# Comparing Systems Using Sample Data

Raj Jain
Washington University in Saint Louis
Saint Louis, MO 63130
Jain@cse.wustl.edu

These slides are available on-line at:

http://www.cse.wustl.edu/~jain/cse567-15/


- Sample Versus Population
- Confidence Interval for The Mean
- Approximate Visual Test
- One Sided Confidence Intervals
- Confidence Intervals for Proportions
- Sample Size for Determining Mean and proportions

# Sample

- Old French word `essample'
  $\Rightarrow$ `sample' and `example'
- One example $\neq$ theory
- One sample $\neq$ Definite statement

# Sample Versus Population

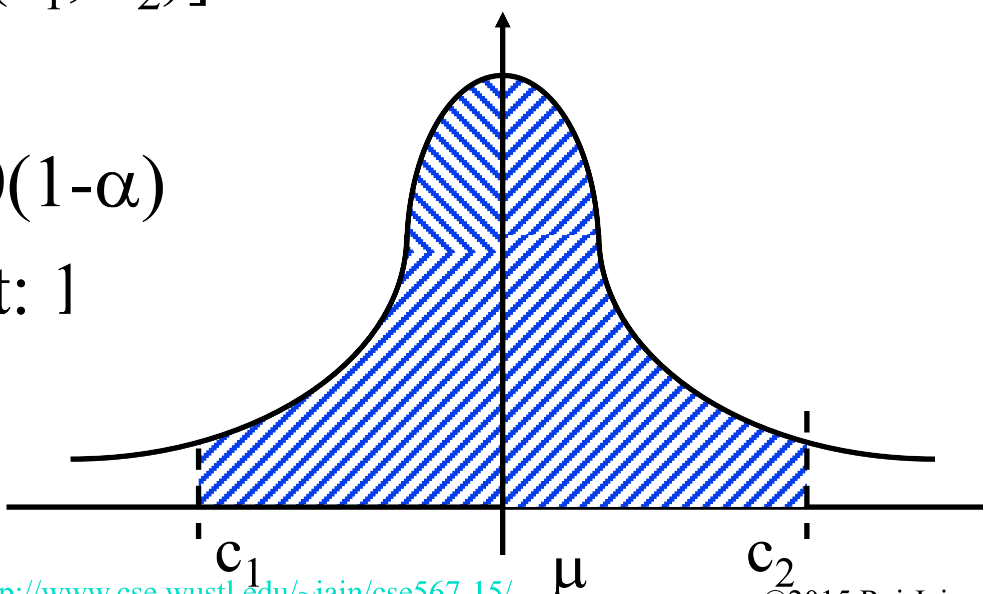
- Generate several million random numbers with mean $\mu$ and standard deviation $\sigma$
  Draw a sample of n observations
  $\bar{x} \neq \mu$
- Sample mean $\neq$ population mean
- Parameters: population characteristics = Unknown = Greek
- Statistics: Sample estimates = Random = English

# Confidence Interval for The Mean

- $k$ samples ⇒ $k$ Sample means
  ⇒ Can't get a single estimate of μ
  ⇒ Use bounds c_{1} and c_{2}:

  $$\text{Probability}\{c_1 \leq \mu \leq c_2\} = 1 - \alpha$$

- Confidence interval: $[(c_1, c_2)]$
- Significance level: α
- Confidence level: 100(1-α)
- Confidence coefficient: 1



# Determining Confidence Interval

- Use 5-percentile and 95-percentile of the sample means to get 90% Confidence interval ⇒ Need many samples.
- Central limit theorem: Sample mean of independent and identically distributed observations:

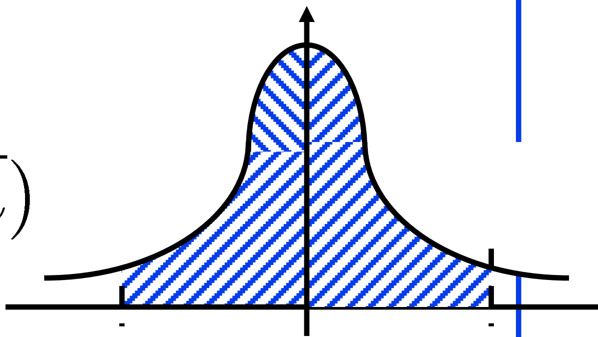
$$\bar{x} \sim N(\mu, \sigma/\sqrt{n})$$

Where μ = population mean, σ = population standard deviation

- Standard Error: Standard deviation of the sample mean = $\sigma/\sqrt{n}$
- 100(1-a)% confidence interval for μ:

$$(\bar{x} - z_{1-\alpha/2} s/\sqrt{n}, \bar{x} + z_{1-\alpha/2} s/\sqrt{n})$$

$z_{1-\alpha/2}$ = (1-α/2)-quantile of N(0,1)



# Example 13.1

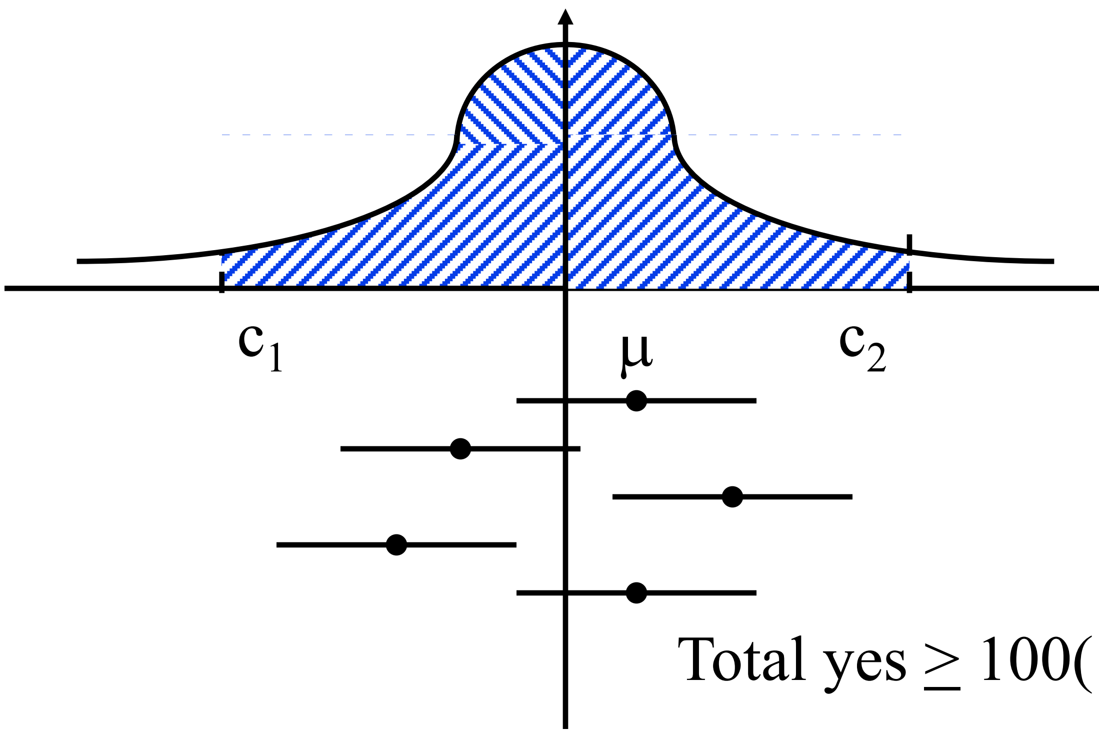
- $\bar{x} = 3.90$, s = 0.95 and $n = 32$
- A 90% confidence interval for the mean $= 3.90 \mp (1.645)(0.95)/\sqrt{32} = (3.62,\ 4.17)$
- We can state with 90% confidence that the population mean is between 3.62 and 4.17 The chance of error in this statement is 10%.

$$\text{A 95\% confidence interval for the mean} = 3.90 \mp (1.960)(0.95)/\sqrt{32} = (3.57,\ 4.23)$$

$$\text{A 99\% confidence interval for the mean} = 3.90 \mp (2.576)(0.95)/\sqrt{32} = (3.46,\ 4.33)$$

# Confidence Interval: Meaning

- If we take 100 samples and construct confidence interval for each sample, the interval would include the population mean in 90 cases.




Total yes $\geq 100(1-\alpha)$

# Confidence Interval for Small Samples

- 100(1-α) % confidence interval for for n < 30:

$$(\bar{x} - t_{[1-\alpha/2;n-1]}s/\sqrt{n}, \bar{x} + t_{[1-\alpha/2;n-1]}s/\sqrt{n})$$

- $t_{[1-\alpha/2;\, n-1]}$ = (1-α/2)-quantile of a t-variate with n-1 degrees of freedom

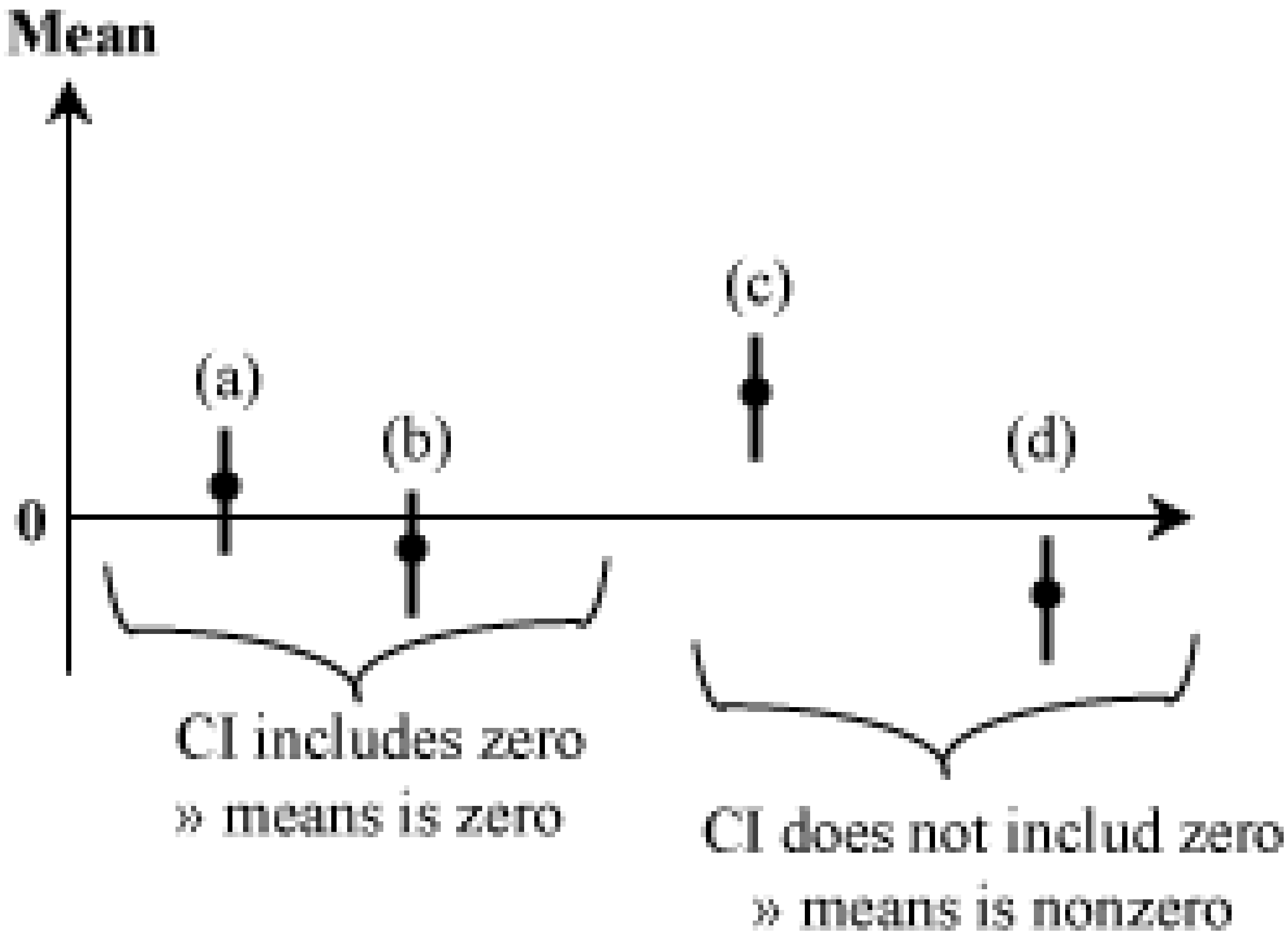
$$x \sim N(\mu, \sigma^2)$$
$$\Rightarrow (\bar{x} - \mu)/(\sigma/\sqrt{n}) \sim N(0, 1)$$
$$(n-1)s^2/\sigma^2 \sim \chi^2(n-1)$$
$$(\bar{x} - \mu)/\sqrt{s^2/n} \sim t(n-1)$$

# Example 13.2

- Sample: -0.04, -0.19, 0.14, -0.09, -0.14, 0.19, 0.04, and 0.09.
- Mean = 0, Sample standard deviation = 0.138.
- For 90% interval: $t_{[0.95;7]} = 1.895$
- Confidence interval for the mean

$$0 \mp 1.895 \times 0.138/\sqrt{8} = 0 \mp 0.0926 = (-0.0926,\ 0.0926)$$

# Testing For A Zero Mean



# Example 13.3

- Difference in processor times: {1.5, 2.6, -1.8, 1.3, -0.5, 1.7, 2.4}.
- Question: Can we say with 99% confidence that one is superior to the other?

  Sample size = $n$ = 7

  Mean = 7.20/7 = 1.03

  Sample variance = (22.84 - 7.20*7.20/7)/6 = 2.57

  Sample standard deviation} = $\sqrt{2.57}$ = 1.60

$$\text{Confidence interval} = 1.03 \mp t * 1.60/\sqrt{7} = 1.03 \mp 0.605t$$

$$100(1-\alpha) = 99, \; \alpha = 0.01, \; 1-\alpha/2 = 0.995$$

$t_{[0.995;\, 6]} = 3.707$

- 99% confidence interval = (-1.21, 3.27)

# Example 13.3 (Cont)

- Opposite signs ⇒ we cannot say with 99% confidence that the mean difference is significantly different from zero.
- Answer: They are same.
- Answer: The difference is zero.

# Example 13.4

- Difference in processor times: {1.5, 2.6, -1.8, 1.3, -0.5, 1.7, 2.4}.
- Question: Is the difference 1?
- 99% Confidence interval = (-1.21, 3.27)
- Yes: The difference is 1

# Homework 13A: Exercise 13.2 (Updated)

- Answer the following for the data of Exercise 12.11:
  - What is the 10-percentile and 90-percentile from the sample?
  - What is the mean number of disk I/Os per program?
  - What is the 90% confidence interval for the mean?
  - What fraction of programs make less than or equal to 25 I/Os and what is the 95% confidence interval for the fraction?
  - What is the one sided 90% confidence interval for the mean?

# Paired vs. Unpaired Comparisons

- **Paired**: one-to-one correspondence between the ith test of system A and the ith test on system B
- Example: Performance on ith workload
- Use confidence interval of the difference
- **Unpaired**: No correspondence
- Example: *n* people on System A, *n* on System B
  ⇒Need more sophisticated method

# Example 13.5

- Performance: {(5.4, 19.1), (16.6, 3.5), (0.6, 3.4), (1.4, 2.5), (0.6, 3.6), (7.3, 1.7)}. Is one system better?
- Differences: {-13.7, 13.1, -2.8, -1.1, -3.0, 5.6}.

$$\text{Sample mean} = -0.32$$
$$\text{Sample variance} = 81.62$$
$$\text{Sample standard deviation} = 9.03$$
$$\text{Confidence interval for the mean} = -0.32 \mp t\sqrt{(81.62/6)}$$
$$= -0.32 \mp t(3.69)$$
$$t_{[0.95,5]} = 2.015$$
$$90\% \text{ confidence interval} = -0.32 \mp (2.015)(3.69)$$
$$= (-7.75,\ 7.11)$$

- Answer: No. They are not different.

# Unpaired Observations

- Compute the sample means:

$$\bar{x}_a = \frac{1}{n_a} \sum_{i=1}^{n_a} x_{ia}$$

$$\bar{x}_b = \frac{1}{n_b} \sum_{i=1}^{n_b} x_{ib}$$

- Compute the sample standard deviations:

$$s_a = \left\{ \frac{\left(\sum_{i=1}^{n_a} x_{ia}^2\right) - n_a \bar{x}_a^2}{n_a - 1} \right\}^{\frac{1}{2}}$$

$$s_b = \left\{ \frac{\left(\sum_{i=1}^{n_b} x_{ib}^2\right) - n_b \bar{x}_b^2}{n_b - 1} \right\}^{\frac{1}{2}}$$

# Unpaired Observations (Cont)

- Compute the mean difference: $(\bar{x}_a - \bar{x}_b)$
- Compute the standard deviation of the mean difference:

$$s = \sqrt{\left(\frac{s_a^2}{n_a} + \frac{s_b^2}{n_b}\right)}$$

- Compute the effective number of degrees of freedom:

$$\nu = \frac{\left(\frac{s_a^2}{n_a} + \frac{s_b^2}{n_b}\right)^2}{\frac{1}{n_a-1}\left(\frac{s_a^2}{n_a}\right)^2 + \frac{1}{n_b-1}\left(\frac{s_b^2}{n_b}\right)^2} - 2$$

- Compute the confidence interval for the mean difference:

$$(\bar{x}_a - \bar{x}_b) \mp t_{[1-\alpha/2;\nu]} s$$

Note: No $\sqrt{\nu}$

# Example 13.6

- Times on System A: {5.36, 16.57, 0.62, 1.41, 0.64, 7.26}
  Times on system B: {19.12, 3.52, 3.38, 2.50, 3.60, 1.74}
- Question: Are the two systems significantly different?
- For system A:

  $$\text{Mean } \bar{x}_a = 5.31$$
  $$\text{Variance } s_a^2 = 37.92$$
  $$n_a = 6$$

- For System B:

  $$\text{Mean } \bar{x}_b = 5.64$$
  $$\text{Variance } s_b^2 = 44.11$$
  $$n_b = 6$$

# Example 13.6 (Cont)

Mean difference $\bar{x}_a - \bar{x}_b = -0.33$

Standard deviation of the mean difference $= 3.698$

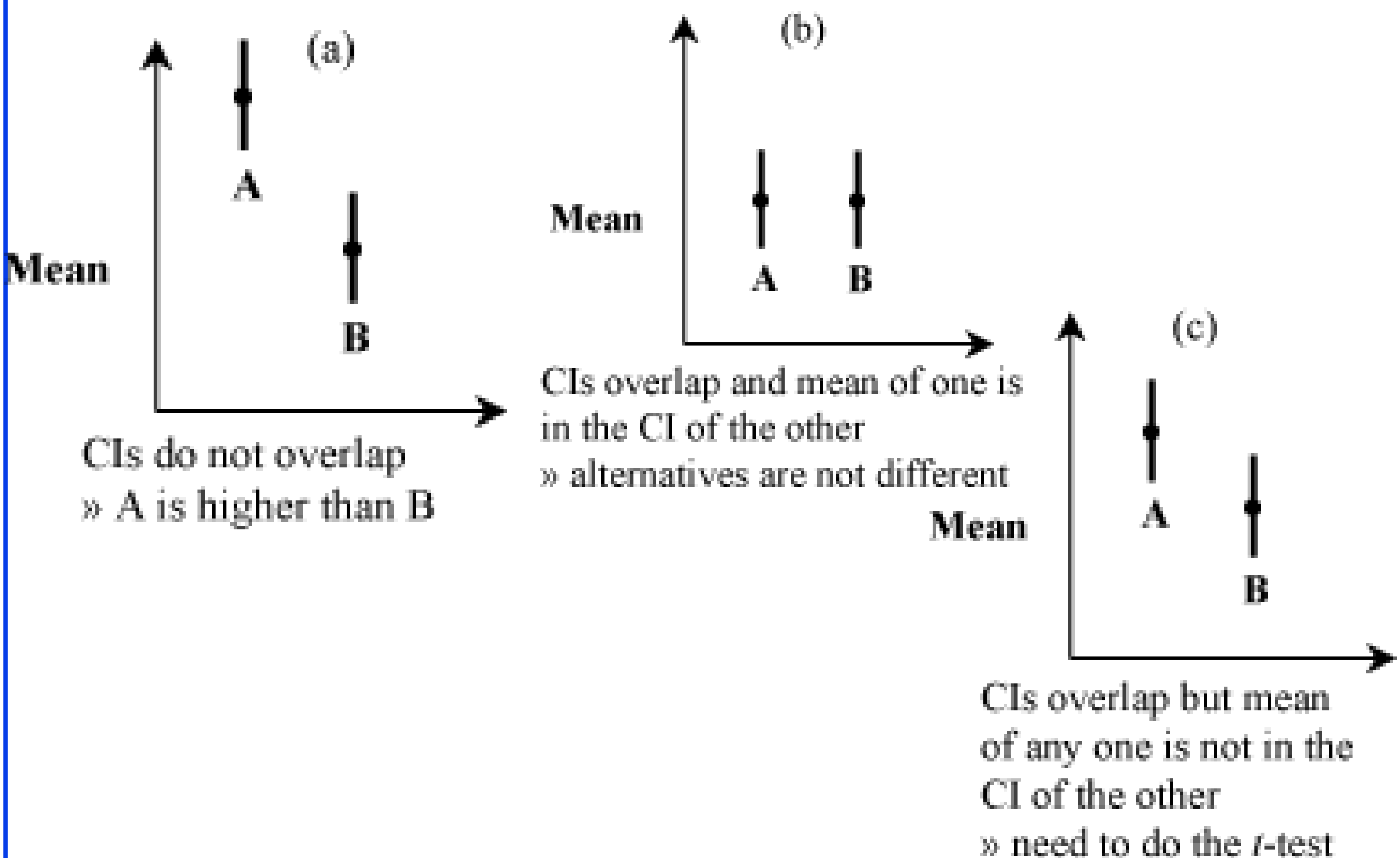
Effective number of degrees of freedom $f = 7.943$

The 0.95-quantile of a t-variate with 8 degrees of freedom $= 1.860$

The 90% confidence interval for the difference $= (-7.21, 6.54)$

- The confidence interval includes zero
  $\Rightarrow$ the two systems are not different.

# Approximate Visual Test



# Example 13.7

- Times on System A: {5.36, 16.57, 0.62, 1.41, 0.64, 7.26}
  Times on system B: {19.12, 3.52, 3.38, 2.50, 3.60, 1.74}
  $t_{[0.95, 5]} = 2.015$
- The 90% confidence interval for the mean of A = 5.31 ∓ (2.015) $\sqrt{(37.92/6)}$
  = (0.24, 10.38)
- The 90% confidence interval for the mean of B = 5.64 ∓ (2.015) $\sqrt{(44.11/6)}$
  = (0.18, 11.10)
- Confidence intervals overlap and the mean of one falls in the confidence interval for the other.
  ⇒ Two systems are not different at this level of confidence.

# What Confidence Level To Use?

- Need not always be 90% or 95% or 99%
- Base on the loss that you would sustain if the parameter is outside the range and the gain you would have if the parameter is inside the range.
- Low loss ⇒ Low confidence level is fine
  E.g., lottery of 5 Million with probability $10^{-7}$
- 90% confidence ⇒ buy nine million tickets
- 0.01% confidence level is fine.
- 50% confidence level may or may not be too low
- 99% confidence level may or may not be too high

# Hypothesis Testing vs. Confidence Intervals

- Confidence interval provides more information
- Hypothesis test = yes-no decision
- Confidence interval also provides  possible range
- Narrow confidence interval ⇒ high degree of precision
- Wide confidence interval ⇒ Low precision
- Example: (-100,100) ⇒ No difference
  (-1,1) ⇒ No difference
- Confidence intervals tell us not only what to say but also how loudly to say it
- CI is easier to explain to decision makers
- CI is more useful.
  E.g.,  parameter range (100, 200)
  vs. Probability of (parameter = 110) = 3%

# One Sided Confidence Intervals

- Two side intervals: 90% Confidence
  $\Rightarrow$ P(Difference > upper limit) = 5%
  $\Rightarrow$ P(Difference < Lower limit) = 5%
- One sided Question: Is the mean greater than 0?
  $\Rightarrow$ One side confidence interval
- One sided lower confidence interval for μ:

$$(\bar{x} - t_{[1-\alpha;n-1]}\frac{s}{\sqrt{n}}, \infty)$$

  Note t at 1-α (not 1-α/2)
- One sided upper confidence interval for μ:

$$\left(-\infty, \bar{x} + t_{[1-\alpha;n-1]}\frac{s}{\sqrt{n}}\right)$$

- For large samples: Use z instead of t

# Example 13.8

- Time between crashes

| System | Number | Mean | Stdv |
|---|---|---|---|
| A | 972 | 124.10 | 198.20 |
| B | 153 | 141.47 | 226.11 |

- Assume unpaired observations
- Mean difference:

$$\bar{x}_A - \bar{x}_B = 124.10 - 141.47 = -17.37$$

- Standard deviation of the difference:

$$s = \sqrt{\left(\frac{s_a^2}{n_a} + \frac{s_b^2}{n_b}\right)} = \sqrt{\frac{(198.20)^2}{972} + \frac{(226.11)^2}{153}} = 19.35$$

- Effective number of degrees of freedom:

# Example 13.8 (Cont)

$$\nu = \frac{\left(\frac{s_a^2}{n_a} + \frac{s_b^2}{n_b}\right)^2}{\frac{1}{n_a - 1}\left(\frac{s_a^2}{n_a}\right)^2 + \frac{1}{n_b - 1}\left(\frac{s_b^2}{n_b}\right)^2} - 2$$

$$= \frac{\left(\frac{(198.20)^2}{972} + \frac{(226.11)^2}{153}\right)^2}{\frac{1}{972-1}\left(\frac{(198.20)^2}{972}\right)^2 + \frac{1}{153-1}\left(\frac{(226.11)^2}{153}\right)^2} - 2$$

$$= 188.56$$

- $\nu > 30 \Rightarrow$ Use z rather than t
- One sided test $\Rightarrow$ Use $z_{0.90}$=1.28 for 90% confidence
- 90% Confidence interval:

  $(-\infty, -17.37+1.28 * 19.35)=(-\infty, 7.402)$
- CI includes zero $\Rightarrow$ System A is not more susceptible to crashes than system B.

# Confidence Intervals for Proportions

- Proportion = probabilities of various categories
  E.g., P(error)=0.01, P(No error)=0.99
- $n_1$ of n observations are of type 1 ⇒

$$\text{Sample proportion} = p = \frac{n_1}{n}$$

$$\text{Confidence interval for the proportion} = p \mp z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}}$$

- Assumes Normal approximation of Binomial distribution
  ⇒ Valid only if $np \geq 10$.
- Need to use binomial tables if $np < 10$
  Can't use t-values

# CI for Proportions (Cont)

- 100(1-α)% one sided confidence interval for the proportion: ‡

$$\left(0, p + z_{1-\alpha}\sqrt{\frac{p(1-p)}{n}}\right) \text{ or } \left(p - z_{1-\alpha}\sqrt{\frac{p(1-p)}{n}}, 1\right)$$

‡ Provided $np \geq 10$.

# Example 13.9

- 10 out of 1000 pages printed on a laser printer are illegible.

  $$\text{Sample proportion} = p = \frac{10}{1000} = 0.01$$

- np≥ 10

  $$\begin{aligned}\text{Confidence interval} &= p \mp z\sqrt{\frac{p(1-p)}{n}} \\ &= 0.01 \mp z\sqrt{\frac{0.01(0.99)}{1000}} = 0.01 \mp 0.003z\end{aligned}$$

- 90% confidence interval = 0.01 ∓ (1.645)(0.003) = (0.005, 0.015)
- 95% confidence interval = 0.01 ∓ (1.960)(0.003) = (0.004, 0.016)

# Example 13.9 (Cont)

- At 90% confidence:
  0.5% to 1.5% of the pages are illegible
  Chances of error = 10%
- At 95% Confidence:
  0.4% to 1.6% of the pages are illegible
  Chances of error = 5%

# Example 13.10

- 40 Repetitions on two systems: System A superior in 26 repetitions
- Question: With 99% confidence, is system A superior?

$$p = 26/40 = 0.65$$

- Standard deviation = $\sqrt{p*(1-p)/n} = 0.075$
- 99% confidence interval = $0.65 \mp (2.576)(0.075)$

$$= (0.46, 0.84)$$

- CI includes 0.5

  ⇒ we cannot say with 99% confidence that system A is superior.
- 90% confidence interval = $0.65 \mp (1.645)(0.075) = (0.53, 0.77)$
- CI does not include 0.5

  ⇒ Can say with 90% confidence that system A is superior.

# Sample Size for Determining Mean

- Larger sample ⇒ Narrower confidence interval \R
  Higher confidence
- Question: How many observations n to get an accuracy of ± r% and a confidence level of 100(1-α)%?

$$\bar{x} \mp z \frac{s}{\sqrt{n}}$$

- r% Accuracy ⇒

$$\mathrm{CI} = (\bar{x}(1 - r/100), \bar{x}(1 + r/100))$$

$$\therefore \bar{x} \mp z \frac{s}{\sqrt{n}} = \bar{x}\left(1 \mp \frac{r}{100}\right)$$

$$z \frac{s}{\sqrt{n}} = \bar{x} \frac{r}{100}$$

$$n = \left(\frac{100zs}{r\bar{x}}\right)^2$$

# Example 13.11

- Sample mean of the response time = 20 seconds

  Sample standard deviation = 5

  Question: How many repetitions are needed to get the response time accurate within 1 second at 95% confidence?

- Required accuracy = 1 in 20 = 5%

  Here, $\bar{x}$ = 20, s= 5, z= 1.960, and r=5,

$$n = \left(\frac{(100)(1.960)(5)}{(5)(20)}\right)^2 = (9.8)^2 = 96.04$$

A total of 97 observations are needed.

# Sample Size for Determining Proportions

$$\text{Confidence interval for the proportion} = p \mp z\sqrt{\left(\frac{p(1-p)}{n}\right)}$$

To get a half-width (accuracy of) r:

$$p \mp r = p \mp z\sqrt{\left(\frac{p(1-p)}{n}\right)}$$

$$r = z\sqrt{\left(\frac{p(1-p)}{n}\right)}$$

$$n = z^2 \frac{p(1-p)}{r^2}$$

# Example 13.12

- Preliminary measurement : illegible print rate of 1 in 10,000.
- Question: How many pages must be observed to get an accuracy of 1 per million at 95% confidence?
- Answer:

$$p = 1/10000 = 1E-4, r = 1E-6, z = 1.960$$

$$n = (1.960)^2 \left( \frac{10^{-4}(1-10^{-4})}{(10^{-6})^2} \right) = 384160000$$

A total of 384.16 million pages must be observed.

# Example 13.13

- Algorithm A loses 0.5% of packets and algorithm B loses 0.6%.
- Question: How many packets do we need to observe to state with 95% confidence that algorithm A is better than the algorithm B?
- Answer:

$$\text{CI for algorithm A} = 0.005 \mp 1.960 \left( \frac{0.005(1-0.005)}{n} \right)^{1/2}$$

$$\text{CI for algorithm B} = 0.006 \mp 1.960 \left( \frac{0.006(1-0.006)}{n} \right)^{1/2}$$

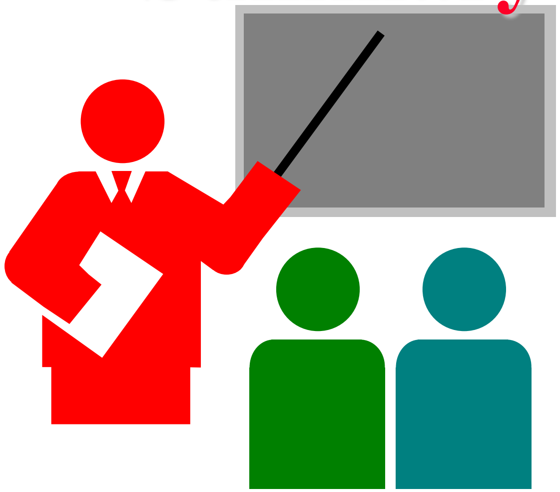
# Example 13.13 (Cont)

- For non-overlapping intervals:

$$0.005 \mp 1.960 \left( \frac{0.005(1-0.005)}{n} \right)^{1/2}$$
$$\leq 0.006 \mp 1.960 \left( \frac{0.006(1-0.006)}{n} \right)^{1/2}$$

- n = 84340 ⇒ We need to observe 85,000 packets.

# Summary



- All statistics based on a sample are random and should be specified with a confidence interval
- If the confidence interval includes zero, the hypothesis that the population mean is zero cannot be rejected
- Paired observations ⇒ Test the difference for zero mean
- Unpaired observations ⇒ More sophisticated test
- Confidence intervals apply to proportions too.

# Homework 13B: Exercise 13.3 (Revised)

- For the code size data of Table 11.2, find 90% confidence intervals for the average code sizes on each processor. Answer the following for RISC-I and Z8002:
  - At what level of significance, can you say that one is better than the other?
  - How many workloads would you need to decide the superiority at 90% confidence? (Compute n to avoid zero in the confidence interval.)