RenR 480/711

Standard error, Normal & t-distribution,

confidence intervals, z-scores and t-values

Recap

Descriptive statistics

→ summary of data (graphical and/or numerical)

Distribution

→ Probability associated with each possible value of a variable

Parameter

→ Population characteristic

Statistic

→ Estimate of parameter using a sample

Recap

Bias

→ systematic deviation of statistic from parameter

Error

→ random deviation of a statistic from parameter

Inferential statistics

→ making predictions, drawing conclusions from data that is subject to random variation

Standard error

→ standard deviation of any statistic

New Vocabulary

New:

- Alpha level

Predetermined probability at which we make a decision

- P-value

The probability that an observed difference is due to random chance

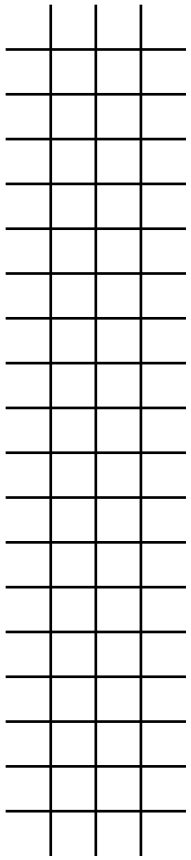
Degrees of freedom

The number of observations in your data that are free to vary when calculating a statistic

Excursion: Degrees of freedom

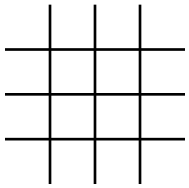
Degrees of freedom are good!

Good



Large n

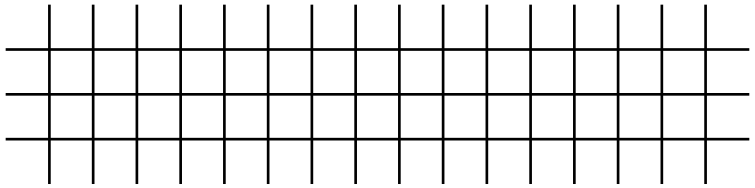
Bad



# of variables
=
# of observations

Equal to calculating mean from 1 data point

Hopeless



# of variables >
# of observations

The larger the ratio of observations to variables, the more complex analyses are possible. Complexity costs df.

Intro to inferential statistics

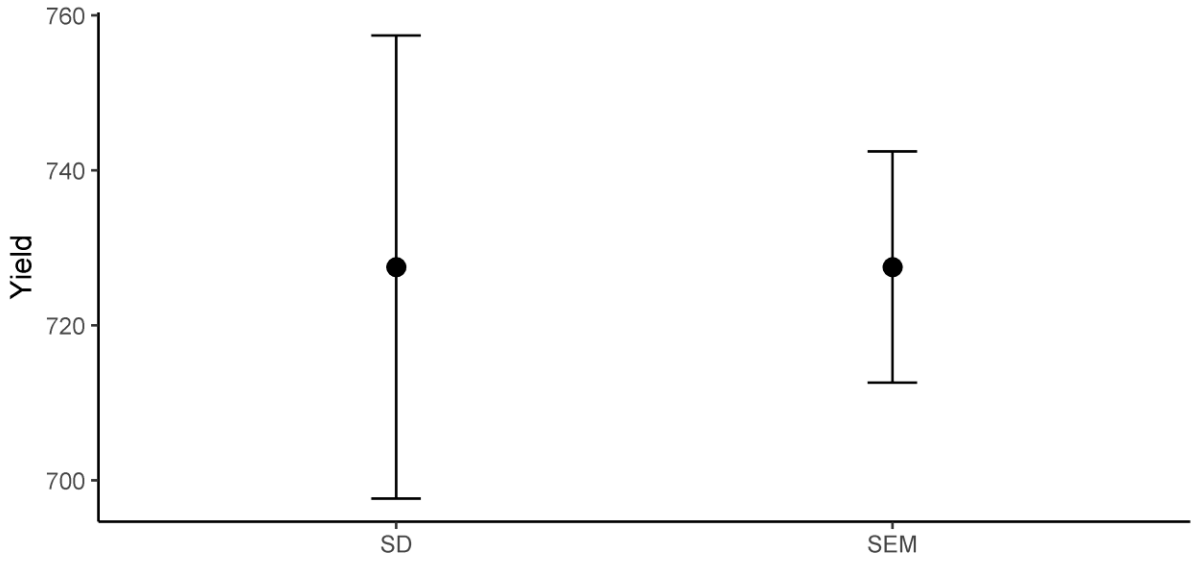
What's the difference between standard deviation and standard error?

Descriptive

Variability of my data

Inferential

Variability of my mean estimate



Meaning:

~68% of the data is within $\pm$ 1 SD

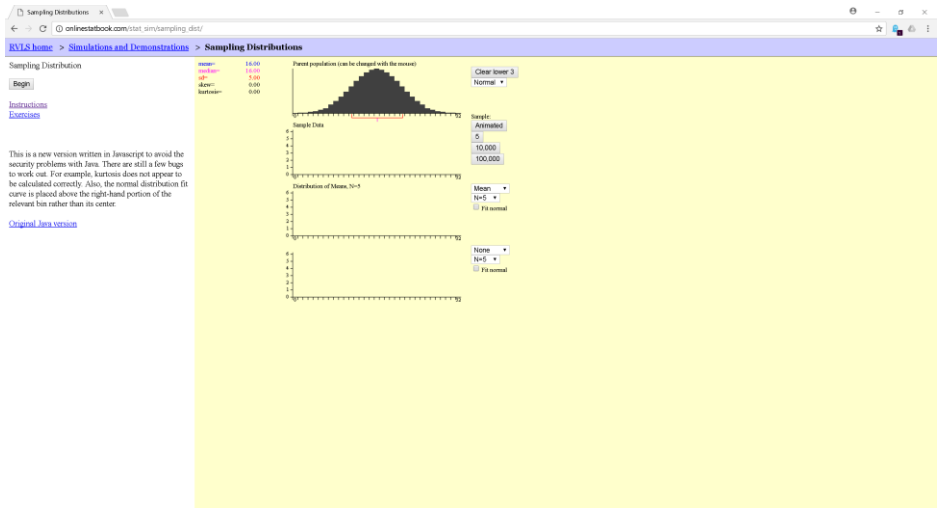
Meaning:

If we would repeat this study over and over ~68% of the means would fall within $\pm$ 1 SE

Central limit theorem

Please open on in our browser:

www.tinyurl.com/clt-simulator



When does the central limit theorem fail?

Another simulation: Standard error of the mean

Please enter the following code:

s1 = rnorm(10, mean=15, sd=5)

mean(s1)

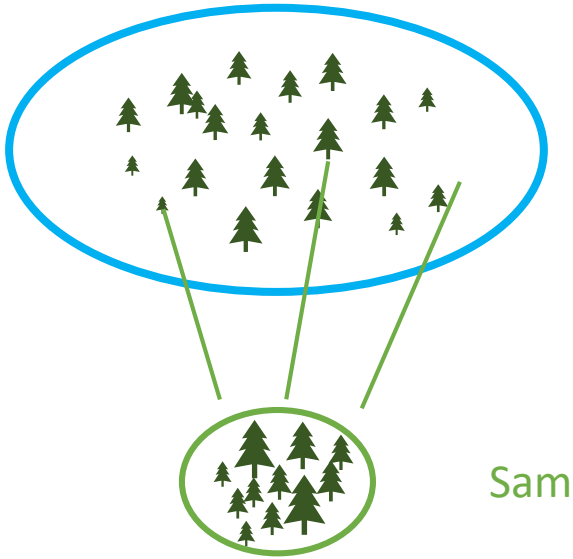
sd(s1)

sd(c(…))

se1 = sd(s1)/sqrt(10)

or

se1 =**your s**/sqrt(10)



Population

$\mu = 15\text{m}$

$\sigma = 5\text{m}$

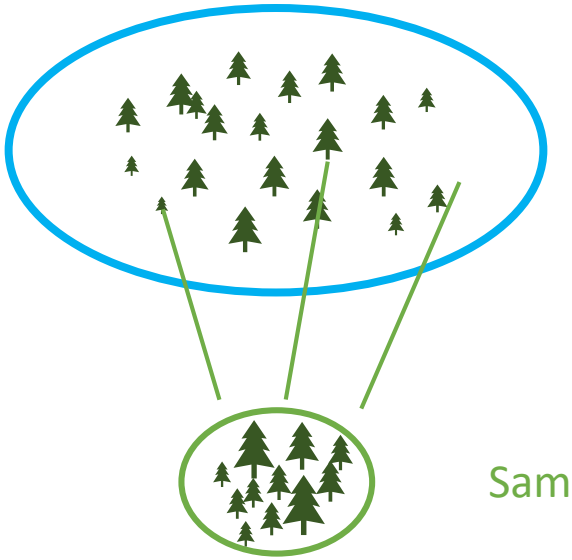
Sample of 10

$\bar{x} = ?$

$s = ?$

$SEM_{\bar{x}} = \frac{s}{\sqrt{n}}$

Standard error of the mean



Population

$\mu = 15\text{m}$

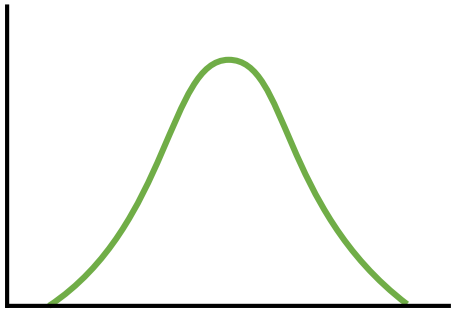
$\sigma = 5\text{m}$

$$SEM_{\bar{x}} = \frac{s}{\sqrt{n}}$$

Sample of 10

$\bar{x} = 15.2\text{m}$

$s = 4.4$

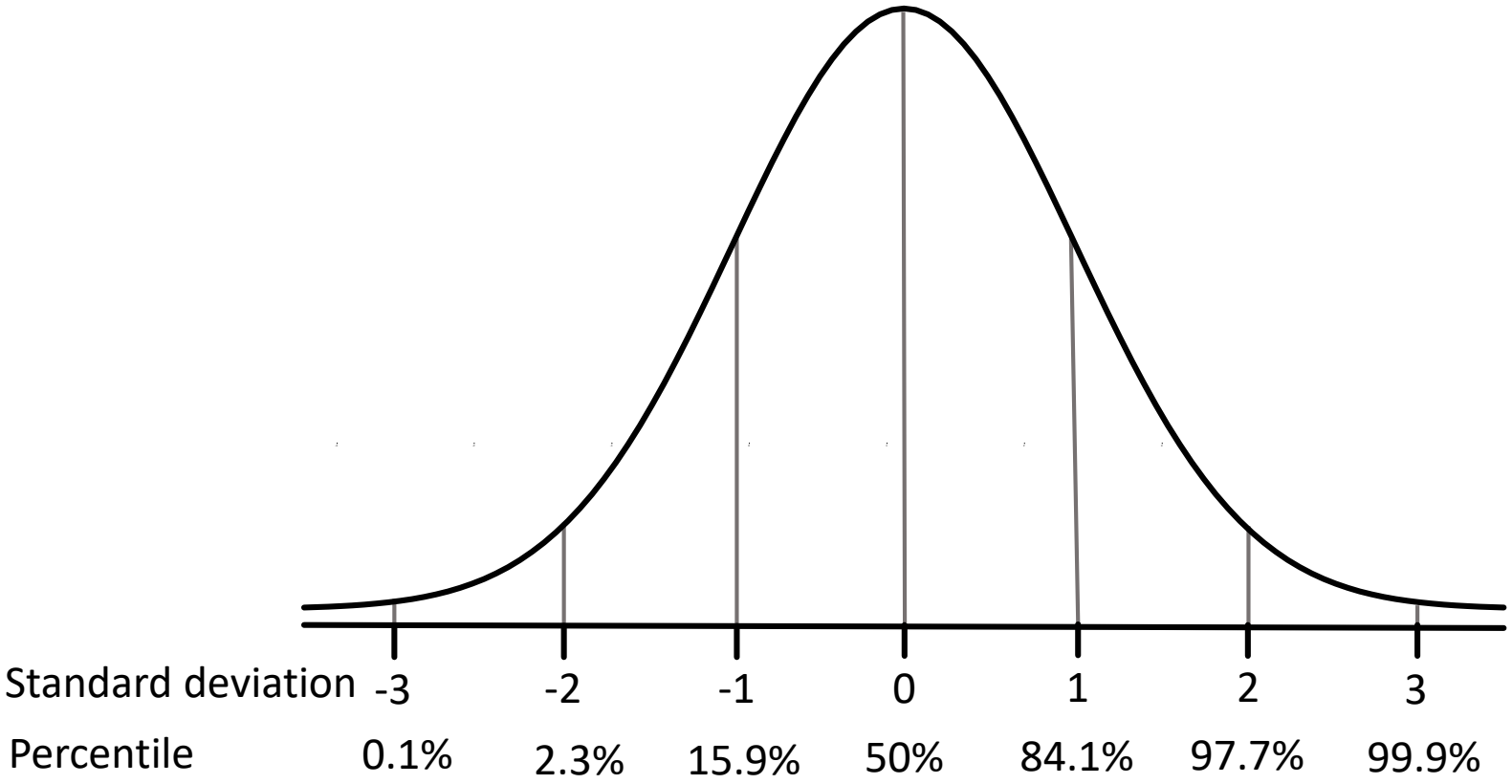


By repeated sampling what do we get?

A distribution of $\bar{x}$! … and it's normal!

Normal distribution

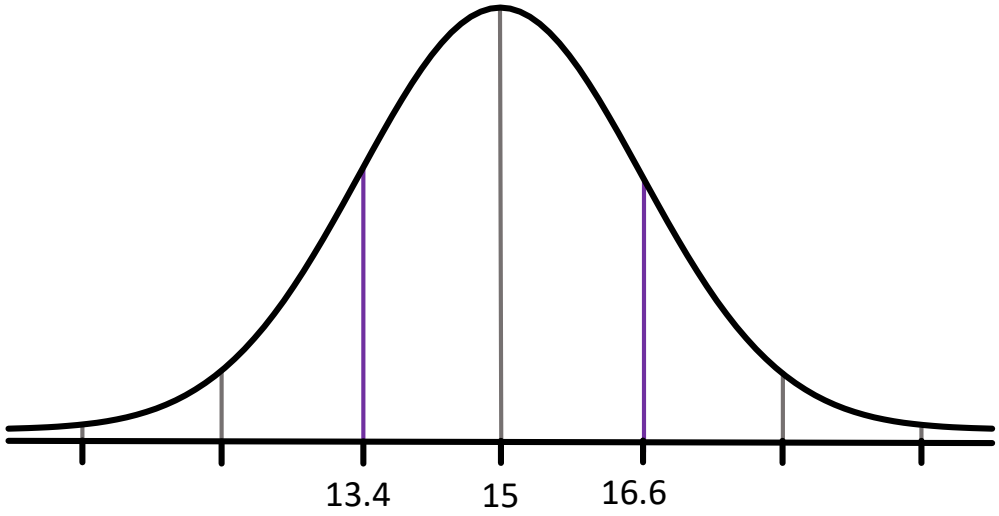
Because we know that sample means are normally distributed, we can now make assumptions based on it.



Standard errors & Confidence intervals

$\bar{x} = 15$

$SEM_{\bar{x}} = 1.6$



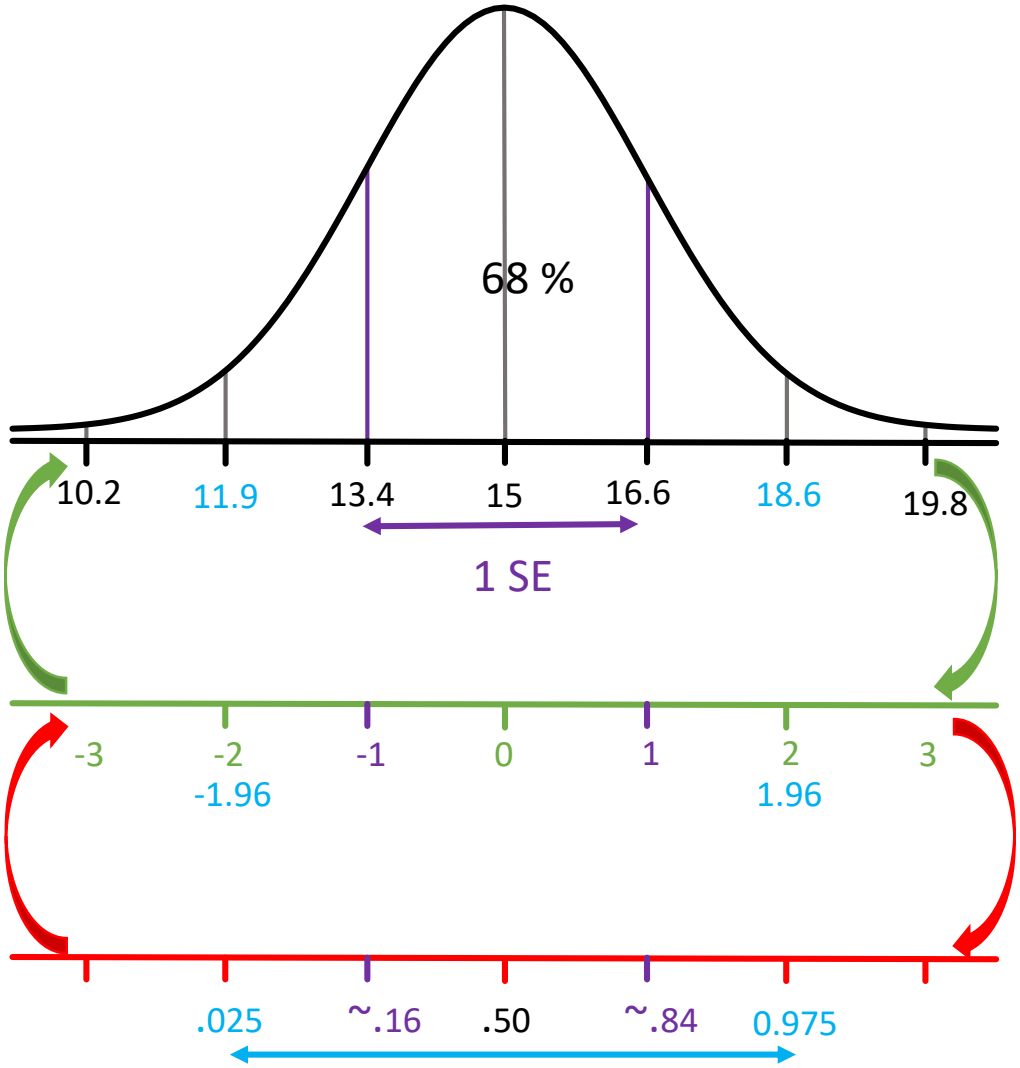
Distribution of the means

Original values

Standard errors & Confidence intervals

$\bar{x} = 15$

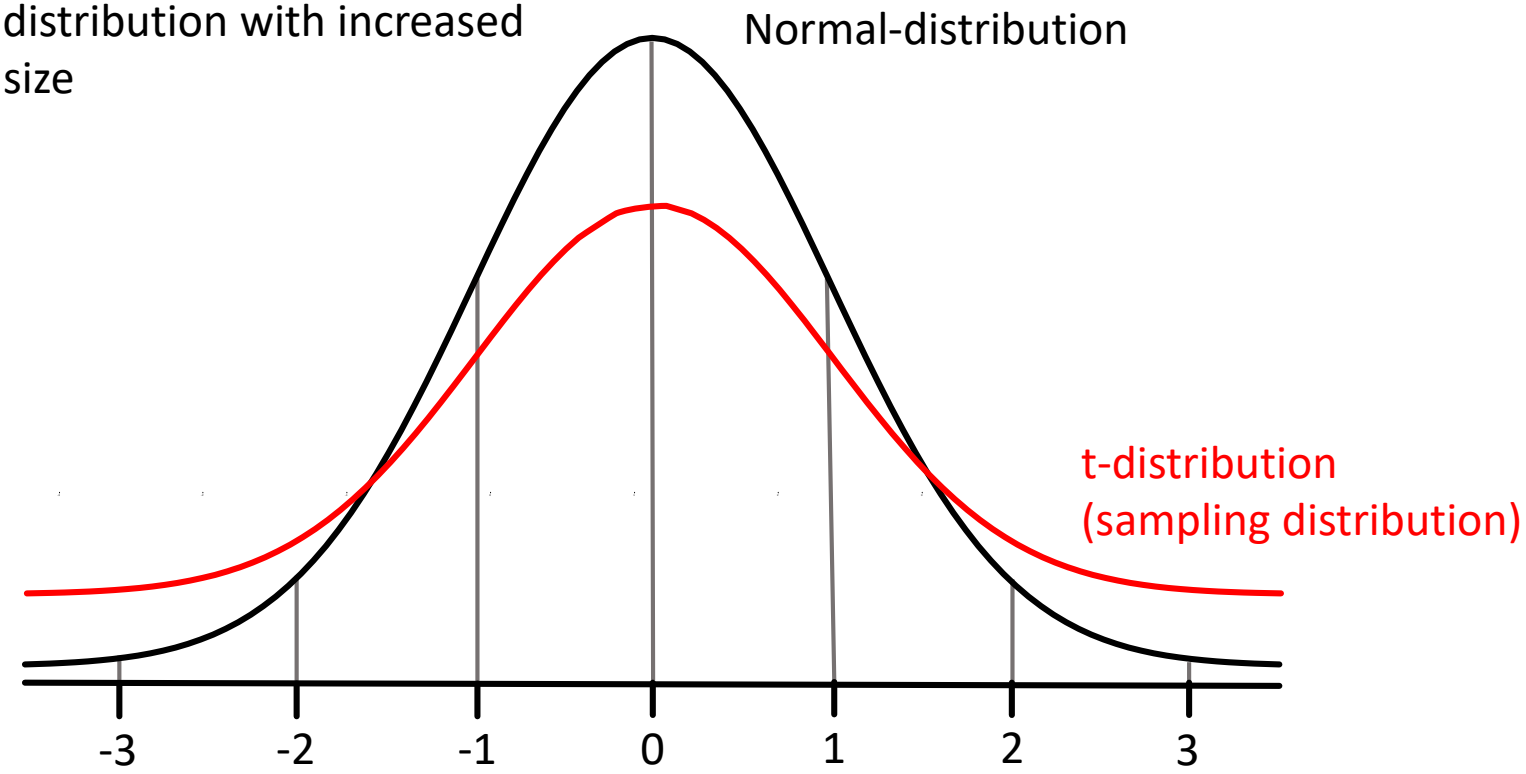
$SEM_{\bar{x}} = 1.6$



T-distribution

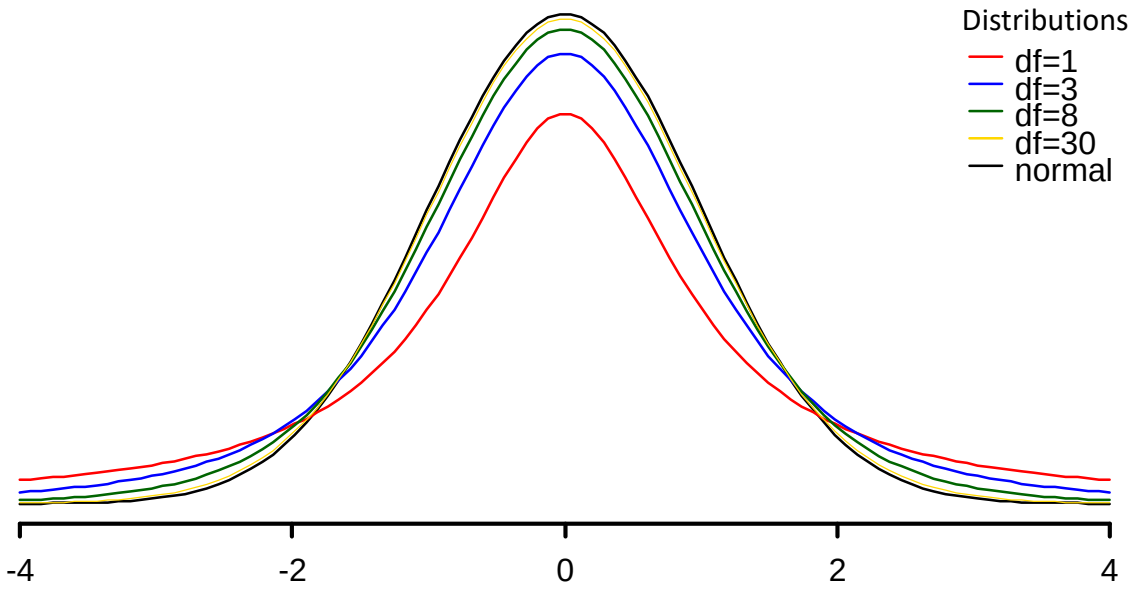
Features:

- Wider tails
- Degrees of freedom: n-1
- Approaches normal the shape of a normal distribution with increased sample size



Fixes biased sample means when sample size is low

T-distribution

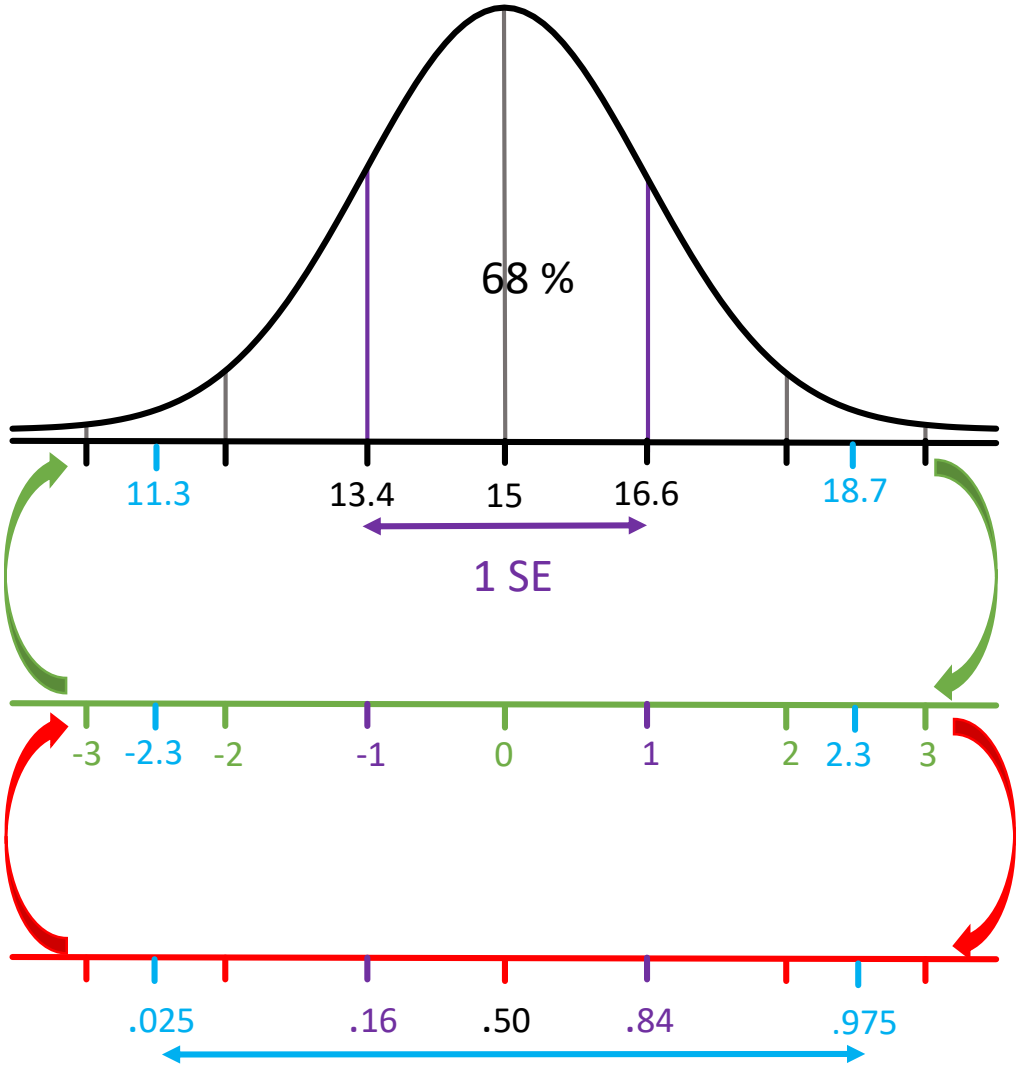


T-distribution

$\bar{x} = 15$

$SEM_{\bar{x}} = 1.6$

n=10



Challenge

Lentil varieties on Farm 1

FARM	VARIETY	YIELD
Farm1	A	720
Farm1	A	740
Farm1	A	690
Farm1	A	760
Farm1	B	515
Farm1	B	480
Farm1	B	545
Farm1	B	492
Farm1	C	540
Farm1	C	502
Farm1	C	510
Farm1	C	505

Can you calculate the 90% CI for Var A, B & C?

Enter the data like this:

VarA = c(720, 740, 690, 760)

VarB = c(515, 480, 545, 492)

VarC = c(540, 502, 510, 505)

1. Decide what p-value you need (1-0.9=0.1; 0.1/2=0.05; so either 0.05 or 0.95)
2. Calculate the t-values → t-value = qt(p-value, df)
3. Convert t-values to real values: value = t-value*SE+ $\bar{x}$

   use VarA_m=mean(VarA) and SE =sd(VarA)/sqrt(length(VarA)) to calculate mean and standard error