

Performance Evaluation Techniques Summer 2005

Statistics Refresher

Dr.-Ing. Andreas Willig

Telecommunication Networks Group (TKN)
Technical University Berlin

awillig@ieee.org

June 30, 2005

Introduction

- Assume there is an iid sequence X_1, X_2, X_3, \dots of random variables having the common distribution $F(\cdot)$
- $F(\cdot)$ is not known to us, but we have a number of numerical *realizations*, *samples* or *observations* $x_1, x_2, x_3, \dots, x_n$ of these random variables and want to determine certain characteristics of $F(\cdot)$ from these observations
- Interesting characteristics can be:
 - $E[X_1]$ and $\text{Var}[X_1]$
 - quantiles

Here we focus on estimating $E[X_1]$ and on determining *confidence intervals* for this

Estimation – General Notion

- We want to know the “true” expectation $\mu = E[X_1]$ of the distribution $F(\cdot)$
- But: we only have observations x_1, x_2, \dots, x_n at hand
- Therefore: we must *estimate* μ from the observations, giving an estimated value $\hat{\mu}$
- Any function $g(\cdot)$ such that

$$\hat{\mu} = g(x_1, x_2, \dots, x_n)$$

is called an *estimator*

Unbiased Estimators

Definition 1 (unbiased estimator). *An estimator is called unbiased, if*

$$\lim_{n \rightarrow \infty} E[\hat{\mu}] = E[\mu]$$

holds, i.e. when adding more observations makes the estimator more precise.

One wants to have unbiased estimators or only those estimators where the bias ($\lim_{n \rightarrow \infty} E[\hat{\mu}] - E[\mu]$) is small compared to $|\mu|$

Estimating the Expected Value

The standard estimator for $E[X_1]$ is given by:

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i \quad (1)$$

It is called *sample mean*. It is unbiased:

$$E[\hat{\mu}] = E\left[\frac{1}{n} \sum_{i=1}^n X_i\right] = \frac{1}{n} \sum_{i=1}^n E[X_i] = \frac{1}{n} \sum_{i=1}^n E[X_1] = E[X_1] = \mu$$

but we will almost surely have:

$$\mu \neq \hat{\mu}$$

Estimating the Variance

Definition 2 (Sample Variance). *The sample variance $S^2(n)$ of the observations x_1, \dots, x_n is given by:*

$$S^2(n) = \frac{1}{n-1} \sum_{i=1}^n (x_i - \hat{\mu})^2 \quad (2)$$

where $\hat{\mu}$ is just the estimation given in Equation 1.

One can show that $S^2(n)$ is an unbiased estimator of $\sigma^2 = \text{Var}[Y]$ being the “true” variance.

The question why $n - 1$ is used instead of n in the denominator is beyond the scope of this lecture ...

Estimating the Variance

One does not know how close the sample mean $\hat{\mu}$ is to the true mean μ . But, the larger n is, the smaller is the variance of the sample mean $\hat{\mu}$ (which is a random variable!):

$$\text{Var} [\hat{\mu}] = \text{Var} \left[\frac{1}{n} \sum_{i=1}^n X_i \right] = \frac{1}{n^2} \sum_{i=1}^n \text{Var} [X_i] = \frac{1}{n^2} n \sigma^2 = \frac{\sigma^2}{n}$$

where we have used the independence assumption for the X_i . For large n the variance gets close to zero and large differences between μ and $\hat{\mu}$ become more and more unlikely.

This does not hold true when the X_i are correlated!!!

Estimating the Mean for Grouped Observations

Sometimes our observations x_1, \dots, x_n have only a finite range \mathcal{R} , i.e. assume only a finite set of values, like for example a subset of the natural numbers $\mathcal{R} = \{i, i + 1, i + 2, \dots, j - 1, j\}$.

Suppose we have not collected the observations, but instead we have counted how often each value $k \in \mathcal{R}$ actually occurred in the set of observations. Let f_i, f_{i+1}, \dots, f_j denote these frequencies. Then $n = f_i + f_{i+1} + \dots + f_j$, and the estimator for $\mu = E[Y]$ is given by:

$$\hat{\mu} = \frac{1}{n} \sum_{k=i}^j k \cdot f_k$$

Estimating the Variance for Grouped Observations

We can derive the estimator $S^2(n) = \text{Var}[X_1]$ as follows:

$$\begin{aligned} S^2(n) &= \frac{1}{n-1} \sum_{l=1}^n (x_l - \hat{\mu})^2 = \frac{1}{n-1} \sum_{l=1}^n (x_l^2 - 2x_l\hat{\mu} + \hat{\mu}^2) \\ &= \frac{1}{n-1} \left(\sum_{l=1}^n x_l^2 + n\hat{\mu}^2 - 2\hat{\mu} \sum_{l=1}^n x_l \right) \\ &= \frac{1}{n-1} \left(\sum_{l=1}^n x_l^2 + n\hat{\mu}^2 - 2n\hat{\mu}^2 \right) \\ &= \frac{1}{n-1} \left(\sum_{l=1}^n x_l^2 - n\hat{\mu}^2 \right) = \frac{1}{n-1} \left(\sum_{k=i}^j k^2 \cdot f_k - n\hat{\mu}^2 \right) \end{aligned}$$

Estimating Probabilities

The estimation of probabilities can be expressed as estimation of a mean value of a Bernoulli random variable

Example: you want to estimate the probability that a packet traveling between two fixed Internet hosts has a delay of more than 200 msec. If $d_1, d_2, d_3, \dots, d_m$ are the observed independent end-to-end delays, then we could form the observations:

$$x_i = \begin{cases} 1 & : d_i > 200 \text{ msec} \\ 0 & : d_i \leq 200 \text{ msec} \end{cases}$$

corresponding to the “true” random variables:

$$X_i = \begin{cases} 1 & : D_i > 200 \text{ msec} \\ 0 & : D_i \leq 200 \text{ msec} \end{cases}$$

(X_i is called an “indicator variable”, it is a Bernoulli random variable). A small calculation shows that $E[X_i] = \Pr[D_i > 200 \text{ msec}]$.

Confidence Intervals for the Sample Mean

- Goal: characterize the estimation error $\mu - \hat{\mu}$ for a set x_1, x_2, \dots, x_n of iid observations
- Precisely: we look for an interval $I_\alpha = [\hat{\mu} - s, \hat{\mu} + s]$ which contains the true mean μ with at least $(1 - \alpha) \cdot 100\%$ probability
 - Such an interval is called a *confidence interval* for the *confidence level* α – typical values for α are $\alpha \in \{0.1, 0.05, 0.01\}$
 - Of course, a small interval (small s) is desirable, since this indicates a “likely high accuracy” of the estimation

Confidence Intervals for the Sample Mean II

- Interpretation:
 - For single experiment (set of observations) μ may / may not be in I_α
 - If you repeat the experiment hundreds of times (each experiment giving distinct observations) and calculate for each experiment i a separate confidence interval $I_\alpha^{(i)}$ then then $(1 - \alpha) \cdot 100\%$ of the confidence intervals $I_\alpha^{(i)}$ contain μ
- The development of confidence intervals rests on the assumption that x_1, x_2, \dots, x_n are realizations of iid random variables; thus, for

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

the *central limit theorem* can be invoked, if n is “large enough” ($n \geq 50$)

Confidence Intervals for the Sample Mean III

- In the following it is explicitly assumed that for n large enough the random variable

$$\frac{\hat{\mu} - \mu}{\sqrt{\frac{\sigma^2}{n}}}$$

has a standard normal distribution

- For given α we can find a value $z_{1-\frac{\alpha}{2}}$ from a table of the standard normal distribution for which:

$$1 - \alpha = \Pr \left[-z_{1-\frac{\alpha}{2}} \leq X \leq z_{1-\frac{\alpha}{2}} \right] = \Phi(z_{1-\frac{\alpha}{2}}) - \Phi(-z_{1-\frac{\alpha}{2}})$$

where $X \sim N(0,1)$ is a random variable with standard normal distribution

Confidence Intervals for the Sample Mean IV

- Here,

$$\Phi(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^t e^{-\frac{y^2}{2}} dy$$

is the distribution function of X

Confidence Intervals for the Sample Mean V

- Now we use the approximation given by the central limit theorem:

$$\begin{aligned} 1 - \alpha &\approx \Pr \left[-z_{1-\frac{\alpha}{2}} \leq \frac{\hat{\mu} - \mu}{\sqrt{\frac{\sigma^2}{n}}} \leq z_{1-\frac{\alpha}{2}} \right] \\ &= \Pr \left[-z_{1-\frac{\alpha}{2}} \sqrt{\frac{\sigma^2}{n}} \leq \hat{\mu} - \mu \leq z_{1-\frac{\alpha}{2}} \sqrt{\frac{\sigma^2}{n}} \right] \\ &= \Pr \left[\hat{\mu} - z_{1-\frac{\alpha}{2}} \sqrt{\frac{\sigma^2}{n}} \leq \mu \leq \hat{\mu} + z_{1-\frac{\alpha}{2}} \sqrt{\frac{\sigma^2}{n}} \right] \end{aligned}$$

Confidence Intervals for the Sample Mean VI

- The unknown variance σ^2 has to be estimated by $S^2(n)$, which makes the approximation weaker.
- The interval in which we can find μ with probability $(1 - \alpha) \cdot 100\%$ is just given by:

$$\left[\hat{\mu} - z_{1-\frac{\alpha}{2}} \sqrt{\frac{\sigma^2}{n}}, \hat{\mu} + z_{1-\frac{\alpha}{2}} \sqrt{\frac{\sigma^2}{n}} \right]$$

The interval is symmetric around $\hat{\mu}$.

- For $n \rightarrow \infty$ the interval gets smaller and vanishes. However, to halve the interval's size, we have to increase n by four times.
- For σ^2 getting larger, the interval gets larger, too

Confidence Intervals for the Sample Mean VII

- To summarize our results so far: after having obtained n samples, we can estimate the confidence interval $I = [l(n, \alpha), u(n, \alpha)]$ which contains the true mean μ with $(1 - \alpha) \cdot 100\%$ probability as follows:

$$l(n, \alpha) = \hat{\mu} - z_{1-\frac{\alpha}{2}} \sqrt{\frac{S^2(n)}{n}} \quad , \quad u(n, \alpha) = \hat{\mu} + z_{1-\frac{\alpha}{2}} \sqrt{\frac{S^2(n)}{n}}$$

- Sources of approximation errors:
 - What does “large enough” mean for n ?
 - The distribution of $\hat{\mu}$ *converges* to a normal distribution, but the convergence speed depends on $F(\cdot)$
 - The true variance σ^2 is approximated by $S^2(n)$, which might for “real”, positively correlated data be significantly smaller than σ^2

Confidence Intervals for the Sample Mean VII

The approximation error might manifest itself in two ways:

- The interval $I = [l(n, \alpha), u(n, \alpha)]$ might be too large, but contains the value μ with the desired confidence level $(1 - \alpha) \cdot 100\%$. This just tends to lengthen simulation times.
- The interval $I = [l(n, \alpha), u(n, \alpha)]$ contains μ with lower probability than $(1 - \alpha) \cdot 100\%$. This might become a serious problem for the credibility of simulation results.

This is just what happens when our observations are positively correlated, because the confidence interval is smaller than it should be.

Confidence Intervals for the Sample Mean VIII

Instead of taking $z_{1-\frac{\alpha}{2}}$ from the standard normal distribution, a more appropriate choice (especially for smaller values of n) is to take the $1 - \frac{\alpha}{2}$ quantiles of the Student-t-distribution with parameter $a = n - 1$ degrees of freedom (can be found in tables, e.g. in [1]), i.e. we replace $z_{1-\frac{\alpha}{2}}$ by $t_{n-1,1-\frac{\alpha}{2}}$ and the confidence interval becomes

$$I_\alpha = [l(n, \alpha), u(n, \alpha)] \quad (3)$$

with:

$$l(n, \alpha) = \hat{\mu} - t_{n-1,1-\frac{\alpha}{2}} \sqrt{\frac{S^2(n)}{n}} \quad (4)$$

$$u(n, \alpha) = \hat{\mu} + t_{n-1,1-\frac{\alpha}{2}} \sqrt{\frac{S^2(n)}{n}} \quad (5)$$

Confidence Intervals for the Sample Mean IX

When the distribution function $F(\cdot)$ belongs to a normal distribution $X_1 \sim N(\mu, \sigma^2)$ then the confidence interval using $t_{n-1, 1-\frac{\alpha}{2}}$ is *exact*.

Using the Student-t distribution we get:

- a larger (and thus more conservative) confidence interval for small n , since $t_{n-1, 1-\frac{\alpha}{2}} \geq z_{1-\frac{\alpha}{2}}$, especially for small n
 - We thus reduce the probability of type-I errors
- For large n more or less the same confidence interval as if we have used $z_{1-\frac{\alpha}{2}}$.

References

- [1] Averill M. Law and W. David Kelton. *Simulation Modeling and Analysis*. McGraw-Hill, third edition, 2000.