Some Statistical Procedures in R

Dr. Paige Rutner

Associate Professor of Practice

Information Systems and Quantitative Sciences

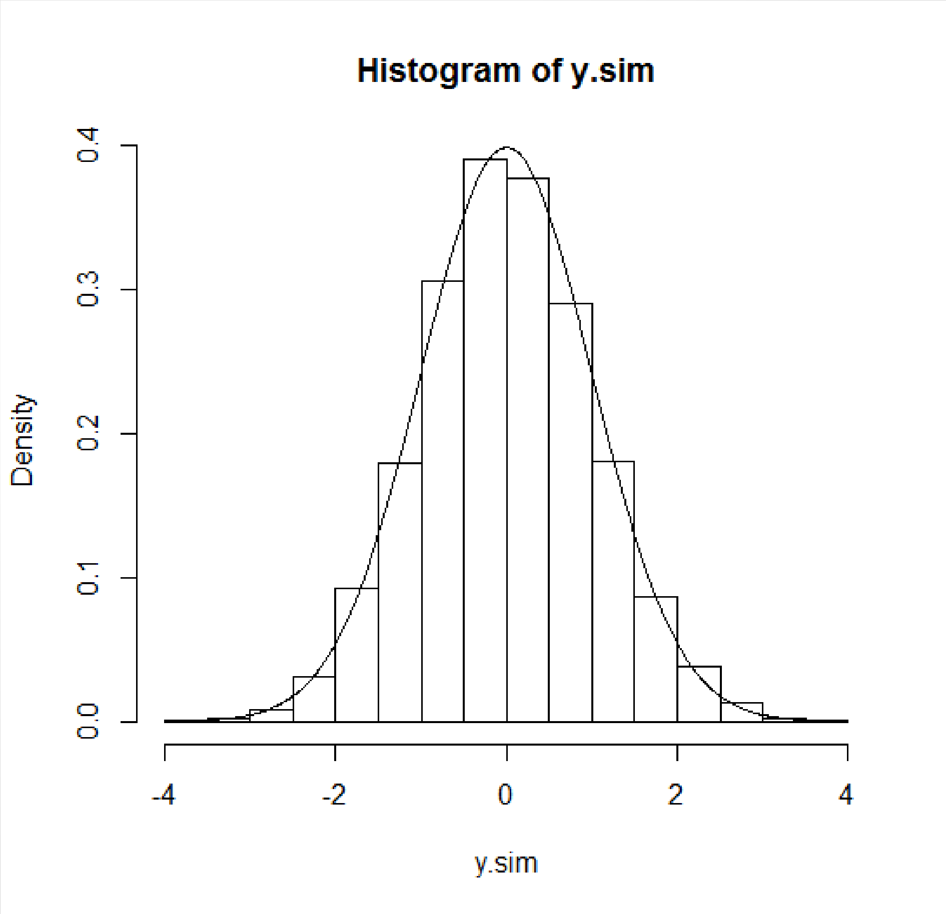
Probability Distribution Functions

- R provides functions to work with many probability distributions. Most distributions have functions in the following form:
  - d***: density function
  - p***: cumulative distribution function, P(X < x)
  - q***: quantile function
  - r***: generate random numbers from distribution

  where *** represents the specific distribution.

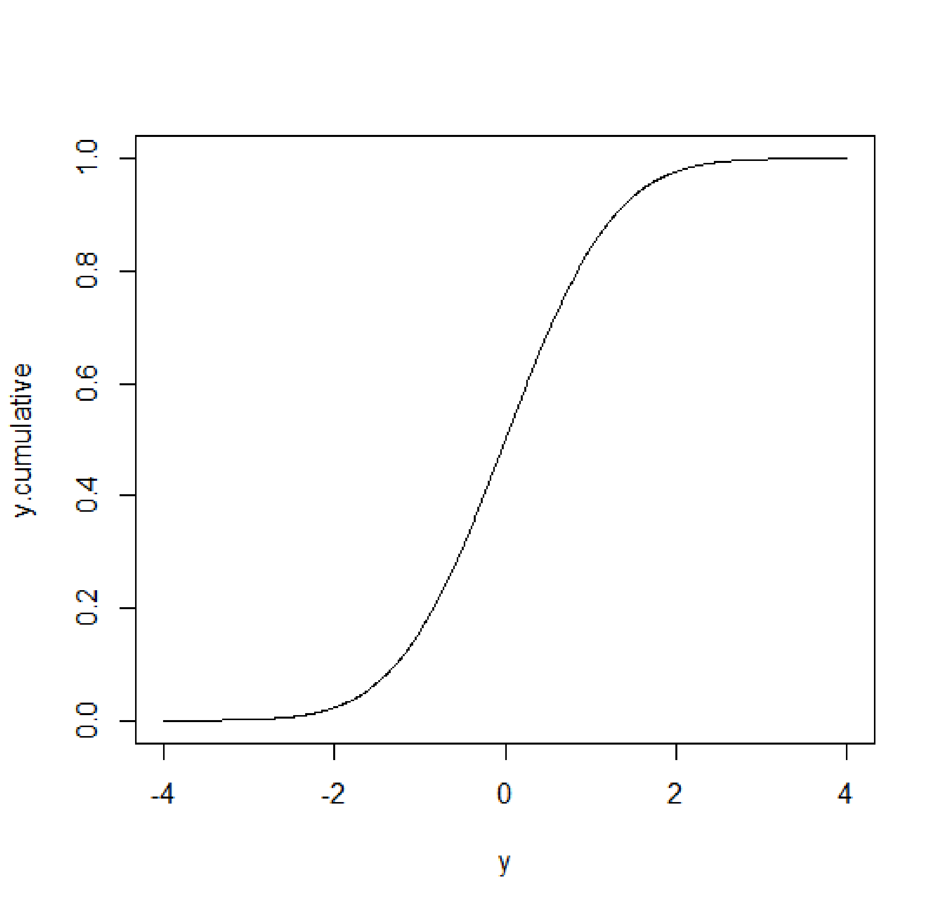
Plotting simulated data and the pdf

```
> # generate and plot data from the standard normal curve
> y.sim=rnorm(10000,type="l")
> hist(y.sim,freq="F")
> # generate y-values and densities to plot a normal curve
> y=seq(-4,4,0.01)
> y.density=dnorm(y,0,1)
> points(y,y.density,type="l")
```



Plotting the cumulative distribution function

```
> # generate cumulative
  probabilities using
  existing y values
> y.cumulative=pnorm(y,0,1)
> # plot against the y
  values
> plot(y, y.cumulative,
  type="l"
```



Some of the distributions included in R

- Continuous
  - unif: Uniform
  - norm: Normal
  - t: t
  - chisq: Chi-square
  - f: F
  - gamma: Gamma
  - exp: Exponential
  - beta: Beta
  - lnorm: Log-normal
- Discrete
  - binom: Binomial
  - geom: Geometric
  - hyper: Hypergeometric
  - nbinom: Negative binomial
  - pois: Poisson

A few examples of computations

- Given a normal distribution with a mean of 2 and a standard deviation of .8, compute $P(X < 1)$
- For the same distribution, compute $P(1 < X < 2.4)$

Sampling Distributions

- A sampling distribution is a distribution of all of the possible values of a sample statistic for a given sample size selected from a population.
- For example, suppose you sample 50 students from your college regarding their mean GPA. If you obtained many different samples of size 50, you will compute a different mean for each sample. We are interested in the distribution of all potential mean GPAs we might calculate for any sample of 50 students.

Uniform distribution

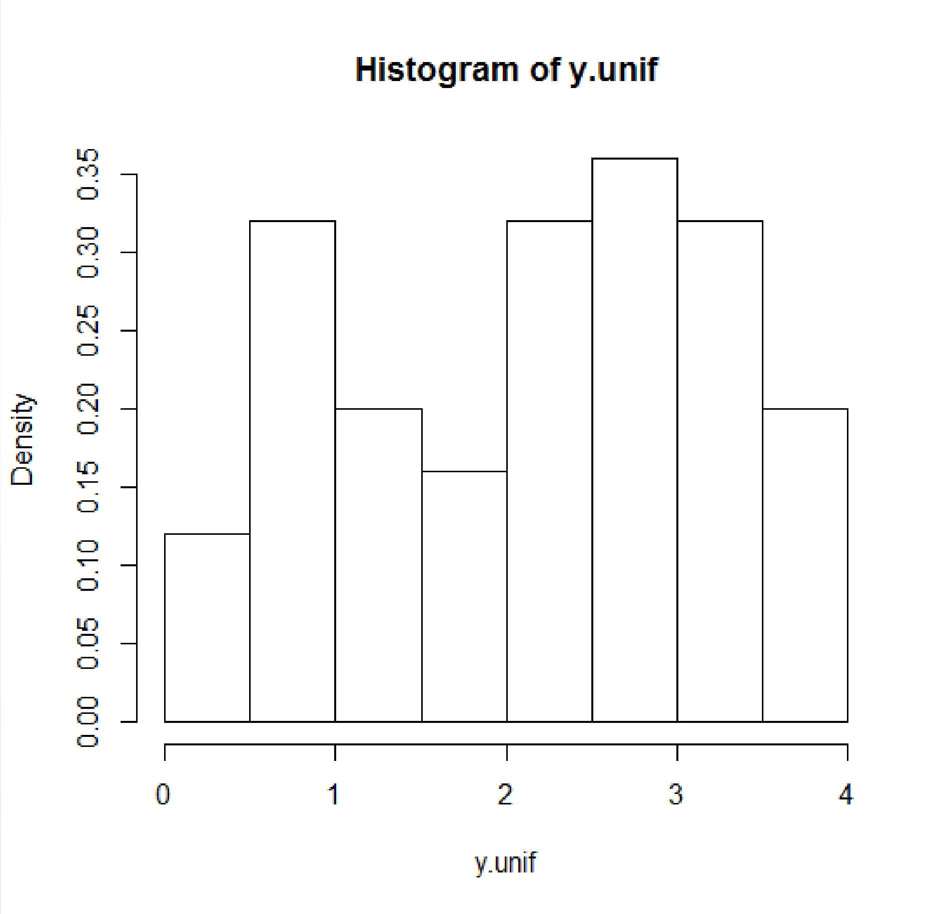
- Suppose student GPAs follow a uniform distribution between 0 and 4.
- What would you expect to be the mean and standard deviation of this distribution?
- Given a uniform distribution with a range of A to B,

$$Mean = \frac{A + B}{2}$$

$$Standard\ deviation = \sqrt{\frac{(B - A)^2}{12}}$$

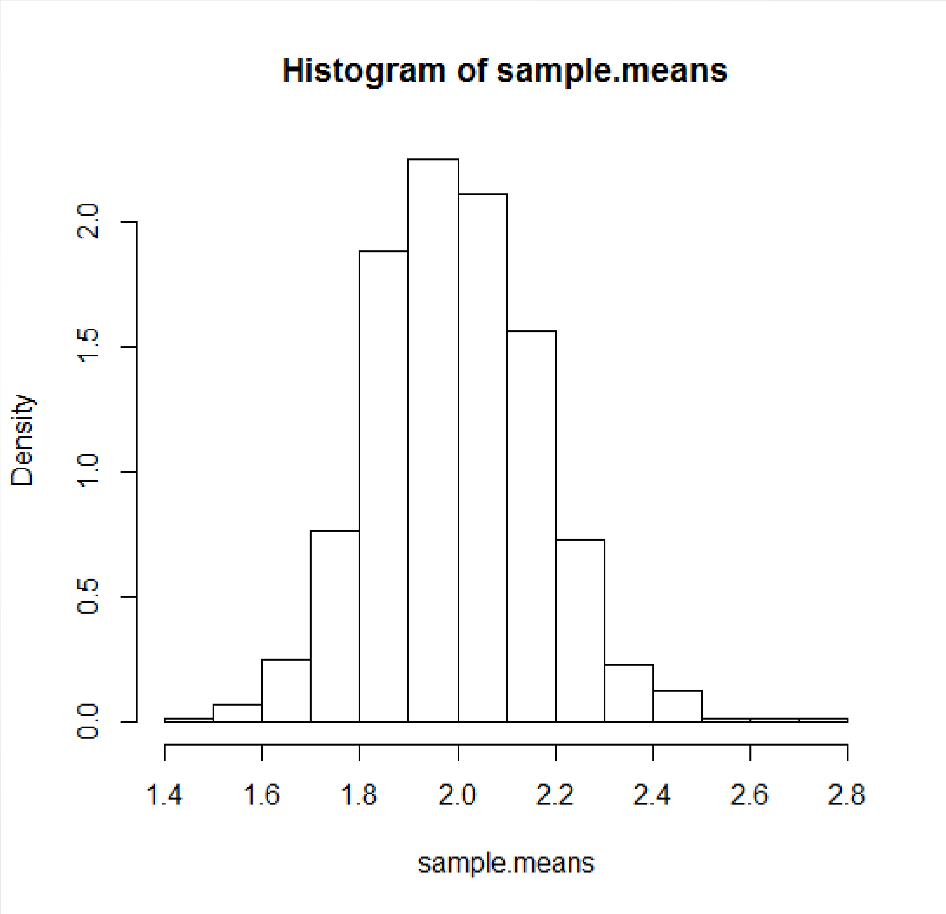
Generate a simulated sample

```
># generate a sample of
 50 values from a
 uniform distribution
 between 0 and 4.0
>y.unif=runif(50,0,4)
># compute the mean and
 plot the histogram of
 the sample
>mean(y.unif)
>hist(y.unif, freq=F)
```



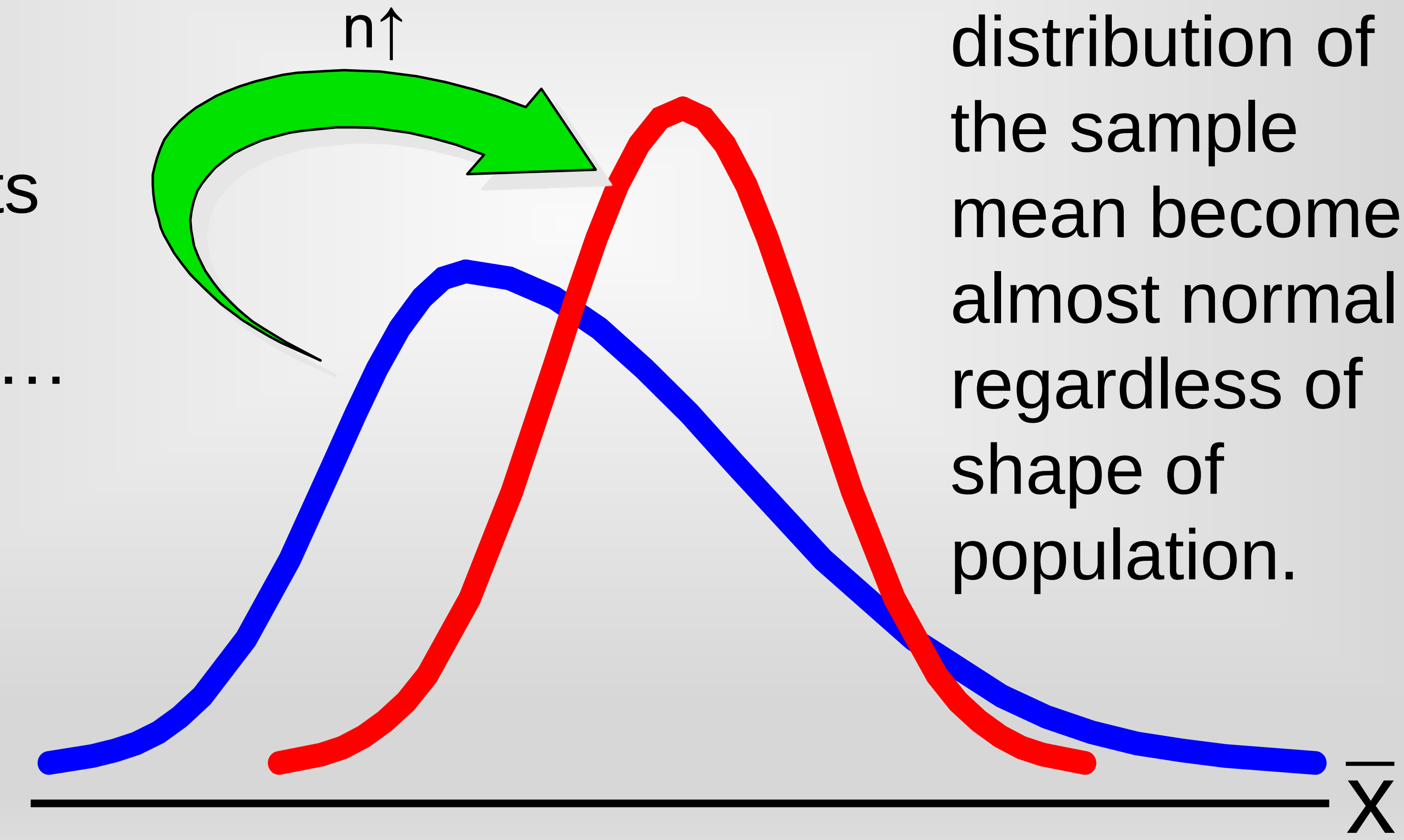
Repeat the sampling process many times . . .

```
># replicate the random
 sample 1000 times
>y.unif=replicate(1000,ru
 nif(50,0,4))
># compute the mean of
 each of the 1000 samples
>sample.means=colMeans(y.
 unif)
># plot the histogram of
 sample means
>hist(sample.means,freq=F
 )
```



Central Limit Theorem

As the sample size gets large enough…



the sampling distribution of the sample mean becomes almost normal regardless of shape of population.

Sampling distribution statistics

- Given a sufficiently large sample size, for a population with mean μ and standard deviation σ, the sampling distribution is approximately normally distributed with:

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

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

How Large is Large Enough?

- For most distributions, $n > 30$ will give a sampling distribution that is nearly normal.
- For fairly symmetric distributions, $n > 15$.
- For a normal population distribution, the sampling distribution of the mean is always normally distributed.

Check simulated data

- $Expected\ Mean = \mu = \frac{0+4}{2} = 2$
- $Expected\ Standard\ Deviation = \sigma = \sqrt{\frac{(4-0)^2}{12}} = 1.155$
- $Expected\ Standard\ Error = \frac{\sigma}{\sqrt{n}} = \frac{1.155}{\sqrt{50}} = 0.162$
- Mean and standard deviation of simulated data

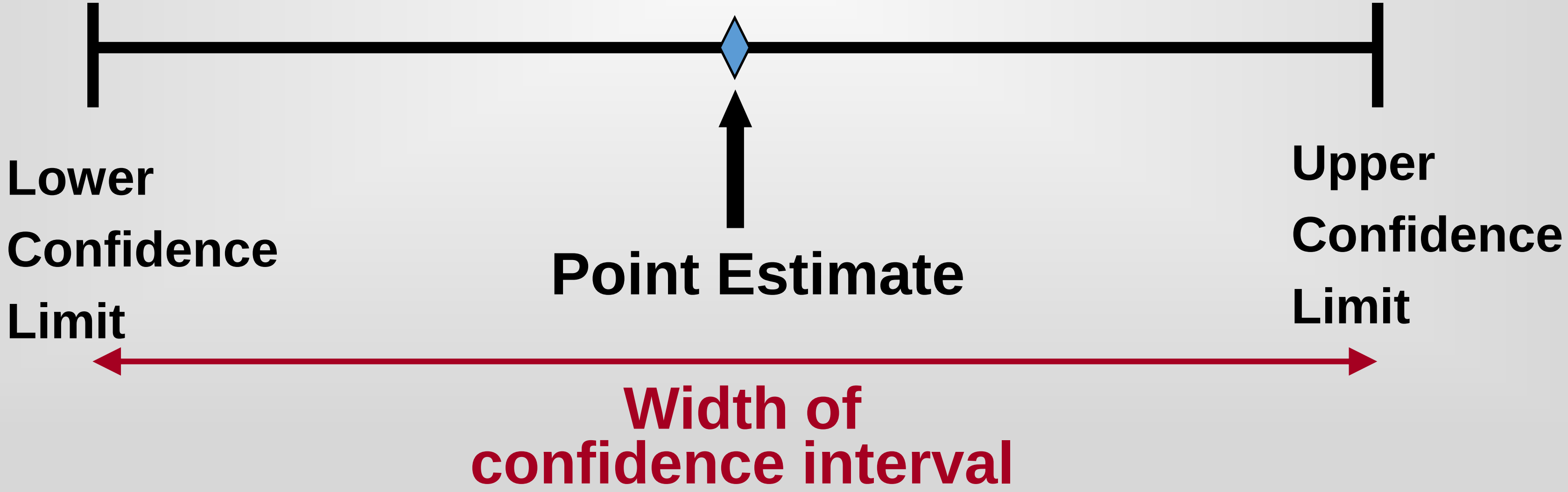
```
> mean(sample.means)
[1] 1.996643
> sd(sample.means)
[1] 0.1687319
```

Another example calculation

- Given a uniform distribution with a mean of 2.0 and a standard deviation of 1.155, compute the probability that a randomly selected sample of *n*=50 would have a mean less than 2.5 or $P(\bar{X} < 2.5)$.
- $z = \frac{x-\mu}{\frac{\sigma}{\sqrt{n}}} = \frac{2.5-2}{\frac{1.155}{\sqrt{50}}} = 3.061$
- P(Z < 3.061) = 0.9989

Point and Interval Estimates

- A point estimate, such as a sample mean, is a single number.
- A confidence interval provides additional information about the variability of the estimate.



Confidence Interval Estimate

- An interval gives a range of values:
  - Takes into consideration variation in sample statistics from sample to sample.
  - Based on observations from 1 sample.
  - Gives information about closeness to unknown population parameters.
  - Stated in terms of level of confidence:
    - e.g. 95% confident, 99% confident.
    - Can never be 100% confident.

General Formula

- The general formula for all confidence intervals is:

Point Estimate ± (Critical Value)(Standard Error)

Where:

- Point Estimate is the sample statistic estimating the population parameter of interest.

- Critical Value is a z score value based on the sampling distribution of the point estimate and the desired confidence level.

- Standard Error is the standard deviation of the point estimate.

Confidence Interval for μ (σ Known)

- Assumptions:
  - Population standard deviation σ is known.
  - Population is normally distributed.
  - If population is not normal, use large sample (n > 30).
- Confidence interval estimate:

$$\bar{X} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

where $\bar{X}$ is the point estimate

$Z_{\alpha/2}$ is the normal distribution critical value for a probability of $\alpha/2$ in each tail

$\sigma/\sqrt{n}$ is the standard error

Example computation

- Still working with the uniform distribution of student GPAs, compute the 95% confidence interval based a sample of 50 students with a mean of 1.9. Recall that the expected population mean and standard deviations were 2 and 1.155, respectively.
- The critical value of Z should result in 0.05/2 area under the curve in each tail of the distribution.

```
> qnorm(0.025)
[1] -1.959964
> qnorm(0.975)
[1] 1.959964
```

Example computation - continued

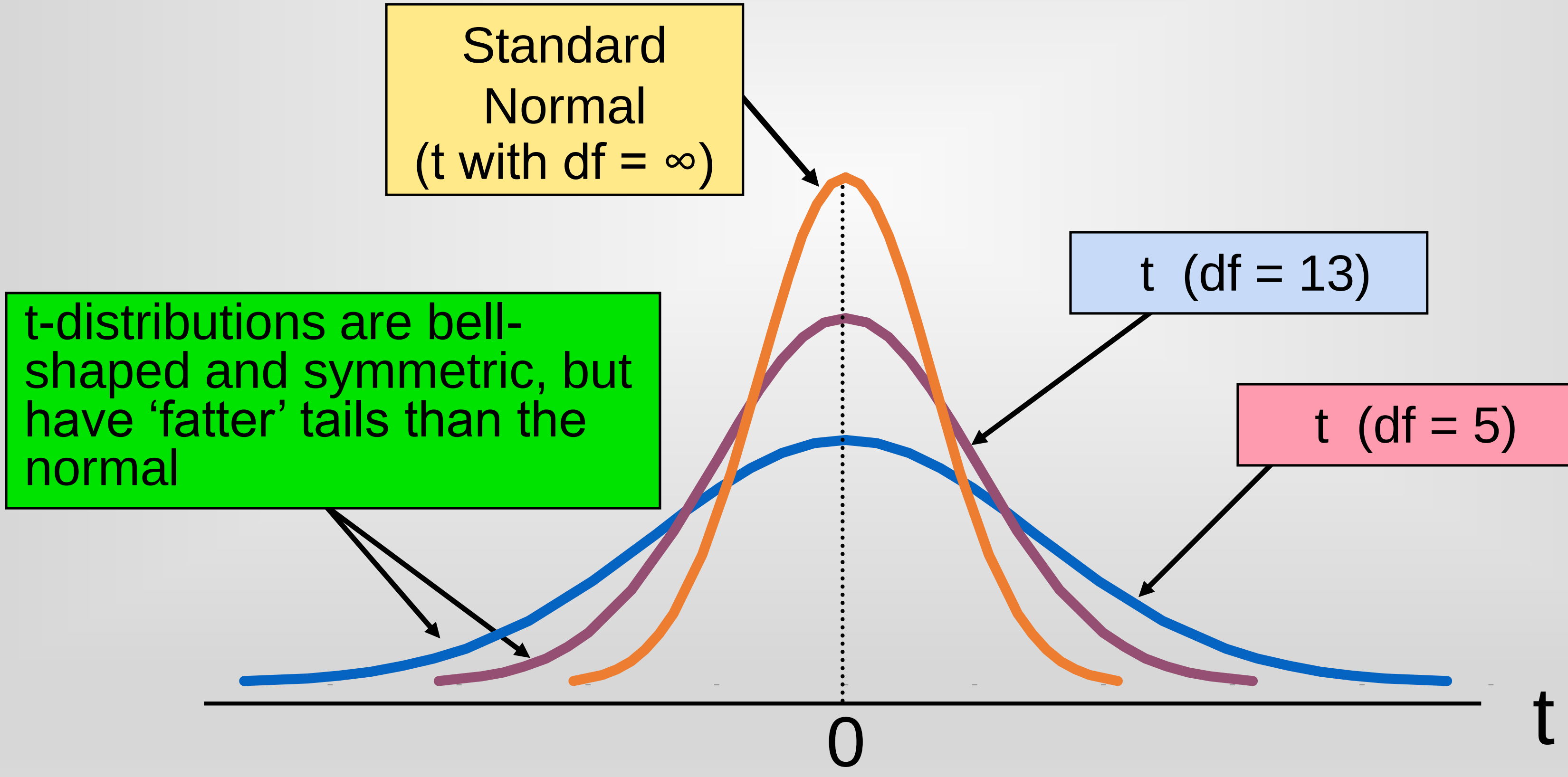
- Margin of error = Critical Value * Standard Error
- $Margin\ of\ error = \pm 1.959964 * \frac{1.155}{\sqrt{50}}$= ±0.3201
- Lower confidence limit = 1.9 – 0.3201 = 1.5799
- Upper confidence limit = 1.9 + 0.3201 = 2.2201
- 1.5799 ≤ μ ≤ 2.2201
- With 95% confidence, we estimate that the true population mean is between 1.5799 and 2.2201.

Confidence Interval for μ (σ Unknown)

- If the population standard deviation σ is unknown, we can substitute the sample standard deviation, S.
- This introduces extra uncertainty, since S is variable from sample to sample.
- So we use the t distribution instead of the normal distribution.

Student's t Distribution

Note: t $\longrightarrow$ Z as n increases



Confidence Interval for μ (σ Unknown)

- Assumptions:
  - Population standard deviation is unknown.
  - Population is normally distributed.
  - If population is not normal, use large sample ($n > 30$).
- Use Student's t Distribution.
- Confidence Interval Estimate:

$$\boxed{\overline{X} \pm t_{\alpha/2} \frac{S}{\sqrt{n}}}$$

(where $t_{\alpha/2}$ is the critical value of the t distribution with $n - 1$ degrees of freedom and an area of $\alpha/2$ in each tail.)

Example computation

- Still working with the uniform distribution of student GPAs, compute the 95% confidence interval based a sample of 50 students with a mean of 2.19 with a standard deviation of 1.14.
- The critical value of t with n-1 degrees of freedom should result in 0.05/2 area under the curve in each tail of the distribution

```
> qt(0.025, df=49)
[1] -2.009575
> qt(0.975, df=49)
[1] 2.009575
```

Example computation - continued

- Margin of error = Critical Value * Standard Error
- $Margin\ of\ error = \ \pm 2.009575 * \frac{1.155}{\sqrt{50}}$= ±0.3268
- Lower confidence limit = 2.19 – 0.3268 = 1.8362
- Upper confidence limit = 1.9 + 0.3268 = 2.5168
- 1.8362 ≤ μ ≤ 2.5168
- With 95% confidence, we estimate that the true population mean is between 1.8362 and 2.5168.