NNSE 784
Advanced Analytics Methods

Instructor: F Doyle (CESTM L210)

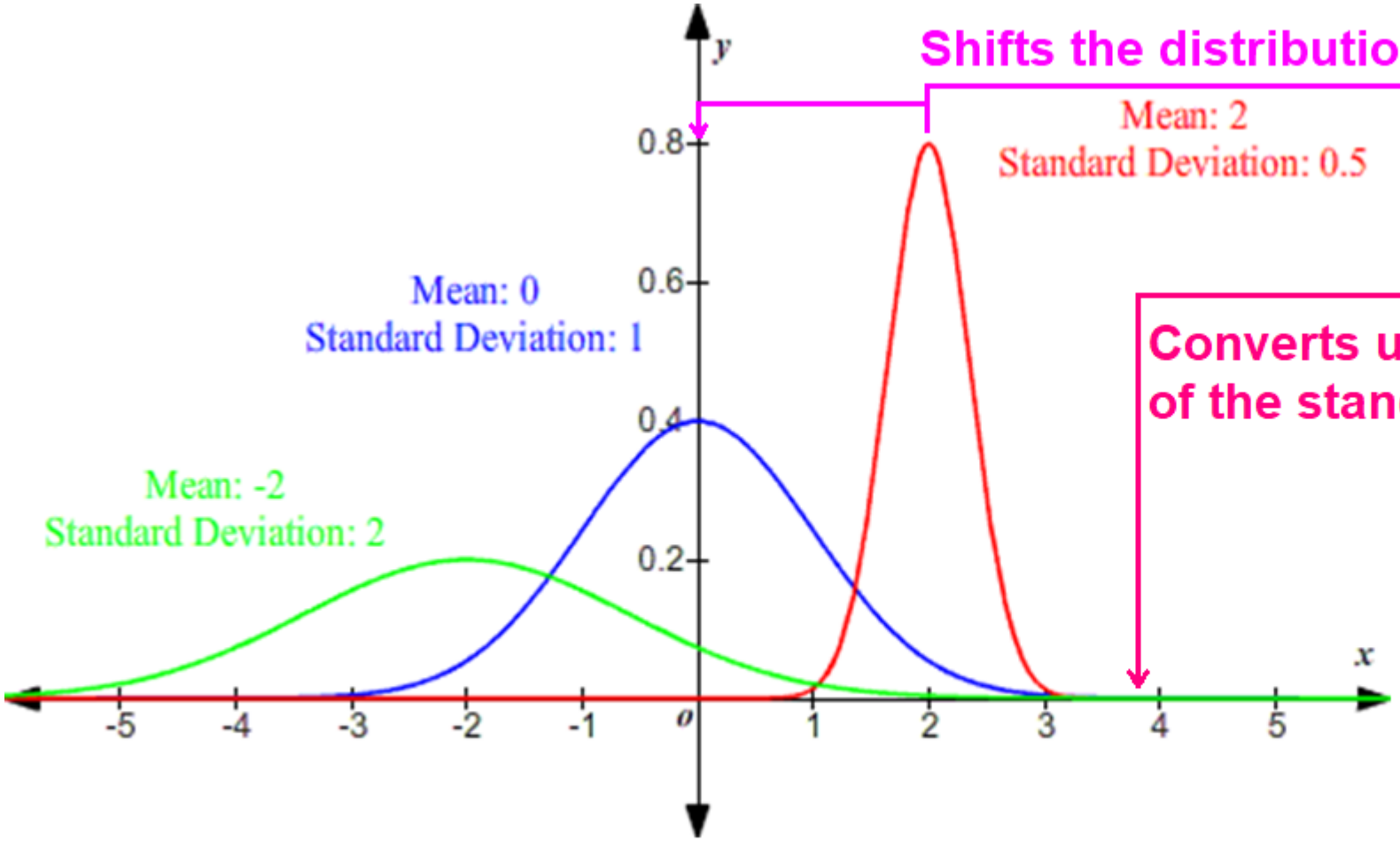
MW 4:30 – 5:50, NFN 203

Slide Set #9
Inferential Statistics: Estimators

Lecture Outline

- Standard Normal Distribution and z-score review
- What are estimators?
- Confidence Interval for a Population Mean
- The *t* Distribution
- Confidence Interval for the Difference Between Two Population Means
- Discuss any points of confusion thus far

What Does The z Distribution Formula Do?



Shifts the distribution mean to zero

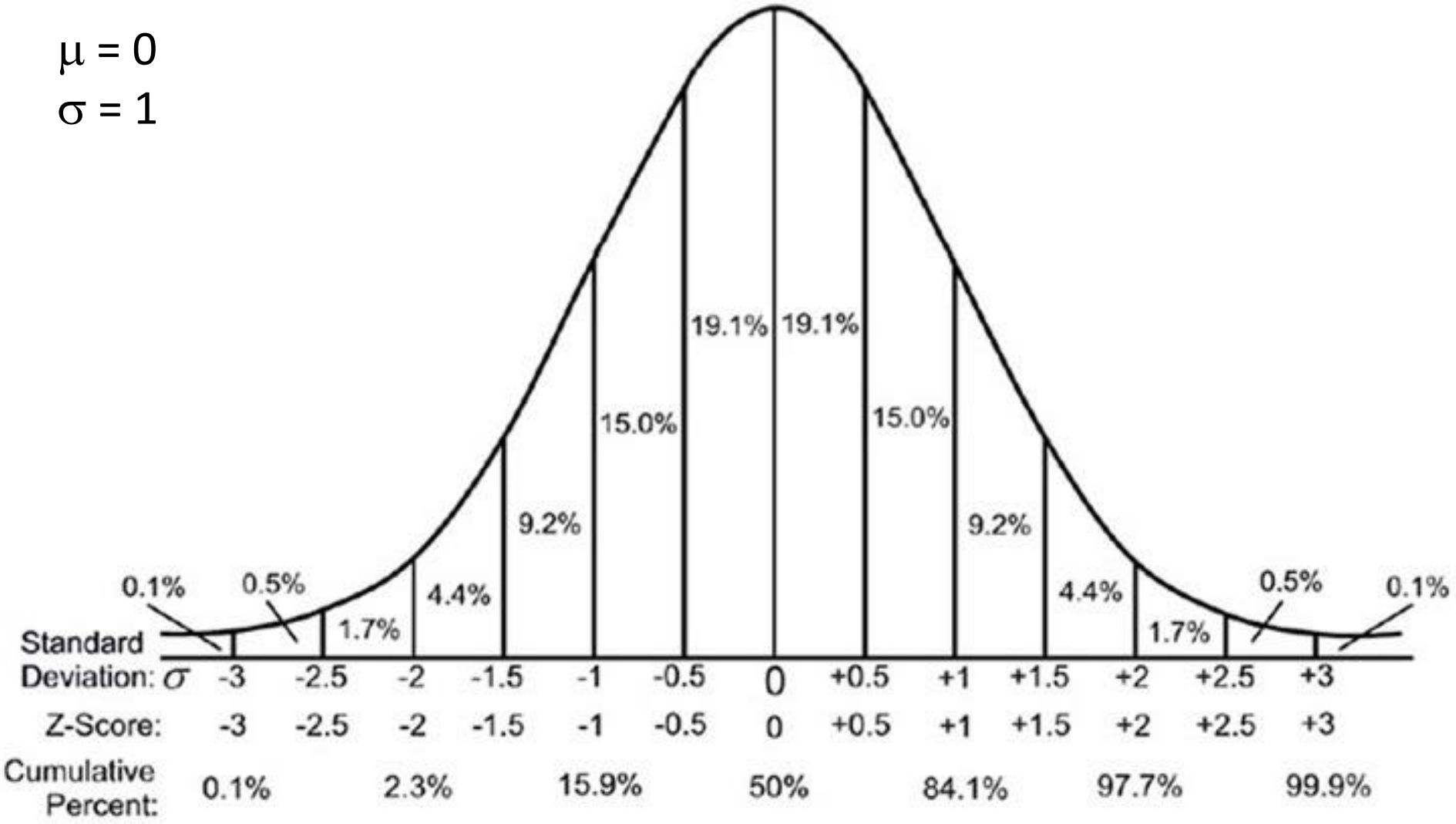
$$z = \frac{x - \mu}{\sigma}$$

Converts units to multiples of the standard deviation

The Standard Normal Distribution

$\mu = 0$

$\sigma = 1$



Descriptive vs Inferential Statistics recap

- Descriptive Statistics – used to summarize and describe a population
  - Central Tendencies – e.g. average values
  - Dispersion – e.g. variance, standard deviation
- Inferential Statistics – draw conclusions about a population based on the data in a sample from that population
  - Estimation – generate approximate values for parameters of the population the samples were taken from
  - Hypothesis tests – with regard to statistics, we can think of a hypothesis as a statement about one or mores populations

Estimation

- The calculation of a statistic from sample data that approximates a corresponding parameter of the population the sample came from
- The statistic used to calculate an estimate is referred to as an "estimator"
- There are two types of estimates
  - ***Point estimate*** a single numeric value used to estimate the corresponding population parameter
  - ***Interval estimate*** two numeric values defining a range that, with a specified degree of confidence, likely include the value of parameter being estimated

Example Estimator

Estimators are usually presented as formulas, for example:

$$\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n}$$

$\bar{x}$ is an estimator of the population mean, μ

We say that an estimator, such as "T" of the parameter θ, is an unbiased estimator of θ if $E(T)=\theta$

E(T) is read as "the expected value of T" - $E(T) = \mu_T$

Sampled Populations and Target Populations

- *Sampled Population* – the population from which one actually draws a sample
- *Target Population* – the population about which one wished to make an inference

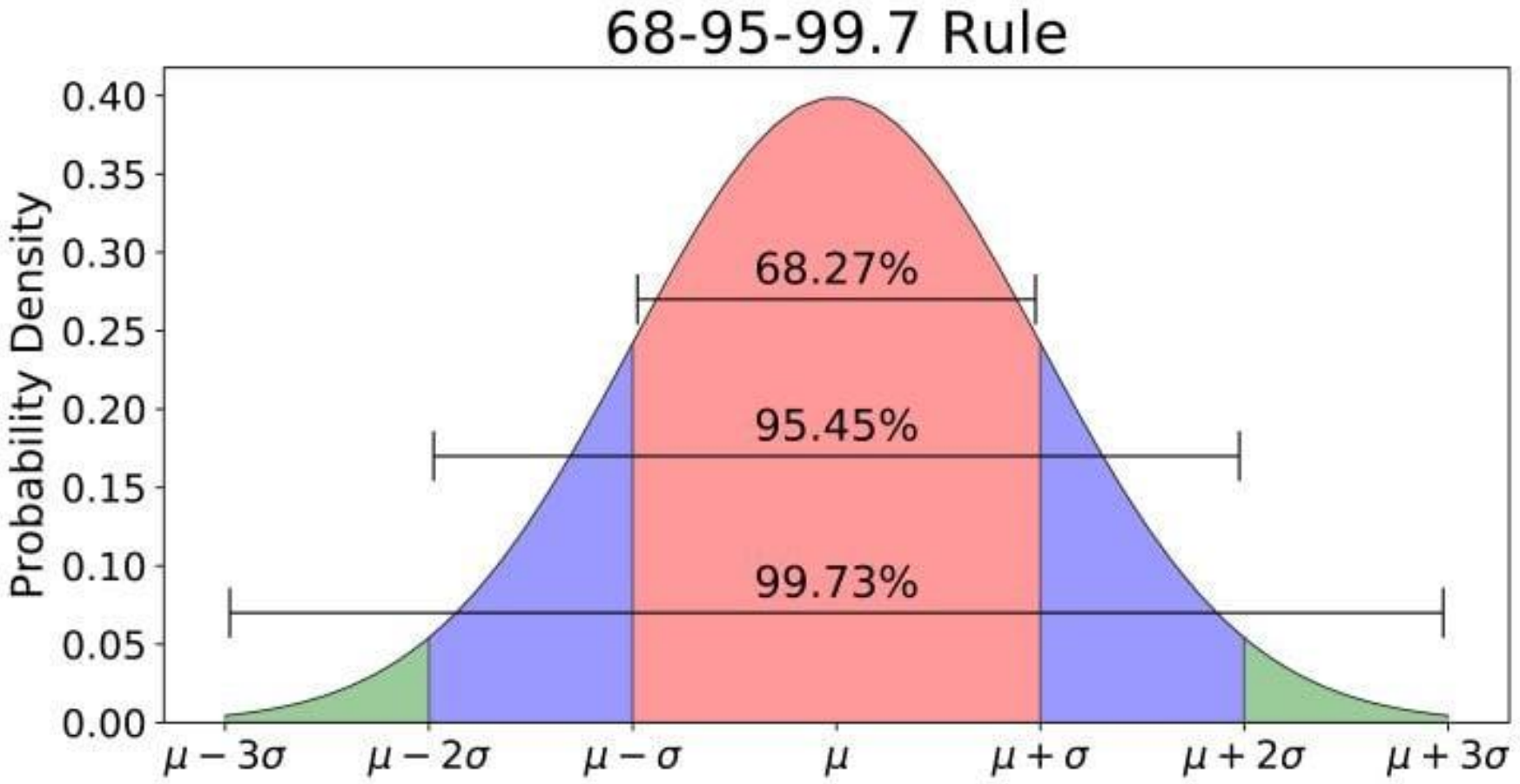
These are not always the same. Statistical inference procedures allow you to make inferences about sampled populations.

<u>If the sampled and target populations are different, you can reach conclusions about the target population only on the basis of non-statistical considerations.</u>

Confidence Interval for a Population Mean

- We can use a sample to calculate $\bar{x}$ and produce a point estimate of μ, however… as random sampling inherently involves chance, $\bar{x}$ cannot be expected to equal μ.
- It would be more meaningful to estimate μ by an interval that communicates information regarding the probable magnitude of μ.
- To obtain an interval estimate, we must utilize our knowledge of sampling distributions.
- We know that the sampling distribution of the sampling mean will be normally distributed if drawn from a normal population (or if n is large enough), so:
  - $\mu_{\bar{x}} = \mu$
  - $\sigma^2_{\bar{x}} = \sigma^2/n$

Area Versus Standard Deviations from the Mean – The Empirical Rule / 68-95-99.7 Rule

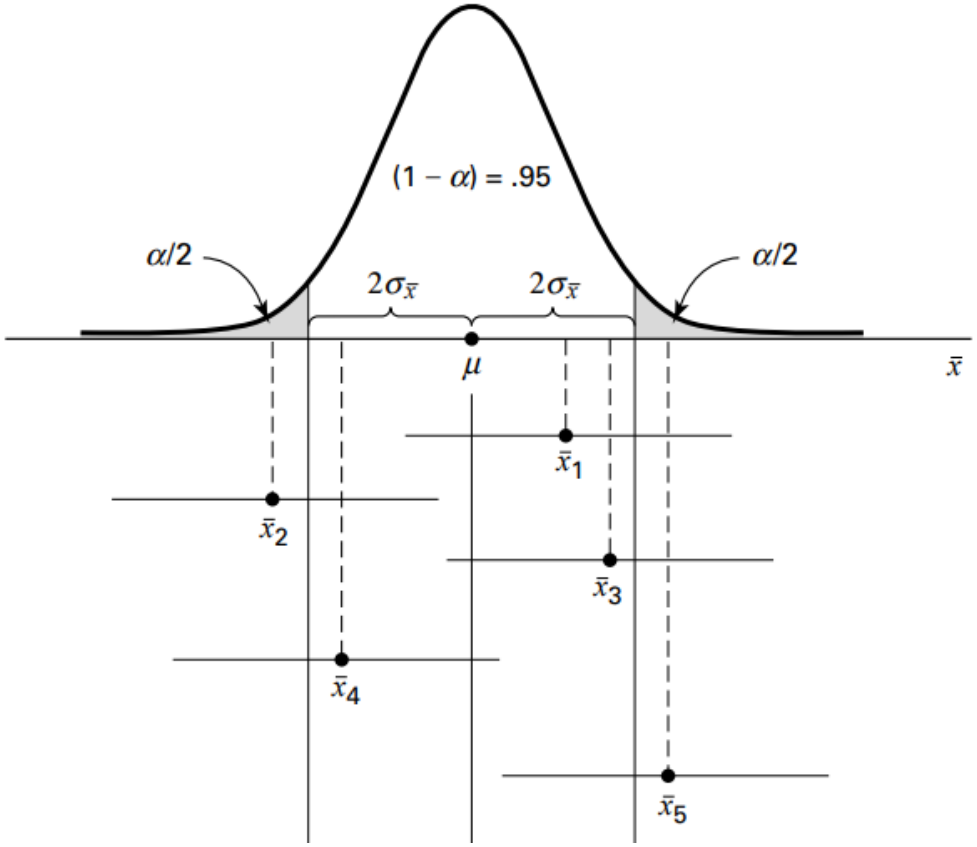


No matter what μ and σ are, the area between μ+/- 1σ is about 68%; the area between μ+/- 2σ is about 95%; and the area between μ+/- 3σ is about 99.7%. Almost all values fall within 3 standard deviations.

Confidence Interval for a Population Mean continued

- We don't know the value for the μ, so $\mu \pm 2\sigma_{\bar{x}}$ doesn't tell us much
- We do, however, have the point estimate of μ, that is, $\bar{x}$

If we construct such intervals about every possible value of $\bar{x}$ computed from all possible samples of size n from the population of interest, we would have a large number of the form $\bar{x} \pm 2\sigma_{\bar{x}}$ with widths equal to the $\pm 2\sigma$ about the unknown μ.
Approximately 95% of these intervals would have centers falling within $\pm 2\sigma_{\bar{x}}$ of μ and therefore contain the value of μ.



Interval Estimate Components

We constructed the 95% confidence interval based on our understanding of the 68-95-99.7 rule and understanding that 95% of values in a normal distribution fall within +or- 2 standard deviations of the mean.

$$\bar{x} \pm 2\sigma_{\bar{x}}$$

The general form is:

estimator ± (reliability coefficient) x (standard error)

So, an interval estimate for μ would be:

$$\bar{x} \pm z_{(1-\alpha/2)}\sigma_{\bar{x}}$$

…where $z_{(1-\alpha/2)}$ is the value of z to the left of which lies $1 - \alpha/2$ of the area under its curve and to the right of which lies $\alpha/2$ of the area under its curve. $(1-\alpha)$ is the ***confidence coefficient*** value we are using to create the interval.

Probabilistic Interpretation of Confidence Interval

In repeated sampling, from a normally distributed population with a known standard deviation, $100(1-\alpha)$ *percent of all intervals of the form* $\bar{x} \pm z_{(1-\alpha/2)}\sigma_{\bar{x}}$ *will in the long run include the population mean* μ.

i.e., 100(.95) or 95

Practical Interpretation of Confidence Interval

When sampling is from a normally distributed population with known standard deviation, we are $100(1-\alpha)$ *percent confident that the single computed interval,* $\bar{x} \pm z_{(1-\alpha/2)}\sigma_{\bar{x}}$, *contains the population mean* μ.

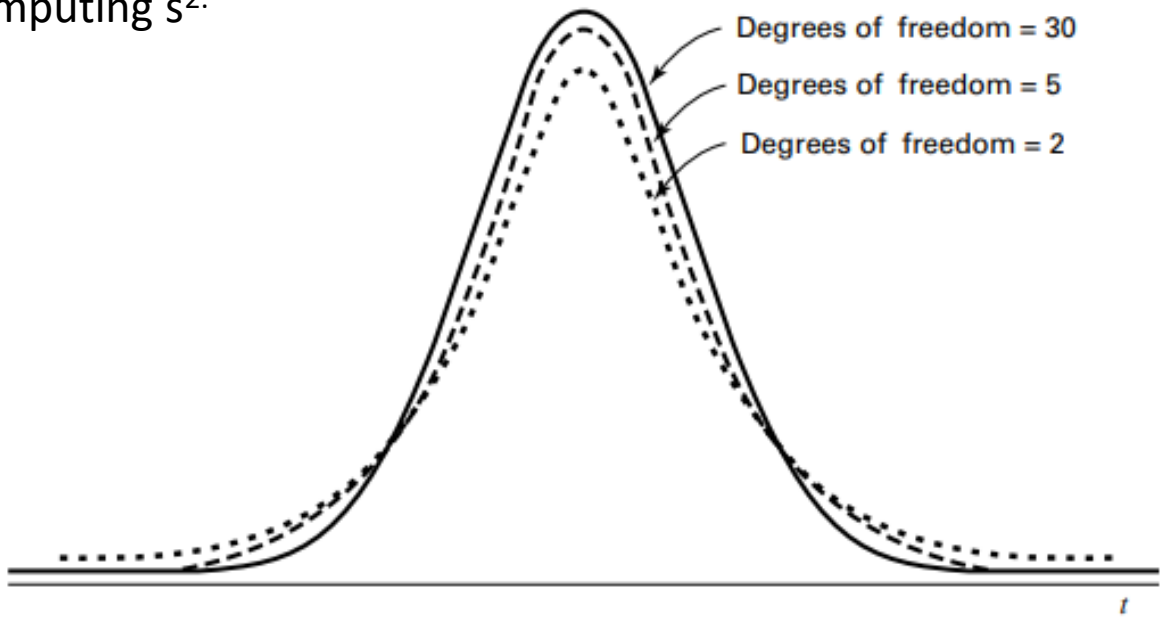
Limitations

- Our procedure for constructing the confidence interval for a population mean required knowledge of the population variance
- It is typically the case that if we do not have the population mean… we also don't have the population variance
- We know $z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$ is normally distributed if the underlying population is normally distributed, or approximately normally distributed if n is large. However, we cannot make use of this if σ is unknown
- However, if n is large (≥ 30) we can make use of the sample standard deviation: $s = \sqrt{\sum (x_i - \bar{x})^2/(n-1)}$
- What about when the sample size is small? We need an alternative procedure.

The *t* Distribution
aka "Student's t Distribution"

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

1. It has a mean of 0
2. It is symmetrical about the mean
3. In general, it has a variance greater than 1, but the variance approaches 1 as sample size becomes large. For df>2, the variance of the t distribution is df/(df-2), where df is the "degrees of freedom"
4. The variable t ranges from $-\infty$ to $+\infty$.
5. The t distribution is really a family of distributions, since there is a different distribution for each sample value of n-1, the divisor used in computing s^2.



Degrees of Freedom (sidenote)

- You will often see “degrees of freedom” (df) mentioned in relation to statistics formulas
- Definitions of this concept are often confusing
- Think of it in terms of a control used to avoid underestimating certain statistics (particularly things like variance)
- Simplified interpretation is that the df corresponds to number of observations we use to estimate a parameter minus the number of intermediate parameters we need to estimate

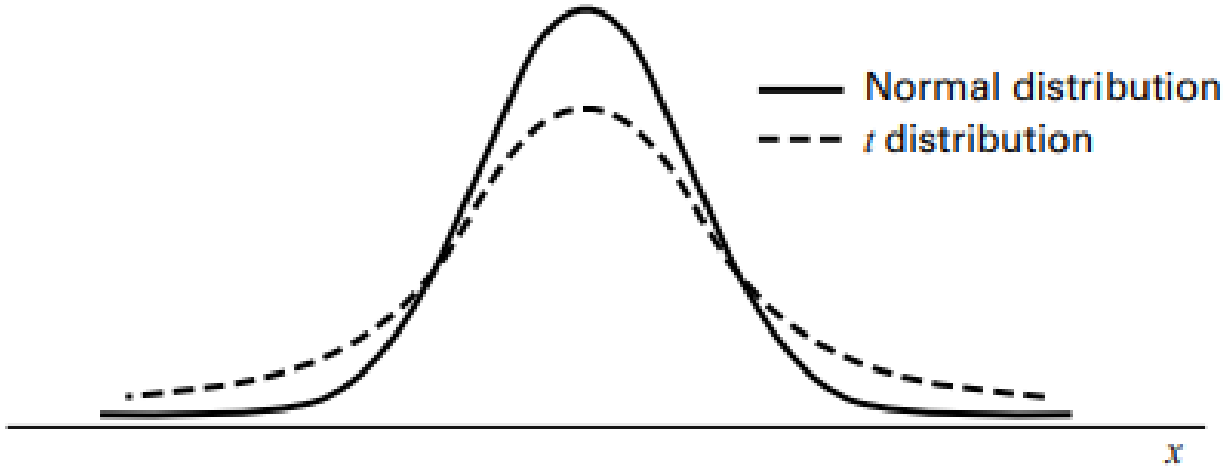
Estimate of standard deviation:

$$SD = \sqrt{\sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}}$$

df = (n-1), because we have already estimated 1 parameter, $\bar{x}$

t Distribution Compared to Normal Distribution

- Compared to the normal distribution, the *t* distribution is less peaked in the center and has thicker tails.
- The *t* distribution approaches the normal distribution as n-1 approaches infinity



Confidence Intervals Using *t*

- Same general procedure as shown previously:

  estimator $\pm$ (reliability coefficient) x (standard error of the estimator)

- What has changed is the source of the reliability coefficient. It is now obtained from the *t* distribution rather than from the standard normal distribution (z). Specifically, *when sampling from a normal distribution whose standard deviation, σ, is unknown, the 100(1 – α) percent confidence interval for the population mean, μ, is given by:*

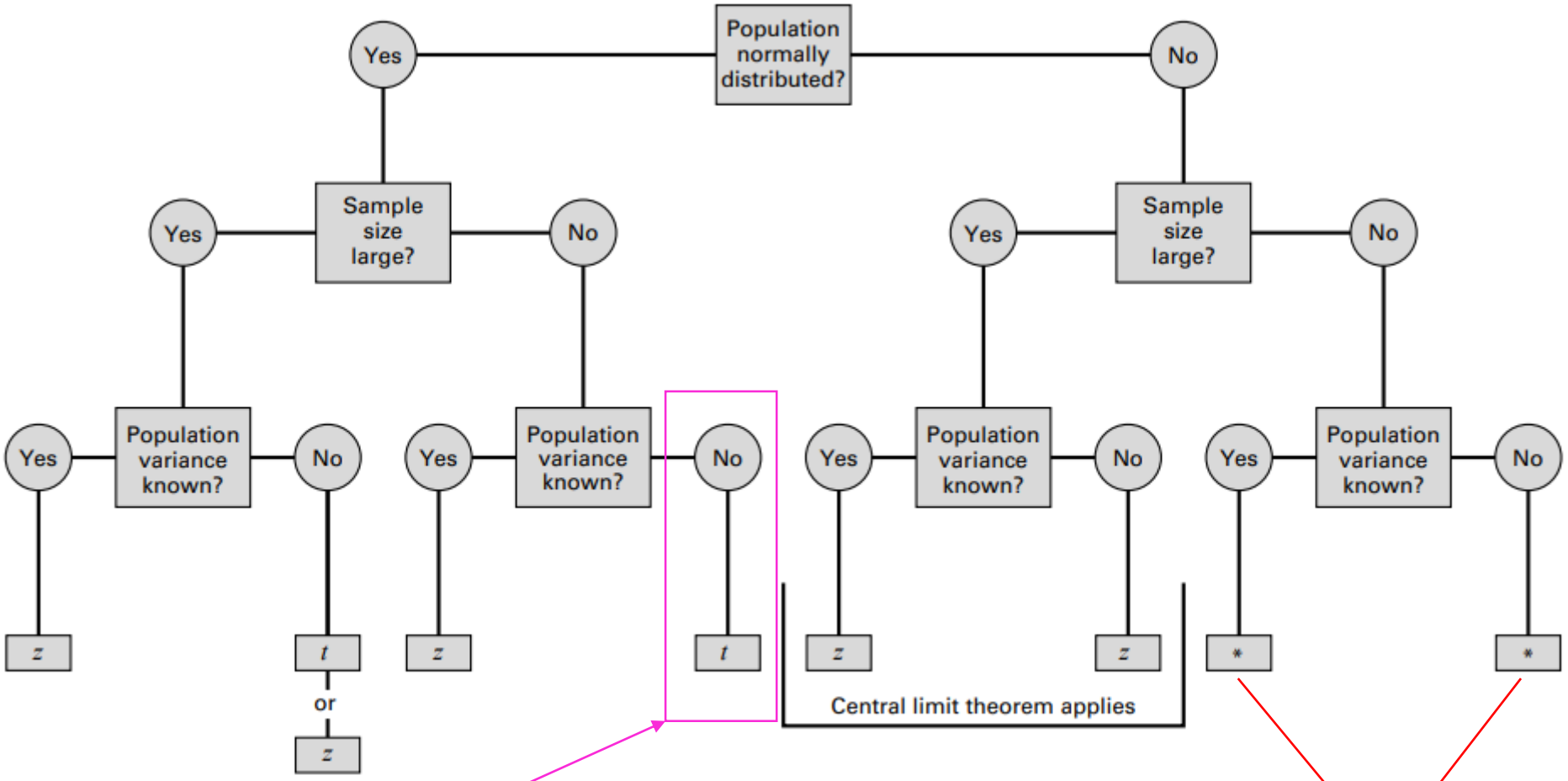
$$\bar{x} \pm t_{(1-\alpha/2)} \frac{s}{\sqrt{n}}$$

Expectations for Using *t*

- The **strictly valid** use of *t* requires that the sample must be drawn from a normal distribution!
- That said, experience has shown that moderate departures from this requirement can be tolerated.
- Consequently, the *t* distribution is used even when it is known that the parent population deviates somewhat from normality.
- "Most researchers require that an assumption of, at least, a mound-shaped population distribution be plausible"

This is a quote from a statistics text!

Deciding Between z and t



With an assumption of normalcy, this may be the most common position in analyzing experimental data

* = “use non-parametric procedure”

Confidence Interval for the Difference Between Two Population Means

- When population variances are known, the 100(1 – α) percent confidence interval for $\mu_1 - \mu_2$ is given by:

$$(\bar{x}_1 - \bar{x}_2) \pm z_{1-\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

t Distribution and the Difference Between Two Population Means

- When population variances are unknown but are assumed to be equal, the 100(1 – α) percent confidence interval for $\mu_1 - \mu_2$ is given by:

$$(\bar{x}_1 - \bar{x}_2) \pm t_{(1-\alpha/2)} \sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}$$

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

Pooled estimate of common variance

- When population variances are unknown and are assumed to **not** be equal, the 100(1 – α) percent confidence interval for $\mu_1 - \mu_2$ is given by:

$$(\bar{x}_1 - \bar{x}_2) \pm t'_{(1-\alpha/2)} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$$t'_{1-\alpha/2} = \frac{w_1 t_1 + w_2 t_2}{w_1 + w_2}$$

Adjusted reliability factor

where $w_1 = s_1^2/n_1, w_2 = s_2^2/n_2, t_1 = t_{1-\alpha/2}$ for $n_1 - 1$ degrees of freedom, and $t_2 = t_{1-\alpha/2}$ for $n_2 - 1$ degrees of freedom.

Deciding Between z, t and t’ for Reliability Factor