Measurements and Statistics

Learning goals: Improved ability to assess the validity of software development-related measures (construct validity) and use of statistical methods.

Supporting texts:

www.moffitt.org/moffittapps/ccj/v4n5/article4.html

Software quality measurement, M. Jørgensen, Advances in Engineering Software 30(12):907-912, 1999.

Introduction to Measurement Theory

- ***When you can measure what you are speaking about, and express it in numbers, you know something about it; but when you cannot measure it, when you cannot express it in numbers, your knowledge of it is of a meager and unsatisfactory kind; it may be the beginning of knowledge, but you have scarcely, in your thoughts, advanced it to the stage of science. (Kelvin)***

- BUT, if you don't know much about it, it is not meaningful to measure it (and learn from the measurements)! So, here we have a problem.

Exercise 1: Which of the above two statements are more correct? If both are correct, how is measurement possible? What does this tell us about the nature of measurement?

Exercise 2: Why do we easily accept some measures (like the measure of length in meters), while others not (like the measure of intelligence thorough IQ-tests)?

Measurement Theory

Def. Empirical Relational System: <E,{R_1..R_n}>, where E is a set of entities and R_1..R_n the set of empirical relations defined on E with respect to a given attribute.

Def. Formal (numerical) Relational System: <N,{S_1..S_n}>, where N is a set of numerals or symbols, and S_1..S_n the set of numerical relations defined on N.

Def. Measure: M is a measure for <E,{R_1..R_n}> with respect to a given attribute iff:

1. M: E -> N

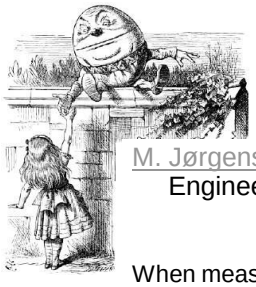
2. $R_i(e_1, e_2, ... e_k) <=> S_i(M(e_1), M(e_2), ... M(e_k))$, for all i.

So, what does complex formalism really mean?

Illustration 1: Why is «meter» a meaningful measure of the height of a person?

- We have an “empirical relational system”.
  - There exists a commonly accepted understanding of the meaning of «height of a person» and of height-relations and operations, such as «person A is taller than person B».
- We have a “formal relational system”.
  - Numbers, relationships, logic, ...
- We have a mapping (function) that connect “height” and numbers so that all relationships in the “real world” are present in the “formal world”, AND, all relationships in the “formal world” are present in the “real world”.
  - For example *(A,B,C,D are persons) and h our measure of height:*
    - A is taller than B in the real world => h(A) = 1.92 meter > h(B) = 1.80 meter
    - h(C) = 1.88 meter > h(D) = 1.87 meter => C is taller than D in the real world.
- In addition, have acceptable methods for the measurement process!

Illustration 2: Measurement of software quality



M. Jørgensen. **Software quality measurement**, Advances in Engineering Software 30(12):907-912, 1999.

When measuring complex phenomena like software quality we frequently have to choose between two evils:

- Use of a definition of software quality close to people's intuition of what software quality is (e.g., "how well software meet the software development stakeholders needs"), which is good for communication purposes, but impossible to measure.
- Use of a definition that enables measurement of software quality (e.g., "errors per lines of code"), but only partly connected to the way the term software quality is used.

Exercise

- Assume that:
  - The management of an organization wants to know whether the process changes have had a positive effect on software maintainability (one possible aspect of software quality) or not.
  - Your are the unfortunate person in charge of measurement of this!
- How would you proceed?

Elementary Statistics

Distributions + Central values

- An essential concept of statistical hypothesis testing is “distribution”.
  - A distribution depict possible outcomes and their likelihood or frequency
  - The height of people is, for example, close to normally distributed
  - The salaries of people have a long tail towards high values and is not normally distributed.
  - The grading of students is meant to be normally distributed.
- Distributions are, among other values, described by their central value and spread.
- Central value examples: Mode (most typical), median (50% probable to exceed), arithmetic mean.
- When evaluating studies:
  - What do we know about the underlying distribution?
  - Is the arithmetic mean likely to be misleading. Should the more outlier robust median be used instead?

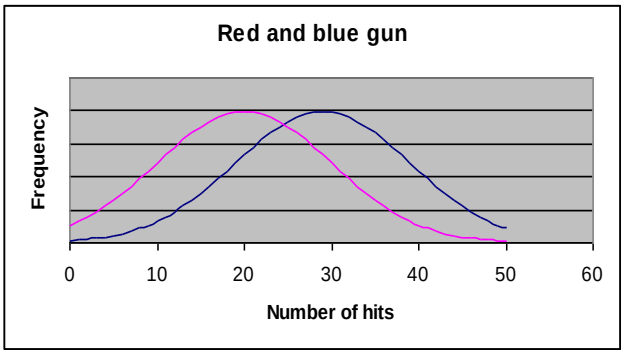
Spread

- **Variance** = $\Sigma (x_i - x_{a.middel})^2/(n-1)$, for i=1..n
- **Standard deviation** = $\sqrt{}$(Variance)
- **Standard error (of the mean)** = standard deviation / $\sqrt{}$(n)
- If we can assume that a distribution is close to a predefined one (e.g., the normal distribution) and that the sample is randomly drawn, we know something the spread may provide us with useful information about the population.
  - Example: Assume a normal distribution of Norwegian 18-year old men and that we have randomly sampled 100 men of that age. We measure a mean height of 177 cm and a standard deviation of 15 cm. Then, we are able to induce that about 66% of Norwegian men of that age is in the interval [177-15 cm; 177+15 cm] = [162 cm; 192 cm]. This interval is the +/- one standard deviation prediction interval.
- Measures of spread are essential tools when statistically testing hypotheses.

Hypothesis testing

- I will not try learning you the, in many aspects, difficult process of statistical hypothesis testing, but instead make you understand the underlying principles and how to evaluate studies based on it.
- Frequently the focus is on the number of observations. The statistical hypothesis testing instruments deals with the number of observations that properly – remember how standard error of mean is defined.
- It is much more important to investigate the samples (external validity), the treatment allocation process (internal validity requires in many cases random allocation) and the measures (construct validity) involved.

Example – Hypothesis testing



Some soldiers are testing the new guns "Red" and "Blue". Assume that the x-axis shows the number of hits of 50 for those guns and the y-axis the frequency of each each number of hits. The distributions show that "Red" and "Blue" has the a mean hit rate of 20 and 29 hits, respectively. Both gun types has a standard deviation of 20. It looks like "Blue" is the best gun, but how sure can we be? How should we evaluate the internal validity of the result?

Example – Hypothesis testing

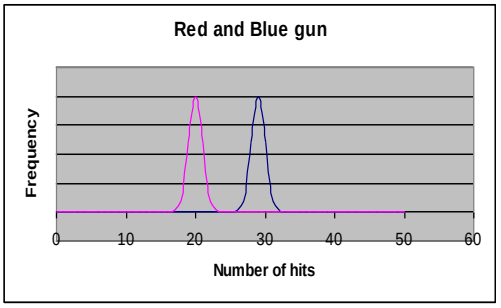
- To test whether "Blue" is better than "Red" we need (in accordance with classic, statistical hypothesis testing) to:
  - Choose a level of significance (alpha). This level of significance says something about how sure we need to be to accept the result that "Blue" is in fact better. This is frequently difficult to decided and, for some really, really strange reason, nearly all researchers end up with the selection of level of significance of 90%, 95% or 99% (corresponds to $p<0.1$, $p<0.05$ and $p<0.01$)
  - Know the number of observations (n).
  - Know the standard deviation of both distributions.
  - Know the sampling process. If for example the soldiers selected the guns themselves, the randomness criteria is violated and the statistical tests cannot be used. Similarly, if there were differences in weather conditions when testing the different gun types, the tests are not meaningful as tests of the guns either.
  - Know the underlying distribution, e.g., whether a normal distribution can be assumed.

Example – Hypothesis testing

n	$t_{0.05}$	95% Interval Red	95% Interval Blue
5	2,13	[1;39]	[10; 48]
10	1,83	[8; 32]	[17; 41]
50	1,68	[15; 25]	[23; 35]
100	1,66	[17; 23]	[26; 32]

- ➔ n = number of observations
- ➔ $\mathbf{t_{0.05}}$ = the value (found in a statistical table) used to calculate the 95% confidence intervals
- ➔ 95% confidence interval Red = (simplified) the interval that with 95% confidence include the actual **mean** value of the Red gun's hit rate.
- ➔ With see from the above table that with 5, 10 and 50 observations the 95% confidence intervals are overlapping. This means that it is more than 5% likely that the difference is due to sampling variance (and not reflecting an actual difference in mean values). Only with 100 observations the difference is significant. To be significant
- ➔ **Formula for the intervals**: mean +/- $\mathbf{t_{0.05*}}$ (standard deviation / √n)

Example – Hypothesis testing



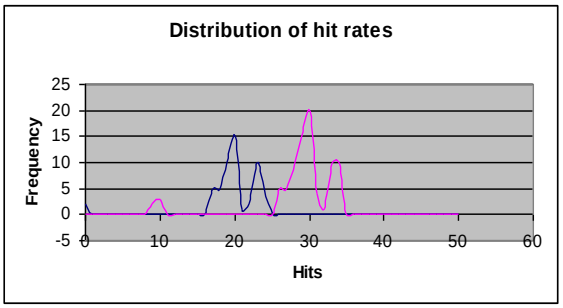
A reduction of the standard deviation from 20 to 2 has a high impact on the number of observations needed to enable 95% confidence. Reduction of variance can, for example, be achieved through more similar shooting skill among the soldiers.

n	$t_{0.05}$	95% Interval Red	95% Interval Blue
5	2,13	[18;22]	[27; 31]
10	1,83	[19; 21]	[28; 30]
40	1,68	[19; 21]	[28; 30]
100	1,66	[20; 20]	[29; 29]

NB: Notice that in this case only 5 observations are more than sufficient for a high degree of confidence in this case.

Example – Hypothesis testing

It is easy to deceive with statistics and the use of statistical hypothesis testing increases rather decreases this problem. To enable the readers to evaluate the robustness of the finding, the observed distribution should be displayed.
Exercise: Is there essential information that would not be revealed in a statistical test if the observation was distributed as shown below?



Correlation and regression analysis

- **Correlation** shows the degree of linear co-variation of variables. Not cause-effect, and not non-linear relationships.
- **Linear regression**: The straight line running among the points of a scatter diagram about which the amount of scatter is smallest, as defined, for example, by the least squares method.
  - Why is the method typically based on minimizing the square?
  - What are there consequences from unusual values (outliers)? Large unusual values, small unusual values?
  - What will happen with relationships that are non-linear?