Lesson 4: Fundamantals of Visual Analytics

Dr. Kam Tin Seong
Assoc. Professor of Information Systems

School of Computing and Information Systems,
Singapore Management University

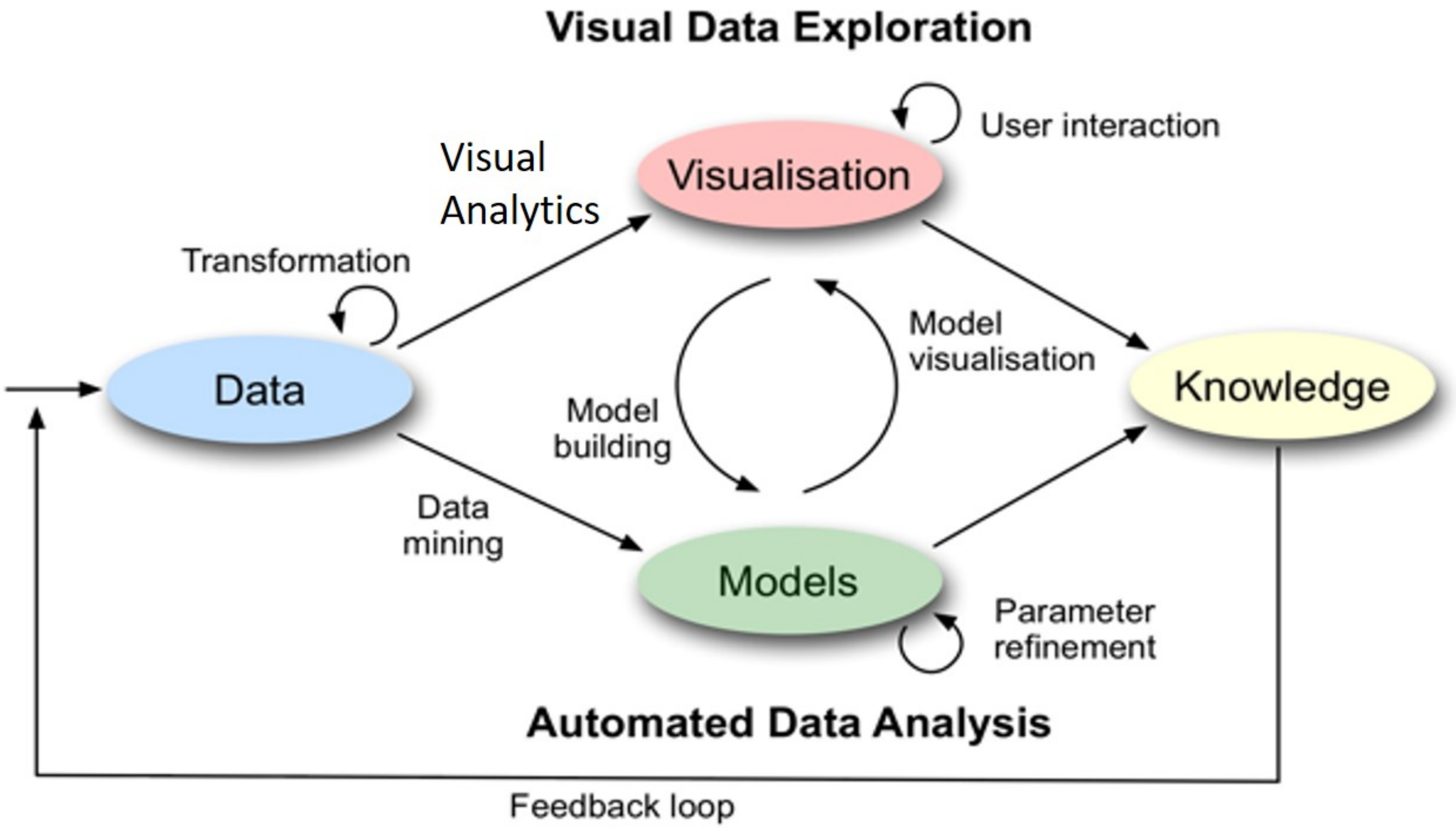
2020-01-20 (updated: 2022-05-04)

What will you learn from this lesson?

- Visual Analytics for Knowledge Discovery
- Visual Analytics Approach for Statistical Testing
- Visual Analytics for Building Better Models
- Visualising Uncertainty
- Variation and Its Discontents

Visually Analytics for Knowledge Discovery

Motivation: To combine data visualisation and statistical modeling



Visual Statistical Testing

To provide alternative statistical inference methods by default.

INTERNATIONAL REVIEW OF SOCIAL PSYCHOLOGY

Delacre, M., et al (2017). Why Psychologists Should by Default Use Welch's *t*-test Instead of Student's *t*-test. *International Review of Social Psychology*, 30(1), 92–101, DOI: https://doi.org/10.5334/irsp.82

RESEARCH ARTICLE

Why Psychologists Should by Default Use Welch's *t*-test Instead of Student's *t*-test

Marie Delacre*, Daniël Lakens† and Christophe Leys*

When comparing two independent groups, psychology researchers commonly use Student's *t*-tests. Assumptions of normality and homogeneity of variance underlie this test. More often than not, when these conditions are not met, Student's *t*-test can be severely biased and lead to invalid statistical inferences. Moreover, we argue that the assumption of equal variances will seldom hold in psychological research, and choosing between Student's *t*-test and Welch's *t*-test based on the outcomes of a test of the equality of variances often fails to provide an appropriate answer. We show that the Welch's *t*-test provides a better control of Type 1 error rates when the assumption of homogeneity of variance is not met, and it loses little robustness compared to Student's *t*-test when the assumptions are met. We argue that Welch's *t*-test should be used as a default strategy.

INTERNATIONAL REVIEW OF SOCIAL PSYCHOLOGY

Delacre, M., et al. (2019). Taking Parametric Assumptions Seriously: Arguments for the Use of Welch's *F*-test instead of the Classical *F*-test in One-Way ANOVA. *International Review of Social Psychology*, 32(1): 13, 1–12. DOI: https://doi.org/10.5334/irsp.198

RESEARCH ARTICLE

Taking Parametric Assumptions Seriously: Arguments for the Use of Welch's *F*-test instead of the Classical *F*-test in One-Way ANOVA

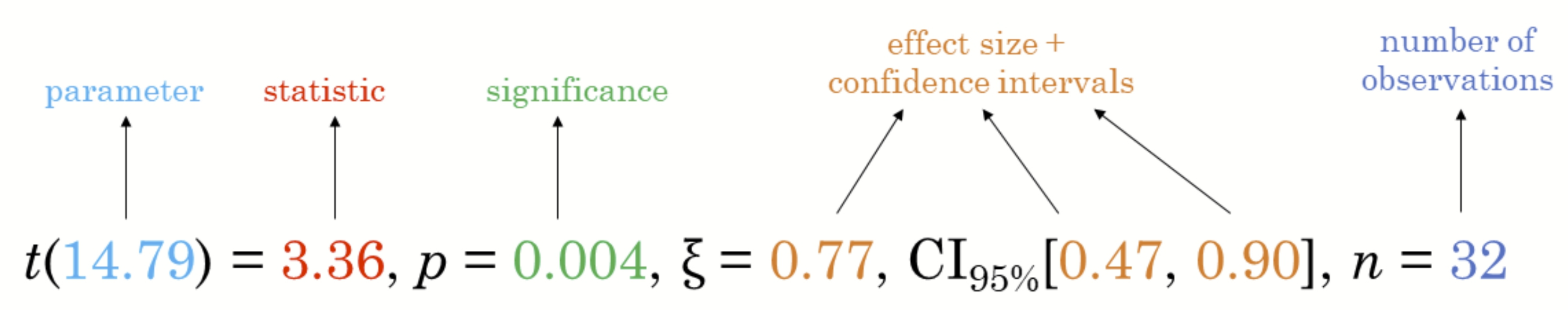
Marie Delacre*, Christophe Leys*, Youri L. Mora* and Daniël Lakens†

Student's *t*-test and classical *F*-test ANOVA rely on the assumptions that two or more samples are independent, and that independent and identically distributed residuals are normal and have equal variances between groups. We focus on the assumptions of normality and equality of variances, and argue that these assumptions are often unrealistic in the field of psychology. We underline the current lack of attention to these assumptions through an analysis of researchers' practices. Through Monte Carlo simulations, we illustrate the consequences of performing the classic parametric *F*-test for ANOVA when the test assumptions are not met on the Type I error rate and statistical power. Under realistic deviations from the assumption of equal variances, the classic *F*-test can yield severely biased results and lead to invalid statistical inferences. We examine two common alternatives to the *F*-test, namely the Welch's ANOVA (*W*-test) and the Brown-Forsythe test (*F**-test). Our simulations show that under a range of realistic scenarios, the *W*-test is a better alternative and we therefore recommend using the *W*-test by default when comparing means. We provide a detailed example explaining how to perform the *W*-test in SPSS and R. We summarize our conclusions in practical recommendations that researchers can use to improve their statistical practices.

Visual Statistical Testing

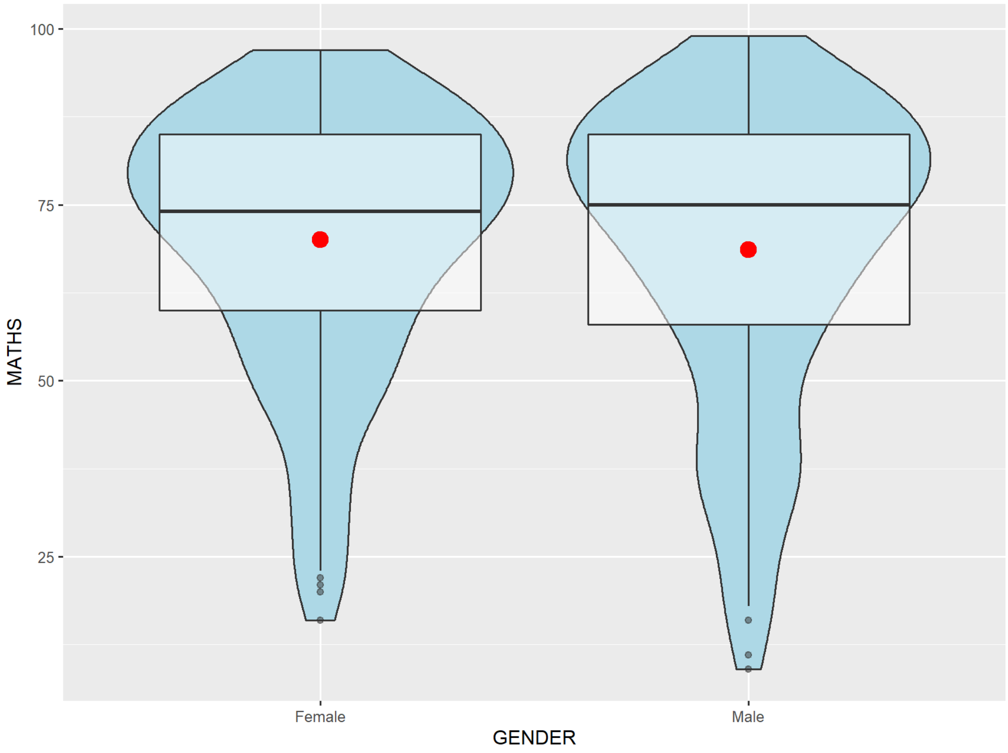
To follow best practices for statistical reporting.

- For all statistical tests reported in the plots, the default template abides by the APA gold standard for statistical reporting. For example, here are results from a robust t-test:

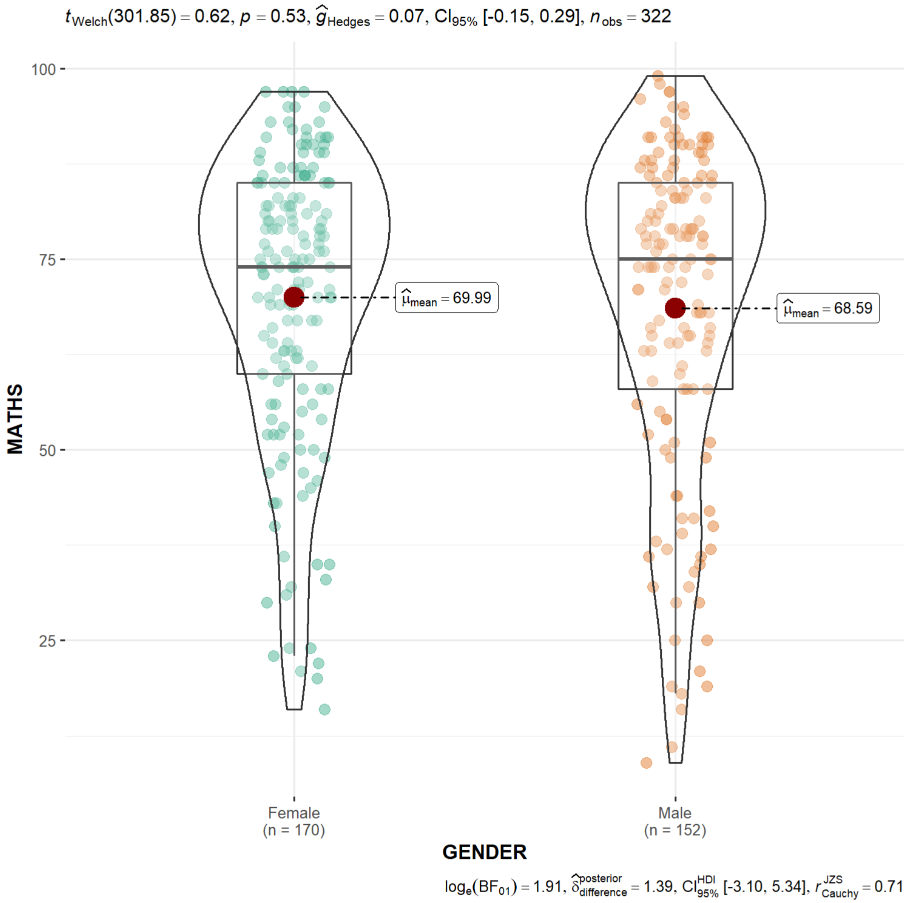


Two-sample means

Boxplot revealing the mean and distribution of two samples.

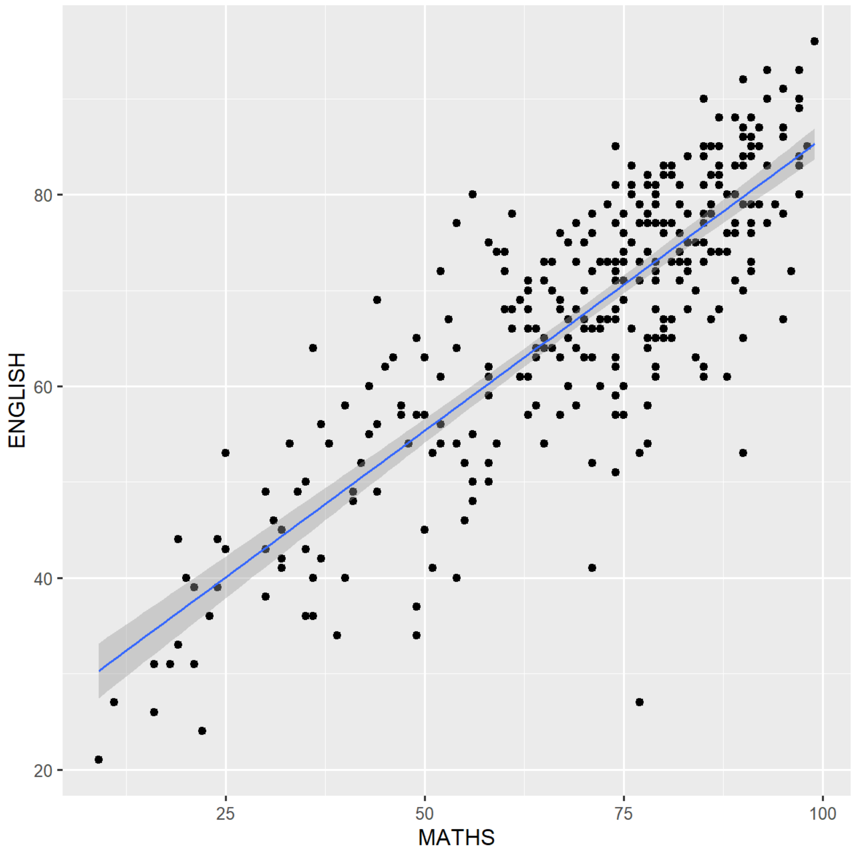


Boxplot with two-sample mean test

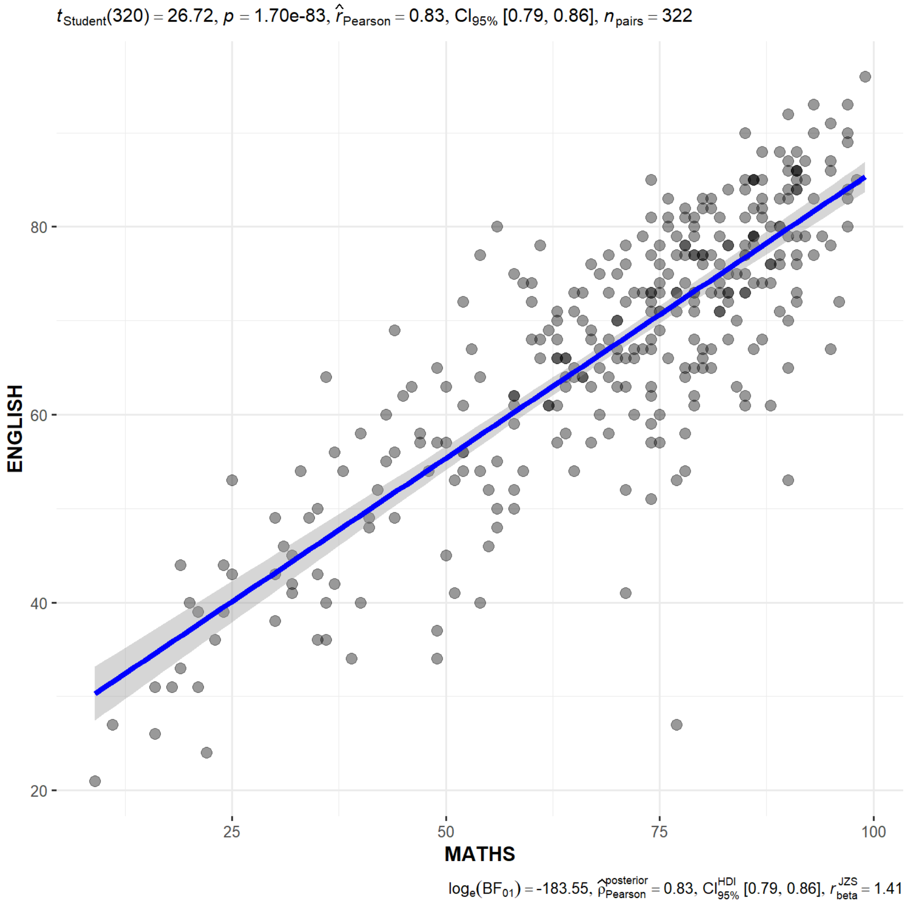


Visually-driven Correlation Analysis

Scatter plot showing the relationship between two continuous variables.

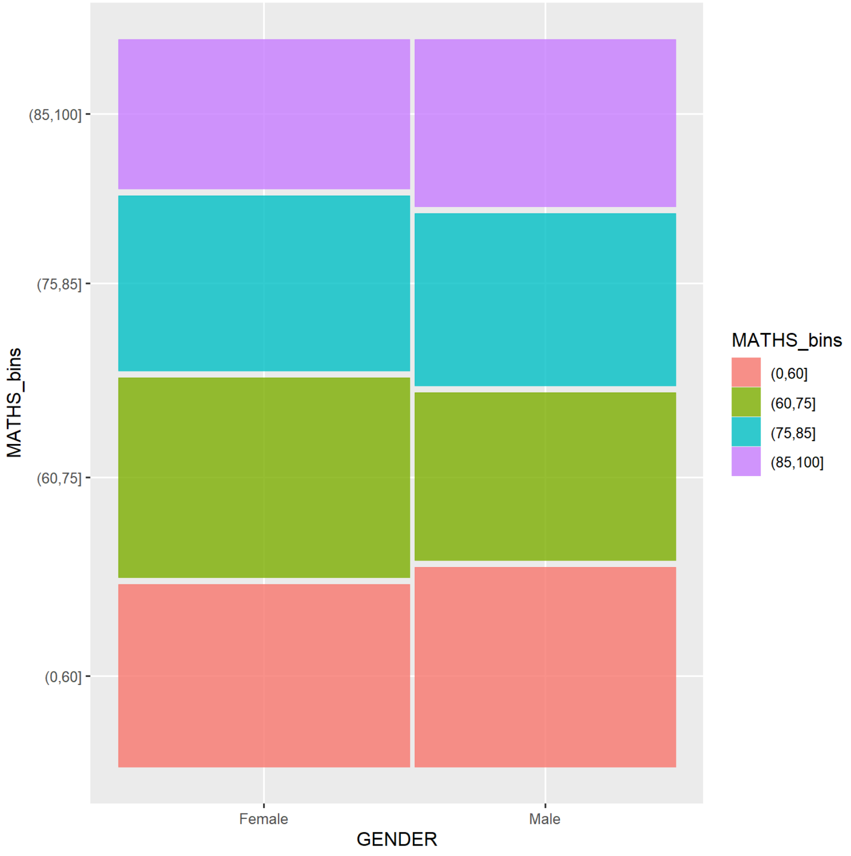


Scatter plot with significant test of correlation.

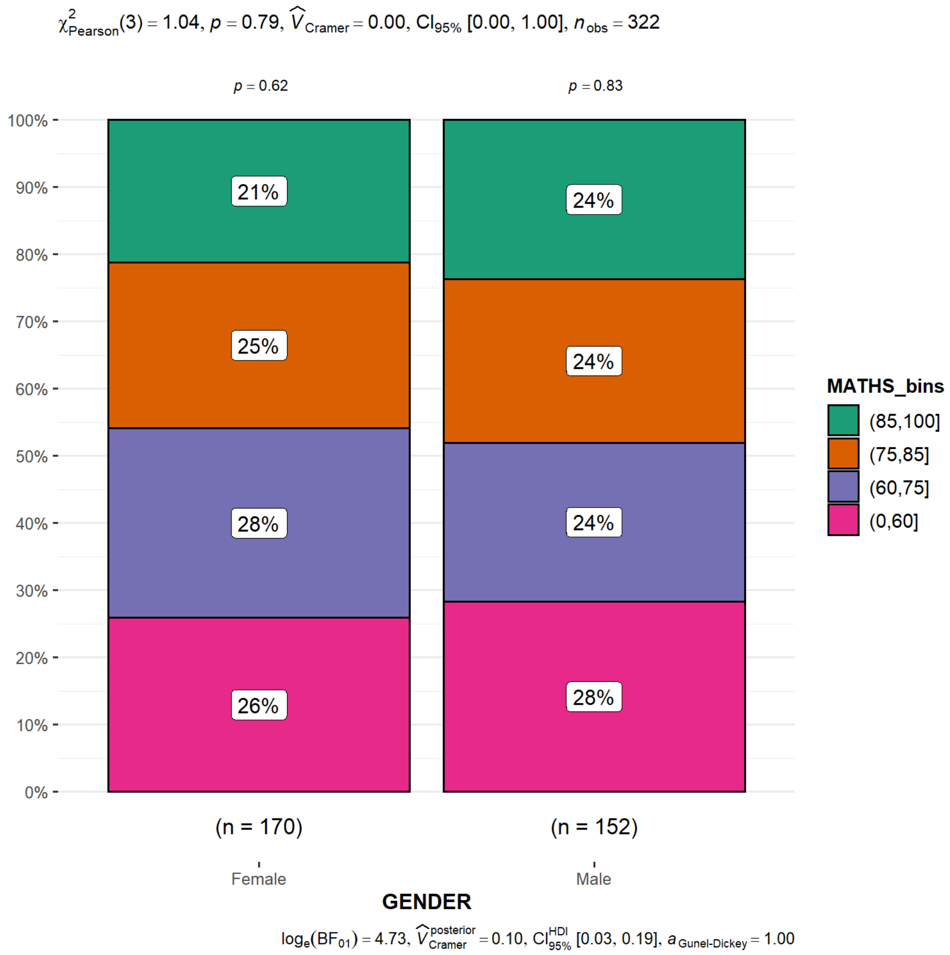


Visually-driven Association (Independent) Analysis

Mosaic plot showing the association between two categorical variables.



Stacked bar chart with significant test of association.



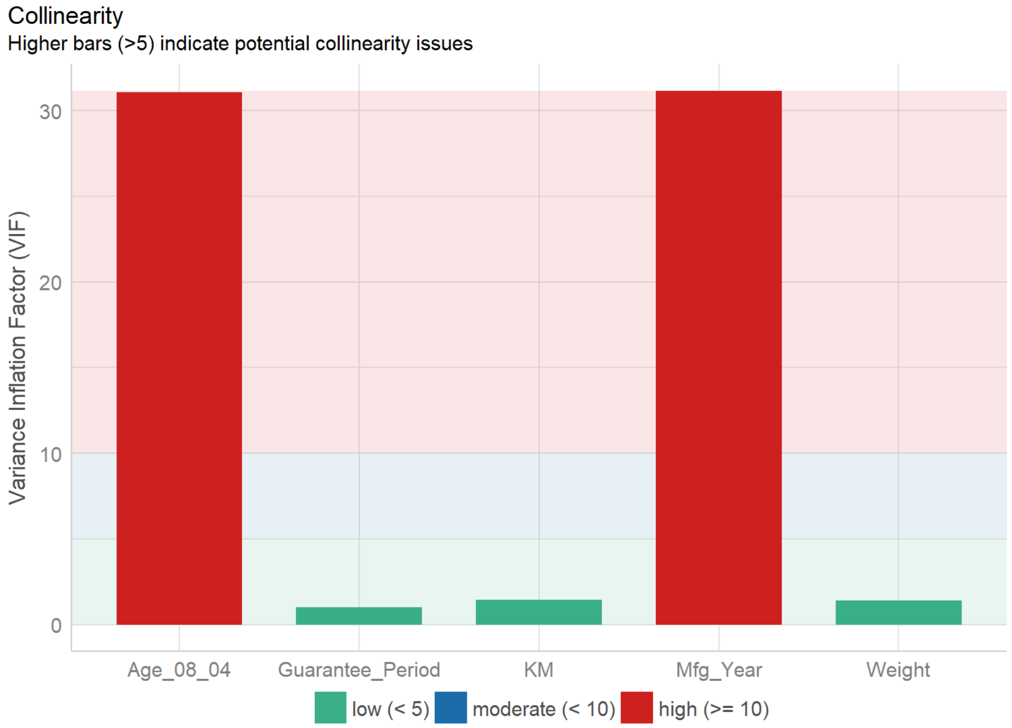
Visual Analytics Appraoch for Building Exploratory Models

Model Diagnostic: checking for multicolinearity:

Conventional statistical report

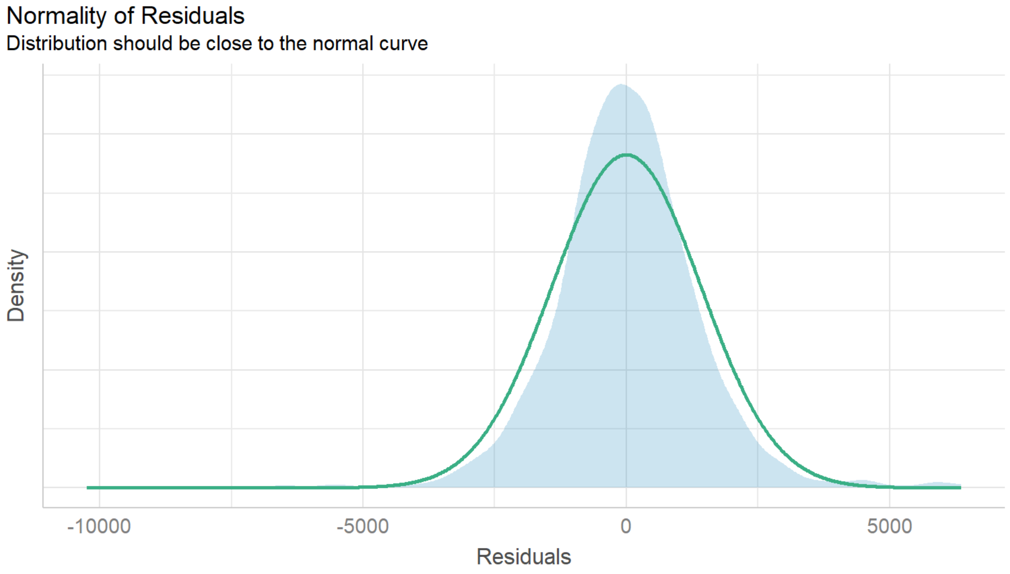
```
## # Check for Multicollinearity
##
## Low Correlation
##
##              Term  VIF Increased SE Tolerance
##                KM 1.46         1.21      0.68
##            Weight 1.41         1.19      0.71
##  Guarantee_Period 1.04         1.02      0.97
##
## High Correlation
##
##        Term   VIF Increased SE Tolerance
##  Age_08_04 31.07         5.57      0.03
##   Mfg_Year 31.16         5.58      0.03
```

Visual Analytics approach

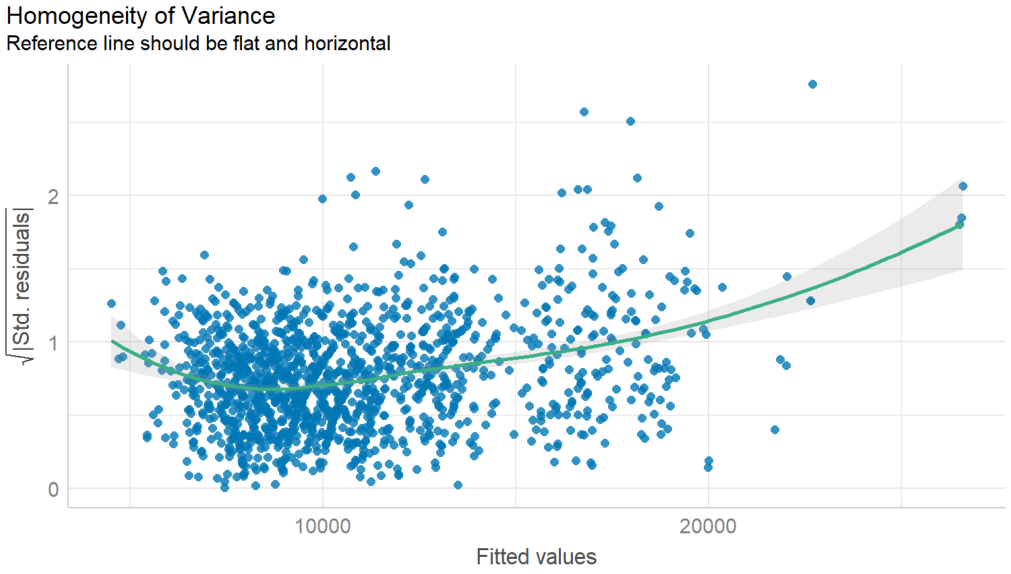


Visual Analytics Appraoch for Building Exploratory Models

Model Diagnostic: Checking normality assumption



Model Diagnostic: Checking model for homogeneity of variances



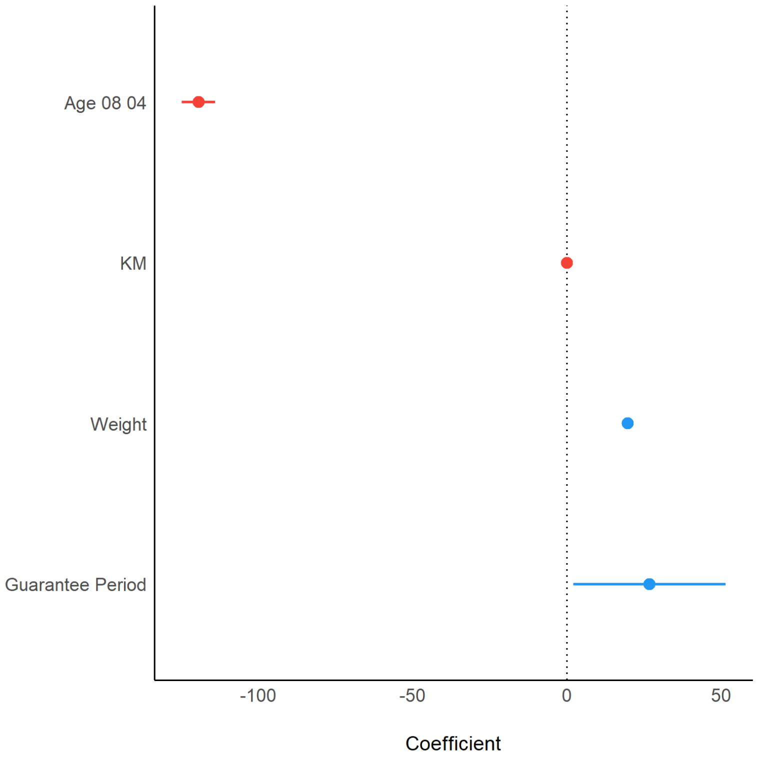
Visual Analytics Appraoch for Building Exploratory Models

Analysing model parameters

Conventional statistical report

```
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -2.186e+03  9.722e+02  -2.248   0.0247 *
## Age_08_04        -1.195e+02  2.760e+00 -43.292   <2e-16 ***
## KM               -2.406e-02  1.201e-03 -20.042   <2e-16 ***
## Weight            1.972e+01  8.379e-01  23.533   <2e-16 ***
## Guarantee_Period  2.682e+01  1.261e+01   2.126   0.0336 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

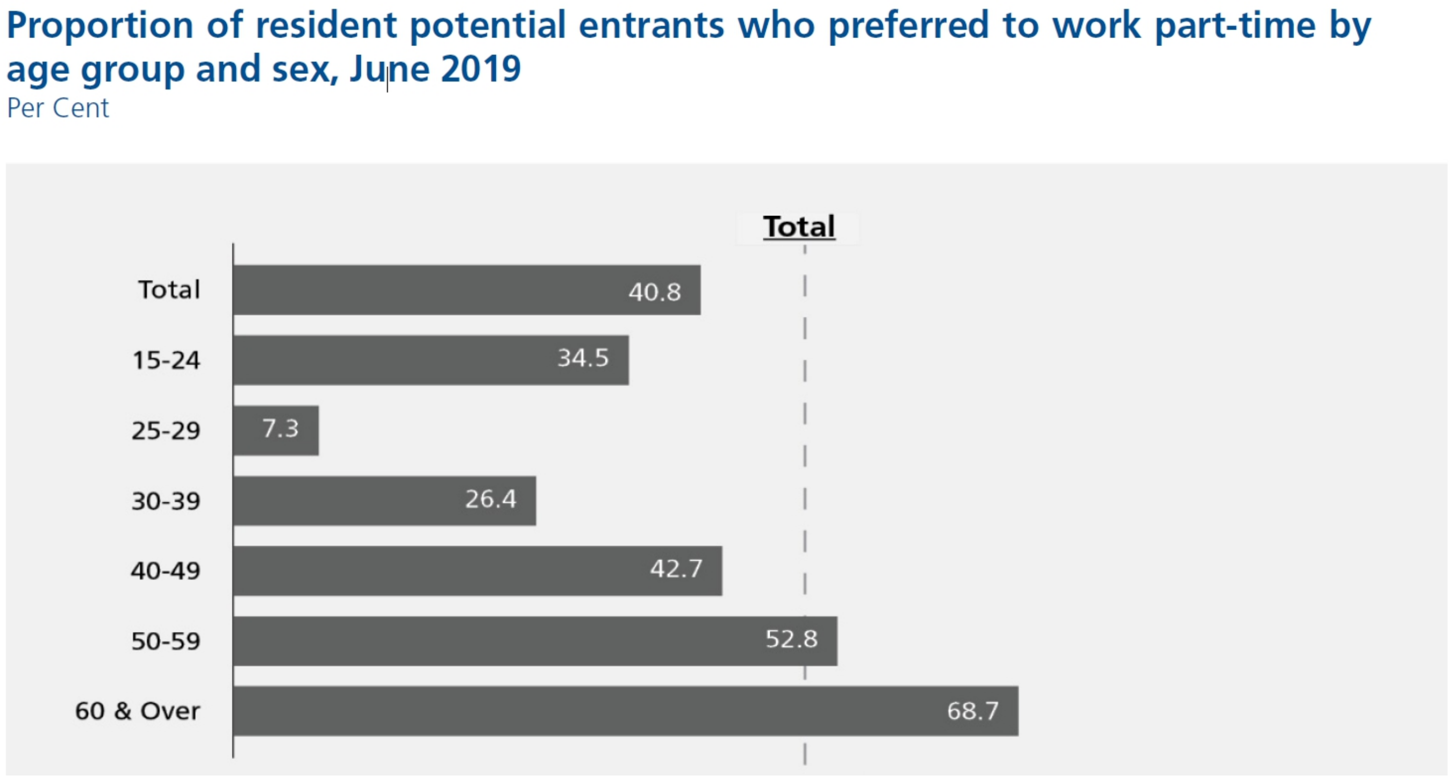
Visual Analytics approach



Visualising Uncertainty

Why it is important?

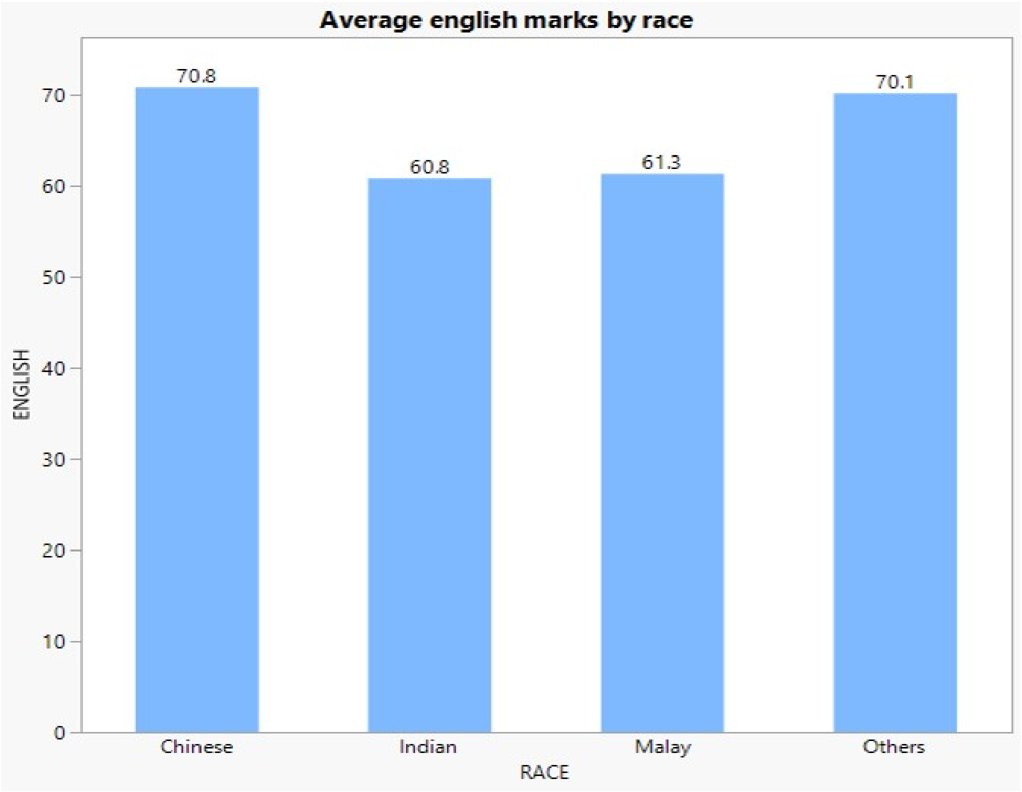
- One of the most challenging aspects of data visualization is the visualization of uncertainty.



Source: Chart 61, LABOUR FORCE IN SINGAPORE 2019, pg. 52

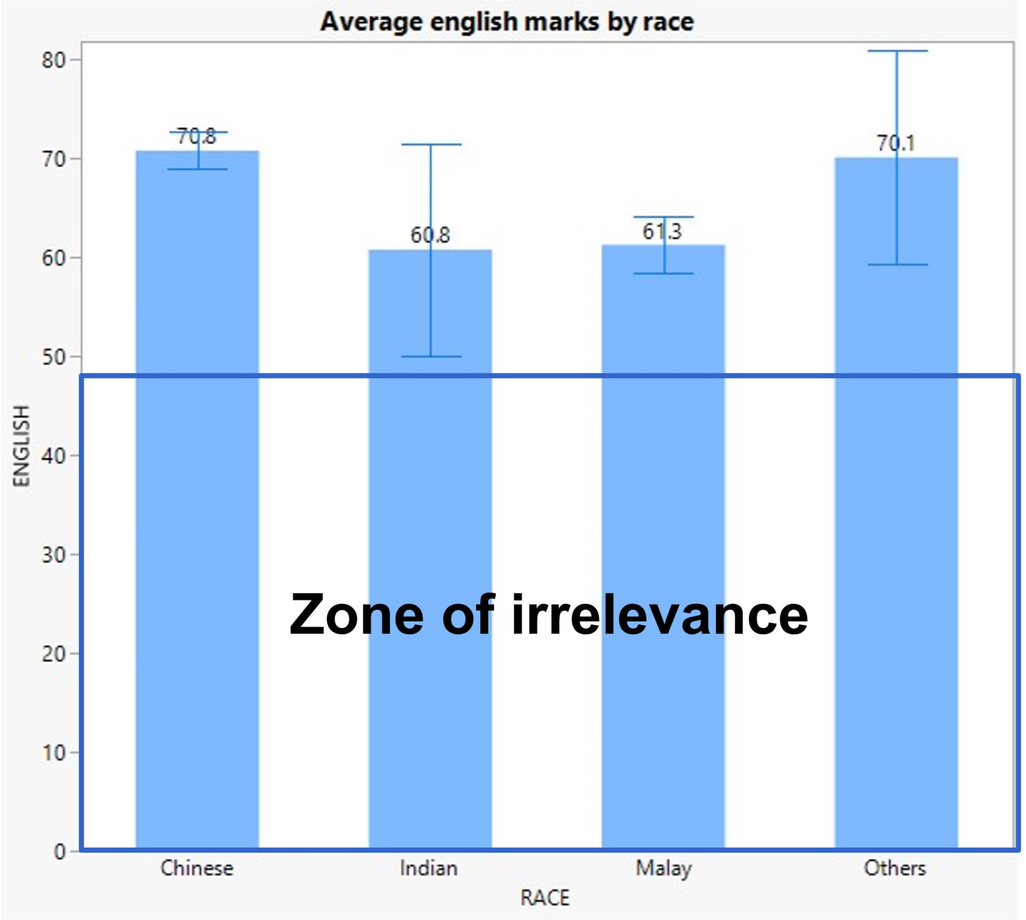
Why one shouldn't use a bar graph, even if the data are normally distributed?

- It is not appropriate to displace average values on bars



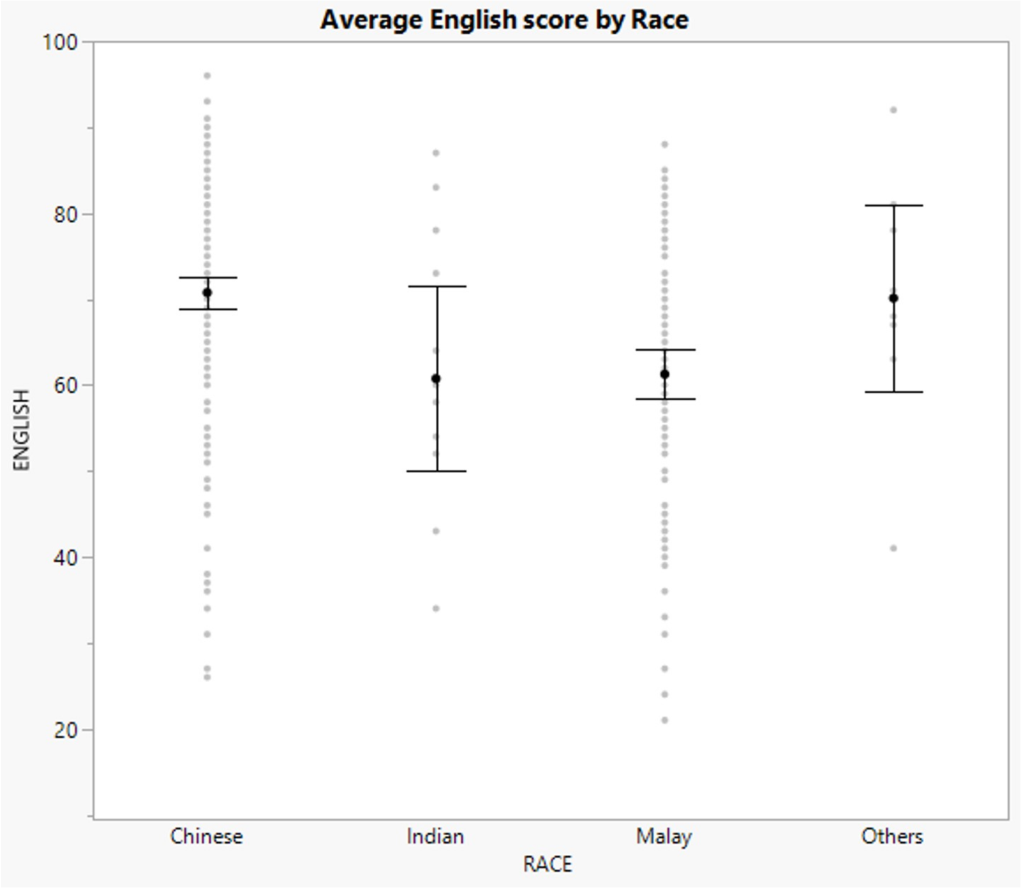
Why Error bar failed?

- Each error bar is constructed using a 95% confidence interval of the mean.

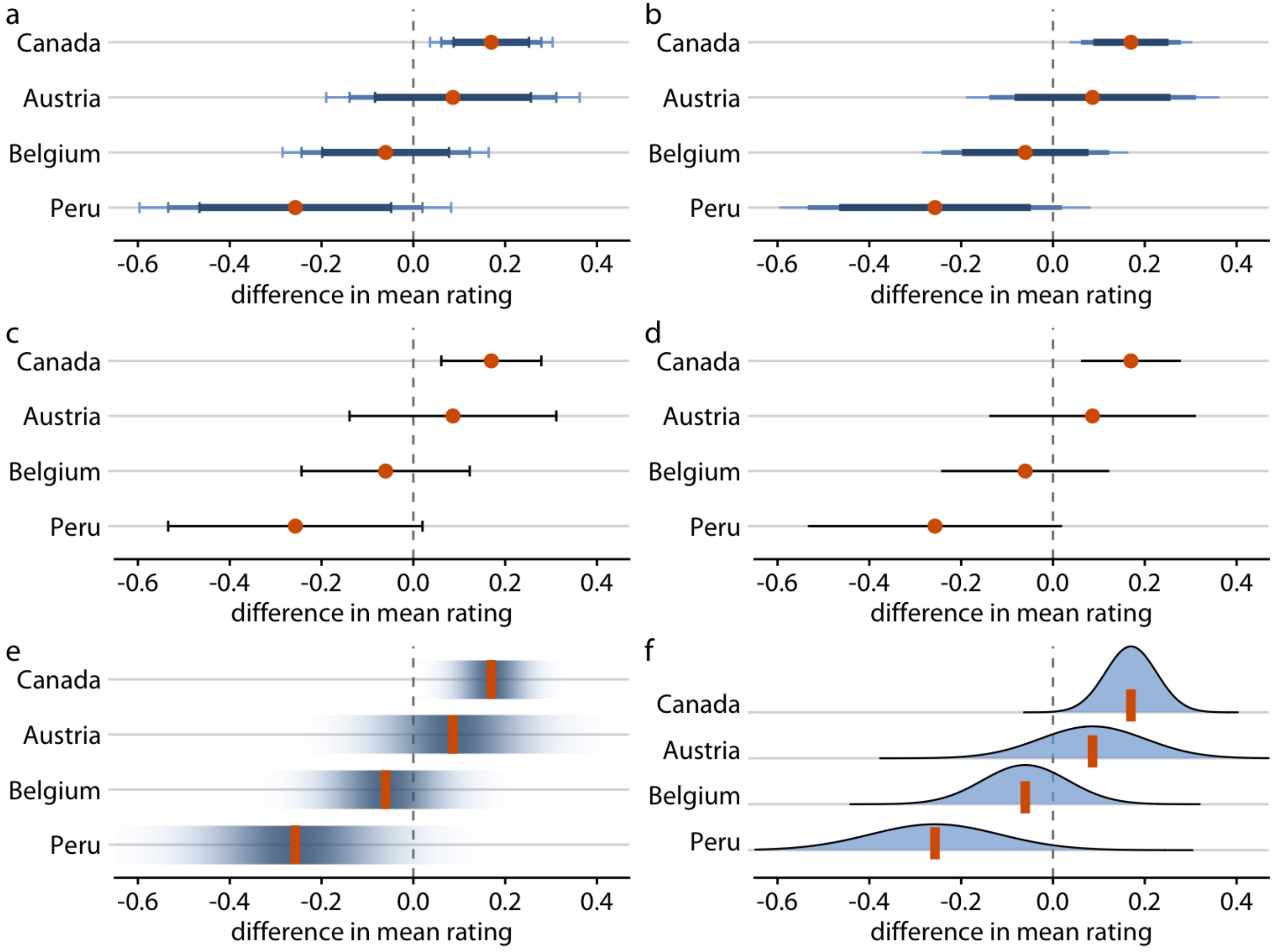


Error bar on a dot plot

- Each error bar is constructed using a 95% confidence interval of the percentage.



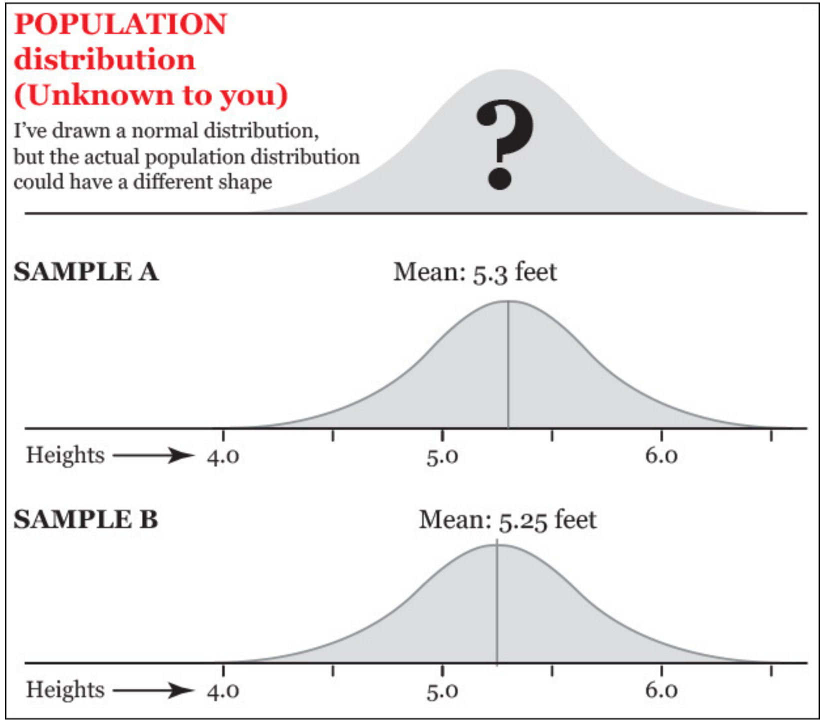
Graphical methods for visualising uncertainty

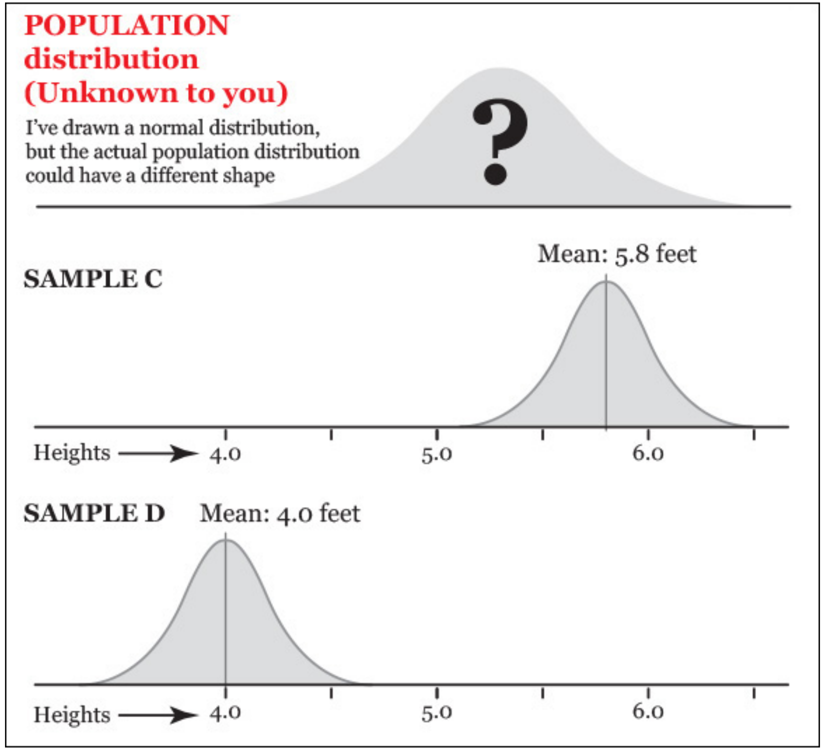


Reference: https://serialmentor.com/dataviz/visualizing-uncertainty.html#visualizing-the-uncertainty-of-point-estimates

Back to Statistics 101

Population and samples

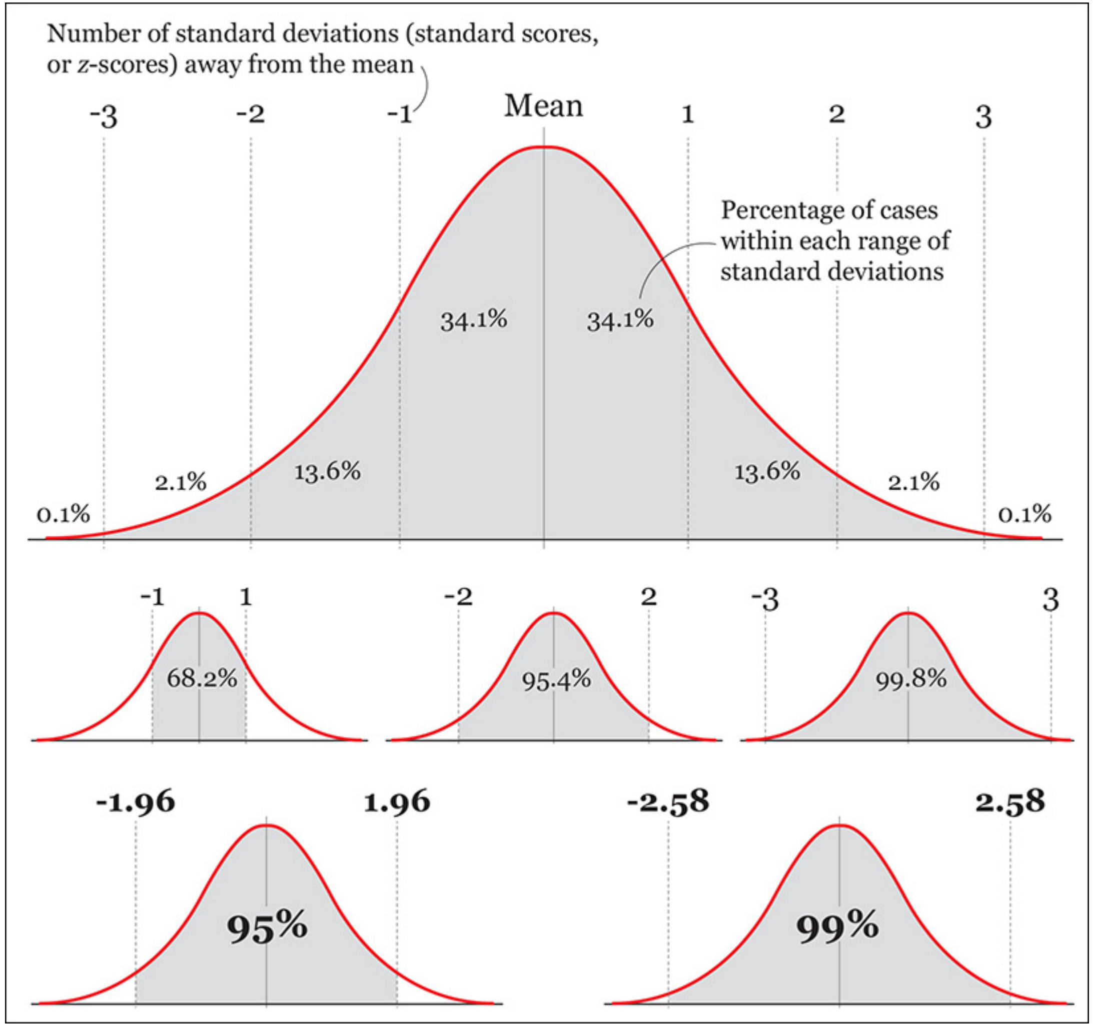




- When drawing many samples from a population, it is possible to obtain a few with means that greatly differ from the population.

Reference: Cairo, A. (2016) The Truthful Art, Chapter 11, New Riders

A reminder of the standard normal distribution



Reference: Cairo, A. (2016) The Truthful Art, Chapter 11, New Riders

The standard error

The formulas of standard deviation and standard error

$$\text{standard deviation } \sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

$$\text{variance} = \sigma^2$$

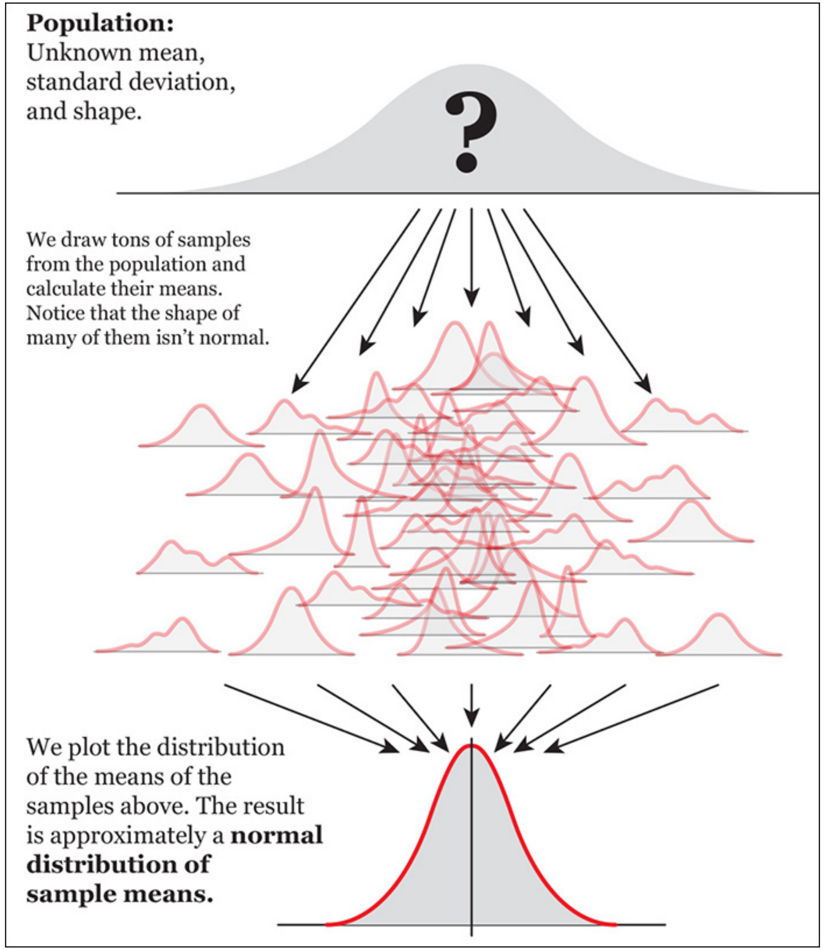
$$\text{standard error } (\sigma_{\bar{x}}) = \frac{\sigma}{\sqrt{n}}$$

where:

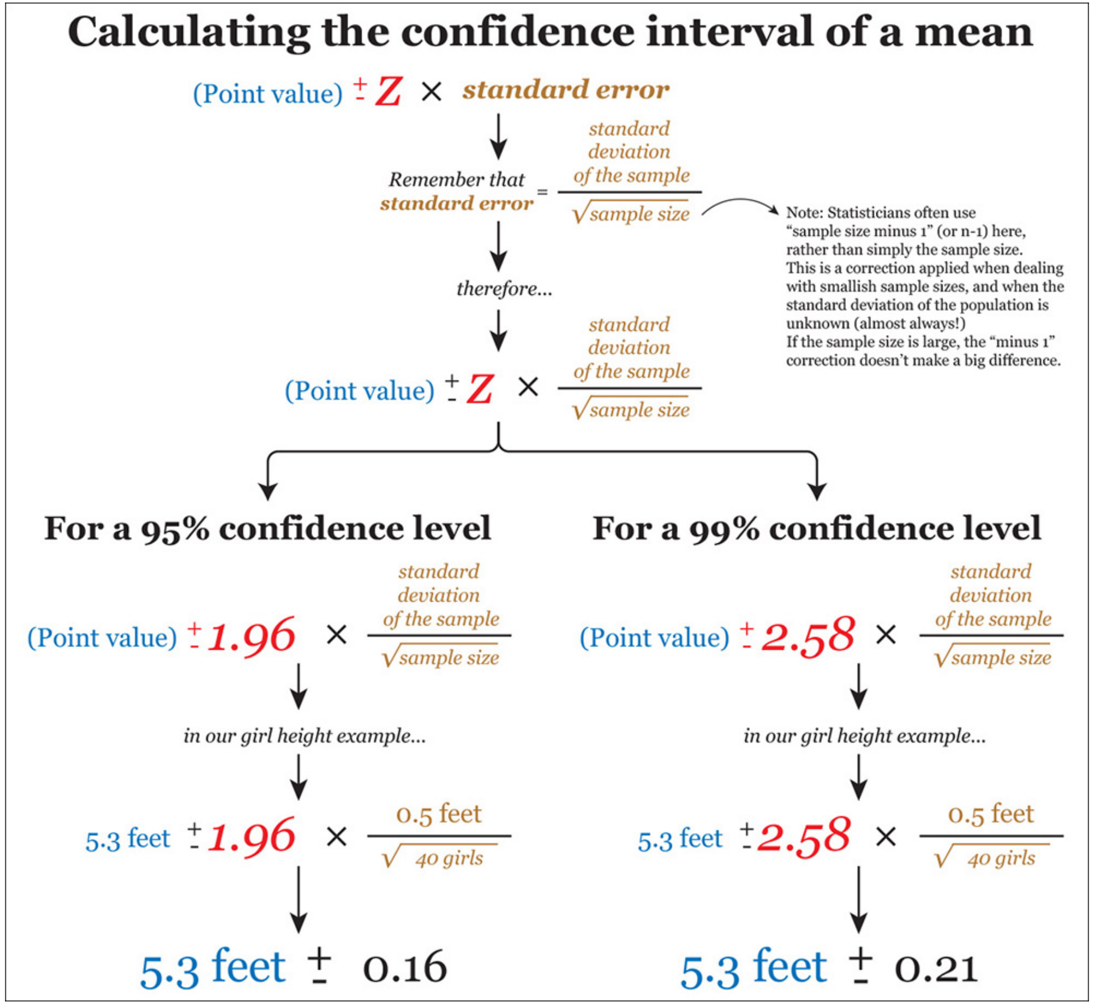
$\bar{x}$ = the sample's mean

n = the sample size

Reference: Cairo, A. (2016) The Truthful Art, Chapter 11, New Riders

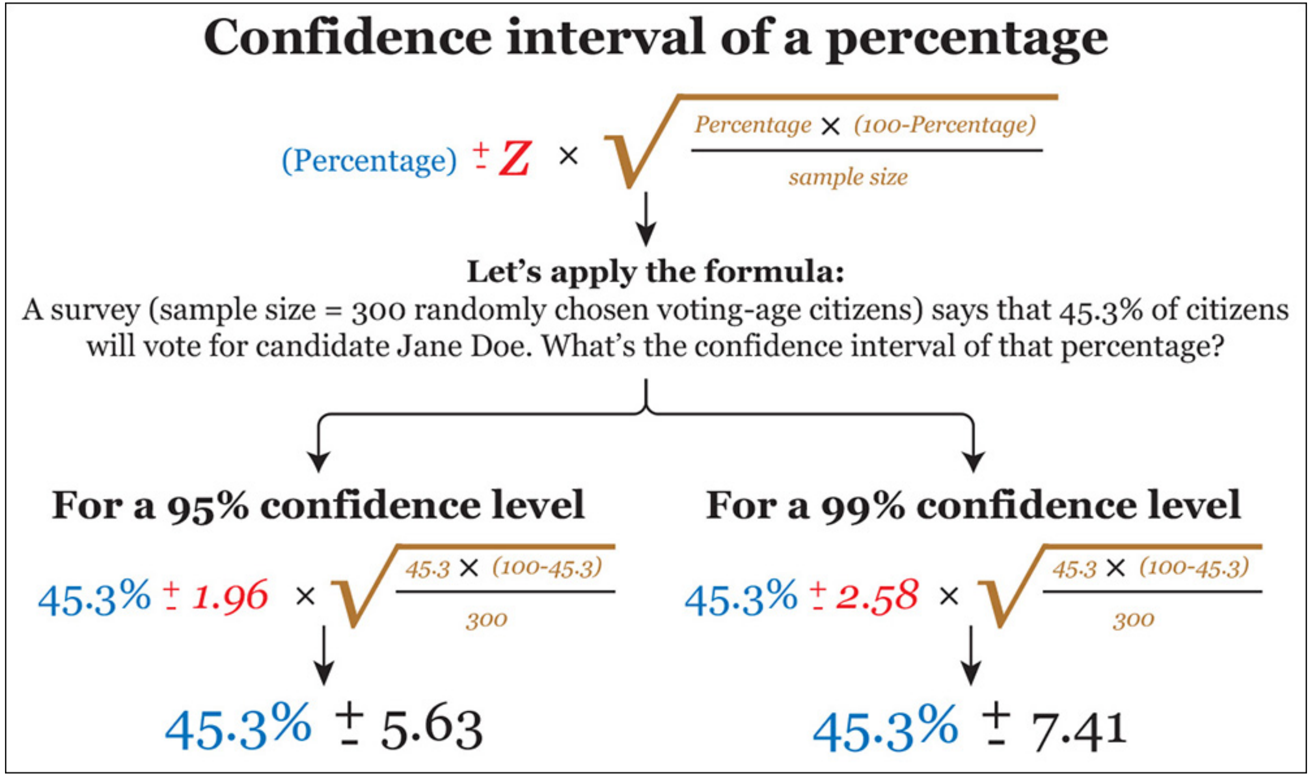


Calculating the confidence interval of a mean



Reference: Cairo, A. (2016) The Truthful Art, Chapter 11, New Riders

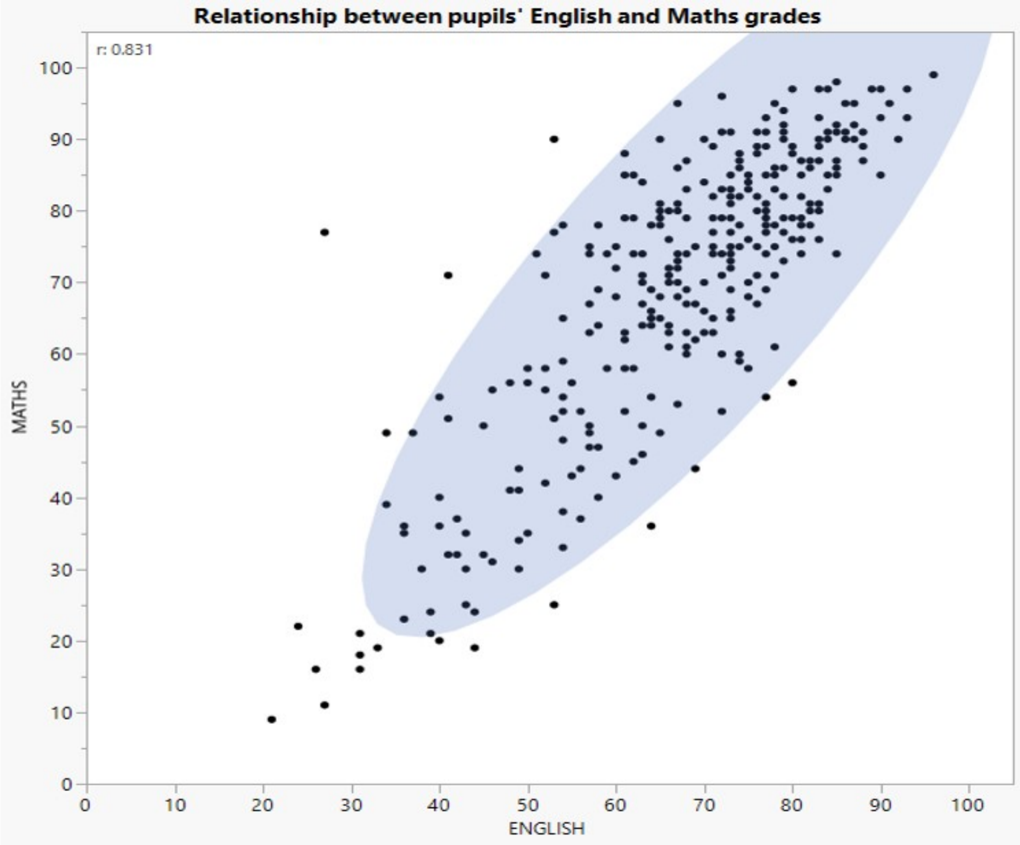
Calculating the confidence interval of a percentage



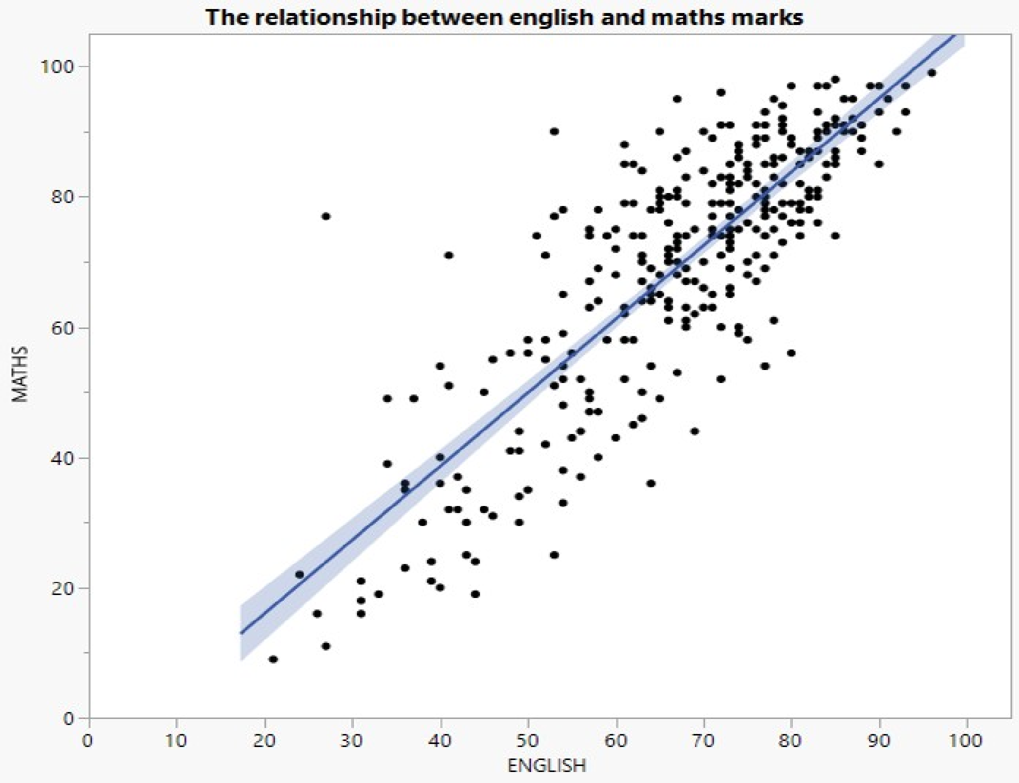
Reference: Cairo, A. (2016) The Truthful Art, Chapter 11, New Riders

2-d graphical methods for visualising uncertainty

Scatter plot with 95% confidence ellipse

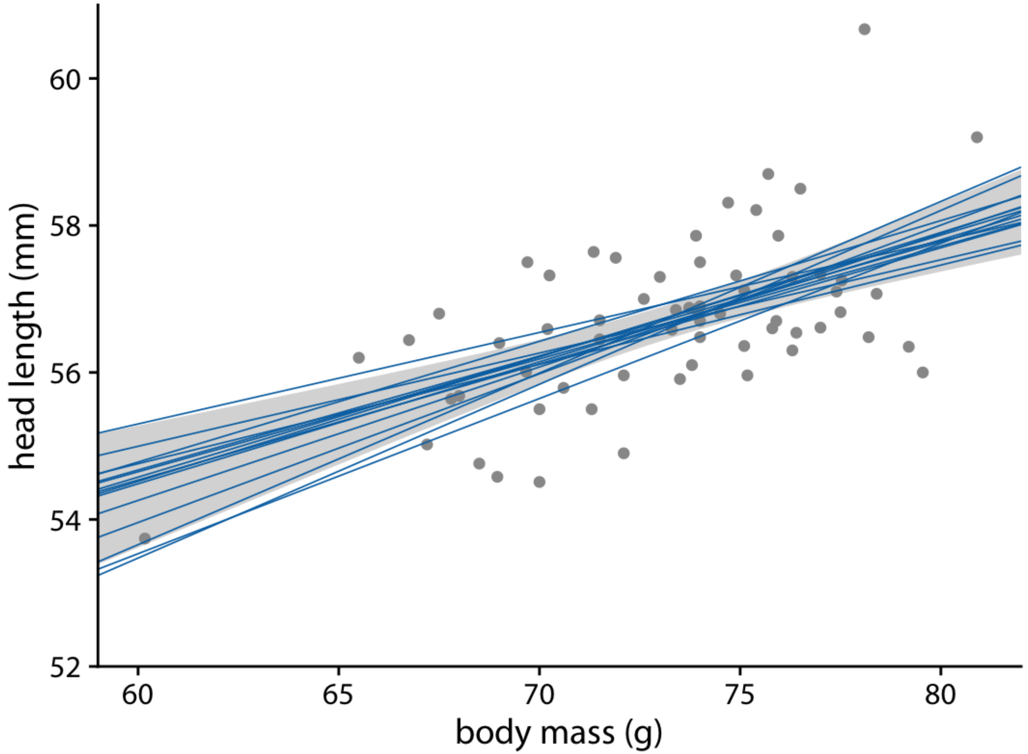


Scatter plot with 95% confidence intervals

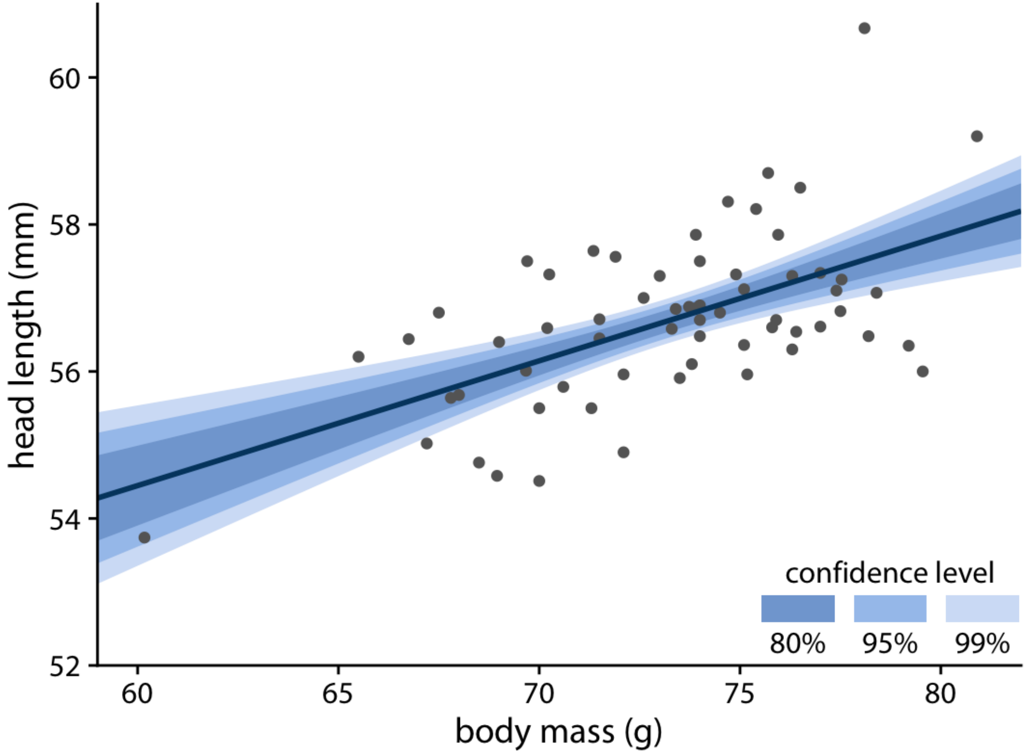


Confidence band of a trend line

Confidence band and fit lines

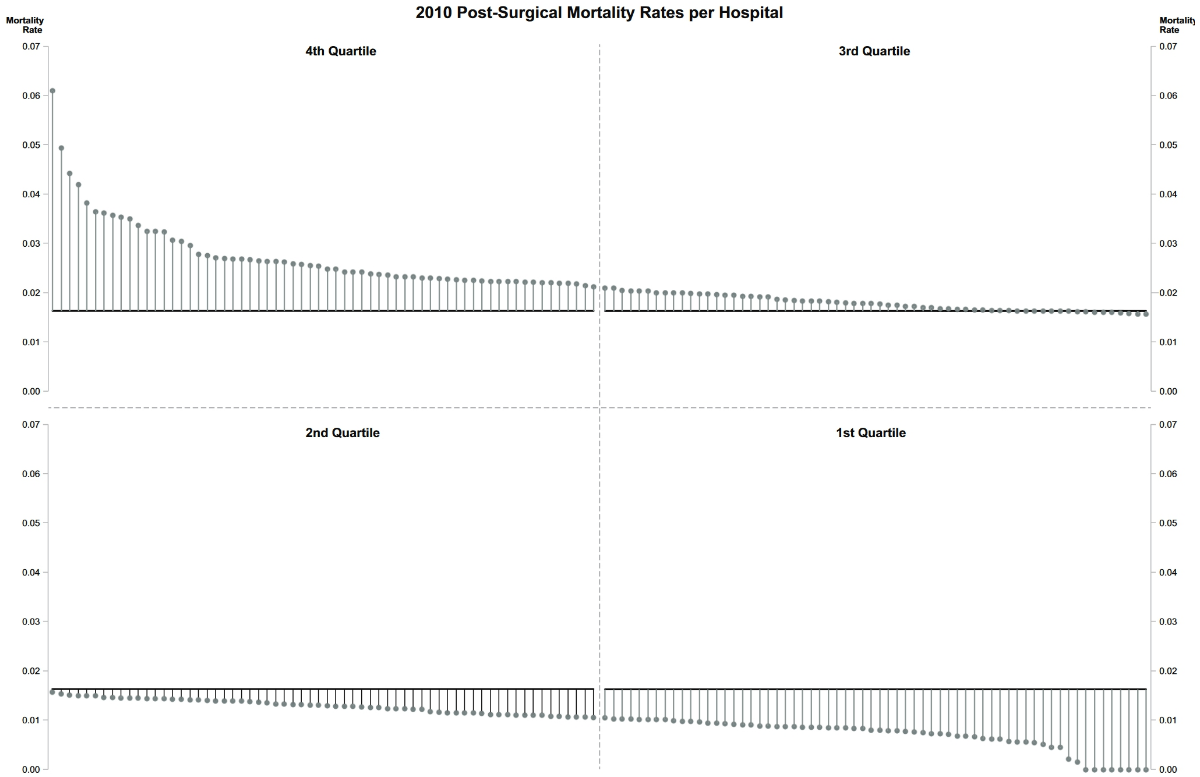


A graded confidence band



Variation and Its Discontents

Random and unfair comparisons



Reference: https://www.perceptualedge.com/articles/visual_business_intelligence/variation_and_its_discontents.pdf

Funnel plots to the rescue

Statistical details

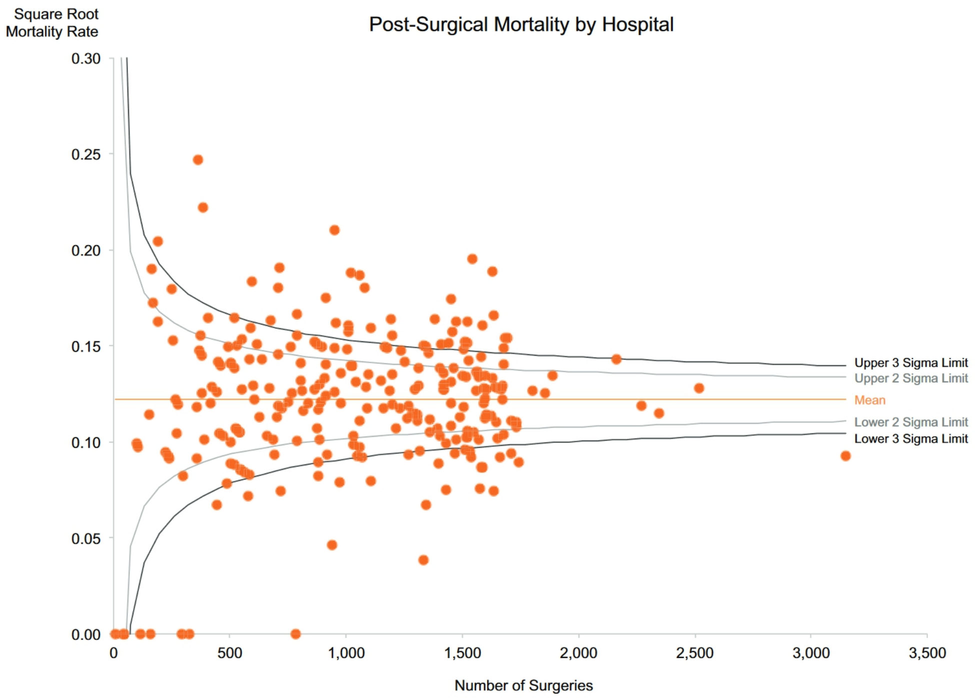
Calculations:

Fit Mean = Sum of Occurrences / Sum of Sample Sizes

$$\text{Two Sigmas (95\% Limit)} = \text{Fit Mean} \pm 1.96 * \sqrt{\frac{\text{Fit Mean} * (1 - \text{Fit Mean})}{x}}$$

$$\text{Three Sigmas (99.8\% Limit)} = \text{Fit Mean} \pm 3.0 * \sqrt{\frac{\text{Fit Mean} * (1 - \text{Fit Mean})}{x}}$$

$$\text{Histogram Values} = \frac{\left(\frac{\text{Occurrences}}{\text{Sample Sizes}}\right) - \text{Fit Mean}}{\sqrt{\frac{\text{Fit Mean} \cdot (1 - \text{Fit Mean})}{\text{Occurrences}}}}$$



Normality assumption

- Before plotting the graph, it is important to check if the values are conformed to normal distribution assumption.

- If the raw values are not conformed to normality assumption, they have to be transformed.