Performance Evaluation:

Comparing Systems Using Sample Data

Hongwei Zhang

http://www.cs.wayne.edu/~hzhang

Statistics are like alienists --- they will testify for both side.

--- Fiorello La Guardia

Acknowledgement: this lecture is partially based on the slides of Dr. Raj Jain.

Outline

- Sample vs. Population
- Confidence Interval for the Mean
- Application & variations of CI
  - Testing for a zero mean
  - Comparing two alternatives
  - One Sided Confidence Intervals
  - Confidence Intervals for Proportions
- What confidence level to use
- Hypothesis testing vs. confidence interval
- Sample Size

Outline

- Sample vs. Population
- Confidence Interval for The Mean
- Application & variations of CI
  - Testing for a zero mean
  - Comparing two alternatives
  - One Sided Confidence Intervals
  - Confidence Intervals for Proportions
- What confidence level to use
- Hypothesis testing vs. confidence interval
- Sample Size

Sample

- Old French word `essample‘

  => `sample' and `example'

  A sample is only an example!

- One example ≠ theory

  One sample ≠ Definite statement

Sample vs. population

- Generate several million random numbers with mean μ and standard deviation σ;

  Draw a sample of n observations:

$$\bar{x} \neq \mu$$

- => Sample mean ≠ population mean

- Parameters: population characteristics
  - Unknown; written in Greek (by convention)

  Statistics: Sample estimates
  - Random variables; written in English

Outline

- Sample vs. Population
- Confidence Interval for The Mean
- Application & variations of CI
  - Testing for a zero mean
  - Comparing two alternatives
  - One Sided Confidence Intervals
  - Confidence Intervals for Proportions
- What confidence level to use
- Hypothesis testing vs. confidence interval
- Sample Size

Confidence interval for μ

- An interval (C1, C2) so that

$$\text{probability}\{C1 \le \mu \le C2\} = 1 - \alpha$$

where α is called the significance level,

100(1-α) is called the confidence level,

1-α is called the confidence coefficient.

Determining confidence interval

- Use 5-percentile and 95-percentile of the sample means to get (approx.) 90% confidence interval => Need many samples
- Central limit theorem:

  Sample mean of independent and identically distributed observations:

$$\bar{x} \sim N(\mu, \sigma^2/n)$$

  where μ = population mean, σ = population standard deviation

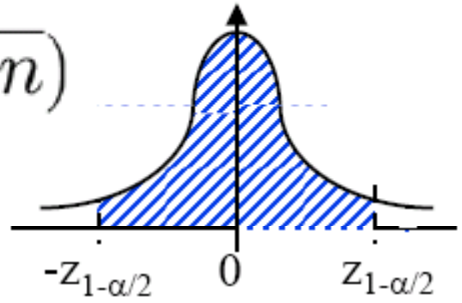
- Standard Error: Standard deviation of the sample mean = $\sigma/\sqrt{n}$

Determining confidence interval (contd.)

- 100(1-α)% 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)



- Why?

Example

- Packet CPU time: $\bar{x} = 3.90$, s = 0.95 and $n = 32$

- A 90% confidence interval for the mean
  $= 3.90 \mp (1.645)(0.95)/\sqrt{32} \quad = (3.62,\ 4.17)$

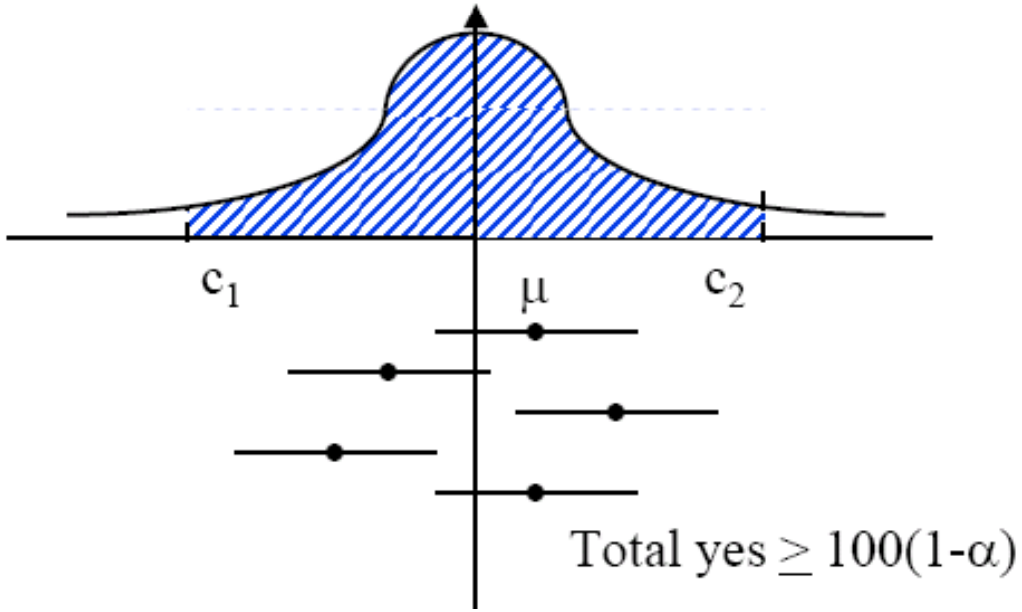
- We can state with 90% confidence that the population mean is between 3.62 and 4.17 The chance of error in this statement is 10%.

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

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

Confidence interval: intuitive meaning

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



Confidence interval for small samples (n<30): ONLY if samples come from normal distribution

- For n < 30, $(\bar{x}-\mu)/\sqrt{s^2/n}$ has a t-distribution with n-1 degrees of freedom

- 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

- Error 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% confidence interval: $t_{[0.95;7]} = 1.895$

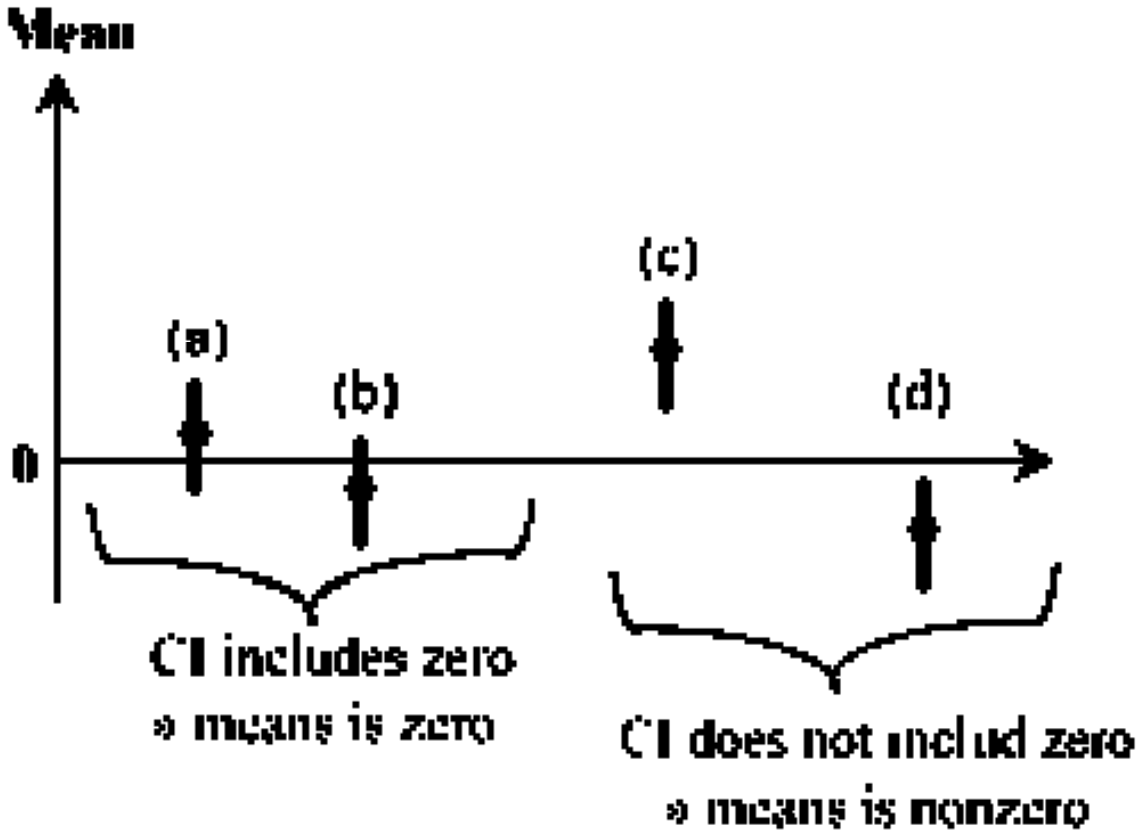
  Thus, confidence interval for the mean

$$0 \mp 1.895 \times 0.138 = 0 \mp 0.262 = (-0.262,\ 0.262)$$

Outline

- Sample vs. Population
- Confidence Interval for The Mean
- Application & variations of CI
  - Testing for a zero mean
  - Comparing two alternatives
  - One Sided Confidence Intervals
  - Confidence Intervals for Proportions
- What confidence level to use
- Hypothesis testing vs. confidence interval
- Sample Size

Testing for a zero mean



Example

- Difference in network 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 = = 1.60

  $$\text{Confidence interval} = 1.03 \mp t * 1.60/\sqrt{7} = 1.03 \mp 0.6t$$

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

  => t[0.995; 6] = 3.70

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

Example (contd.)

- Opposite signs => we cannot say with 99% confidence that the mean difference is significantly different from zero.

- Answer: They are same; or the difference is zero.

Another question: testing for a constant

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

Outline

- Sample vs. Population
- Confidence Interval for The Mean
- Application & variations of CI
  - Testing for a zero mean
  - Comparing two alternatives
  - One Sided Confidence Intervals
  - Confidence Intervals for Proportions
- What confidence level to use
- Hypothesis testing vs. confidence interval
- Sample Size

Comparing two alternatives: Paired vs. unpaired observations

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

Example of "paired observations"

- 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}.

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

- Answer: No. They are not different.

Unpaired observations: t-test

- 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 (contd.)

- 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:

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

Example

- 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:

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

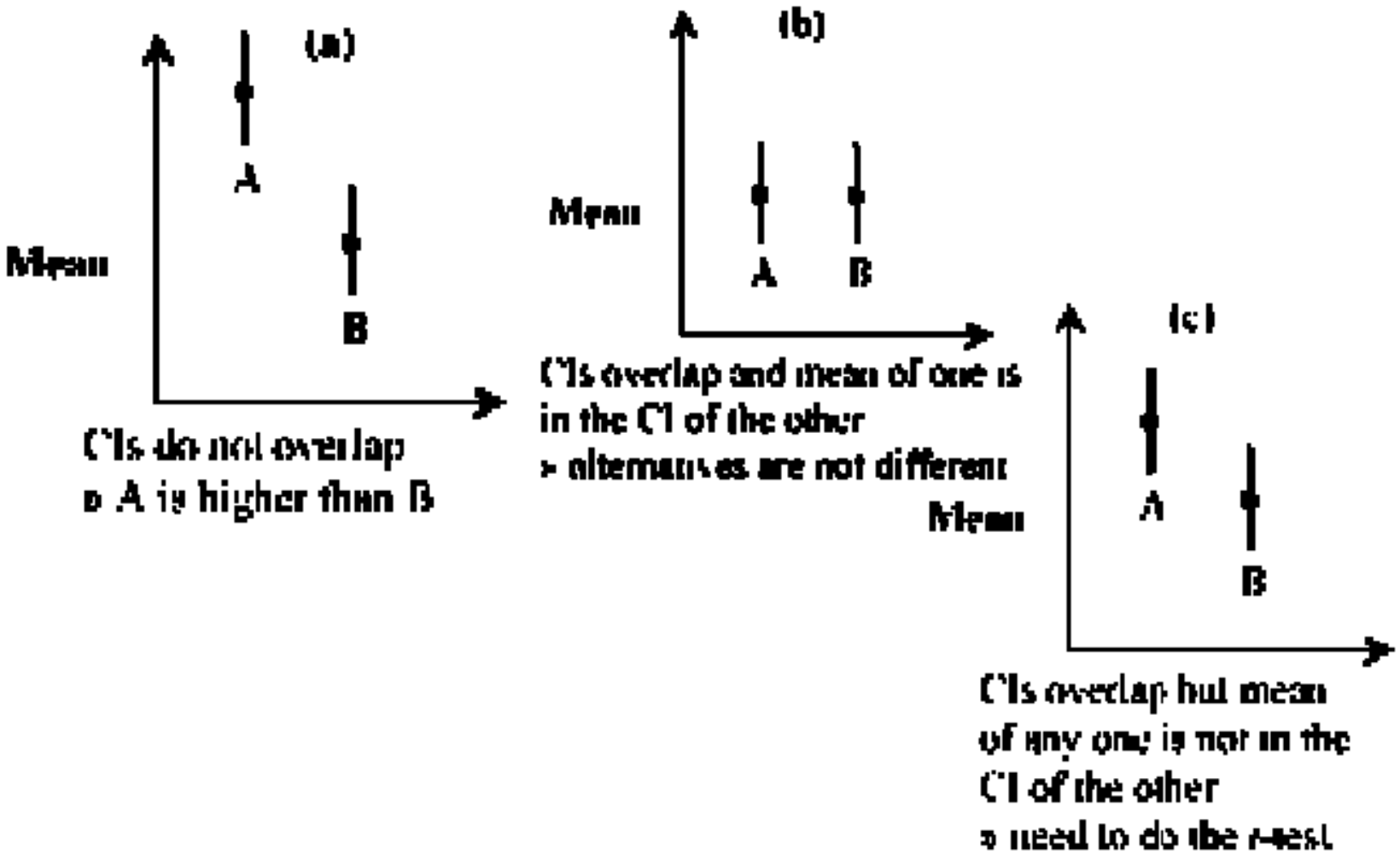
Example (contd.)

$$\text{Mean difference } \bar{x}_a - \bar{x}_b = -0.33$$
$$\text{Standard deviation of the mean difference} = 3.698$$
$$\text{Effective number of degrees of freedom } f = 11.921$$
$$\text{The 0.95-quantile of a t-variate with 12 degrees of freedom} = 1.71$$
$$\text{The 90\% confidence interval for the difference} = (-6.92, 6.26)$$

- The confidence interval includes zero

  => the two systems are not different.

Approximate visual test



Example

- 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 = $5.31 \mp (2.015)$ = (0.24, 10.38)

  The 90% confidence interval for the mean of B = $5.64 \mp (2.015)$ = (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.

Small samples, and they do not come from normal distribution?

- Non-parametric methods

- Reference:

  M. Hollander, D. A. Wolfe, “Nonparametric Statistical Methods”, 2nd edition, Wiley, 1999

Outline

- Sample vs. Population
- Confidence Interval for The Mean
- Application & variations of CI
  - Testing for a zero mean
  - Comparing two alternatives
  - One Sided Confidence Intervals
  - Confidence Intervals for Proportions
- What confidence level to use
- Hypothesis testing vs. confidence interval
- Sample Size

One sided confidence interval

- Two side intervals: 90% Confidence

  => P(Difference > upper limit) = 5%

  => P(Difference < Lower limit) = 5%

- One sided Question: Is the mean greater than 0?

  => One sided confidence interval

- *One sided lower confidence interval* for μ:

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

  Note t at 1-α (not 1-α/2)

- *One sided upper confidence interval* for μ: $\left(\bar{x}, \bar{x} + t_{[1-\alpha;n-1]}\frac{s}{\sqrt{n}}\right)$

- For large samples: Use z instead of t

Example

- 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 (contd.)

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

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

$$= 191.05$$

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

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

Outline

- Sample vs. Population
- Confidence Interval for The Mean
- Application & variations of CI
  - Testing for a zero mean
  - Comparing two alternatives
  - One Sided Confidence Intervals
  - Confidence Intervals for Proportions
- What confidence level to use
- Hypothesis testing vs. confidence interval
- Sample Size

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 =>

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

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

Assumed Normal approximation of Binomial distribution

- Valid only if $np \geq 10$.

- If $np < 10$, computation is complex and need to use binomial tables; cannot use t-values

CI for proportions (contd.)

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

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

Provided that $np \geq 10$.

Example

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

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

- np$\geq$ 10

$$\text{Confidence interval} = p \mp z\sqrt{\frac{p(1-p)}{n}}$$

$$= 0.01 \mp z\sqrt{\frac{0.01(0.99)}{1000}} = 0.01 \mp 0.003z$$

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

Example (contd.)

- 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%

Another example: test proportion for a constant

- 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
  $\Rightarrow$ 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
  $\Rightarrow$ Can say with 90% confidence that system A is superior.

Outline

- Sample vs. Population
- Confidence Interval for The Mean
- Application & variations of CI
  - Testing for a zero mean
  - Comparing two alternatives
  - One Sided Confidence Intervals
  - Confidence Intervals for Proportions
- What confidence level to use
- Hypothesis testing vs. confidence interval
- Sample Size

What confidence level to use?

- Need not always be 90% or 95% or 99%
- Base on the loss of drawing wrong conclusions and the gain of drawing correct conclusions
  - Low loss => Low confidence level is fine
    - E.g., lottery of 5 Million with probability 10^{-7};

      90% confidence => buy nine million tickets (each ticket costs $1);

      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

Outline

- Sample vs. Population
- Confidence Interval for The Mean
- Application & variations of CI
  - Testing for a zero mean
  - Comparing two alternatives
  - One Sided Confidence Intervals
  - Confidence Intervals for Proportions
- What confidence level to use
- Hypothesis testing vs. confidence interval
- Sample Size

Hypothesis testing vs. confidence intervals

- Closely related
  - E.g., If the value of the parameter specified by the null hypothesis is contained in the 95% confidence interval, then the null hypothesis cannot be rejected at the 0.05 level. If the value specified by the null hypothesis is not in the interval then the null hypothesis can be rejected at the 0.05 level.

Hypothesis testing vs. confidence intervals (contd.)

- 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%

Outline

- Sample vs. Population
- Confidence Interval for The Mean
- Application & variations of CI
  - Testing for a zero mean
  - Comparing two alternatives
  - One Sided Confidence Intervals
  - Confidence Intervals for Proportions
- What confidence level to use
- Hypothesis testing vs. confidence interval
- Sample Size

Sample size for determining mean

- Larger sample ⇒ Narrower confidence interval & 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 ⇒

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

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

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

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

Example

- 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 proportion

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

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

- 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.

Sample size for comparing two alternatives: non-overlapping CIs

- 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:

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

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

Sample size for non-overlapping CIs (contd.)

- 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 $\Rightarrow$ We need to observe 85,000 packets.

Note: sufficient condition, not necessary condition

Summary

- Sample vs. Population
- Confidence Interval for The Mean
- Application & variations of CI
  - Testing for a zero mean
  - Comparing two alternatives
  - One Sided Confidence Intervals
  - Confidence Intervals for Proportions
- What confidence level to use
- Hypothesis testing vs. confidence interval
- Sample Size