UCLA STAT 251

Statistical Methods for the Life and Health Sciences

- **Instructor:** **Ivo Dinov,**

Asst. Prof. In Statistics and Neurology

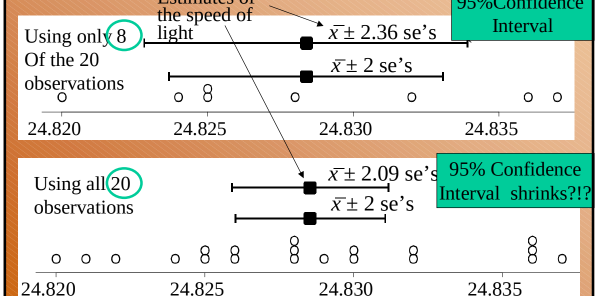
University of California, Los Angeles, Winter 2003

http://www.stat.ucla.edu/~dinov/

Confidence Intervals

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

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

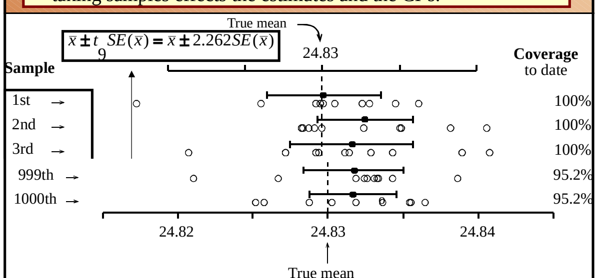


A 95% confidence interval

- 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, the 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

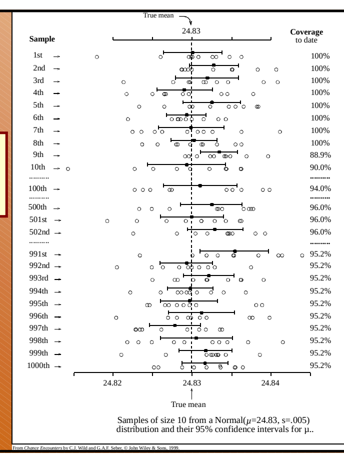
estimate $\pm$ ***t*** ***standard errors***

- 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 the process of taking samples effects the estimates and the CI's.



Samples of size 10 from a Normal(μ=24.83, s=.005) distribution and their 95% confidence intervals for μ..

Most of the table



How many of the previous samples contained the true mean?

Summary - CI for population mean

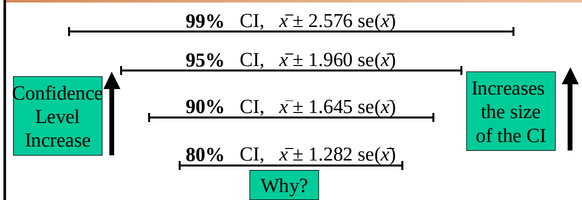
Confidence Interval for the true (population) mean μ:
sample mean* $\pm$ *t standard errors

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

Value of the Multiplier, *t*, for a 95% CI

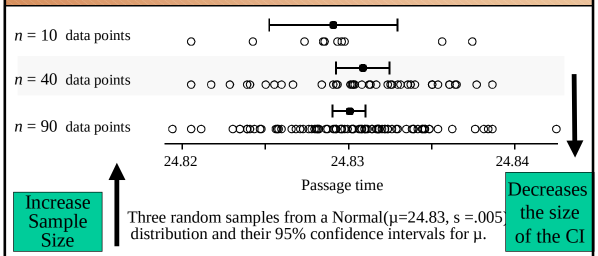
df :	7	8	9	10	11	12	13	14	15	16	17
t :	2.365	2.306	2.262	2.228	2.201	2.179	2.160	2.145	2.131	2.120	2.110
df :	18	19	20	25	30	35	40	45	50	60	∞
t :	2.101	2.093	2.086	2.060	2.042	2.030	2.021	2.014	2.009	2.000	1.960

Effect of increasing the confidence level



The greater the confidence level, the wider the interval

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* $\pm$ *z standard errors

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

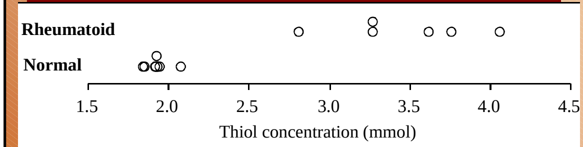
Example – higher blood thiol concentrations associated with rheumatoid arthritis?!?

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



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?

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 $$\bar{x}_1 - \bar{x}_2 \pm t\,\text{se}(\bar{x}_1 - \bar{x}_2)$$

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

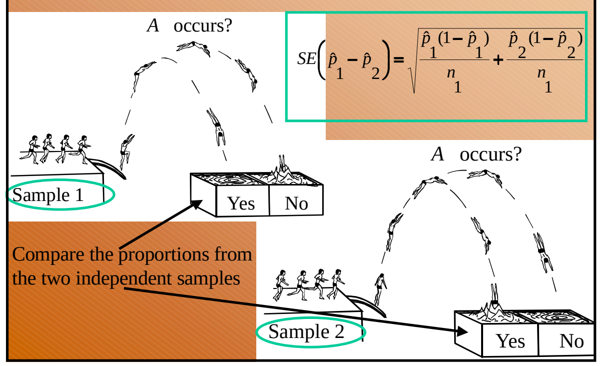
$$\hat{p}_1 - \hat{p}_2 \pm z\,\text{se}(\hat{p}_1 - \hat{p}_2)$$

Big Question ???

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

1. Proportions from 2 independent samples

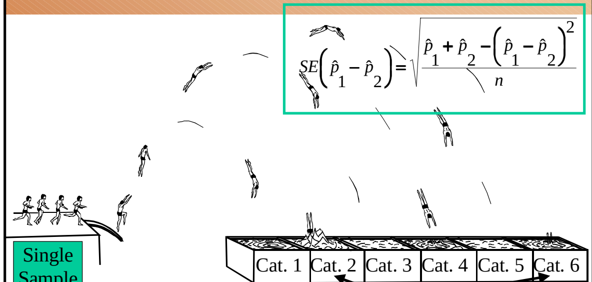
$$SE\left(\hat{p}_1 - \hat{p}_2\right) = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_1}}$$



Compare the proportions from the two independent samples

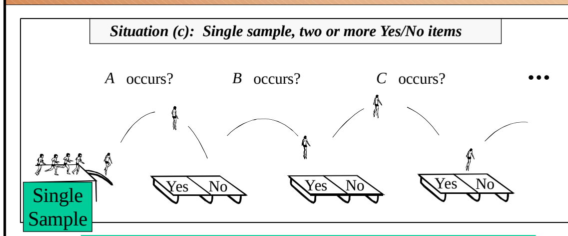
2. Single sample, several response categories

$$SE\left(\hat{p}_1 - \hat{p}_2\right) = \sqrt{\frac{\hat{p}_1 + \hat{p}_2 - \left(\hat{p}_1 - \hat{p}_2\right)^2}{n}}$$



Compare different proportions from the same sample

3. Single sample, several yes/no items



From Chance Encounters by C.J. Wild and G.A.F. Seber, © John Wiley & Sons, 2000.

$$SE\left(\hat{p}_1 - \hat{p}_2\right) = \sqrt{\frac{\text{Min}\left(\hat{p}_1 + \hat{p}_2 ; 2 - \hat{p}_1 - \hat{p}_2\right) - \left(\hat{p}_1 - \hat{p}_2\right)^2}{n}}$$

SE's for the 3 cases of differences in proportion

(a) **Proportions from two independent samples** of sizes n_1 and n_2, respectively

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

(b) **One sample of size n, several response categories**

$$se(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{\hat{p}_1 + \hat{p}_2 - (\hat{p}_1 - \hat{p}_2)^2}{n}}$$

(c) **One sample of size n, many Yes/No items**

$$se(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{Min(\hat{p}_1 + \hat{p}_2, \hat{q}_1 + \hat{q}_2) - (\hat{p}_1 - \hat{p}_2)^2}{n}}$$

where $\hat{q}_1 = 1 - \hat{p}_1$ and $\hat{q}_2 = 1 - \hat{p}_2$

1998 Back-to-school survey – smoking/drinking 2000 students participating in the survey

Characteristics by Smoking and Drinking Status

(Table entry is % of group saying yes)	Smoker	Nonsmoker	Drinker	Nondrinker
Get mostly A's or B's?	41	68		
Read 1 or more hours/day?	54	72	56	75
Get drunk at least once a month?	63	10		
Have smoked Marijuana?	79	14	52	12
Likely to try illegal drug in future?	42	14	35	11
n	130	870	260	740

1 Sample many yes/no Answers -
compare proportions of smokers who get a drink at least once a month(63) With proportion of smokers who have smoked marijuana (79). Smokers group, say, are not forced to choose 1 category out of a set of categories as in the previous situation.

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 (basically, What's n, so that m >= margin?)

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

Covariance – a measure of LINEAR association between two variables, X & Y

$$E(aY_1 + bY_2) = aE(Y_1) + bE(Y_2)$$
$$= a\mu_1 + b\mu_2$$
$$Var(aY_1 + bY_2 + c) =$$
$$a^2Var(Y_1) + b^2Var(Y_2) + 2abCov(Y_1; Y_2)$$

$$Cov(Y_1; Y_2) = E[(Y_1 - \mu_1)(Y_2 - \mu_2)]$$

HypothesisTesting4a.ppt and then:

F_Chi2_dist_Ch4_6.pdf Variance estimates/CI's