

Statistical estimation: point vs. interval estimation

- Up until this point, we have assumed that the values of the parameters of a probability distribution are known.
- In the real world, the values of these population parameters are usually not known.
- Instead, we must try to say something about the way in which a random variable is distributed using the information contained in a sample of observations.

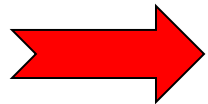
- The process of drawing conclusions about an entire population based on the data in a sample is known as statistical inference.
- Methods of inference usually fall into one of two broad categories: estimation or hypothesis testing.
- For now, we will focus on using the observations in a sample to estimate a population parameter.

Estimation

- Is concerned with estimating the values of specific population parameters based on sample statistics.
- is about using information in a sample to make estimates of the characteristics (parameters) of the source population.

Example

- A sample survey revealed:
 - Proportion of smokers among a certain group of population aged 15 to 24.
 - Mean of SBP among sampled population
 - Prevalence of HIV-positive among people involved in the study



The next question is what can we predict about the characteristics of the population from which the sample was drawn

Estimation, Estimator & Estimate

- ♣ Estimation is the computation of a statistic from sample data, often yielding a value that is an approximation (guess) of its target, an unknown true population parameter value.
- ♣ The statistic itself is called an estimator and can be of two types - point or interval.
- ♣ The value or values that the estimator assumes are called estimates.

- Two methods of estimation are commonly used: point estimation and interval estimation.
- Point estimation involves the calculation of a single number to estimate the population parameter.
- Interval estimation specifies a range of reasonable values for the parameter.

Point versus Interval Estimators

- ♣ An estimator that represents a "single best guess" is called a point estimator.
- ♣ When the estimate is of the form of a "range of plausible values", it is called an interval estimator.
- ♣ Thus,
 - A point estimate is of the form: [Value],
 - Whereas, an interval estimate is of the form: [lower limit, upper limit]

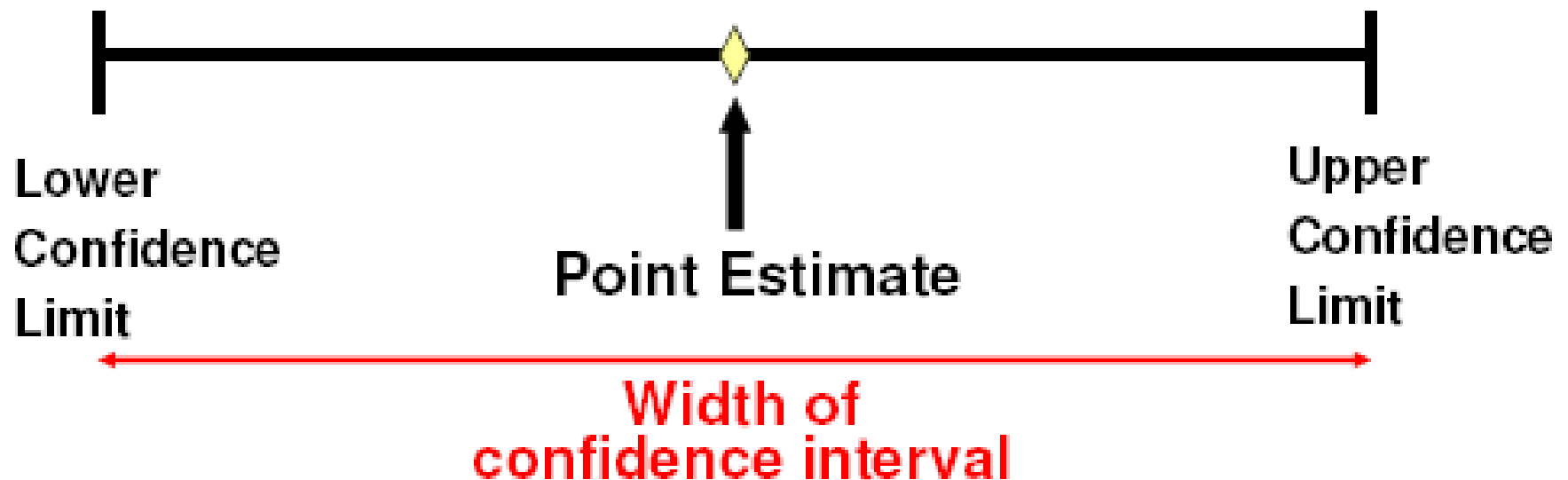
Example -

The sample mean $\bar{X}_n$, calculated using data in a sample of size n , is a point estimator of the population mean μ . If $\bar{X}_n = 10$, the value 10 is called a point estimate of the population mean μ .

Sample mean ($\bar{X}$) is an unbiased estimator of population mean

$$E (\bar{X}) = \mu$$

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



Properties of good estimates

A. Unbiased Estimator

- A sample statistic whose mean is equal to the population parameter it estimates is unbiased.
- The "long run" average of the statistic is equal to the population parameter value.
- The sample mean and median are unbiased estimators of the population mean μ .

B. Minimum Variance

Illustration for data from a normal distribution

1. The unbiased estimators are the sample mean $\bar{X}$ and median $\tilde{X}$
2. $\text{variance} [\bar{X}] < \text{variance} [\tilde{X}]$

Choose the sample mean $\bar{X}$. It is the minimum variance unbiased estimator.

For a random sample of data from a normal probability distribution, $\bar{X}$ is the minimum variance unbiased estimator of the population mean μ .

Estimating the Sampling Error

- Any estimates derived from samples are subject to the sampling error.
- This comes from the fact that only a part of the population was observed, instead of the whole.
- A different samples could have come up with different results.
- The amount of variation that exists among the estimates from the different possible samples is the sampling error.

- The set of sample means in repeated random samples of size n from a given population has a variance of σ^2/n .
- The standard deviation of this set of sample means is $\sigma/\sqrt{n}$ and is referred to as the standard error of the mean (sem) or the standard error.
- The sem is estimated by $s/\sqrt{n}$ if σ is unknown.

- The sampling error is dependent on sample size (n), the variability of individual sample points (σ), sampling and estimation methods.
- As n increases, the sample mean ($\bar{x}$) and the sample variance s^2 approach the values of the true population parameters, μ and σ^2 , respectively.

Example

Suppose that the mean $\pm$ sd of DBP on 20 old males is 78.5 ± 10.3 mm Hg.

1. What is our best estimate of μ ?
2. What is the sem?
3. Compare the sem with the sd.

- The following table gives the se for mean of DBP for different sample sizes.

n	sem
1	10.3
20	2.3
100	1.0

- Our best estimate of μ is 78.5.
- The sem of this estimate is $10.3/\sqrt{20} = 2.3$
- The sem (2.3) is much smaller than sd (10.3).

1. Point Estimate

- A single numerical value used to estimate the corresponding population parameter.

Sample Statistics are Estimators	of Population Parameters
Sample mean,	μ
Sample variance, S^2	σ^2
Sample proportion,	P or π
Sample Odds Ratio, OR	OR
Sample Relative Risk, RR	RR
Sample correlation coefficient,	ρ

2. Interval Estimation

- **Interval estimation** specifies a range of reasonable values for the population parameter based on a point estimate.
- **A confidence interval is a particular type of interval estimator.**

Confidence Intervals

- Give a plausible range of values of the estimate likely to include the “true” (population) value with a given confidence level.
- An interval estimate provides more information about a population characteristic than does a point estimate.
- Such interval estimates are called confidence intervals.

- CIs also give information about the precision of an estimate.
- How much uncertainty is associated with a point estimate of a population parameter?
- When sampling variability is high, the CI will be wide to reflect the uncertainty of the observation.
- Wider CIs indicate less certainty.

- CIs can also answer the question of *whether or not* an association exists or a treatment is beneficial or harmful. (analogous to p-values...)

e.g., if the CI of an odds ratio includes the value 1.0 we cannot be confident that exposure is associated with disease.

- **A CI in general:**

- Takes into consideration variation in sample statistics from sample to sample
- Based on observation from 1 sample, gives information about closeness to unknown population parameters
- Stated in terms of level of confidence
 - Never 100% sure

General Formula:

The general formula for all CIs is:

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

point estimate $\pm$ (measure of how confident we want to be) $\times$ (standard error)

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

Standard error of the statistic.

Lower limit = Point Estimate - (Critical Value)
x (Standard Error)

Upper limit = Point Estimate + (Critical Value)
x (Standard Error)

- A wide interval suggests imprecision of estimation.
- Narrow CI widths reflects large sample size or low variability or both.
- *Note: Measure of how confident we want to be = critical value = confidence coefficient*

Confidence Level

- Confidence Level
 - Confidence in which the interval will contain the unknown population parameter
- A percentage (less than 100%)
 - Example: 95%
- Also written $(1 - \alpha) = .95$

Definition: 95% CI

1. Probabilistic interpretation:

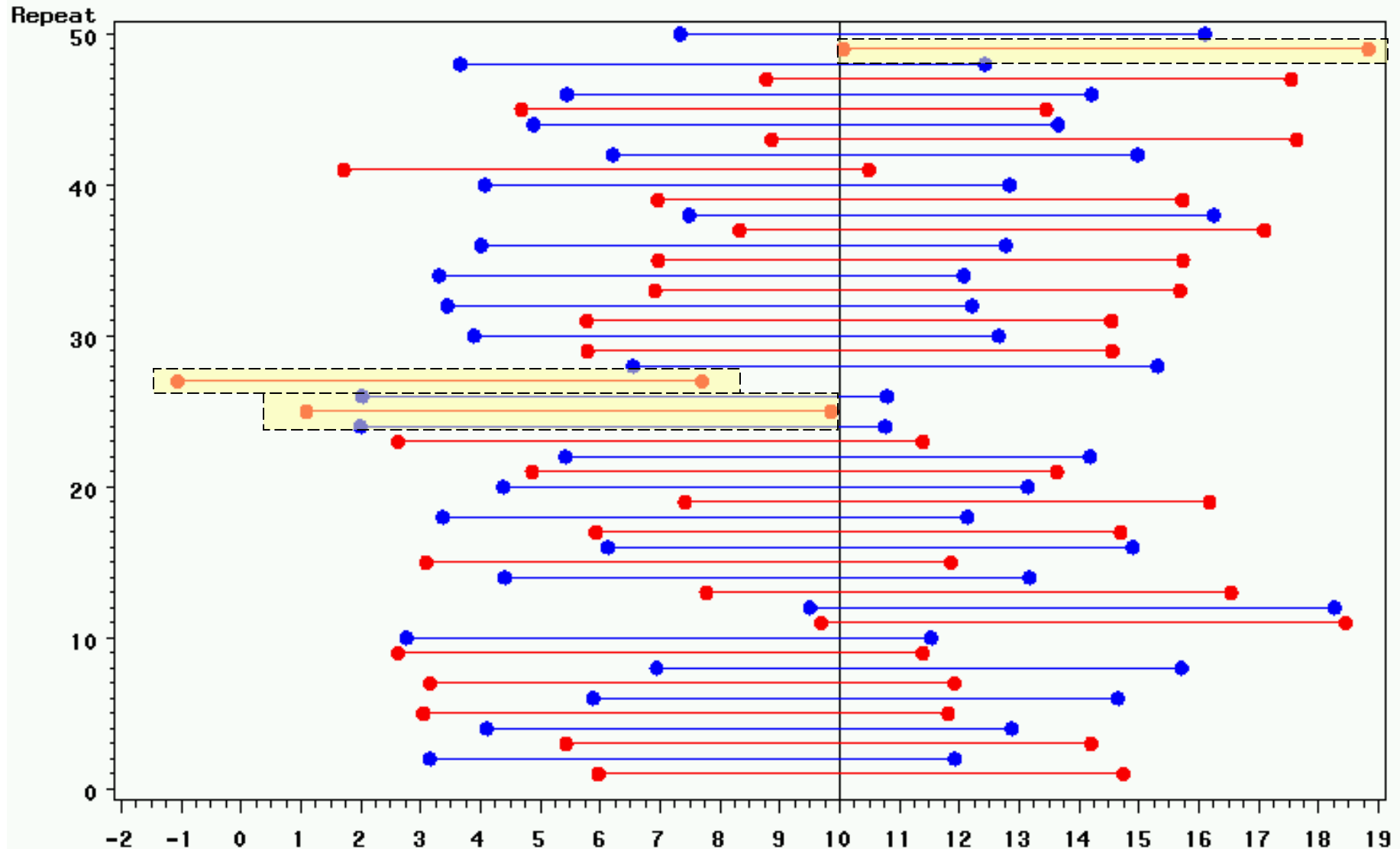
- If all possible random samples (an infinite number) of a given sample size (e.g. 10 or 100) were obtained and if each were used to obtain its own CI, then 95% of all such CIs would contain the unknown population parameter; the remaining 5% would not.
- It is incorrect to say *“There is a 95% probability that the CI contains the unknown population parameter”*.

2. Practical interpretation

- When sampling is from a normally distributed population with known standard deviation, we are $100(1-\alpha)$ [e.g., 95%] confident that the single computed interval contains the unknown population parameter.

Parent population=normal; n=5

Use known standard deviation; Use Z-scores

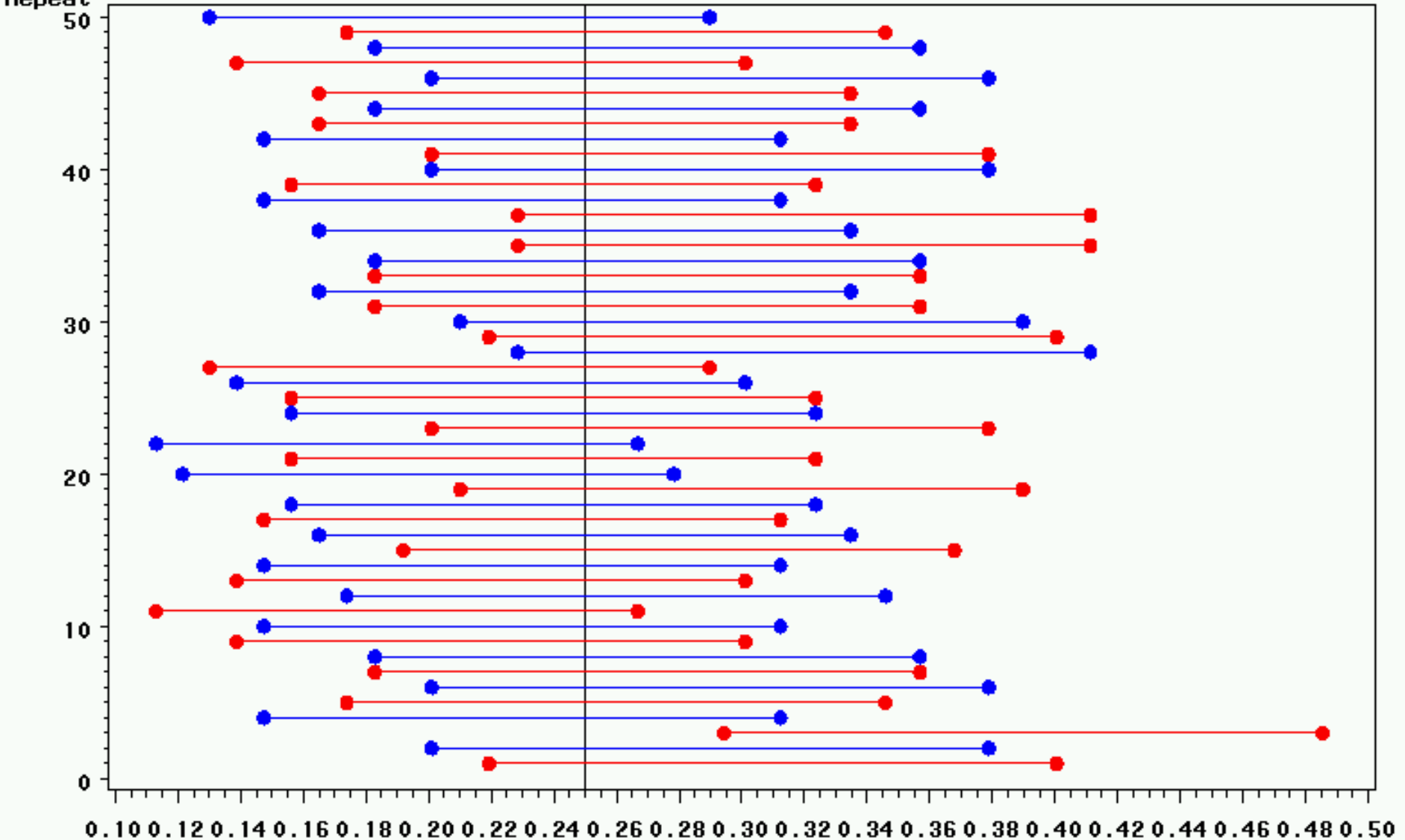


Keep only the confidence intervals that missed the true mean



Parent population=Bernoulli (.25); n=100

Use sample p



Limits= $\pm 1.96 \times \text{Estimated SE}$,

Repeat

Single miss=2% error
rate

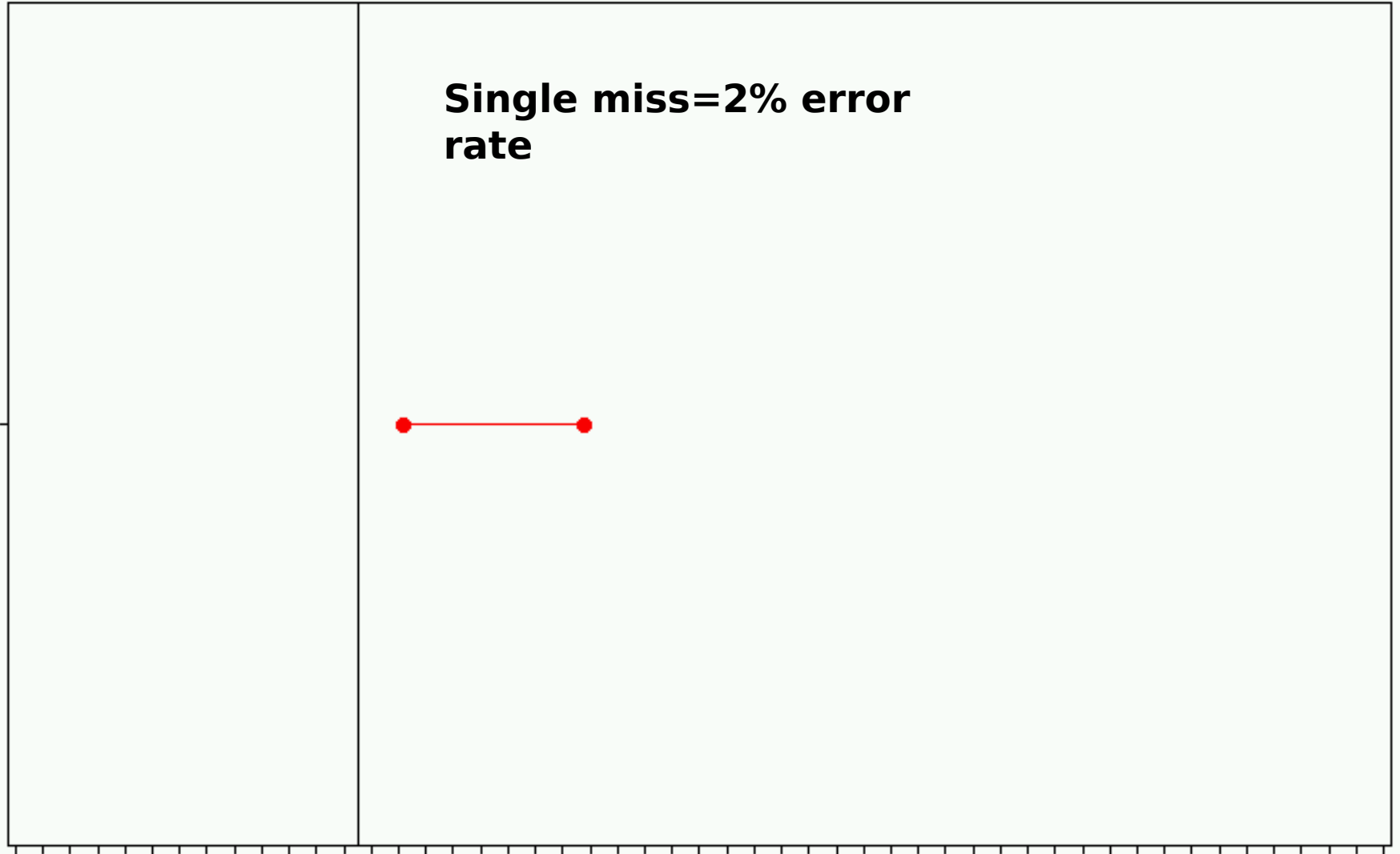
3

0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0

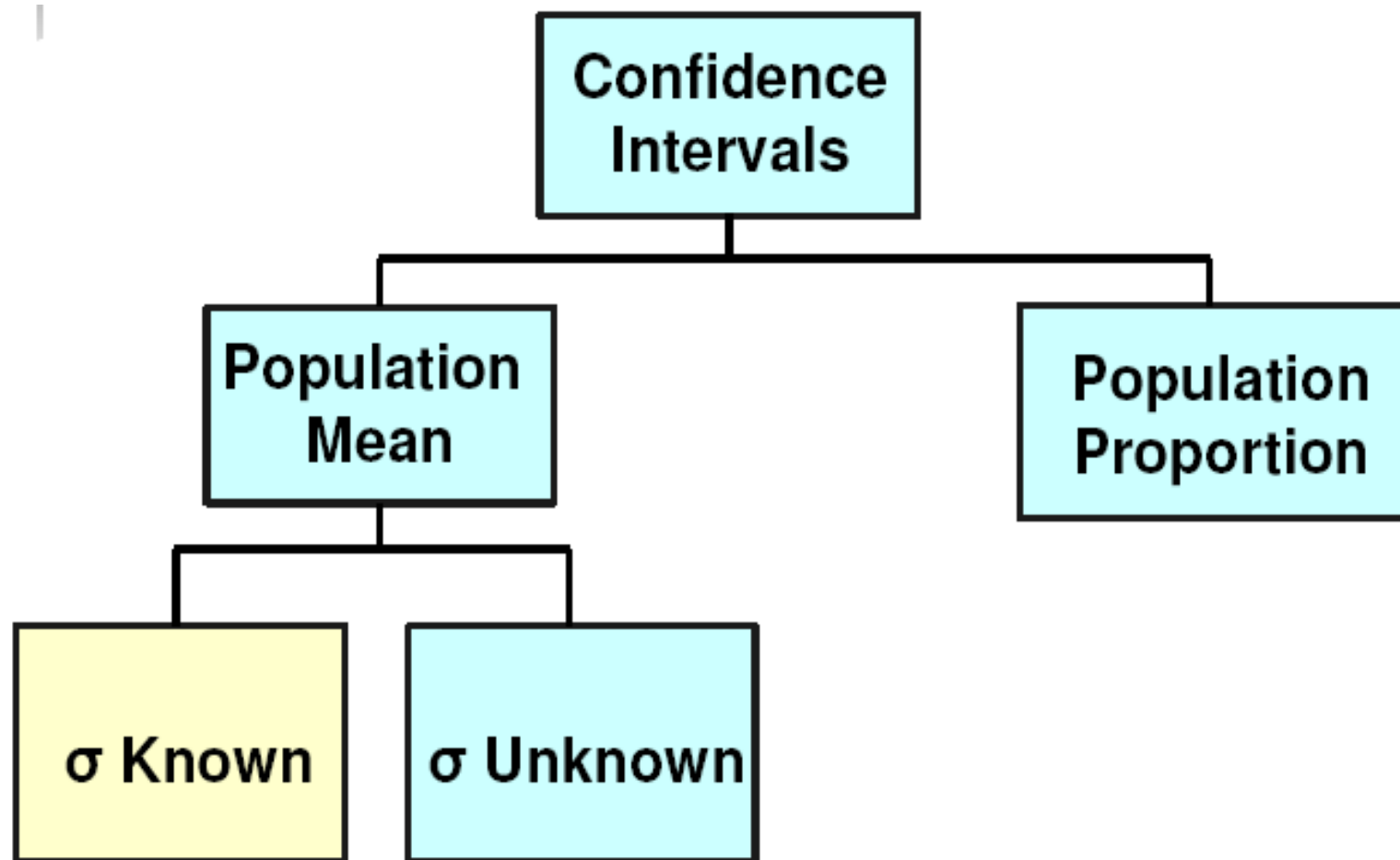
True Proportion

01/14/26

confidence intervals using Z when sigma unknown, n=200, simulation run 50 times



Estimation for Single Population



1. CI for a Single Population Mean (normally distributed)

A. Known variance (large sample size)

- There are 3 elements to a CI:
 1. Point estimate
 2. SE of the point estimate
 3. Confidence coefficient
- Consider the task of computing a CI estimate of μ for a population distribution that is normal with σ *known*.
- Available are data from a random sample of size = n .

■ Assumptions

- Population standard deviation (σ) is known
- Population is normally distributed
- If population is not normal, use large sample

- A 100(

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

s:

- α is to be chosen by the researcher, most common values of α are 0.05, 0.01 and 0.1.

1. The Point Estimate of μ is the Sample Mean $\bar{X}$

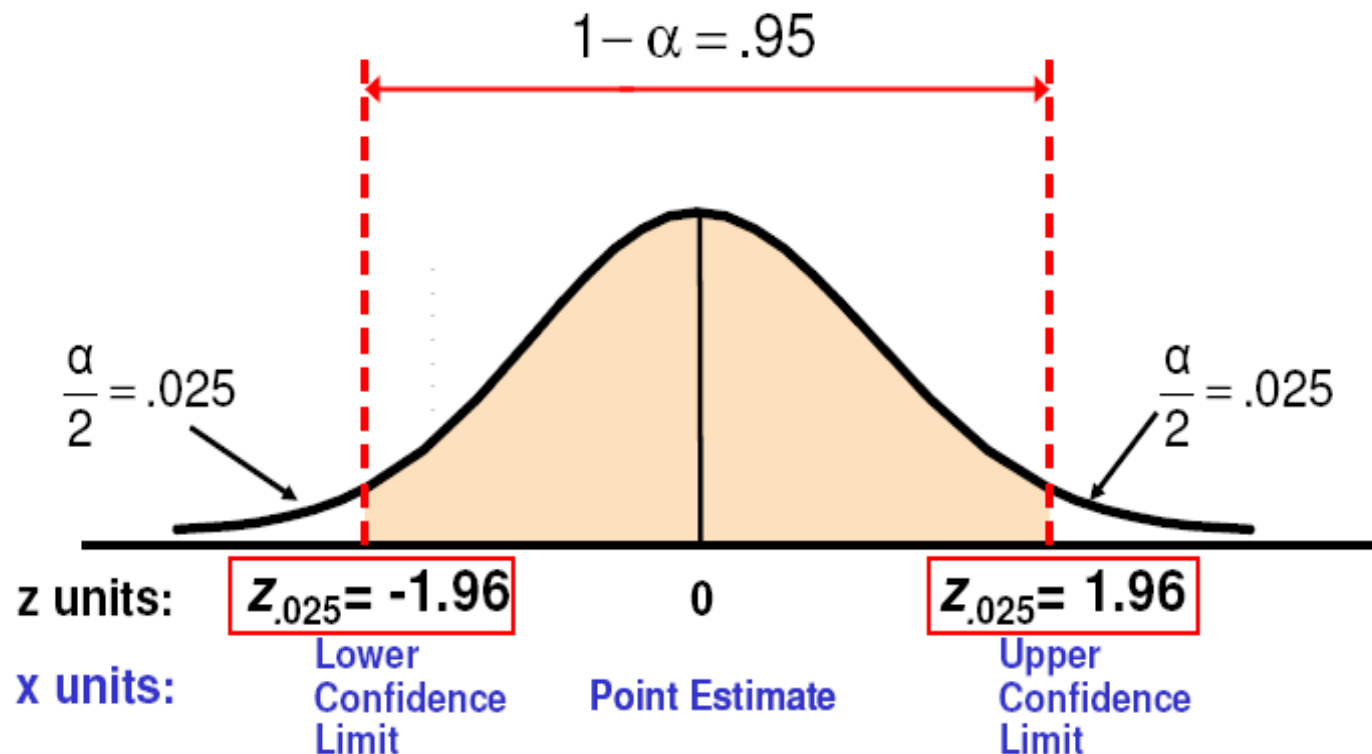
2. The Standard Error of $\bar{X}_n$ is $\sigma/\sqrt{n}$

3. Commonly used CLs are 90%, 95%, and 99%

Confidence Level	Confidence Coefficient, $1 - \alpha$	z value, $Z_{\alpha/2}$
80%	.80	1.28
90%	.90	1.645
95%	.95	1.96
98%	.98	2.33
99%	.99	2.57
99.8%	.998	3.08
99.9%	.999	3.27

Finding the Critical Value

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

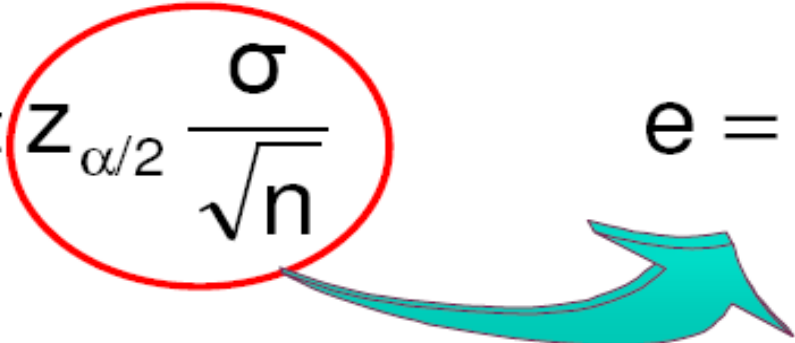


Margin of Error

(Precision of the estimate)

- **Margin of Error (e):** the amount added and subtracted to the point estimate to form the confidence interval

Example: Margin of error for estimating μ , σ known:

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

$$e = z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

Factors Affecting Margin of Error

$$e = z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

- The CI for mean or margin of error is determined by n , s , and α .
- ***As n increases, the CI decreases.***
- ***As s increases, the length of CI increases.***
- ***As the confidence level increases (α decreases), the length of CI increases.***

Example:

1. Waiting times (in hours) at a particular hospital are believed to be approximately normally distributed with a variance of 2.25 hr.
 - a. A sample of 20 outpatients revealed a mean waiting time of 1.52 hours. Construct the 95% CI for the estimate of the population mean.
 - b. Suppose that the mean of 1.52 hours had resulted from a sample of 32 patients. Find the 95% CI.
 - c. What effect does larger sample size have on the CI?

a.

$$1.52 \pm 1.96 \frac{\sqrt{2.25}}{\sqrt{20}} = 1.52 \pm 1.96(.33) \\ = 1.52 \pm .65 = (.87, 2.17)$$

- We are 95% confident that the true mean waiting time is between 0.87 and 2.17 hrs.
- Although the true mean may or may not be in this interval, 95% of the intervals formed in this manner will contain the true mean.
- An incorrect interpretation is that there is 95% probability that this interval contains the true population mean.

b.

$$1.52 \pm 1.96 \frac{\sqrt{2.25}}{\sqrt{32}} = 1.52 \pm 1.96(.27) \\ = 1.52 \pm .53 = (.99, 2.05)$$

c. The larger the sample size makes the CI narrower (more precision).

- When constructing CIs, it has been assumed that the standard deviation of the underlying population, σ , is known
- What if σ is not known?
- In practice, if the population mean μ is unknown, then the standard deviation, σ , is probably unknown as well.

- In this case, the SE of the population can be replaced by the SE of the sample if the sample size is large enough ($n > 30$). With large sample size, we assume a normal distribution.
- Example: It was found that a sample of 35 patients were 17.2 minutes late for appointments, on the average, with SD of 8 minutes. What is the 90% CI for μ ? Ans: (15.0, 19.4).
- Since the sample size is fairly large (> 30) and the population SD is unknown, we assume the distribution of sample mean to be normally distributed based on the CLT and the sample SD to replace population σ .

B. Unknown variance

(small sample size, $n \leq 30$)

- What if the σ for the underlying population is unknown and the sample size is small?
- As an alternative we use *Student's t distributic*

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

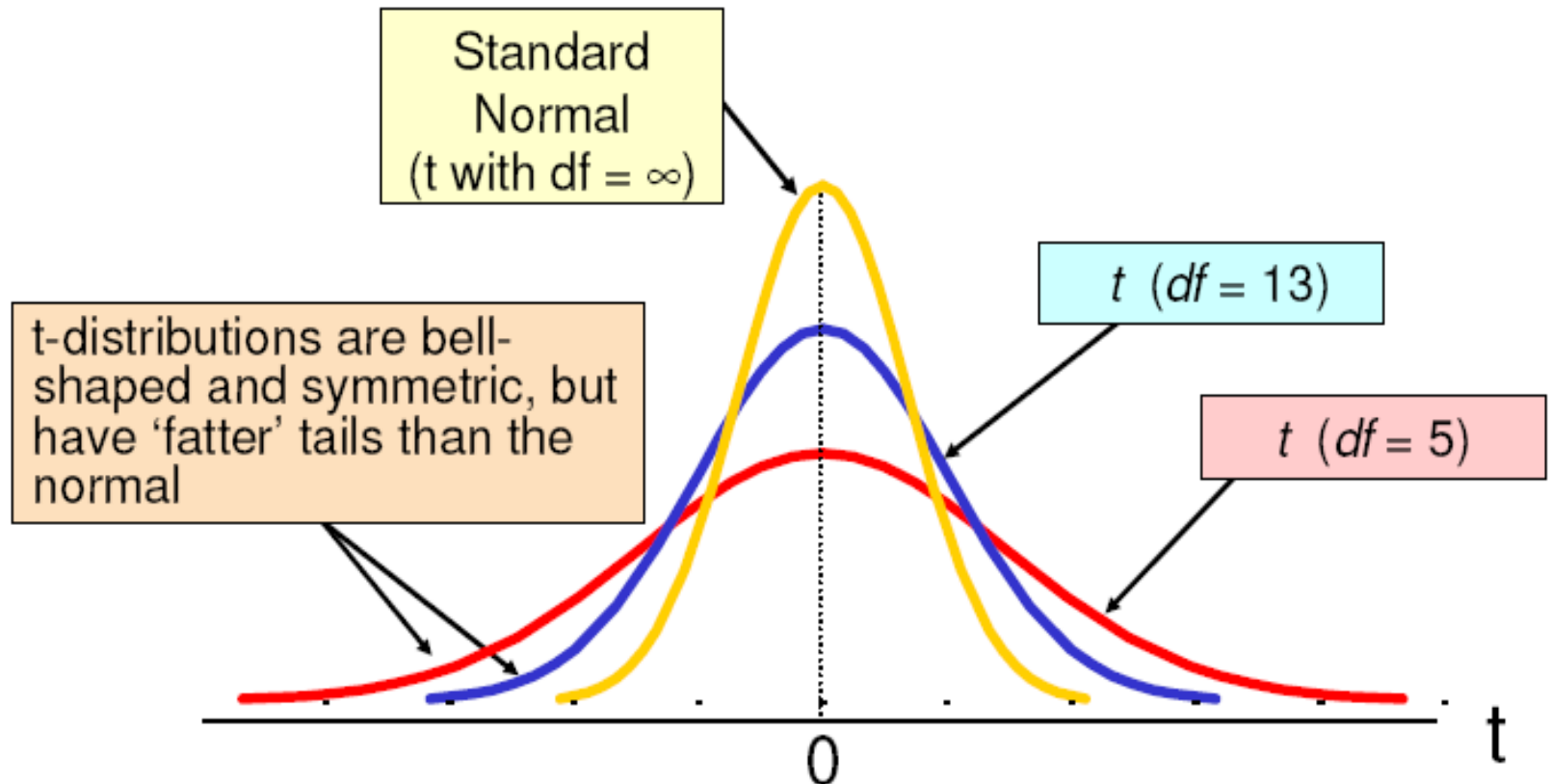
t - score = $t_{DF=n-1} = \frac{\bar{X} - \mu}{s/\sqrt{n}}$ is distributed Student's t with degrees of freedom = (n-1)

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

$$\bar{x} \pm t_{\alpha/2} \frac{s}{\sqrt{n}}$$

Student's t Distribution

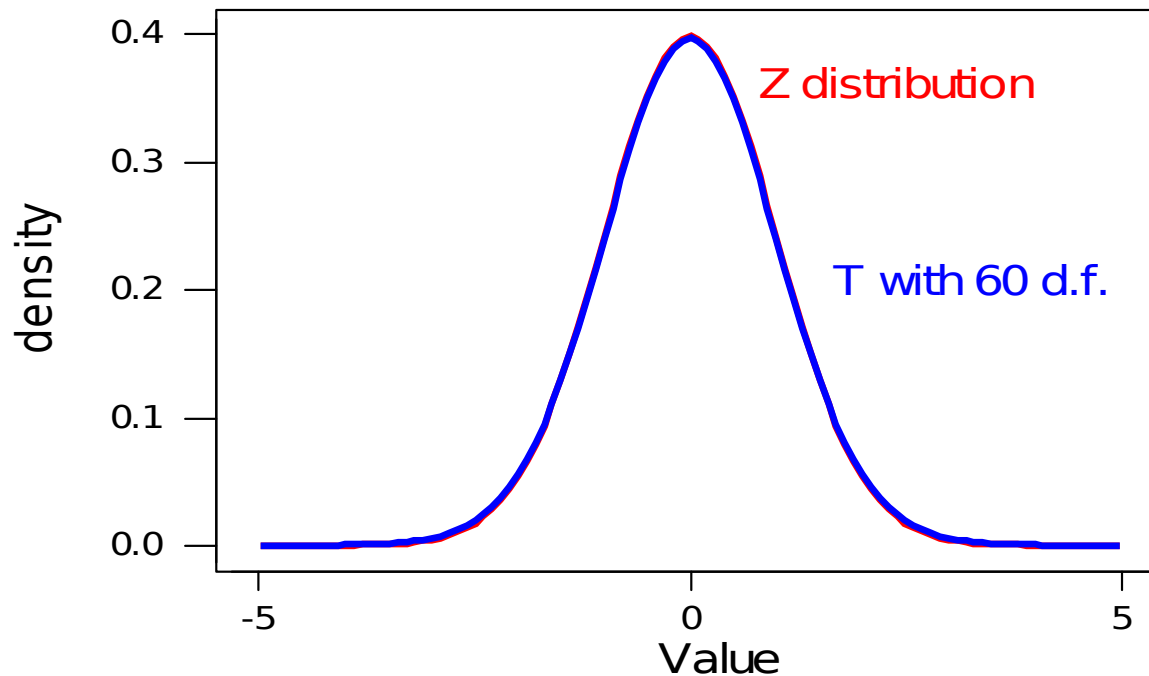
- The t is a family of distributions
- Bell Shaped
- Symmetric about zero (the mean)
- Flatter than the Normal (0,1). This means
 - The variability of a t is greater than that of a Z that is normal(0,1)
 - Thus, there is more area under the tails and less at center
 - Because variability is greater, resulting confidence intervals will be wider.



- Note: t approaches z as n increases

What happens as sample gets larger?

T-distribution and Standard Normal Z distribution



As the df gets larger, the student's t-distribution looks more and more like the SND with mean=0 and variance=1.

What happens to CI as sample gets larger?

$$\bar{X} \pm Z \left(\frac{s}{\sqrt{n}} \right)$$

For large samples:

$$\bar{X} \pm t \left(\frac{s}{\sqrt{n}} \right)$$

Z and t values become almost identical, so CIs are almost identical.

Degrees of Freedom (df)

df = Number of observations that are free to vary after

sample mean has been calculated

Example: Suppose the mean of 3 numbers is 8.0

Let $x_1 = 7$

Let $x_2 = 8$

What is x_3 ?



If the mean of these three values is 8.0, then x_3 **must be 9** (i.e., x_3 is not free to vary)

Here, $n = 3$, so degrees of freedom = $n - 1 = 3 - 1 = 2$

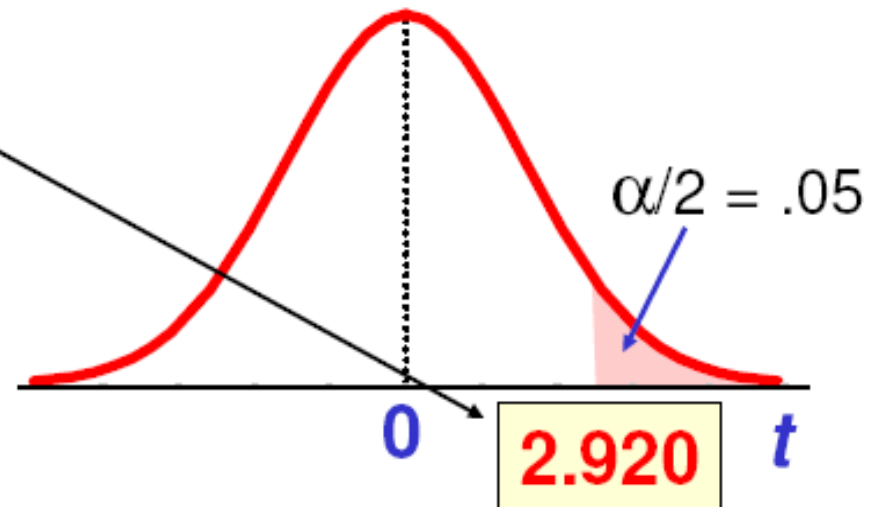
(2 values can be any numbers, but the third is not free to vary for a given mean)

Student's t Table

Upper Tail Area			
df	.25	.10	.05
1	1.000	3.078	6.314
2	0.817	1.886	2.920
3	0.765	1.638	2.353

Let: $n = 3$
 $df = n - 1 = 2$
 $\alpha = .10$
 $\alpha/2 = .05$

The body of the table contains t values, not probabilities



t distribution values

- With comparison to the Z value

<u>Confidence Level</u>	<u>t</u> <u>(10 d.f.)</u>	<u>t</u> <u>(20 d.f.)</u>	<u>t</u> <u>(30 d.f.)</u>	<u>z</u>
.80	1.372	1.325	1.310	1.28
.90	1.812	1.725	1.697	1.64
.95	2.228	2.086	2.042	1.96
.99	3.169	2.845	2.750	2.57

Note: $t \rightarrow z$ as n increases

Example

A random sample of size $n=20$ durations (minutes) of cardiac bypass surgeries has a mean duration of $\bar{X} = 267$ minutes, and variance $s^2 = 36,700$ minutes². Assuming the underlying distribution is normal with unknown variance, construct a 90% CI estimate of the unknown true mean, μ .

- Standard error $= \frac{S}{\sqrt{n}} = \frac{\sqrt{36,700}}{\sqrt{20}} = 42.7$ minutes
- t-value at 90% CL at 19 df = 1.729

Putting this all together –

$$\begin{aligned}\text{Lower limit} &= (\text{point estimate}) - (\text{conf coeff.}) (\text{SE of point estimate}) \\ &= 267 - (1.729)(42.7) \\ &= 193.17\end{aligned}$$

$$\begin{aligned}\text{Upper limit} &= (\text{point estimate}) + (\text{conf coeff}) (\text{SE of point estimate}) \\ &= 267 + (1.729)(42.7) \\ &= 340.83\end{aligned}$$

Thus, a 90% confidence interval for the true mean duration of surgery is (193.2, 340.8) minutes.

Exercise

- Compute a 95% CI for the mean birth weight based on $n = 10$, sample mean = 116.9 oz and $s = 21.70$.
- From the t Table, $t_{9, 0.975} = 2.262$
- Ans: (101.4, 132.4)

2. CIs for single population proportion, p

- An interval estimate for the population proportion (p) can be calculated by adding an allowance for uncertainty to the sample proportion ($\bar{p}$)
- Is based on three elements of CI.
 - **Point estimate**
 - **SE of point estimate**
 - **Confidence coefficient**

- Recall that the distribution of the sample proportion is approximately normal if the sample size is large, with standard deviation

$$\sigma_p = \sqrt{\frac{p(1-p)}{n}}$$

- We will estimate this with sample data:

$$s_{\bar{p}} = \sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$$

- Upper and lower confidence limits for the population proportion are calculated with the formula

$$\bar{p} \pm z_{\alpha/2} \sqrt{\frac{\bar{p}(1 - \bar{p})}{n}}$$

- where
 - z is the standard normal value for the level of confidence desired
 - $\bar{p}$ is the sample proportion
 - n is the sample size

Lower limit = Point Estimate - (Critical Value) x (Standard Error of Estimate)

Upper limit = Point Estimate + (Critical Value) x (Standard Error of Estimate)

Hence,

$$\left(\hat{p} - 1.96 \sqrt{\frac{\hat{p} \hat{q}}{n}}, \hat{p} + 1.96 \sqrt{\frac{\hat{p} \hat{q}}{n}} \right)$$

is an approximate 95% CI for the true proportion p .

Example 1

- A random sample of 100 people shows that 25 are left-handed. Form a 95% CI for the true proportion of left-

1. $\bar{p} = 25/100 = .25$

2. $S_{\bar{p}} = \sqrt{\bar{p}(1-\bar{p})/n} = \sqrt{.25(.75)/n} = .0433$

3. $.25 \pm 1.96 (.0433)$

$$0.1651 \dots 0.3349$$

Interpretation

- We are 95% confident that the true percentage of left-handers in the population is between
16.51% and 33.49%.
- Although this range may or may not contain the true proportion, 95% of intervals formed from samples of size 100 in this manner will contain the true proportion.

Changing the sample size

- Increases in the sample size reduce the width of the confidence interval.

Example:

- If the sample size in the above example is doubled to 200, and if 50 are left-handed in the sample, then the interval is still centered at .25, but the width shrinks to

.19 31

Example 2

- It was found that 28.1% of 153 cervical-cancer cases had never had a Pap smear prior to the time of case's diagnosis. Calculate a 95% CI for the percentage of cervical-cancer cases who never had a Pap test.

A 95% CI is given by

- $$\begin{aligned}\hat{p} \pm 1.96\sqrt{\frac{\hat{p}\hat{q}}{n}} &= .281 \pm 1.96\sqrt{\frac{.281(.719)}{153}} \\ &= .281 \pm .071 = (.210, .352).\end{aligned}$$

Example 3

- Suppose that among 10,000 female operating-room nurses, 60 women have developed breast cancer over five years. Find the 95% CI for p based on point estimate.

Solution

Point estimate = $60/10,000 = 0.006$

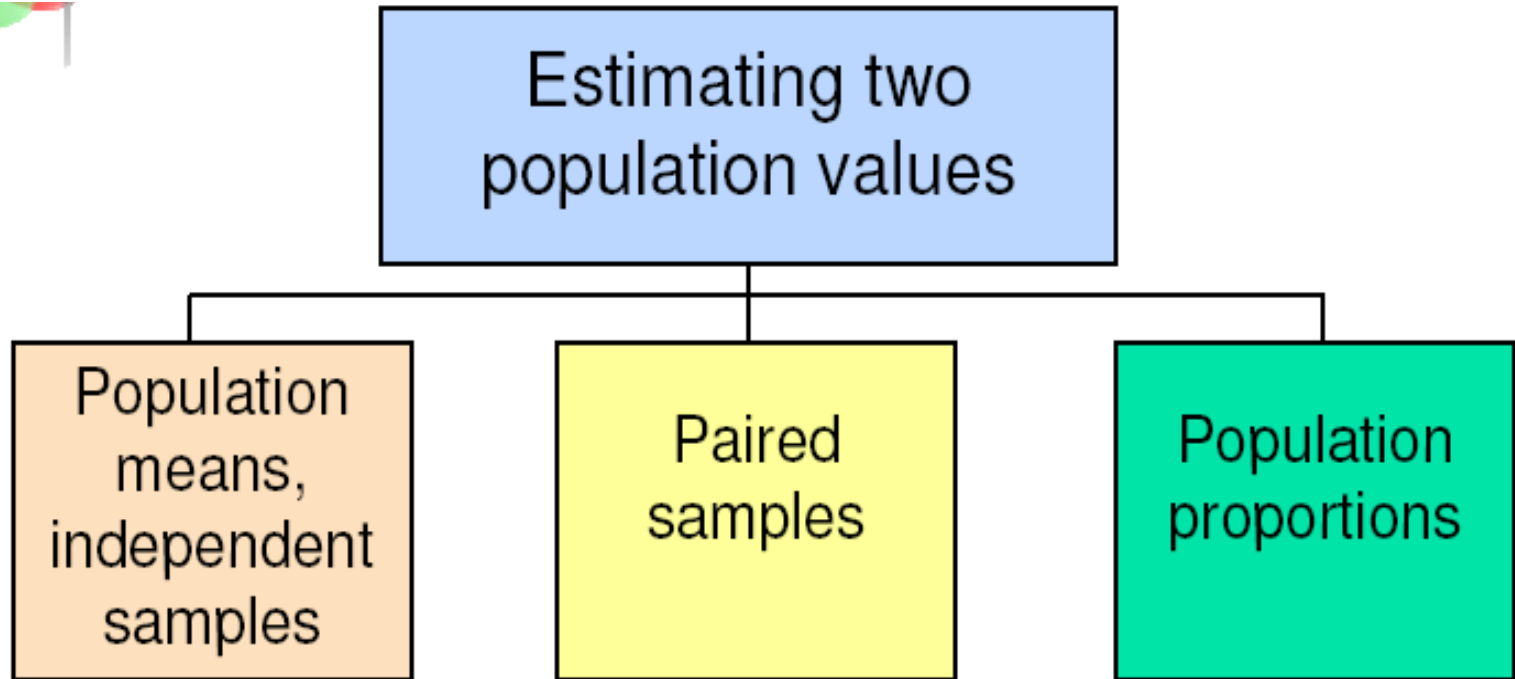
The 95% CI for p is given by the interval:

The 95% CI for p is:

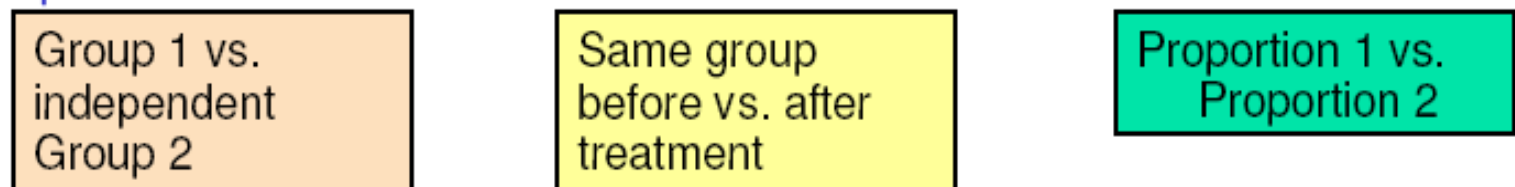
$$\left(\hat{p} - 1.96 \sqrt{\frac{\hat{p} \hat{q}}{n}}, \hat{p} + 1.96 \sqrt{\frac{\hat{p} \hat{q}}{n}} \right)$$

$$\begin{aligned} \left(.006 - 1.96 \sqrt{\frac{.006(.994)}{10,000}}, .006 + 1.96 \sqrt{\frac{.006(.994)}{10,000}} \right) &= (.006 - .0015, .006 + .0015) \\ &= (.0045, .0075) \end{aligned}$$

Estimation for Two Populations



Examples:



3. CI for the difference between population means (normally distributed)

A. Known variances (2 independent samples)

- When σ_1 and σ_2 are known and both populations are normal or both sample sizes are at least 30 the test statistic is a z-
...and the standard error of $\bar{X}_1 - \bar{X}_2$ is

$$\sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

The point estimate for the difference is

$$\bar{X}_1 - \bar{X}_2$$

The confidence interval for $\mu_1 - \mu_2$ is:

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

Assumptions

- Samples are randomly and independently drawn
- Population distributions are normal or both sample sizes are ≥ 30
- Population standard deviations are known

Illustration

- A researcher performs a drug trial involving two independent groups.
 - A **control** group is treated with a placebo while, separately; the **intervention** group is treated with an active agent.
 - Interest is in a comparison of the **mean control** response with the **mean intervention** response under the assumption that the responses are independent.

Examples

- We are interested in the similarity of the two groups.
 - 1) *Is mean blood pressure the same for males and females?*
 - 2) *Is body mass index (BMI) similar for breast cancer cases versus non-cancer patients?*
 - 3) *Is length of stay (LOS) for patients in*
 - Thus, evidence of similarity of the two groups is reflected in a difference between means that is “near” zero.
 - Focus is on $[\mu_{\text{Group 1}} - \mu_{\text{Group 2}}]$

Example

- Researchers are interested in the difference between serum uric acid levels in patients with and without Down's syndrome.
- Patients without Down's syndrome
 - $n=12$, sample mean=4.5 mg/100ml, $\sigma^2=1.0$
- Patients with Down's syndrome
 - $n=15$, sample mean=3.4 mg/100ml, $\sigma^2=1.5$
- Calculate the 95% CI.
- $SE = 0.43$, $95\% \text{ CI} = 1.1 \pm 1.96 (0.43) = (0.26, 1.94)$
- WE are 95% confident that the true difference between the two population means is between 0.26 and 1.94.

B. Unknown variances (Independent samples)

I. Population variances equal (large sample)

- **Assumptions:**
 - **Samples are randomly and independently drawn**
 - **Both sample sizes are ≥ 30**
 - **Population standard deviations are unknown**

Forming confidence estimates:

- Use sample standard deviation s to estimate σ , and the test statistic is a z-value

The confidence interval for
 $\mu_1 - \mu_2$ is:

$$\left(\bar{x}_1 - \bar{x}_2 \right) \pm z_{\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Example

- The mean CD4+ cells for 112 men with HIV infection was 401.8 with a SD of 226.4. For 75 men without HIV, the mean and SD were 828.2 and 274.9, respectively. Calculate a 99% CI for the difference between population means.
- SE of the difference b/n two means = 38.28
- 99% CI = $426.4 \pm 2.58 (38.28)$
= (327.6, 525.2)

II. Population variances equal (small sample)

- Assumptions:
 - Populations are normally distributed
 - *The populations have equal variances*
 - *Samples are independent*
 - *Both sample sizes are <30*
 - *Population standard deviations are unknown*
- * If $0.5 \leq s_1^2/s_2^2 \leq 2$ then we assume that the population variances are equal.**

Forming confidence estimates:

- The population variances are assumed equal, so use the two sample standard deviations and pool them to estimate σ
- The test statistic is a t value with $(n1 + n2 - 2)$ degrees of freedom
- The pooled estimate (s^2_p) is the weighted average of the two sample variances.

- The pooled standard deviation is :

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

- The standard error of the estimate is given by:

$$s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

The confidence interval for
 $\mu_1 - \mu_2$ is:

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

Where $t_{\alpha/2}$ has $(n_1 + n_2 - 2)$ d.f.,

and

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

Example 1

- A study was conducted to compare the serum iron levels of children with cystic fibrosis to those of healthy children. Serum iron levels were measured for random samples of $n_1 = 9$ healthy children and $n_2 = 13$ children with cystic fibrosis.

For the 9 healthy children, $\bar{x}_1 = 18.9 \mu\text{mol/l}$ and $s_1 = 5.9 \mu\text{mol/l}$

For the 13 children with cystic fibrosis, $\bar{x}_2 = 11.9 \mu\text{mol/l}$ and $s_2 = 6.3 \mu\text{mol/l}$

- The two underlying populations of serum iron levels are independent and normally distributed.

For now, assume that $\sigma_1^2 = \sigma_2^2 = \sigma^2$

$$\begin{aligned}s_p^2 &= \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \\&= \frac{(9 - 1)(5.9)^2 + (13 - 1)(6.3)^2}{9 + 13 - 2} \\&= 37.74\end{aligned}$$

The difference in sample means $\bar{x}_1 - \bar{x}_2 = 7.0$ can be used as a point estimate for the true difference in population means $\mu_1 - \mu_2$

We could also construct a confidence interval for $\mu_1 - \mu_2$

A t-value at 95% CL with 20 df is 2.086

$$(\bar{x}_1 - \bar{x}_2) \pm 2.086 \sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}$$

or

$$(18.9 - 11.9) \pm 2.086 \sqrt{37.74 \left(\frac{1}{9} + \frac{1}{13} \right)}$$

or

$$(1.4, 12.6)$$

is a 95% confidence interval for $\mu_1 - \mu_2$

Example 2

- Birth weights of children born to 14 heavy smokers (group 1) and to 15 non-smokers (group 2) were sampled from live births at a large teaching hospital. For the heavy smokers, sample mean = 3.17 kg, SD = 0.46 and for non-smokers, sample mean = 3.63 kg and SD = 0.36.
- $Sp = 0.4121$, $SE = 0.1531$, t-value at 27 df = 2.05
- 95% CI = (0.14, 0.77)

III. Population variances unequal (small sample)

- **The confidence interval for $\mu_1 - \mu_2$ is:**

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2, d'} \sqrt{\left(\frac{s_1^2}{n_1}\right) + \left(\frac{s_2^2}{n_2}\right)}$$

- Where the degree of freedom (d') is given by:

$$d' = \frac{\left[(s_1^2/n_1) + (s_2^2/n_2) \right]^2}{\left[\frac{(s_1^2/n_1)^2}{(n_1-1)} + \frac{(s_2^2/n_2)^2}{(n_2-1)} \right]}$$

Round d' down to the nearest integer — this is d''

Example

For the tuberculosis meningitis example, a random sample of $n_1 = 37$ HIV infected patients has mean age at diagnosis $\bar{x}_1 = 27.9$ years and standard deviation $s_1 = 5.6$ years

A sample of $n_2 = 19$ uninfected patients has mean age at diagnosis $\bar{x}_2 = 38.8$ years and standard deviation $s_2 = 21.7$ years

$$\begin{aligned}
 d' &= \frac{\left[(s_1^2/n_1) + (s_2^2/n_2)\right]^2}{\left[\frac{(s_1^2/n_1)^2}{(n_1-1)} + \frac{(s_2^2/n_2)^2}{(n_2-1)}\right]} \\
 &= \frac{\left[(5.6^2/37) + (21.7^2/19)\right]^2}{\left[\frac{(5.6^2/37)^2}{(37-1)} + \frac{(21.7^2/19)^2}{(19-1)}\right]} \\
 &= 19.24
 \end{aligned}$$

and

$$d'' = 19$$

- For a t distribution with 19 df at 95% CL t -value is 2.093.
- Therefore, a 95% confidence interval would take the form

$$(\bar{x}_1 - \bar{x}_2) \pm 2.093 \sqrt{\left(\frac{s_1^2}{n_1}\right) + \left(\frac{s_2^2}{n_2}\right)}$$

- Using the data from two samples of patients with tuberculosis meningitis, the 95% CI for $\mu_1 - \mu_2$ is

$$(27.9 - 38.8) \pm 2.093 \sqrt{\left(\frac{5.6^2}{37}\right) + \left(\frac{21.7^2}{19}\right)}$$

or

$$(-21.5, -0.3)$$

C. Paired Samples

- Tests Means of 2 **Related** Populations
 - △ Paired or matched samples
 - △ Repeated measures (before/after)
 - △ Use **difference** between paired values:
$$\mathbf{d} = \mathbf{x_1 - x_2}$$
- Eliminates variation among subjects
- Assumptions:
 - Both populations are normally distributed,
 - Or, if not normal, use large samples.

Paired Data

- Paired data arises when each individual (more specifically, each unit of measurement) in a sample is measured twice.
- Measurement might be "pre/post", "before/after", "right/left", "parent/child", etc.

Examples of paired data

- 1) Blood pressure prior to and following treatment,
 - 2) Number of cigarettes smoked per week measured prior to and following participation in a smoking cessation program,
 - 3) Number of sex partners in the month prior to and in the month following an HIV education campaign.
- Notice in each of these examples that the two occasions of measurement are linked by virtue of the two measurements being made on the same individual.
 - Longitudinal or follow-up study

Paired differences

- If two measurements of the same phenomenon (eg. blood pressure, # cigarettes/week, etc) X and Y are measured on an individual and if each is normally distributed, then their difference is also distributed normal.
- The interest in the difference between two measurements

The i^{th} paired difference is d_i , where

$$d_i = x_{1i} - x_{2i}$$

The point estimate for the population mean paired difference is $\bar{d}$:

$$\bar{d} = \frac{\sum_{i=1}^n d_i}{n}$$

The sample standard deviation is

$$s_d = \sqrt{\frac{\sum_{i=1}^n (d_i - \bar{d})^2}{n-1}}$$

n is the number of pairs in the paired sample

The confidence interval for $\bar{d}$ is

$$\bar{d} \pm t_{\alpha/2} \frac{s_d}{\sqrt{n}}$$

- Where $t_{\alpha/2}$ has $n-1$ df.

Example

- Ten hypertensive patients are screened at a neighborhood health clinic and are given methyl dopa, a strong antihypertensive medication for their condition.
- They are asked to come back 1 week later and have their blood pressures measured again.
- Suppose the initial and follow-up SBPs (mm Hg) of the patients are given below.

Patient number	Initial SBP	Follow-up SBP
1	200.0	188.0
2	194.0	212.0
3	236.0	186.0
4	163.0	150.0
5	240.0	200.0
6	225.0	222.0
7	203.0	190.0
8	180.0	154.0
9	177.0	180.0
10	240.0	225.0

1. What is the mean and s_d of the difference?
2. What is the standard error of the mean?
3. Assume that the difference is normally distributed, construct a 95% CI for μ .

Answer

- We have the following data and summary statistics

Patient number	$d_i(I - F)$	d_i^2
1	12.0	144
2	-18.0	324
3	50.0	2500
4	13.0	169
5	40.0	1600
6	3.0	9
7	13.0	169
8	26.0	676
9	-3.0	9
10	15.0	225
Total	151.0	5825

$$\bar{d} = \frac{\sum d_i}{n} = \frac{151}{10} = 15.1$$

$$s = \sqrt{\frac{\sum (d_i - \bar{d})^2}{n-1}} = \sqrt{\frac{3544.9}{9}} = \sqrt{393.878} = 19.85$$

$$se = \frac{s}{\sqrt{n}} = \frac{19.85}{\sqrt{10}} = 6.28$$

A 95% CI for μ is provided by

$$\begin{aligned} & \left(\bar{d} - \frac{t_{9, .975} s}{\sqrt{n}}, \bar{d} + \frac{t_{9, .975} s}{\sqrt{n}} \right) \\ &= [15.1 - 2.262(6.28), 15.1 + 2.262(6.28)] \\ &= (15.1 - 14.2, 15.1 + 14.2) = (0.9, 29.3). \end{aligned}$$

4. Two Population Proportions

- We are often interested in comparing proportions from 2 populations:
 - Is the incidence of disease A the same in two populations?
 - Patients are treated with either drug D, or with placebo. Is the proportion “improved” the same in both groups?

Goal: Form a confidence interval for or test a hypothesis about the difference between two population proportions, $p_1 - p_2$

Assumptions:

$$n_1 p_1 \geq 5 \quad , \quad n_1 (1 - p_1) \geq 5$$

$$n_2 p_2 \geq 5 \quad , \quad n_2 (1 - p_2) \geq 5$$

The point estimate for the difference is $\bar{p}_1 - \bar{p}_2$

Confidence Interval for Two Population Proportions

- SE of the difference =

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

- The confidence interval for $p_1 - p_2$ is:

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

The following formula is also equally used

- An approximate 95% confidence interval takes the form

$$(\hat{p}_1 - \hat{p}_2) \pm 1.96 \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}$$

Example

- In a clinical trial for a new drug to treat hypertension, $N_1 = 50$ patients were randomly assigned to receive the new drug, and $N_2 = 50$ patients to receive a placebo. 34 of the patients receiving the drug showed improvement, while 15 of those receiving placebo showed improvement.
- Compute a 95% CI estimate for the difference between proportions improved.

- $p_1 = 34/50 = 0.68$, $p_2 = 15/50 = 0.30$
- The point estimate for the difference is:

$$= [0.68 - 0.30] = 0.38$$
- SE of the difference =
$$\sqrt{\frac{.68(.32)}{50} + \frac{.30(.70)}{50}} = 0.0925$$
- 95% CI
 - **Lower = (point estimate) - ($Z_{\alpha/2}$) (SE)**

$$= 0.38 - (1.96)(0.0925) = 0.20$$
 - **Upper = (point estimate) + ($Z_{\alpha/2}$) (SE)**

$$= 0.38 + (1.96)(0.0925) = 0.56$$
- 95% CI = (0.20, 0.56)

CI for the true OR

- The odds ratio is defined as the odds of disease among exposed individuals divided by the odds of disease among the unexposed

$$OR = \frac{p_1/q_1}{p_2/q_2}$$

- It is estimated by:

$$\begin{aligned}\widehat{OR} &= \frac{\hat{p}_1/\hat{q}_1}{\hat{p}_2/\hat{q}_2} \\ &= \frac{[a/(a+b)]/[b/(a+b)]}{[c/(c+d)]/[d/(c+d)]} \\ &= \frac{ad}{bc}\end{aligned}$$

- Upper and lower confidence limits for the natural log of the odds ratio are calculated with the formula:

$$\ln \hat{OR} \pm Z \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$$

- Exponentiate the upper and lower confidence limits for the

natural log of the OR:

$$e^{\ln \hat{OR} - Z \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}}, e^{\ln \hat{OR} + Z \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}}$$

- where
 - Z is the standard normal value for the level of confidence desired
 - a-d are the cells of a 2x2 table

Thank you!