

# **Lesson 4:**

# **Fundamentals of Visual Analytics**

Dr. Kam Tin Seong  
Assoc. Professor of Information Systems (Practice)  
School of Computing and Information Systems,  
Singapore Management University

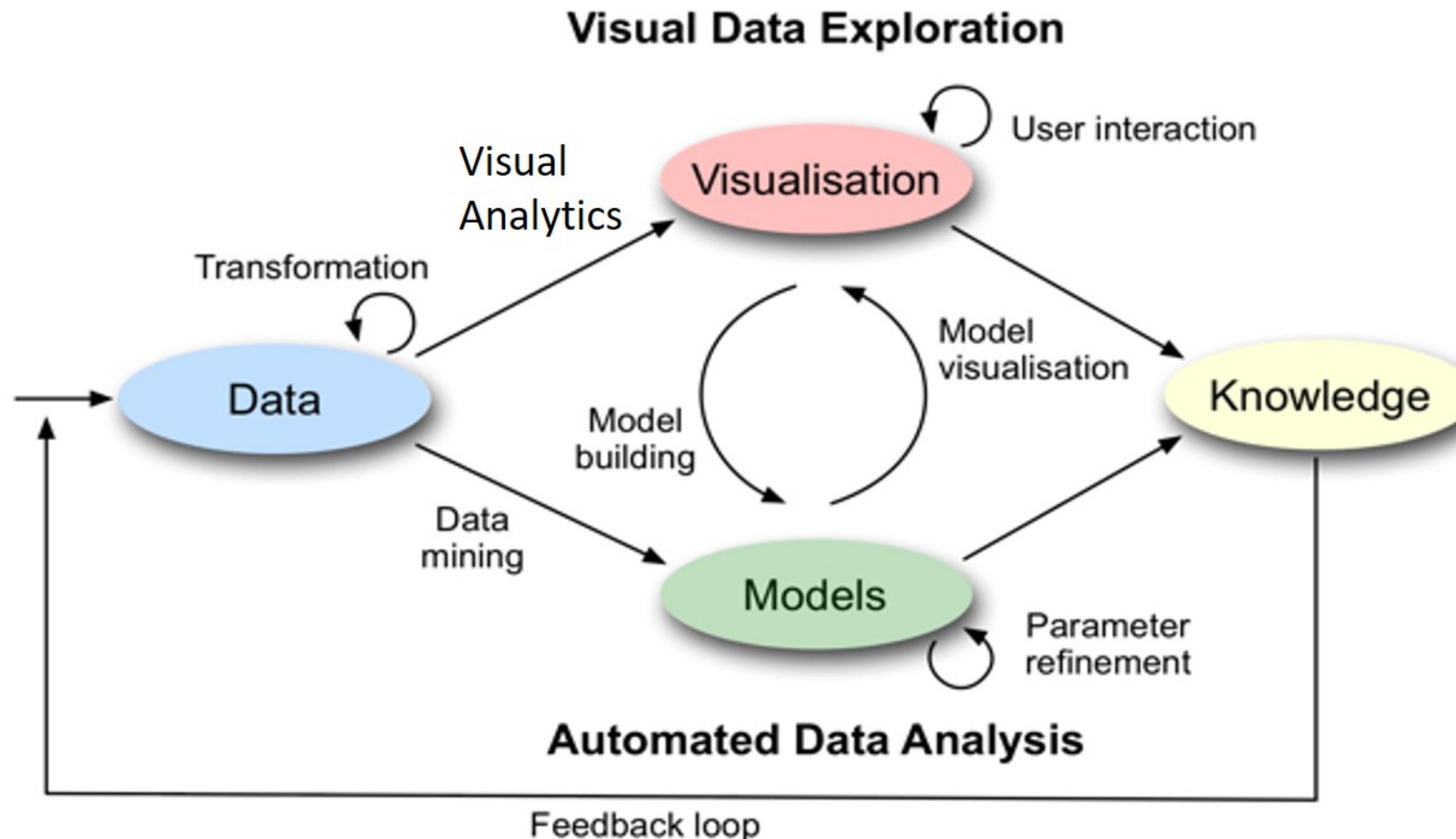
23 Jan 2023

# 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.



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<sup>†</sup> 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.



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<sup>†</sup>

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*<sup>\*</sup>-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:

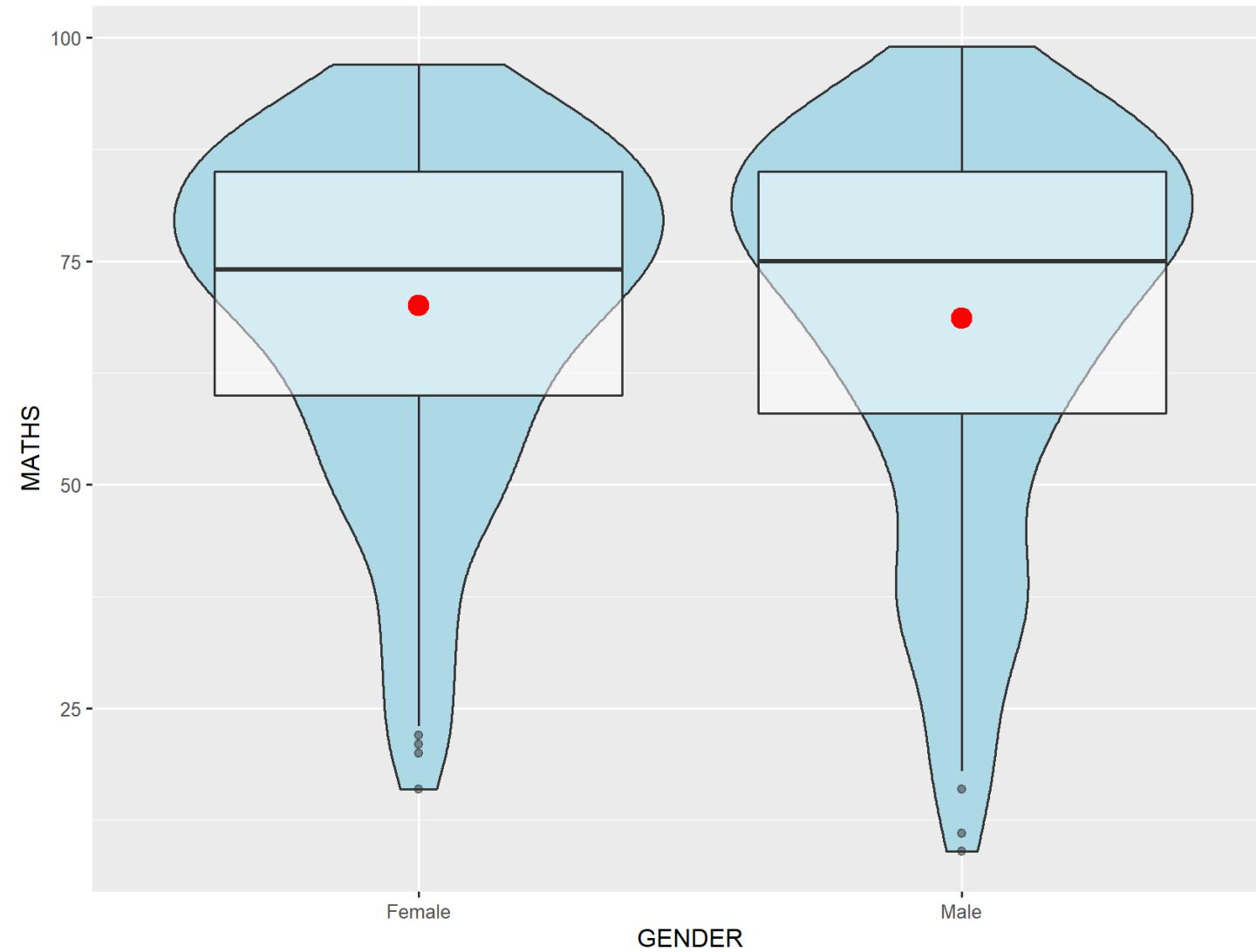
parameter      statistic      significance      effect size + confidence intervals      number of observations

The diagram illustrates the components of a statistical report. On the left, four labels are listed vertically: 'parameter', 'statistic', 'significance', and 'number of observations'. Arrows point from each of these labels to specific parts of the following text. The 'parameter' arrow points to the sample size '14.79'. The 'statistic' arrow points to the t-value '3.36'. The 'significance' arrow points to the p-value '0.004'. The 'number of observations' arrow points to the value '32'. In the center, the text displays the results:  $t(14.79) = 3.36, p = 0.004, \xi = 0.77, \text{CI}_{95\%}[0.47, 0.90], n = 32$ . The labels 'effect size + confidence intervals' are positioned above the text '0.77, CI<sub>95%</sub>[0.47, 0.90]'.

$t(14.79) = 3.36, p = 0.004, \xi = 0.77, \text{CI}_{95\%}[0.47, 0.90], n = 32$

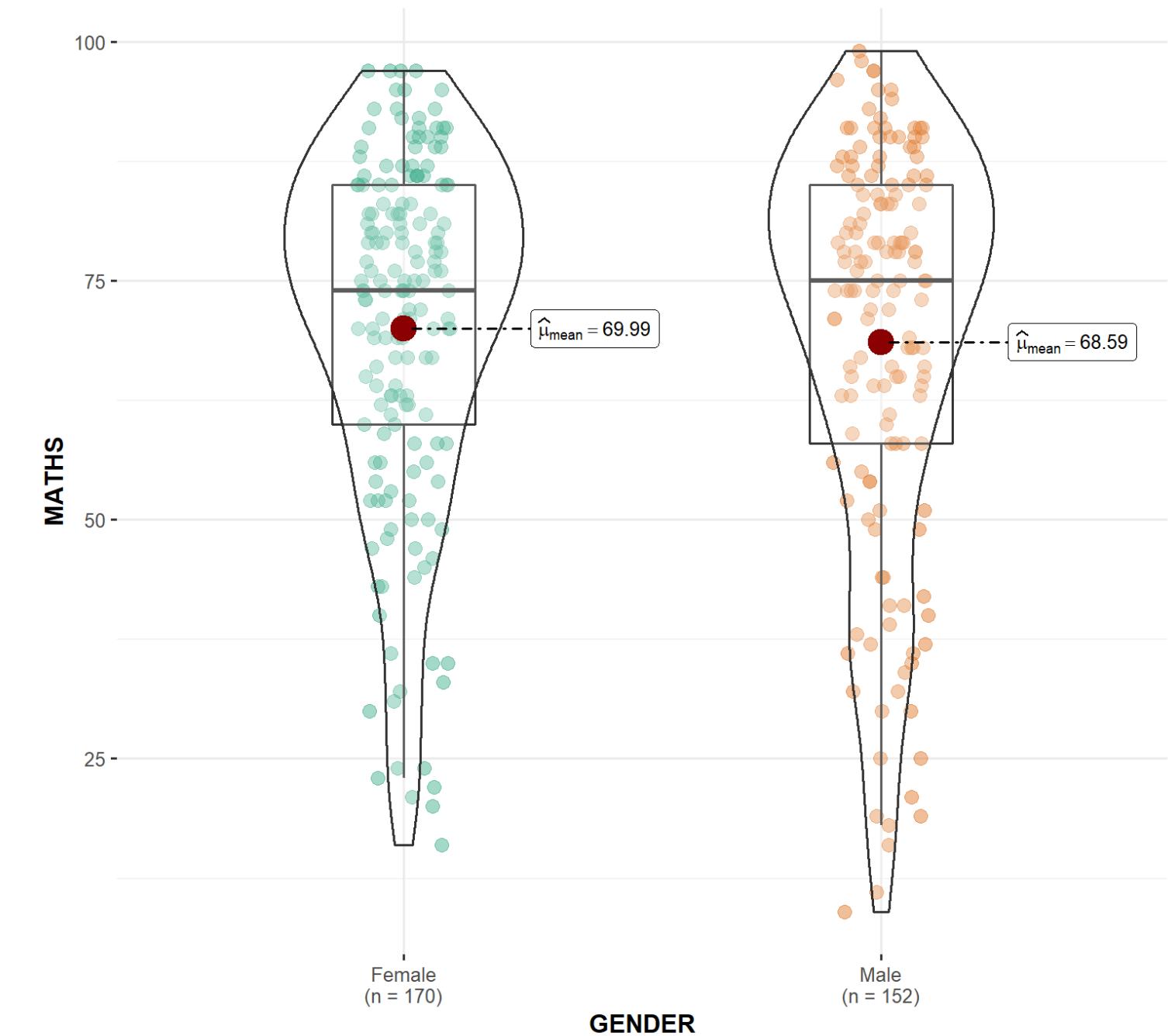
# Two-sample means

Boxplot revealing the mean and distribution of two samples.



Boxplot with two-sample mean test

$$t_{\text{Welch}}(301.85) = 0.62, p = 0.53, \hat{g}_{\text{Hedges}} = 0.07, \text{CI}_{95\%} [-0.15, 0.29], n_{\text{obs}} = 322$$

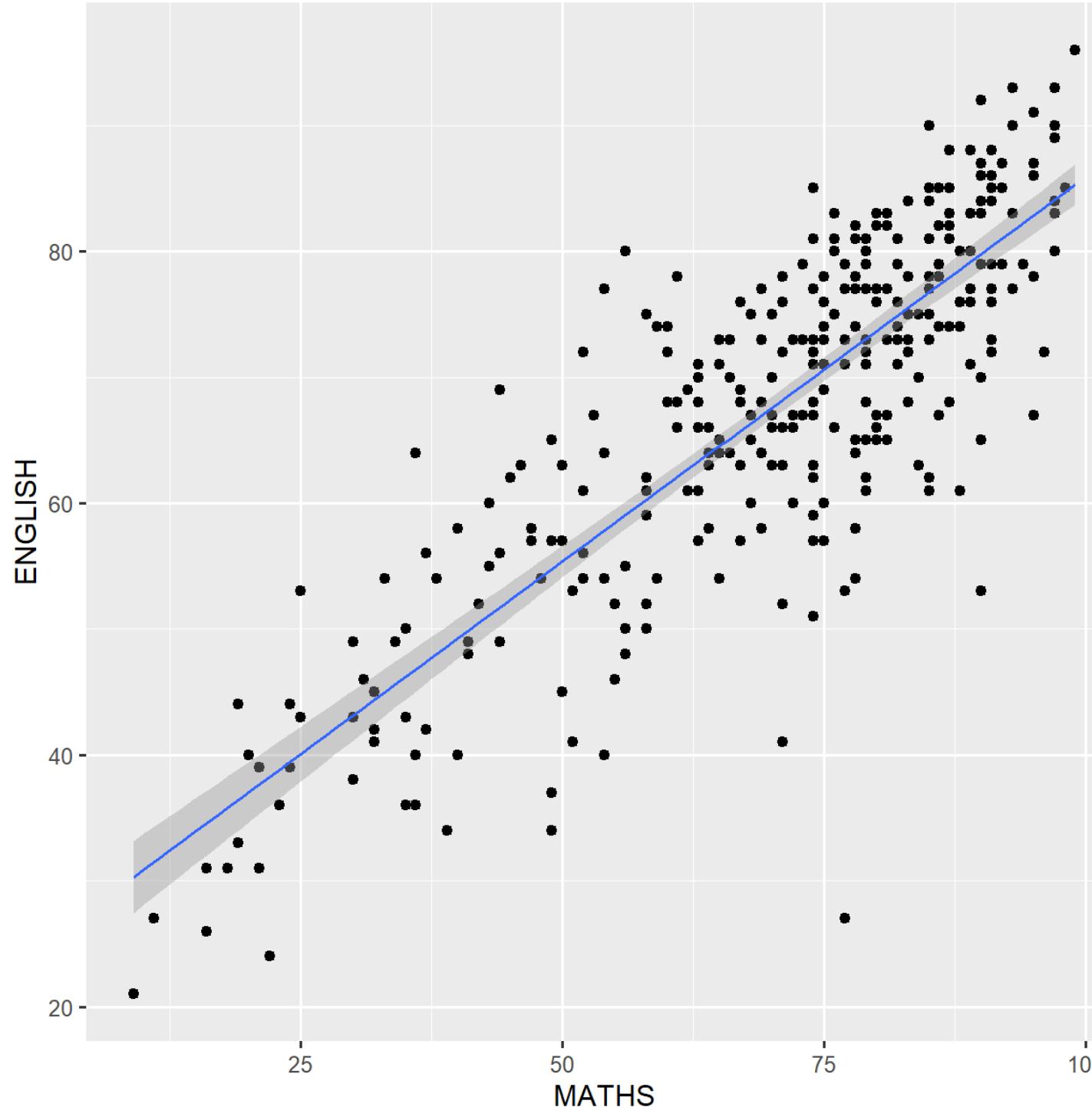


$$\log_e(BF_{01}) = 1.91, \hat{\delta}_{\text{difference}}^{\text{posterior}} = 1.39, \text{CI}_{95\%}^{\text{HDI}} [-3.10, 5.34], r_{\text{Cauchy}}^{\text{JZS}} = 0.71$$



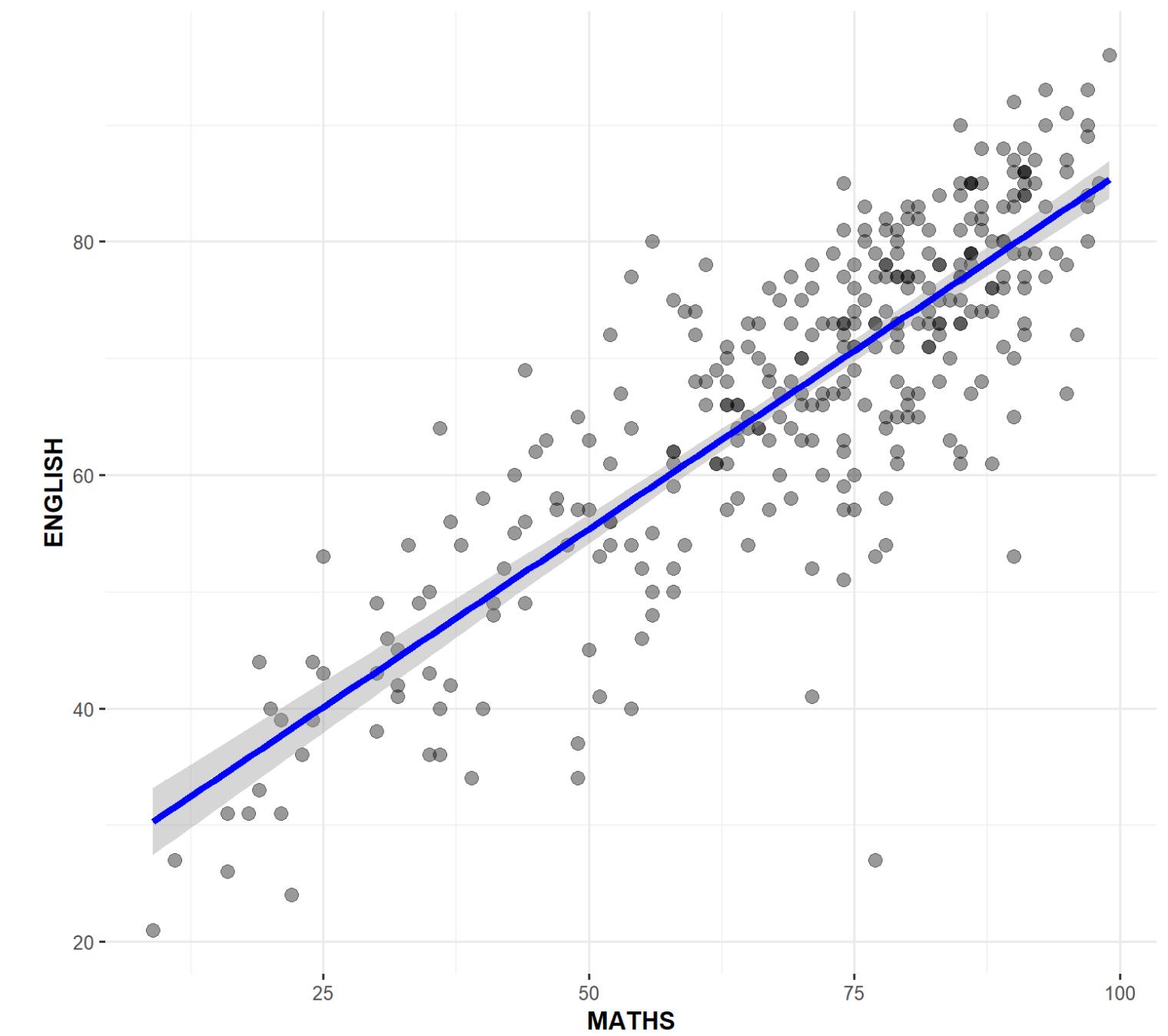
# Visually-driven Correlation Analysis

Scatter plot showing the relationship between two continuous variables.



Scatter plot with significant test of correlation.

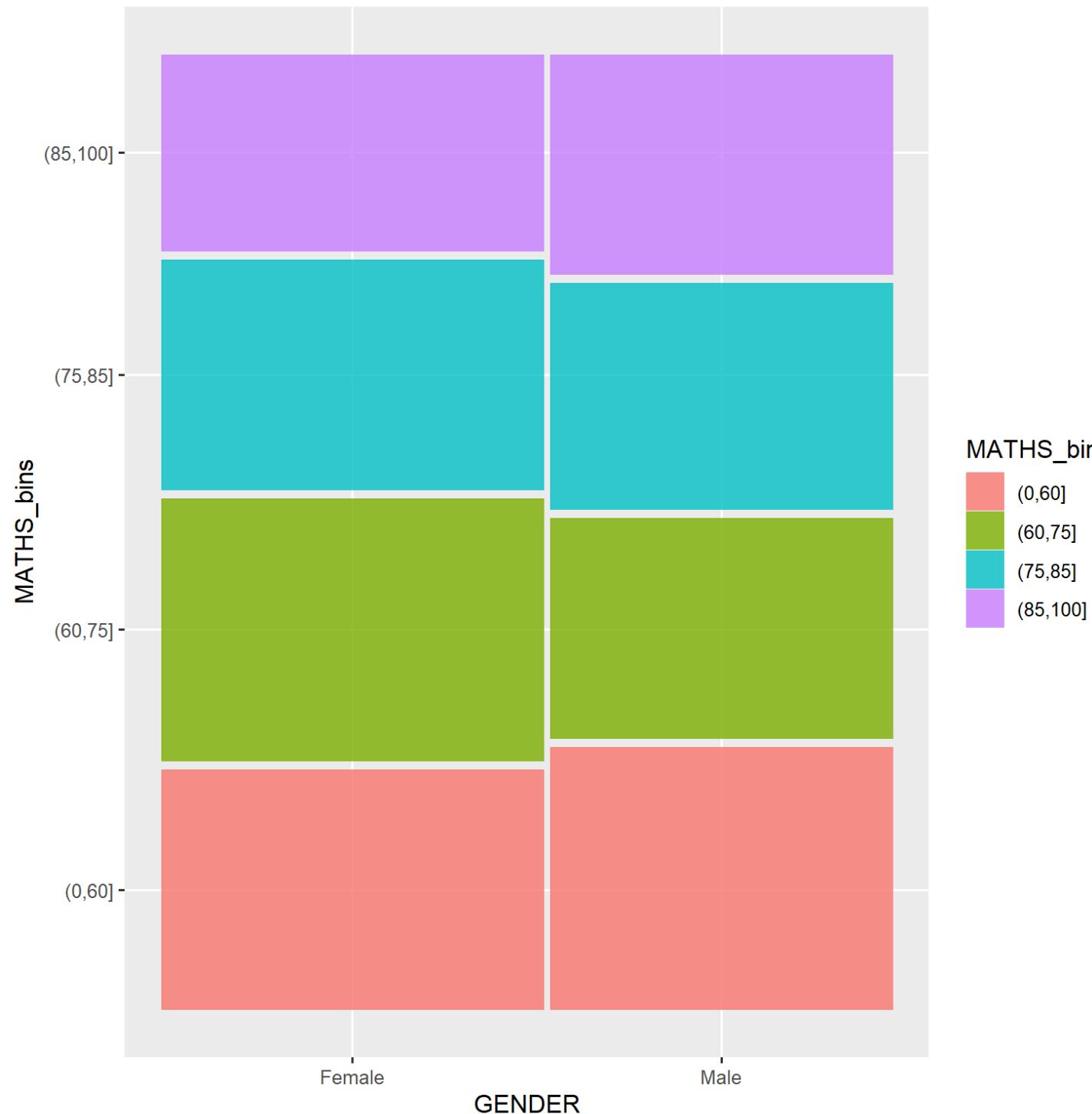
$t_{\text{Student}}(320) = 26.72, p = 1.70e-83, \hat{r}_{\text{Pearson}} = 0.83, \text{CI}_{95\%} [0.79, 0.86], n_{\text{pairs}} = 322$



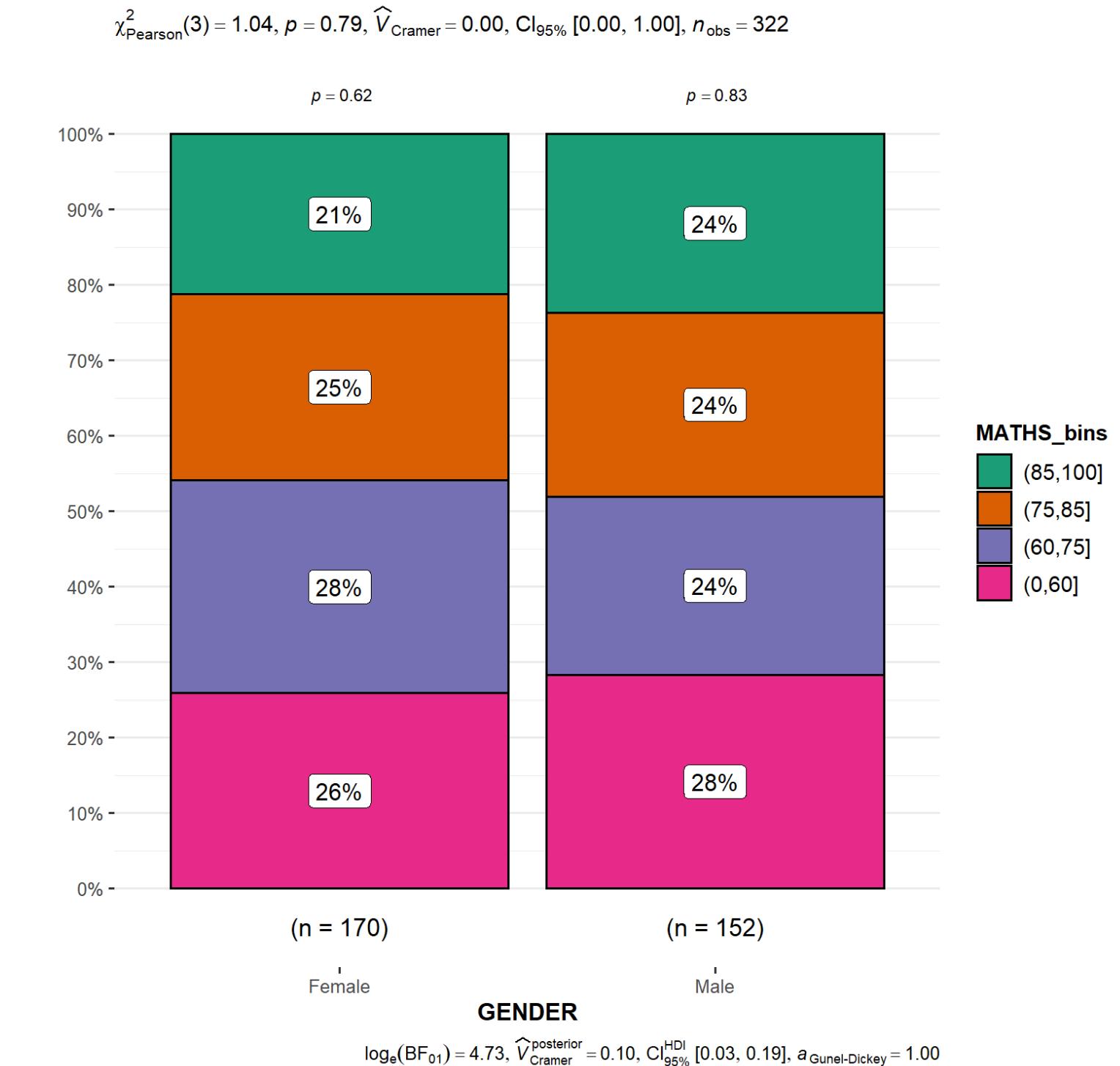
$\log_e(BF_{01}) = -183.55, \hat{\rho}_{\text{Pearson}}^{\text{posterior}} = 0.83, \text{CI}_{95\%}^{\text{HDI}} [0.79, 0.86], r_{\text{beta}}^{\text{JZS}} = 1.41$

# Visually-driven Association (Independent) Analysis

Mosaic plot showing the association between two categorical variables.



Stacked bar chart with significant test of association.



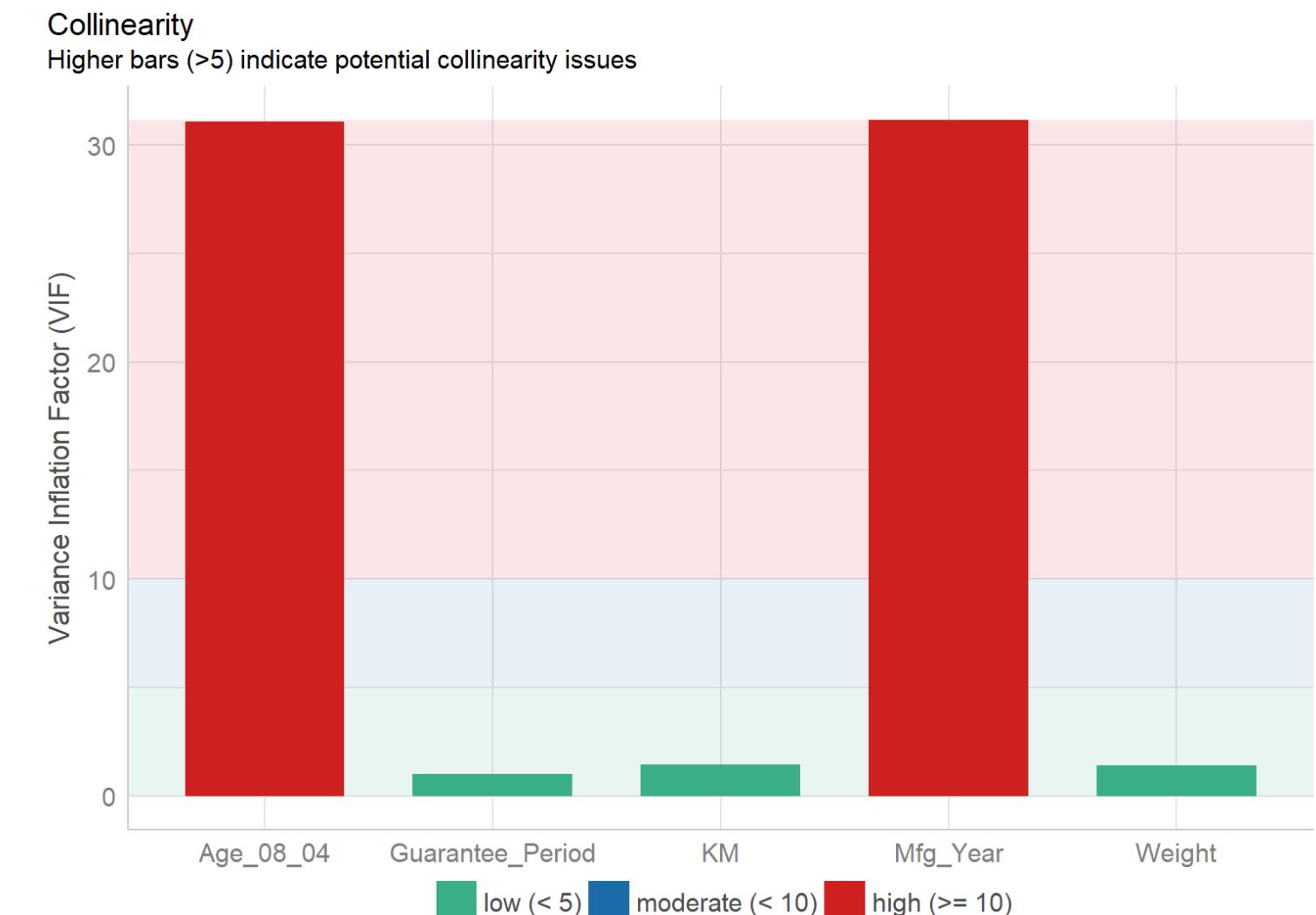
# Visual Analytics Approach for Building Exploratory Models

## Model Diagnostic: checking for multicollinearity:

### 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
```

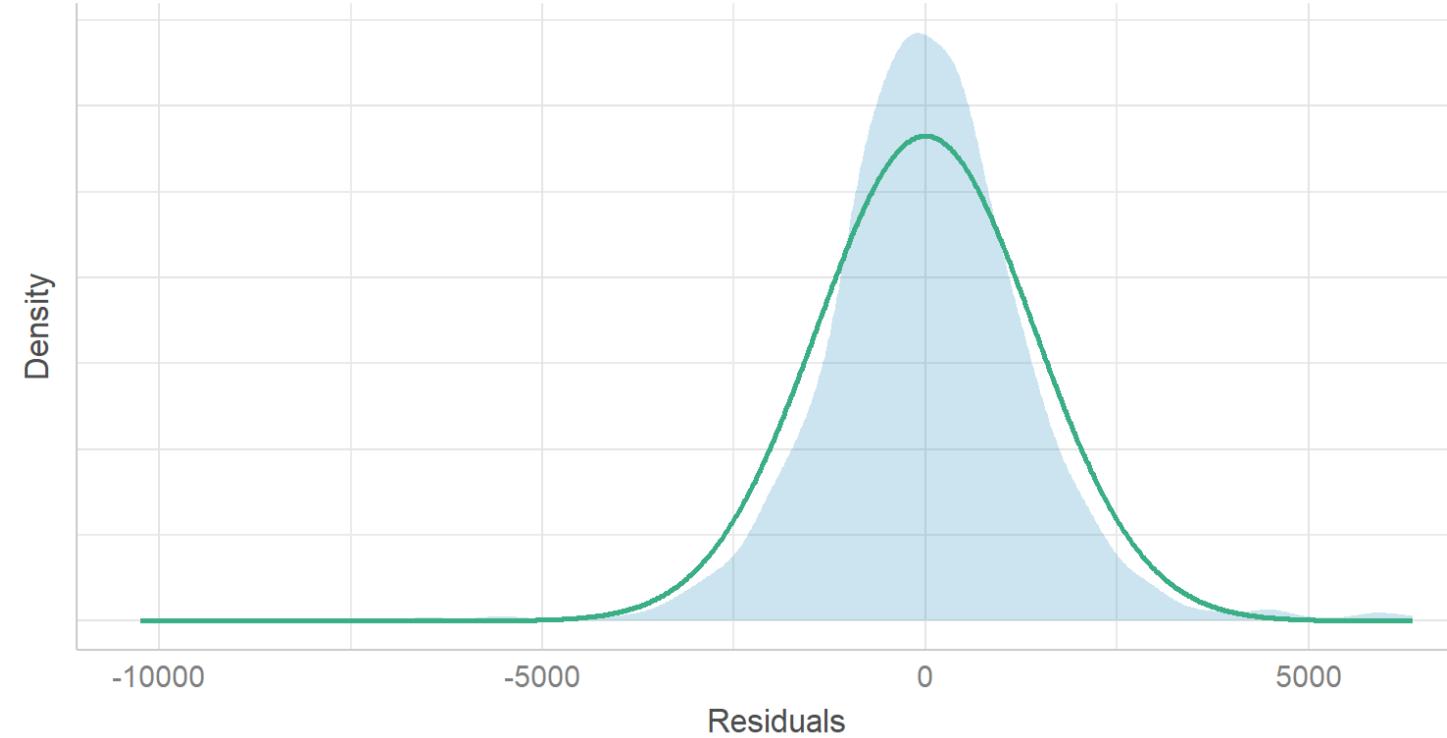
### Visual Analytics approach



# Visual Analytics Approach for Building Exploratory Models

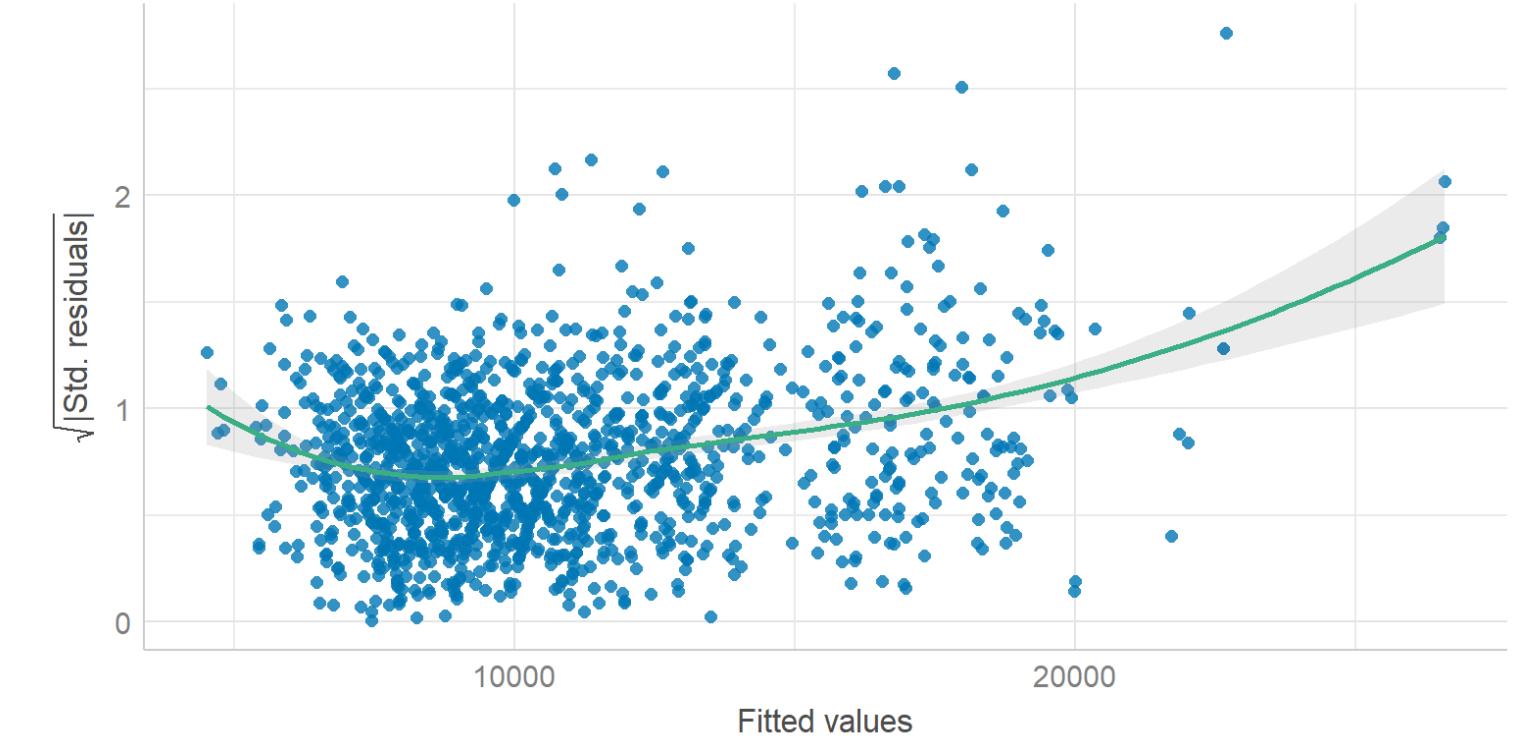
Model Diagnostic: Checking normality assumption

Normality of Residuals  
Distribution should be close to the normal curve



Model Diagnostic: Checking model for homogeneity of variances

Homogeneity of Variance  
Reference line should be flat and horizontal



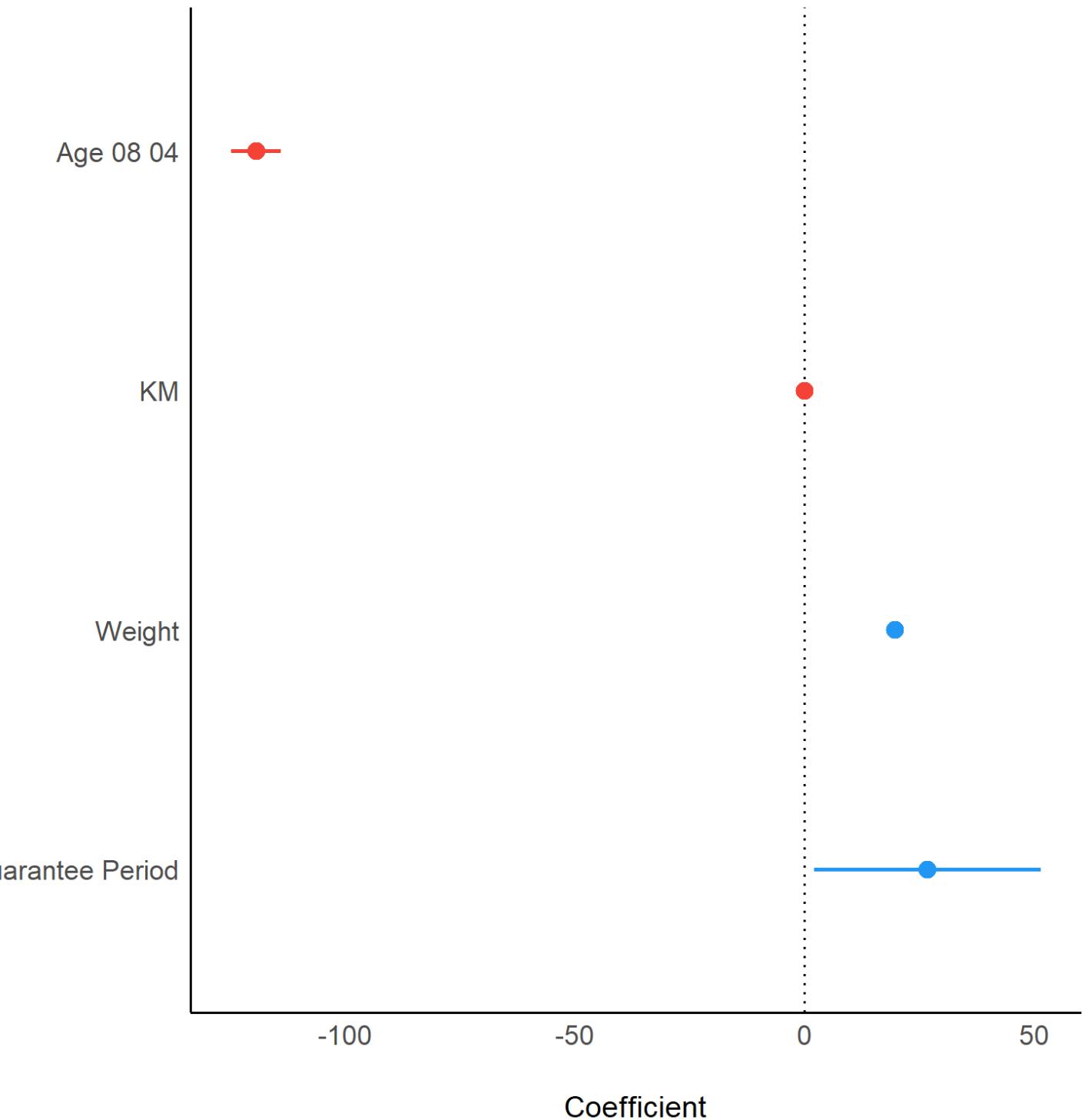
# Visual Analytics Approach for Building Exploratory Models

## 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

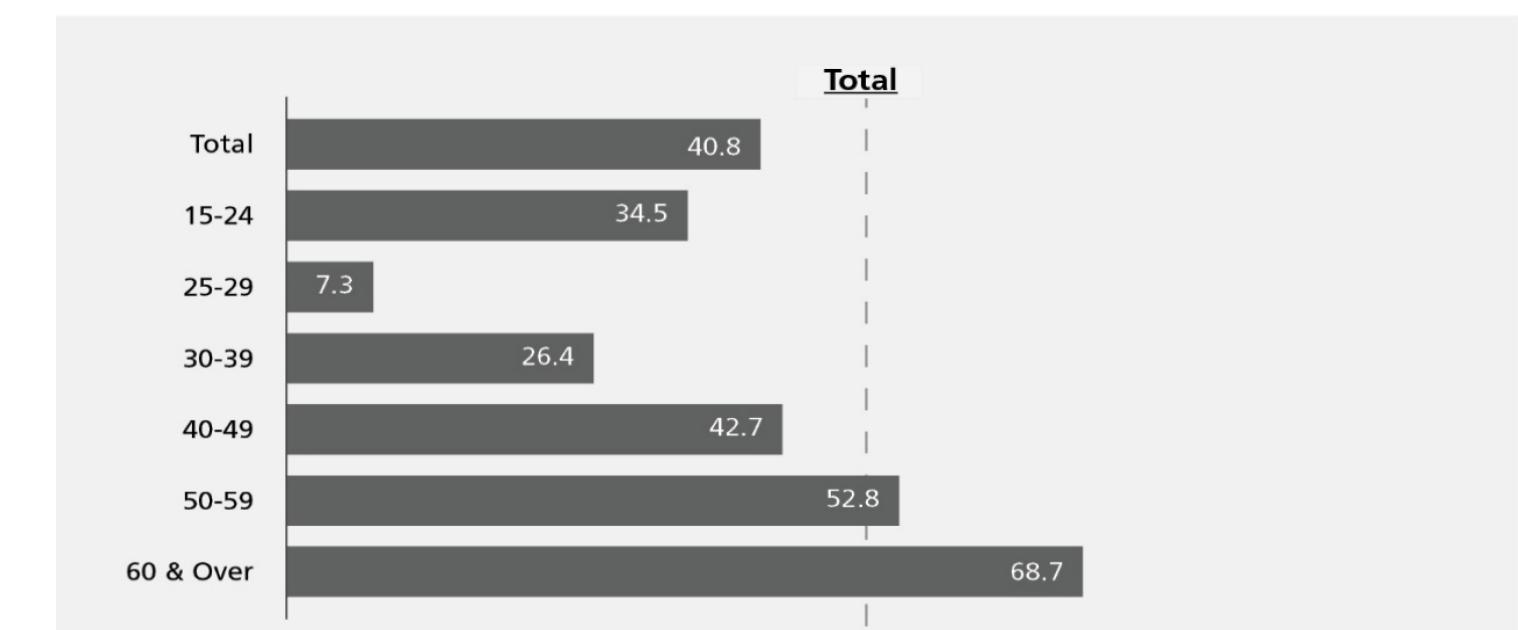


# Visualising Uncertainty

## Why it is important?

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

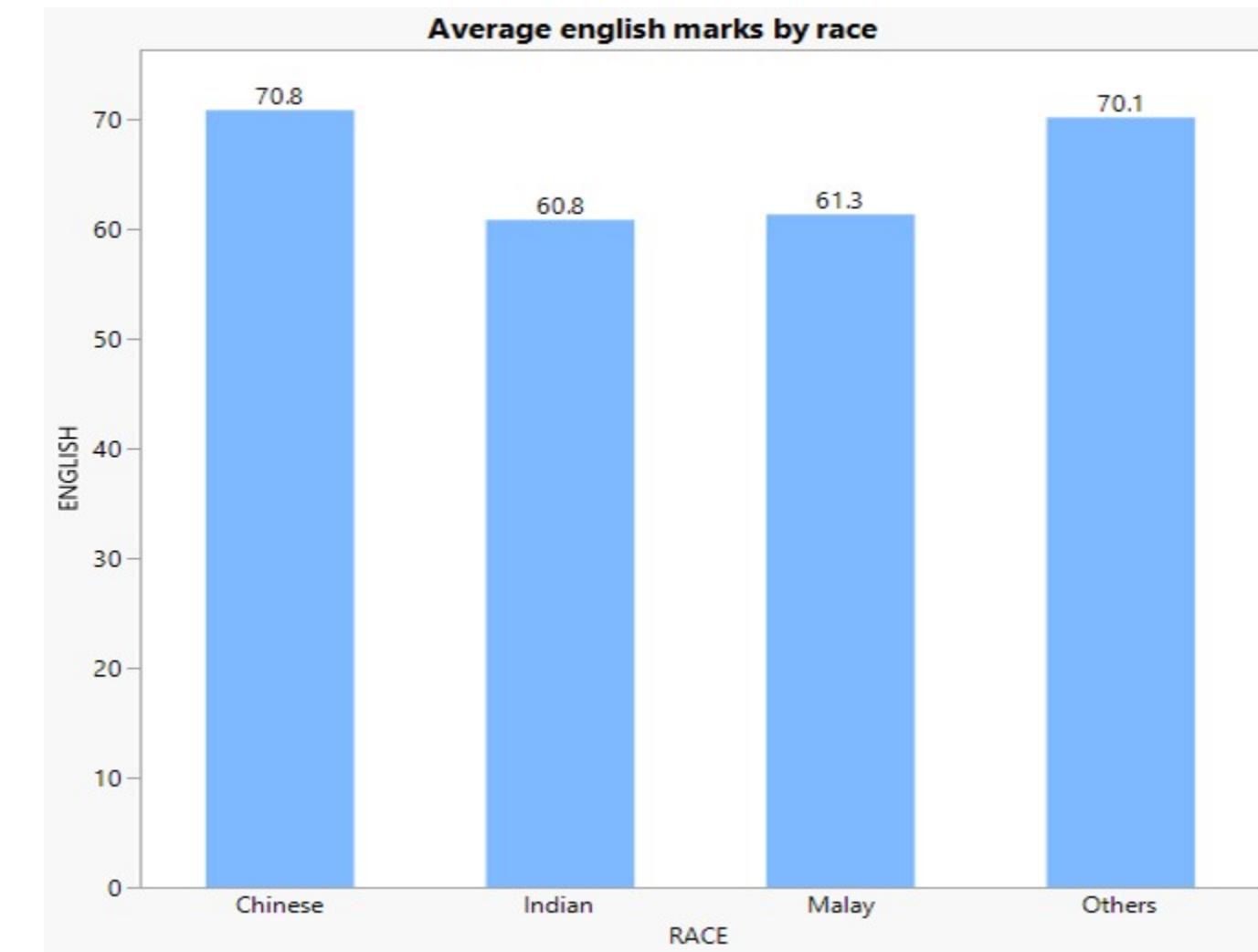
Proportion of resident potential entrants who preferred to work part-time by age group and sex, June 2019  
Per Cent



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

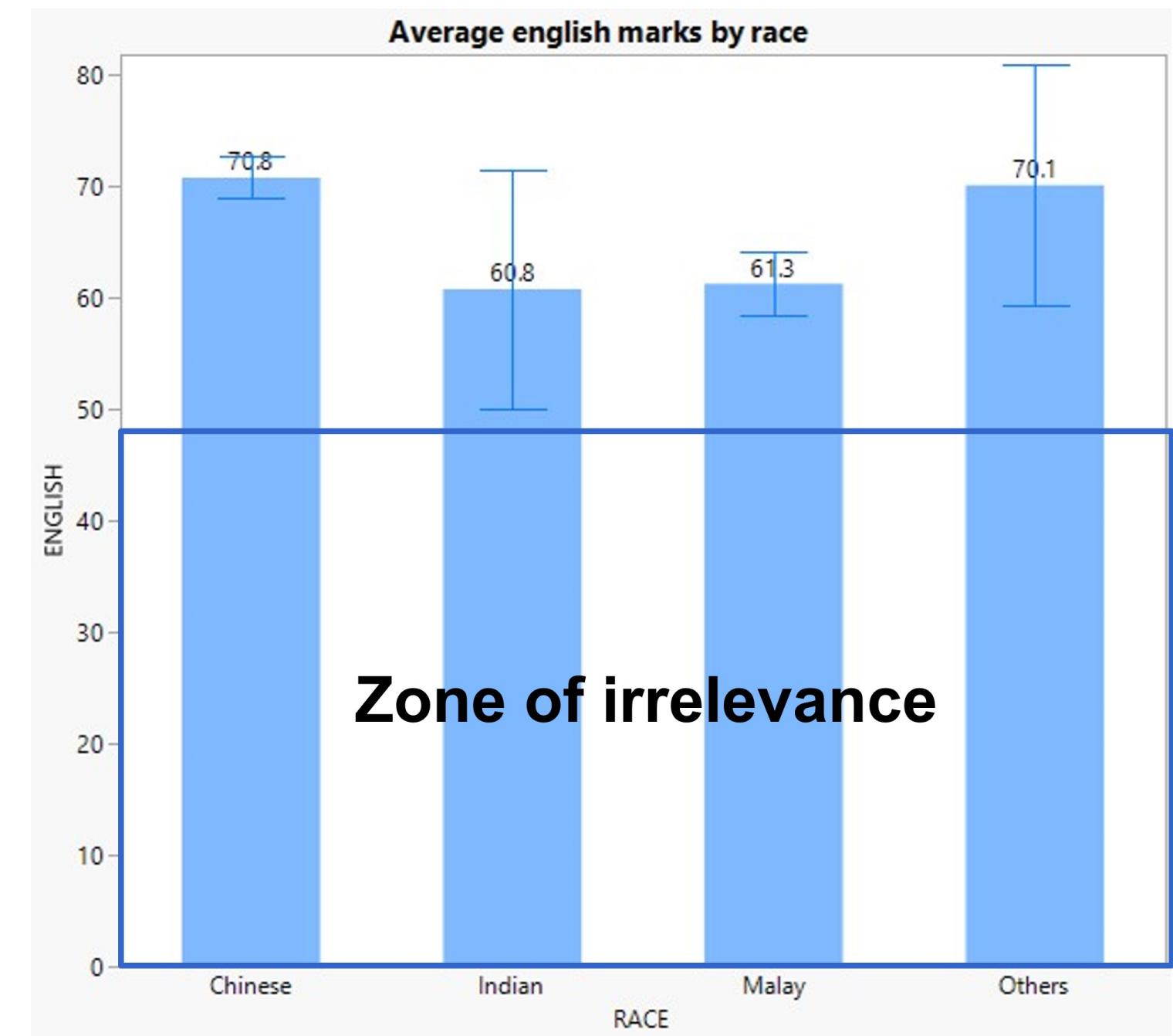
# Why one shouldn't use a bar graph, even if the data are normally distributed?

- It is not appropriate to displace average values on bars.



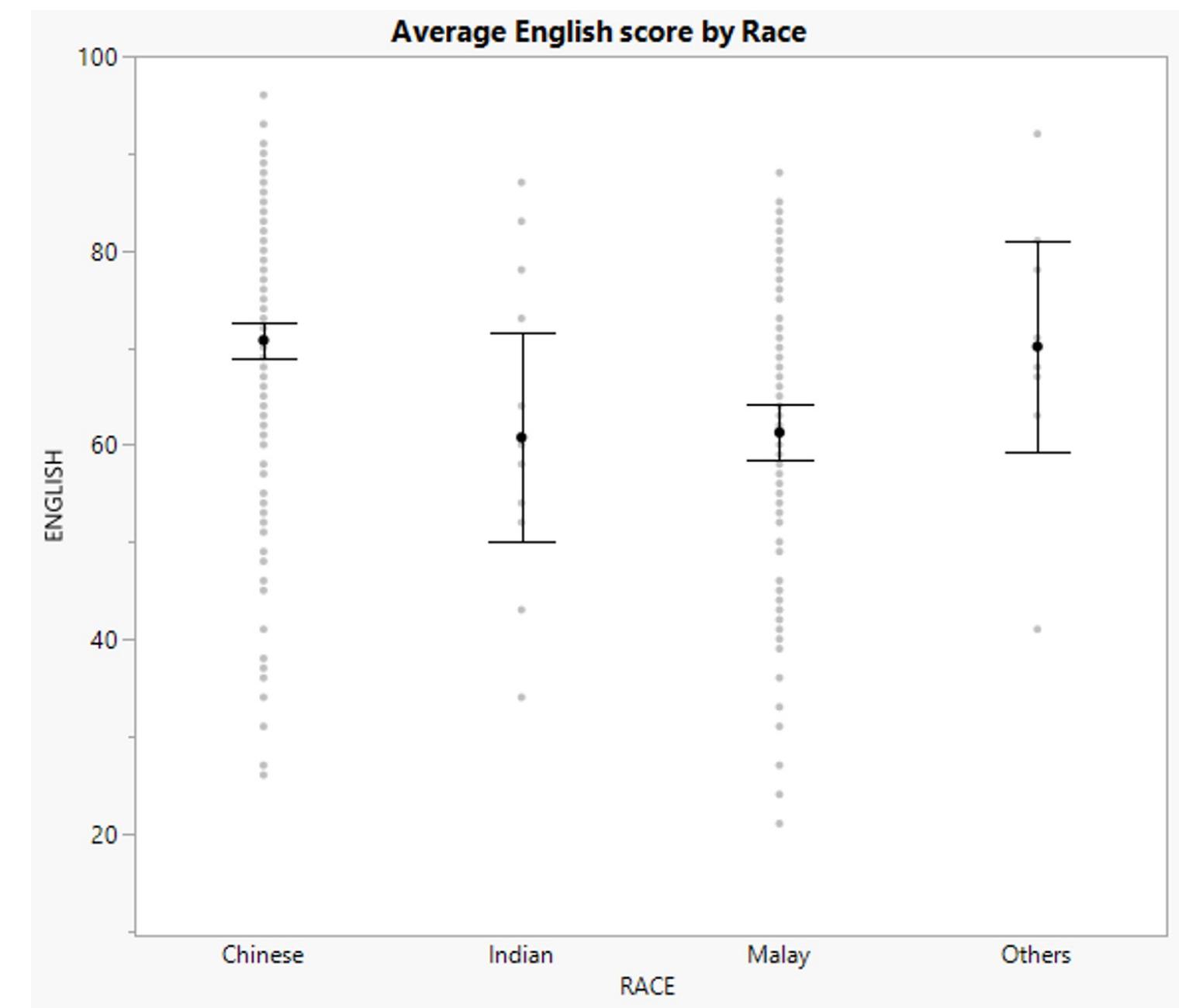
# Why Error bar failed?

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

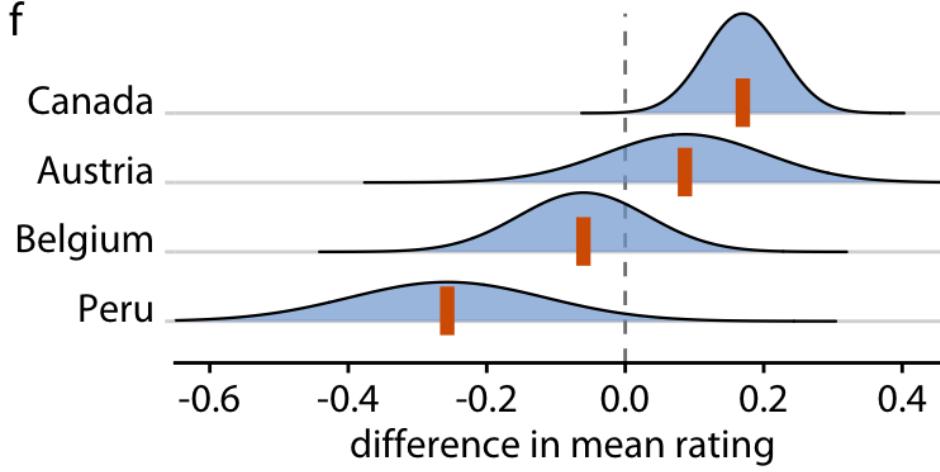
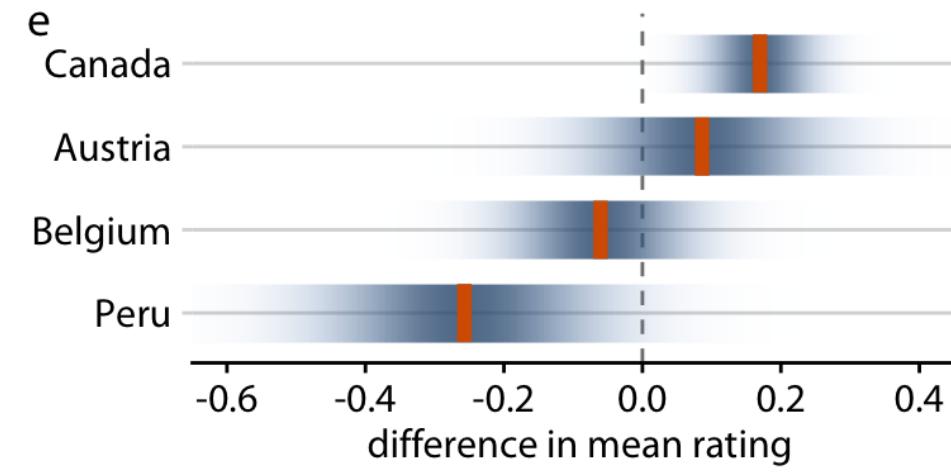
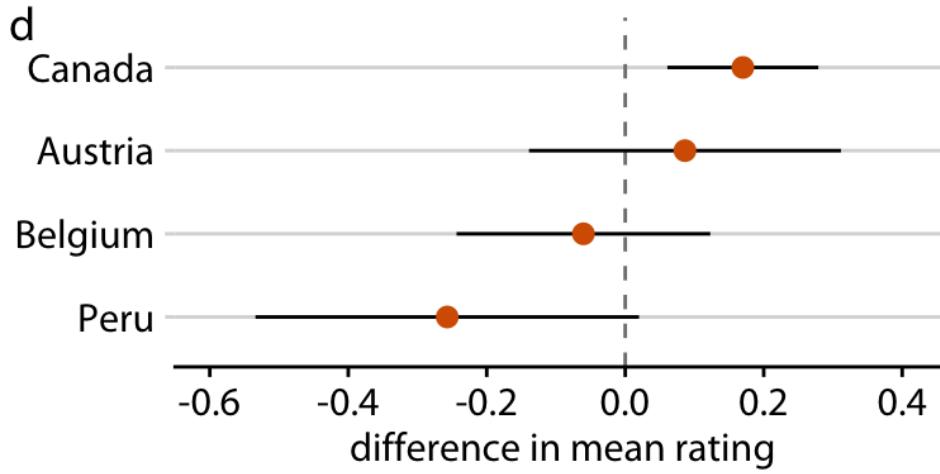
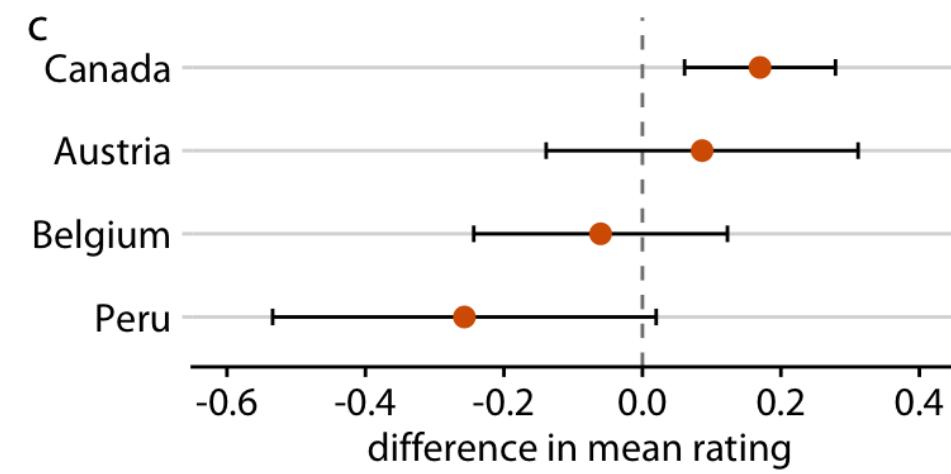
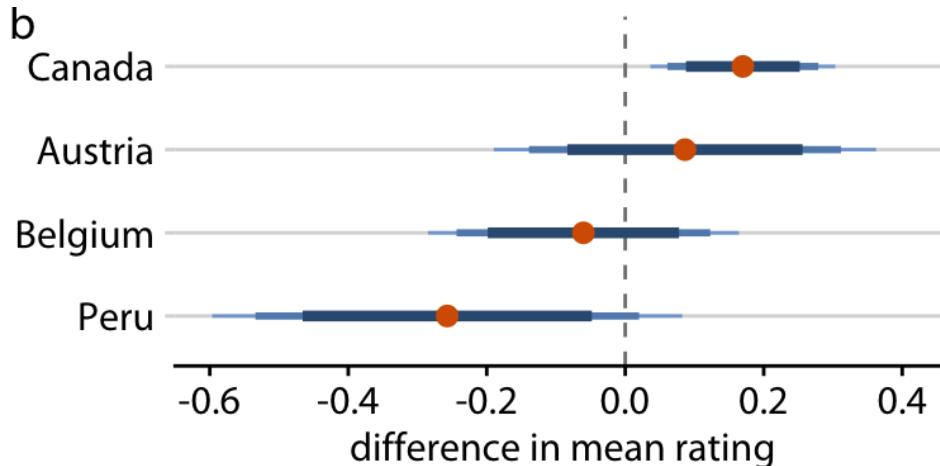
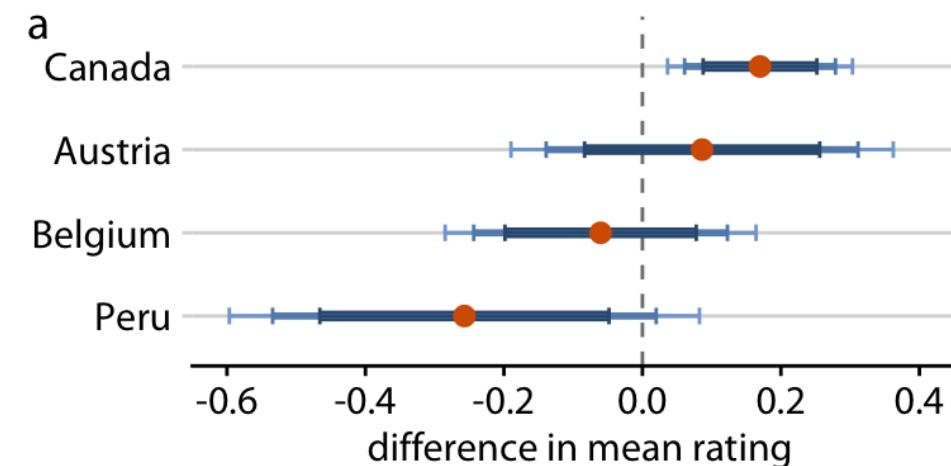


# Error bar on a dot plot

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



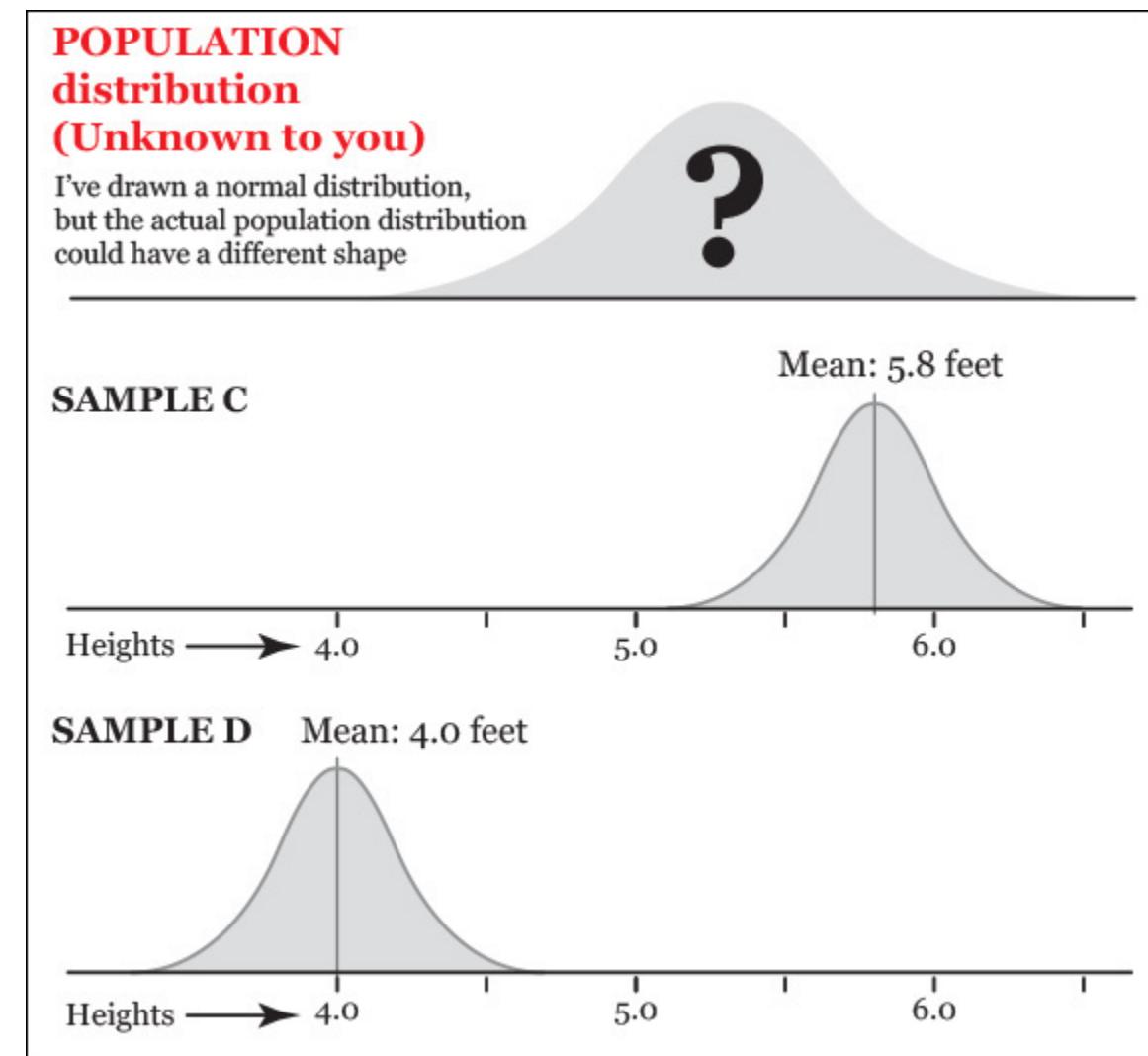
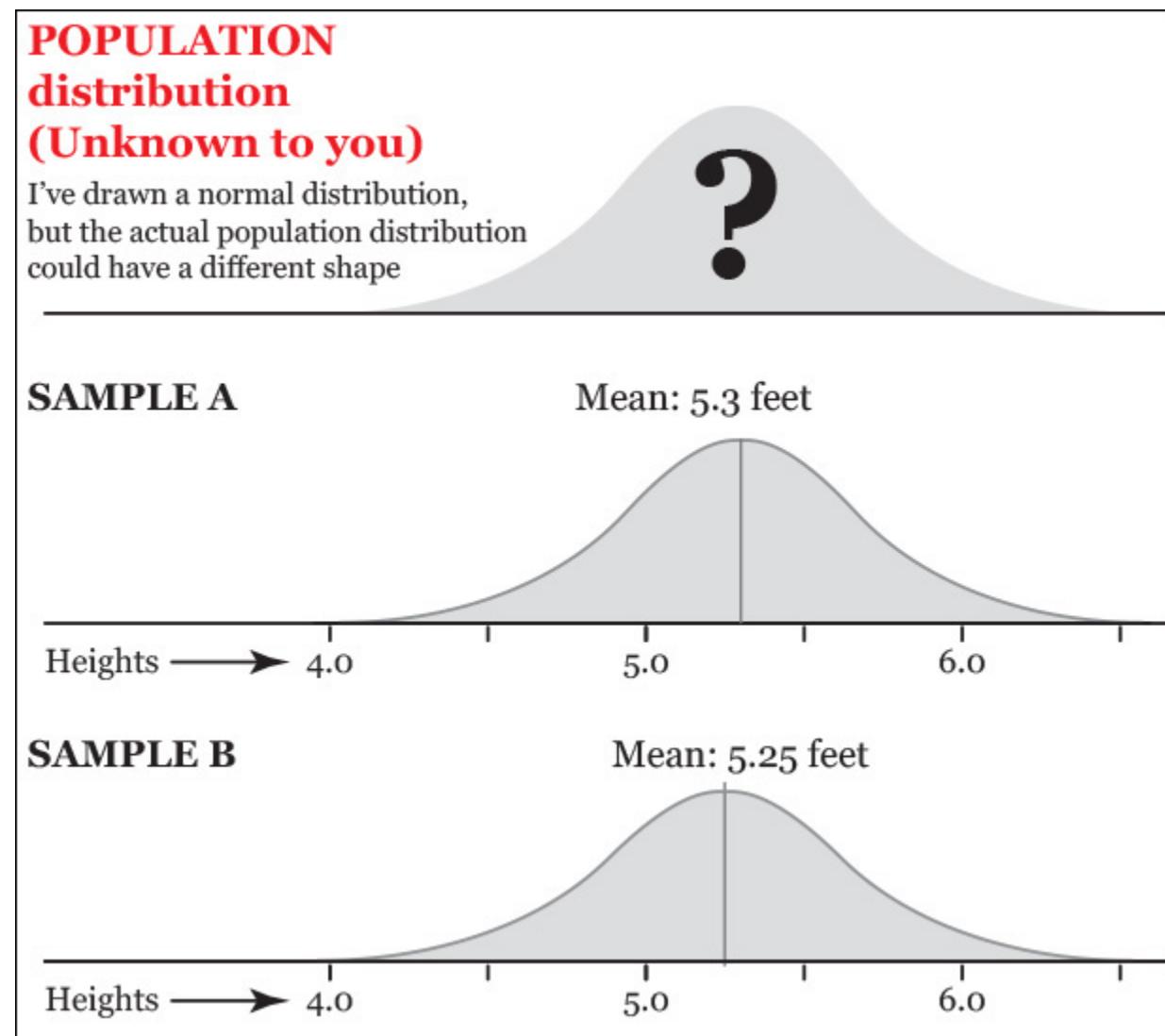
# Graphical methods for visualising uncertainty



Reference: [Visualizing the uncertainty of point estimates](#)

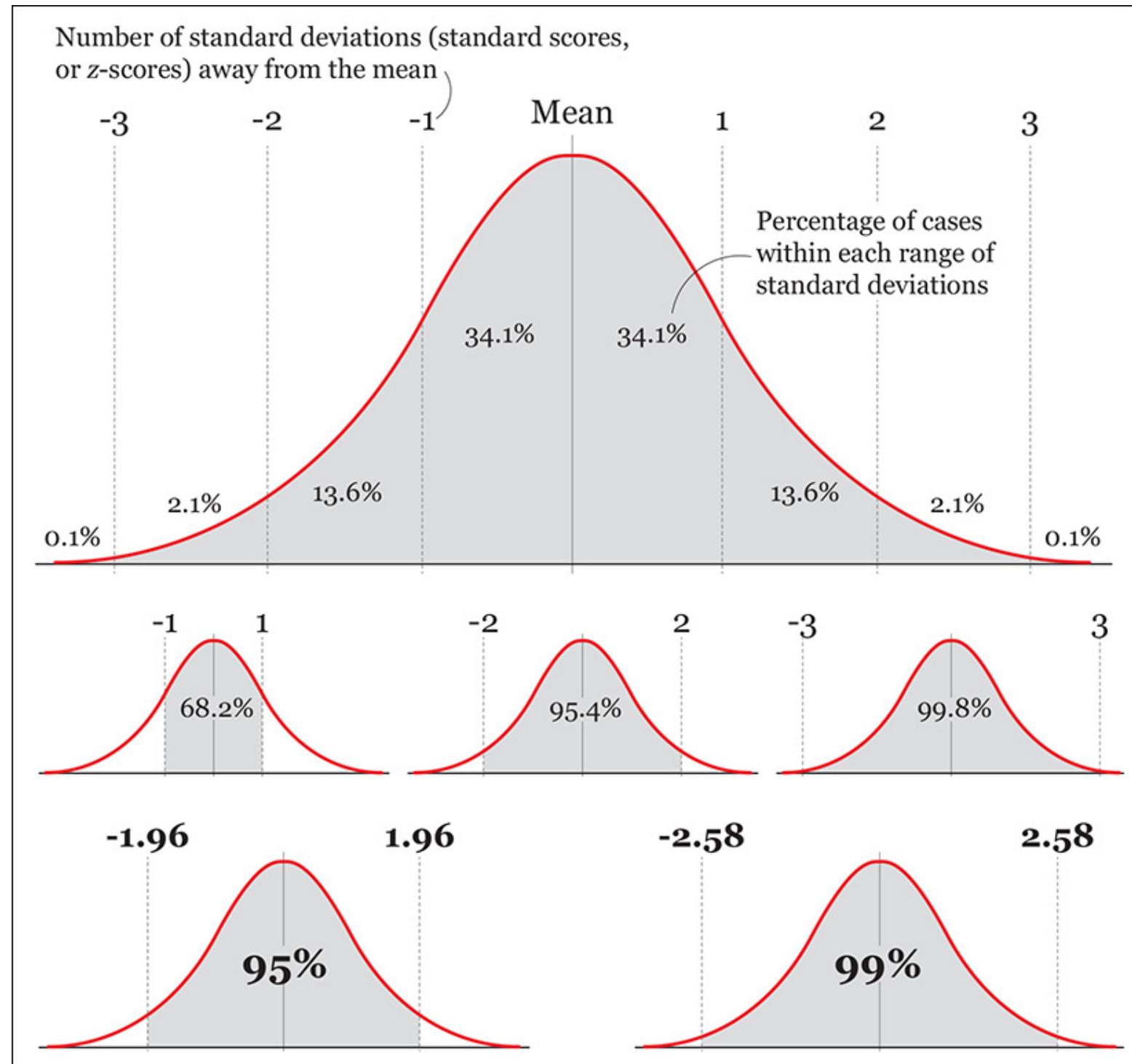
# Back to Statistics 101

## 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.

# 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

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

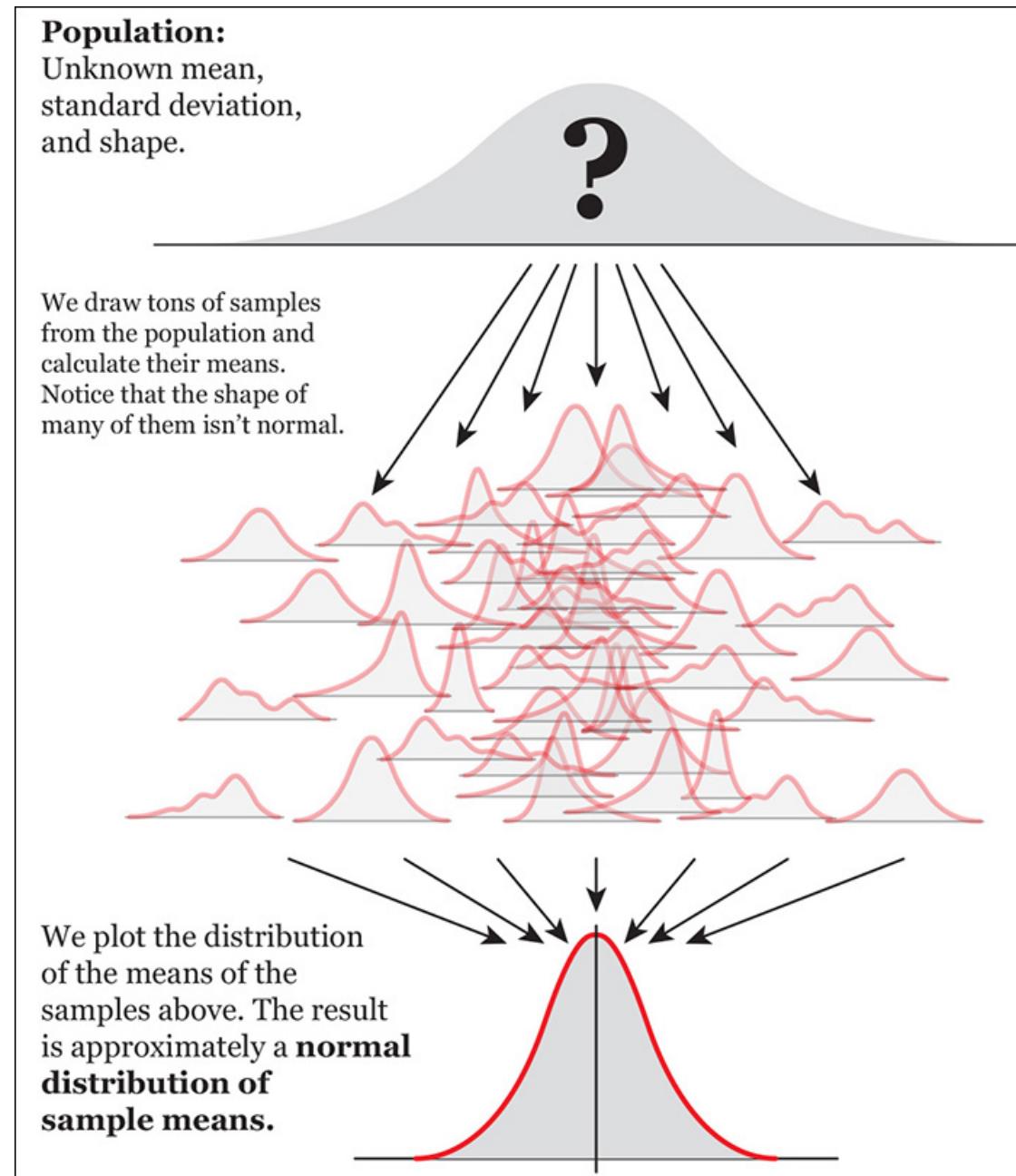
$$\text{variance} = \sigma^2$$

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

**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

## Calculating the confidence interval of a mean

$$(\text{Point value}) \pm Z \times \text{standard error}$$

Remember that  
**standard error** =  $\frac{\text{standard deviation of the sample}}{\sqrt{\text{sample size}}}$

therefore...

$$(\text{Point value}) \pm Z \times \frac{\text{standard deviation of the sample}}{\sqrt{\text{sample size}}}$$

Note: Statisticians often use "sample size minus 1" (or  $n-1$ ) here, rather than simply the sample size. This is a correction applied when dealing with smallish sample sizes, and when the standard deviation of the population is unknown (almost always!) If the sample size is large, the "minus 1" correction doesn't make a big difference.

### For a 95% confidence level

$$(\text{Point value}) \pm 1.96 \times \frac{\text{standard deviation of the sample}}{\sqrt{\text{sample size}}}$$

in our girl height example...

$$5.3 \text{ feet} \pm 1.96 \times \frac{0.5 \text{ feet}}{\sqrt{40 \text{ girls}}}$$

$$5.3 \text{ feet} \pm 0.16$$

### For a 99% confidence level

$$(\text{Point value}) \pm 2.58 \times \frac{\text{standard deviation of the sample}}{\sqrt{\text{sample size}}}$$

in our girl height example...

$$5.3 \text{ feet} \pm 2.58 \times \frac{0.5 \text{ feet}}{\sqrt{40 \text{ girls}}}$$

$$5.3 \text{ feet} \pm 0.21$$

# Calculating the confidence interval of a percentage

## Confidence interval of a percentage

$$(\text{Percentage}) \pm Z \times \sqrt{\frac{\text{Percentage} \times (100-\text{Percentage})}{\text{sample size}}}$$

Let's apply the formula:

A survey (sample size = 300 randomly chosen voting-age citizens) says that 45.3% of citizens will vote for candidate Jane Doe. What's the confidence interval of that percentage?

For a 95% confidence level

$$45.3\% \pm 1.96 \times \sqrt{\frac{45.3 \times (100-45.3)}{300}}$$

$\downarrow$

$$45.3\% \pm 5.63$$

For a 99% confidence level

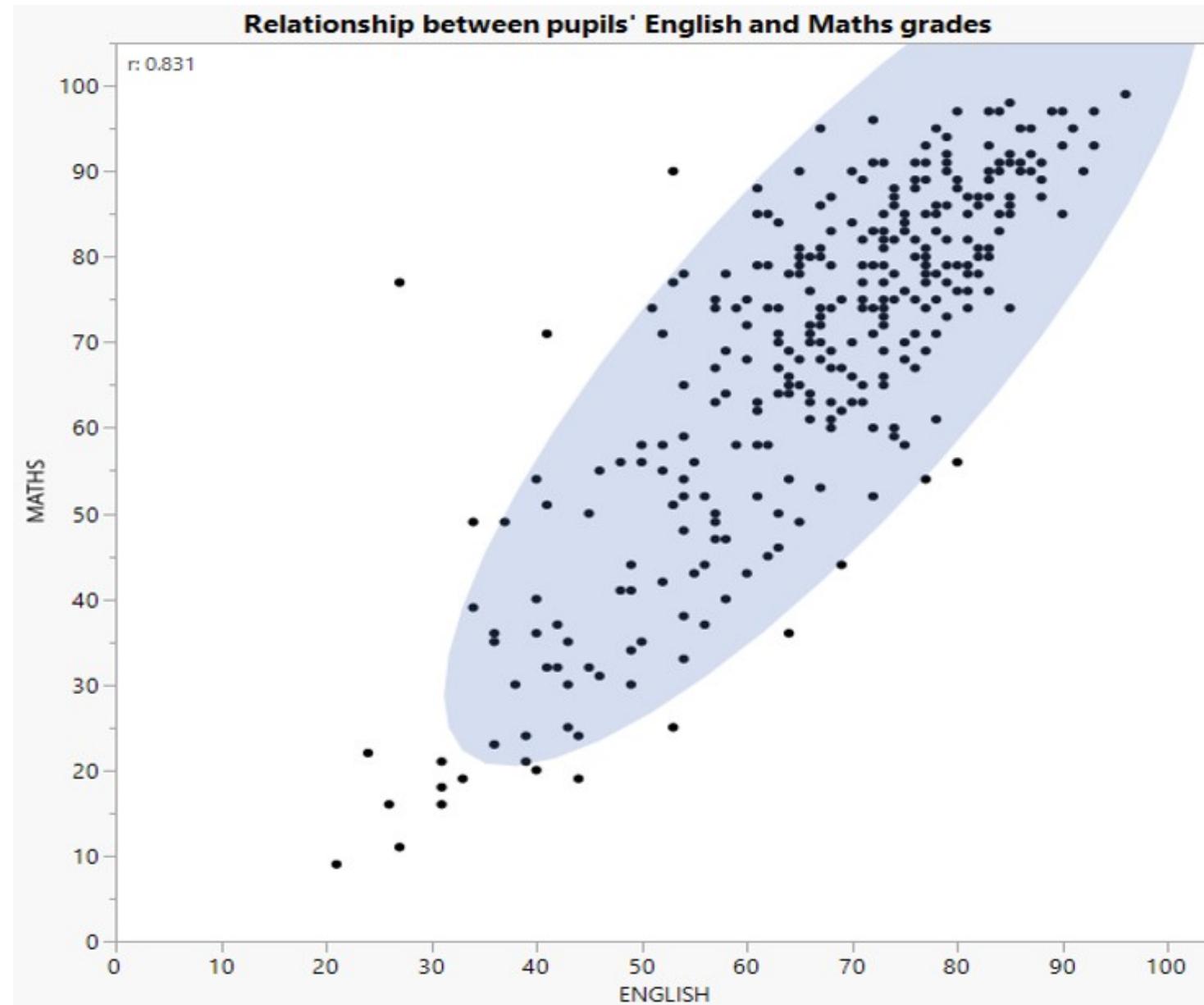
$$45.3\% \pm 2.58 \times \sqrt{\frac{45.3 \times (100-45.3)}{300}}$$

$\downarrow$

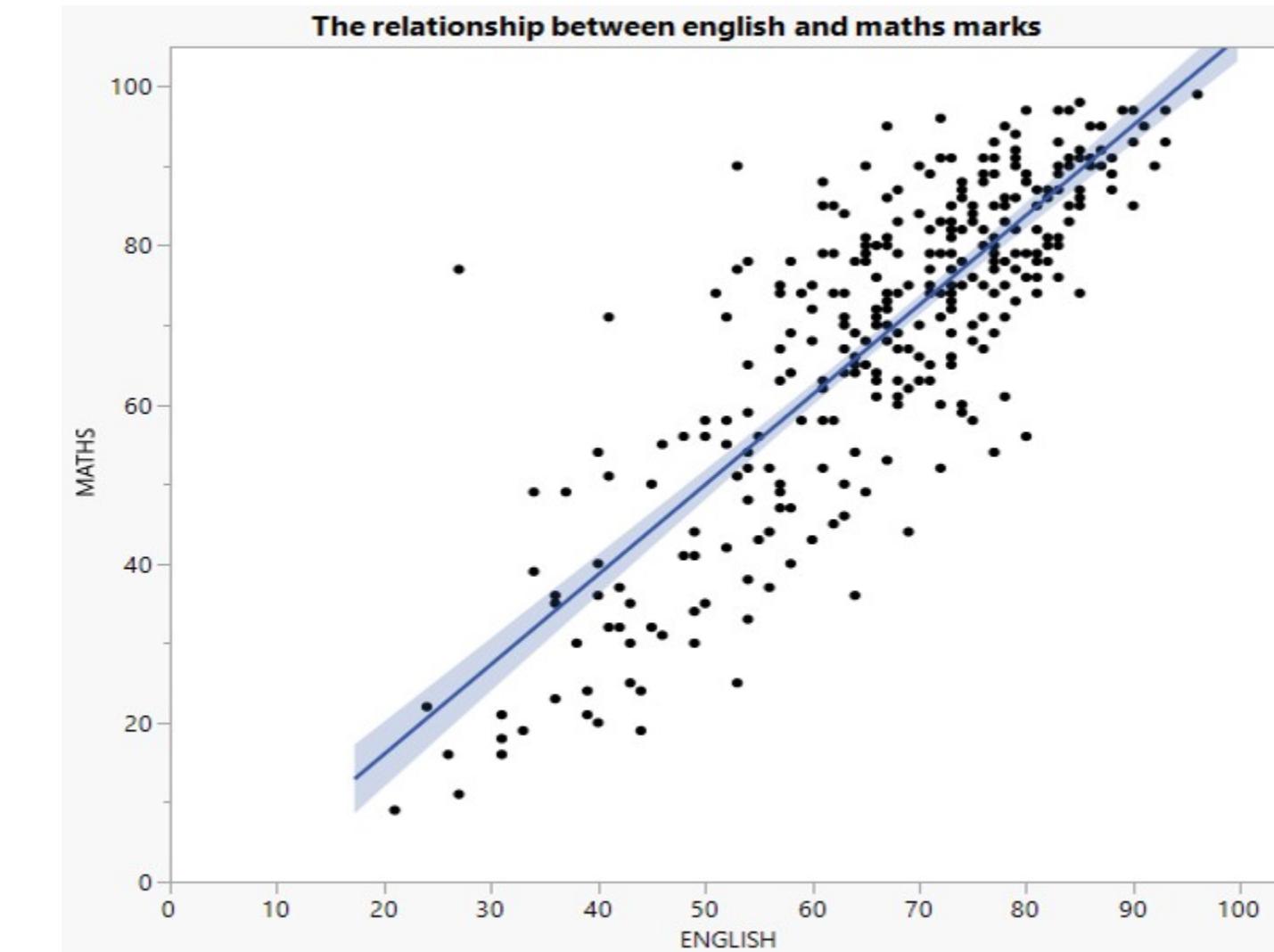
$$45.3\% \pm 7.41$$

# 2-d graphical methods for visualising uncertainty

Scatter plot with 95% confidence ellipse

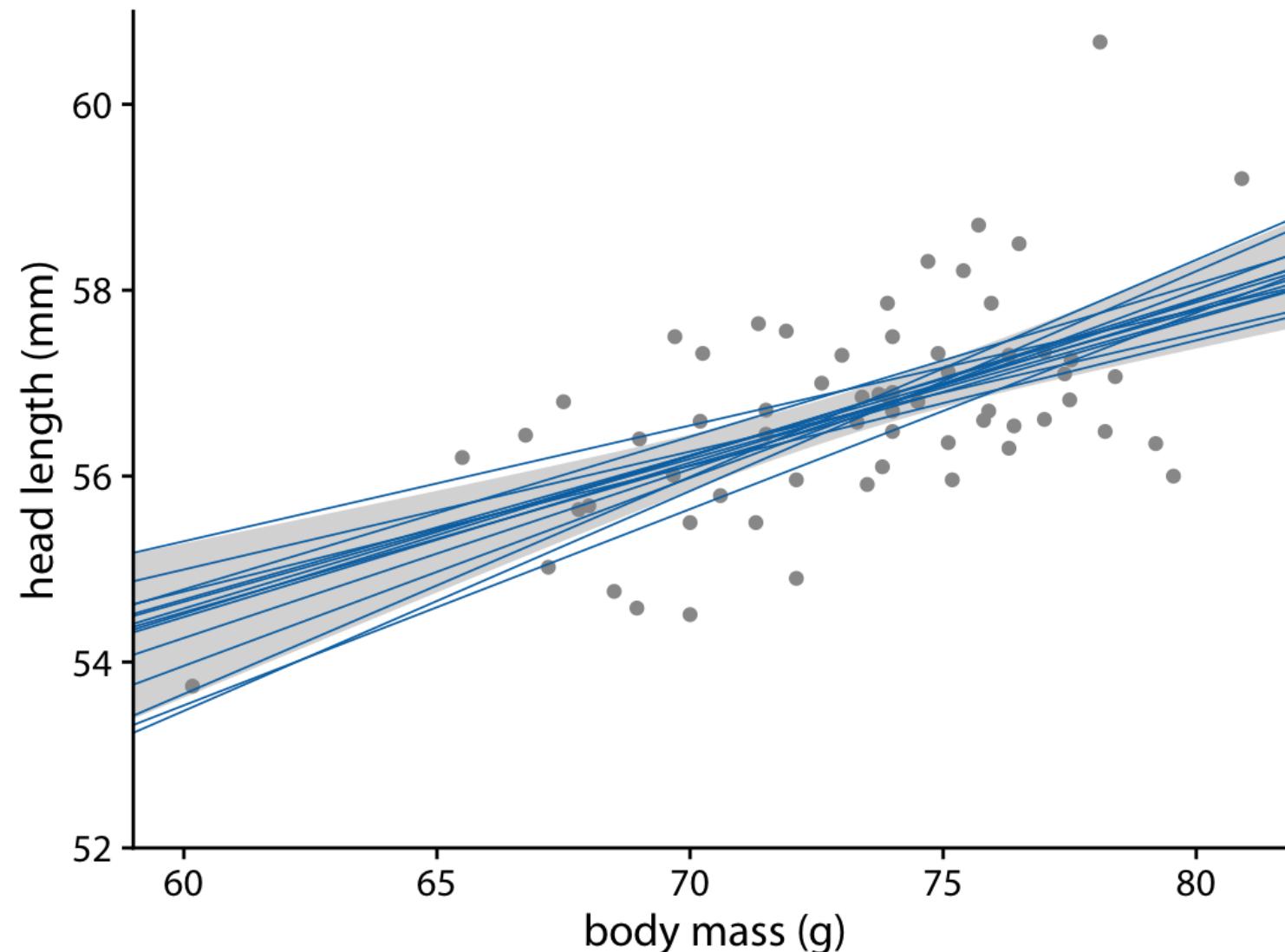


Scatter plot with 95% confidence intervals

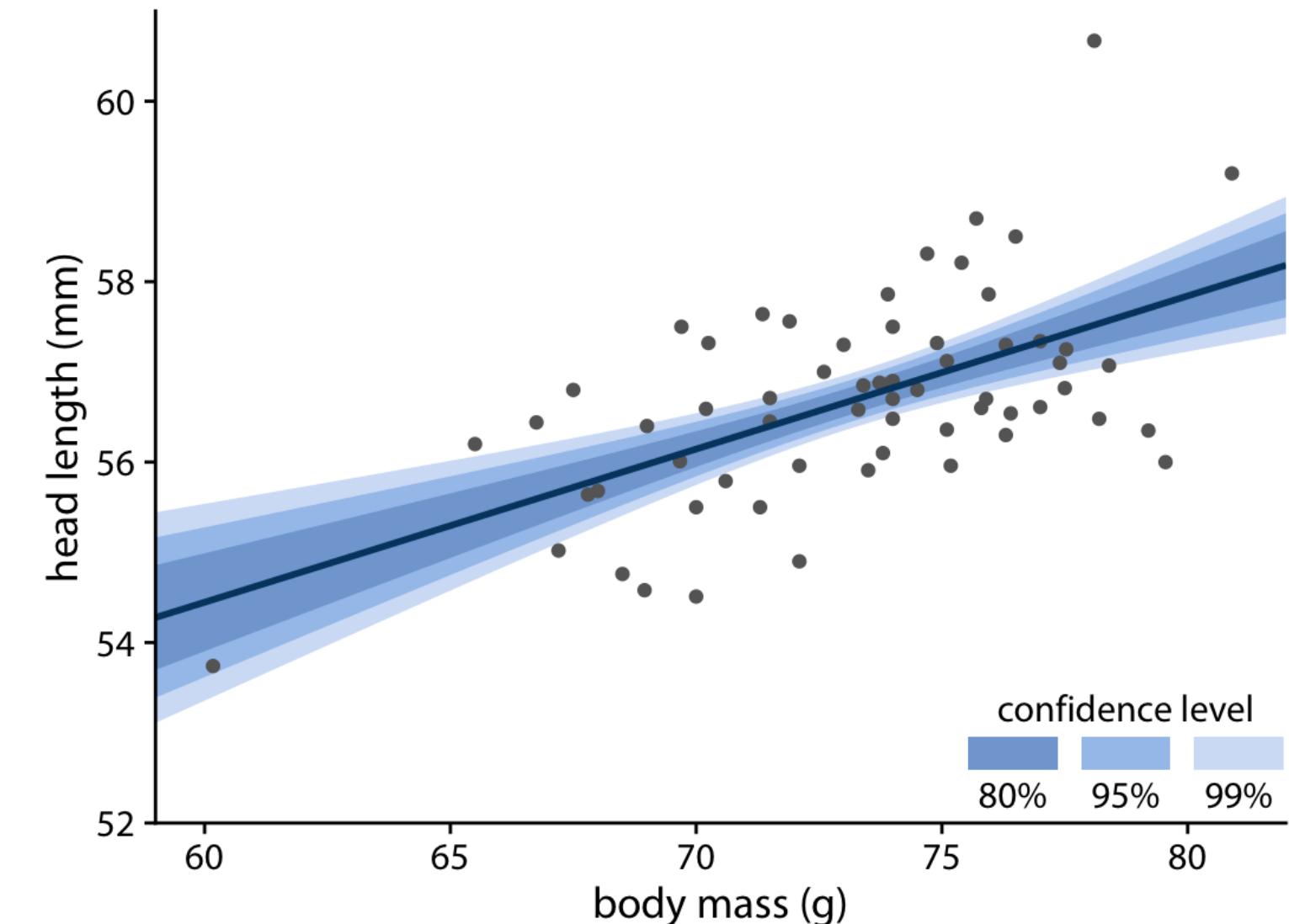


# Confidence band of a trend line

## Confidence band and fit lines

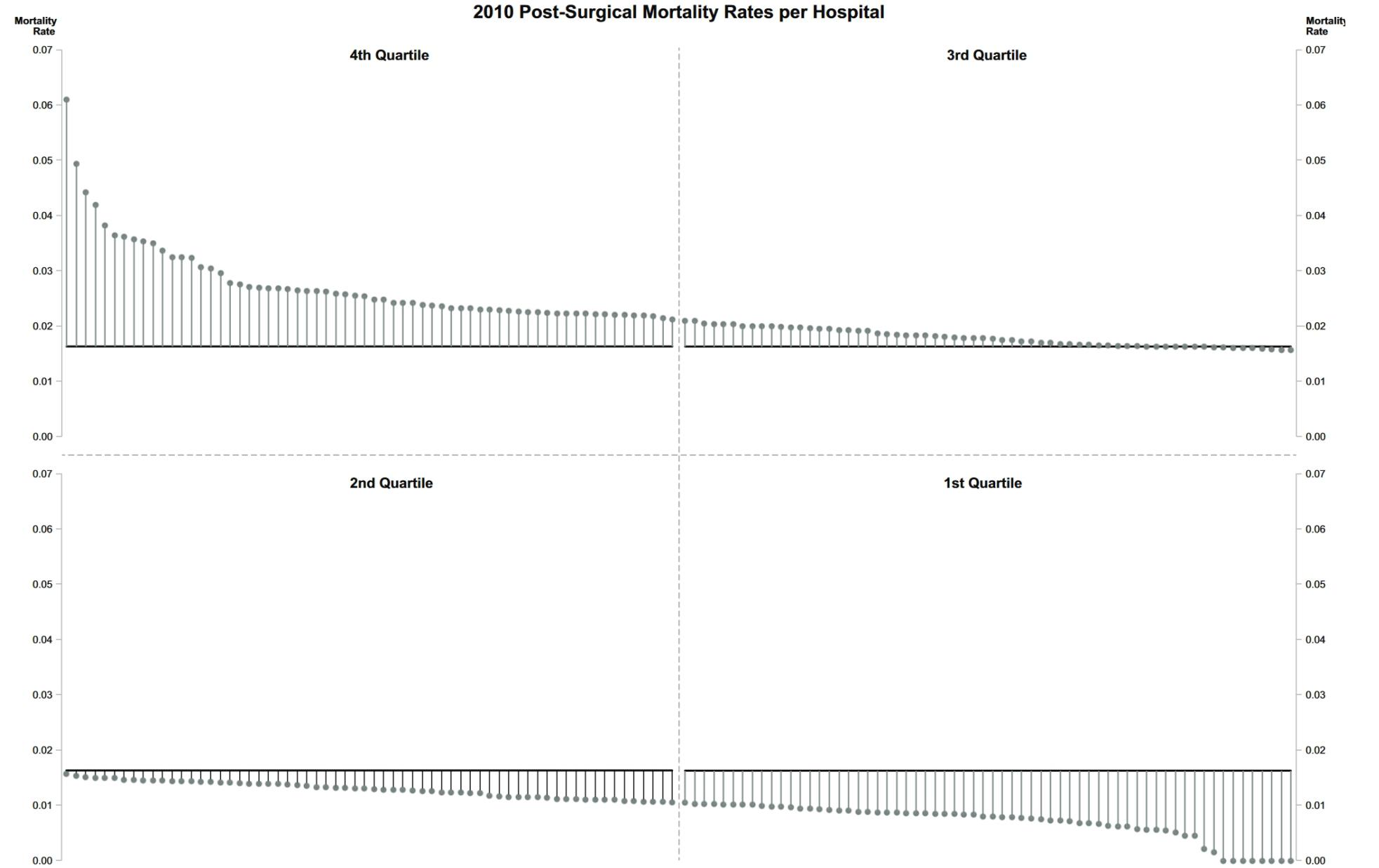


## A graded confidence band



# Variation and Its Discontents

## Random and unfair comparisons



Reference: [Variation and its discontents](#)

# Funnel plots to the rescue

## Statistical details

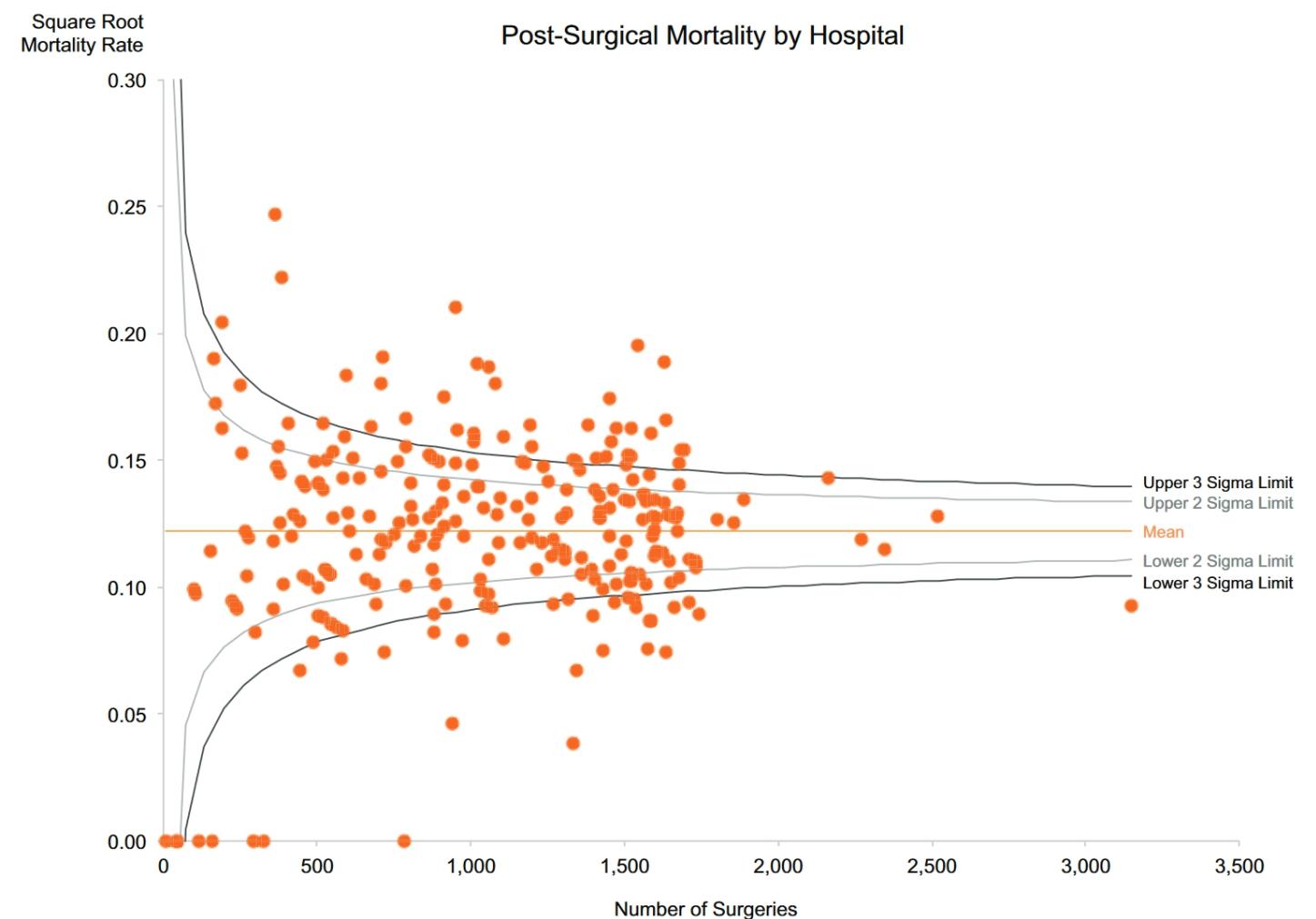
Calculations:

$$\text{Fit Mean} = \frac{\text{Sum of Occurrences}}{\text{Sum of Sample Sizes}}$$

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

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

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



# 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.

