Comparing Systems Using Sample Data

CS 700

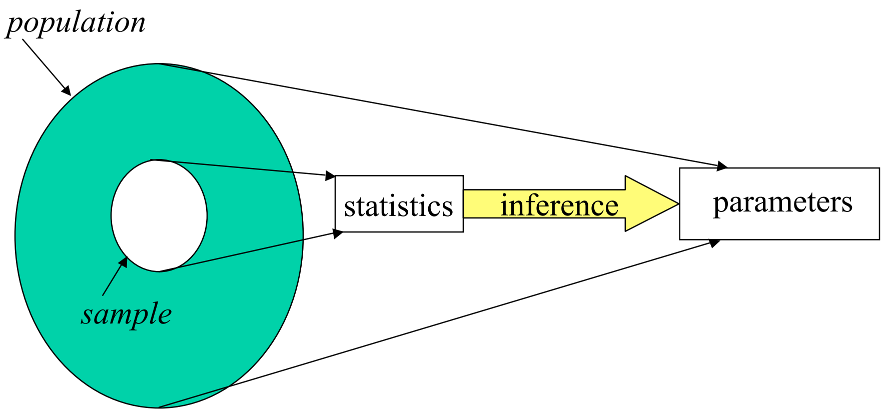
1

Acknowledgement

These slides are based on presentations created and copyrighted by Prof. Daniel Menasce (GMU)

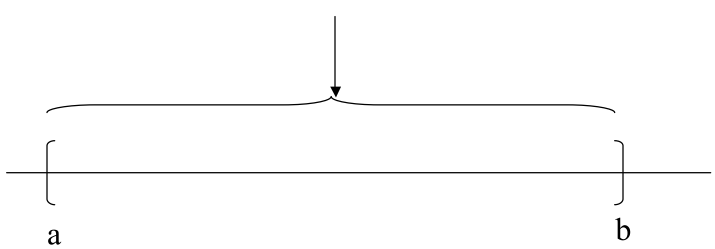
2

Statistical Inference



3

Interval Estimate



The interval estimate of the population parameter will have a specified confidence or probability of correctly estimating the population parameter.

4

Properties of Point Estimators

- Example of point estimator: sample mean.
- Properties:
  - Unbiasedness: the expected value of all possible sample statistics (of given size *n*) is equal to the population parameter.

$$E[\overline{X}] = \mu$$

$$E[s^2] = \sigma^2$$

  - Efficiency: precision as estimator of the population parameter.
  - Consistency: as the sample size increases the sample statistic becomes a better estimator of the population parameter.

5

Unbiasedness of the Mean

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

$$E[\overline{X}] = \frac{E\left[\sum_{i=1}^{n} X_i\right]}{n} = \frac{\sum_{i=1}^{n} E[X_i]}{n} =$$

$$\frac{\sum_{i=1}^{n} \mu}{n} = \frac{n\mu}{n} = \mu$$

6

Sample size= **15** **1.7% of population**

	Sample 1	Sample 2	Sample 3			
	0.0739	0.0202	0.2918			
	0.1407	0.1089	0.4696			
	0.1257	0.0242	0.8644			
	0.0432	0.4253	0.1494			
	0.1784	0.1584	0.4242			
	0.4106	0.8948	0.0051			
	0.1514	0.0352	1.1706			
	0.4542	0.1752	0.0084			
	0.0485	0.3287	0.0600			
	0.1705	0.1697	0.7820			
	0.3335	0.0920	0.4985			
	0.1772	0.1488	0.0988			
	0.0242	0.2486	0.4896			
	0.2183	0.4627	0.1892			
	0.0274	0.4079	0.1142	E[sample]	Population	Error
Sample Average	0.1718	0.2467	0.3744	0.2643	0.2083	26.9%
Sample Variance	0.0180	0.0534	0.1204	0.0639	0.0440	45.3%
Efficiency (average)	18%	18%	80%			
Efficiency (variance)	59%	21%	173%			

7

Sample size = **87** **10% of population**

	Sample 1	Sample 2	Sample 3			
	0.5725	0.3864	0.4627			
	0.0701	0.0488	0.2317			
	0.2165	0.0611	0.1138			
	0.6581	0.0881	0.0047			
	0.0440	0.5866	0.2438			
	0.1777	0.3419	0.0819			
	0.2380	0.1923	0.6581			
	0.0102	0.9460	0.0714			
	0.4325	0.0445	0.2959		Population	% Rel. Error
Sample Average	0.2239	0.2203	0.2178	0.2206	0.2083	5.9%
Sample Variance	0.0452688	0.0484057	0.0440444	0.0459	0.0440	4.3%
Efficiency (average)	7.5%	5.7%	4.5%			
Efficiency (variance)	2.9%	10.0%	0.1%			

8

Confidence Interval Estimation of the Mean

- Known population standard deviation.
- Unknown population standard deviation:
  - Large samples: sample standard deviation is a good estimate for population standard deviation. OK to use normal distribution.
  - Small samples and original variable is normally distributed: use *t* distribution with *n-1* degrees of freedom.

9

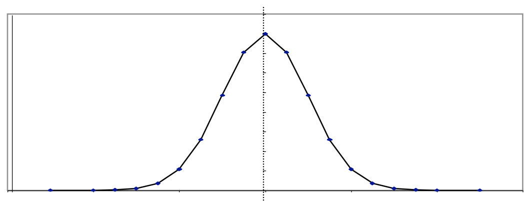
Confidence Interval Estimation of the Mean

$$\Pr[c_1 \leq \mu \leq c_2] = 1 - \alpha$$

(c_1,c_2): confidence interval
α: significance level (e.g., 0.05)
$1-\alpha$: confidence coefficient (e.g., 0.95)
$100(1-\alpha)$: confidence level (e.g., 95%)

10

μ



Sample	Interval include μ?
1	YES
2	YES
3	NO
…	
100	YES

100 (1 – α) of the 100 samples include the population mean μ.

11

Central Limit Theorem

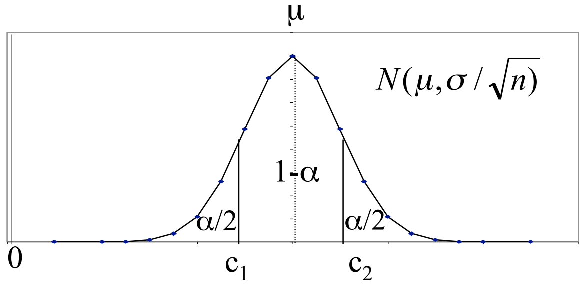
- If the observations in a sample are independent and come from the same population that has mean μ and standard deviation σ then the sample mean for **large** samples has a normal distribution with mean μ and standard deviation $\sigma/\sqrt{n}$.

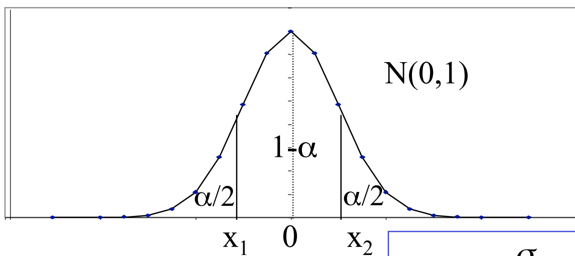
$$\bar{x} \sim N(\mu, \sigma / \sqrt{n})$$

- The standard deviation of the sample mean is called the *standard error.*

12

$c_2 = Q(1-\alpha/2)$
$c_1 = Q(\alpha/2)$





$x_2 = z_{1-\alpha/2}$
$x_1 = z_{\alpha/2} = -z_{1-\alpha/2}$

$$c_2 = \mu + \frac{\sigma}{\sqrt{n}} z_{1-\alpha/2}$$

$$c_1 = \mu + \frac{\sigma}{\sqrt{n}} z_{\alpha/2} = \mu - \frac{\sigma}{\sqrt{n}} z_{1-\alpha/2}$$

13

Confidence Interval (large (n>30) samples)

- 100 (1-α)% confidence interval for the population mean:

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

$\bar{x}$: sample mean
s: sample standard deviation
n: sample size
$z_{1-\alpha/2}$: (1-α/2)-quantile of a unit normal variate (N(0,1)).

14

	0.4325	0.0445	0.2959		**Population**
Sample Average	0.2239	0.2203	0.2178	0.2206	**0.2083**
Sample Variance	0.0452688	0.0484057	0.0440444	0.0459	**0.0440**
Efficiency (average)	7.5%	5.7%	4.5%		
Efficiency (variance)	2.9%	10.0%	0.1%		
95% interval lower	0.1792	0.1740	0.1737		
95% interval upper	0.2686	0.2665	0.2619	0.0894	
Mean in interval	YES	YES	YES		
99% interval lower	0.1651	0.1595	0.1598		
99% interval upper	0.2826	0.2810	0.2757	0.1175	
Mean in interval	YES	YES	YES		
90% interval lower	0.1864	0.1815	0.1807		
90% interval upper	0.2614	0.2591	0.2548	0.0750	
Mean in interval	YES	YES	YES		

In Excel:
_ interval = CONFIDENCE(1-0.95,s,n)

α

interval size

15

Confidence Interval (small samples, normally distributed population)

- 100 (1-α)% confidence interval for the population mean:

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

$\bar{x}$: sample mean
s: sample standard deviation
n: sample size
$t_{[1-\alpha/2;n-1]}$: critical value of the *t* distribution with *n-1* degrees of freedom for an area of $\alpha/2$ for the upper tail.

16

Student's t distribution

$$t(v) \sim \frac{N(0,1)}{\sqrt{\chi^2(v)/v}}$$

v: number of degree of freedom.

$\chi^2(v)$: chi-square distribution with v degrees of freedom. Equal to the sum of squares of v unit normal variates.

• the pdf of a t-variate is similar to that of a N(0,1).
• for $v > 30$ a t distribution can be approximated by N(0,1).

17

Confidence Interval (small samples)

- For samples from a normal distribution $N(\mu,\sigma^2)$, has a N(0,1) distribution and $(\bar{X}-\mu)/(\sigma/\sqrt{n})$ has a chi-square distribution with n-1 degrees of freedom $(n-1)s^2/\sigma^2$
- Thus, $(\bar{X}-\mu)/\sqrt{s^2/n}$ has a t distribution with n-1 degrees of freedom

18

Using the *t* Distribution. Sample size= 15.

	0.0274	0.4079	0.1142	E[sample]	Population	Error
Sample Average	0.1718	0.2467	0.3744	0.2643	0.2083	26.9%
Sample Variance	0.0180	0.0534	0.1204	0.0639	0.0440	45.3%
Efficiency (average)	18%	18%	80%			
Efficiency (variance)	59%	21%	173%			
95% interval lower	0.0975	0.1187	0.1823			
95% interval upper	0.2462	0.3747	0.5665			
Mean in inteval	YES	YES	YES			

95%,n-1 critical value 2.145

In Excel: TINV(1-0.95,15-1)

α

Confidence Interval for the Variance

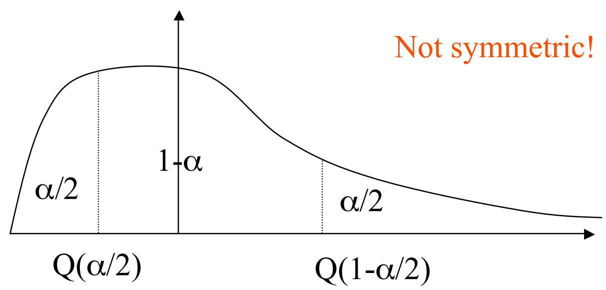
- If the original variable is normally distributed then the chi-square distribution can be used to develop a confidence interval estimate of the population variance.
- The (1-α)% confidence interval for σ^2 is

$$\frac{(n-1)s^2}{\chi_U^2} \le \sigma^2 \le \frac{(n-1)s^2}{\chi_L^2}$$

χ_L^2 : lower critical value of χ^2

χ_U^2 : upper critical value of χ^2

Chi-square distribution



21

95% confidence interval for the population variance for a sample of size 100 for a N(3,2) population.

2.91903	average
4.71435	variance
2.17126	std deviation
73.36110	lower critical value of chi-square for 95%
128.42193	upper critical value of chi-square for 95%

lower bound for confidence interval for the variance	3.634277
upper bound for confidence interval for the variance	6.361966

1-α/2

In Excel:
CHIINV (0.975, 99)
CHIINV (0.025, 99)

α/2

The population variance (4 in this case) is in the interval (3.6343, 6.362) with 95% confidence.

22

Confidence Interval for the Variance

If the population is not normally distributed, the confidence interval, especially for small samples, is not very accurate.

Confidence Interval for Proportions

- For categorical data:
  - E.g. file types {html, html, gif, jpg, html, pdf, ps, html, pdf …}
  - If n_1 of n observations are of type html, then the sample proportion of html files is $p = n_1/n$.
- The population proportion is π.
- Goal: provide confidence interval for the population proportion π.

Confidence Interval for Proportions

- The sampling distribution of the proportion formed by computing *p* from all possible samples of size *n* from a population of size *N* with replacement tends to a normal with mean π and standard error $\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}}$.
- The normal distribution is being used to approximate the binomial. So, $n\pi \geq 10$.

25

Confidence Interval for Proportions

- The (1-α)% confidence interval for π is

$$(p - z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}}, p + z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}})$$

p: sample proportion.
n: sample size
$z_{1-\alpha/2}$: (1-α/2)-quantile of a unit normal variate (N(0,1)).

26

Confidence Interval for Proportions

- One thousand entries are selected from a Web log. Six hundred and fifty correspond to gif files. Find 90% and 95% confidence intervals for the proportion of files that are gif files.

Sample size (n)	1000		
No. gif files in sample	650		
Sample proportion (p)	0.65		
n*p	650	> 10	OK

90% confidence interval

alpha	0.1
1-alpha/2	0.95
z0.95	1.645
Lower bound	0.625
Upper bound	0.675

In Excel:
NORMSINV(1-0.1/2)

95% confidence interval

alpha	0.05
1-alpha/2	0.975
z0.975	1.960
Lower bound	0.620
Upper bound	0.680

NORMSINV(1-0.05/2)

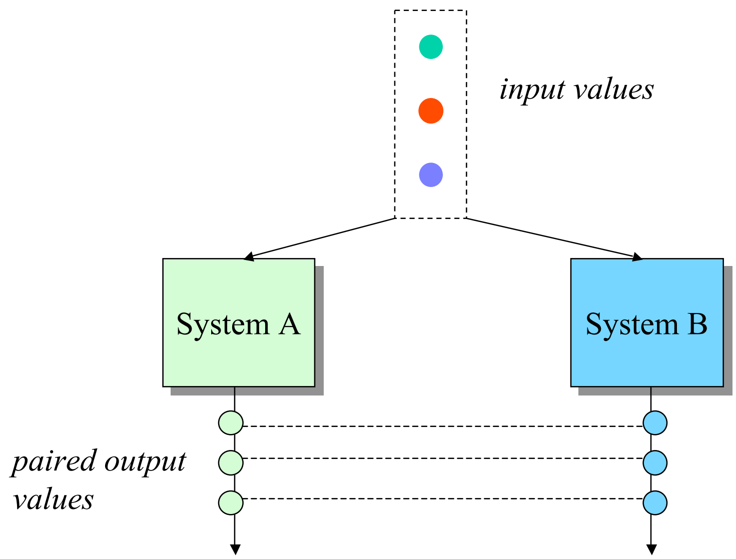
27

Comparing Alternatives

- Suppose you want to compare two cache replacement policies under similar workloads.
- Metric of interest: cache hit ratio.
- Types of comparisons:
  - Paired observations
  - Unpaired observations.

28

Paired Observations



29

Example of Paired Observations

- Six similar workloads were used to compare the cache hit ratio obtained under object replacement policies A and B on a Web server. Is A better than B?

Workload	Cache Hit Ratio		
	Policy A	**Policy B**	**A-B**
1	0.35	0.28	0.07
2	0.46	0.37	0.09
3	0.29	0.34	-0.05
4	0.54	0.60	-0.06
5	0.32	0.22	0.10
6	0.15	0.18	-0.03
	Sample mean		0.02000
	Sample variance		0.00552
	Sample standard dev.		0.07430

30

Example of Paired Observations

Sample mean	0.02000
Sample variance	0.00552
Sample standard dev.	0.07430

In Excel: TINV(1-0.9,5)

0.95 quantile of t-variable with 5 degrees of freedom	2.015
90% confidence interval	
lower bound	-0.0411
upper bound	0.0811

0.0743

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

0.02 2.015

6

31

Example of Paired Observations

Sample mean	0.02000
Sample variance	0.00552
Sample standard dev.	0.07430

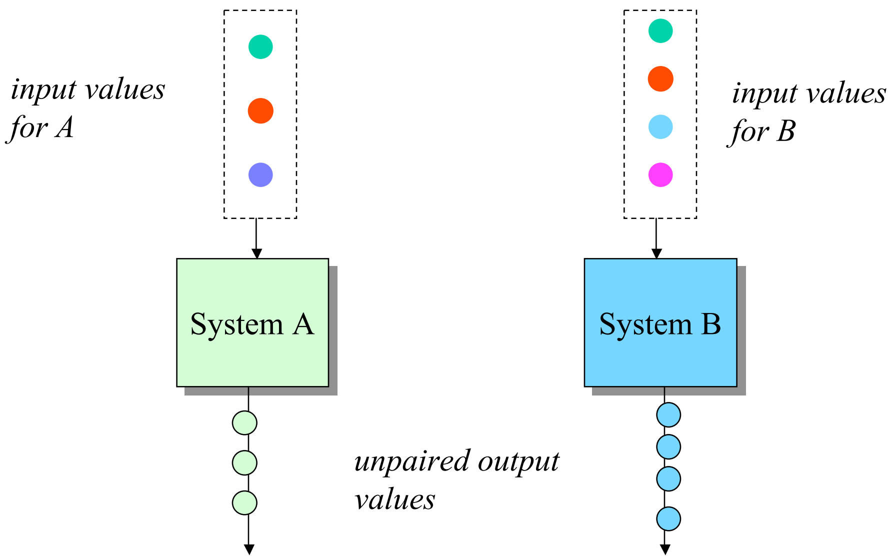
In Excel: TINV(1-0.9,5)

0.95 quantile of t-variable with 5 degrees of freedom	2.015
90% confidence interval	
lower bound	-0.0411
upper bound	0.0811

The interval includes zero, so we cannot say that policy A is better than policy B.

32

Unpaired Observations



33

Inferences concerning two means

- For large samples, we can statistically test the equality of the means of two samples by using the statistic

$$Z = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1}} + \sqrt{\frac{\sigma_2^2}{n_2}}}$$

  - Z is a random variable having the standard normal distribution.
  - We need to check if the confidence interval of Z at a given level includes zero
  - We can approximate the population variances above with sample variances when n_1 and n_2 are greater than 30

34

Inferences concerning two means (cont'd)

- For small samples, if the population variances are unknown, we can test for equality of the two means using the t-statistic below, provided we can assume that both populations are normal with equal variances

$$t = \frac{\bar{X}_1 - \bar{X}_2}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

  - t is a random variable having the t-distribution with $n_1 + n_2 - 2$ degrees of freedom and S_p is the square root of the pooled estimate of the variance of the two samples

$$S_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)}$$

35

Inferences concerning two means (cont'd)

- The pooled-variance t test can be used if we assume that the two population variances are equal
  - In practice, we can use it if one sample variance is less than 4 times the variance of the other sample
- If this is not true, we need another test
  - Smith-Satterthwaite test described in Jain (with some errors)

36

Unpaired Observations (t-test)

1. Size of samples for A and B: n_A and n_B
2. Compute sample means:

$$\bar{x}_A = \frac{1}{n_A}\sum_{i=1}^{n_A} x_{iA}$$

$$\bar{x}_B = \frac{1}{n_B}\sum_{i=1}^{n_B} x_{iB}$$

37

Unpaired Observations (t-test)

3. Compute the sample standard deviations:

$$s_A = \sqrt{\frac{\left(\sum_{i=1}^{n_A} x_{iA}^2\right) - n_A(\bar{x}_A)^2}{n_A - 1}}$$

$$s_B = \sqrt{\frac{\left(\sum_{i=1}^{n_B} x_{iB}^2\right) - n_B(\bar{x}_B)^2}{n_B - 1}}$$

38

Unpaired Observations (t-test)

4. Compute the mean difference: $\bar{x}_a - \bar{x}_b$
5. Compute the standard deviation of the mean difference:

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

6. Compute the effective number of degrees of freedom.

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

39

Unpaired Observations (t-test)

7. Compute the confidence interval for the mean difference:

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

8. If the confidence interval includes zero, the difference is not significant at $100(1-\alpha)\%$ confidence level.

40

Example of Unpaired Observations

- Two cache replacement policies A and B are compared under similar workloads. Is A better than B?

Workload	Cache Hit Ratio	
	Policy A	Policy B
1	0.35	0.49
2	0.23	0.33
3	0.29	0.33
4	0.21	0.55
5	0.21	0.65
6	0.15	0.18
7	0.42	0.29
8		0.35
9		0.44
Mean	0.2657	0.4011
St. Dev	0.0934	0.1447

41

Example of Unpaired Observations

na	7
nb	9
mean diff	-0.135
st.dev diff.	0.059776
Eff. Deg. Freed.	15
alpha	0.1
1-alpha/2	0.95
t[1-alpha/2,v]	1.782287

for 90% confidence interval

In Excel: TINV(1-0.9,15)

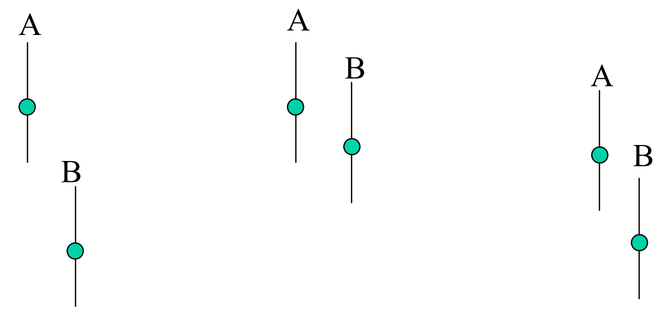
90% Confidence Interval

lower bound	-0.24193
upper bound	-0.02886

At a 90% confidence level the two policies are not identical since zero is not in the interval. With 90% confidence, the cache hit ratio for policy A is smaller than that for policy B. So, policy B is better at that confidence level.

42

Approximate Visual Test



CIs do not overlap: A is higher than B

CIs overlap and mean of A is in B's CI: A and B are similar

CIs overlap and mean of A is not in B's CI: need to do t-test

43

Example of Visual Test

Workload	Cache Hit Ratio	
	Policy A	Policy B
1	0.35	0.49
2	0.23	0.33
3	0.29	0.33
4	0.21	0.55
5	0.21	0.65
6	0.15	0.18
7	0.42	0.29
8		0.35
9		0.44
Mean	0.2657	0.4011
St. Dev	0.0934	0.1447

na	7		
nb	9		
alpha	0.1	for	90% confidence interval
1-alpha/2	0.95		
	Policy A	**Policy B**	
t[1-alpha/2,v]	1.9432	1.8595	
90% Confidence Interval			
lower bound	0.197	0.311	
upper bound	0.334	0.491	

CIs overlap but mean of A is not in CI of B and vice-versa. Need to do a t-test.

44

Non-parametric tests

- The unpaired t-tests can be used if we assume that the data in the two samples being compared are taken from normally distributed populations
- What if we cannot make this assumption?
  - We can make some normalizing transformations on the two samples and then apply the t-test
  - Some non-parametric procedure such as the Wilocoxon rank sum test that does not depend upon the assumption of normality of the two populations can be used

45

One-sided Confidence Intervals

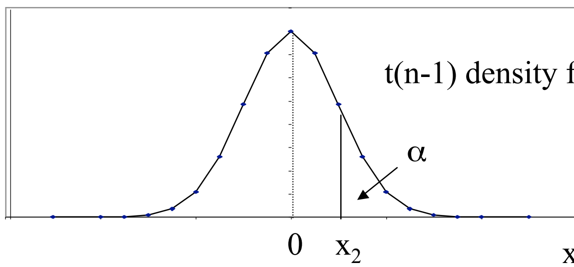
- Useful to test the hypothesis that the mean is greater (or smaller) than a certain value.

$$\Pr[\mu \geq c_1] = 1 - \alpha$$

$$\Pr[\mu \leq c_2] = 1 - \alpha$$

46

$c_2 = Q(1-\alpha)$



$x_2 = t_{1-\alpha;n-1}$

$$c_2 = \bar{x} + \frac{s}{\sqrt{n}} t_{[1-\alpha;n-1]}$$

47

One-sided Confidence Intervals

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

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

For large samples, use z-values instead of t-values

48

Determining Sample Size

- Large samples imply high confidence.
- Large samples require more data collection effort.
- How to determine the sample size *n* to estimate the population parameter with accuracy *r%* and confidence level of *100 (1-* α*)%?*

Determining the Sample Size for the Mean

- Perform a set of measurements to estimate the sample mean and the sample variance.
- Determine the sample size to obtain proper accuracy as follows:

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

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

Determining the Sample Size for the Mean

- A preliminary test shows that the sample mean of the response time is 5 sec and the sample standard deviation is 1.5. How many repetitions are needed to get the response time within 2% accuracy at 95% confidence level?

$r = 2 \quad \bar{x} = 5 \quad s = 1.5$

$z = 1.96$

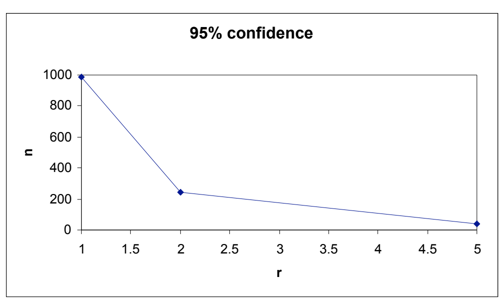
$$n = \left(\frac{100 \times 1.96 \times 1.5}{2 \times 5}\right)^2 = 864.36$$

865 repetitions would be Needed!

51

Determining the Sample Size for the Mean

Accuracy (r)	Confidence Level (1-alpha)	X	S	Sample size
1	0.95	5	0.8	984
2	0.95	5	0.8	246
5	0.95	5	0.8	40
1	0.9	5	0.8	693
2	0.9	5	0.8	174
5	0.9	5	0.8	28



52

Hypothesis testing vs estimating confidence intervals

- Textbooks on statistics devote a chapter to hypothesis testing
  - Example: Hypothesis test for a zero mean
  - Hypothesis test has a yes-no answer so either a hypothesis is accepted or rejected
  - Jain argues that confidence intervals provide more information
    - The difference between two systems has a confidence interval of (-100,100) vs a confidence interval of (-1,1)
    - In both cases, the interval includes zero but the width of the interval provides additional information

53

Computing Important Quantiles in Excel

$z_{1-\alpha/2}$ = (1-α/2)-quantile of a unit normal variate (N(0,1)):
= NORMINV (1-α/2,0,1) = NORMSINV(1- α/2)
Half-interval = CONFIDENCE (α,σ,n)

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

χ^2_L : lower critical value of χ^2 = CHIINV (1–α/2,n-1)

χ^2_U : upper critical value of χ^2 = CHIINV (α/2, n-1)

54