Standard and Normal

Cohen Chapter 4

EDUC/PSY 6600

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

Exploring Quantitative Data

Building on what we've already discussed:

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.

Let's Start with Density Curves

A density curve is a curve that:

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

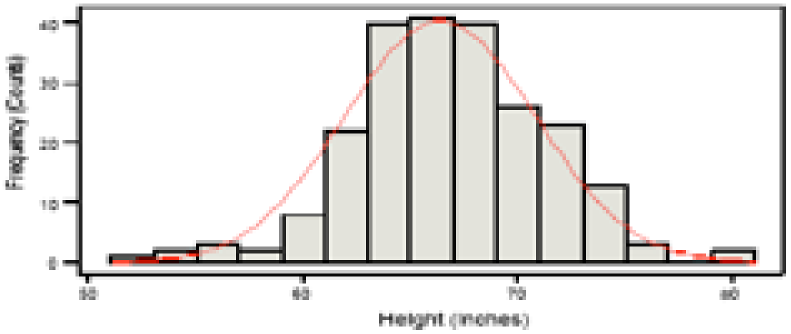
It describes the overall pattern of a distribution and highlights proportions of observations as the area.

Density Curves and Normal Distributions

Heights (inches)

Mean = 66.3 inches

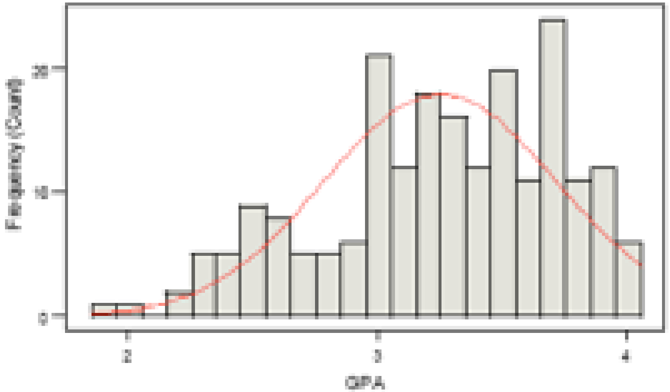
Median = 66 inches



GPA

Mean = 3.25

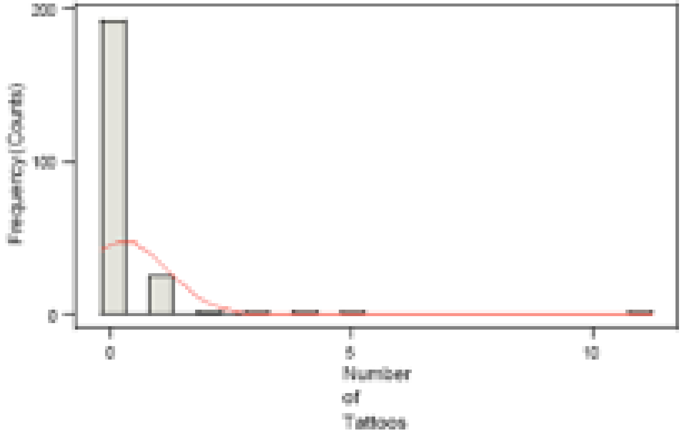
Median = 3.3

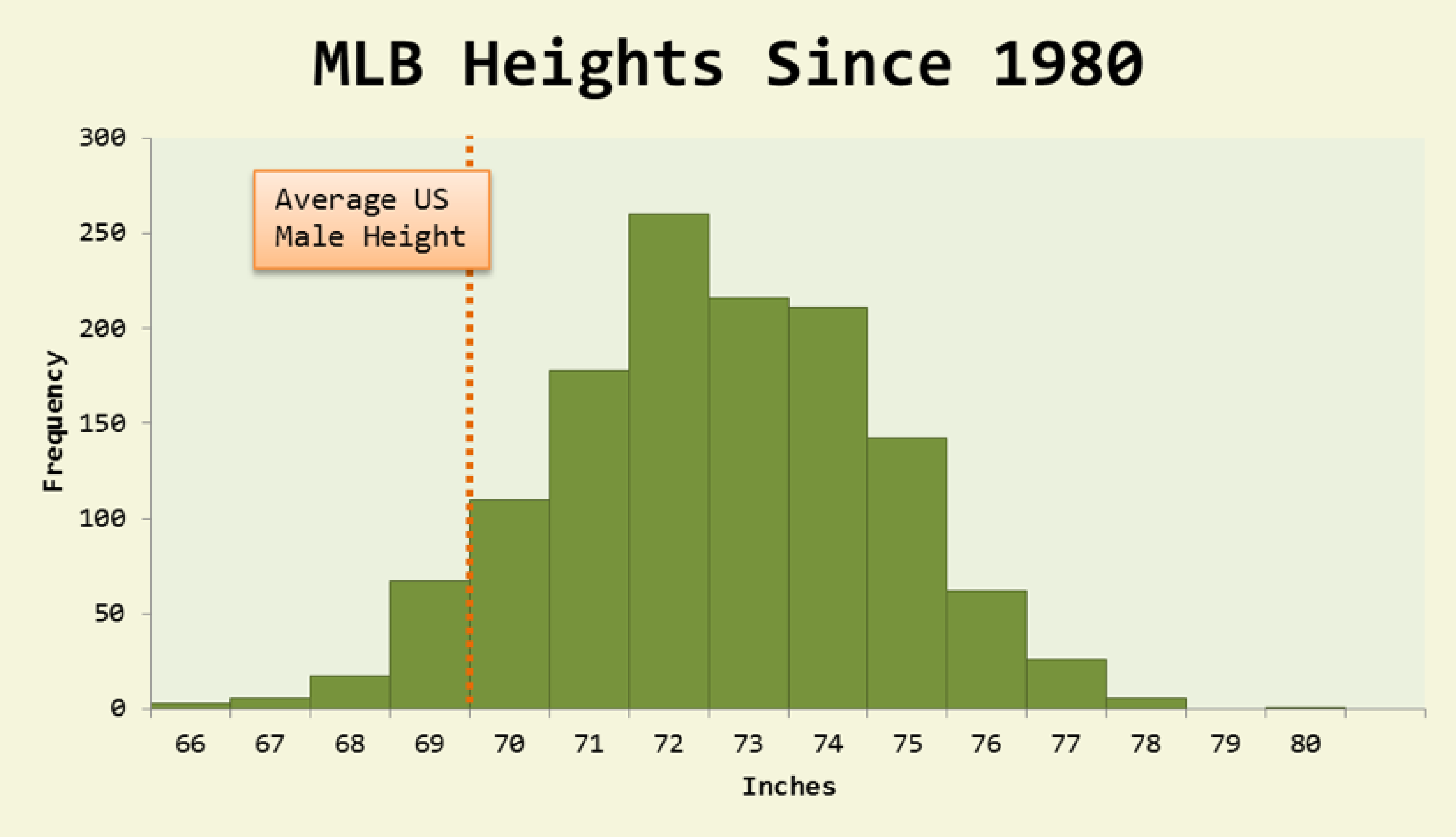


Number of Tattoos

Mean = .23

Median = 0

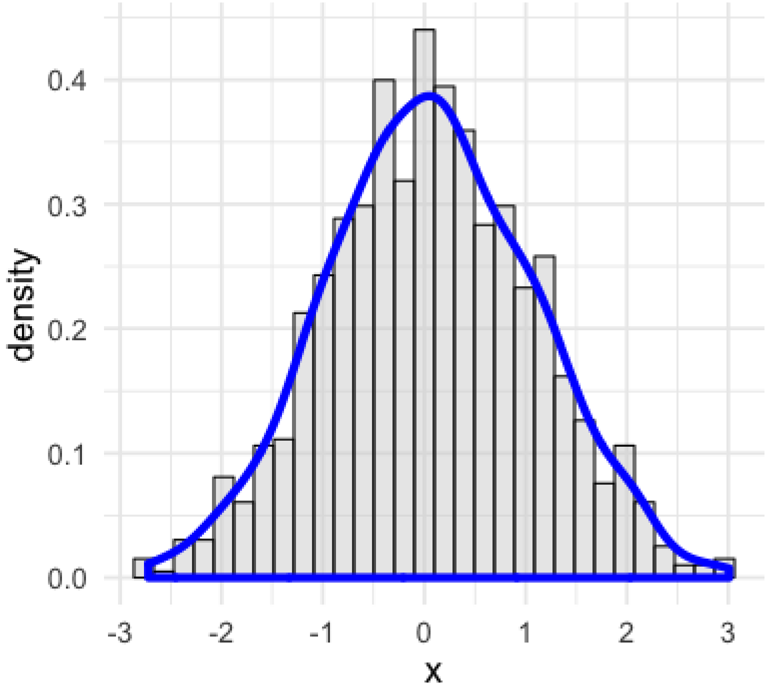



MLB Heights Since 1980
Average US
Male Height
Frequency
300
250
200
150
100
50
0
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
Inches

Normal Distribution

Many dependent variables are assumed to be normally distributed

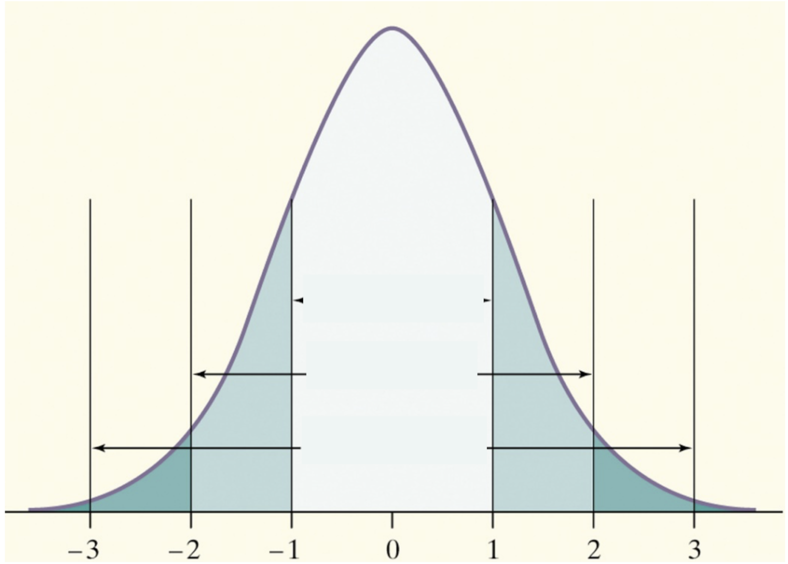
- Many statistical procedures assume this
  - Correlation, regression, t-tests, and ANOVA
- Also called the Gaussian distribution
  - for Karl Gauss

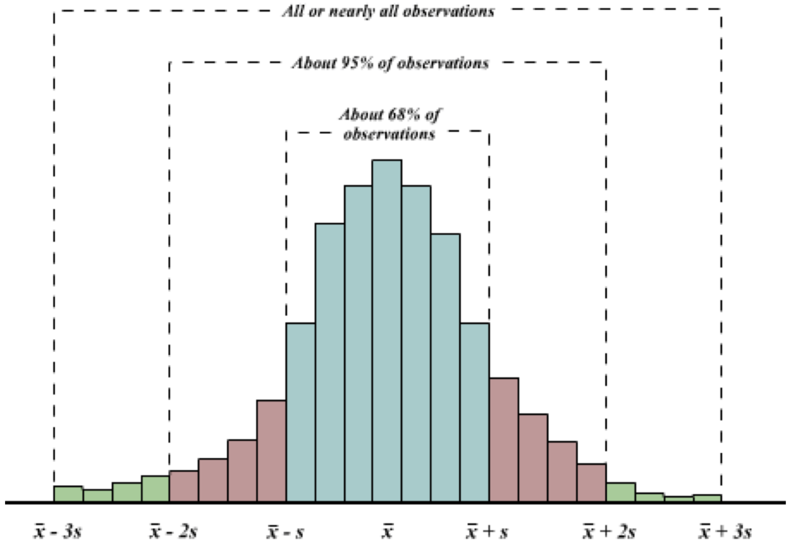


The 68-95-99.7 Rule

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.

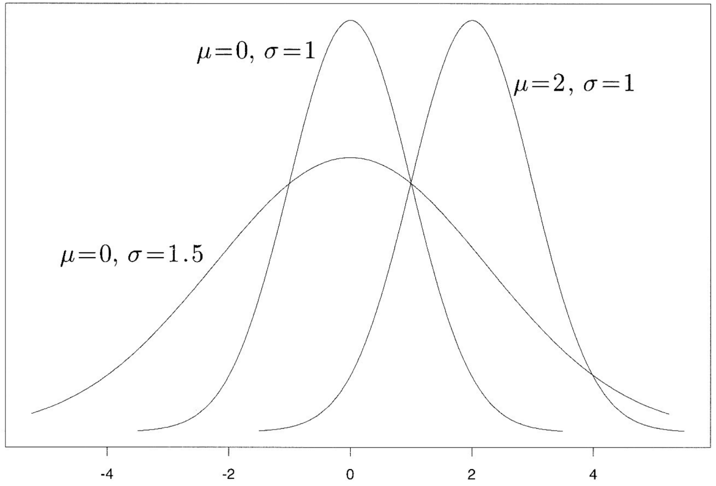




Each μ and σ combination produces differently shaped **normal distribution**

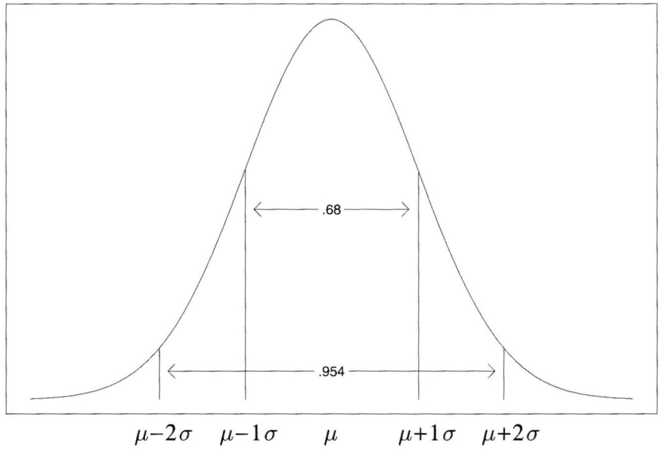
- Family of distributions
- Probability generating function for normal distribution:

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



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

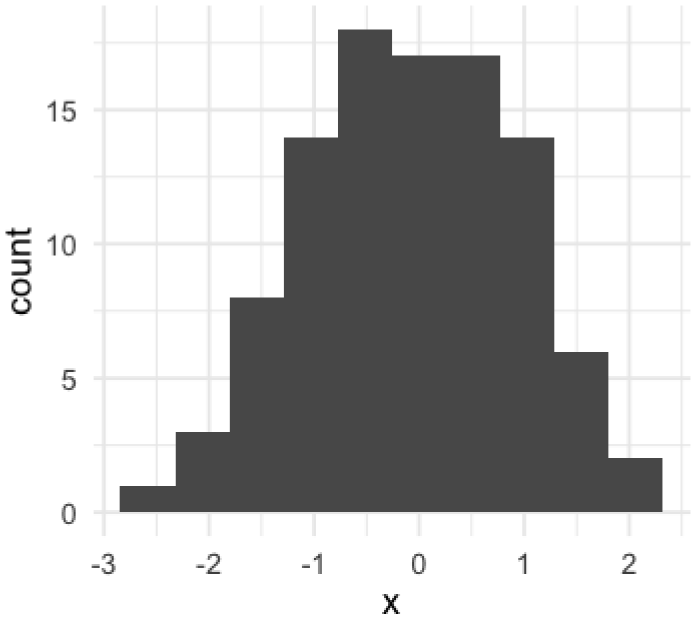
- No matter the exact shape, the properties in terms of area under the curve per *SD* unit are the same!



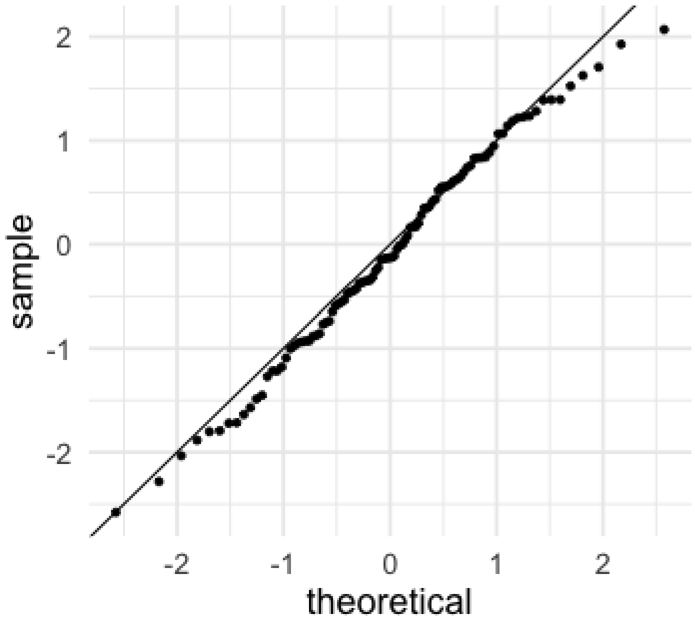
Do We Have a Normal Distribution?

Check Plot!

Bell shaped curve?



Points on the line?



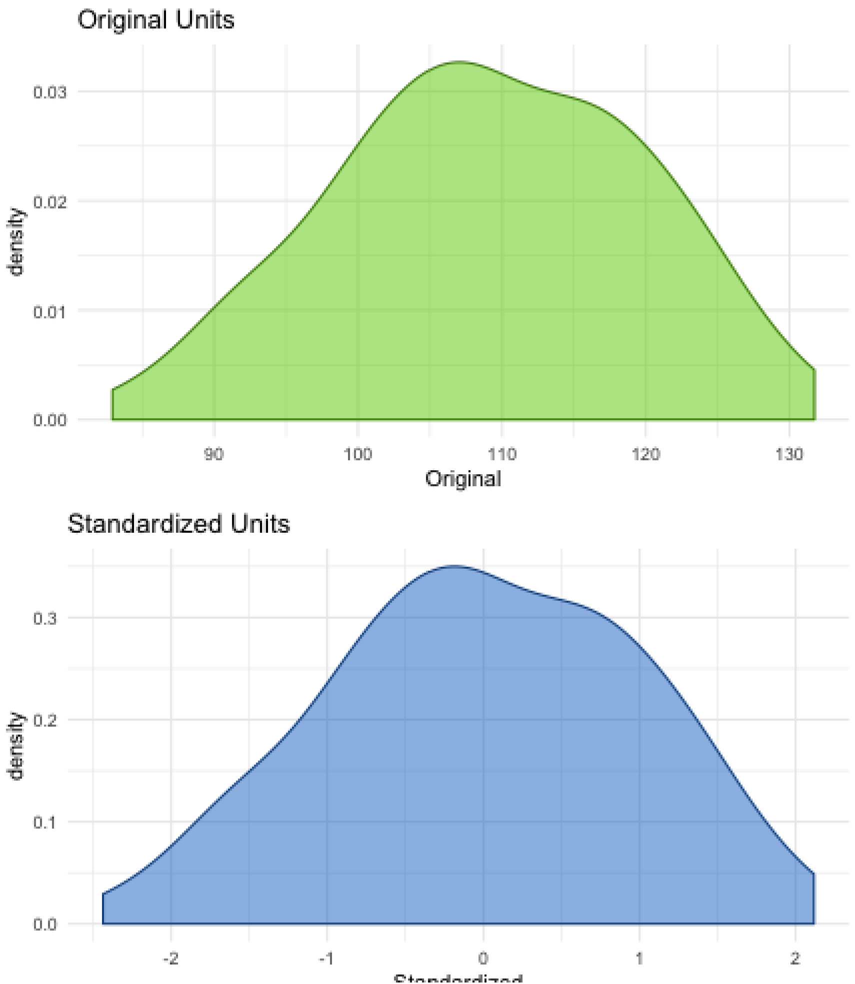
Z-Scores, Computation

Standardizing

Convert a value to a standard score ("z-score")

- First subtract the mean
- Then divide by the standard deviation

$$z = \frac{X - \mu}{\sigma} = \frac{X - \bar{X}}{s}$$



Z-Scores, Units

- z-scores are in SD units
- Represent SD distances away from the mean (M = 0)
  - if z-score = -0.50 then it is $\frac{1}{2}$ of SD below mean
- Can compare z-scores from 2 or more variables originally measured in differing units

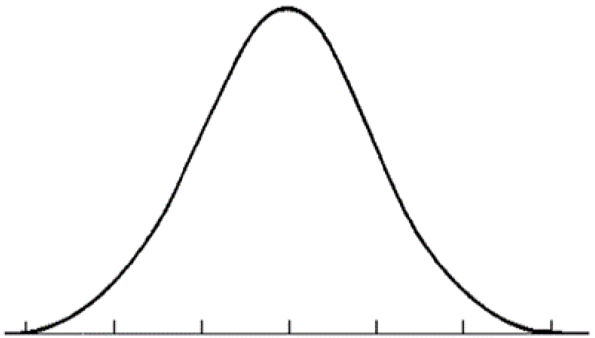
Note: Standardizing does NOT "normalize" the data

Let's Apply This to an Exmple Situation

Example: Draw a Picture

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

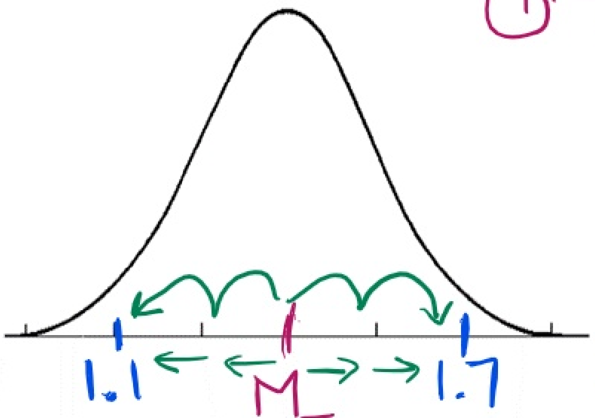
Assuming this data is normally distributed, can you calculate the MEAN and STANDARD DEVIATION?



Example: Draw a Picture

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

Assuming this data is normally distributed, can you calculate the MEAN and STANDARD DEVIATION?



① the mean is in the middle

$$M = \frac{1.1 + 1.7}{2} = \frac{2.8}{2}$$

meters

$$M = 1.4$$

② 95% = ± 2 SD's

$$SD = \frac{1.7 - 1.1}{4} = \frac{0.6}{4}$$

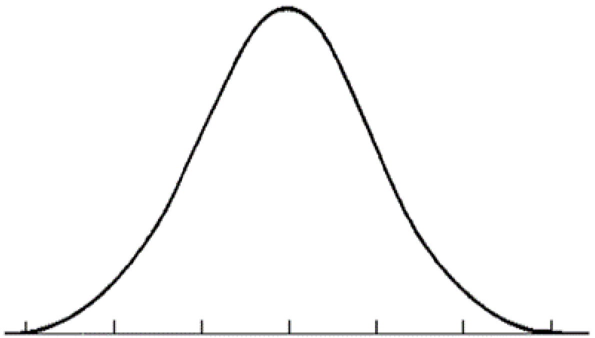
$$SD = 0.15$$

Example: Calculate a z-Score

You have a friend who is 1.85 meters tall.

Class: M = 1.4 meters, SD = 0.15 meters

How far is 1.85 from the mean? How many standard deviations is that?

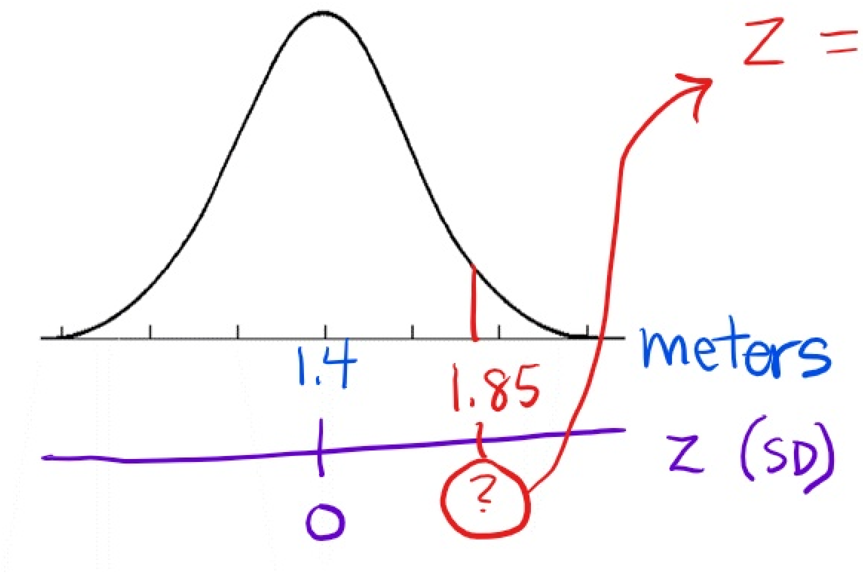


Example: Calculate a z-Score

You have a friend who is 1.85 meters tall.

Class: M = 1.4 meters, SD = 0.15 meters

How far is 1.85 from the mean? How many standard deviations is that?

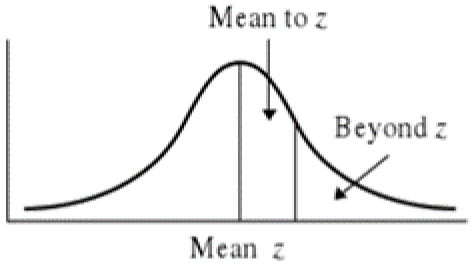


$$z = \frac{X - \mu}{\sigma} = \frac{1.85 - 1.40}{0.15}$$

$$= \frac{0.45 \text{ meters}}{0.15}$$

$$= \boxed{3 \text{ SD's above avg}}$$

Using the z-Table



z	Mean to z	Beyond z	z	Mean to z	Beyond z
.00	.0000	.5000	.41	.1591	.3409
.01	.0040	.4960	.42	.1628	.3372
.02	.0080	.4920	.43	.1664	.3336
.03	.0120	.4880	.44	.1700	.3300
.04	.0160	.4840	.45	.1736	.3264
.05	.0199	.4801	.46	.1772	.3228
.06	.0239	.4761	.47	.1808	.3192
.07	.0279	.4721	.48	.1844	.3156
.08	.0319	.4681	.49	.1879	.3121
.09	.0359	.4641	.50	.1915	.3085
.10	.0398	.4602	.51	.1950	.3050
.11	.0438	.4562	.52	.1985	.3015
.12	.0478	.4522	.53	.2019	.2981
.13	.0517	.4483	.54	.2054	.2946
.14	.0557	.4443	.55	.2088	.2912
.15	.0596	.4404	.56	.2123	.2877
.16	.0636	.4364	.57	.2157	.2843

z	Mean to z	Beyond z	z	Mean to z	Beyond z
2.18	.4854	.0146	2.72	.4967	.0033
2.19	.4857	.0143	2.73	.4968	.0032
2.20	.4861	.0139	2.74	.4969	.0031
2.21	.4864	.0136	2.75	.4970	.0030
2.22	.4868	.0132	2.76	.4971	.0029
2.23	.4871	.0129	2.77	.4972	.0028
2.24	.4875	.0125	2.78	.4973	.0027
2.25	.4878	.0122	2.79	.4974	.0026
2.26	.4881	.0119	2.80	.4974	.0026
2.27	.4884	.0116	2.81	.4975	.0025
2.28	.4887	.0113	2.82	.4976	.0024
2.29	.4890	.0110	2.83	.4977	.0023
2.30	.4893	.0107	2.84	.4977	.0023
2.31	.4896	.0104	2.85	.4978	.0022
2.32	.4898	.0102	2.86	.4979	.0021
2.33	.4901	.0099	2.87	.4979	.0021
2.34	.4904	.0096	2.88	.4980	.0020
2.35	.4906	.0094	2.89	.4981	.0019
2.36	.4909	.0091	2.90	.4981	.0019
2.37	.4911	.0089	2.91	.4982	.0018
2.38	.4913	.0087	2.92	.4982	.0018
2.39	.4916	.0084	2.93	.4983	.0017
2.40	.4918	.0082	2.94	.4984	.0016
2.41	.4920	.0080	2.95	.4984	.0016
2.42	.4922	.0078	2.96	.4985	.0015
2.43	.4925	.0075	2.97	.4985	.0015
2.44	.4927	.0073	2.98	.4986	.0014
2.45	.4929	.0071	2.99	.4986	.0014
2.46	.4931	.0069	3.00	.4987	.0013
2.47	.4932	.0068	3.20	.4993	.0007
2.48	.4934	.0066			
2.49	.4936	.0064	3.40	.4997	.0003
2.50	.4938	.0062			
2.51	.4940	.0060	3.60	.4998	.0002
2.52	.4941	.0059			
2.53	.4943	.0057	3.80	.4999	.0001
2.54	.4945	.0055			
2.55	.4946	.0054	4.00	.49997	.00003

Examples: Standardizing Scores

Assume: School's population of students heights are normal (M = 1.4m, SD = 0.15m)

1. The z-score for a student 1.63 m tall = __
2. The height of a student with a z-socre of -2.65 = __
3. The Pecentile Rank of a student that is 1.51 m tall = __
4. The 90th percentile for students heights = __

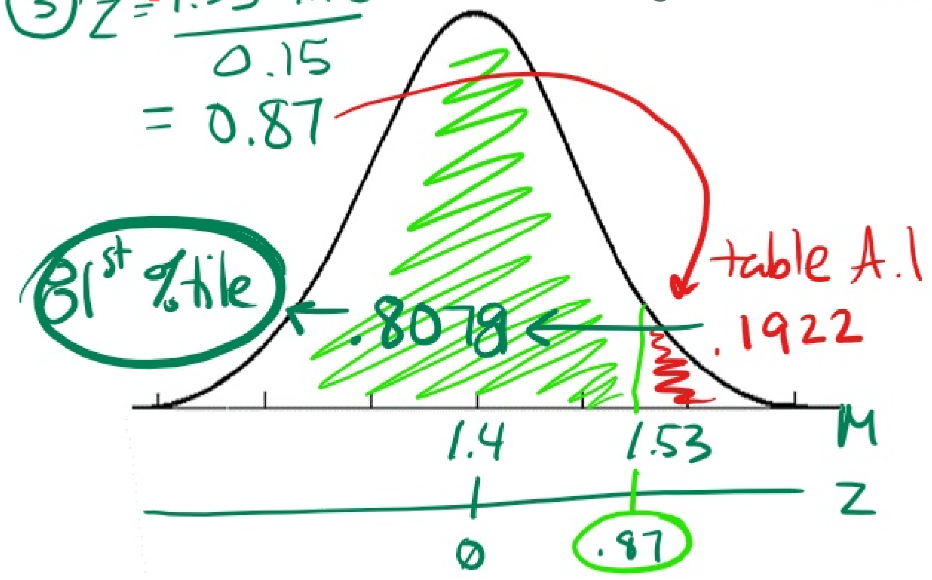
Examples: Standardizing Scores

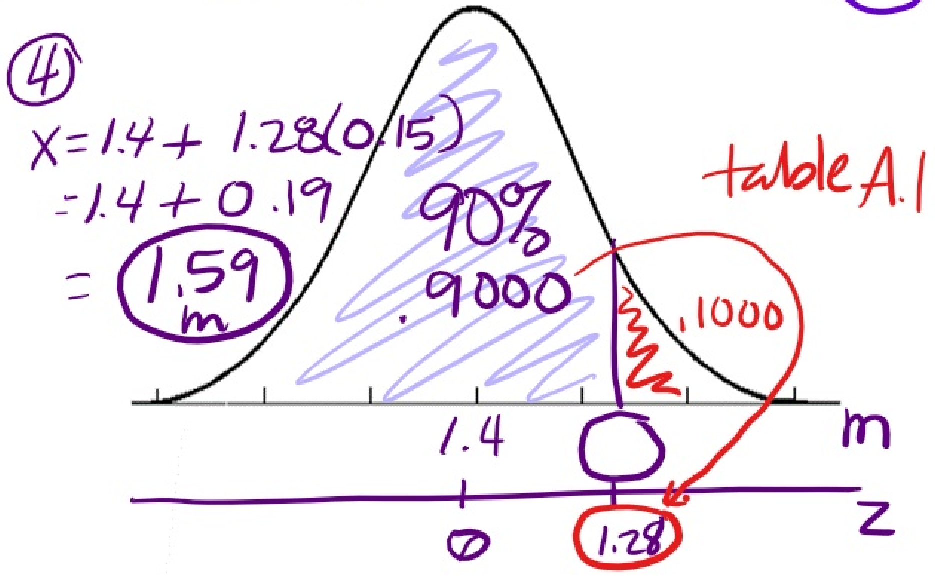
Assume: School's population of students heights are normal (M = 1.4m, SD = 0.15m)

1. The z-score for a student 1.63 m tall = __
2. The height of a student with a z-socre of -2.65 = __
3. The Pecentile Rank of a student that is 1.51 m tall = __
4. The 90th percentile for students heights = __

① $z = \frac{1.63 - 1.40}{0.15} = \frac{0.23}{0.15} = 1.53$

② $1.40 - 2.65(0.15) = 1.40 - .3975 = 1.0$





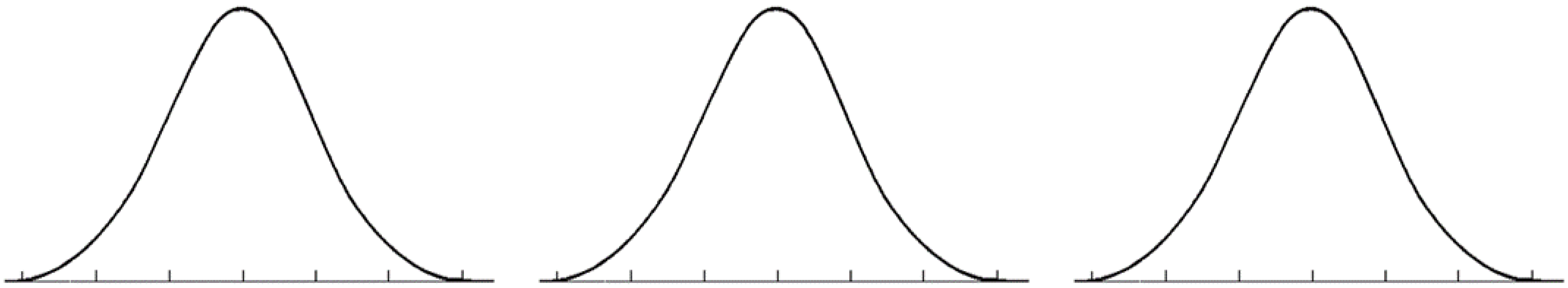
Examples: Find the Probability That...

Assume: School's population of students heights are normal (M = 1.4m, SD = 0.15m)

(1) More than 1.63 m tall

(2) Less than 1.2 m tall

(3) between 1.2 and 1.63 tall



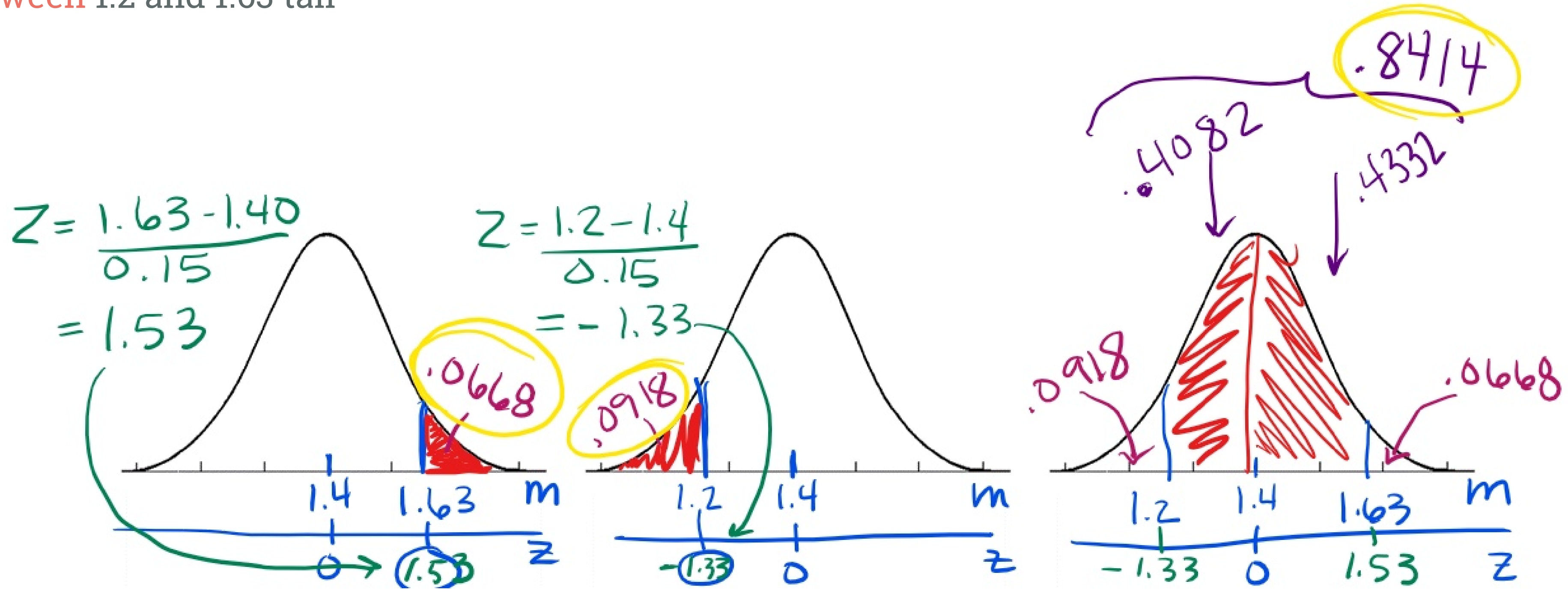
Examples: Find the Probability That...

Assume: School's population of students heights are normal (M = 1.4m, SD = 0.15m)

(1) More than 1.63 m tall

(2) Less than 1.2 m tall

(3) between 1.2 and 1.63 tall

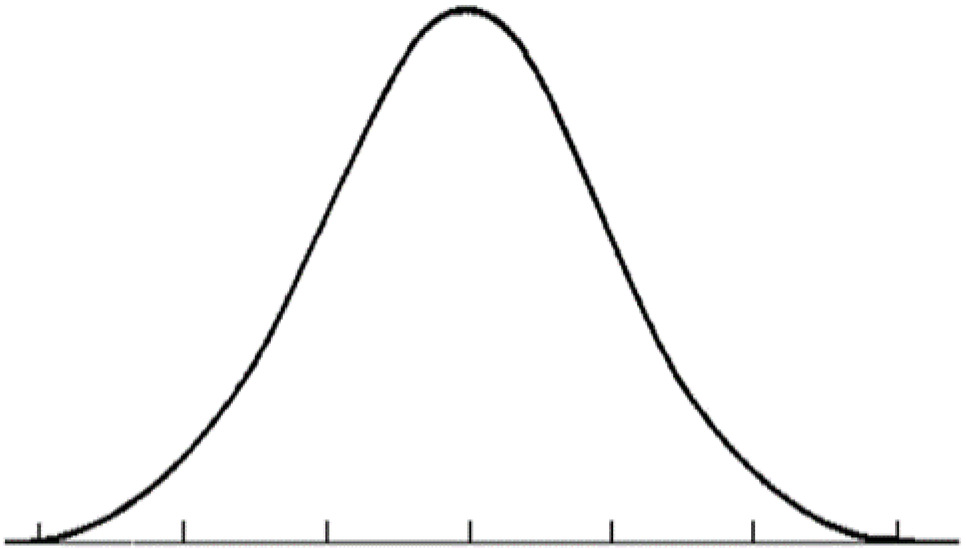


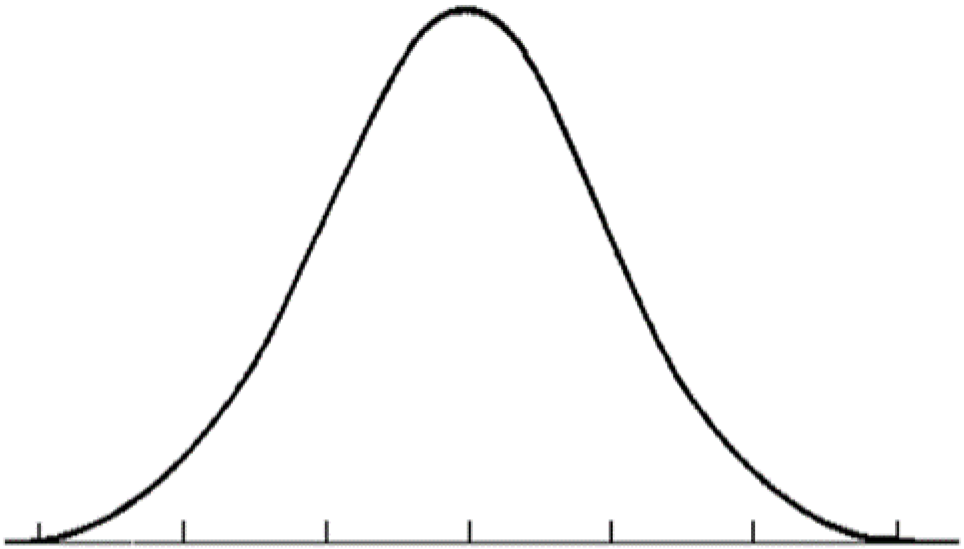
Examples: Percentiles

Assume: School's population of students heights are normal (M = 1.4m, SD = 0.15m)

(1) The perentile rank of a 1.7 m tall Student = __

(2) The height of a studnet in the 15th percentile = __



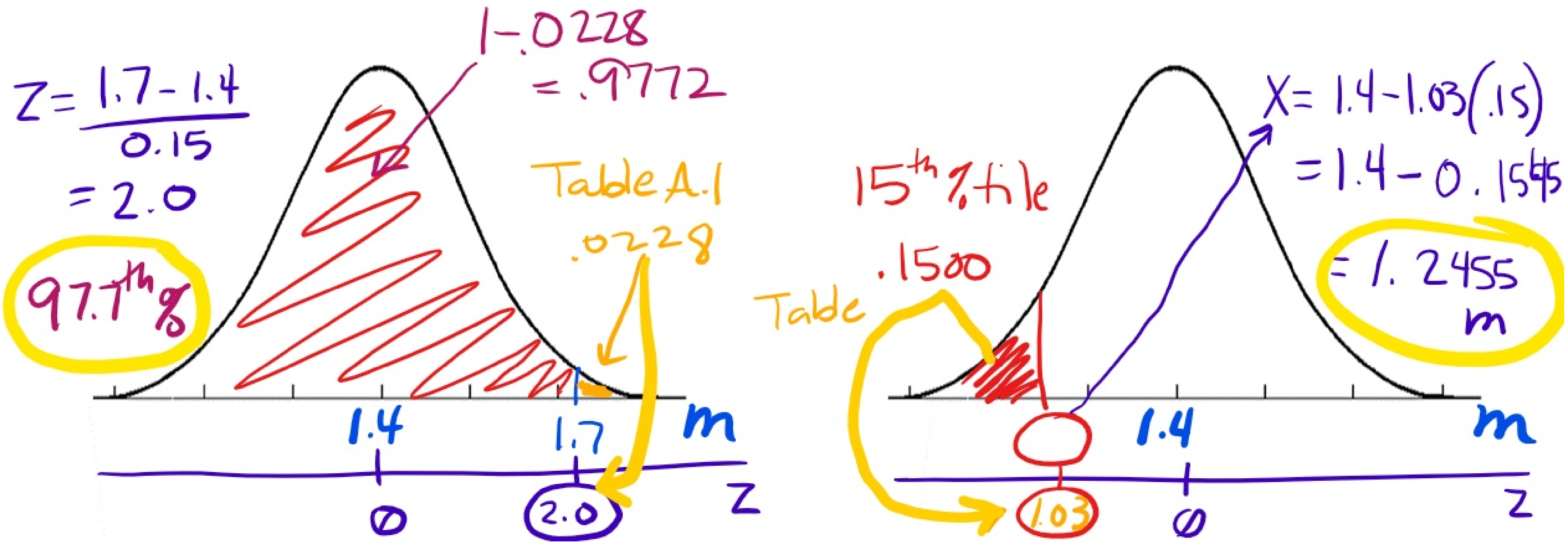


Examples: Percentiles

Assume: School's population of students heights are normal (M = 1.4m, SD = 0.15m)

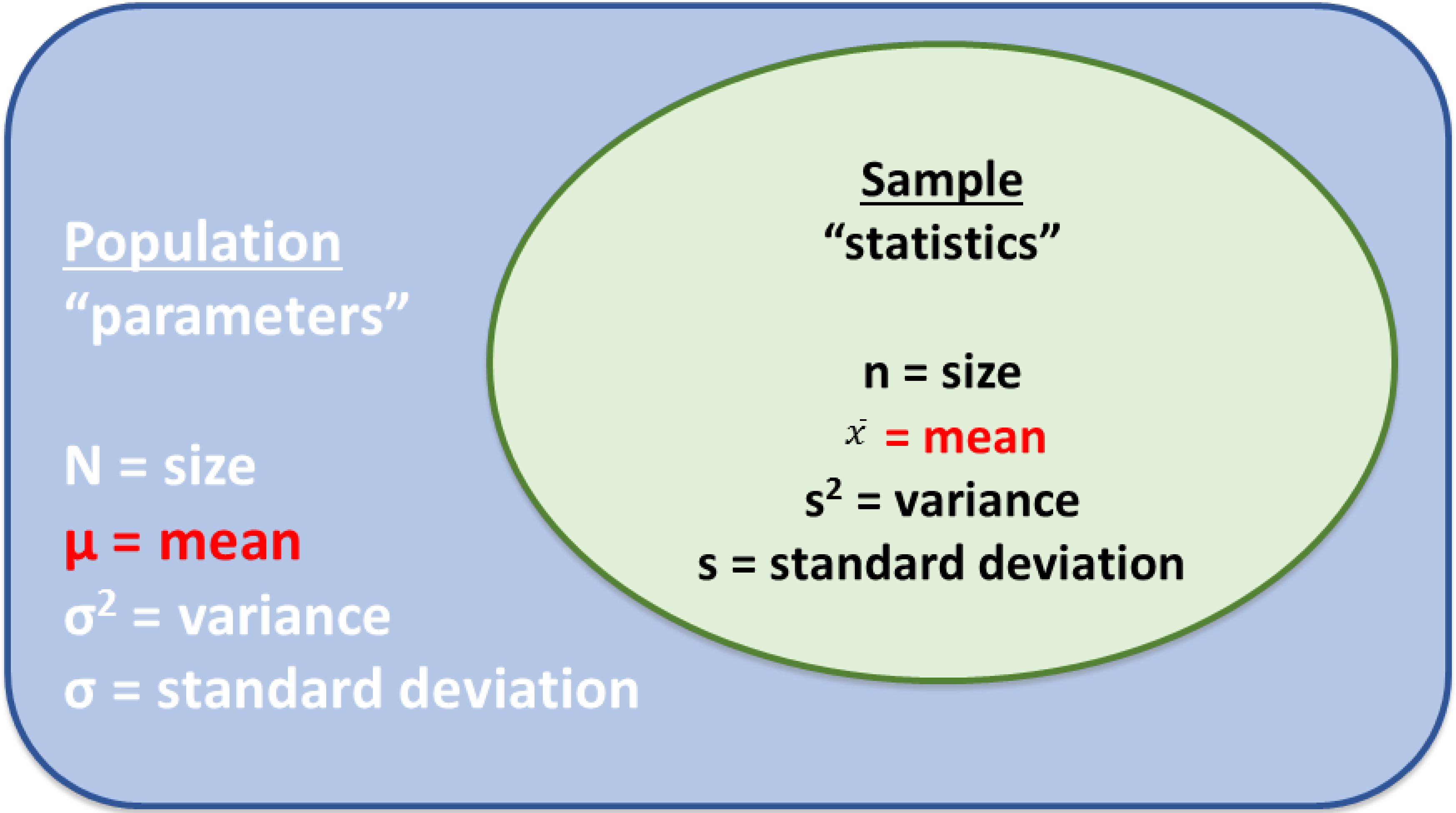
(1) The perentile rank of a 1.7 m tall Student = __

(2) The height of a studnet in the 15th percentile = __



Into Theory Mode Again

Parameters vs. Statistics



Statistical Estimation

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

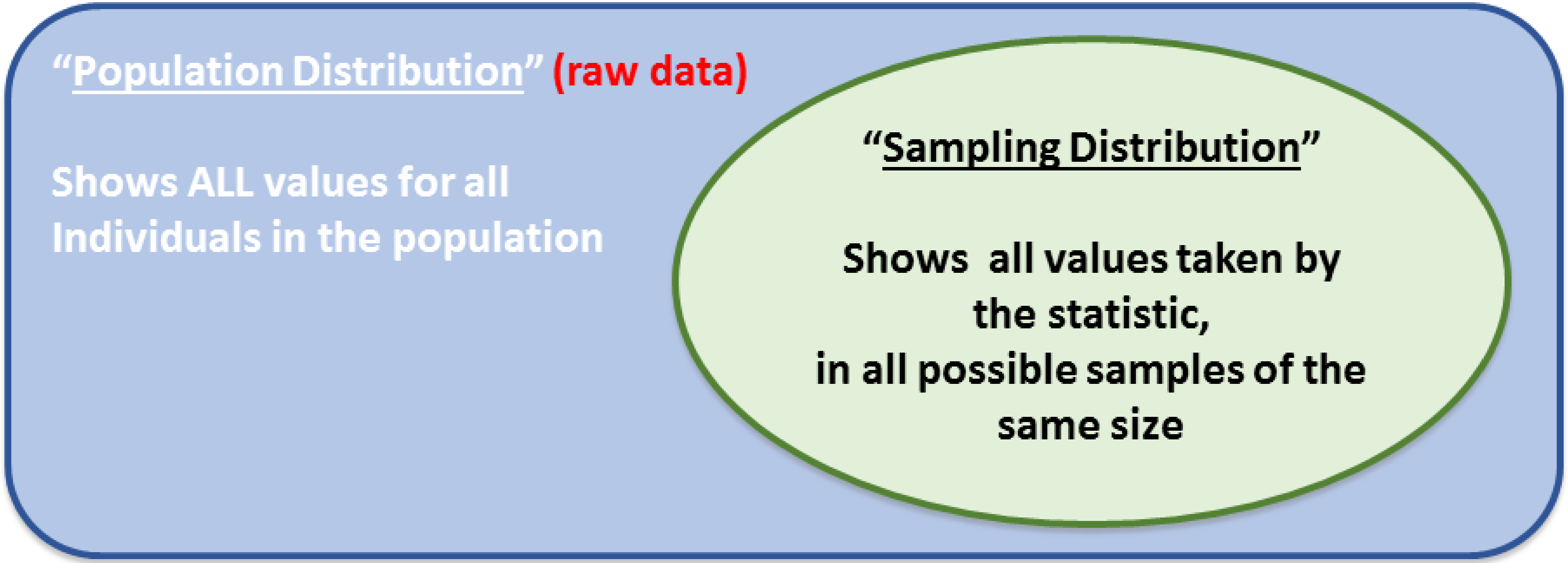
- Different random samples yield different statistics. We need to be able to describe the sampling distribution of possible statistic values in order to perform statistical inference.

- 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 mu.

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.



http://shiny.stat.calpoly.edu/Sampling_Distribution/

Sampling Distribution for the MEAN

The MEAN of a sampling distribution for a sample mean is just as likely to be above or below the population mean, even if the distribution of the raw data is skewed.

The STANDARD DEVIATION of a sampling distribution for a sample mean is is SMALLER than the standard deviation for the population by a factor of the square-root of n.

IF
individual observations
Mean μ_x &
Standard Deviation σ_x

SRS size n

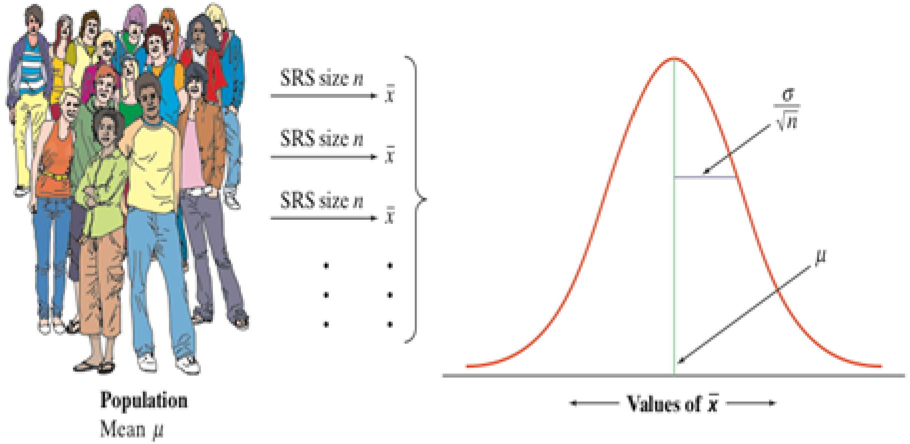
THEN
sample mean
Mean $\mu_{X\text{-}bar} = \mu_X$ &
Standard Deviation $\sigma_{X\text{-}bar} = \sigma_X / \sqrt{n}$

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

Normally Distributed Population

If the population is NORMALLY distributed:

IF
individual observations
have the
N(μ,σ) distribution



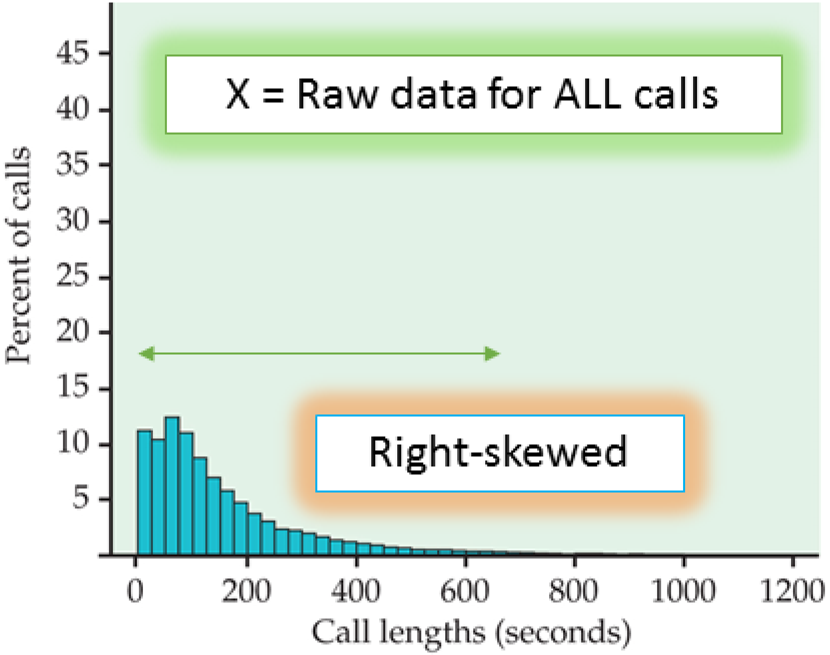
THEN
the **sample mean**
of an SRS of size *n*
has the ***N*(μ, $\sigma/\sqrt{n}$)**
distribution

"SE"
Standard error

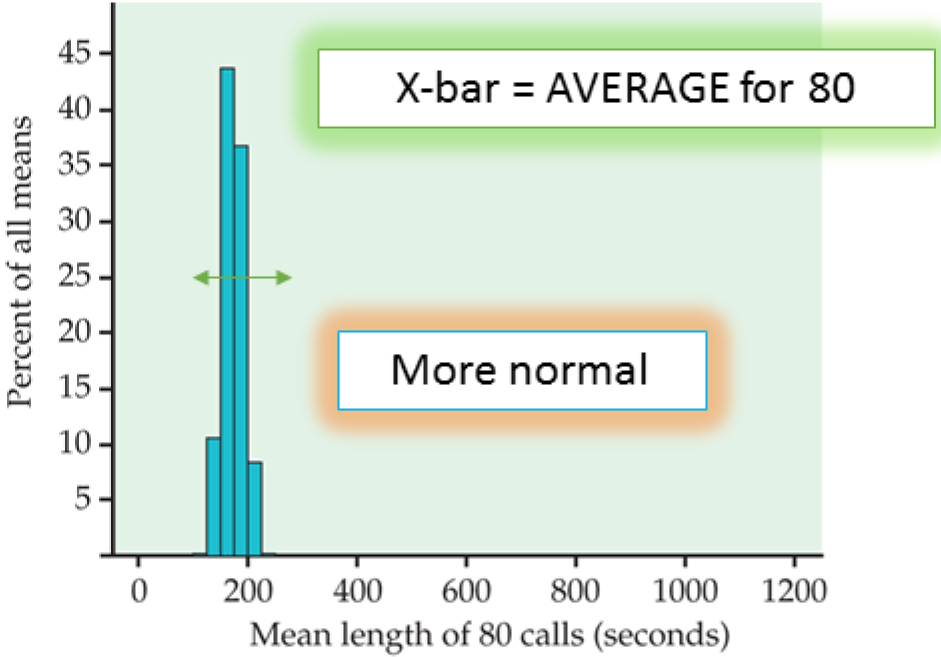
SE for mean = SD divided by square root of the sample size

Skewed Population

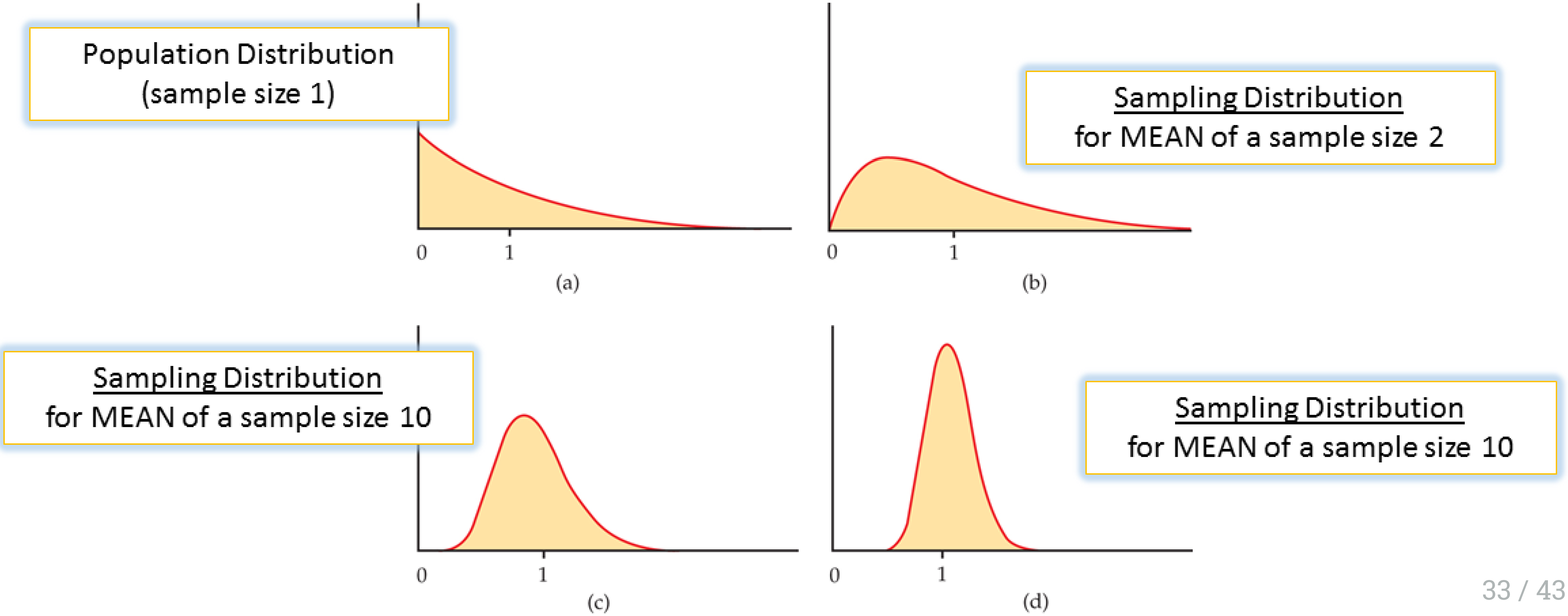
The distribution of lengths of all customer service calls received by a bank in a month.



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



The Central Limit Theorem



The Central Limit Theorem

When a sample size (n) is large, the sampling distribution of the sample MEAN is approximately normally distributed about the mean of the population with the stadard deviation less than than of the population by a factor of the square root of n.

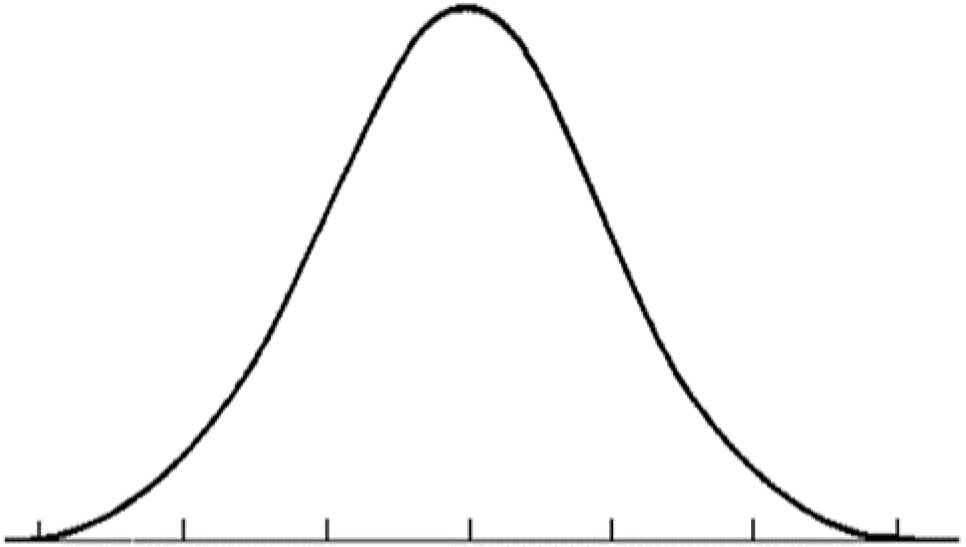
Back to the Example Situation

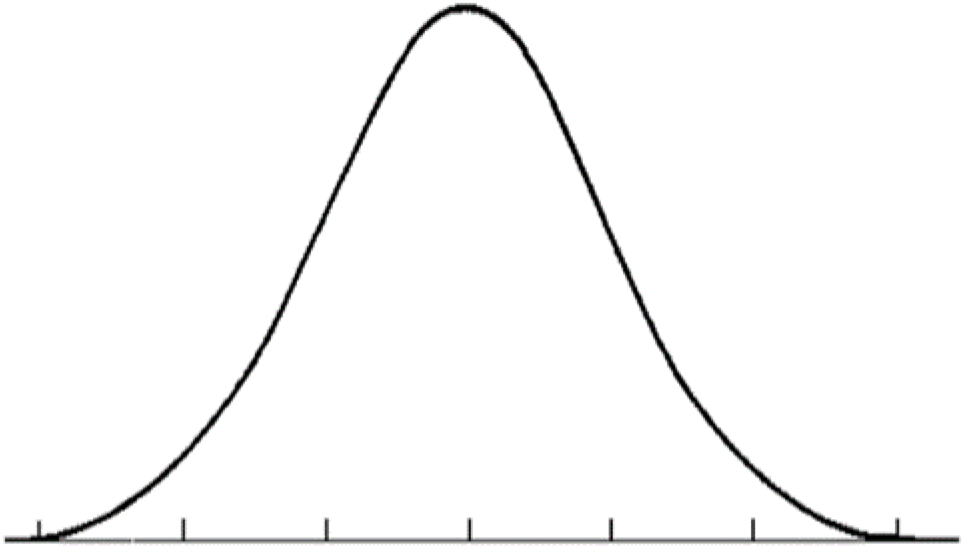
Examples: Probabilities

Assume: School's population of students heights are normal (M = 1.4m, SD = 0.15m)

(1) The probability a randomly selected student is more than 1.63 m tall = __

(2) The probability a randomly selected sample of 16 students average more than 1.63 m tall = __





Examples: Probabilities

Assume: School's population of students heights are normal (M = 1.4m, SD = 0.15m)

(1) The probability a randomly selected student is more than 1.63 m tall = __

(2) The probability a randomly selected sample of 16 students average more than 1.63 m tall = __

- Image needed here

Let's Apply This to the Cancer Dataset

Read in the Data

```r
library(tidyverse)     # Loads several very helpful 'tidy' packages
library(rio)           # Read in SPSS datasets
library(furniture)     # Nice tables (by our own Tyson Barrett)
library(psych)         # Lots of nice tid-bits
```

```r
cancer_raw <- rio::import("cancer.sav")
```

Read in the Data

```r
library(tidyverse)     # Loads several very helpful 'tidy' packages
library(rio)           # Read in SPSS datasets
library(furniture)     # Nice tables (by our own Tyson Barrett)
library(psych)         # Lots of nice tid-bits
```

```r
cancer_raw <- rio::import("cancer.sav")
```

And Clean It

```r
cancer_clean <- cancer_raw %>%
  dplyr::rename_all(tolower) %>%
  dplyr::mutate(id = factor(id)) %>%
  dplyr::mutate(trt = factor(trt,
                             labels = c("Placebo",
                                        "Aloe Juice"))) %>%
  dplyr::mutate(stage = factor(stage))
```

Standardize a variable with `scale()`

```
cancer_clean %>%
  furniture::table1(age)
```

```
      Mean/Count (SD/%)
      n = 25
age
      59.6 (12.9)
```

```
cancer_clean %>%
  dplyr::mutate(agez = (age - 59.6) / 12.9) %
  dplyr::mutate(ageZ = scale(age))%>%
  dplyr::select(id, trt, age, agez, ageZ) %>%
  head()
```

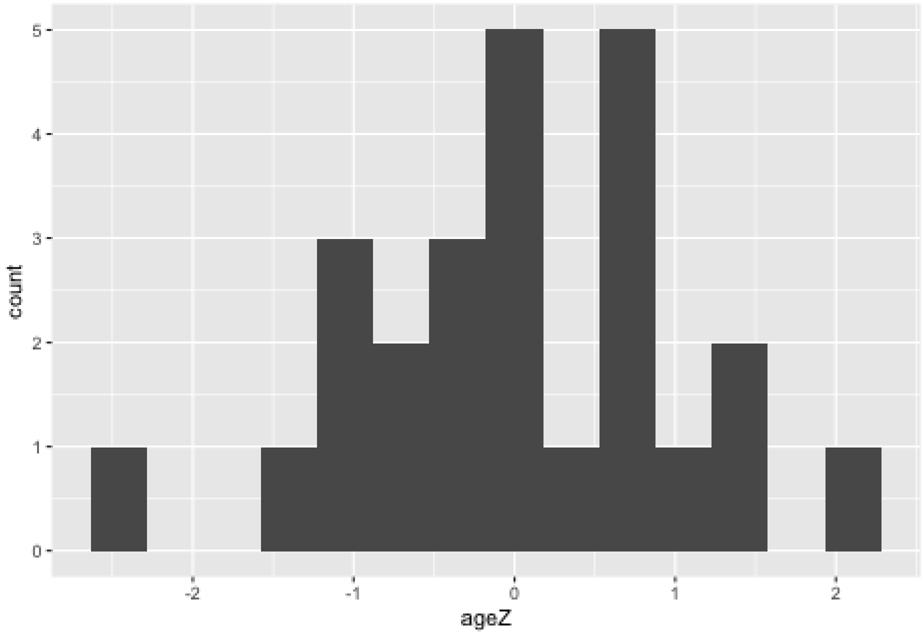
```
# A tibble: 6 x 5
  id    trt       age    agez ageZ[,1]
  <fct> <fct>   <dbl>   <dbl>    <dbl>
1 1     Placebo    52 -0.589   -0.591
2 5     Placebo    77  1.35     1.34
3 6     Placebo    60  0.0310   0.0278
4 9     Placebo    61  0.109    0.105
5 11    Placebo    59 -0.0465  -0.0495
6 15    Placebo    69  0.729    0.724
```

Standardize a variable - not normal

```
cancer_clean %>%
  dplyr::mutate(ageZ = scale(age)) %>%
  furniture::table1(age, ageZ)
```

```
      Mean/Count (SD/%)
      n = 25
age
      59.6 (12.9)
ageZ
      -0.0 (1.0)
```

```
cancer_clean %>%
  dplyr::mutate(ageZ = scale(age)) %>%
  ggplot(aes(ageZ)) +
  geom_histogram(bins = 14)
```



Questions?

Next Topic

Intro to Hypothesis Testing: 1 Sample z-test