# UCLA STAT 13
## Introduction to Statistical Methods for the Life and Health Sciences

● **Instructor:** Ivo Dinov,
   Asst. Prof. of Statistics and Neurology
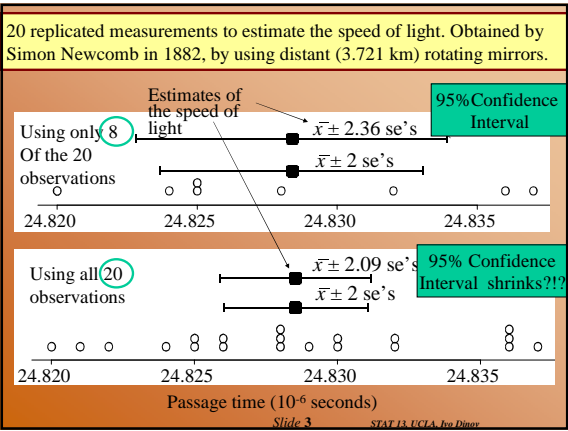
● **Teaching Assistants:** Chris Barr & Ming Zheng

University of California, Los Angeles, Fall 2004
http://www.stat.ucla.edu/~dinov/courses_students.html

---

## Chapter 8: 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.
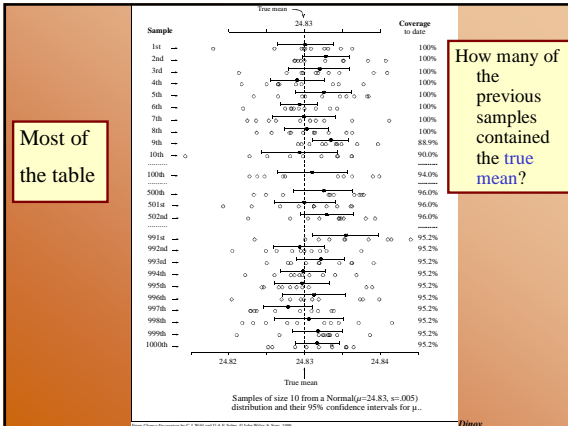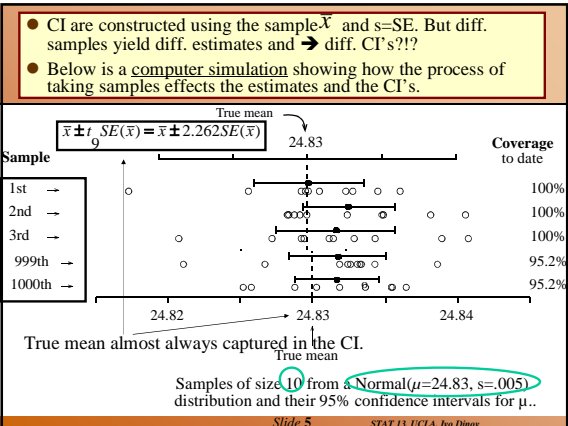


Passage time ($10^{-6}$ seconds)

---

## 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

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

**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 |

---

● 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.



$$\bar{x} \pm t_9 \ SE(\bar{x}) = \bar{x} \pm 2.262 SE(\bar{x})$$

True mean almost always captured in the CI.

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

---

Most of the table



How many of the previous samples contained the true mean?

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

## Summary - CI for population mean

Confidence Interval for the true (population) mean $\mu$:

**sample mean $\pm\, t$ standard errors**

or $\quad \bar{x} \pm t\,\mathrm{se}(\bar{x}),\ $ where $\mathrm{se}(\bar{x}) = \dfrac{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 | $\infty$ |
| $t$ : | 2.101 | 2.093 | 2.086 | 2.060 | 2.042 | 2.030 | 2.021 | 2.014 | 2.009 | 2.000 | 1.960 |

*Slide 7*        STAT 13, UCLA, Ivo Dinov

---

. summarize

| Variable | Obs | Mean | Std. Dev. | Min | Max |
|---|---|---|---|---|---|
| make | 0 | | | | |
| price | 74 | 6165.257 | 2949.496 | 3291 | 15906 |
| mpg | 74 | 21.2973 | 5.785503 | 12 | 41 |
| rep78 | 69 | 3.405797 | .9899323 | 1 | 5 |
| hdroom | 74 | 2.993243 | .8459948 | 1.5 | 5 |
| trunk | 74 | 13.75676 | 4.277404 | 5 | 23 |
| weight | 74 | 3019.459 | 777.1936 | 1760 | 4840 |
| length | 74 | 187.9324 | 22.26634 | 142 | 233 |
| turn | 74 | 39.64865 | 4.399354 | 31 | 51 |
| displ | 74 | 197.2973 | 91.83722 | 79 | 425 |
| gratio | 74 | 3.014865 | .4562871 | 2.19 | 3.89 |
| foreign | 74 | .2972973 | .4601885 | 0 | 1 |

*Slide 8*        STAT 13, UCLA, Ivo Dinov

---

## Output from other statistics packages

**Minitab Output**

**T Confidence Intervals**

```
Variable      N     Mean    StDev   SE Mean      95.0 % CI
Passage Times 20  24.8286  0.0051   0.0011  ( 24.8262, 24.8309)
```

*Minitab from menus*:
Stat → Basic Statistics → 1-Sample t
Check "Confidence interval" in dialogue box

*Excel from menus*:
Tools → Data Analysis
Choose "Descriptive Statistics",
Check "Summary" and "Confidence
Level for Mean" in dialogue box

CI =
24.82855 ± 0.0023983  ← ± term for CI

**Selected Excel Output**

| Passage Times | |
|---|---|
| Mean | 24.82855 |
| Standard Error | 0.0011459 |
| .... | .... |
| Standard Deviation | 0.0051245 |
| .... | .... |
| Count | 20 |
| Confidence Level(95.0%) | 0.0023983 |

$\bar{x}$ →

Computer output for Newcomb's passage-time data.

*Slide 9*        STAT 13, UCLA, Ivo Dinov

---

## Effect of increasing the confidence level

**99%** CI, $\bar{x} \pm 2.576\ \mathrm{se}(\bar{x})$

**95%** CI, $\bar{x} \pm 1.960\ \mathrm{se}(\bar{x})$

**90%** CI, $\bar{x} \pm 1.645\ \mathrm{se}(\bar{x})$

**80%** CI, $\bar{x} \pm 1.282\ \mathrm{se}(\bar{x})$

Confidence Level Increase

Increases the size of the CI

Why?

**Figure 8.1.3**  The greater the confidence level, the wider the interval

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

*Slide 10*        STAT 13, UCLA, Ivo Dinov

---

## Effect of <u>increasing</u> the sample size

$n = 10$ data points

$n = 40$ data points

$n = 90$ data points

Passage time

Increase Sample Size

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

Decreases the size of the CI

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

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

*Slide 11*        STAT 13, UCLA, Ivo Dinov

---

## Why ↑ in sample-size ↓ CI?

Confidence Interval for the true (population) mean $\mu$:

**sample mean $\pm\, t$ standard errors**

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

*Slide 12*        STAT 13, UCLA, Ivo Dinov

2

## CI for a population proportion

Confidence Interval for the true (population) proportion $p$:

*sample proportion $\pm$ $z$ standard errors*

or $\quad \hat{p} \pm z \, se(\hat{p})$, where $se(\hat{p}) = \sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}$ , Section 7.3.

---

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

**TABLE 8.4.1 Thiol Concentration (mmol)**

| | Normal | Rheumatoid |
|---|---|---|
| **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? | 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 |

---

## Example – higher blood thiol concentrations with rheumatoid arthritis
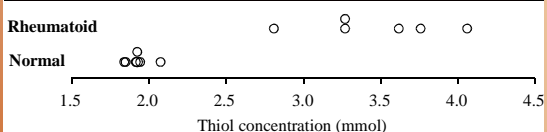


**Figure 8.4.1** Dot plot of Thiol concentration data.

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?

---

## Difference between means

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

*Difference between sample means
$\pm$ $t$ standard errors of the difference*

or $\quad \boxed{\bar{x}_1 - \bar{x}_2 \pm t \, se(\bar{x}_1 - \bar{x}_2)}$

---

## Difference between proportions

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

*Difference between sample proportions
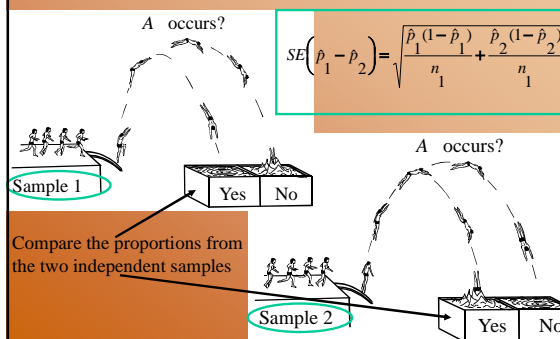$\pm$ $z$ standard errors of the difference*

**Big Question ???**

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

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

---

## Proportions from 2 independent samples



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

A occurs?

Sample 1 — Yes / No

Compare the proportions from the two independent samples

A occurs?

Sample 2 — Yes / No

## Single sample, several response categories

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

| Single Sample | | | | | | |
|---|---|---|---|---|---|---|
| | Cat. 1 | Cat. 2 | Cat. 3 | Cat. 4 | Cat. 5 | Cat. 6 |

Compare different proportions from the same sample

*Slide 19*    STAT 13, UCLA, Ivo Dinov

---

## Example – 1996 US Presidential Election

| State | n | Pre-election Polls | | | | Election Results | | |
|---|---|---|---|---|---|---|---|---|
| | | Clinton | Doll | Perot | Other/Undecided | Clinton | Doll | Perot |
| New Jersey | 1,000 | 51 | 33 | 8 | 8 | 53 | 36 | 9 |
| New York | 1,000 | 59 | 25 | 7 | 9 | 59 | 31 | 8 |
| Connecticutt | 1,000 | 51 | 29 | 11 | 9 | 52 | 35 | 10 |

### Single sample, several response categories

How far is Clinton ahead of Dole in NJ?
Diff.proportions= 18%
CI: [12% : 24%]
Actual diff 53-36=17

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

$$\text{estimate} \pm z \times SE = \hat{p}_1 - \hat{p}_2 \pm 1.96 \times SE\left(\hat{p}_1 - \hat{p}_2\right) =$$

$$\hat{p}_1 - \hat{p}_2 \pm 1.96 \times \sqrt{\frac{\hat{p}_1 + \hat{p}_2 - \left(\hat{p}_1 - \hat{p}_2\right)^2}{n}} =$$

$$0.18 \pm 1.96 \times 0.02842 = [12\% : 24\%]$$

*Slide 23*    STAT 13, UCLA, Ivo Dinov

---

## Random sample of 1,000 people is taken from 5 countries to assess efficacy, cost and quality of health care

| (Table entry is % agreeing) | Australia | Canada | N.Z. | UK | U.S. |
|---|---|---|---|---|---|
| Difficulties getting needed care | 15 | 20 | 18 | 15 | 28 |
| Recent changes will harm quality | 28 | 46 | 38 | 12 | 18 |
| System should be rebuilt | 30 | 23 | 32 | 14 | 33 |
| No bills not covered by insurance | 7 | 27 | 12 | 44 | 8 |

**2 independent Samples ($n_1$, $n_2$) compare proportions of people agreeing to a particular health care statement.**

**1 Sample, many response categories - compare proportions of New Zealanders either agreeing (Yes) or disagreeing (No) with a SET of statements.**

*Slide 26*    STAT 13, UCLA, Ivo Dinov

---

## 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    $\hat{q}_2 = 1 - \hat{p}_2$

*Slide 28*    STAT 13, UCLA, Ivo Dinov

---

## 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 <u>margin of error no greater than</u> *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**?)

*Slide 29*    STAT 13, UCLA, Ivo Dinov

---

## 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

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

● $\sigma^*$ is an estimate of the variability of individual observations

● *z* is the multiplier appropriate for the confidence level

*Slide 30*    STAT 13, UCLA, Ivo Dinov

4

# Chapter 8 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, | $\mu$ | $\bar{x}$ | $\dfrac{s_x}{\sqrt{n}}$ | $n-1$ |
| Proportion, | $p$ | $\hat{p}$ | $\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}$ | $\infty$ |
| Difference in means, | $\mu_1 - \mu_2$ | $\bar{x}_1 - \bar{x}_2$ | $\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}$ | $\mathrm{Min}(n_1\text{-}1, n_2\text{-}1)$ |
| Difference in proportions, | $p_1\text{-}p_2$ | $\hat{p}_1 - \hat{p}_2$ | (see Table 8.5.5) | $\infty$ |

$df = \infty$ means we use a multiplier obtained from the Normal(0,1) distribution.
CIs 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.