"How do all these unusuals strike you, Watson?"

"Their cumulative effect is certainly considerable, and yet each of them is quite possible in itself"

Sherlock Holmes and Dr. Watson

The Adventure of Abbey Grange

COHEN CHAP 4. STANDARD & NORMAL

For EDUC/PSY 6600

EXPLORING QUANTITATIVE DATA

We now have a kit of graphical and numerical tools for describing distributions. We also have a strategy for exploring data on a single quantitative variable. Now, we'll add one more step to the strategy.

Exploring Quantitative Data

1. Always plot your data: make a graph.
2. Look for the overall pattern (shape, center, and spread) and for striking departures such as outliers.
3. Calculate a numerical summary to briefly describe center and spread.
4. Sometimes the overall pattern of a large number of observations is so regular that we can describe it by a smooth curve.

DENSITY CURVES & NORMAL DISTRIBUTIONS

Heights (inches)

Mean = 66.3 inches

Median = 66 inches

GPA

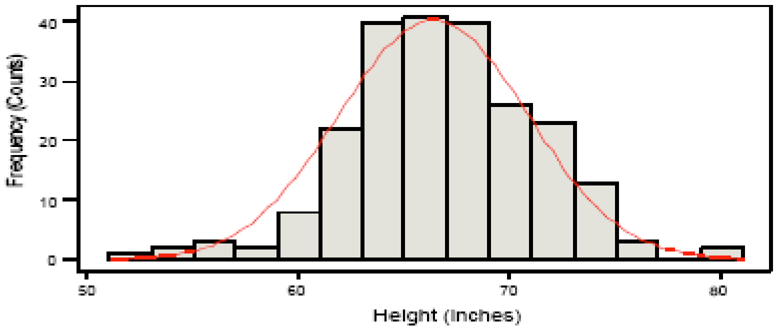
Mean = 3.25

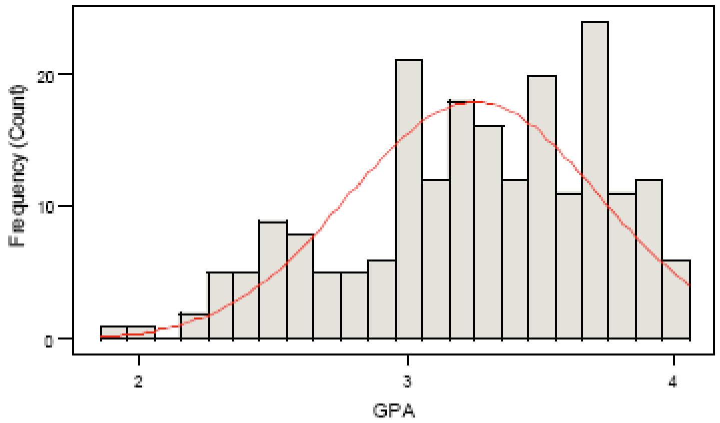
Median = 3.3

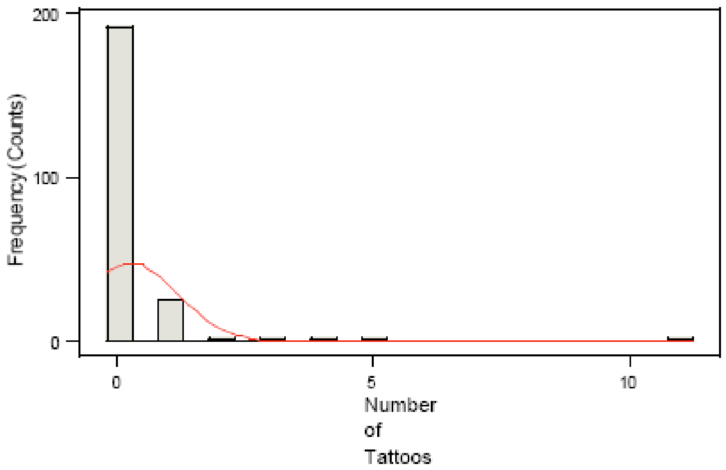
Number of Tattoos

Mean = .23

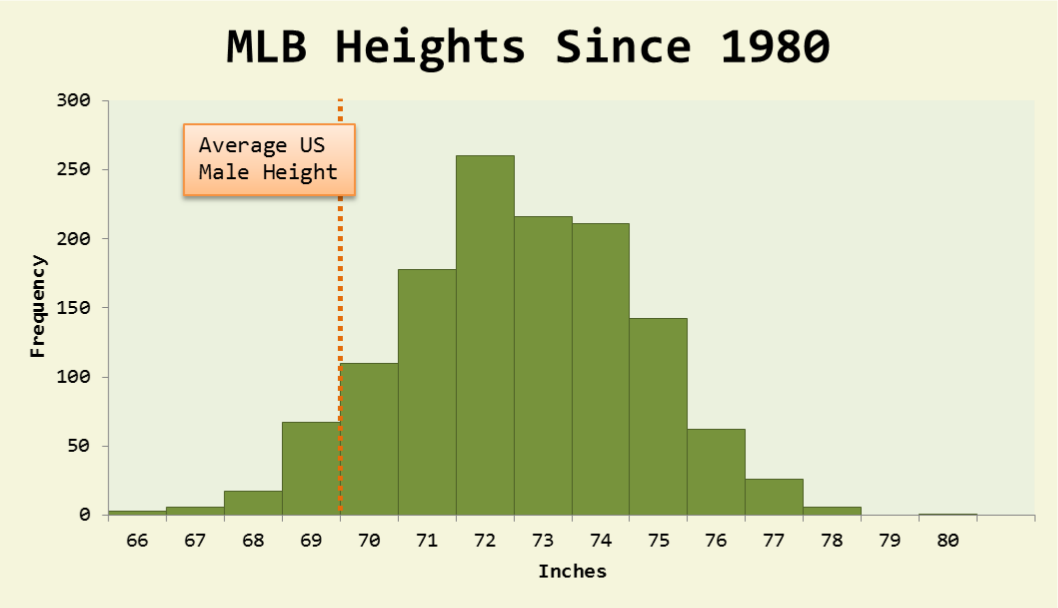
Median = 0







DENSITY CURVES & NORMAL DISTRIBUTIONS



A **density curve** is a curve that:

- is always on or above the horizontal axis
- has an area of exactly 1 underneath it

A density curve describes the overall pattern of a distribution. The area under the curve and above any range of values on the horizontal axis is the proportion of all observations that fall in that range.

NORMAL DISTRIBUTION

Many dependent variables are assumed normally distributed

- Use statistical procedures where data are assumed normally distributed
  - Correlation, regression, t-tests, and ANOVA

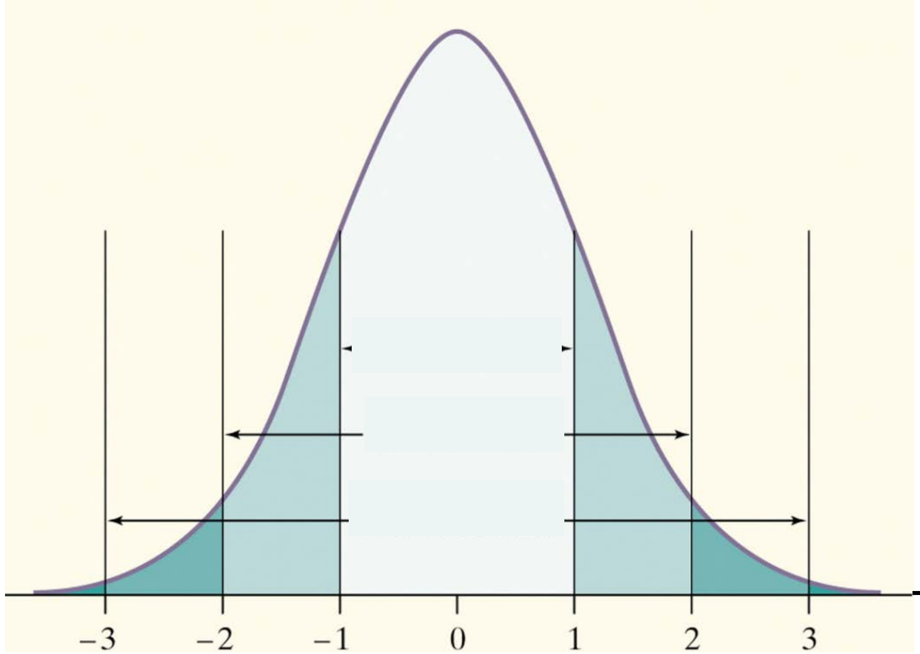
Gaussian distribution

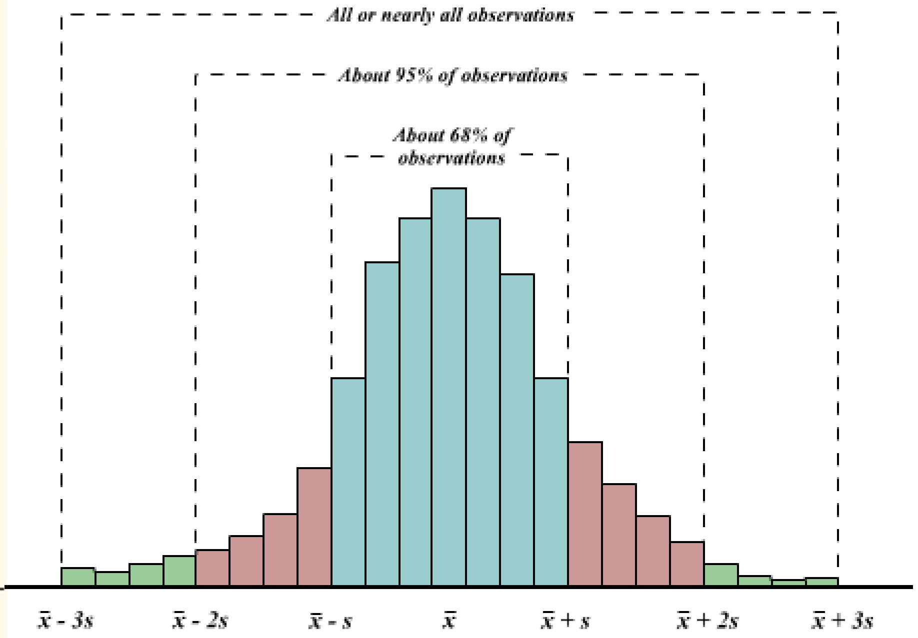
- Karl Gauss

The 68-95-99.7 Rule

In the Normal distribution with mean μ and standard deviation σ:

- Approximately **68%** of the observations fall within σ of μ.
- Approximately **95%** of the observations fall within 2σ of μ.
- Approximately **99.7%** of the observations fall within 3σ of μ.





NORMAL DISTRIBUTIONS

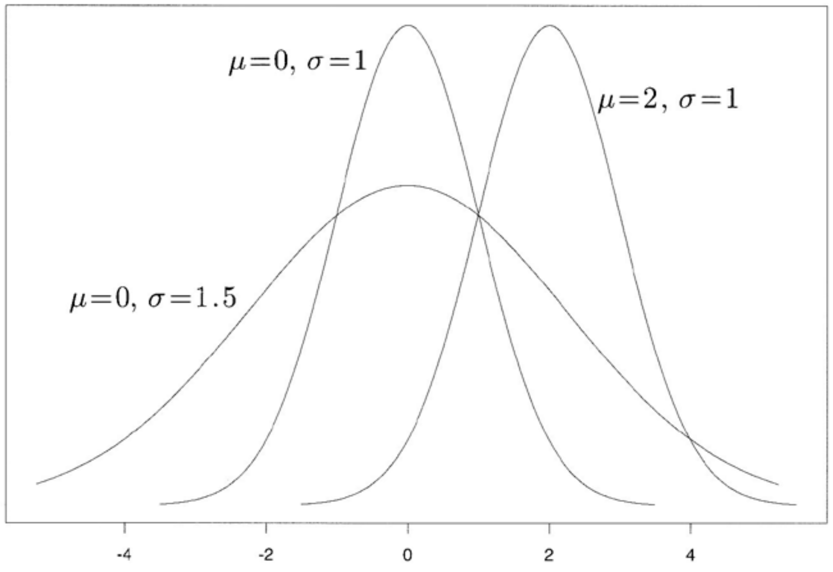
Each μ and σ combination produces differently shaped normal distribution

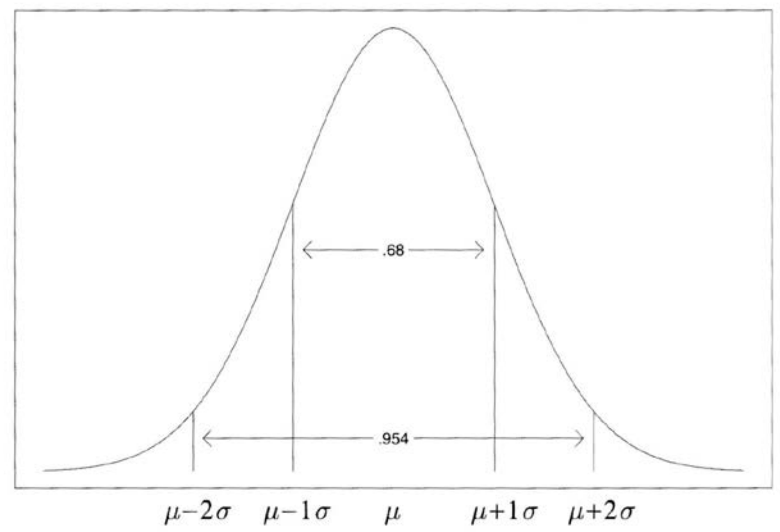
- Family of distributions
- Probability generating function for normal distribution:

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

If we know μ and σ for given variable in given population we can, for given value of X, compute the density (frequency) of that value and thus determine its probability

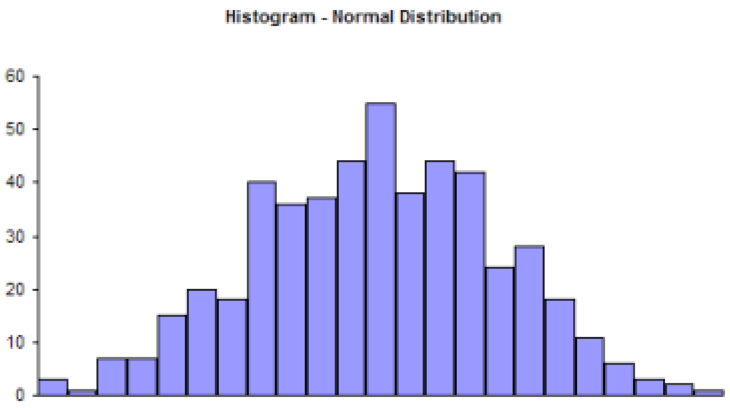
- No matter the exact shape, the properties in terms of area under the curve per *SD* unit are the same!



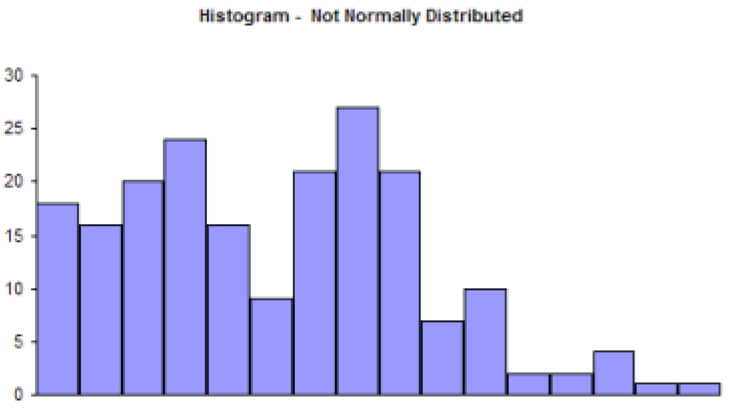


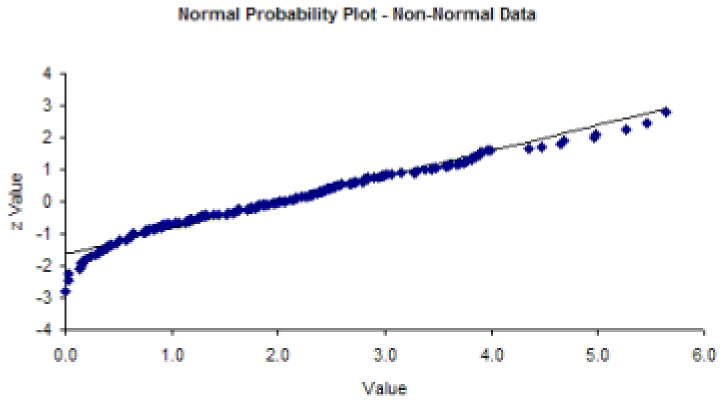
DENSITY CURVES & NORMAL DISTRIBUTIONS

How can you tell if the data is normally distributed? A Q-Q plot!



Normally distributed data will have all a Q-Q plots with the dots all in a straight line





Z-SCORES

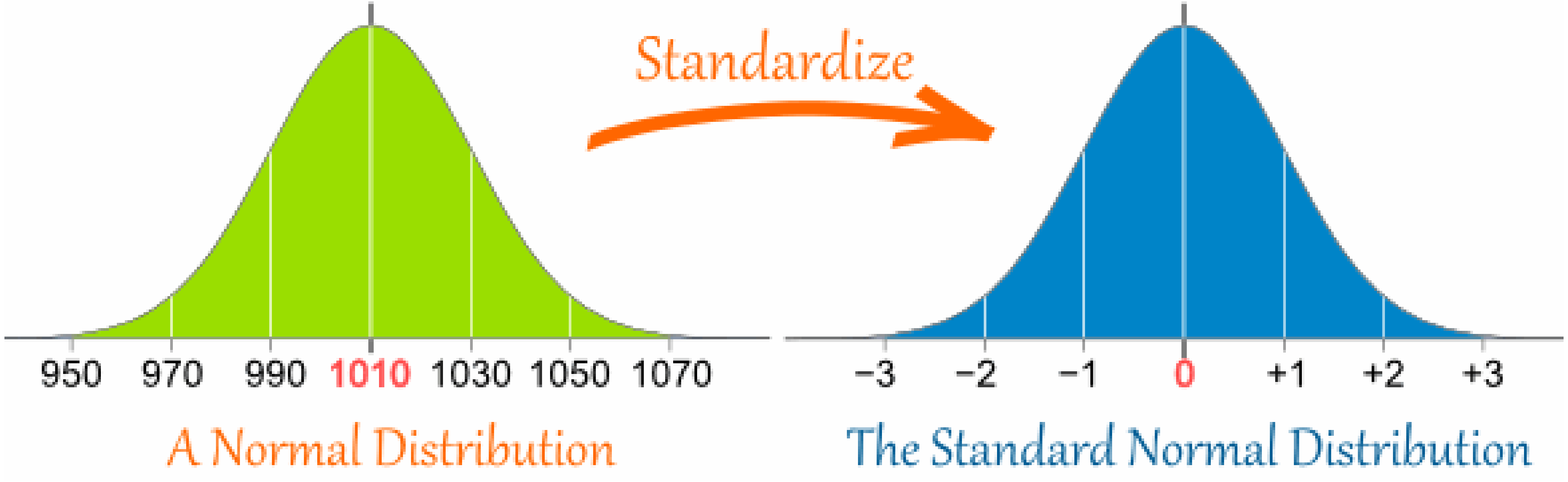
So to convert a value to a Standard Score ("z-score"):

- first subtract the mean,
- then divide by the Standard Deviation

And doing that is called "Standardizing":

z-scores are in *SD* units
Represent *SD* distances
away from *M* (= 0)
z-score = -0.50 → ___ *SD* _____ *M*

Can compare *z*-scores from 2 or more variables originally measured in differing units

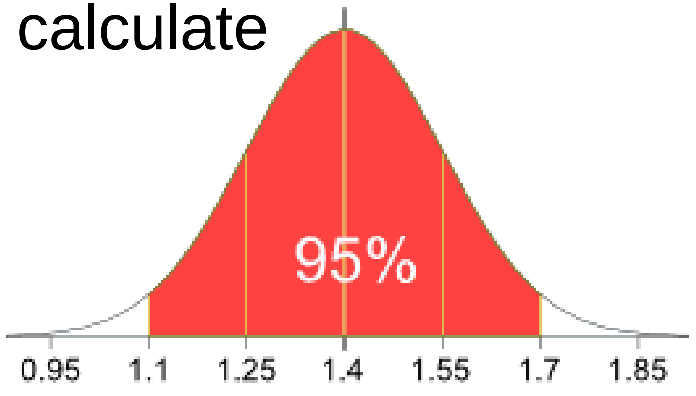


Standardizing does NOT "normalize" data

EXAMPLE: DRAW A PICTURE

95% of students at school are between 1.1m and 1.7m tall

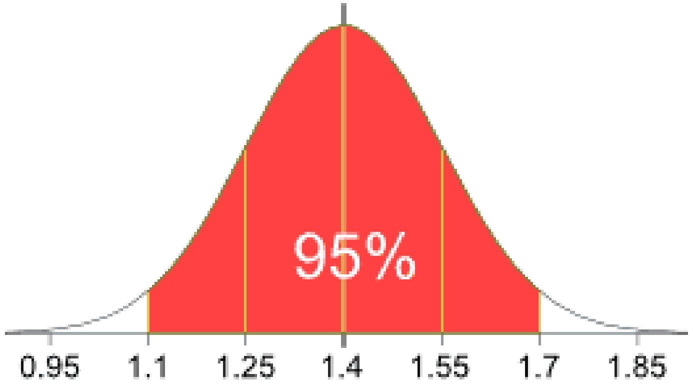
Assuming this data is **normally distributed** can you calculate the mean and standard deviation?



EXAMPLE: CALCULATE Z-SCORE

You have a friend who is 1.85m tall

How far is 1.85 from the mean?



How many standard deviations is that?

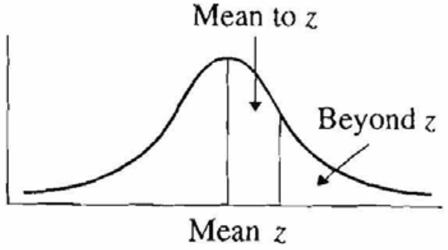
USING Z-SCORES IN THE TABLE

Statistical texts: z or standard normal distribution table

- Only ½ distribution presented in table (symmetrical)
- Add negative sign for z-scores below M

z-scores used to determine area under curve

- Between given z-score and M (0)
- Between given z-score and tail of distribution
- Between 2 z-scores



z	Mean to z	Beyond z	z	Mean to z	Beyond z
.00	.0000	.5000	.43	.1664	.3336
.01	.0040	.4960	.44	.1700	.3300
.02	.0080	.4920	.45	.1736	.3264
.03	.0120	.4880	.46	.1772	.3228
.04	.0160	.4840	.47	.1808	.3192
.05	.0199	.4801	.48	.1844	.3156
.06	.0239	.4761	.49	.1879	.3121
.07	.0279	.4721	.50	.1915	.3085
.08	.0319	.4681	.51	.1950	.3050
.09	.0359	.4641	.52	.1985	.3015
.10	.0398	.4602	.53	.2019	.2981
.11	.0438	.4562	.54	.2054	.2946
.12	.0478	.4522	.55	.2088	.2912
.13	.0517	[illegible]	[illegible]	[illegible]	[illegible]

EXAMPLES: STANDARDIZING SCORES

Assume: School's **population** of student's heights have M (μ) = 1.4 m & SD (σ) = 0.15 m

1. The z-score for student 1.63 m tall = ______

2. Height of student with a z-score of -2.65 = _______

3. The PR of a student that is 1.51 m tall = _______

4. The 90th percentile for student heights = _______

EXAMPLES: FINDING PROBABILITIES

Assume: School's **population** of student's heights have M (μ) = 1.4 m & SD (σ) = 0.15 m

Probability a randomly chosen student is…

more than 1.63 m tall	**less than** 1.2 m tall	**between** 1.2 & 1.63 m tall

EXAMPLES: USING PERCENTILES

Assume: School's **population** of student's heights have M (μ) = 1.4 m & SD (σ) = 0.15 m

What is the **Percentile Rank (PR)** for a student with a height of 1.7 m?	What height correspond to the **15 percentile** in student height?

OTHER NORMAL DISTRIBUTIONS

Which one? Convention and tradition

Name & formula	μ	σ
SAT		
T		
IQ: Standford-Binet		
IQ: Wechsler		

EXAMPLES: CONVERT SCORES

1. Z = -0.2 → _____ SAT score
2. SAT = 520 → _____ z score

3. Z = 1.3 → _____ T score
4. T-score = 38 → _____ z score

5. Z = -3.1 → _____ W-IQ score
6. W-IQ = 127 → _____ z score

PARAMETERS & STATISTICS

Population

"parameters"

N = size
μ = mean
σ^2 = variance
σ = standard deviation

Sample

"statistics"

n = size
$\bar{x}$ = mean
s^2 = variance
s = standard deviation

EXAMPLE: SLEEP

Implies for the entire population

A recent survey describes the distribution of total sleep time among college students as **approximately Normal** with a **mean of 7.02** hours and **standard deviation of 1.15** hours.

Select a college student at random and obtain his or her sleep time. The result is a **random variable *X***. Prior to the random sampling, we don't know the sleep time of the chosen college student, but we do know that in **repeated sampling *X* will have the same *N*(7.02, 1.15)** distribution that describes the pattern of sleep time in the entire population. We call *N*(7.02, 1.15) the ***population distribution***.

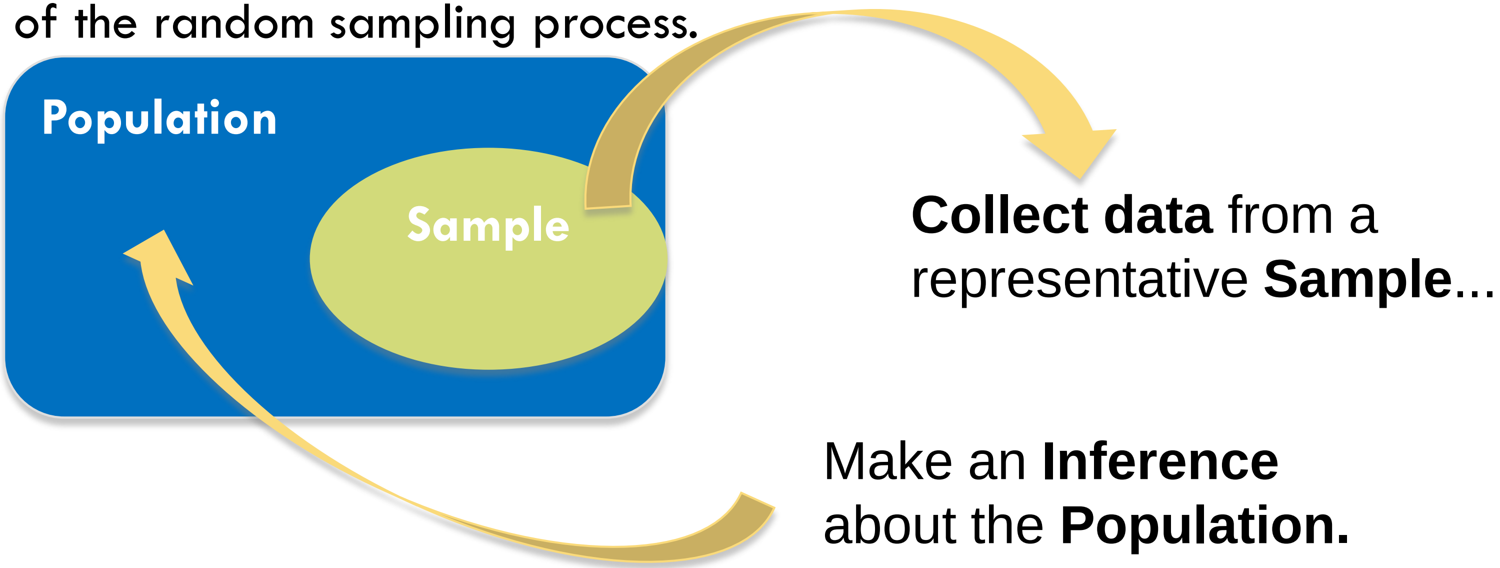
N (___ , ___) means the distribution is **NORMALLY** distributed, with MEAN ___ and STANDARD DEVIATION ___

STATISTICAL ESTIMATION

The process of **statistical inference** involves using information **from a sample** to draw conclusions about a wider population.

Different random samples yield different statistics. We need to be able to describe the **sampling distribution** of **possible statistic values** in order to perform statistical inference.

We can think of a **statistic** as a **random variable** because it takes numerical values that describe the outcomes of the random sampling process.



SAMPLING DISTRIBUTION

The **law of large numbers** assures us that if we measure ***enough*** subjects, the statistic x-bar will eventually get ***very close to*** the unknown parameter μ.

If we took every one of the possible samples of a certain size, calculated the sample mean for each, and graphed all of those values, we'd have a **sampling distribution.**

"Population Distribution" (raw data)

Shows ALL values for all Individuals in the population

"Sampling Distribution"

Shows all values taken by the statistic, in all possible samples of the same size

SAMPLING DISTRIBUTION FOR THE MEAN

Mean of a sampling distribution of a sample mean: just as likely to be above or below μ, *even if* the distribution of the ***raw data*** is skewed.

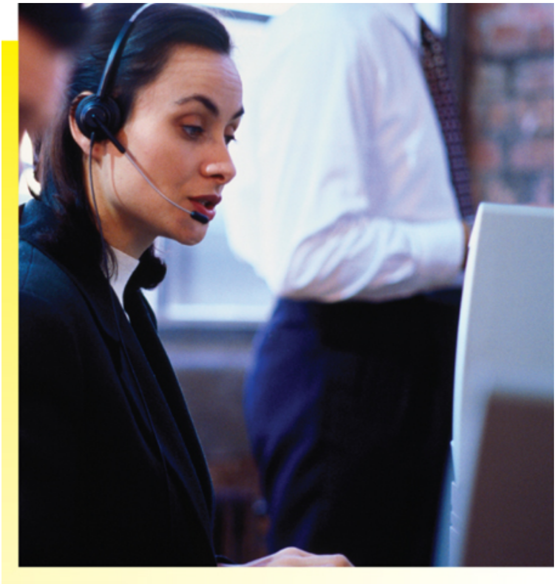
Standard deviation of a sampling distribution of a sample mean: is *smaller* than the standard deviation of the population ***by a factor of*** $\sqrt{n}$.

➔ Averages are less variable than individual observations!

IF		THEN
individual observations	SRS size n ➔	**sample mean**
Mean μ_x &		Mean $\mu_{X\text{-}bar} = \mu_X$ &
Standard Deviation σ_x		Standard Deviation $\sigma_{X\text{-}bar} = \sigma_X / \sqrt{n}$

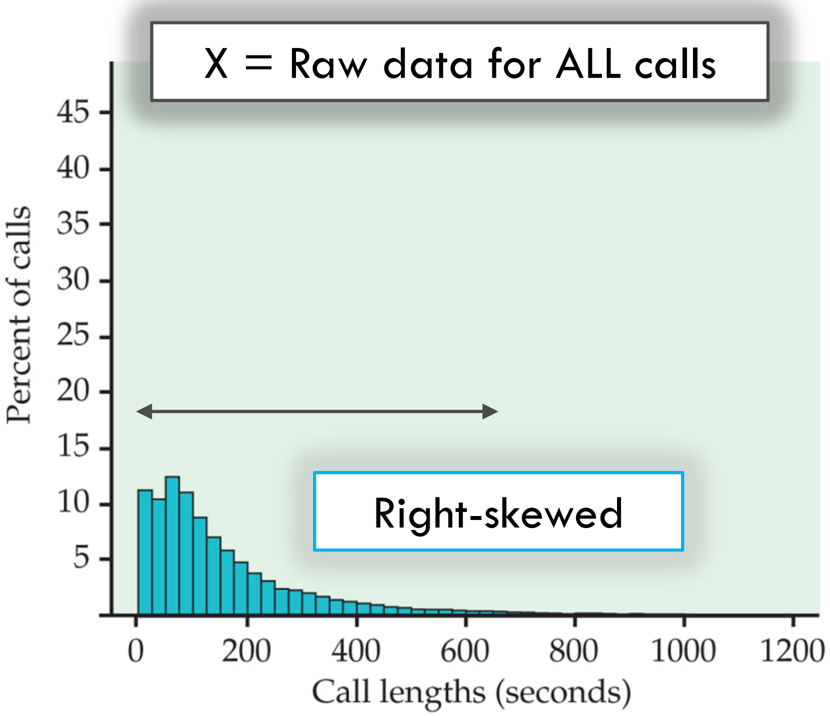
Note : These facts about the mean and standard deviation of $\bar{x}$ are true *no matter what shape the population distribution has*.

EXAMPLE: BANK CALLS

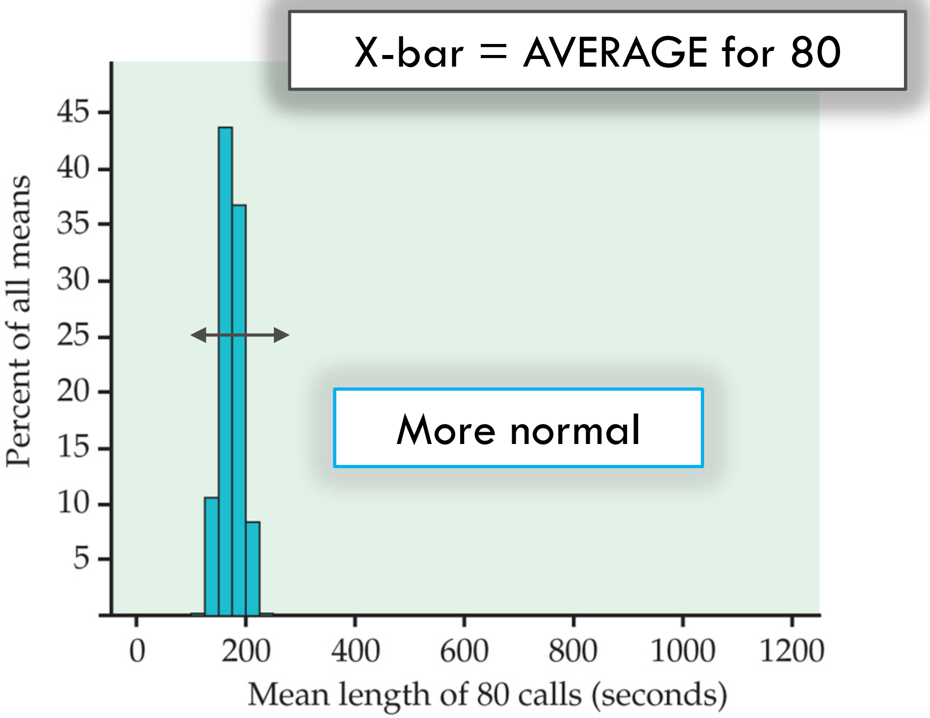


(a) The distribution of lengths of **all** customer service calls received by a bank in a month.

(b) The distribution of the **sample** means x-bar for 500 random samples of **size 80** from this population. The scales and histogram classes are exactly the same in both panels



(a)

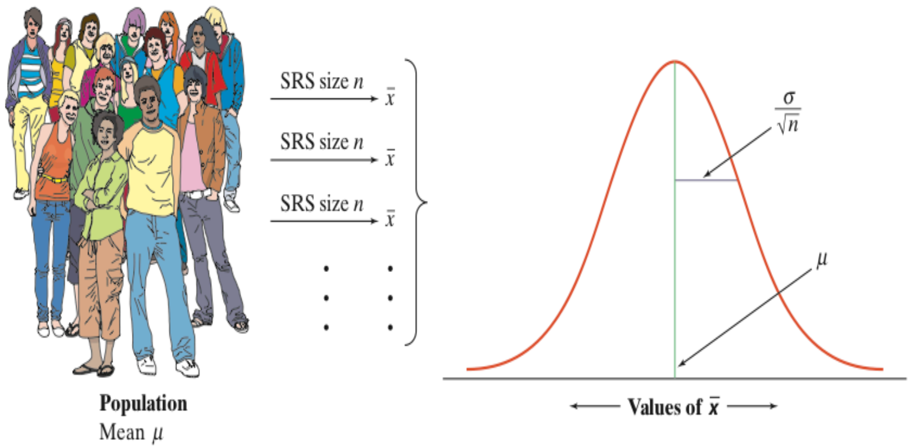


(b)

SAMPLING DISTRIBUTION FOR THE MEAN

What if the **population** distribution was **NORMAL**?

IF
individual observations
have the
$N(\mu,\sigma)$ distribution



THEN
the **sample mean**
of an SRS of size n
has the $N(\mu, \sigma/\sqrt{n})$
distribution

"SE"
Standard
error

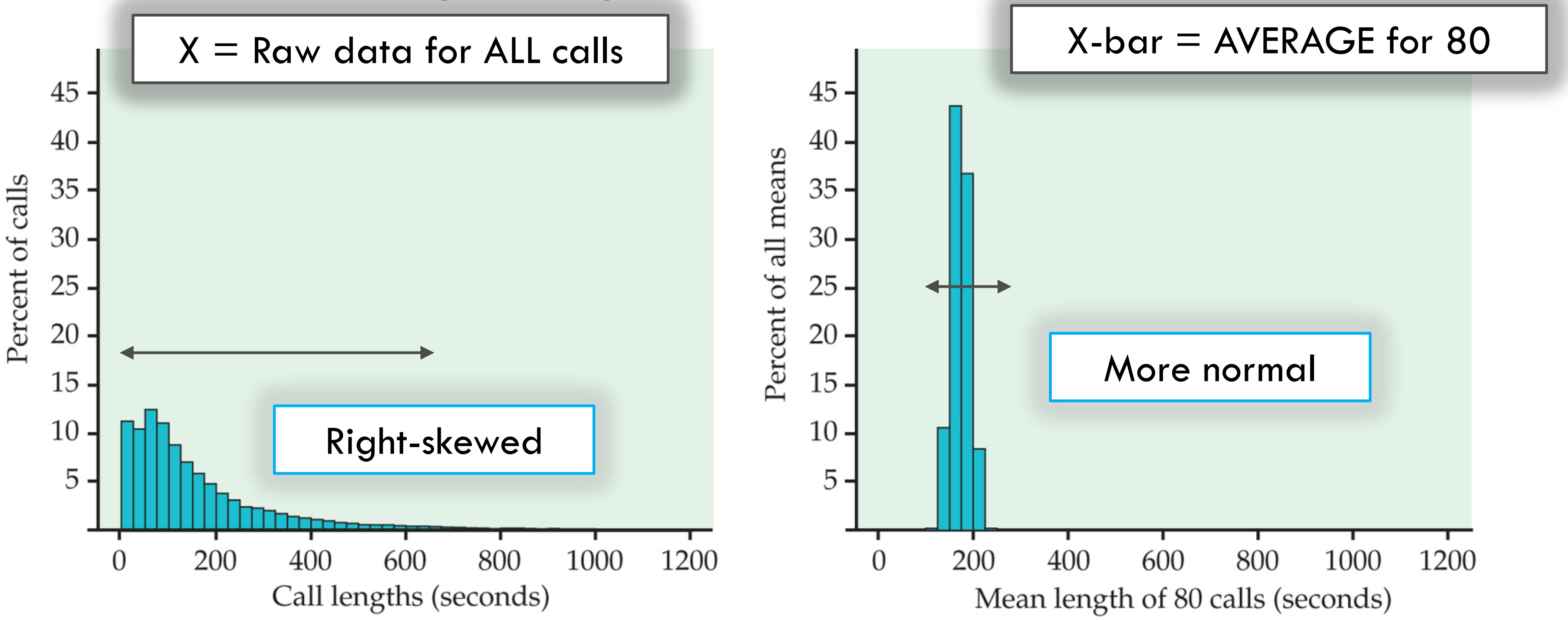
SE for mean
= SD divided
by square
root of the
sample size

What if the **population** distribution is **NOT** normal, or even discrete?

Draw an SRS of size n from any population with mean μ and finite standard deviation σ. The **central limit theorem (CLT)** says that when n is large, the sampling distribution of the sample mean $\bar{x}$ is approximately Normal:

$$\bar{x} \text{ is approximately } N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

EXAMPLE: BANK CALLS



(a) (b)

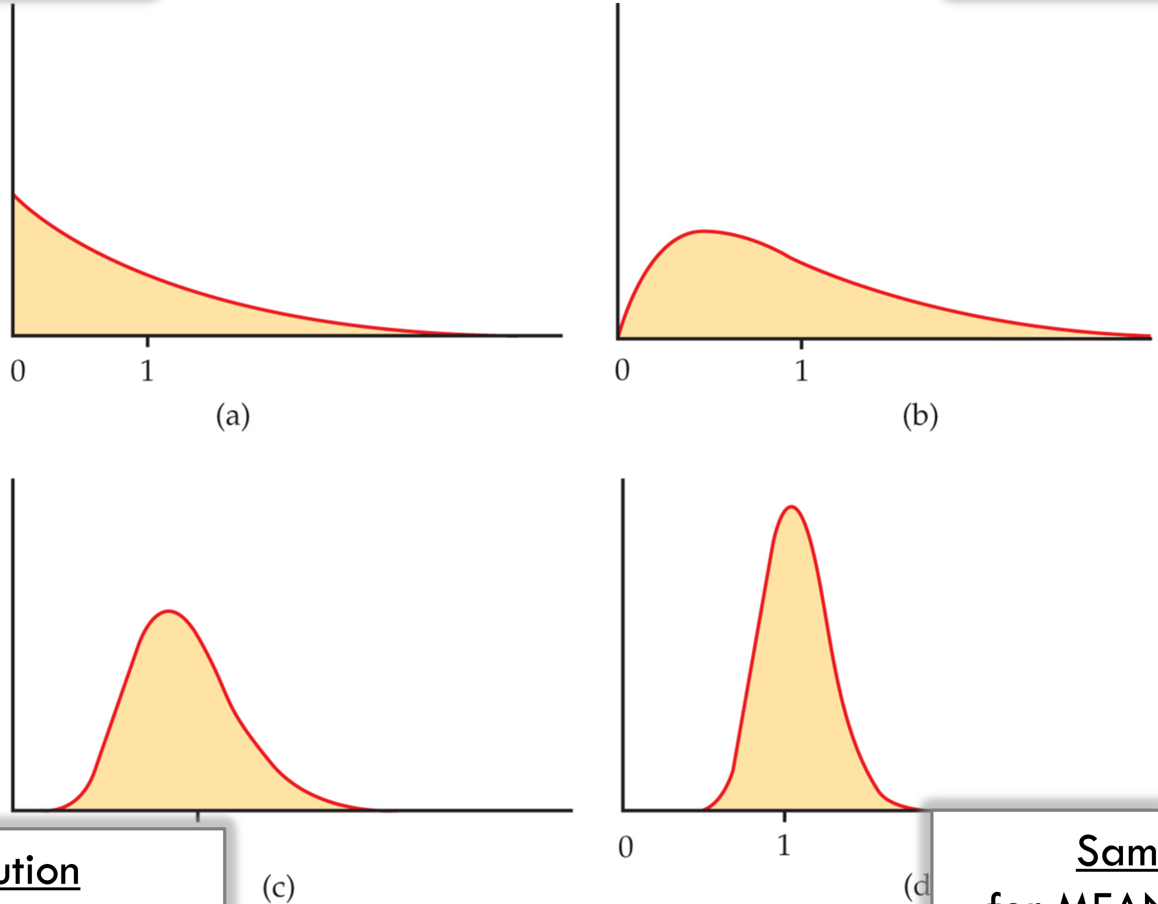
The standard deviation of the population of service call lengths is σ_x = 184.81 sec. *The length of a single call will often be far from the population mean.*
What is the standard deviation for a SRS sample of 80 calls?

If we choose an SRS of 20 calls, the standard deviation of their mean length is…

THE CENTRAL LIMIT THEOREM

Population Distribution
(sample size 1)

<u>Sampling Distribution</u>
for MEAN of a sample size 2



<u>Sampling Distribution</u>
for MEAN of a sample size 10

<u>Sampling Distribution</u>
for MEAN of a sample size 10

EXAMPLES: FINDING PROBABILITIES

Assume: School's **population** of student's heights have M (μ) = 1.4 m & SD (σ) = 0.15 m

Probability the height of **a** randomly chosen **student** is…	Probability the **MEAN** of a randomly chosen **GROUP of 16** students is…
more than 1.63 m tall	**more than** 1.63 m tall

EXAMPLE: CANCER DATASET

```
GET FILE='C:\Users\A00315273\Box Sync\Teaching\Educ6600\Dataset\Cancer.sav'.
DATASET NAME DataSet1 WINDOW=FRONT.

VARIABLE LABELS
ID "Patient identification number"
TRT "Treatment Group"
AGE "Patient's Incoming Age"
WEIGHIN "Patient's Incoming Weight in pounds"
STAGE "Patient's Stage of Cancer".

VALUE LABELS TRT 0 "control" 1 "aleo treatment".

RECODE
TRT AGE WEIGHIN STAGE TOTALCIN TOTALCW2 TOTALCW4 TOTALCW6
   (SYSMIS = 999).
EXECUTE.

VALUE LABELS
TRT
   0 "control" 1 "aleo treatment" 999 "missing"/
AGE WEIGHIN STAGE TOTALCIN TOTALCW2 TOTALCW4 TOTALCW6
   999 "missing".

MISSING VALUES
TRT AGE WEIGHIN STAGE TOTALCIN TOTALCW2 TOTALCW4 TOTALCW6
   (999).
```

Available on Canvas
Save to your computer
Edit the path to match

SPSS: CREATE A Z-SCORE VARIABLE

```
* first find M and SD.

FREQUENCIES AGE
   /FORMAT NOTABLE
   /STATISTICS MEAN STDDEV.
```

Statistics

AGE Patient's Incoming Age

N	Valid	25
	Missing	0
Mean		59.64
Std. Deviation		12.932

```
* then create the new variable.

COMPUTE zAGE = (AGE - 59.64) / 12.932.
EXECUTE.


* Check to see if it looks ok.

FREQUENCIES AGE zAGE
   /FORMAT NOTABLE
   /STATISTICS MEAN STDDEV.
```

Statistics

		AGE Patient's Incoming Age	zAGE
N	Valid	25	25
	Missing	0	0
Mean		59.64	.0000
Std. Deviation		12.932	1.00001

	ID	TRT	AGE	VEIGHIN	STAGE	TOTAL...	TOTALCW2	TOTALCW4	TOTALCW6	zAGE
4	6	0	60	137	4	7	9	17	19	.03
5	9	0	61	180	1	6	7	9	3	.11
6	11	0	59	176	2	6	7	16	13	-.05
7	12	1	56	227	4	6	10	11	9	-.28
8	14	1	42	163	1	4	6	8	7	-1.36
9	15	0	69	168	1	6	6	6	11	.72
10	16	1	44	261	2	6	11	11	14	-1.21
11	21	0	67	186	1	6	11	11	10	.57
12	22	1	27	225	1	6	7	6	6	-2.52
13	24	1	68	226	4	12	11	12	9	.65
14	26	0	56	158	3	6	11	15	15	-.28
15	31	0	61	213	1	6	9	6	8	.11
16	34	1	77	164	2	5	7	13	12	1.34
17	35	0	51	189	1	6	4	8	7	-.67
18	37	1	86	140	1	6	7	7	7	2.04
19	39	0	46	149	4	7	8	11	11	-1.05
20	41	0	65	157	1	6	6	9	6	.41
21	42	1	73	182	0	8	11	16	999	1.03
22	44	1	67	187	1	5	7	7	7	.57
	45	0	67	186	1	8	8	9	10	.57
	50	1	60	164	2	6	8	16	999	.03
	58	1	54	173	4	7	8	10	8	-.44

SPSS: TRANSFORMING VARIABLES

* This is useful IF you have a variable that is POSITIVELY SKEWED, since the methods we will learn all require your variables are NORMALLY distributed.

```
* one version is a square root.

COMPUTE sqrt_AGE = SQRT(AGE).
EXECUTE.


* another option is the (natural) logrithm.

COMPUTE ln_AGE = ln(AGE).
EXECUTE.
```