Advanced Software Engineering

Data Analysis (HypothesisTesting)

Aim

Know the basics of analysis of data from empirical studies

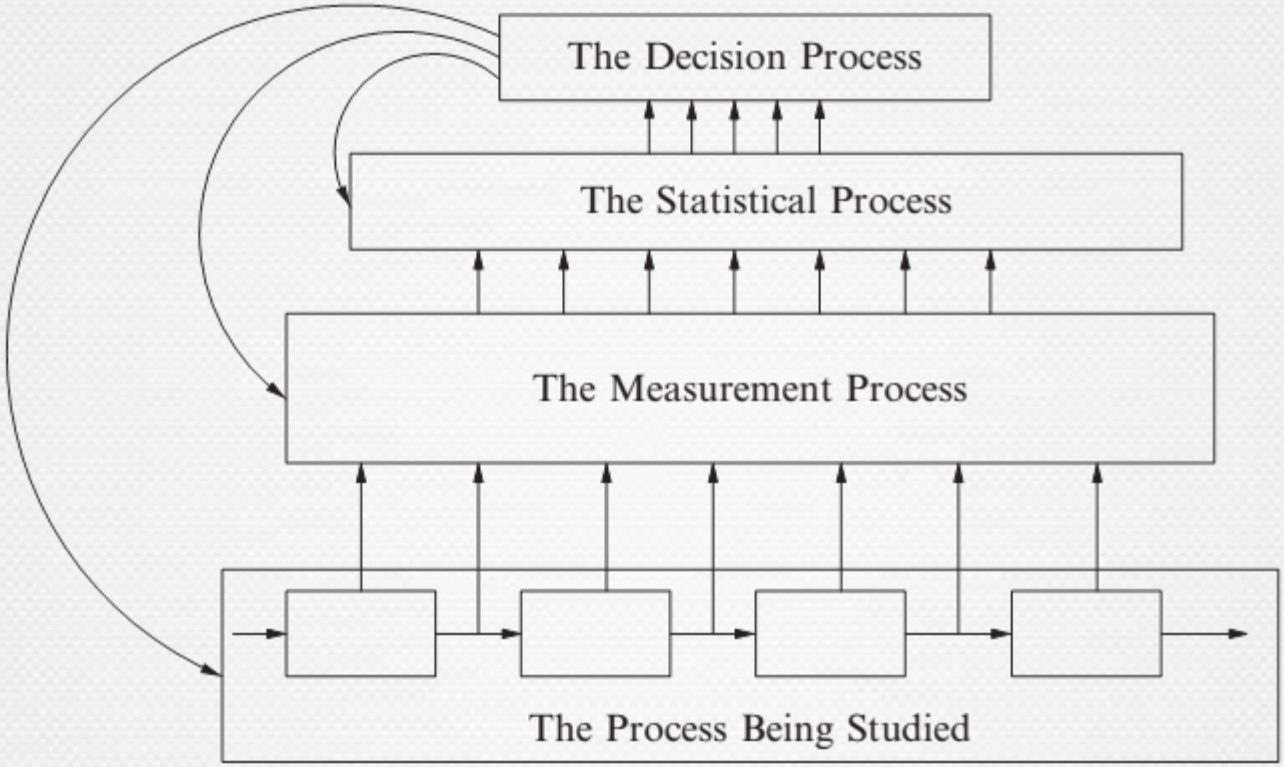
Know how to use statistical methods to test SE Hypotheses

Ref: Shull – Chapter 5 – Rosenberg, J., Statistical Methods and Measurement

Voelz, V.A., Hypothesis Testing

Key terms/Ideas

- Measurement = process of assigning labels (typically numbers) to an attribute of an object or action where the label (e.g. some number) tells us about a characteristic of the attribute
  - Result is a called a measurement scale or metric
  - Metrics are for interpretation (obviously)
    - If scale is wrong interpretation will be wrong.
- "Statistics" – term comes from collection of demographic and economic information important to the government e.g. size of the population, the birth rate, annual crop yield – all used for decision making
- Descriptive statistics applies to such measures, whether simple or complex, that describe some variable quantity of interest.
- Inferential statistics – to allow conclusions to be drawn from the comparison of the observed values of descriptive statistics to other real or hypothesized values.
  - require some assumptions in order to work, and much of
  - statistical theory is devoted to making those assumptions as flexible as possible in order to fit real-world situations


The Decision Process
The Statistical Process
The Measurement Process
The Process Being Studied

Context of Measurement

- Measurements from Observation (e.g. via a case study)
  - Observation: hard to sample = possible bias from observer
  - Experiments = precise measurements under controlled conditions, *but* very specific i.e. lack of generalisability
- "Leading" or "Lagging" indicators
  - Leading – good for forecasting
    - Cause now but effect later
    - Measure of quality e.g. cohesion metric
  - Lagging - more common and appear after the event
    - Cause was long ago, effect now
    - E.g. customer-reported software defects
    - E.g. customer satisfaction measure could be years after company's products and services typically by several years
- good measurements are actionable
  - You can do something with them

Types of Metric

- Simple metrics
  - include counts (e.g., LOC, support cost); categories (e.g., problem types), and rankings(e.g., problem severity).
- Compound
  - in terms of two or more metrics
  - Ratios (e.g. defects per KLOC; defects per thousand units),
  - Rates (time-based ratios e.g. number of problem reports per month),
  - Proportions or percentages(e.g., proportion of customers responding “very satisfied” to a survey question),
  - linear algebraic - (e.g., mean repair cost
  - Indices (dimensionless measures typically based on a sum and then standardized to some baseline value)

Measurement Theory

- Nominal = unordered categories, and no mathematical manipulation makes sense.
- Ordinal. The scale values are ordered, but the intervals between the values are not necessarily of the same size, so only order-preserving manipulations such as ranking make sense.
- Interval. The scale values are ordered and have equal intervals, but there is no zero point, so only sums and differences make sense.
- Ratio. The scale values are ordered and have equal intervals with a zero point, so any mathematical manipulation makes sense.

Using Measurement Theory

- Nominal categories - makes no sense to calculate the mean e.g. mean of a set of ID numbers!
- Subjective rating scale values (such as defect severity as VH/H/M/L/VL) – mean of limited use since difference between categories isn't fixed
- Note: serial numbers ordered chronologically ordered can be treated as an ordinal, rather than nominal
- Precision should be appropriate e.g. significant digits

Measurement Reliability

- Reliability: Any measurements must be consistent across repeated observations in the same circumstances.
  - Easy with physical measurements
  - Hard if *any* subjectivity e.g. Rating scales
  - Assess with Cronbach's coefficient alpha
    - Measure of correlation among repeated measurements;
- E.g.
  - 5 point Likert scale to answer "Does the tool improve efficiency?"
  - 6 questions given – how much do they agree/ are they consistent?

- 1A: The new tool improves quality of code
- 1B: The code from the new tool is better
- 1C: Code produced using the new tool has fewer defects and meets requirements better
- 1D: The new tool produces a better product
- 1E: The new tool improves the old approach
- 1F: The new tool reduces errors

- 1= “Strongly Disagree” 2=“Disagree”3=“Undecided” 4=“Agree”, 5=“Strongly Agree.”

- k = number of items (or questions in this case)
- s_i^2 = the variance associated with item i,
- s_t^2= variance associated with the total (or sum) of all k item scores.
- Roughly a ratio of variances
  - 1 is perfect

$$\alpha = \frac{k}{k-1}\left(1 - \frac{1}{s_{\mathrm{T}}^2}\sum_{i=1}^{k} s_i^2\right)$$

Cronbach's alpha	Internal consistency
α ≥ 0.9	Excellent
0.7 ≤ α < 0.9	Good
0.6 ≤ α < 0.7	Acceptable
0.5 ≤ α < 0.6	Poor
α < 0.5	Unacceptable

- s_1^2=VAR.S(E2:E16)

$$\alpha = \frac{K}{K-1}\left(1 - \frac{\sum_{i=1}^{K} \sigma_{Y_i}^2}{\sigma_X^2}\right)$$

- $\alpha = \frac{6}{6-1}\left(1 - \frac{.2667}{8.2095}\right)$
- =0.7114

Respondent	1A	1B	1C	1D	1E	1F	
1	5	5	5	5	5	5	30
2	5	5	5	5	5	5	30
3	5	4	5	4	4	5	27
4	4	5	4	5	5	5	28
5	5	5	5	5	5	4	29
6	4	4	3	4	5	5	25
7	5	5	5	5	5	5	30
8	4	5	4	4	4	5	26
9	5	5	4	5	5	5	29
10	4	2	3	5	2	4	20
11	4	4	4	4	5	5	26
12	4	4	1	4	5	5	23
13	4	4	5	5	5	5	28
14	5	5	5	4	5	5	29
15	5	5	5	5	4	5	29
VAR.S	0.2667	0.6952	1.3143	0.2571	0.6857	0.1238	3.3429
						VAR.s	8.2095238

- >7 =Good internal consistency

Validity of a Metric

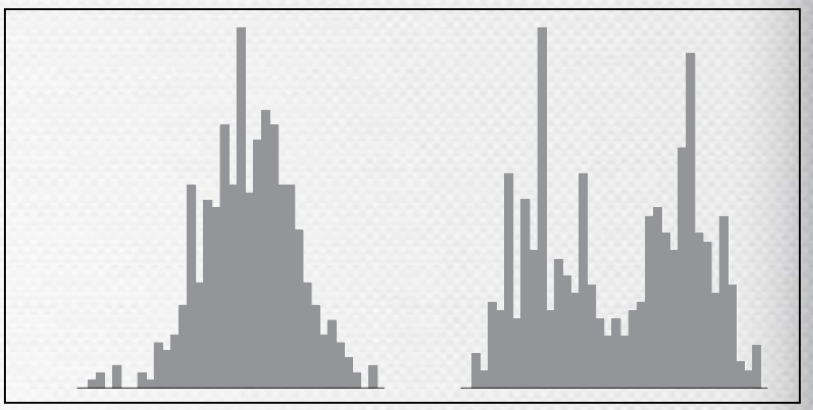
- NB: different from validity of an experiment
- **Content validity** is the degree to which the metric reflects the domain it is intended to measure. E.g. Measuring coupling to assess test coverage would make no sense.
- **Criterion validity** is the degree to which a metric reflects the measured object's relationship to some criterion. For example, a complexity metric should assign high values to programs which are known to be highly complex. (aka discrimination validity)
- **Construct validity** is the degree to which a metric actually measures the conceptual entity of interest. E.g. IQ via intelligence tests

Statistical Analysis

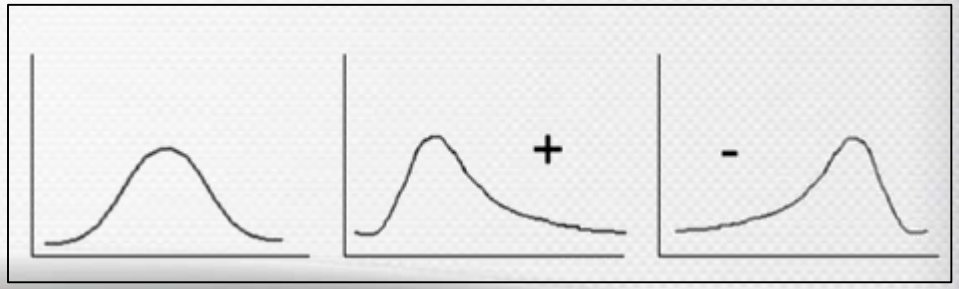
- Need to understand models of the underlying data-generating process
  - Parametric models assume specific functional forms e.g. a Normal Distribution
  - Knowing the model makes the analysis possible via assumptions
    - E.g. use student t test on a Normal distribution to test for a difference between two means.
  - Non-parametric models make few assumptions
  - How do we know – there are tests!

Looking at the data

- Mean and standard deviation don't tell the whole story
  - E.g. same mean and σ could be very different
- Also if distribution is skewed
  - Mean and Median very different
- Precision of an estimate
  - measured via standard error (std dev)



- Usually 2 x standard errors either side of the estimate (e.g. mean) = 95% confidence interval
- i.e. 95% of the time the true value of the parameter being estimated will fall in that interval
- Wide interval = low precision



Common Statistics

- Central Tendency
  - Mode – most frequent value
  - Median – 50th percentile/ mid point
  - Arithmetic Mean – most commonly used (nb affected by extreme values = not good for skewed distributions
  - Geometric mean (the nth root of the product of the n data values) more accurate for observed rates e.g. errors per day.

Common Statistics

- Dispersion
  - Range = difference between the highest and lowest values
    - Doesn't tell you much!
  - Variance
    - When using the mean as the statistic
      - (square root of the) sum of the squared deviations about the mean, scaled by the sample size
    - When using the median as the statistic
      - the first and third quartiles (i.e., the 25th and 75th percentiles) or the semi-interquartile range
    - Coefficient of variation (CV) = standard deviation divided by the mean (measure of spread)
      - CV for cyclomatic complexity with the CV for module length (gives comparison of dispersion even though different units)

Common Statistics

- Association (aka correlation)
  - Various correlation measures are available
  - *r*, *(*or Pearson's *r*) = how strongly are 2 variables related (based on their covariance of the variables divided by the product of their (sample) standard deviations.
  - Excel =CORREL(A2:A11,B2:B11)
  - Result gives positive or negative number = positive or negative correlation. The bigger the number the more correlation
  - Also - Spearman's rank correlation coefficient e.g. used in comparing two rankings (ordinal data, non-parametric)
  - Kendall tau rank correlation coefficient(ordinal data, non-parametric)

Sort Field 1	Sort Field 2	Sort Field 3	Sort Field 4	Spearman coefficient	t statistic
Benefit	Cost	Risk Reduction Efficacy	Development Risk	0.76	5.18
Benefit	Cost	Development Risk	Risk Reduction Efficacy	0.81	6.02
Benefit	Risk Reduction Efficacy	Cost	Development Risk	0.78	5.40
Benefit	Risk Reduction Efficacy	Development Risk	Cost	0.79	5.63
Benefit	Development Risk	Cost	Risk Reduction Efficacy	0.81	5.97
Benefit	Development Risk	Risk Reduction Efficacy	Cost	0.79	5.68
Cost-Benefit	Risk Reduction Efficacy	Development Risk	-	0.33	1.53
Cost-Benefit	Development Risk	Risk Reduction Efficacy	-	0.06	0.27

example

- Correlation between Sorts and a Project Manager's Sort

Sort Field 1	Sort Field 2	Sort Field 3	Sort Field 4	Spearman coefficient	t statistic
Benefit	Cost	Risk Reduction Efficacy	Development Risk	0.76	5.18
Benefit	Cost	Development Risk	Risk Reduction Efficacy	0.81	6.02
Benefit	Risk Reduction Efficacy	Cost	Development Risk	0.78	5.40
Benefit	Risk Reduction Efficacy	Development Risk	Cost	0.79	5.63
Benefit	Development Risk	Cost	Risk Reduction Efficacy	0.81	5.97
Benefit	Development Risk	Risk Reduction Efficacy	Cost	0.79	5.68
Cost-Benefit	Risk Reduction Efficacy	Development Risk	-	0.33	1.53
Cost-Benefit	Development Risk	Risk Reduction Efficacy	-	0.06	0.27

Common Statistics

- Comparison
  - Model vs chance
    - i.e. null hypothesis versus what we would get randomly
    - basic method - compare the difference in the average values for two groups (treatments) with the amount of dispersion in the groups' values
      - E.g. a difference of 10 units is more significant if the two groups' values ranged from 30 to 40 than if they ranged from 300 to 400. (common sense)
        - » For 300-400, we would easily expect a 10-unit difference to appear in two successive samples drawn from exactly the same population by chance.
  - Errors
    - Type 1 (α) – incorrectly rejecting the null hypothesis i.e. saying there is a difference but there is not (False positive)
    - Type 2 (β) – incorrectly NOT rejecting the null hypothesis i.e. saying there is no difference but there is (False Negative)
    - Assess α & β as probabilities
    - $\alpha = 0.05$ is usually considered acceptable
    - $1-\beta$ = power of a test i.e. how good it is at detecting a difference if there is one

- Well known Comparison Tests
  - two-sample t-test for Normal Distributions
  - (one-way) analysis of variance (ANOVA), with its F-test - Normal Distributions
  - Wilcoxon/Mann-Whitney test (non-parametric)
    - Also good for ordinal (ranked) data
  - Kruskal-Wallis test(non-parametric)
    - Also good for ordinal (ranked) data
  - one-sample t-test – an observed mean versus hypothesised mean (i.e not a group mean)
  - Z-test – as in one-sample t-test but larger samples (>30).
- p-values
  - Above tests provide these
  - $p = P(x \geq x^{\dagger})$ = probability of observing a value of x that is at least x or larger
  - So decide on an acceptable α (usually 0.05) and if $p<0.05$ then we reject the null hypothesis

Testing a Hypothesis

1. pick an appropriate significance level, α, and then
2. calculate the *p*-value of your observed statistic.
3. If the *p*-value is less than α, then reject the null hypothesis.

Advanced Software Engineering

Data Analysis (HypothesisTesting)

Aim

Know the basics of analysis of data from empirical studies

Know how to use statistical methods to test SE Hypotheses

Ref: Shull – Chapter 5 – Rosenberg, J., Statistical Methods and Measurement

Voelz, V.A., Hypothesis Testing