COMPARING SYSTEMS USING SAMPLE DATA

Gaia Maselli
maselli@di.uniroma1.it

SAPIENZA
UNIVERSITÀ DI ROMA

Introduction

- Summarizing sample data ➔ one sample set

- Comparing two systems ➔ two sample sets

Confidence intervals and sample size

- We use *estimation*: the process of estimating the value of a parameter from information obtained from a sample

Sample versus population

Example

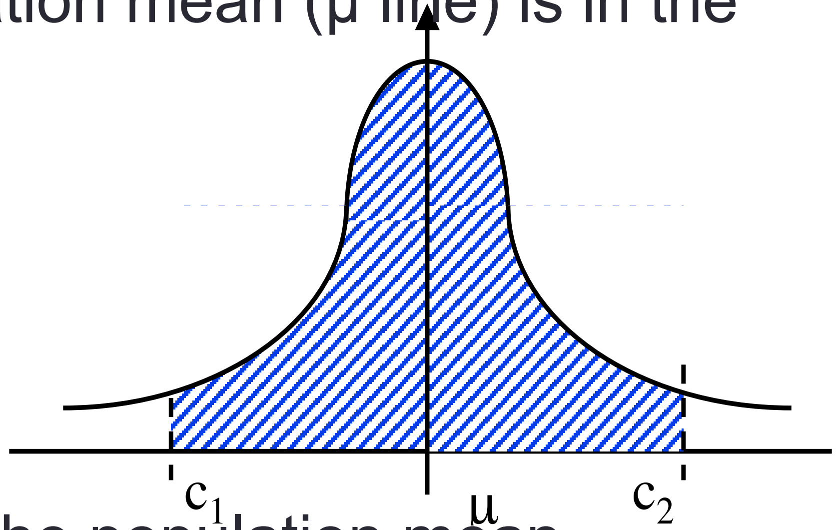
- We want to estimate the average age of students in CS
- We take 100 students and find average mean (23.3 years)
- There is a probability of $\bar{x}$ being right and a probability of being wrong based on the sample
- 23.3 is a *sample mean* , that can be used to estimate the *population mean μ*
- Sample mean will be somewhat different from the population mean

Confidence interval for the mean

- Suppose you take one sample and calculate the mean
- Then you take another sample
- Would the mean be the same?
- NO
- If you take a third sample you get a different mean, etc.
- You can plot the distribution of the sample mean
- Each sample mean is an estimate of the population mean (it has a variability around it)
- How good are the estimates? (accuracy of estimates)
- How can we get a single estimate of the population mean from *k* estimates?
- It is not possible to get a perfect estimate of the population mean from any finite number of finite size samples
- But we can get probabilistic bounds (interval containing the population mean with some probability)

Confidence interval for the mean

- We can get two bounds, c_1 and c_2, such that there is **high** probability, $1\text{-}\alpha$, that the population mean (μ line) is in the interval (c_1,c_2):



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

- (c_1,c_2): *confidence interval* for the population mean
- *α: significance level*
- *100(1-α)*: *confidence level* (percentage typically near 100%)
- *1-α: confidence coefficient* (e.g., 0.05 or 0.1)

How to determine the confidence interval from *k* samples

- One way to determine the 90% confidence interval would be to use the 5-percentile and 95-percentile of the sample means as the bounds
- Example: we take *k* samples, find sample means, sort them out in an increasing order, and take the [1+0.05(k-1)]th and [1+0.95(k-1)]th element of the sorted set
- But we have to get k samples…

How to determine the confidence interval from *one* sample

- If we want to determine the confidence interval without gathering many samples, but from just one sample
- It is possible because of the **central limit theorem**: if the observations in a sample $\{x_1, x_2, \ldots, x_n\}$ are independent and come from the same population that has a mean μ and a standard deviation σ, then the sample mean for large samples is approximately *normally distributed* with mean μ and standard deviation $\sigma/\sqrt{n}$

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

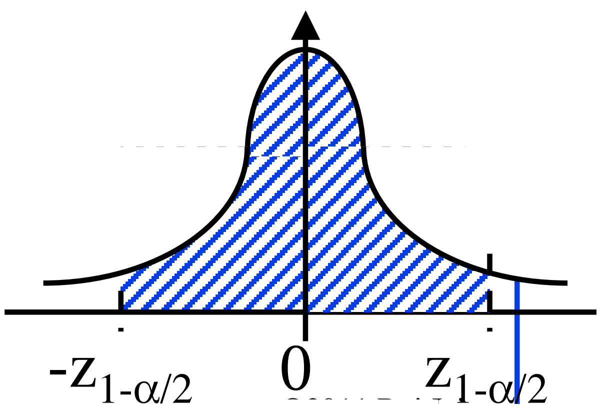
How to determine the confidence interval from *one* sample

- A 100(1-α)% confidence interval for the population mean is given by

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

- $\bar{x}$ is the sample mean
- *s* is the sample standard deviation
- $Z_{1-\alpha/2}$ is the (1-α/2)-quantile of a normal variate

(quantiles are listed in table A.2 of the book)



How to determine the confidence interval from *one* sample - example

- Given the sample {3.1, 4.2, 2.8, 5.1, 2.8, 4.4, 5.6, 3.9, 3.9, 2.7, 4.1, 3.6, 3.1, 4.5, 3.8, 2.9, 3.4, 3.3, 2.8, 4.5, 4.9, 5.3, 1.9, 3.7, 3.2, 4.1, 5.1, 3.2, 3.9, 4.8, 5.9, 4.2}
- The mean is $\bar{x} = 3.90$ (calculated)
- The standard deviation is $s = 0.95$ (calculated)
- *n=32* (known)
- A 90% confidence interval for the mean is

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

$$3.90 \mp (1.645)(0.95) / \sqrt{32} = (3.62, 4.17)$$

- We can state with 90% confidence that the population mean is between 3.62 and 4.17

(1-α/2)-quantile of a unit normal variate

- 90% confidence interval
- $\alpha=0.1$
- $\alpha/2=0.05$
- $1-\alpha/2 = 0.95$
- Check *z* value on table of quantiles of the Unit Normal Distribution

$$z_{1-\alpha/2} = 1.645$$

z_p

TABLE A.2 Quantiles of the Unit Normal Distribution

p	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.5	0.000	0.025	0.050	0.075	0.100	0.126	0.151	0.176	0.202	0.228
0.6	0.253	0.279	0.305	0.332	0.358	0.385	0.412	0.440	0.468	0.496
0.7	0.524	0.553	0.583	0.613	0.643	0.674	0.706	0.739	0.772	0.806
0.8	0.842	0.878	0.915	0.954	0.994	1.036	1.080	1.126	1.175	1.227

p	0.000	0.001	0.002	0.003	0.004	0.005	0.006	0.007	0.008	0.009
0.90	1.282	1.287	1.293	1.299	1.305	1.311	1.317	1.323	1.329	1.335
0.91	1.341	1.347	1.353	1.359	1.366	1.372	1.379	1.385	1.392	1.398
0.92	1.405	1.412	1.419	1.426	1.433	1.440	1.447	1.454	1.461	1.468
0.93	1.476	1.483	1.491	1.499	1.506	1.514	1.522	1.530	1.538	1.546
0.94	1.555	1.563	1.572	1.580	1.589	1.598	1.607	1.616	1.626	1.635
0.95	1.645	1.655	1.665	1.675	1.685	1.695	1.706	1.717	1.728	1.739
0.96	1.751	1.762	1.774	1.787	1.799	1.812	1.825	1.838	1.852	1.866
0.97	1.881	1.896	1.911	1.927	1.943	1.960	1.977	1.995	2.014	2.034
0.98	2.054	2.075	2.097	2.120	2.144	2.170	2.197	2.226	2.257	2.290

p	0.0000	0.0001	0.0002	0.0003	0.0004	0.0005	0.0006	0.0007	0.0008	0.0009
0.990	2.326	2.330	2.334	2.338	2.342	2.346	2.349	2.353	2.357	2.362
0.991	2.366	2.370	2.374	2.378	2.382	2.387	2.391	2.395	2.400	2.404
0.992	2.409	2.414	2.418	2.423	2.428	2.432	2.437	2.442	2.447	2.452
0.993	2.457	2.462	2.468	2.473	2.478	2.484	2.489	2.495	2.501	2.506
0.994	2.512	2.518	2.524	2.530	2.536	2.543	2.549	2.556	2.562	2.569
0.995	2.576	2.583	2.590	2.597	2.605	2.612	2.620	2.628	2.636	2.644
0.996	2.652	2.661	2.669	2.678	2.687	2.697	2.706	2.716	2.727	2.737
0.997	2.748	2.759	2.770	2.782	2.794	2.807	2.820	2.834	2.848	2.863
0.998	2.878	2.894	2.911	2.929	2.948	2.968	2.989	3.011	3.036	3.062
0.999	3.090	3.121	3.156	3.195	3.239	3.291	3.353	3.432	3.540	3.719

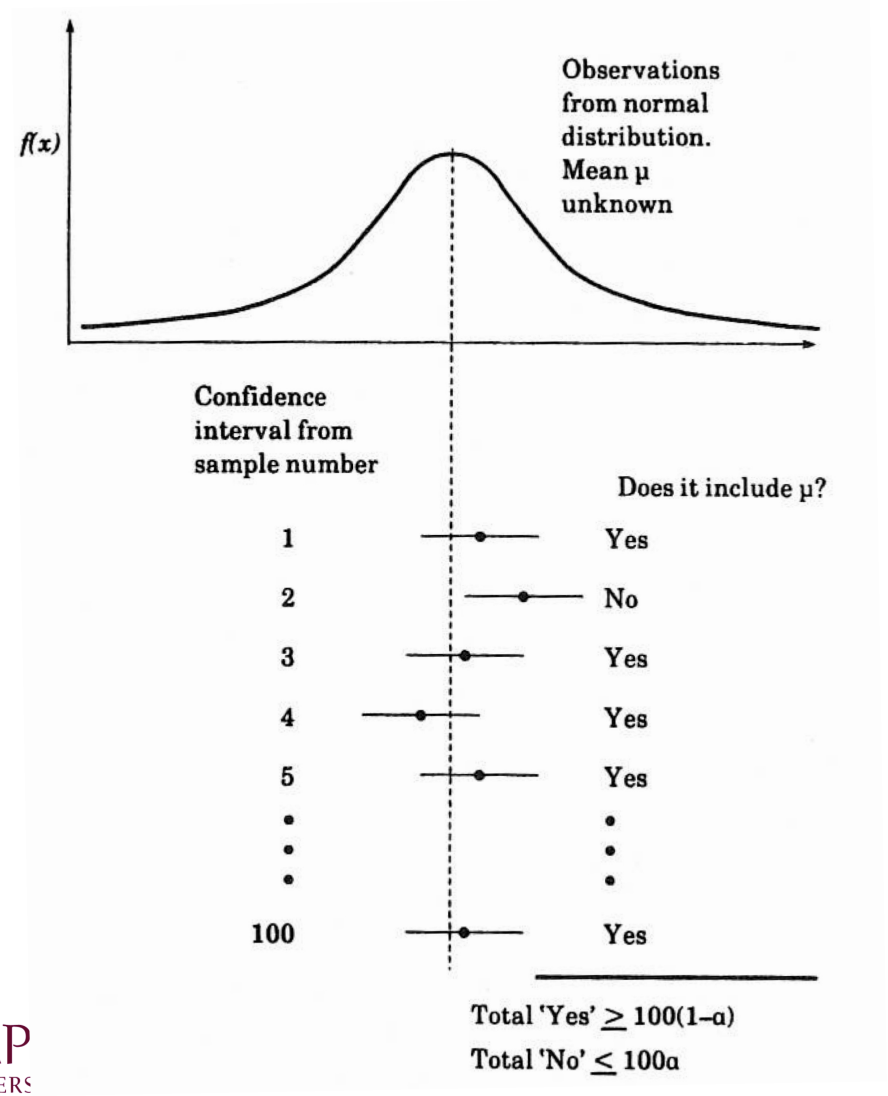
Exercises

A 95% confidence interval for the mean = ?

A 99% confidence interval for the mean = ?

Confidence interval: meaning

- Stating with 90% confidence that the population mean is between c_1 and c_2 means that the chance of error is 10%
- If we take 100 samples and construct a confidence interval for each sample, in 90 cases the interval would include the population mean and in 10 cases the interval would not include the population mean



Confidence interval for small samples (n<30)

- 100(1-α) % confidence interval for n < 30 is given by

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

- Where $t_{[1-\alpha/2;n-1]}$ is the (1-α/2)-quantile of a *t*-variate with *n-1* degrees of freedom
- The interval is based on the fact that for samples from a normal population N(μ,σ²), $(\bar{x}-\mu)/(\sigma/\sqrt{n})$ has a N(0,1) distribution and $(n-1)s^2/\sigma^2$ has a *chi-square distribution* with *n-1* degrees of freedom, and therefore $(\bar{x}-\mu)/(\sqrt{s^2/n})$ has a *t distribution* with n-1 degrees of freedom

SAPIENZA
UNIVERSITÀ DI ROMA

Example

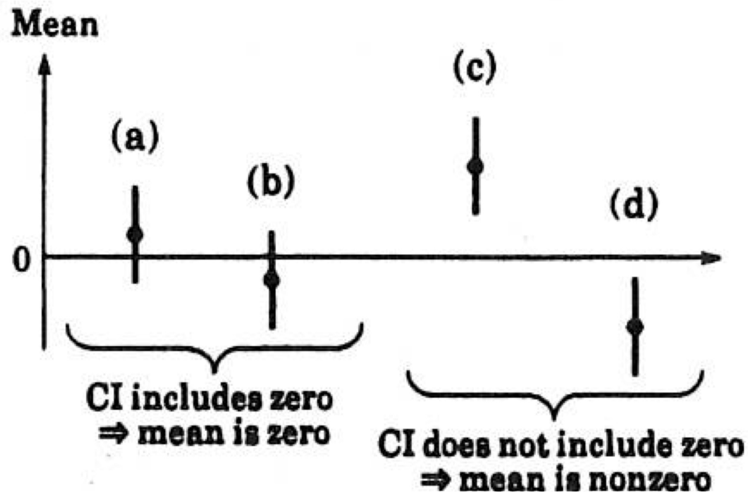
- The difference between the values measured on a system and those predicted by a model is called modeling error.
- The modeling error for eight predictions of a model were found to be -0.04, -0.19, 0.14, -0.09, -0.14, 0.19, 0.04, 0.09
- The mean of these values is zero and their sample standard deviation is 0.138.
- The $t_{[0.95;7]}$ from Table A.4 is 1.895
- Thus the confidence interval for the mean error is

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

SAPIENZA
UNIVERSITÀ DI ROMA

Testing for a zero mean

- A common use of confidence intervals is to check if a measured value is significantly different from zero
- If the measured value passes out test of difference with a probability greater than or equal to the specified level of confidence, 100(1-α)%, then the value is significantly different from zero
- The test consists of determining a confidence interval and simply checking if the interval includes zero



Example

- 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

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

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

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

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

Example (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.

Testing if a mean is different from a value *a*

- The procedure for testing for a zero mean applies equally well to any other value as well
- To test if the mean is equal to a given value *a*, a confidence interval is constructed as before, and if the interval includes *a*, then the hypothesis that the mean is equal to *a* cannot be rejected at the given level of confidence

- Example: if I get a confidence interval (-1.21,3.27) at 99% confidence level and a=1, then as the interval includes 1 the mean can be 1.

Comparing two alternatives

Paired vs. unpaired comparisons

- Paired: if we conduct n experiments on each of the two systems such that there is a 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, n on System B (or the same but in different order)
  - ⇒Need more sophisticated method

Paired observations

- *n* paired observations
- The analysis of paired observation is straightforward
- The two samples are treated as one sample of *n* pairs
- For each pair, the difference in performance can be computed
- A confidence interval can be constructed for the difference
- If the CONFIDENCE INTERVAL includes ZERO

  ⇒ the systems are NOT SIGNIFICANTLY DIFFERENT

Example

- 6 similar workloads were used on two systems.
- 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

- Suppose we have two samples of size n_a and n_b for alternatives A and B, respectively
- The observations are unpaired in the sense that there is no correspondence between *i*th observations in the two samples
- There is a procedure called ***t*-test** to determine the confidence interval for the difference in mean performance

Unpaired observations: *t*-test

1. Compute the sample means

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

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

2. 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: *t*-test

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

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

5. Compute the effective number of degrees of freedom

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

Unpaired observations: *t*-test

6. Compute the confidence interval for the mean difference

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

(1-α/2)-quantile of a t-variate with *v* degrees of freedom

7. If the confidence interval includes zero, the difference is not significant at 100(1-α)% confidence level
8. If the confidence interval does not include zero, then the sign of the mean difference indicates which system is better

Example

- The processor time required to execute a task was measured on two systems. The times on system A were {5.36, 16.57, 0.62, 1.41, 0.64, 7.26}. The times on system B were {19.12, 3.52, 3.38, 2.50, 3.60, 1.74}. 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 (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 $f = 11.921$

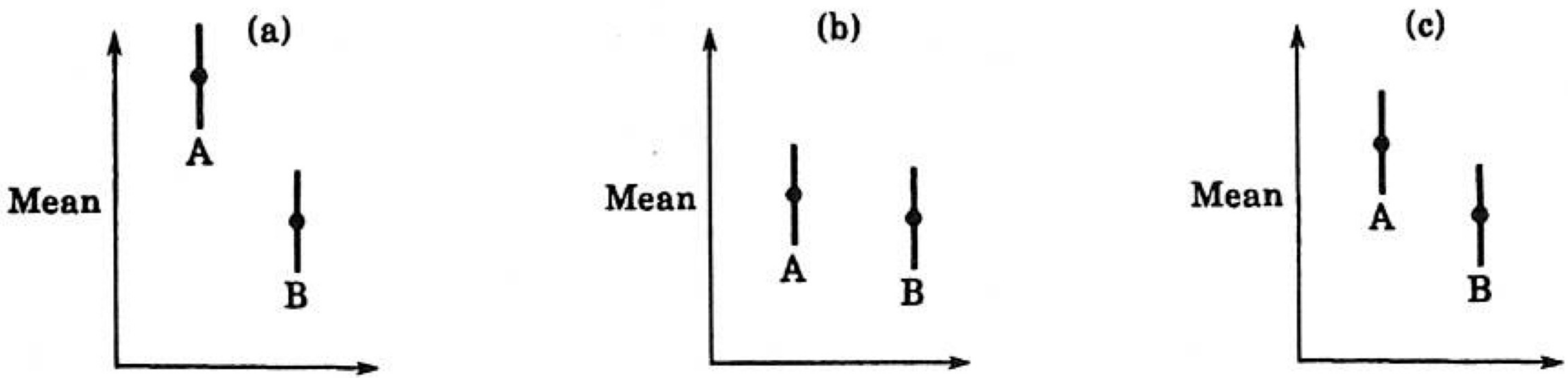
The 0.95-quantile of a t-variate with 12 degrees of freedom $= 1.71$

The 90% confidence interval for the difference $= (-6.92, 6.26)$

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

Approximate visual test

- A simpler visual test to compare two unpaired samples
- Simply compute the confidence interval for each alternative separately and compare them



CIs do not overlap
⇒ A is higher than B

CIs overlap and mean of one is in the CI of the other
⇒ alternatives are not different

CIs overlap but mean of any one is not in the CI of the other
⇒ need to do the *t*-test

In the case of the last 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 = $5.31 \mp (2.015)\sqrt{(37.92/6)}$
  $= (0.24, 10.38)$
- The 90% confidence interval for the mean of B = $5.64 \mp (2.015)\sqrt{(44.11/6)}$
  $= (0.18, 11.10)$
- Confidence intervals overlap and the mean of one falls in the confidence interval for the other.
  $\Rightarrow$ Two systems are not different at this level of confidence.

SAPIENZA UNIVERSITÀ DI ROMA

What confidence level to use

- Can be 90% or 95% or 99% or any other value
- 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.

One-sided confidence intervals

- Two side intervals: 90% Confidence
- ⇒ P(Difference > upper limit) = 5%
- ⇒ P(Difference < Lower limit) = 5%
- Sometimes only one-sided comparison is desired
- Example: is the mean greater than a certain value (e.g., zero)?
- One-sided lower confidence interval for μ is given by

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

- One-sided upper confidence interval for μ is given by

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

- For large samples use z instead of t

SAPIENZA
UNIVERSITÀ DI ROMA

Confidence interval for proportions

- For categorical variables, we have probabilities associated with various categories
- Estimation of proportions is very similar to estimation of means
- Each sample of n observations gives a sample proportion
- We need to obtain a confidence interval to get a bound
- Given that n_1 of n observations are of type 1, a confidence interval for the proportion is obtained as follows
- Sample proportion=p=n_1/n
- Confidence interval for proportion = $p \mp z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}}$
- $z_{1-\alpha/2}$ is the (1-α/2)-quantile of a unit normal variate
- Condition: np>=10

SAPIENZA
UNIVERSITÀ DI ROMA

Determining sample size

- The confidence level of conclusions drawn from a set of measured data depends upon the size of the data set
- The larger the sample, the higher is the associated confidence
- But larger samples require more effort and resources
- Analyst's goal: to find the smallest sample size that will provide the desired confidence
- There are formulas for determining the sample sizes required to achieve a given level of accuracy and confidence
- We consider three different cases

1. Single system measurement
2. Proportion determination
3. Two-system comparison

- In each case, a small set of preliminary measurements are done to estimate the variance, which is used to determine the sample size required for the given accuracy

Sample size for determining the mean of a single system

- We want to estimate the mean performance of a system with an accuracy of ±r% and a confidence level of 100(1-α)%
- The number of observations *n* required to achieve this goal can be determined as follows:
- For sample size = n, the 100(1-α)% confidence interval of the population mean is

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

- The desired accuracy of r% implies that the confidence interval should be

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

- Equating the desired interval with that obtained with n observations we can determine *n*

Sample size for determining the mean of a single system (Cont)

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

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

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

z is the normal variate of the desired confidence level

Example

- Based on a preliminary test, the sample mean of the response time is 20 seconds, and the sample standard deviation is 5. How many repetitions are needed to get the response time accurate within 1 second at 95% confidence?

Required confidence =1 in 20 = 5% $\bar{x}\frac{r}{100}=1 \quad 20\frac{r}{100}=1 \quad r=\frac{100}{50}=5$

X=20, s=5, z=1.960, 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

Sapienza Università di Roma

Sample size for determining proportions

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

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

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

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

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

SAPIENZA
UNIVERSITÀ DI ROMA

Sample size for comparing two alternatives

- Two packet-forwarding algorithms were measured. Preliminary measurements showed that:

- 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 comparing two alternatives

- 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.