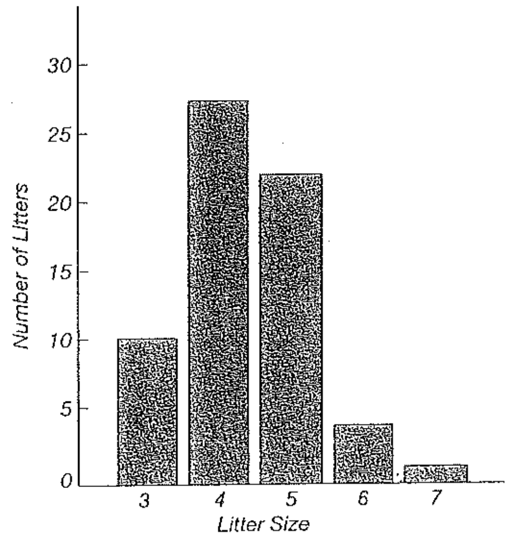


Figure 1.4 A bar graph of the fox litter data of Example 1.3. An example of a bar graph for discrete, ratio-scale data.

EXAMPLE 1.4a **Number of aphids observed per clover plant. A frequency table of discrete, ratio-scale data.**

Number of aphids on a plant	*Number of plants observed*	*Number of aphids on a plant*	*Number of plants observed*
0	3	20	17
1	1	21	18
2	1	22	23
3	1	23	17
4	2	24	19
5	3	25	18
6	5	26	19
7	7	27	21
8	8	28	18
9	11	29	13
10	10	30	10
11	11	31	14
12	13	32	9
13	12	33	10
14	16	34	8
15	13	35	5
16	14	36	4
17	16	37	1
18	15	38	2
19	14	39	1
		40	0
		41	1

Total number of observations = 424

Graphical Examples (continued)

EXAMPLE 1.4b Number of aphids observed per clover plant. A frequency table grouping the discrete, ratio-scale data of Example 1.4a.

Number of aphids on a plant	*Number of plants observed*
0–3	6
4–7	17
8–11	40
12–15	54
16–19	59
20–23	75
24–27	77
28–31	55
32–35	32
36–39	8
40–43	1

Total number of observations = 424

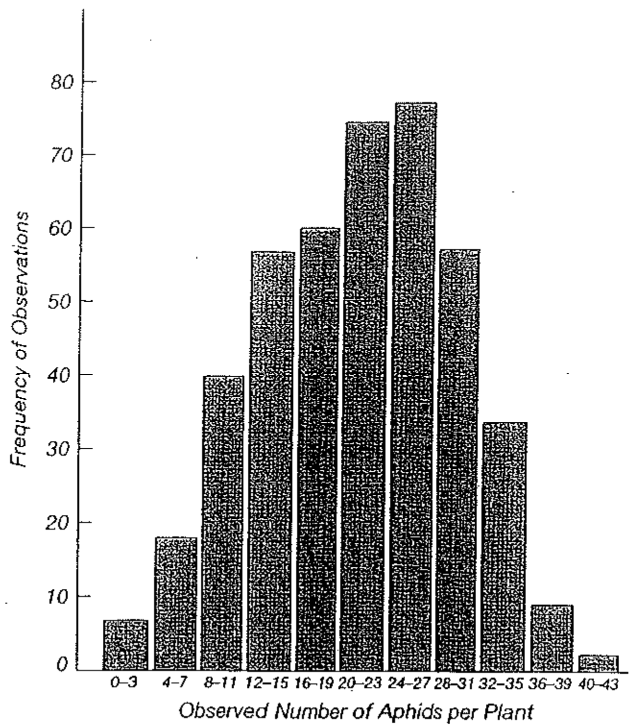


Figure 1.5 A bar graph of the aphid data of Example 1.4b. An example of a bar graph for grouped discrete, ratio-scale data.

Graphical Examples (continued)

EXAMPLE 1.5 Determinations of the amount of phosphorus in leaves. A frequency table of continuous data.

Phosphorus mg/g of leaf)	*Frequency (i.e., number of determinations)*	**Cumulative frequency** *Starting with low values*	*Starting with high values*
8.15–8.25	2	2	130
8.25–8.35	6	8	128
8.35–8.45	8	16	122
8.45–8.55	11	27	114
8.55–8.65	17	44	103
8.65–8.75	17	61	86
8.75–8.85	24	85	69
8.85–8.95	18	103	45
8.95–9.05	13	116	27
9.05–9.15	10	126	14
9.15–9.25	4	130	4

Total frequency = 130

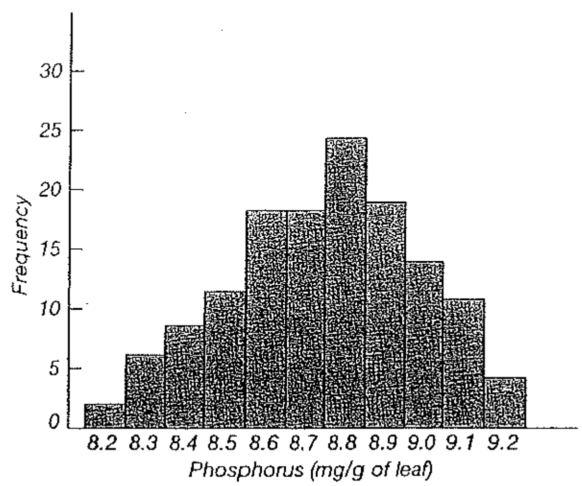


Figure 1.6a A histogram of the leaf phosphorus data of Example 1.5. An example of a histogram for continuous data.

1.3.2 Numerical Descriptive Methods

In addition to **summarizing** a variable **graphically** to look for **patterns**, it's also useful to **summarize** it **numerically**, to capture **most of the information it contains** in **fewer than** n **numbers**.

People have found **two main types** of **numerical** summary useful: measures of <u>**center**</u> (or <u>**central tendency**</u>), and measures of <u>**spread**</u> (or <u>**dispersion**</u>, or <u>**variability**</u>).

$\boxed{\text{Measures of center.}}$ The three most useful are the **mean**, the **median** (and other **quantiles**, or **percentiles**), and the **mode**.

The <u>**mean**</u> is, I'm sure, an old familiar object: with a variable like **butterfly wing length** above ($n = 24$)

$$\begin{array}{cccccc} 4.4, & 3.6, & 4.1, & \ldots, & 4.0, & 3.9 \\ y_1, & y_2, & y_3, & \ldots, & y_{23}, & y_{24} \\ y_1, & y_2, & y_3, & \ldots, & y_{n-1}, & y_n \end{array}$$

to get the **mean** (for a variable y, let's call the mean $\bar{y}$) I just **add up** all the data values and **divide** by **how many there are**:

$$\bar{y} = \frac{4.4 + 3.6 + 4.1 + \cdots + 4.0 + 3.9}{24} = \frac{95}{24} \doteq 3.96\,\text{cm}. \quad (1)$$

Symbolically, using the idea of **summation notation**, this can be written more **succinctly** as

$$\begin{aligned} \bar{y} &= \frac{y_1 + y_2 + y_3 + \cdots + y_{n-1} + y_n}{n} \\ &= \frac{1}{n}(y_1 + y_2 + \cdots + +y_n) = \frac{1}{n}\sum_{i=1}^{n} y_i. \end{aligned} \quad (2)$$

Measures of Center

It turns out that the **mean** has the **graphical interpretation** of the **center of gravity** of the data: if you visualize the **histogram** of the variable y as made of **bricks** that are sitting on a **number line** made of **plywood**, which in turn is put **on top of a saw-horse**, the mean is **the place where the histogram would exactly balance**.

The <u>**median**</u> $\tilde{y}$ of a column of numbers $(y_1, \ldots, y_n)$ is defined to be the **middle** value in the list in which $(y_1, \ldots, y_n)$ has been **sorted** from smallest to largest — if n is an **odd** number this is **uniquely defined**; if n is **even** there is **no single middle number** and people define the median to be the **mean** of the two middle values.

By examining the **sorted list** of the **butterfly wing lengths** on page 12 above ($n = 24$), you can see that the **middle two numbers** are both **4.0**, so the **median** $\tilde{y}$ of that variable is **4.0cm** (**not very far** from the **mean** in this case).

Since **histograms** graphically display the **frequency distribution** of a variable, the **graphical interpretation** of the **median** must be that it's the place where **half of the data is to the left of that place** and **half to the right** in the histogram.

The **median** is a **special case** of the general idea of finding places in the distribution where a particular **percentage** of the data is **to the left** of that place — these are called <u>**quantiles**</u> or <u>**percentiles**</u>.

By definition the **median** is the **50th percentile**, but it's also useful sometimes to look at **other percentiles** — for example, the **25th percentile** (also called the **first quartile**) is the place where $\frac{1}{4}$ of the data is to the **left** of that place, and similarly the **75th percentile** (the **third quartile**) is the place where $\frac{1}{4}$ of the data is to the **right** of that place.

Mean, Median, Mode

The mode is defined to be the **point of highest frequency** in the distribution of a variable, so by definition its **graphical interpretation** is just the **highest point in the histogram** (**Note:** there are **many different possible histograms** for the same set of data [as a function of **how wide you choose the bars** to be and **where you decide to start the first bar**], and the mode is **sensitive** to these choices).

Some distributions have **more than one mode** — these are called **multimodal** (a common special case is **two modes**, which defines a **bimodal** distribution), and distributions with only one mode are therefore **unimodal**; usually multimodality means that there are **two or more subgroups** in the sample that should perhaps be studied **separately**.

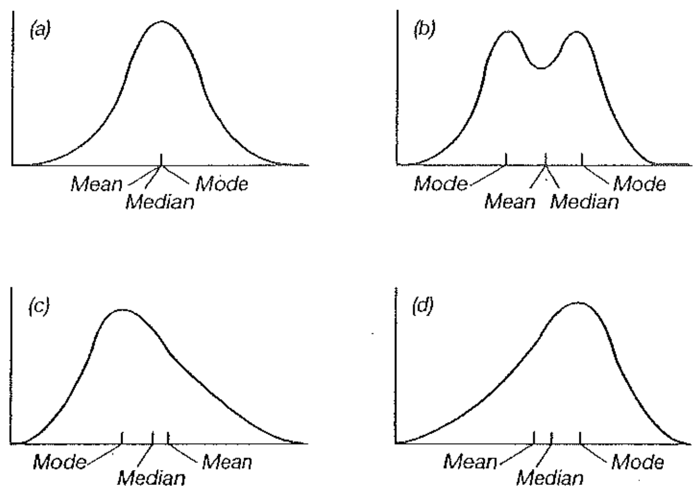


Figure 3.2 Frequency distributions showing measures of central tendency. Values of the variable are along the abscissa (horizontal axis), and the frequencies are along the ordinate (vertical axis). Distributions (a) and (b) are symmetrical, (c) is positively skewed, and (d) is negatively skewed. Distributions (a), (c), and (d) are unimodal, and distribution (b) is bimodal. In a unimodal asymmetric distribution, the median lies about one-third the distance between the mean and the mode.*

Distributions that are composed of **mirror images to the left and right of a central folding point** are called **symmetric**; combining two of these terms, a **fairly common distributional shape** is **symmetric unimodal** (like Figure (a) above) — for such distributions the mean, median and mode all **coincide** with the **point of symmetry**, which makes choosing a measure of center for them **easy**.

Outliers; Sensitivity Analysis

EXAMPLE 3.3 **Life expectancy for two hypothetical species of birds in captivity.**

Species A X_i (mo)	*Species B* X_i (mo)
34	34
36	36
37	37
39	39
40	40
41	41
42	42
43	43
79	44
	45

$n = 9$
$\mathcal{M} = X_{(n+1)/2} = X_{(9+1)/2}$
$= X_5 = 40$ mo
$\bar{X} = 43.4$ mo

$n = 10$
$\mathcal{M} = X_{(n+1)/2} = X_{(10+1)/2}$
$= X_{5.5} = 40.5$ mo
$\bar{X} = 40.1$ mo

The parts of a distribution to the **left** and **right** of the **center** are called the left and right tails of the distribution, respectively.

Notice in the data set for species A above that the **largest observation** in the **right tail** (a life expectancy of **79** months (mo)) is **much larger** than the others — data values in either tail that are **far from the bulk of the data** are called outliers.

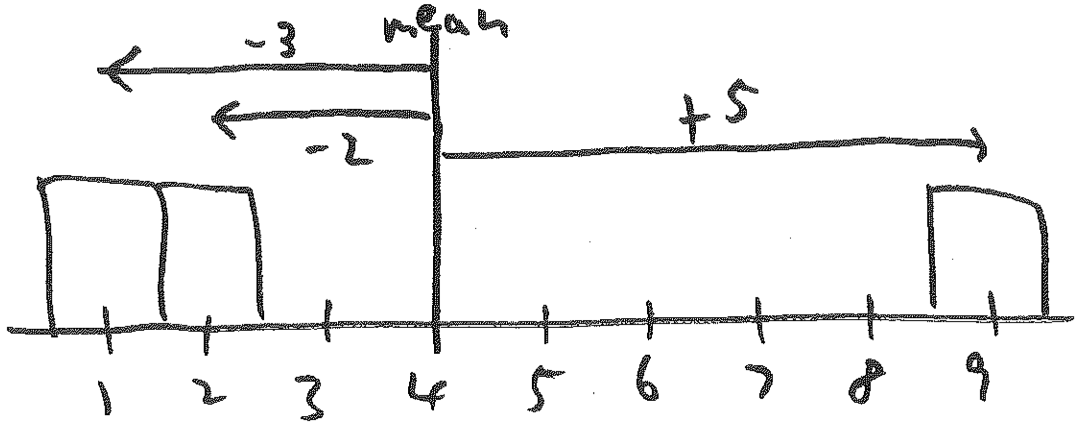
You can see that with data set A the **mean** has been **quite strongly influenced** by the **outlier** (the **median** is **40** mo, and the mean has been **pulled by the outlier** all the way up to **43.3** mo) — this observation (the fact that the **mean is more sensitive to outliers than the median**) is part of a **general** phenomenon.

We'll see examples later of **what to do about outliers** — one of the **simplest ways** to address them is by means of sensitivity analysis (**repeat** your **main analyses with** and **without** the outliers and see if you get **more or less the same results** — if so, **great**; if not, you have to **think harder** about whether there's a **good scientific reason** to **discard** the outliers).

Measures of Spread

Measures of spread. The two most useful are the **variance** and the **standard deviation** (SD).

To see what's going on in **measuring** the **spread** of a list of numbers, consider the tiny **fake data set** with $n = 3$ values $(y_1, \dots, y_n) = (1, 2, 9)$, whose **mean** $\bar{y}$ is **4** and whose **histogram** looks like this:



We might **intuitively define** the **spread** of a list of numbers to be the **typical amount** by which the numbers **differ** from their **center**; how should this idea be **quantified**?

One way might begin by calculating the <u>**deviations**</u> $(y_i - \bar{y})$ of each **observation** y_i from the **mean** $\bar{y}$:

$$\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{bmatrix} 1 \\ 2 \\ 9 \end{bmatrix} \xrightarrow[\bar{y} = 4]{\text{subtract}} \begin{pmatrix} y_1 - \bar{y} \\ \vdots \\ y_n - \bar{y} \end{pmatrix} = \begin{bmatrix} 1 - 4 = -3 \\ 2 - 4 = -2 \\ 9 - 4 = +5 \end{bmatrix}$$

The **deviations** represent the **amount by which each number differs from the center** (as measured by the **mean**); all we need to finish off the calculation is to **summarize** them (to get the **"typical"** deviation).

Variance

One way to **summarize** the deviations would be to take their **mean**, but this **doesn't work**: with the example above you get **0**, and in fact by the way the mean is defined you would **always get 0** no matter what the numbers $(y_1, \ldots, y_n)$ are (it's not hard to **prove** this **algebraically**) — the problem is **cancellation** of $+$ and $-$ deviations.

One way to **avoid cancellation** is to take the **mean** of the **absolute values** of the deviations — this is the **mean absolute deviation** (MAD), which here comes out $\frac{|1-4|+|2-4|+|9-4|}{3} = \frac{3+2+5}{3} \doteq 3.3$; this does seem to correspond to the **typical length** of the **arrows** in the **sketch** above.

For **technical reasons** having to do with **calculus** and **theoretical statistics**, the MAD is **not used much**; the two **most frequently used** measures of spread are based on the **other way** to **avoid cancellation**: taking (more or less) the mean of the **squares** of the deviations.

Again for **technical reasons** (which will be explained later), when the data set you're measuring the spread of is a **(random) sample** from a **population**, in calculating this **"mean"** people prefer to divide not by the number of data values n but by $(n-1)$; the resulting quantity is called the **(sample) variance**, usually abbreviated s^2:

$$s^2 = \frac{(y_1 - \bar{y})^2 + \ldots + (y_n - \bar{y})^2}{n-1} = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2. \qquad (3)$$

Thinking of the little **fake data set** as a **sample**, the **(sample) variance** comes out $\frac{(1-4)^2+(2-4)^2+(9-4)^2}{3-1} = \frac{3^2+2^2+5^2}{2} = \mathbf{19}$, which seems **too big** as a measure of the **typical length** of the **arrows** in the **sketch** above, and moreover **the units of the variance are wrong**: if the data values were measurements of **money** ($, say), the variance would come out in $\$^2$.

Standard Deviation

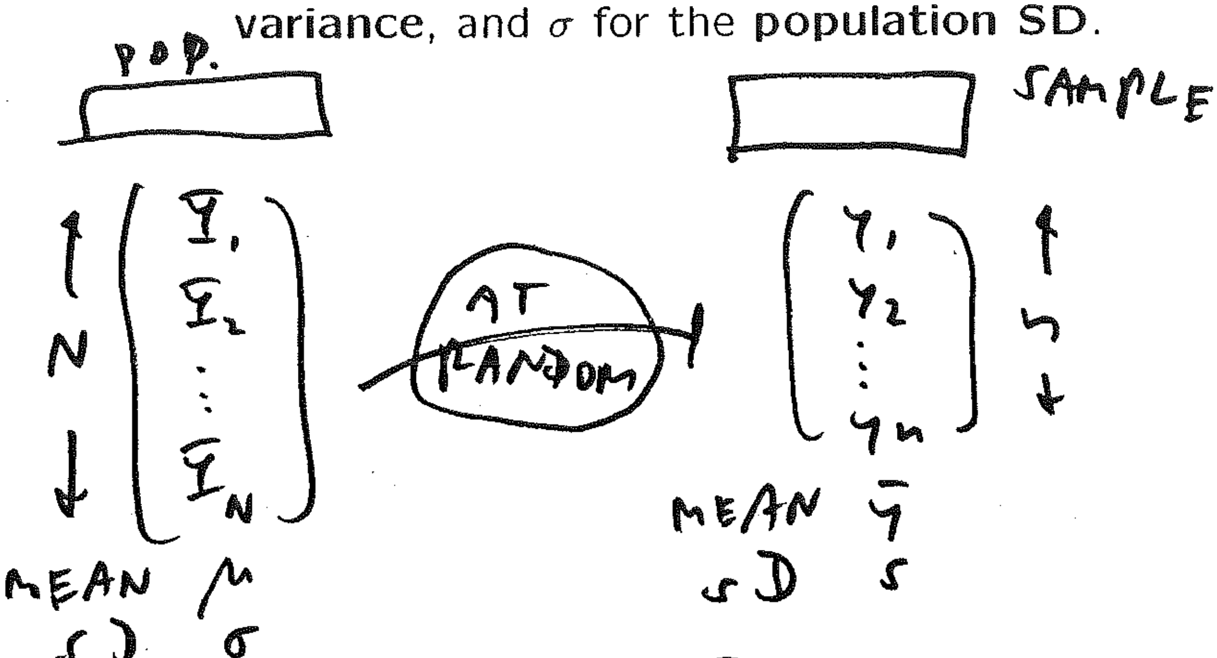
Both of these problems can be **solved** by taking the **square root** of the variance, which is called the **standard deviation** (**SD** for short), usually abbreviated s; for the **sample** this is

$$s = \sqrt{\frac{(y_1 - \bar{y})^2 + \ldots + (y_n - \bar{y})^2}{n-1}} = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2}. \quad (4)$$

With the little **fake data set** the **sample SD** comes out $s = \sqrt{19} \doteq 4.4$, which seems **about right** as a **summary of spread** for this list of numbers.

Samples and populations. Suppose that the **data set** $(y_1, \ldots, y_n)$ is a **random sample** of size n from a **population** of N values, and to keep from getting **confused** between the population and the sample let's call the **population** values $(Y_1, \ldots, Y_N)$; we've already agreed to call the **sample mean, variance** and **SD** $\bar{y}$, s^2 and s, respectively.

With this notation it would be **natural** to call the **population mean, variance** and **SD** $\bar{Y}$, S^2 and S, respectively, but instead people typically use **Greek letters** and write μ for the **population mean**, σ^2 for the **population variance**, and σ for the **population SD**.



Graphical Interpretation of the SD

In this setup, to make things even more **mysterious**, people define the **population variance** and **SD** not by dividing by $(N-1)$ but by N:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(Y_i - \mu)^2 \text{ and } \sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(Y_i - \mu)^2}. \tag{5}$$

The **reason** for all of this **mystery** will be explained later when we talk about **sampling distributions**; for now it's enough to notice that when n is **large** it **hardly matters** in practice whether you divide by n or $(n-1)$ in calculating the **sample SD** s.

$\boxed{\text{Graphical interpretation of the SD.}}$ SDs are a **pain** to **compute** by **hand** or with a **calculator**, and it's easy to **make mistakes** when doing so, so it would be good to have a **simple way** to **roughly approximate** the **SD** of a list of numbers **by looking at its histogram** — this is provided by something called the **Empirical Rule**.

Empirical Rule: For **almost any** list of numbers, if you **start at the mean** and **go one SD either way**, you'll **capture** about $\frac{2}{3}$ of the data (the theoretical figure is **68%**); if you **start at the mean** and go **2 SDs either way** you'll get **most** of the data (the theoretical number to remember is **95%**); and if you **start at the mean** and go **3 SDs either way** you'll get **almost all** of the data (the theoretical figure is **99.7%**).

Looking back at the **butterfly wing length** example, for instance (the **sorted data values** are on page 12, and the **histogram** is on page 13; recall that the **mean** is about **4 cm**), you can see that if you guessed the **SD** was about **0.1 cm** that would be **too small** (since **starting at 4.0** and

going 0.1 either way, down to 3.9 and up to 4.1, would only get you **half** of the data), and if you guessed $s = 0.5$ that would be **too big**, because the interval from (**mean** − **1 SD**) = (4.0 − 0.5) = 3.5 to (**mean** + **1 SD**) = (4.0 + 0.5) = 4.5 includes **all** of the data; a bit of **trial and error** should convince you that the **SD** is around **0.3 cm**, and actually computing it yields $s = 0.29$ cm.

Using the normal distribution descriptively. There's a **distribution** that's quite **special** in many ways (you've probably already met it) — it's the **symmetric unimodal** distribution called the **normal** (or **Gaussian**) **distribution**.

Actually there's **not just one normal distribution**, there are **infinitely many** of them: for any values $\bar{y}$ and s you can imagine for the **mean** and **SD** (respectively) of a **single-variable data set** (thought of as a **sample** from a **population**), there's a **normal distribution** with that **mean** and **SD**.

Looking at the **histogram** of the **butterfly wing length data** (page 13), which is roughly **symmetric** and **unimodal**, you can imagine someone **approximating** it with a **smooth curve** drawn through or near the tops of the bars — the **curve** (or **function**) corresponding to the **normal distribution** with mean $\bar{y} = 4.0$ **cm** and SD $s = 0.29$ **cm** has the **equation** (with y as the values of the variable running along the **horizontal** axis)

$$f(y) = \frac{1}{s\sqrt{2\pi}} \exp\left[-\frac{(y - \bar{y})^2}{2s^2} \right]. \qquad (6)$$

The Normal Curve

The idea behind **using the normal distribution (also called the normal curve) descriptively** is as follows.

Suppose you wanted to **answer** a **question** like "What **percentage** of the butterflies in this data set had **wing lengths smaller than 3.56 cm?**"

The **exact** way to answer this question is just to **count** how many butterflies satisfied this criterion (2, as it turns out: the ones with wing lengths of **3.3** and **3.5 cm**) and see what **percentage** this is of the **total sample size**; here, the answer would be $\frac{2}{24} \doteq 0.083 = \mathbf{8.3\%}$.

To use the **normal curve** to get an **approximate** answer to this same question, we can reason as follows.

All histograms express **frequency** information, but it turns out that there are **three different possible choices** of the **vertical** axis for histograms, and each choice expresses the idea of frequency in a **different** way.

The histogram back on page 13 was a **raw frequency** histogram — the **vertical axis** just plotted the **raw frequencies** (counts) of data values in each bar.

Another idea would be to **divide** the **raw frequencies** by the **sample size**, to produce **relative frequencies**, and plot them on the vertical scale instead — this would give you a **relative frequency histogram**, which would have the **same shape** as the raw frequency picture (technically speaking, in moving from raw to relative frequency the vertical axis has undergone a **linear change of scale**, and that has **no effect** on the shape of the basic distribution).

The Normal Curve (continued)

There's a **third way** to plot histograms: to draw the **vertical axis** on what's called the density scale, which is chosen so that

(a) **relative frequency** is expressed by computing the **area** under the histogram or curve, and

(b) the **total area** under the histogram or curve is therefore **100%** (or **1**).

When considered as an **approximation** to a **histogram**, the **normal curve** (as it turns out) is (by definition) drawn on the **density scale**, so it can be used to **approximate relative frequencies** (like the **percentage** of the data values **less than 3.56**) by working out the **area to the left of 3.56** under the **normal curve** with the **same mean and SD** as the data.

Q: How do you work out the **area under a normal curve**?

A: Formally speaking, the **calculus** technique of **integrating** the function in equation (6) from $-\infty$ to 3.56 will give you the right answer, but it turns out that the **Gaussian density function** in (6) **cannot be integrated in closed form** (it has no **anti-derivative**), so the **back-up technique** is called **numerical integration**: you use **numerical** (rather than **symbolic**) methods to make a **table** in which the area under the normal curve to the left of some number z is computed for **lots of different choices of** z.

Q: But you said earlier that there isn't **just one normal curve**, there are **infinitely** many of them, one for each choice of the **mean** and **SD**, so with this approach wouldn't you have to create **infinitely many tables**?

The Normal Curve (continued)

A: Ah, yes, **embarrassing** but **good question**; to **rescue** this idea we have to appeal to a **remarkable fact** about the **normal distribution**:

Fact: **Every** normal curve (no matter what it's mean and SD is) satisfies the (theoretical version of the) **Empirical rule** not just approximately but **exactly**; in other words, if you **start at the mean** and go **1 SD either way**, the area under **any** normal curve will be **68%**; **2 SDs** either way, **95%**; and **3 SDs** either way, **99.7%**.

This means that it's enough to make a **table** for **only one normal curve** — by convention, it's called the standard normal curve — and then **relate** whatever normal curve you're interested in **back** to the **standard curve**.

Q: What did people choose for the **mean** and **SD** of the **standard normal curve**?

A: Well, the **mean** could be anywhere from $-\infty$ to ∞, and the **simplest number** between these **extremes** is probably **0**, and the **SD** could be anywhere from 0 to ∞ (why can't an SD be negative?), and the **simplest number** in this range is probably **1**, so the **standard normal curve** (by definition) has **mean 0** and **SD 1**.

The **table** on the next two pages (Table **A–2** in T&T, also available on the **inside back cover** of the book) gives **areas to the left** of a bunch of places z under the standard normal curve; for example, to work out the area to the left of **−1.27** you look in the **row** marked **−1.2** and the **column** marked **0.07** (because ignoring the minus sign and putting the 1.2 and the 0.07 together you get 1.27), and the table says the **area** is **0.1020**, which could also be expressed as **10.20%** (in practice this would typically be **rounded** to **10.2%** or **10%**, because the normal curve is only being used as an **approximation** to the actual histogram).

The (Standard) Normal Table

POSITIVE z Scores

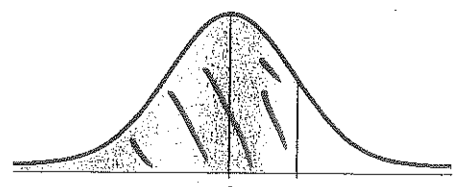


TABLE A-2 (continued) Cumulative Area from the LEFT

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998
3.50 and up	.9999									

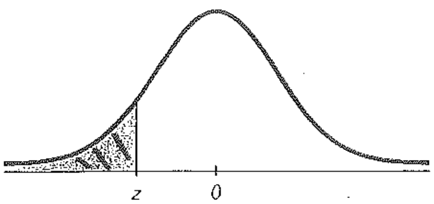
NOTE: For values of z above 3.49, use 0.9999 for the area.
*Use these common values that result from interpolation:

z score	Area
1.645	0.9500
2.575	0.9950

Common Critical Values

Confidence Level	Critical Value
0.90	1.645
0.95	1.96
0.99	2.575

The (Standard) Normal Table



NEGATIVE z Scores

TABLE A-2 Standard Normal (z) Distribution: Cumulative Area from the LEFT

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−3.50 and lower	.0001									
−3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
−3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
−3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
−3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
−3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
−2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
−2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
−2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
−2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
−2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051 *	.0049	.0048
−2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
−2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
−2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
−2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
−2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
−1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
−1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
−1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
−1.6	.0548	.0537	.0526	.0516	.0505 *	.0495	.0485	.0475	.0465	.0455
−1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
−1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
−1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
−1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
−1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
−1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
−0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
−0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
−0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
−0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
−0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
−0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
−0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
−0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
−0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
−0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641

NOTE: For values of z below −3.49, use 0.0001 for the area.
*Use these common values that result from interpolation:

z score	Area
−1.645	0.0500
−2.575	0.0050

The Normal Curve (continued)

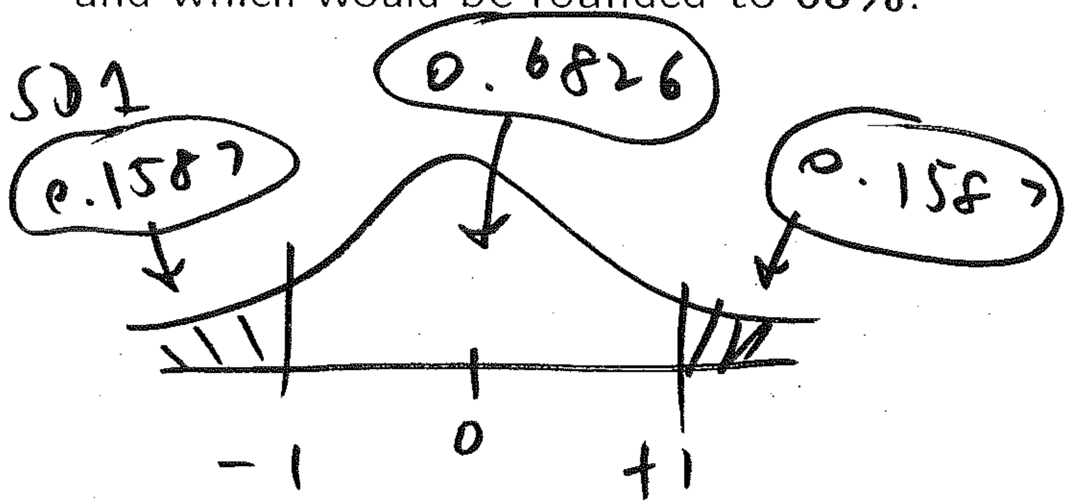
Q: The **table** only gives areas to the **left** of z under the **standard normal curve**; what if I want the **middle area** between $-z$ and $+z$, or the area to the **right** of z?

A: You can use **two basic facts** about the normal curve to work out **any other kind of area** you want:

- Since **all normal curves** are (by definition) drawn on the **density scale**, the **total area** under any normal curve is **100%** (or 1), and
- **All normal** curves are **symmetric** around their **means**.

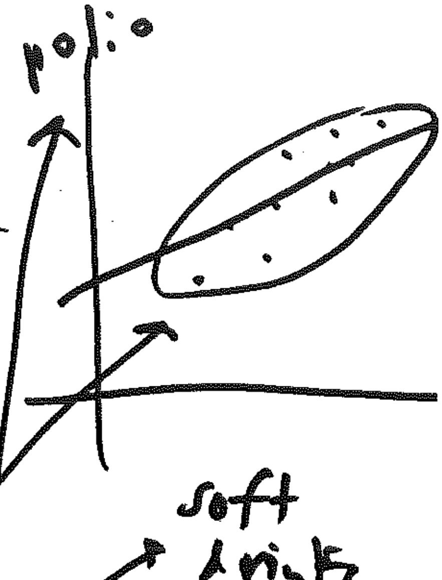
For example, to work out the area between -1 and $+1$ under the standard normal curve, this corresponds to going **1 SD either way from the mean (0)**, so we know from the **theoretical figure** in the **Empirical Rule** that the **answer** should be **68%**.

Reasoning from the **table**, you can look up the **area to the left of -1** and get **0.1587**; by symmetry the **area to the right of $+1$** must also be **0.1587**; and because the **total area is 1** the **area in the middle** must be $1 - 2(0.1587) = \mathbf{0.6826}$, which we might express as **68.26%** and which would be rounded to **68%**:



additional notes for chapter 1

season	# cases soft d.	# cases polio
W	Low	Low
Sp	Medium	Medium
Su	High	High
F	Medium	Medium



scatterplot

positive association between variables
soft drinks & polio

Some relationships are <u>causal</u>, some are just associations; how can we tell the difference?

X — Y : X is associated with Y

$u \rightarrow v$: u causes v

---

cwd or not?	
N	0
N	0
Y	1
N	0
⋮	⋮
Y	1

deer ↑ 128 ↓

1 row for each deer

sum of 1s & 0s = # of yeses ← 38

yes : $\frac{2}{128}$

---

mean of 1s & 0s = $\frac{\text{sum}}{\text{\# deer}} = \frac{2}{128}$ = proportion of yes population $\mathcal{P}$

ALL UCSC DEER ON 27 JUL 08

sample $\mathcal{S}$

THE OBSERVED DEER

$1 = Y$
$0 = N$

pop size $\downarrow$ $N = ?$

CWD ?

$$\begin{bmatrix} 0 \\ 0 \\ 1 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix}$$

$\xrightarrow{\text{~~representative~~}}$

CWD ?

$$\begin{bmatrix} 1_s \\ k \\ 0_s \end{bmatrix}$$

sample size $\downarrow$ $n = ?$

MEAN $\theta = p = ?$ ← parameter

mean $\hat{\theta} = \bar{y}$

↕

a good estimate of θ

1 column for each variable

1 row for each subject

might be between 100 & 1000

representative

principle: make sample, & unsample similar in all relevant ways

pop

Disc. sec 1 (DS) (19) ① better diagnosis → more cases of cancer

Ans 7

② population growth

old = x

new = y

2 ways to compare: ① absolute comparison:

(new - old) = 462,000 - 331,000

= +131,000 ← increase of 131,000 cancer deaths from 1970 to 1986

② relative comparison: $\frac{\text{new} - \text{old}}{\text{old}}$

3 significant figures (sigfigs)

$$= \frac{+131{,}000}{331{,}000} = 0.3957704 \doteq 0.396$$

3 sigfigs

= 39.6% ≐ 40%

increase in reported deaths from cancer; US population grew from 1970 to 1986 but probably not that fast

(40)

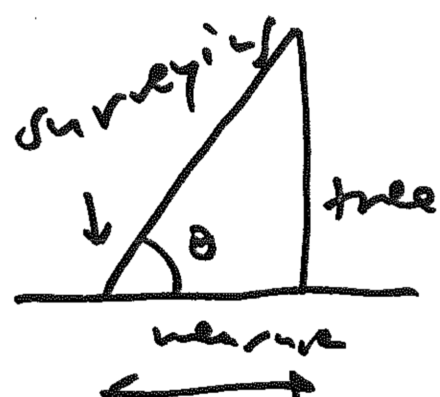
better measure: 5-year survival rate

1(b) * circumference 1 m from ground * height (length of shadow) * volume: if tree were cylinder, $V = \pi \underline{\underline{r}}^2 \underline{\underline{h}}$ (r = radius, h = height); can get radius from circumference

cheapest: circumference

most expensive: volume

height? trig: angle & base of Δ



if cone: $V = \frac{1}{3}\pi \underline{\underline{r}}^2 \underline{\underline{h}}$

1(c) ⊛ climate (temperature, humidity ⊛ rent/mortgage : cost of living (transportation, food, ...) ⊛ population density ⊛ crime rate ⊛ quality of schools

---

Q: how do you trade these off against each other? A: need to use judgment to rank variables in importance to you : decision

theory | (1d) weight : $\frac{\text{new} - \text{old}}{\text{old}}$ ✓

height : silly $\frac{\text{weight}}{\text{height}}$ also good

I

1e

$$\frac{39 \text{ beats}}{\cancel{30}\ \cancel{\text{sec}}} \cdot \frac{\cancel{60}^{\,2}\ \cancel{\text{sec}}}{1 \text{ min}} = 78 \frac{\text{beats}}{\text{min}}$$

$$\frac{80 \text{ beats}}{65\ \cancel{\text{sec}}} \cdot \frac{60\ \cancel{\text{sec}}}{1 \text{ min}} \doteq 74 \frac{\text{beats}}{\text{min}}$$

---

with a reliable measurement method, more data = more accuracy

---

so 2nd method is more accurate (65 sec of data vs. 30 sec)

~~sample~~

2 $\begin{bmatrix} 1 \\ 2 \\ 9 \end{bmatrix}$ $n = 3$ $\longrightarrow$ $\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix}$ ← subscript

mean 4 mean $\bar{y} = \dfrac{y_1 + y_2 + y_3}{3}$

L- 43

in general $\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \updownarrow n$

mean $\bar{y} = \dfrac{y_1 + y_2 + \dots + y_n}{n}$

$= \frac{1}{n}(y_1 + y_2 + \dots + y_n)$

need shorthand for sums:

capital sigma (summation sign)

$$y_1 + y_2 + \dots + y_n = \sum_{i=1}^{n} y_i$$

index of summation

sample mean $\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$

L-44

2 a(i) $\sum_{i=1}^{3} 1 = 1 + 1 + 1 = 3$

(ii) $$\sum_{i=1}^{n} 1 = \underbrace{1 + 1 + \dots + 1}_{n} = n$$

(iii) $$\sum_{i=1}^{5} i = 1 + 2 + \dots + 5 = 15$$

---

sample,
unsample
similar in all
relevant ways

$\longleftrightarrow$

sampling L-45
method
representative

in 1920s Fisher (UK) & Neyman (Poland, UK, US) figured out a good way to encourage representativeness of the pop by the sample: choose sample at random; 2 ways to do this:

---

at random with replacement $\leftrightarrow$ **i**ndependent **i**dentically **d**istributed (IID) sampling

$$\begin{array}{c} \text{pop} \\ \begin{pmatrix} 1 \\ 2 \\ 9 \end{pmatrix} \end{array} \xrightarrow[\text{with repl.}]{\text{at random}} \begin{array}{c} \text{possible sample} \\ \begin{bmatrix} 9 \\ 9 \end{bmatrix} \end{array} \quad n = 2$$

put draw back in pop before drawing again

---

at random without replacement $\leftrightarrow$ **s**imple **r**andom **s**ampling (SRS)

↳ 46

don't put draws back in

pop

$$\begin{bmatrix} 1 \\ 2 \\ 9 \end{bmatrix}$$

at random without repl

sample

$$\begin{bmatrix} 9 \\ 1 \end{bmatrix} n = 2$$

---

math easier with IID but learn about pop faster with SRS ;

if n (sample size) is a lot smaller than N (pop size) then SRS $\doteq$ IID

is approx same as

pop

$N = 10{,}000$

$$\begin{bmatrix} \\ \\ \\ \end{bmatrix}$$

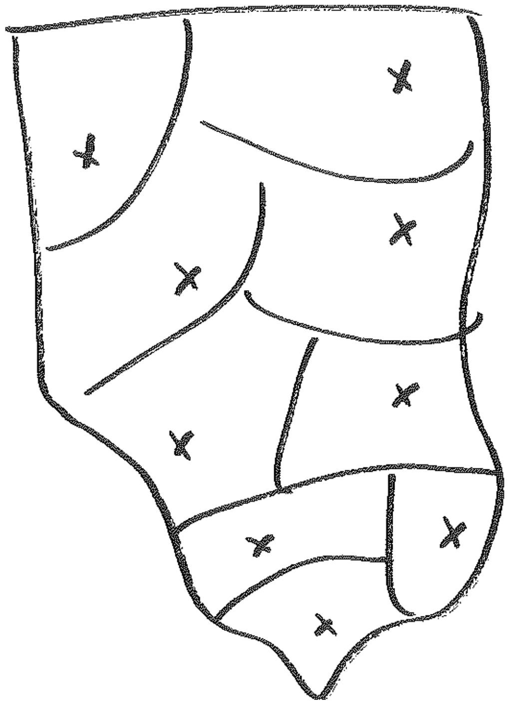
sample

$$\begin{bmatrix} \\ \end{bmatrix} n = 2$$

---

in real world people use SRS but usually $n \ll N$ so they use math from IID because it's simpler

47



partition of campus (n partition sets)

variable	values it could take on
	categorical, dichotomous
eye color	blue, brown
	nominal
hair color	black, brown, red, white
	qualitative, ordinal = ordered categorical
success in maze running	very slow, slow, moderate, fast, very fast
	quantitative, continuous
plant size: height (cm)	7.42…, 8.20…,
	discrete
plant size: # of leaves	46, 29, 5

(sample) wing length (cm)

$$\begin{bmatrix} 4.4 \\ 3.6 \\ \vdots \\ 3.9 \end{bmatrix} \quad n = 24 \quad \xrightarrow{\text{sort}} \quad \begin{bmatrix} 3.3 \\ 3.5 \\ 3.6 \\ 3.6 \\ \vdots \\ 4.5 \end{bmatrix}$$

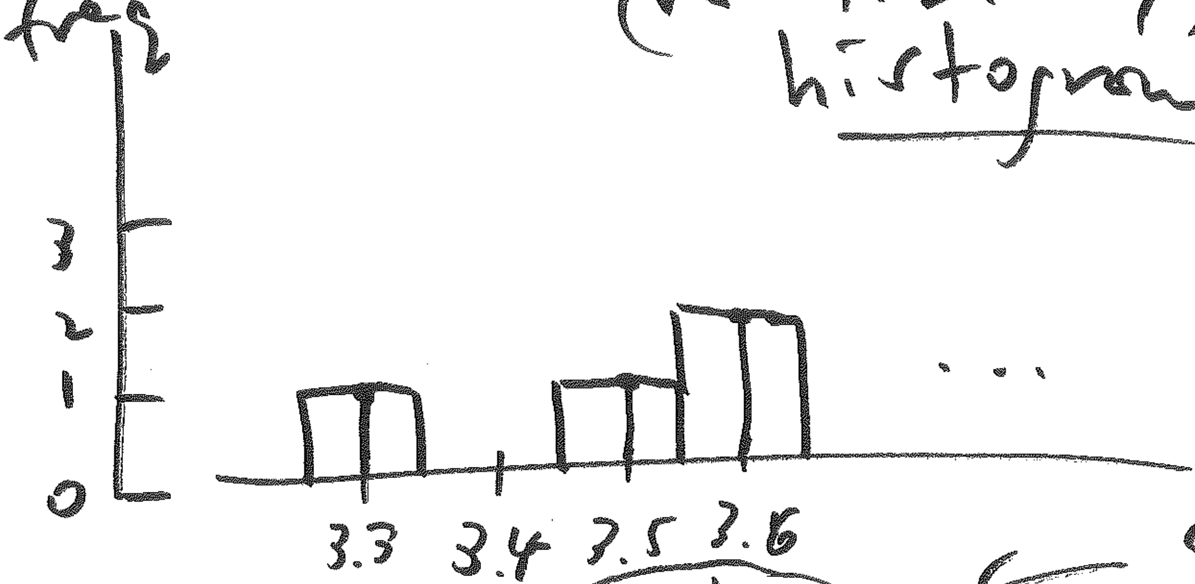
value	(count) frequency
3.3	1
3.4	0
3.5	1
3.6	2
⋮	⋮
4.5	1

sum $n = 24$

<u>raw</u> <u>frequency</u> <u>distribution</u> of butterfly wing lengths

(or just <u>distribution</u>)

(raw frequency)
histogram

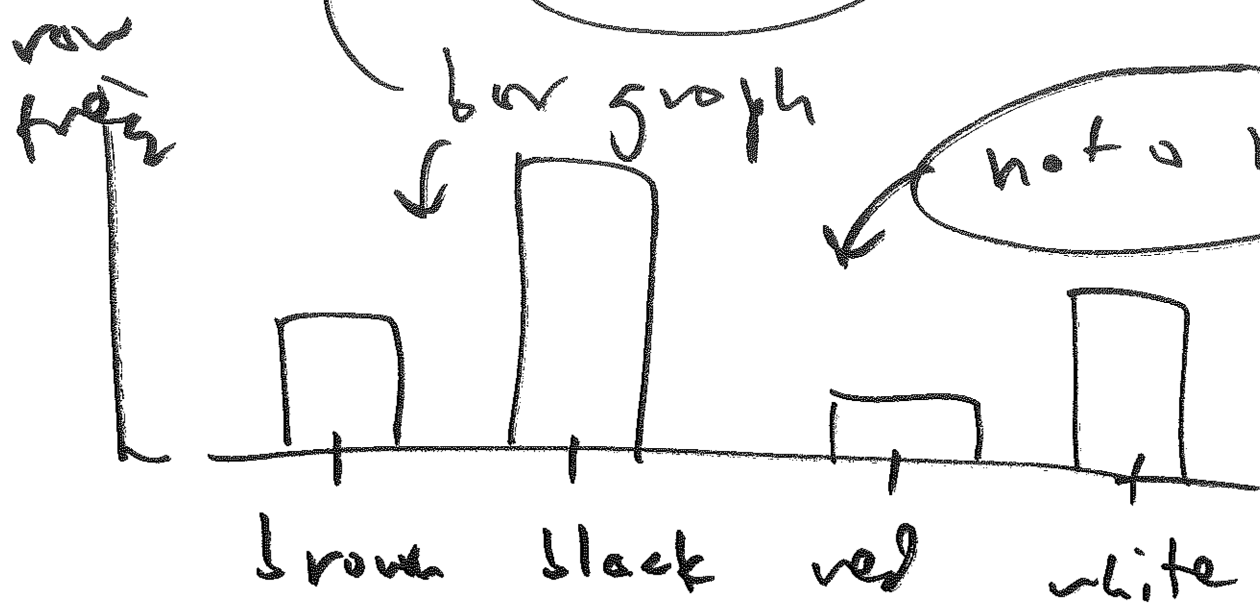


quant
variable

values

bar graph

not a hist



not
quant

histogram shapes

(a)

symmetric unimodal

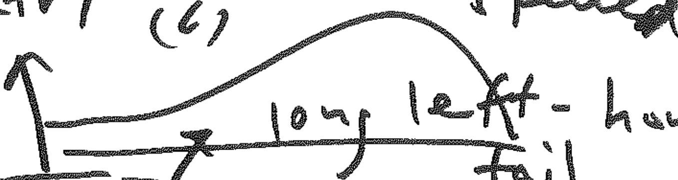
point of symmetry

(b)

skewed

long right-hand tail

(c)



skewed

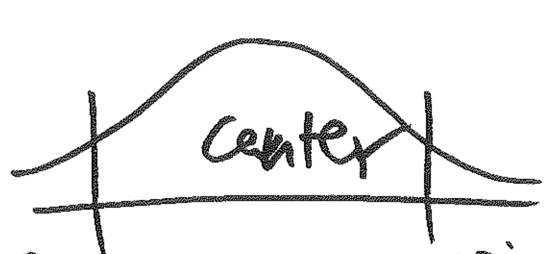
long left-hand tail

unimodal

(d)

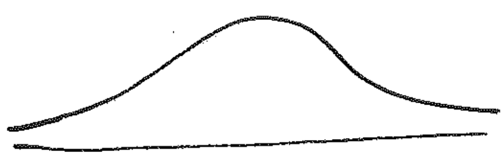
skewed

multimodal, in this case bimodal

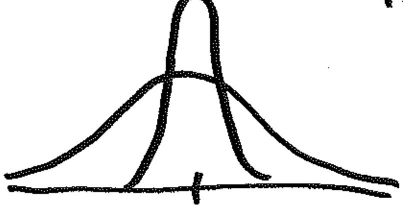


left tail

right tail

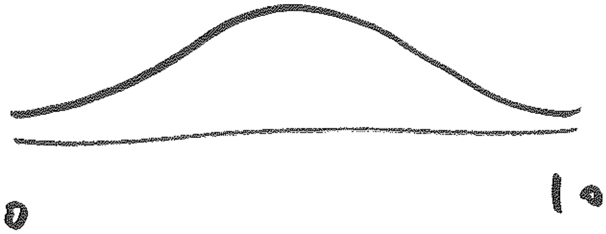


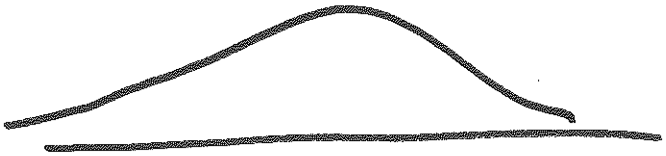
same shape, same spread, different center



same shape, same center, different spread

L-51





same center,
same spread,
different shape

center

wing length

$$\begin{bmatrix} 4.4 \\ \vdots \\ 3.9 \end{bmatrix} \quad n = 24$$

y

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{24} \end{bmatrix} \leftarrow y_n \qquad \updownarrow n$$

subscript

mean $\dfrac{4.4 + \dots + 3.9}{24} \doteq 3.96$ cm

y

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \quad \updownarrow n$$

mean $\bar{y} = \dfrac{y_1 + \dots + y_n}{n} = \dfrac{1}{n}(y_1 + \dots + y_n)$

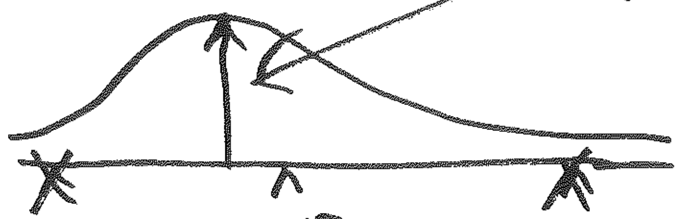
"y bar"

$$= \frac{1}{n} \sum_{i=1}^{n} y_i$$

index of summation

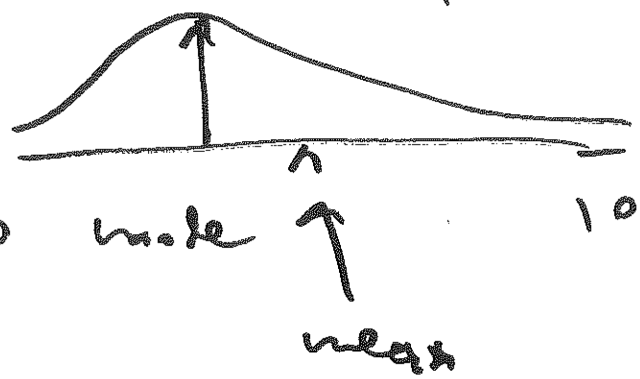
L-52

mode = place of highest frequency

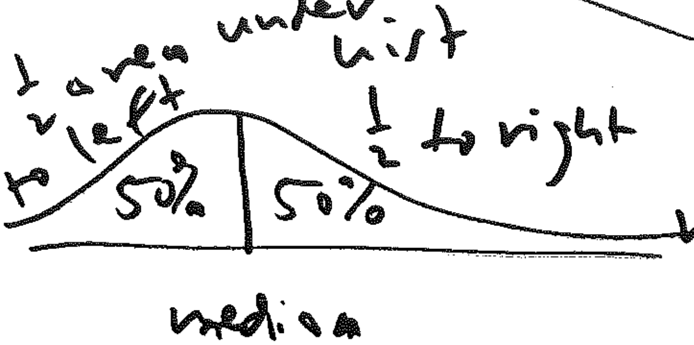


mean = balance point

median = middle of data sorted from smallest to largest



= place where $\frac{1}{2}$ data below, $\frac{1}{2}$ above



odd ↓ $n=3$ $\begin{bmatrix}9\\1\\2\end{bmatrix} \xrightarrow{\text{sort}} \begin{bmatrix}1\\2\\9\end{bmatrix}$

median 2

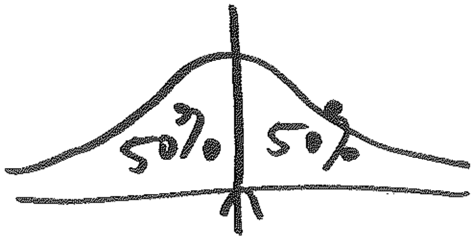
mean 4

even ↓ $n=4$ $\begin{bmatrix}9\\1\\2\\6\end{bmatrix} \xrightarrow{\text{sort}} \begin{bmatrix}1\\2\\6\\9\end{bmatrix}$

median $\frac{2+6}{2} = 4$

L: 53

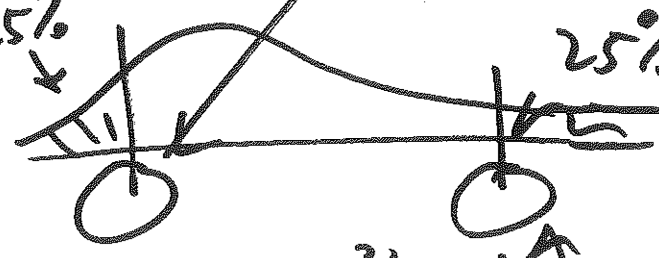
butterfly:
mean 3.96
median 4
mode 4



symmetric unimodal

pt of symmetry = mean = median = mode

25th percentile (quantile)



75th percentile

with this language median = 50th percentile disc sec

(DS)1 #2(a)(iv)

Ans ?

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} n$$

mean $\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$

$$\left(\sum_{i=1}^{n} y_i\right) - \left(\sum_{j=1}^{n} y_j\right)$$

$$= (y_1 + y_2 + \ldots + y_n)$$

$$- (y_1 + y_2 + \ldots + y_n) = 0$$

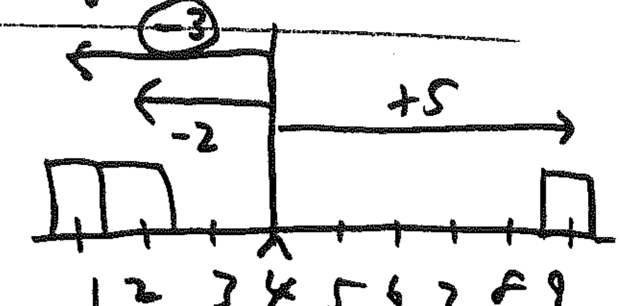
(v) $\sum_{i=1}^{n} (y_i + c) = \underbrace{(y_1 + c) + (y_2 + c) + \ldots + (y_n + c)}_{n}$

$= (y_1 + \ldots + y_n) + (c + \ldots + c) = \left(\sum_{i=1}^{n} y_i\right) + nc$

(b) $\sum_{i=1}^{n} c y_i = (c y_1) + (c y_2) + \ldots + (c y_n)$

$= c(y_1 + \ldots + y_n) = c \sum_{i=1}^{n} y_i$ ✓

(c) $\begin{bmatrix} 1 \\ 2 \\ 9 \end{bmatrix} (n = 3) = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$



mean $\bar{y} = 4$

L- 55

deviations from mean $\begin{bmatrix} 1-4 \\ 2-4 \\ 9-4 \end{bmatrix} = \begin{bmatrix} y_1 - \bar{y} \\ y_2 - \bar{y} \\ \vdots \\ y_n - \bar{y} \end{bmatrix} = \begin{bmatrix} -3 \\ -2 \\ +5 \end{bmatrix}$

mean 0

$$\sum_{i=1}^{n}(y_i-\bar{y}) = (y_1-\bar{y})+(y_2-\bar{y})+\dots+(y_n-\bar{y})$$

$$= (y_1+y_2+\dots+y_n) - \overbrace{(\bar{y}+\bar{y}+\dots+\bar{y})}^{n}$$

$$= \sum_{i=1}^{n} y_i - n\bar{y} = \sum_{i=1}^{n} y_i - \cancel{n}\left(\frac{1}{\cancel{n}}\sum_{i=1}^{n} y_i\right)$$

$$= \left(\sum_{i=1}^{n} y_i\right) - \left(\sum_{i=1}^{n} y_i\right) = 0$$

3.(a) 6 or more foxes?

$$\begin{array}{cc} \begin{bmatrix} N \\ Y \\ \vdots \\ N \end{bmatrix} & \begin{array}{c} (2) \\ (9) \\ \\ 1 \end{array} \end{array}$$

1 row for each litter

golden rule: visualize the data

nominal: $\overline{N \quad Y}$ or $\overline{Y \quad N}$

qual

ordinal: Y means more foxes

dichotomous ✓

1 = Y, 0 = N

$$\begin{bmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{bmatrix}$$

quant

discrete

dichotomous

ratio: 0 = absence of 6 or more foxes

sample

(b) phosphate concentration

1 row for each stream location

$n = 60$

quant

continuous

ratio

(c) temp (°C) at which chirprate ~~falls below 100/min~~

$\begin{bmatrix} 24.8 \\ 27.2 \\ 23.5 \\ \vdots \end{bmatrix}$

1 row for each night of observation

$n = 44$

75°F

$°F = \frac{9}{5}°C + 32°$

$20°C = 68°F$

$24°C =$

$25°C = 77°F$

quant

cont

interval

(d) kind of animal

turtle
snake
turtle
⋮
⋮
mammal

1 row for each vertebrate animal

$n = ?$

qual

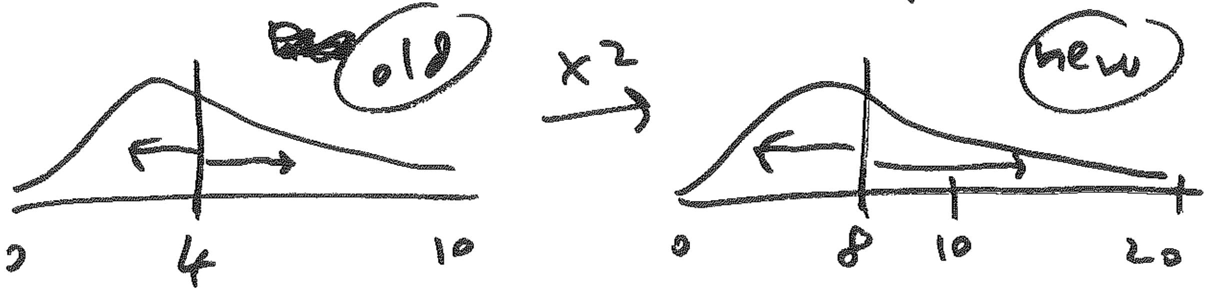
nominal

not dich

$\leftarrow$ 57

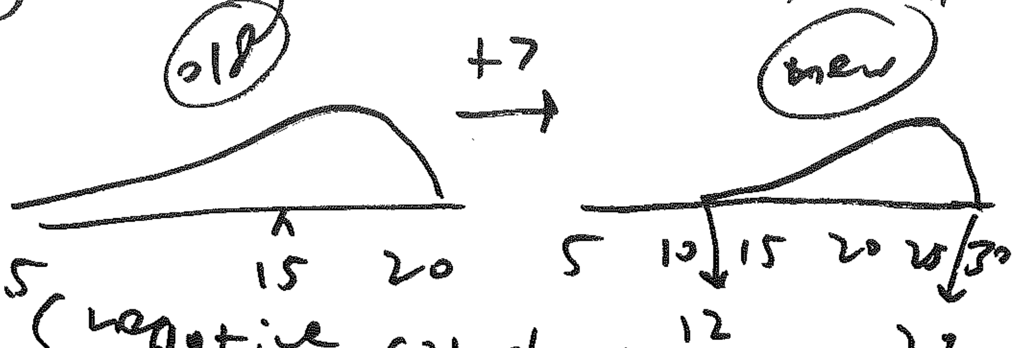
DS 2 | #2 $°F = \frac{9}{5}°C + 32°$

multiplying a dataset by a positive constant has no effect on basic shape of hist:

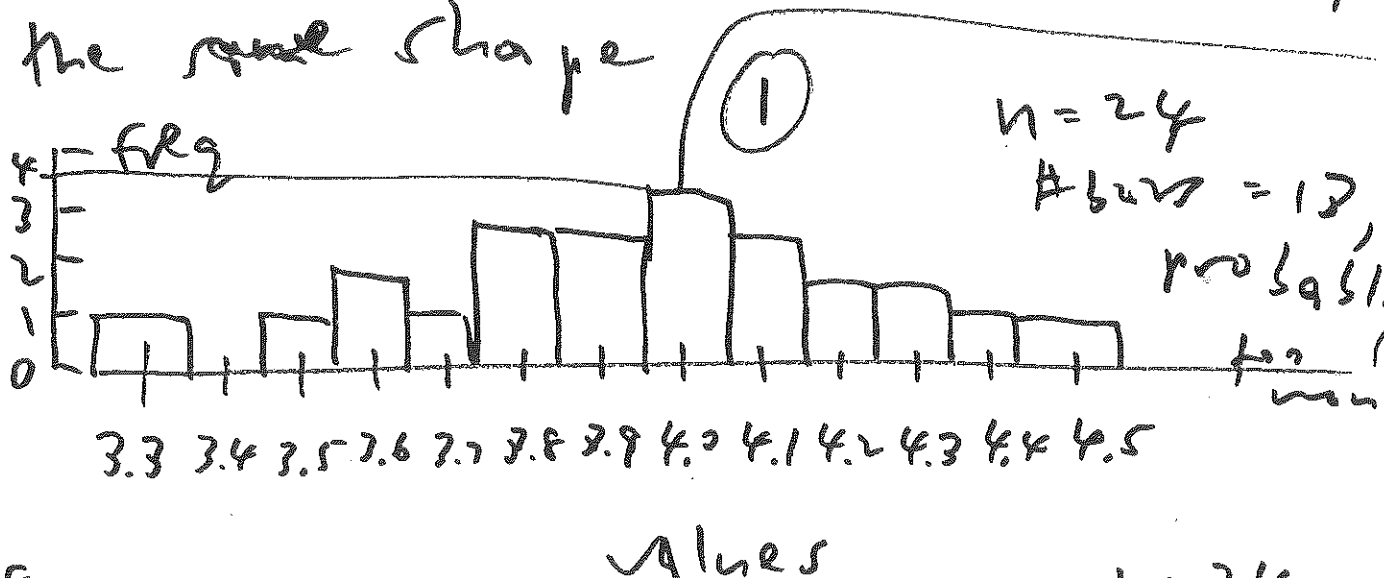


if mult by 2 mean gets multiplied by 2 & spread also goes up

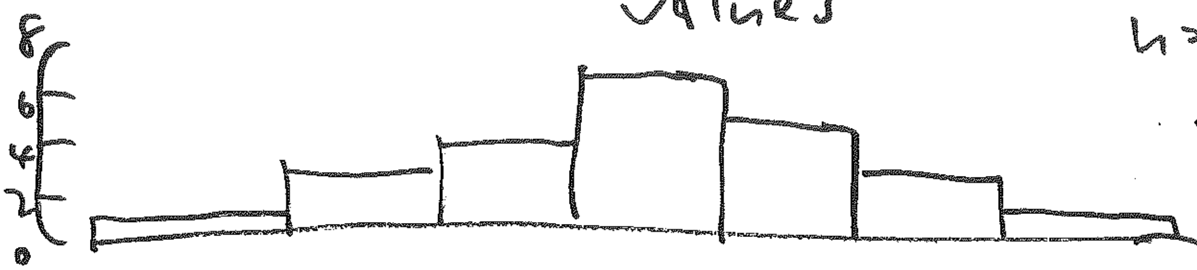
adding a constant to all data values just shifts hist to left (negative constant) or right (positive constant) by the amount of the constant, which again has no effect on shape of hist; the mean goes up (const ⊕) or down (const ⊖) by the amount of the constant, but spread is unchanged.



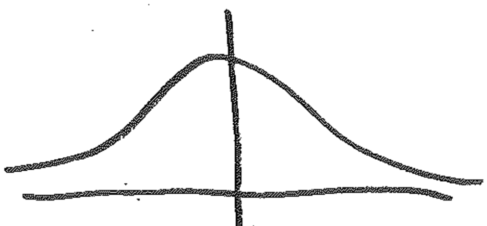
So if I measure temp in °F & you use °C our hist will have exactly the ~~same~~ shape



n = 24
# bars = 13,
probably too many

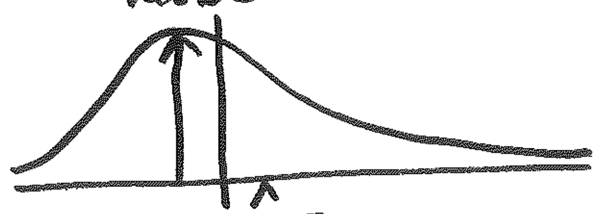


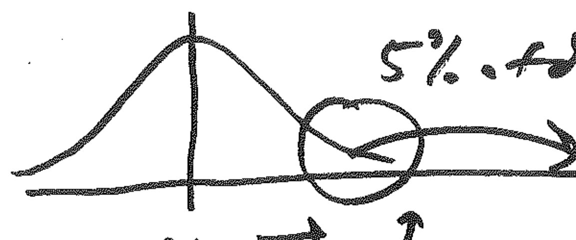
n = 24,
# bars = 7
better



pt of symmetry

mean
median
mode





mean →
median
stays same

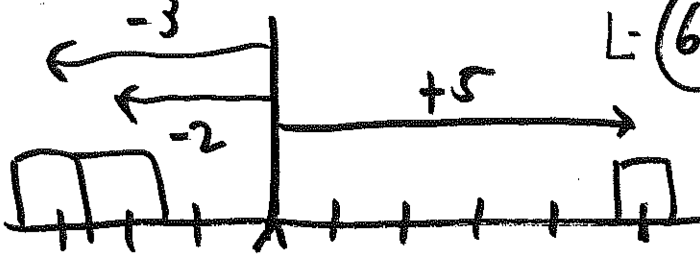
mean most sensitive measure of center to outliers; median more robust to unusual data values

L-60

measures of spread

$$\begin{bmatrix}1\\2\\9\end{bmatrix}$$

mean 4



spread: typical length of arrows

$$\begin{bmatrix} 1 \\ 2 \\ 9 \end{bmatrix} \longrightarrow \begin{bmatrix} 1-4 \\ 2-4 \\ 9-4 \end{bmatrix} = \begin{bmatrix} -3 \\ -2 \\ +5 \end{bmatrix}$$

mean 4 (deviations from mean) mean 0

want to get rid of +, − signs:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \longrightarrow \begin{bmatrix} y_1-\bar{y} \\ y_2-\bar{y} \\ \vdots \\ y_n-\bar{y} \end{bmatrix}$$

mean $\bar{y}$

(a) absolute values:

$$\begin{bmatrix} -3 \\ -2 \\ +5 \end{bmatrix} \longrightarrow \begin{bmatrix} |-3| \\ |-2| \\ |+5| \end{bmatrix}$$

$$= \begin{bmatrix} 3 \\ 2 \\ 5 \end{bmatrix}$$

mean 3.3

in general $\frac{1}{n}\sum_{i=1}^{n} |y_i - \bar{y}|$ (AAD)

average absolute deviation (mean) (MAD)

(b) squares:

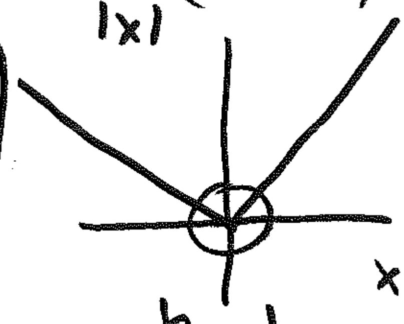
$$\begin{bmatrix} (-3)^2 \\ (-2)^2 \\ (+5)^2 \end{bmatrix} = \begin{bmatrix} 9 \\ 4 \\ 25 \end{bmatrix}$$

mean 12.7

$$\begin{bmatrix} 81 \\ 82 \\ 89 \end{bmatrix}$$

mean 84

L = 6



pathology at 0

new idea $\begin{bmatrix} 9 \\ 4 \\ 5 \end{bmatrix}$ take $\sqrt{\ }$ of this:

mean 12.7

sample standard deviation (SD) $= \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2} = s$ (units right)

mysterious idea

$\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 = s^2 =$ sample variance (units wrong, math nice)

graphical interpretation of SD: **Empirical Rule**: start at mean of almost any data set, go 1 (2) [3] SD either way: you will usually capture about 2/3 (most) [almost all] of the data, no matter what hist. looks like

official story:

# SDs either way	informal	formal
1	about 2/3	68%
2	most	95%
3	almost all	99.7%

why divide by (n−1), not n, in computing sample SD s?

y_1 free (1)
y_2 free (2)
$\vdots$
y_n not free (9)

n = 3

mean 4

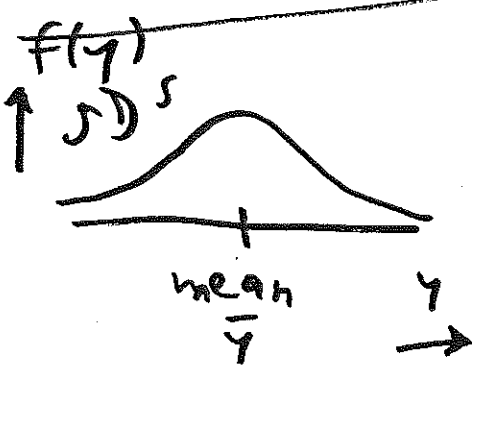
y_1 free
y_2 free
$\vdots$
y_{n-1} free
y_n not free

mean $\bar{y}$

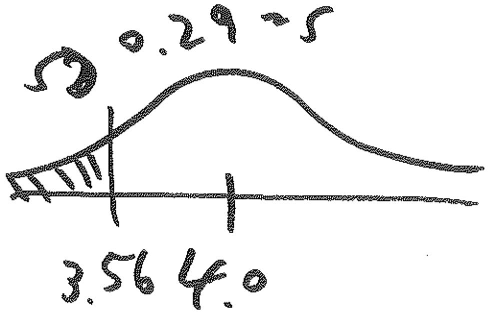
3 ↑ n ↓

a dataset with n obs. only has (n−1) **degrees of freedom** for measuring spread

normal (Gaussian) curve



$$f(y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(y-\mu)^2}$$



q: what % of data was below 3.56 cm?

$\underline{A_1}$: 2 of 24 values are < 3.56, so this % is $\frac{2}{24}$

exact

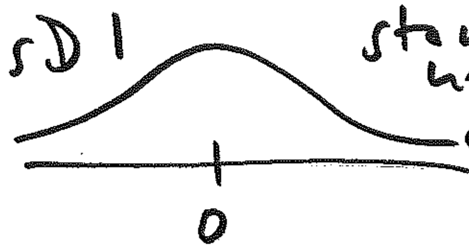
$\doteq 8.3\%$

A_2: exact

what % of data in list is < 3.56?
also get 8.3%

$\underline{A_3}$: (approximate)

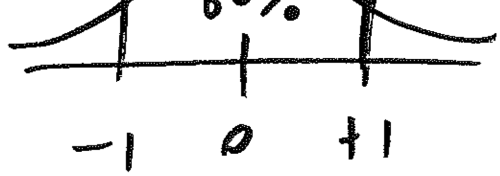
use normal curve with same mean & SD as data

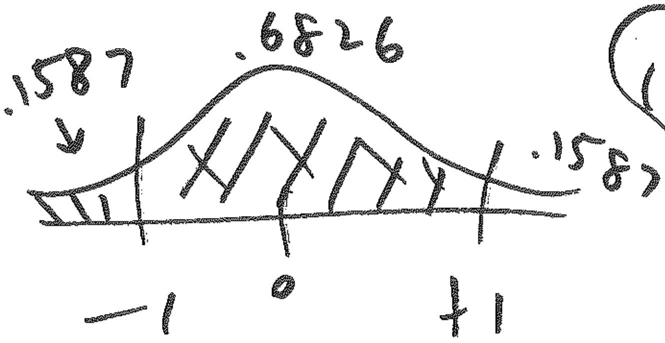


$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2} \geq 0$$

$$\begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \quad \begin{bmatrix} c \\ c \\ \vdots \\ c \end{bmatrix}$$

SD 0 0



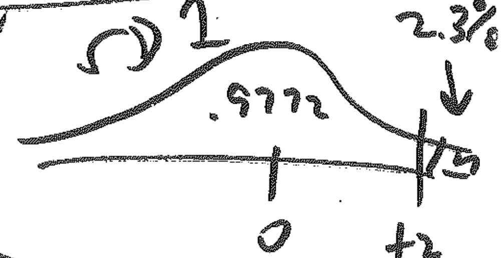


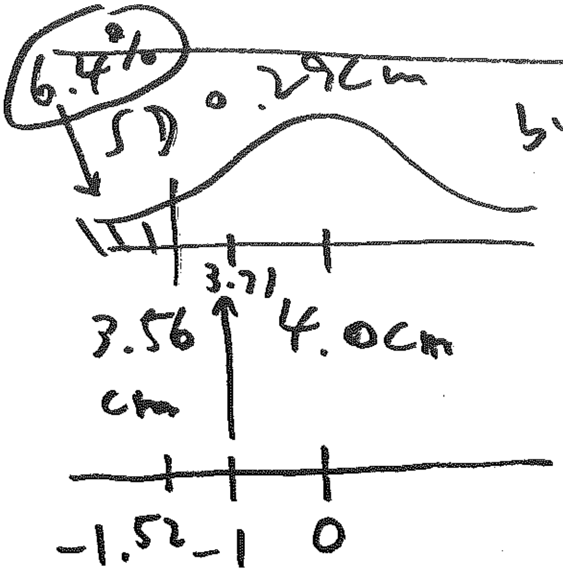
(a) normal curve symmetric

(b) total area under curve is 1

$\longleftarrow 1 \longrightarrow$

area between -1 & +1 is about 68%





butterfly wing length

raw units axis (y)

standard units (SU) axis (z)

$$SU = \frac{\# - \text{mean}}{SD} = \frac{3.56\text{ cm} - 4.0\text{ cm}}{0.29\text{ cm}} = z \text{ score}$$

$= -1.52$ L-65

converting to standard units

$\boxed{z = \frac{y - \bar{y}}{s}}$

$\boxed{y = \bar{y} + z \cdot s}$ converting to raw units

Disc
Sec
2

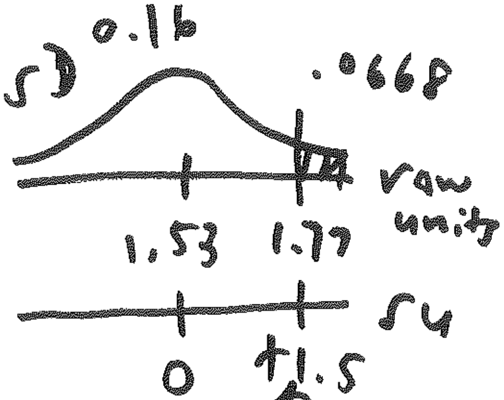
entzyme activity

1.7
1.4
~~2.9~~ ← impossible

n = 159

mean $\bar{y}$ = 1.53
SD s = 0.16

#3
(a)

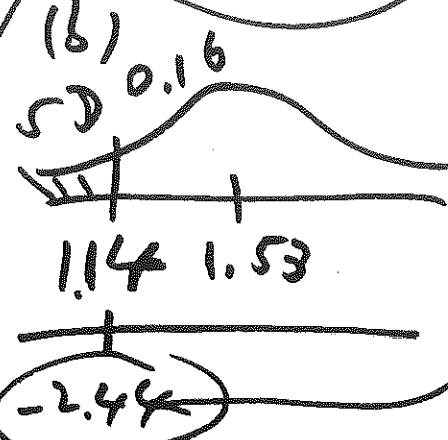


$\frac{1.77 - 1.53}{0.16} = +1.5$

$\frac{2.9 - 1.53}{0.16} = +8.6$ in su

6.7% or 7%

(b)

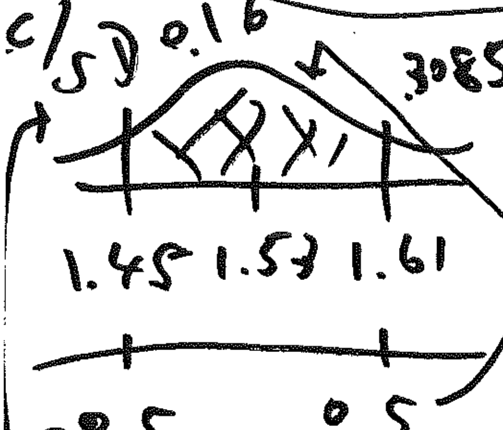


$\frac{1.14 - 1.53}{0.16} = -2.44$

.0073

expected
# of small frogs = (0.0073)(159)
= 1.16 ≐ 1

c)



.3085

$\frac{1.61 - 1.53}{0.16} = +0.5$

.383 = 1 − 2(.3085)
≐ 38%

L-66