DM812
METAHEURISTICS

Lecture 3

Empirical Methods for Configuring and Tuning

Marco Chiarandini

Department of Mathematics and Computer Science
University of Southern Denmark, Odense, Denmark

Outline

Outline

Statistics

Field of mathematics that studies the probability of events on the basis of inference from empirical data.

Descriptive statistics resumes and visualizes data (Exploratory data analysis)

Inferential statistics makes inference or prediction about the populations from which samples are drawn.
  Population: total of subjects that share something in common
  Sample: set of subjects drawn from populations

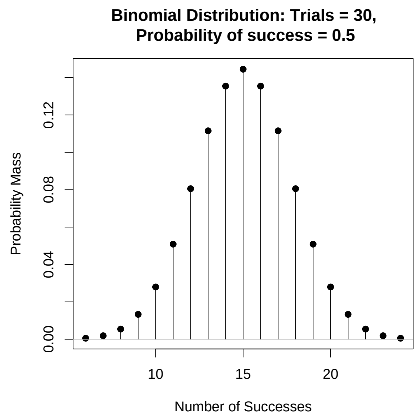
Data:

- quantitative (numerical) discrete or continuous (presence of an order)
- qualitative or categorical

Probability Distributions

Binomial distribution

$$P[x = v] = \binom{n}{v} p^v (1-p)^{n-v}$$



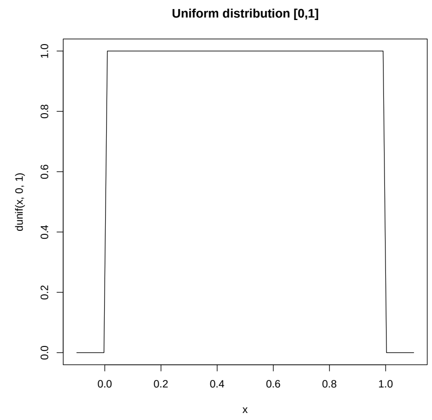
p probability of successes
x number of successes
The binomial distribution indicates the probability for each set of outcomes, *i.e.*, $v = \{1, \ldots, n\}$ successes.

One parameter: p

Introduction
Inferential Statistics

Uniform distribution (continuous)

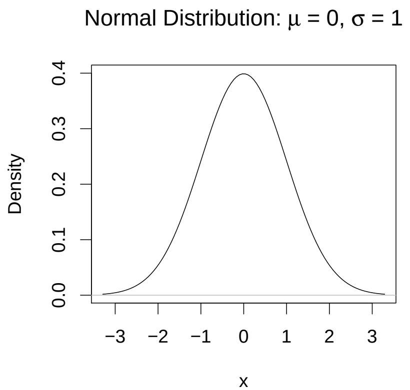
$$f(x) = \frac{1}{b-a}$$



Introduction
Inferential Statistics

Normal distribution (continuous)

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



Theoretical importance

Defined by two parameters: $N(\mu, \sigma)$.

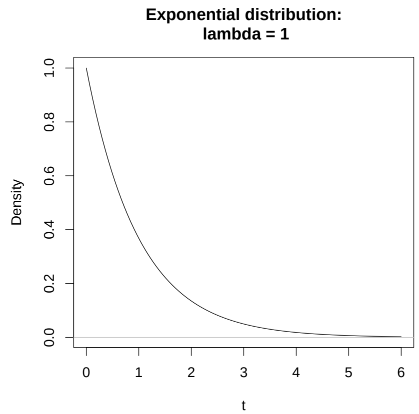
$N(0, 1)$ is the standardized version.

In $N(0, 1)$ 68.27% of data fall within $\mu \pm \sigma$

Introduction
Inferential Statistics

Exponential distribution (continuous)

$$f(t) = \lambda e^{-\lambda t}$$

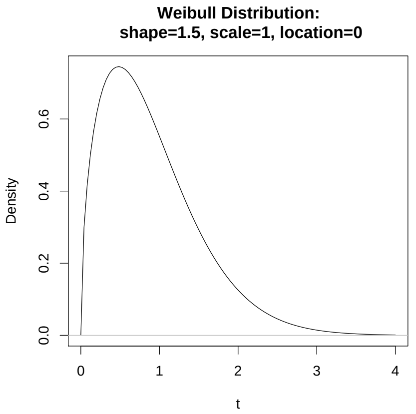


It has the memory-less property, *i.e.*, the probability of a new event to happen within a fixed time does not depend on the time passed so far.

Defined by one parameter: $E[X] = \frac{1}{\lambda}$.

Weibull distribution (continuous)

$$f(x) = \frac{\beta}{\eta}\left(\frac{t-\gamma}{\eta}\right)^{\beta-1} e^{-\left(\frac{t-\gamma}{\eta}\right)^{\beta}}$$



Used in life data and reliability analysis

Defined by three parameters:
β (shape), η (scale), γ (location)

Others (theoretically relevant)

- $\chi^2(n)$: chi-squared distribution with n degrees of freedom:
  distribution of $\sum_i X_n^2$ where $X_1, \ldots, X_n$ are independently, standard normally distributed variables
- $t(r)$: Student t-distribution with r degrees of freedom:
  distribution of $X_1/\sqrt{X_2/r}$ with $X_1 \sim N(0,1)$ and $X_2 \sim \chi^2(r)$ independently distributed variables
- $F(r_1, r_2)$: Fisher distribution with r_1 and r_2 degrees of freedom:
  distribution of $(X_1/r_1)/(X_2/r_2)$ with $X_1 \sim \chi^2$ and $X_2 \sim \chi^2$ independently distributed variables

Outline

Inferential Statistics

- We work with samples (instances, solution quality)
- But we want sound conclusions: generalization over a given population (all possible instances)
- Thus we need statistical inference

$$\begin{array}{ccc} \text{Random Sample} & & \text{Population} \\ X^n & \xrightarrow{\text{Inference}} & P(x,\theta) \\ \text{Statistical Estimator } \widehat{\theta} & & \text{Parameter } \theta \end{array}$$

Since the analysis is based on finite-sized sampled data, statements like

> *"the cost of solutions returned by algorithm $\mathcal{A}$ is smaller than that of algorithm $\mathcal{B}$"*

must be completed by

> *"at a level of significance of 5%".*

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.

A Motivating Example

- There is a competition and two stochastic algorithms $\mathcal{A}_1$ and $\mathcal{A}_2$ are submitted.
- We run both algorithms once on n instances.
  On each instance either $\mathcal{A}_1$ wins $(+)$ or $\mathcal{A}_2$ wins (-) or they make a tie $(=)$.

Questions:

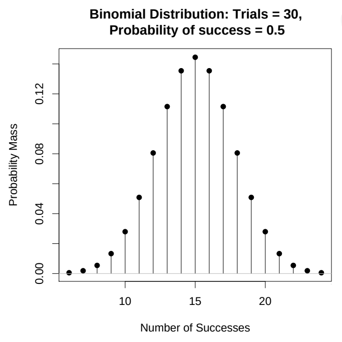
1. If we have only 10 instances and algorithm $\mathcal{A}_1$ wins 7 times how confident are we in claiming that algorithm $\mathcal{A}_1$ is the best?
2. How many instances and how many wins should we observe to gain a confidence of 95% that the algorithm $\mathcal{A}_1$ is the best?

A Motivating Example

- p: probability that $\mathcal{A}_1$ wins on each instance $(+)$
- n: number of runs without ties
- Y: number of wins of algorithm $\mathcal{A}_1$

If each run is indepenedent and consitent:

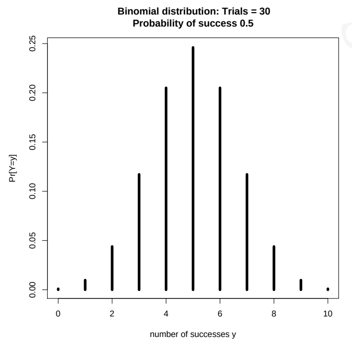
$$Y \sim B(n,p): \qquad \Pr[Y = y] = \binom{n}{y} p^y (1-p)^{n-y}$$



1. If we have only 10 instances and algorithm $\mathcal{A}_1$ wins 7 times how confident are we in claiming that algorithm $\mathcal{A}_1$ is the best?

Under these conditions, we can check how unlikely the situation is if it were $p(+) \leq p(-)$.

If $p = 0.5$ then the chance that algorithm $\mathcal{A}_1$ wins 7 or more times out of 10 is 17.2%: quite high!



2 How many instances and how many wins should we observe to gain a confidence of 95% that the algorithm $\mathcal{A}_1$ is the best?

To answer this question, we compute the 95% quantile, *i.e.*, $y : \Pr[Y \geq y] < 0.05$ with $p = 0.5$ at different values of n:

n	10	11	12	13	14	15	16	17	18	19	20
y	9	9	10	10	11	12	12	13	13	14	15

This is an application example of sign test, a special case of binomial test in which $p = 0.5$

Outline

Inferential Statistics

General procedure:

- Assume that data are consistent with a null hypothesis H_0 (e.g., sample data are drawn from distributions with the same mean value).
- Use a statistical test to compute how likely this is to be true, given the data collected. This "likely" is quantified as the p-value.
- Accept H_0 as true if the p-value is larger than an user defined threshold called level of significance α.
- Alternatively (p-value $< \alpha$), H_0 is rejected in favor of an alternative hypothesis, H_1, at a level of significance of α.

Inferential Statistics

Two kinds of errors may be committed when testing hypothesis:

$$\alpha = P(\text{type I error}) = P(\text{reject } H_0 \mid H_0 \text{ is true})$$
$$\beta = P(\text{type II error}) = P(\text{fail to reject } H_0 \mid H_0 \text{ is false})$$

General rule:

1. specify the type I error or level of significance α
2. seek the test with a suitable large statistical power, i.e., $1 - \beta = P(\text{reject } H_0 \mid H_0 \text{ is false})$

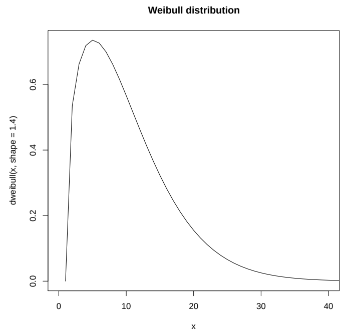
Theorem: Central Limit Theorem

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.*,

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

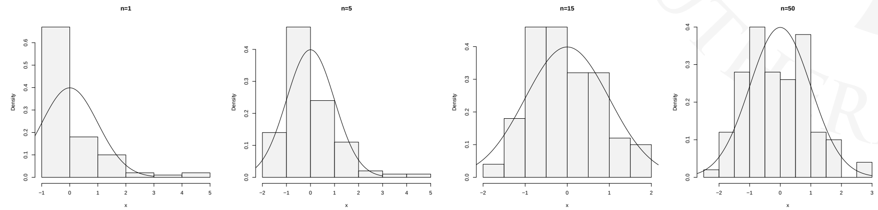
- Consequences:
  - allows inference from a sample
  - allows to model errors in measurements: $X = \mu + \epsilon$
- Issues:
  - n should be *enough* large
  - μ and σ must be known

Introduction
Inferential Statistics
Basics of Inferential Statistics
Experimental Designs



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

Samples of size 1, 5, 15, 50 repeated 100 times



Introduction
Inferential Statistics
Basics of Inferential Statistics
Experimental Designs

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

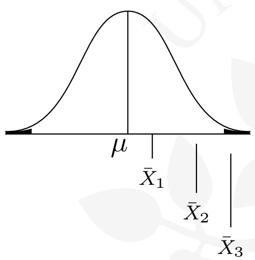
$$z = \frac{X - \mu}{S\sqrt{n}} \approx t_{n-1}$$

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

Inference Procedures

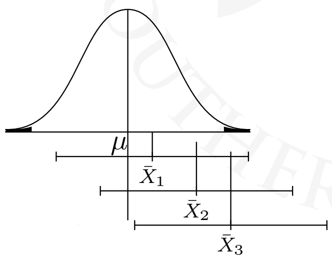
Introduction
Inferential Statistics
Basics of Inferential Statistics
Experimental Designs

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.



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

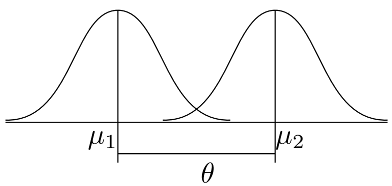
A confidence interval contains all those values that a parameter θ is likely to assume with probability $1-\alpha$: $Pr(\hat{\theta}_1 < \theta < \hat{\theta}_2) = 1 - \alpha$

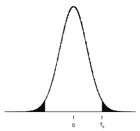


$$Pr\left\{\bar{X} - z_1\frac{\theta}{\sqrt{n}} \leq \mu \leq \bar{X} + z_2\frac{\theta}{\sqrt{n}}\right\} = 1 - \alpha$$

Statistical Tests

The Procedure of Test of Hypothesis

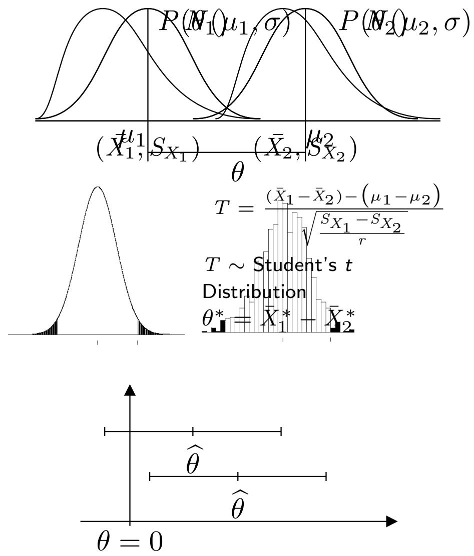




1. Specify the parameter θ and the test hypothesis,
$$\theta = \mu_1 - \mu_2 \qquad \begin{cases} H_0 : \theta = 0 \\ H_1 : \theta \neq 0 \end{cases}$$
2. Obtain $P(\theta|\theta = 0)$, the null distribution of θ
3. Compare $\hat{\theta}$ with the $\alpha/2$-quantiles (for two-sided tests) of $P(\theta|\theta = 0)$ and reject or not H_0 according to whether $\hat{\theta}$ is larger or smaller than this value.

Statistical Tests

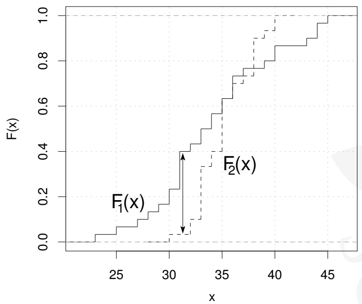
The Confidence Intervals Procedure



1. Specify the parameter θ and the test hypothesis,
$$\theta = \mu_1 - \mu_2 \qquad \begin{cases} H_0 : \theta = 0 \\ H_1 : \theta \neq 0 \end{cases}$$
2. Obtain $P(\theta, \theta = 0)$, the null distribution of θ in correspondence of the observed estimate $\hat{\theta}$ of the sample X
3. Determine $(\hat{\theta}^-, \hat{\theta}^+)$ such that $Pr\{\hat{\theta}^- \leq \theta \leq \hat{\theta}^+\} = 1 - \alpha$.
4. Do not reject H_0 if $\theta = 0$ falls inside the interval $(\hat{\theta}^-, \hat{\theta}^+)$. Otherwise reject H_0.

Kolmogorov-Smirnov Tests

The test compares empirical cumulative distribution functions.



It uses maximal difference between the two curves, $sup_x|F_1(x) - F_2(x)|$, and assesses how likely this value is under the null hypothesis that the two curves come from the same data

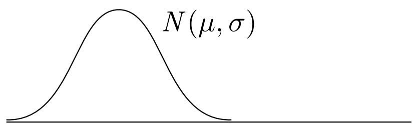
The test can be used as a two-samples or single-sample test (in this case to test against theoretical distributions: goodness of fit)

The test can be done in R with `ks.test`.

Parametric *vs* Nonparametric

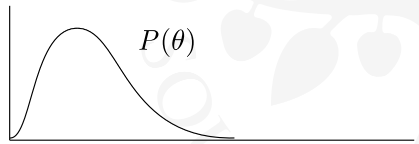
Parametric assumptions:

- independence
- homoschedasticity
- normality



Nonparametric assumptions:

- independence
- homoschedasticity



- Rank based tests
- Permutation tests
  - Exact
  - Conditional Monte Carlo

Parametric *vs* Nonparametric

- Parametric assumptions seem to be saliently violated when dealing with optimization algorithms
- Nonparametric rank based tests are based on asymptotic (large sample) theory
- Parametric tests are typically more powerful than nonparametric
- With few data permutations tests are an alternative but less powerful than parametric.

Hence:

- When from diagnostic investigation, assumptions seem satisfied (e.g., with large samples), parametric methods are more powerful and should be preferred.
  Otherwise, consider data transformations ($\log x, x^2, \sqrt{x}$)
- Alternatively, nonparametric methods based on ranks are helpful and also remove scale and location problems due to the instances.

Outline

Preparation of the Experiments

Variance reduction techniques

- Same pseudo random seed

Sample Sizes

- If the sample size is large enough (infinity) any difference in the means of the factors, no matter how small, will be significant
- Real *vs* Statistical significance
  Study factors until the improvement in the response variable is deemed small
- Desired statistical power + practical precision ⇒ sample size

Note: If resources available for N runs then the optimal design is one run on N instances [Birattari, 2004]

The Design of Experiments

- Statement of the objectives of the experiment
  - Comparison of different algorithms
  - Impact of algorithm components
  - How instance features affect the algorithms
- Identification of the sources of variance
  - Treatment factors (qualitative and quantitative)
  - Controllable nuisance factors ⇐ blocking
  - Uncontrollable nuisance factors ⇐ measuring
- Definition of factor combinations to test
  Easiest design: Unreplicated or Replicated Full Factorial Design
- 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.

Experimental Design

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	$X_{b11}, \ldots, X_{b1r}$	$X_{b21}, \ldots, X_{b2r}$		$X_{bk1}, \ldots, X_{bkr}$

Multiple Comparisons

$H_0: \mu_1 = \mu_2 = \mu_3 = \ldots$ $\quad H_1: \{$at least one differs$\}$

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

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

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

Adjustment methods

- Protected versions: global test + no adjustments
- Bonferroni $\alpha = \alpha_{EX}/c$ (conservative)
- Tukey Honest Significance Method (for parametric analysis)
- Holm (step-wise)
- Other step procedures

Post-hoc analysis: Once the effect of factors has been recognized a finer grained analysis is performed to distinguish where important differences are.