Applied Statistics

Feb 9 2021

Prof.ssa Chiara Seghieri,

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

In questa lezione..

In questa lezione faremo un ripasso veloce dei seguenti concetti base della teoria della stima:

- Distribuzioni campionarie
- Stima puntuale
- Stima per intervalli di confidenza
- Verifica di ipotesi statistiche

Random variables

A Random Variable X is a function that maps outcomes of a random process to real values.

Examples: tossing a coin, you want to know how many sixes you get if you roll the die a certain number of times. Your random variable, X could be equal to 1 if you get a six and 0 if you get any other number.

Or Y=sum of upward faces after rolling the dice 5 times

- Survey to people about their voting preferences, the percentage of the sample that responds “Yes” is also a random variable (the percentage will be slightly differently every time you poll).

Random variables are most often used in conjunction with a probability of a random event happening. Say you wanted to see the probability of obtaining less than 30 after rolling the dice 5 times. You could write it as: $P(Y<30)$.

Random Variables: continuous & discrete

Random variables can be **discrete** or **continuous**

Discrete random variables have a countable (or finite) number of outcomes.

Examples: Dead/alive, satisfied/not satisfied, etc.

Continuous random variables have an infinite continuum of possible values.

Examples: blood pressure, weight, the speed of a car, QI.

Random Variable and Probability Distribution

The **probability distribution** of a random variable is the collection of possible outcomes along with their probabilities:

- Discrete case: $\Pr(X = x) = p_\theta(x)$
- Continuous case: $\Pr(a \leq X \leq b) = \int_a^b p_\theta(x)dx$

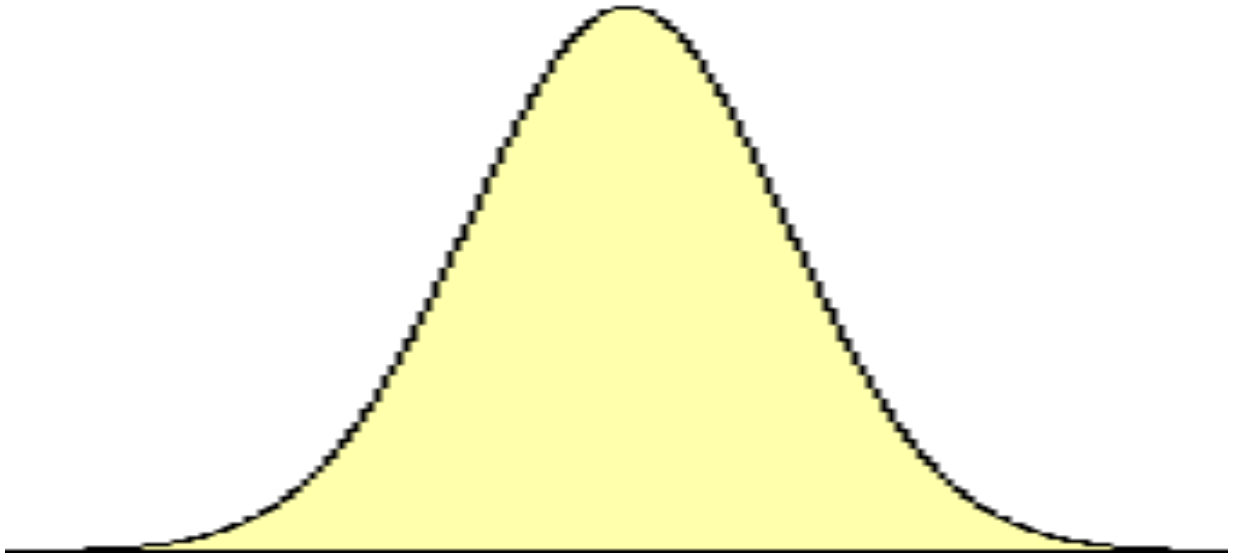
Discrete Random Variables

- Discrete random variables can be summarized by listing all values along with the probabilities
  - Called a **probability distribution**
- Example: number of members in US families

X	2	3	4	5	6	7
P(X)	0.413	0.236	0.211	0.090	0.032	0.018

Continuous Random Variables

- Continuous random variables have a **non-countable** number of values
- Can't list the entire probability distribution, so we use a **density curve** instead of a histogram
- Eg. Normal density curve:

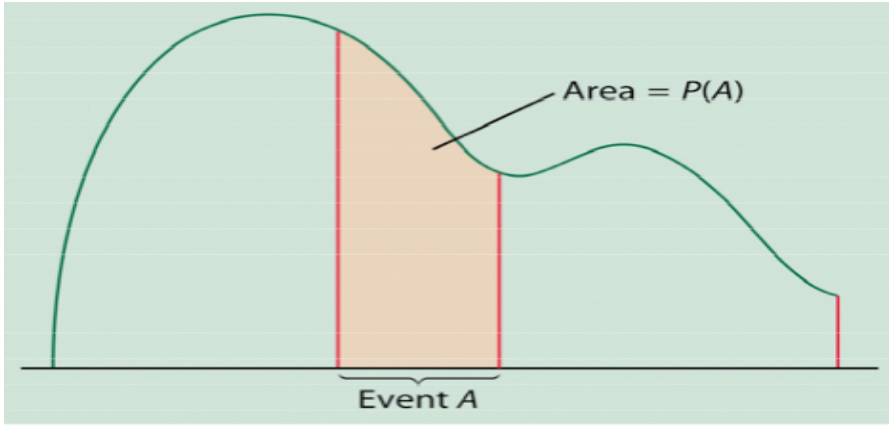


Continuous case

- The probability function that accompanies a continuous random variable is a continuous mathematical function that integrates to 1.
- Probabilities are given for a range of values, rather than a particular value (e.g., the probability of getting a math score between 29 and 30 is 2%).
- The probabilities associated with continuous functions are just areas under the curve (integrals!).

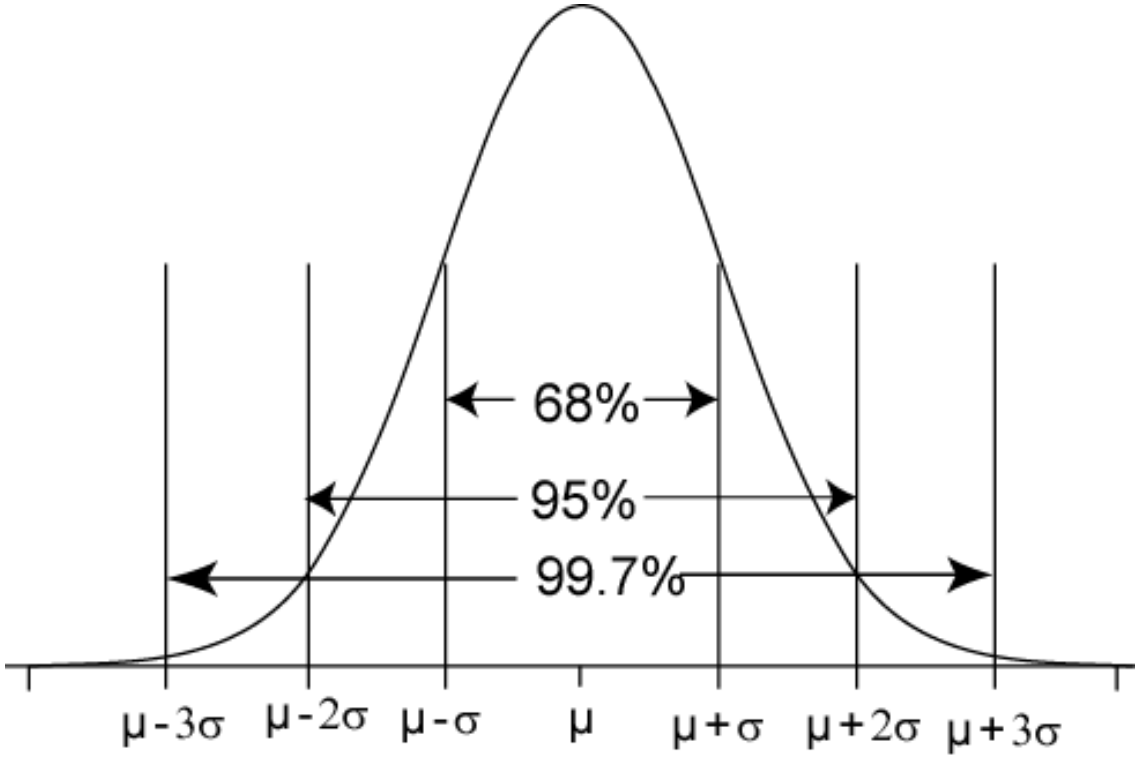
Calculating Continuous Probabilities

- Continuous case: we have to use **integration** to calculate the area under the density curve:



68-95-99.7 Rule for Normal Distributions

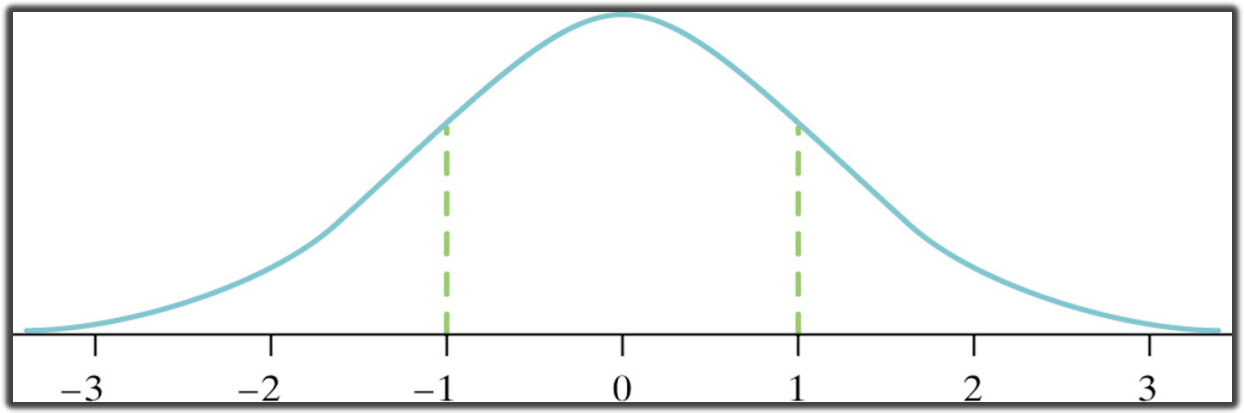
- **68%** of the AUC within ±1**σ** of μ
- **95%** of the AUC within ±2**σ** of μ
- **99.7%** of the AUC within ±3**σ** of μ



The Standard Normal Distribution

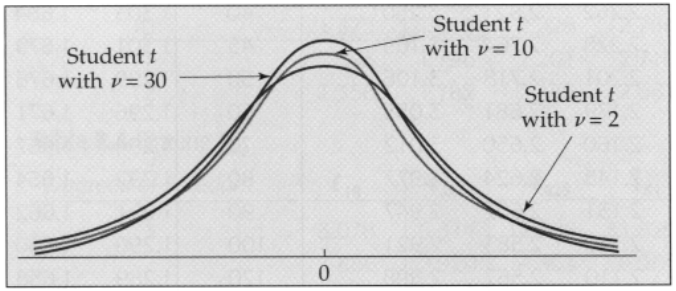
- The **standard Normal distribution** is the Normal distribution with mean 0 and standard deviation 1.
- Shown as ***N***(0,1)
- If a variable x has any Normal distribution $N(\mu, \sigma)$, with mean μ and standard deviation σ, then the standardized variable

$$z = \frac{x - \mu}{\sigma}$$

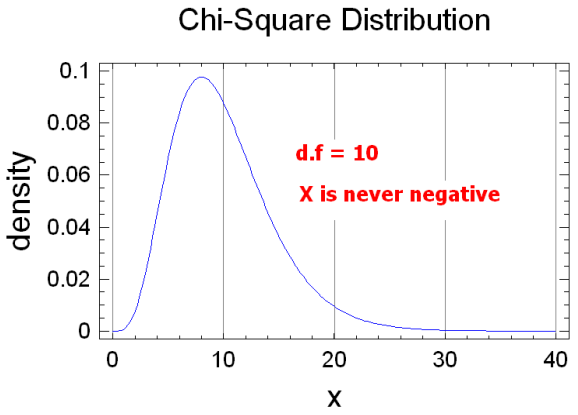


Three other important continuous distributions which are extensively used:

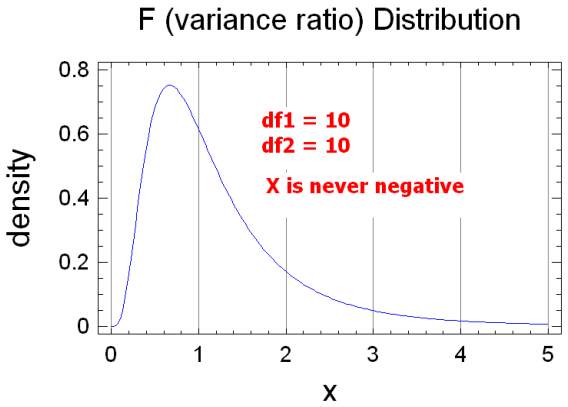
Student *t* Distribution



Chi-Squared Distribution



F Distribution



Sampling Distribution: Introduction

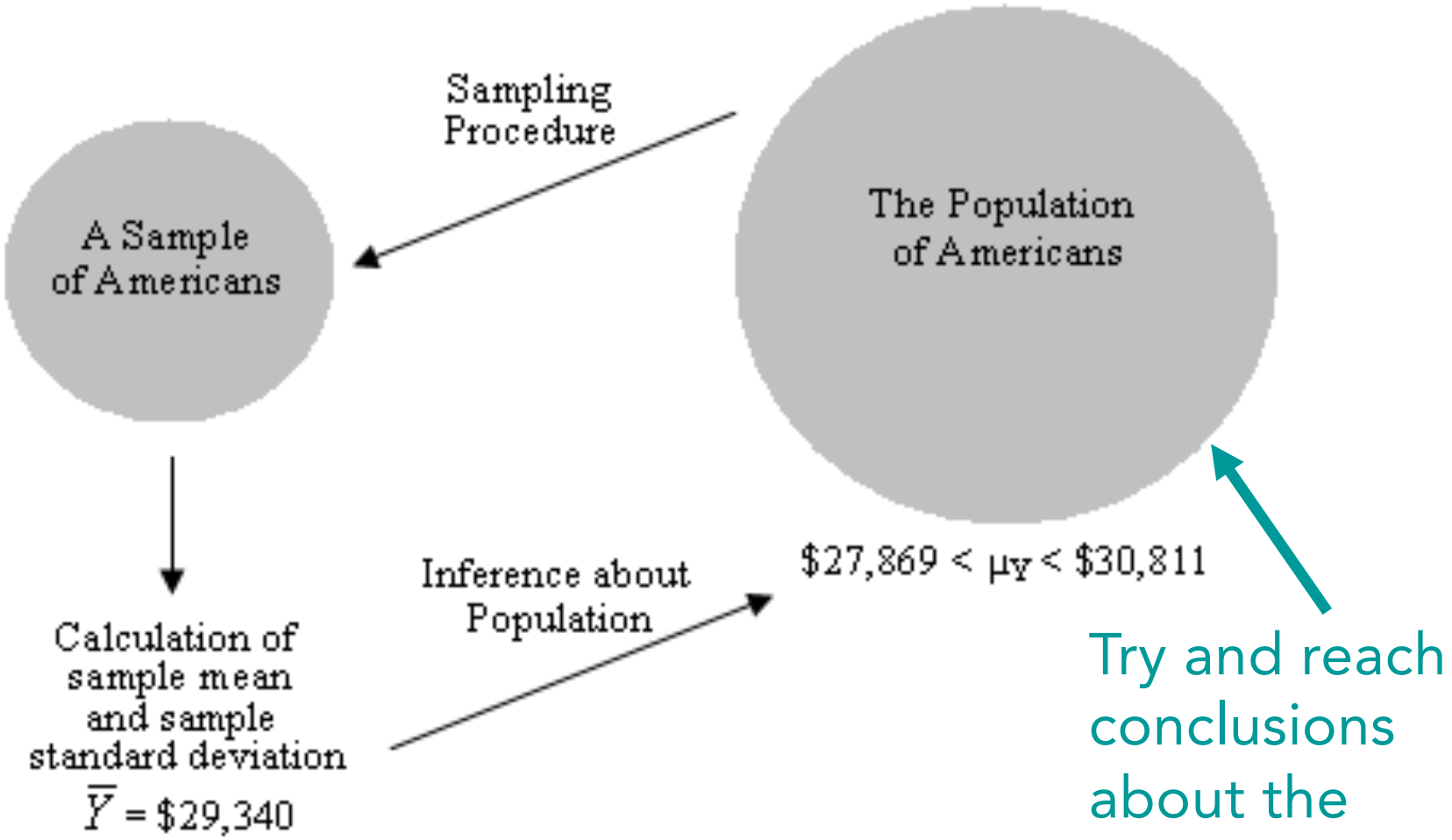
- In real life calculating parameters of populations is prohibitive because populations are very large.
- Rather than investigating the whole population, we take a sample, calculate a **statistic** (function of the sampled observations) related to the **parameter** of interest obtain the estimate (value that thatches the statistic in the sample) and make an inference.
- The **sampling distribution** of the **statistic** is the tool that tells us how close is the statistic to the parameter.

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. Called a ***confidence interval***.

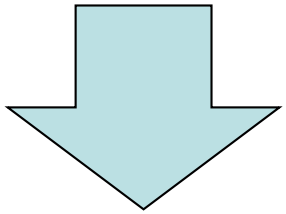

Sampling
Procedure
A Sample
of Americans
The Population
of Americans
Calculation of
sample mean
and sample
standard deviation
$\overline{Y}$ = $29,340
Inference about
Population
$27,869 < μY < $30,811
Try and reach
conclusions
about the
population

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, s, 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.

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 close is the statistic to the population parameter.

The sampling distribution of a statistic

Three important points :

- the numerical value of a statistic cannot be expected to give us the exact value of the population parameter; $\hat{\theta} = \theta + \varepsilon$
- 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.

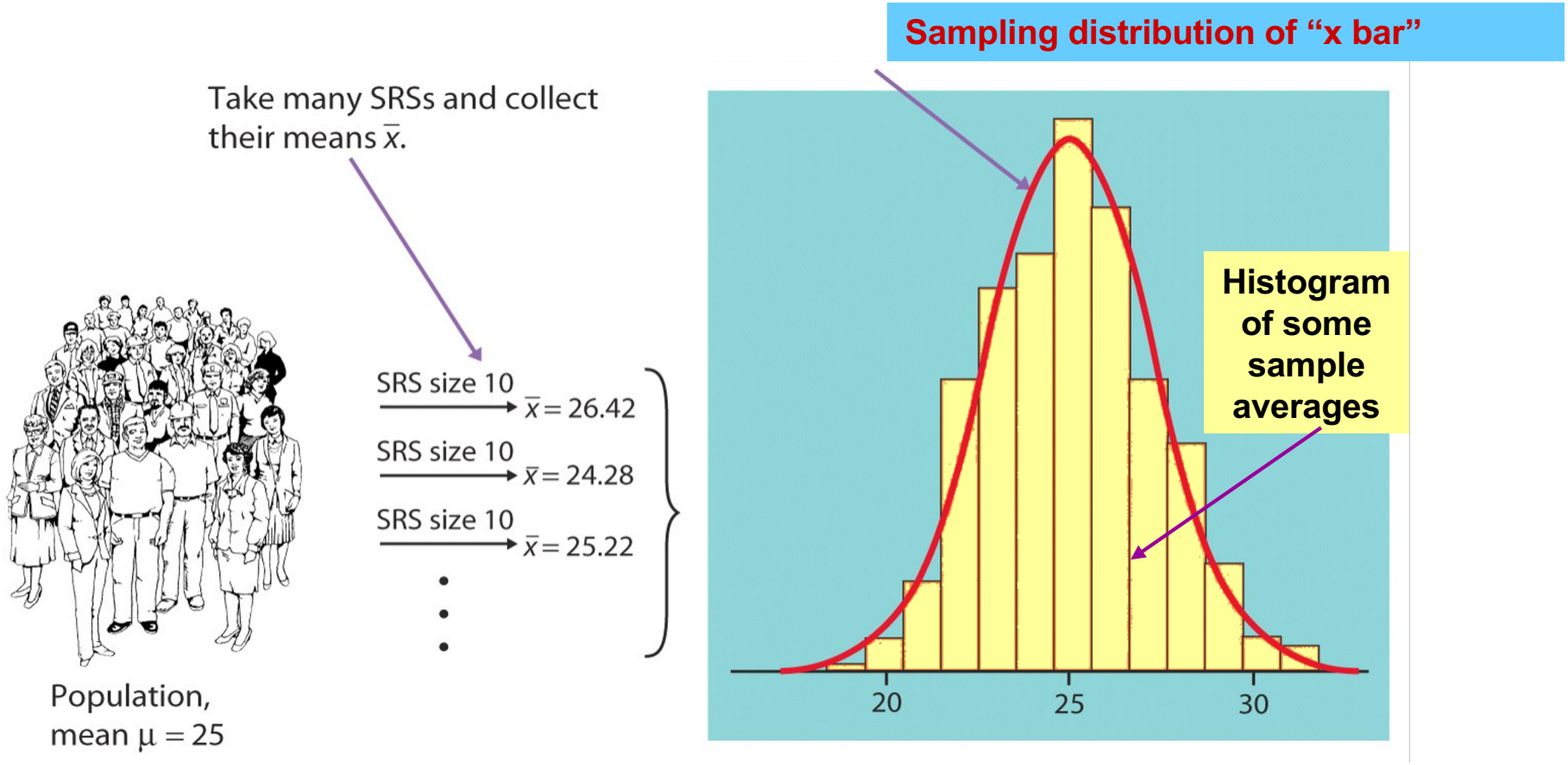
The sampling distribution:

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

Sampling distribution of the sample mean

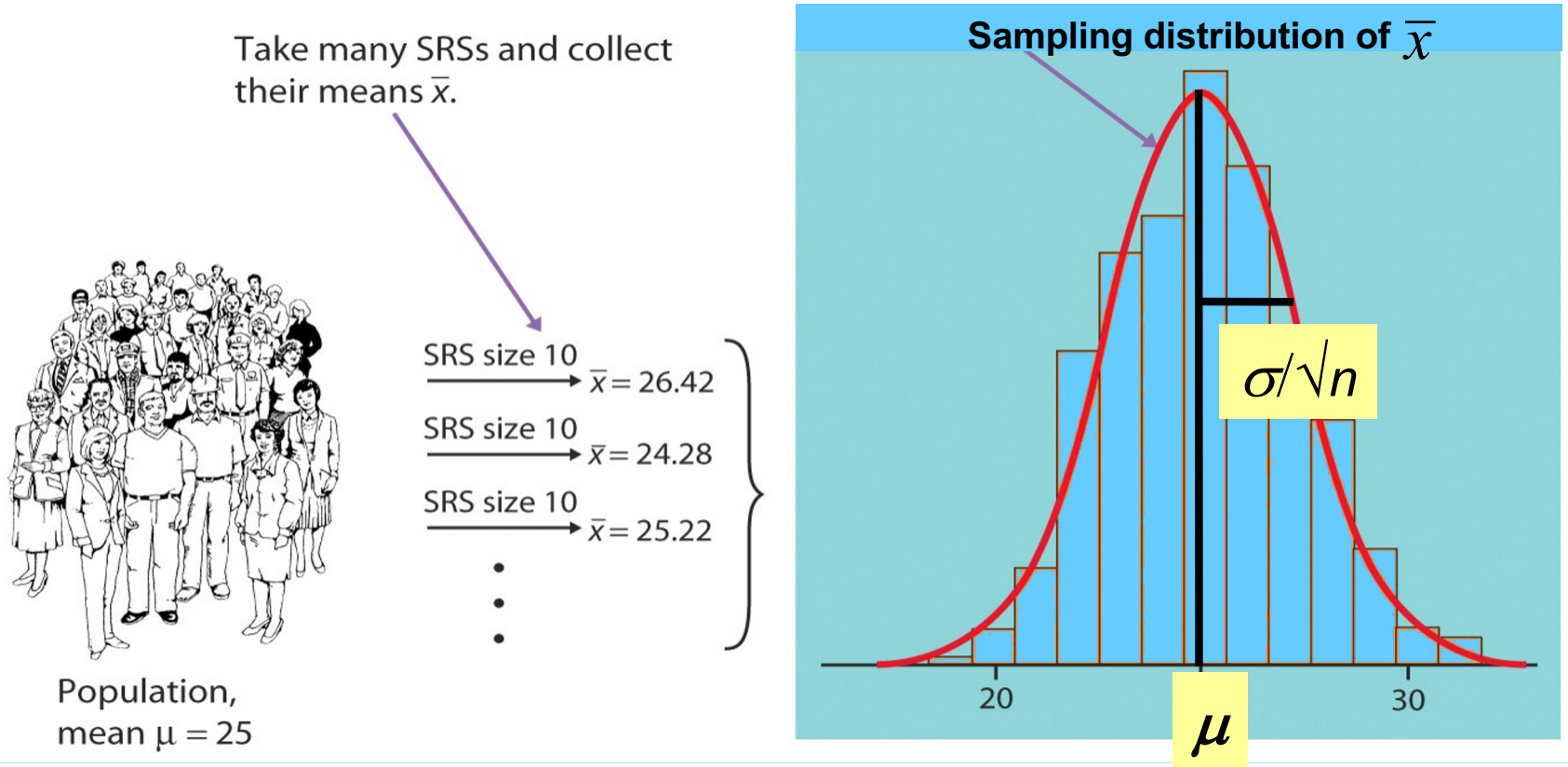
We take many random samples of a given size n from a population with mean μ and standard deviation σ.

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



For any population with mean μ and standard deviation σ:

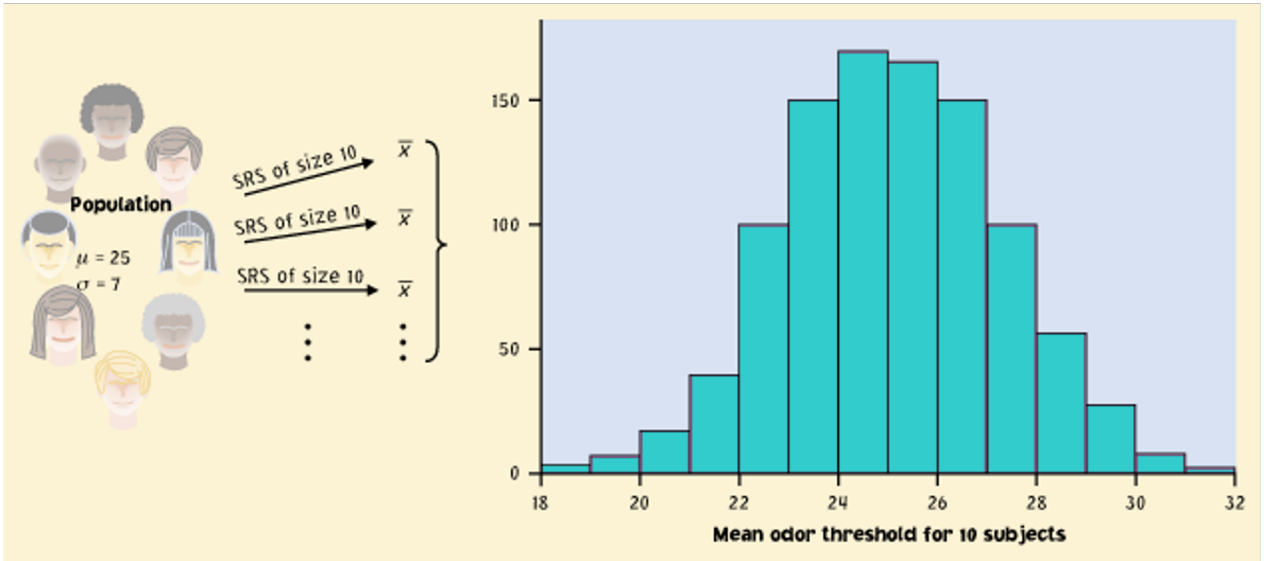
- The mean, or center of the sampling distribution of $\bar{x}$, is equal to the population mean μ.
- The standard deviation of the sampling distribution is $\sigma/\sqrt{n}$, where n is the sample size, it is a measure of how much error there is in the sampling process.



For normally distributed populations

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

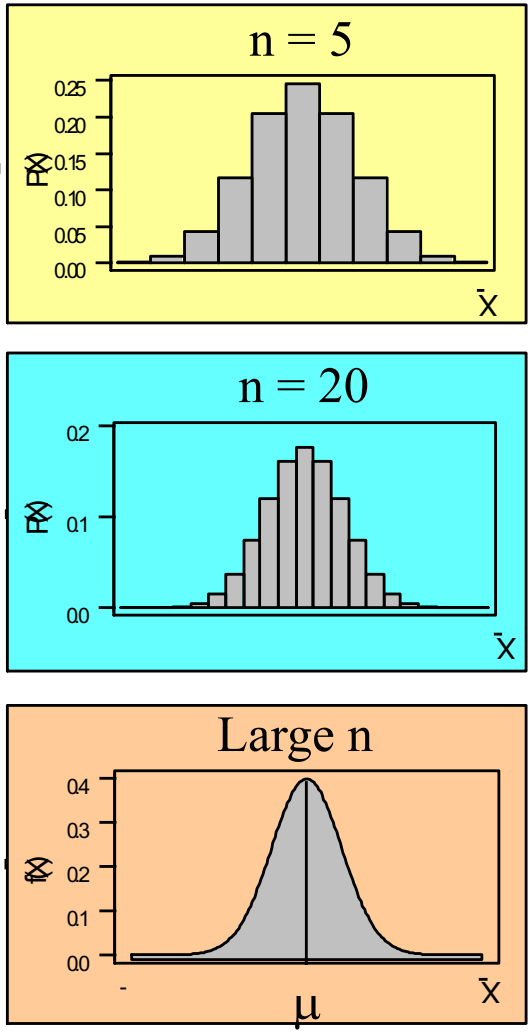
If the population is $N(\mu,\sigma)$, then the sample means distribution is $N(\mu,\sigma/\sqrt{n})$.



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

- 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 $\bar{X}$ 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 the population distribution, the larger n must be before the shape of the sampling distribution is close to normal.
- 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.

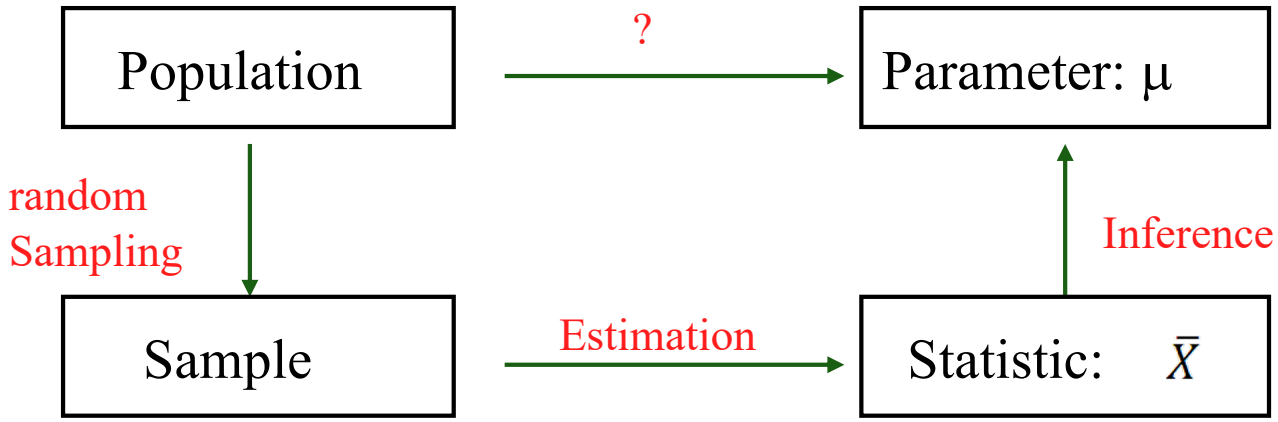
The sampling distribution model for a sample proportion p

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* $= \sqrt{\frac{pq}{n}}$

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

where q = 1 – p

Confidence Intervals



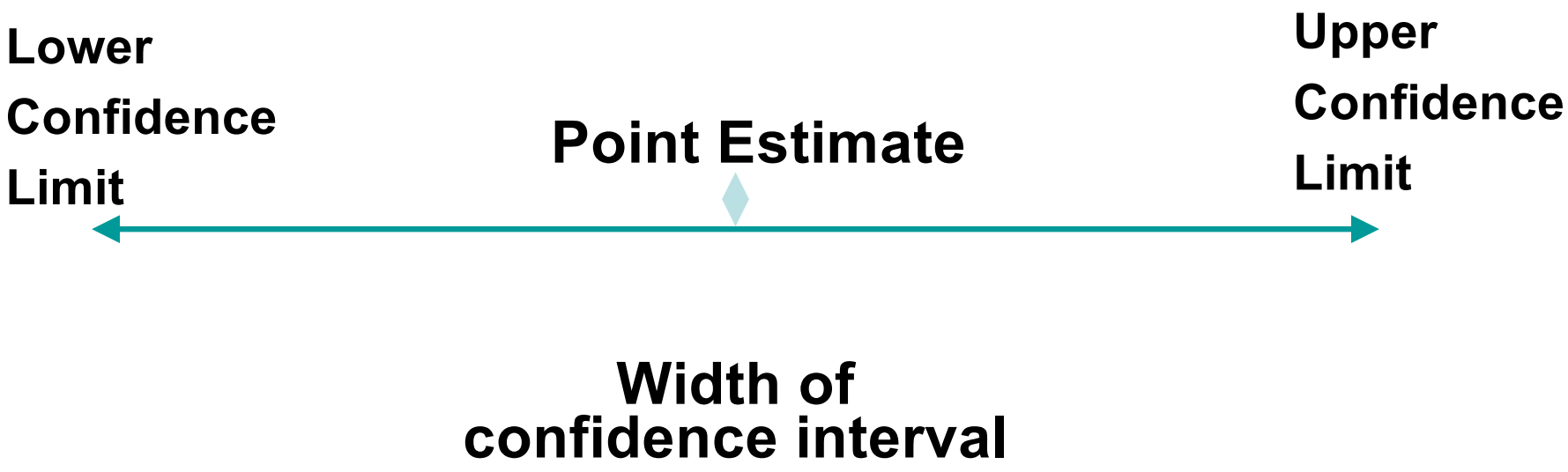
Sample mean $\bar{x}$ is the best (point) estimate of μ

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

Point and Interval Estimates

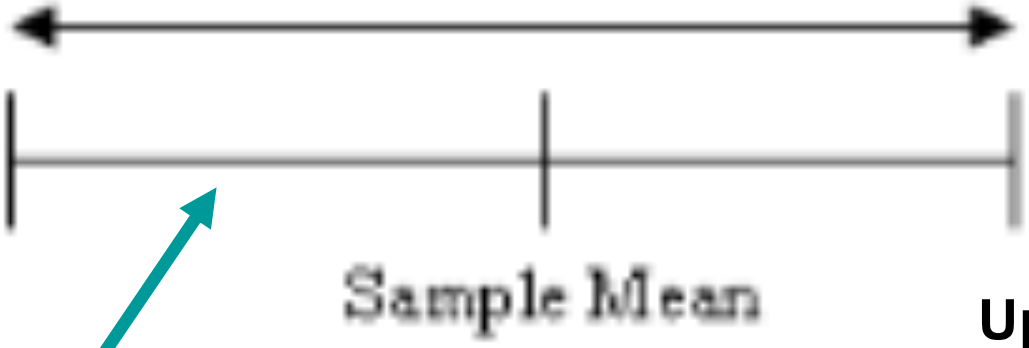
- A point estimate is a single number,
- a confidence interval provides additional information about the variability of the estimate



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



Lower Confidence Limit

Upper Confidence Limit

This is our Confidence Interval, centered on the sample mean

In other words…

In general, we have a sample mean, $\overline{x}$, and are using it as a guess or estimate for the population mean μ. Hence, it would be useful if we could have an interval based on $\overline{x}$, that is, with $\overline{x}$ in the middle.

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

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:

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

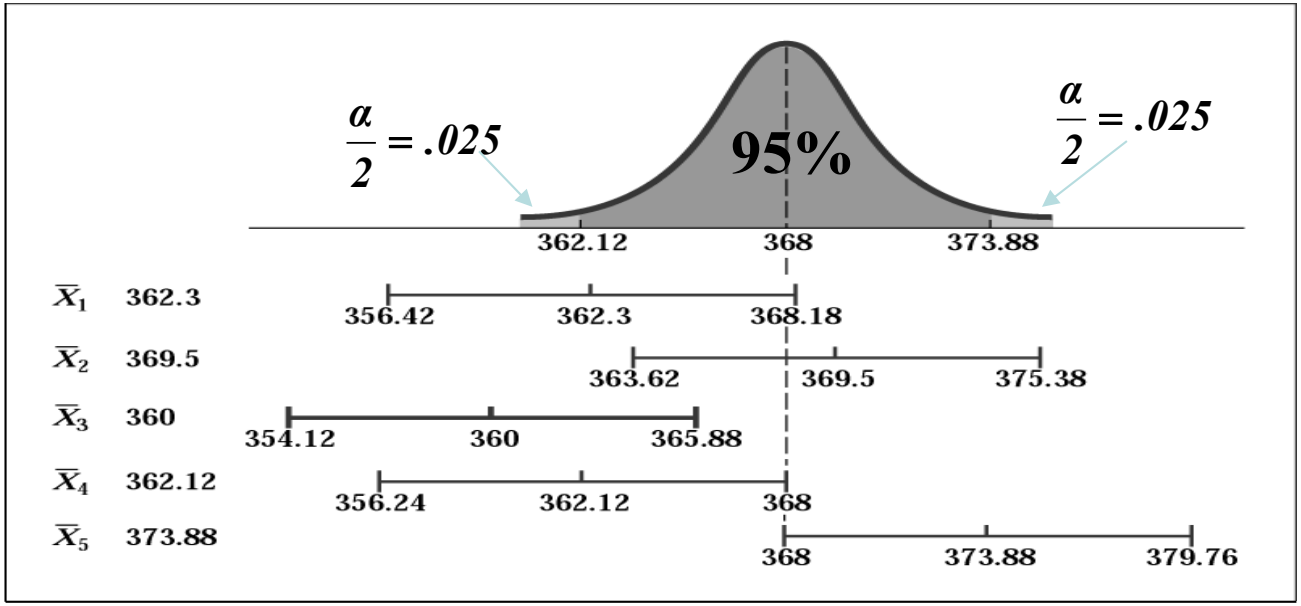
- Rearrange the expression:

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

- This tells us that 95% of the time the true mean should lie within $\pm 1.96(\sigma/\sqrt{n})$ of the sample mean

Confidence Interval μ for (σ is known)

Let's hypothsize 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.

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 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)

Confidence Intervals

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: $z_{\alpha/2} = 1.65$

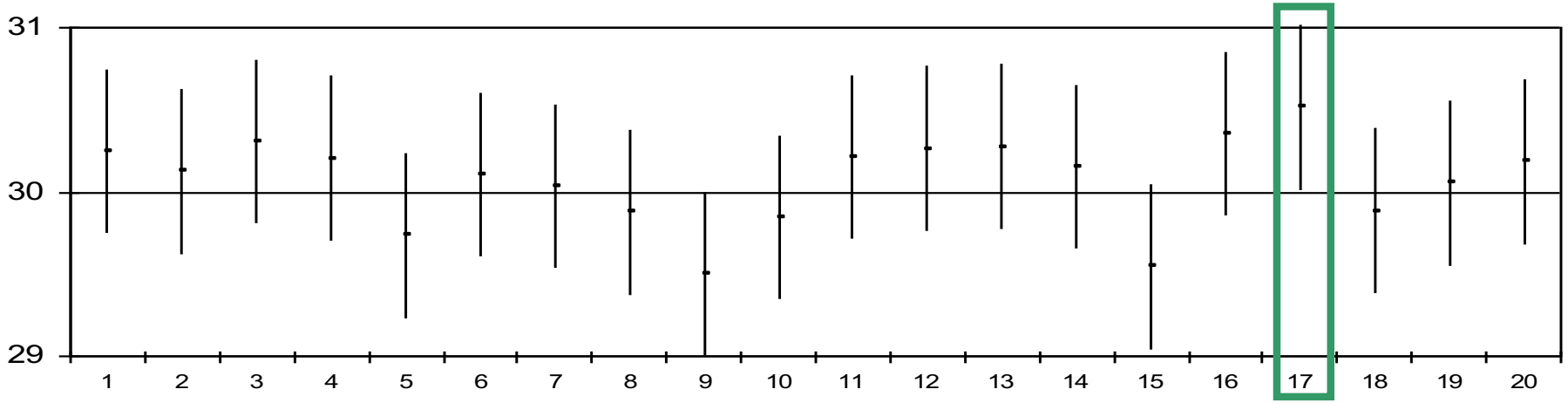
For a 95% confidence interval: $z_{\alpha/2} = 1.96$

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 inetvals conatining contengano 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.

Confidence Interval for μ (σ Unknown)

- Assumptions
  - Population standard deviation is unknown
  - Population is normally distributed
  - If population is not normal,
- Use Student's t Distribution
- Confidence Interval Estimate:

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

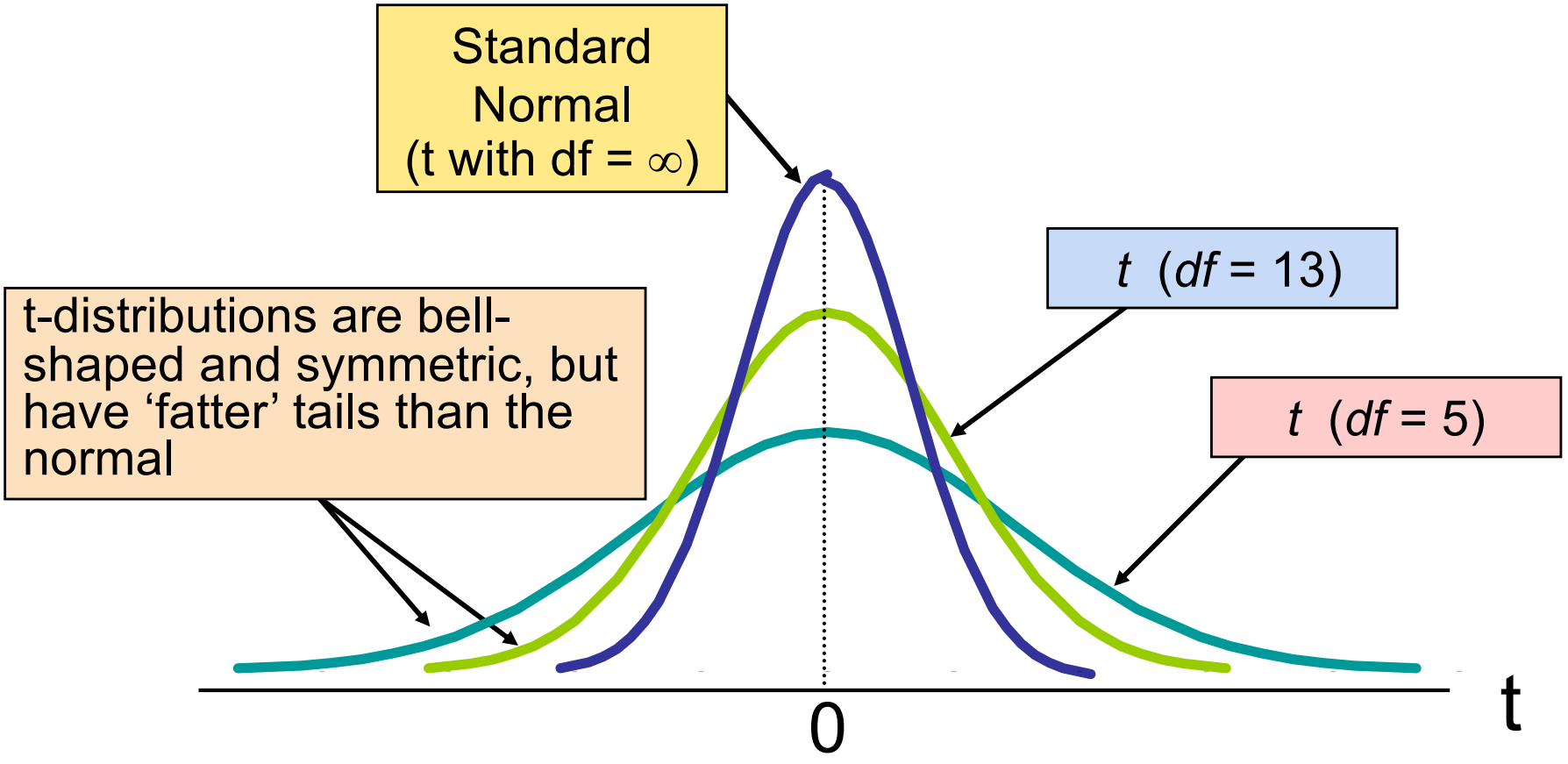
(where $t_{\alpha/2}$ is the critical value of the t distribution with n -1 degrees of freedom and an area of $\alpha/2$ in each tail)

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

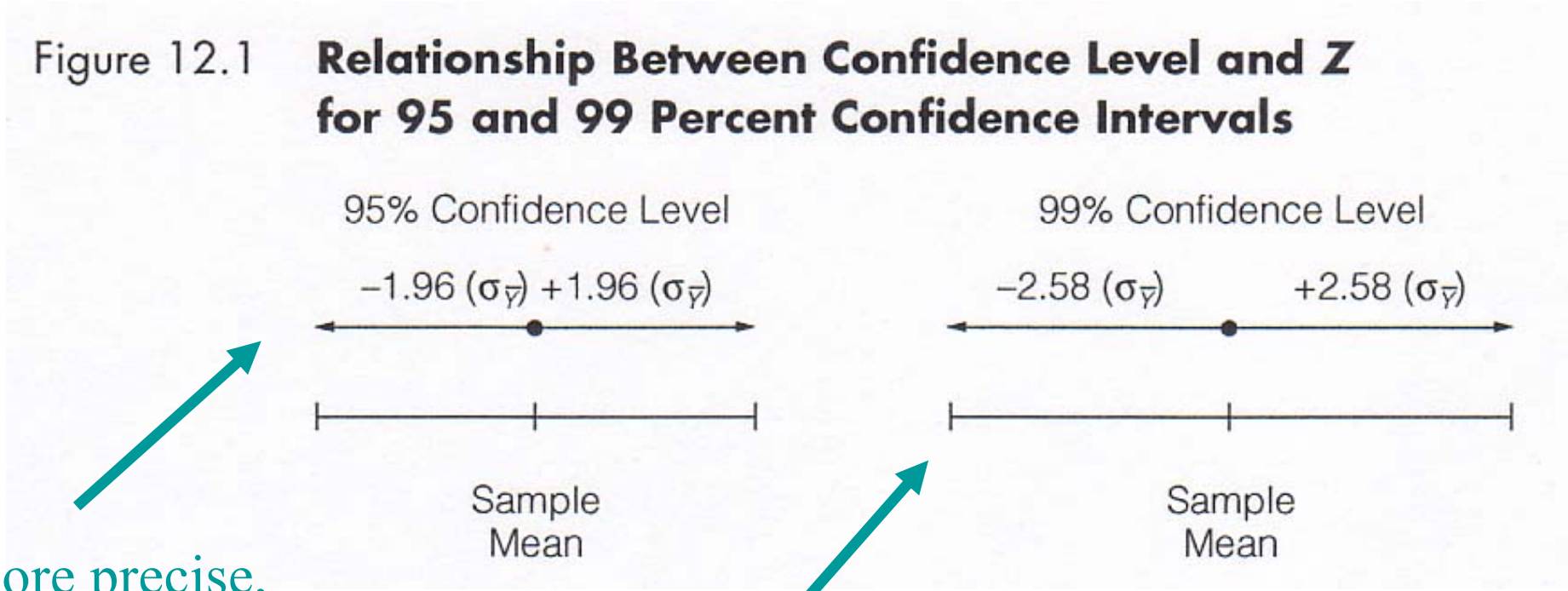
Student's t Distribution

Note: t ⟶ Z as n increases



- 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.

Confidence Interval Width



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**