ADVANCED QUANTITATIVE METHODS CLINIC

Master's in Sustainability Leadership, Cambridge Institute for Sustainability Leadership

Sreekumar Thaithara Balan

Monday 3rd August, 2015

Department of Physics and Astronomy,
University College London
sbalan@star.ucl.ac.uk

OUTLINE

OUTLINE

- Software for data analysis ($\sim$15mins)
- Data visualisation ($\sim$20mins)
- Descriptive statistics ($\sim$20mins)
- Inferential statistics ($\sim$20mins)
- Regression ($\sim$20mins)
- Discussion ($\sim$10mins)

SOFTWARE FOR DATA ANALYSIS

REPOSITORY

Supplementary materials can found at
`https://github.com/tbs1980/CISLQuantWorkshop`

SOFTWARE FOR DATA ANALYSIS

A large list can be found in `Wikipedia`. Some widely used ones are below.

- `Python`, `https://www.python.org/`
- R, `https://cran.r-project.org/`
- `Excel`, `https://products.office.com/en-us/excel`
- SPSS, `http://www-01.ibm.com/software/analytics/spss/`

I will demonstrate the examples using `Python`. If you have no prior experience, no problem, there will be plenty of help.

SOFTWARE SET-UP

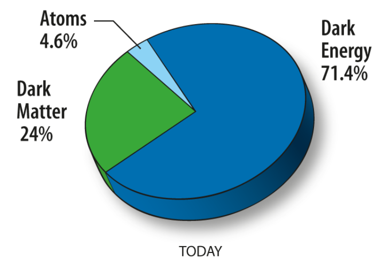
We need at least one of the statistical software mentioned in the previous slide. Please follow the instructions below

- `http://docs.continuum.io/anaconda/install`
- `https://cran.r-project.org/`
- `https://products.office.com/en-us/excel`
- `http://www-01.ibm.com/software/analytics/spss/`

DATA VISUALISATION

examples

- bar-charts
- histograms
- scatter-plots
- error-bars
- pie-charts
- many more!



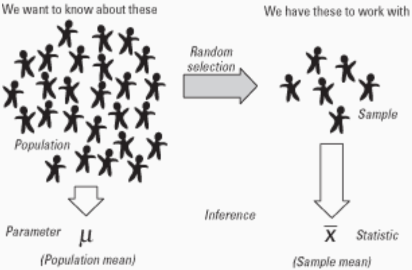
- We have several examples in the repository.
- Please follow the instructions in `https://github.com/tbs1980/CISLQuantWorkshop/tree/master/AdvancedQuantitativeMethodsClinic`.
- clustering
- outliers
- normalisation

DESCRIPTIVE STATISTICS

- Definitions
- Frequency distributions
- Central tendency and variability

Glossary

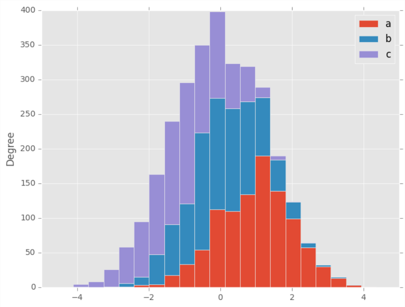
- Population
- Samples
- Variable
- Data
- Parameter
- Statistic



FREQENCY DISTRIBUTIONS

Defined by

- Size
- Range
- Bins-size
- Normalisation



How to characterise a distribution?

- What is a measure of central tendency?
- Mean, median and mode

The mean μ of samples $\{x_1, x_2, \cdots, x_n\}$ can be computed as

$$\mu = \frac{\sum_i x_i}{n} \tag{1}$$

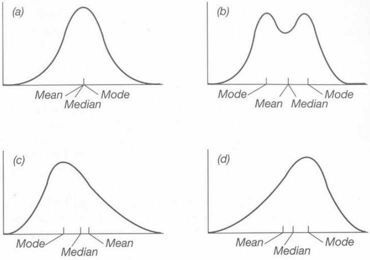


Figure 3.2 Frequency distributions showing measures of central tendency. Values of the variable are along the abscissa (horizontal axis), and the frequencies are along the ordinate (vertical axis). Distributions (a) and (b) are symmetrical, (c) is positively skewed, and (d) is negatively skewed. Distributions (a), (c), and (d) are unimodal, and distribution (b) is bimodal. In a unimodal asymmetric distribution, the median lies about one-third the distance between the mean and the mode.*

How to measure variations?

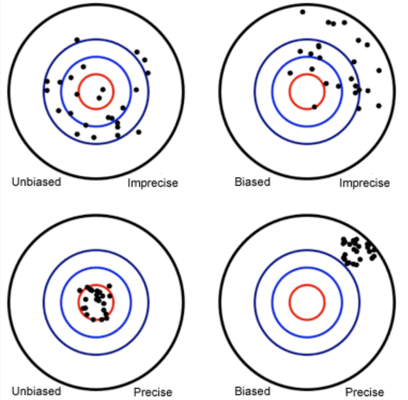
- Are you a good shooter?
- Variance and standard deviation
- Population and samples

The (biased) samples variance is defined as

$$\sigma^2 = \frac{\sum_i (x_i - \mu)^2}{n} \qquad (2)$$

We also define sums of squares SS as

$$SS = \sum x^2 \qquad (3)$$



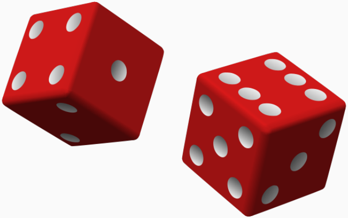
- How do we characterise skewed distributions?
- Concept of moments
- Distributions outside law of large numbers
- Examples can be found at `https://github.com/tbs1980/CISLQuantWorkshop/tree/master/AdvancedQuantitativeMethodsClinic/examples`
- Use the rest of the time for examples/discussion.

INFERENTIAL STATISTICS

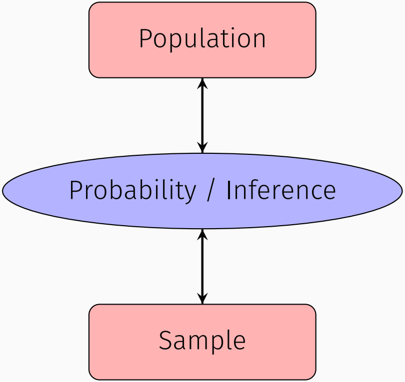
- Probability
- The Normal distribution
- Sample means and their distribution
- Introduction to hypothesis testing

Frequency or degree of belief?

- Frequency
- Desired outcome
- Random sample



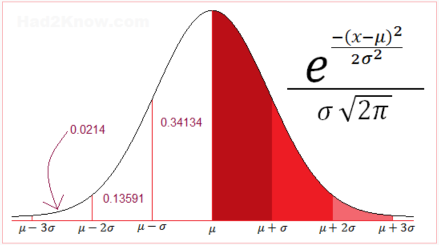
- What kind of samples are likely to obtained from the population?
- What can we say about the population given a sample?



Characteristics

- Mean μ
- Standard deviation σ
- Why is it important?
- Distribution of sample means

$$\Pr(x|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right) \tag{4}$$

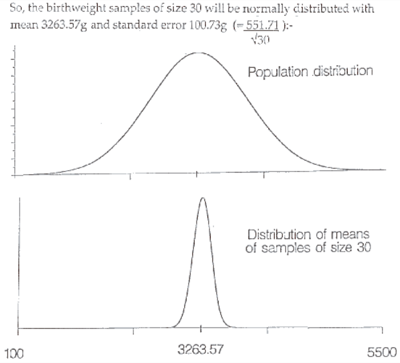


Characteristics

- Sampling error
- Distribution of sample means
- Expected value
- Standard error
- Law of large numbers

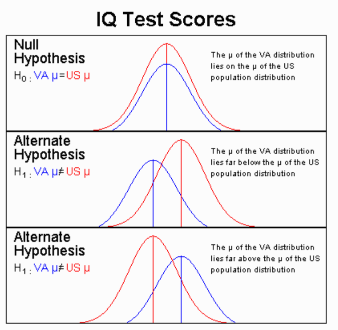
The standard error is defined as

$$\sigma_M = \frac{\sigma}{n} \tag{5}$$



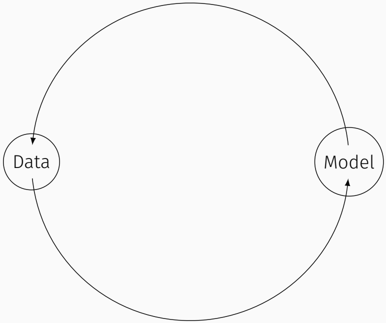
Baic idea

- Known versus unknown
- Null versus alternative hypothesis
- Decision criteria
- Level of significance
- Critical region
- Uncertainty and errors
- Statistical significance



Questions

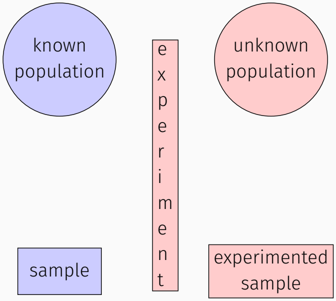
- Can we observe meaningful patterns in the data
- Are the findings statistically significant?
- Does the model adequately describe the data?
- Is there evidence for an alternative hypothesis?



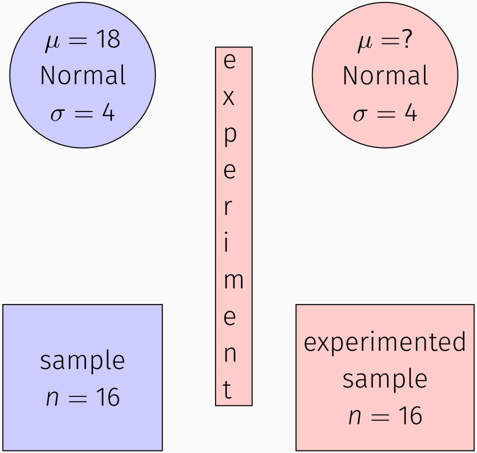
- Known characteristics of a population
- Selected sample for research
- Characteristics of the sample after experiment
- How do they compare?

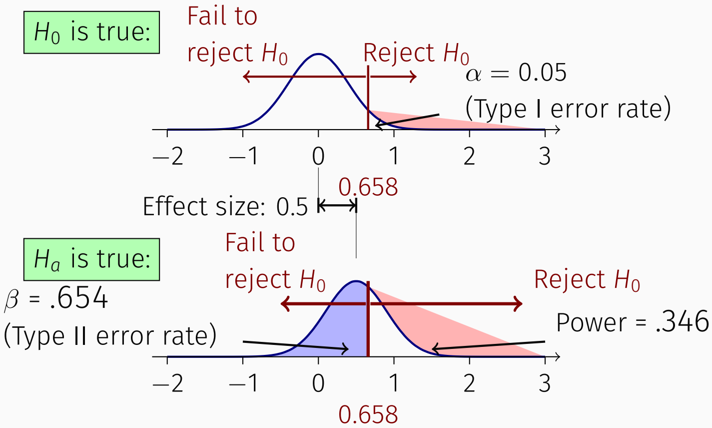
We deifne z-score as

$$z = \frac{M - \mu}{\sigma_M} \tag{6}$$



COMPARING CHANGES IN μ




H_0 is true:
Fail to reject H_0
Reject H_0
$\alpha = 0.05$
(Type I error rate)
−2
−1
0
1
2
3
0.658
Effect size: 0.5
H_a is true:
β = .654
(Type II error rate)
Fail to reject H_0
Reject H_0
Power = .346
−2
−1
0
1
2
3
0.658

- What if the H_0 is false?
- How accurate is the σ invariance assumption?
- How will we choose the level of significance?
- Examples can be found at `https://github.com/tbs1980/CISLQuantWorkshop/tree/master/AdvancedQuantitativeMethodsClinic/examples`
- Use the rest of the time for examples/discussion.

INFERENCES ABOUT POPULATION MEANS

- t-statistic
- ANalysis Of VAriacne (ANOVA)

Motivation

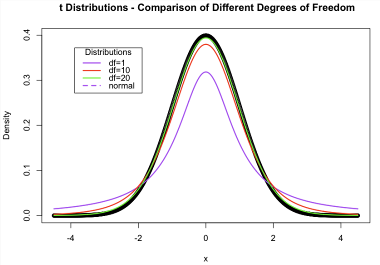
- Population σ unknown
- How can we compare means?
- Estimated standard error, s_M
- Degrees of freedom, df

We define the t-statistics as

$$t = \frac{M - \mu}{s_M} \tag{7}$$

and percentage of variance as

$$r^2 = \frac{t^2}{t^2 + df} \tag{8}$$



COMPARING TWO SETS OF DATA

How do we compare?

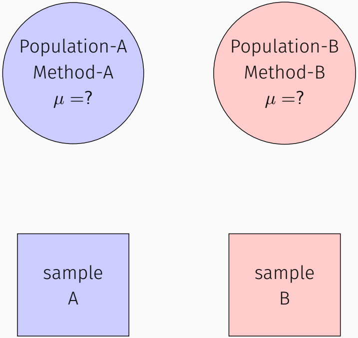
- Are there differences between population means?

$$t = \frac{(M_1 - M_2) - (\mu_1 - \mu_2)}{s_{(M_1 - M_2)}} \tag{9}$$

$$s_{(M_1 - M_2)} = \sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}} \tag{10}$$

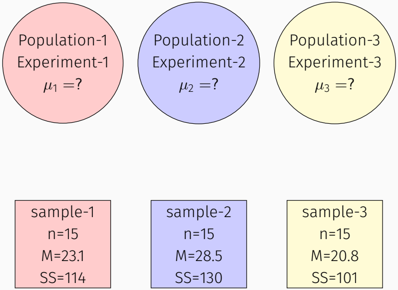
$$s_p = \frac{SS_1 + SS_2}{df} \tag{11}$$

$$df = df_1 + df_2 \tag{12}$$



ANLYSIS OF VARIACNE (ANOVA)

- Different samples for different experiments
- Are the results statistically different?
- How to differentiate between random and systematic variations?
- $H_0 : \mu_1 = \mu_2 = \mu_3$
- H_1 : at least one mean differences



ANOVA: THE FUNDAMENTAL CONCEPT

Measuring variability

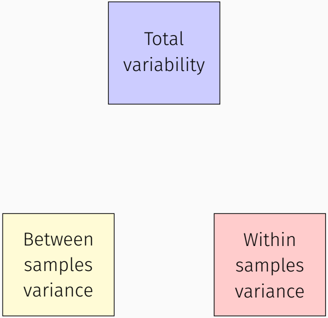
Recap: we can write the t-statistic as

$$t = \frac{\text{difference between sample means}}{\text{standard error}} \tag{13}$$

We define F-statistic as

$$F = \frac{\text{variance between sample means}}{\text{intrinsic variance}} \tag{14}$$

- systematic effects
- random, unsystematic factors



formulae

$$SS = \sum x^2 - \frac{(\sum x)^2}{N} \tag{15}$$

$$s^2 = \frac{SS}{df} \tag{16}$$

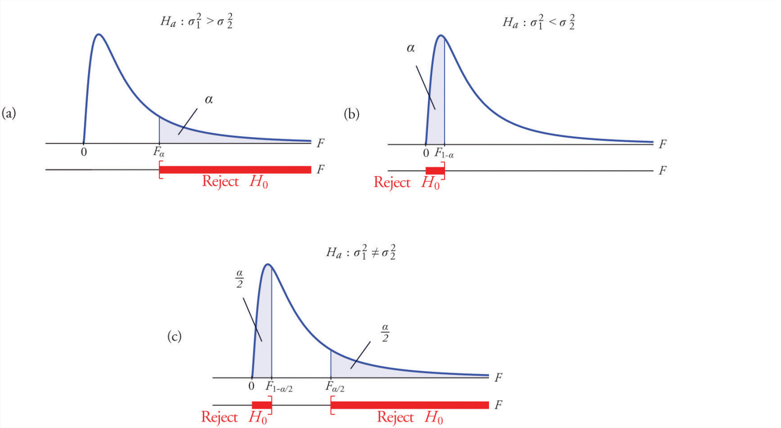
$$F = \frac{s^2_{\text{between}}}{s^2_{\text{within}}} \tag{17}$$

Total
$SS = \sum x^2 - \frac{G}{N}x$
$df = N - 1$

Between
$SS = \sum \frac{T^2}{n} - \frac{G}{N}x$
$df = k - 1$

Within
intrinsic $\sum SS$
$df = N - k$

F-STATISTIC



(a)
$H_a : \sigma_1^2 > \sigma_2^2$
α
0
F_α
F
F
Reject H_0
(b)
$H_a : \sigma_1^2 < \sigma_2^2$
α
0
$F_{1-\alpha}$
F
F
Reject H_0
(c)
$H_a : \sigma_1^2 \neq \sigma_2^2$
$\frac{\alpha}{2}$
$\frac{\alpha}{2}$
0
$F_{1-\alpha/2}$
$F_{\alpha/2}$
F
F
Reject H_0
Reject H_0

EXAMPLES/DISCUSSION

- Link to examples `https://github.com/tbs1980/CISLQuantWorkshop/tree/master/AdvancedQuantitativeMethodsClinic/examples`
- Use the rest of the time for examples/discussion.

REGRESSION

REGRESSION

- Correlation
- Parametric-regression
- χ^2-test

Covariability

- linear relationship between two variables x and y
- positive and negative

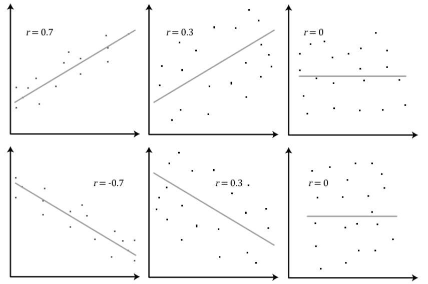
Pearson correlation is defined as

$$r = \frac{\text{co-variability of X and Y}}{\text{individual variability of X and Y}} \tag{18}$$

$$SP_{xy} = \sum xy - \frac{\sum x \sum y}{n} \tag{19}$$

$$r = \frac{S_{xy}}{\sqrt{SS_{xx} SS_{yy}}} \tag{20}$$

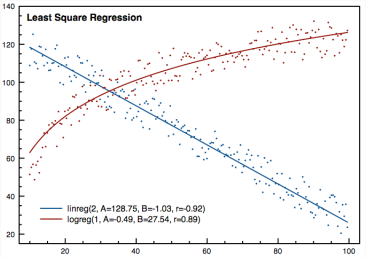
Visualisation



- outliers
- correlation and causation

Basics

- Model Vs data
- Predicted Vs observed
- Parametric Vs non-parametric
- Signal and noise
- Residuals
- $y = f(x) + n$
- least-squares



Computation

Linear model can be written as

$$y = a + bx \tag{21}$$

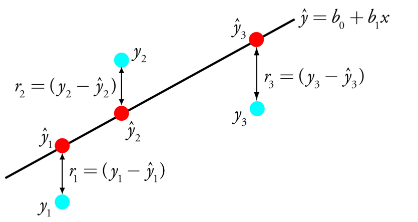
a and b can be calculated as

$$b = \frac{SS_{yy}}{SS_{xx}} \tag{22}$$

$$a = M_y - bM_x \tag{23}$$

The standard error is

$$err_{std} = \sqrt{\frac{SS_{\text{residual}}}{df}} = \frac{\sum(y - \hat{y})^2}{n - 2} \tag{24}$$



SIGNIFICANCE TEST

Definitions

Related to Pearson correlation as

$$\text{predicted variability} = r^2 SS_{yy} \tag{25}$$

$$\text{unpredicted variability} = (1 - r^2) SS_{yy} \tag{26}$$

Use F-test for significance test

$$MS_{\text{regression}} = \frac{SS_{\text{regression}}}{df_{\text{regression}}} = \frac{SS_{\text{regression}}}{1} \tag{27}$$

$$MS_{\text{residual}} = \frac{SS_{\text{residual}}}{df_{\text{residual}}} = \frac{SS_{\text{residual}}}{n-2} \tag{28}$$

$$F = \frac{MS_{\text{regression}}}{MS_{\text{residual}}} \tag{29}$$

ANOVA

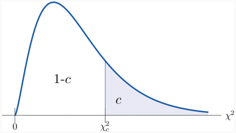
SS_{yy}
$df = n - 1$

$SS_{\text{regression}}$
$r^2 SS_{yy}$
$df = 1$

$SS_{\text{residuals}}$
$(1 - r^2) SS_{yy}$
$df = n - 2$

Goodness of fit

- How good is your model?
- observed and expected frequency
- χ^2 distribution



We define the χ^2 as

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e} \qquad (30)$$

EXAMPLES/DISCUSSION

- Link to examples `https://github.com/tbs1980/CISLQuantWorkshop/tree/master/AdvancedQuantitativeMethodsClinic/examples`
- Use the rest of the time for examples/discussion.

DISCUSSION AND SUMMARY

×	✓
majority respondents	60% of the respondents
shows large variation	sample variance is
large scatter	no correlation $r = 0.001$
treatment has an impact	the hypothesis tests shows that
cannot be sure	ANOVA shows that
we can confident	the probability oddes from F-test was
no difference between methods	t-statistic was calculated to be

SUMMARY

- Descriptive and inferential statistics
- Hypotheis testing
- Regression
- Best practices

REFERENCES

- Gravetter and Wallnau,*Statistics for the Behavioral Sciences*, 2013
- Wilcox, *Introduction to Robust Estimation and Hypothesis Testing*, 2012
- Grimmet, *Probability and Random Processes*, 2001
- Jaynes, *Probability Theory, The Logic of Science*, 2003
- Sivia and Skilling, *Data Analysis, A Bayesian Tutorial*, 2006