# General knowledge on Statistics and Data Analysis

**CLEMENT TWUMASI**

Email: twumasic@cardiff.ac.uk
School of Mathematics, Cardiff University, UK

November 27, 2020

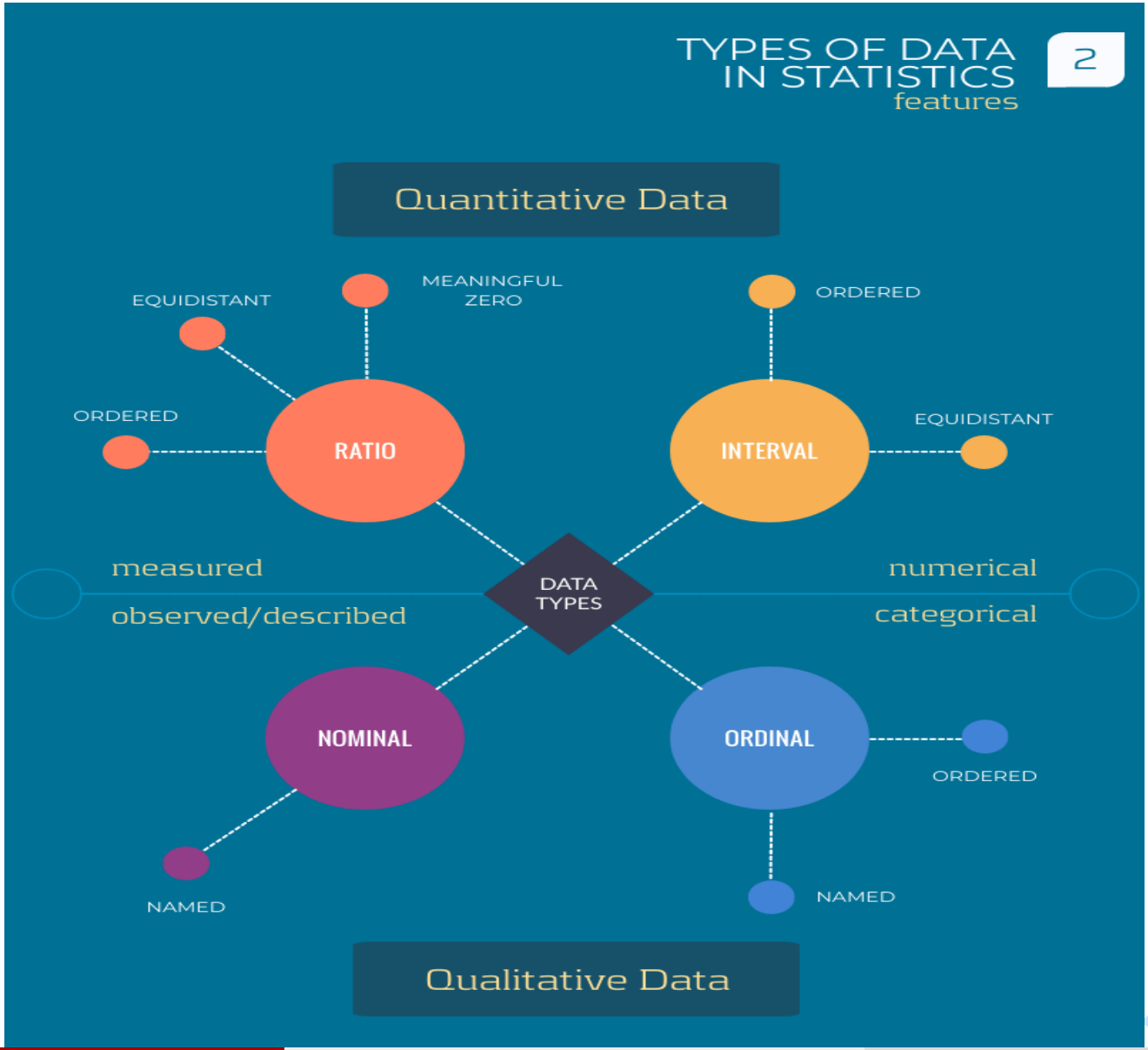
## Outline of presentation

- Brief introduction to statistics & its related fields.
- Useful information about Data Analysis.
- Regression analyses and variable selection techniques.
- Statistical software available and their advantages.

# Brief introduction to statistics & its related fields

- What is Statistics as a field of Mathematics?
- Any related fields to Statistics (Operational Research, Econometrics, Bio-statistics, Data Science, etc)?
- Two main statistical methods in Data Analysis (Descriptive & Inferential statistics).
- Brief introduction of useful terminologies (parameters, statistic, types of variables/data, scales of measurement, etc).
- Types of statistical tests & models (parametric and non-parametric methods)

# Brief introduction to statistics & its related fields
## Con't



# Data structure
## (cross-sectional data, time series data, panel data, etc)

Table 1. Example of cross sectional data

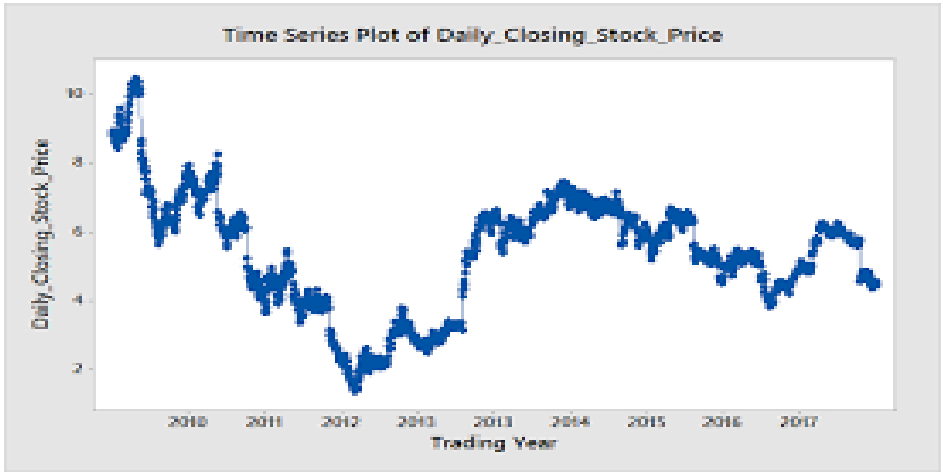
| indiv ($i$) | year | wage | edu | exper | female |
|---|---|---|---|---|---|
| 1 | 1990 | 3.10 | 11 | 2 | 1 |
| 2 | 1990 | 3.24 | 12 | 22 | 1 |
| . | . | . | . | . | . |
| 100 | 1990 | 5.30 | 12 | 7 | 0 |

Table 2. Example of pooled cross sectional data

| house ($i$) | year ($t$) | hprice | bdrms | bthrms | sqrft |
|---|---|---|---|---|---|
| 1 | 2000 | 85,500 | 3 | 2.0 | 1600 |
| 2 | 2000 | 67,300 | 3 | 2.5 | 1400 |
| . | . | . | . | . | . |
| 100 | 2000 | 134,000 | 4 | 2.5 | 2000 |
| 101 | 2010 | 243,000 | 4 | 3.0 | 2600 |
| 102 | 2010 | 65,000 | 2 | 1.0 | 1250 |

Table 3. Example of panel data (aka, longitudinal data)

| obs. | $i$ | $t$ | murder rate | pop density | police |
|---|---|---|---|---|---|
| 1 | 1 | 2000 | 9.3 | 2.24 | 440 |
| 2 | 1 | 2001 | 11.6 | 2.38 | 471 |
| 3 | 2 | 2000 | 7.6 | 1.61 | 75 |
| 4 | 2 | 2001 | 10.3 | 1.73 | 75 |
| . | . | . | . | . | |
| 199 | 100 | 2000 | 11.1 | 11.1 | 520 |
| 200 | 100 | 2001 | 17.2 | 17.2 | 493 |



- Mean tests (Parametric tests: Paired t-tests, independent t-test, ANOVA, Bonferroni pairwise comparison tests, Repeated Measures ANOVA, MANOVA,etc).
- Median tests (Non-parametric tests: Wilcoxon sign test, Mann-Whitney, Kruskal-Wallis test, Bonferroni-Dunn's test, Friedman test, Multivariate Kruskal-Wallis test,etc).
- Test of proportions for one or more groups, etc.
- Correlation test/analysis

- GLM (OLS regression, binomial & multinomial regression, poisson regression, quasi poisson regression, negative binomial regression, etc) [cross-sectional data]
- Generalized Linear Mixed Models (for all types of dependent variables) [these class of models are for longitudinal data]
- Generalized Additive Models (fixed and mixed-effect types; both cross-sectional & longitudinal data) and Panel regression models [these class of models for longitudinal data]
- Machine learning algorithms (Classification tree, Random forest, Gradient Boosting Machine, etc) [both cross-sectional & longitudinal data]

**Method of variable selection:** Stepwise regression, Penalized regression (Ridge, LASSO and Elastic net), Recursive Feature Selection.

**NB:** Beware of multicollinearity and its effects.

# Statistical software available
# (their advantages over the other)

- SPSS, STATA, EVIEWS, MINITAB, GRETL, XLSTAT, etc,
- R
- Python
- Julia
- SAS
- Matlab,
- C language , C++, etc.