Performance of Computer System

Sample Data

Dr. Lina Xu
lina.xu@ucd.ie

School of Computer Science,
University College Dublin

October 18, 2019

Sample vs. Population

Suppose we generate a set S containing several million random numbers. We will call this set the population.

- Denote population mean with μ
- Denote population std deviation with ρ

Draw a sample of n numbers $\{x_1, x_2, ..., x_n\}$ from S

- Denote sample mean with $\bar{x}$
- Denote the STD of the sample with s

No guarantee that $\bar{x} = \mu$ and $s = \rho$
($\bar{x}$, s) of the samples are estimates of the population parameters (μ, ρ)

Sample vs. Population

Conventions

- population characteristics: parameters (μ, ρ) (Greek alphabet)
- sample estimates: statistics ($\bar{x}$, s) (Roman alphabet)

Normal Distribution

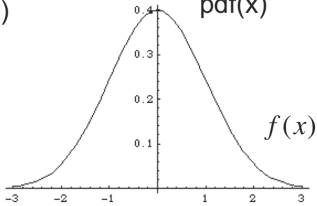
N(μ, ρ) most commonly used distribution in data analysis

$$\text{pdf} = \quad f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}, -\infty \le x \le \infty$$

μ = mean
σ = std dev

(also known as a Gaussian distribution)

N(μ=0,σ=1)
unit normal
distribution



$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

Confidence Interval for the Mean

- k samples of a population may yield k different sample means
- No sample or finite set of samples will necessarily give a perfect estimate for the population mean μ
- Instead, we use probability bounds for an estimate of μ (the population mean)
  - $P(c_1 \leq \mu \leq c_2] = 1 - \alpha$
- Confidence interval (c1,c2)

Confidence Interval Example

- Say you were interested in the mean weight of 10-year-old girls living in the United States.
- Since it would have been impractical to weigh all the 10-year-old girls in the United States, you took a sample of 16 and found that the mean weight was 90 pounds.
- This sample mean of 90 is a point estimate of the population mean.
- A point estimate by itself is of limited usefulness because it does not reveal the uncertainty associated with the estimate; you do not have a good sense of how far this sample mean may be from the population mean.
- For example, can you be confident that the population mean is within 5 pounds of 90? You simply do not know.

Confidence Interval Example

- Confidence intervals provide more information than point estimates.
- An example of a 95% confidence interval is shown below:
  - $72.85 < \mu < 107.15 \Rightarrow P(72.85 < \mu < 107.15) = 95\%$
- There is good reason to believe that the population mean lies between these two bounds of 72.85 and 107.15 since 95% of the time confidence intervals contain the true mean.

Confidence Interval for the Mean

- k samples of a population may yield k different sample means
- No sample or finite set of samples will necessarily give a perfect estimate for the population mean μ
- Instead, we use probability bounds for an estimate of μ (the population mean)

$P(c_1 \leq \mu \leq c_2] = 1 - \alpha$

- Confidence interval (c1,c2)
- Significance level: α
- Confidence coefficient: 1- α
- Confidence level (a percentage): 100 (1- α)

Understanding Confidence Intervals

Why use them?

- provide a way to decide if measurements are meaningful – characterise potential error in sample mean
- enable comparisons in the presence of experimental error

Understand their limitations!

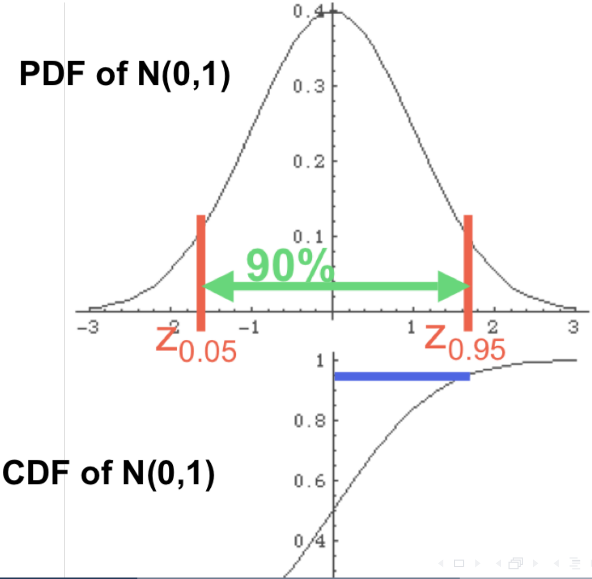
- at 95% confidence, confidence intervals for 5% of sample means do not include the population mean μ

Computing (c1,c2) for Population Mean μ–The hard way

- Collect a large number of samples
- To compute a 90% confidence interval for a population mean μ
  - Take k samples of the population (each sample is a set)
  - Compute the set of sample means (one for each sample)
  - Sort the set of sample means
  - Select the $[1 + .05(k-1)]_{th}$ element as c_1
  - Select the $[1 + .95(k-1)]_{th}$ element as c_2
  - 90% confidence interval for μ is $(c_1\ ,\ c_2\ )$
- $90\% = 100(1\text{-}\ \alpha)\ ,\ \alpha = 0.1$
- $0.05 = \alpha/2$
- $0.95 = 1\text{-}\ \alpha/2$

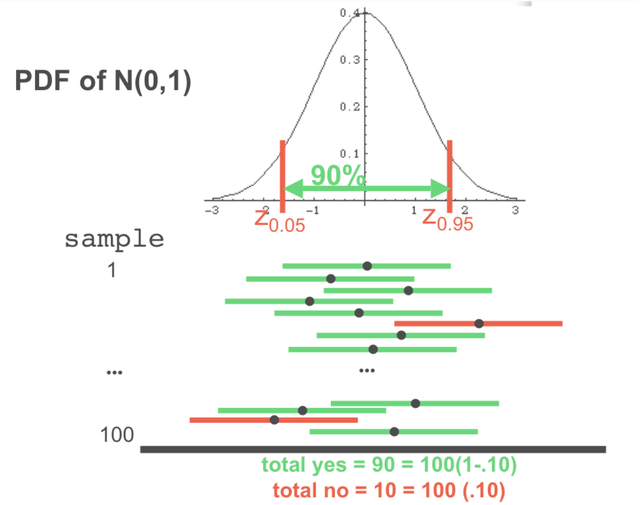
Confidence Interval of a Normal Distribution

- Example: 90% confidence interval, $\alpha = 0.10$



Confidence Interval of a Normal Distribution

- Example: 90% confidence interval, $\alpha = 0.10$



The Central Limit Theorem

- If observations $x_1, x_2, ..., x_n$ are – independent
  - from the same population
  - the population has mean μ
  - the population has STD ρ
- Then sample mean $\bar{x}$ for large samples is approximately normally distributed
  - $\bar{x} = \sim N(\mu, \frac{\rho}{\sqrt{n}})$
- If we define: STD error = STD of sample mean
- If population std deviation is ρ, STD error is $\frac{\rho}{\sqrt{n}}$
- From this expression, it is easy to see that as the sample size n increases, the standard error decreases.

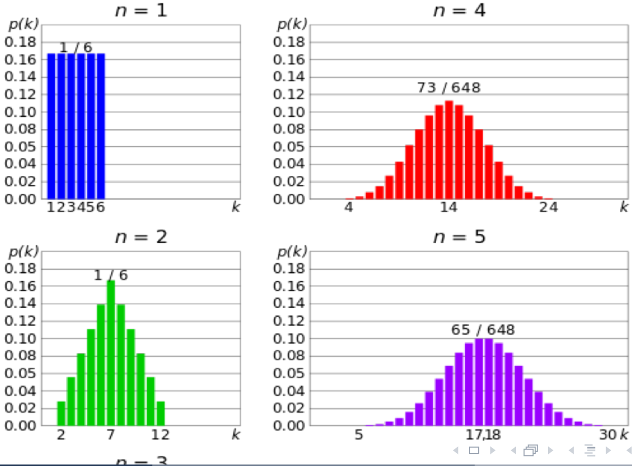
The Central Limit Theorem-Large Sample

How large is "large enough"? The answer depends on two factors.

- Requirements for **accuracy**. The more closely the sampling distribution needs to resemble a normal distribution, the more sample points will be required.
- The shape of the **underlying population**. The more closely the original population resembles a normal distribution, the fewer sample points will be required.

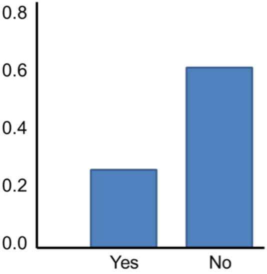
The Central Limit Theorem – Example I

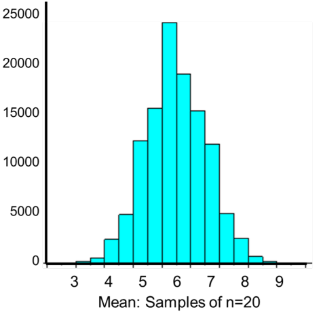
- One of the simplest types of test: rolling a fair die.
- The more times you roll the die, the more likely the shape of the distribution of the means tends to look like a normal distribution graph.



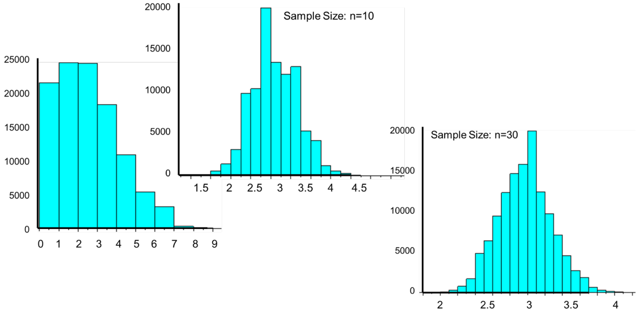
The Central Limit Theorem – Example II

- Success of a medical procedure: yes or no with 30% of the population classified as a success as shown below.
- The distribution of sample means based on samples of size n=20.





The Central Limit Theorem – Example III



Computing (c1,c2) for Population Mean μ–Normally used

- Fortunately, it is not necessary to gather too many samples. It is possible to determine the confidence interval from just one sample because of the central limit theorem.
- When you compute a confidence interval on the mean, you compute the mean of a sample in order to estimate the mean of the population.
- Clearly, if you already knew the population mean, there would be no need for a confidence interval.
- However, to explain how confidence intervals are constructed, we are going to work backwards and begin by assuming characteristics of the population.

Confidence Interval Example

- Assume that the weights of 10-year-old children are normally distributed with a mean of $\mu = 90$ and a standard deviation of $\rho = 36$.
- Then the sample distribution will be normally distribution with a mean of μ and a standard deviation of $\frac{\rho}{\sqrt{n}}$, where n is the size of the sample.
  - Supposedly n=9
  - Then $\frac{\rho}{\sqrt{n}} = 12$
- The shaded area represents the middle 95% of the distribution and stretches from 66.48 to 113.52.
  - 90 - 1.96*12= 66.48
  - 90 + 1.96*12 = 113.52
- The value of 1.96 is based on the fact that 95% of the area of a normal distribution is within 1.96 standard deviations of the mean;

Confidence Interval Example

- Now let's work from the sample data! Consider the probability that a sample mean computed from a random sample is within 23.52 (= 1.96*12)) units of the population mean of 90.
- Since 95% of the distribution is within 23.52 of 90, the probability that the mean from any given sample will be within 23.52 of 90 is 0.95.
- This means that if we repeatedly compute the mean (M) from a sample, and create an interval ranging from M - 23.52 to M + 23.52, this interval will contain the population mean 95% of the time.

Computing (c_1, c_2) for Population Mean μ

The easy way (for a large sample, n $>$ 30)

- By the central limit theorem, a $100(1-\alpha)\%$ confidence interval for μ
- $(\bar{x} - z_{1-\alpha/2}\frac{s}{\sqrt{n}}, \bar{x} + z_{1-\alpha/2}\frac{s}{\sqrt{n}})$
  - Where $\bar{x}$ is the sample mean !
  - s is the sample std deviation
  - n is the sample size
  - α is the significance level, $100(1-\alpha)\%$ is the confidence level
  - $z_{1-\alpha/2}$ is the $(1-\alpha/2)$ quantile of the unit normal variate
  - Normal distribution table (Z-values)

Confidence Interval Example

- Given a (large) sample with the following characteristics
  - 32 elements (n = 32)
  - sample mean x = 3.90
  - sample std deviation s = 0.71
- A 90% confidence interval for the mean can be computed as
  - $(\bar{x} - z_{1-\alpha/2}\frac{s}{\sqrt{n}}, \bar{x} + z_{1-\alpha/2}\frac{s}{\sqrt{n}})$
  - $(\bar{x} - z_{0.95}\frac{s}{\sqrt{n}}, \bar{x} + z_{0.95}\frac{s}{\sqrt{n}})$
- Recall $z_{1-\alpha/2}$ is approximately 1.645

Computing (c1,c2) for Population Mean μ

The easy way (for a small sample, n $<$ 30)

- For large set
  - $(\bar{x} - z_{1-\alpha/2}\frac{s}{\sqrt{n}}, \bar{x} + z_{1-\alpha/2}\frac{s}{\sqrt{n}})$
- For small set
  - $(\bar{x} - t_{[1-\alpha/2;n-1]}\frac{s}{\sqrt{n}}, \bar{x} + t_{[1-\alpha/2;n-1]}\frac{s}{\sqrt{n}})$
  - Using t-distribution table
  - For instance, if our sample size were n, then the number of degrees of freedom to be used in calculations would be n - 1.
  - To calculate the degrees of freedom (df) for a sample size of n=8 we would subtract 1 from 8 (df=8-1=7).
  - For the previous example, a 90% confidence interval is 1.895

When to use t distribution table rather than normal distribution table

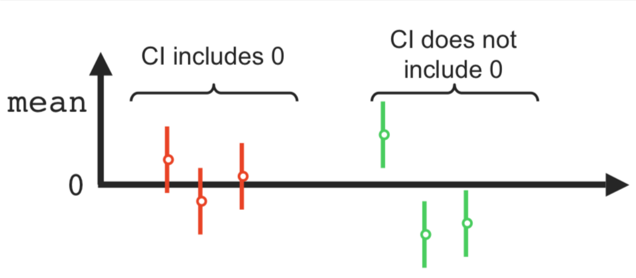
- You must use the t-distribution table when working problems when the population standard deviation (ρ) is not known and the sample size is small (n$<$30).
- If ρ known, then use normal.
- If ρ not known:
  - If n is large, then use normal.
  - If n is small, then use t-distribution.

Small vs. Large Samples

- Why the difference when computing confidence for small vs. large samples?
- As n increases, t-distribution approaches normal distribution

Testing for a Zero Mean

- Is a measured value significantly different from zero?
  - common use of confidence intervals
- When comparing random measurement with zero, must do so probabilistically
- If value different from zero with probability 100(1-α)%, then value is significantly different from zero



Example: Testing for a Zero Mean

- Difference in running time of two sorting algorithms A and B was measured on several different input sequences
- Differences are 1.5, 2.6, -1.8, 1.3, -.5, 1.7, 2.4
- Can we say with 99% confidence that 1 algorithm is superior?

Example: Testing for a Zero Mean

Example properties

- n $= 7$, $\bar{x} = 1.03$, STD $= 1.6$
- $\alpha = .01$, $\alpha/2 = .005$
- confidence interval

Example: Testing for a Zero Mean

Confidence interval includes 0; thus, cannot say with 99% confidence that the mean difference between A & B is significantly different from 0

$$(1.03 - t_{[1-.005;6]} * 1.60/\sqrt{7}, 1.03 + t_{[1-.005;6]} * 1.60/\sqrt{7})$$

$$(1.03 - (3.707) * 1.60/\sqrt{7}, 1.03 + (3.707) * 1.60/\sqrt{7})$$

$$= (-1.21, 3.27)$$

Paired Observations

- Conduct n experiments on each of 2 systems
  - system a: $\{a_1, a_2, ..., a_n\}$
  - system b: $\{b_1, b_2, ..., b_n\}$
- If one-one correspondence between tests on both systems – observations are said to be "paired"
- Treat the samples for 2 systems as one sample of n pairs
- For each pair, compute difference in performance
  - $a_1 - b_1, a_2 - b_2, ..., a_n - b_n$
- Construct a confidence interval for the mean difference
- Is the confidence interval includes 0, systems not significantly different

Unpaired Observations (t-test)

- Two samples, one size n_a, the other size n_b
- Compute mean of each sample: $\bar{x_a}$, $\bar{x_b}$
- Compute STD of each sample: s_a, s_b
- Compute mean difference $\bar{x_a} - \bar{x_b}$
- Compute std deviation of mean difference $\sqrt{\frac{s_a^2}{n_a} + \frac{s_b^2}{n_b}}$
- Effective number of degrees of freedom
  - $\mathbf{v} = \frac{(\mathbf{s_a^2/n_a + s_b^2/n_b})^2}{\frac{1}{\mathbf{n_a+1}}(\frac{\mathbf{s_a^2}}{\mathbf{n_a}})^2 + \frac{1}{\mathbf{n_b+1}}(\frac{\mathbf{s_b^2}}{\mathbf{n_b}})^2} - \mathbf{2}$

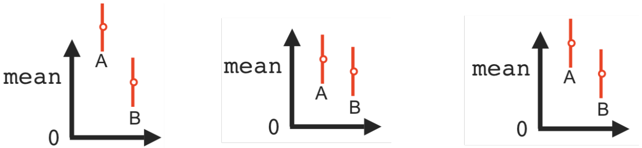
  Confidence interval for mean difference
  - $(\bar{x_a} - \bar{x_b}) \pm t_{[1-\alpha/2;v]}s$

Notes on Unpaired Observations

- Preceding slide made following assumptions
  - two samples of unequal size
  - standard deviations not assumed equal
  - small sample sizes
  - normal populations

Approximate Visual Test

- Simple visual test to compare unpaired samples
  - CI no overlap $A > B$
  - CI overlap; means in CI of other; alternatives not different
  - CI overlap; mean A not in CI B; need t-test



What Confidence Level to Use?

- Typically use confidence of 90% or 95%
- Need not always be that high
- Choice of confidence level is based on cost of loss if wrong!
- If loss is high compared to gain, use high confidence
- If loss is negligible compared to gain, low confidence OK

Consider, for example, a lottery in which a ticket costs one dollar but pays five million dollars to the winner. Suppose the probability of winning is 10-7 or one in ten million. To win the lottery with 90% confidence would require one to buy nine million tickets. It is clear that no one would be willing to spend that much for winning just five million.

One-sided Confidence Intervals

- Sometimes only a one-sided confidence interval is needed
- Example: want to test if mean $> \mu_0$
- In this case, one-sided lower confidence interval for μ needed
  - $(\bar{x} - t_{[1-\alpha;n-1]}s/\sqrt{n}, \bar{x})$
- For large samples, use z-values (unit normal distribution) rather than t-distribution

Determining Sample Size

- Confidence level from a sample depends on sample size
- The larger the sample, the higher the confidence
- Goal: determine smallest sample yielding desired accuracy

Sample Size for Determining Mean

- For a sample size n, the $100(1-\alpha)\%$ confidence interval of μ is $\bar{x} \pm z_{1-\alpha/2}\frac{s}{\sqrt{n}}$
- For a desired accuracy of r%, the confidence interval must be $\bar{x} \pm \bar{x}\frac{r}{100}$
- Thus, $z_{1-\alpha/2}\frac{s}{\sqrt{n}} = \bar{x}\frac{r}{100}$ and $n = \left\lceil \left(\frac{100 z_{1-\alpha/2} s}{r\bar{x}}\right)^2 \right\rceil$
- In a preliminary test, sample mean of response time is 20 seconds and std dev. = 5 seconds. How many repetitions are needed to estimate the mean response time within 2s at 95% confidence Required accuracy r = 2 in 20 = 10%
  - $n = \left\lceil \left(\frac{100 z_{1-\alpha/2} s}{r\bar{x}}\right)^2 \right\rceil = \left\lceil \left(\frac{100(1.96)(5)}{(10)(20)}\right)^2 \right\rceil = \lceil 24.01 \rceil = 25$