AMS 7

April 3, 2017

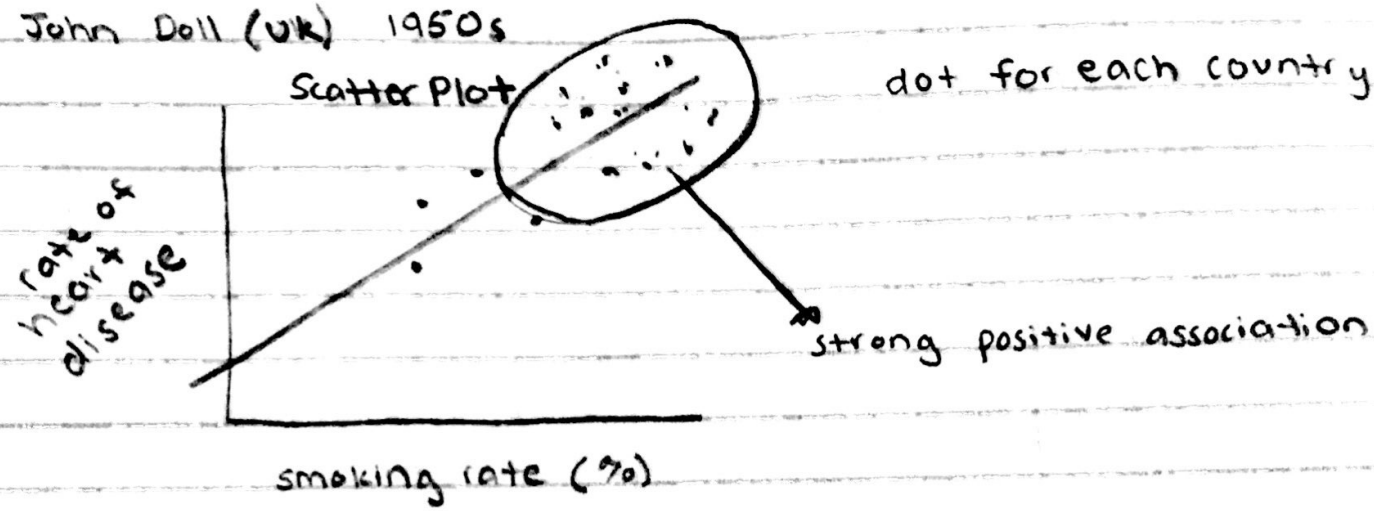
This time - intro

Next time - descriptive methods

Webcasts webcast.ucsc.edu

↳ lectures will be found here

Dr. John Doll (UK) 1950s

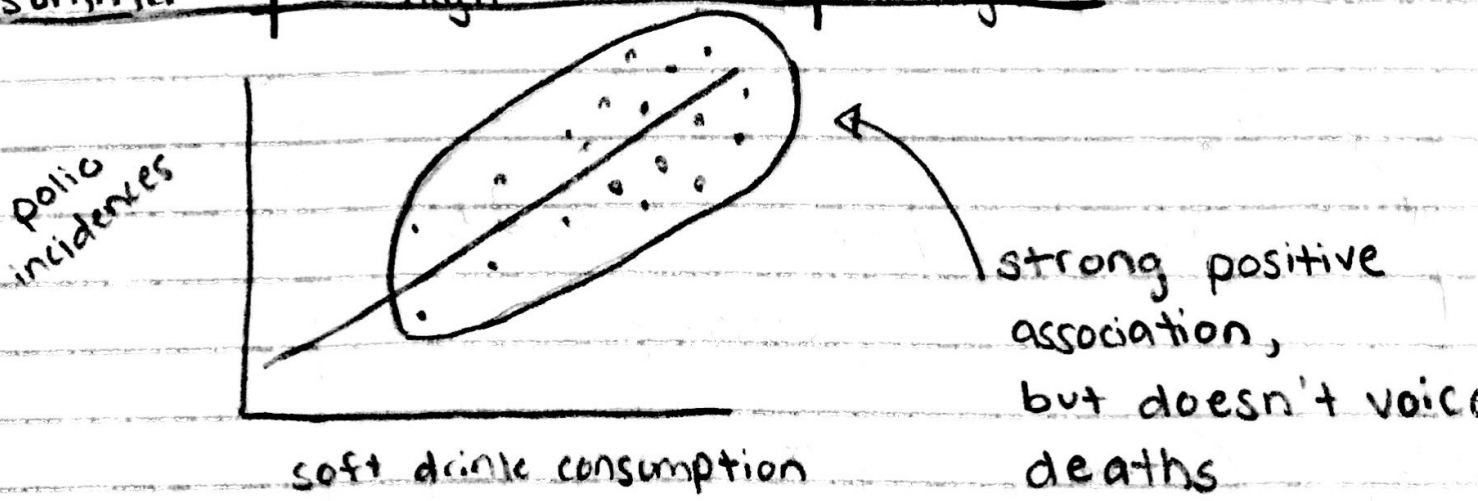


April 5, 2017

This Time - descriptive methods

Next Time - descriptive Methods

Season	soft drink consumption	polio
Fall	medium	medium
Winter	low	low
Spring	medium	medium
Summer	high	high

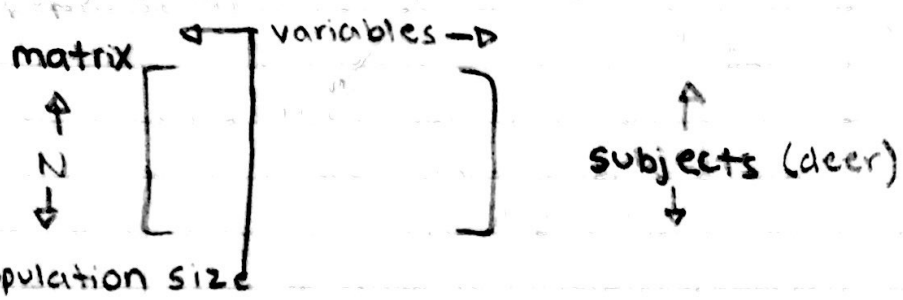


Lecture 1:

Stats - the study of uncertainty: how to measure it, and what to do about it

uncertainty - state of incomplete or imperfect info about something of interest to you

percentage θ (theta = p) P = population



N: population size

$N = 800$

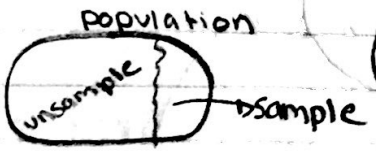
disease ?
no/N - 0
no/N - 0
no/N - 0
yes/Y - 1
no/N - 0
⋮
no/N - 0

- population of all deer living @ UCSC on 3 March 2017
- 1 row for each deer

mean $p = \theta = ?$ → unknown; → portion in pop. w/ disease

means of 1s to 0s keep track automatically of **percent** of 1s

mean $\bar{y} = \hat{p} = \hat{\theta}$ (y bar, p hat, theta hat)



at random → IID

goal of sampling: representitiveness: want sample and unsample to be similar in all relevant ways

$n = 85$

disease = 3

$\frac{3}{85}$ = 3.5% has disease

Sample (the observed deer)

disease ?

1s + 0s

→ sample size $n = 85$

(IID)

- at random with replacement (Independent identically distributed) (SRS)
- at random w/out replacement (simple random sampling)

SRS - is more informative than IID, but IID has easier **math**

if n is a lot smaller than N, SRS and IID are approximately equal to eachother

SRS $\doteq$ IID

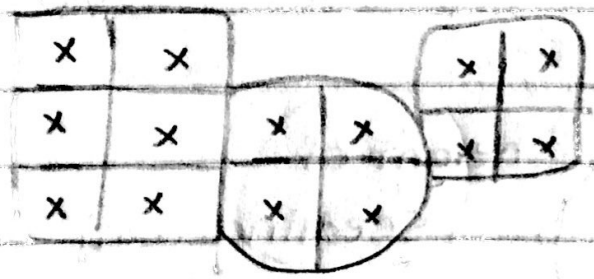
April 7, 2017

* **Sections** start Monday, April 10th

webcast
username: ams-7-01 → zero
password: ready14slow

Lecture 2

- variables (things that can be measured on pop. subjects)
  - variable can be either [yes, no]
- parameter (numerical summary of a population



haphazard = **random**

variables	possible values	
- eye color in humans	→ brown, blue, black, green	qualitative: no unique place for its value on a number line
- hair color in hamsters	→ brown, black, dichotomous	
- success at maze running	→ very slow, slow, moderate, fast, very fast (ordered categorical or ordinal)	qualitative: order to category
- seedling height (cm)	→ 7.42 cm, 7.017 cm	quantitative: values have unique places on the number line
- # leaves (adult)	→ 46, 111	

plot size

same distance

very slow < slow < moderate < fast < very fast

ratio scale variable

has a true zero

quantitative: continuous — height of seedling (cm)

6.98742 7.017 7.42

no gaps between possible values

quantitative: discrete — # of leaves

46 47 48 111

possible gaps

constant size interval + ratio variable

qualitative

Three variables measured:

Animal ID	Age	Hair Color
45	2	0
333	0	2
.	.	.
.	.	.
.	.	.

can't take mean of qualitative variables

0 = brown
2 = red

1 column for each variable

1 row for each anima

mean

no point, not meaningful

meaningful

not meaningful

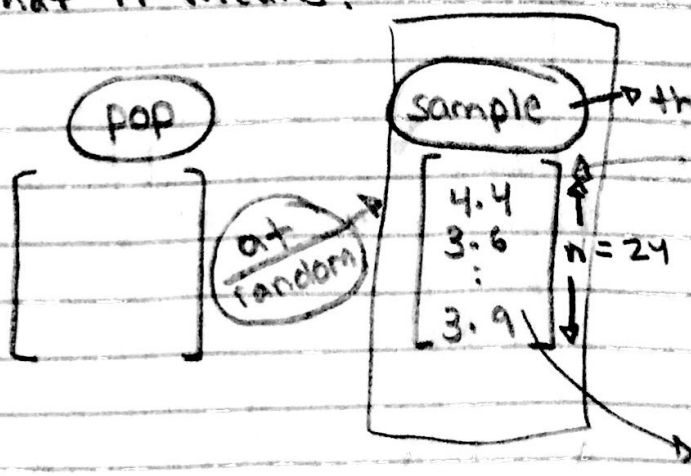
April 10, 2017

This time: histograms, bar graphs
Next time: measures of center spread

*discussion start this week, go! (quiz)

I. intro (week 1)
II. descriptive methods (graphical numerical)

you have a data set, how to gain insight to what it means?



the observed butterflies description here

*1 row for each butterfly

wavelength (cm)

raw data

4.4
3.6
4.1
⋮
3.9

original order is irrelevant

sort →

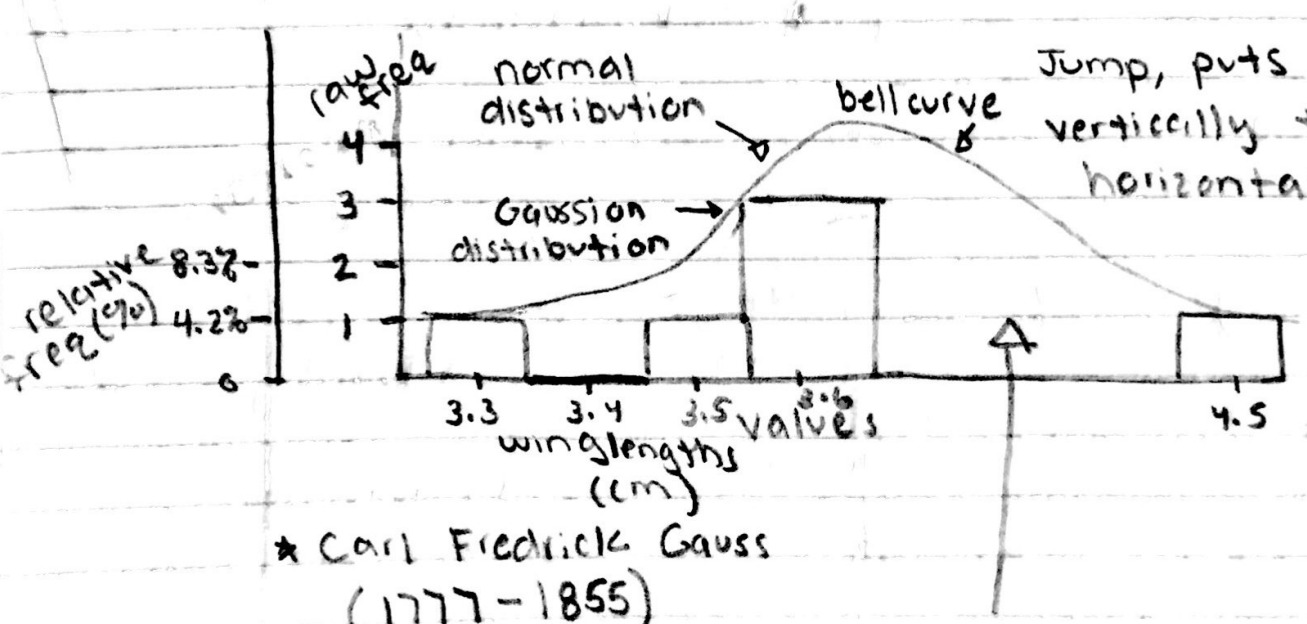
3.3 ← minimum
3.5
3.6
3.6
⋮
4.5 ← maximum

2 significant figures (sigfigs)

values	count = raw frequency
3.3	1
3.4	0
3.5	1
3.6	3
⋮	⋮
4.5	1
	n=24

raw frequency distribution

variables are often named with values such as x, y, and z ex. y = wavelengths



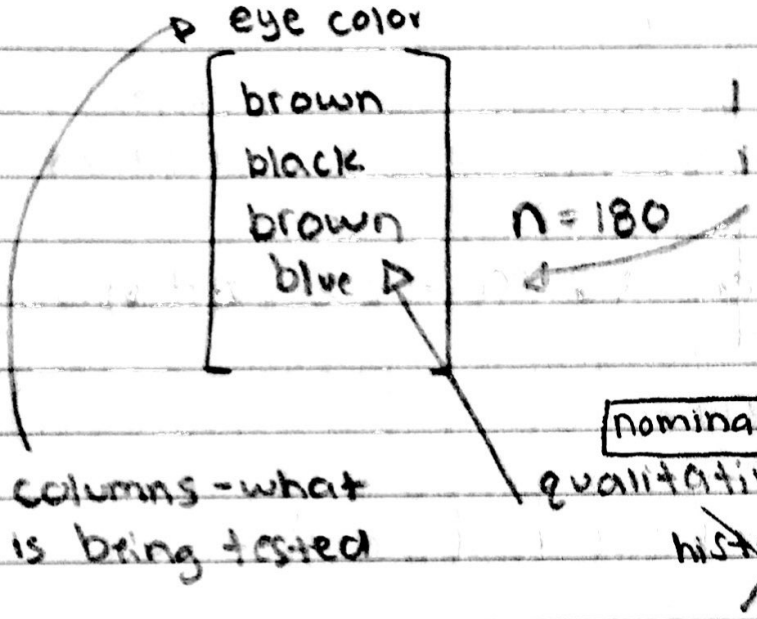
★ Carl Fredrick Gauss (1777–1855)

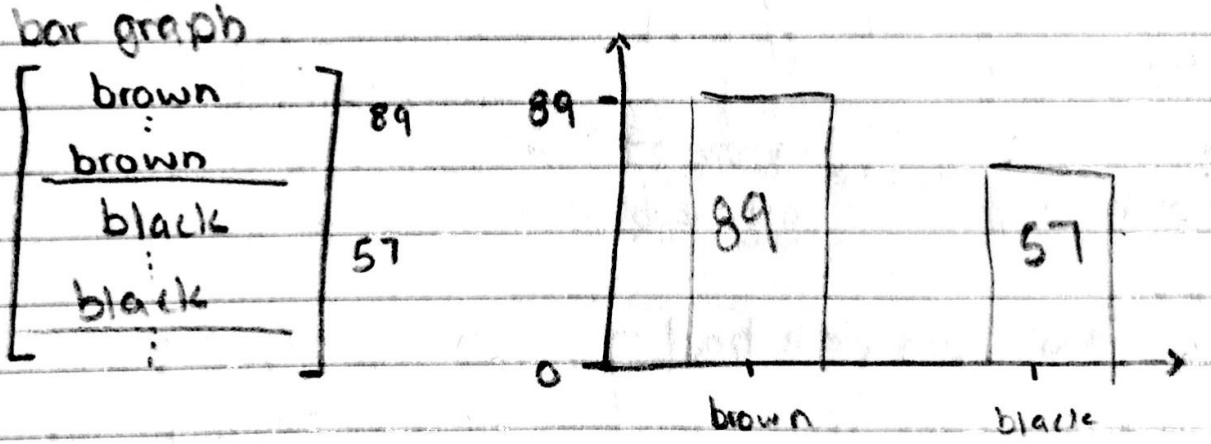
raw frequency histogram (a kind of bar graph)

wing length ← quantitative
*conceptually continuous, but made discrete by rounding in measurement process

histograms: can only be made for quantitative variables (either discrete or continuous)

* visualize the raw data set



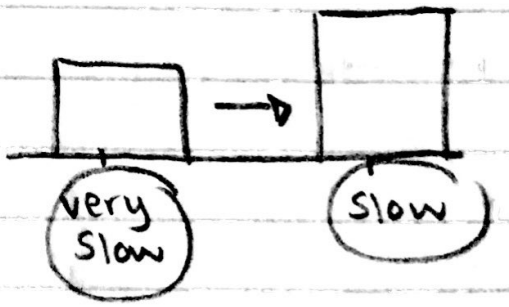


bar graphs = qualitative
histograms = quantitative

value	raw freq.	relative freq (%)
3.3	1	1/24 × 100% = 4.2%
3.4	0	0%
3.5	1	1/24 × 100% = 4.2%
3.6	2	2/24 × 100% = 8.3%

* 3 diff. vertical scales for histograms
① raw frequency vs ② relative freq vs ③ density

maze running

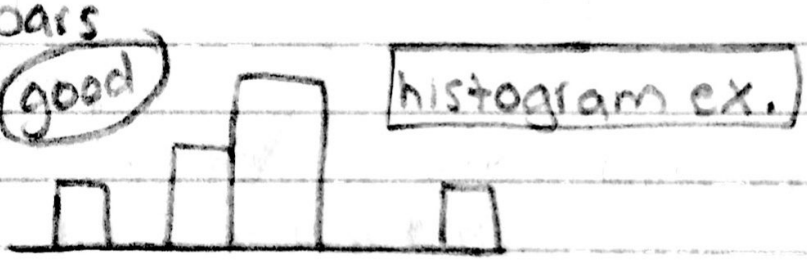


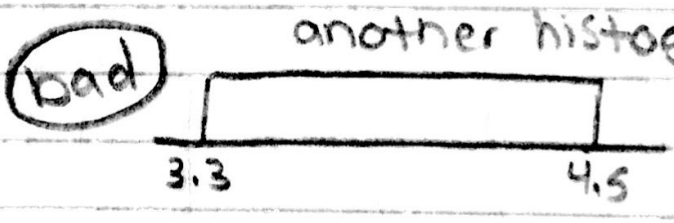
qualitative, ordinal

Question: Is there just one unique histogram fo
any given data set?
Answer: no, you have to make a good choice
of # of bars

butterfly:
n=24

bar
13
bar





all sense of
distribution shape
is lost

too few bars

too few bars → bad → all sense of distribution shape is lost

just right → good

too many bars → can't see the forest for the tree (too much detail)

* to cut # of bars in half, combine adjacent bars

location (sparrows)

vine
building eave
vine
⋮

n = 211

1 row for each nest (nesting site)

qualitative, nominal, non dich.

bar graph = yes
histogram = no

pigment type (sunfish)

0
2
5
0
⋮

n =

1 row for each sunfish

qualitative = even when we code 0, ... 4 → ordinal

bar graph: (yes)
histogram = (no)

April 12, 2017

This time: numerical measures of center spreads
Next time: the **normal curve**

ex. litter size in foxes

of pups
4
7
3
4

← 1 row for each litter

- quantitative, discrete, ratio
- pups between possible values represent impossibility

histogram? (yes)
mean? (yes)

ex. aphids on clover

of aphids
14
36
⋮

n=424

1 row for each clover plant

histogram? (yes)
mean? (yes)

- quantitative, discrete, ratio

ex.

phosphorous concentration (y)
8.42
9.08
⋮

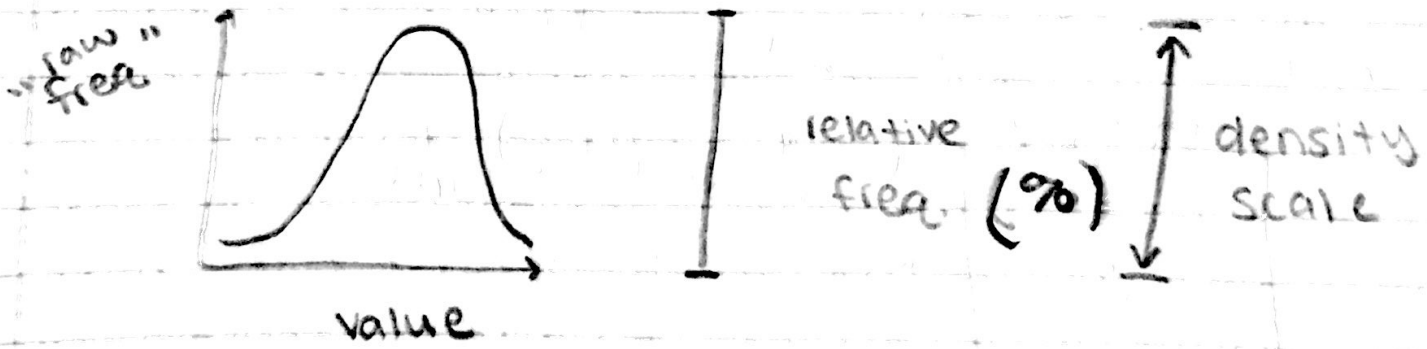
n=130

1 row for each leaf

- quantitative, conceptually continuous, ratio

histogram? (yes)
mean? (yes)

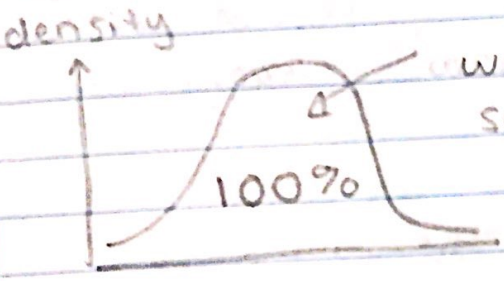
$$\frac{y_{max} - y_{min}}{\#\text{ bars}} = \text{bar width}$$



* unless otherwise noted, all histograms in this class will be drawn on the density scale

density scale

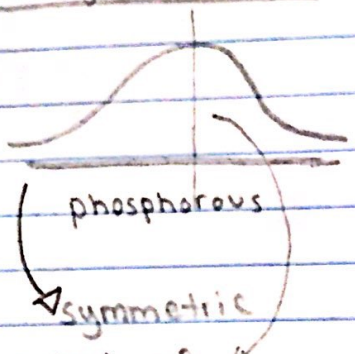
relative freq. = area under histogram (curve)



with histogram on density scale, total area under curve = 1 or = 100%

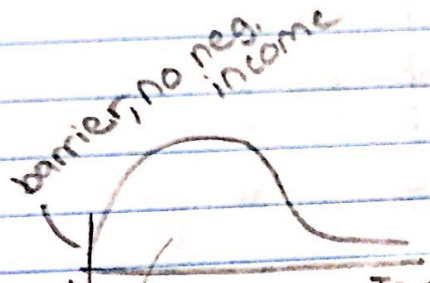
histogram shapes

unimodal

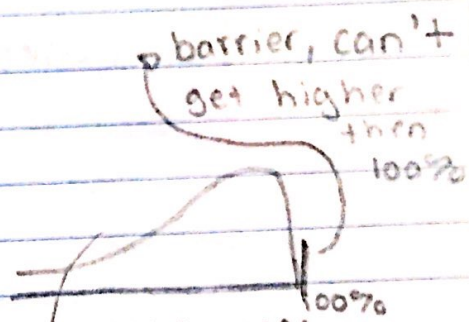


symmetric

point of symmetry

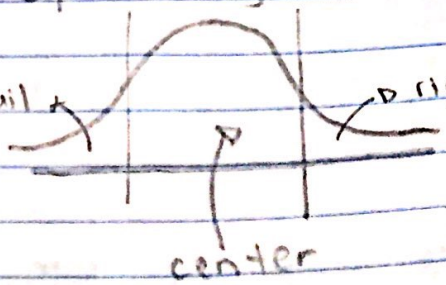


long right hand tail



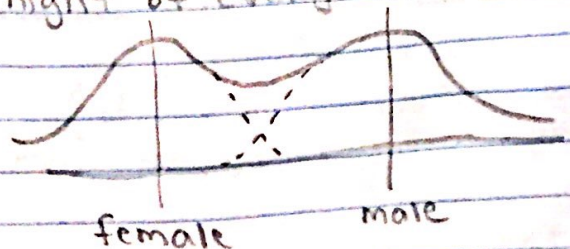
asymmetric = skewed

* long left-hand tail



ex. hight of everyone in class today

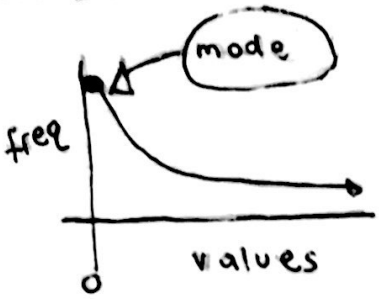
bimodal or multi-modal



mode = a) a point of highest frequency

quantitative (numerical) measures the center of a distribution

① mode



mode is not necessarily central, but always typical in terms of frequency

② mean

butterflies — wing length (cm)

$$\begin{bmatrix} 4.4 \\ 3.6 \\ \vdots \\ 3.9 \end{bmatrix} \quad n = 24$$

mean = ?

generic data set

variable y

y_1
y_2
$\vdots$
y_n

subscript

n = sample size = 24

y_{24}

mean $\bar{y}$ "y-bar"

$$\frac{y_1 + y_2 \cdots + y_n}{n}$$

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

---

summation notation

capital sigma

largest index

smallest index

index of summation (i, j, k)

(expand out theorem)

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

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

↑ sample mean of variable y

* in this data set

$$\bar{y} = \frac{95.0}{24}$$

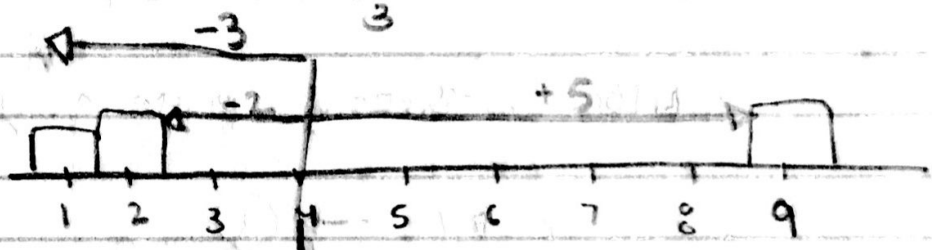
$$= 3.96\,cm$$

graphical interpretation of the mean

fake data set

$$\begin{matrix} y_1 \\ y_2 \\ y_3 \end{matrix}\begin{bmatrix} 1 \\ 2 \\ 9 \end{bmatrix} \quad n = 3$$

mean $= \frac{12}{3} = 4$



$$\begin{bmatrix} 1 \\ 2 \\ 9 \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ y_n \end{bmatrix} \xrightarrow[\bar{y} = y]{\text{subtract}} \begin{bmatrix} 1-4=-3 \\ 2-4=-2 \\ 9-4=+5 \end{bmatrix} \quad n=3$$

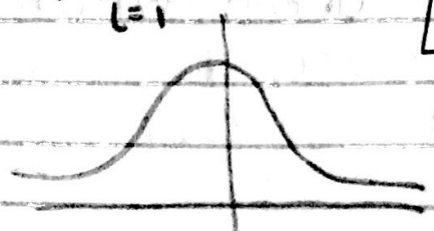
mean

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

theorem:
let n be any positive integer $y_1, \ldots y_n$ be any real numbers, then

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

mean = balance point



= mode = mean =

April 14, 2017

This time: standard derivitive, the normal curve
Next time: controlled experiments + observational studies

Read: DD
ch 1-3 (A), 1-6 (B), LN pp. 1-84?

No more ecommons → HW is to be submitted to canvas.ucsc.edu

Official Note taker = Kaitlyn Abercrombie

Bring reader and lecture notes to class

-the mean is the balance point

ex. L15 (butterfly winglength)

quantitative → y

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

mean = $\bar{y}$
median = $\tilde{y}$

median ex.

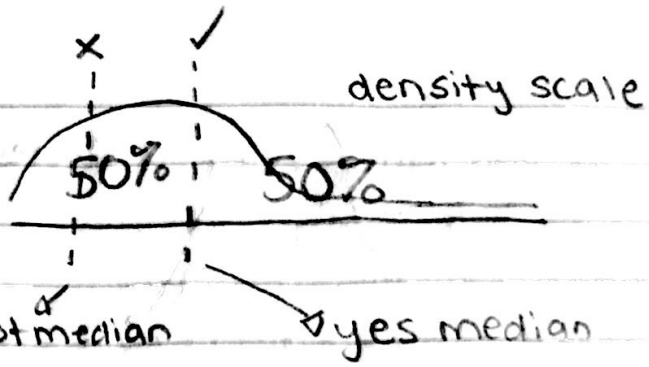
n(odd)

$\begin{bmatrix} 9 \\ 2 \\ 1 \end{bmatrix}$ sort → $\begin{bmatrix} 1 \\ (2) \\ 9 \end{bmatrix}$ → median = middle #, after data set has been sorted

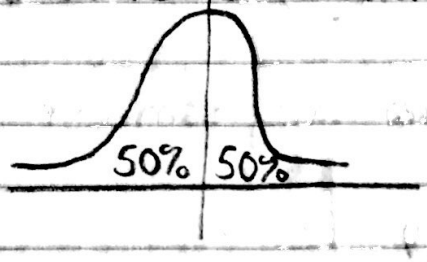
$\begin{bmatrix} 9 \\ 1 \\ 4 \\ 2 \end{bmatrix}$ → $\begin{bmatrix} 1 \\ (2 \\ 4) \\ 9 \end{bmatrix}$ n=4 → median = 3 (mean of middle 2 numbers)

n(even)

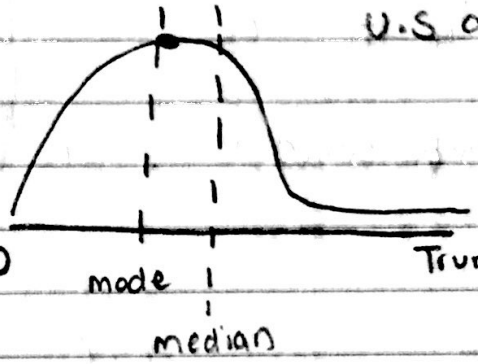
median - 50%/50% point in data

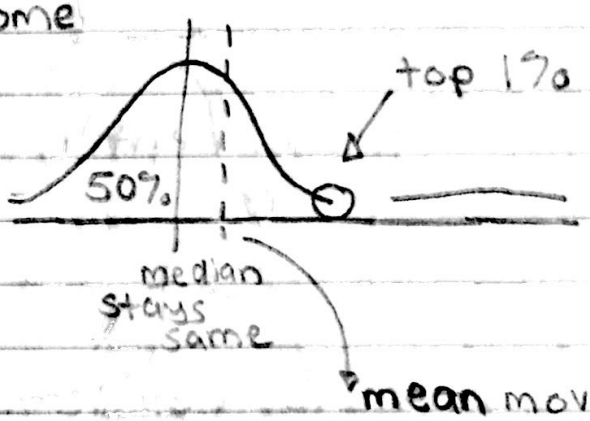


Symmetric | unimodal

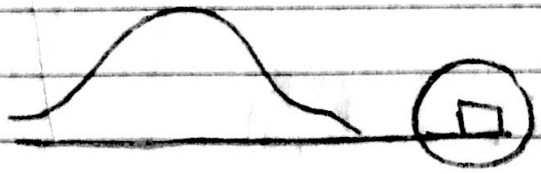


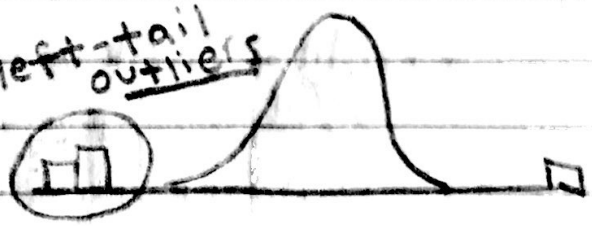
mode = point of symmetry
= mean
= median





right-tail outlier

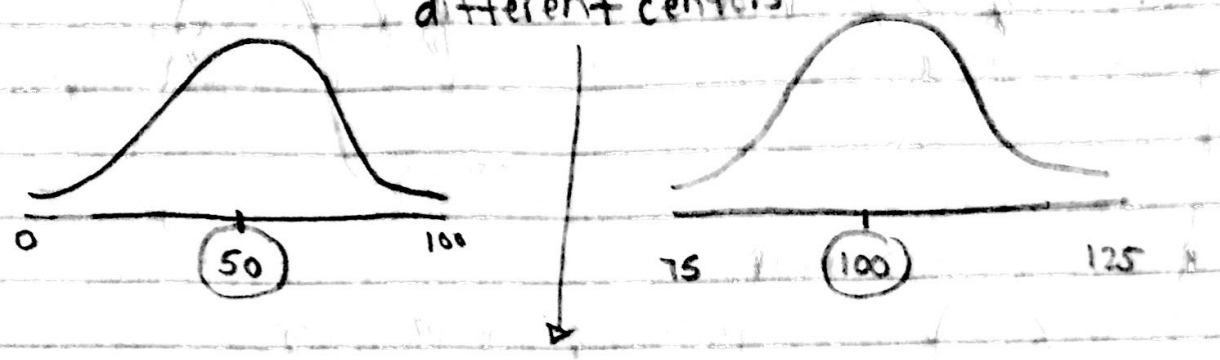




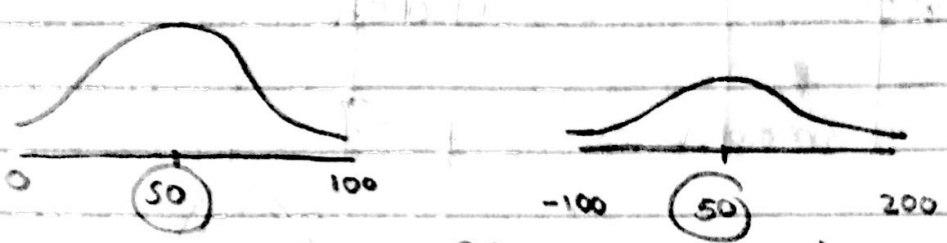
median is insensitive to outliers, but mean can be highly sensitive to outliers

+ Look up page L-23-24 for more information
Ex. L-25 (Life expectancy of birds in captivity)

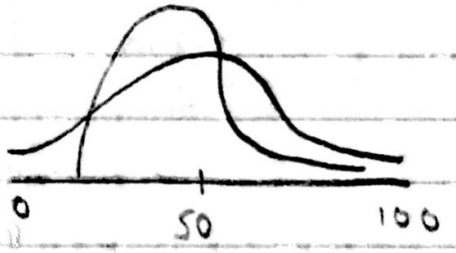
different centers



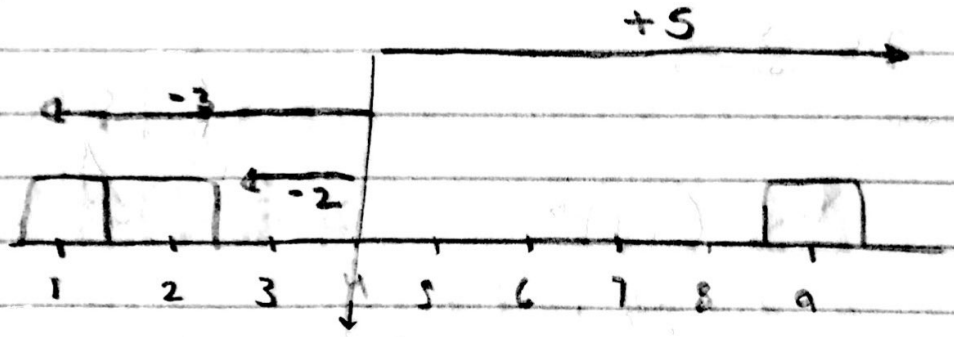
same spread, and basically same shape



same center, different spread, same basic shape



- same center
- same spread
- different shape



$$\begin{bmatrix} y \\ 1 \\ 2 \\ 9 \end{bmatrix} \quad n=3$$

mean $\bar{y} = 4$

$$\begin{bmatrix} 1 \\ 2 \\ 9 \end{bmatrix} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$$

subtract $\bar{y} = 4$ from all data values

$$\begin{bmatrix} -3 \\ -2 \\ +5 \end{bmatrix} = \begin{bmatrix} y_1 - \bar{y} \\ \vdots \\ y_n - \bar{y} \end{bmatrix}$$

mean: 0 0

to avoid ⊕ + ⊖ cancellation,

1) absolute values

derivitives from the means

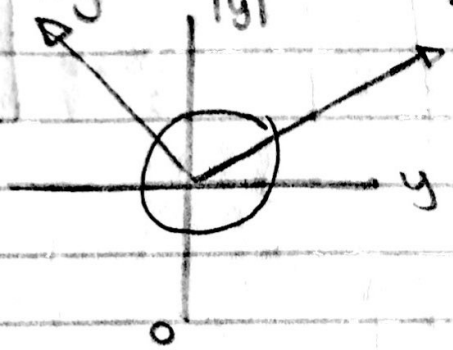
$$\begin{bmatrix} |-3| \\ |-2| \\ |5| \end{bmatrix} = \begin{bmatrix} 3 \\ 2 \\ 5 \end{bmatrix} \quad \begin{bmatrix} |y_1 - \bar{y}| \\ \vdots \\ |y_n - \bar{y}| \end{bmatrix}$$

mad = mean absolute dev

mean → 3.3 mean

MAD (sir arthur Eddington ~ 1910)

↳ not used much today



$|y|$ is not differentiable at 0

$$\begin{bmatrix} \$1 \\ \$2 \\ \$9 \end{bmatrix} \xrightarrow[\$4]{\text{subtract}} \begin{bmatrix} \$-3 \\ \$-2 \\ \$+5 \end{bmatrix} \xrightarrow{\text{square}} \begin{bmatrix} (-3)^2 = 9\$ \\ (-2)^2 = 4\$ \\ (5)^2 = 25\$ \end{bmatrix}$$

mean $\$4$ — mean = $12.7\$^2$

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \xrightarrow[\bar{y}]{\text{subtract}} \begin{bmatrix} y_1 - \bar{y} \\ \vdots \\ y_n - \bar{y} \end{bmatrix} \xrightarrow{\text{square}} \begin{bmatrix} (y_1 - \bar{y})^2 \\ \vdots \\ (y_n - \bar{y})^2 \end{bmatrix}$$

mean $\bar{y}$

$(-3)^2 + (-2)^2 + (+5)^2$ = (sample) variance (s^2)

3 − 1

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

(sample) Standard deviation (SD) $= \sqrt{\text{(sample) variance}} = s$

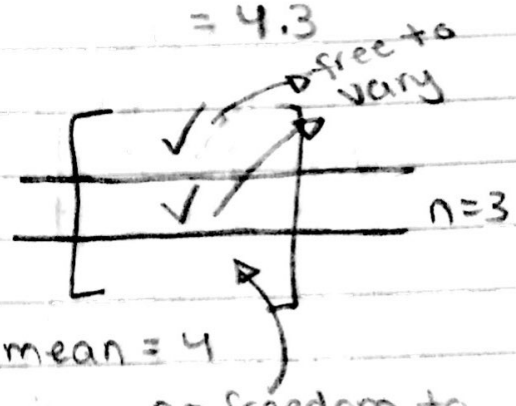
$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

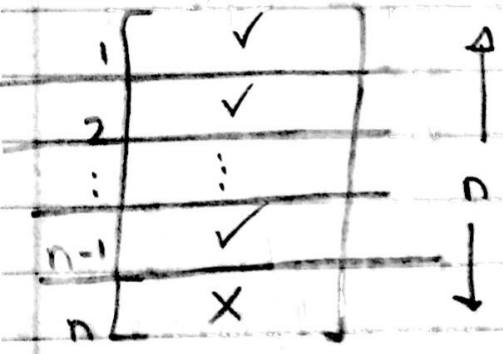
here $s = \sqrt{\frac{38}{2}} = 4.3$

why $(n-1)$ not n!

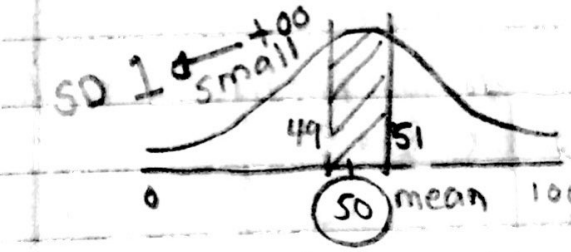
$\begin{bmatrix} 1 \\ 2 \\ 9 \end{bmatrix}$ subtract mean $\begin{bmatrix} -3 \\ -2 \\ +5 \end{bmatrix}$

mean $\bar{y} = 4$



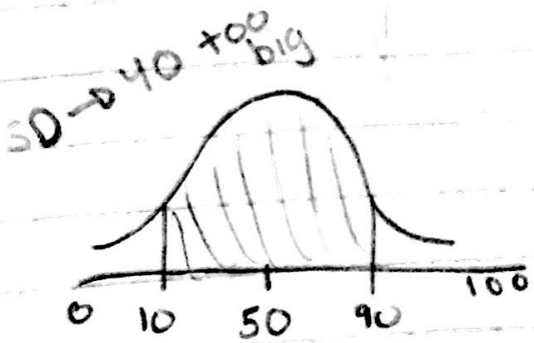


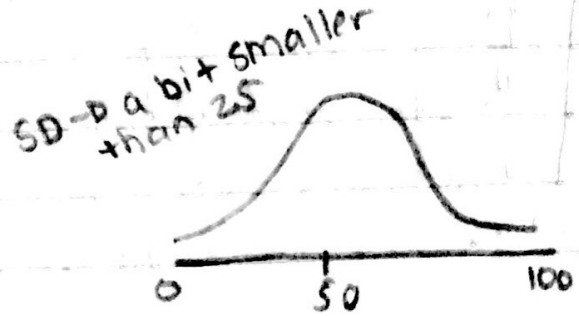
* a data set with n obs. only has $(n-1)$ degrees of freedom for measuring spread



graphical interpretation of SD
* empirical rule

- for virtually any data set, if you start at the mean ± so $\begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}$ $SD_{(s)}$ either way, you will capture
  - about 2/3 → 68%
  - most 95% → 99.7%
  - almost all → of the data in that interval





April 17, 2017

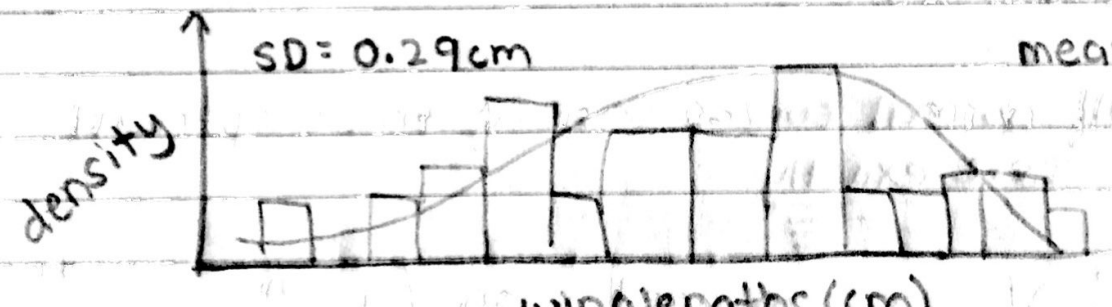
This time: normal curve
Next time: experimental design

read: DD ch. 1-3(A)
ch. 1-6(B)
read: LN pp. 1-94

New HW due date: Friday 21 April @ 11:59pm

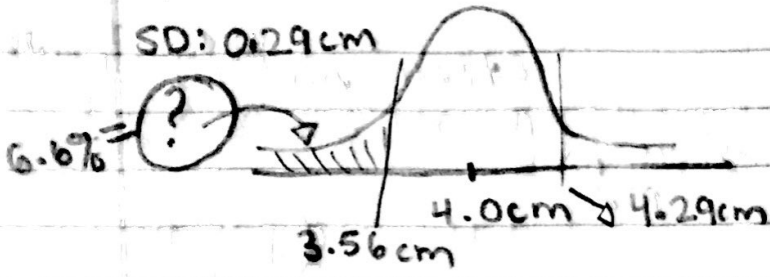
Get course materials packet soon if you don't already have it: will need it for discussion section week two

SG office hours: wed. 3-5pm SG BE 119, not 319



Q: what % of butterflies in data set had wing lengths ≤ 3.56cm?
A_1: (exact) relative freq. from data: $2/24 = 0.083$ or 8.3%
A_2: (exact) area under histogram < 3.56cm $= 2/24 = 8.3\%$

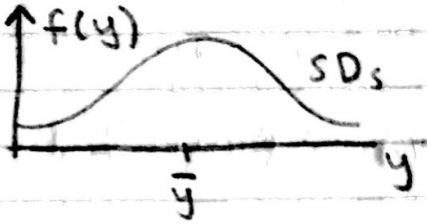
Quetelet (1796-1874)
A_3: (approximate) work out area under bell curve (normal curve) that best approximates histogram to left of 3.56cm



raw units for y is above average

$z = -1.517$
$= -1.52$

normal curve density formula

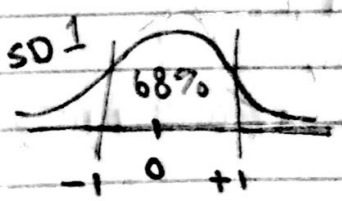


$$f(y)^2 \frac{1}{s\sqrt{2\pi}} \exp\left[\frac{-1}{2s^2}(y-\bar{y})^2\right]$$

not integrable

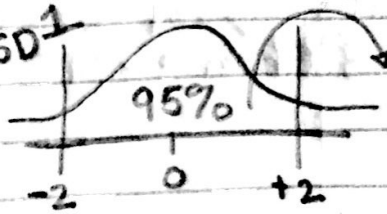
Isaac Newton = numerical integration
Gottfried Leibniz

Fact: All normal curves satisfy the Emperical Rule exactly



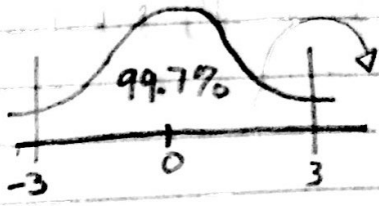
standard normal curve

$$SD = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y-\bar{y})^2}$$



SD = 0 ?
-negative SDs impossible

$$\begin{bmatrix} 4 \\ \vdots \\ 4 \end{bmatrix}$$ SD: 0



$0 \leq SD < \infty$

Look up pages L-34 through L-35

Facts about normal curve density scale:

① it's symmetric

② total area under curve = 100%

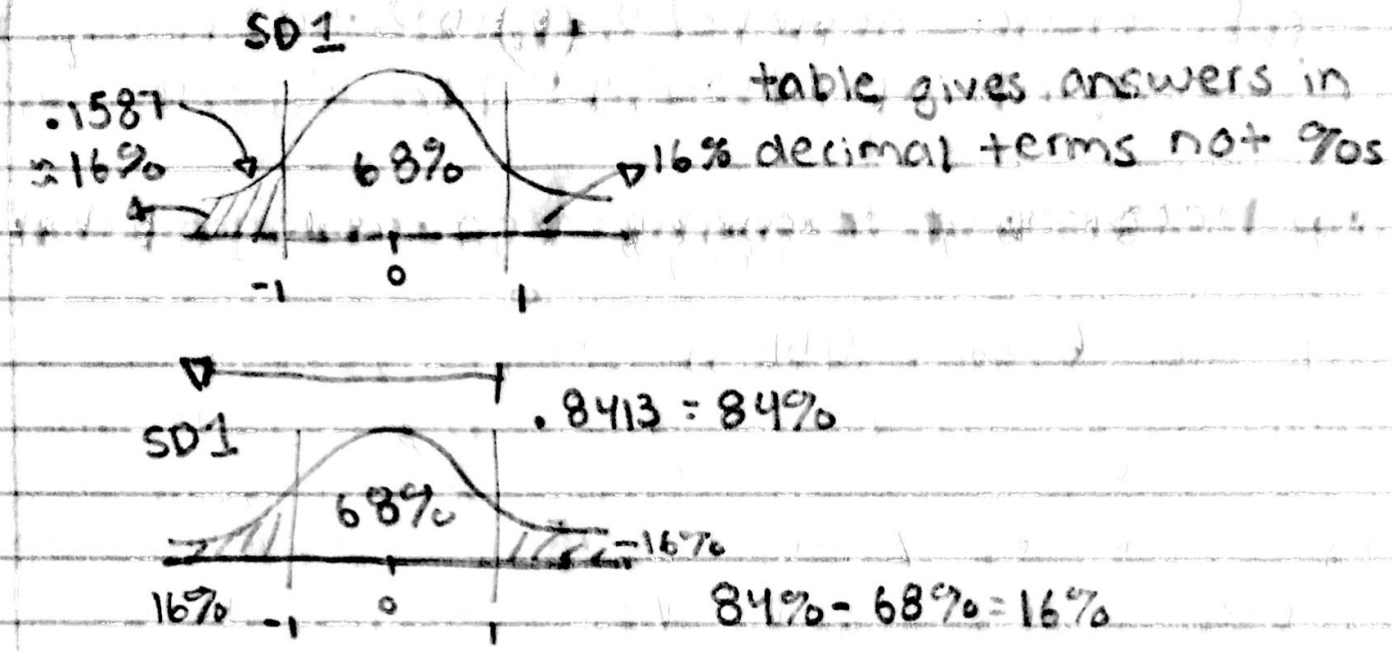


table gives answers in decimal terms not %s

84% - 68% = 16%

Converting from raw units to standard units

$$sw = \frac{\# - \text{mean}}{SD}$$

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

z scores (sw) are unitless

$$\frac{3.56 - 4.0}{0.29} = \frac{-0.44}{0.29} = -1.517$$

area under normal curve = normal approximation

6.6%

normal approx. rather crude to actual rel. freq.

8.3%

Convert y from standard to raw units

$$y = \bar{y} + s \cdot z$$

formula sheet:

Reader - 22 → 28

L-69 R-41

read: science article R-41 → R-50 now, + again in week 10

to decrease uncertainty → get more good data