'Statistical Inference'
Prof Gino Verleye, Ph.D.
University Gent.

- Introduction: what is Statistical inference?
- Ideal world versus reality: data quality issues and solutions
- The nature of sample parameters
- Statistical inference implies making errors
- Decision tree for classic analysis procedures
- Software : innovations that make your life easier

What does Statistical Inference mean?

- Statistical practice happens on three levels:
  - 1° Descriptive Statistics: compute parameters (means, fractions, crosstables, correlations,…) on the data at hand
  - 2° Inferential Statistics: bring in the notion of probability samples: learn from the sample data and project towards the population
  - 3° Statistical Modeling: compute complex relationships between manifest and even latent variables

Example from Human Factors & Medicine program in NATO

Measuring and modelling Psychosocial Resilience in civil populations

Descriptive level

- Do men and woman have the same levels of fear in public places?
  - Fear: metric variable
  - Gender: Nominal 2category variable

Breakdown Table of Descriptive Statistics (nato1.STA)
Smallest N for any variable: 1021

GENDER	FEARPUB Means	FEARPUB N	FEARPUB Std.Dev.
male	2,793307	508	0,814740
female	2,975146	513	0,893494
All Grps	2,884672	1021	0,859624

Inferential level

- Which error do I make when saying that the gender difference is for real in the population?

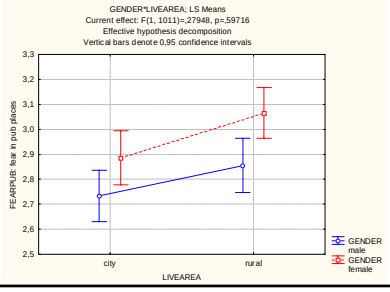
Analysis of Variance (nato1.STA)
Marked effects are significant at p < ,05000

Variable	SS Effect	df Effect	MS Effect	SS Error	df Error
FEARPUB	8,439757	1	8,439757	745,2929	1019

- Implies statistical testing with the appropriate method: ANOVA or t-test in this case

Model level

Joint modeling of gender and area



Ideal world versus reality: data quality issues and solutions

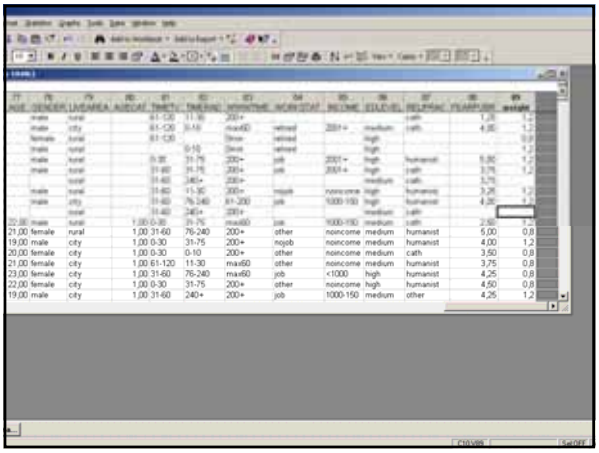
- Ideal world: no missing data and normal distributions (required by many inferential methods)
- Real world:
  - Missing data: item & unit level
  - Non normality is a fact

Missing data

- Unit missing:
  - example: male and females 50%-50%
  - Sample data : 40%-60%

  Issue : we can not be representative :
  too much female information

  Weighting : boost males, downsize females

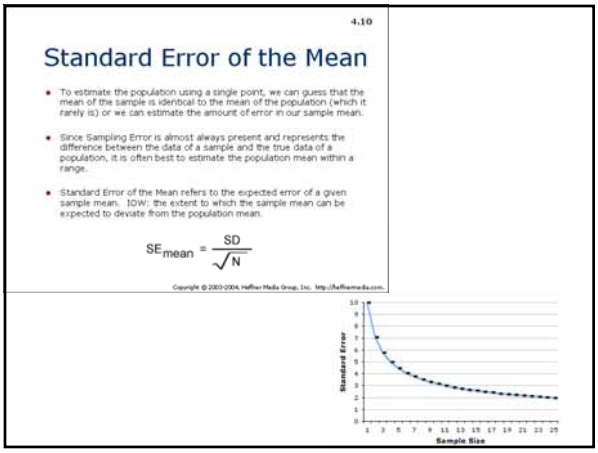


Missing data

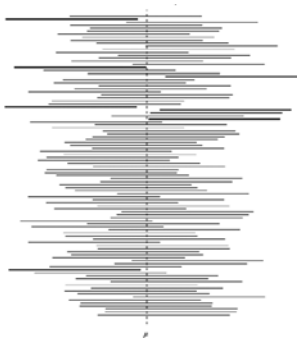
- Item missing:
  - It is not the fraction of non observed data that counts
  - It is the NATURE of the missing data process:
    - Completely at random: OK but realistic?
    - At random: very often the case, OK for stats
    - Non ignorable: very often : troublesome, dramatic
  - Solution: smart imputation, special estimation methods, look out: usualy listwise deletion: a case is entirely dropped if one variable is missing

The nature of sample parameters

- They vary from sample to sample
- But : we only have one sample in practice
- Mathematical statistics learn:
  - If N>30, sample parameters are drown from a normal distribution, with known error margin: the standard error: standard deviation of the parameter estimate
  - Error decreases with N: intuitive logical
  - Requires random sampling!!!

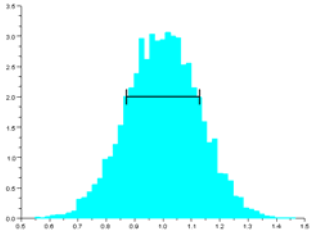


For infinite number of samples:

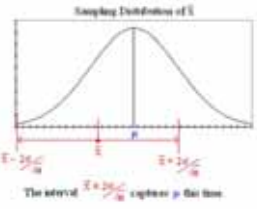


Confidence Interval

Where is the area in which we will find the true mean with 95% probability?



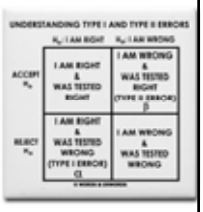
In practice:



Hypothesis testing…implies making errors

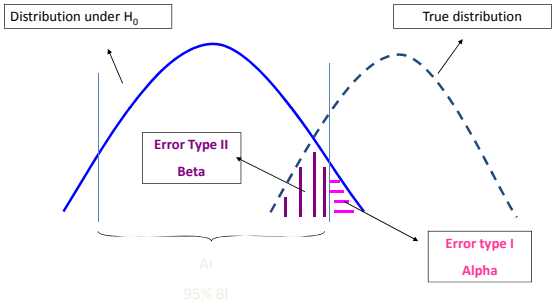
- Ho: true mean= x
- H1: true mean <> x

- Alpha: we control it
- Beta: unknown: only simulate



Alpha & Beta are the prices you pay to infer from sample to population

Graphically



Which methods to apply?

- Relation between 2 categorical variables: cross table with chi2 test
- Compare 2 means: t-test
- Compare >2 means : one way anova
- Compute linear relationship: correlation

- Relation between 2 categorical variables: cross table with chi2 test

Summary Frequency Table (nato1.STA)
Marked cells have counts > 1000
(Marginal summaries are not marked)

	GENDER	TIMETV 0-30	TIMETV 31-60	TIMETV 61-120	TIMETV 120+
Count	male	107	135	165	104
Row Percent		20,94%	26,42%	32,29%	20,35%
Total Percent		10,40%	13,12%	16,03%	10,11%
Count	female	83	105	170	160
Row Percent		16,02%	20,27%	32,82%	30,89%
Total Percent		8,07%	10,20%	16,52%	15,55%
Count	All Grps	190	240	335	264
Total Percent		18,46%	23,32%	32,56%	25,66%

Summary Table: Expected Frequencies (nato1.STA)
Marked cells have counts > 1000
Pearson Chi-square: 18,6882, df=3, p=,000317

GENDER	TIMETV 0-30	TIMETV 31-60	TIMETV 61-120	TIMETV 120+	Row Totals
male	94,3537	119,1837	166,3605	131,1020	511,000
female	95,6463	120,8163	168,6395	132,8980	518,000
All Grps	190,0000	240,0000	335,0000	264,0000	1029,000

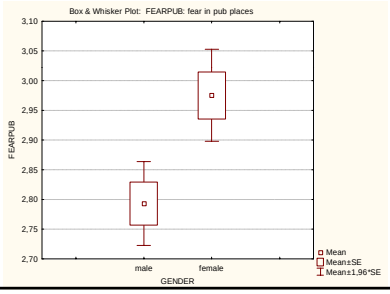
Summary Table: Observed minus Expected Frequencies (nato1.S
Marked cells have counts > 1000
Pearson Chi-square: 18,6882, df=3, p=,000317

GENDER	TIMETV 0-30	TIMETV 31-60	TIMETV 61-120	TIMETV 120+	Row Totals
male	12,6463	15,8163	1,36054	-27,1020	-0,000000
female	-12,6463	-15,8163	1,36054	27,1020	0,000000
All Grps	0,0000	0,0000	0,00000	0,0000	0,000000

- Compare 2 means: t-test

T-tests; Grouping: GENDER: gender (nato1.STA)
Group 1: male
Group 2: female

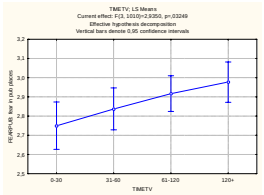
Variable	Mean male	Mean female	t-value	df	p	Valid N male	Valid N female	Std.Dev. male	Std.Dev. female	F-ratio Variances	p Variances
FEARPUB	2,793307	2,975146	-3,39695	1019	0,000708	508	513	0,814740	0,893494	1,202666	0,037548



- Compare >2 means : one way anova

Univariate Tests of Significance for FEARPUB (nato1.
Sigma-restricted parameterization
Effective hypothesis decomposition

Effect	SS	Degr. of Freedom	MS	F	p
Intercept	8023,575	1	8023,575	10926,04	0,000000
TIMETV	6,466	3	2,155	2,94	0,032492
Error	741,697	1010	0,734		



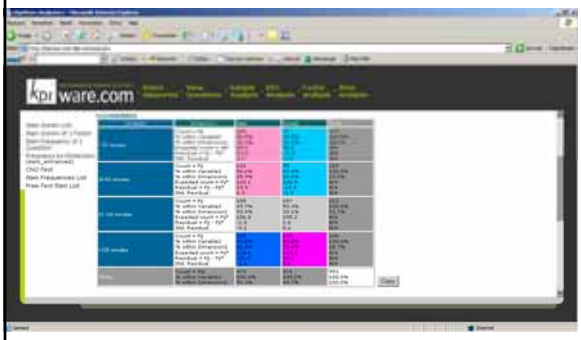
- Compute linear relationship: correlation

Correlations (nato1.STA)
Marked correlations are significant at p < .05000
N=972 (Casewise deletion of missing data)

Variable	MENTDIST	FEARPUB	FEARLAT	TIMETV	TIMER
MENTDIST	1,00	-0,12	-0,21	0,08	
FEARPUB	-0,12	1,00	0,49	0,10	
FEARLAT	-0,21	0,49	1,00	0,08	
TIMETV	0,08	0,10	0,08	1,00	
TIMERAD	0,04	0,03	0,07	0,13	
WWWTIME	-0,13	-0,15	-0,11	-0,15	

The other way: drobots.com

- In order to apply statistics correctly, one needs a lot of knowledge and some experience
- Statistical anxiety exists
- Fast analysis & reporting
- Manifest & latent levels
- Internet solution



Group comparison

Impact modeling