Introduction to Bioinformatics (GS011143)

2025 Spring

Statistics Lecture 1

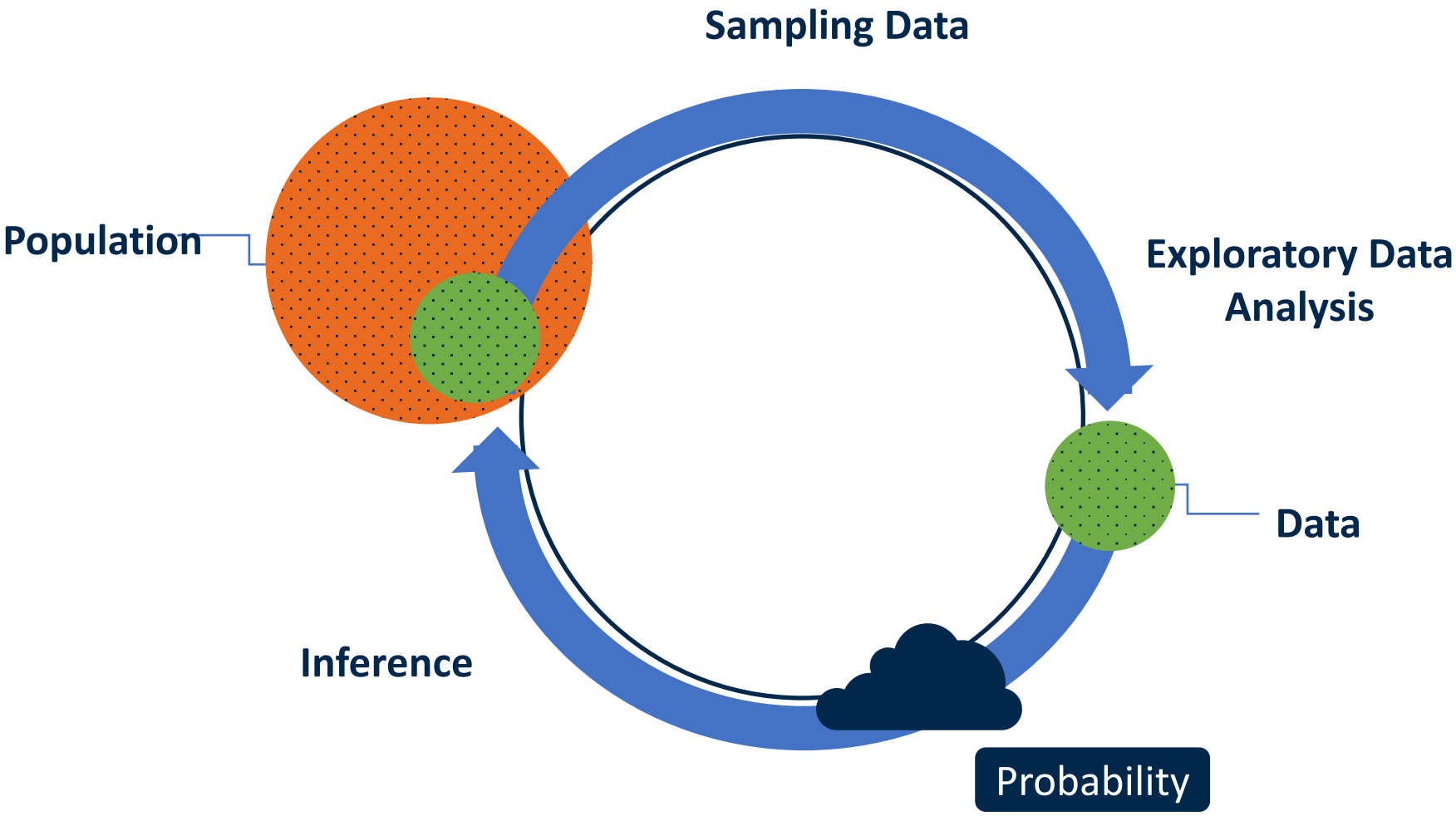
Instructor: Lulu Shang

(Dept. of Biostatistics, MDACC)

Outline

- Lecture 1
  - Descriptive Statistics
  - Probability Distributions
  - Mean and Confidence Interval
  - Introduction to R
- Lecture 2
  - Statistical Inference: Tests for means
  - High-dimensional Analysis: Multiple testing & Feature selection
- Lecture 3
  - Principal Component Analysis (PCA)
  - Clustering

The Big Picture

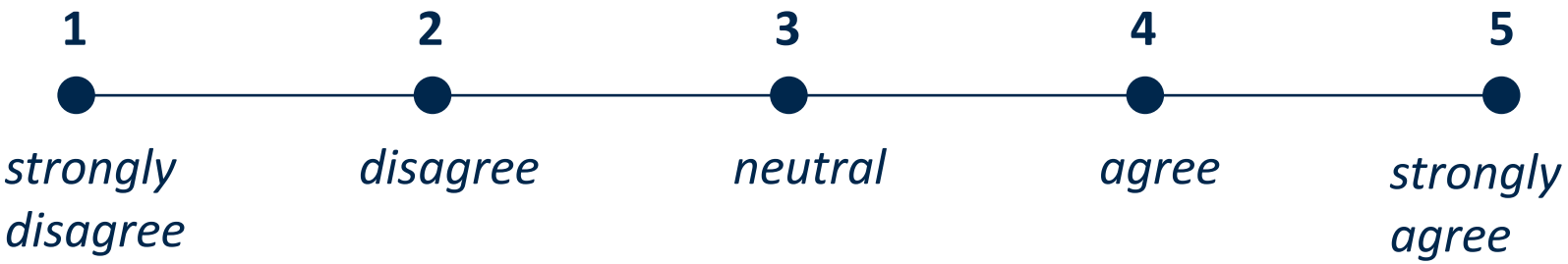


Adapted from: http://bolt.mph.ufl.edu/6050-6052/unit-3/#review

Types of Data

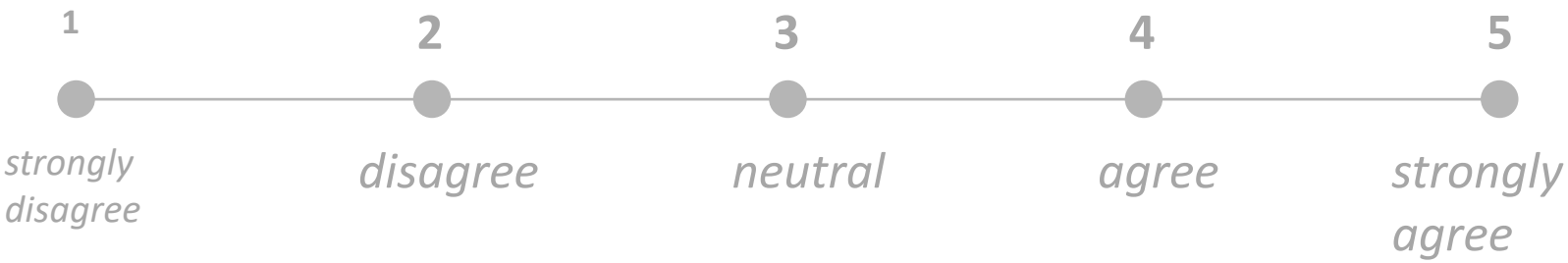
Qualitative data

Ordinal: Data that have a natural order



Qualitative data

Ordinal: Data that have a natural order



Nominal: Data that do NOT have a natural order

- Race
- Gender
- Political Affiliation

Quantitative data

Continuous

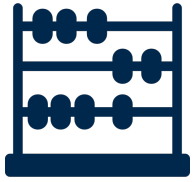
- *Age*
- *Family Income*
- *Blood pressure*
- *Pain measured by 0-100 VAS*

Quantitative data

Continuous

- *Age*
- *Family Income*
- *Blood pressure*
- *Pain measured by 0-100 VAS*

Count



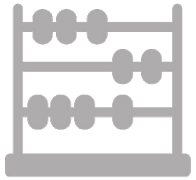
- *Number missing teeth*
- *Number of texts sent/day*
- *Number depressive episodes/week*

Quantitative data

Continuous

- *Age*
- *Family Income*
- *Blood pressure*
- *Pain measured by 0-100 VAS*

Count



- *Number missing teeth*
- *Number of texts sent/day*
- *Number depressive episodes/week*

Binary

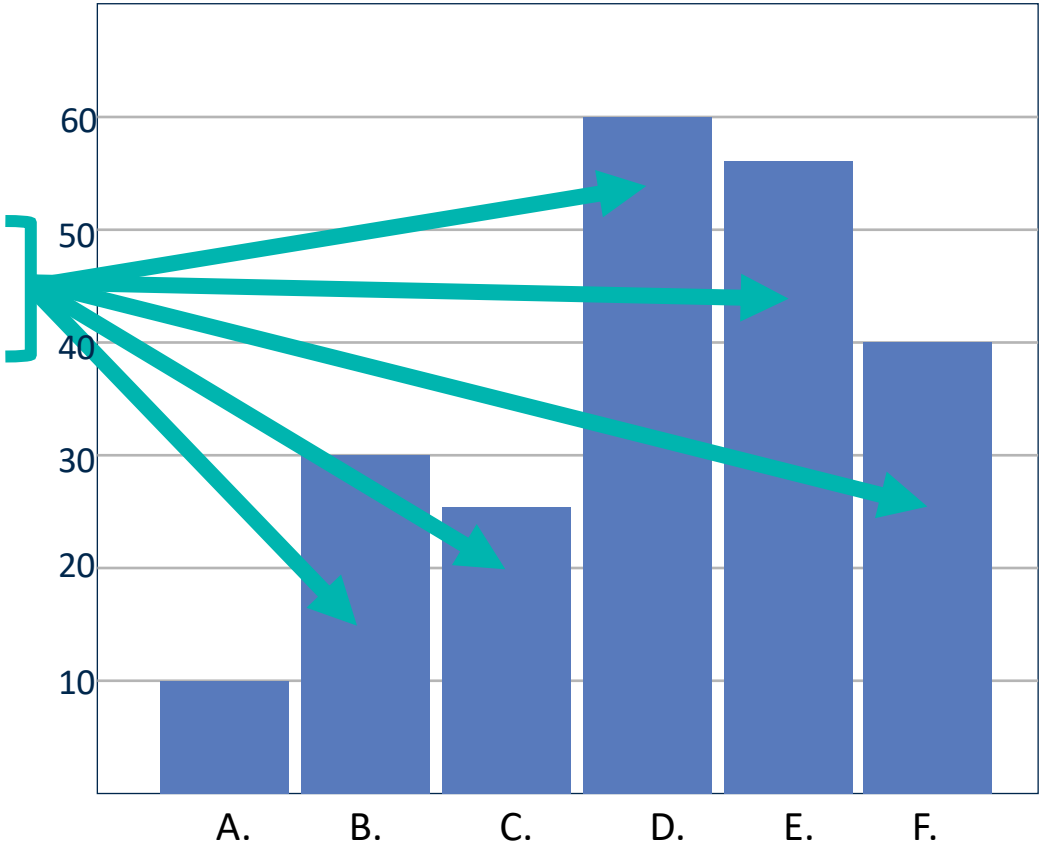
- *Missing any teeth (yes/no)*
- *Developed throat cancer (yes/no)*
- *Experienced pain (yes/no)*

Graphically Summarize the Data

Barplots

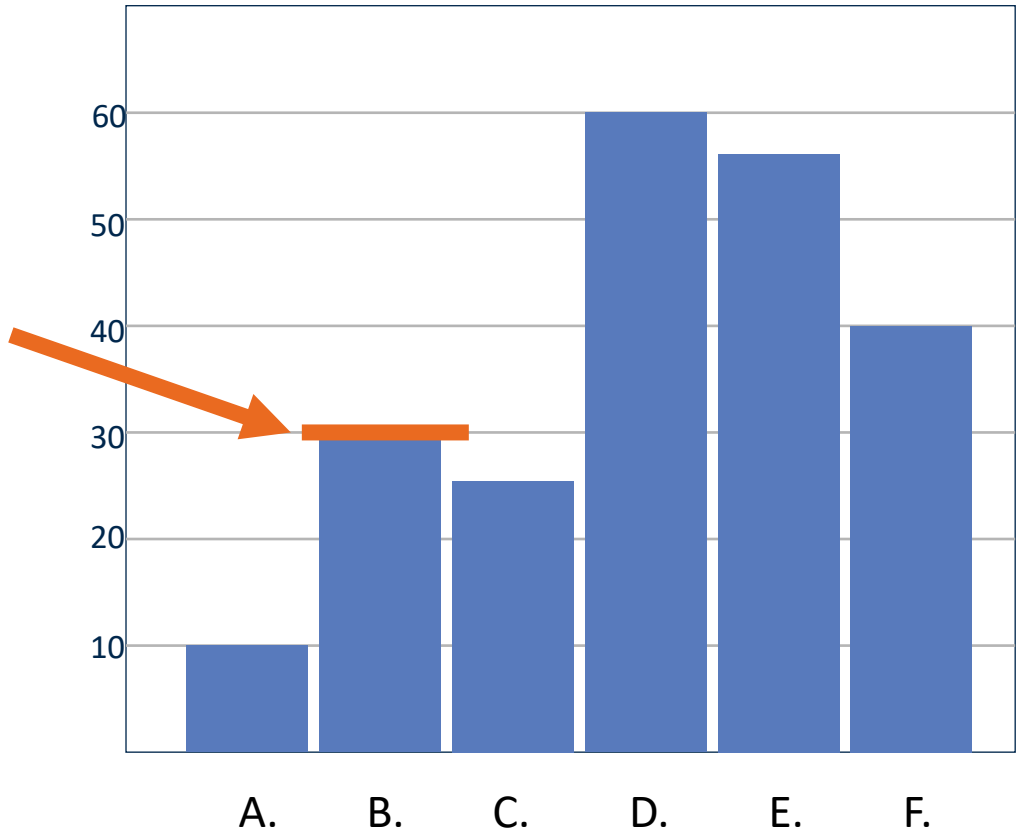
Barplot: summarizes one categorical variable, whether ordinal or nominal

Each bar corresponds to each specific category or value that the variable can have



Barplots

The height of each bar denotes how many individuals are in the same category or have the same value for the variable

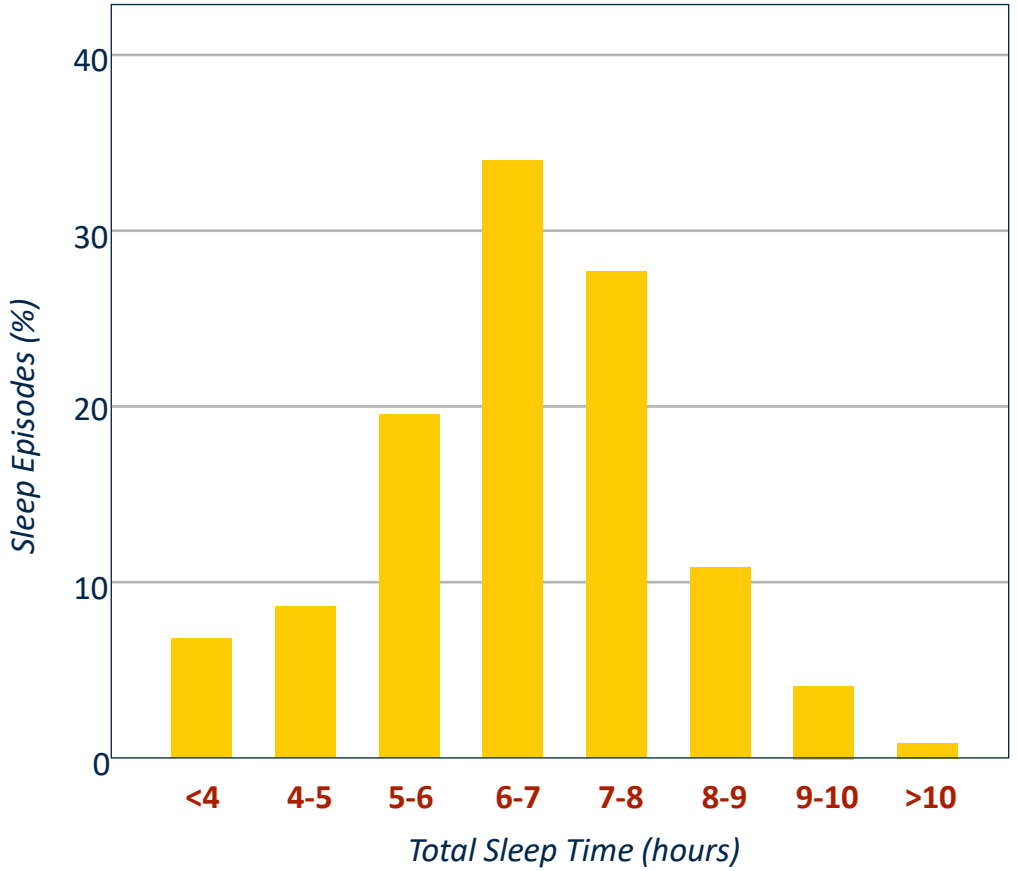


Histograms

Histogram: a barplot applied to a continuous variable

We divide the range of possible values of the variable into intervals

- Each interval must be the same width
- Each interval is now a **category**
- These categories are called ***bins***

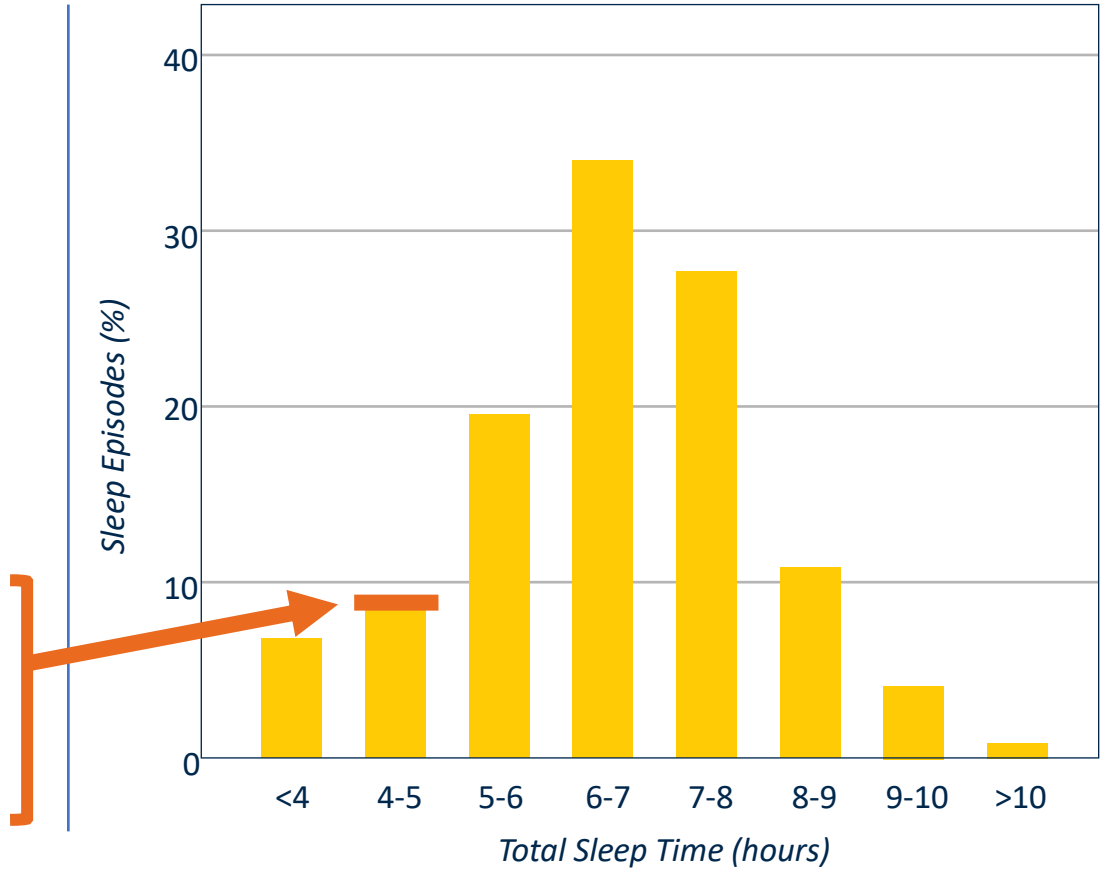


Histograms

Histogram: a barplot applied to a continuous variable

We divide the range of possible values of the variable into intervals

We then have a series of bars, one for each bin. The height of each bar is the number (or percentage) of individuals with a value in that bin



Graphical frequency distribution

- Active COVID-19 cases as of 01/01/2021 (source: https://coronavirus.jhu.edu)

Numerically Summarize the Data

The Sample Mean

$$\text{Sample Mean} = \frac{\text{sum of observations}}{\text{number of observations}}$$

The Sample Mean

If we have n observations, and our data points are labelled as $X_1, X_2, \ldots, X_n$

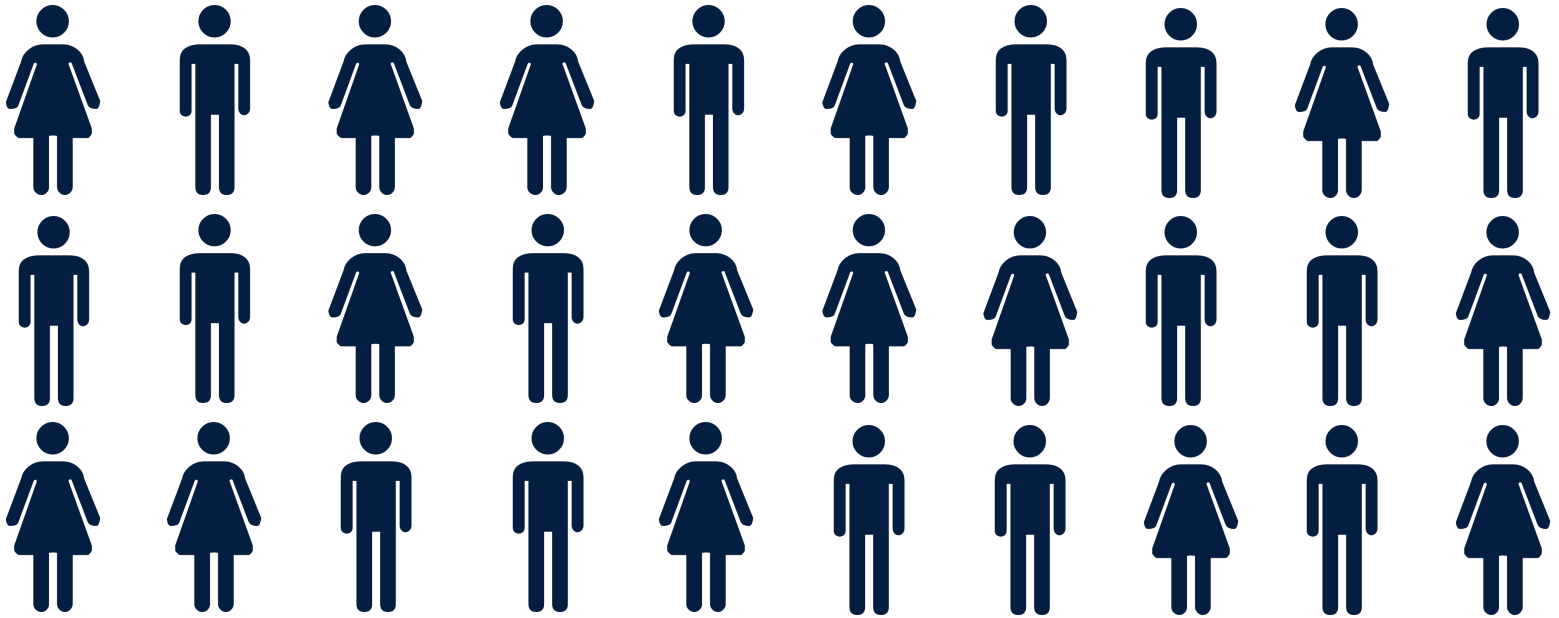
$$\text{Sample Mean} = \frac{\boldsymbol{X_1 + X_2 + \cdots + X_n}}{\boldsymbol{n}}$$

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

$$= \overline{X}$$

Consider this example

We have 30 patients who received antibiotics while being hospitalized

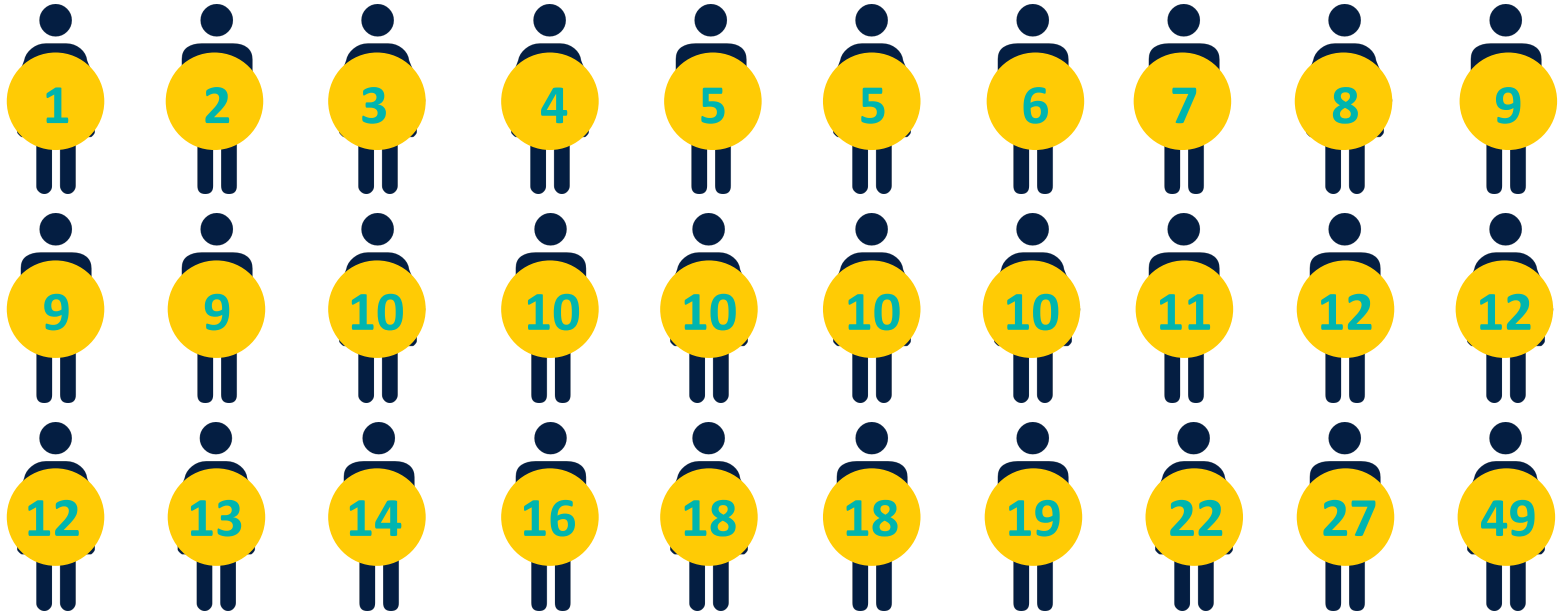


Let X_i = length of stay (days) for patient *i* = 1, 2, … , 30

Consider this example

We have 30 patients who received antibiotics while being hospitalized

The ordered values are:



Let X_i = length of stay (days) for patient *i* = 1, 2, … , 30

Let $\overline{X}$ = mean length of stay in data

$$\overline{X} = \frac{X_1 + X_2 + \cdots + X_{30}}{30}$$

$$= \frac{1}{30}\sum_{i=1}^{30} X_i = 12$$

$X_i - \overline{X}$ ⟵ **deviation of length of stay for subject *i* from the average of all subjects**

However: $\frac{1}{30}\sum_{i=1}^{30}\left(X_i - \overline{X}\right) = 0$

Sample Standard Deviation

To avoid this problem, we compute the average of the **squared** deviations and then take the square root of that average:

$$S_x = \sqrt{\frac{1}{(30-1)}\sum_{i=1}^{30}\left(X_i - \overline{X}\right)^2} = 9.1$$

S_x is known as the **sample standard deviation** of the length of stay in our data

If we do not take the square root of the sum, we have a value known as the **sample variance**, which we denote as S_X^2

Probability Distributions

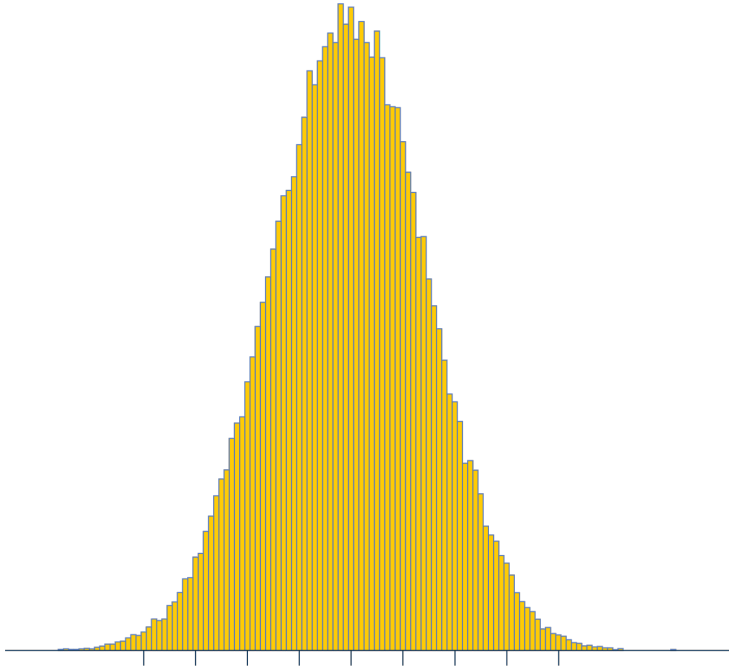
Relative Frequency

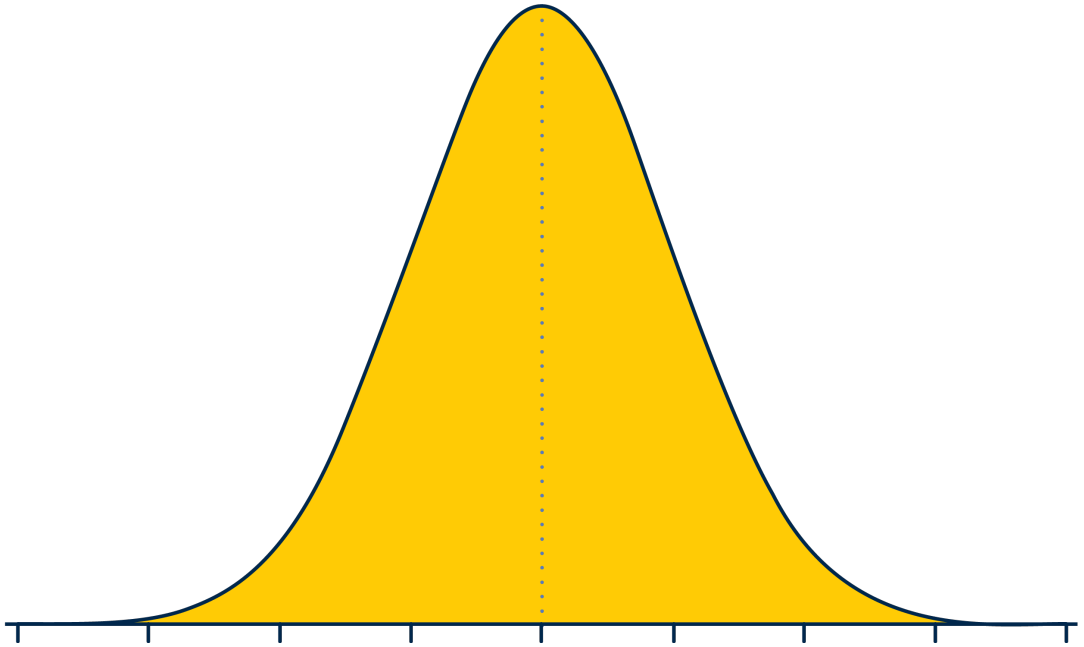
- The *relative frequency* for an interval is the proportion of the total number of observations appearing in that interval.
- This is calculated by dividing the number of values within an interval by the total number of values in the table

Location	Cases	Deaths
Harris County	**238K**	**3,402**
Dallas County	193K	2,007
Tarrant County	150K	1,494
Bexar County	116K	1,726
El Paso County	98,540	1,666
Hidalgo County	51,662	2,201
Travis County	50,595	549

Calculate relative frequency from this table.

Empirical Distribution vs Theoretical Distribution

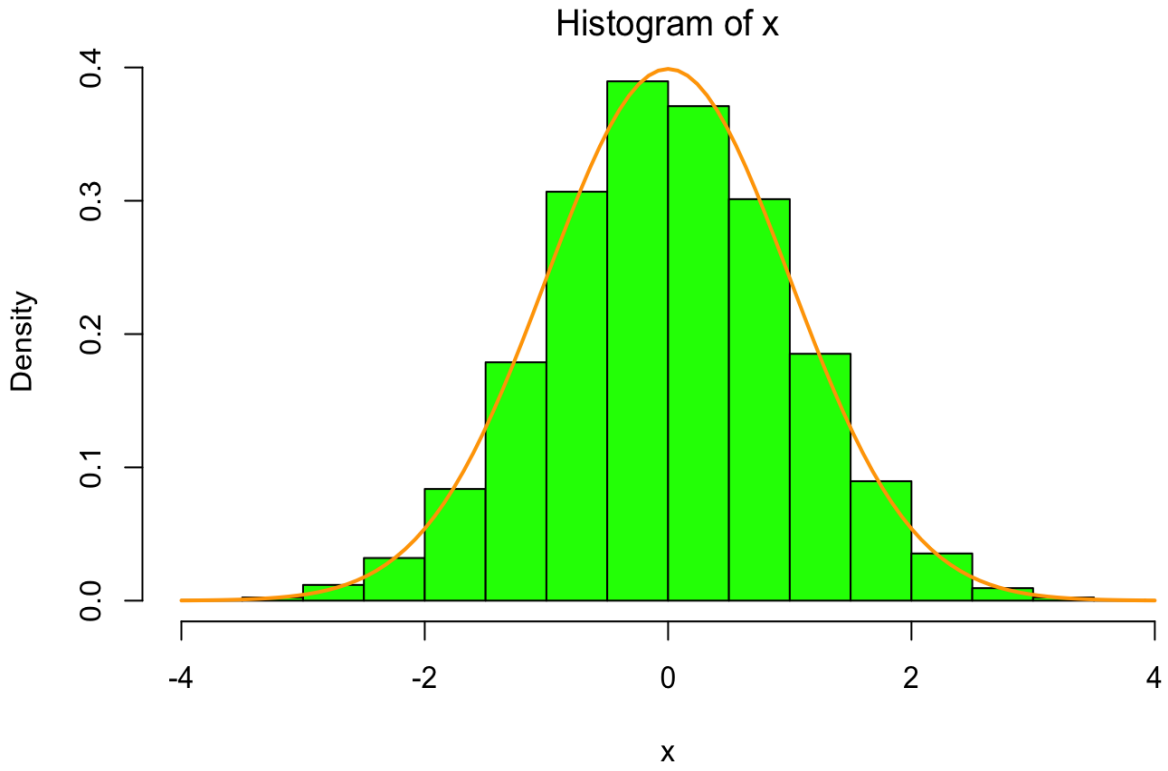




- Theoretical distribution is a distribution that is derived from certain principles or assumptions by logical and mathematical reasoning, as opposed to empirical distributions derived from real-world data.

Empirical distributions for continuous variables

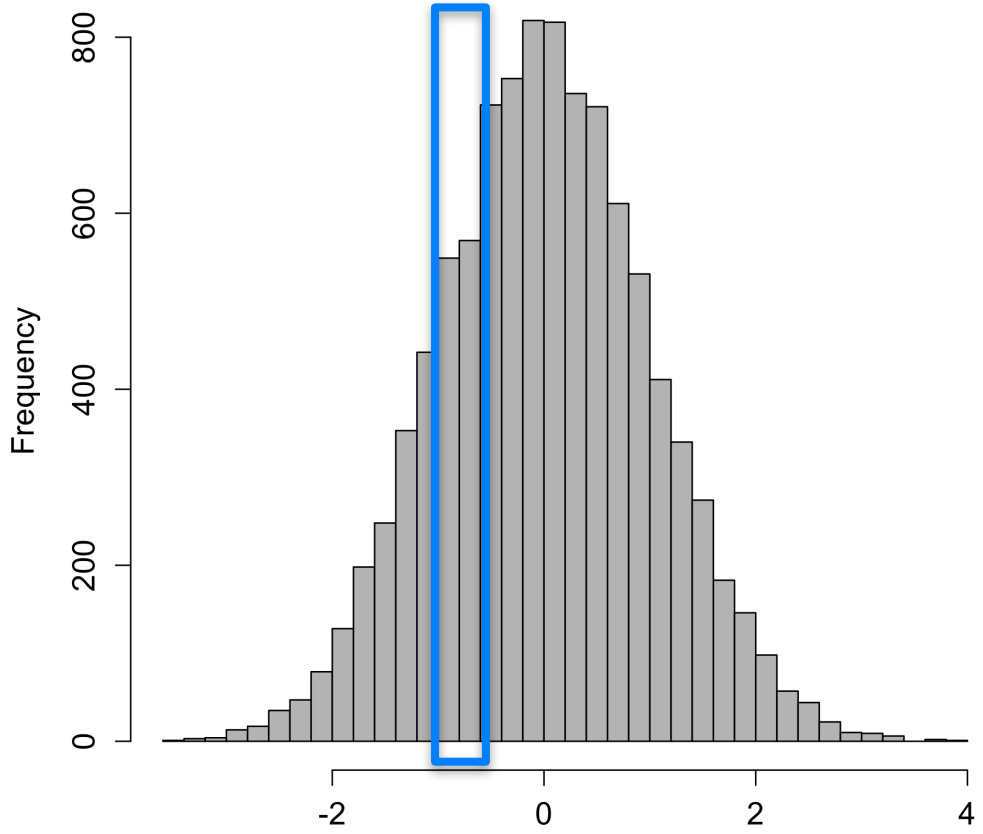
- In a given range of observed values, we define a frequency.
- The shape of frequencies are *empirical distribution*
- The probability of observing a value in a given range is the empirically observed (relative) frequency
- The distribution can be shown using *Histogram* (Try ?hist in R)
- The area of each bar is proportional to the frequency (or relative frequency= density) of the categories.



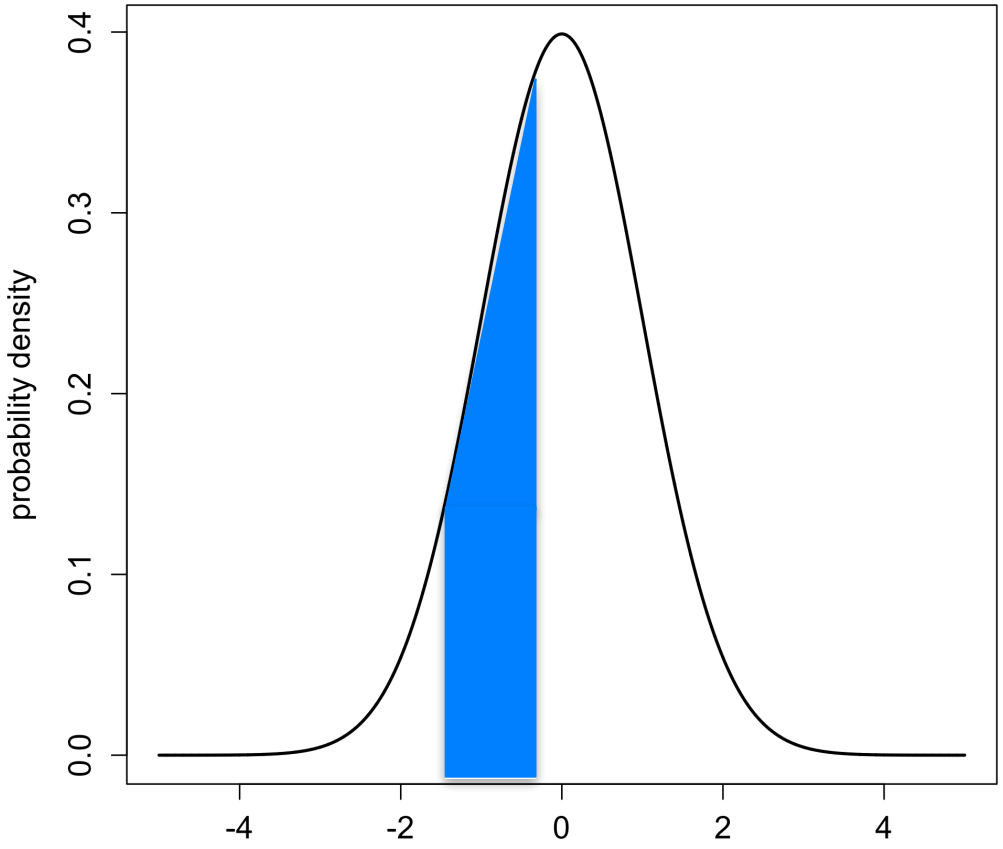
Continuous Distributions

- A continuous distribution describes the probabilities of possible values of a continuous random variable (infinite and uncountable)
- Density functions/curves, like histograms, can have any shape. The area under the density curve is always 1.
- How do you find the area of interest in the curves?
  - Integration!

$$P(x_a < x < x_b) = \int_{x_a}^{x_b} f(x)dx$$



Empirically observed frequency
(count the number of values observed)

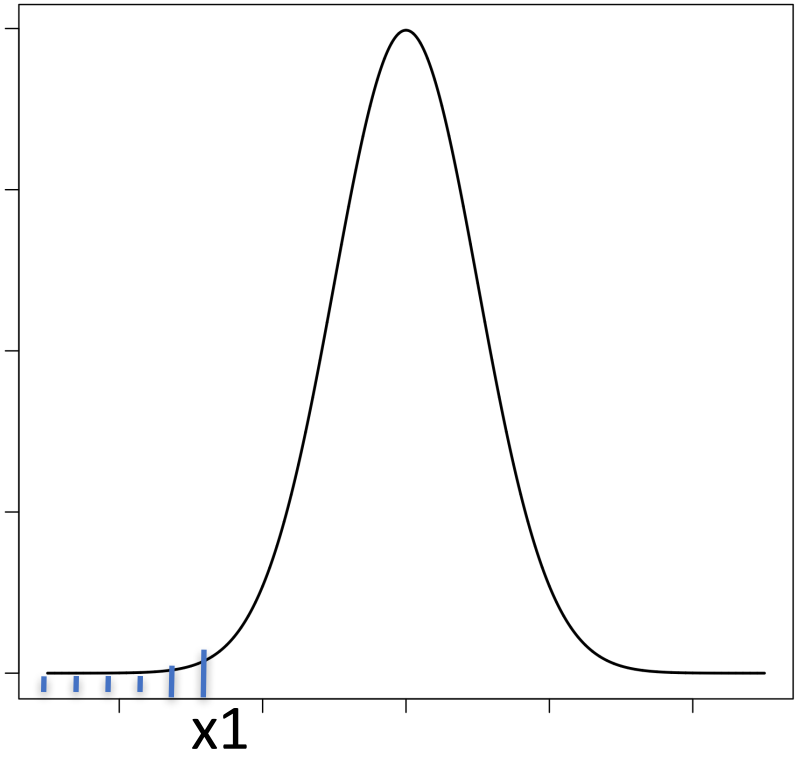


Analytical probability density
(area under the curve)

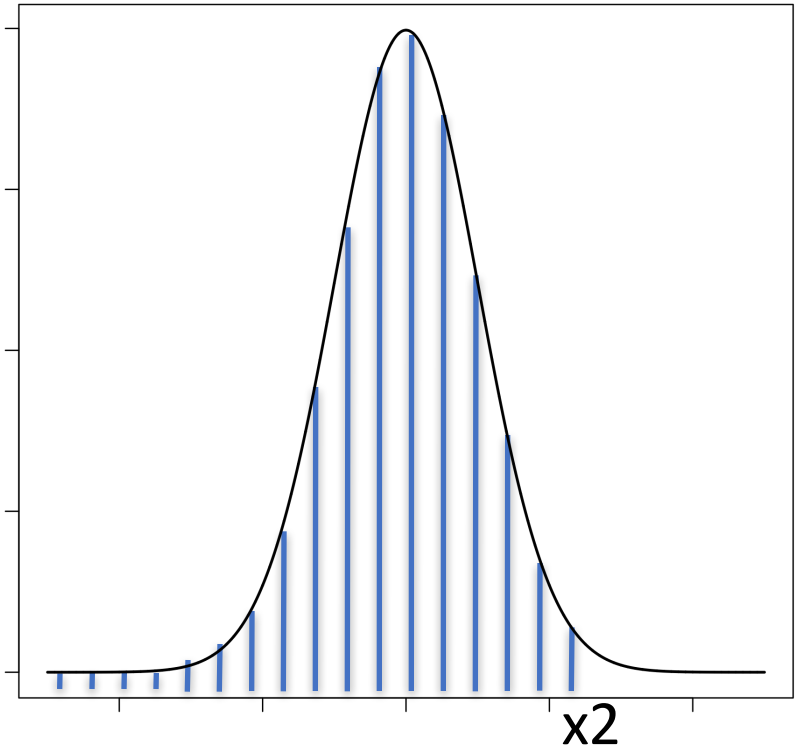
$$P(x_a < x < x_b) = \int_{x_a}^{x_b} f(x)dx$$

Finding tail probabilities (P given x)

P (X< x1) = …



P (X < x2) = …

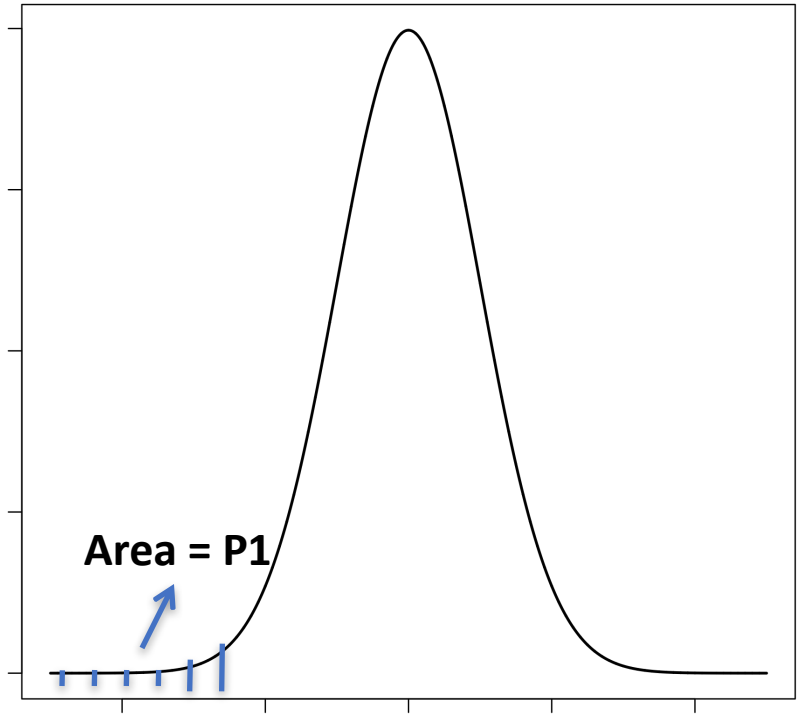


In R, pnorm (x =x1, mean = …, sd = …)

Input: quantile and parameters

Finding quantiles (x given P)

P (X <x …) = P1

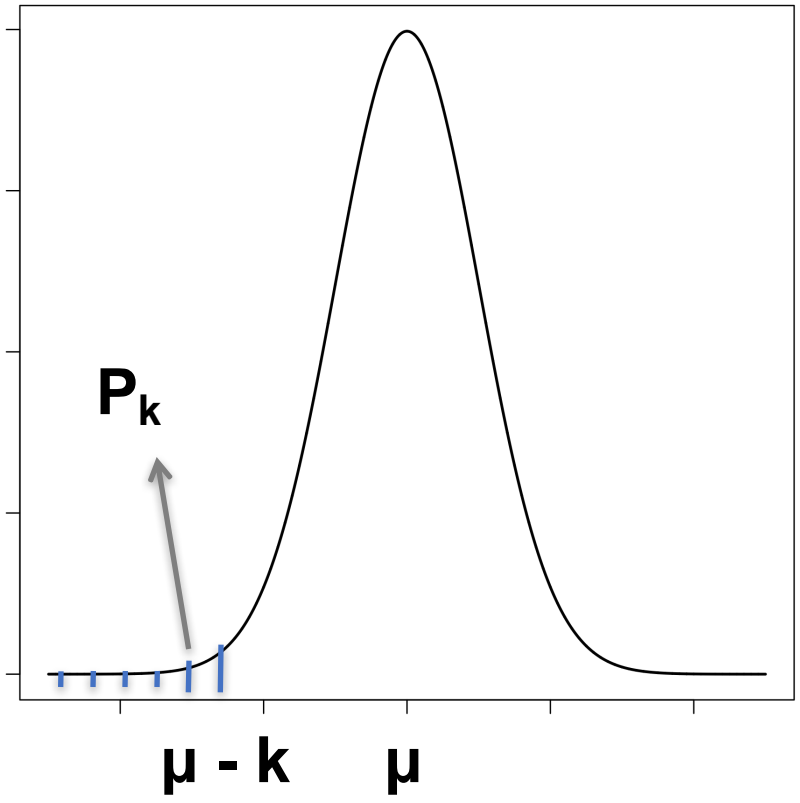


In R, qnorm (p =P1, mean = …, sd = …)

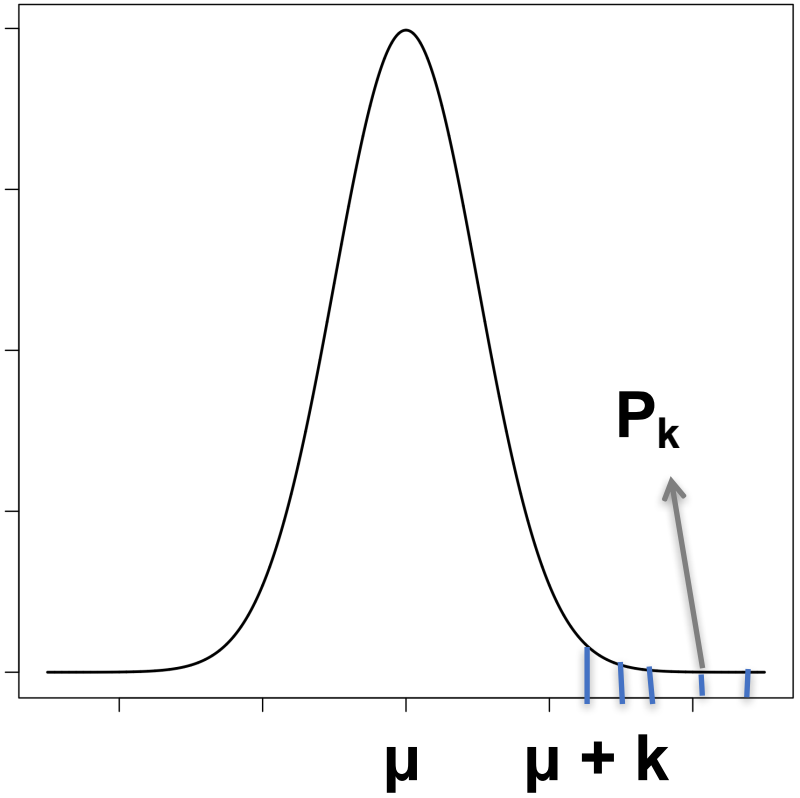
Input: left tail probability and parameters

Use the property of symmetry

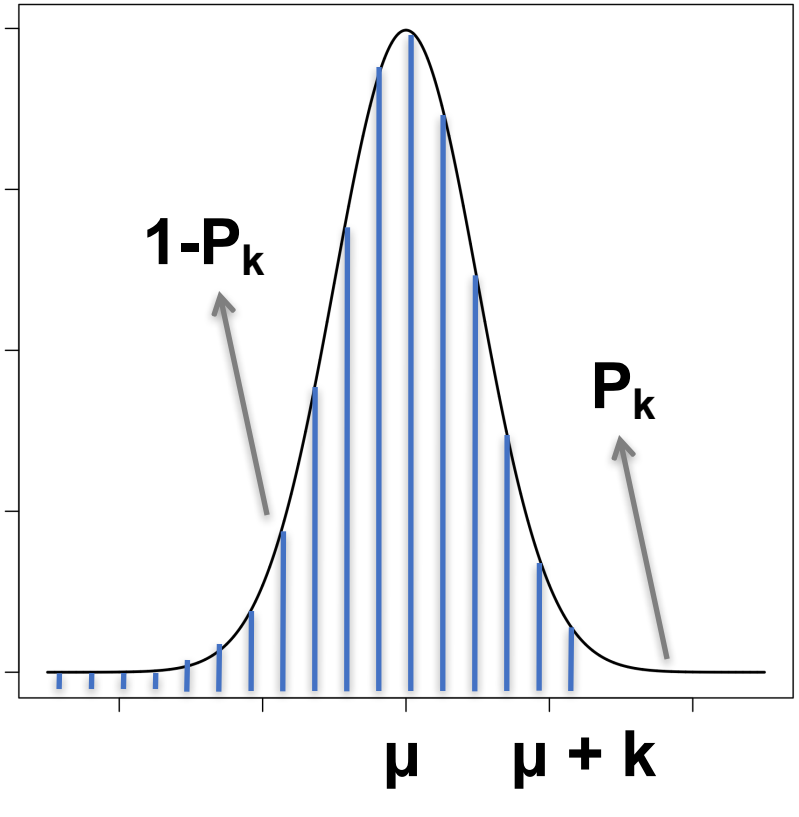
$P(x < \mu - k) = P_k$



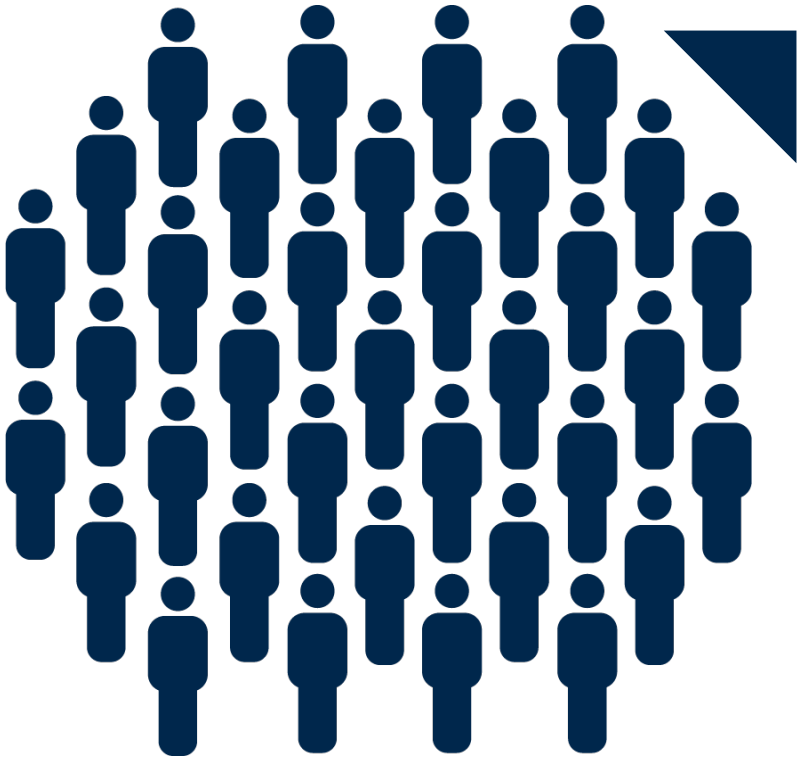
$P(x > \mu + k) = P_k$



$P(x < \mu + k) = 1 - P_k$



Normal Distribution



Population parameters

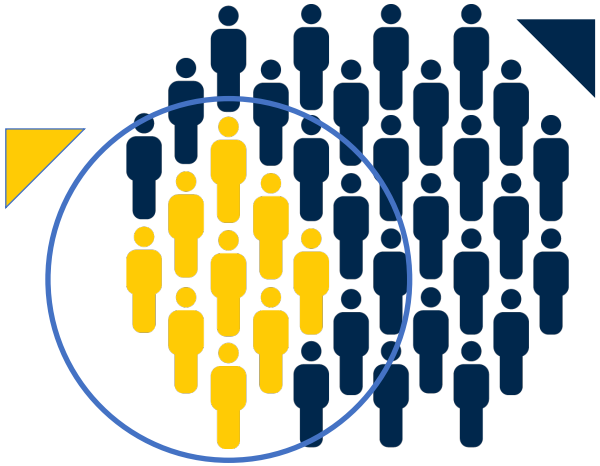
μ = population mean

σ = population standard deviation

Sample statistics

$\overline{X}$ = sample mean

S_X = sample standard deviation



Population parameters

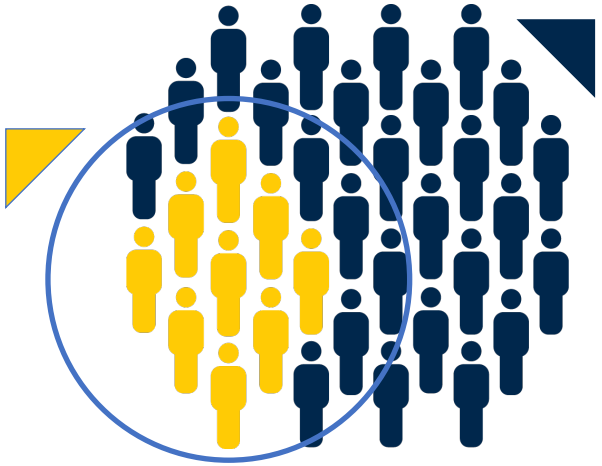
μ = population mean

σ = population standard deviation

Sample statistics

$\overline{X}$ = sample mean

S_X = sample standard deviation



Population parameters

μ = population mean

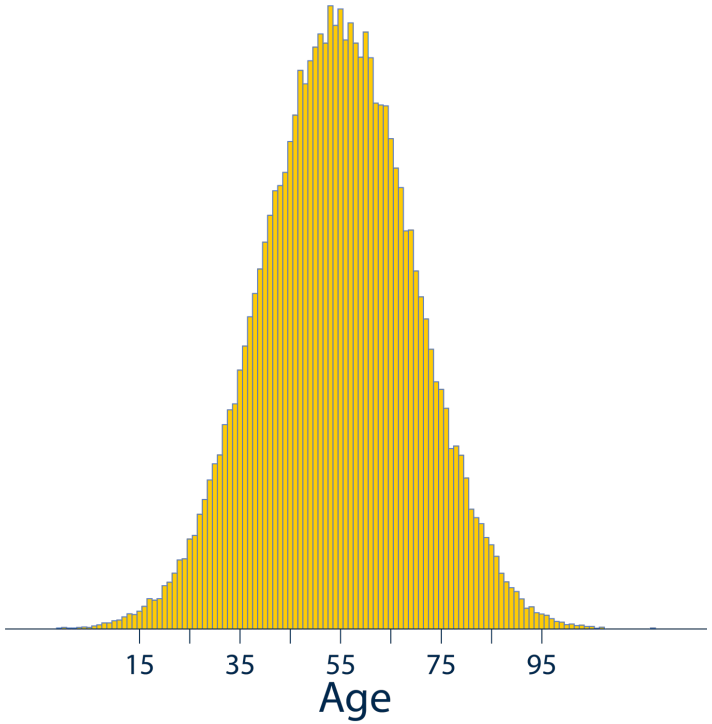
σ = population standard deviation

$\overline{X}$ is an estimate of μ

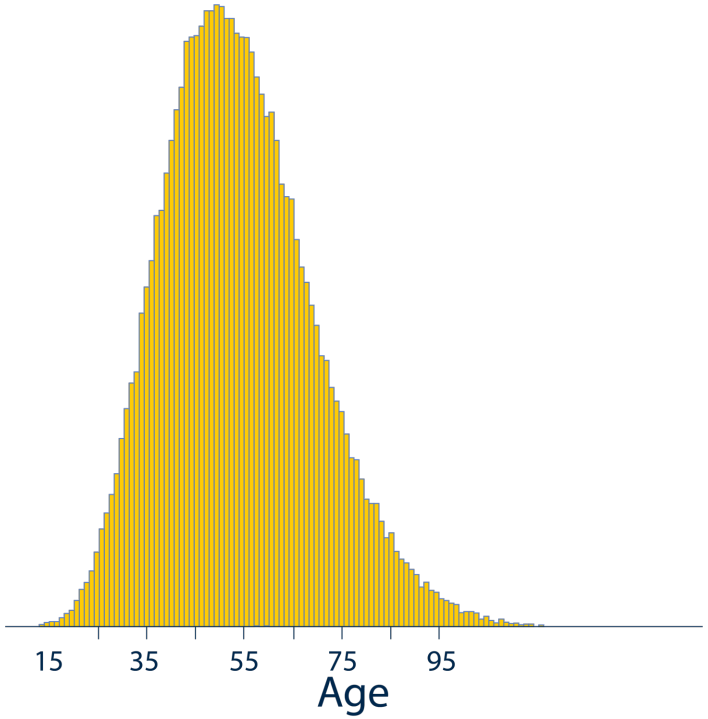
S_X is an estimate of σ

Many populations can have the same values for μ and σ, but have different shapes.

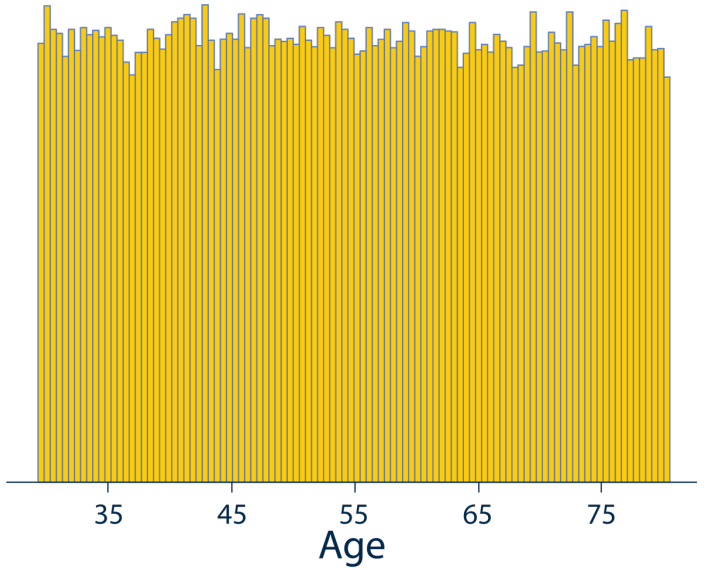
μ = 55
σ = 15



μ = 55
σ = 15

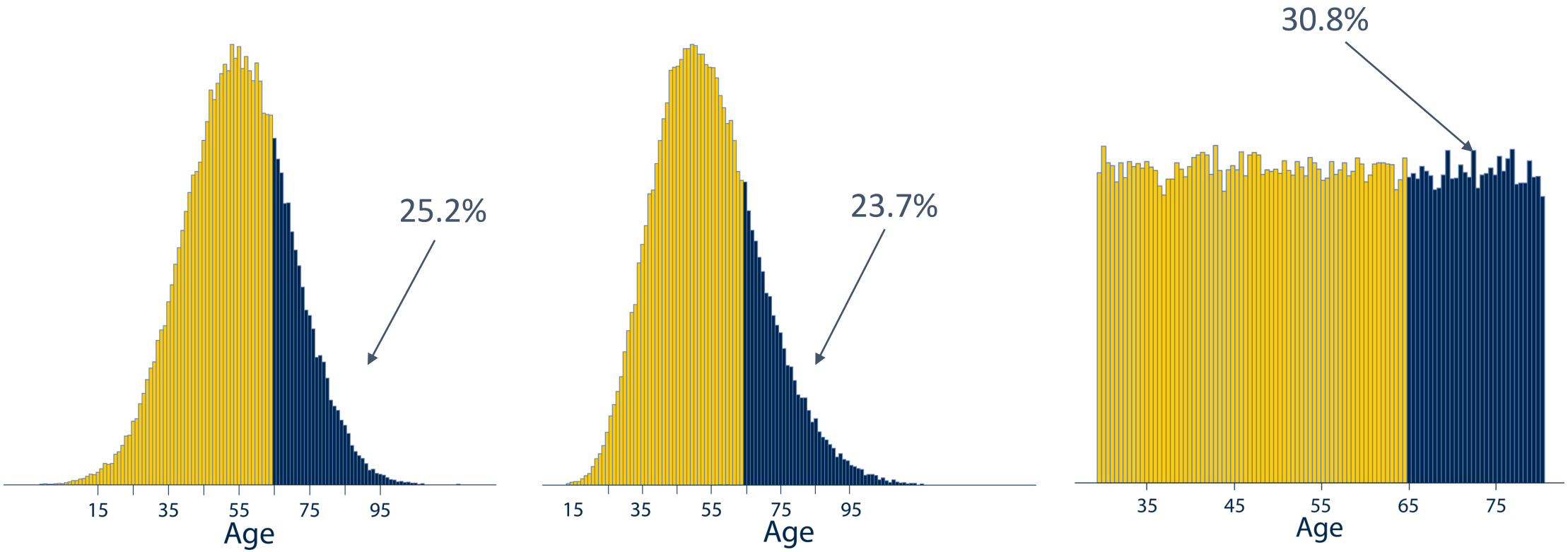


μ = 55
σ = 15

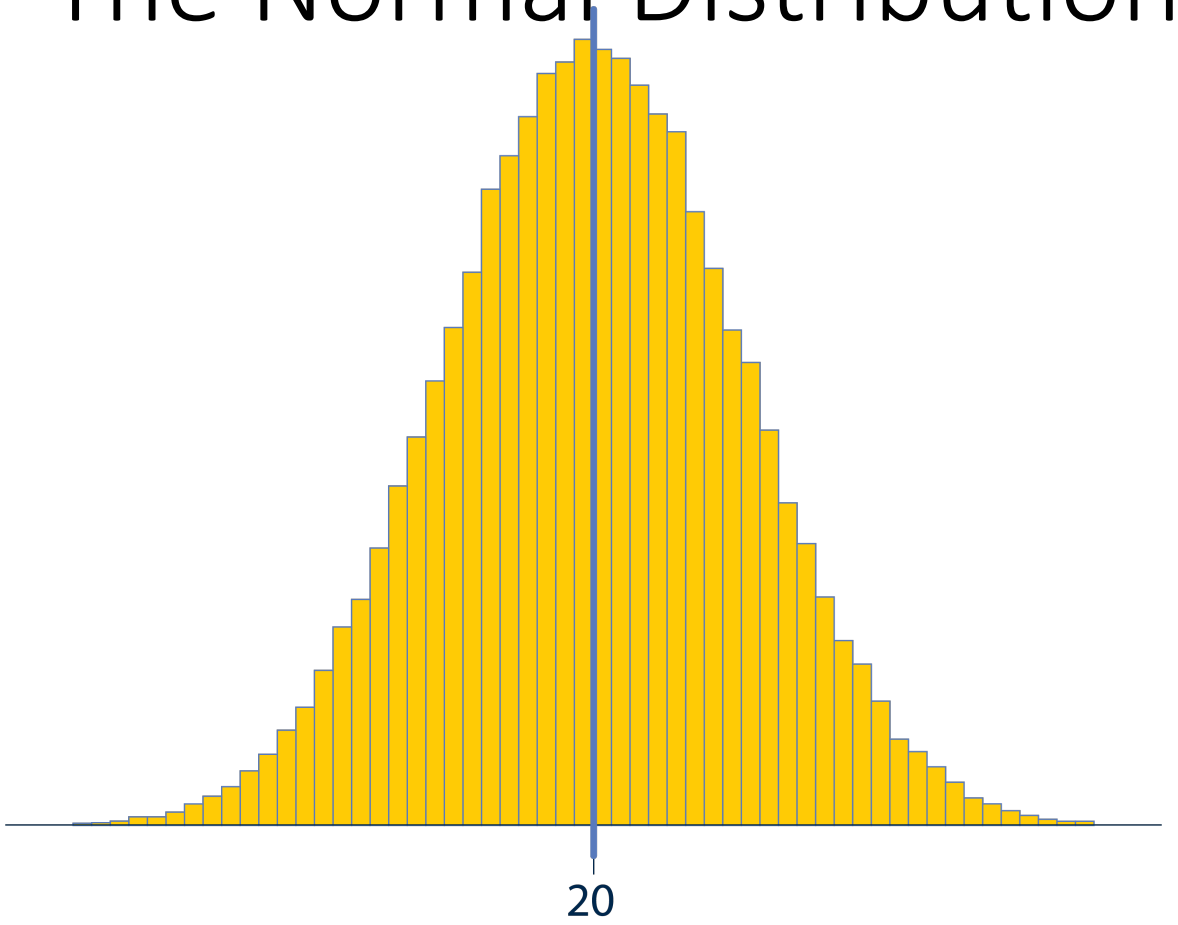


Different shapes lead to different probability statements!

= proportion of ages above 65

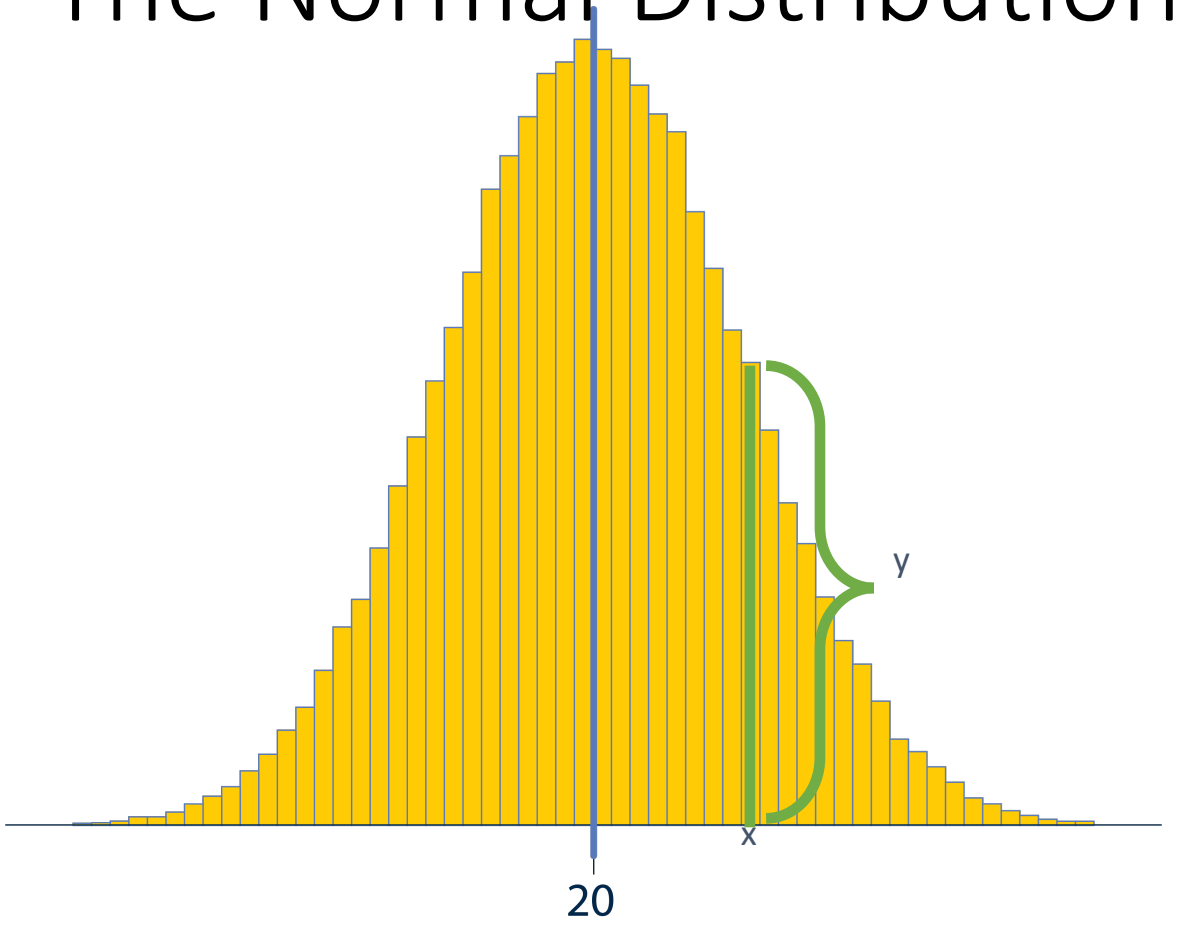


The Normal Distribution



- Bell-shaped
- Symmetric
  (mean = median)
- Mean: μ
  Standard Deviation: σ
- Notation: $N(\mu,\sigma)$

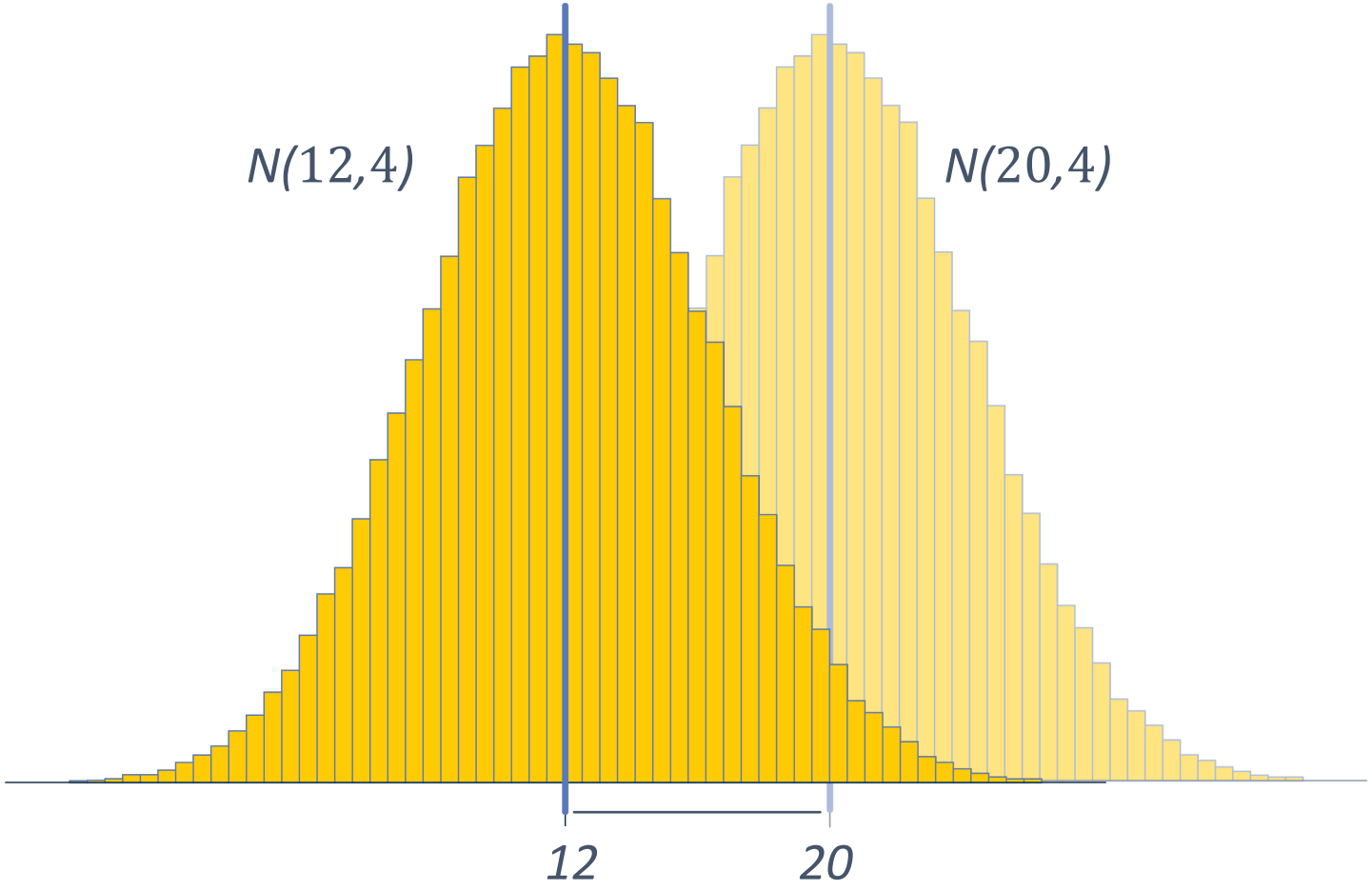
The Normal Distribution



$$y = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left[-\left(\frac{x-\mu}{\sigma}\right)^2\right]$$

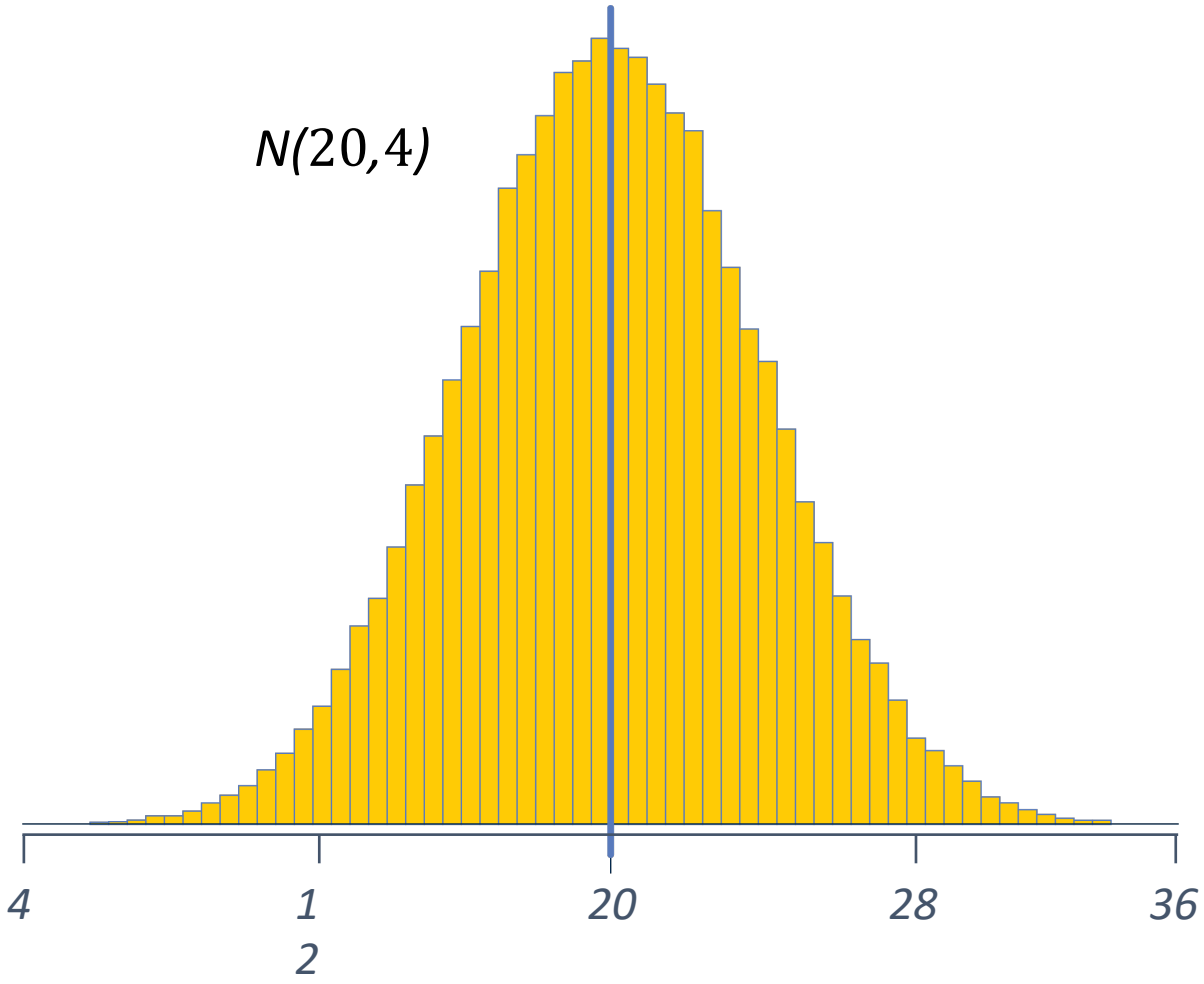
Fact 1

All normal distributions with the same value of σ have the same scale or spread



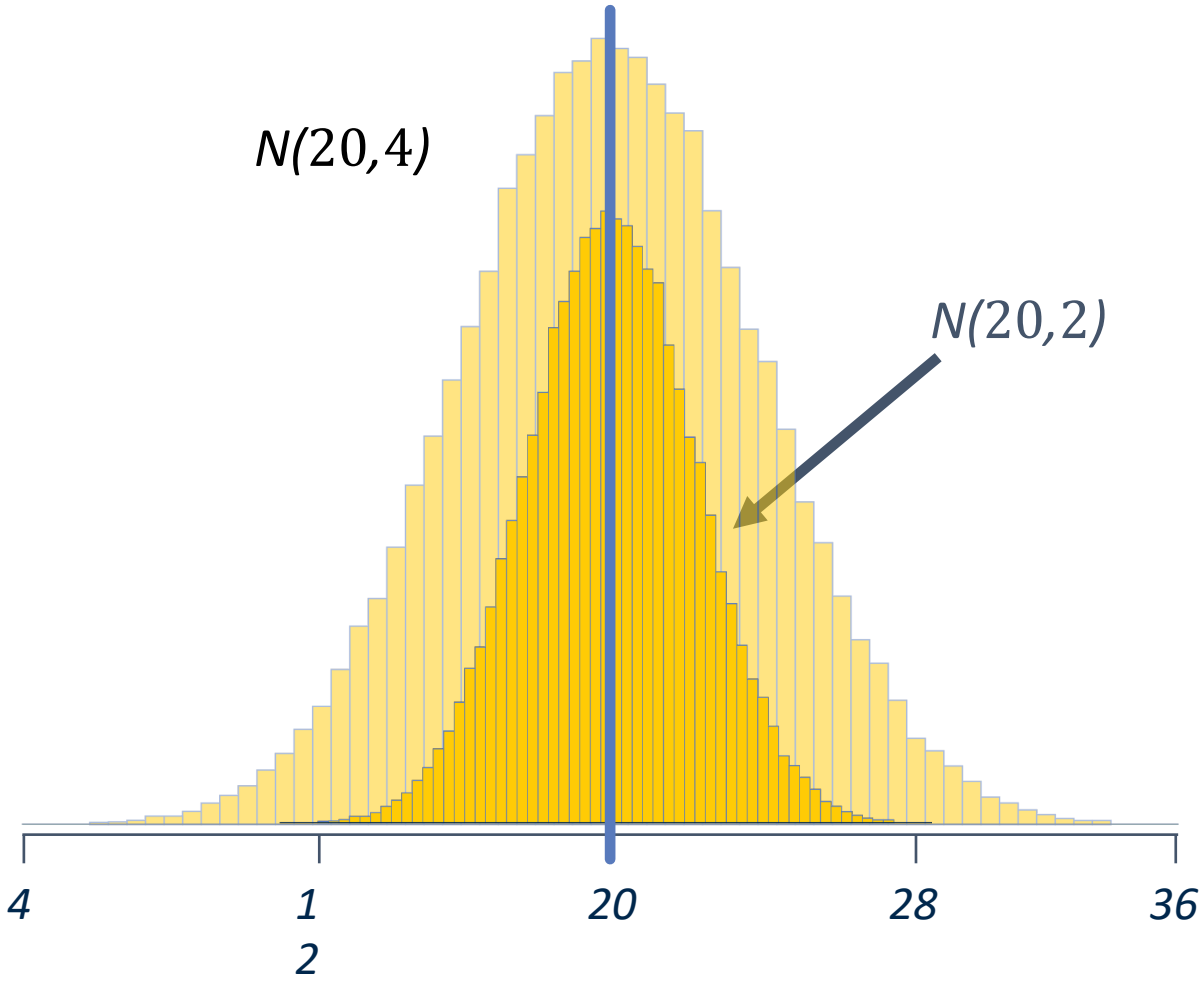
Fact 2

All normal distributions with the same value of μ have the same location



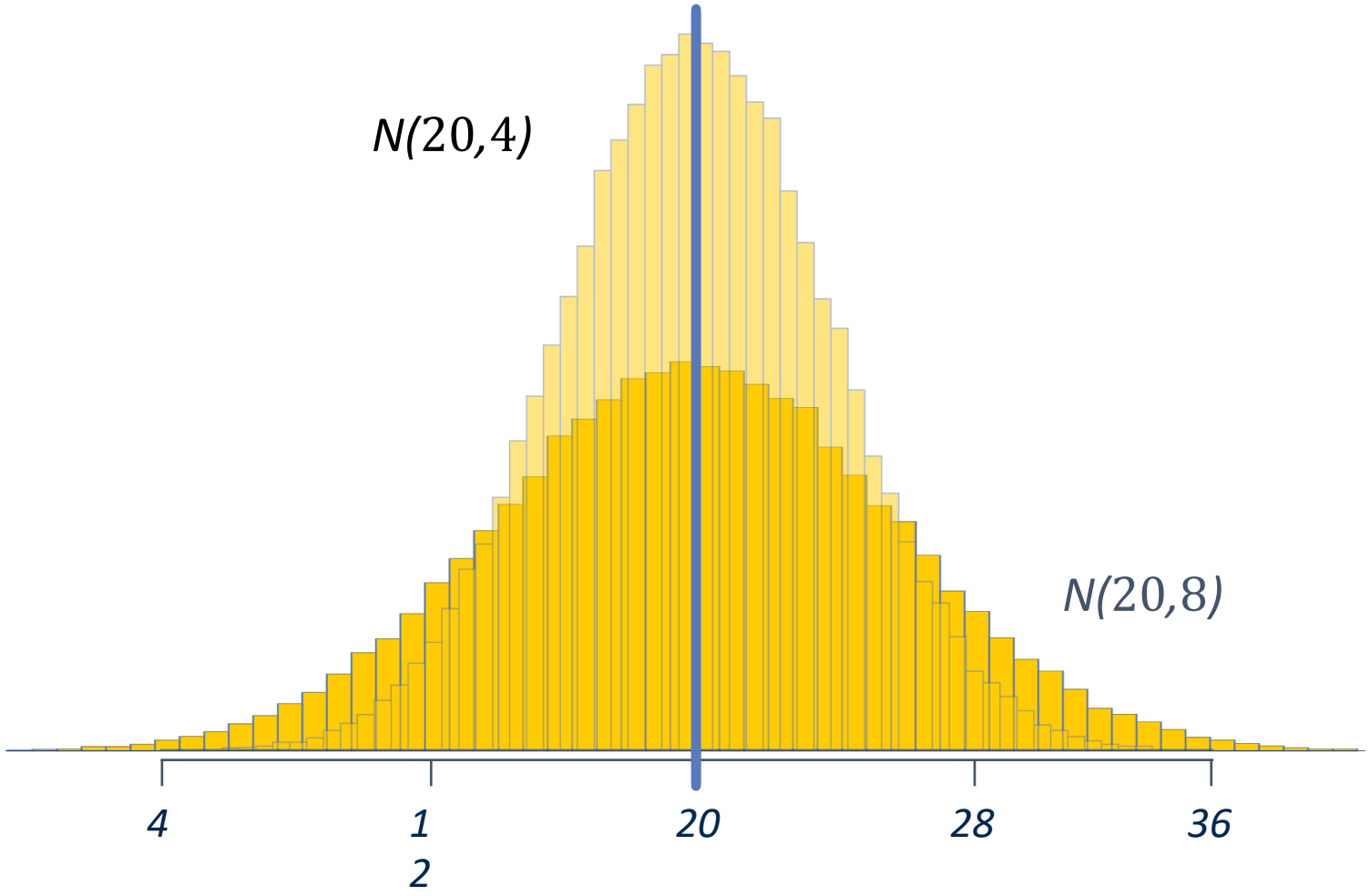
Fact 2

All normal distributions with the same value of μ have the same location



Fact 2

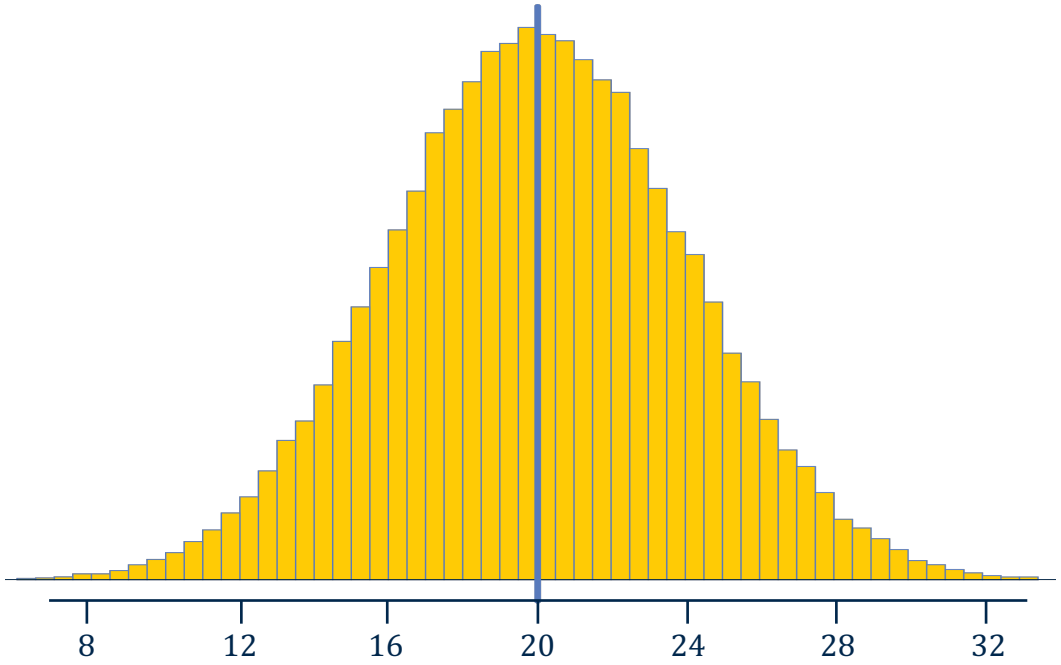
All normal distributions with the same value of μ have the same location



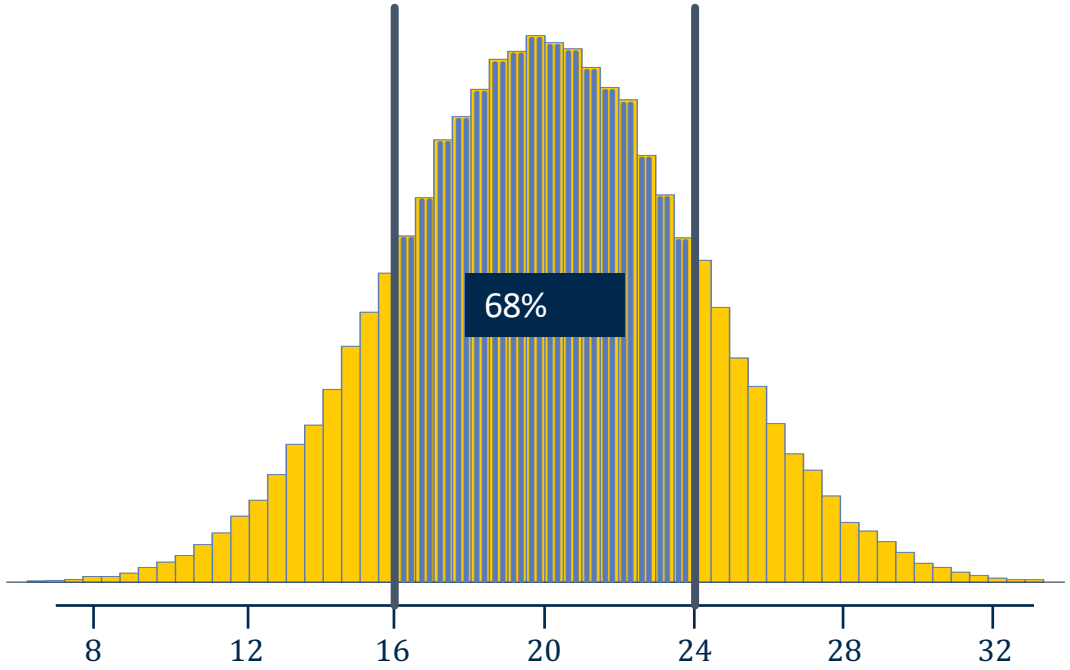
Fact 3a

68% of a normal distribution is within 1 standard deviation of its mean

$\mu = 20$ and $\sigma = 4$



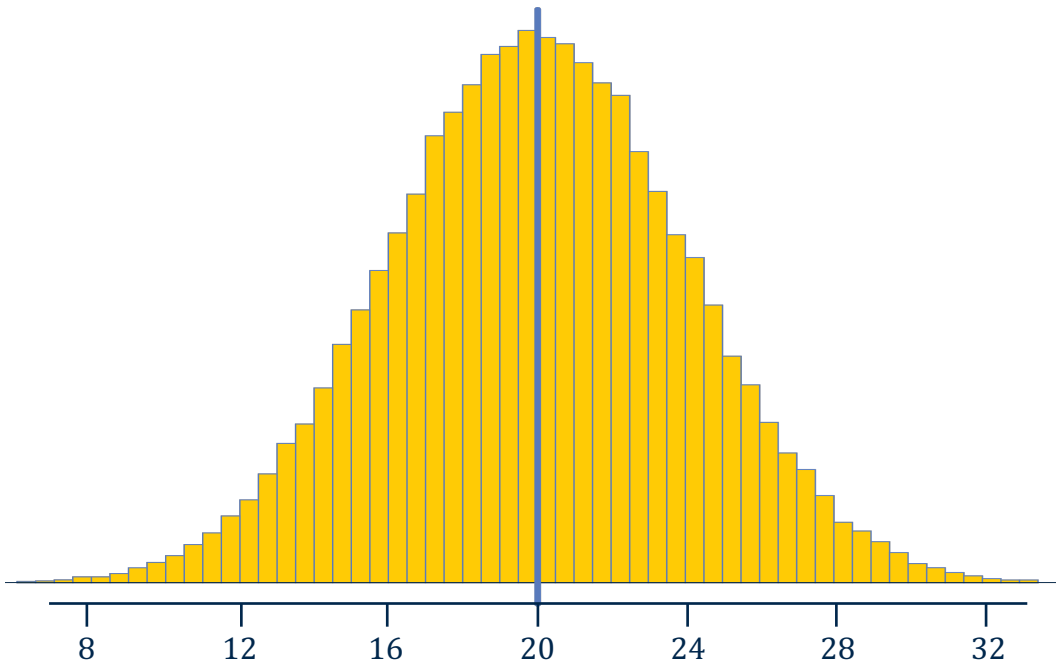
$(\mu - \sigma, \mu + \sigma) = (16, 24)$



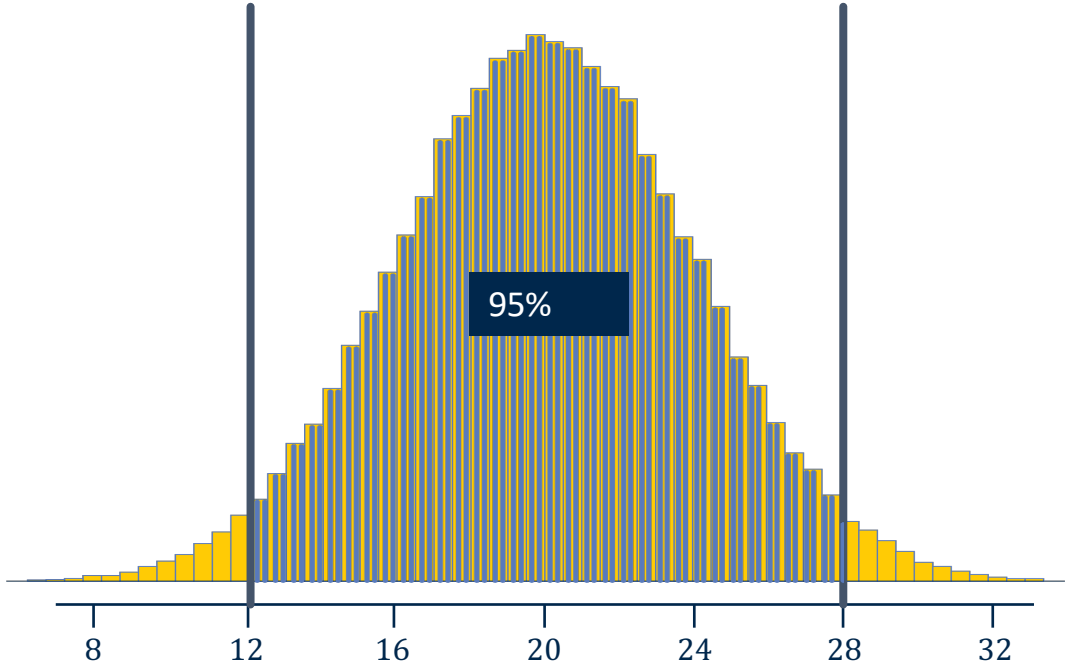
Fact 3b

95% of a normal distribution is within 2 standard deviations of its mean

$\mu = 20$ and $\sigma = 4$



$(\mu - 2\sigma, \mu + 2\sigma) = (12, 28)$

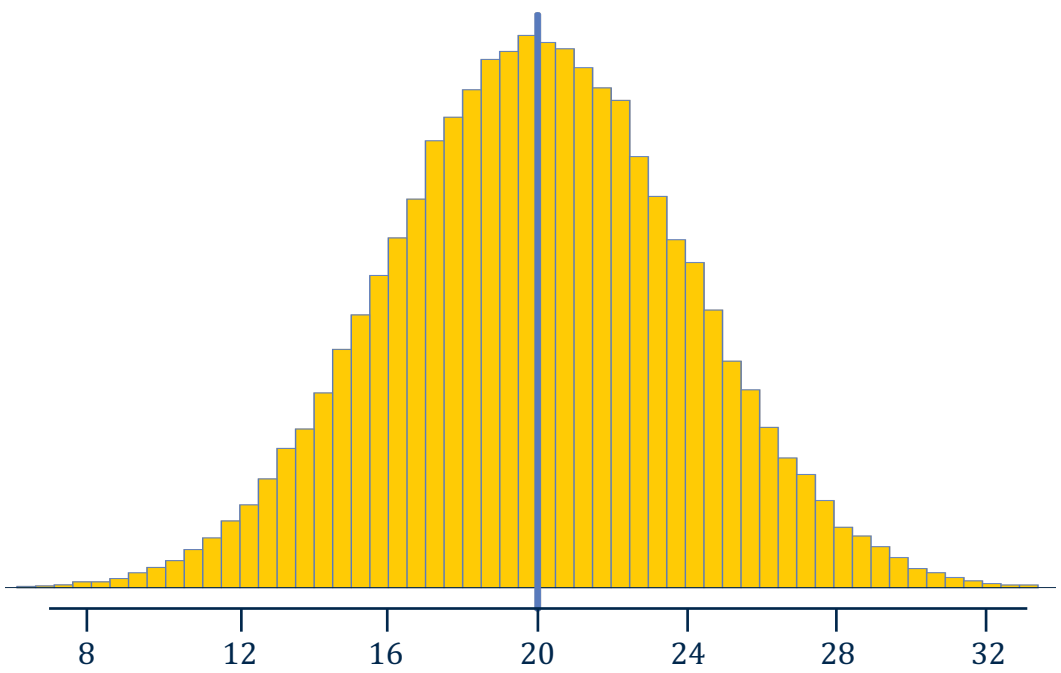


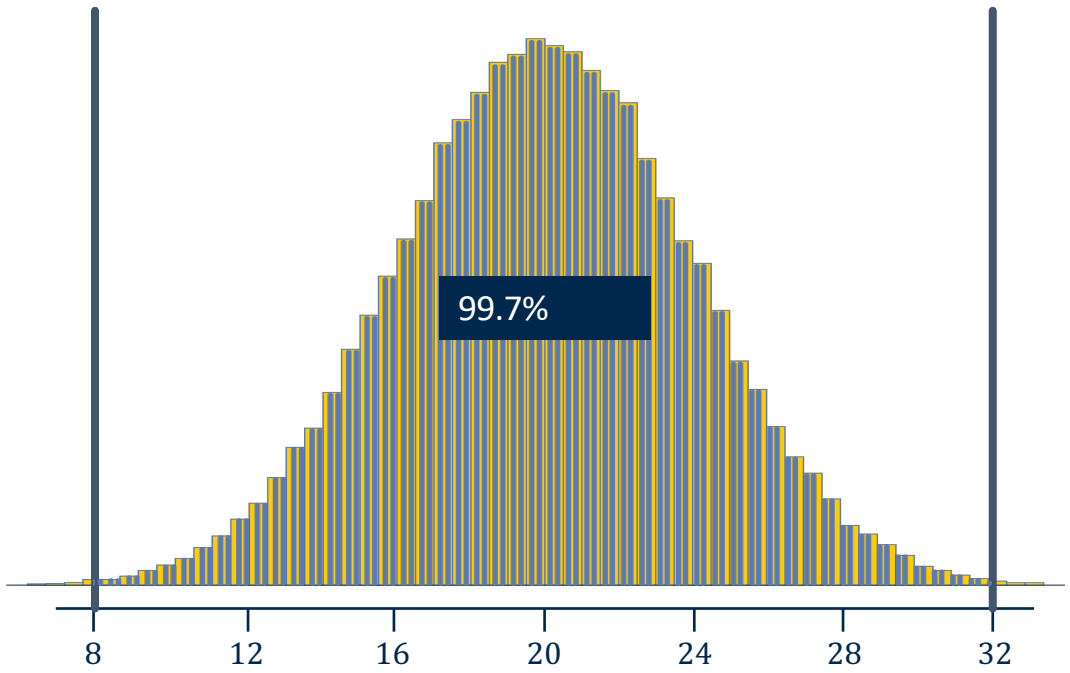
Fact 3c

Nearly 100% (actually 99.7%) of a normal distribution is within 3 standard deviations of its mean

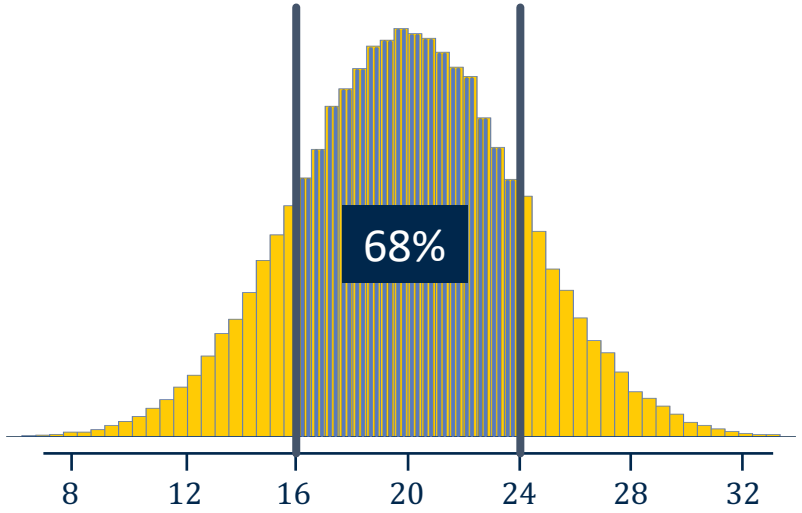
$\mu = 20$ and $\sigma = 4$

$(\mu - 3\sigma, \mu + 3\sigma) = (8, 32)$

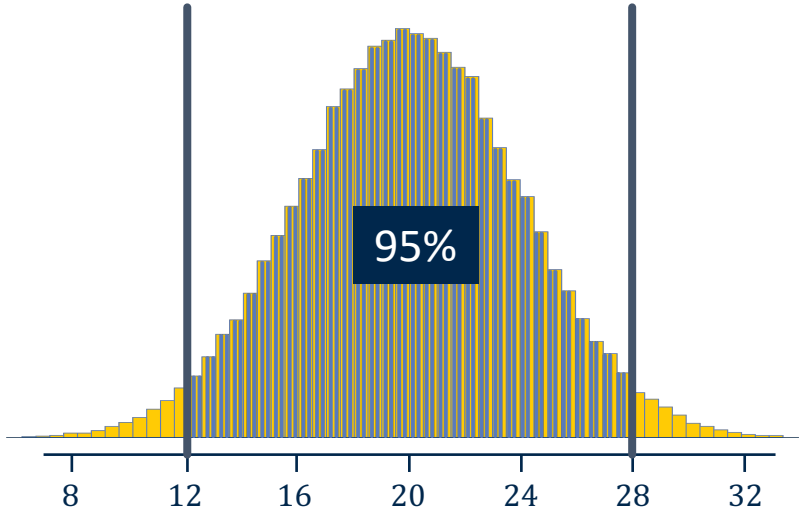




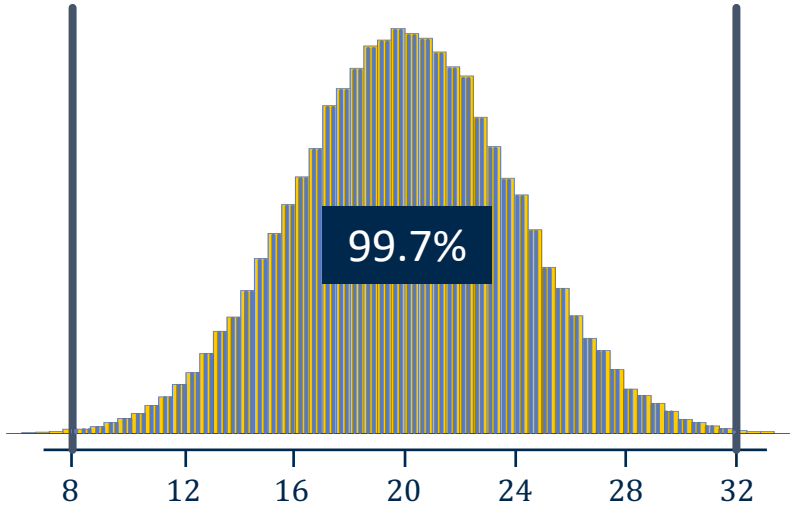
The “68-95-100” Heuristic



$(\mu - \sigma, \mu + \sigma) = (16, 24)$



$(\mu - 2\sigma, \mu + 2\sigma) = (12, 28)$



$(\mu - 3\sigma, \mu + 3\sigma) = (8, 32)$

Assuming the population is normal also gives us two facts about our data:

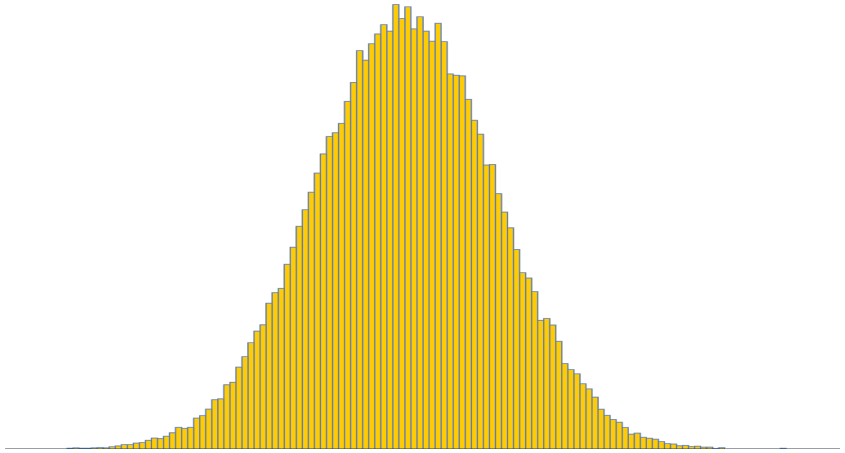
Fact 4a: The sample mean $\overline{X}$ is a "good" (unbiased) estimate for the population mean μ

Fact 4b: The sample standard deviation S_X is a "good" (unbiased) estimate for the population standard deviation σ

Assume we have a population with a **mean μ** and a **standard deviation σ**

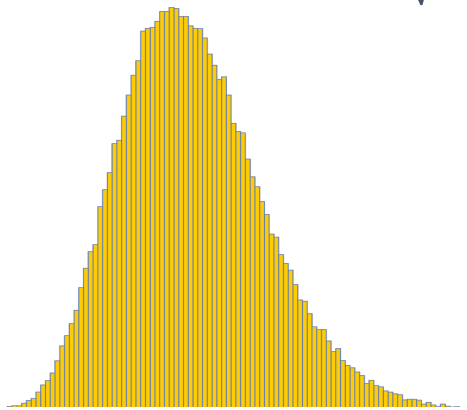
The population is **normal**

Fact 5: The **sampling distribution** of the sample mean $\overline{X}$ is a normal distribution with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$



The population is **not normal**

Fact 6: The **sampling distribution** of the sample mean $\overline{X}$ is *approximately* a normal distribution with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$



- Assume we have a population with a **mean μ** and a **standard deviation σ**

The population is **normal**

Fact 5: The **sampling distribution** of the sample mean $\overline{X}$ is a normal distribution with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$

The population is **not normal**

Fact 6: The **sampling distribution** of the sample mean $\overline{X}$ is *approximately* a normal distribution with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$

Central Limit Theorem (CLT)
We will introduce it in later slides

Z-scores

$$Z = \frac{X - \mu}{\sigma}$$

- The z-score is a standard normal variable, following normal distribution with mean zero and unit standard deviation
- The z-score is used to transform normally distributed variables with mean μ and SD σ into a variable that follows standard normal distribution
- Z ~ N(0,1)
- When we standardize by finding z-scores, we change the the normal distribution by moving the location (mean moves to zero) and changing the scale (SD moves to 1)
- Check Workshop!

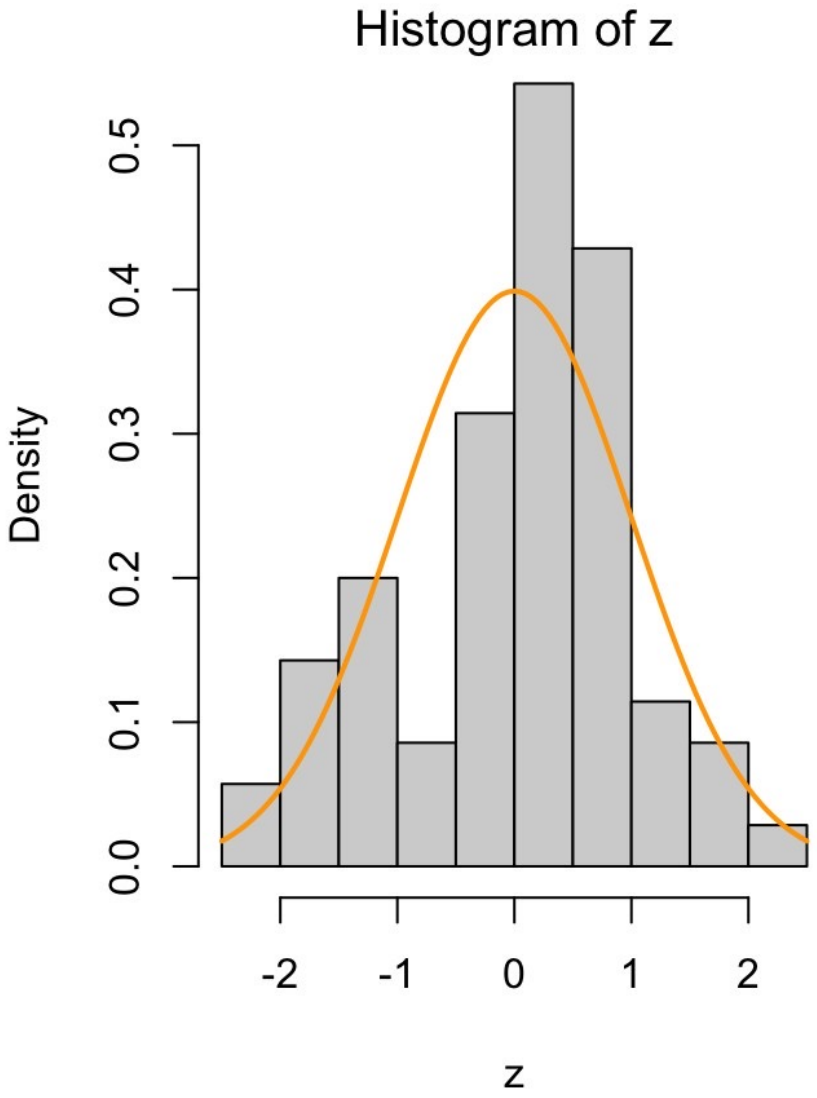
Quantile-Quantile plot (Q-Q plot)

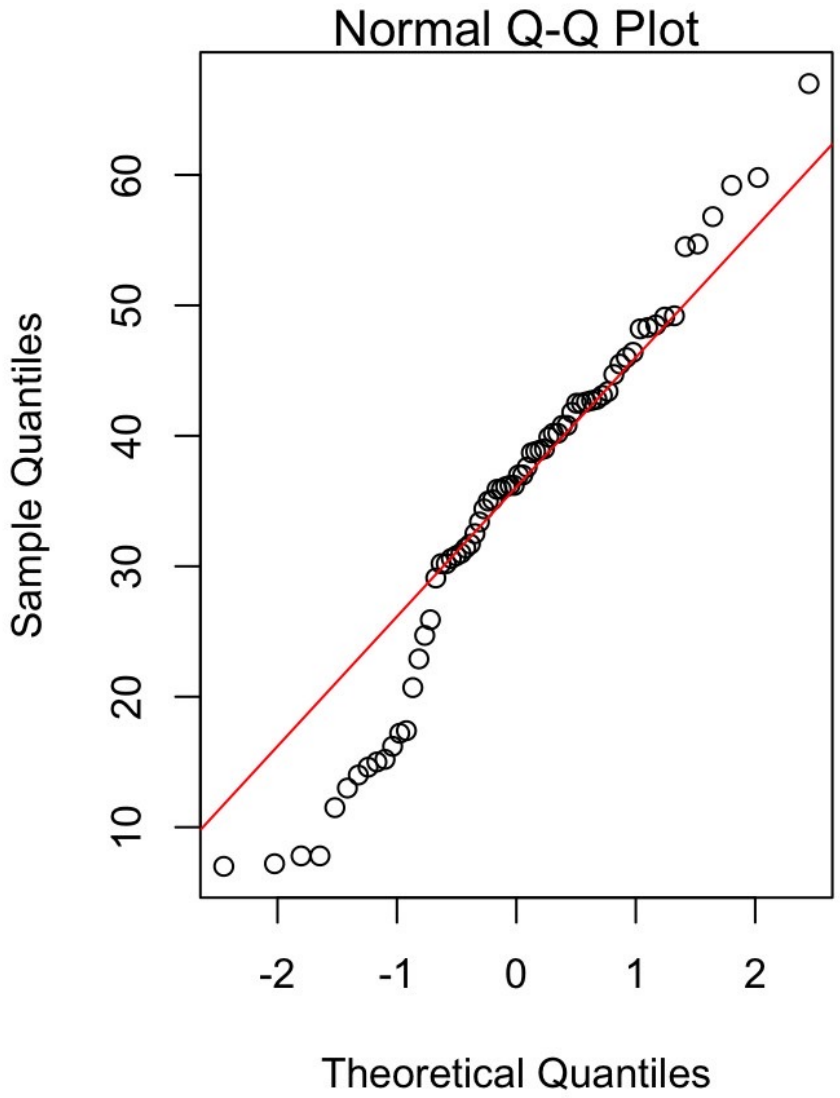
- Q-Q plot is designed to compare two probability distributions by plotting their quantiles against each other.
- Many statistical methods are developed under normality assumption
- Q-Q plot for normality check is called normal Q-Q plot
  - We obtain data and a statistical method with normality assumption will be used
  - We need to check if the method is ok to be applied to our data.
  - Try Q-Q plot which is a scatterplot for quantiles from data vs. the normal distribution (theoretical)

Normal Q-Q plot

Example data:

The average amount of rainfall in inches for each of the 70 states

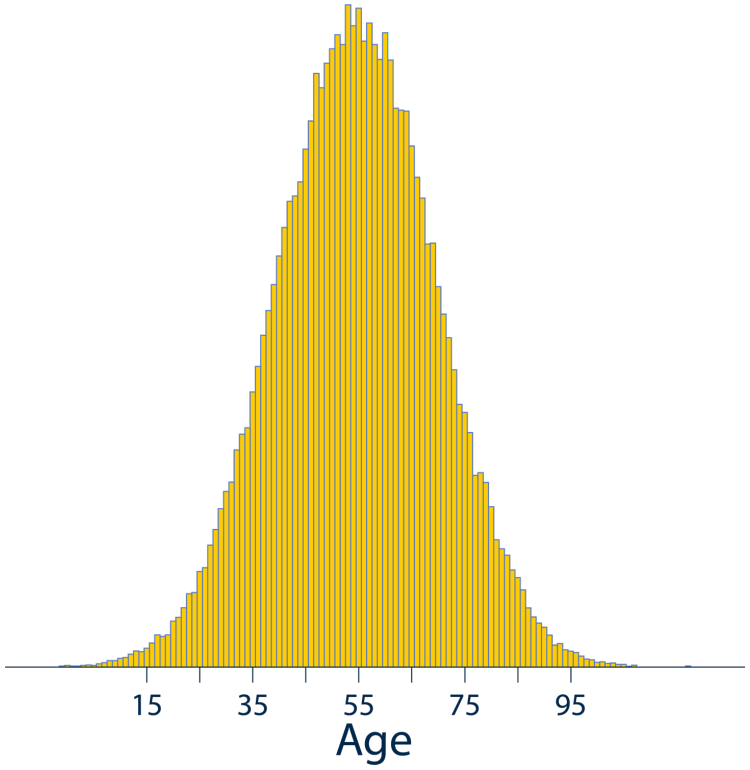




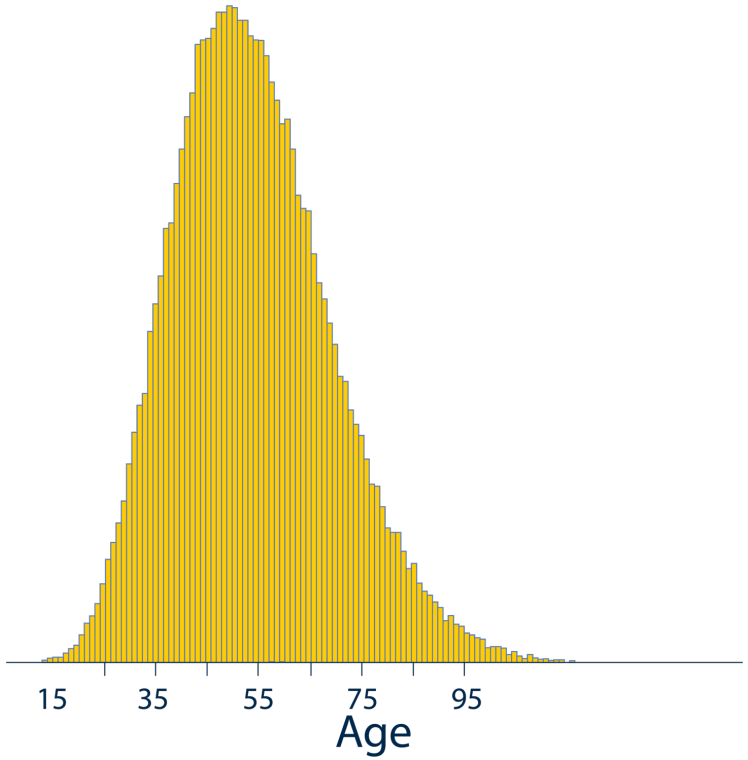
Central Limit Theorem

Recall the three population distributions we looked at earlier:

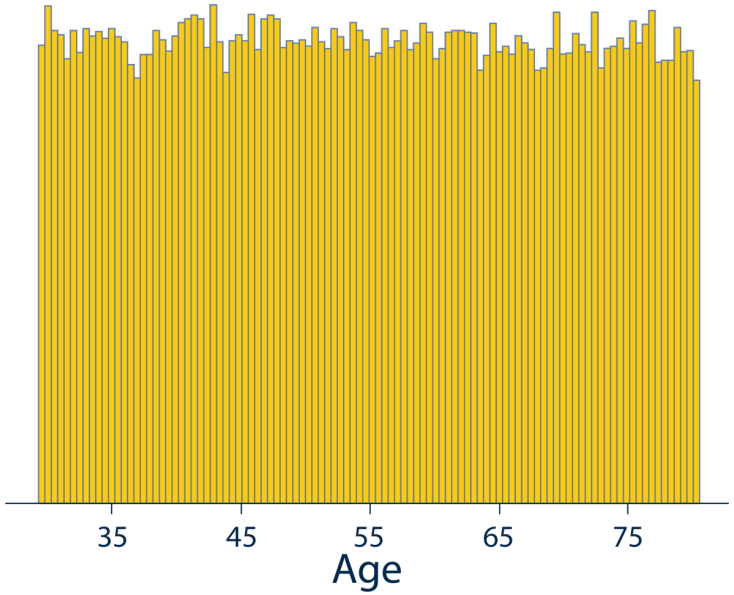
μ = 55
σ = 15



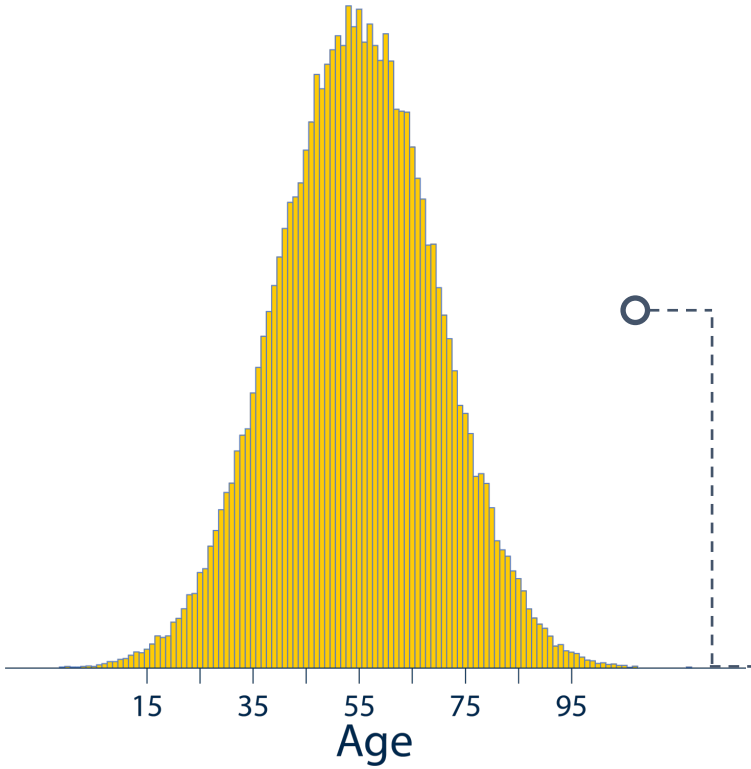
μ = 55
σ = 15



μ = 55
σ = 15



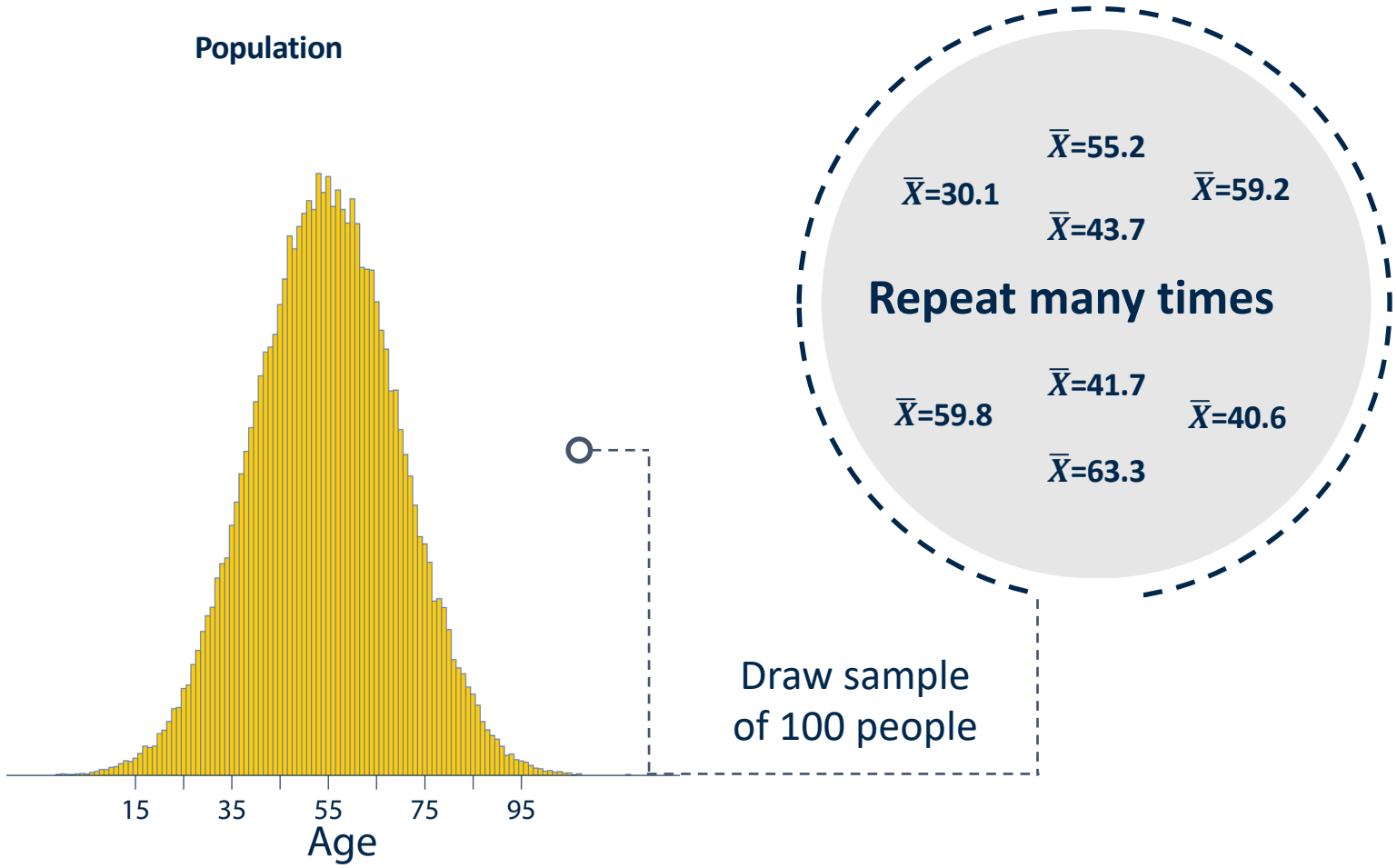
Population

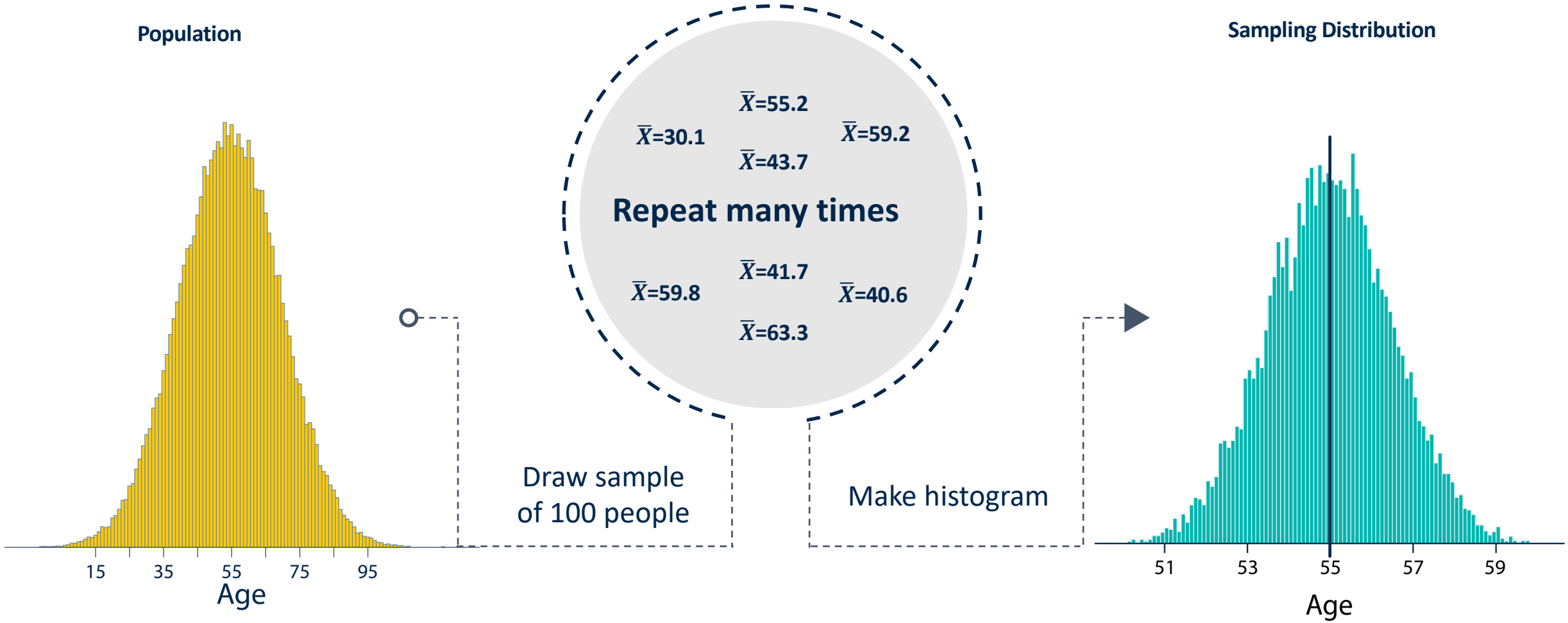


Draw sample
of 100 people

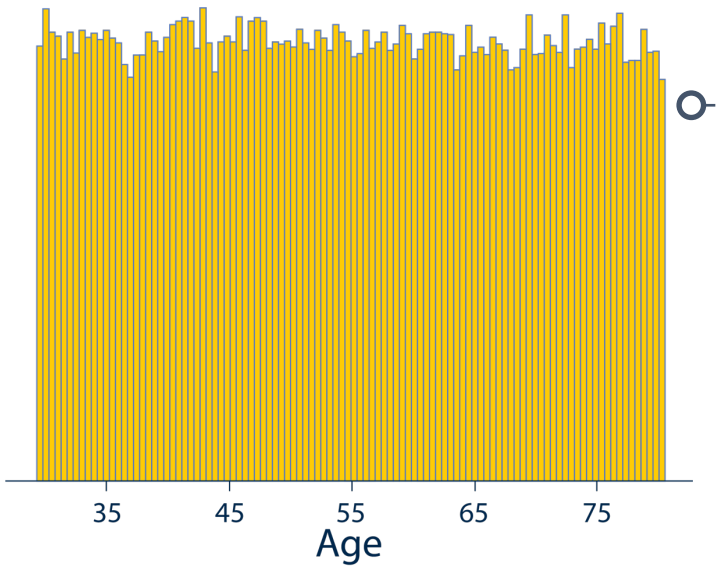
$\overline{X}$=30.1

Compute sample mean


Population
15
35
55
75
95
Age
Draw sample
of 100 people
$\bar{X}$=55.2
$\bar{X}$=30.1
$\bar{X}$=59.2
$\bar{X}$=43.7
Repeat many times
$\bar{X}$=41.7
$\bar{X}$=59.8
$\bar{X}$=40.6
$\bar{X}$=63.3


Population
15
35
55
75
95
Age
$\bar{X}$=30.1
$\bar{X}$=55.2
$\bar{X}$=59.2
$\bar{X}$=43.7
Repeat many times
$\bar{X}$=59.8
$\bar{X}$=41.7
$\bar{X}$=40.6
$\bar{X}$=63.3
Draw sample
of 100 people
Make histogram
Sampling Distribution
51
53
55
57
59
Age

Population



Draw sample
of 100 people

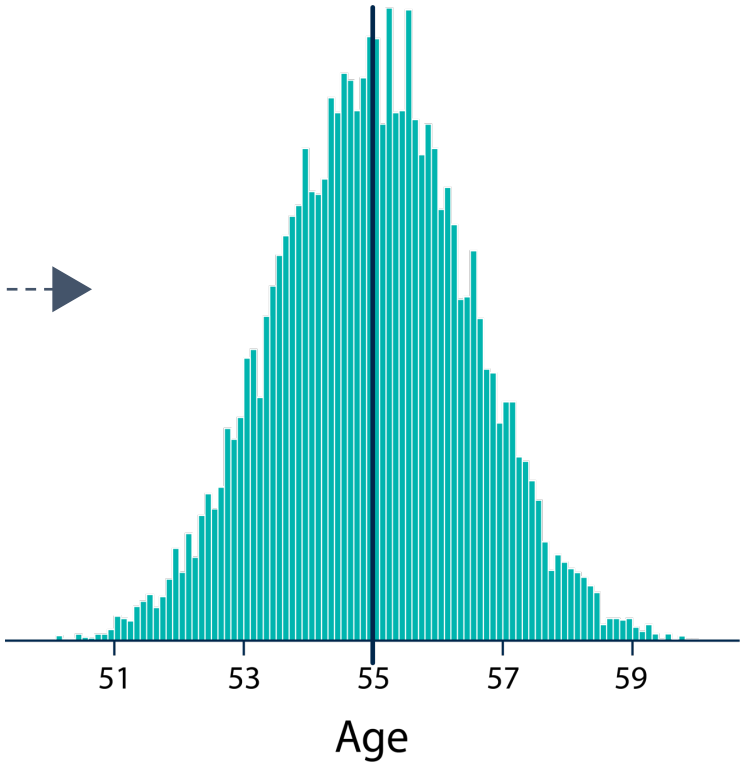
$\bar{X}$=55.2
$\bar{X}$=30.1
$\bar{X}$=59.2
$\bar{X}$=43.7

Repeat many times

$\bar{X}$=41.7
$\bar{X}$=59.8
$\bar{X}$=40.6
$\bar{X}$=63.3

Make histogram

Sampling Distribution



Central Limit Theorem

- The CLT is a very powerful theorem.
  - The **population** distribution **does not** have to be normal!
  - Any shape for the **population** will lead to a normally shaped distribution of **sample means**, if we have “enough” data!

Central Limit Theorem

Let $x_1, x_2, \ldots, x_n$ be the random sample n from a population distribution with mean μ and standard deviation σ. Then the sampling distribution of $\overline{x}$ approaches to normal distribution with mean μ and standard deviation $\sigma/\sqrt{n}$ as $n \to \infty$.

- The mean of the sampling distribution is identical to the population mean
- The standard deviation of the distribution of the sample mean is the population standard deviation divided by square root of the sample size
- For the sample size large enough, the shape of the sampling distribution is approximately normal → holds for any population distributions with finite variances

Confidence interval

- **By the CLT:**

Sample means have a normal distribution with mean μ and standard deviation $\sigma/\sqrt{n}$

By the properties of a normal distribution:

95% of sample means cover two standard deviations on either side of their mean.

Thus, 95% of sample means will lie in the interval

$$(\mu - \frac{2\sigma}{\sqrt{n}}, \mu + \frac{2\sigma}{\sqrt{n}})$$

In other words

$$\text{Prob}\left(\mu - \frac{2\sigma}{\sqrt{n}} \leq \overline{X} \leq \mu + \frac{2\sigma}{\sqrt{n}}\right) = 0.95$$

Confidence Intervals

We now do some algebra to get μ in-between the "$\leq$" symbols:

$$\text{Prob}\left(\mu - \frac{2\sigma}{\sqrt{n}} \leq \overline{X} \leq \mu + \frac{2\sigma}{\sqrt{n}}\right) = 0.95$$

Confidence Intervals

We now do some algebra to get μ in-between the "$\leq$" symbols:

$$\text{Prob}\left(\mu - \frac{2\sigma}{\sqrt{n}} \leq \overline{X} \leq \mu + \frac{2\sigma}{\sqrt{n}}\right) = 0.95$$

$$\text{Prob}\left(-\frac{2\sigma}{\sqrt{n}} \leq \overline{X} - \mu \leq \frac{2\sigma}{\sqrt{n}}\right) = 0.95$$

Subtract μ

Confidence Intervals

We now do some algebra to get μ in-between the "$\leq$" symbols:

$$\text{Prob}\left(\mu - \frac{2\sigma}{\sqrt{n}} \leq \overline{X} \leq \mu + \frac{2\sigma}{\sqrt{n}}\right) = 0.95$$

$$\text{Prob}\left(-\frac{2\sigma}{\sqrt{n}} \leq \overline{X} - \mu \leq \frac{2\sigma}{\sqrt{n}}\right) = 0.95$$

Subtract $\overline{X}$

$$\text{Prob}\left(-\overline{X} - \frac{2\sigma}{\sqrt{n}} \leq -\mu \leq -\overline{X} + \frac{2\sigma}{\sqrt{n}}\right) = 0.95$$

Confidence Intervals

We now do some algebra to get μ in-between the "$\leq$" symbols:

$$\text{Prob}\left(\mu - \frac{2\sigma}{\sqrt{n}} \leq \overline{X} \leq \mu + \frac{2\sigma}{\sqrt{n}}\right) = 0.95$$

$$\text{Prob}\left(-\frac{2\sigma}{\sqrt{n}} \leq \overline{X} - \mu \leq \frac{2\sigma}{\sqrt{n}}\right) = 0.95$$

Multiply by -1

$$\text{Prob}\left(-\overline{X} - \frac{2\sigma}{\sqrt{n}} \leq -\mu \leq -\overline{X} + \frac{2\sigma}{\sqrt{n}}\right) = 0.95$$

$$\text{Prob}\left(\overline{X} - \frac{2\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + \frac{2\sigma}{\sqrt{n}}\right) = 0.95$$

Confidence Intervals

We now do some algebra to get μ in-between the "$\leq$" symbols:

$$\text{Prob}\left(\mu - \frac{2\sigma}{\sqrt{n}} \leq \overline{X} \leq \mu + \frac{2\sigma}{\sqrt{n}}\right) = 0.95$$

$$\text{Prob}\left(-\frac{2\sigma}{\sqrt{n}} \leq \overline{X} - \mu \leq \frac{2\sigma}{\sqrt{n}}\right) = 0.95$$

$$\text{Prob}\left(-\overline{X} - \frac{2\sigma}{\sqrt{n}} \leq -\mu \leq -\overline{X} + \frac{2\sigma}{\sqrt{n}}\right) = 0.95$$

$$\text{Prob}\left(\overline{X} - \frac{2\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + \frac{2\sigma}{\sqrt{n}}\right) = 0.95$$

We have changed the interval to now describe a range of values of μ for a given value of $\overline{X}$

Confidence Intervals

This interval still has one unknown value σ.

- We use our best guess, which is the sample standard deviation S_X
- Recall that $S_X/\sqrt{n}$ is known as the **standard error of the mean (SEM)**

Thus, we have a **95% confidence interval** for μ:

$$\left(\overline{X} - 2\frac{S_X}{\sqrt{n}}\ ,\ \overline{X} + 2\frac{S_X}{\sqrt{n}}\right)$$

$$= \left(2350 - 2\frac{570}{\sqrt{100}}\ , 2350 + 2\frac{570}{\sqrt{100}}\right)$$

$$= (2236, 2464)$$

100(1-α)% CI for mean

Confidence interval for μ when σ is known.

- We want to evaluate how good is the point estimator for μ ,

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

- From the CLT, $\bar{x} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$ and the z-statistic (z-score) can be constructed by $z = \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}}$, which follows N(0,1) approximately.
- Find 95% = 100(1- α)% CI,

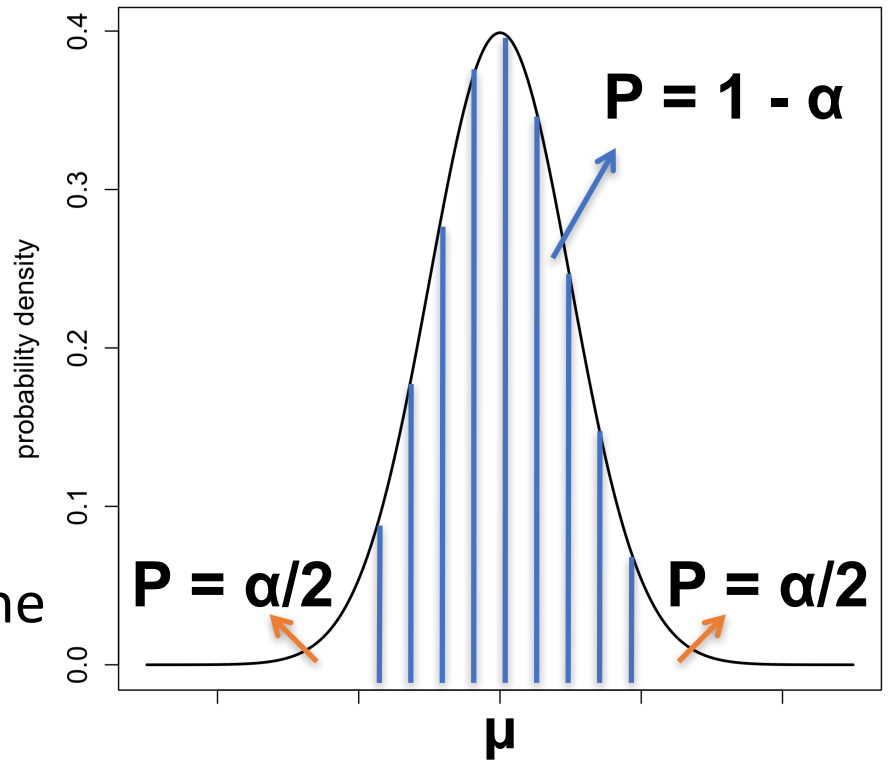
$$\mathrm{P}(-\mathrm{z}_{\frac{\alpha}{2}} < \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}} < \mathrm{z}_{\frac{\alpha}{2}}) = 1- \alpha$$

Where $\mathrm{z}_{\frac{\alpha}{2}}$ is the quantile of probability 1- $\frac{\alpha}{2}$

CI is the bounds for the parameter of interest, we need to rewrite:

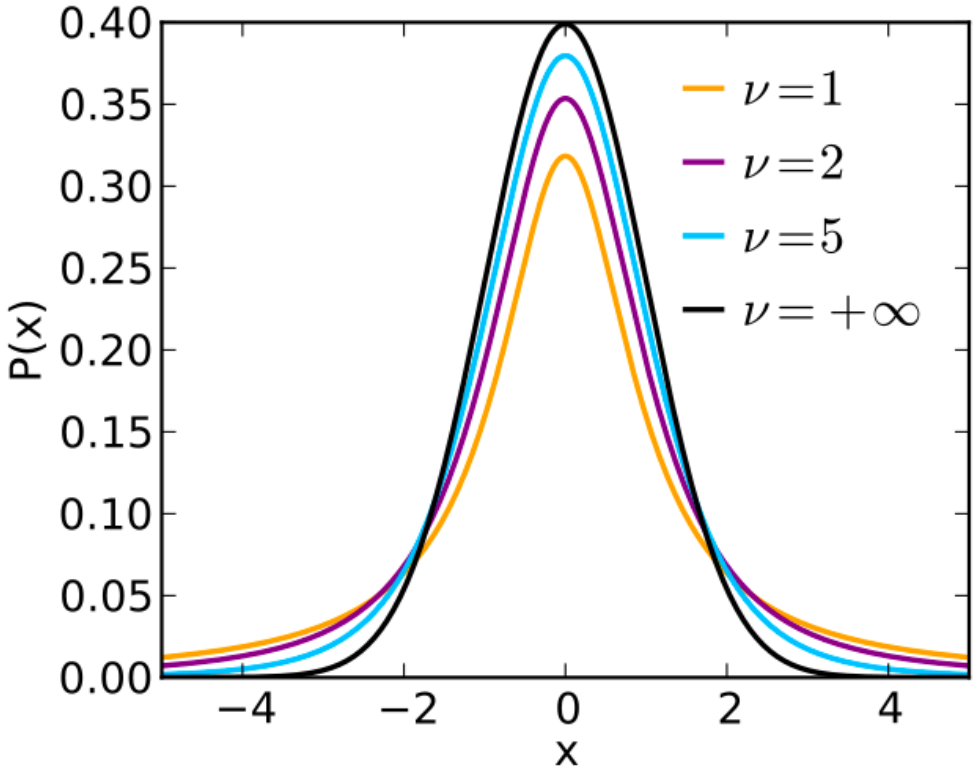
$$P\left(-z_{\frac{\alpha}{2}} < \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}} < z_{\frac{\alpha}{2}}\right) = P\left(\bar{x} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right)$$

- Step 1: Set the significance level α (usually 0.05)
- Step 2: Find the 100(1-$\frac{\alpha}{2}$) % quantile from standard normal distribution. E.g., In R, qnorm(p=0.975)
- Be careful interpreting the interval!
  - If we select a number m of random samples from the population and use them to calculate m different Cis for μ, then approximately 95% of the intervals would cover the true Population mean μ, and 5% would not.
  - WRONG: There is 95% chance that μ lies in the In the interval. μ is fixed!



t-distribution

Student's t-distribution



- The Student's t-distribution to account for the additional variability due to estimating σ with s
- The t distribution looks a lot like the normal except that it has fatter tails
- The parameter for t distribution is called degrees of freedom (df)
- As the df (denoted by ν in the figure)gets bigger, the t distribution looks more and more like the normal

t distribution for CI

- The df measure the amount of information available in the data to estimate σ
- The statistic $t = \frac{\bar{x}-\mu}{s/\sqrt{n}}$ has a t distribution with $n-1$ df (denoted by t_{n-1}). We use 1 df by estimating the sample mean $\bar{x}$
- Thus, n gets larger → s gets to be a better estimate of σ → the distribution of the t statistic looking more like the normal
- With large enough n, normal approximation can be used to construct CI.
- CI from t distribution is wider, accounting for the uncertainty on σ

What if σ is unknown?

- We use CI given by

$$\mathrm{P}\left(\bar{x} - \mathrm{t}_{n-1,\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + \mathrm{t}_{n-1,\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right)$$

Where $\mathrm{t}_{n-1,\frac{\alpha}{2}}$ is the quantile of probability 1- $\frac{\alpha}{2}$ from t_{n-1}

- In R, use qt(p=0.975,df=n-1)

workshop slides: https://github.com/shangll123/GS011143

Enjoy R programming and workshop!
We will discuss statistics using R in lectures 2&3