Normal distribution

Slides developed by Mine Çetinkaya-Rundel of OpenIntro, modified by Leah Dorazio for use with AHSS.

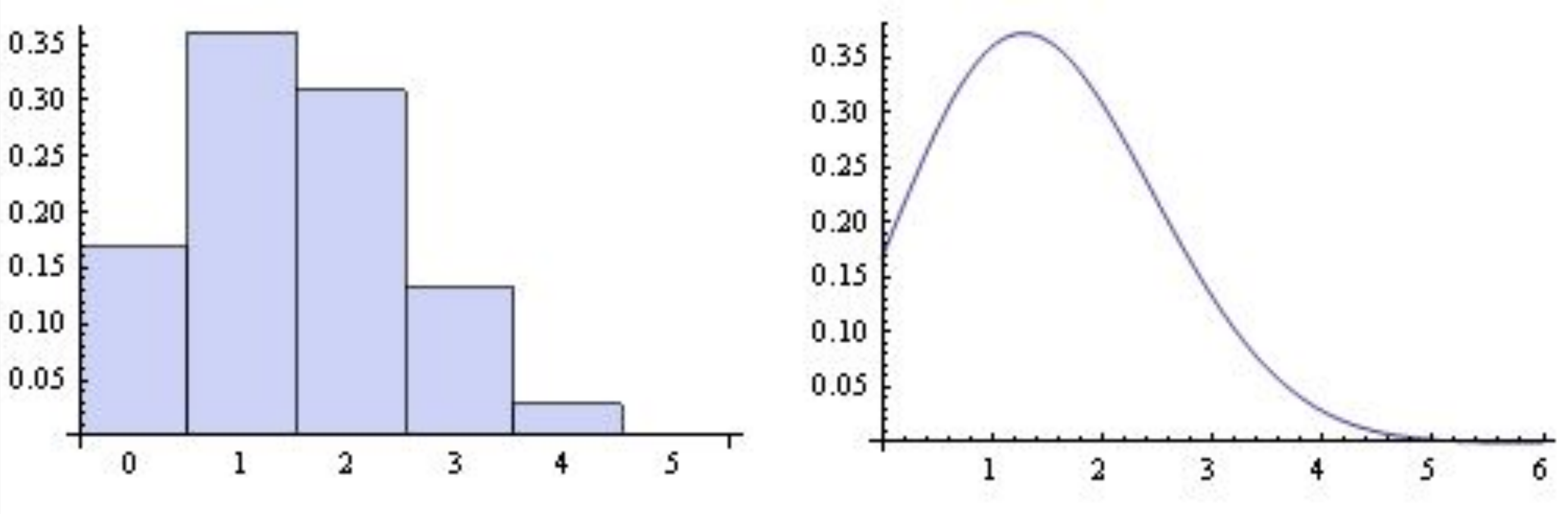
From discrete to continuous...

Discrete

- sum of probabilities must = 1

Continuous

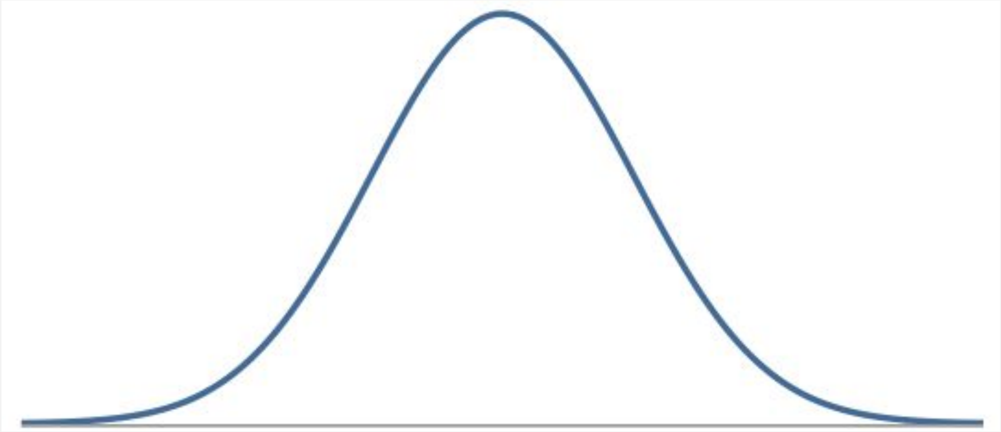
- total *area* must = 1
- probability of a specific value = 0, e.g. $P(X = 2) = 0$
- only intervals have probability, e.g. $P(1 < X < 2) = ?$



The Normal distribution...

- is the most well known continuous distribution
- Is unimodal and symmetric, bell shaped curve
- has mean μ and standard deviation σ
- has tails that extend infinitely in both directions

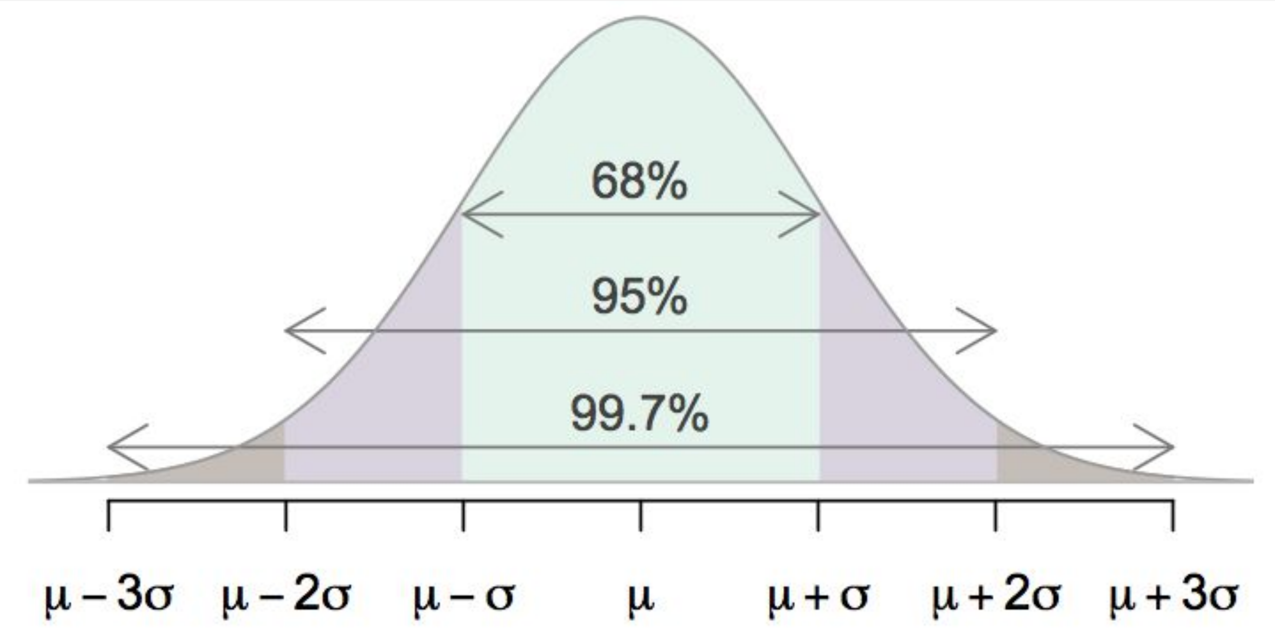
Many variables are nearly normal, but none are *exactly* normal



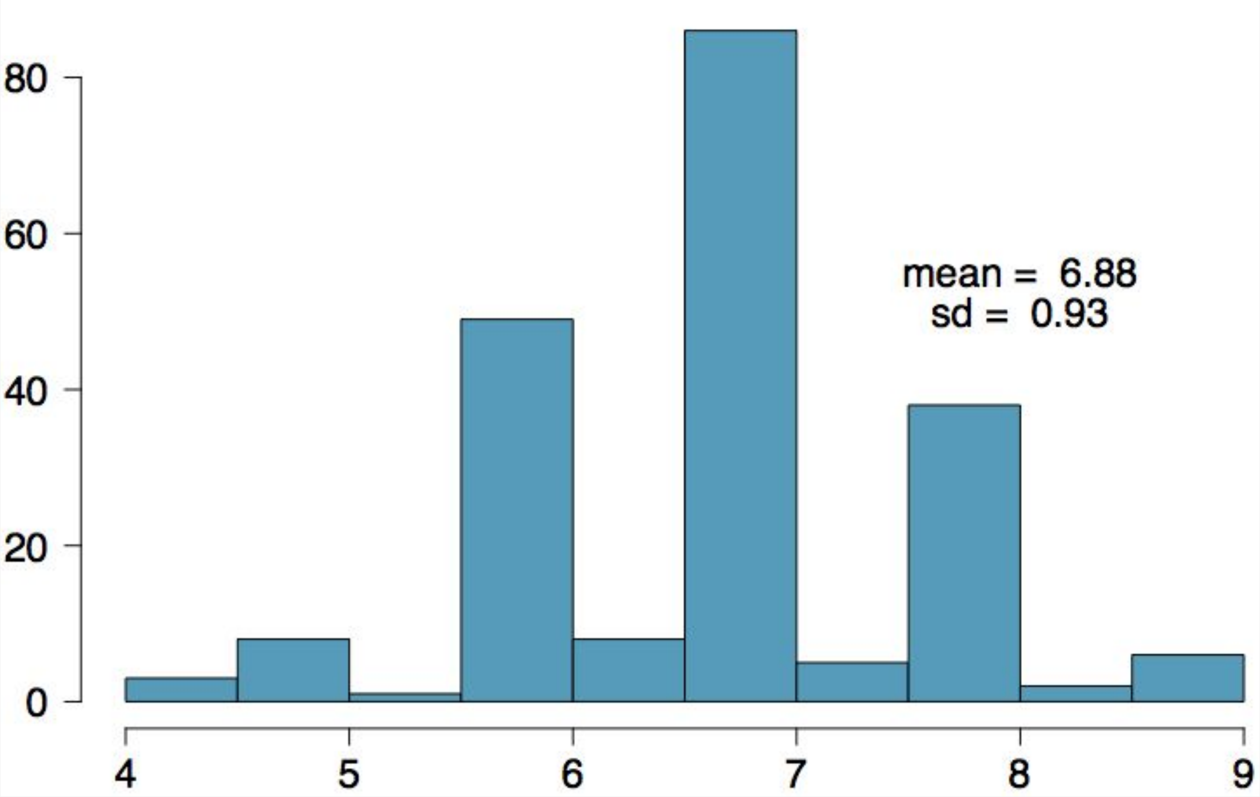
The source of the 68-95-99.7 Rule

For nearly normally distributed data,

- about 68% falls within 1 SD of the mean,
- about 95% falls within 2 SDs of the mean,
- about 99.7% falls within 3 SDs of the mean.

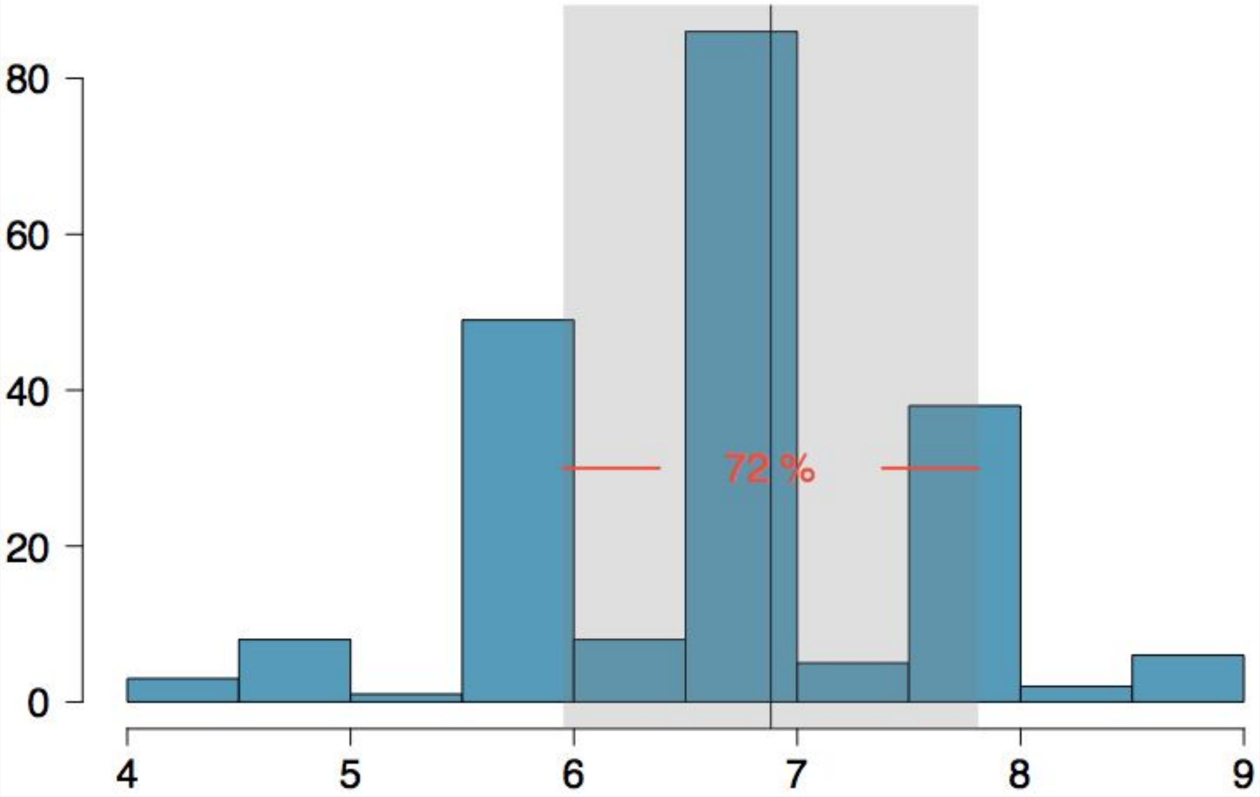


Number of hours of sleep on school nights



Mean = 6.88 hours, SD = 0.92 hrs

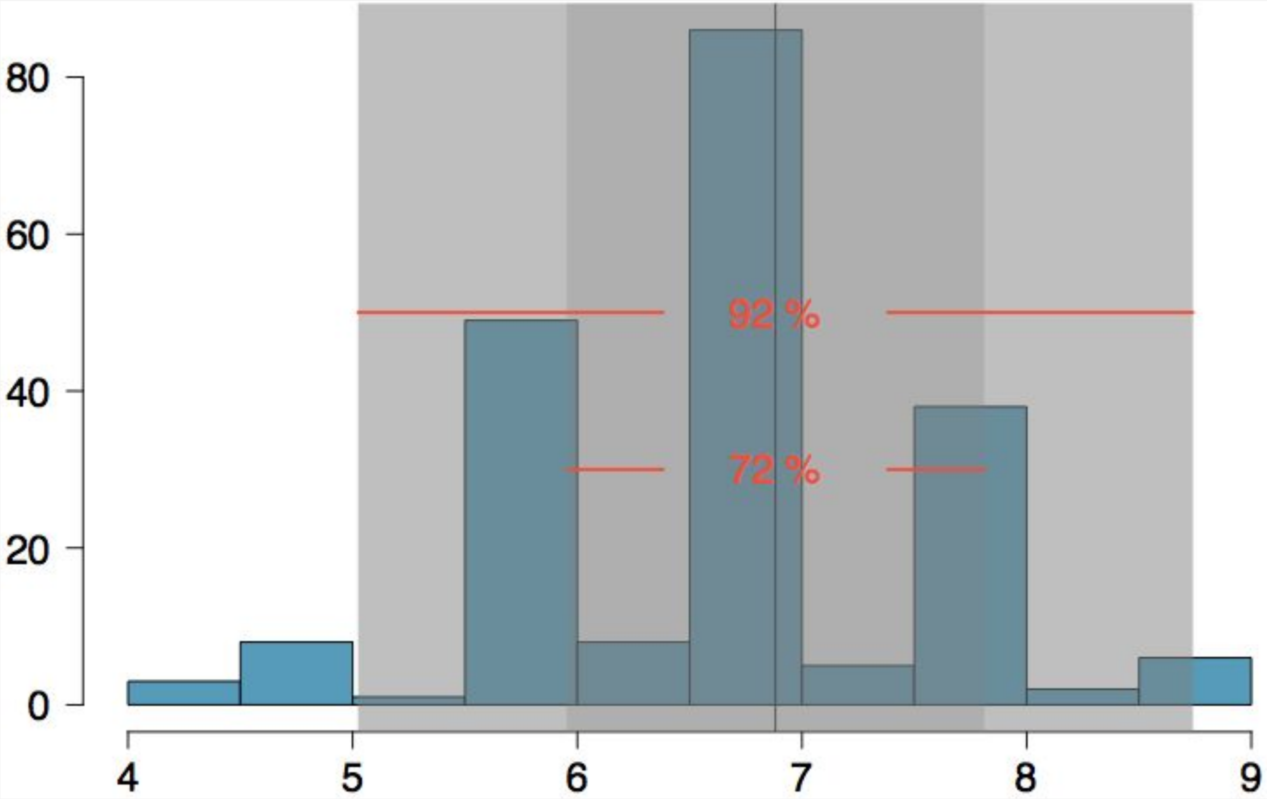
Number of hours of sleep on school nights



Mean = 6.88 hours, SD = 0.92 hrs

72% of the data are within 1 SD of the mean: 6.88 ± 0.93

Number of hours of sleep on school nights

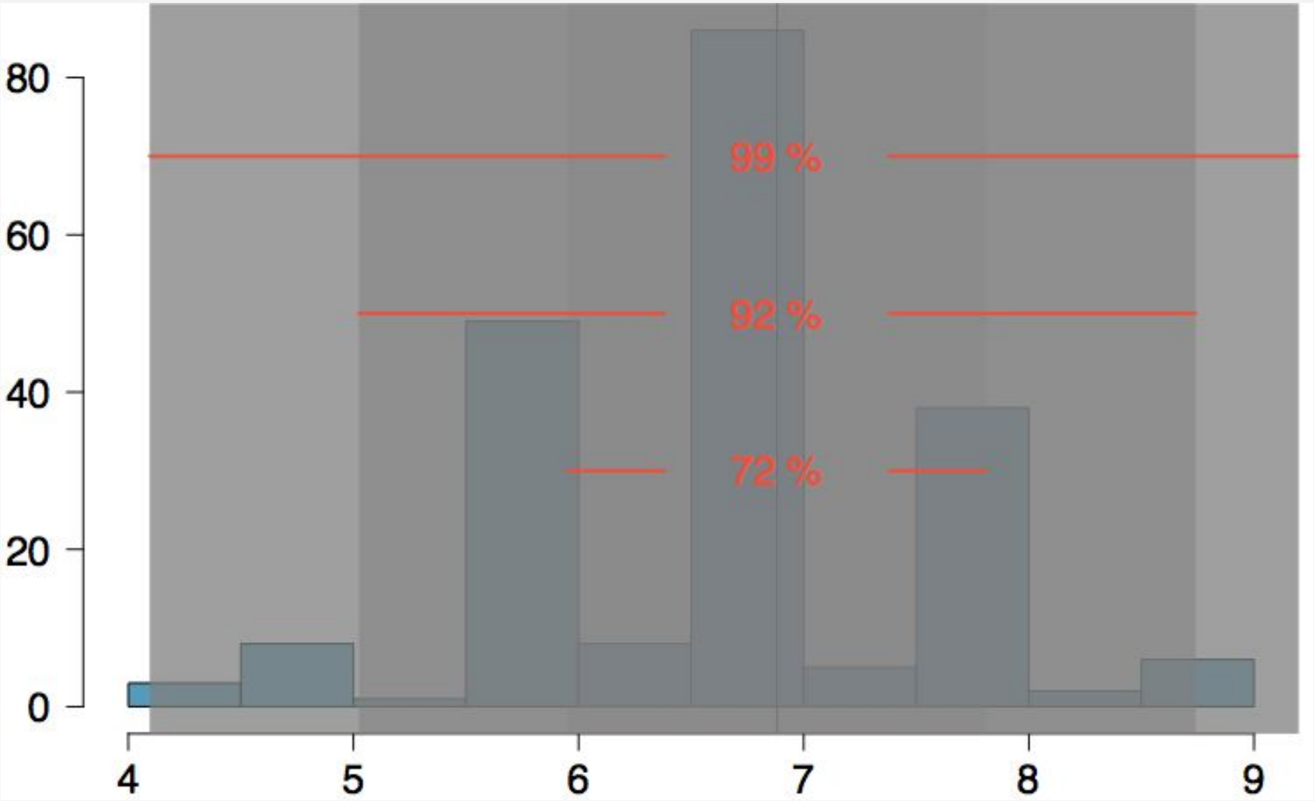


Mean = 6.88 hours, SD = 0.92 hrs

72% of the data are within 1 SD of the mean: 6.88 ± 0.93

92% of the data are within 1 SD of the mean: 6.88 ± 2 x 0.93

Number of hours of sleep on school nights



Mean = 6.88 hours, SD = 0.92 hrs

72% of the data are within 1 SD of the mean: 6.88 ± 0.93

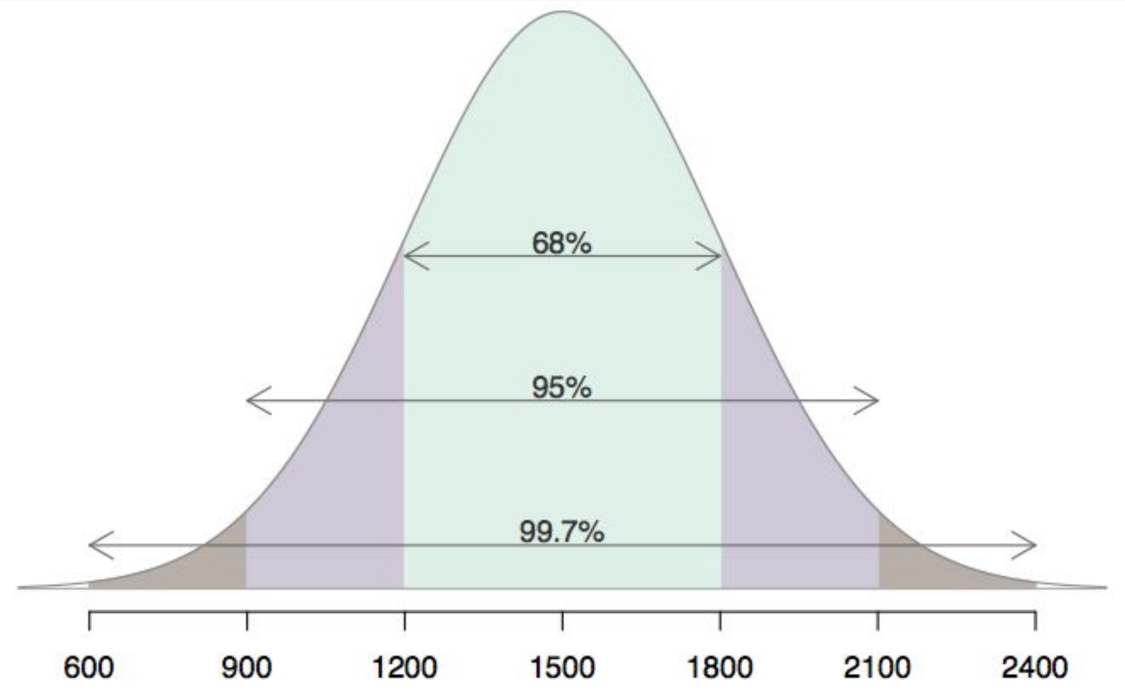
92% of the data are within 1 SD of the mean: 6.88 ± 2 x 0.93

99% of the data are within 1 SD of the mean: 6.88 ± 3 x 0.93

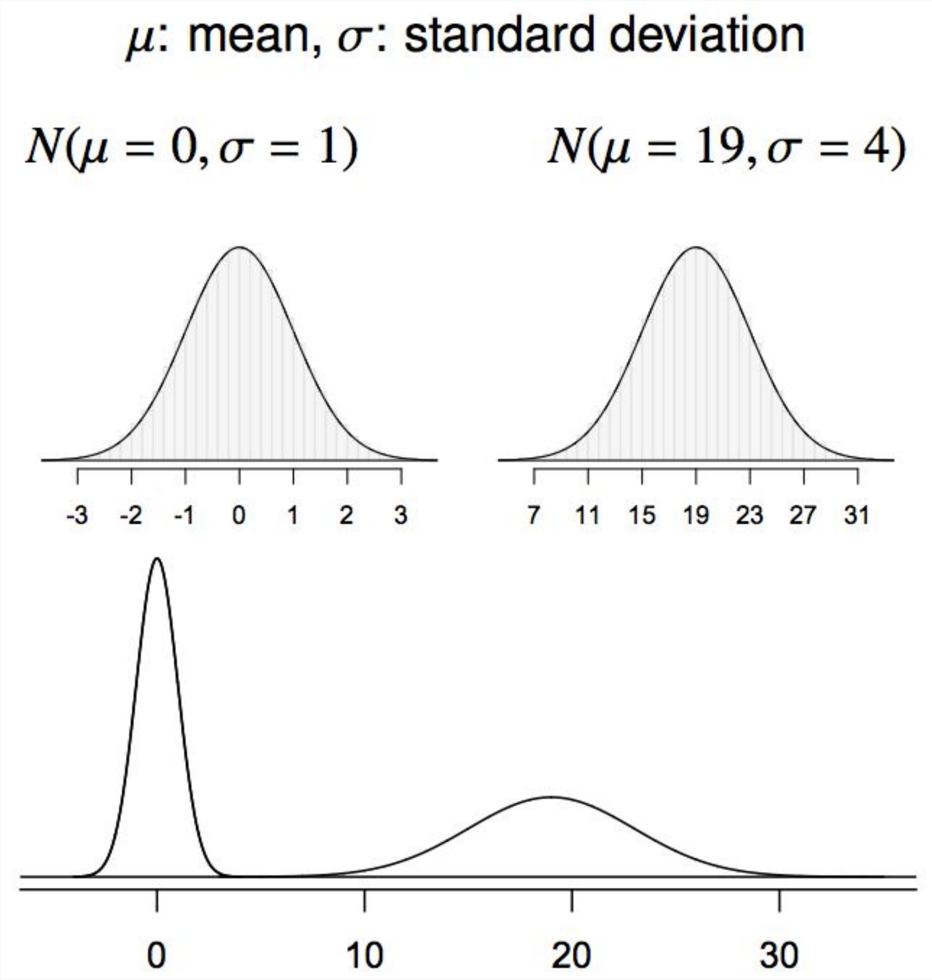
Describing variability using the 68-95-99.7 Rule

SAT scores are distributed nearly normally with mean 1500 and standard deviation 300.

- ~68% of students score between 1200 and 1800 on the SAT.
- ~95% of students score between 900 and 2100 on the SAT.
- ~99.7% of students score between 600 and 2400 on the SAT.



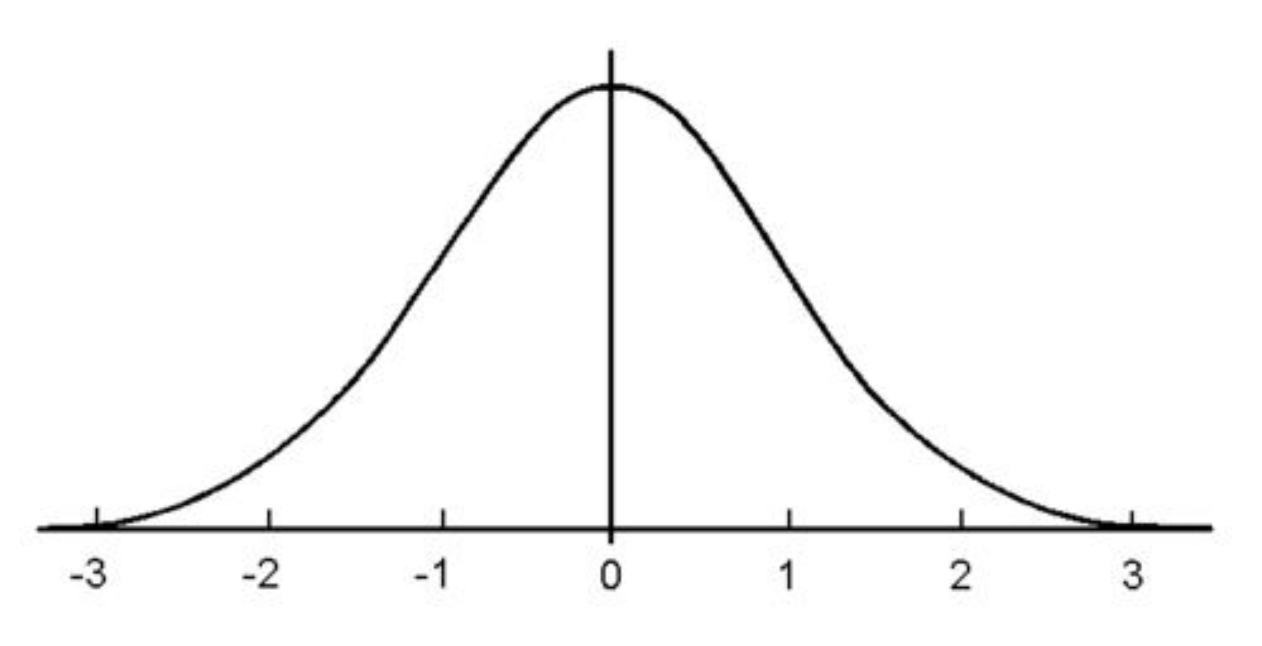
Normal distributions with different parameters



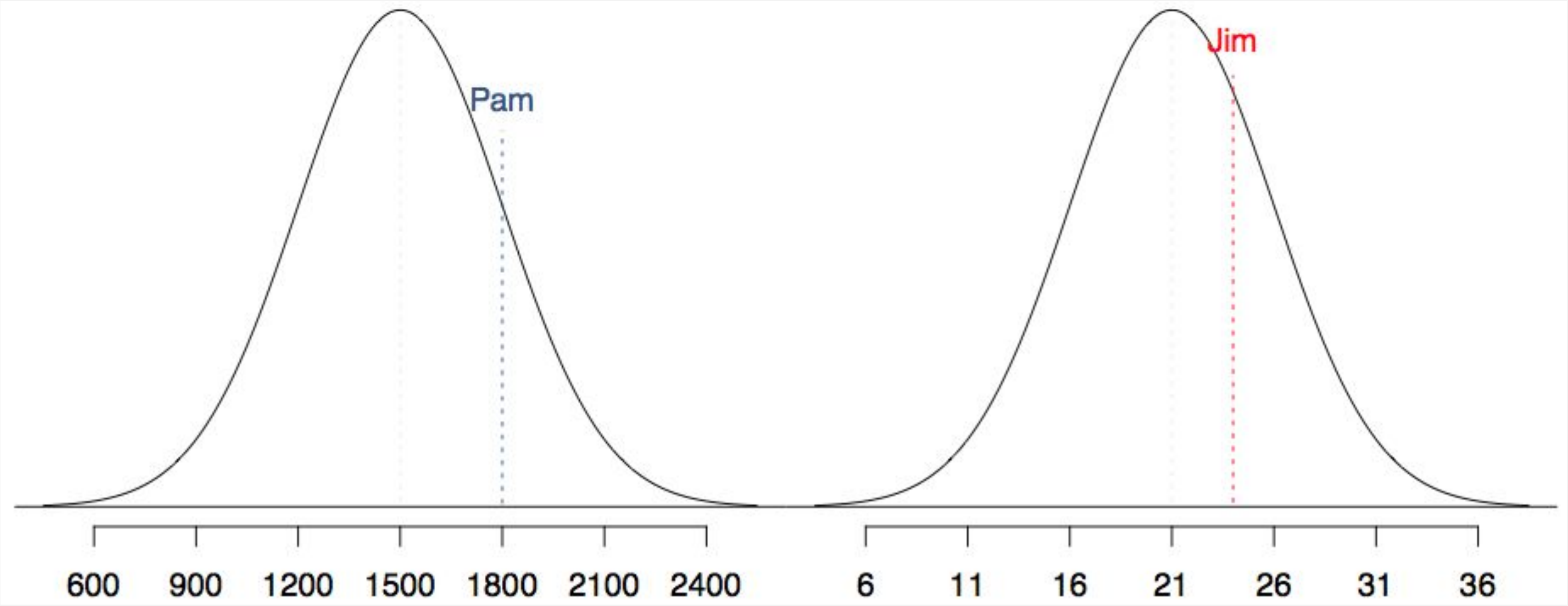
The Standard Normal Curve

What units are on the horizontal axis?

- Z-scores!



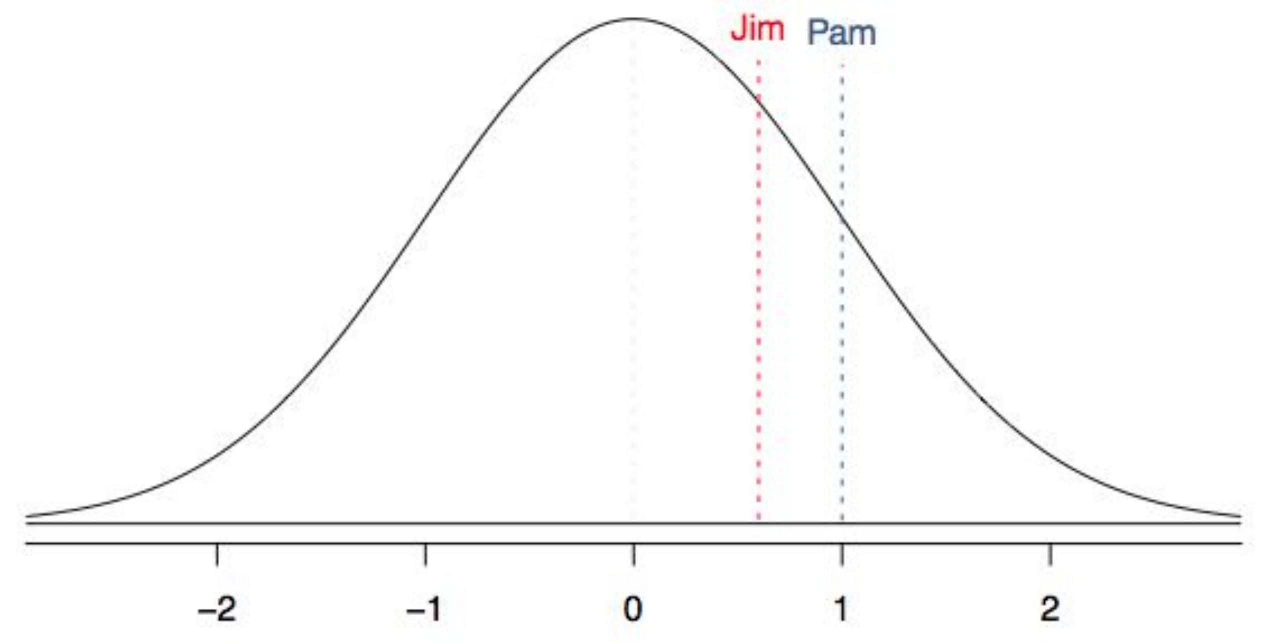
SAT scores are distributed nearly normally with mean 1500 and standard deviation 300. ACT scores are distributed nearly normally with mean 21 and standard deviation 5. A college admissions officer wants to determine which of the two applicants scored better on their standardized test with respect to the other test takers: Pam, who earned an 1800 on her SAT, or Jim, who scored a 24 on his ACT?



Standardizing with Z-scores

Since we cannot just compare these two raw scores, we instead compare their Z-scores, that is, how many SDs beyond the mean their raw score is.

- Pam's Z-score = (1800 - 1500) / 300 = 1. She is 1 standard deviation above the mean.
- Jim's Z-score = (24 - 21) / 5 = 0.6. He is 0.6 standard deviations above the mean.
- Therefore, Pam did better.



Standardizing with Z scores (cont.)

These are called standardized scores, or Z scores.

- Z score of an observation is the *number of standard deviations* it falls above or below the mean.

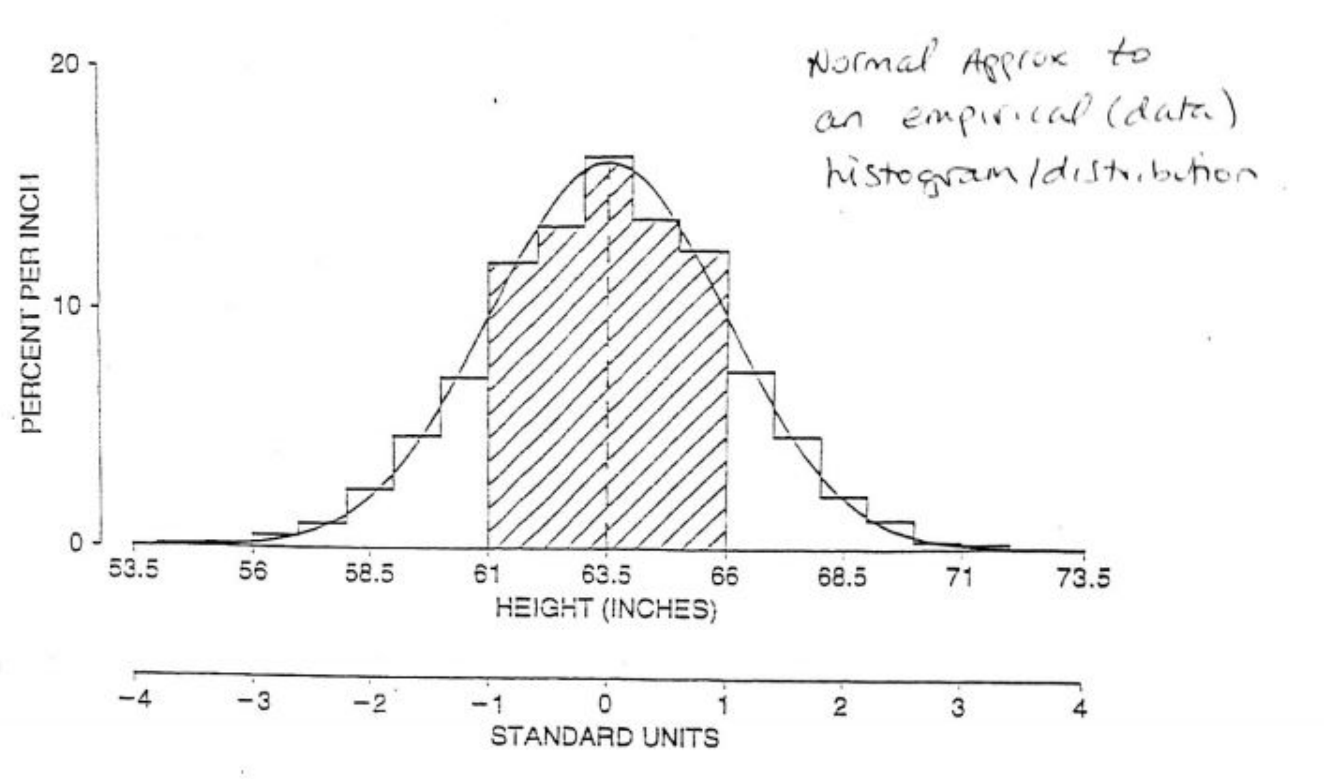
$$Z = \frac{\text{(observation - mean)}}{\text{SD}}$$

- Z scores are defined for distributions of any shape, but only when the distribution is normal can we use Z scores to calculate percentiles.
- Observations that are more than 2 SD away from the mean ($|Z| > 2$) are generally considered unusual.

Normal approximation for data

Step 1. Convert *x* value to Z-score

Step 2: Use calculator to find corresponding area under standard normal curve



Finding areas under the standard normal curve

$P(Z < -1) =$

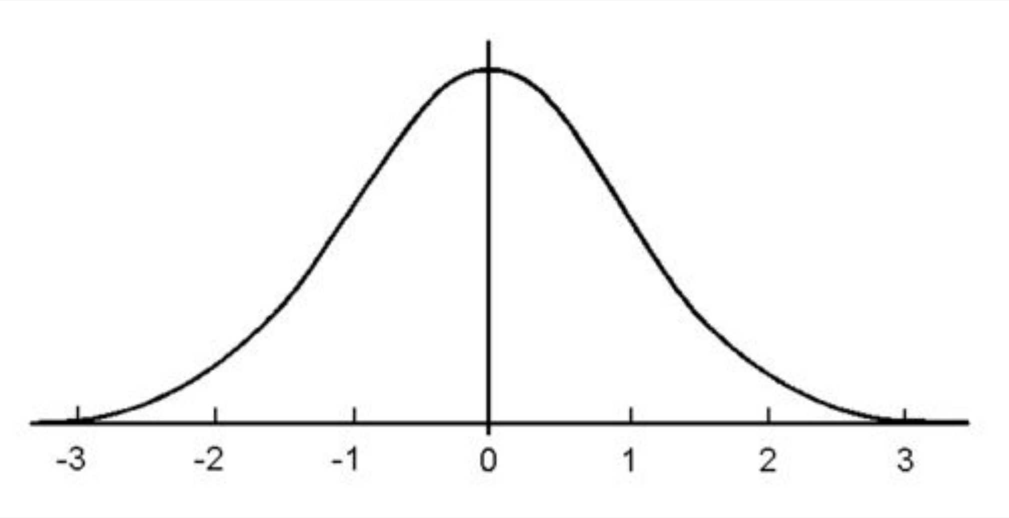
$P(Z < -1.5) =$

$P(Z > 2.1) =$

$P(-1.2 < Z < 2.1) =$

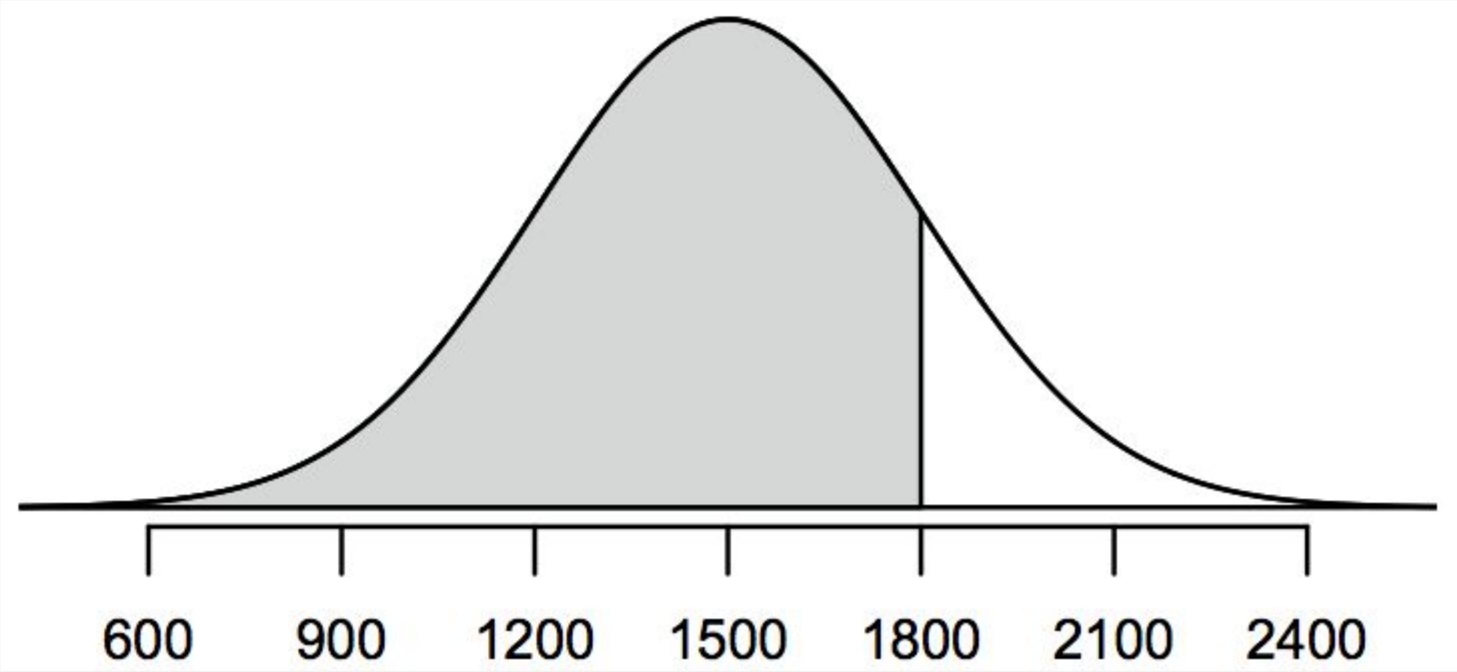
2nd VARS, 2:normcdf(
lower: ?
upper: ?
Paste

TI-83: do 2:normcdf(lower, upper)



Percentiles

- Percentile is the percentage of observations that fall below a given data point.
- Graphically, percentile is the area below the probability distribution curve to the left of that observation.



Finding percentiles from the standard normal curve

What Z-score corresponds to the 50th percentile?

i.e. $P(Z < ?) = 0.5$ Z =

What Z-score corresponds to the 20th percentile?

i.e. $P(Z < ?) = 0.2$ Z =

What Z-score has 70% of the area to the *right* of it?

i.e. $P(Z < ?) =$ **0.3** Z =

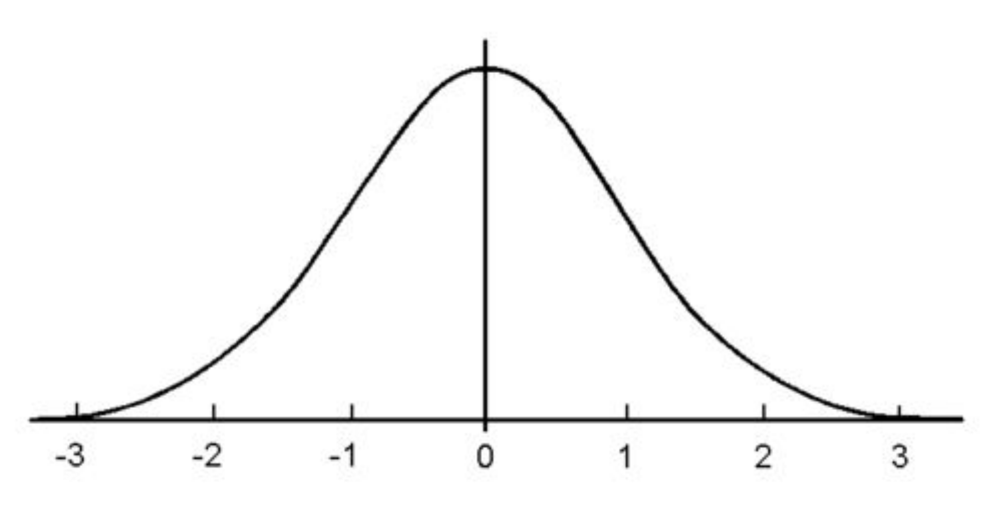
2nd VARS: 3: invNorm(

area = ? (enter percentile as a decimal)

Paste

TI-83 do:

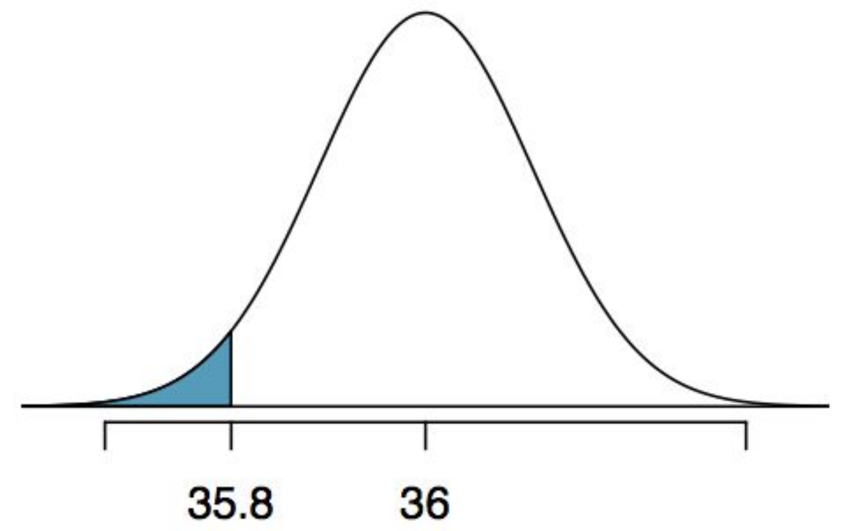
3:invNorm(percentile as a decimal)



Quality control

At Heinz ketchup factory the amounts which go into bottles of ketchup are supposed to be normally distributed with mean 36 oz. and standard deviation 0.11 oz. If the amount of ketchup in the bottle is below 35.8 oz. then the bottle fails the quality control inspection. What is the probability that a randomly selected bottle fails quality control (that is, what percent of bottles fail quality control)?

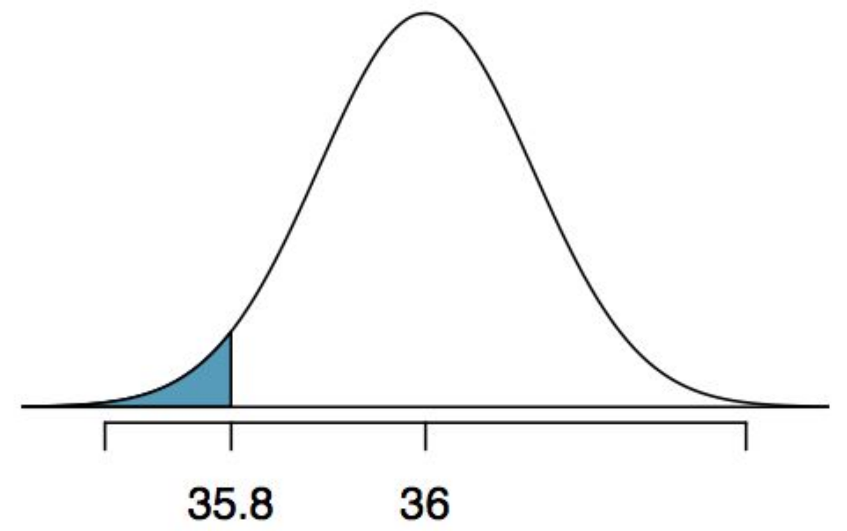
- *Let X = amount of ketchup in a bottle:* $\mu = 36$, $\sigma = 0.11$. $P(X < 35.8) = ?$



Quality control

At Heinz ketchup factory the amounts which go into bottles of ketchup are supposed to be normally distributed with mean 36 oz. and standard deviation 0.11 oz. If the amount of ketchup in the bottle is below 35.8 oz. then the bottle fails the quality control inspection. What is the probability that a randomly selected bottle fails quality control (that is, what percent of bottles fail quality control)?

- *Let X = amount of ketchup in a bottle: μ = 36, σ = 0.11. P(X < 35.8) = ?*



$$Z = \frac{35.8 - 36}{0.11} = -1.82$$

$P(Z < -1.82) = 0.034 = 3.4\%$

Practice

If the amount of ketchup in the bottle is below 35.8 oz. or above 36.2 oz., then the bottle *fails* the quality control inspection. Recall $\mu = 36, \sigma = 0.11.$

What percent of bottles pass the quality control inspection?

(a) 1.82%

(b) 3.44%

(c) 6.88%

(d) 93.12%

(e) 96.56%

Practice

If the amount of ketchup in the bottle is below 35.8 oz. or above 36.2 oz., then the bottle *fails* the quality control inspection. Recall $\mu = 36, \sigma = 0.11$.

What percent of bottles pass the quality control inspection?

(a) 1.82%

(b) 3.44%

(c) 6.88%

(d) 93.12%

(e) 96.56%

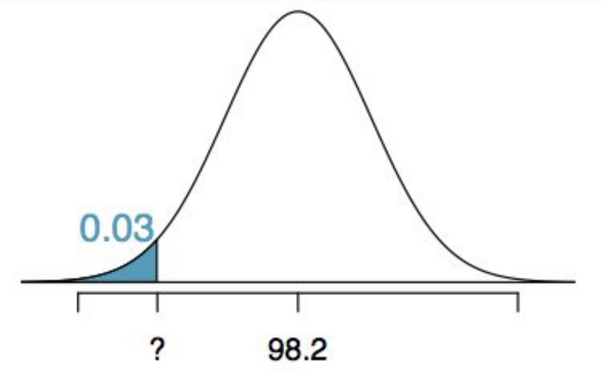
$$Z_{35.8} = \frac{35.8 - 36}{0.11} = -1.82$$

$$Z_{36.2} = \frac{36.2 - 36}{0.11} = 1.82$$

$$P(35.8 < X < 36.2) = P(-1.82 < Z < 1.82) = 0.9312$$

Finding cutoff points

Body temperatures of healthy humans are distributed nearly normally with mean 98.2ºF and standard deviation 0.73ºF. What is the cutoff for the lowest 3% of human body temperatures?



Note: On the TI you can use invNorm. Enter the percentile as a decimal and it will return the corresponding Z-score.

Z=invNorm(0.03)= -1.88

$$Z = \frac{obs - mean}{SD} \rightarrow \frac{x - 98.2}{0.73} = -1.88$$

$$x = (-1.88 \times 0.73) + 98.2 = 96.8^\circ F$$

Mackowiak, Wasserman, and Levine (1992), A Critical Appraisal of 98.6 Degrees F, the Upper Limit of the Normal Body Temperature, and Other Legacies of Carl Reinhold August Wunderlick.

Practice

Body temperatures of healthy humans are distributed nearly normally with mean 98.2°F and standard deviation 0.73°F. What is the cutoff for the *highest* 10% of human body temperatures?

(a) 97.3°F

(b) 99.1°F

(c) 99.4°F

(d) 99.6°F

Practice

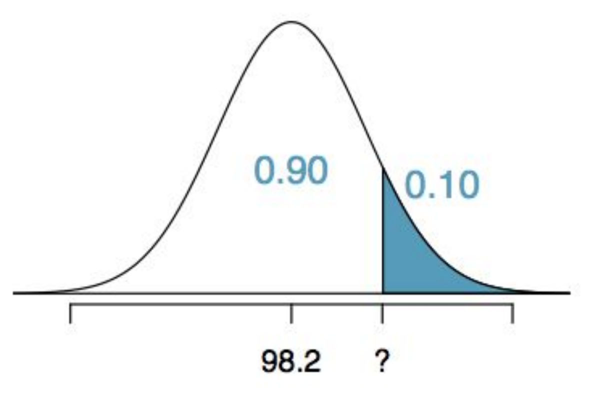
Body temperatures of healthy humans are distributed nearly normally with mean 98.2°F and standard deviation 0.73°F. What is the cutoff for the *highest* 10% of human body temperatures?

(a) 97.3°F

(b) 99.1°F

(c) 99.4°F

(d) 99.6°F



We are looking for the Z score that corresponds to the *90th percentile*. In this case Z = invNorm(0.9) = 1.28.

$$Z = \frac{obs - mean}{SD} \rightarrow \frac{x - 98.2}{0.73} = 1.28$$

$$x = (1.28 \times 0.73) + 98.2 = 99.1$$

Practice

Which of the following is false?

1. Majority of Z scores in a right skewed distribution are negative.
2. In skewed distributions the Z score of the mean might be different than 0.
3. For a normal distribution, IQR is less than 2 x SD.
4. Z scores are helpful for determining how unusual a data point is compared to the rest of the data in the distribution.

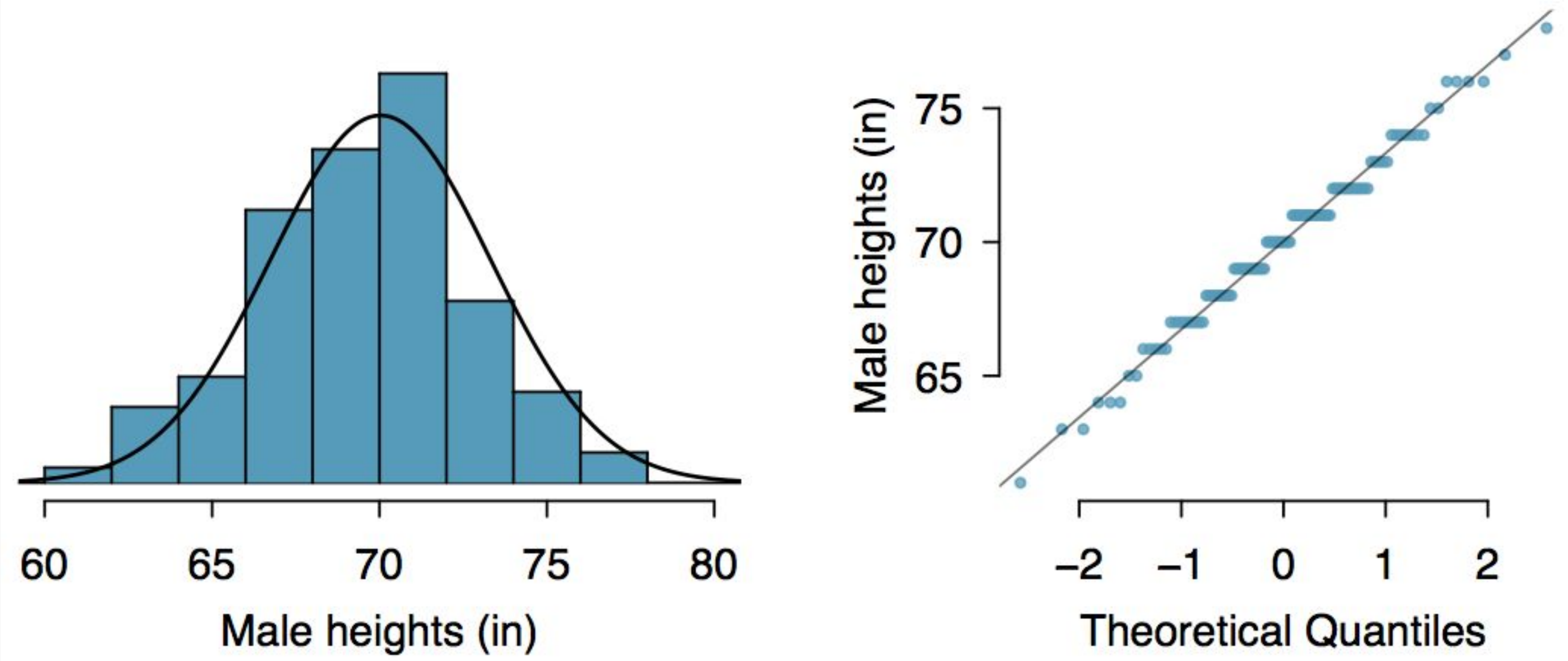
Practice

Which of the following is false?

1. Majority of Z scores in a right skewed distribution are negative.
2. *In skewed distributions the Z score of the mean might be different than 0.*
3. For a normal distribution, IQR is less than 2 x SD.
4. Z scores are helpful for determining how unusual a data point is compared to the rest of the data in the distribution.

Normal probability plot

A histogram and normal probability plot of a sample of 100 male heights.

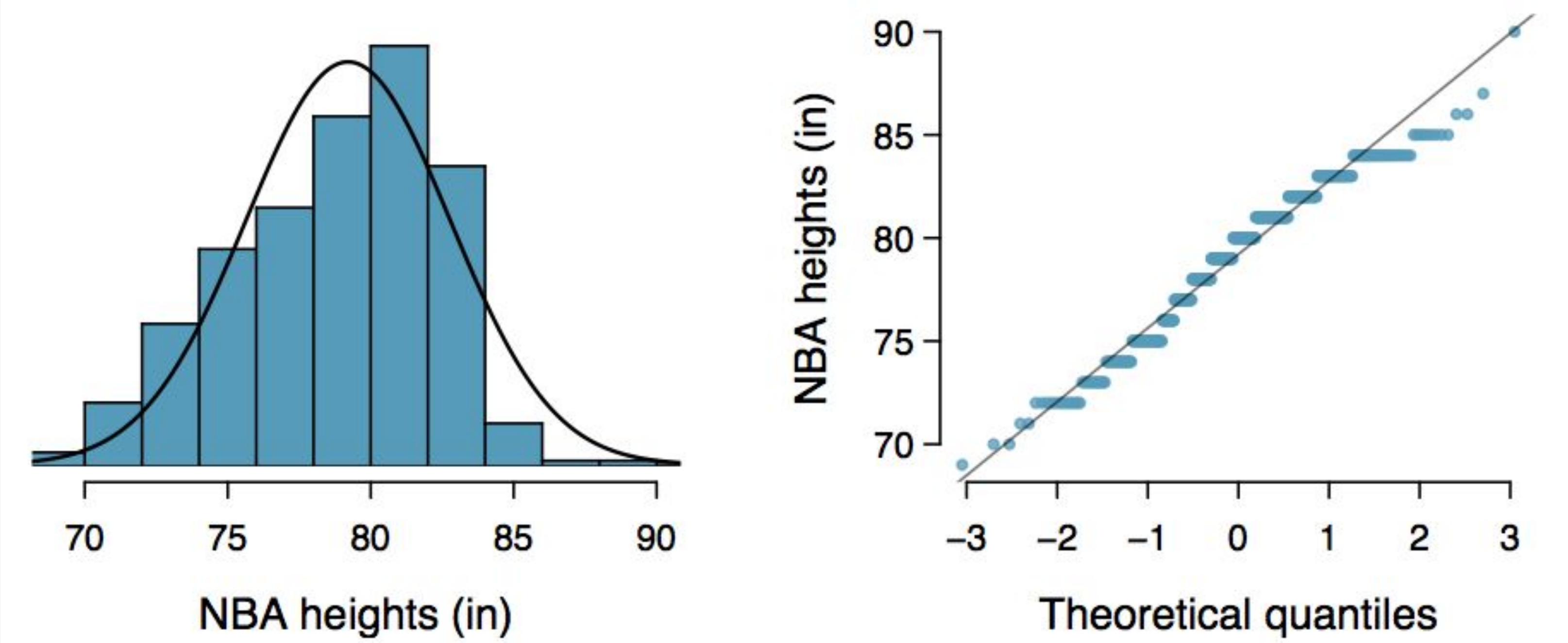


Anatomy of a normal probability plot

- Data are plotted on the y-axis of a normal probability plot, and theoretical quantiles (following a normal distribution) on the x-axis.
- If there is a linear relationship in the plot, then the data follow a nearly normal distribution.
- Constructing a normal probability plot requires calculating percentiles and corresponding z-scores for each observation, which is tedious. Therefore we generally rely on software when making these plots.

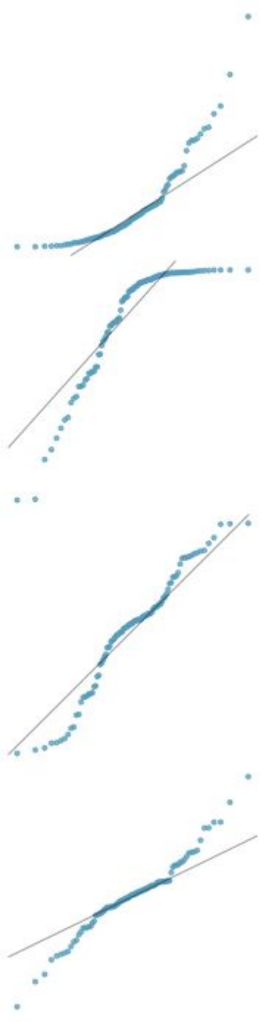
Practice

Below is a histogram and normal probability plot for the NBA heights from the 2008-2009 season. Do these data appear to follow a normal distribution?



Why do the points on the normal probability have jumps?

Normal probability plot and skewness (optional)



Right skew - Points bend up and to the left of the line.

Left skew - Points bend down and to the right of the line.

Short tails (narrower than the normal distribution) - Points follow an S shaped-curve.

Long tails (wider than the normal distribution) - Points start below the line, bend to follow it, and end above it.

Sampling distribution of a sample mean

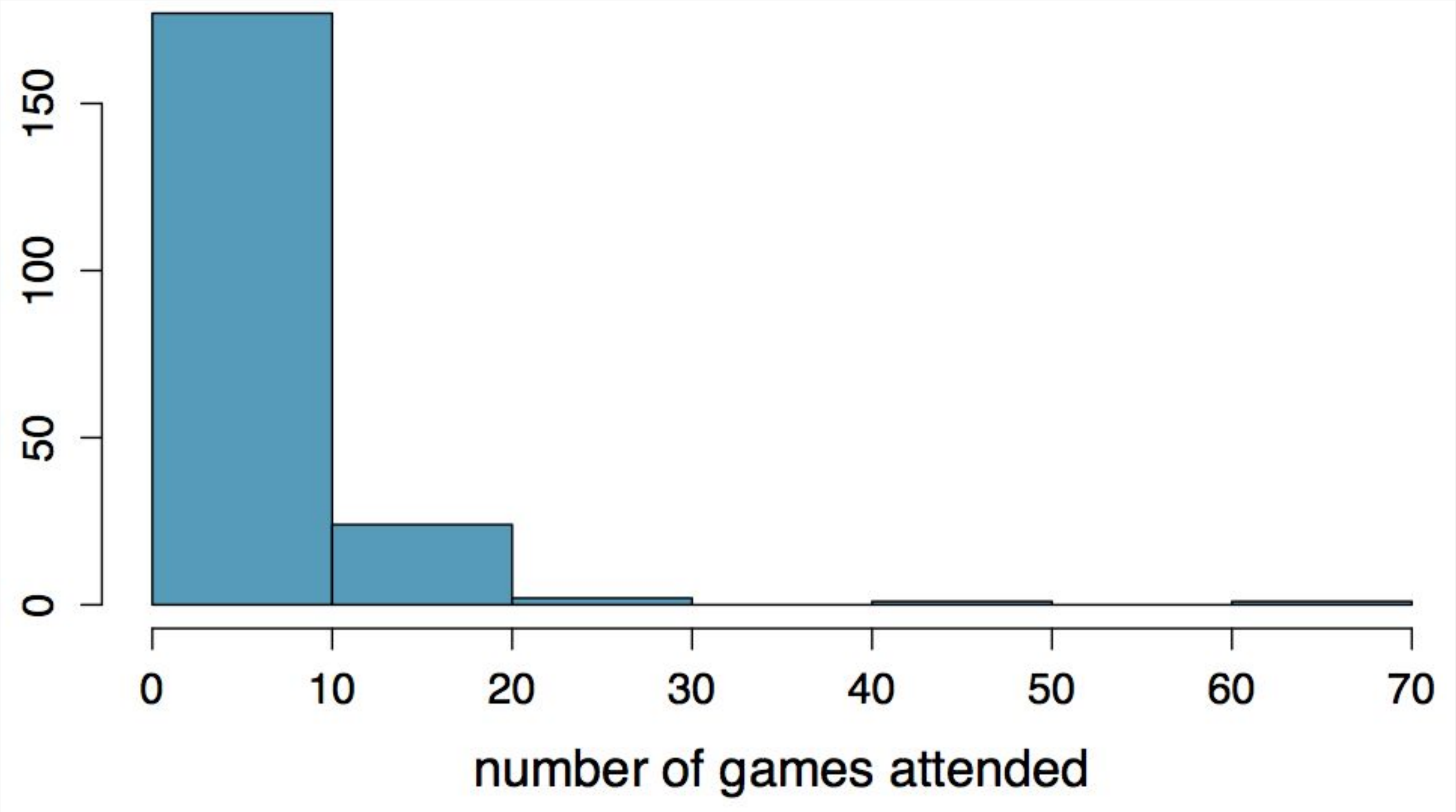
To save and make a local (editable) copy, do: File, Make a copy.

Advanced High School Statistics

Slides developed by Mine Çetinkaya-Rundel of OpenIntro, modified by Leah Dorazio for use with AHSS.

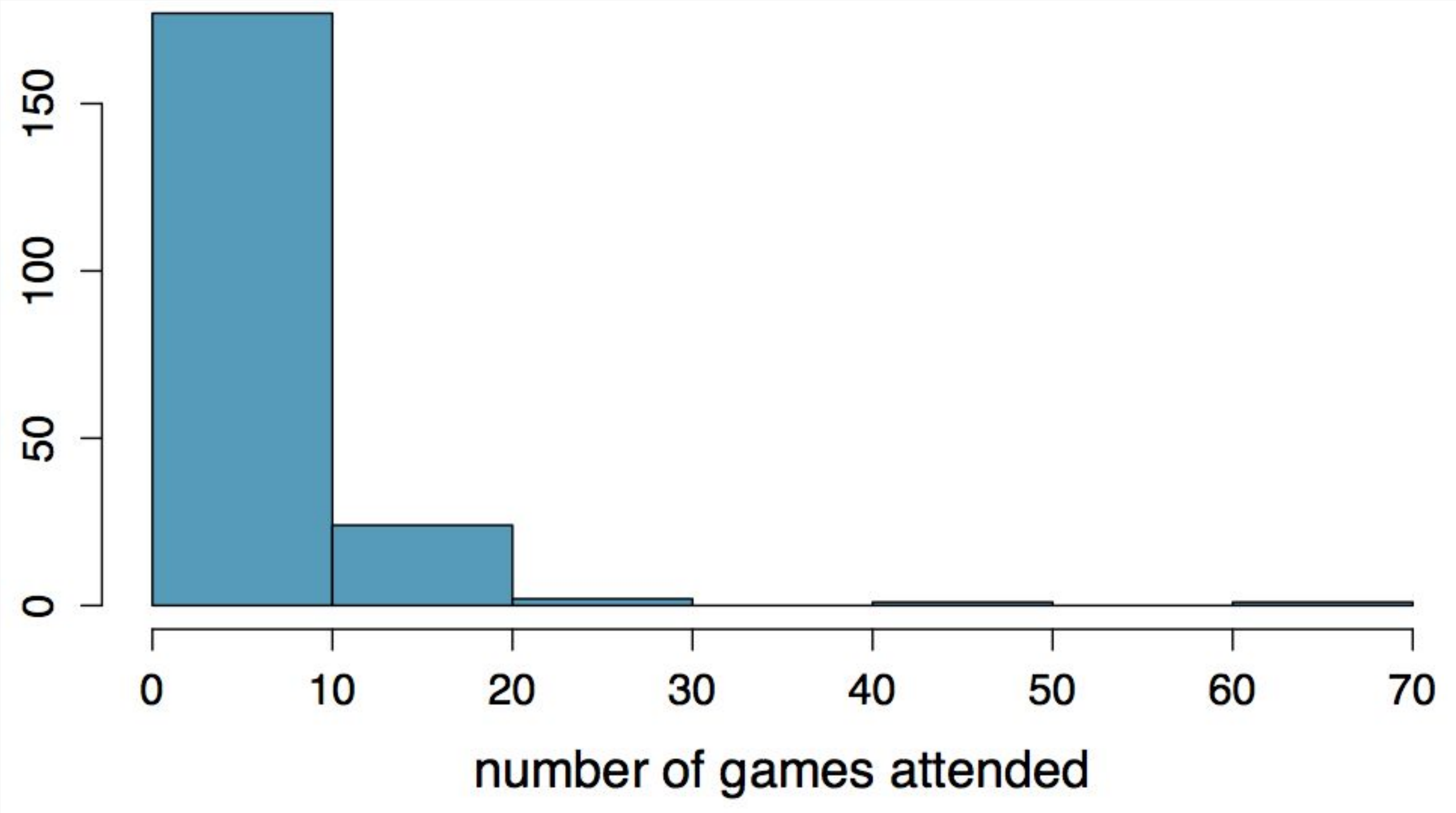
Number of basketball games attended (population)

First, let's look at the population data for the number of basketball games attended. Is this distribution right-skewed, left-skewed, or approximately symmetric?



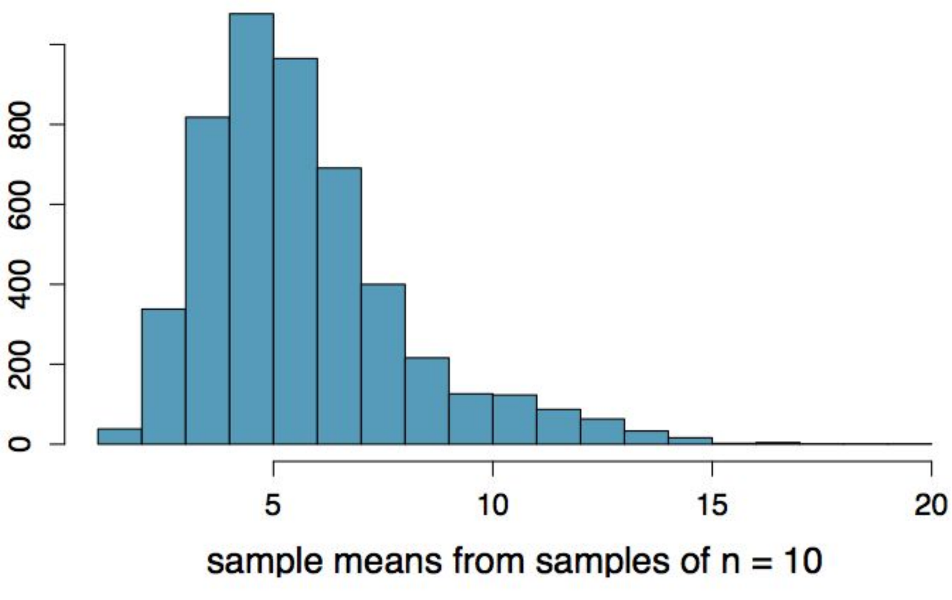
Number of basketball games attended (population)

First, let's look at the population data for the number of basketball games attended. Is this distribution right-skewed, left-skewed, or approximately symmetric?



Average number of games attended

Now, let's look at the sampling distribution of the sample mean for a sample of size n = 10:



What does each observation in the sampling distribution represent?

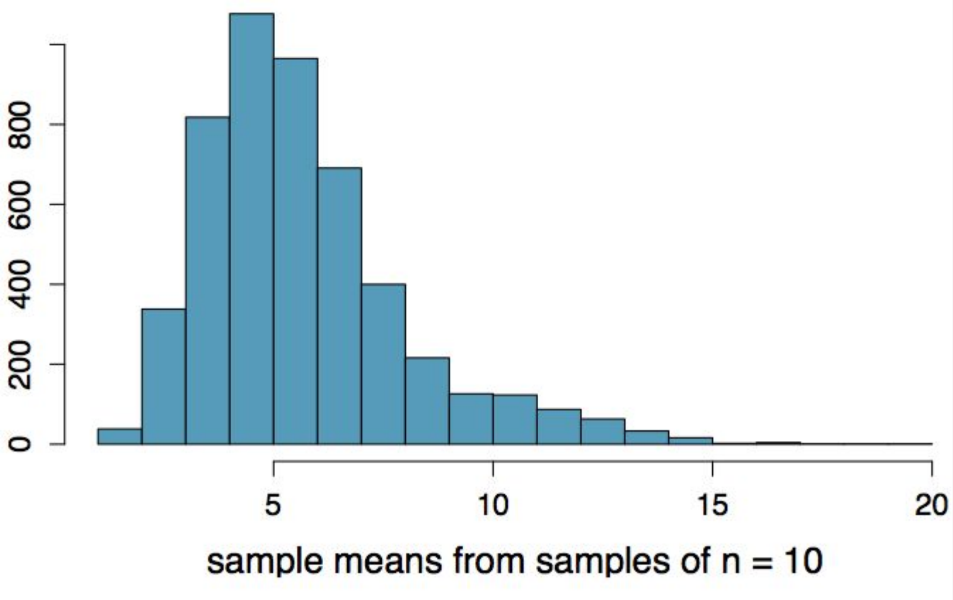
Sample mean ($\bar{x}$) of samples of size n = 10.

Is the variability of the sampling distribution smaller or larger than the variability of the population distribution? Why?

Smaller, sample means will vary less than individual observations.

Average number of games attended

Now, let's look at the sampling distribution of the sample mean for a sample of size n = 10:



What does each observation in the sampling distribution represent?

One sample mean ($\bar{x}$) from a sample of size 10.

Sample mean ($\bar{x}$) of samples

Is the variability of the sampling distribution smaller or larger than the variability of the population distribution? Why?

Smaller. Sample means will vary less than individual observations.

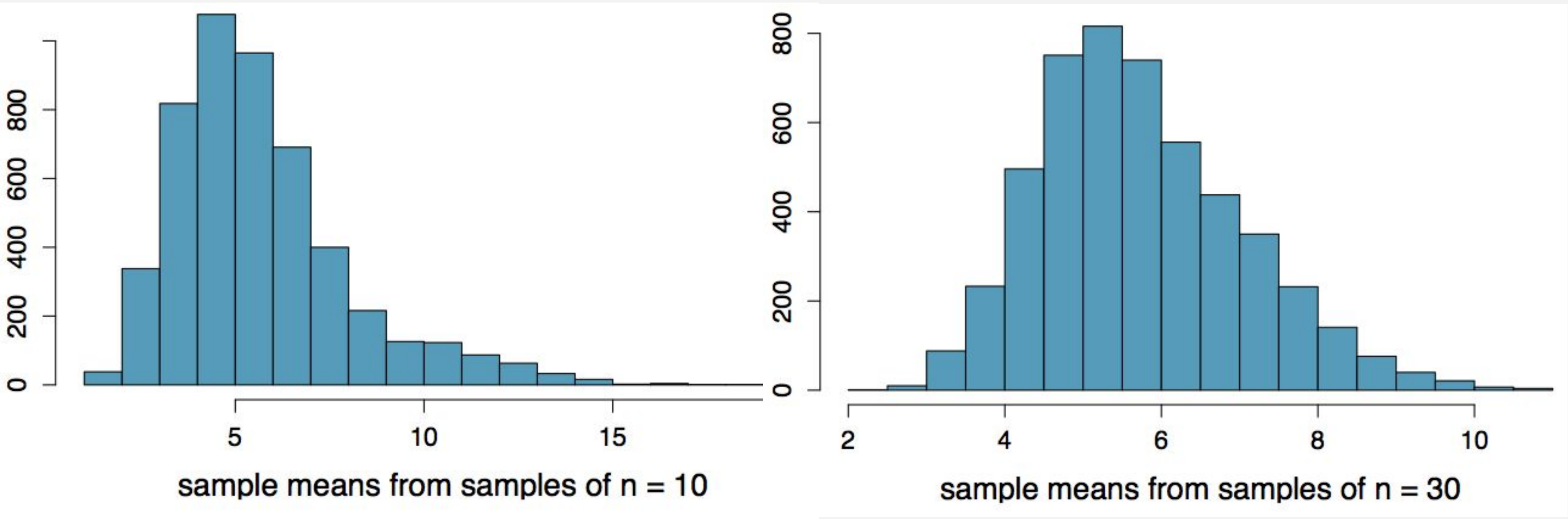
Smaller, sample means will

Average number of games attended

Now, let’s compare the sampling distribution of the mean for a sample of size 10 versus a sample of size 30.

How did the center, spread, and shape of the sampling distribution change going from n = 10 to n = 30?

Sample mean (x̄) of samples of size n = 10.



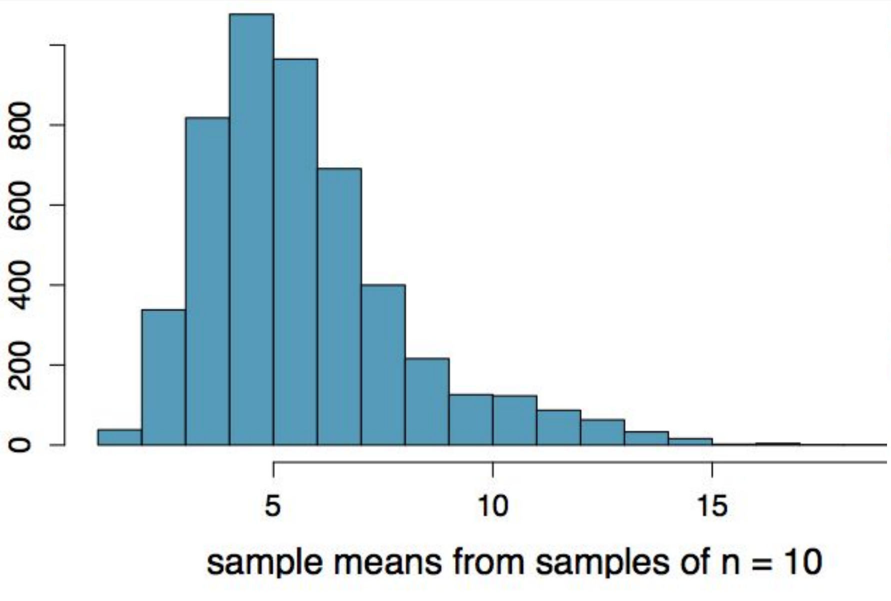
Note that the scales for the two histograms are different!

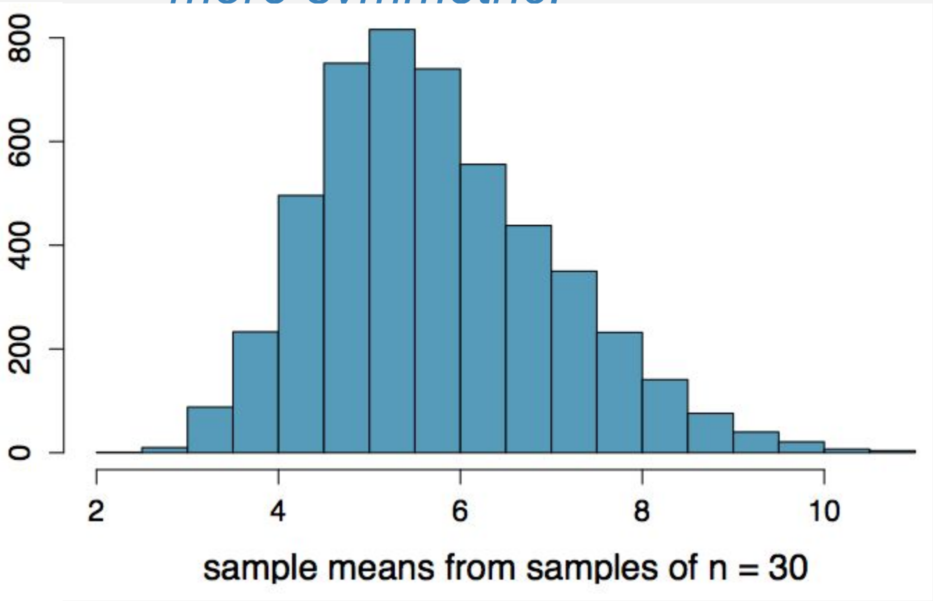
Average number of games attended

Now, let’s compare the sampling distribution of the mean for a sample of size 10 versus a sample of size 30.

How did the center, spread, and shape of the sampling distribution change going from n = 10 to n = 30?

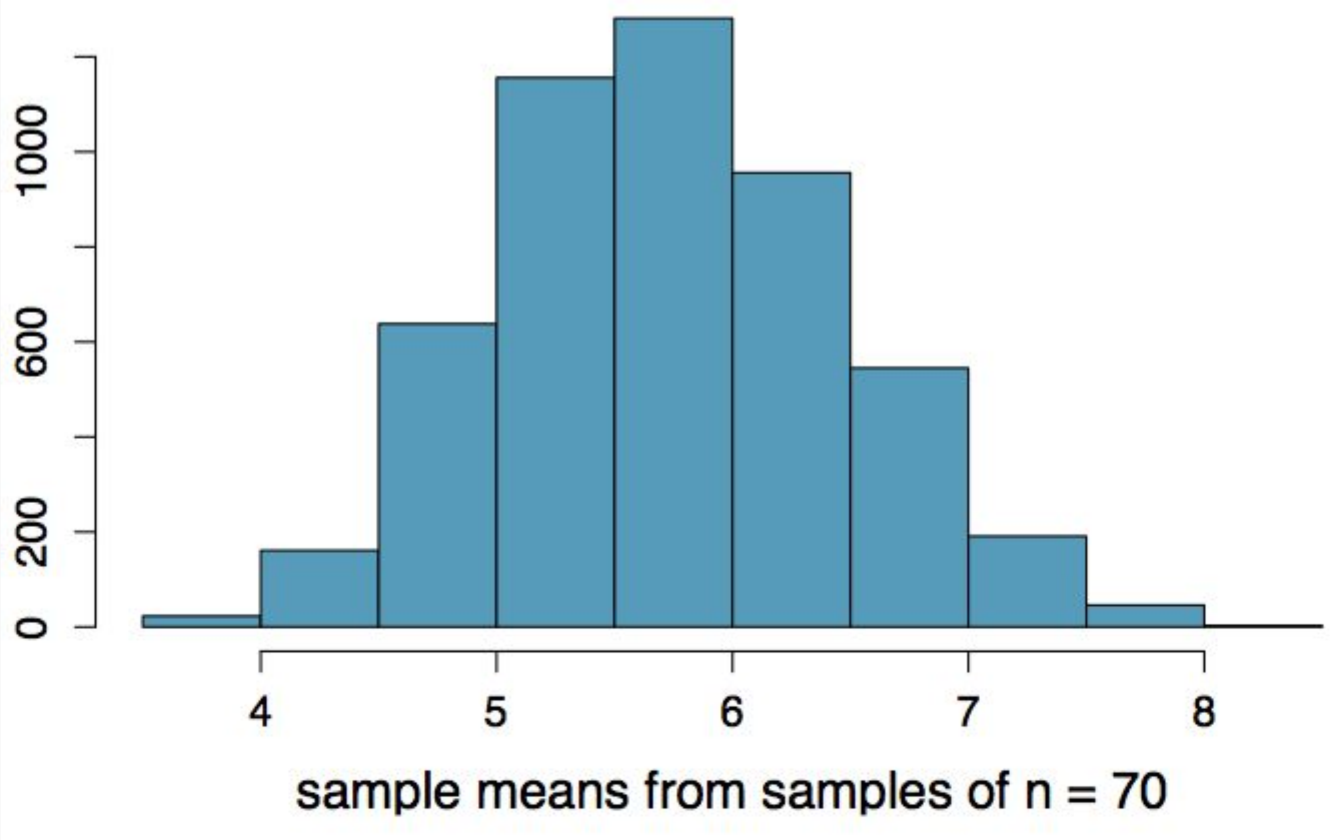
Center is about the same, spread is smaller, shape is more symmetric.





Average number of games attended

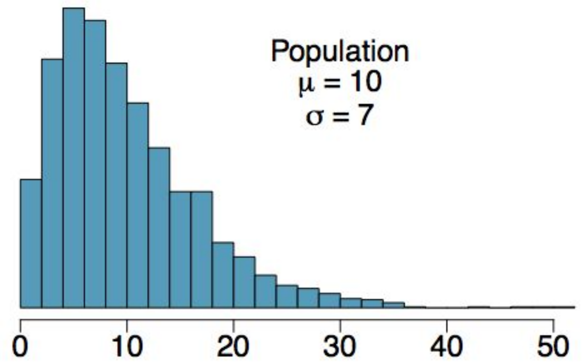
Sampling distribution, n = 70:



Practice

At top: distribution for a population (μ = 10, σ = 7),
Determine which plot (A, B, or C) goes with each of the following:

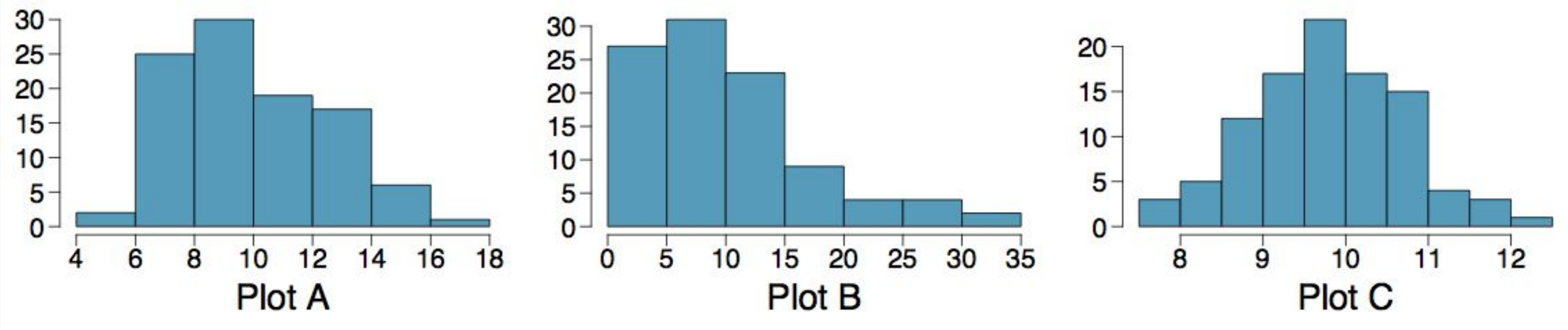
1. a single random sample of 100 observations from this population,
2. a distribution of 100 sample means from random samples with size 7, and
3. a distribution of 100 sample means from random samples with size 49.



1 →

2 →

3 →

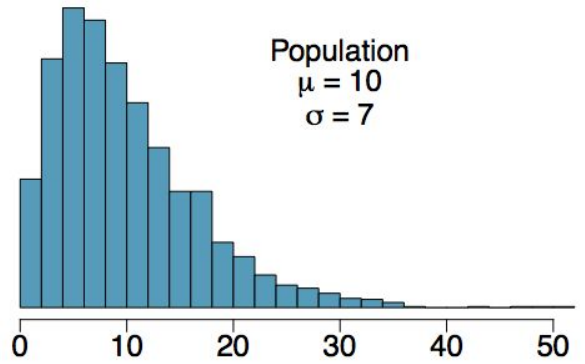


Note: the scales of the histograms are different!

Practice

At top: distribution for a population (μ = 10, σ = 7),
Determine which plot (A, B, or C) goes with each of the following:

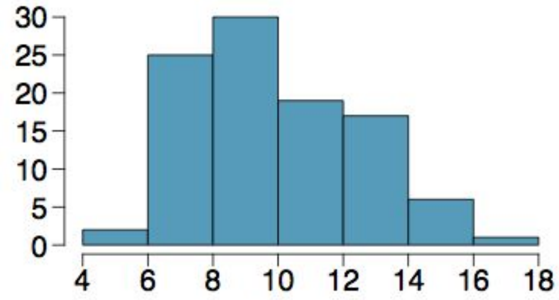
1. a single random sample of 100 observations from this population,
2. a distribution of 100 sample means from random samples with size 7, and
3. a distribution of 100 sample means from random samples with size 49.



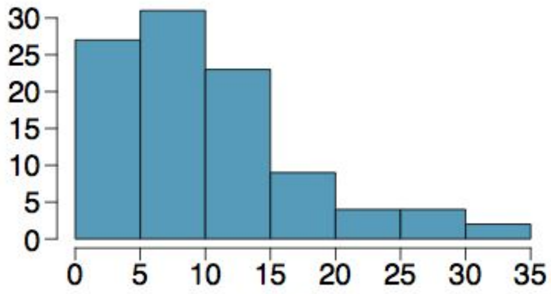
1 → Plot B (SD similar to SD of population)

2 → Plot A

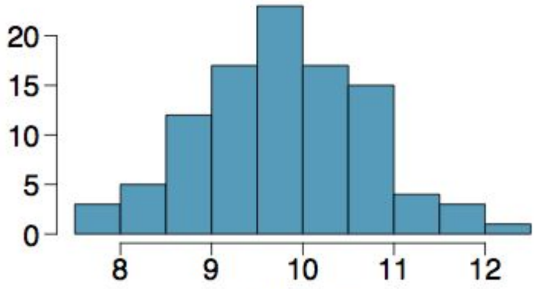
3 → Plot C (smallest spread)



Plot A



Plot B



Plot C

Note: the scales of the histograms are different!

The Central Limit Theorem

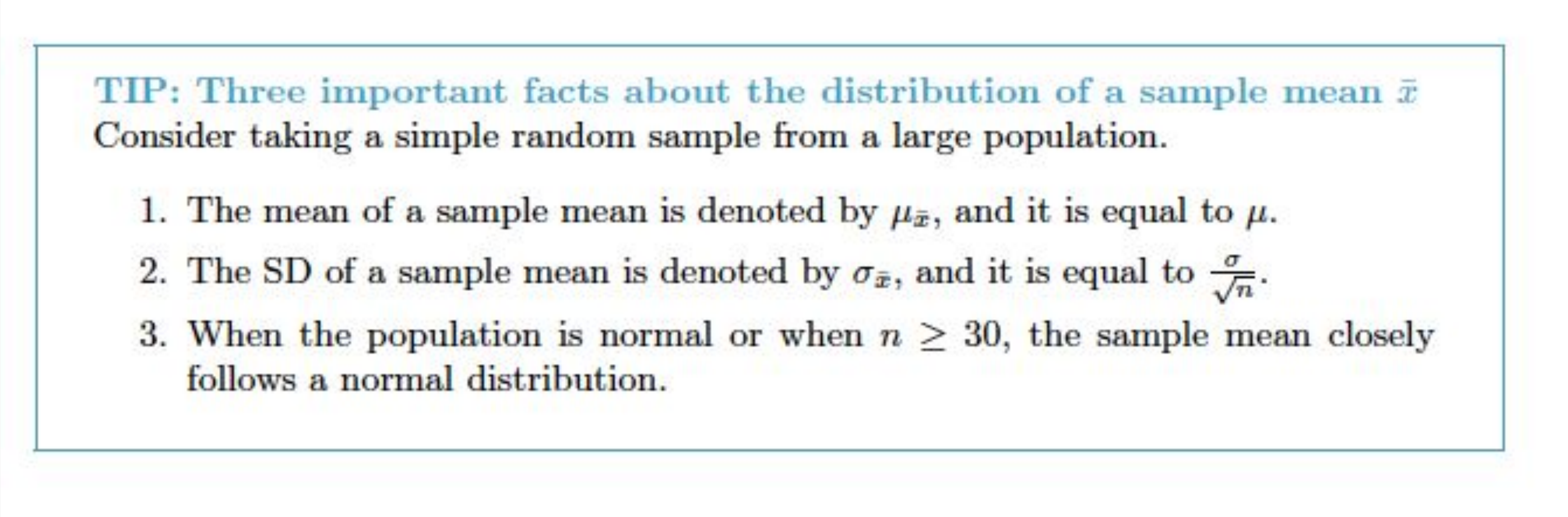
The Central Limit Theorem (CLT) is the most fundamental/famous theorem in Statistics. What follows is an *informal* rendition of the CLT for sample averages.

CLT for sample averages: As the sample size *n* gets larger, the distribution of the sample mean becomes more normal (no matter what the distribution of the population looks like)

Caveats: You are taking a *random* sample from a *large* population with a fixed mean and sd.

For an excellent animation demonstrating the Central Limit Theorem for sample averages go to: http://onlinestatbook.com/stat_sim/sampling_dist/index.html

4.2 Summary



TIP: Three important facts about the distribution of a sample mean $\bar{x}$

Consider taking a simple random sample from a large population.

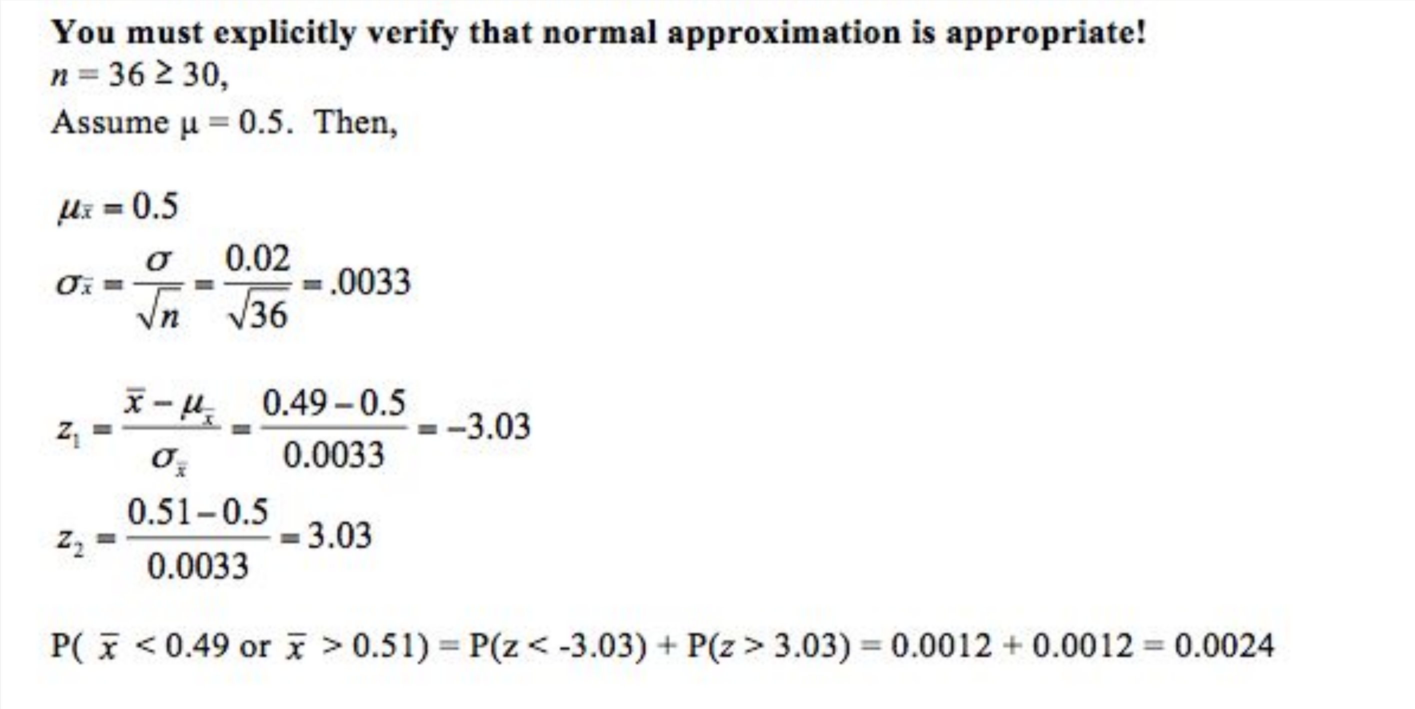
1. The mean of a sample mean is denoted by $\mu_{\bar{x}}$, and it is equal to μ.
2. The SD of a sample mean is denoted by $\sigma_{\bar{x}}$, and it is equal to $\frac{\sigma}{\sqrt{n}}$.
3. When the population is normal or when $n \geq 30$, the sample mean closely follows a normal distribution.

Practice

A manufacturing process is designed to produce bolts with 0.5-in diameter. Once each day, a random sample of 36 bolts is selected and the diameters recorded. If the resulting sample mean is less than 0.49-in or greater than 0.51-in, the process is shut down for adjustment. The standard deviation for diameter is 0.02-in. What is the probability that the manufacturing line will be shut down unnecessarily? [Hint, find the probability of finding an xbar in the shut-down range when the true process mean really 0.5-in].

Practice

A manufacturing process is designed to produce bolts with 0.5-in diameter. Once each day, a random sample of 36 bolts is selected and the diameters recorded. If the resulting sample mean is less than 0.49-in or greater than 0.51-in, the process is shut down for adjustment. The standard deviation for diameter is 0.02-in. What is the probability that the manufacturing line will be shut down unnecessarily? [Hint, find the probability of finding an xbar in the shut-down range when the true process mean really 0.5-in].



You must explicitly verify that normal approximation is appropriate!

$n = 36 \geq 30$,

Assume $\mu = 0.5$. Then,

$$\mu_{\bar{x}} = 0.5$$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{0.02}{\sqrt{36}} = .0033$$

$$z_1 = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{0.49 - 0.5}{0.0033} = -3.03$$

$$z_2 = \frac{0.51 - 0.5}{0.0033} = 3.03$$

$P(\bar{x} < 0.49 \text{ or } \bar{x} > 0.51) = P(z < -3.03) + P(z > 3.03) = 0.0012 + 0.0012 = 0.0024$