CS 147:
Computer Systems Performance Analysis
Comparing Systems and Analyzing Alternatives

CS 147:
Computer Systems Performance Analysis
Comparing Systems and Analyzing Alternatives

Overview

Overview
Finding Confidence Intervals
Basics
Using the z Distribution
Using the t Distribution
Comparing Alternatives
Paired Observations
Unpaired Observations
Proportions
Special Considerations
Sample Sizes

Comparing Systems Using Sample Data

- It's not usually enough to collect data
- Usually we also want to say what's better

Comparing Systems Using Sample Data

- It's not usually enough to collect data
- Usually we also want to say what's better

Review

- How tall is Fred?

Review

- How tall is Fred?

Review

- How tall is Fred?
  - Suppose 90% of humans are between 155 and 190 cm

Review
- How tall is Fred?
  - Suppose 90% of humans are between 155 and 190 cm

Review

- How tall is Fred?
  - Suppose 90% of humans are between 155 and 190 cm

$\therefore$ Fred is between 155 and 190 cm

Review
- How tall is Fred?
  - Suppose 90% of humans are between 155 and 190 cm
∴ Fred is between 155 and 190 cm

Review

- How tall is Fred?
  - Suppose 90% of humans are between 155 and 190 cm

  $\therefore$ Fred is between 155 and 190 cm
- We are 90% *confident* that Fred is between 155 and 190 cm

Review

- How tall is Fred?
  - Suppose 90% of humans are between 155 and 190 cm

  $\therefore$ Fred is between 155 and 190 cm
- We are 90% *confident* that Fred is between 155 and 190 cm

Confidence Interval of Sample Mean

- Knowing where 90% of sample means fall, we can state a *90% confidence interval*
- Key is *Central Limit Theorem*:
  - Sample means are normally distributed
  - Only if samples independent
  - Mean of sample means is population mean μ
  - Standard deviation (*standard error*) is $\sigma/\sqrt{n}$

Confidence Interval of Sample Mean

- Knowing where 90% of sample means fall, we can state a *90% confidence interval*
- Key is *Central Limit Theorem*:
  - Sample means are normally distributed
  - Only if samples independent
  - Mean of sample means is population mean μ
  - Standard deviation (*standard error*) is $\sigma/\sqrt{n}$

Estimating Confidence Intervals

- Two formulas for confidence intervals
  - Over 30 samples from any distribution: z-distribution
  - Small sample from normally distributed population: t-distribution
- Common error: using t-distribution for non-normal population
  - Central Limit Theorem often saves us

Estimating Confidence Intervals
- Two formulas for confidence intervals
  - Over 30 samples from any distribution: z-distribution
  - Small sample from normally distributed population: t-distribution
- Common error: using t-distribution for non-normal population
  - Central Limit Theorem often saves us

The z Distribution

- Interval on either side of mean:

$$\overline{x} \mp z_{1-\frac{\alpha}{2}} \left(\frac{s}{\sqrt{n}} \right)$$

- Significance level α is small for large confidence levels
- Tables of z are tricky: be careful!

The z Distribution

- Interval on either side of mean:

$$\overline{x} \mp z_{1-\frac{\alpha}{2}} \left(\frac{s}{\sqrt{n}} \right)$$

- Significance level α is small for large confidence levels
- Tables of z are tricky: be careful!

Example of z Distribution

- 35 samples: 10, 16, 47, 48, 74, 30, 81, 42, 57, 67, 7, 13, 56, 44, 54, 17, 60, 32, 45, 28, 33, 60, 36, 59, 73, 46, 10, 40, 35, 65, 34, 25, 18, 48, 63
- Sample mean $\overline{x} = 42.1$. Standard deviation $s = 20.1$. $n = 35$.
- 90% confidence interval is

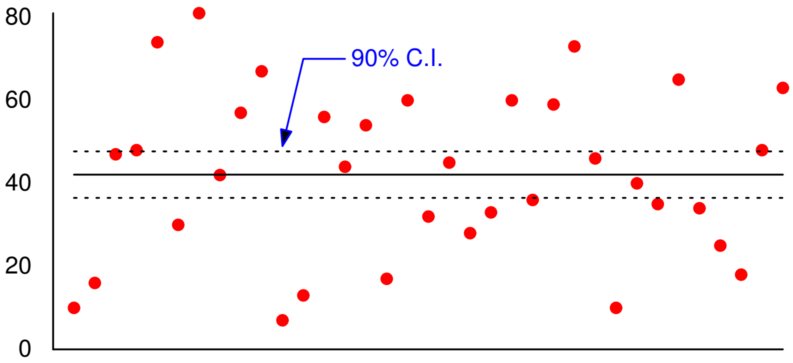
$$42.1 \mp (1.6456)\frac{20.1}{\sqrt{35}} = (36.5, 47.4)$$

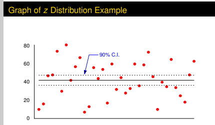
Example of z Distribution

- 35 samples: 10, 16, 47, 48, 74, 30, 81, 42, 57, 67, 7, 13, 56, 44, 54, 17, 60, 32, 45, 28, 33, 60, 36, 59, 73, 46, 10, 40, 35, 65, 34, 25, 18, 48, 63
- Sample mean $\overline{x} = 42.1$. Standard deviation $s = 20.1$. $n = 35$.
- 90% confidence interval is

$$42.1 \mp (1.6456)\frac{20.1}{\sqrt{35}} = (36.5, 47.4)$$

Graph of z Distribution Example





The t Distribution

- Formula is almost the same:

$$\overline{x} \mp t_{[1-\frac{\alpha}{2};n-1]}\left(\frac{s}{\sqrt{n}}\right)$$

- Usable only for normally distributed populations!
- But works with small samples

The t Distribution

- Formula is almost the same:

$$\overline{x} \mp t_{[1-\frac{\alpha}{2};n-1]}\left(\frac{s}{\sqrt{n}}\right)$$

- Usable only for normally distributed populations!
- But works with small samples

Example of t Distribution

- 10 height samples: 148, 166, 170, 191, 187, 114, 168, 180, 177, 204
- Sample mean $\overline{x} = 170.5$. Standard deviation $s = 25.1$, $n = 10$.
- 90% confidence interval is

$$170.5 \mp (1.833)\frac{25.1}{\sqrt{10}} = (156.0, 185.0)$$

- 99% interval is (144.7, 196.3)

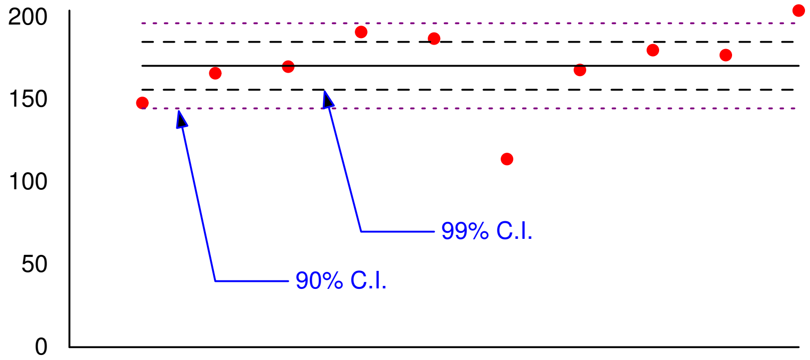
Example of t Distribution

- 10 height samples: 148, 166, 170, 191, 187, 114, 168, 180, 177, 204
- Sample mean $\overline{x} = 170.5$. Standard deviation $s = 25.1$, $n = 10$.
- 90% confidence interval is

$$170.5 \mp (1.833)\frac{25.1}{\sqrt{10}} = (156.0, 185.0)$$

- 99% interval is (144.7, 196.3)

Graph of *t* Distribution Example



Getting More Confidence

- Asking for a higher confidence level widens the confidence interval
  - Counterintuitive?
- How tall is Fred?
  - 90% sure he's between 155 and 190 cm
  - We want to be 99% sure we're right
  - So we need more room: 99% sure he's between 145 and 200 cm

Getting More Confidence
- Asking for a higher confidence level widens the confidence interval
  - Counterintuitive?
- How tall is Fred?
  - 90% sure he's between 155 and 190 cm
  - We want to be 99% sure we're right
  - So we need more room: 99% sure he's between 145 and 200 cm

Making Decisions

- Why do we use confidence intervals?
  - Summarizes error in sample mean
  - Gives way to decide if measurement is meaningful
  - Allows comparisons in face of error
- But remember: at 90% confidence, 10% of sample C.I.s do not include population mean
  - In other words, 10% of experiments give wrong answer!

Making Decisions

- Why do we use confidence intervals?
  - Summarizes error in sample mean
  - Gives way to decide if measurement is meaningful
  - Allows comparisons in face of error
- But remember: at 90% confidence, 10% of sample C.I.s do not include population mean
  - In other words, 10% of experiments give wrong answer!

Testing for Zero Mean

- Is population mean significantly $\neq 0$?
- If confidence interval includes 0, answer is *no*
- Can test for any value (mean of sums is sum of means)
- Our height samples are consistent with average height of 170 cm

Testing for Zero Mean

- Is population mean significantly $\neq 0$?
- If confidence interval includes 0, answer is *no*
- Can test for any value (mean of sums is sum of means)
- Our height samples are consistent with average height of 170 cm

Testing for Zero Mean

- Is population mean significantly $\neq 0$?
- If confidence interval includes 0, answer is *no*
- Can test for any value (mean of sums is sum of means)
- Our height samples are consistent with average height of 170 cm
  - Also consistent with 160 and 180!

Testing for Zero Mean
- Is population mean significantly $\neq 0$?
- If confidence interval includes 0, answer is *no*
- Can test for any value (mean of sums is sum of means)
- Our height samples are consistent with average height of 170 cm
  - Also consistent with 160 and 180!

Comparing Alternatives

- Often need to find better system
  - Choose fastest computer to buy
  - Prove our algorithm runs faster
- Different methods for paired/unpaired observations
  - *Paired* if i^{th} test on each system was same
  - *Unpaired* otherwise

Comparing Alternatives
- Often need to find better system
  - Choose fastest computer to buy
  - Prove our algorithm runs faster
- Different methods for paired/unpaired observations
  - *Paired* if i^{th} test on each system was same
  - *Unpaired* otherwise

Comparing Paired Observations

- For each test calculate performance difference
- Calculate confidence interval for differences
- If interval includes zero, systems aren't different
  - If not, sign indicates which is better

Comparing Paired Observations

- For each test calculate performance difference
- Calculate confidence interval for differences
- If interval includes zero, systems aren't different
  - If not, sign indicates which is better

Example: Comparing Paired Observations

- Do home baseball teams outscore visitors?
- Sample from 9-4-96:

H	4	5	0	11	6	6	3	12	9	5	6	3	1	6
V	2	7	7	6	0	7	10	6	2	2	4	2	2	0
H-V	2	-2	-7	5	6	-1	-7	6	7	3	2	1	-1	6

- Mean 1.4, 90% interval (-0.75, 3.6)
  - Can't tell from this data
  - 70% interval is (0.10, 2.76)
    - But tuning the interval to the data is guaranteed to produce wrong answers ("data snooping")

2015-06-15

CS147
└ Comparing Alternatives
  └ Paired Observations
    └ Example: Comparing Paired Observations

Example: Comparing Paired Observations
- Do home baseball teams outscore visitors?
- Sample from 9-4-96:
H 4 5 0 11 6 6 3 12 9 5 6 3 1 6
V 2 7 7 6 0 7 10 6 2 2 4 2 2 0
H-V 2 -2 -7 5 6 -1 -7 6 7 3 2 1 -1 6
- Mean 1.4, 90% interval (-0.75, 3.6)
  - Can't tell from this data
  - 70% interval is (0.10, 2.76)
    - But tuning the interval to the data is guaranteed to produce wrong answers ("data snooping")

Comparing Unpaired Observations

Start with confidence intervals

- If no overlap:
  - Systems are different and higher mean is better (for HB metrics)
- If overlap and at least one CI contains other's mean:
  - Systems are not different at this level
- If overlap and neither mean is in other CI
  - Must do *t*-test

Comparing Unpaired Observations

Start with confidence intervals
- If no overlap:
  - Systems are different and higher mean is better (for HB metrics)
- If overlap and at least one CI contains other's mean:
  - Systems are not different at this level
- If overlap and neither mean is in other CI
  - Must do *t*-test

The *t*-Test (1)

1. Compute sample means $\overline{x}_a$ and $\overline{x}_b$
2. Compute sample standard deviations s_a and s_b
3. Compute mean difference = $\overline{x}_a - \overline{x}_b$
4. Compute standard deviation of difference:

$$s = \sqrt{\frac{s_a^2}{n_a} + \frac{s_b^2}{n_b}}$$

The *t*-Test (1)
1. Compute sample means $\overline{x}_a$ and $\overline{x}_b$
2. Compute sample standard deviations s_a and s_b
3. Compute mean difference = $\overline{x}_a - \overline{x}_b$
4. Compute standard deviation of difference:
$$s = \sqrt{\frac{s_a^2}{n_a} + \frac{s_b^2}{n_b}}$$

The t-Test (2)

1. Compute effective degrees of freedom:

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

2. Compute the confidence interval:

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

3. If interval includes zero, no difference

The t-Test (2)
1. Compute effective degrees of freedom:
$$\nu = \frac{\left(s_a^2/n_a + s_b^2/n_b\right)^2}{\frac{1}{n_a+1}\left(\frac{s_a^2}{n_a}\right) + \frac{1}{n_b+1}\left(\frac{s_b^2}{n_b}\right)} - 2$$
2. Compute the confidence interval:
$$(\overline{x}_a - \overline{x}_b) \mp t_{[1-\alpha/2;\nu]}s$$
3. If interval includes zero, no difference

Comparing Proportions

- If k of n trials give a certain result, then confidence interval is:

$$\frac{k}{n} \mp z_{1-\alpha/2} \frac{\sqrt{k - k^2/n}}{n}$$

- If interval includes 0.5, can't say which outcome is statistically meaningful
- Must have $k \geq 10$ to get valid results

Comparing Proportions

- If k of n trials give a certain result, then confidence interval is:

$$\frac{k}{n} \mp z_{1-\alpha/2} \frac{\sqrt{k - k^2/n}}{n}$$

- If interval includes 0.5, can't say which outcome is statistically meaningful
- Must have $k \geq 10$ to get valid results

Selecting a Confidence Level

- Depends on cost of being wrong
- 90%, 95% are common values for scientific papers
- Generally, use highest value that lets you make a firm statement
  - But you must choose before you analyze data
  - And it's better to be consistent throughout a given paper

Selecting a Confidence Level

- Depends on cost of being wrong
- 90%, 95% are common values for scientific papers
- Generally, use highest value that lets you make a firm statement
  - But you must choose before you analyze data
  - And it's better to be consistent throughout a given paper

Hypothesis Testing

- The *null hypothesis* (H_0) is common in statistics
  - Confusing due to double negative
  - Gives less information than confidence interval
  - Often harder to compute
- Should understand that *rejecting* null hypothesis implies result *is* meaningful

Hypothesis Testing

- The *null hypothesis* (H_0) is common in statistics
  - Confusing due to double negative
  - Gives less information than confidence interval
  - Often harder to compute
- Should understand that *rejecting* null hypothesis implies result *is* meaningful

One-Sided Confidence Intervals

- Two-sided intervals test for mean being outside a certain range (see “error bands” in previous graphs)
- One-sided tests useful if only interested in one limit
- Use $z_{1-\alpha}$ or $t_{1-\alpha;n}$ instead of $z_{1-\alpha/2}$ or $t_{1-\alpha/2;n}$ in formulas

One-Sided Confidence Intervals

- Two-sided intervals test for mean being outside a certain range (see “error bands” in previous graphs)
- One-sided tests useful if only interested in one limit
- Use $z_{1-\alpha}$ or $t_{1-\alpha;n}$ instead of $z_{1-\alpha/2}$ or $t_{1-\alpha/2;n}$ in formulas

Sample Sizes

- Bigger sample sizes give narrower intervals
  - Smaller values of t, ν as n increases
  - $\sqrt{n}$ in formulas
- But sample collection is often expensive
  - What is minimum we can get away with?

Sample Sizes

- Bigger sample sizes give narrower intervals
  - Smaller values of t, ν as n increases
  - $\sqrt{n}$ in formulas
- But sample collection is often expensive
  - What is minimum we can get away with?

Choosing a Sample Size

- To get a given percentage error, $\pm r\%$ of the mean:

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

- Here, z represents either z or t as appropriate
- For a proportion $p = k/n$:

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

Choosing a Sample Size
- To get a given percentage error, $\pm r\%$ of the mean: $n = \left(\frac{100zs}{r\bar{x}}\right)^2$
- Here, z represents either z or t as appropriate
- For a proportion $p = k/n$: $n = z^2 \frac{p(1-p)}{r^2}$

Choosing a Sample Size for Comparisons

- Want to demonstrate system A is better than B (or vice versa)
- Must use same number of samples n for both systems
- Then we need:

$$\hat{n} \geq \left(\frac{z_{1-\alpha/2}(s_A + S_B)}{\overline{x}_B - \overline{x}_A} \right)^2$$

- For proportions, use p_A for $\overline{x}_A$, and $\sqrt{p_A(1 - p_A)}$ for s_A, etc.

Choosing a Sample Size for Comparisons

- Want to demonstrate system A is better than B (or vice versa)
- Must use same number of samples n for both systems
- Then we need:

$$\hat{n} \geq \left(\frac{z_{1-\alpha/2}(s_A + S_B)}{\overline{x}_B - \overline{x}_A} \right)^2$$

- For proportions, use p_A for $\overline{x}_A$, and $\sqrt{p_A(1 - p_A)}$ for s_A, etc.

Example of Choosing Sample Size

- Five runs of a compilation took 22.5, 19.8, 21.1, 26.7, 20.2 seconds
- How many runs to get $\pm 5\%$ confidence interval at 90% confidence level?
- $\overline{x} = 22.1$, $s = 2.8$, $t_{0.95;4} = 2.132$
- $n = \left(\frac{(100)(2.132)(2.8)}{(5)(22.1)} \right)^2 = 5.4^2 = 29.2$
- Note that first 5 runs can't be reused!

Example of Choosing Sample Size

- Five runs of a compilation took 22.5, 19.8, 21.1, 26.7, 20.2 seconds
- How many runs to get $\pm 5\%$ confidence interval at 90% confidence level?
- $\overline{x} = 22.1$, $s = 2.8$, $t_{0.95;4} = 2.132$
- $n = \left(\frac{(100)(2.132)(2.8)}{(5)(22.1)} \right)^2 = 5.4^2 = 29.2$
- Note that first 5 runs can't be reused!