Applied Statistics
Lecture 2

Prof.ssa Chiara Seghieri, Dott.ssa Costanza Tortù

Laboratorio di Management e Sanità, Istituto di Management, L'EMbeDS
Scuola Superiore Sant'Anna, Pisa
c.seghieri@santannapisa.it
c.tortu@santannapisa.it

Outline

1. Introduction to Statistical Inference
2. Parameters, estimators, estimates
3. Sampling distribution of estimators
4. Desiderable properties of estimators
5. Different Approaches to point estimate
6. Point Estimates and Confidence Intervals
7. Width of a confidence interval

1) Introduction to Statistical Inference

Inaccurate representation of reality

Without observing the **whole population** of interest, it is impossible to have a complete knowledge on the phenomenon we are investigating.

Working with a **sample** implies being aware that the best you can get is an **inaccurate representation** of reality. In addition, results are affected by measurement errors.

All the findings we can obtain are affected by **uncertainty!**

$$\hat{\theta} = \theta + \varepsilon$$

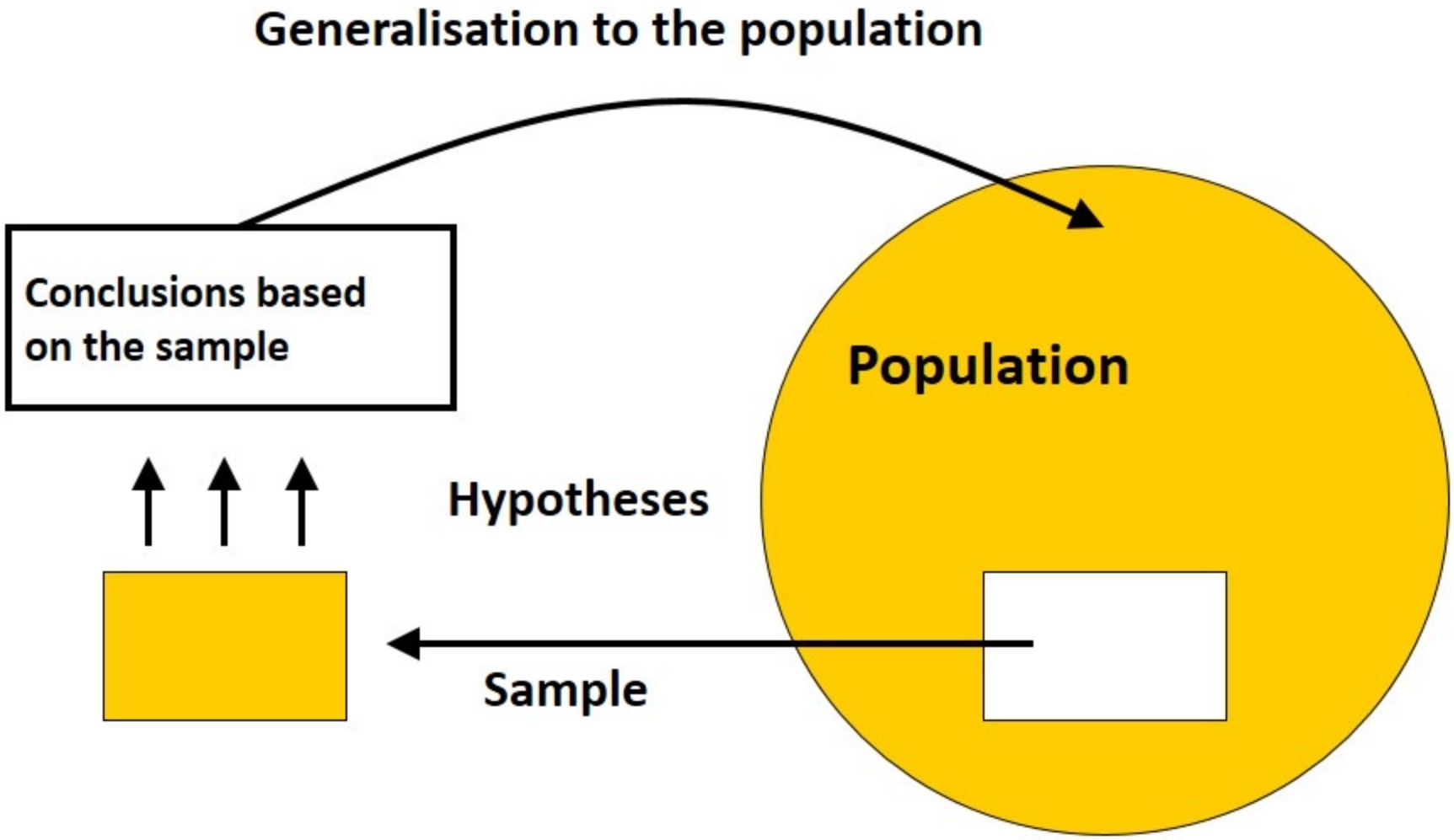
θ True value of the parameter

$\hat{\theta}$ Estimated value of the parameter

ε Error

The error component can't be determined. We have to **infer** on reality while being aware that this is uncertain.

The idea of statistical inference

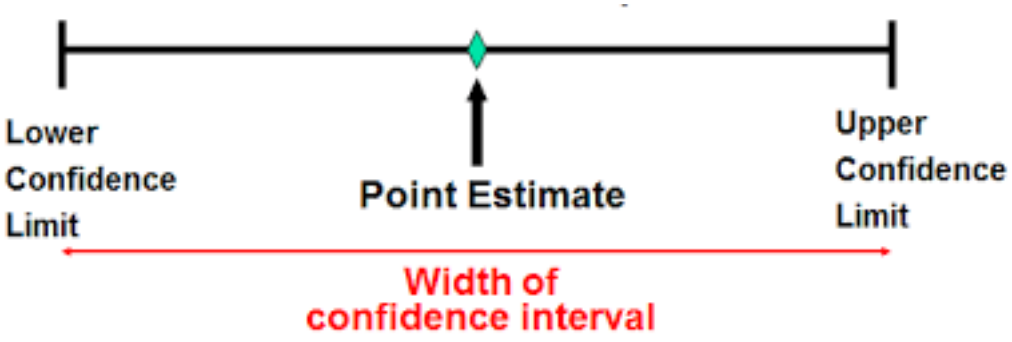


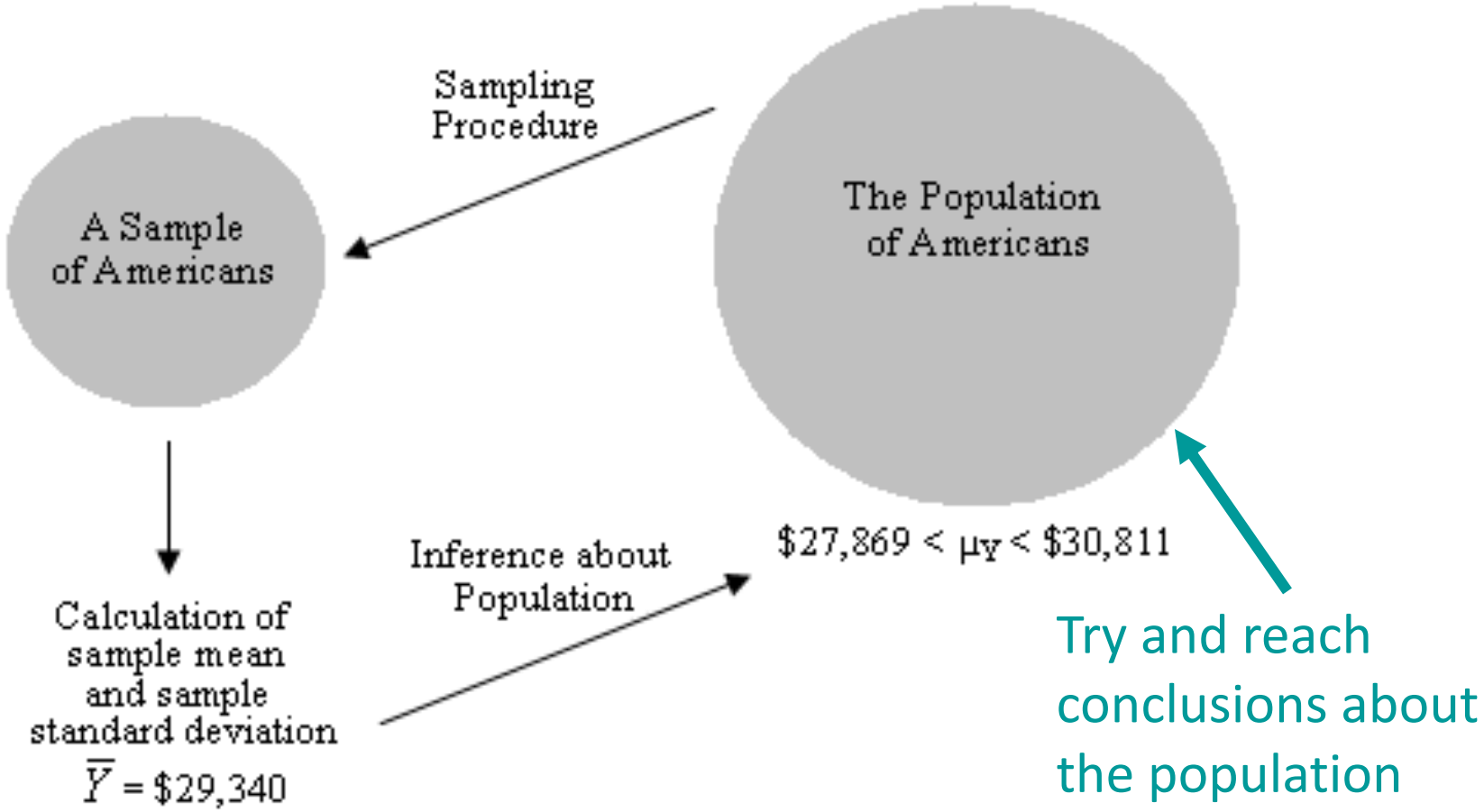
Statistical Inference: Estimation

Goal: How can we use sample data to estimate values of population parameters?

Point estimate: A single statistic value that is the "best guess" for the unknow parameter value of the population.

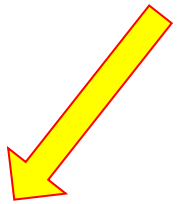
Interval estimate: An interval of numbers around the point estimate, that has a fixed "confidence level" of containing the parameter value. This interval is known as a ***confidence interval***.




Sampling
Procedure
A Sample
of Americans
The Population
of Americans
Inference about
Population
$27,869 < μY < $30,811
Calculation of
sample mean
and sample
standard deviation
$\overline{Y}$ = $29,340
Try and reach
conclusions about
the population

2) Estimands, estimators, estimates

Parameter: synthetical statistical measure that is observed in a population (mean, variance, probability of an event).
It is fixed but might not be observed in reality, it is **estimated** by sample observations.



Estimator/Statistic: function of the sample observations that allows the researchers to understand

Estimate: value of the estimator associated to a particular sample.

Common Estimators: statistics used to estimate population parameters

- The sample mean $\bar{x}$ is the most common estimator of the population mean (μ).

- The sample variance s^2, is the most common estimator of the population variance (σ^2).

-The sample standard deviation, is the most common estimator of the population standard deviation (σ).

- The sample proportion $\hat{p}$, is the most common estimator of the population proportion (p).

3) Sampling distribution of estimators

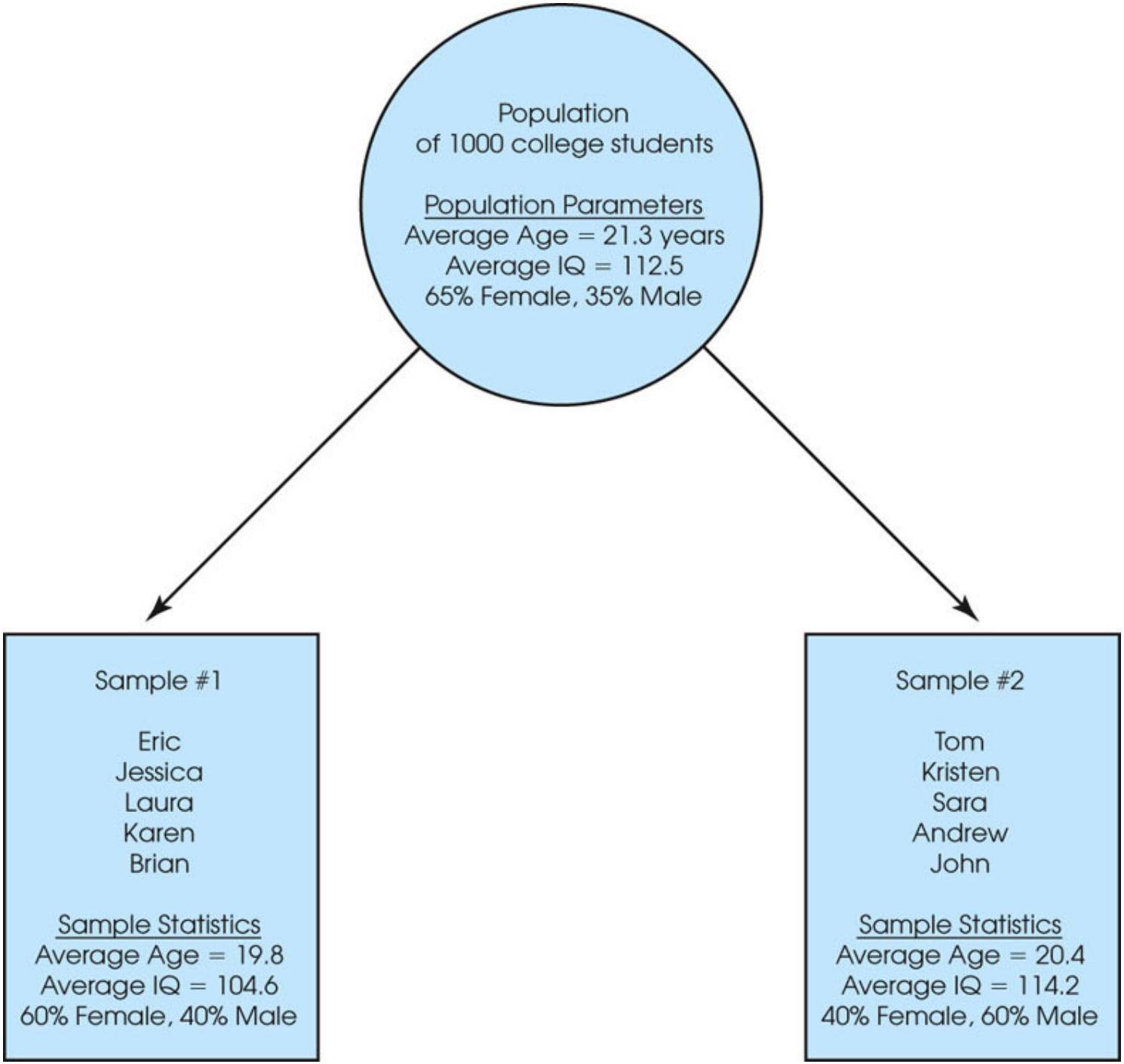
The sampling distribution of a statistic

To sum up:

- the numerical value of a statistic cannot be expected to give us the exact value of the population parameter;
- the observed value of a statistic depends on the particular sample that happens to be selected;
- there will be some variability in the values of a statistic over different occasions of sampling.

The sampling distribution of the estimator:

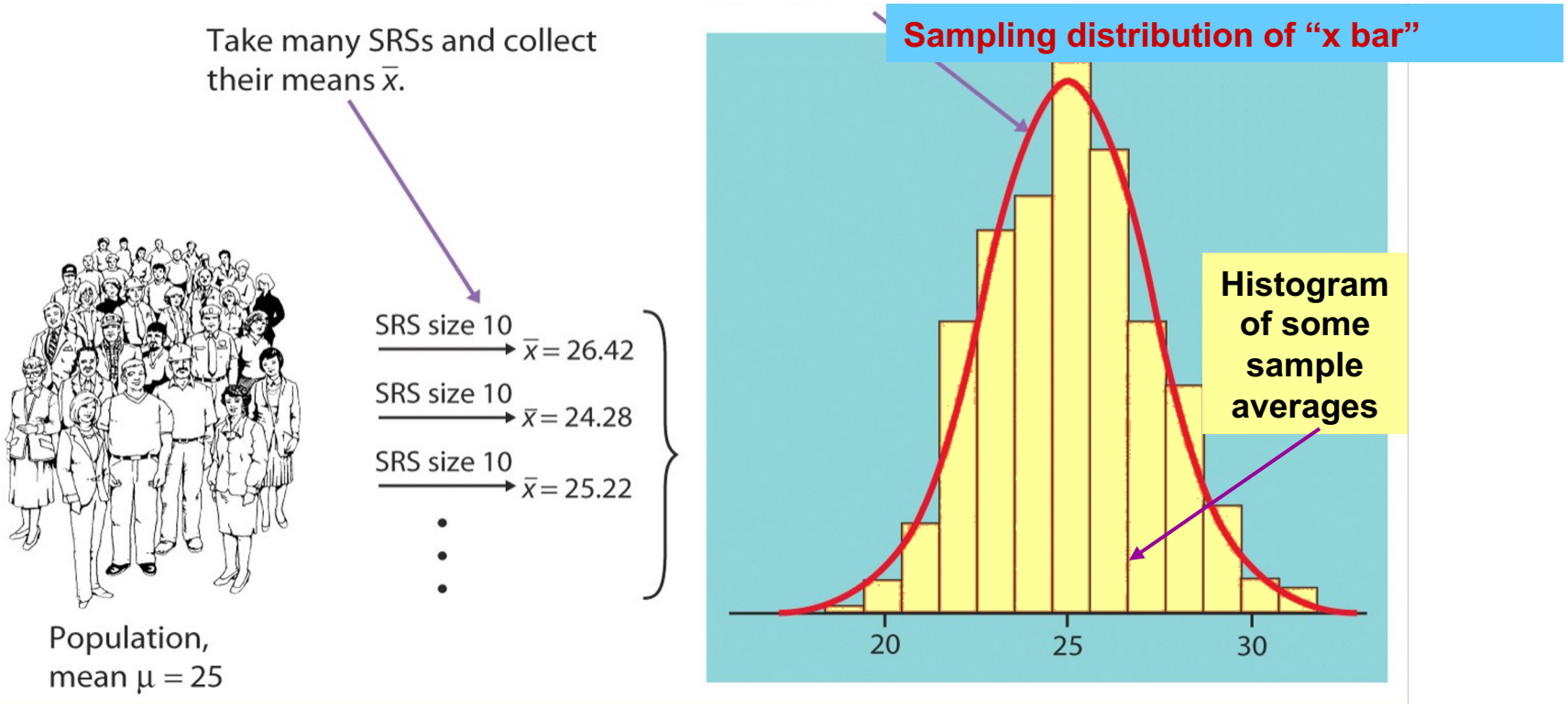
The **distribution of values taken by the statistic in all possible samples** of the **same size** from the **same population**.


Population
of 1000 college students
Population Parameters
Average Age = 21.3 years
Average IQ = 112.5
65% Female, 35% Male
Sample #1
Eric
Jessica
Laura
Karen
Brian
Sample Statistics
Average Age = 19.8
Average IQ = 104.6
60% Female, 40% Male
Sample #2
Tom
Kristen
Sara
Andrew
John
Sample Statistics
Average Age = 20.4
Average IQ = 114.2
40% Female, 60% Male

Sampling distribution of the sample mean

We take many random samples of a given size n from a population with **mean μ and standard deviation σ.**

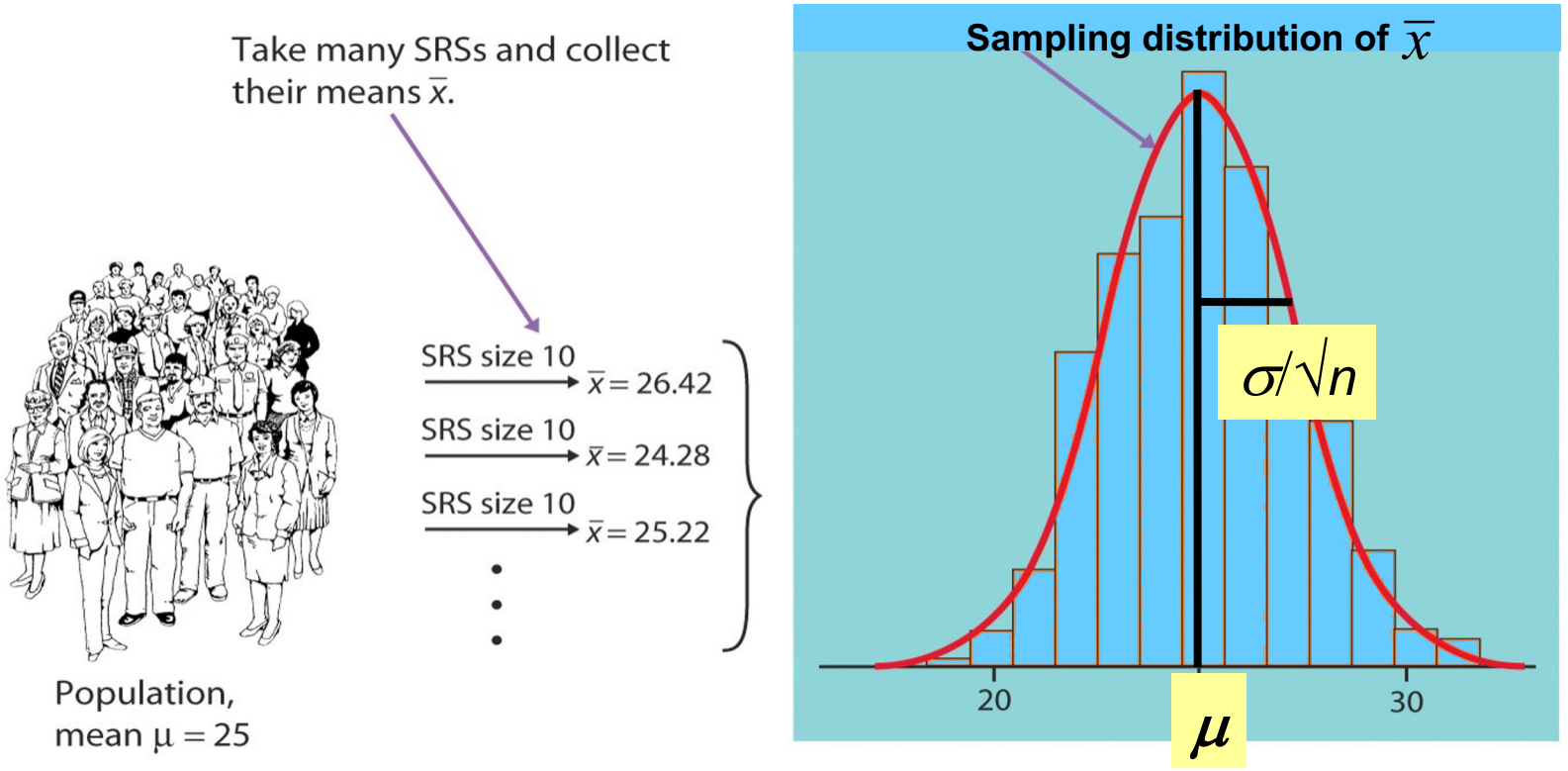
Some sample means will be above the population mean μ and some will be below, making up the sampling distribution.



For any population with a given mean and a given standard deviation :

- The **mean,** or center of the sampling distribution of $\overline{x}$, is equal to the population mean.

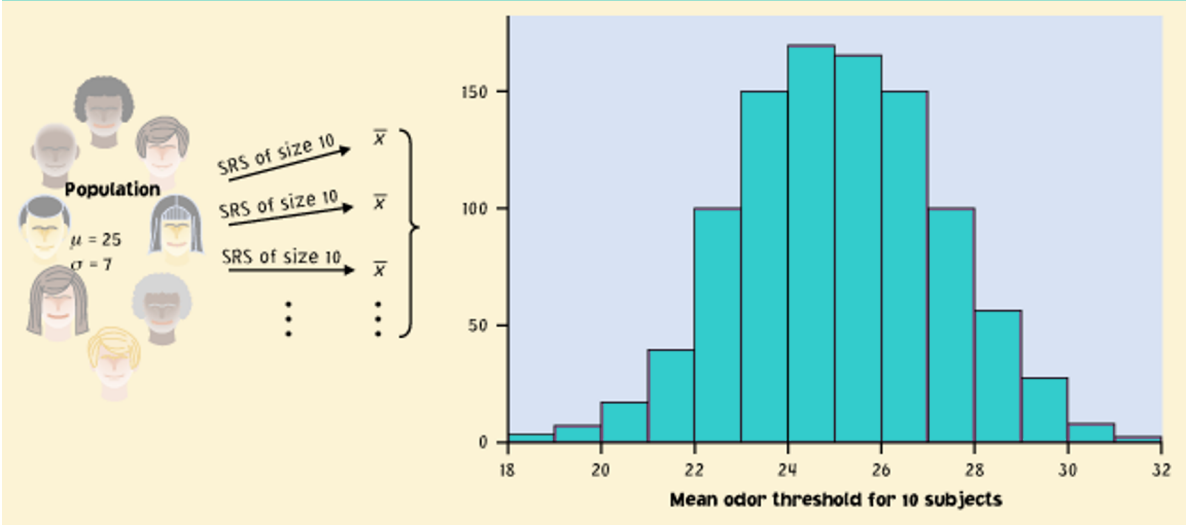
- The **standard deviation** of the sampling distribution is *the ratio between sampling variance and the square root of n,* where *n* is the sample size, it is a measure of how much error there is in the sampling process.



For normally distributed populations

When a variable in a population is normally distributed, then the sampling distribution of $\overline{x}$ for all possible samples of size n is also normally distributed.

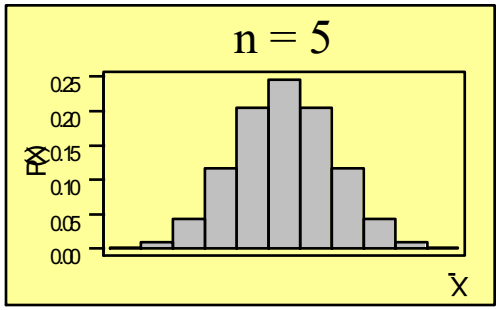
If the population is $N(\mu,\sigma)$, then the sample means distribution is $N(\mu,\sigma/\sqrt{n})$.

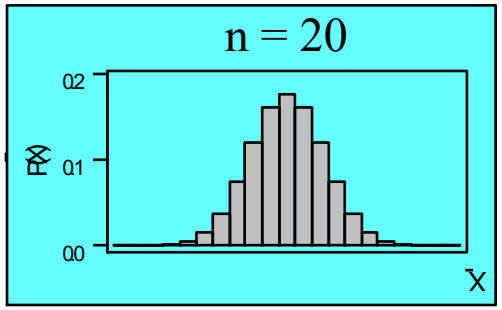


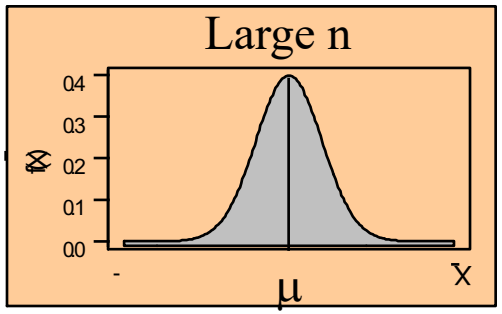
The Central Limit Theorem

When sampling (random sample!) from a population with mean μ and standard deviation σ, the sampling distribution of the sample mean will tend to be a normal distribution with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$ as the sample size becomes large (usually $n > 30$).

For "large enough" n: $\bar{X} \sim N(\mu, \sigma^2/n)$







No matter the shape of the population, the distribution of the sample means tends toward Normality.

Sampling Distribution of Sample Mean

The **center** of the sampling distribution of the sample mean is the population mean μ

- Over all samples, the sample mean will, *on average*, be equal to the population mean (no guarantees for 1 sample!)

The **spread** of the sampling distribution of the sample mean is also called Standard Error $SD(\bar{X}) = \frac{\sigma}{\sqrt{n}}$

- As sample size increases, variance of the sample mean decreases!

Central Limit Theorem: if the sample size is large enough, then the sample mean has an approximately **Normal distribution**

- This is true *no matter what the shape of the distribution of the original data!*

The Central Limit Theorem tells us:

- Even if a population distribution is skewed, we know that the sampling distribution of the mean is normally distributed (for large sample size).

- As the sample size gets larger, the *mean of the sampling distribution becomes equal to the population mean*.

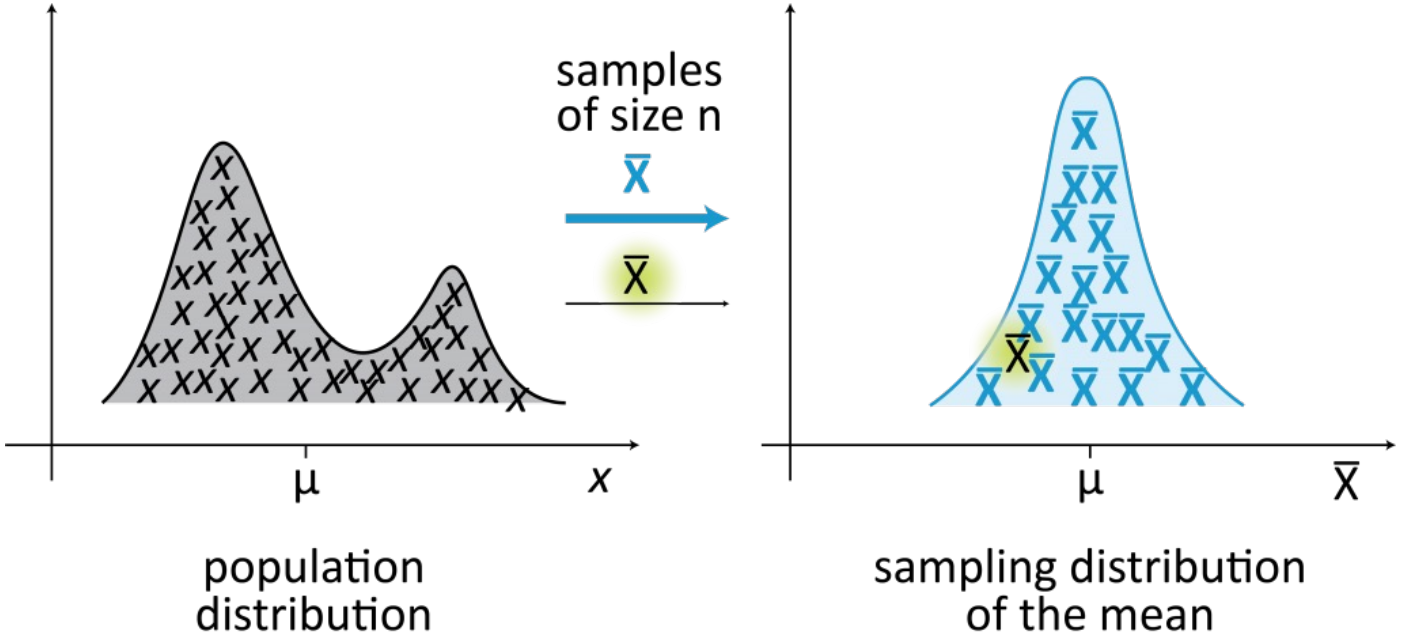
- As the sample size gets larger, the standard error of the mean decreases in size (which means that the variability in the sample estimates from sample to sample decreases as n increases).

- It is important to remember that researchers do not typically conduct repeated samples of the same population. Instead, they use the knowledge of theoretical sampling distributions to construct confidence intervals around estimates.

•The more skewed is the population distribution, the is larger n must be before the shape of the sampling distribution is close to normal.

• *Rule of thumb:* in practice, the sampling distribution is usually close to normal when the sample size n is at least about 30.

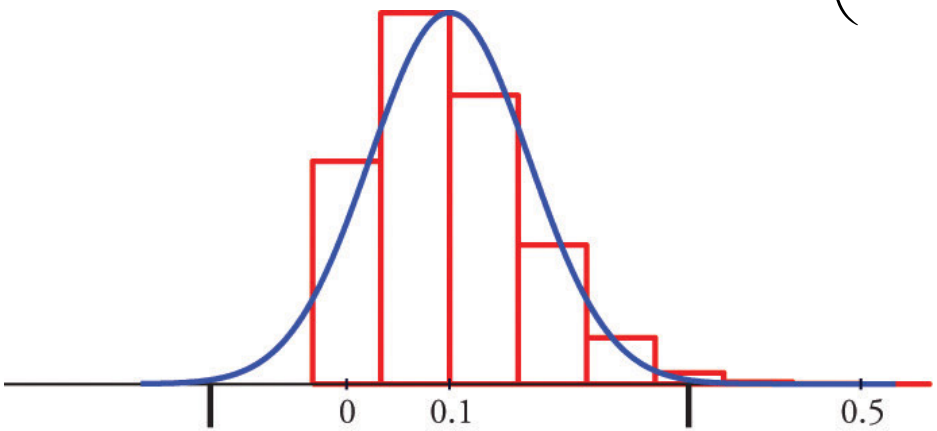
• If the population distribution is approximately normal, then the sampling distribution is approximately normal for all sample sizes.



The sampling distribution model for a sample proportion p

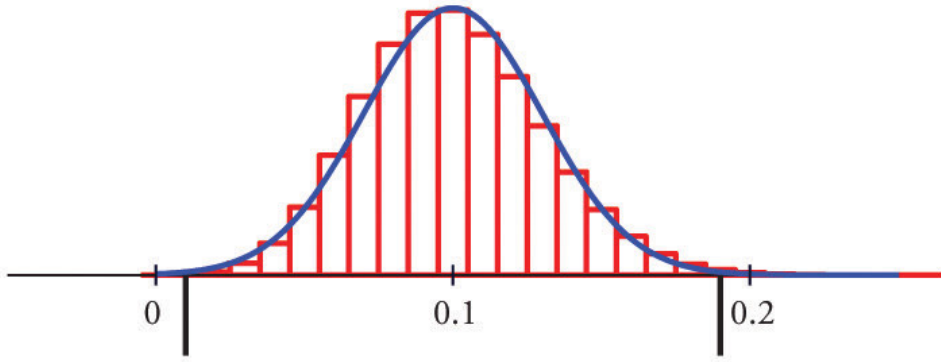
Provided that the sample size n is large enough, the sampling distribution of p is modeled by a **normal** distribution with mean = p and standard deviation that equals $\sqrt{\frac{pq}{n}}$

$$\hat{p} \sim N\left(p, \sqrt{\frac{pq}{n}}\right)$$



$p-3\sqrt{\frac{p(1-p)}{n}} = -0.13$ $p+3\sqrt{\frac{p(1-p)}{n}} = 0.33$

(a) $p = 0.1, \quad n = 15$



$p-3\sqrt{\frac{p(1-p)}{n}} = 0.01$ $p+3\sqrt{\frac{p(1-p)}{n}} = 0.19$

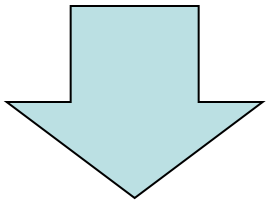
(b) $p = 0.1, \quad n = 100$

4) Desiderable properties of estimators

What is a good estimator

We want good estimates from the sample (as closed as possible to the **unknown** population parameter).

What is a good estimator? What properties should it have?

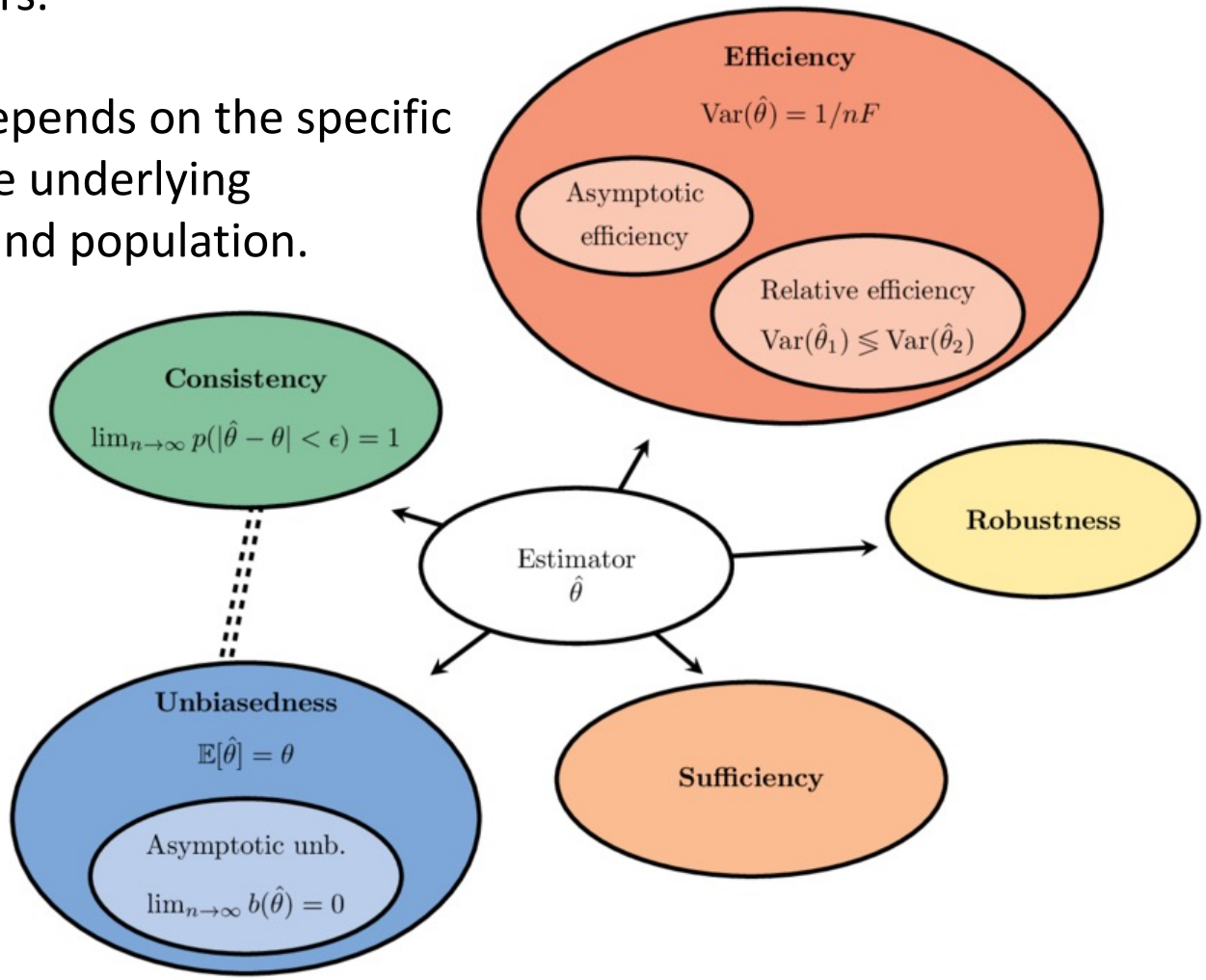


We use the distribution of the statistics (the **sampling distribution**) to verify how well the statistic performs in estimating the population parameter.

Desiderable properties of estimators – fancy overview -

- These properties are used to evaluate and compare different estimators.

- The choice of estimator depends on the specific goals of the analysis and the underlying characteristics of the data and population.



Desiderable properties of estimators (1)

1. **Unbiasedness:** An estimator is considered unbiased if the expected value (mean) of the estimator is equal to the true population parameter it is estimating. Mathematically, an estimator $\widehat{\theta}$ is unbiased for a parameter θ if $E(\widehat{\theta}) = \theta$.

2. **Consistency:** An estimator is consistent if it converges in probability to the true population parameter as the sample size increases. In other words, as the sample size becomes larger, the estimator becomes more and more accurate. Mathematically, for any $\varepsilon > 0$, $\lim (n \to \infty) P(|\widehat{\theta} - \theta| > \varepsilon) = 0$.

3. **Efficiency:** An efficient estimator has the smallest possible variance. In other words, it provides precise and reliable estimates.

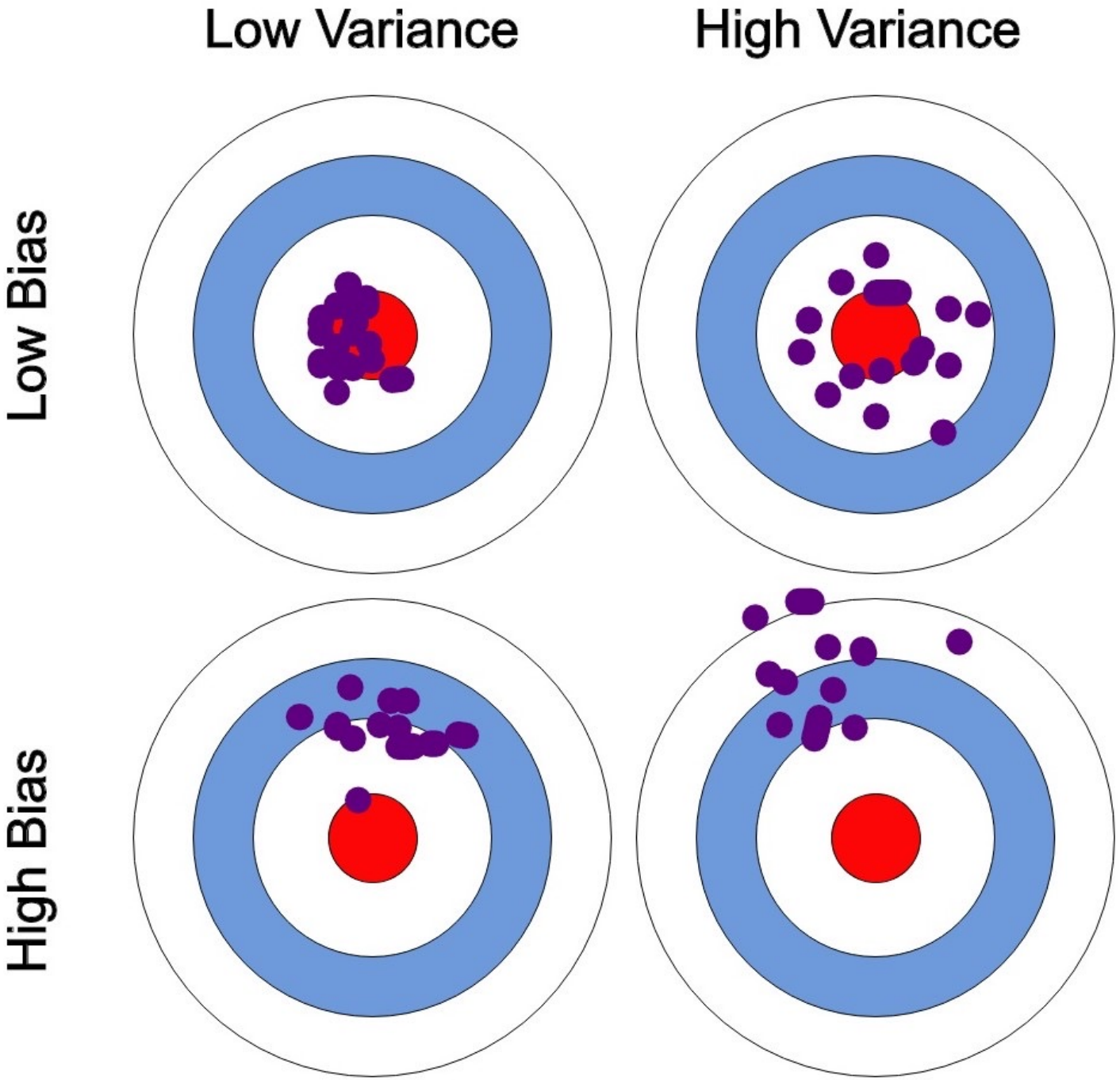
4. **Sufficiency:** A sufficient statistic contains all the information in the sample relevant to the parameter of interest. Estimators based on sufficient statistics are often more efficient than those that are not.

Desiderable properties of estimators (2)

5 **Robustness:** A robust estimator is one that is not highly sensitive to violations of underlying statistical assumptions or to the presence of outliers in the data. Robust estimators can still provide reasonable estimates even in the presence of data deviations from model assumptions.

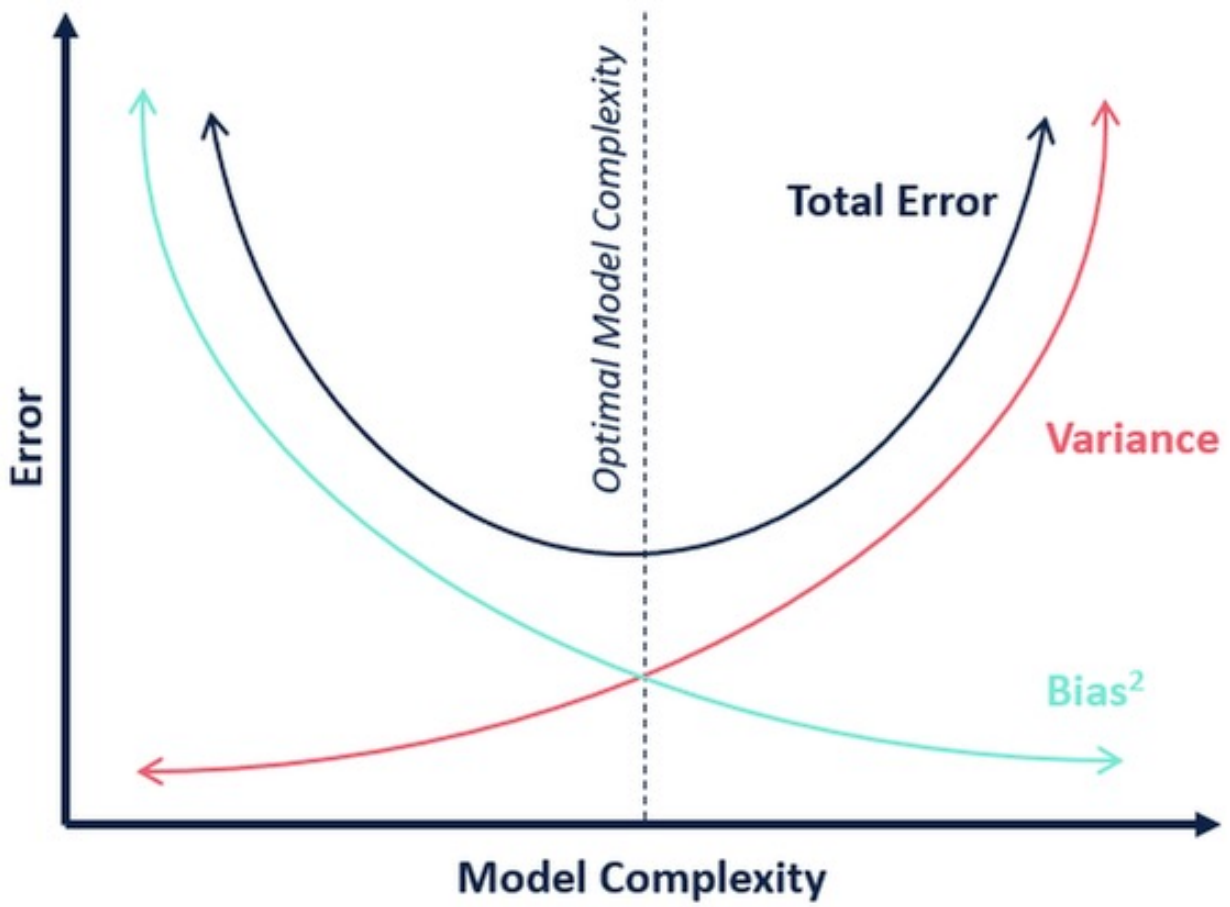
6 **Asymptotic Normality:** Some estimators have the property of asymptotic normality. This means that as the sample size becomes large, the sampling distribution of the estimator approaches a normal distribution, which makes it useful for making inferences using techniques like the Central Limit Theorem.

7 **Interpretability:** While not a mathematical property, interpretability is important. An estimator should be easy to understand and explain to others, especially when conveying the results to non-statisticians or decision-makers.


Low Variance
High Variance
Low Bias
High Bias

The Bias – Variance trade off

Bias-Variance Trade-Off: there is often a trade-off between bias and variance in estimators. Some estimators may have lower bias but higher variance, while others may have higher bias but lower variance. The choice of estimator depends on the specific requirements of the analysis.



MACHINE LEARNING!

5) Different approaches to obtain point estimates

Different approaches to obtain point estimates

A — Minimization of mean square error

B — Maximum Likelihood

C — Method of Moments

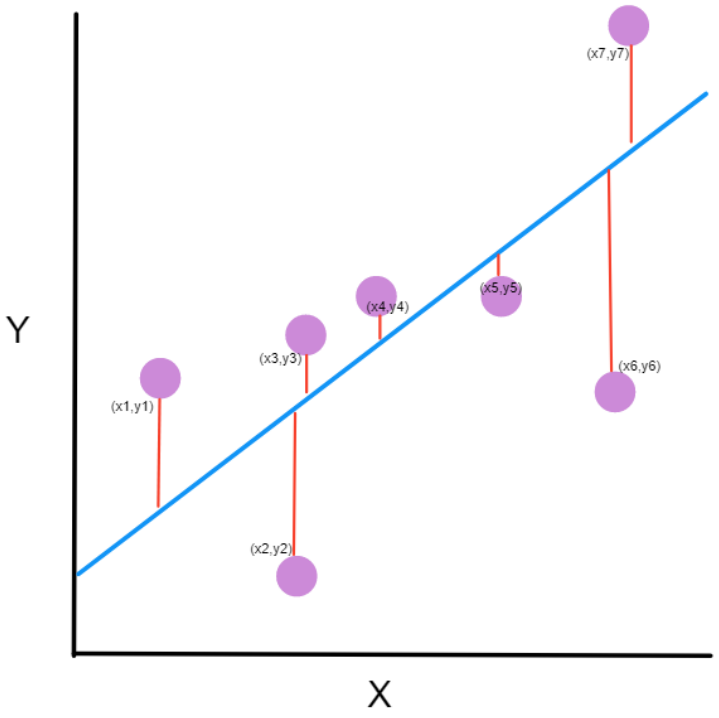
Different approaches to obtain point estimates

A **Minimization of mean square error**

The goal is to minimize the **mean square error (MSE).**

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

Mean — Error — Squared



Issue: in most cases there is not an estimator which minimizes the MSE for all possible values of the parameter

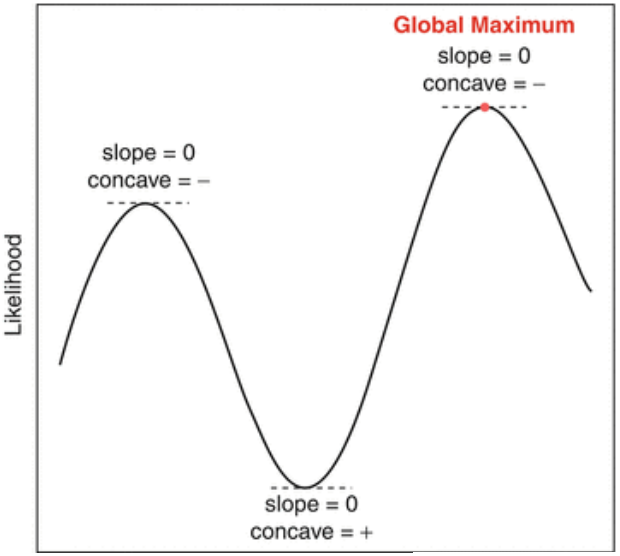
⇩

We choose the best estimator within specific sub-domains of the parameter's domain.

Different approaches to obtain point estimates

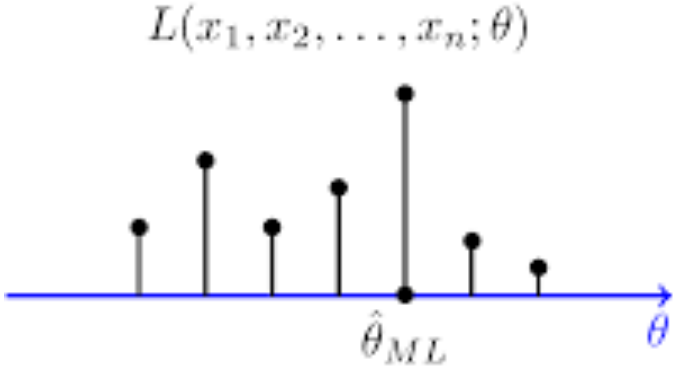
B

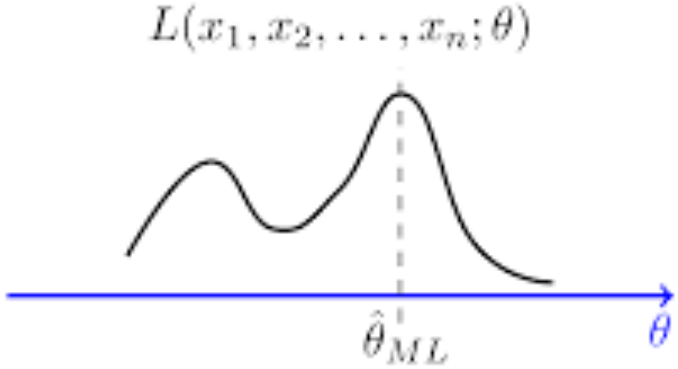
Maximum Likelihood



The goal is to find the value of the parameter which maximizes the Likelihood function

$$\mathrm{L}(\theta) = \prod_{i=1}^{n} f(X_i|\theta)$$





Different approaches to obtain point estimates

C

Method of Moments

If you need to estimate k parameters you use the moments of the function and you solve a system made up by k equations (empirical moments = theoretical moments)

Theoretical moments

$$\begin{aligned}
\mu_1 &\equiv \mathrm{E}[\,X] = g_1(\theta_1, \theta_2, \ldots, \theta_k), \\
\mu_2 &\equiv \mathrm{E}[\,X^2] = g_2(\theta_1, \theta_2, \ldots, \theta_k), \\
&\vdots \\
\mu_k &\equiv \mathrm{E}[\,X^k] = g_k(\theta_1, \theta_2, \ldots, \theta_k).
\end{aligned}$$

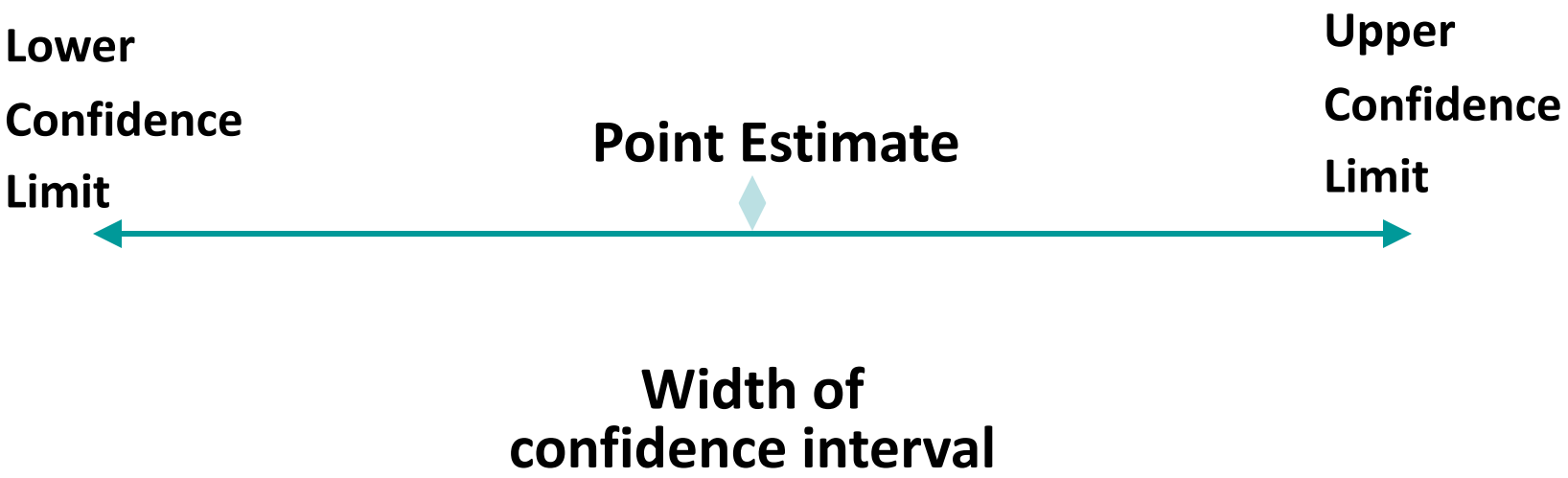
Empirical moments

$$\begin{aligned}
\widehat{\mu}_1 &= g_1(\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_k), \\
\widehat{\mu}_2 &= g_2(\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_k), \\
&\vdots \\
\widehat{\mu}_k &= g_k(\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_k).
\end{aligned}$$

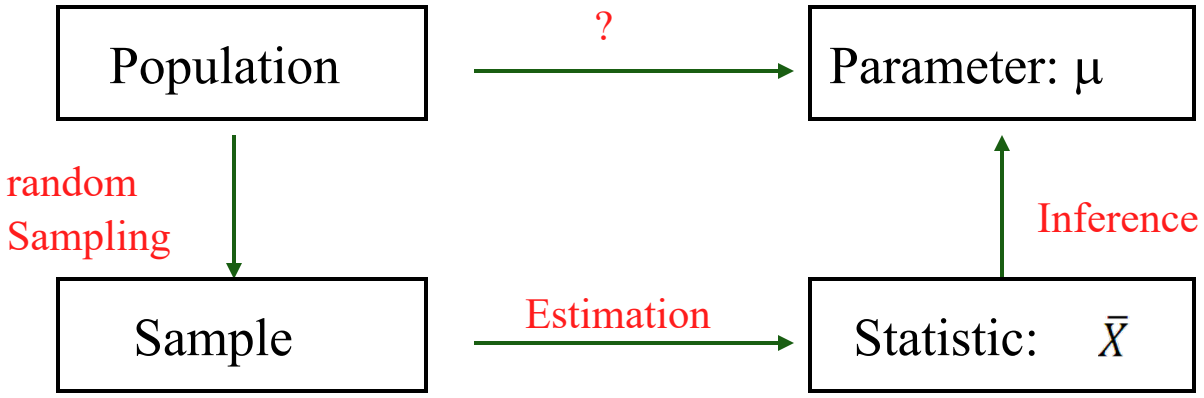
6) Point estimates and Confidence Intervals

Point and Interval Estimates

- A point estimate is a single number,
- a confidence interval provides additional information about the variability of the estimate



Confidence Intervals



Sample mean $\overline{x}$ is the best (point) estimate of μ

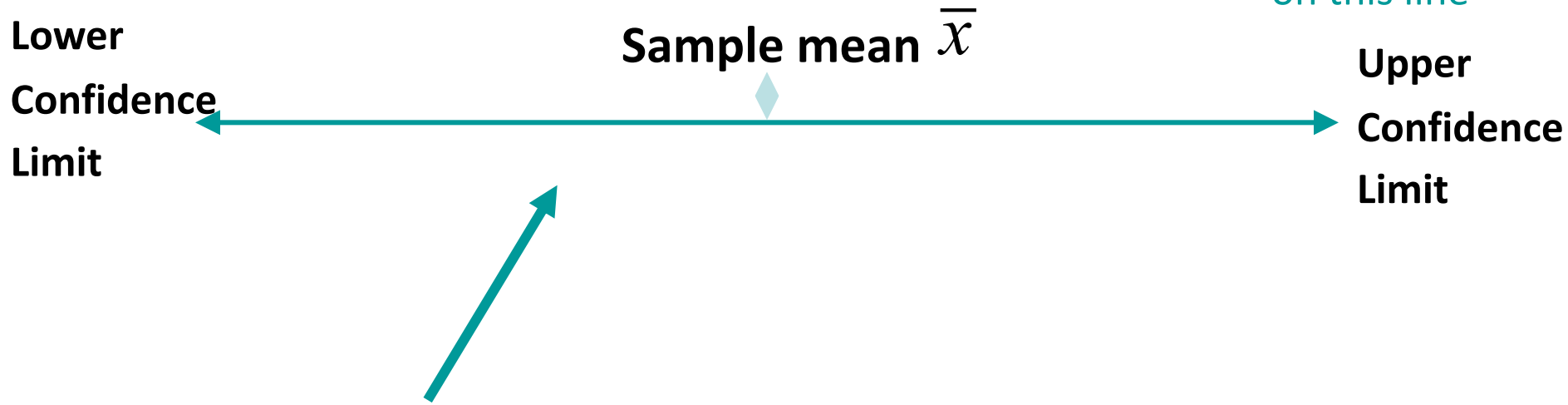
However, we realize that the sample mean is probably not exactly equal to population mean, and that we would get a different value of the sample mean in another sample.

We use the sample mean $\overline{x}$ as the center of an entire **interval of likely values** for our population mean μ

Confidence Interval for the population mean

We want to construct an estimate of where the population mean falls based on our sample statistics

The actual population parameter falls somewhere on this line



This is our Confidence Interval, centered on the sample mean

We know that, thanks to the CLT, the sampling distribution approaches a normal curve in which 95% of all sample means are in the interval

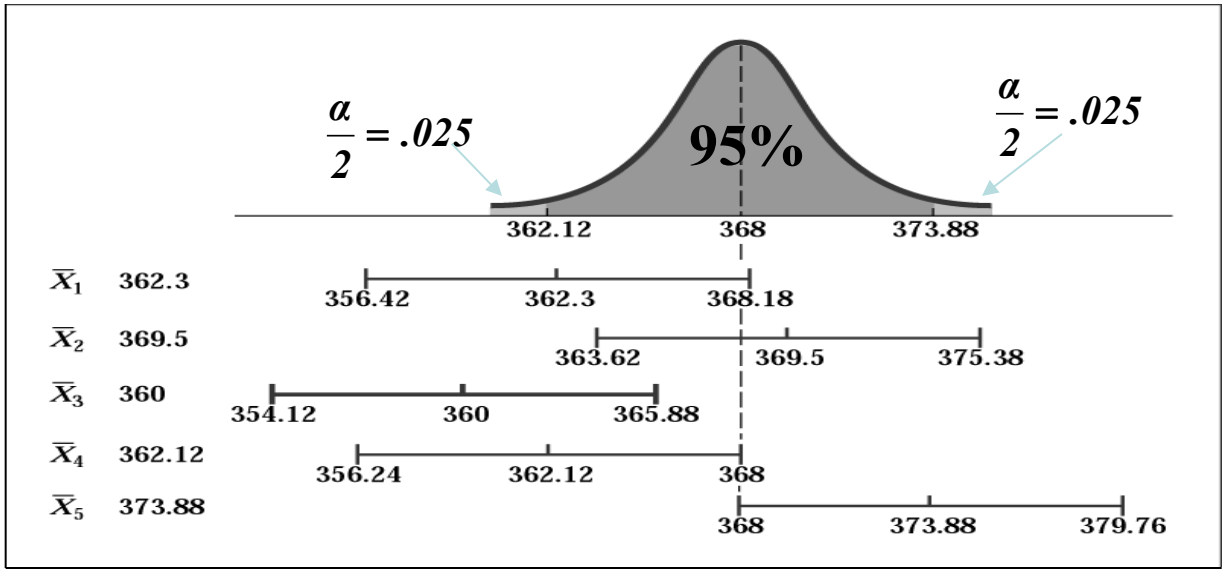
$$\mu - 1.96\frac{\sigma}{\sqrt{n}} < \bar{x} < \mu + 1.96\frac{\sigma}{\sqrt{n}}$$

With a little algebraic manipulation, we can rewrite this inequality and obtain:

$$pr\left[\left(\bar{x} - 1.96\frac{\sigma}{\sqrt{n}}\right) \leq \mu \leq \left(\bar{x} + 1.96\frac{\sigma}{\sqrt{n}}\right)\right] = 0.95$$

Confidence Interval for μ (σ is known)

Let's assume that μ (unknown value of the population parameter) is 368. We want to estimate a confidence interval for μ through the sample mean. Let's hypothesize that we draw from the population n samples and calculate the sample means $\overline{X}_1,\ldots,\overline{X}_n$ and confidence intervals



Not all the samples produce confidence intervals that include μ. However, we know that 95% of the possible samples provide estimates that include μ.

Confidence Interval

An interval gives a range of values:

- Based on observations from 1 sample
- Takes into consideration variation in sample statistics from sample to sample
- Gives information about closeness to the unknown population parameter
- The success rate (proportion of all samples whose intervals contain the parameter) is known as the confidence level
- A 95% confidence interval will contain the true parameter for 95% of all samples
  - Can never be 100% confident

Be careful to the interpretation!

- In practice we select only one sample and - since μ is unknown - we cannot conclude whether our conclusions from the sample about the population are correct or not (whether the confidence intervals includes μ or not).

- In other words, we will never know whether the selected sample is one of the 95% samples that produce correct estimates, we can only say that a priori we have a 95% confidence that the confidence interval produced by the sample includes μ.

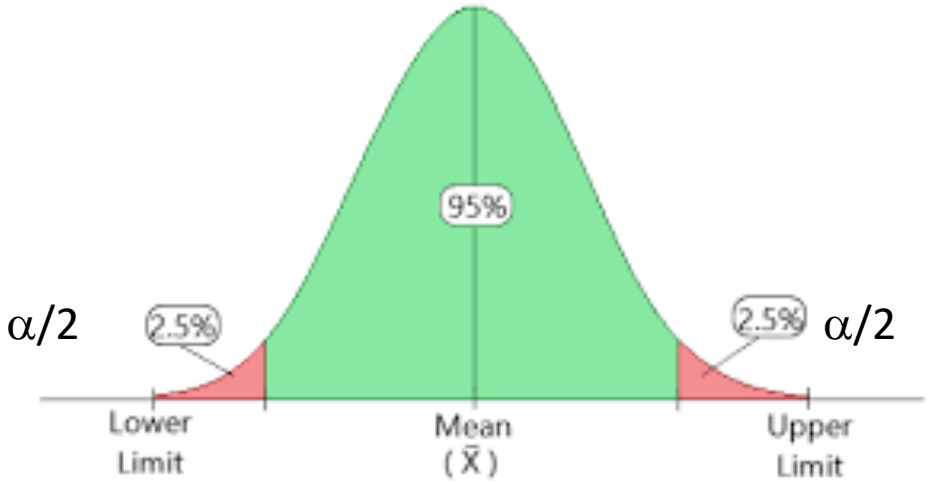
Suppose confidence level = (1 - α)= 95%

α is the proportion of the distribution in the two tails areas outside the confidence

α interval.

- A relative frequency interpretation:
  - If all possible samples of size n are taken and their means and intervals are estimated, 95% of all the intervals will include the **true value of that the unknown parameter**

- A specific interval either will contain or will not contain the true parameter (due to the 5% risk)



To sum up

The value of the statistic in my sample (eg., mean, odds ratio, etc.)

point estimate ± (measure of how confident we want to be) × (standard error)

From a Z table or a T table, depending on the sampling distribution of the statistic.

Standard error of the statistic.

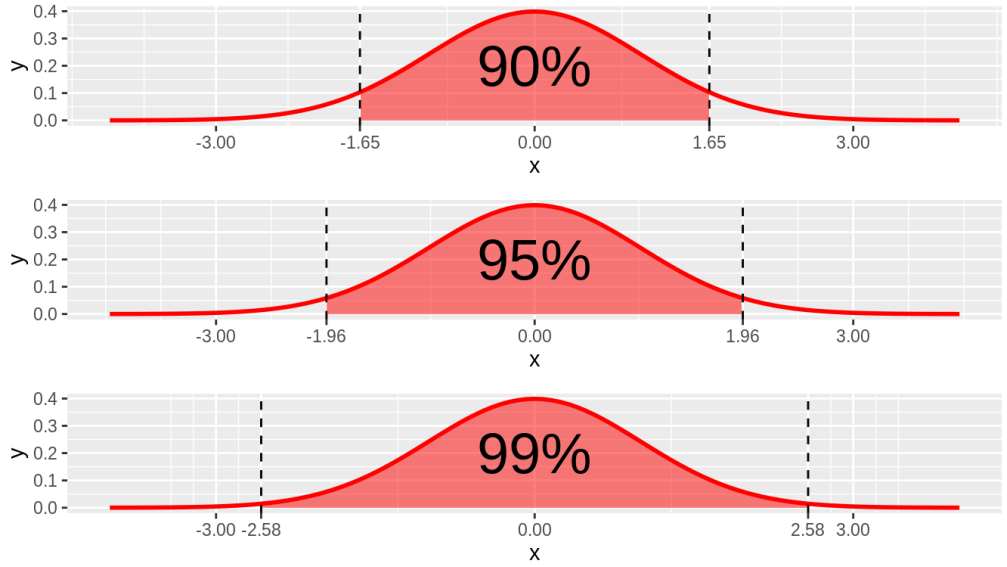
Formula for the Confidence Interval of the Mean for a Specific *α*

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

For a 90% confidence interval: $z_{\alpha/2} = 1.65$

For a 95% confidence interval: $z_{\alpha/2} = 1.96$

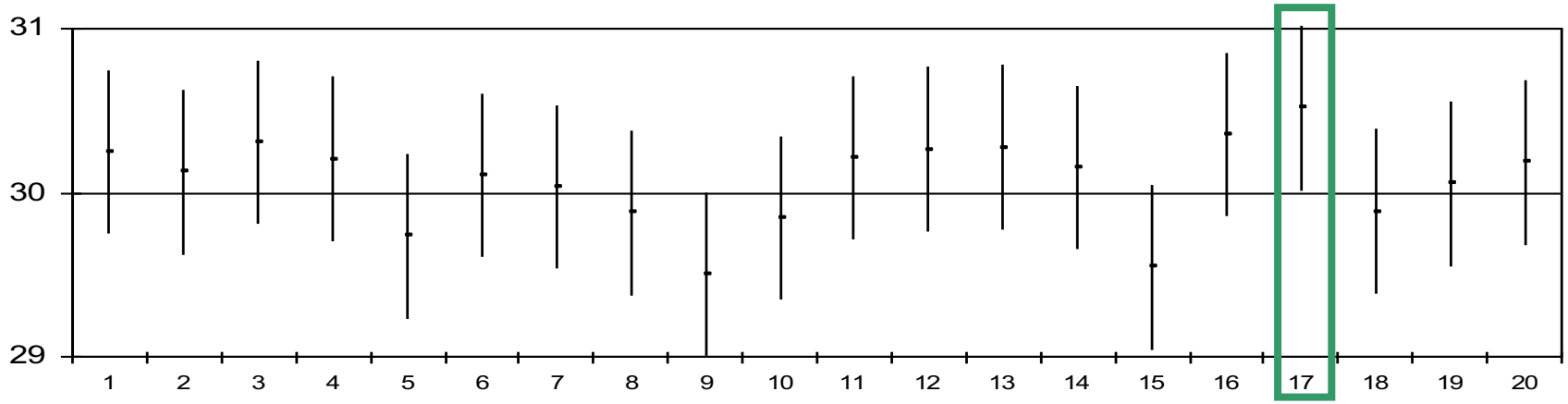
For a 99% confidence interval: $z_{\alpha/2} = 2.58$



Example

The researcher knows that the expenditure for a consumption of a product X is normally distributed in the population with μ =30 euro and σ=2,55. As countercheck the researcher asks to 20 different companies to make a survey based on a random sample of 100 individuals to estimate X only knowing that X is normally distributed with σ = 2,55 and that the confidence level is 95% (1- α=0,95). Each company will slelect a different sample therefore the sample mean will vary from sample to sample.

The first company will observe a mean=30,24, with an inetrval: 30,24 $\pm$ 1,96(2,55/10)=[29,74; 30,74]. The second will observe a mean=30,12, [29,62; 30,62].....Given that the probability that the confidence interval includes μ **is fixed and equal to 95%, ther reseracher will expect 19 out of 20 confidence intervals conatining** μ



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.

- Assumptions
  - Population standard deviation is unknown
  - Population is normally distributed
  - If population is not normal,
- Use Student's t Distribution
- Confidence Interval Estimate: $\overline{X} \pm t_{\alpha/2} \frac{S}{\sqrt{n}}$

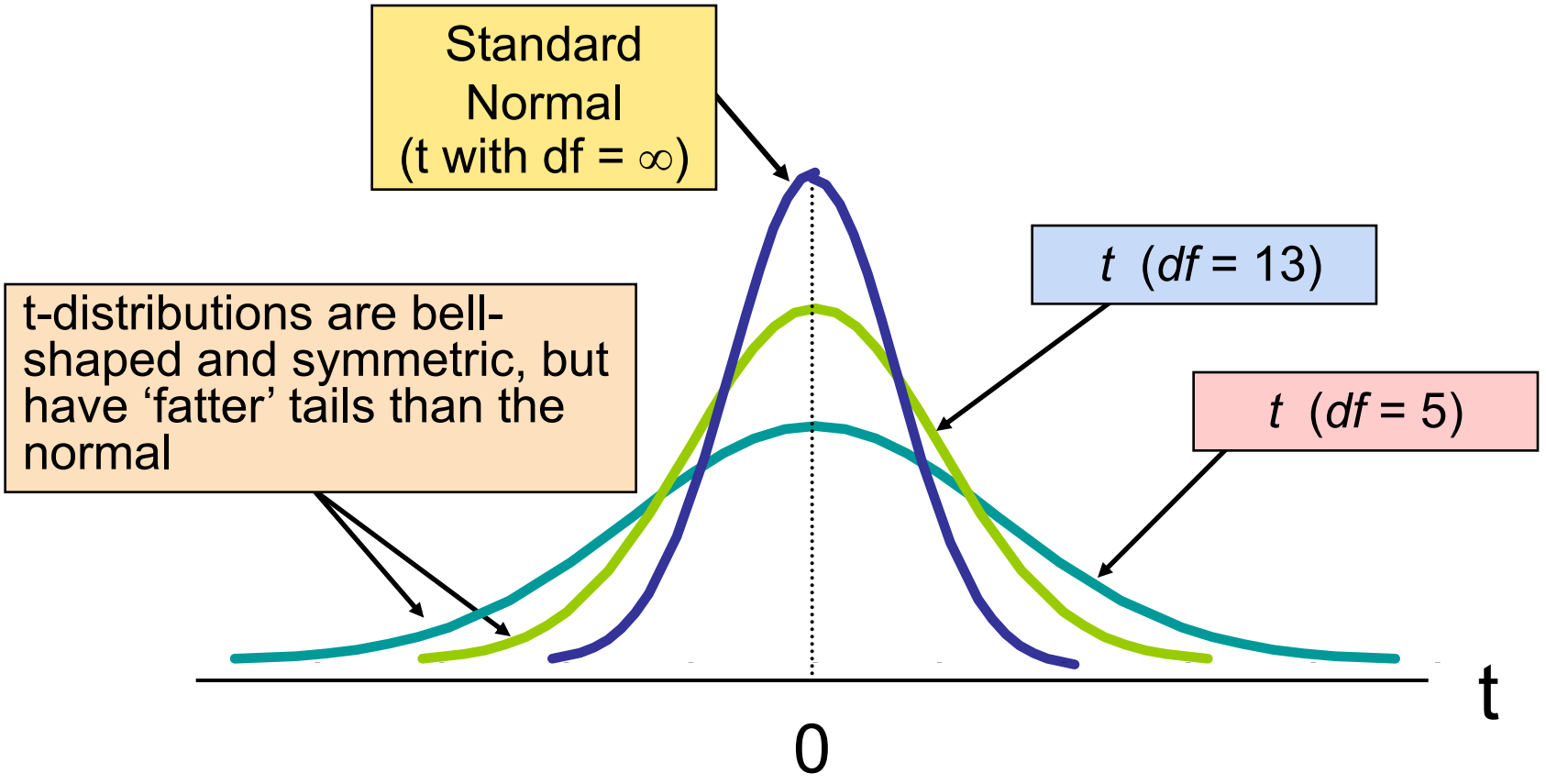
(where $t_{\alpha/2}$ is the critical value of the t distribution with n -1 degrees of freedom and an area of **α**/2 in each tail)

Properties of Student's t distribution

- Similar to Standard normal distribution
  - Symmetric
  - unimodal
  - Centred at zero
- has slightly fatter tails to reflect the uncertainty added by estimating σ with *s.*
- As the sample size increases (degrees of freedom increases) the t distribution approaches the standard normal distribution

Student's t Distribution

Note: t ⟶ Z as n increases



Clarification

- If the underlying data are not normally distributed AND n is small**, the means do not follow a t-distribution (so using a ttest will result in erroneous inferences).
- Data transformation or non-parametric tests should be used instead.
- **How small is too small? No hard and fast rule—depends on the true shape of the underlying distribution.

7) Width of a confidence interval

Confidence Interval Width

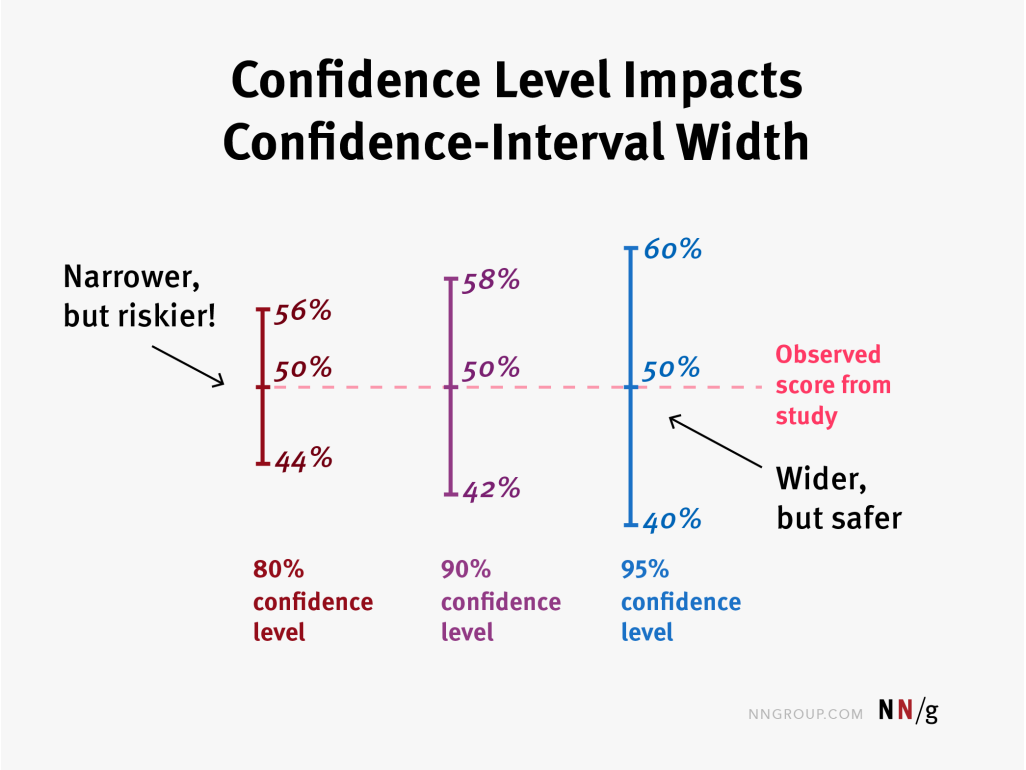
The three elements that determine the width of a confidence interval

- The variance (standard error) of the phenomenon
- The confidence level
- The sample size n

TRADE - OFF

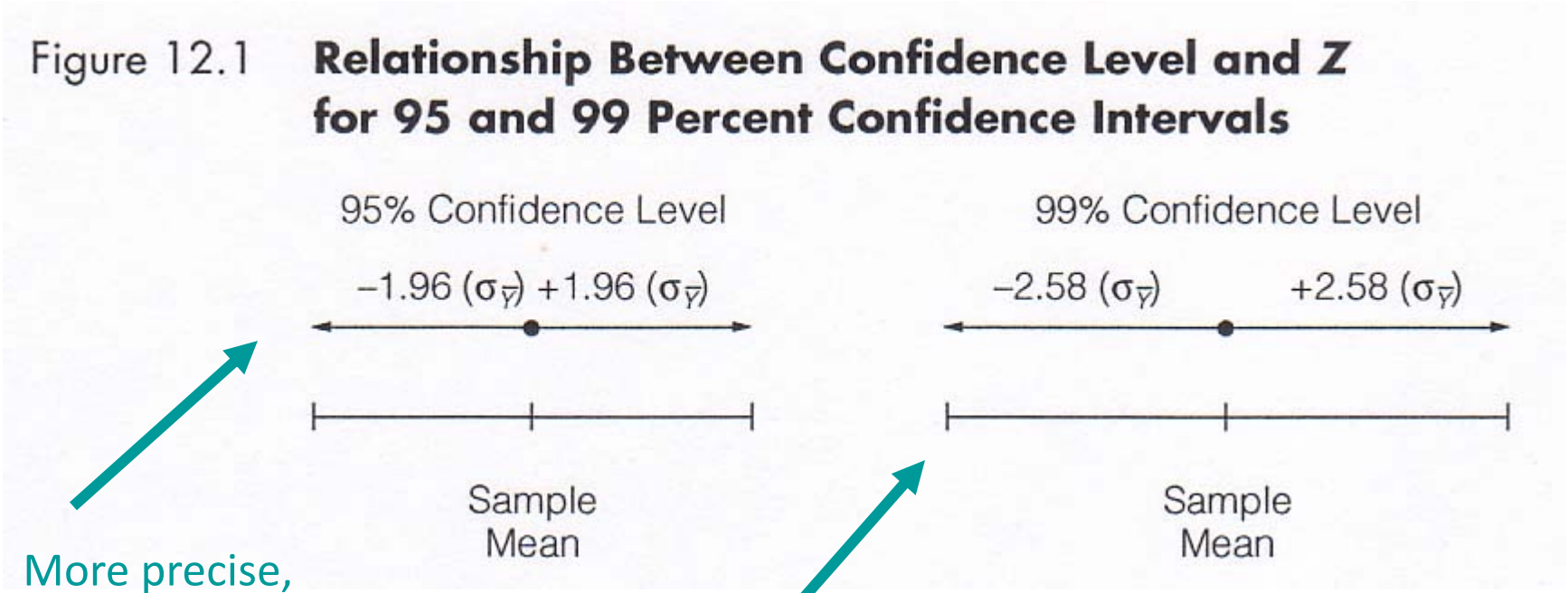
Reliability

Precision



Confidence Interval Width

Figure 12.1 **Relationship Between Confidence Level and *Z* for 95 and 99 Percent Confidence Intervals**



More precise, less confident

More confident, less precise

Confidence Interval Width

$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

- **Sample Size –** Larger samples result in smaller standard errors, and therefore, in sampling distributions that are more clustered around the population mean. A more closely clustered sampling distribution indicates that our confidence intervals will be narrower and more precise.

Confidence Interval Width

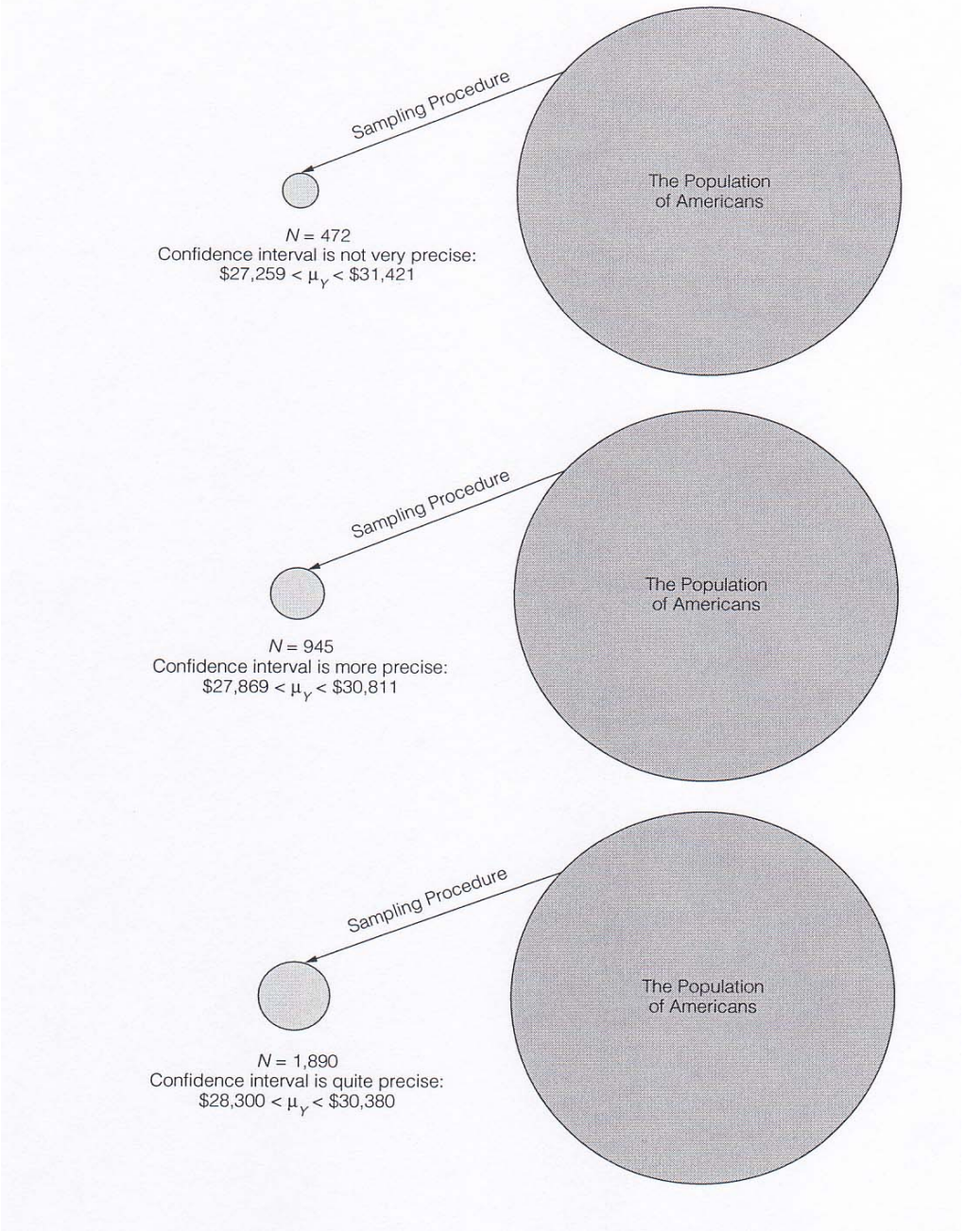
$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

Standard Deviation – Smaller sample standard deviations result in smaller, more precise confidence intervals.

(*Unlike sample size and confidence level, the researcher plays no role in determining the standard deviation of a sample.)*

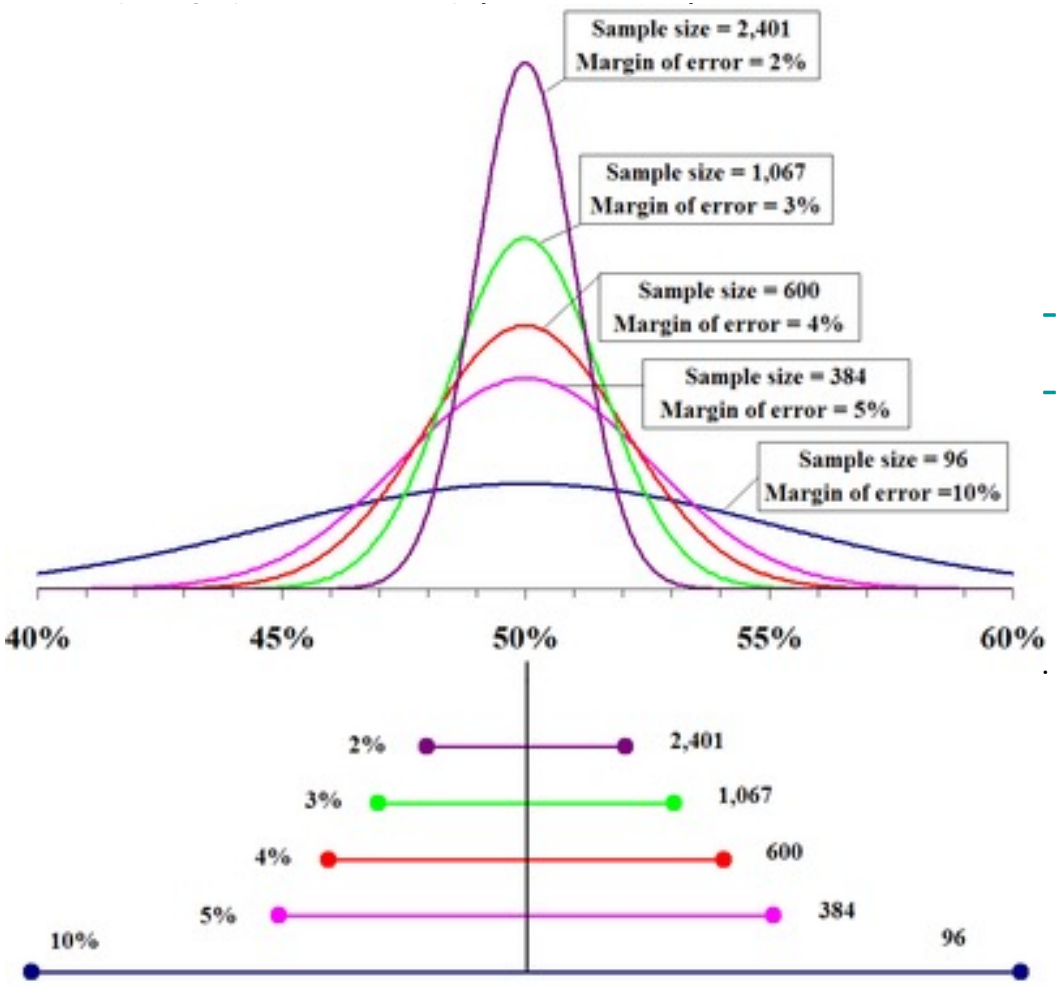
Example: Sample Size and Confidence Intervals

Figure 12.5 **The Relationship Between Sample Size and Confidence Interval Width**



How to determine the sample size n?

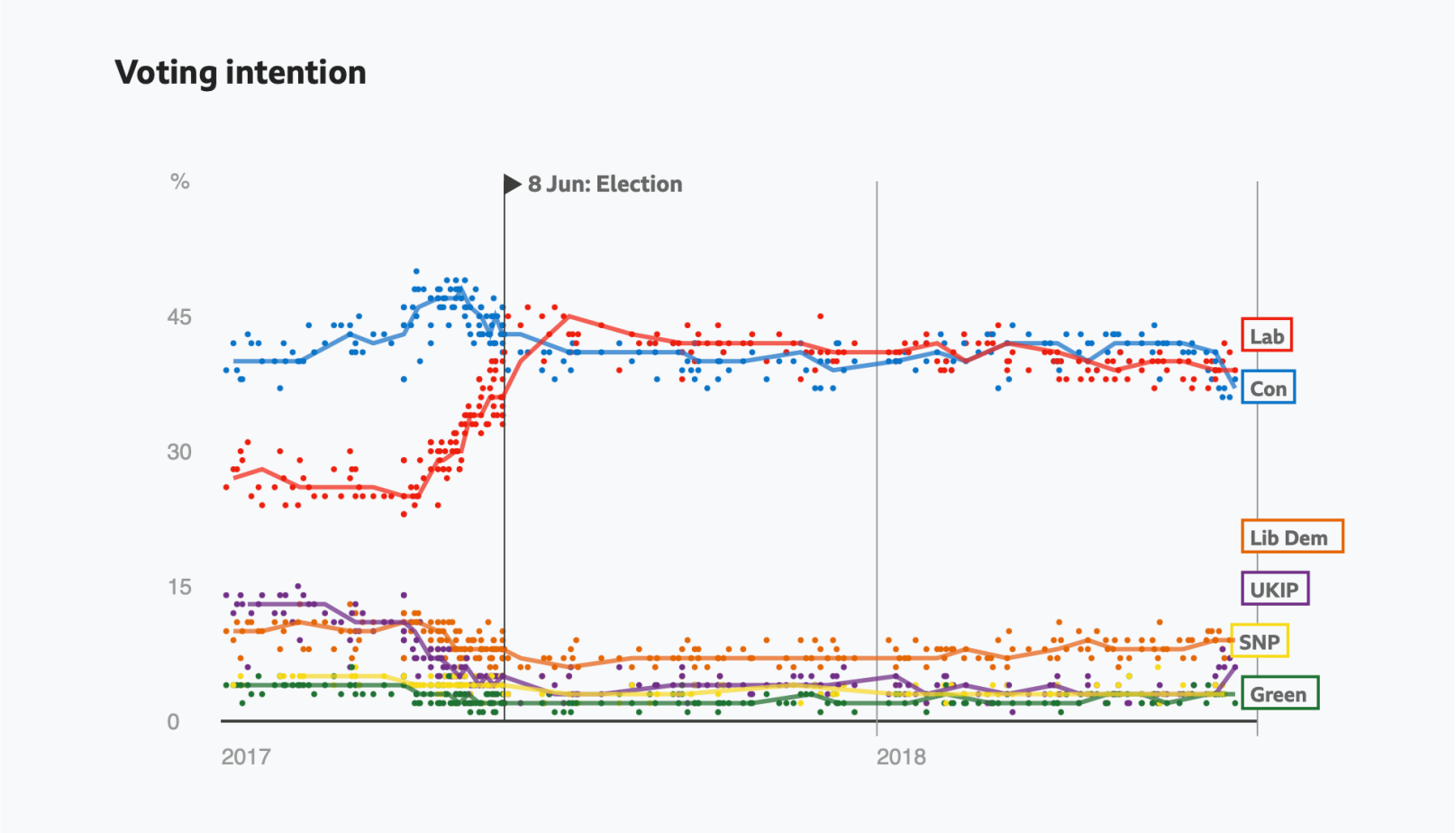
What if the sample size is not predetermined – If the researcher can a priori choose the sampling size n, then she/he will choose the number n that guarantees a given confidence level (reliability), but also determines a given



$$n = (2z\sigma/A)\text{^}2$$

- A is the width of the interval
- what if the standard error is unknown?

CI and uncertainity
Poll trends: UK Election 2017



The **trend lines** are calculated as a rolling median of the seven latest polls. Polling companies generally claim that 95% of the time, a poll of 1,000 people will be accurate within a margin of error of +/-3%.

https://www.bbc.com/news/election-2017-39856354

“….Of course, this is only accurate if the polling company really did take a random sample, and everyone replied, and they all had an opinion and they all told the truth. So, although we can calculate margins of error, we must remember that they only hold if our assumptions are roughly correct. But can we rely on these assumptions?”

it is reasonable to take a precautionary approach and communicate uncertainty which also depends on non sampling errors (systematic bias which wont decrease if we increase the sample size). The CI does not contain systematic bias (we need experts to evaluate non sampling errors).

What happens when we have all the existing data available?

The UK government, each year, reports the number of homicides (if we exclude errors in counting) these are just descriptive indexes.

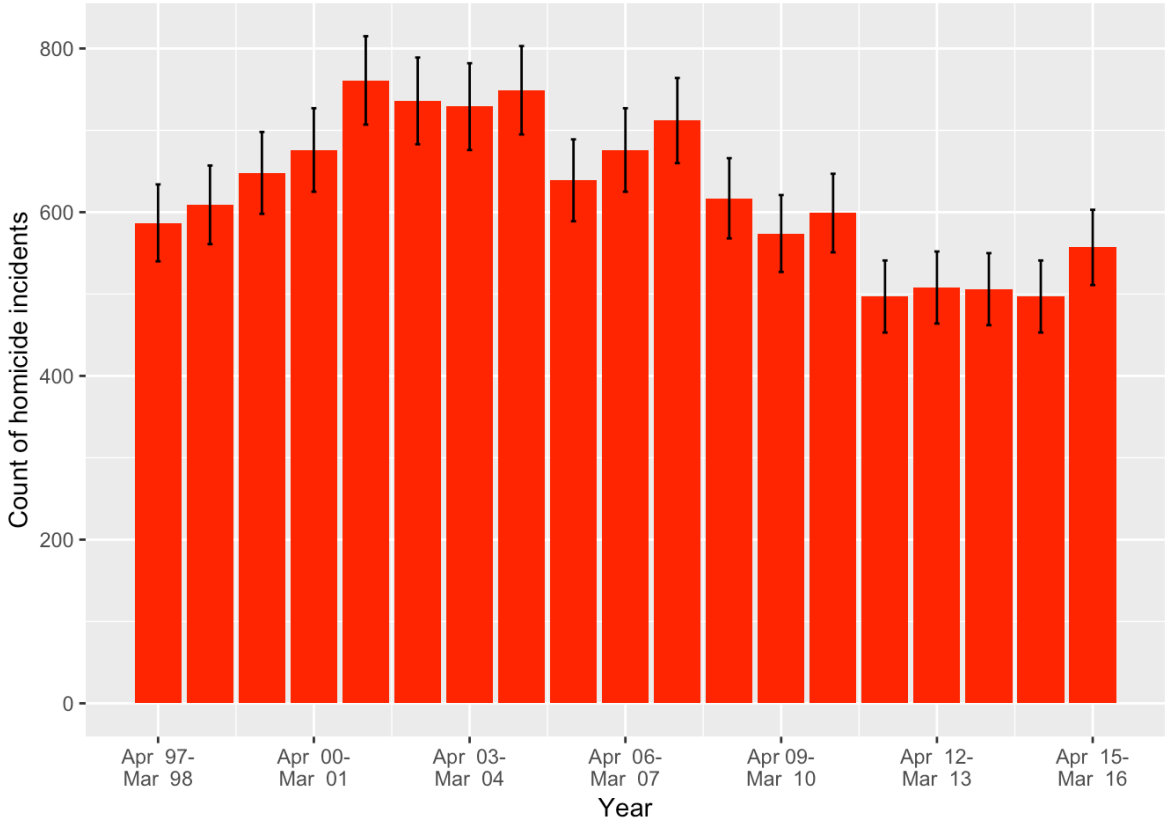
Now imagine you want to study the time trend. We know that the homicides increased between 2014 and 2015 (497 vs 517). Is this a true increase?

We need to use a probabilistic model, we assume that the daily number of homicides is a random observation distributed as a Poisson from a metaphoric population.
This means that the total yearly number of homicides can be considered as one observation from a Poisson distribution with mean μ (the ”real” yearly value). We want to understand whether μ varies from year to year.

497 is the estimates of μ, therefore the CI is (453;541)

Can we conclude that the homicides increased and that this variation is not due to random fluctuations?

Number of homicides in England and Wales between 1998 and 2016, and 95% CIs for the underlying 'true' homicide rate:



Even when we observe all the data we can still calculate CI which represents uncertainty over the parameters of a metaphoric population!