

'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	FEARPUB	FEARPUB
	Means	N	Std. Dev.
male	2.793307	508	0.814740
female	2.971546	513	0.893494
All Grps	2.884672	1021	0.859624

Inferential level

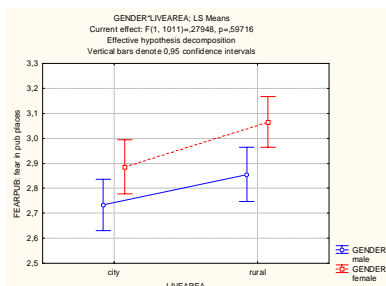
- Which error do I make when saying that the gender difference is for real in the population?

Analysis of Variance (nato1_STA)					
Marked effects are significant at p < .05000					
Variable	SS	df	MS	SS	df
	Effect	Effect	Effect	Error	Error
FEARPUB	8.439737	1	8.439737	7.452928	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

AGE	GENDER	EDUCATION	INCOME	RELIGION	PLACEMENT	WEIGHT
11-15	male	200+	noninc	other	medium	1.0
16-17	male	0-10	noninc	other	medium	1.0
18-19	male	0-10	noninc	other	medium	1.0
20-24	male	0-10	noninc	other	medium	1.0
25-29	male	0-10	noninc	other	medium	1.0
30-34	male	0-10	noninc	other	medium	1.0
35-39	male	0-10	noninc	other	medium	1.0
40-44	male	0-10	noninc	other	medium	1.0
45-49	male	0-10	noninc	other	medium	1.0
50-54	male	0-10	noninc	other	medium	1.0
55-59	male	0-10	noninc	other	medium	1.0
60-64	male	0-10	noninc	other	medium	1.0
65-69	male	0-10	noninc	other	medium	1.0
70-74	male	0-10	noninc	other	medium	1.0
75-79	male	0-10	noninc	other	medium	1.0
80-84	male	0-10	noninc	other	medium	1.0
85-89	male	0-10	noninc	other	medium	1.0
90-94	male	0-10	noninc	other	medium	1.0
95-99	male	0-10	noninc	other	medium	1.0
100+	male	0-10	noninc	other	medium	1.0
11-15	female	200+	noninc	other	medium	1.0
16-17	female	0-10	noninc	other	medium	1.0
18-19	female	0-10	noninc	other	medium	1.0
20-24	female	0-10	noninc	other	medium	1.0
25-29	female	0-10	noninc	other	medium	1.0
30-34	female	0-10	noninc	other	medium	1.0
35-39	female	0-10	noninc	other	medium	1.0
40-44	female	0-10	noninc	other	medium	1.0
45-49	female	0-10	noninc	other	medium	1.0
50-54	female	0-10	noninc	other	medium	1.0
55-59	female	0-10	noninc	other	medium	1.0
60-64	female	0-10	noninc	other	medium	1.0
65-69	female	0-10	noninc	other	medium	1.0
70-74	female	0-10	noninc	other	medium	1.0
75-79	female	0-10	noninc	other	medium	1.0
80-84	female	0-10	noninc	other	medium	1.0
85-89	female	0-10	noninc	other	medium	1.0
90-94	female	0-10	noninc	other	medium	1.0
95-99	female	0-10	noninc	other	medium	1.0
100+	female	0-10	noninc	other	medium	1.0

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: usually listwise deletion: a case is entirely dropped if one variable is missing

The nature of sample parameters

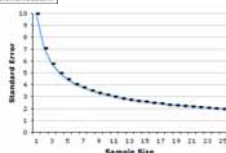
- They vary from sample to sample
- But : we only have one sample in practice
- Mathematical statistics learn:
 - If $N > 30$, sample parameters are drawn from a normal distribution, with known error margin: the standard error: standard deviation of the parameter estimate
 - Error decreases with N :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

Standard Error of the Mean

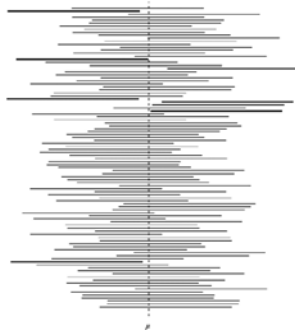
- To estimate the population using a single point, we can guess that the mean of the sample is identical to the mean of the population (which it rarely is) or we can estimate the amount of error in our sample mean.
- Since Sampling Error is almost always present and represents the difference between the data of a sample and the true data of a population, it is often best to estimate the population mean within a range.
- Standard Error of the Mean refers to the expected error of a given sample mean. SDW: the extent to which the sample mean can be expected to deviate from the population mean.

$$SE_{\text{mean}} = \frac{SD}{\sqrt{N}}$$

Copyright © 2003-2004, Haller Media Group, Inc. <http://hallermedia.com>

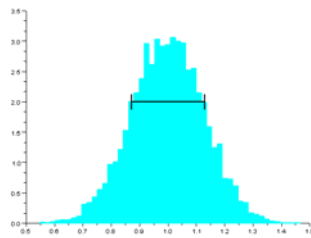


For infinite number of samples:

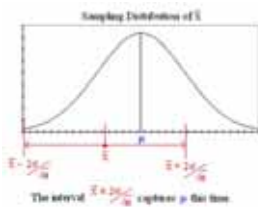


Confidence Interval

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



In practice:



Levels of Confidence	
99%	$z = 2.576$
95%	$z = 1.96$
90%	$z = 1.645$

Hypothesis testing...implies making errors

- H_0 : true mean = x
- H_1 : true mean $\neq x$

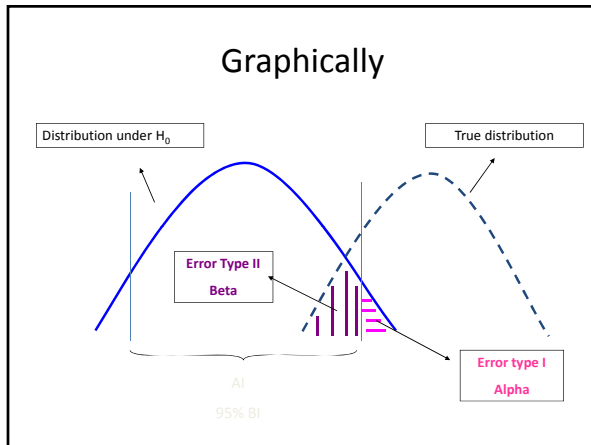


- Alpha: we control it
- Beta: unknown: only simulate

		H_0 : I AM RIGHT	H_0 : I AM WRONG
ACCEPT H_0	I AM RIGHT & WAS TESTED RIGHT	I AM WRONG & WAS TESTED RIGHT (TYPE II ERROR) β	
REJECT H_0	I AM RIGHT & WAS TESTED WRONG (TYPE I ERROR) α	I AM WRONG & WAS TESTED WRONG	

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 χ^2 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				Row Totals
		0-30	31-60	61-120	120+	
Count	male	107	135	165	104	
Row Percent		23.84%	26.42%	32.29%	30.39%	
Total Percent		10.40%	13.12%	16.07%	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 Gps	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				Row Totals
	0-30	31-60	61-120	120+	
male	84.3537	119.1837	166.3605	131.1020	511.0000
female	55.6463	120.8163	168.6395	132.8980	518.0000
All Gps	150.0000	240.0000	335.0000	264.0000	1029.0000

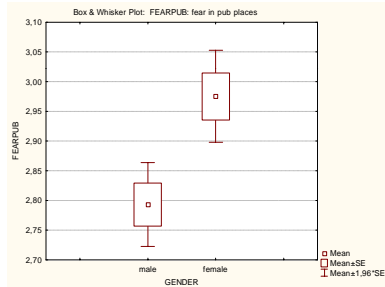
Summary Table: Observed minus Expected Frequencies (nato1.S)
Marked cells have counts > 1000
Pearson Chi-square: 18.6882, df=3, p=.000317

GENDER	TIMETV				Row Totals
	0-30	31-60	61-120	120+	
male	-12.6463	15.8163	-1.36054	-27.1020	-0.000000
female	12.6463	-15.8163	1.36054	27.1020	0.000000
All Gps	0.00000	0.00000	0.00000	0.00000	0.000000

- Compare 2 means: t-test

T-tests, Grouping: GENDER: gender (nato1.STA)

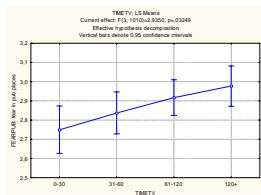
Variable	Group 1: male		Group 2: female		t-value	df	p	Valid N	Valid N	Std. Dev.	Std. Dev.	F-ratio	p
	Mean	Mean	Mean	Mean									
FEARPUB	2.750007	2.975146	-1.39506	1019	0.000708	208	0.13	0.814740	0.893494	1.202666	0.007148		



- Compare >2 means : one way anova

Univariate Tests of Significance for FEARPUB (nato1.S)
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 (naot_STA)					
Marked correlations are significant at p < .05000					
N=972 (Casewise deletion of missing data)					
Variable	MENTDIS1	FEARPUB	FEARLAT	TIMETV	TIMEA
MENTDIS1	1.00				
FEARPUB		1.00			
FEARLAT			1.00		
TIMETV				1.00	
TIMEA					1.00

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



NATO Intelligence against terrorism



The screenshot shows a software interface with a sidebar on the left containing menu items like 'Home', 'Reports', 'Tools', and 'Help'. The main area displays a table with columns for 'Item', 'Status', 'Priority', and 'Date'. The table contains several rows of data, with some cells highlighted in blue and others in pink.

