Dicii

Performance Modeling of Computer Systems and Networks

Prof. Vittoria de Nitto Personè

Interval Estimation

Università degli studi di Roma Tor Vergata
Department of Civil Engineering and Computer Science Engineering

1

model development

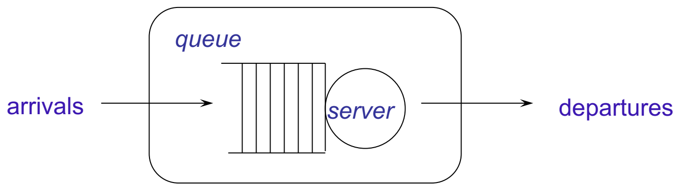
Algorithm 1.1: how to develop a model

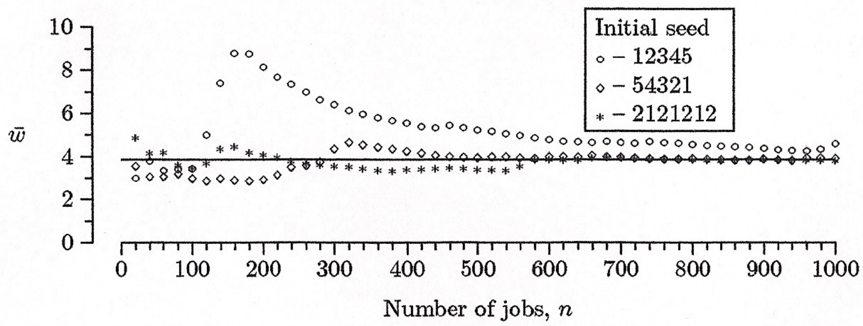
1. Goals and objectives
2. *Conceptual* model (cm)
3. Convert cm into a *specification* model (sm)
1. Convert sm into a *computational* model (cptm)
2. Verify
3. Validate

2

Algorithm 1.2: using the resulting model

7. Design simulations experiments
   - What parameters should be varied?
   - perhaps many combinatoric possibilities
8. Make production runs
   - Record initial conditions, input parameters
   - Record statistical output
9. Analyze the output
   - Random components → statistical analysis (means, standard deviations, percentiles, histograms etc.)
10. Make decisions
    - The step9 results drive the decisions → actions
    - Simulation should be able to correctly predict the outcome of these actions (→ further refinements)
11. Document the results
    - summarize the gained insights in specific observations and conjectures useful for subsequent similar system models





• The accumulated average wait was printed every 20 job

Consider a sample $x_1, x_2, \ldots, x_n$ (continuous or discrete) with

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Consider a piecewise constant sample path

$$x(t) = \begin{cases} x_1 & t_0 < t \le t_1 \\ x_2 & t_1 < t \le t_2 \\ . & . \\ . & . \\ x_n & t_{n-1} < t \le t_n \end{cases}$$

$$\bar{x} = \frac{1}{\tau}\int_0^{\tau} x(t)dt = \frac{1}{t_n}\sum_{i=1}^{n} x_i \delta_i \qquad s^2 = \frac{1}{\tau}\int_0^{\tau}(x(t) - \bar{x})^2 dt = \frac{1}{t_n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \delta_i$$

Central limit theorem

If $X_1, X_2, \ldots, X_n$ is an iid sequence of random variables (RVs) with

- common mean μ
- common standard deviation σ

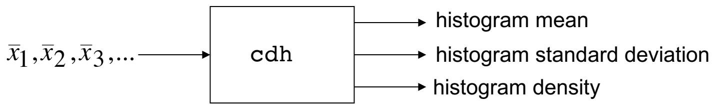
and if $\bar{X}$ is the (sample) mean of these RVs $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$
then $\bar{X}$ approaches a *Normal*(μ, $\sigma / \sqrt{n}$)
as $n \to \infty$

Sample Mean Distribution

- Choose one of the random variate generators in `rvgs` to generate a <u>sequence</u> of random variable <u>samples</u> with fixed sample size $n > 1$
- with the n-point samples indexed $j=1, 2, \ldots$, the corresponding sample mean $\bar{x}$ and sample standard deviation s can be calculated using Welford's algorithm

$$\underbrace{x_1, x_2, \ldots, x_n}_{\bar{x}_1, s_1}, \underbrace{x_{n+1}, x_{n+2}, \ldots, x_{2n}}_{\bar{x}_2, s_2}, \underbrace{x_{2n+1}, x_{2n+2}, \ldots, x_{3n}}_{\bar{x}_3, s_3}, x_{3n+1}$$

- A continuous-data histogram can be created using program `cdh`

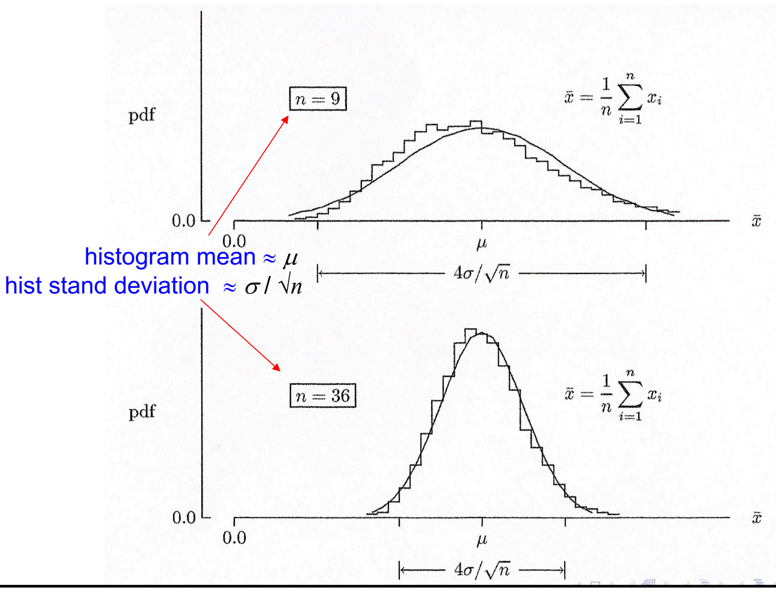


Properties of Sample Mean Histogram

If we denote with μ and σ the theoretical mean and standard deviation respectively of the random variates

- independent of n
  - the histogram mean is approximately μ
  - the histogram standard deviation is approximately $\sigma / \sqrt{n}$
- if n is sufficiently large,
  - the histogram density approximates the *Normal*$(\mu, \sigma / \sqrt{n})$ pdf

10000 n-point *Exponential*(μ) samples



Example

- The histogram density corresponding to the 36-point sample means is closely matched by the pdf of a *Normal*(μ, $\sigma / \sqrt{n}$) RV

  for *Exponential*(μ) samples, n=36 is large enough for the sample mean to be approximately *Normal*(μ, $\sigma / \sqrt{n}$)

- The histogram density corresponding to the 9-point sample means matches relatively well, but with a skew to the left

  - n=9 is not large enough

Example (cont.)

- Essentially all of the sample means are within an interval of width of $4\sigma / \sqrt{n}$ centered about μ
- because $n \to \infty$ as $\sigma / \sqrt{n} \to 0$, if n is large, all the sample means will be close to μ
- In general:
  - the accuracy of the *Normal*(μ, $\sigma / \sqrt{n}$) pdf approximation is dependent on the shape of a fixed population pdf
  - If the samples are drawn from a population with
    - a highly asymmetric pdf (like the *Exponential*(μ) pdf): n may need to be as large as 30 or more for good fit
    - a pdf symmetric about the mean (like the *Uniform*(a,b) pdf): n as small as 10 or less may produce a good fit

Examples of Linear Data Transformations

- suppose $x_1, x_2, \ldots, x_n$ measured in seconds
  - to convert to minutes, let $x'_i = x_i/60$

  (a=1/60, b=0)

$$\bar{x}' = \frac{45}{60} = 0.75 \qquad s' = \frac{15}{60} = 0.25$$ (minutes)

- *standardize* data (a=1/s, b= $-\bar{x}/s$)

$$x'_i = \frac{1}{s} x_i - \frac{\bar{x}}{s} \qquad x'_i = \frac{x_i - \bar{x}}{s}$$

Then

$$\bar{x}' = 0 \qquad s' = 1$$

Used to avoid problems with very large (or small) valued samples

Standardized Sample Mean Distribution

We can standardize the sample means $\bar{x}_1, \bar{x}_2, \bar{x}_3, \ldots$ by subtracting μ and dividing the result by $\sigma / \sqrt{n}$ to form the standardized sample means $z_1, z_2, z_3, \ldots$ defined by

$$z_j = \frac{\bar{x}_j - \mu}{\sigma / \sqrt{n}} \qquad j = 1, 2, 3, \ldots$$

- Generate a continuous-data histogram for the standardized sample means by program `cdh`

$\bar{z}_1, \bar{z}_2, \bar{z}_3, \ldots \longrightarrow$ `cdh` $\longrightarrow$ histogram mean; histogram standard deviation; histogram density

13

Properties of Standardized Sample Mean Histogram

- independent of n
  - the histogram mean is approximately 0
  - the histogram standard deviation is approximately 1
- if n is sufficiently large,
  - the histogram density approximates the *Normal*$(0,1)$ pdf

14

t-Statistic Distribution

Definition

- each sample mean $\bar{x}_j$ is a point estimate of μ
- each sample variance s_j^2 is a point estimate of σ^2
- each sample standard deviation s_j is a point estimate of σ

Want to replace *population* standard deviation σ with *sample* standard deviation s_j in standardization equation

$$z_j = \frac{\bar{x}_j - \mu}{\sigma / \sqrt{n}} \qquad j = 1,2,3,\ldots$$

$$\frac{s_j}{\sqrt{n-1}}$$

- Calculate the *t*-statistic

$$t_j = \frac{\bar{x}_j - \mu}{s_j / \sqrt{n-1}} \qquad j = 1,2,3,\ldots$$

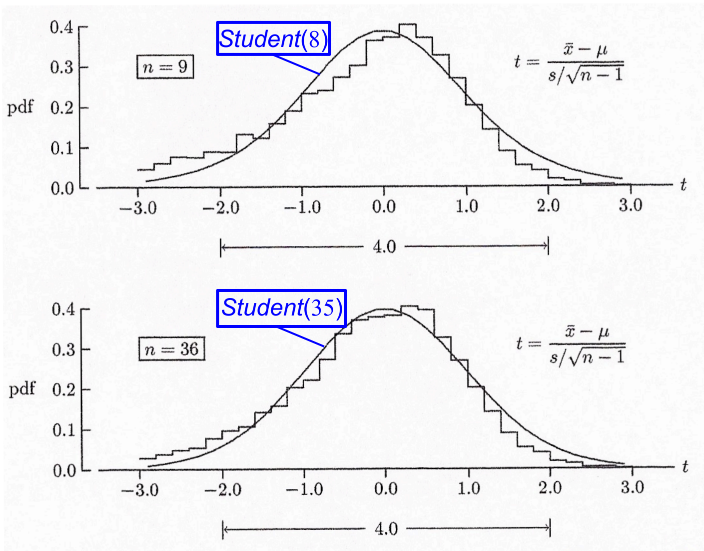
- Generate a continuous-data histogram using `cdh`

$$t_1, t_2, t_3, \ldots \longrightarrow \boxed{\texttt{cdh}} \longrightarrow \begin{cases} \text{histogram mean} \\ \text{histogram standard deviation} \\ \text{histogram density} \end{cases}$$

Properties of *t*-statistic histogram

- if $n > 2$, the histogram mean is approximately 0
- if $n > 3$, the histogram standard deviation is approximately $\sqrt{(n-1)/(n-3)}$
- if n is sufficiently large, the histogram density approximates the pdf of a *Student*$(n-1)$ random variable

Example (cont.)



Example (cont.)

- The histogram mean and standard deviation are approximately 0.0 and $\sqrt{(n-1)/(n-3)} \cong 1.0$ respectively
- The histogram density corresponding to the 36-point sample means matches the pdf of a *Student*(35) RV relatively well
- The histogram density corresponding to the 9-point sample means matches the pdf of a *Student*(8) RV, but not as well

Interval Estimation

Theorem 2

If $x_1, x_2, \ldots, x_n$ is an independent random sample from a "source" of data with unknown mean μ, if $\bar{x}$ and s are the mean and standard deviation of this sample, and if n is large, it is approximately true that

$$t = \frac{\bar{x} - \mu}{s / \sqrt{n-1}}$$

is a *Student*$(n-1)$ random variate

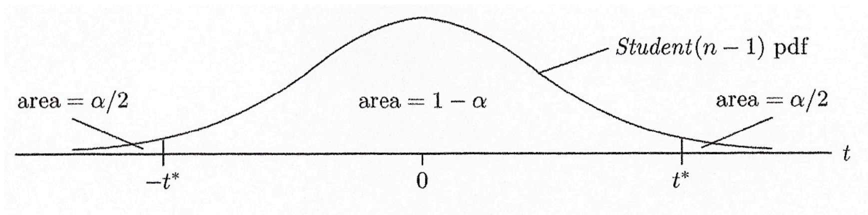
- provides the justification for estimating an interval that is likely to contain the mean μ
- as $n \to \infty$, the *Student*$(n-1)$ distribution becomes indistinguishable from *Normal*$(0,1)$

Suppose

- T is a *Student*(n-1) random variable
- α is a “confidence parameter” with $0.0 < \alpha < 1.0$

Then there exists a corresponding positive real number t^*

$$\Pr(-t^* \le T \le t^*) = 1 - \alpha$$



Interval Estimation

- suppose μ is unknown. Since $t \approx$ *Student*(n-1)

$$-t^* \le \frac{\bar{x}-\mu}{s/\sqrt{n-1}} \le t^*$$

will be approximately true with probability 1- α

- right inequality:

$$\frac{\bar{x}-\mu}{s/\sqrt{n-1}} \le t^* \Leftrightarrow \bar{x}-\mu \le \frac{t^* s}{\sqrt{n-1}} \Leftrightarrow \bar{x} - \frac{t^* s}{\sqrt{n-1}} \le \mu$$

- left inequality:

$$-t^* \le \frac{\bar{x}-\mu}{s/\sqrt{n-1}} \Leftrightarrow -\frac{t^* s}{\sqrt{n-1}} \le \bar{x}-\mu \Leftrightarrow \mu \le \bar{x} + \frac{t^* s}{\sqrt{n-1}}$$

So, with probability 1- α (approximately),

$$\bar{x} - \frac{t^* s}{\sqrt{n-1}} \le \mu \le \bar{x} + \frac{t^* s}{\sqrt{n-1}}$$

Theorem 3

If

$x_1, x_2, \ldots, x_n$ is an independent random sample from a "source" of data with unknown mean μ

- if $\bar{x}$ and s are the sample mean and sample standard deviation
- n is large

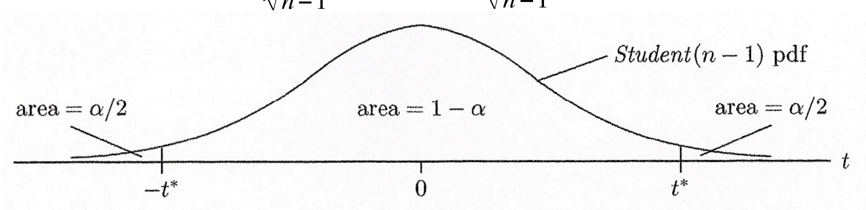
Then, given a confidence parameter α with $0.0 < \alpha < 1.0$, there exists an associated positive real number t^* such that

$$Pr\left(\bar{x} - \frac{t^* s}{\sqrt{n-1}} \leq \mu \leq \bar{x} + \frac{t^* s}{\sqrt{n-1}}\right) \cong 1 - \alpha$$

Example

- If α = 0.05, we are 95% confident that μ lies somewhere between

$$\bar{x} - \frac{t^* s}{\sqrt{n-1}} \quad \text{and} \quad \bar{x} + \frac{t^* s}{\sqrt{n-1}}$$



- for a fixed sample size n and level of confidence $1 - \alpha$, use `rvms` to determine t^* = *idfStudent*($n - 1$, $1 - \alpha/2$)
- ex. n = 30, α = 0.05 $\rightarrow t^*$ = *idfStudent*(29, 0.975) $\cong$ 2.045

Definition

- The interval defined by the two endpoints $\bar{x} \pm \frac{t^* s}{\sqrt{n-1}}$

  is a $(1-\alpha)$x100% confidence *interval estimate* for μ

- $(1-\alpha)$ is the *level of confidence* associated with this interval estimate and t^* is the *critical value* of t

Algorithm

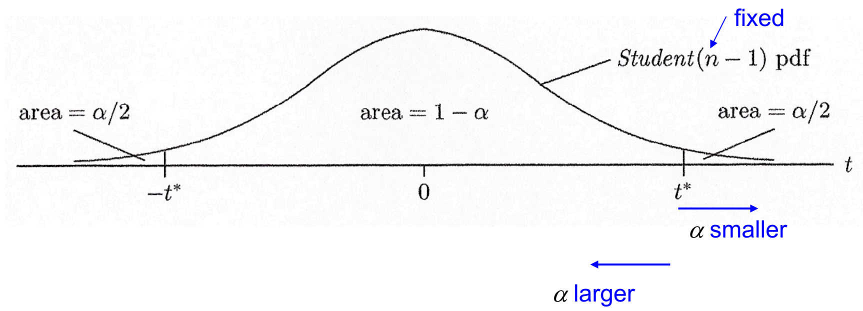
To calculate an interval estimate for the unknown mean μ of the population from which a random sample $x_1, x_2, \ldots, x_n$ was drawn:

- pick a level of confidence $1-\alpha$ (tipically α =0.05)
- calculate the sample mean $\bar{x}$ and standard deviation s (use Welford's algorithm)
- calculate the critical value t^* = *idfStudent*$(n-1, 1-\alpha/2)$
- calculate the interval endpoints $\bar{x} \pm \frac{t^* s}{\sqrt{n-1}}$

If n is sufficiently large, then you are $(1-\alpha)$x100% confident that the mean μ lies within the interval. The midpoint of the interval is $\bar{x}$

Tradeoff - Confidence Versus Sample Size

- For a fixed sample size
  - More confidence can be achieved only at the expense of a larger interval
  - A smaller interval can be achieved only at the expense of less confidence



Example

- The random sample of size $n = 10$:

1.051	6.438	2.646	0.805	1.505
0.546	2.281	2.822	0.414	1.307

is drawn from a population with unknown mean μ

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \qquad s = \sqrt{s^2}$$

$$\bar{x} = 1.982$$

Example

- The random sample of size $n = 10$:

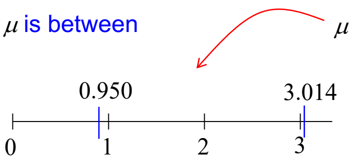
1.051	6.438	2.646	0.805	1.505
0.546	2.281	2.822	0.414	1.307

  is drawn from a population with unknown mean μ

- $\bar{x} = 1.982$ and $s = 1.690$
- to calculate a 90% confidence interval estimate:
  - determine $t^* = idfStudent(9, 0.95) \simeq 1.833$
  - interval: $1.982 \pm (1.833)(1.690/\sqrt{9}) = 1.982 \pm 1.032$

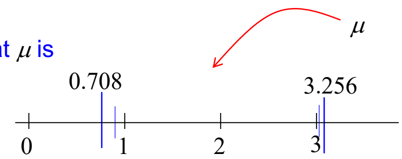
$$\bar{x} \pm \frac{t^* s}{\sqrt{n-1}}$$

- we are approximately 90% confident that μ is between 0.950 and 3.014

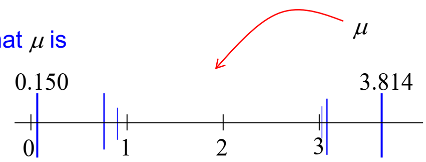


Example (cont.)

- To calculate a 95% confidence interval estimate:
  - determine: $t^* = idfStudent(9, 0.975) \simeq 2.262$
  - interval: $1.982 \pm (2.262)(1.690/\sqrt{9}) = 1.982 \pm 1.274$
- We are approximately 95% confident that μ is between 0.708 and 3.256



- To calculate a 99% confidence interval estimate:
  - determine: $t^* = idfStudent(9, 0.995) \simeq 3.250$
  - interval: $1.982 \pm (3.250)(1.690/\sqrt{9}) = 1.982 \pm 1.832$
- We are approximately 99% confident that μ is between 0.150 and 3.814
- Note: n=10 is <u>not</u> large



1. starting from a sample $x_1, x_2, \ldots, x_n$

- Program `estimate` automates the interval estimation process
- A typical application: estimate the value of an unknown population mean μ by using n replications to generate an independent random variate sample $x_1, x_2, \ldots, x_n$
- Function Generate() represents a discrete-event or Monte Carlo simulation program that returns a random variate output x

Using the Generate Method

```
for (i = 1; i <= n; i++)
        xi = Generate();
return x1, x2, . . . , xn;
```

- Given a level of confidence $1-\alpha$, program `estimate` can be used with $x_1, x_2, \ldots, x_n$ to compute an interval estimate for μ

estimate.c

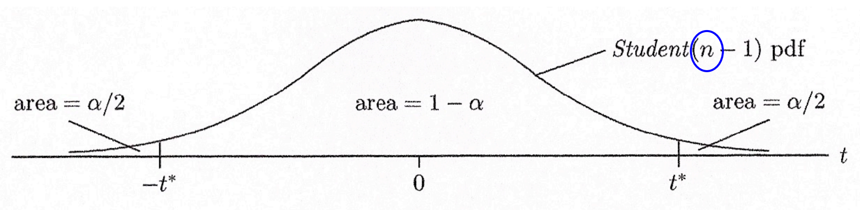
$$\bar{x}_i = \bar{x}_{i-1} + \frac{1}{i}(x_i - \bar{x}_{i-1}) \qquad v_i = v_{i-1} + \left(\frac{i-1}{i}\right)(x_i - \bar{x}_{i-1})^2$$

```
#include <math.h>
#include <stdio.h>
#include "rvms.h"
#define LOC 0.95              /* level of confidence, */
                                                 /* use 0.95 for
95% confidence */
  int main(void)
{ long   n    = 0;              /* counts data points */
  double sum  = 0.0;
  double mean = 0.0;
  double data;
  double stdev;
  double u, t, w;
  double diff;
  while (!feof(stdin)) {    /* use Welford's one-pass method */
    scanf("%lf\n", &data); /* to calculate the sample mean  */
    n++;                    /* and standard deviation        */
    diff  = data - mean;
    sum  += diff * diff * (n - 1.0) / n;
    mean += diff / n;
  }
  stdev  = sqrt(sum / n)
```

$t^* = idfStudent(n-1, 1-\alpha/2)$

```
if (n > 1) {
  u = 1.0 - 0.5 * (1.0 - LOC);      /* interval parameter  */
  t = idfStudent(n - 1, u);         /* critical value of t */
  w = t * stdev / sqrt(n - 1);      /* interval half width */
  printf("\nbased upon %ld data points", n);
  printf(" and with %d%% confidence\n", (int) (100.0 * LOC + 0.5));
  printf("the expected value is in the interval");
  printf("%10.2f +/- %6.2f\n", mean, w);  }
else
  printf("ERROR - insufficient data\n");
return (0);}
```


Tradeoff - Confidence Versus Sample Size



The only way to make the interval smaller without lessening the level of confidence is to increase the sample size

$$\bar{x} \pm \frac{t^* s}{\sqrt{n-1}}$$

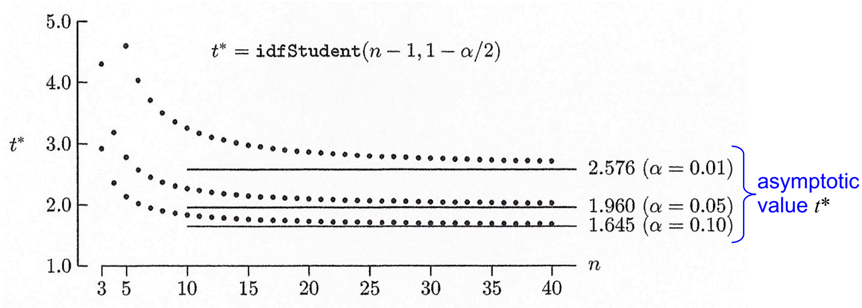
- Good news: with simulation, we can collect more data
- Bad news: interval size decreases with $\sqrt{n}$, not n

How Much More Data Is Enough?

- How large should n be to achieve an interval estimate $\bar{x} \pm w$ where w is user-specified?

- Answer: Use Welford's Algorithm with the algorithm p. 28 to iteratively collect data until a specified interval width is achieved

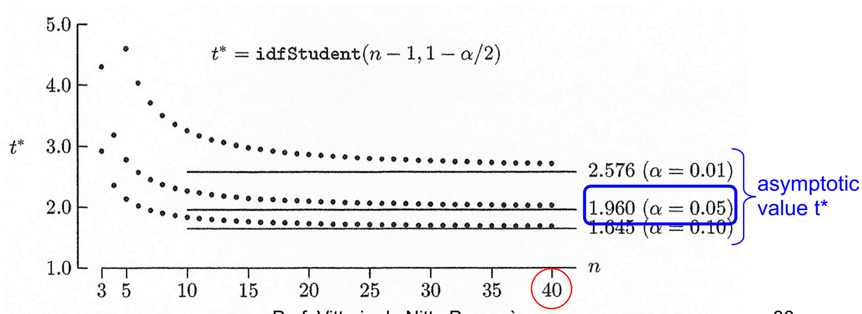
Note: if n is large then t^* is essentially independent of n



Asymptotic Value of t^*

- The asymptotic (large n) value of t^* is
$$t^*_\infty = \lim_{n\to\infty} idfStudent(n-1, 1-\alpha/2) = idfNormal(0.0, 1.0, 1-\alpha/2)$$
- Unless α is very close to 0.0, if $n > 40$, the asymptotic value t^*_∞ can be used

- If $n > 40$ and wish to construct a 95% confidence interval estimate, $t^*_\infty = 1.960$ can be used in the algorithm on p.28



Example

- Given a reasonable guess for s and a user-specified half-width parameter w, if t^*_∞, is used in place of t^*

  n can be determined by solving $w = \frac{t^* s}{\sqrt{n-1}}$ for n:

  $$n = \left\lfloor \left(\frac{t^*_\infty s}{w} \right)^2 \right\rfloor + 1$$

  provided $n > 40$

- For example, if s=3.0 and want to estimate μ with 95% confidence to within ± 0.5, a value of $n = 139$ should be used

Example

$$n = \left\lfloor \left(\frac{t^*_\infty s}{w} \right)^2 \right\rfloor + 1$$

- If a reasonable guess for s is not available, w can be specified as a proportion of s thereby eliminating s from the previous equation

- For example, if w is 10% of s and 95% confidence is desired, $n = 385$ should be used to estimate μ to within $\pm w$

  $(w/s = 0.1)$

See in the book algorithm 8.1.2 to obtain confidence interval starting from the sample $x_1, x_2, \ldots, x_n$ or from the half-width parameter w respectively

Discrete Simulation
Interval Estimation

The meaning of confidence

Incorrect:

"For this 95% confidence interval, the probability that μ is within this interval is 0.95"

- Why incorrect?
  - μ is not a random variable; it is constant (but unknown)
  - the interval endpoints are random

Correct:

"If I create many 95% confidence intervals, approximately 95% of them should contain μ"

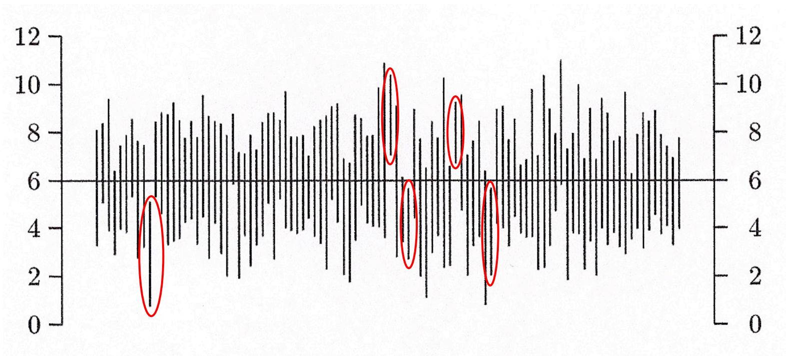
Prof. Vittoria de Nitto Personè 39

39

Discrete Simulation
Interval Estimation

Example

- 100 samples of size n=9 drawn from *Normal*(6,3) population
- For each sample, construct a 95% confidence interval
- 95 intervals contain μ=6
- Three intervals "too low", two intervals "too high"



Prof. Vittoria de Nitto Personè 40

40

Exercise

- Exercises 8.1.1, 8.1.5

- Consider case study 1 or case study 2, at your choice. Derive the sample mean histogram from one run (as in the picture in slide 8) and for two different sizes for the samples. Compare the obtained results with reference to the Exponential sample mean histograms seen in this lecture (slide p.10).