Verification of the Dataset used to Build the Models

(Chapter 5 – Software Project Estimation)

Alain Abran

(Tutorial Contribution: Dr. Monica Villavicencio)

Topics covered

1. Introduction
2. Verification of the direct inputs
3. Graphical analysis – one dimensional
4. Analysis of the distribution of the input variables
5. Graphical analysis – two dimensional
6. Size inputs derived from a conversion formula

5.1 Introduction

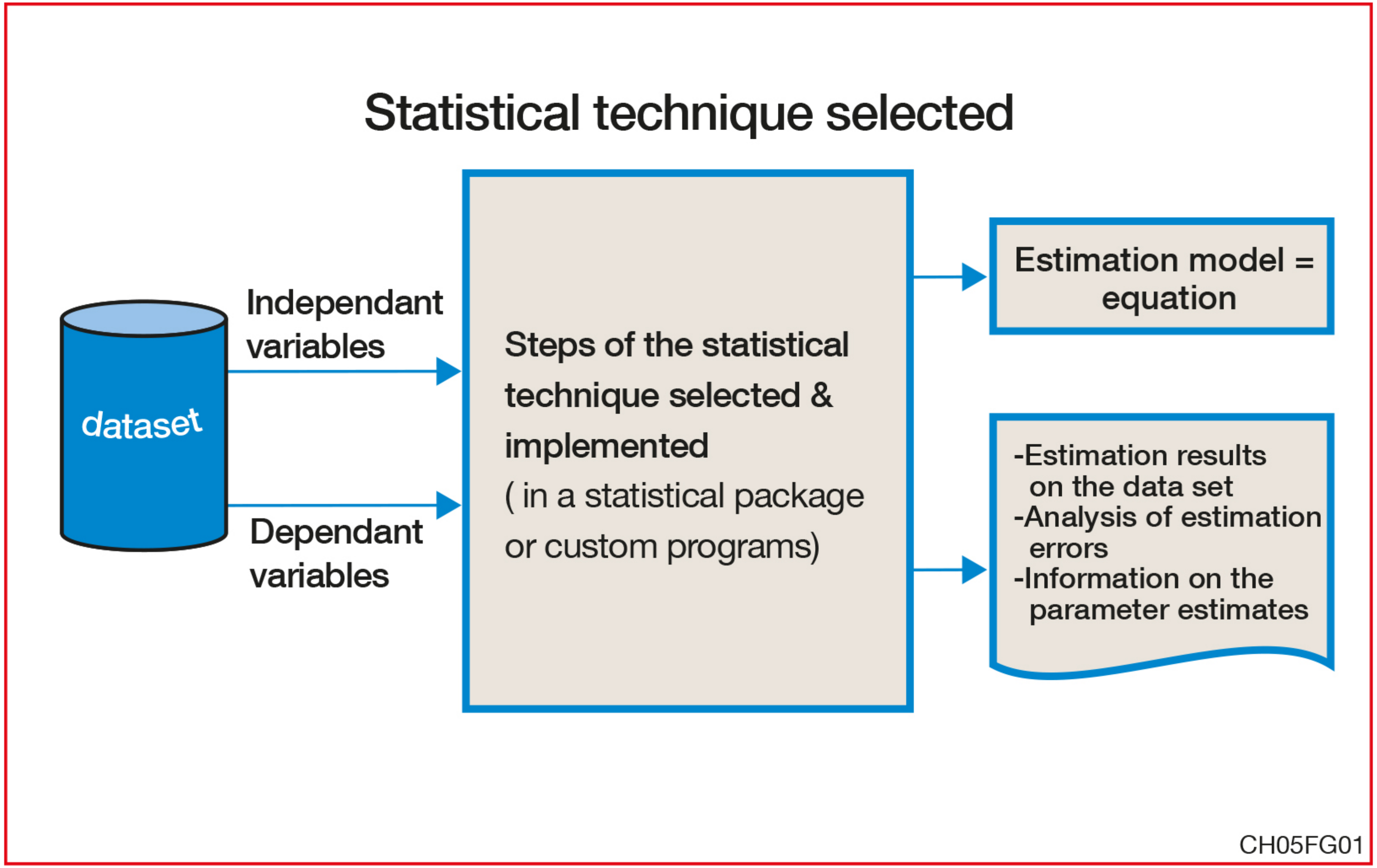
Productivity models based on statistical techniques consist of:

- Inputs (dataset of independent and dependent variables).
- Steps specific to the statistical technique used.
- Outcomes that include:
  - the mathematical equation of the productivity model built;
  - the estimation results on the dataset used;
  - the information on the estimation errors of the model on the original dataset.

Proper use of statistical techniques to build a productivity model

1. Verification of the characteristics of the input variables in the dataset.
2. Verification of the proper execution of the steps of the statistical techniques.
3. Verification of the characteristics of the output variable

Building a productivity model



Statistical technique selected

dataset
Independant variables
Dependant variables
Steps of the statistical technique selected & implemented
(in a statistical package or custom programs)
Estimation model = equation
-Estimation results on the data set
-Analysis of estimation errors
-Information on the parameter estimates
CH05FG01

5.2 Verification of Direct Inputs

Verification of direct inputs

1. Verification of the data definitions and data quality
2. Verification of the measurement scale types
3. Verification of the characteristics of the output variable

Verification of data definitions & quality

or

http://alifeoflight.com/garbage-in-garbage-out/

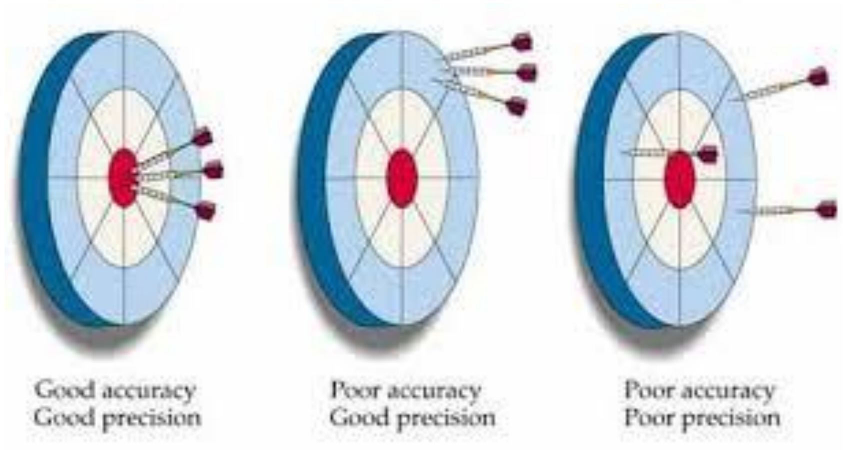
Garbage in = garbage out

Verification of measurement scale types

- Nominal
- Ordinal (ranks: cannot be added or multiplied)
- Interval
- Ratio

Quality criteria for attributes (ISO VIM 2007)

- Accuracy
- Precision
- Repeatability
- Reproducibility
- Etc.



http://gis.stackexchange.com/questions/26410/need-to-learn-more-about-precision-and-accuracy

Well designed measurement methods ....

- are expressed in a single measurement unit.
- use standard-etalons.

Exercise:

- Which software measures:
  - Meet these criteria ....?
  - DO not meet these criteria..................?

5.3 Graphical analysis – one dimensional

Graphical analysis – One dimensional

One-dimensional graphical analysis will typically provide the user with an intuitive feel about the data collected, one data field at a time.

Distribution of data points

Criteria to look for when analyzing the input data:

- Ranges of values: min, max.
- Dispersion of the values and the density of the ranges of data values.
- Gaussian distribution: skewness, kurtosis, normality tests, etc.
- Candidate outliers on a single dimension.
- Etc.

Dataset: Effort and Functional Size (N=21)

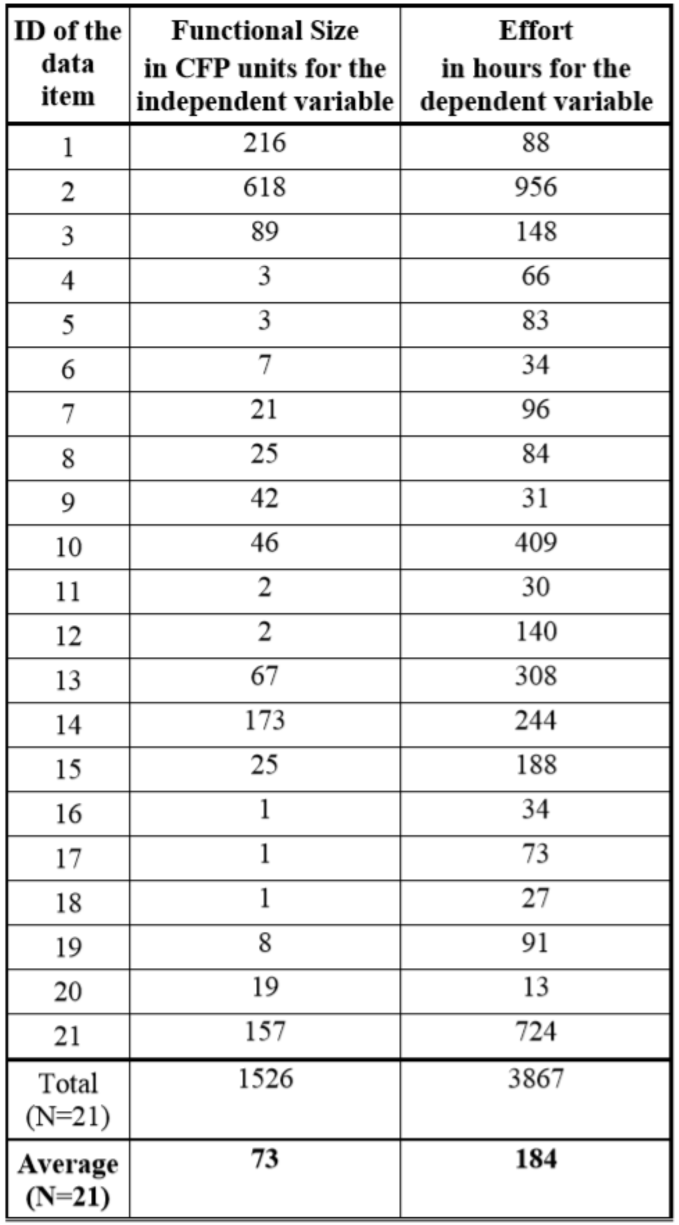
For this dataset of 21 values, find:

1- Ranges of values: min, max?

2- Dispersion of the values and the density of the ranges of data values?

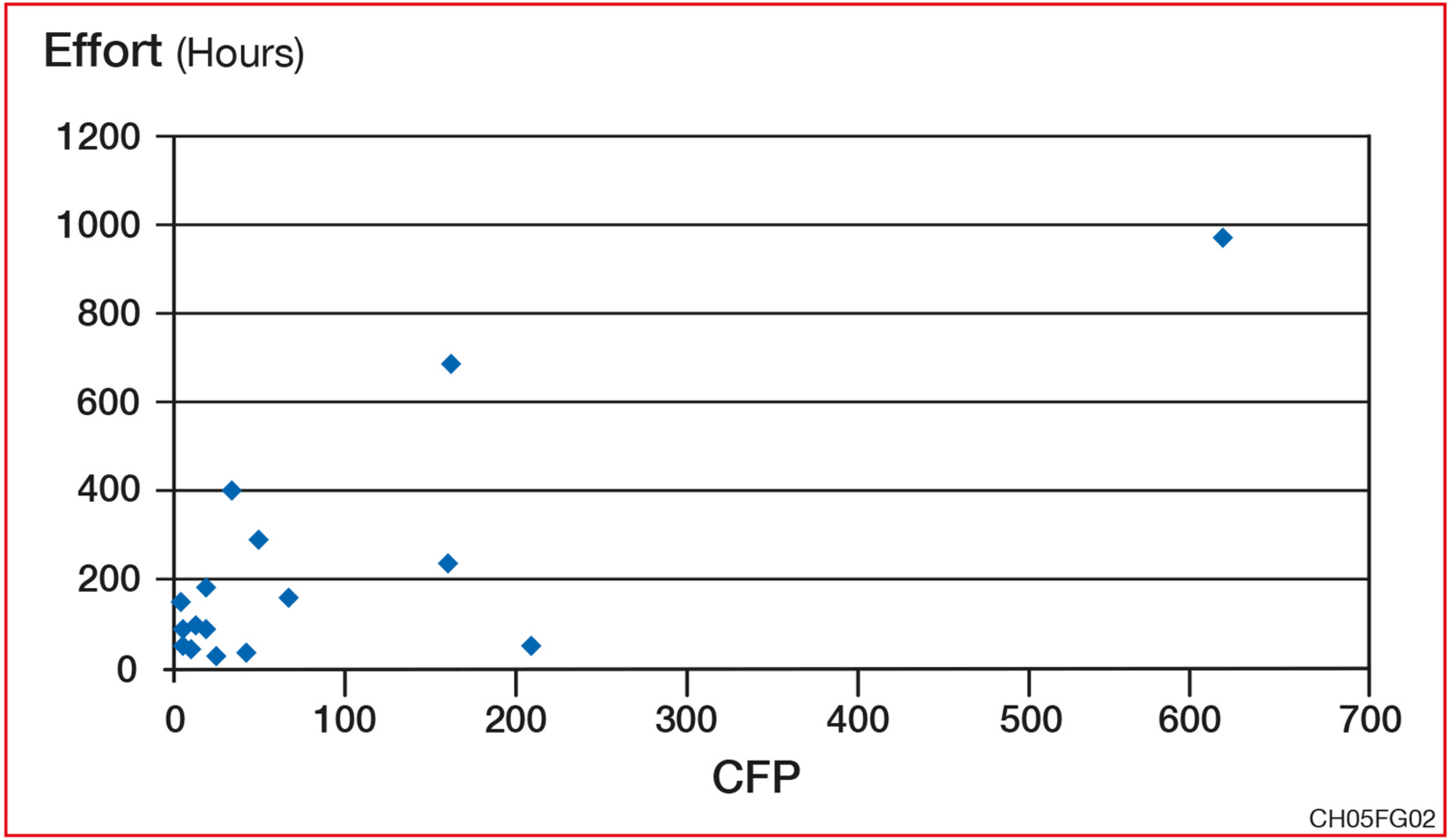
3- Gaussian distribution: skewness, kurtosis, normality tests, etc.?

4- Candidate outliers on a single dimension?

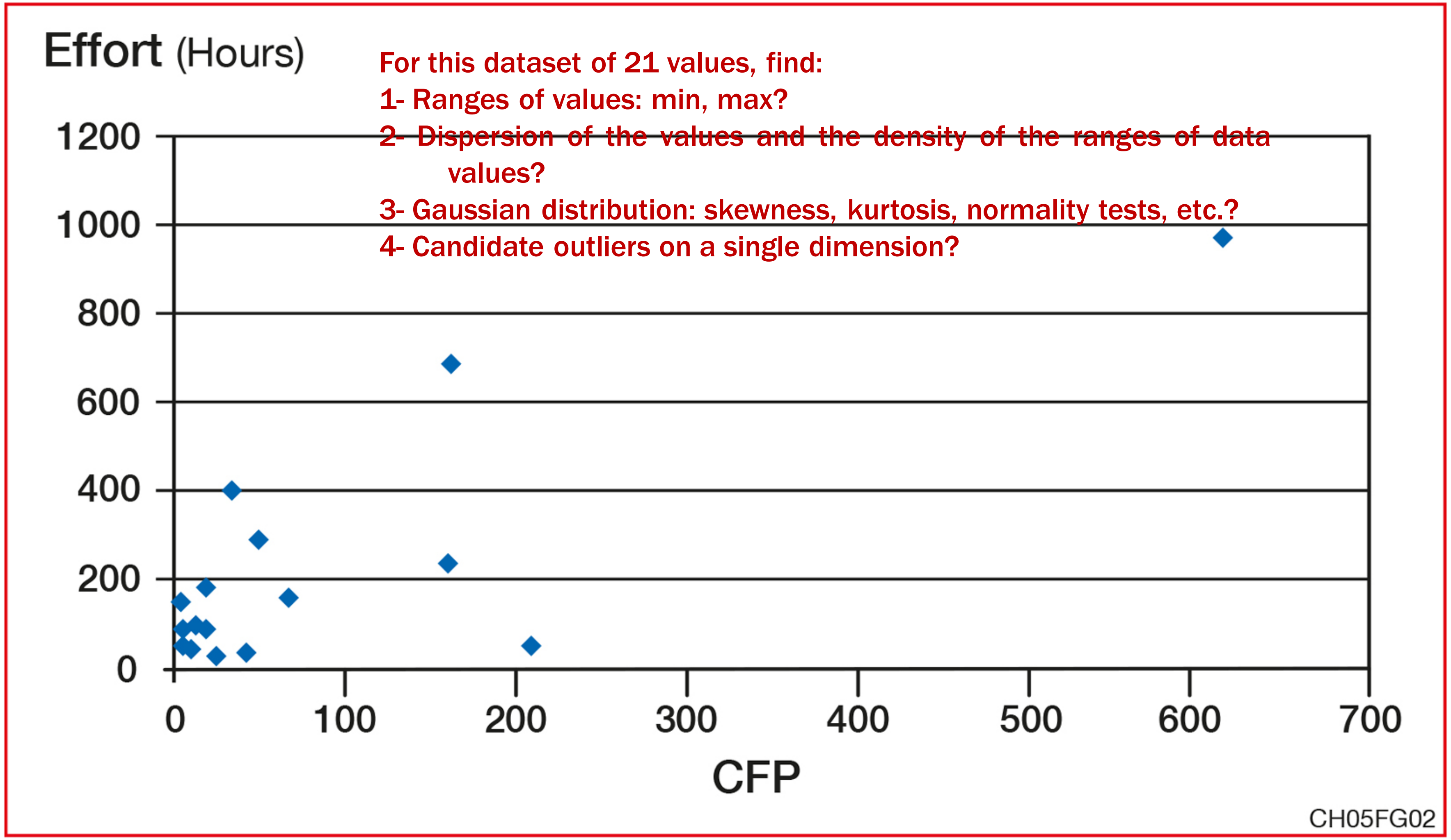


ID of the data item	Functional Size in CFP units for the independent variable	Effort in hours for the dependent variable
1	216	88
2	618	956
3	89	148
4	3	66
5	3	83
6	7	34
7	21	96
8	25	84
9	42	31
10	46	409
11	2	30
12	2	140
13	67	308
14	173	244
15	25	188
16	1	34
17	1	73
18	1	27
19	8	91
20	19	13
21	157	724
Total (N=21)	1526	3867
Average (N=21)	**73**	**184**

Graph of the dataset



Graph of the dataset



For this dataset of 21 values, find:

1- Ranges of values: min, max?

2- Dispersion of the values and the density of the ranges of data values?

3- Gaussian distribution: skewness, kurtosis, normality tests, etc.?

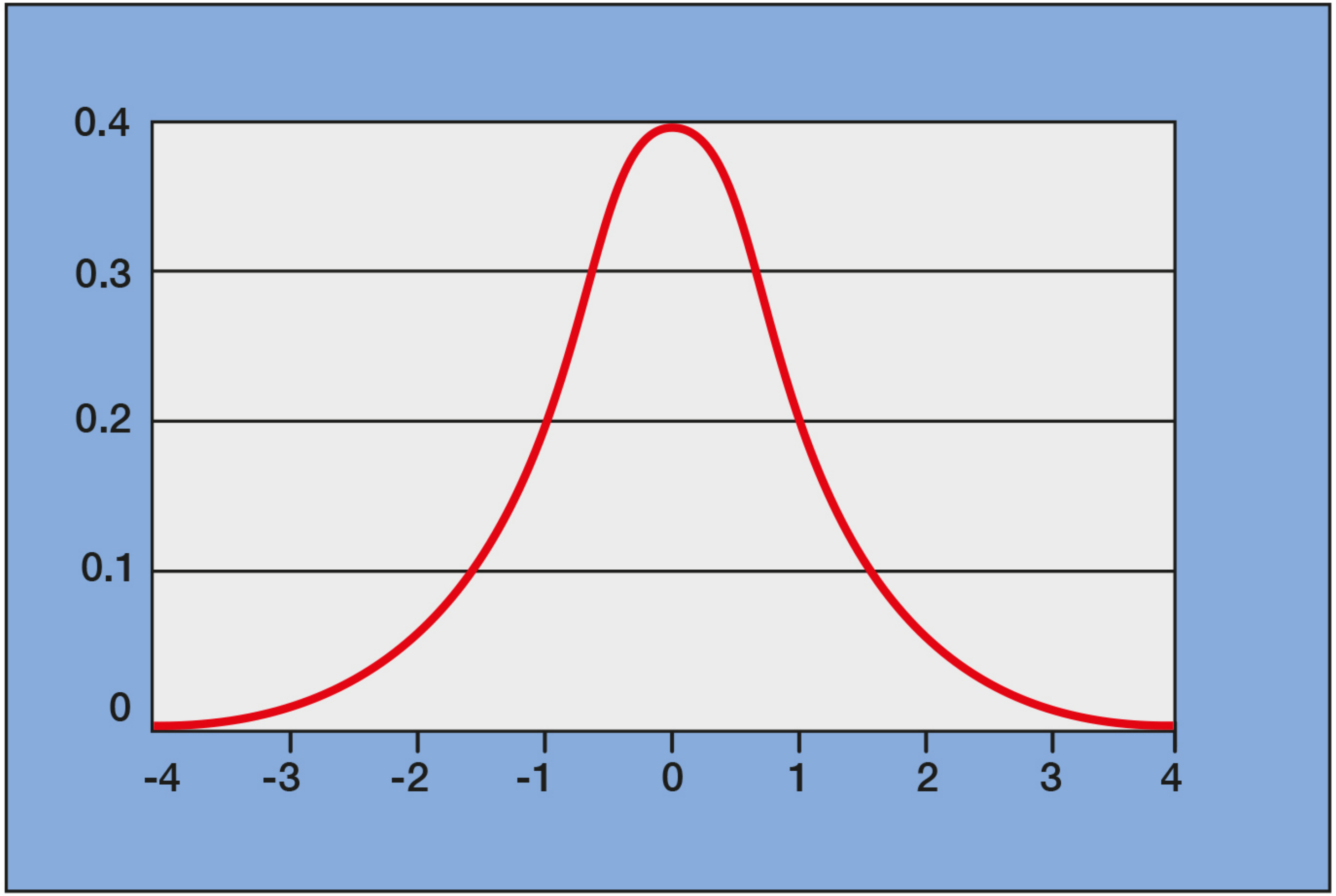
4- Candidate outliers on a single dimension?

5.4 Analysis of the distribution of the input variables

Tests used to verify the normality (Gaussion distribution) of a data variable

- Standard deviations
- Skewness and kurtosis
- Normal distribution and statistical outliers
- Grubbs test
- Kolmogorov-Smirnov test
- Shapiro-Wilk normality test

Example of a Gaussian distribution

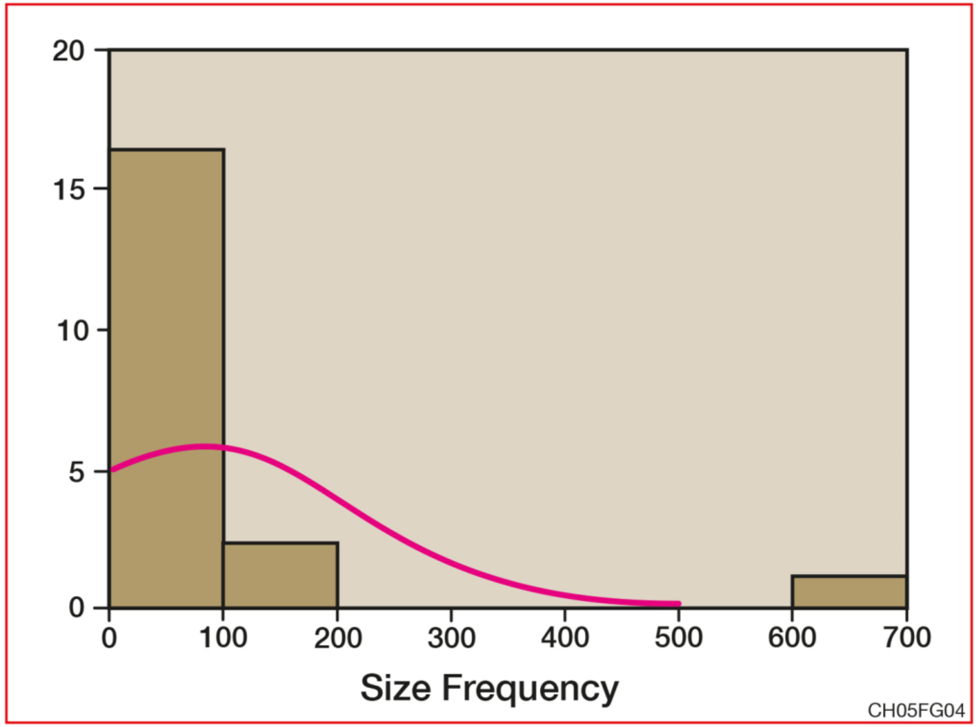


Identification of outliers

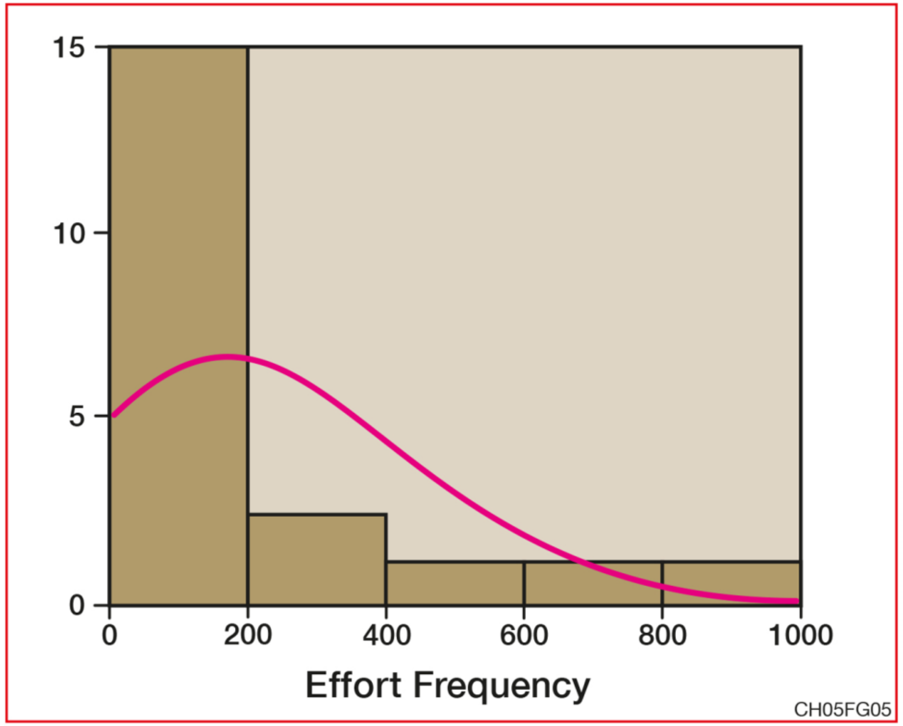
- Outliers: Values of the data that are significantly far from the average of the population of the dataset.
- Candidate outliers:
  - Typically at least 1 or 2 orders of magnitude larger than a data point closer to it.
    - Can be identified from a graphical representation.

Frequency distribution

Size (Independent variable); N=212



Effort (Dependent variable); N=21



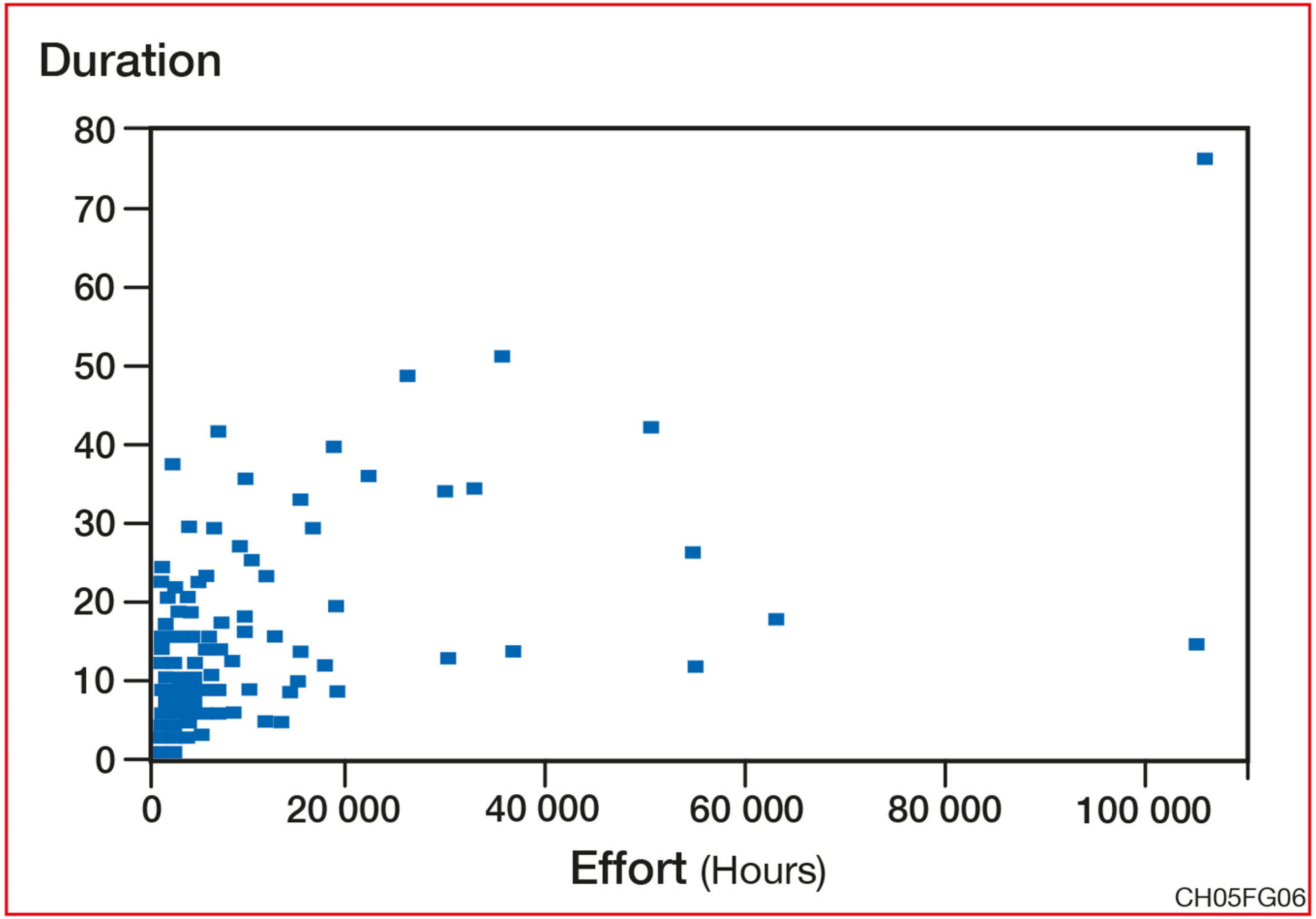
Statistical tests to identify outliers

- The Grubbs test, also referred to as the ESD method (Extreme Studentized Deviate).
  - The studentized values measure how many standard deviations each value is from the sample mean.
    - When the P-value for the Grubbs test is less than 0.05, that value is a significant outlier at the 5.0% significance level.
    - Values with a modified Z score greater than 3.5 in absolute value may well be outliers.
  - http://www.graphpad.com/quickcalcs/Grubbs1.cfm

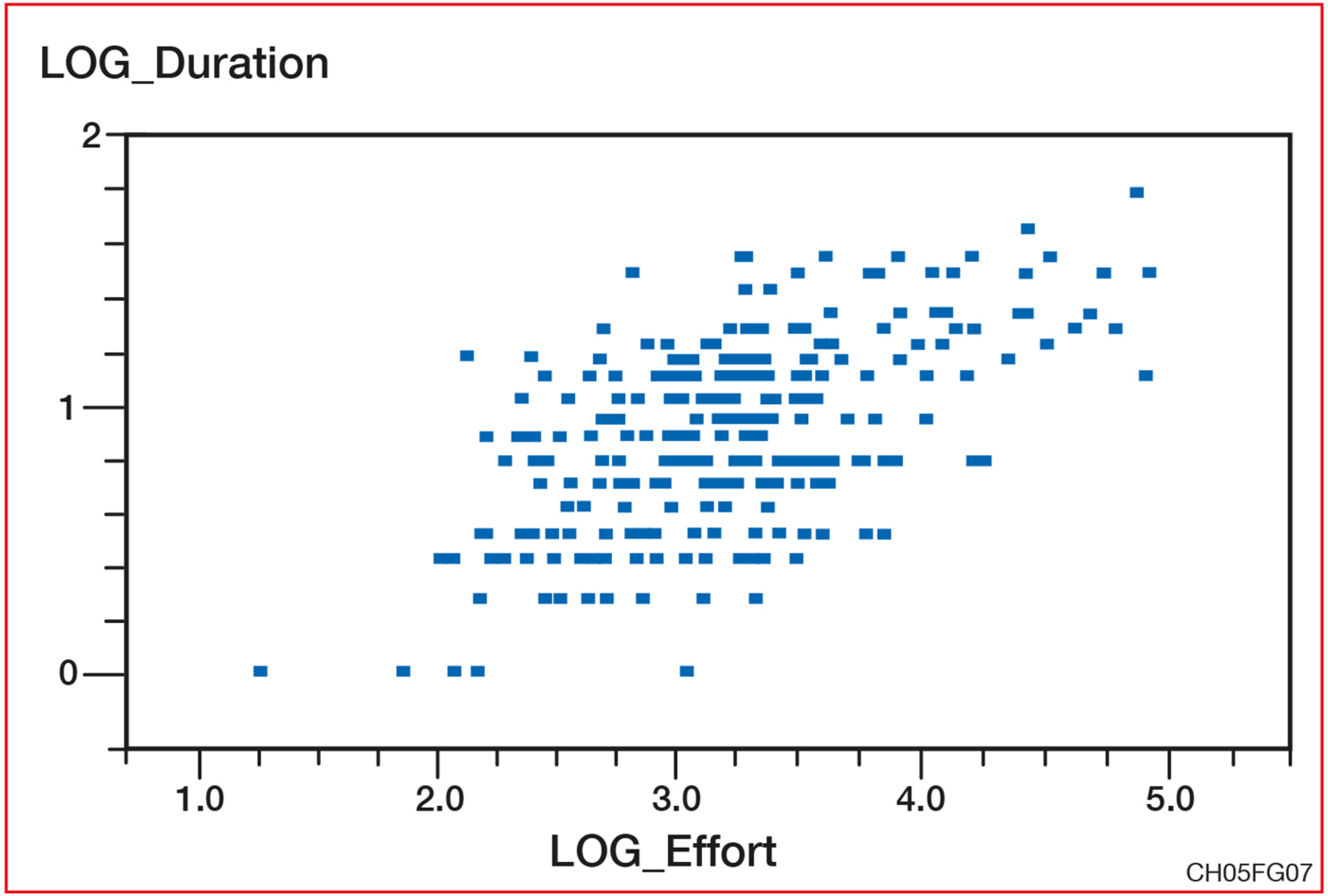
Log transformations

- Used when variables are not normally distributed
  - Weak support for linear regression → mathematical transformations
- The log transform is often used to obtain normal distributions for either the size or the effort variable, or both.

Scatter plot (n=312)



Log transformation



5.5 Graphical analysis – two dimensional

Graphical analysis

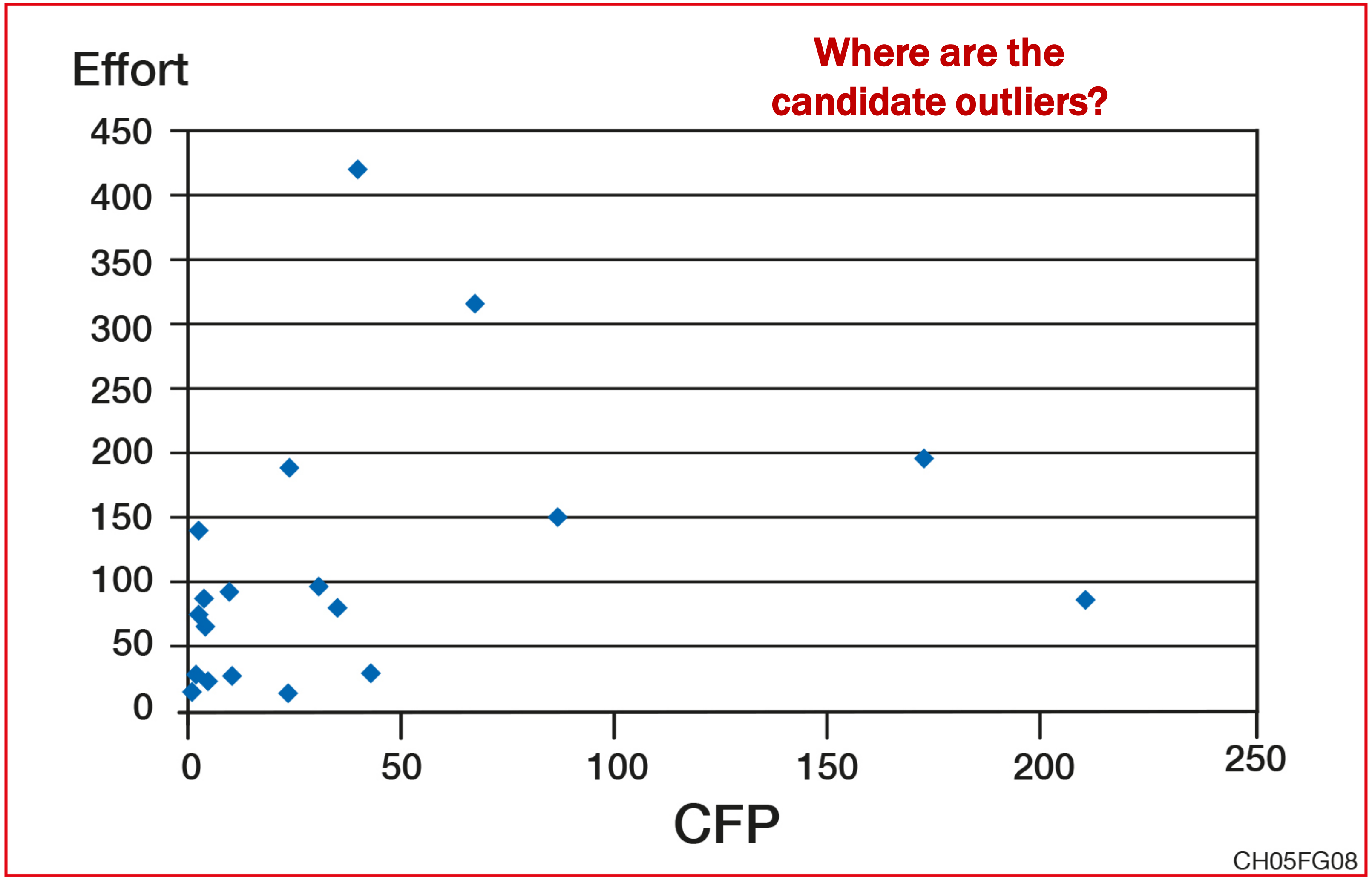
Multi-dimensional graphical analysis will typically provide an intuitive feel for the relationships between the :

dependent variable

&

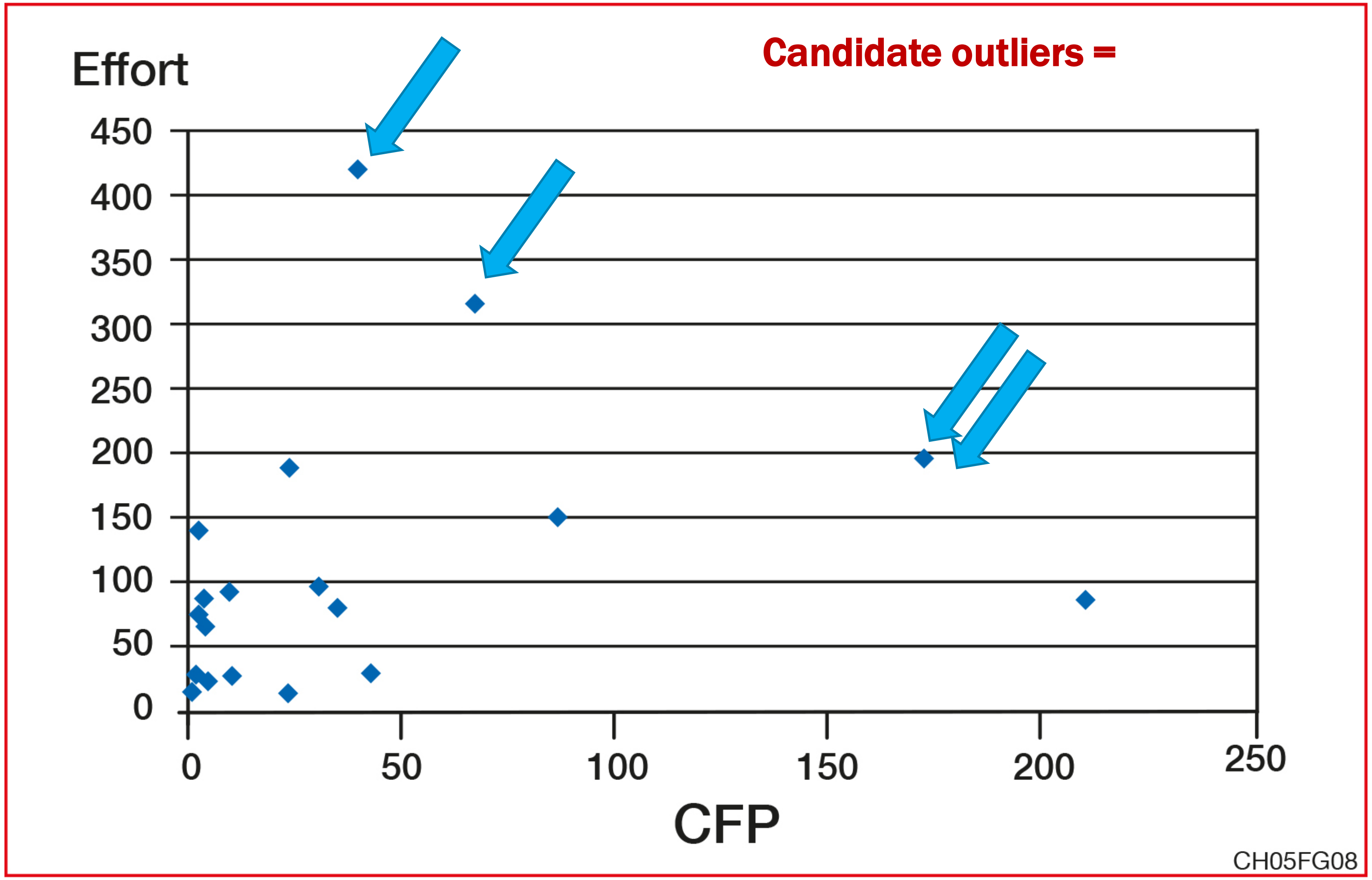
the independent variable.

Dataset including outliers

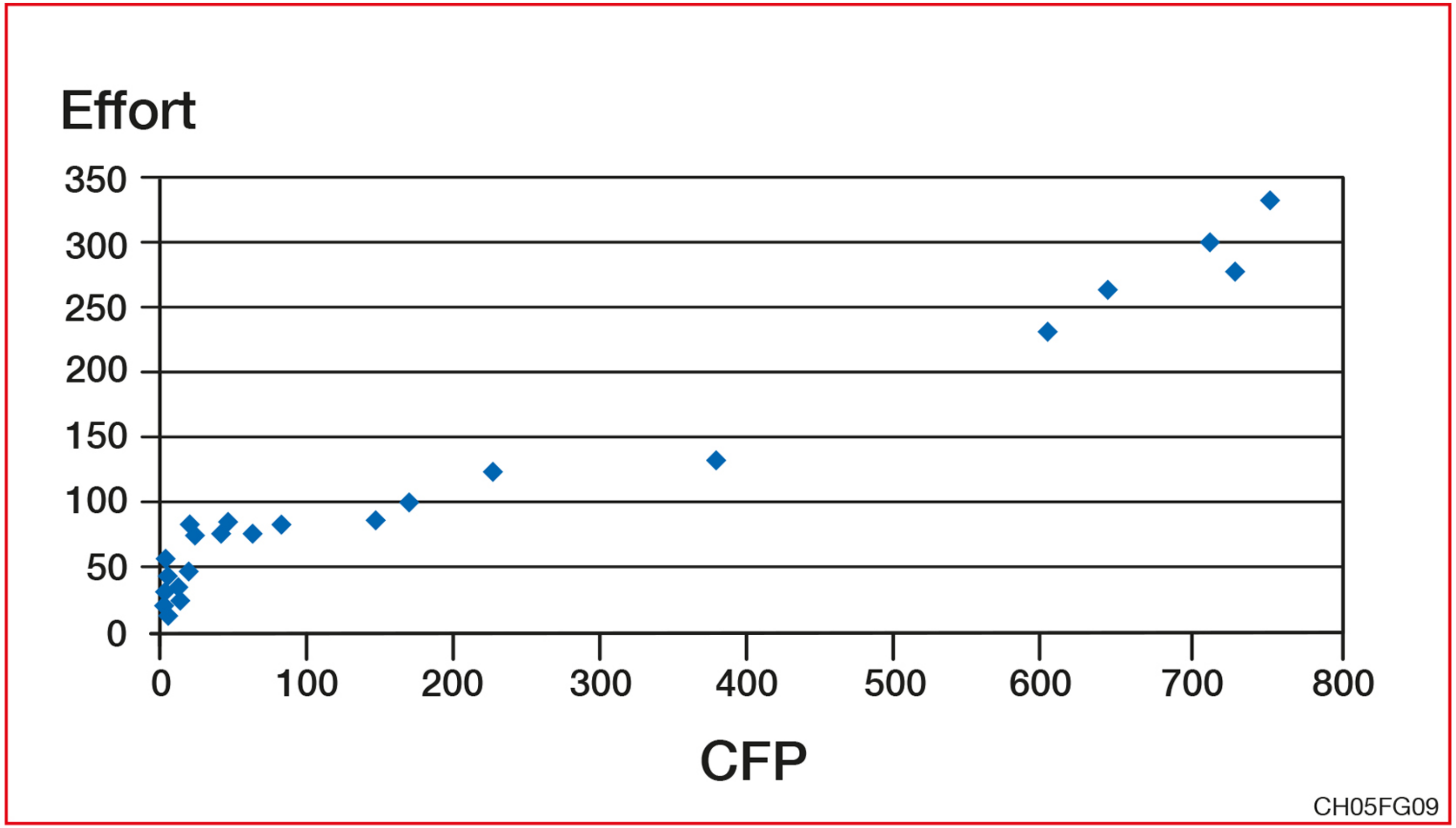


Where are the candidate outliers?

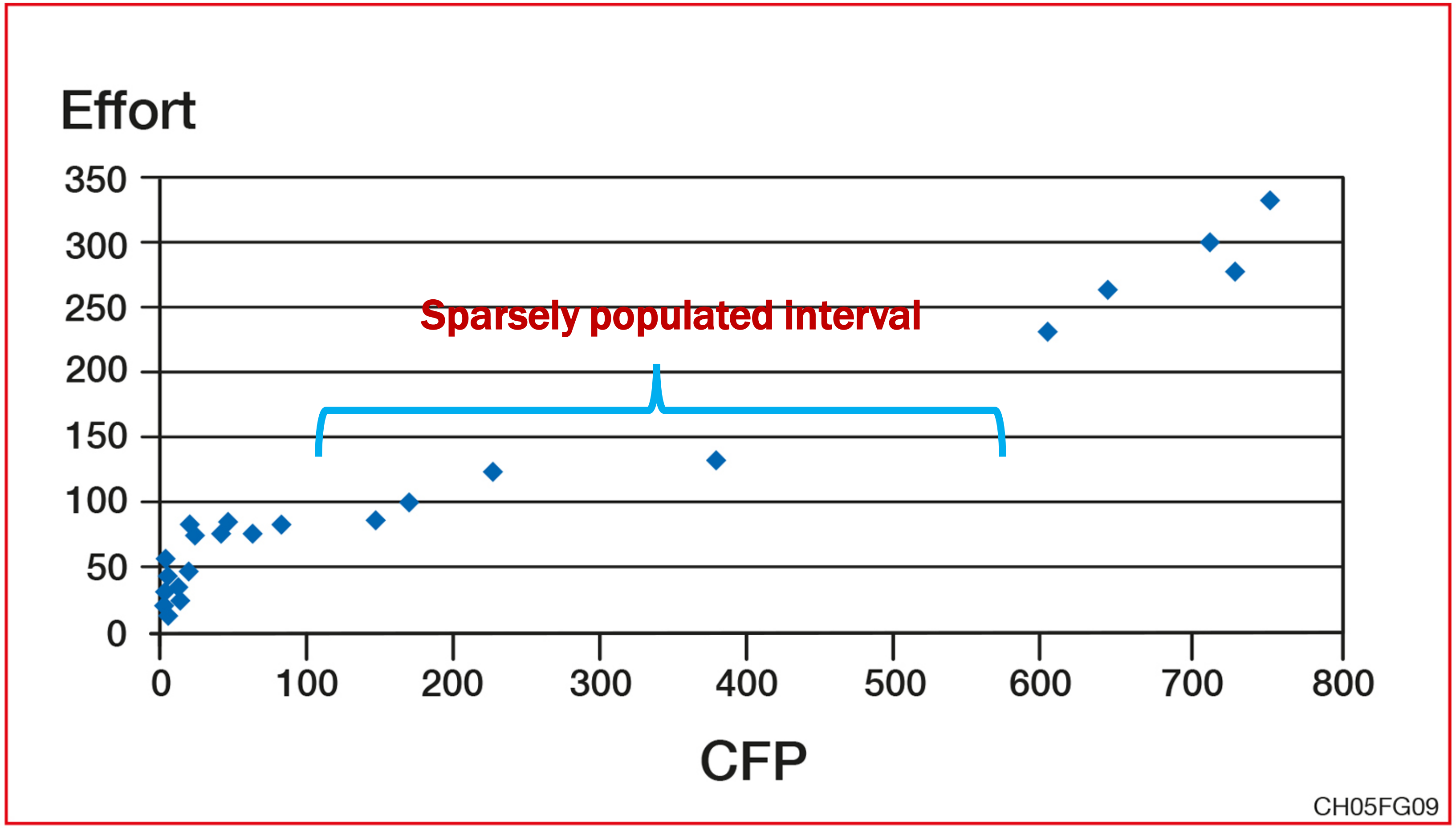
Dataset including outliers



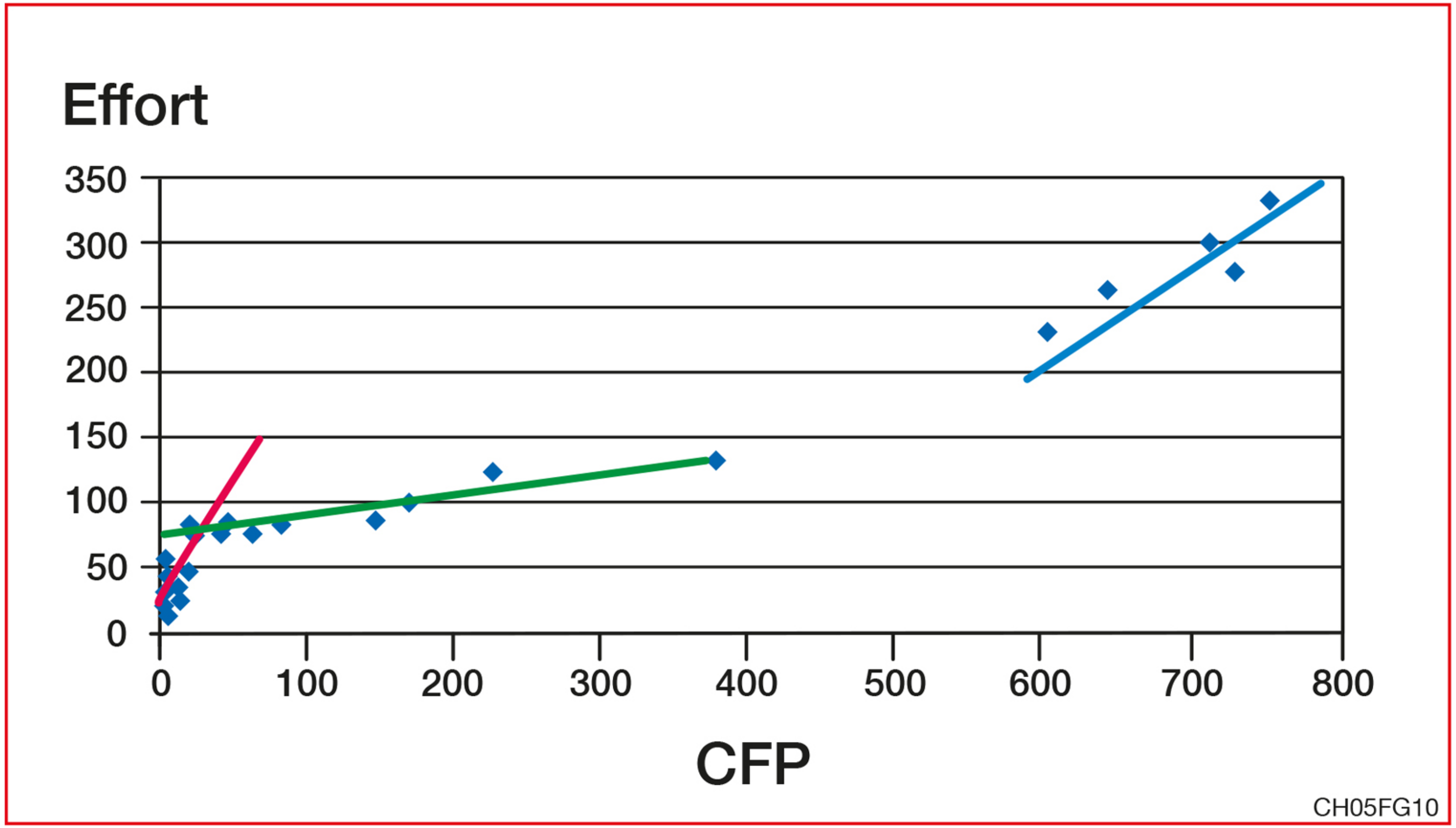
Dataset with sparsely populated size intervals



Dataset with sparsely populated size intervals



Dataset with sparsely populated size intervals



5.6 Size inputs derived from a conversion formula

Conversions

LOC → FP
FP → LOC

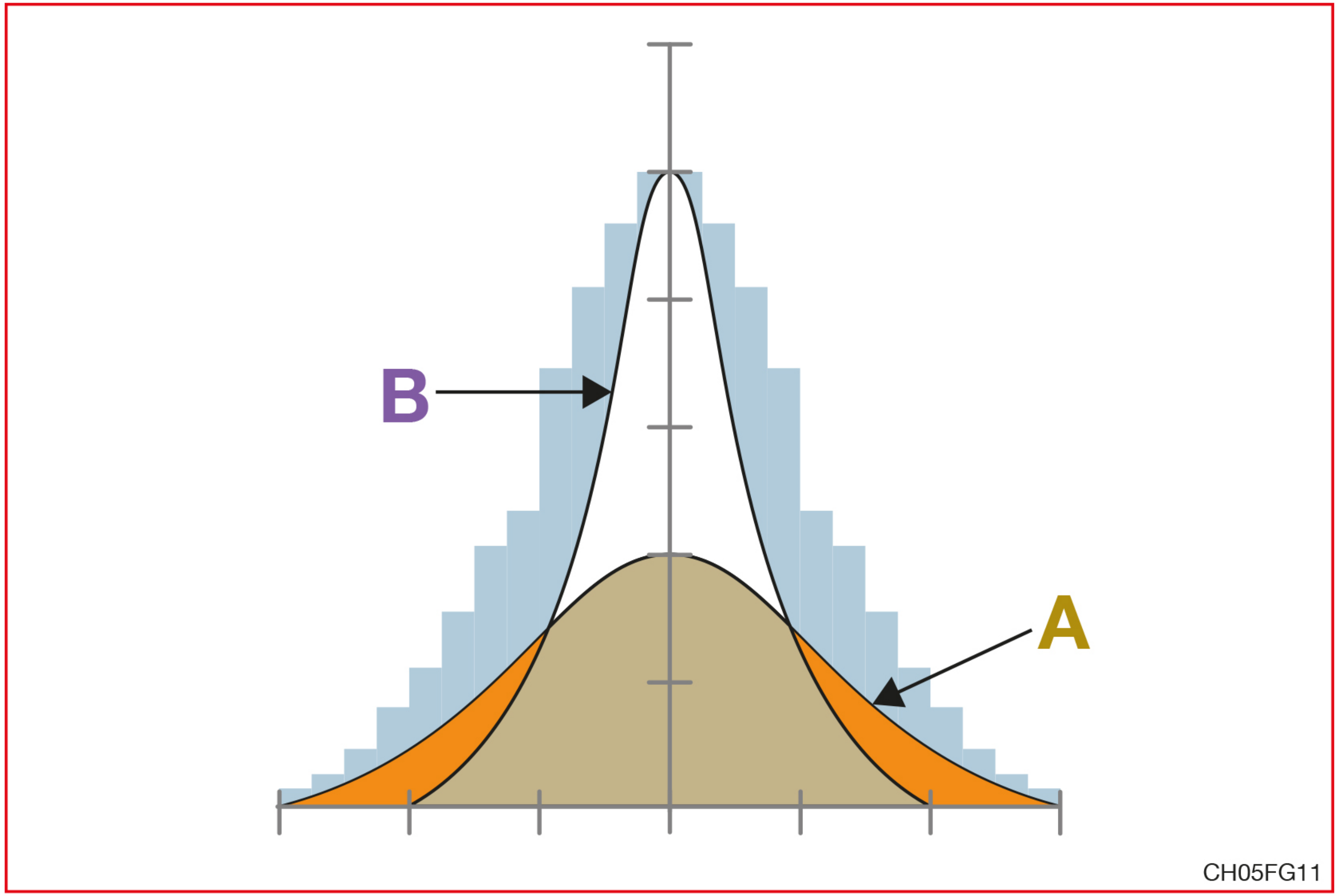
BE CAREFUL:
These FP-LOC conversion ratios from Lines of code (in any programming language) are taken from ‘averages’ from:

- Unknown sample size
- Unknown variability across these average
- No documentation on these samples and data points

FP-LOC conversion ratios do not reduce uncertainty – they may as well increase it!

RISK: Do not use any average for decision making without a reasonable knowledge of the sample from which the average is derived.

The kurtosis in a normal distribution



Exercises

1. When you are building a productivity model, why do you need to verify each input to your model?
2. Can models and techniques recognize bad input data?
3. Give five examples of bad or poor data input to productivity models.
4. Give two examples of independent and dependent variables used as inputs to productivity models.
5. Why is the scale type of an input variable important in a model?
6. Give an example of the misuse of a scale type in an input variable.
7. What mathematical operations can you perform with the CMMi® maturity level numbers?
8. Productivity models are built with numbers and they produce numbers. Under what conditions can you add and multiply numbers in a software productivity model?
9. Why is it important to use international standards for the measurement of input variables for estimation purposes?

Exercises

10. List some of the verification procedures that can be implemented to improve the quality of the inputs to a productivity model?
11. How do you verify that the data points for a variable have a normal distribution?
12. Why is a normal distribution important for models?
13. What is a statistical outlier in a dataset?
14. How do you identify a statistical outlier in a dataset?
15. Discuss the perils for practitioners of using models built on the basis of log-transforms.
16. Are there outliers in the dataset in Table 5.1? Use both a graphical analysis and some statistical tests to support your answer.
17. What are the necessary conditions for the use of the LOC-Function Point conversion ratios by programming languages?

Term Assignments

1. Measure the functional size of the software project you are currently working on. What is the most probable error range in your measurement result? Comment.

2. Look at the documentation associated with the estimate preparation of 3 recent projects in your organization. Comment on the quality of the inputs to the estimation model used (including the inputs for estimating based on expert judgment). What lessons can be learned from this?

3. Carry out a literature review of software estimation models. Of those that you have come across, which use measurement units recognized as international standards? When the estimation models do not use standards for their measurement units, what is the impact in practice?

Term assignments

4. Access three estimation models available on the Web. Document the basis for experimentation provided by the designers of each of these publicly available estimation models. To what point can you trust them? Explain your point of view to your management and customers. Look back at the reasons supporting that point of view and classify them as: engineering-based or opinion-based.

5. Access some of the Web information on conversion factors from Lines of Code to Function Points (or vice versa). What is their documented quality? How much more uncertainty is introduced into your estimation model if you use any of these conversion factors?