Lingnan University 嶺南大學 香港 Hong Kong

School of Data Science 數據科學學院

CDS 533
Statistics for Data Science

Instructor: Lisha Yu

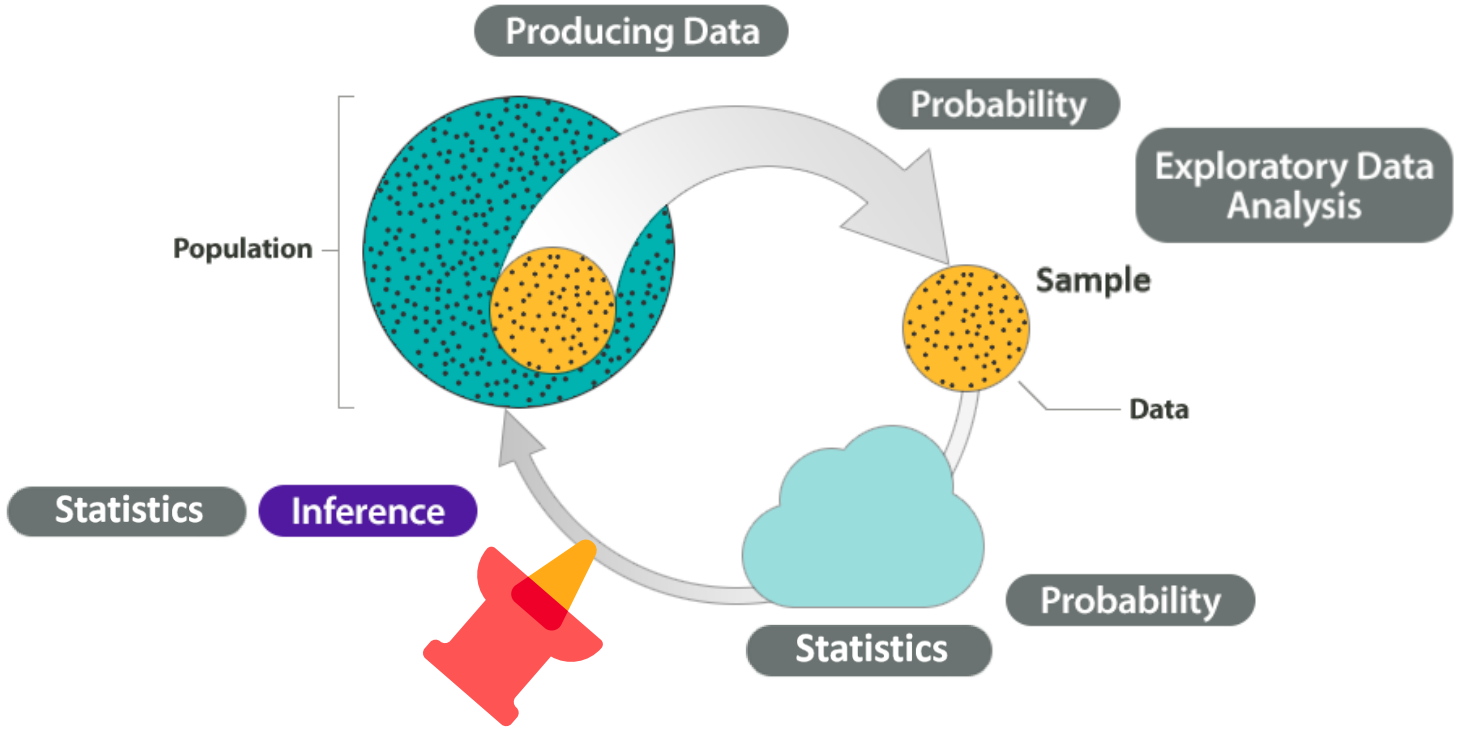
Division of Artificial Intelligence

School of Data Science

Lingnan University

Fall 2024

Big Picture of Statistics



Statistical Inference (Estimation)

Inferential Statistics

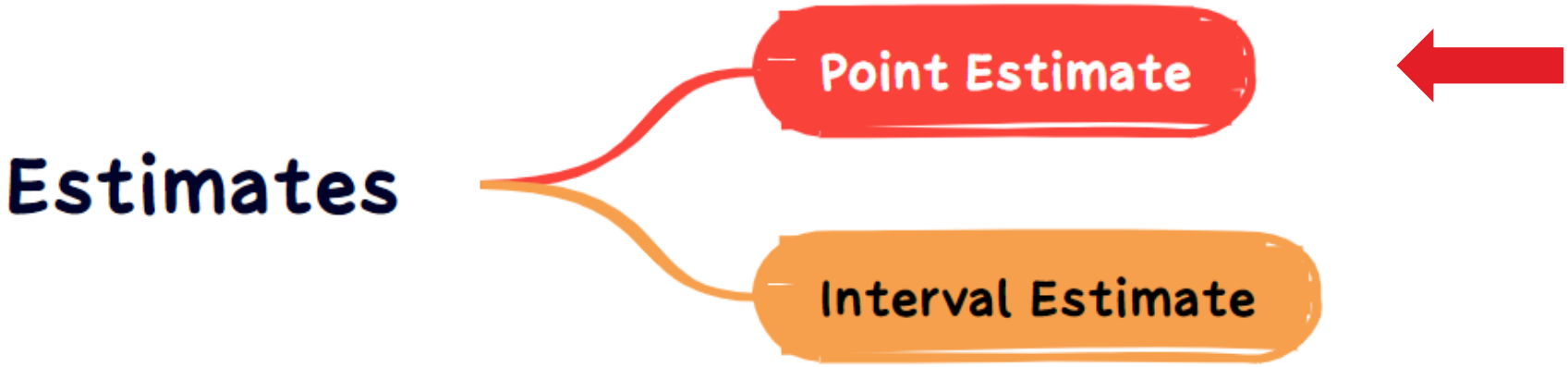
Inferential statistics

- Draw conclusion from data
- Sample
  - Describe data
- Use **sample statistic** to infer **population parameter**
  - Estimation
  - Hypothesis testing

Two Types of Estimates

Two Types of Estimate

Estimation is specific observed numerical value used to estimate an unknown population parameter.



Parameter		Statistics
μ	Mean	$\bar{x}$
σ	Standard deviation	s
P	Proportion	p

Point Estimate

Point estimation is a single value (or point) used to approximate a population parameter.

Average height = 185cm

Estimator

Estimator refers to a particular statistic used to construct a point estimate.

The estimator is a random variable, while the estimate is a particular number.

How to 

estimator?

Properties of Estimator

- X: random variable
- θ: a parameter of interest; ***unknown***

$$X_1, X_2, \dots, X_n \text{ independent and identically distributed (iid) sample}$$

- $\hat{\theta} = \theta(X_1, X_2, \dots, X_n)$: estimator of θ
- Previously, we found *good(?)* estimator(s) for θ *or its function* $f(\theta)$.

Goal

- Check how *good* are these estimator(s). Or are they ***good*** at all?
- If more than one *good* estimator is available, which one is ***better***?

Unbiased

Unbiased estimator

$\hat{\theta}$ is an **Unbiased Estimator(UE)** of θ for all $\theta \in \Omega$ if

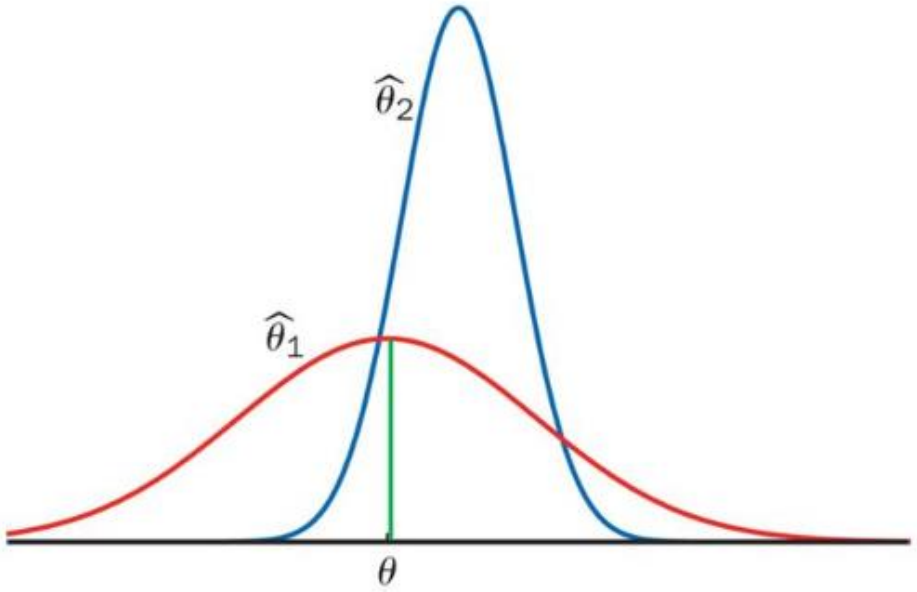
$$E[\hat{\theta}] = \theta \; for \; all \; \theta \in \Omega$$

Otherwise, it is a **Biased Estimator** of θ.

$$b(\hat{\theta}) = E[\hat{\theta}] - \theta \qquad (\textbf{bias of } \hat{\theta})$$

The bias of an unbiased estimator is zero:

$$\hat{\theta} \; unbiased \iff b(\hat{\theta}) = 0$$



Consistency

Consistency

$\hat{\theta}$ is a **Consistent Estimator(CE)** for parameter θ for all $\theta \in \Omega$ if

$$\hat{\theta} \xrightarrow{p} \theta.$$

$$P\left|\hat{\theta} - \theta\right| \geq \epsilon \to 0 \quad as \quad n \to \infty$$

For large n, a consistent estimator tends to be closer to the unknown population parameter.

Example

$$E(\bar{X}) = \mu \Rightarrow \bar{X} \text{ is an UE of } \mu.$$

$$\bar{X} \xrightarrow{p} \mu$$

$$\Rightarrow \bar{X} \text{ is a CE of } \mu.$$

Efficient

Efficient

An **Efficient Estimator** tends to fall closer to θ, on the average, than other estimators.

Mean Square Error (MSE)

$$MSE(\hat{\theta}) = E\left[\hat{\theta} - \theta\right]^2$$

The mean squared error satisfies

$$\begin{aligned}
\mathrm{MSE}(\hat{\theta}) &= E(\hat{\theta} - \theta)^2 = E\{[\hat{\theta} - E(\hat{\theta})] + [E(\hat{\theta}) - \theta]\}^2 \\
&= E[\hat{\theta} - E(\hat{\theta})]^2 + [E(\hat{\theta}) - \theta]^2 + 2[E(\hat{\theta}) - E(\hat{\theta})][E(\hat{\theta}) - \theta] \\
&= E[\hat{\theta} - E(\hat{\theta})]^2 + [E(\hat{\theta}) - \theta]^2 = \mathrm{var}(\hat{\theta}) + [E(\hat{\theta}) - \theta]^2.
\end{aligned}$$

Efficient

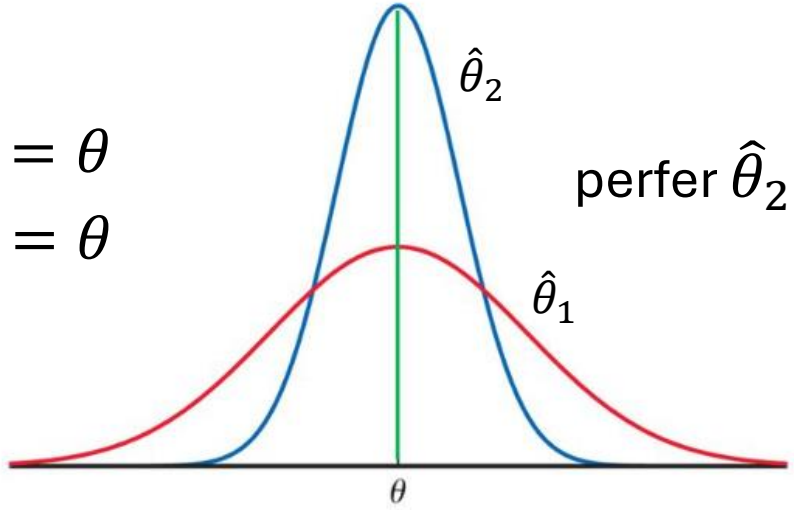
Efficient

$$MSE(\hat{\theta}) = E[\hat{\theta} - \theta]^2 = var(\hat{\theta}) + (bias)^2$$

For unbiased estimators,

$$MSE(\hat{\theta}) = \text{var}(\hat{\theta})$$

$E[\hat{\theta}_1] = \theta$
$E[\hat{\theta}_2] = \theta$



perfer $\hat{\theta}_2$ because $\text{var}(\hat{\theta}_2) < \text{var}(\hat{\theta}_1)$

The best one with smallest MSE is a **minimum variance unbiased estimator**.

Efficient

Example

$$X_1, X_2, \ldots, X_n \sim \text{iid } \mathrm{N}(\mu, \sigma^2)$$

Unbiased estimators

- Sample mean, $\bar{X}$
- Sample median

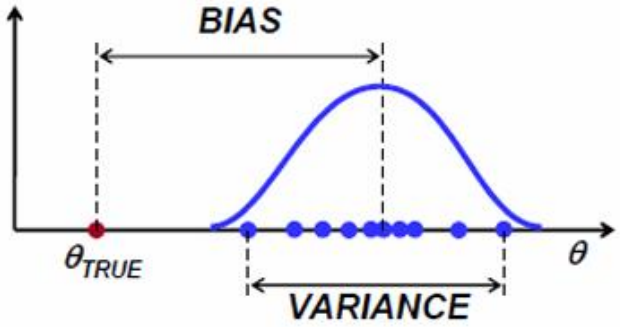
$$SE(sample\ median) = 1.25\sigma/\sqrt{n}$$

$\bar{X}$ is the **minimum variance unbiased estimator** for μ for normal populations, whereas the sample median is an inefficient estimator.

Properties of Estimator

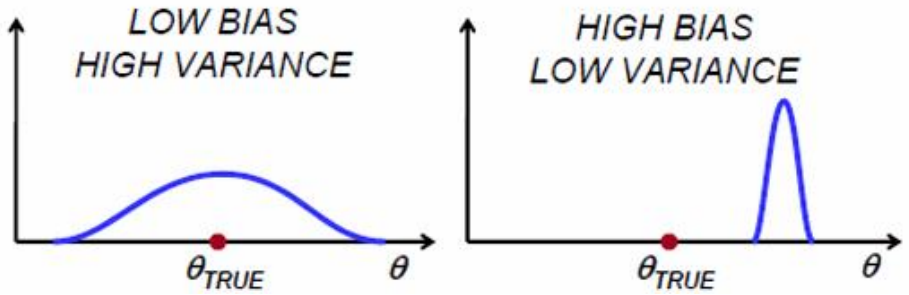
In summary, a good estimator $\hat{\theta}$ of θ is

- **consistent**
- at least **asymptotically unbiased** and **asymptotically efficient**



The bias-variance tradeoff

- In most cases, you can only decrease one of them at the expense of the other



Example

- S^2 is unbiased estimator
- $E(\hat{\sigma}^2) \rightarrow \sigma^2$
- s is biased

Maximum Likelihood Estimator

Maximum likelihood estimator(MLE)

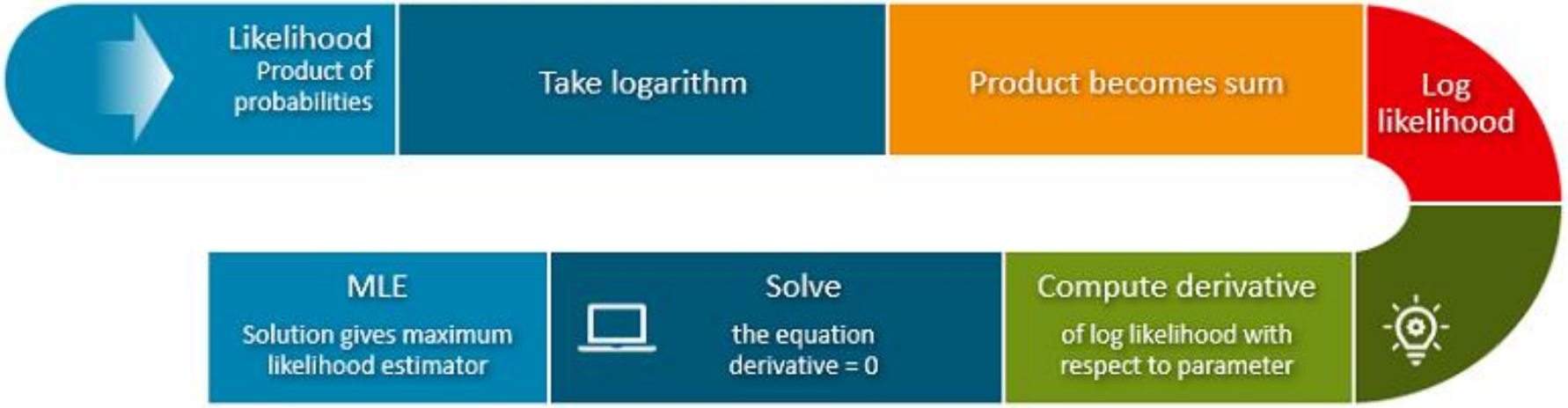
$$L(\theta; x_1, \ldots, x_n) = f_\theta(x_1) \ldots f_\theta(x_n)$$

$$L(\hat{\theta}_n; x_1, \ldots, x_n) = \max_\theta L(\theta; x_1, \ldots, x_n)$$

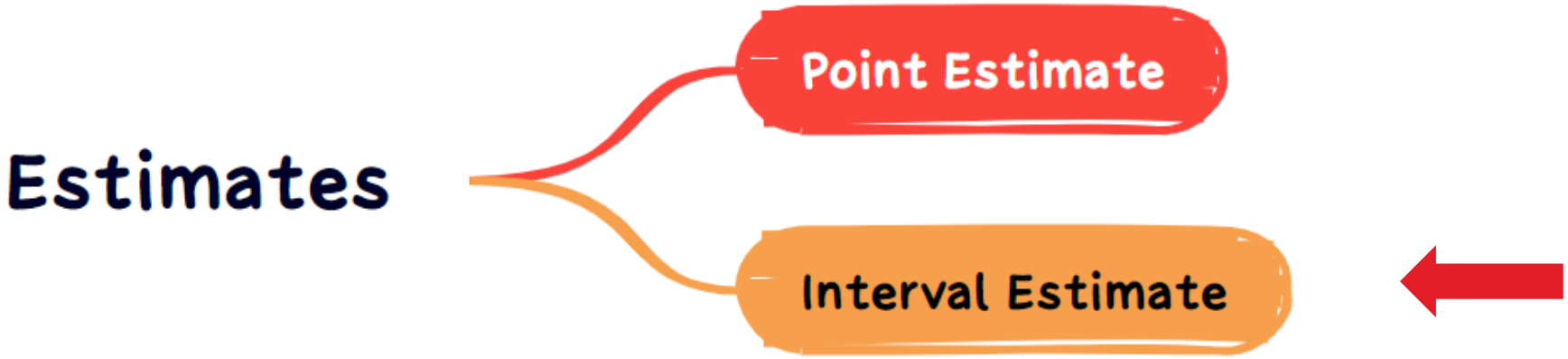
$$s(X;\theta) = \frac{\partial \log f_\theta(X;\theta)}{\partial \theta} \quad \text{(score function)}$$

Properties

- Asymptotically unbiased
- Consistent
- Asymptotically efficient

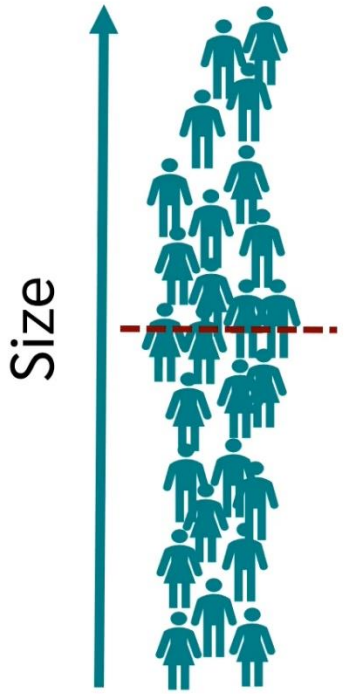


Two Types of Estimate



Interval Estimate

Motivation Example



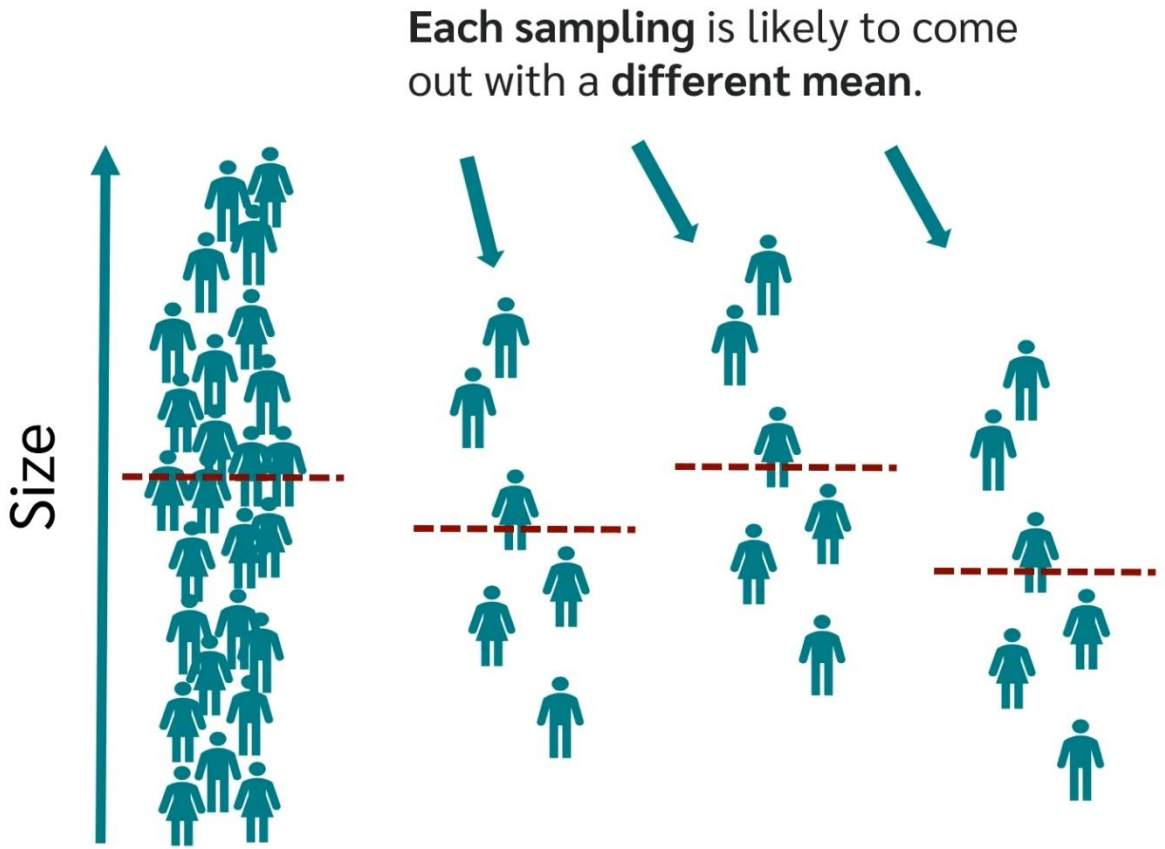
You want to know the **height** of all professional **basketball players** in the **USA**.

One sample average height = 185cm

?

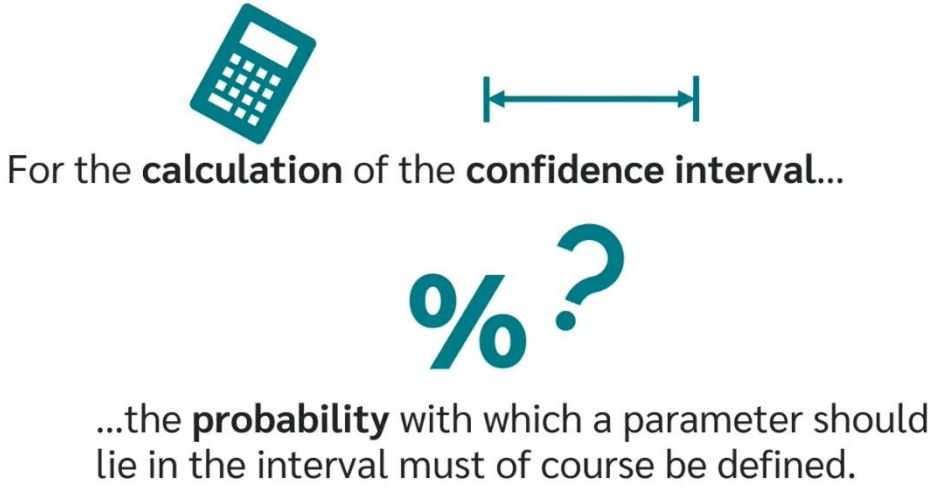
Interval Estimate

Interval estimation is an interval of numbers around the point estimate used to approximate a population parameter.



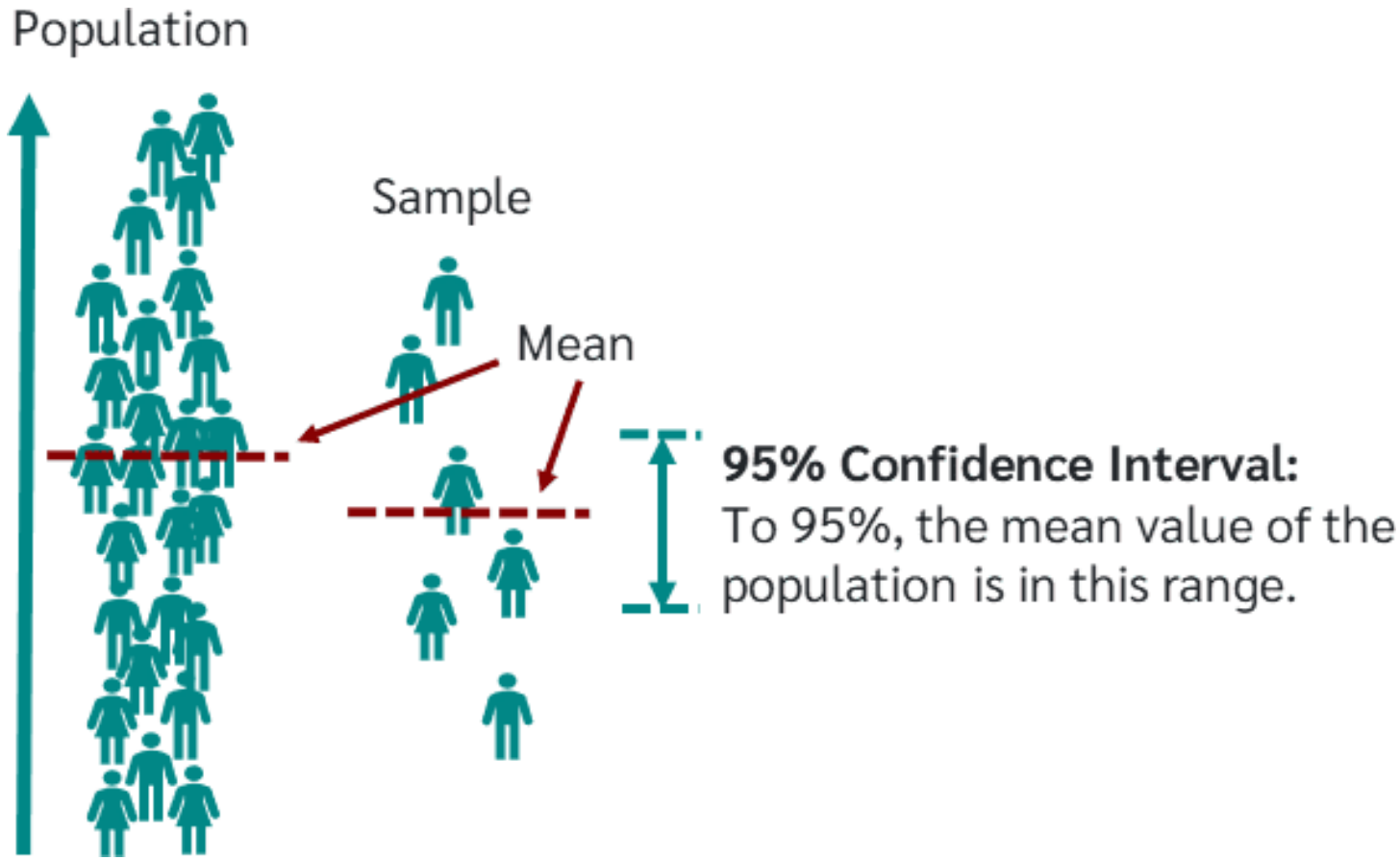
Interval Estimate

What is high probability?



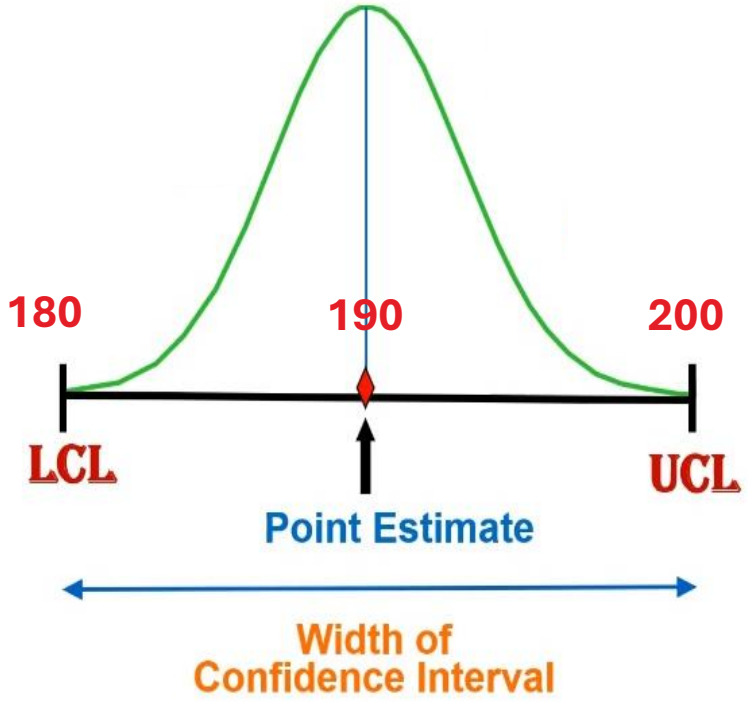
Interval Estimate

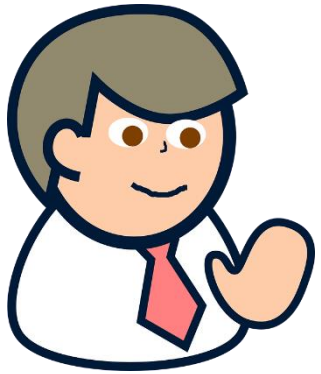
Confidence interval is a range within which the true parameter value is believed to lie with a specified degree of confidence.



Confidence Interval

CI Animation: https://www.zoology.ubc.ca/~whitlock/Kingfisher/CIMean.htm





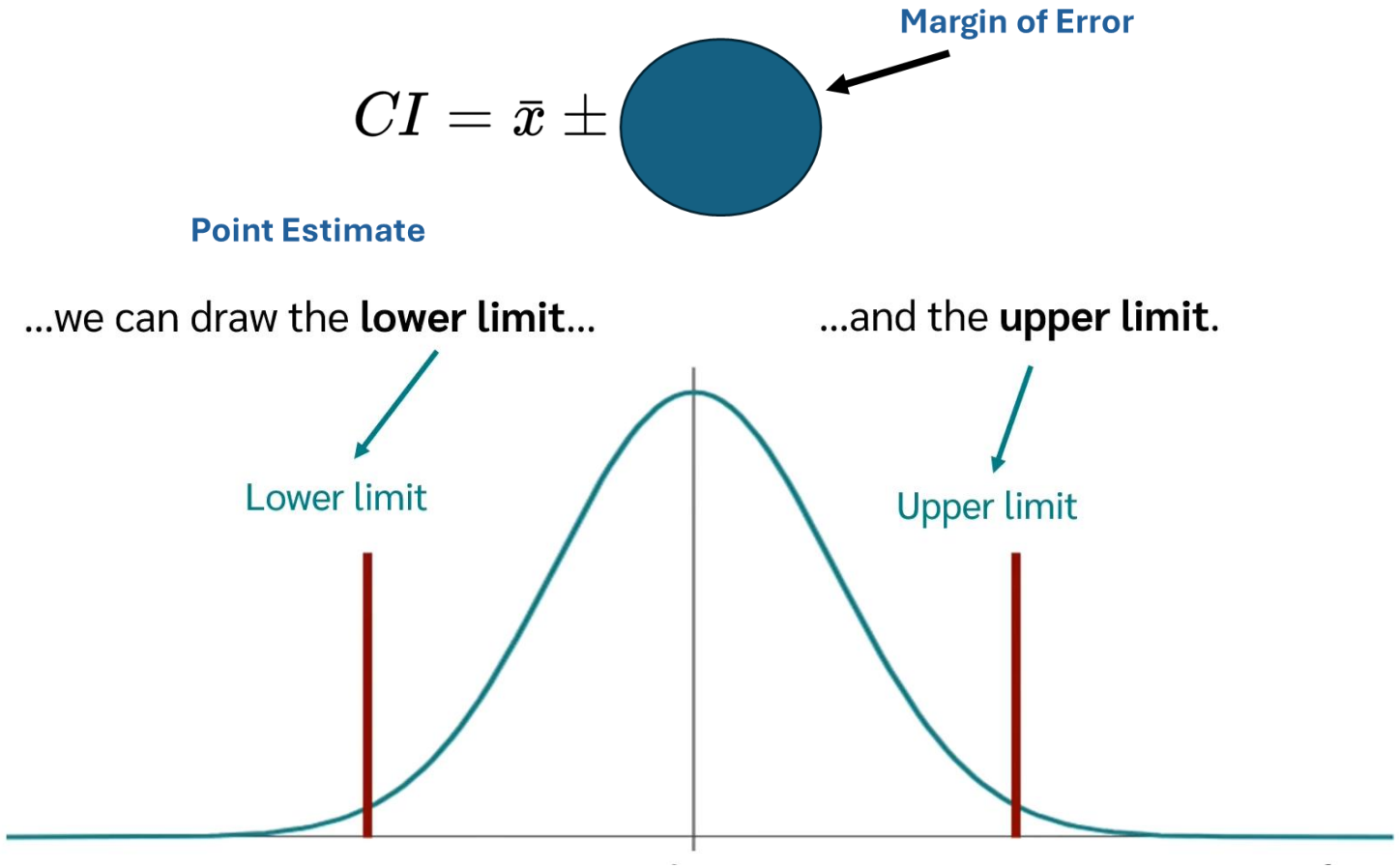
I am **95% confident** that population mean is between **180 and 200.**

There is a **95% chance** that the population mean is between 180 and 200.

95% of sample mean will fall between 180 and 200.

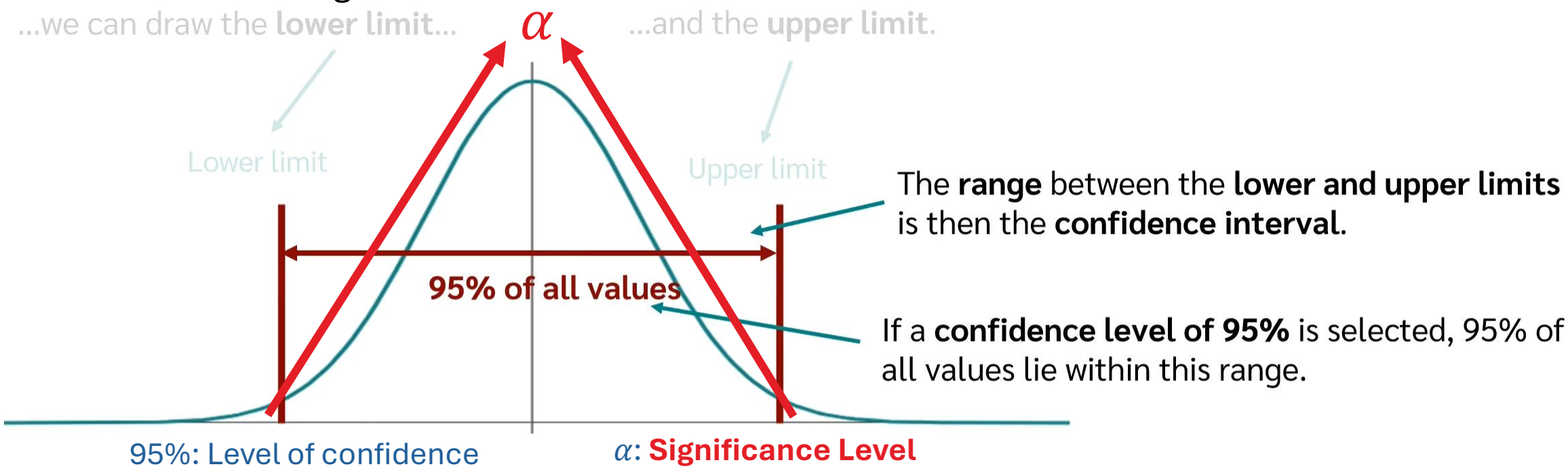
Construction of Confidence Interval

- An interval estimate consists of a range of values with an upper & lower limit
- The population parameter is expected to lie within with a certain level of confidence



Construction of Confidence Interval

- Typical used 90%, 95%, 99%
- Probability denoted by
  - $(1 - \alpha)$ known as the level of confidence
  - α is the significance level



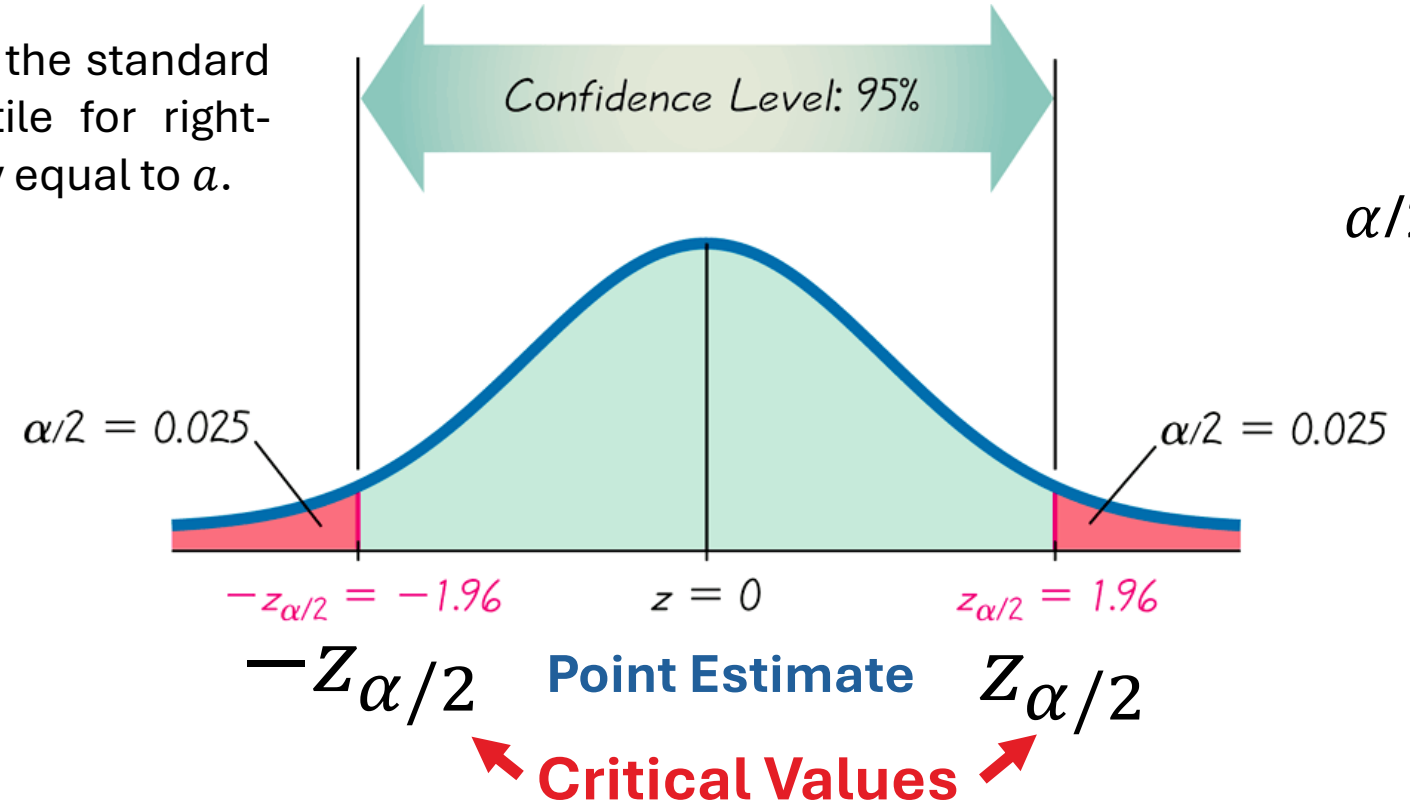
Level of confidence c is the probability that the interval estimate contains the population parameter.

How to determine the exact width of CI using α?

Critical Value

Critical Value is the value of the test statistic which defines the upper and lower bounds of a confidence interval.

Let z_α denote the standard normal quantile for right-tail probability equal to α.



$\alpha = 5\%$

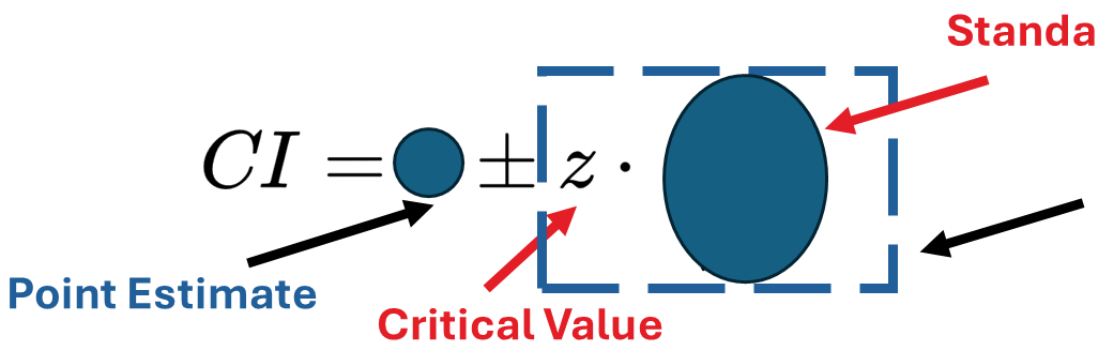
$\alpha/2 = 2.5\% = .025$

*Or other distributions depending on sampling distribution.

Margin of Error

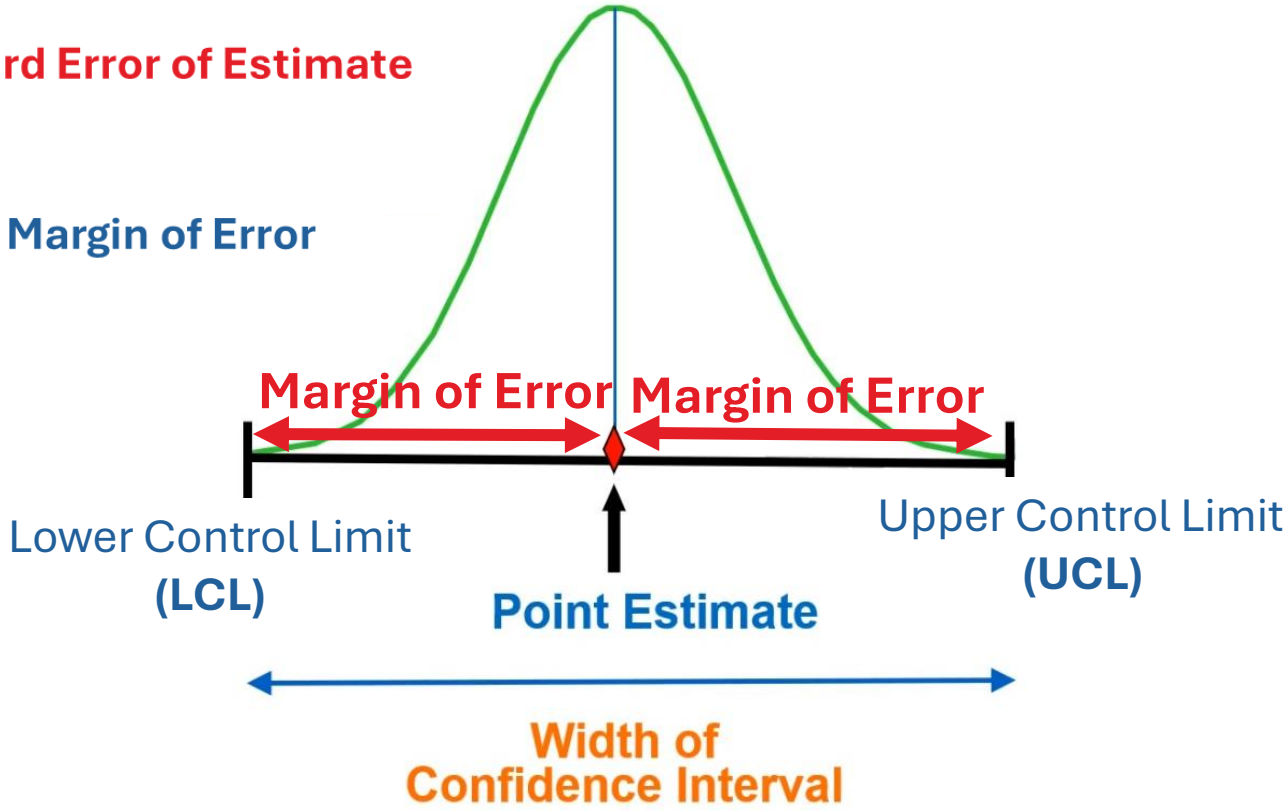
Margin of Error denoted by **E**, is the maximum likely difference (with probability 1 – α, such as 0.95) between the observed estimate and the true population parameter.

- It can be found by multiplying the critical value and the standard error.

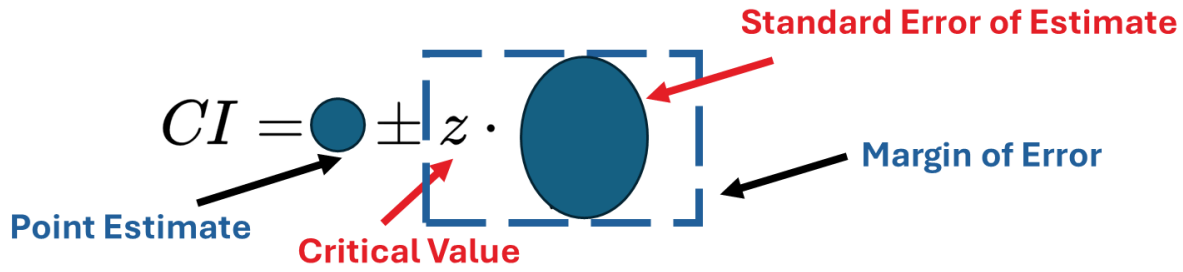


Confidence Interval Width = 2 × Margin of Error

Margin of Error = $\frac{UCL-LCL}{2}$



Confidence Interval

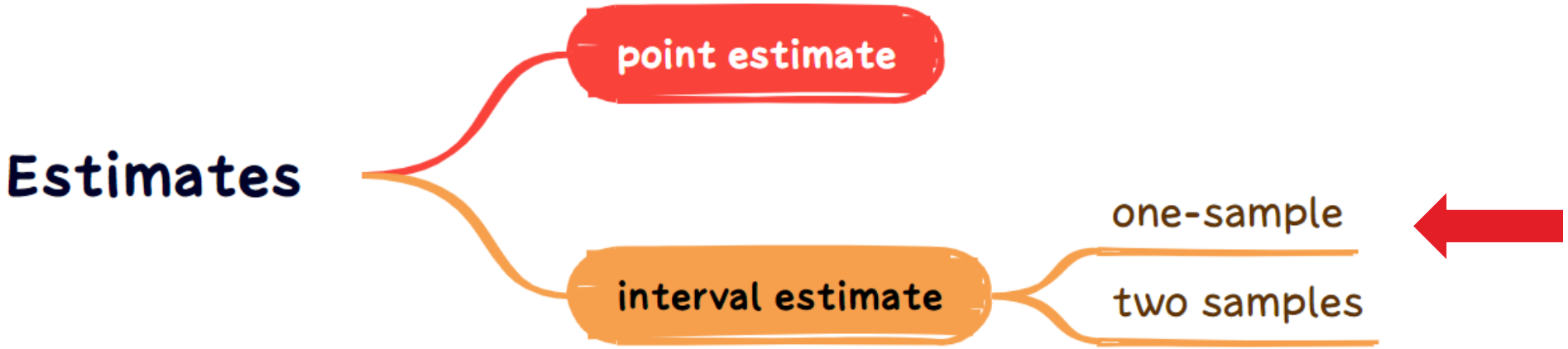


General form of a confidence interval

$$Point\ Estimate\ \pm \text{Margin of Error}$$

$$Point\ Estimate\ \pm Critical\ Value \times Standard\ Error\ of\ Estimator$$

Confidence Interval



One-sample: Mean: Large Sample

Mean with $n \geq 30$ and σ is known

μ = population mean

$\bar{x}$ = sample mean

n = number of sample

E = Margin of Error

- Standard error of sample mean: $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$
- Critical value $z_{\alpha/2}$ separating an area of $\alpha/2$ in the right tail of the standard normal distribution.

$$\bar{x} \pm E \longrightarrow \bar{x} - Z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) < \mu < \bar{x} + Z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$$

One-sample: Mean: Large Sample

A survey of 30 emergency room patients found that the average waiting time for treatment was 174.3 minutes. Assuming that the population standard deviation is 46.5 minutes, find the best point estimate of the population mean and the 99% confidence of the population mean.

(152.439, 196.161)

Confidence Level	α	Critical Value, $z_{\alpha/2}$
90%	0.10	1.645
95%	0.05	1.96
99%	0.01	2.575

One-sample: Mean: Small Sample

If σ is **known**, sample mean follow a normal distribution $\bar{x} \sim N(\mu, \frac{\sigma^2}{n})$

Recall Central Limit Theorm, $n \geq 30$

$$\bar{x} \sim N(\mu, \frac{\sigma^2}{n})$$

Then what if $n < 30$ ...or σ is unknow?

If $n<30$ or σ is unknow, sample mean follows

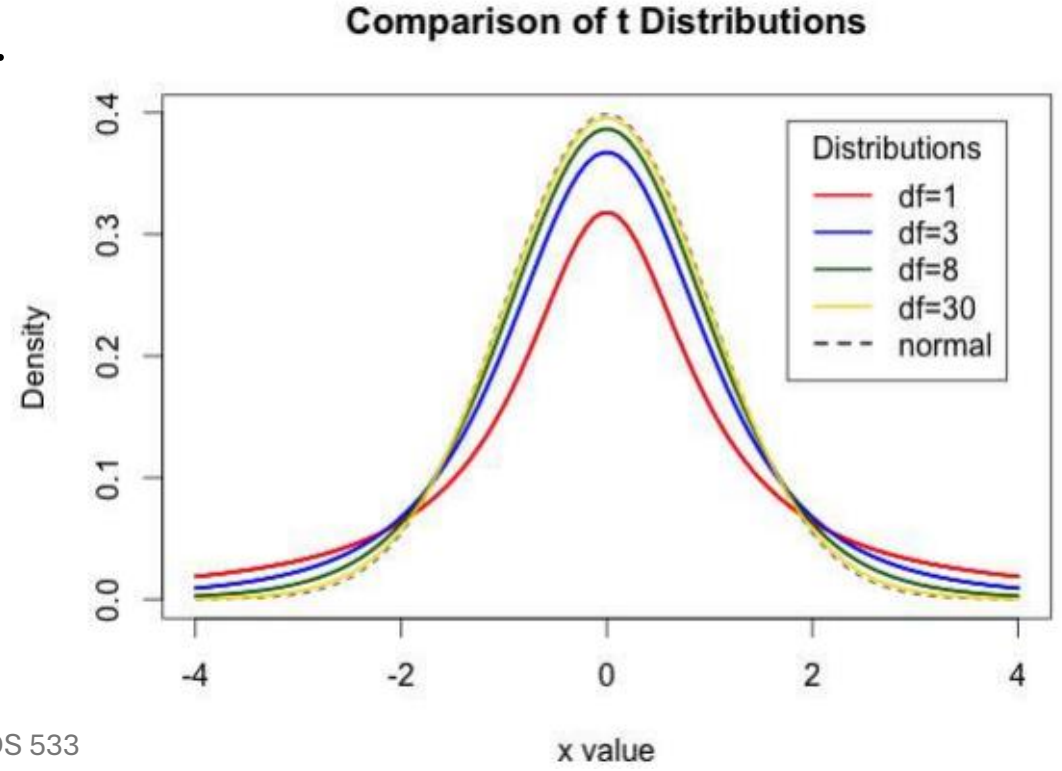
t-distribution with degree of freedom n-1, $t_{n-1} = \frac{\bar{x}-\mu}{\frac{s}{\sqrt{n}}}$

t distribution

Properties of the t-distribution

1. The *t*-distribution is bell-shaped and symmetric about the mean $t = 0$.
2. Its shapes are determined by a parameter called the degrees of freedom, df = n – 1.
3. As the sample size increases, $n \geq 30$, t-distribution approaches the standard normal distribution.
4. The total area under a t-curve is 1 or 100%.

The tails in the *t*-distribution are "thicker" than those in the standard normal distribution.



One-sample: Mean: Small Sample

Mean with $n < 30$ and σ is unknown

μ = population mean

$\bar{x}$ = sample mean

n = number of sample

E = Margin of Error

- Standard error of sample mean: $s_{\bar{x}} = \frac{s}{\sqrt{n}}$
- Critical value $t_{\alpha/2}$ separating an area of $\alpha/2$ in the right tail of the standard normal distribution.
- degree of freedom: n-1

$$\bar{x} \pm E \longrightarrow \bar{x} - t_{n-1,\alpha/2}\left(\frac{s}{\sqrt{n}}\right) < \mu < \bar{x} + t_{\mathrm{n}-1,\alpha/2}\left(\frac{s}{\sqrt{n}}\right)$$

One-sample: Mean: Small Sample

Ten randomly selected people were asked how long they slept at night. The mean time was 7.1 hours, and the standard deviation was 0.78 hour. Find the 95% confidence interval of the mean time. Assume the variable is normally distributed.

(6.542,7.658)

TABLE A-3 *t* Distribution: Critical *t* Values

	Area in One Tail				
	0.005	0.01	0.025	0.05	0.10
Degrees of Freedom	**Area in Two Tails**				
	0.01	0.02	0.05	0.10	0.20
1	63.657	31.821	12.706	6.314	3.078
2	9.925	6.965	4.303	2.920	1.886
3	5.841	4.541	3.182	2.353	1.638
4	4.604	3.747	2.776	2.132	1.533
5	4.032	3.365	2.571	2.015	1.476
6	3.707	3.143	2.447	1.943	1.440
7	3.499	2.998	2.365	1.895	1.415
8	3.355	2.896	2.306	1.860	1.397
9	3.250	2.821	2.262	1.833	1.383
10	3.169	2.764	2.228	1.812	1.372

One-sample: Proportion

Proportion

p = population proportion

$\hat{p}$ = sample proportion

n = number of sample values

E = Margin of Error

- Standard error of sample proportions: $\sigma_{\hat{p}} = \sqrt{\frac{\hat{p}\hat{q}}{n}}$
- Critical value $z_{\alpha/2}$ separating an area of $\alpha/2$ in the right tail of the standard normal distribution.

$$\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}} < p < \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}$$

One-sample: Proportion

A survey of 1404 respondents found that 323 students paid for their education by student loans. Find the 90% confidence of the true proportion of students who paid for their education by student loans.

(0.211, 0.249)

Confidence Level	α	Critical Value, $z_{\alpha/2}$
90%	0.10	1.645
95%	0.05	1.96
99%	0.01	2.575

One-sample: Variance

Variance

The formula for the **sample variance** is

$$s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}$$

Chi-Square Distribution

In a normally distributed population with variance σ^2, if we randomly select independent samples of size n and, for each sample, compute the sample variance s^2. The sample statistic (variance) $\chi^2 = (n-1)\ s^2 / \sigma^2$ has a sampling distribution called the **chi-square distribution.** The sample statistic is

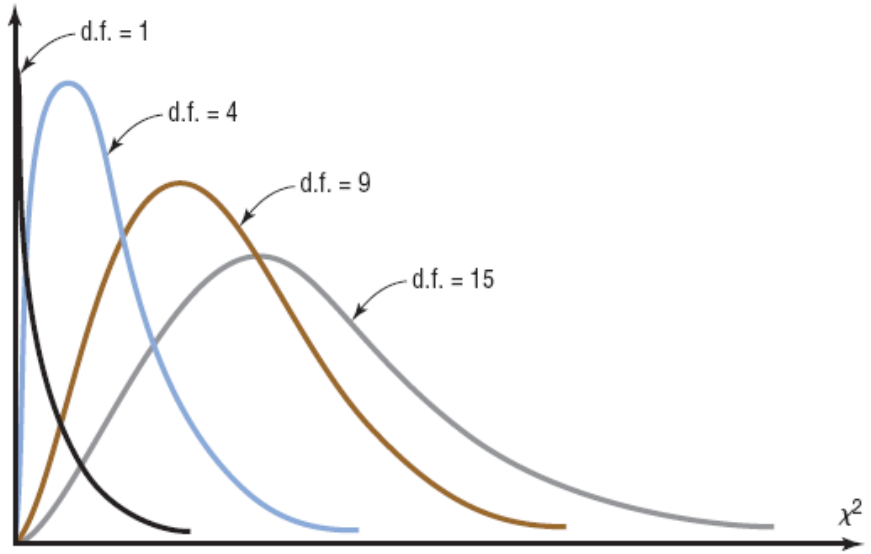
$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$$

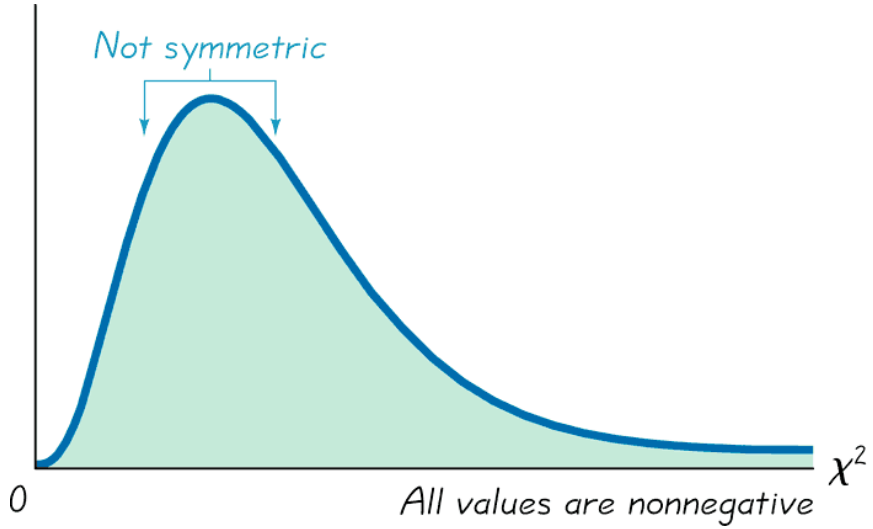
With degrees of freedom df = n – 1, formally denoted as χ^2_{n-1}.

Chi-square distribution

Properties of the Chi-Square Distribution

1. The chi-square variable is similar to the t variable in that its distribution is a family of curves based on the number of **degrees of freedom**.
2. A Chi-Square variable cannot be negative, and the distributions are skewed to the right.
3. As the number of degrees of freedom increases, the distribution becomes more symmetric.

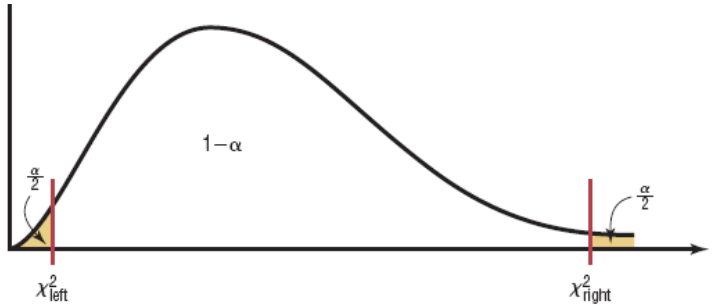




One-sample: Variance

Population Variance

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$$



Not symmetric: Confidence interval estimate of σ^2 does not fit a format of $s^2 - E < \sigma^2 < s^2 + E$, so we must do **separate calculations for the upper and lower confidence interval limits.**

For a confidence level 1-α, we have $P(\chi^2_{1-\alpha/2} \leq \frac{(n-1)s^2}{\sigma^2} \leq \chi^2_{\alpha/2}) = 1 - \alpha$

$$\frac{(n-1)s^2}{\chi^2_R} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_L}, \text{df} = n - 1$$

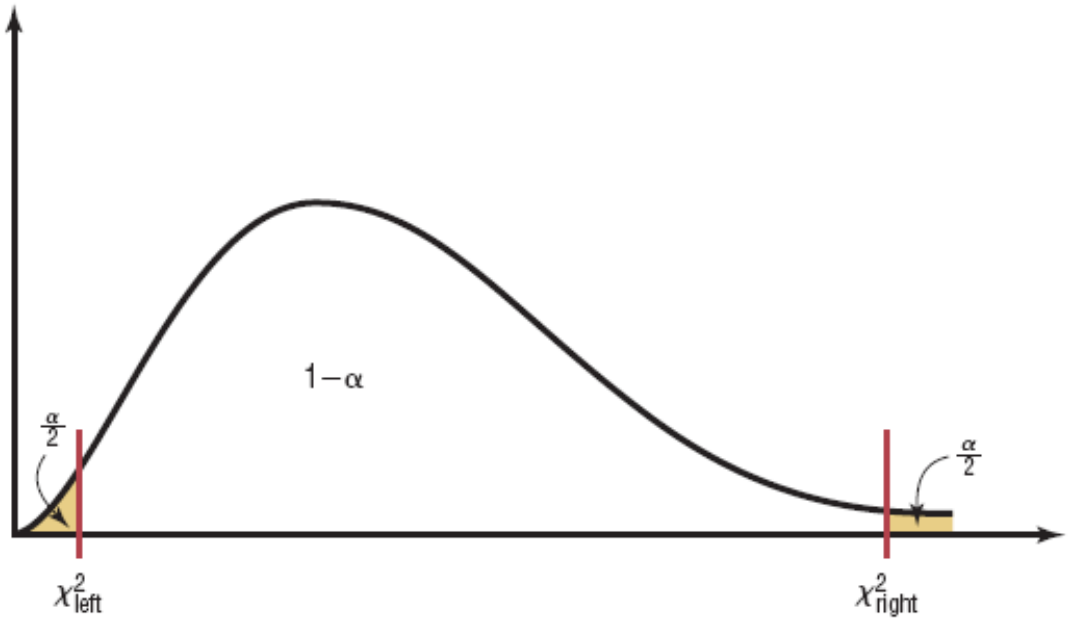
$\chi^2_{n-1,1-\alpha/2}$: lower critical value (χ^2_L)

$\chi^2_{n-1,\alpha/2}$: upper critical value (χ^2_R)

Standard Deviation: $\sqrt{\frac{(n-1)s^2}{\chi^2_R}} < \sigma < \sqrt{\frac{(n-1)s^2}{\chi^2_L}}$

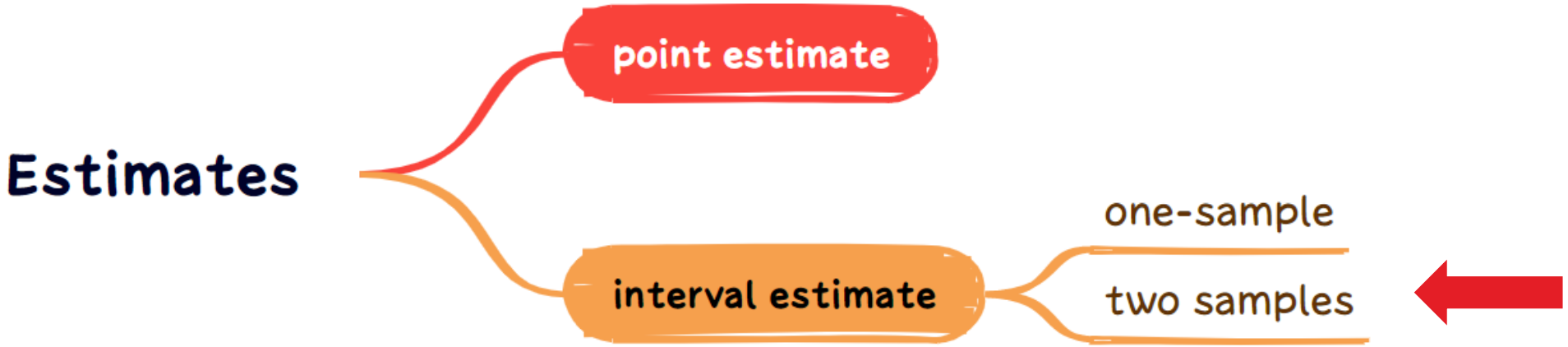
One-sample: Variance

A group of 22 subjects took an IQ test. The subjects had a standard deviation IQ score of 14.3. Construct a 95% confidence interval estimate of σ^2, the variance of the population from which the sample was obtained. (Assume Normally distributed data)



(121.036,417.611)

Confidence Interval



Two-samples

	Population 1	Sample 1	Population 2	Sample 2
Mean	μ_1	$\bar{x}_1$	μ_2	$\bar{x}_2$
Variance	σ^2_1	s^2_1	σ^2_2	s^2_2
Std dev	σ_1	s_1	σ_2	s_2
Size	N_1	n_1	N_2	n_2
Proportion	P_1		P_2	

Differences as parameters!

- $X_1 \sim N(\mu_1, \sigma^2)$
- $X_2 \sim N(\mu_2, \sigma^2)$

With unknown values for the parameters and with $\sigma_1^2 = \sigma_2^2 = \sigma^2$ or not equal.

Two-samples: Comparing means

$\sigma_1^2 \neq \sigma_2^2$

μ_1 = population 1 mean; μ_2 = population 2 mean

$\bar{x}_1$ = sample 1 mean; $\bar{x}_2$ = sample 2 mean

n_1 = number of sample 1; n_2 = number of sample 2

- Standard error of mean difference:

$$se = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

A $100(1 - \alpha)\%$ confidence interval for $\mu_1 - \mu_2$ is

$$(\bar{x}_1 - \bar{x}_2) - Z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} < \mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + Z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

*Use s_1^2 , s_2^2 and $t_{\alpha/2}$ if σ_1^2 and σ_1^2 are unknown with d.f.=smaller of n_1 or n_2.

Two-samples: Comparing means

$\sigma_1^2 = \sigma_2^2$

μ_1 = population 1 mean; μ_2 = population 2 mean

$\bar{x}_1$= sample 1 mean; $\bar{x}_2$= sample 2 mean

n_1 = number of sample 1; n_2 = number of sample 2

- Standard error of mean difference (**pooled estimate** s):

$$se = s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}, \text{ where } s_p = \sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}$$

- **Sp** is the **pooled estimate of the common standard deviation**

A $100(1-\alpha)\%$ confidence interval for $\mu_1 - \mu_2$ is

$$(\bar{x}_1 - \bar{x}_2) - t_{n_1+n_2-2;\alpha/2}\,(se) < \mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + t_{n_1+n_2-2;\,\alpha/2}(se)$$

Two-samples: Comparing means

A plant that operates two shifts per week would like to consider the difference in productivity for the two shifts. The number of units that each shift produces on each of the 5 working days is recorded in the following table:

	Monday	Tuesday	Wednesday	Thursday	Friday
Shift 1	263	288	290	275	255
Shift 2	265	278	277	268	244

Assuming that the number of units produced by each shift is normally distributed and that the population standard deviations for the two shifts are equal construct a 99% confidence interval for the difference in mean productivity for the two shifts and comment on the result.

[–23,0221;38,6221]

Two-samples: Comparing proportions

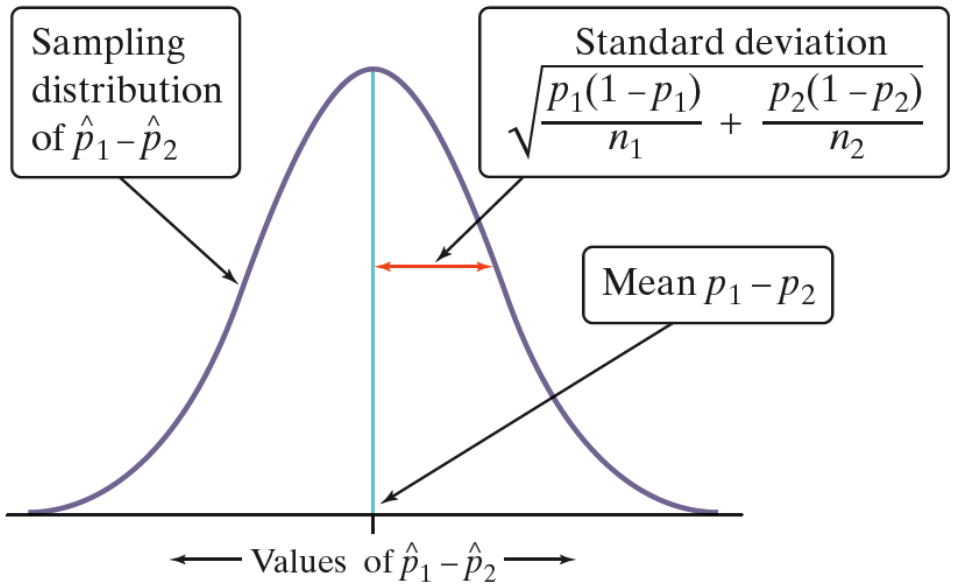
P_1 = population 1 proportion; P = population 2 proportion

$\hat{p}_1$ = sample 1 proportion; $\hat{p}_2$ = sample 2 proportion

n_1 = number of sample 1; n_2 = number of sample 2

- Standard error of proportion difference :

$$se = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$



A $100(1 - \alpha)\%$ confidence interval for $P_1 - P_2$ is

$$(\hat{p}_1 - \hat{p}_2) - Z_{\alpha/2}(se) < P_1 - P_2 < (\hat{p}_1 - \hat{p}_2) + Z_{\alpha/2}(se)$$

Two-samples: Comparing proportions

Two groups of males are polled concerning their interest in a new electric razor that has four cutting edges. A sample of 64 males under the age of 40 indicated that only 12 were interested while in a sample of 36 males over the age of 40, only 8 indicated an interest. Construct a 95% confidence interval for the difference between age group populations.

[–0,2008;0,1314]

Two-samples: Comparing variance

One sample case:

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$$

The **F distribution** is the ratio of two variance estimates:

$$F = \frac{s_1^2}{s_2^2}$$

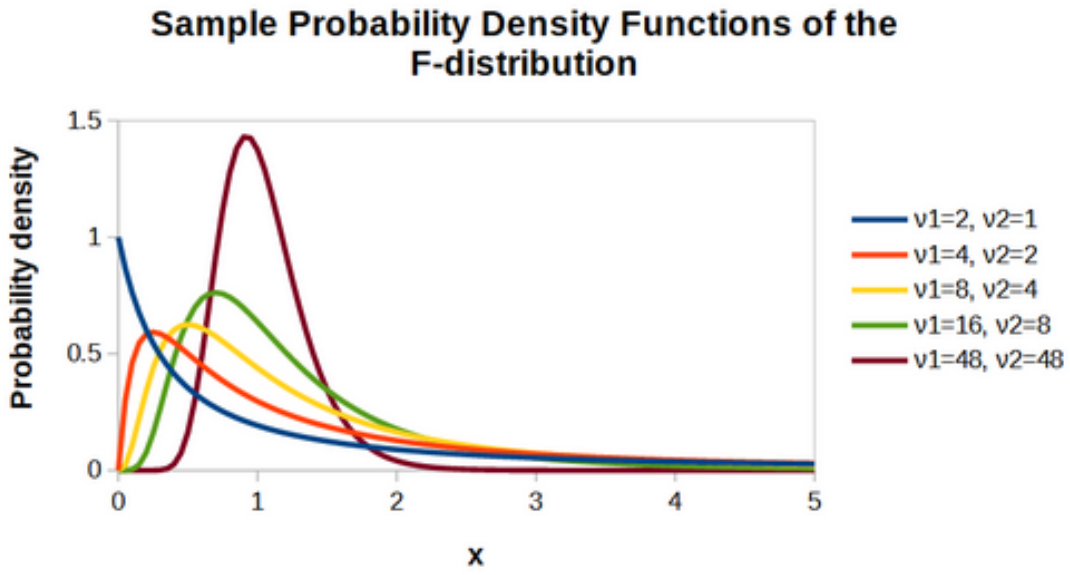
Also the **ratio of two Chi-squares**, each divided by its degrees of freedom:

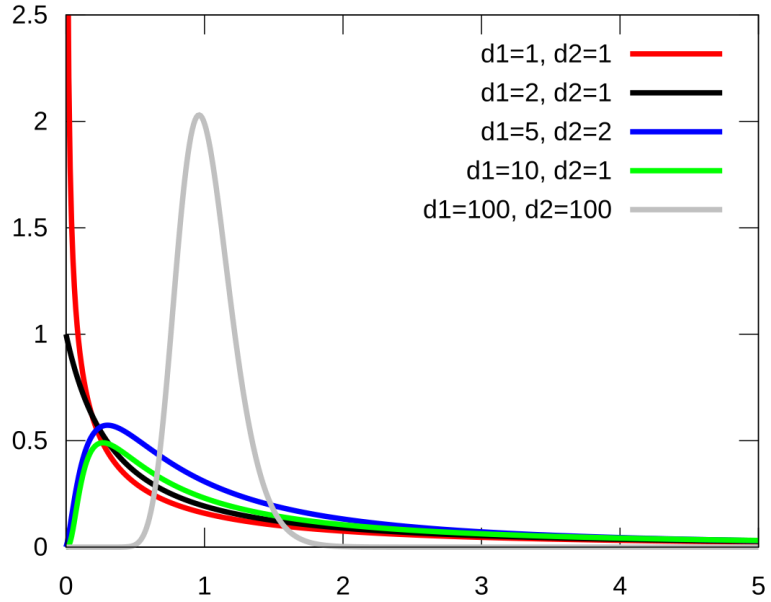
$$F = \frac{\chi^2_{(n_1-1)}/(n_1-1)}{\chi^2_{(n_2-1)}/(n_2-1)}$$

F distribution

Properties of the F Distribution

1. The F-distribution are generally not symmetric and skewed to the right.
2. The value of F can be 0 or positive, but they cannot be negative.
3. There is a different F distribution for each pair of **degrees of freedom** for the numerator and denominator.
4. It shape depends upon the **degrees of freedom** in the numerator and denominator.





Tw-samples: Comparing variance

Confidence Interval for $\frac{\sigma_1^2}{\sigma_2^2}$

Not symmetric: Confidence interval estimate of σ^2 does not fit a format of $s^2 - E < \sigma^2 < s^2 + E$, so we must do **separate calculations for the upper and lower confidence interval limits.**

For a confidence level 1-α, we have

$$P(F_{1-\alpha/2}(n_2-1, n_1-1) \leq \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} \leq F_{\alpha/2}(n_2-1, n_1-1)) = 1-\alpha$$

$$\frac{s_1^2}{s_2^2}\frac{1}{F_{\alpha/2}(n_1-1, n_2-1)} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{s_1^2}{s_2^2}F_{\alpha/2}(n_2-1, n_1-1)$$

where $\frac{1}{F_{\alpha/2}(n_1-1,n_2-1)} = F_{1-\alpha/2}(n_2-1, n_1-1)$

Two-samples: Comparing variance

A criminologist is interested in comparing the consistency of the lengths of sentences given to people convicted of robbery by two judges. A random sample of 17 people convicted of robbery by judge 1 showed a standard deviation of 2.53 years, while a random sample of 21 people convicted by judge 2 showed a standard deviation of 1.34 years. Construct a 95% confidence interval for the ratio of the two populations variances. Does the data suggest that the variances of the lengths of sentences by the two judges differ? Motivate your answer.

[1,3979;9,5536]

Two-samples: Conclusion

- **If the two confidence intervals do not overlap,** we can conclude that there is a statistically significant difference in the two population values at the given level of confidence; or alternatively

- **If the confidence interval for the difference does not contain zero,** we can conclude that there is a statistically significant difference in the two population values at the given level of confidence. (except for F)

Determine the Sample Size

Everything have done so far has assumed that a sample has **ALREADY** been taken

- We often need to know how large a sample should we take to construct the confidence interval
- Many factors can affect sample size such as budget, time and ease of selection
- We will now look at how to determine the proper sample size (from a statistical perspective)

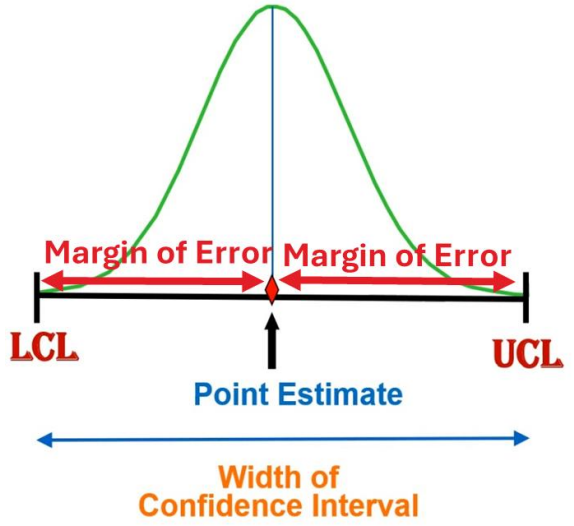
Sample Size for Estimating Parameter

Everything have done so far has assumed that a sample has **ALREADY** been taken

- We often need to know how large a sample should we take to construct the confidence interval
- Many factors can affect sample size such as budget, time and ease of selection
- We will now look at how to determine the proper sample size (from a statistical perspective)

Sample Size for Estimating Mean

- Confidence level $(1 - \alpha)$
- Accepted sampling error E
- Need to know σ, else use s



$$Margin\ of\ Error = Critical\ Value \times Standard\ Error\ of\ Estimate$$

$$n = \left(\frac{z_{\alpha/2} \cdot \sigma}{E}\right)^2 \quad \text{or} \quad n = \left(\frac{t_{\alpha/2} \cdot s}{E}\right)^2$$

NOTE: Sample size n is required to be a whole number, always round UP to the next largest integer.

Sample Size for Estimating Proportion

- Confidence level $(1-\alpha)$
- Accepted sampling error E
- Need to know P, else use $\hat{p}$

$$Margin\ of\ Error = Critical\ Value \times Standard\ Error\ of\ Estimate$$

$$n = \frac{(Z_{\frac{\alpha}{2}})^2 p(1-p)}{E^2}$$

When no estimate of $\hat{p}$ is not known: $\hat{p} = \hat{q} = 0.5$ $\quad n = \frac{(Z_{\alpha/2})^2 0.25}{E^2}$

NOTE: Sample size n is required to be a whole number, always round UP to the next largest integer.

Sample Size for Estimating Proportion

How many statistics students must be randomly selected for IQ tests if we want 95% confidence that the sample mean is within 3 IQ points of the population mean? Assume normal distribution with $\sigma = 15$.

97

Compute Confidence Interval

Depending on the situation, we have to use a different approach

- **Exact method**: based on a known distribution of $\hat{\theta}$
- **Asymptotic method**: based on asymptotic normality of the MLE

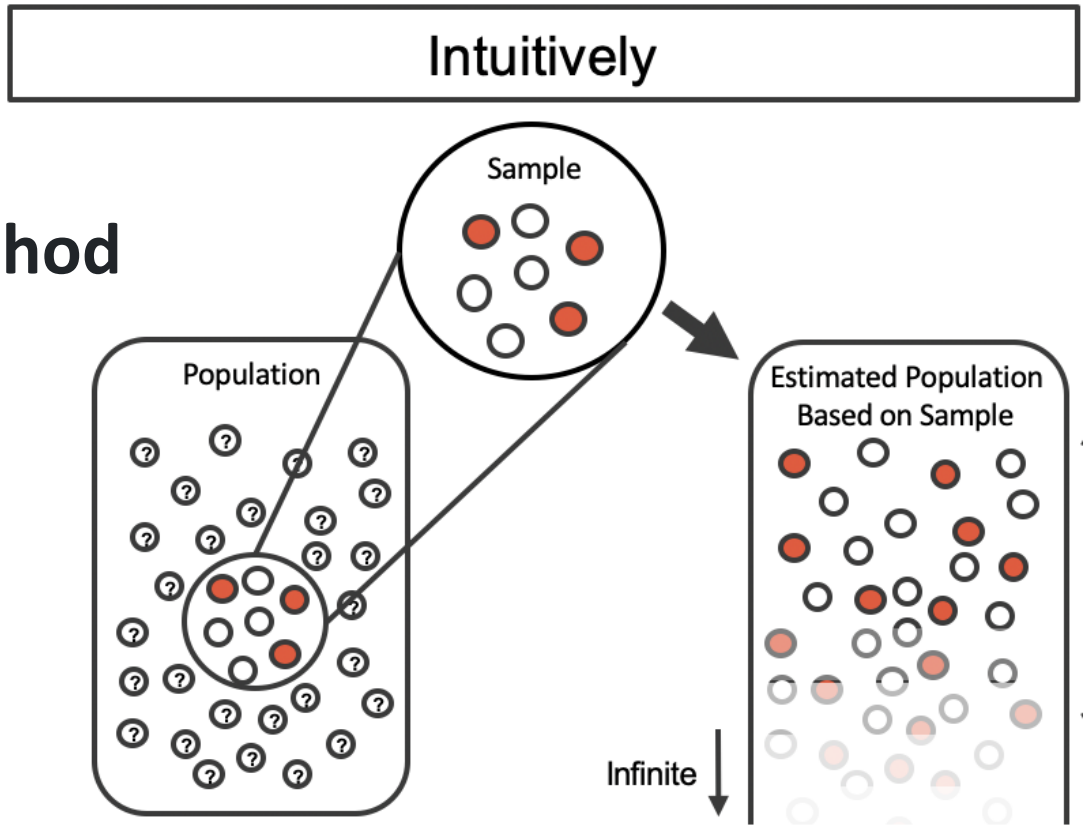
➢ **Bootstrap method**: more elaborate resampling technique

The Bootstrap

- Some statistics are too complex to have a simple standard error formula.
- Even if the statistic has an approximately normal sampling distribution, without the standard error we cannot use the confidence interval formula of a point estimate plus and minus margin of error.

For such cases, a computational resampling method – bootstrap is a simple and powerful alternative.

The Bootstrap

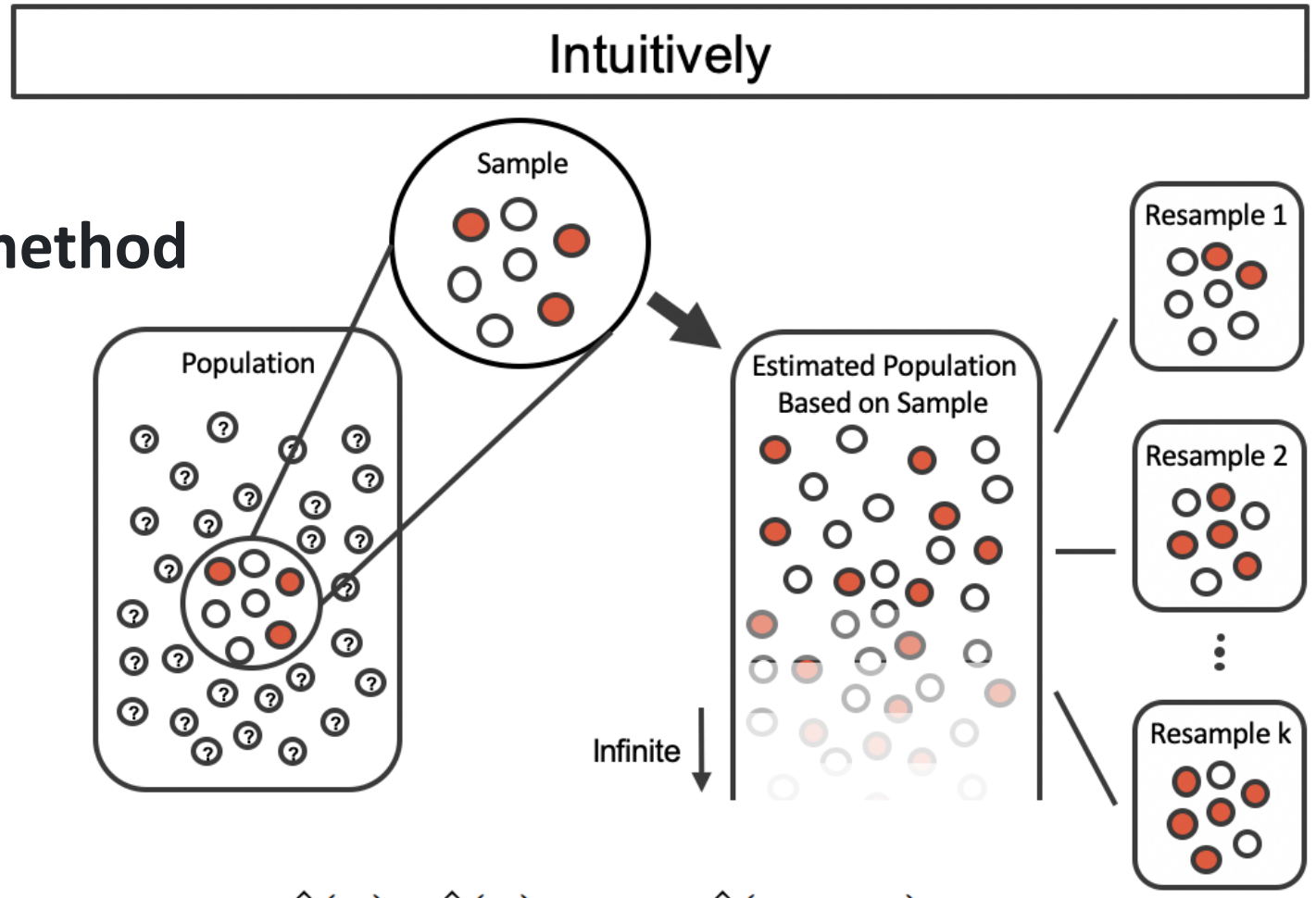


Traditional method

instead of using a "resample from the population" approach, bootstrapping uses a "resample from the sample" approach.

The Bootstrap

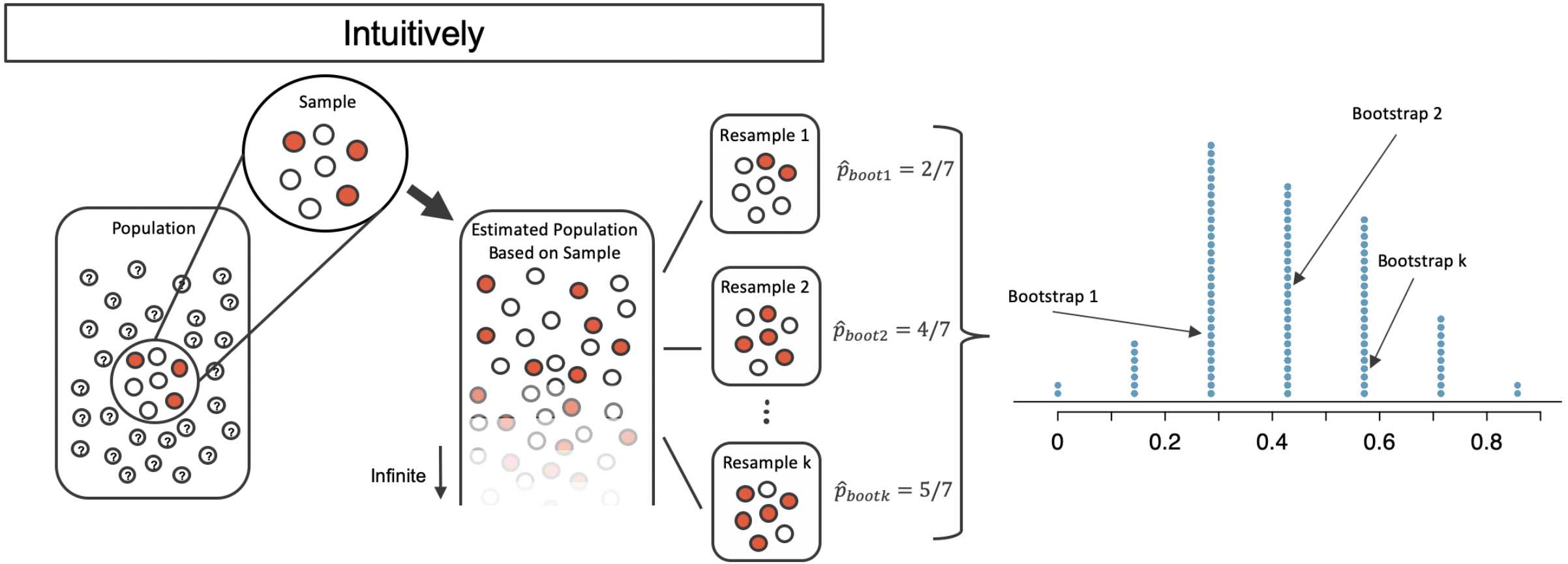
Bootstrap method



$$\{\hat{\theta}^{(1)}, \hat{\theta}^{(2)}, \ldots, \hat{\theta}^{(10000)}\} \text{ of } \theta$$

The Bootstrap

Bootstrap method



The Bootstrap

Which of the following statements is most likely accurate in relation to bootstrap analysis?...

A.Bootstrap analysis aims to deduce statistics about population parameters from a singular sample.

B.Bootstrap analysis involves the repeated extraction of samples of equal size, with replacement, from the initial population.

C.During bootstrap analysis, it is necessary for analysts to determine probability distributions for primary risk factors that govern the underlying random variables.

The Bootstrap

Pros

- Can be used for non-parametric statistics
- It approximates whole distribution of $\hat{\theta}$
- Relatively accurate for computing variance of $\hat{\theta}$
- …

Cons

- If n is very small bootstrap may fail
- …

Summary

Parameters being estimated	Begin Derivation with	Distribution (D.F.)	100(1-α)% conf. interval
$\mu(\sigma^2$ known $)$	$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$	Z	$\bar{X} \pm z_{\alpha/2}\,\sigma/\sqrt{n}$
$\mu(\sigma^2$ unknown$)$	$\frac{\bar{X}-\mu}{S/\sqrt{n}}$	T_{n-1}	$\bar{X} \pm t_{\alpha/2}\,S/\sqrt{n}$
σ^2	$(n-1)S^2/\sigma^2$	χ^2_{n-1}	$L_1 = \frac{(n-1)S^2}{\chi^2_{\alpha/2}}, L_2 = \frac{(n-1)S^2}{\chi^2_{1-\alpha/2}}$
p (proportion)	$\frac{\hat{p}-p}{\sqrt{p(1-p)/n}}$	$\sim Z$	$\hat{p} \pm z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}$
$\mu_1-\mu_2$ $(\sigma_1^2,\sigma_2^2$ known$)$	$\frac{(\bar{X}_1-\bar{X}_2)-(\mu_1-\mu_2)}{\sqrt{\sigma_1^2/n_1+\sigma_2^2/n_2}}$	Z	$(\bar{X}_1-\bar{X}_2) \pm z_{\alpha/2}\sqrt{\sigma_1^2/n_1+\sigma_2^2/n_2}$
$\mu_1-\mu_2$ $(\sigma_1^2,\sigma_2^2$ unknown & equal$)$	$\frac{(\bar{X}_1-\bar{X}_2)-(\mu_1-\mu_2)}{\sqrt{S_p^2(1/n_1+1/n_2)}}$	$T_{n_1+n_2-2}$	$(\bar{X}_1-\bar{X}_2) \pm t_{\alpha/2}\sqrt{S_p^2(1/n_1+1/n_2)}$ $S_p^2=\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}$
$\mu_1-\mu_2$ $(\sigma_1^2,\sigma_2^2$ unknown & unequal$)$	$\frac{(\bar{X}_1-\bar{X}_2)-(\mu_1-\mu_2)}{\sqrt{S_1^2/n_1+S_2^2/n_2}}$	T_γ	$(\bar{X}_1-\bar{X}_2) \pm t_{\alpha/2}\sqrt{S_1^2/n_1+S_2^2/n_2}$ $\gamma=\left[S_1^2/n_1+S_2^2/n_2\right]^2 \Big/ \left[\frac{(S_1^2/n_1)^2}{n_1-1}+\frac{(S_2^2/n_2)^2}{n_2-1}\right]$

Lab Time

Review

www.kahoot.it