LOCAL SEARCH METHODS
APPLICATIONS AND ENGINEERING

Lecture 12

Empirical Methods
Statistical Inference

Marco Chiarandini

Outline

1. Statistical Inference
2. Two Sample Tests
3. Design of Experiments for Algorithms
4. Analysis of Variance

Outline

1. Statistical Inference
2. Two Sample Tests
3. Design of Experiments for Algorithms
4. Analysis of Variance

Where do we need inference?

- in the prediction of algorithm results
- in the comparison between algorithms
- in the analysis of the impact of algorithmic factors
- in the goodness of fit of distributions

Statistical Inference

Random Sample	Inference	Population
X^n	$\longrightarrow$	$P(x, \theta)$
Statistical Estimator $\widehat{\theta}$		Parameter θ

Deductive approach: test hypotheses on the parameters of the population

Inductive approach: estimate unknown parameters of the population (through confidence intervals)

Parameter Estimation

Estimator $\hat{\theta}(X_1, \ldots, X_n)$ makes a guess on the parameter (Es. $\bar{X}$)

Estimate is the actual value $\hat{\theta}(x_1, \ldots, x_n)$

Properties of an estimator:

- unbiased: $E[\hat{\theta}] = \theta$ (*e.g.*, $E[\bar{X}] = \mu$)
- consistent
- efficient (uncertainty must decrease with size, *e.g.*, Var$[\bar{X}] = \sigma^2/n$)
- sufficient

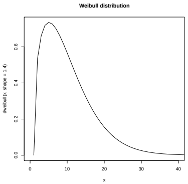
Note: The *best* result $b_N = \min_i c_i$ is not a good estimator. It is biased and not efficient.

Theorem: Central Limit Theorem

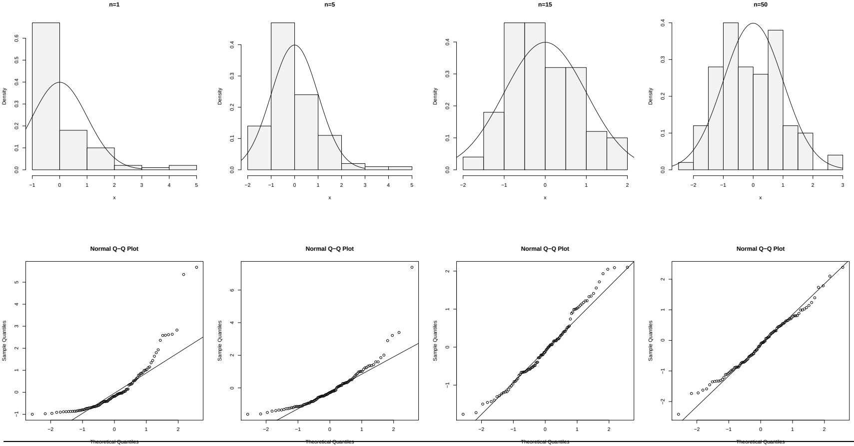
If X^n is a random sample from an **arbitrary** distribution with mean μ and variance σ then the average $\bar{X}^n$ is asymptotically normally distributed, *i.e.*,

$$\bar{X}^n \approx N(\mu, \frac{\sigma^2}{n}) \qquad \text{or} \qquad z = \frac{\bar{X}^n - \mu}{\sigma/\sqrt{n}} \approx N(0,1)$$

- Consequences:
  - allows inference from a sample
  - allows to model errors in measurements: $X = \mu + \epsilon$
- Issues:
  - n should be *enough* large
  - μ and σ must be known



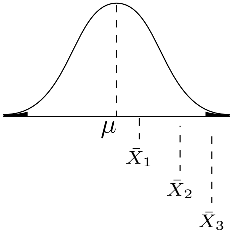
$$z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$



Inference: Hypothesis Testing and Confidence Intervals

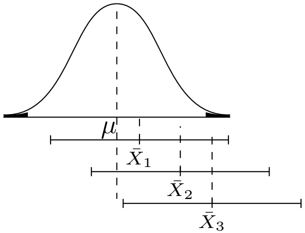
A test of hypothesis determines how likely an sampled estimate $\hat{\theta}$ is to occur under some assumptions on the parameter θ of the population.

$$Pr\left\{\mu - z_1 \frac{\delta}{\sqrt{n}} \leq \bar{X} \leq \mu + z_2 \frac{\delta}{\sqrt{n}}\right\} = 1 - \alpha$$



A confidence interval contains all those values that a parameter θ is likely to assume with probability $1 - \alpha$: $Pr(\hat{\theta}_1 < \theta < \hat{\theta}_2) = 1 - \alpha$

$$Pr\left\{\bar{X} - z_1 \frac{\delta}{\sqrt{n}} \leq \mu \leq \bar{X} + z_2 \frac{\delta}{\sqrt{n}}\right\} = 1 - \alpha$$



And if the variance is unknown...

then we substitute σ with its estimator $\hat{\sigma} = S$

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$$

but then

$$z = \frac{X - \mu}{S\sqrt{n}} \approx t_{n-1}$$

i.e., z approximates a t-student distribution.

Terminology

Statistical Hypotheses:

- H_0: null hypothesis
- H_1 alternative hypothesis (one-sided, two-sided)

Within the testing procedure two types of errors are possible:

- error of Type I when the null hypothesis is rejected although it is true
  the level α specifies a priori the assumed likelihood of this error
  α is called level of significance
- error of Type II occurs when a false null hypothesis is not rejected
  it is denoted by β

The power of the test is the likelihood of rejecting a false null hypothesis, $1 - \beta$.

The power of a test depends on the test statistics, the sample size, the alternative hypothesis.

Outline

Two Sample Matched Pairs Case

It corresponds to one run on various instances design

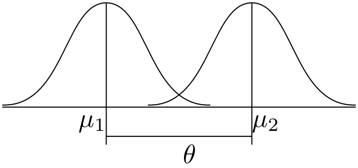
- Student t test `t.test()`
- Wilcoxon test `wilcox.test()`
- Binomial test `binom.test()`
- Permutation solution

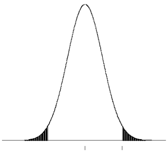
Two Sample Case

It corresponds to a several runs on one single instance design

- Student t test `t.test()`
- Kruskall Wallis `kruskal.test()`
- Permutation solution

Hypothesis Testing Procedure





1. Specify the parameter θ and the test hypothesis, *e.g.*,

$$\theta = \mu_1 - \mu_2 \qquad \begin{cases} H_0 : \theta = 0 \\ H_1 : \theta \neq 0 \end{cases}$$

2. Obtain $P(\theta | \theta = 0)$, the null distribution of θ
3. Compare $\hat{\theta}$ with the upper (in case of one-sided tests) α-quantile of $P(\theta|\theta = 0)$ and accept or reject H_0 according to whether $\hat{\theta}$ is smaller or larger than this value.

Outline

Steps in the Design of Experiments for Algorithms

- Statement of the objectives of the experiment
  - Comparison of different algorithms
  - Impact of algorithm components
  - How instance affect the algorithms
- Identification of the sources of variance
  - Treatment factors, Nuisance factors
  - Factorial Experiment, Experimental unit
  - Complete block design (within-subject matched pairs)
- Definition of test instances. How many instances how many runs?
- Selection of the factor combinations to test
- Running a pilot experiment and refine the design
  - Bugs and no external biases
  - Ceiling or floor effects
  - Rescaling levels of quantitative factors
  - Detect the number of experiments needed to obtained the desired power.

Relevant Experimental Designs

Algorithms $\Rightarrow$ Treatment Factor; Instances $\Rightarrow$ Blocking Factor

Design A: One run on various instances (Unreplicated Factorial)

	Algorithm 1	**Algorithm 2**	...	**Algorithm k**
Instance 1	X_{11}	X_{12}		X_{1k}
$\vdots$	$\vdots$	$\vdots$		$\vdots$
Instance b	X_{b1}	X_{b2}		X_{bk}

Design B: Several runs on various instances (Replicated Factorial)

	Algorithm 1	**Algorithm 2**	...	**Algorithm k**
Instance 1	$X_{111}, \ldots, X_{11r}$	$X_{121}, \ldots, X_{12r}$		$X_{1k1}, \ldots, X_{1kr}$
Instance 2	$X_{211}, \ldots, X_{21r}$	$X_{221}, \ldots, X_{22r}$		$X_{2k1}, \ldots, X_{2kr}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$
Instance b	$X_{b11}, \ldots, X_{b1r}$	$X_{b21}, \ldots, X_{b2r}$		$X_{bk1}, \ldots, X_{bkr}$

Two issues:

- Need to test more than two algorithms
- Need to test more than one factor

Outline

1. Statistical Inference
2. Two Sample Tests
3. Design of Experiments for Algorithms
4. Analysis of Variance

Multisamples – Analysis of Variance

$H_0: \ \mu_1 = \mu_2 = \mu_3 = \dots \qquad H_1: \{\text{at least one differ}\}$

Applying t-test to all pairs the error of Type I is not α but higher:

$$\alpha_{EX} = 1 - \alpha^c$$

Eg, for $\alpha = 0.05$ and $c = 3$ $\alpha_{EX} = 0.14$!

Single Factor Analysis of Variance

$$X_{it} = \mu_i + \epsilon_{it}$$

$$X_{it} = \mu + \alpha_i + \epsilon_{it}$$

$$\sum_{i=1}^{k}\sum_{t=1}^{r}(X_{it} - \bar{X}_{..})^2 = \sum_{i=1}^{k}\sum_{t=1}^{r}(X_{it} - \bar{X}_{i.})^2 + \sum_{i=1}^{k} r(\bar{X}_{i.} - \bar{X}_{..})^2$$

(decomposition in *within-group* and *between-group* sum of squares)

$$MST = \frac{\sum_{i=1}^{k} r\left(\bar{X}_{i.} - \bar{X}_{..}\right)^2}{k-1} \qquad MSE = \frac{\sum_{i=1}^{k}\sum_{t=1}^{r}\left(X_{it} - \bar{X}_{i.}\right)^2}{N-k}$$

(MST *mean square per treatment* and MSE *mean square per error*)

$$F = \frac{MST}{MSE}$$

$F \sim F_{k-1,r-k}$, *i.e.*, the F-ratio approximates a the Fisher distribution F_{df_1,df_2} with df_1 and df_2 degrees of freedom.

Single Factor analysis of Variance (contd.)

- Parametric analysis: ANOVA through F-ratio and Fisher test
  - independent
  - normally distributed
  - homoschedastic
- Nonparametric analysis: Kruskall Wallis
  Assumptions
  - independent
  - homoschedastic

Two Factors Analysis of Variance

Model:

$$X_{hi} = \mu + \alpha_i + \theta_h + \epsilon_{hi}$$

Mean square:

$$MST = \frac{b \sum_{i=1}^{k} (\bar{X}_{.i} - \bar{X}_{..})^2}{k-1}; \quad MSE = \frac{\sum_{h=1}^{b} \sum_{i=1}^{k} \left(X_{hi} - \bar{X}_{h.} - \bar{X}_{.i} + \bar{X}_{..} \right)^2}{bk - b - k + 1}$$

Statistical Tests

- ANOVA through F-ratio and Fisher test
- Friedman test

Two Factors Repeated Measures Analysis of Variance

Model:

$$X_{hit} = \mu + \alpha_i + \theta_h + \epsilon_{hit}$$

or alternatively:

$$X_{hit} = \mu + \alpha_i + \theta_h + \alpha\theta_{hi} + \epsilon_{hit}$$

Statistical Tests

- ANOVA through F-ratio and Fisher test
- Friedman test

If the interactions between factors are of interest interaction plots are useful to visualize them