Preliminaries of Medical Statistics
: A Brief Example via Original Study

Sangjun Park, MS

Korea University, College of Medicine

Division of Foregut Surgery

jack2020@korea.ac.kr

010-5603-3066

KU MEDICINE KOREA UNIVERSITY MEDICINE

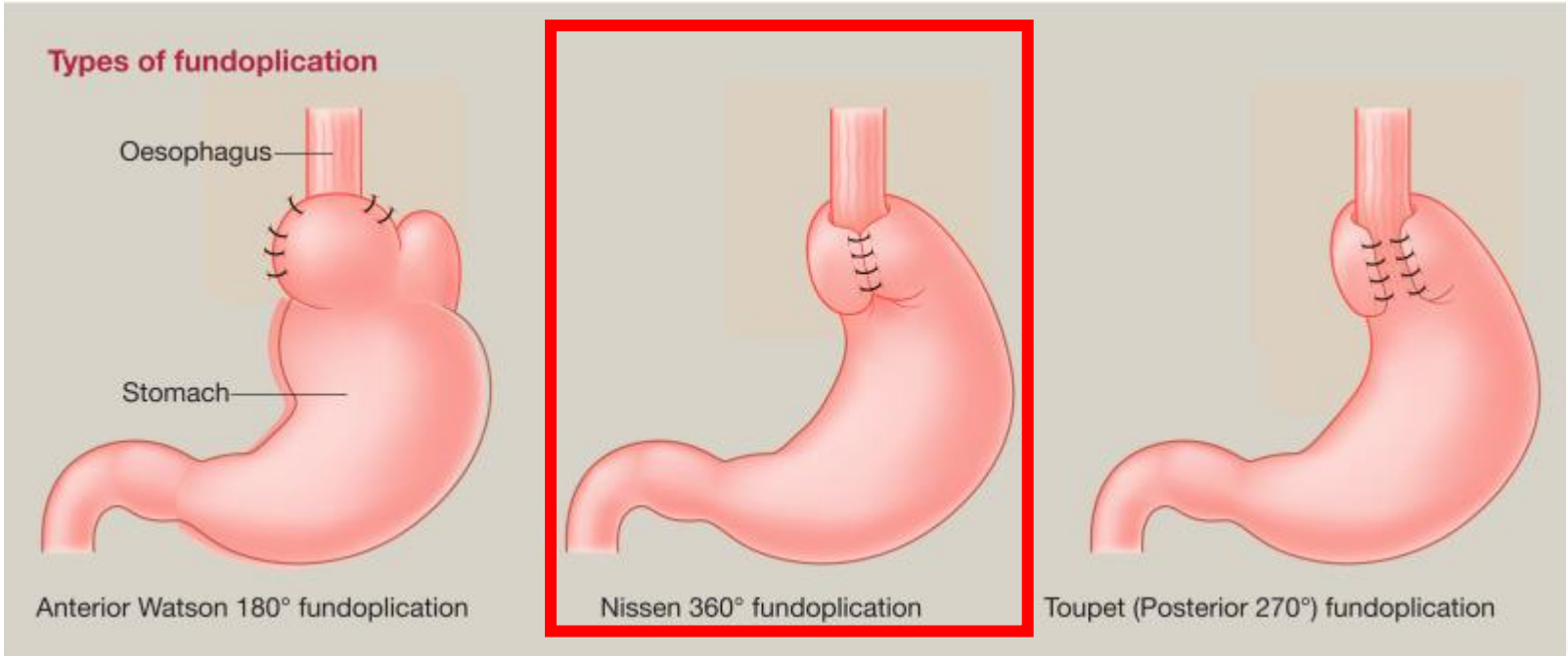
October 2020

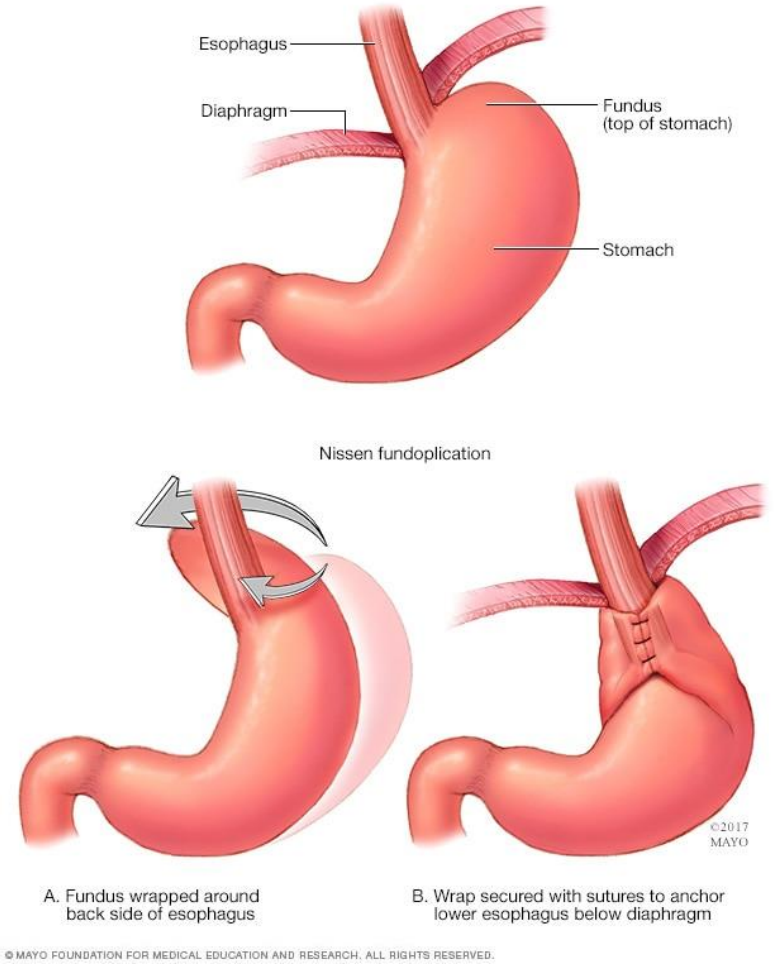
Outline of Statistics

- Data handling techniques → Proof of Concept
- Major Fields
  - Design of Experiments – Randomization, Blinding, Sample size determination
    - Clinical Trials (RCT)
  - Sampling - Distribution estimation, Bayesian Statistics
    - Quality control
    - Survey methodology
  - Descriptive Statistics – Exploratory data analysis (EDA)
  - Inferential Statistics – Point/Interval Estimation, Hypothesis Rejection
  - Predictive Modeling - Classical regression w/ variants, Clustering, Multivariate analysis, Forecasting, Classification, Clustering, Machine learning, etc.
  - Theoretical Statistics – Matrix calculus, Linear algebra, Probability theory, Mathematical Statistics, Real analysis, Measure theory, Differential equations, Numerical Analysis, etc.
- Know your tools before using them!

Motive Background

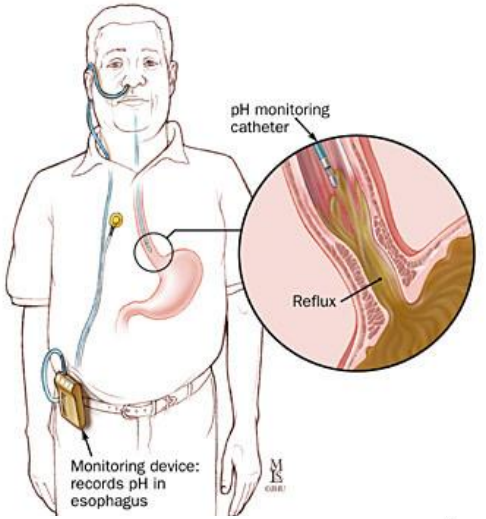
- GERD (Gastroesophageal Reflux Disease)
- Medical Tx. Proton-Pump Inhibitor
- Surgical Tx. Laparoscopic Nissen Fundoplication

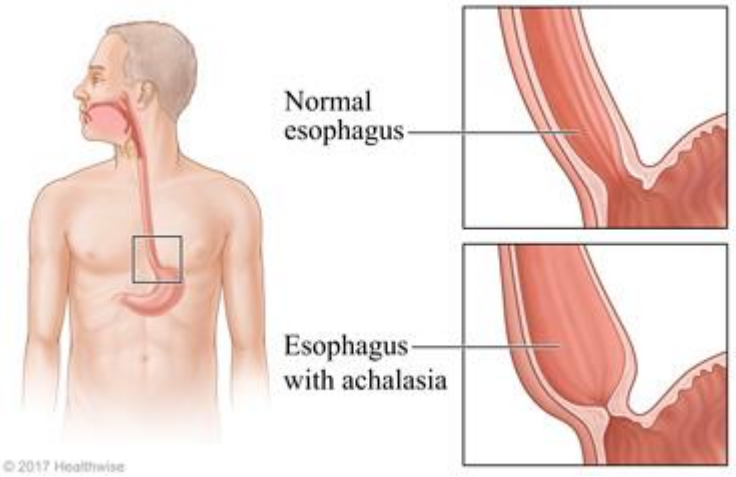


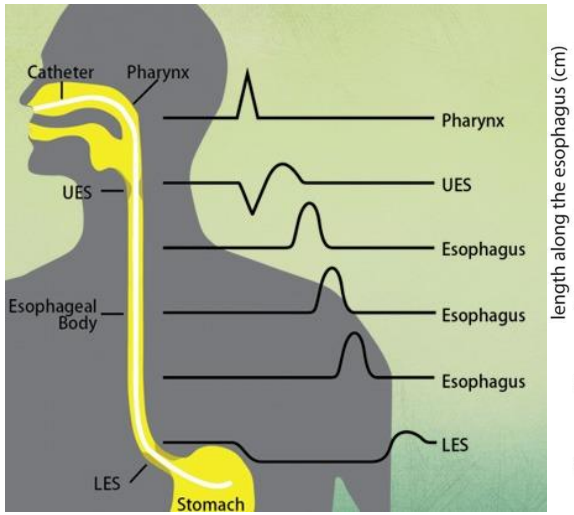


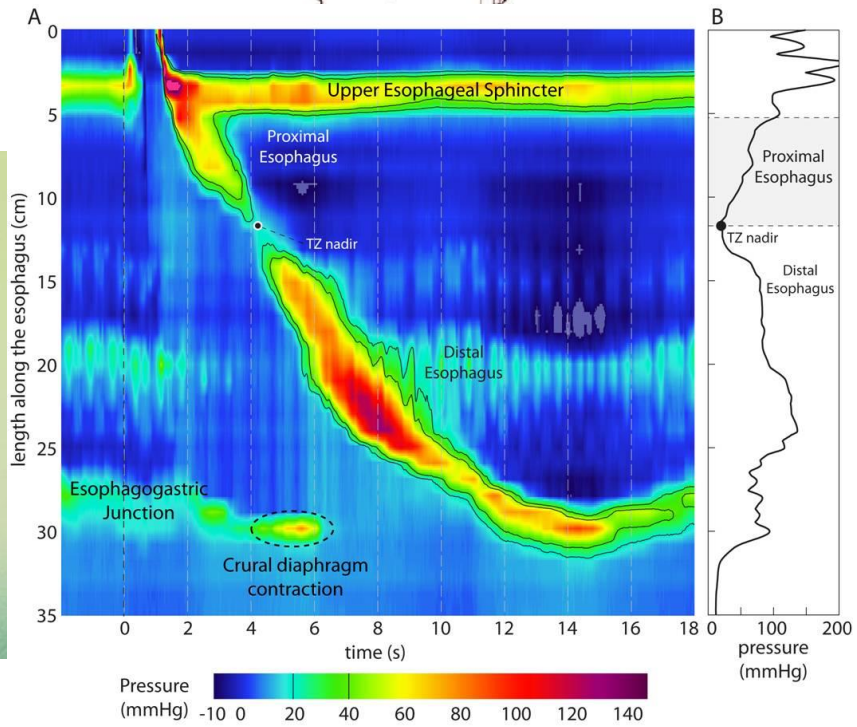
Motive Background

- Upper GI Endoscopy -> Detection of Esophagitis
- 24h pH Monitoring -> DDx. of GERD, NERD, …
- Esophageal Manometry -> Only for achalasia?



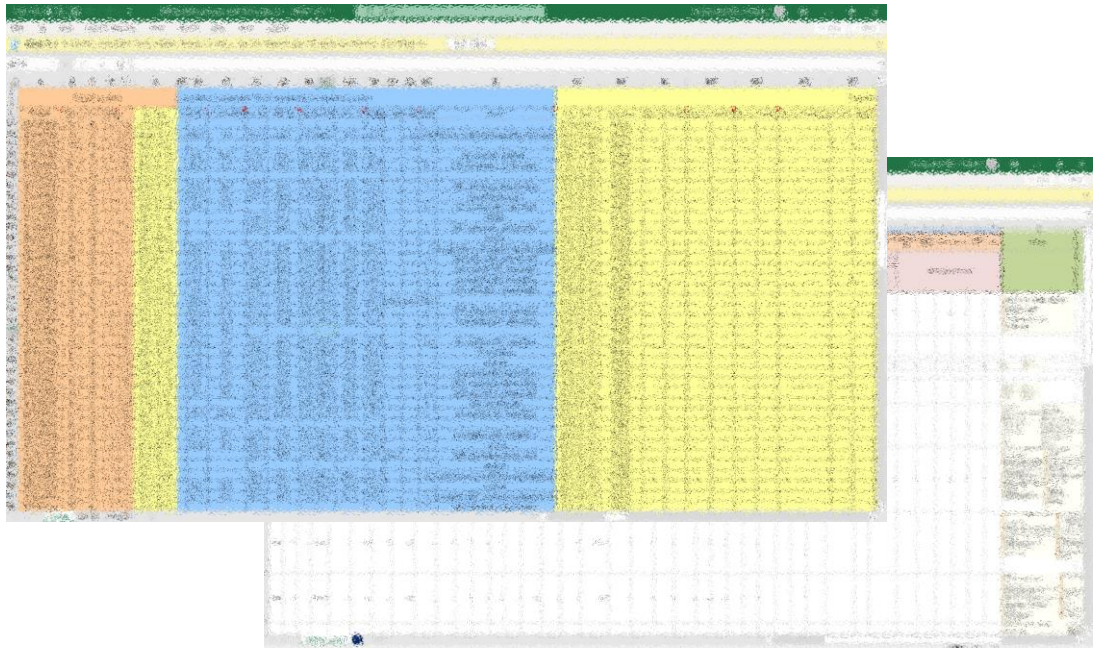


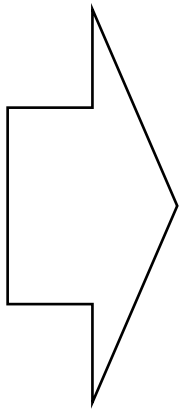


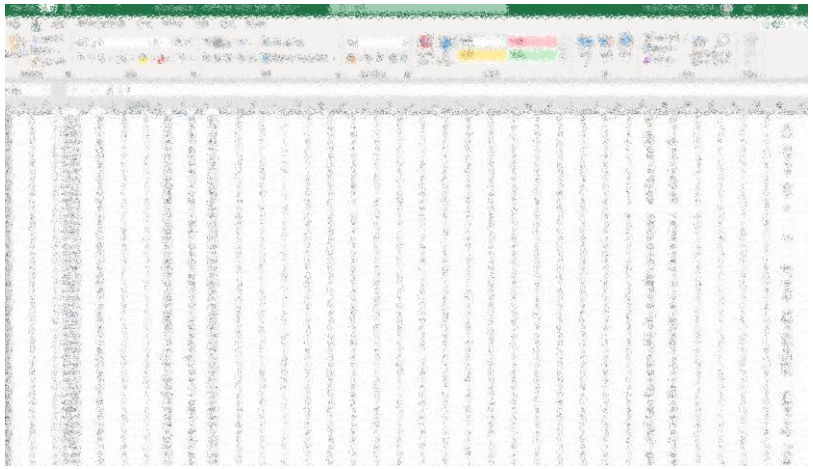


Data Preprocessing

- The literal example of manual labor. Tedious but inevitable.
- Make sure no mistakes happen!
- Tips) Include all missing cells/NA/0 values. Make separate versions for redo. All factors should be columnized. No "tables" are allowed. etc…

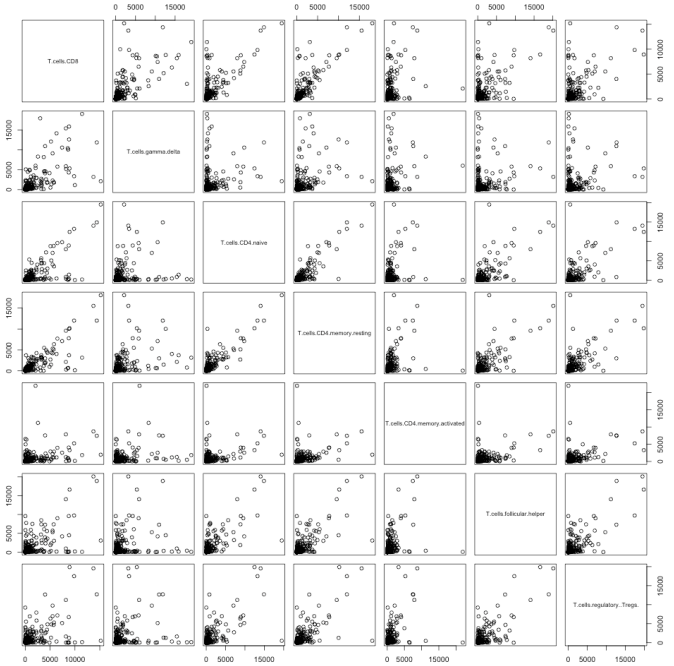


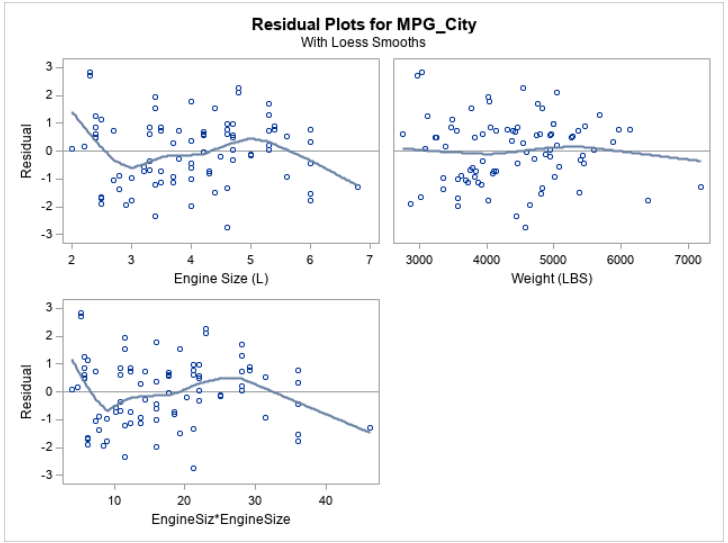


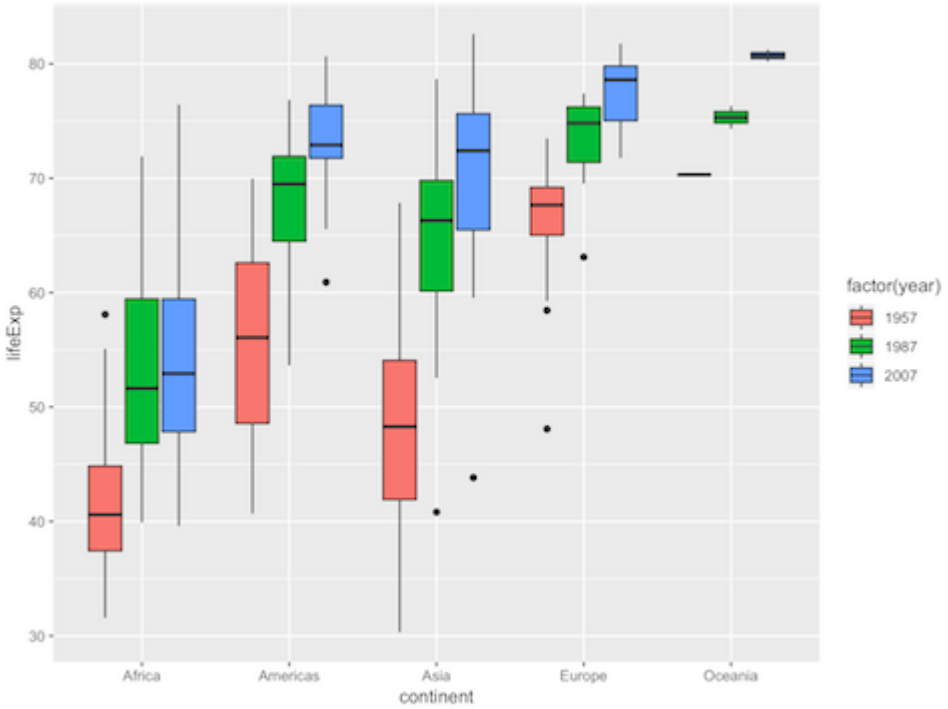


Exploratory Data Analysis (EDA)

- Analysis of the main characteristics and structural features of the data
- Resistance, Residuals, Re-expression, Revelation (Hoaglin, Mosteller and Tukey, 1982)
- Quick glance of intuition through graphing!







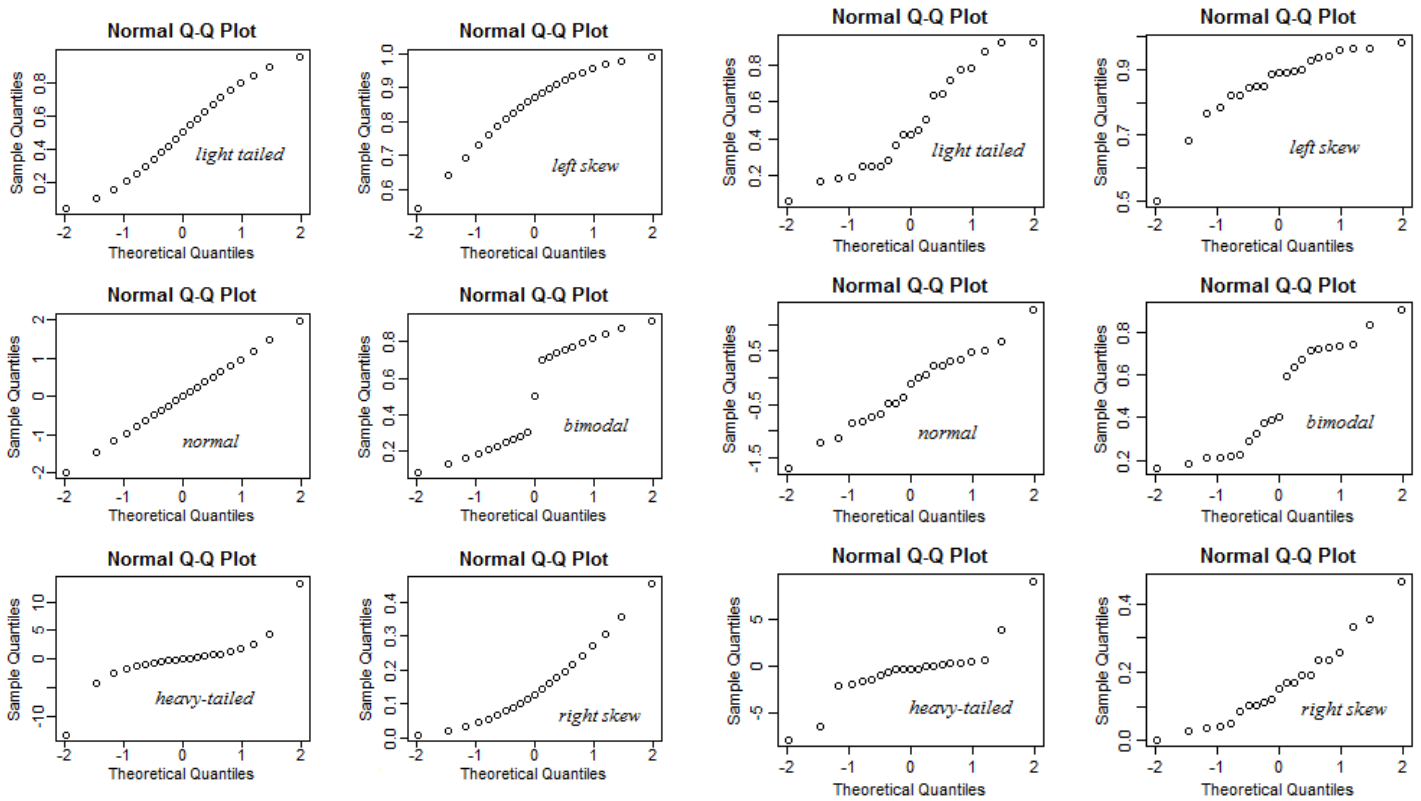
EDA i.e., Q-Q (quantile-quantile) plot

- Comparison of quantiles between two distributions
- Frequently used for checking the normality of the data
- **Normality**
  - **Yes -> Parametric Test**
  - **No -> Nonparametric Test**
  - **Uncertain (Outliers, Slight Deviances)**
    -> Robust Test
- What if the data is too small?

  ->Shapiro-Wilk Test!

  (*Not good for large datasets)

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

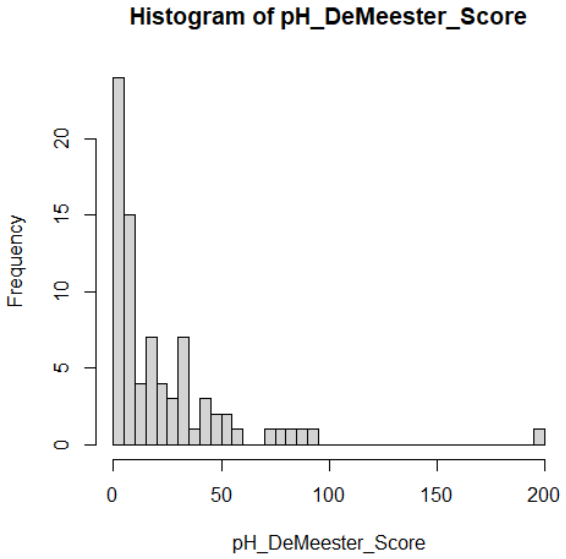


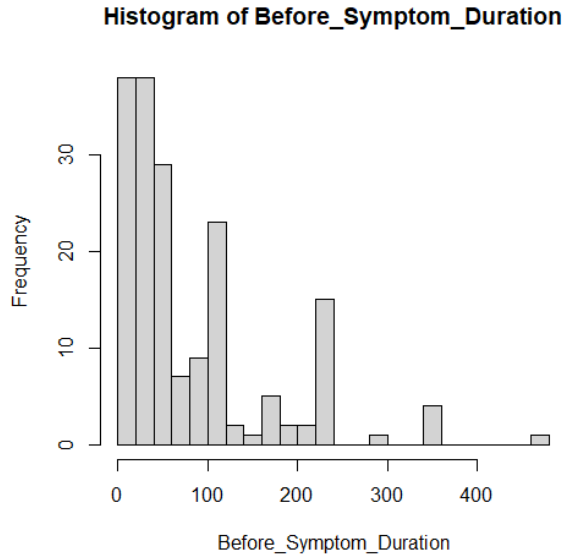
Large Dataset (Interpretable)

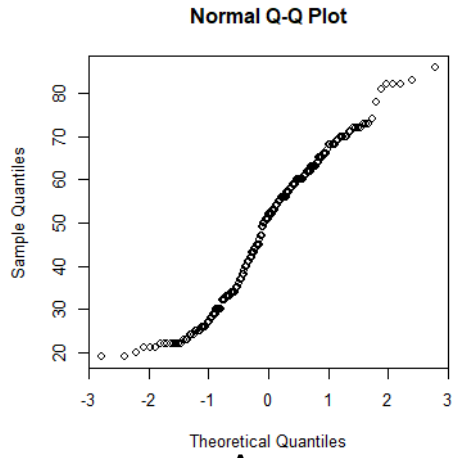
Small Dataset (Ambiguous)

EDA i.e., Q-Q (quantile-quantile) plot

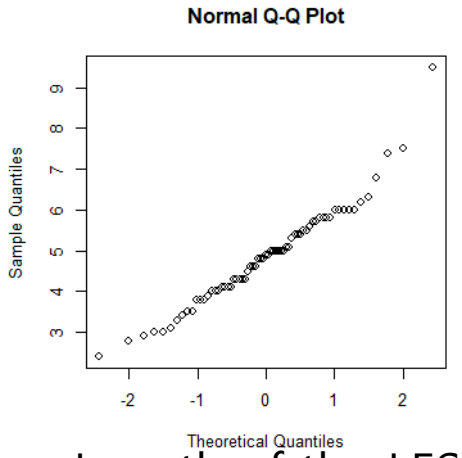
- Age, Length of LES variables seem OK
  -> Parametric Statistics
- DeMeester Score, Duration of GERD
  -> Right-skewed!
  -> Should use nonparametric statistics



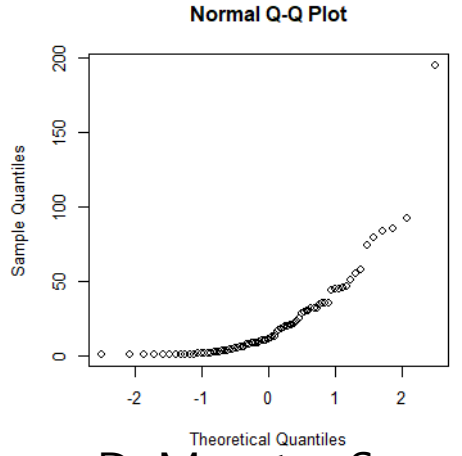




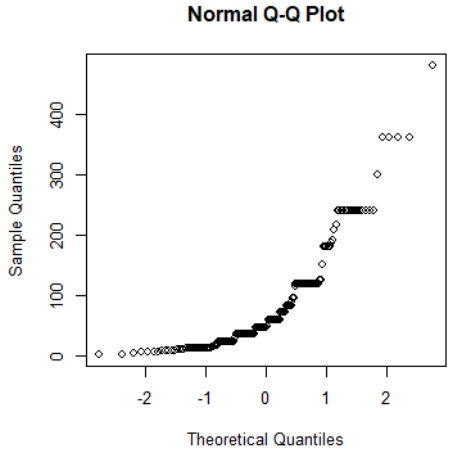
Age



Length of the LES



DeMeester Score



Duration of GERD

Importance of Statistical Assumptions

- **Always check the tests' assumptions!**
  - Normality
  - i.i.d. (independently identically distributed)
  - Homoscedasticity
  - Linearity
  - Sphericity
  - …
- Wrong tests lead to wrong conclusions, and too much robustness leads to weak power.
- Rao-Blackwell Theorem, Neyman-Pearson Lemma, Wilks' Theorem,…

Statistical analysis[a]	Independence	Measurement Level of variable(s)[b]		Normality	Linearity	Variance
		Dependent	Independent			
CHI-SQUARE						
Single sample	Independent observations	Nominal+	N/A			
2+ samples	Independent observations	Nominal+	Nominal+			
***T*-TEST**						
Single sample	Independent observations	Continuous	N/A	Univariate		
Dependent sample	Independent paired observations	Continuous	N/A	Univariate		
Independent sample	Independent observations	Continuous	Dichotomous	Univariate		Homogeneity of variance
OVA-RELATED TESTS						
ANOVA	Independent observations	Continuous	Nominal	Univariate		Homogeneity of variance
ANCOVA[c]	Independent observations	Continuous	Nominal	Univariate	✓	Homogeneity of variance
RM ANOVA	Independent repeated observations	Continuous	Nominal (opt.)	Multivariate	✓	Sphericity
MANOVA	Independent observations	Continuous	Nominal	Multivariate	✓	Homogeneity of covariance matrix
MANCOVA[c]	Independent observations	Continuous	Nominal	Multivariate	✓	Homogeneity of covariance matrix
REGRESSION						
Simple linear	Independent observations	Continuous	Continuous	Bivariate	✓	
Multiple linear	Independent observations	Continuous	Continuous	Multivariate	✓	Homoscedasticity
Canonical correlation	Independent observations	Continuous	Continuous	Multivariate	✓	Homoscedasticity

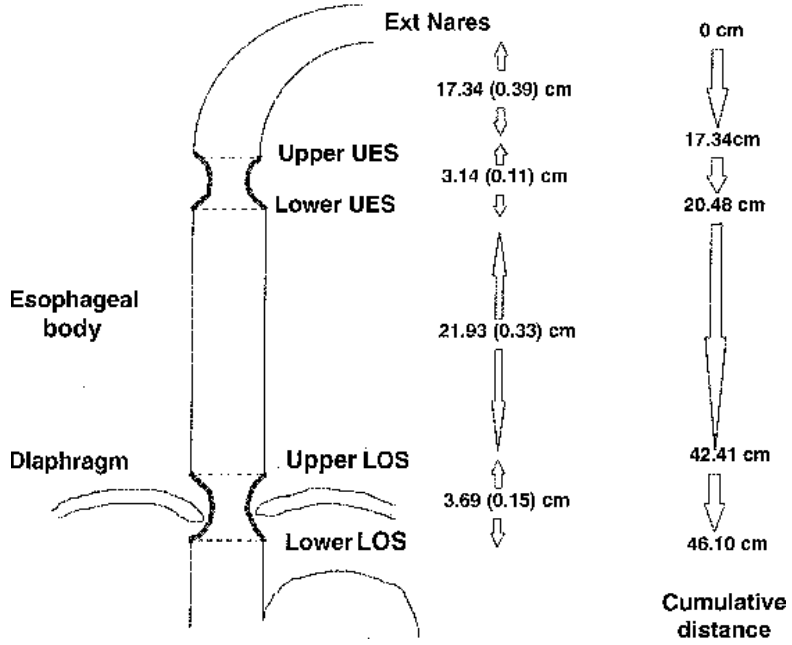
[a]Across all analyses, data are assumed to be randomly sampled from the population. [b]Data are assumed to be reliable. [c]ANCOVA and MANCOVA also assumes homogeneity of regression and continuous covariate(s). Continuous refers to data that may be dichotomous, ordinal, interval, or ratio (cf. Tabachnick and Fidell, 2001).

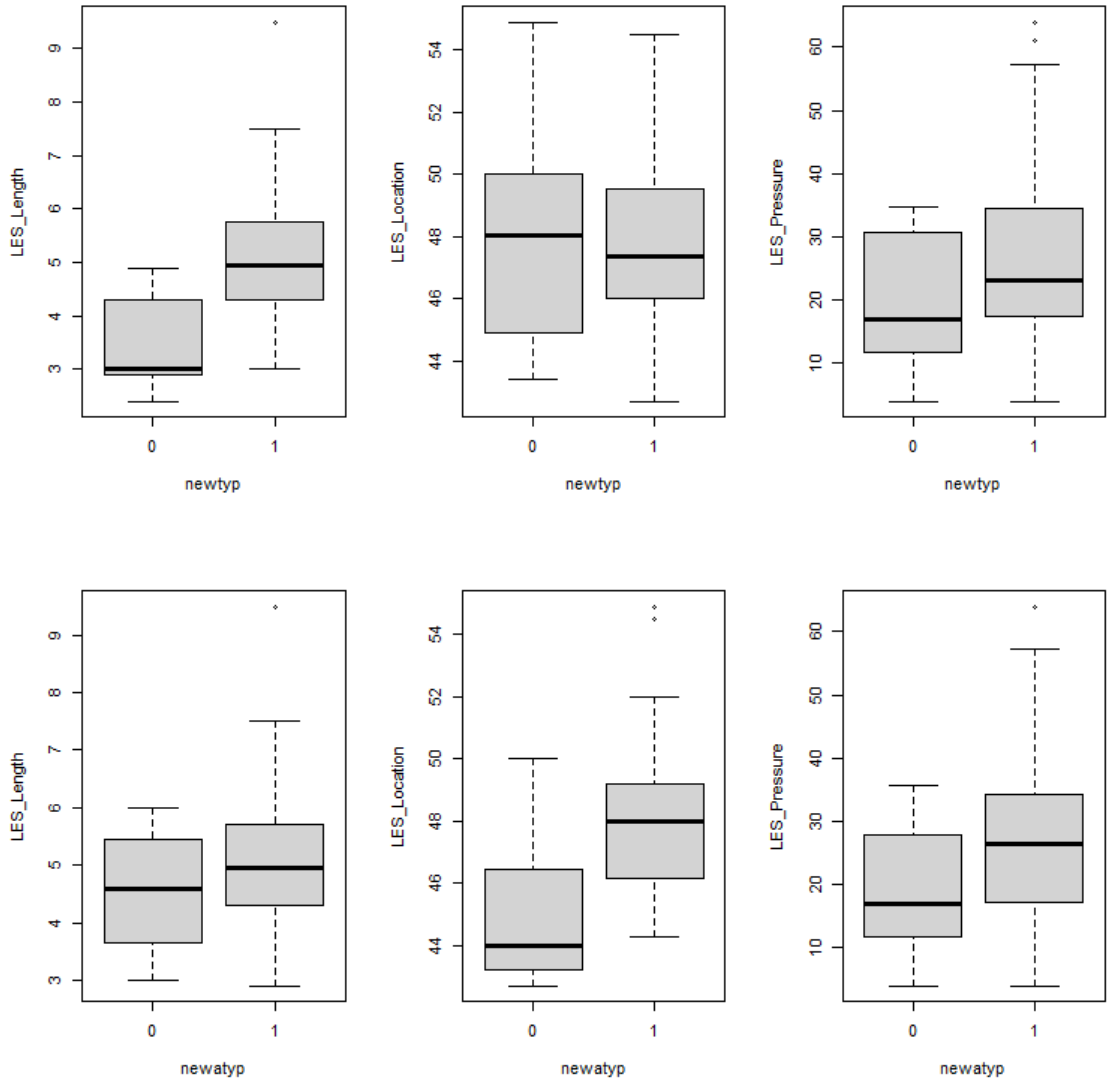
Nimon, Kim. (2012). Statistical Assumptions of Substantive Analyses Across the General Linear Model: A Mini-Review. Frontiers in psychology. 3. 322. 10.3389/fpsyg.2012.00322.

Parametric Tests	Nonparametric Tests
T-test	Mann-Whitney U-test
ANOVA	Kruskal-Wallis
Repeated-ANOVA	MANOVA, Friedman test, Durbin test
Chi-squared Test	Fisher's Exact test
Pearson Correlation	Spearman's rank Correlation
Proportional Hazards	Cox regression (Semi-parametric) Kaplan-Meier (Non-parametric)
⋮	⋮

EDA i.e., Boxplot

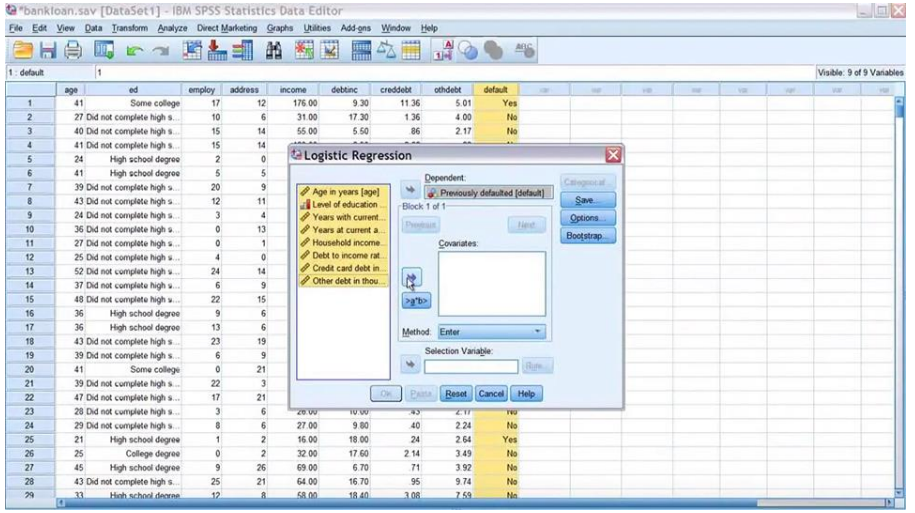
- Easy to check factors with significant difference between multiple groups
- Resolution of typical symptoms seems to be related to the length of LES
- Resolution of atypical symptoms seems to be related to the location of LES



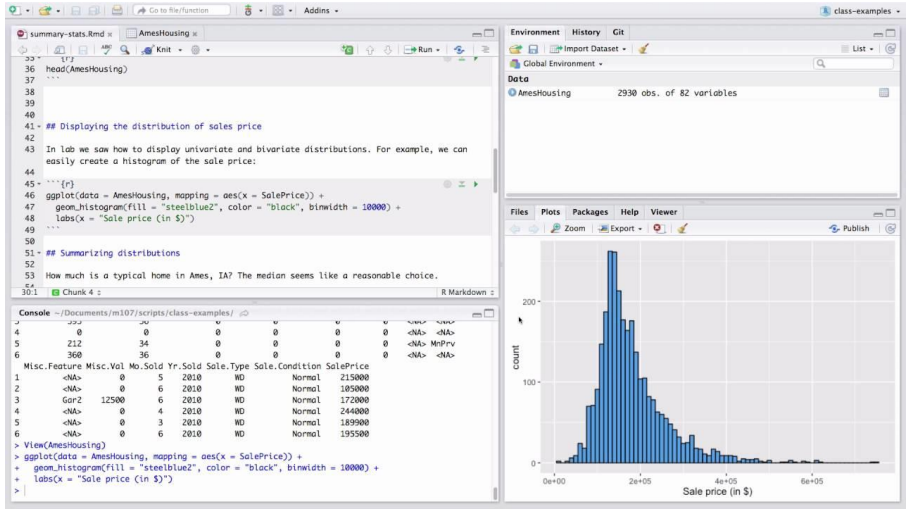


Statistical Softwares

- SPSS
  - Propriety Software – Use the university portal or get the student discount version
  - "Clickable", Easy-to-use, No programming language needed!
  - Frequently used in the medical field
- SAS
  - Mostly used in limited corporate settings
  - SQL for large database managements?
- R
  - Open-source software with various pre-made packages
  - Has extreme popularity due to its fast computing ability and flexibility (with programming skills)
  - Install with Rstudio (GUI, Best IDE)
- Python?
  - Pandas (Raw Data Handling)
  - Numpy (Matrix Calculation)
  - Scipy, PyStan, PyTorch (Machine Learning)



SPSS



R (Rstudio)

Recommendation of Reference Materials

- Matrix Calculus – Magnus
- Linear Algebra – Lay (Introductory, Only has all you need to know), Strang (Moderate), Friedberg (Slightly advanced, For mathematics majors), Hoffman (Hard-core with abstract algebra)
- Probability Theory – Ross (Introductory), Billingsley (Graduate, Measure-theory, Extremely rigorous, Prereq. Real Analysis)
- Mathematical Statistics – Hogg (Moderate, Useful in gaining insight), Casella (Graduate)
- Exploratory Data Analysis – Tukey (Old but classic)
- Categorical Data Analysis – Agresti (Two versions, Introductory and Advanced)
- Nonparametric Statistics – Conover
- Regression Analysis – Chatterjee (Moderate)
- Experimental Design – Montgomery (Moderate)
- Survival Analysis – Kleinbaum, Collett
- Meta-Analysis – Borenstein
- Multivariate Statistics – Johnson (Moderate provided that one is already familiar with linear algebra and matrix calculus)
- Bayesian Statistics (Learn sampling methods first) – Gelman (Bible of BDA), Hoff
- Machine Learning – Deep learning (I. Goodfellow, Standard text for neural network engineering), An introduction to statistical learning (G. James, Moderate), The elements of statistical learning (Hastie, Graduate, Mostly deals with advanced regression techniques), Pattern recognition and Machine learning (Bishop, Graduate), Convex optimization (Boyd, Entirely Mathematical)

Good Luck!

All models are wrong, but some are useful.

- George E. P. Box (1919-2013)