UCLA STAT 10

Introduction to Statistical Reasoning

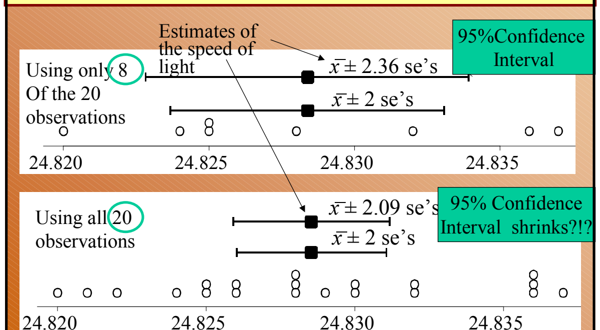
- **Instructor:** Ivo Dinov,
  Asst. Prof. In Statistics and Neurology
- **Teaching Assistants:** Yan Xiong, Will Anderson,
  UCLA Statistics

University of California, Los Angeles, Winter 2002
http://www.stat.ucla.edu/~dinov/

Chapter 21: Confidence Intervals

- Introduction
- Means
- Proportions
- Comparing 2 means
- Comparing 2 proportions
- How big should my study be?

20 replicated measurements to estimate the speed of light. Obtained by Simon Newcomb in 1882, by using distant (3.721 km) rotating mirrors.

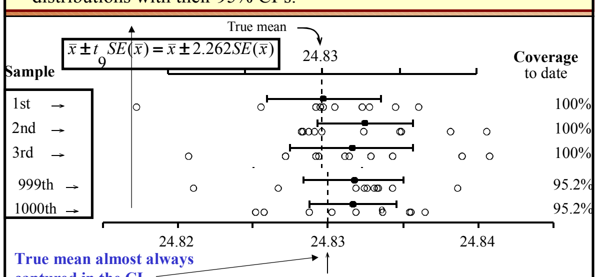


A 95% confidence interval

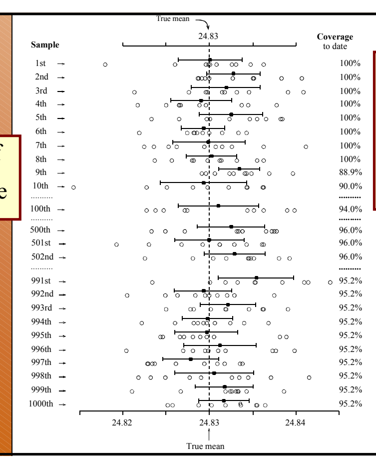
- A type of interval that contains the true value of a parameter for 95% of samples taken is called a ***95% confidence interval*** for that parameter, ends of the CI are called *confidence limits*.
- (For the situations we deal with) a confidence interval (CI) for the true value of a parameter is given by

  estimate $\pm$ z ***standard errors***

- CI are constructed using the sample $\bar{x}$ and s=SE. But diff. samples yield diff. estimates and → diff. CI's?!?
- Below is a computer simulation showing how process of taking samples effects the estimates and the CI's.
- 1000 samples of size 10 obs's from a Normal(m=24.83, s=0.005) distributions with their 95% CI's.



Most of the table



How many of the previous samples contained the true mean?

The Z and the T scores/values

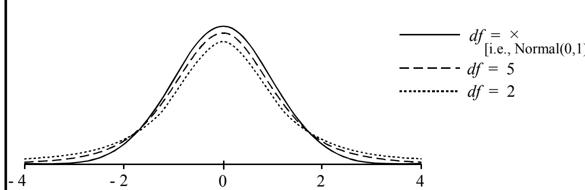
- Remember Z = (X – mean)/SD ~ std. Normal (0, 1)
- We know how to read the Standard Normal (Z) tables
- How about if the SD is unknown?
- We estimate it from the data, using the **sample-SD**
- Then we'd like to do the same we did for normal standardization –

T = (X – mean) / sample_SD

- But is T standard normally distributed? Almost!
- T = (X–mean)/sample_SD ~ Student's T (df=n-1)

The Z and the T scores/values

- T = (X–mean)/sample_SD ~ Student's T (df=n-1)



Summary - CI for population mean

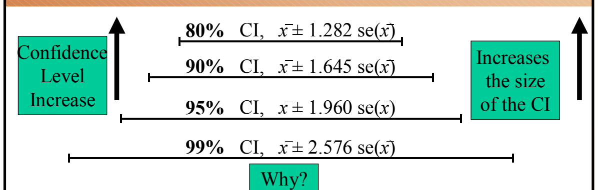
Confidence Interval for the true (population) mean μ:
sample mean $\pm$ ***t standard errors***

or $\bar{x} \pm t\ \text{se}(\bar{x})$, where $\text{se}(\bar{x}) = \frac{s_x}{\sqrt{n}}$ and $df = n-1$

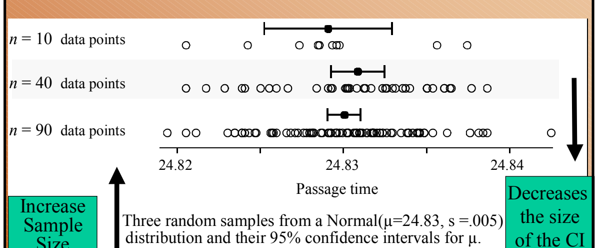
TABLE 8.1.1 Value of the Multiplier, t, for a 95% CI

df :	7	8	9	10	11	12	13	14	15	16	17
t :	2.365	2.306	2.262	2.228	2.201	2.179	2.160	2.145	2.131	2.120	2.110
df :	18	19	20	25	30	35	40	45	50	60	∞
t :	2.101	2.093	2.086	2.060	2.042	2.030	2.021	2.014	2.009	2.000	1.960

Effect of increasing the confidence level



Effect of increasing the sample size



Three random samples from a Normal(μ=24.83, s =.005) distribution and their 95% confidence intervals for μ.

To ***double the precision*** we need ***four times*** as many observations.

Why ↑ in sample-size ↓ CI?

Confidence Interval for the true (population) mean μ:
sample mean $\pm$ ***t standard errors***

or $\bar{x} \pm t\ \text{se}(\bar{x})$, where $\text{se}(\bar{x}) = \frac{s_x}{\sqrt{n}}$ and $df = n-1$

CI for a population proportion

Confidence Interval for the true (population) proportion p:

sample proportion ± z standard errors

or $\hat{p} \pm z\ \mathrm{se}(\hat{p})$, where $\mathrm{se}(\hat{p}) = \sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}$

Example – higher blood thiol concentrations with rheumatoid arthritis

Thiol Concentration (mmol)

	Normal	Rheumatoid
	1.84	2.81
	1.92	4.06
	1.94	3.62
	1.92	3.27
	1.85	3.27
	1.91	3.76
	2.07	
Sample size	7	6
Sample mean	1.92143	3.46500
Sample standard deviation	0.07559	0.44049

Research question: Is the change in the Thiol status in the lysate of packed blood cells substantial to be indicative of a non trivial relationship between Thiol-levels and rheumatoid arthritis?

Example – higher blood thiol concentrations with rheumatoid arthritis

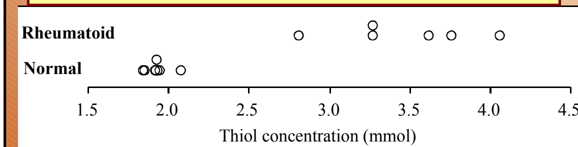


Figure 8.4.1 Dot plot of Thiol concentration data.

Two groups of subjects are studied: 1. NC (normal controls) 2. RA (rheumatoid arthritis).

Observations: 1. The avg. levels of thiol seem diff. in NC & RA 2. NC and RA groups are separated completely.

Question: Is there statistical evidence that thiol-level correlates with the disease?

CI's for difference between means

Confidence Interval for a difference between population means $(\mu_1 - \mu_2)$:

Difference between sample means

± t standard errors of the difference

or $\bar{x}_1 - \bar{x}_2 \pm t\ \mathrm{se}(\bar{x}_1 - \bar{x}_2)$

CI's for difference between proportions

Confidence Interval for a difference between population proportions $(p_1 - p_2)$:

Difference between sample proportions

± z standard errors of the difference

or $\hat{p}_1 - \hat{p}_2 \pm z\ \mathrm{se}(\hat{p}_1 - \hat{p}_2)$

But how do we compute the $SE(\hat{p}_1 - \hat{p}_2)$ for different cases?

Sample size - proportion

- **For a 95% CI,** $\text{margin} = 1.96 \times \sqrt{\hat{p}(1-\hat{p})/n}$
- **Sample size for a desired margin of error:** For a margin of error no greater than m, use a sample size of approximately

$$n = \left(\frac{z}{m}\right)^2 \times p^*(1-p^*)$$

- p^* is a guess at the value of the proportion -- err on the side of being too close to 0.5
- z is the multiplier appropriate for the confidence level
- m is expressed as a proportion (between 0 and 1), not a percentage

Sample size -- mean

- **Sample size for a desired margin of error:**
  For a margin of error no greater than m, use a sample size of approximately

$$n = \left(\frac{z \sigma^*}{m} \right)^2$$

- σ^* is an estimate of the variability of individual observations
- z is the multiplier appropriate for the confidence level

Summary

Confidence intervals

- We construct an interval estimate of a parameter to summarize our level of uncertainty about its true value.
- The uncertainty is a consequence of the sampling variation in point estimates.
- If we use a method that produces intervals which contain the true value of a parameter for 95% of samples taken, the interval we have calculated from our data is called a 95% confidence interval for the parameter.
- Our confidence in the particular interval comes from the fact that the method works 95% of the time (for 95% CI's).

TABLE 8.7.1 Standard Errors and Degrees of Freedom

Parameter		Estimate	Standard error of estimate	df
Mean,	μ	$\bar{x}$	$\frac{s_x}{\sqrt{n}}$	$n-1$
Proportion,	p	$\hat{p}$	$\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$	∞
Difference in means,	$\mu_1-\mu_2$	$\bar{x}_1-\bar{x}_2$	$\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$	Min(n_1-1,n_2-1)
Difference in proportions,	p_1-p_2	$\hat{p}_1-\hat{p}_2$	(see Table 8.5.5)	∞

$df = \infty$ means we use a muitiplier obtained from the Normal(0,1) distribution.
CI's work well when sample sizes are big enough to satisfy the 10% rule in Appendix A3.
Applies to means from independent samples.
df given is a conservative approximation for hand calculation (see Section 10.2).

Summary cont.

- For a great many situations,

  an (approximate) confidence interval is given by

  estimate $\pm$ ***t standard errors***

  The size of the multiplier, t, depends both on the desired confidence level and the degrees of freedom (df).

  [With proportions, we use the Normal distribution (i.e., $df=\infty$) and it is conventional to use z rather than t to denote the multiplier.]
- The *margin of error* is the quantity added to and subtracted from the estimate to construct the interval (i.e. t standard errors).

Summary cont.

- If we want greater confidence that an interval calculated from our data will contain the true value, we have to use a wider interval.
- To double the precision of a 95% confidence interval (i.e.halve the width of the confidence interval), we need to take 4 times as many observations.

CLT Example – CI shrinks by half by quadrupling the sample size!

- If I ask 30 of you the question "Is 5 credit hour a reasonable load for Stat10?", and say, 15 (50%) said ***no***. Should we change the format of the class?
- Not really – the 2SE interval is about **[0.32 ; 0.68].** So, we have little concrete evidence of the proportion of students who think we need a change in Stat 10 format,

$$\hat{p} \pm 2\times \mathrm{SE}(\hat{p}) = 0.5 \pm 2\times\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.5 \pm \text{-}0.18$$

- If I ask all 300 Stat 10 students and 150 say ***no*** (still 50%), then 2SE interval around 50% is: **[0.44 ; 0.56].**
- **So, large sample is much more useful and this is due to CLT effects, without which, we have no clue how useful our estimate actually is …**