Practical Statistics for Human-Computer Interaction

Xin Yi

HCI Lab
Tsinghua University

Content

We used repeated measures ANOVAs and paired two-tailed *t*-tests for our analyses. All *post hoc* pairwise comparisons following the ANOVAs were protected against Type I error using a Bonferroni adjustment. Reported fractional degrees of freedom (*df*s) are from Greenhouse-Geisser adjustments. When parametric tests were not appropriate because the data violated the assumption of normality, we applied nonparametric equivalents, such as the Wilcoxon signed-rank test. We report significant findings at $p < .05$.

Findlater, L., Wobbrock, J. O., & Wigdor, D. (2011, May). Typing on flat glass: examining ten-finger expert typing patterns on touch surfaces. In *Proc. CHI'11* (pp. 2453-2462). ACM.

We used repeated measures ANOVAs and paired two-tailed *t*-tests for our analyses. All *post hoc* pairwise comparisons following the ANOVAs were protected against Type I error using a Bonferroni adjustment. Reported fractional degrees of freedom (*df*s) are from Greenhouse-Geisser adjustments. When parametric tests were not appropriate because the data violated the assumption of normality, we applied nonparametric equivalents, such as the Wilcoxon signed-rank test. We report significant findings at $p < .05$.

Findlater, L., Wobbrock, J. O., & Wigdor, D. (2011, May). Typing on flat glass: examining ten-finger expert typing patterns on touch surfaces. In *Proc. CHI'11* (pp. 2453-2462). ACM.

- The mean distance in the visible keyboard condition is more than the 0.9" of space between visual key centers ($t_{19} = 5.30$, $p < .001$).
- There was a main effect of keyboard for both x- and y-directions (x-direction: $F_{1,19} = 10.77$, $p = .004$; y-direction: $F_{1,19} = 39.28$, $p < .001$).
- We examine the highest-order effect in detail: a three-way interaction of keyboard × finger × row ($F_{2.8,34.3} = 5.70$, $p = .002$).
- A Wilxocon signed-rank test was not significant: $z = 1.45$, $p = .147$.
- Pairwise comparisons showed the keys assigned to the little finger had significantly greater x-direction deviation than the ring ($p = .033$) and middle fingers ($p = .024$), while comparison to the index finger was only a trend ($p = .075$).

Findlater, L., Wobbrock, J. O., & Wigdor, D. (2011, May). Typing on flat glass: examining ten-finger expert typing patterns on touch surfaces. In *Proc. CHI'11* (pp. 2453-2462). ACM.

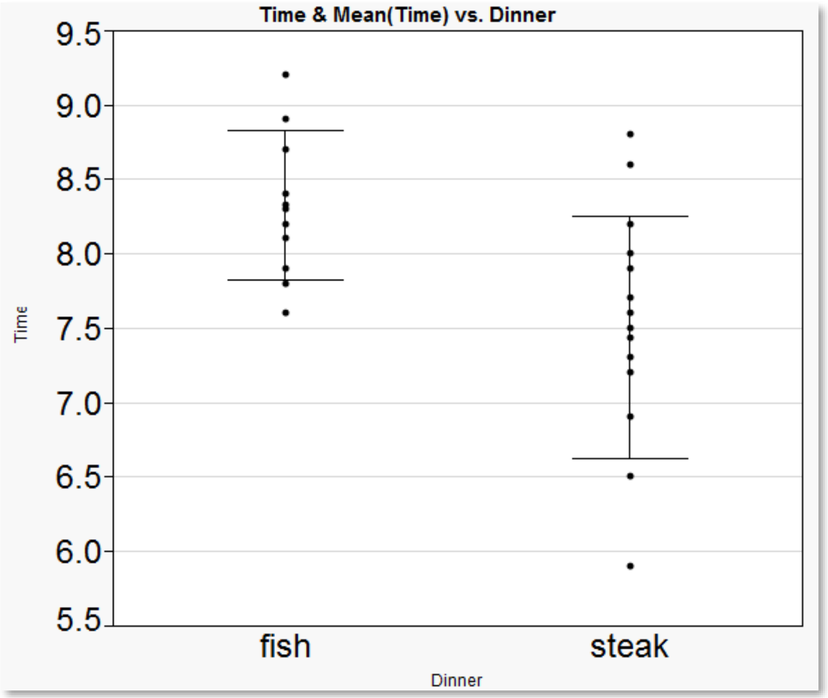
Statistics Overview

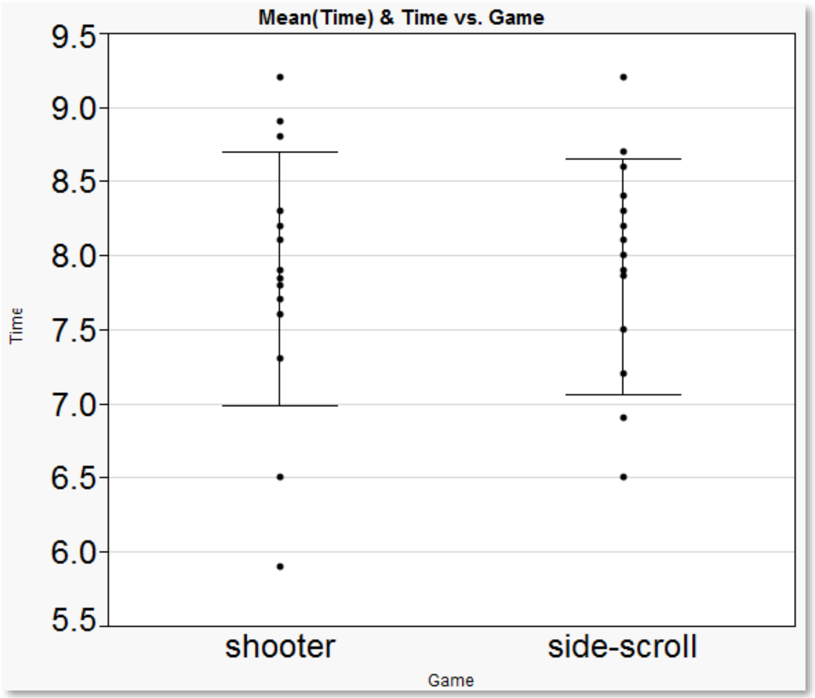
Let's look at a problem:

A researcher wanted to know whether the dinner a computer gamer ate the night before would cause a significant difference in their videogame playing performance the next day. Further, the researcher wanted to know if such differences occurred for first-person shooter games and side-scrolling games.

He recruited 30 computer gamers to take part, asking 16 to eat an 8 oz. steak dinner and 14 to eat an 8 oz. fish dinner the night before. The next morning, 15 subjects played a first-person shooter videogame and 15 other subjects played a side-scrolling videogame. The time taken to complete a given level in each game was recorded as the sole continuous measure of performance.

Wobbrock, J. O. Practical statistics for human-computer interaction. In *Annual Workshop of the HCI Consortium (HCIC'11).*





What's your conclusion?
How can you prove it?

Statistics Overview

Statistical significance

In statistics, **statistical significance** (or a statistically significant result) is attained when a ***p-value*** is less than the **significance level**.

In any experiment or observation that involves drawing a sample from a population, there is always the possibility that an observed effect would have occurred due to sampling error alone. But if the *p-value* is less than the significance level (e.g., $p < 0.05$), then an investigator may conclude that the observed effect actually reflects the characteristics of the population rather than just sampling error. An investigator may then report that the result attains **statistical significance**, thereby rejecting the **null hypothesis**.

http://en.wikipedia.org/wiki/Statistical_significance

Statistics Overview

Some Terms

- Factor (Independent variable)
- Level
- Measure (Dependent variable)
- Trial
- Covariate

- Within-subjects factor
- Between-subjects factor
- Within-subjects design
- Between-subjects design
- Mixed factorial design
- Balanced design

- Main effect
- Interaction
- Carryover effect
- Counterbalancing
- Nominal variable (Categorical variable)
- Ordinal variable
- Continuous variable (Scalar variable)

A researcher wants to know about the effects of body posture on mobile text entry. She implements 3 text entry methods in a custom iPhone test bed: a virtual QWERTY keyboard, Palm OS Graffiti, and a phone keypad simulation running Tegic's T9.

She recruits 20 subjects, having each one train for 30 minutes with one method chosen at random. Then she has each subject enter 20 test phrases in each of 3 postures—standing, walking, and jogging—the order of which was randomly determined. The outcomes of interest are words per minute and error rate.

Wobbrock, J. O. Practical statistics for human-computer interaction. In *Annual Workshop of the HCI Consortium (HCIC'11).*

Some Terms

- Factor (Independent variable)
- Level
- Measure (Dependent variable)
- Trial
- Covariate

- Within-subjects factor
- Between-subjects factor
- Within-subjects design
- Between-subjects design
- Mixed factorial design
- Balanced design

- Main effect
- Interaction
- Carryover effect
- Counterbalancing
- Nominal variable (Categorical variable)
- Ordinal variable
- Continuous variable (Scalar variable)

A factor (independent variable) is an experimental variable systematically changed to examine its effects, if any, upon an outcome of interest.

There are two factors, *body Posture* and *text entry Method*.

Some Terms

- Factor (Independent variable)
- Level
- Measure (Dependent variable)
- Trial
- Covariate

- Within-subjects factor
- Between-subjects factor
- Within-subjects design
- Between-subjects design
- Mixed factorial design
- Balanced design

- Main effect
- Interaction
- Carryover effect
- Counterbalancing
- Nominal variable (Categorical variable)
- Ordinal variable
- Continuous variable (Scalar variable)

A level is a particular value that a factor can assume.

The levels of *Posture* are standing, walking, and jogging.

The levels of *Method* are QWERTY, Graffiti, and T9.

Every factor must have at least two levels.

Statistics Overview

Some Terms

- Factor (Independent variable)
- Level
- Measure (Dependent variable)
- Trial
- Covariate

- Within-subjects factor
- Between-subjects factor
- Within-subjects design
- Between-subjects design
- Mixed factorial design
- Balanced design

- Main effect
- Interaction
- Carryover effect
- Counterbalancing
- Nominal variable (Categorical variable)
- Ordinal variable
- Continuous variable (Scalar variable)

A measure (dependent variable) is an experimental measure, response, or outcome of interest.

There are two dependent variables, *words per minute* and *error rate*.

Some Terms

- Factor (Independent variable)
- Level
- Measure (Dependent variable)
- Trial
- Covariate

- Within-subjects factor
- Between-subjects factor
- Within-subjects design
- Between-subjects design
- Mixed factorial design
- Balanced design

- Main effect
- Interaction
- Carryover effect
- Counterbalancing
- Nominal variable (Categorical variable)
- Ordinal variable
- Continuous variable (Scalar variable)

A trial is the experimental unit of activity over which one measure is taken.

The entry of one text entry phrase would constitute one trial.

Some Terms

- Factor (Independent variable)
- Level
- Measure (Dependent variable)
- Trial
- Covariate

- Within-subjects factor
- Between-subjects factor
- Within-subjects design
- Between-subjects design
- Mixed factorial design
- Balanced design

- Main effect
- Interaction
- Carryover effect
- Counterbalancing
- Nominal variable (Categorical variable)
- Ordinal variable
- Continuous variable (Scalar variable)

A covariate is a measurable feature of an experiment that, like a factor, may affect the dependent variable. Unlike a factor, however, a covariate is not manipulated; its levels take on their "natural" preset values and often cannot be changed.

E.g. the gender and age of each subject would be covariates. Other covariates include how fast each subject walks or jogs, or the outside temperature present while each subject entered his phrases.

Statistics Overview

Some Terms

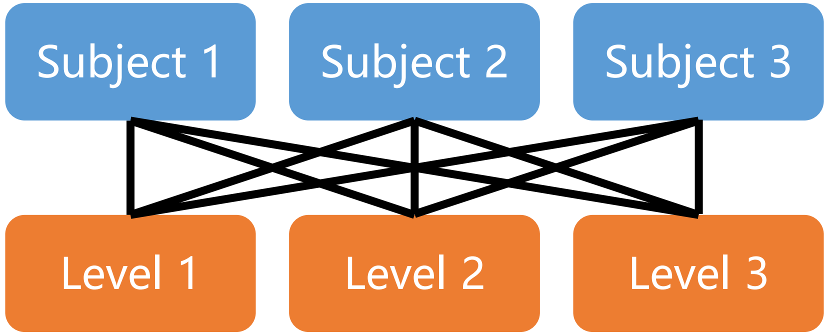
- Factor (Independent variable)
- Level
- Measure (Dependent variable)
- Trial
- Covariate

- Within-subjects factor
- Between-subjects factor
- Within-subjects design
- Between-subjects design
- Mixed factorial design
- Balanced design

- Main effect
- Interaction
- Carryover effect
- Counterbalancing
- Nominal variable (Categorical variable)
- Ordinal variable
- Continuous variable (Scalar variable)

A within-subjects factor is one for which all levels are experienced by each subject.

Since each subject entered phrases in all body postures, *Posture* is a within-subjects factor.



Statistics Overview

Some Terms

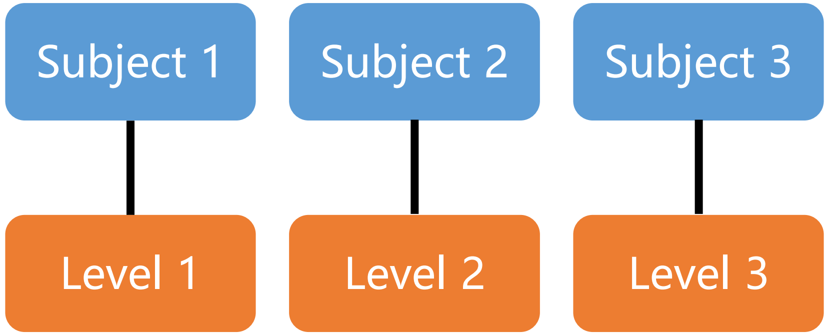
- Factor (Independent variable)
- Level
- Measure (Dependent variable)
- Trial
- Covariate

- Within-subjects factor
- Between-subjects factor
- Within-subjects design
- Between-subjects design
- Mixed factorial design
- Balanced design

- Main effect
- Interaction
- Carryover effect
- Counterbalancing
- Nominal variable (Categorical variable)
- Ordinal variable
- Continuous variable (Scalar variable)

A between-subjects factor is one for which only one level is experienced by each subject.

Since each subject used only one of three possible text entry methods, *Method* is a between-subjects factor.



Statistics Overview

Some Terms

- Factor (Independent variable)
- Level
- Measure (Dependent variable)
- Trial
- Covariate

- Within-subjects factor
- Between-subjects factor
- Within-subjects design
- Between-subjects design
- Mixed factorial design
- Balanced design

- Main effect
- Interaction
- Carryover effect
- Counterbalancing
- Nominal variable (Categorical variable)
- Ordinal variable
- Continuous variable (Scalar variable)

A within-subjects design is an experiment in which all factors are within-subjects factors.

If all of our subjects had used all three text entry methods, the experiment could be said to have used a within-subjects design.

Some Terms

- Factor (Independent variable)
- Level
- Measure (Dependent variable)
- Trial
- Covariate

- Within-subjects factor
- Between-subjects factor
- Within-subjects design
- Between-subjects design
- Mixed factorial design
- Balanced design

- Main effect
- Interaction
- Carryover effect
- Counterbalancing
- Nominal variable (Categorical variable)
- Ordinal variable
- Continuous variable (Scalar variable)

A between-subjects design is an experiment in which all factors are between-subjects factors.

If each subject had only experienced one posture along with only one text entry method, the experiment could be said to have used a between-subjects design.

Some Terms

- Factor (Independent variable)
- Level
- Measure (Dependent variable)
- Trial
- Covariate

- Within-subjects factor
- Between-subjects factor
- Within-subjects design
- Between-subjects design
- Mixed factorial design
- Balanced design

- Main effect
- Interaction
- Carryover effect
- Counterbalancing
- Nominal variable (Categorical variable)
- Ordinal variable
- Continuous variable (Scalar variable)

A mixed factorial design is an experiment in which there are within-subjects factors and between-subjects factors.

Our example experiment uses a mixed factorial design.

Some Terms

- Factor (Independent variable)
- Level
- Measure (Dependent variable)
- Trial
- Covariate

- Within-subjects factor
- Between-subjects factor
- Within-subjects design
- Between-subjects design
- Mixed factorial design
- Balanced design

- Main effect
- Interaction
- Carryover effect
- Counterbalancing
- Nominal variable (Categorical variable)
- Ordinal variable
- Continuous variable (Scalar variable)

A balanced design means that each level of each factor had assigned to it the same number of subjects.

As each of the 20 subjects were assigned to one of the 3 text entry methods, our example experiment does not exhibit a balanced design.

Some Terms

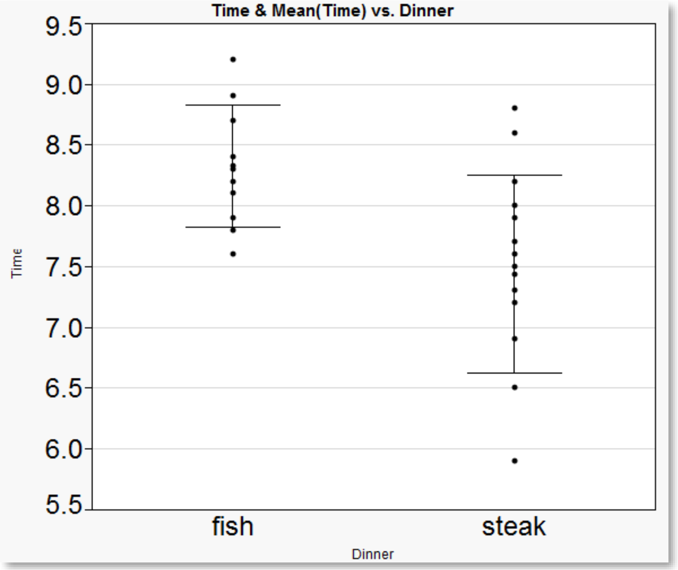
- Factor (Independent variable)
- Level
- Measure (Dependent variable)
- Trial
- Covariate

- Within-subjects factor
- Between-subjects factor
- Within-subjects design
- Between-subjects design
- Mixed factorial design
- Balanced design

- Main effect
- Interaction
- Carryover effect
- Counterbalancing
- Nominal variable (Categorical variable)
- Ordinal variable
- Continuous variable (Scalar variable)

A main effect refers to a finding of statistical significance for a factor in an experiment.

If *Posture* exerts a significant effect on words per minute, we would say "we have a main effect of *Posture* on text entry speed".



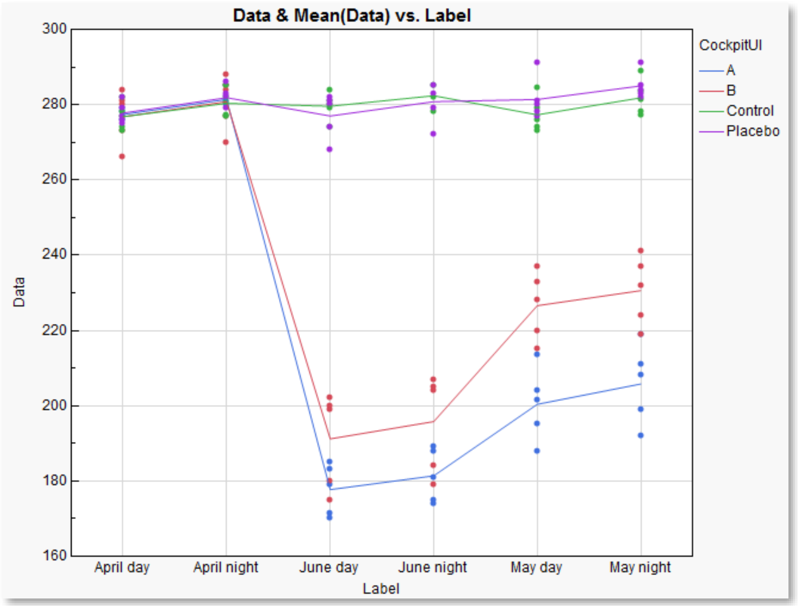
Some Terms

- Factor (Independent variable)
- Level
- Measure (Dependent variable)
- Trial
- Covariate

- Within-subjects factor
- Between-subjects factor
- Within-subjects design
- Between-subjects design
- Mixed factorial design
- Balanced design

- Main effect
- Interaction
- Carryover effect
- Counterbalancing
- Nominal variable (Categorical variable)
- Ordinal variable
- Continuous variable (Scalar variable)

An interaction refers to the interplay of two or more factors such that the effect of a level of a factor depends upon the level of another.



Statistics Overview

Some Terms

- Factor (Independent variable)
- Level
- Measure (Dependent variable)
- Trial
- Covariate

- Within-subjects factor
- Between-subjects factor
- Within-subjects design
- Between-subjects design
- Mixed factorial design
- Balanced design

- Main effect
- Interaction
- Carryover effect
- Counterbalancing
- Nominal variable (Categorical variable)
- Ordinal variable
- Continuous variable (Scalar variable)

Carryover effects threaten to confound any within-subjects factor as effects from one level of the factor change the results for a subsequent level of the factor.

In our example, a carryover effect may exist for *Posture*, if, say, subjects who jog first are then tired when entering phrases in the standing or walking postures.

In general, common carryover effects involve fatigue, learning, and motivational changes.

Carryover effects do not apply to between-subjects factors.

Some Terms

- Factor (Independent variable)
- Level
- Measure (Dependent variable)
- Trial
- Covariate

- Within-subjects factor
- Between-subjects factor
- Within-subjects design
- Between-subjects design
- Mixed factorial design
- Balanced design

- Main effect
- Interaction
- Carryover effect
- Counterbalancing
- Nominal variable (Categorical variable)
- Ordinal variable
- Continuous variable (Scalar variable)

Counterbalancing is the process by which the levels of a within-subjects factor are administered among the subjects to avoid carryover effects from systematically confounding the results.

In our example, the levels of *Posture* were counterbalanced by having their order chosen randomly for each subject.

Randomization is one approach to counterbalancing; other approaches involve deliberately issuing all possible orders in an experiment (called "fully counterbalanced") or using a systematic partial ordering, *e.g.*, a Latin Square. It is good practice to test for order effects to ensure that counterbalancing worked.

Some Terms

- Factor (Independent variable)
- Level
- Measure (Dependent variable)
- Trial
- Covariate

- Within-subjects factor
- Between-subjects factor
- Within-subjects design
- Between-subjects design
- Mixed factorial design
- Balanced design

- Main effect
- Interaction
- Carryover effect
- Counterbalancing
- Nominal variable (Categorical variable)
- Ordinal variable
- Continuous variable (Scalar variable)

A nominal variable (categorical variable) is a factor or measure that takes on one of an unordered assortment of values.

In our example, both *Posture* and *Method* are nominal variables.

Some Terms

- Factor (Independent variable)
- Level
- Measure (Dependent variable)
- Trial
- Covariate

- Within-subjects factor
- Between-subjects factor
- Within-subjects design
- Between-subjects design
- Mixed factorial design
- Balanced design

- Main effect
- Interaction
- Carryover effect
- Counterbalancing
- Nominal variable (Categorical variable)
- Ordinal variable
- Continuous variable (Scalar variable)

An ordinal variable is a factor or measure that takes on one of an ordered assortment of values. Although ordered, no assumption is made that the ordering is linear, i.e., that the gaps between successive values are known or regular.

In our example, if the experimenter had subjects fill out 7-point Likert-type scales with subjective ratings of their opinions, these data would be codified with ordinal variables.

Some Terms

- Factor (Independent variable)
- Level
- Measure (Dependent variable)
- Trial
- Covariate

- Within-subjects factor
- Between-subjects factor
- Within-subjects design
- Between-subjects design
- Mixed factorial design
- Balanced design

- Main effect
- Interaction
- Carryover effect
- Counterbalancing
- Nominal variable (Categorical variable)
- Ordinal variable
- Continuous variable (Scalar variable)

A continuous variable (scalar variable) is a factor or measure that takes on a number whose relationship to and distance from other numbers is known.

In our example, both words per minute and text entry error rate are continuous dependent variables.

Practical Statistics for Human-Computer Interaction

Practical Statistics for Human-Computer Interaction

Statistical Hypothesis Testing

A **statistical hypothesis** is a scientific hypothesis that is testable on the basis of observing a process that is modeled via a set of random variables. A **statistical hypothesis test** is a method of statistical inference used for testing a statistical hypothesis.

A test result is called **statistically significant** if it has been predicted as unlikely to have occurred by sampling error alone, according to a threshold probability — **the significance level**. Hypothesis tests are used in determining what outcomes of a study would lead to a rejection of the null hypothesis for a pre-specified level of significance. In the Neyman-Pearson framework, the process of distinguishing between the **null hypothesis** and the **alternative hypothesis** is aided by identifying two conceptual types of errors **(type I & type II)**, and by specifying parametric limits on e.g. how much type I error will be permitted.

http://en.wikipedia.org/wiki/Statistical_hypothesis_testing

Procedure

1. There is an initial research hypothesis of which the truth is unknown.
2. State the relevant **null** and **alternative hypotheses**.
3. Consider the **statistical assumptions** being made about the sample in doing the test; for example, assumptions about the statistical independence or about the form of the distributions of the observations.
4. Decide which test is appropriate, and state the relevant **test statistic** T.
5. Select a **significance level (α)**, a probability threshold below which the null hypothesis will be rejected. Common values are 5% and 1%.
6. The distribution of the test statistic under the null hypothesis partitions the possible values of T into those for which the null hypothesis is rejected— the so-called **critical region**—and those for which it is not. The probability of the critical region is **α**.
7. Compute from the observations the observed value t_{obs} of the test statistic T.
8. Calculate the ***p*-value**. This is the probability, under the null hypothesis, of sampling a test statistic at least as extreme as that which was observed.
9. Reject the null hypothesis, in favor of the alternative hypothesis, if and only if the *p-value* is less than the significance level (the selected probability) threshold (equivalently, if the observed test statistic is in the critical region).

http://en.wikipedia.org/wiki/Statistical_hypothesis_testing

Interpretation

If the *p*-value is less than the required significance level (equivalently, if the observed test statistic is in the critical region), then we say "The null hypothesis is rejected at the given level of significance".

If the *p*-value is not less than the required significance level (equivalently, if the observed test statistic is outside the critical region), then the test has no result.

http://en.wikipedia.org/wiki/Statistical_hypothesis_testing

Some Terms

- Null hypothesis
- Alternative hypothesis
- Test statistic
- Region of acceptance
- Region of rejection (Critical region)
- Critical value

- Type-I and Type-II error
- Significance level of a test (α)
- Power of test (1 - β)
- Degrees of freedom
- Effect size
- *P*-value

- One-tailed test / Two-tailed test
- One-sample test / Two-sample test
- Paired test
- Parametric test
- Nonparametric test
- Post-hoc test

1. A female colleague of Fisher claimed to be able to tell whether the tea or the milk was added first to a cup. Fisher proposed to give her 8 cups, 4 of each variety, in random order. One could then ask what the probability was for her getting the number she got correct, but just by chance.

2. A defendant is considered not guilty as long as his or her guilt is not proven. The prosecutor tries to prove the guilt of the defendant. Only when there is enough charging evidence the defendant is convicted.

http://en.wikipedia.org/wiki/Statistical_hypothesis_testing

Some Terms

- Null hypothesis
- Alternative hypothesis
- Test statistic
- Region of acceptance
- Region of rejection (Critical region)
- Critical value

- Type-I and Type-II error
- Significance level of a test (α)
- Power of test (1 - β)
- Degrees of freedom
- Effect size
- *P*-value

- One-tailed test / Two-tailed test
- One-sample test / Two-sample test
- Paired test
- Parametric test
- Nonparametric test
- Post-hoc test

A simple hypothesis associated with a contradiction to a theory one would like to prove.

The null hypothesis was that the Lady had no such ability.

"the defendant is not gugilty" is the null hypothesis, and is for the time being accepted.

Some Terms

- Null hypothesis
- Alternative hypothesis
- Test statistic
- Region of acceptance
- Region of rejection (Critical region)
- Critical value

- Type-I and Type-II error
- Significance level of a test (α)
- Power of test (1 - β)
- Degrees of freedom
- Effect size
- *P*-value

- One-tailed test / Two-tailed test
- One-sample test / Two-sample test
- Paired test
- Parametric test
- Nonparametric test
- Post-hoc test

A hypothesis (often composite) associated with a theory one would like to prove.

The alternative hypothesis was that the Lady had such ability.

"the defendant is guilty" is the alternative hypothesis, and is the hypothesis one hopes to support.

Some Terms

- Null hypothesis
- Alternative hypothesis
- Test statistic
- Region of acceptance
- Region of rejection (Critical region)
- Critical value

- Type-I and Type-II error
- Significance level of a test (α)
- Power of test (1 - β)
- Degrees of freedom
- Effect size
- *P*-value

- One-tailed test / Two-tailed test
- One-sample test / Two-sample test
- Paired test
- Parametric test
- Nonparametric test
- Post-hoc test

A test statistic is a single measure of some attribute of a sample (i.e. a statistic) used in statistical hypothesis testing.

An important property of a test statistic is that its sampling distribution under the null hypothesis must be calculable, either exactly or approximately, which allows *p*-values to be calculated.

The test statistic was a simple count of the number of successes in selecting the 4 cups.

http://en.wikipedia.org/wiki/Test_statistic

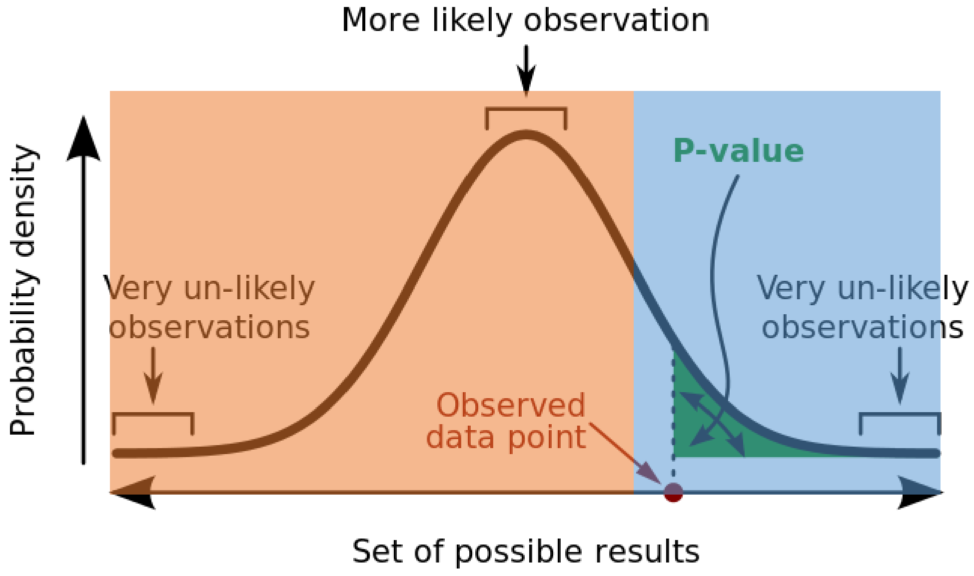
Some Terms

- Null hypothesis
- Alternative hypothesis
- Test statistic
- Region of acceptance
- Region of rejection (Critical region)
- Critical value

- Type-I and Type-II error
- Significance level of a test (α)
- Power of test (1 - β)
- Degrees of freedom
- Effect size
- *P*-value

- One-tailed test / Two-tailed test
- One-sample test / Two-sample test
- Paired test
- Parametric test
- Nonparametric test
- Post-hoc test

The set of values of the test statistic for which we fail to reject the null hypothesis.



Statistics Overview

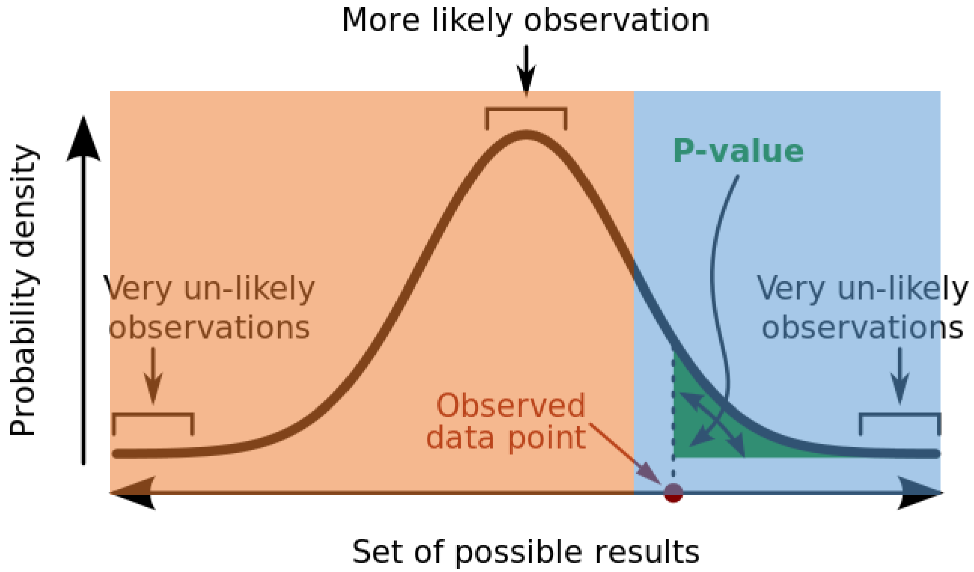
Some Terms

- Null hypothesis
- Alternative hypothesis
- Test statistic
- Region of acceptance
- Region of rejection (Critical region)
- Critical value

- Type-I and Type-II error
- Significance level of a test (α)
- Power of test (1 - β)
- Degrees of freedom
- Effect size
- *P*-value

- One-tailed test / Two-tailed test
- One-sample test / Two-sample test
- Paired test
- Parametric test
- Nonparametric test
- Post-hoc test

The set of values of the test statistic for which the null hypothesis is rejected.



Statistics Overview

Some Terms

- Null hypothesis
- Alternative hypothesis
- Test statistic
- Region of acceptance
- Region of rejection (Critical region)
- Critical value

- Type-I and Type-II error
- Significance level of a test (α)
- Power of test (1 - β)
- Degrees of freedom
- Effect size
- *P*-value

- One-tailed test / Two-tailed test
- One-sample test / Two-sample test
- Paired test
- Parametric test
- Nonparametric test
- Post-hoc test

The critical region was the single case of 4 successes of 4 possible based on a conventional probability criterion (< 5%; 1 of 70 ≈ 1.4%). If the lady correctly identified every cup, that would be considered a statistically significant result.

Test Statistic	Formula	Probability
0	1/C(8,4)	1/70
1	C(4,1)*C(4,3)/C(8,4)	16/70
2	C(4,2)*C(4,2)/C(8,4)	36/70
3	C(4,1)*C(4,3)/C(8,4)	16/70
4	1/C(8,4)	1/70

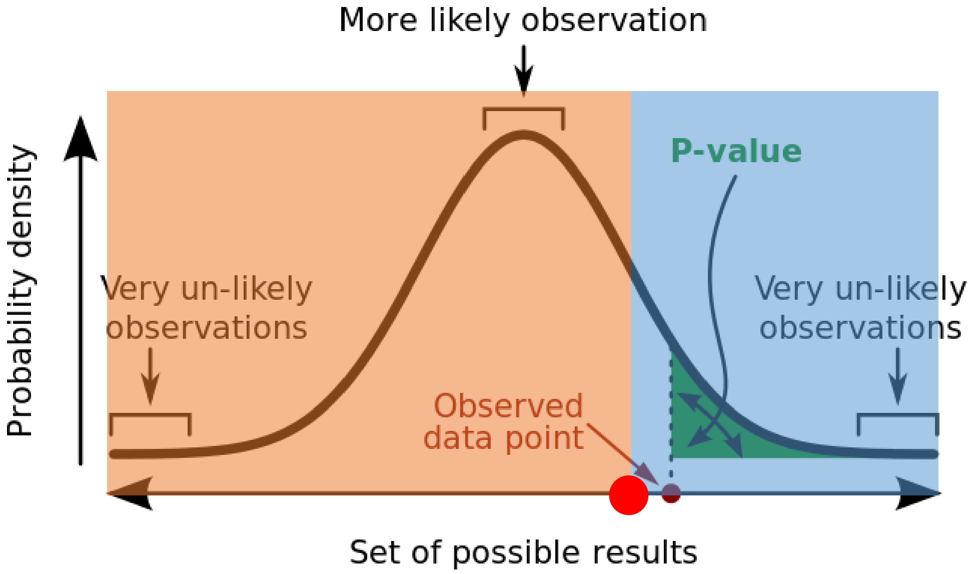
Some Terms

- Null hypothesis
- Alternative hypothesis
- Test statistic
- Region of acceptance
- Region of rejection (Critical region)
- Critical value

- Type-I and Type-II error
- Significance level of a test (α)
- Power of test (1 - β)
- Degrees of freedom
- Effect size
- *P*-value

- One-tailed test / Two-tailed test
- One-sample test / Two-sample test
- Paired test
- Parametric test
- Nonparametric test
- Post-hoc test

The threshold value delimiting the regions of acceptance and rejection for the test statistic.



Statistics Overview

Some Terms

- Null hypothesis
- Alternative hypothesis
- Test statistic
- Region of acceptance
- Region of rejection (Critical region)
- Critical value

- Type-I and Type-II error
- Significance level of a test (α)
- Power of test (1 - β)
- Degrees of freedom
- Effect size
- *P*-value

- One-tailed test / Two-tailed test
- One-sample test / Two-sample test
- Paired test
- Parametric test
- Nonparametric test
- Post-hoc test

A type-I error is the incorrect rejection of a true null hypothesis (a "false positive"), or detecting an effect that is not present.

A type-II error is the failure to reject a false null hypothesis (a "false negative"), or failing to detect an effect that is present.

The terms "type-I error" and "type-II error" are often used interchangeably with the general notion of false positives and false negatives in binary classification.

http://en.wikipedia.org/wiki/Type_I_and_type_II_errors

Some Terms

- Null hypothesis
- Alternative hypothesis
- Test statistic
- Region of acceptance
- Region of rejection (Critical region)
- Critical value

- Type-I and Type-II error
- Significance level of a test (α)
- Power of test (1 - β)
- Degrees of freedom
- Effect size
- *P*-value

- One-tailed test / Two-tailed test
- One-sample test / Two-sample test
- Paired test
- Parametric test
- Nonparametric test
- Post-hoc test

The conviction of an innocent person is called error of the first kind, and the occurrence of this error is controlled to be rare.

As a consequence of this asymmetric behaviour, the error of the second kind (acquitting a person who committed the crime), is often rather large.

	H_0 is true: **Innocent**	H_1 is true: **Guilty**
Accept null hypothesis: **Innocent**	Correct Inference (True negative)	Type-II Error (False negative)
Reject null hypothesis: **Guilty**	Type-I Error (False positive)	Right decision (True positive)

Some Terms

- Null hypothesis
- Alternative hypothesis
- Test statistic
- Region of acceptance
- Region of rejection (Critical region)
- Critical value

- Type-I and Type-II error
- Significance level of a test (α)
- Power of test (1 - β)
- Degrees of freedom
- Effect size
- *P*-value

- One-tailed test / Two-tailed test
- One-sample test / Two-sample test
- Paired test
- Parametric test
- Nonparametric test
- Post-hoc test

The test's probability of incorrectly rejecting the null hypothesis.

α is the rate of type-I error.

$$\alpha = FP/(FP + TN)$$

$F_{1,19} = 39.28$, $p < .001$

	H_0 is true: **Innocent**	H_1 is true: **Guilty**
Accept null hypothesis: **Innocent**	Correct Inference (True negative)	Type-II Error (False negative)
Reject null hypothesis: **Guilty**	Type-I Error (False positive)	Right decision (True positive)

Statistics Overview

Some Terms

- Null hypothesis
- Alternative hypothesis
- Test statistic
- Region of acceptance
- Region of rejection (Critical region)
- Critical value

- Type-I and Type-II error
- Significance level of a test (α)
- Power of test (1 - β)
- Degrees of freedom
- Effect size
- *P*-value

- One-tailed test / Two-tailed test
- One-sample test / Two-sample test
- Paired test
- Parametric test
- Nonparametric test
- Post-hoc test

The test's probability of correctly rejecting the null hypothesis.

The complement of the Type-II error rate, β.

$$1-\beta = TP/(TP+FN)$$

	H_0 is true: **Innocent**	H_1 is true: **Guilty**
Accept null hypothesis: **Innocent**	Correct Inference (True negative)	Type-II Error (False negative)
Reject null hypothesis: **Guilty**	Type-I Error (False positive)	Right decision (True positive)

Some Terms

- Null hypothesis
- Alternative hypothesis
- Test statistic
- Region of acceptance
- Region of rejection (Critical region)
- Critical value

- Type-I and Type-II error
- Significance level of a test (α)
- Power of test (1 - β)
- Degrees of freedom
- Effect size
- *P*-value

- One-tailed test / Two-tailed test
- One-sample test / Two-sample test
- Paired test
- Parametric test
- Nonparametric test
- Post-hoc test

In statistics, the number of degrees of freedom is the number of values in the final calculation of a statistic that are free to vary.

Several commonly encountered statistical distributions (Student's t, Chi-Squared, F) have parameters that are commonly referred to as degrees of freedom. This terminology simply reflects that in many applications where these distributions occur, the parameter corresponds to the degrees of freedom of an underlying random vector.

$$F_{1,19} = 39.28,\ p < .001$$

http://en.wikipedia.org/wiki/Degrees_of_freedom_(statistics)

Statistics Overview

Some Terms

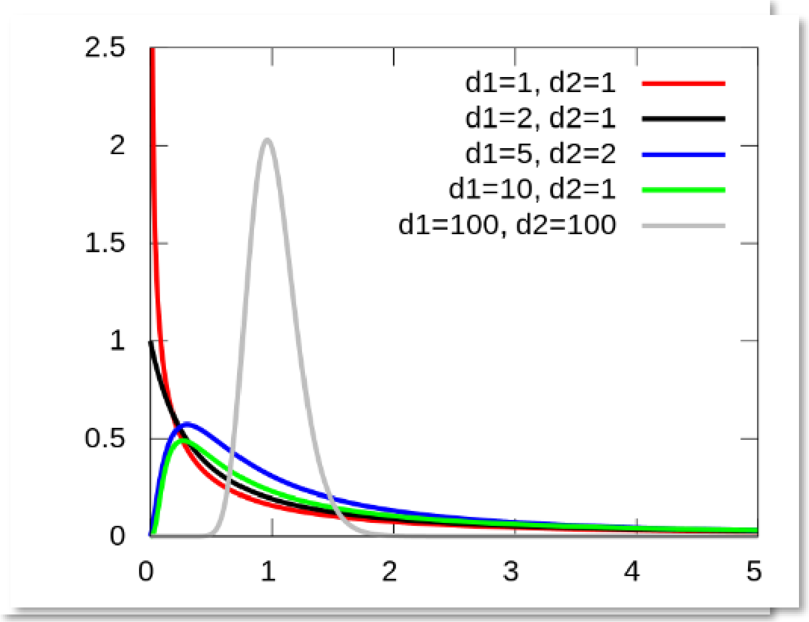
- Null hypothesis
- Alternative hypothesis
- Test statistic
- Region of acceptance
- Region of rejection (Critical region)
- Critical value

- Type-I and Type-II error
- Significance level of a test (α)
- Power of test (1 - β)
- Degrees of freedom
- Effect size
- *P*-value

- One-tailed test / Two-tailed test
- One-sample test / Two-sample test
- Paired test
- Parametric test
- Nonparametric test
- Post-hoc test

In the application of these distributions to linear models, the degrees of freedom parameters can take only integer values. The underlying families of distributions allow fractional values for the degrees-of-freedom parameters, which can arise in more sophisticated uses.

$$F_{2.8,34.3} = 5.70, \; p < .05$$



http://en.wikipedia.org/wiki/Student%27s_t-distribution
http://en.wikipedia.org/wiki/Chi-squared_distribution
http://en.wikipedia.org/wiki/F-distribution

Statistics Overview

Some Terms

- Null hypothesis
- Alternative hypothesis
- Test statistic
- Region of acceptance
- Region of rejection (Critical region)
- Critical value

- Type-I and Type-II error
- Significance level of a test (α)
- Power of test (1 - β)
- Degrees of freedom
- Effect size
- *P*-value

- One-tailed test / Two-tailed test
- One-sample test / Two-sample test
- Paired test
- Parametric test
- Nonparametric test
- Post-hoc test

In statistics, an effect size is a quantitative measure of the strength of a phenomenon.

Examples of effect sizes are the correlation between two variables, the regression coefficient, the mean difference.

For each type of effect size, a larger absolute value always indicates a stronger effect.

The reporting of effect sizes facilitates the interpretation of the substantive, as opposed to the statistical, significance of a research result.

http://en.wikipedia.org/wiki/Effect_size

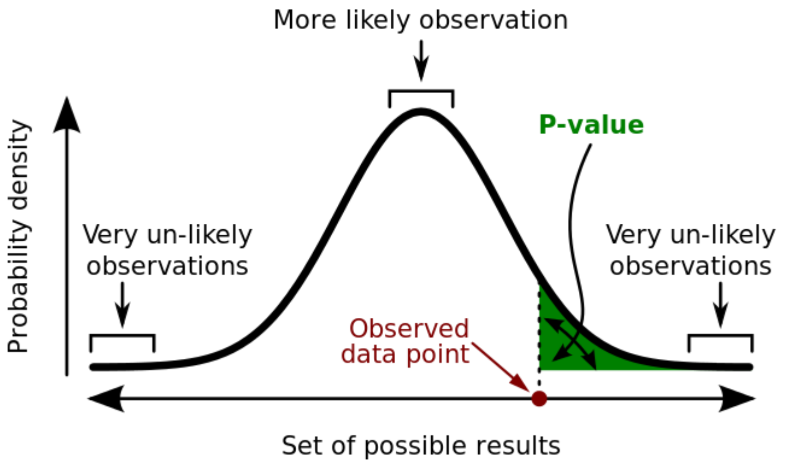
Some Terms

- Null hypothesis
- Alternative hypothesis
- Test statistic
- Region of acceptance
- Region of rejection (Critical region)
- Critical value

- Type-I and Type-II error
- Significance level of a test (α)
- Power of test (1 - β)
- Degrees of freedom
- Effect size
- *P*-value

- One-tailed test / Two-tailed test
- One-sample test / Two-sample test
- Paired test
- Parametric test
- Nonparametric test
- Post-hoc test

P-value is the probability of obtaining the observed sample results, or "more extreme" results, when the null hypothesis is actually true (where "more extreme" is dependent on the way the hypothesis is tested).



A **p-value** (shaded green area) is the probability of an observed (or more extreme) result assuming that the null hypothesis is true.

http://en.wikipedia.org/wiki/P-value

Statistics Overview

Some Terms

- Null hypothesis
- Alternative hypothesis
- Test statistic
- Region of acceptance
- Region of rejection (Critical region)
- Critical value

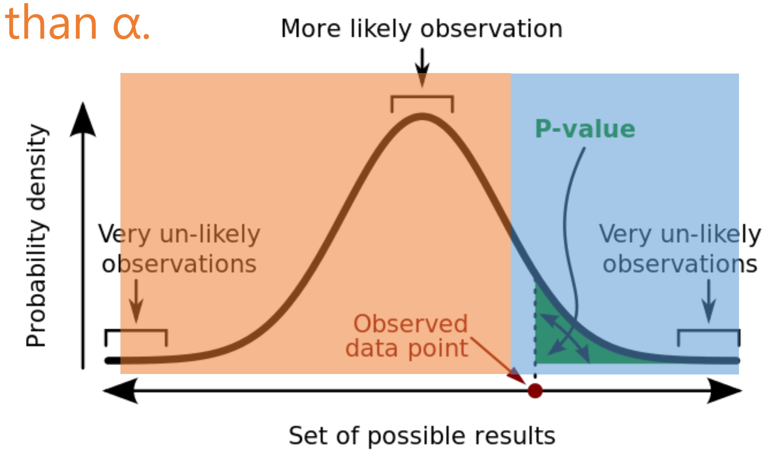
- Type-I and Type-II error
- Significance level of a test (α)
- Power of test (1 - β)
- Degrees of freedom
- Effect size
- *P*-value

- One-tailed test / Two-tailed test
- One-sample test / Two-sample test
- Paired test
- Parametric test
- Nonparametric test
- Post-hoc test

http://en.wikipedia.org/wiki/P-value

If the *p*-value is equal to or smaller than the significance level (α), it suggests that the observed data are inconsistent with the assumption that the null hypothesis is true, and thus that hypothesis must be rejected and the alternative hypothesis is accepted as true.

When the *p*-value is calculated correctly, such a test is guaranteed to control the Type-I error rate to be no greater than α.



Statistics Overview

Some Terms

- Null hypothesis
- Alternative hypothesis
- Test statistic
- Region of acceptance
- Region of rejection (Critical region)
- Critical value

- Type-I and Type-II error
- Significance level of a test (α)
- Power of test (1 - β)
- Degrees of freedom
- Effect size
- *P*-value

- One-tailed test / Two-tailed test
- One-sample test / Two-sample test
- Paired test
- Parametric test
- Nonparametric test
- Post-hoc test

A two-tailed test is used if deviations of the estimated parameter in either direction from some benchmark value are considered theoretically possible; in contrast, a one-tailed test is used if only deviations in one direction are considered possible.

In a one-tailed test, "extreme" is decided beforehand as either meaning "sufficiently small" or meaning "sufficiently large" – values in the other direction are considered not significant. In a two-tailed test, "extreme" means "either sufficiently small or sufficiently large", and values in either direction are considered significant.

http://en.wikipedia.org/wiki/One-_and_two-tailed_tests

Some Terms

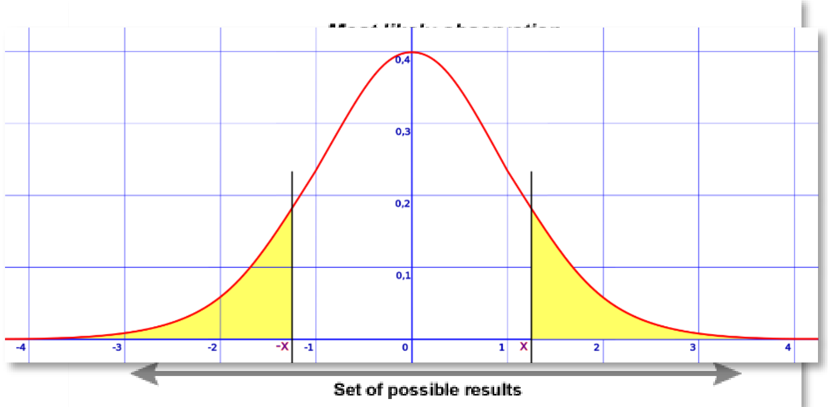
- Null hypothesis
- Alternative hypothesis
- Test statistic
- Region of acceptance
- Region of rejection (Critical region)
- Critical value

- Type-I and Type-II error
- Significance level of a test (α)
- Power of test (1 - β)
- Degrees of freedom
- Effect size
- *P*-value

- One-tailed test / Two-tailed test
- One-sample test / Two-sample test
- Paired test
- Parametric test
- Nonparametric test
- Post-hoc test

http://en.wikipedia.org/wiki/One-_and_two-tailed_tests

For a given test statistic there is a single two-tailed test, and two one-tailed tests, one each for either direction. Given data of a given significance level in a two-tailed test for a test statistic, in the corresponding one-tailed tests for the same test statistic it will be considered either twice as significant (half the *p*-value), if the data is in the direction specified by the test, or not significant at all (*p*-value above 0.5), if the data is in the direction opposite that specified by the test.



Statistics Overview

Some Terms

- Null hypothesis
- Alternative hypothesis
- Test statistic
- Region of acceptance
- Region of rejection (Critical region)
- Critical value

- Type-I and Type-II error
- Significance level of a test (α)
- Power of test (1 - β)
- Degrees of freedom
- Effect size
- *P*-value

- One-tailed test / Two-tailed test
- One-sample test / Two-sample test
- Paired test
- Parametric test
- Nonparametric test
- Post-hoc test

In coin flipping, the null hypothesis is a sequence of Bernoulli trials with probability 0.5, yielding a random variable *X* which is 1 for heads and 0 for tails, and a common test statistic is the sample mean (of the number of heads). Assume we have observed (HHHHH).

If testing for whether the coin is biased towards heads, a one-tailed test would be used – only large numbers of heads would be significant. $p = 1/32 < 0.05$, and thus would be significant (rejecting the null hypothesis) if using 0.05 as the cutoff.

However, if testing for whether the coin is biased towards heads or tails, a two-tailed test would be used, $p = 2/32 > 0.05$. This would not be significant (not rejecting the null hypothesis) if using 0.05 as the cutoff.

http://en.wikipedia.org/wiki/One-_and_two-tailed_tests

Some Terms

- Null hypothesis
- Alternative hypothesis
- Test statistic
- Region of acceptance
- Region of rejection (Critical region)
- Critical value

- Type-I and Type-II error
- Significance level of a test (α)
- Power of test ($1 - \beta$)
- Degrees of freedom
- Effect size
- *P*-value

- One-tailed test / Two-tailed test
- One-sample test / Two-sample test
- Paired test
- Parametric test
- Nonparametric test
- Post-hoc test

One-sample tests are appropriate when a sample is being compared to the population from a hypothesis. The population characteristics are known from theory or are calculated from the population.

Two-sample tests are appropriate for comparing two samples, typically experimental and control samples from a scientifically controlled experiment.

http://en.wikipedia.org/wiki/Statistical_hypothesis_testing

Some Terms

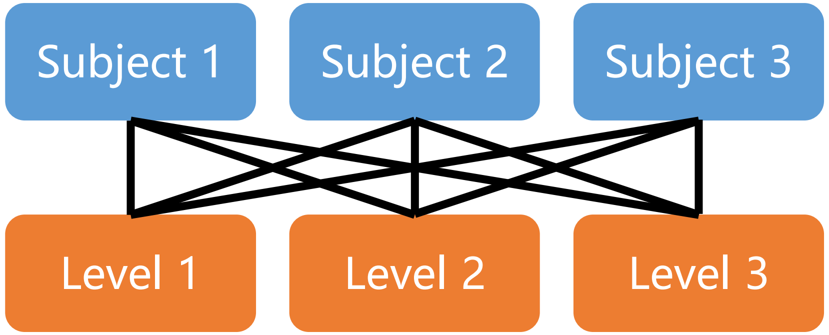
- Null hypothesis
- Alternative hypothesis
- Test statistic
- Region of acceptance
- Region of rejection (Critical region)
- Critical value

- Type-I and Type-II error
- Significance level of a test (α)
- Power of test (1 - β)
- Degrees of freedom
- Effect size
- *P*-value

- One-tailed test / Two-tailed test
- One-sample test / Two-sample test
- Paired test
- Parametric test
- Nonparametric test
- Post-hoc test

http://en.wikipedia.org/wiki/Statistical_hypothesis_testing

Paired tests are appropriate for comparing two samples where it is impossible to control important variables. Rather than comparing two sets, members are paired between samples so the difference between the members becomes the sample. Typically the mean of the differences is then compared to zero. The common example scenario for when a paired difference test is appropriate is when a single set of test subjects has something applied to them and the test is intended to check for an effect.



Statistics Overview

Some Terms

- Null hypothesis
- Alternative hypothesis
- Test statistic
- Region of acceptance
- Region of rejection (Critical region)
- Critical value

- Type-I and Type-II error
- Significance level of a test (α)
- Power of test ($1 - \beta$)
- Degrees of freedom
- Effect size
- *P*-value

- One-tailed test / Two-tailed test
- One-sample test / Two-sample test
- Paired test
- Parametric test
- Nonparametric test
- Post-hoc test

Parametric statistics assumes that the data have come from a type of probability distribution and makes inferences about the parameters of the distribution. (e.g. *F* test)

The difference between parametric model and non-parametric model is that the former has a fixed number of parameters, while the latter grows the number of parameters with the amount of training data.

http://en.wikipedia.org/wiki/Parametric_statistics

Some Terms

- Null hypothesis
- Alternative hypothesis
- Test statistic
- Region of acceptance
- Region of rejection (Critical region)
- Critical value

- Type-I and Type-II error
- Significance level of a test (α)
- Power of test (1 - β)
- Degrees of freedom
- Effect size
- *P*-value

- One-tailed test / Two-tailed test
- One-sample test / Two-sample test
- Paired test
- Parametric test
- Nonparametric test
- Post-hoc test

Nonparametric statistics are statistics not based on parameterized families of probability distributions.
Unlike parametric statistics, nonparametric statistics make no assumptions about the probability distributions of the variables being assessed.

Nonparametric methods are widely used for studying populations that take on a ranked order (such as movie reviews receiving one to four stars). The use of non-parametric methods may be necessary when data have a ranking but no clear numerical interpretation, such as when assessing preferences.

http://en.wikipedia.org/wiki/Nonparametric_statistics

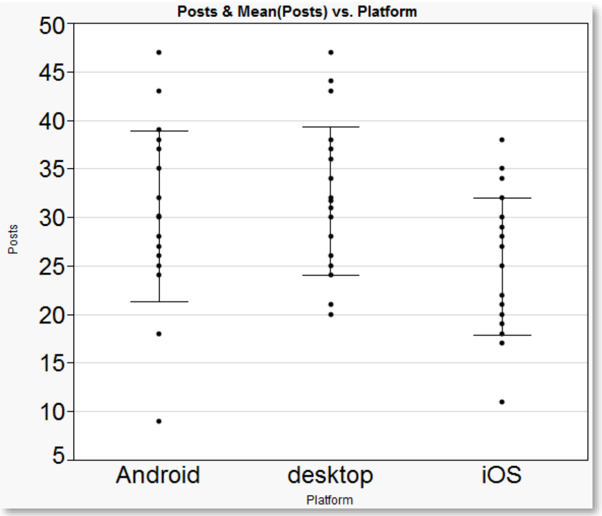
Some Terms

- Null hypothesis
- Alternative hypothesis
- Test statistic
- Region of acceptance
- Region of rejection (Critical region)
- Critical value

- Type-I and Type-II error
- Significance level of a test (α)
- Power of test (1 - β)
- Degrees of freedom
- Effect size
- *P*-value

- One-tailed test / Two-tailed test
- One-sample test / Two-sample test
- Paired test
- Parametric test
- Nonparametric test
- Post-hoc test

The omnibus F-test does not tell us whether all three levels of *Platform* are different from one another, or whether just two levels (and which two?) are different. For this, we need *post hoc* comparisons, which are justified only when the omnibus F-test is significant.



$F_{2,57} = 4.03, p < .05$

Wobbrock, J. O. Practical statistics for human-computer interaction. In *Annual Workshop of the HCI Consortium (HCIC'11)*.

Statistics Overview

Some Terms

- Null hypothesis
- Alternative hypothesis
- Test statistic
- Region of acceptance
- Region of rejection (Critical region)
- Critical value

- Type-I and Type-II error
- Significance level of a test (α)
- Power of test ($1 - \beta$)
- Degrees of freedom
- Effect size
- *P*-value

- One-tailed test / Two-tailed test
- One-sample test / Two-sample test
- Paired test
- Parametric test
- Nonparametric test
- Post-hoc test

Statistical inference logic is based on rejecting the null hypotheses if the likelihood of the observed data under the null hypotheses is low. The problem of multiplicity arises from the fact that as we increase the number of hypotheses in a test, we also increase the likelihood of witnessing a rare event, and therefore, the chance to reject the null hypotheses when it's true (type-I error).

For example, if one test is performed at the 5% level, there is only a 5% chance of incorrectly rejecting the null hypothesis if the null hypothesis is true. However, for 100 tests where all null hypotheses are true, the expected number of incorrect rejections is 5. If the tests are independent, the probability of at least one incorrect rejection is 99.4%. These errors are called false positives or Type-I errors.

http://en.wikipedia.org/wiki/Bonferroni_correction
http://en.wikipedia.org/wiki/Multiple_comparisons_problem

Some Terms

- Null hypothesis
- Alternative hypothesis
- Test statistic
- Region of acceptance
- Region of rejection (Critical region)
- Critical value

- Type-I and Type-II error
- Significance level of a test (α)
- Power of test (1 - β)
- Degrees of freedom
- Effect size
- *P*-value

- One-tailed test / Two-tailed test
- One-sample test / Two-sample test
- Paired test
- Parametric test
- Nonparametric test
- Post-hoc test

A Bonferroni correction divides α by the number of post hoc comparisons.

In this case, with 3 post hoc comparisons, we would use α = .05 / 3 = .0166.

	P-value	Without Correction (α = .05)	With Correction (α = .0166)
Android-Desktop	.5221	*Not significant*	*Not significant*
Desktop-iOS	.0087	*Significant*	*Significant*
Android-iOS	.0427	*Significant*	*Not significant*

Wobbrock, J. O. Practical statistics for human-computer interaction. In *Annual Workshop of the HCI Consortium (HCIC'11).*

Statistics Overview

Practical Statistics for Human-Computer Interaction

Tests of Significance

Parametric tests

- Student's *t* test
- Paired-samples *t* test
- One-way ANOVA
- N-way ANOVA
- Repeated measures ANOVA
- MANOVA
- Tukey-Kramer HSD test

Categories, counts and proportion tests

- One-sample Pearson Chi-Square (χ^2) test
- Two-sample Pearson Chi –Square (χ^2) test
- Fisher's exact test

Nonparametric tests

- Mann-Whitney *U* test
- Wilcoxon rank-sum test
- Wilcoxon signed-rank test
- Kruskal-Wallis test
- Friedman test

Auxiliary tests

- Shapiro-Wilk *W* test
- Kolmogorov-Smirnov *D* test
- Mauchly's sphericity test
- Levene's test

Wobbrock, J. O. Practical statistics for human-computer interaction. In *Annual Workshop of the HCI Consortium (HCIC'11).*

Tests of Significance

No. Factors	No. Levels	Between-subjects or within-subjects	Parametric Test	Semiparametric or Nonparametric Equivalent
1	2	Between	independent-samples t	Mann-Whitney U
1	2	Within	paired-samples t	Wilcoxon signed-rank
1	3+	Between	one-way ANOVA	Kruskal-Wallis
1	3+	Within	repeated measures ANOVA	Friedman
2+	2+ ea.	Between only (cannot do within)	n-way ANOVA	GZLMs
2+	2+ ea.	Within (can also do between)	repeated measures ANOVA	ART or, GLMMs or, GEEs.

Wobbrock, J. O. Practical statistics for human-computer interaction. In*Annual Workshop of the HCI Consortium (HCIC'11).*

Tests of Significance

Parametric tests

- Student's *t* test
- Paired-samples *t* test
- One-way ANOVA
- N-way ANOVA
- Repeated measures ANOVA
- MANOVA
- Tukey-Kramer HSD test

Categories, counts and proportion tests

- One-sample Pearson Chi-Square (χ^2) test
- Two-sample Pearson Chi –Square (χ^2) test
- Fisher's exact test

Nonparametric tests

- Mann-Whitney *U* test
- Wilcoxon rank-sum test
- Wilcoxon signed-rank test
- Kruskal-Wallis test
- Friedman test

Auxiliary tests

- Shapiro-Wilk *W* test
- Kolmogorov-Smirnov *D* test
- Mauchly's sphericity test
- Levene's test

Wobbrock, J. O. Practical statistics for human-computer interaction. In *Annual Workshop of the HCI Consortium (HCIC'11).*

Tests of Significance

Parametric tests

- Student's *t* test
- Paired-samples *t* test
- One-way ANOVA
- N-way ANOVA
- Repeated measures ANOVA
- MANOVA
- Tukey-Kramer HSD test

Categories, counts and proportion tests

- One-sample Pearson Chi-Square (χ^2) test
- Two-sample Pearson Chi –Square (χ^2) test
- Fisher's exact test

Nonparametric tests

- Mann-Whitney *U* test
- Wilcoxon rank-sum test
- Wilcoxon signed-rank test
- Kruskal-Wallis test
- Friedman test

Auxiliary tests

- Shapiro-Wilk *W* test
- Kolmogorov-Smirnov *D* test
- Mauchly's sphericity test
- Levene's test

Wobbrock, J. O. Practical statistics for human-computer interaction. In *Annual Workshop of the HCI Consortium (HCIC'11).*

Student's *t* test

- Usage

  Compare the mean of two independent samples.

- Assumptions

  1. Dependent variable should be measured on a continuous scale.
  2. Independent variable should consist of two categorical, independent groups.
  3. There should be no significant outliers.
  4. Dependent variable should be approximately normally distributed for each group of the independent variable.
  5. There needs to be homogeneity of variances.

- Report of results

$$t_{19} = 5.30, \ p < .001$$

http://en.wikipedia.org/wiki/Student%27s_t-test
https://statistics.laerd.com/spss-tutorials/independent-t-test-using-spss-statistics.php

Paired-samples *t* test

- Usage

  Compare the mean of two paired samples.

- Assumptions

  1. Dependent variable should be measured on a continuous scale.
  2. Independent variable should consist of two categorical, related groups.
  3. There should be no significant outliers.
  4. The distribution of the differences in the dependent variable between the two related groups should be approximately normally distributed.

- Report of results

$$t_{19} = 5.30,\ p < .001$$

http://en.wikipedia.org/wiki/Student%27s_t-test
https://statistics.laerd.com/spss-tutorials/dependent-t-test-using-spss-statistics.php

One-way ANOVA

- Usage

  Typically, compare means of three or more samples.

  When there are only two means to compare, the t-test and the F-test are equivalent. The relation between ANOVA and t is given by $F = t^2$.

- Assumptions

  1. Dependent variable should be measured on a continuous scale.
  2. Independent variable should consist of two or more categorical, independent groups.
  3. There should be no significant outliers.
  4. Dependent variable should be approximately normally distributed for each category of the independent variable.
  5. There needs to be homogeneity of variances.

- Report of results

$$F_{1,19} = 39.28,\ p < .001$$

http://en.wikipedia.org/wiki/One-way_analysis_of_variance
https://statistics.laerd.com/spss-tutorials/one-way-anova-using-spss-statistics.php

N-way ANOVA

- Usage

  An extension of the one-way ANOVA, that examined the influence of n categorical independent variable on one continuous dependent variable. Not only aims at assessing the main effect of each independent variable but also if there is any interaction between them.

- Assumptions

  1. Dependent variable should be measured on a continuous scale.
  2. Independent variables should each consist of two or more categorical, independent groups.
  3. There should be no significant outliers.
  4. Dependent variable should be approximately normally distributed for each combination of the groups of the independent variables.
  5. There needs to be homogeneity of variances for each combination of the groups of the independent variables.

- Report of results

$$F_{1,19} = 39.28,\ p < .001$$

http://en.wikipedia.org/wiki/Two-way_analysis_of_variance
https://statistics.laerd.com/spss-tutorials/two-way-anova-using-spss-statistics.php

Repeated measures ANOVA

- Usage

  Compare means of two or more paired samples.

- Assumptions

  1. Dependent variable should be measured on a continuous scale.

  2. Independent variable should consist of two or more categorical, independent groups.

  3. There should be no significant outliers.

  4. Dependent variable should be approximately normally distributed for each group of the independent variable.

  5. Sphericity: The variances of the differences between all combinations of related groups (levels) are equal. (This assumption only applies if there are more than 2 levels of the independent variable)

- Report of results

$$F_{1,19} = 39.28,\ p < .001$$

http://en.wikipedia.org/wiki/Repeated_measures_design
https://statistics.laerd.com/statistical-guides/repeated-measures-anova-statistical-guide.php

MANOVA

- Usage

  Determine whether there are any differences between independent groups on more than one continuous dependent variable.

- Assumptions

  Normality: Responses for a given group are i.i.d normal random variables.

  1. Dependent variables should be measured on a continuous scale.

  2. Independent variable should consist of two or more categorical, independent groups.

  3. You need to have more cases in each group than the number of dependent variables you are analyzing.

  3. There are no univariate or multivariate outliers.

  4. There is multivariate normality.

  5. There is a linear relationship between each pair of dependent variables for each group of the independent variable.

  6. There is a homogeneity of variance-covariance matrices.

  7. There is no multicollinearity.

- Report of results

$$F_{1,19} = 39.28,\ p < .001$$

http://en.wikipedia.org/wiki/Multivariate_analysis_of_variance
https://statistics.laerd.com/spss-tutorials/one-way-manova-using-spss-statistics.php

Tukey-Kramer HSD test

- Usage

  A single-step multiple comparison procedure. can be used on raw data or in conjunction with an ANOVA (Post-hoc analysis) to find means that are significantly different from each other.

- Assumptions

  No more than ANOVA.

- Report of results

$$p < .001$$

http://en.wikipedia.org/wiki/Tukey%27s_range_test

Tests of Significance

Parametric tests

- Student's *t* test
- Paired-samples *t* test
- One-way ANOVA
- N-way ANOVA
- Repeated measures ANOVA
- MANOVA
- Tukey-Kramer HSD test

Categories, counts and proportion tests

- One-sample Pearson Chi-Square (χ^2) test
- Two-sample Pearson Chi –Square (χ^2) test
- Fisher's exact test

Nonparametric tests

- Mann-Whitney *U* test
- Wilcoxon rank-sum test
- Wilcoxon signed-rank test
- Kruskal-Wallis test
- Friedman test

Auxiliary tests

- Shapiro-Wilk *W* test
- Kolmogorov-Smirnov *D* test
- Mauchly's sphericity test
- Levene's test

Wobbrock, J. O. Practical statistics for human-computer interaction. In*Annual Workshop of the HCI Consortium (HCIC'11)*.

Tests of Significance

Parametric tests

- Student's *t* test
- Paired-samples *t* test
- One-way ANOVA
- N-way ANOVA
- Repeated measures ANOVA
- MANOVA
- Tukey-Kramer HSD test

Categories, counts and proportion tests

- One-sample Pearson Chi-Square (χ^2) test
- Two-sample Pearson Chi –Square (χ^2) test
- Fisher's exact test

Nonparametric tests

- Mann-Whitney *U* test
- Wilcoxon rank-sum test
- Wilcoxon signed-rank test
- Kruskal-Wallis test
- Friedman test

Auxiliary tests

- Shapiro-Wilk *W* test
- Kolmogorov-Smirnov *D* test
- Mauchly's sphericity test
- Levene's test

Wobbrock, J. O. Practical statistics for human-computer interaction. In*Annual Workshop of the HCI Consortium (HCIC'11).*

Shapiro-Wilk *W* test

- Usage

  Test the normality in frequentist statistics.

- Assumptions

  None.

- Report of results

$$W = .97,\ p = .39$$

http://en.wikipedia.org/wiki/Shapiro%E2%80%93Wilk_test
https://statistics.laerd.com/spss-tutorials/testing-for-normality-using-spss-statistics.php

Kolmogorov-Smirnov *D* test

- Usage

  Compare a continuous, one-dimensional probability distribution with a reference probability distribution (one-sample K–S test), or compare two samples (two-sample K–S test).

- Assumptions

  None.

- Report of results

$$D = .09,\ p = .15$$

http://en.wikipedia.org/wiki/Kolmogorov%E2%80%93Smirnov_test

Mauchly's sphericity test

- Usage

  Test the sphericity of a repeated measures ANOVA. If sphericity is violated, a decision must be made as to whether a univariate or multivariate analysis is selected. If a univariate method is selected, the degrees of freedom must be appropriately corrected (e.g. the Greenhouse-Geisser correction).

- Assumptions

  None.

- Report of results

$$\chi^2(2) = 16.8,\ p < .001$$

http://en.wikipedia.org/wiki/Mauchly%27s_sphericity_test
https://statistics.laerd.com/statistical-guides/sphericity-statistical-guide.php

Levene's test

- Usage

  Assess the equality of variances for a variable calculated for two or more groups.

- Assumptions

  None.

- Report of results

$$F_{1,19} = 39.28,\ p < .001$$

http://en.wikipedia.org/wiki/Levene%27s_test

Parametric tests

- Student's *t* test
- Paired-samples *t* test
- One-way ANOVA
- N-way ANOVA
- Repeated measures ANOVA
- MANOVA
- Tukey-Kramer HSD test

Categories, counts and proportion tests

- One-sample Pearson Chi-Square (χ^2) test
- Two-sample Pearson Chi –Square (χ^2) test
- Fisher's exact test

Nonparametric tests

- Mann-Whitney *U* test
- Wilcoxon rank-sum test
- Wilcoxon signed-rank test
- Kruskal-Wallis test
- Friedman test

Auxiliary tests

- Shapiro-Wilk *W* test
- Kolmogorov-Smirnov *D* test
- Mauchly's sphericity test
- Levene's test

Wobbrock, J. O. Practical statistics for human-computer interaction. In *Annual Workshop of the HCI Consortium (HCIC'11).*

Tests of Significance

Tests of Significance

Parametric tests

- Student's *t* test
- Paired-samples *t* test
- One-way ANOVA
- N-way ANOVA
- Repeated measures ANOVA
- MANOVA
- Tukey-Kramer HSD test

Categories, counts and proportion tests

- One-sample Pearson Chi-Square (χ^2) test
- Two-sample Pearson Chi –Square (χ^2) test
- Fisher's exact test

Nonparametric tests

- Mann-Whitney *U* test
- Wilcoxon rank-sum test
- Wilcoxon signed-rank test
- Kruskal-Wallis test
- Friedman test

Auxiliary tests

- Shapiro-Wilk *W* test
- Kolmogorov-Smirnov *D* test
- Mauchly's sphericity test
- Levene's test

Wobbrock, J. O. Practical statistics for human-computer interaction. In *Annual Workshop of the HCI Consortium (HCIC'11).*

Mann-Whitney *U* test

- Usage

  Compare the mean or rank of two independent samples when the dependent variable is either ordinal or continuous, but not normally distributed.

- Assumptions

  1. Dependent variable should be measured at the ordinal or continuous level.
  2. Independent variable should consist of two categorical, independent groups.

- Report of results

$$U = 135.50,\ p < 0.05$$

http://en.wikipedia.org/wiki/Mann%E2%80%93Whitney_U_test
Wobbrock, J. O. Practical statistics for human-computer interaction. In *Annual Workshop of the HCI Consortium (HCIC'11).*
https://statistics.laerd.com/spss-tutorials/mann-whitney-u-test-using-spss-statistics.php

Wilcoxon rank-sum test

- Usage

  Compare the mean or rank of two independent samples when the dependent variable is either ordinal or continuous, but not normally distributed.

- Assumptions

  1. Dependent variable should be measured at the ordinal or continuous level.
  2. Independent variable should consist of two categorical, independent groups.

- Report of results

$$W = 345.50,\ p < 0.05$$

http://en.wikipedia.org/wiki/Mann%E2%80%93Whitney_U_test
Wobbrock, J. O. Practical statistics for human-computer interaction. In *Annual Workshop of the HCI Consortium (HCIC'11).*

Wilcoxon signed-rank test

- Usage

  Compare the mean or rank of two paired samples.

- Assumptions

  1. Dependent variable should be measured at the ordinal or continuous level.
  2. Independent variable should consist of two categorical, related groups.
  3. Distribution of the differences between the two related groups needs to be symmetrical in shape.

- Report of results

$$Z = -1.73,\ p < 0.05$$

http://en.wikipedia.org/wiki/Wilcoxon_signed-rank_test
Wobbrock, J. O. Practical statistics for human-computer interaction. In *Annual Workshop of the HCI Consortium (HCIC'11).*
https://statistics.laerd.com/spss-tutorials/wilcoxon-signed-rank-test-using-spss-statistics.php

Kruskal-Wallis test

- Usage

  Determine if there are statistically significant differences between two or more groups of an independent variable on a continuous or ordinal dependent variable.

- Assumptions

  1. Dependent variable should be measured at the ordinal or continuous level.
  2. Independent variable should consist of two or more categorical, independent groups.

- Report of results

$$\chi^2(2) = 16.8,\ p < .001$$

http://en.wikipedia.org/wiki/Kruskal%E2%80%93Wallis_one-way_analysis_of_variance

Wobbrock, J. O. Practical statistics for human-computer interaction. In *Annual Workshop of the HCI Consortium (HCIC'11)*.

https://statistics.laerd.com/spss-tutorials/kruskal-wallis-h-test-using-spss-statistics.php

Friedman test

- Usage

  Test for differences between groups when the dependent variable being measured is ordinal. It can also be used for continuous data that has violated the assumptions necessary to run the one-way ANOVA with repeated measures.

- Assumptions

  1. One group that is measured on three or more different occasions.
  2. Dependent variable should be measured at the ordinal or continuous level.
  3. Samples do NOT need to be normally distributed.

- Report of results

$$\chi^2(2) = 16.8,\ p < .001$$

http://en.wikipedia.org/wiki/Repeated_measures_design
Wobbrock, J. O. Practical statistics for human-computer interaction. In *Annual Workshop of the HCI Consortium (HCIC'11).*
https://statistics.laerd.com/spss-tutorials/friedman-test-using-spss-statistics.php

Tests of Significance

Parametric tests

- Student's *t* test
- Paired-samples *t* test
- One-way ANOVA
- N-way ANOVA
- Repeated measures ANOVA
- MANOVA
- Tukey-Kramer HSD test

Categories, counts and proportion tests

- One-sample Pearson Chi-Square (χ^2) test
- Two-sample Pearson Chi –Square (χ^2) test
- Fisher's exact test

Nonparametric tests

- Mann-Whitney *U* test
- Wilcoxon rank-sum test
- Wilcoxon signed-rank test
- Kruskal-Wallis test
- Friedman test

Auxiliary tests

- Shapiro-Wilk *W* test
- Kolmogorov-Smirnov *D* test
- Mauchly's sphericity test
- Levene's test

Wobbrock, J. O. Practical statistics for human-computer interaction. In *Annual Workshop of the HCI Consortium (HCIC'11).*

Tests of Significance

Parametric tests

- Student's *t* test
- Paired-samples *t* test
- One-way ANOVA
- N-way ANOVA
- Repeated measures ANOVA
- MANOVA
- Tukey-Kramer HSD test

Categories, counts and proportion tests

- One-sample Pearson Chi-Square (χ^2) test
- Two-sample Pearson Chi –Square (χ^2) test
- Fisher's exact test

Nonparametric tests

- Mann-Whitney *U* test
- Wilcoxon rank-sum test
- Wilcoxon signed-rank test
- Kruskal-Wallis test
- Friedman test

Auxiliary tests

- Shapiro-Wilk *W* test
- Kolmogorov-Smirnov *D* test
- Mauchly's sphericity test
- Levene's test

Wobbrock, J. O. Practical statistics for human-computer interaction. In *Annual Workshop of the HCI Consortium (HCIC'11).*

One-sample Pearson Chi-Square (χ^2) test

- Usage

  Determine whether the distribution of cases in a single categorical variable follows a known or hypothesized distribution.

- Assumptions

  1. One categorical variable.
  2. The groups of the categorical variable must be mutually exclusive.
  3. There must be at least 5 expected frequencies in each group of the categorical variable.

- Report of results

$$\chi^2(2, N=30) = 4.80,\ p < .05$$

http://en.wikipedia.org/wiki/Pearson%27s_chi-squared_test
Wobbrock, J. O. Practical statistics for human-computer interaction. In *Annual Workshop of the HCI Consortium (HCIC'11).*
https://statistics.laerd.com/spss-tutorials/chi-square-goodness-of-fit-test-in-spss-statistics.php

Two-sample Pearson Chi-Square (χ^2) test

- Usage

  Compare counts from two independent groups. Also, there is three-sample test, etc.

- Assumptions

  1. Two variables should be categorical.
  2. Two variable should consist of two or more categorical, independent groups.

- Report of results

$$\chi^2(2, N=30) = 4.80,\ p < .05$$

http://en.wikipedia.org/wiki/Pearson%27s_chi-squared_test
Wobbrock, J. O. Practical statistics for human-computer interaction. In *Annual Workshop of the HCI Consortium (HCIC'11).*
https://statistics.laerd.com/spss-tutorials/chi-square-test-for-association-using-spss-statistics.php

Fisher's exact test

- Usage

  Compare counts from independent groups.

- Assumptions

  1. 2 variables x 2 levels.
  2. Also valid when cell counts are low.

- Report of results

$$\chi^2(2, N=30) = 4.80,\ p < .05$$

http://en.wikipedia.org/wiki/Fisher%27s_exact_test
Wobbrock, J. O. Practical statistics for human-computer interaction. In*Annual Workshop of the HCI Consortium (HCIC'11).*

We used repeated measures ANOVAs and paired two-tailed *t*-tests for our analyses. All *post hoc* pairwise comparisons following the ANOVAs were protected against Type I error using a Bonferroni adjustment. Reported fractional degrees of freedom (*df*s) are from Greenhouse-Geisser adjustments. When parametric tests were not appropriate because the data violated the assumption of normality, we applied nonparametric equivalents, such as the Wilcoxon signed-rank test. We report significant findings at $p < .05$.

Findlater, L., Wobbrock, J. O., & Wigdor, D. (2011, May). Typing on flat glass: examining ten-finger expert typing patterns on touch surfaces. In *Proc. CHI'11* (pp. 2453-2462). ACM.

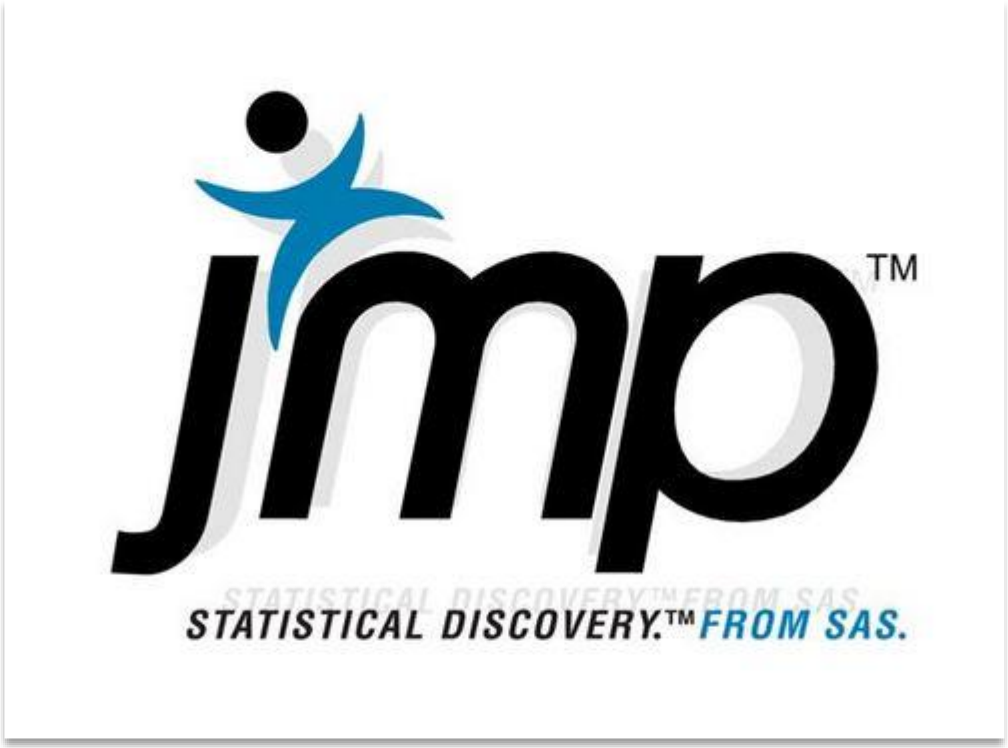
Tests of Significance

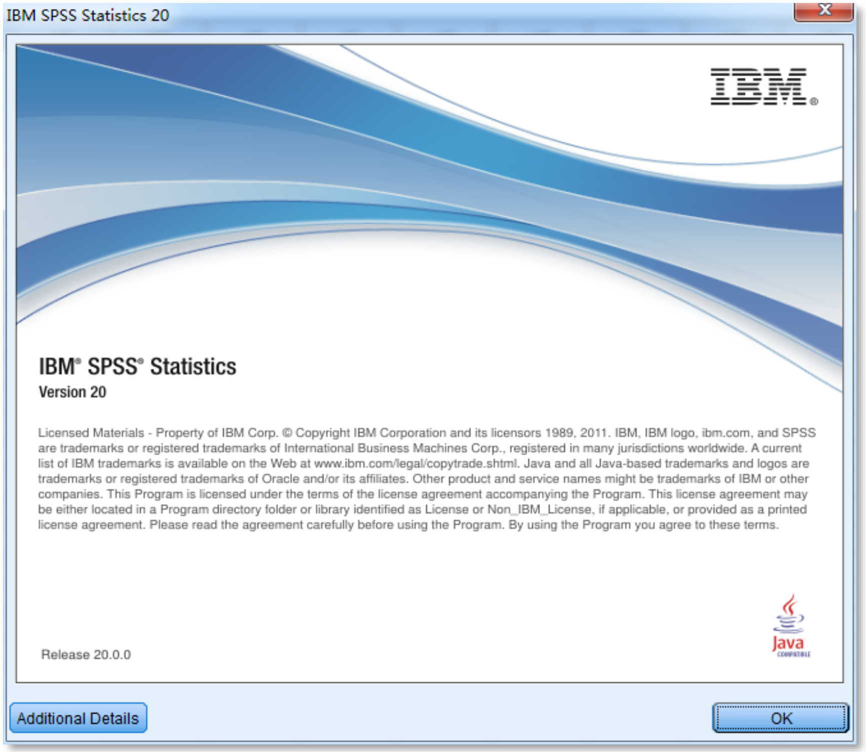
- The mean distance in the visible keyboard condition is more than the 0.9" of space between visual key centers ($t_{19} = 5.30$, $p < .001$).
- There was a main effect of keyboard for both x- and y-directions (x-direction: $F_{1,19} = 10.77$, $p = .004$; y-direction: $F_{1,19} = 39.28$, $p < .001$).
- We examine the highest-order effect in detail: a three-way interaction of keyboard × finger × row ($F_{2.8,34.3} = 5.70$, $p = .002$).
- A Wilxocon signed-rank test was not significant: $z = 1.45$, $p = .147$.
- Pairwise comparisons showed the keys assigned to the little finger had significantly greater x-direction deviation than the ring ($p = .033$) and middle fingers ($p = .024$), while comparison to the index finger was only a trend ($p = .075$).

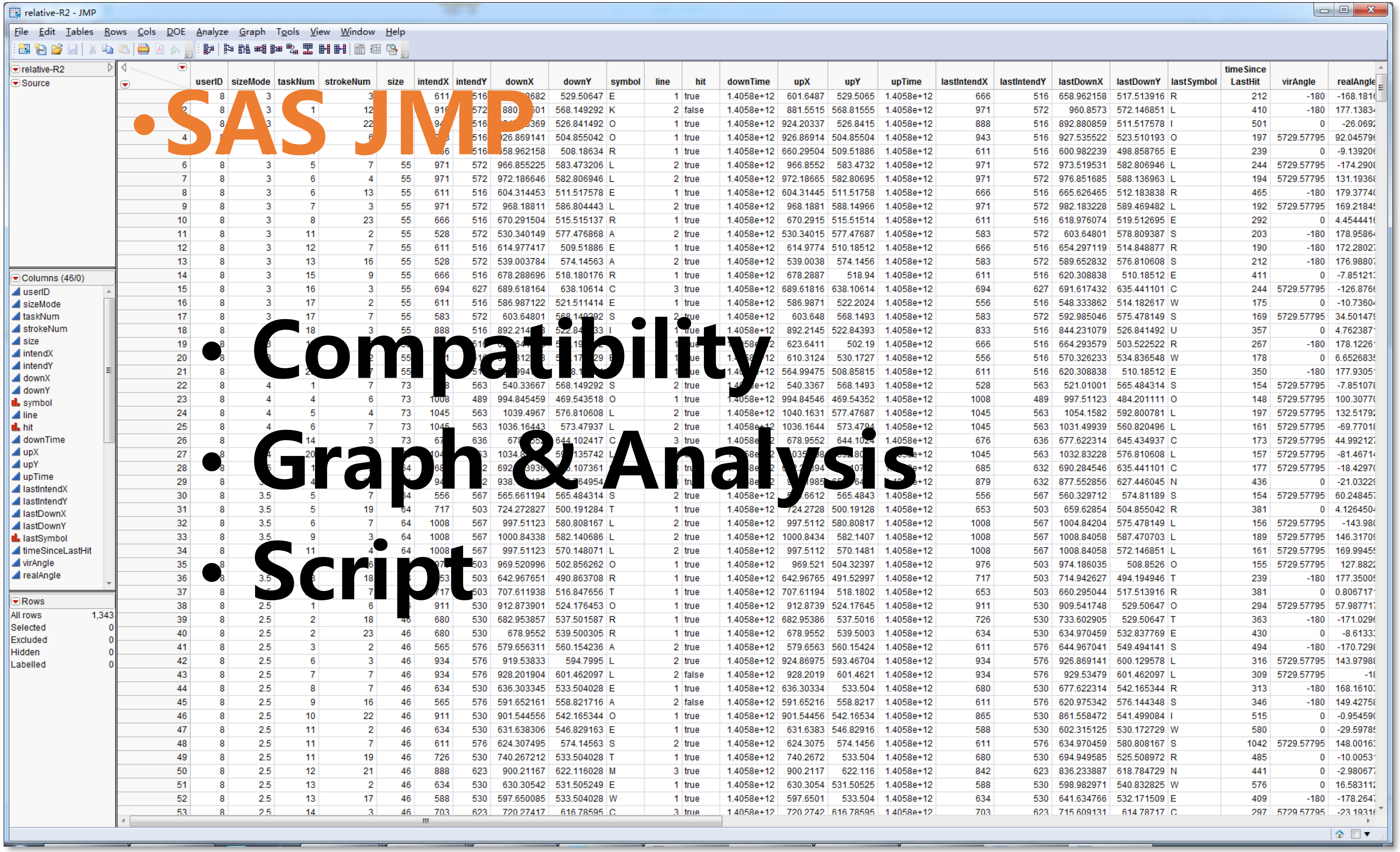
Findlater, L., Wobbrock, J. O., & Wigdor, D. (2011, May). Typing on flat glass: examining ten-finger expert typing patterns on touch surfaces. In *Proc. CHI'11* (pp. 2453-2462). ACM.

Tests of Significance

Software


jmp™
STATISTICAL DISCOVERY.™ FROM SAS.


IBM SPSS Statistics 20
IBM
IBM® SPSS® Statistics
Version 20
Licensed Materials - Property of IBM Corp. © Copyright IBM Corporation and its licensors 1989, 2011. IBM, IBM logo, ibm.com, and SPSS are trademarks or registered trademarks of International Business Machines Corp., registered in many jurisdictions worldwide. A current list of IBM trademarks is available on the Web at www.ibm.com/legal/copytrade.shtml. Java and all Java-based trademarks and logos are trademarks or registered trademarks of Oracle and/or its affiliates. Other product and service names might be trademarks of IBM or other companies. This Program is licensed under the terms of the license agreement accompanying the Program. This license agreement may be either located in a Program directory folder or library identified as License or Non_IBM_License, if applicable, or provided as a printed license agreement. Please read the agreement carefully before using the Program. By using the Program you agree to these terms.
Release 20.0.0
Java
Additional Details
OK



- SAS JMP
  - Compatibility
  - Graph & Analysis
  - Script

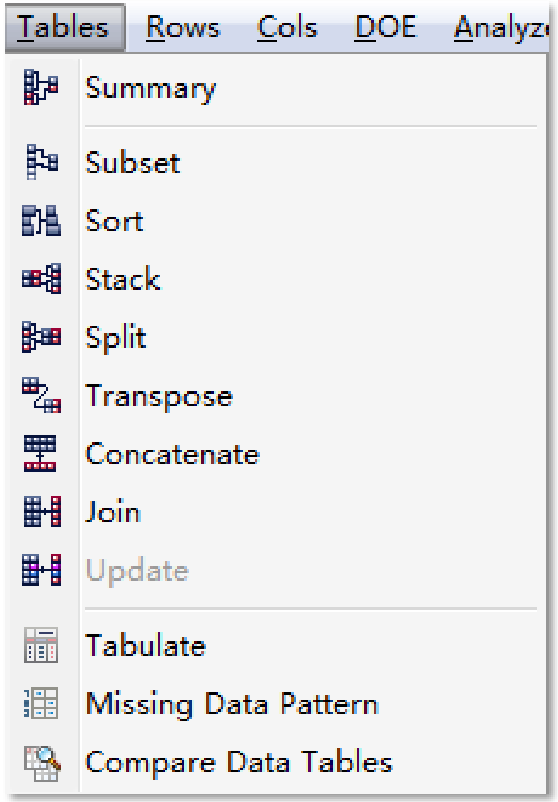
http://www.jmp.com/en_us/software/jmp.html

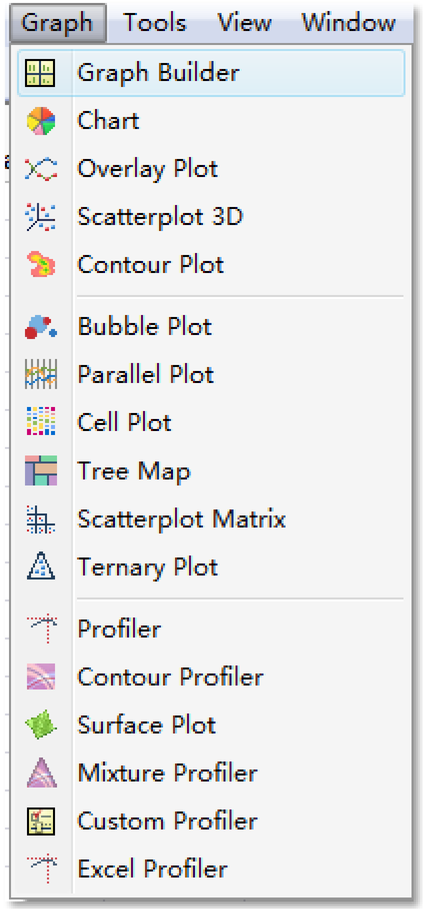
Software & Example

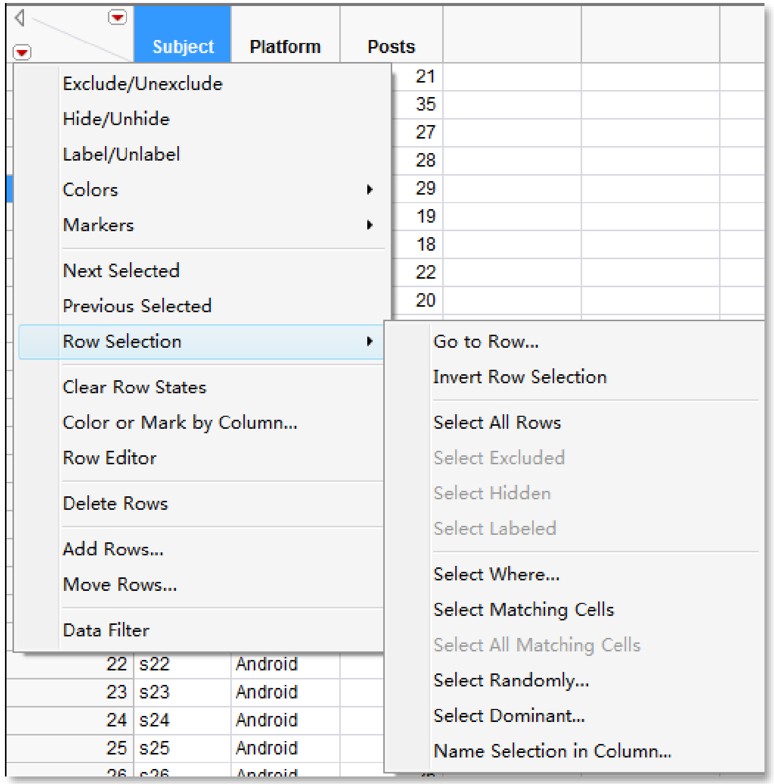
- **Compatibility**

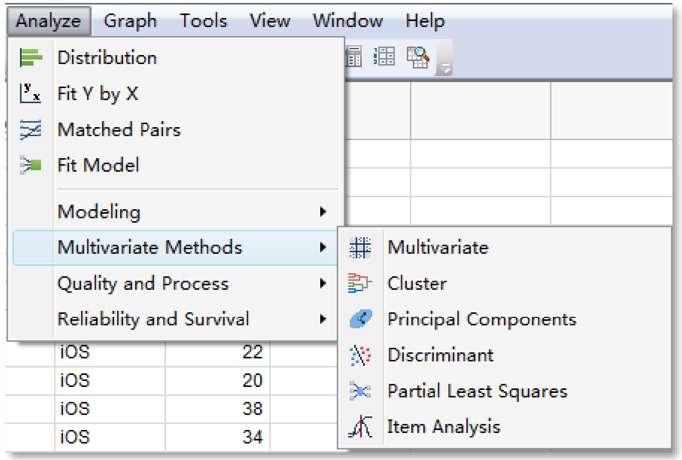
Data Files (*.jmp;*.sas7bdat;*.sd7;*.xpt;*.stx;*.ssd01;*.saseb$data;*.ssd;*.sd2;*.sd5)
JMP Files (*.jmp;*.jsl;*.jrn;*.jrp;*.jmpprj;*.jmpmenu)
JMP Data Tables (*.jmp)
Excel Files (*.xls;*.xlsx;*.xlsm)
Text Files (*.txt;*.csv;*.dat;*.tsv)
JMP Scripts (*.jsl)
JMP Journals (*.jrn)
JMP Reports (*.jrp)
JMP Projects (*.jmpprj)
JMP Add-In Files (*.jmpaddin;*.jmpaddindef;*.def)
JMP Menu Files (*.jmpmenu;*.jmpcust)
JMP Application Files (*.jmpappsource;*.jmpapp)
SAS Data Sets (*.sas7bdat;*.sd7;*.sd2;*.sd5;*.ssd01;*.saseb$data;*.ssd;*.xpt;*.stx)
SAS Program Files (*.sas)
R Code (*.R)
HTML Files (*.htm;*.html)
FACS Files (*.fcs)
SPSS Data Files (*.sav)
xBase Data Files (*.dbf)
Shapefiles (*.shp)
Minitab Portable Worksheet Files (*.mtp)
All Files (*.*)

- **Graph & Analysis**



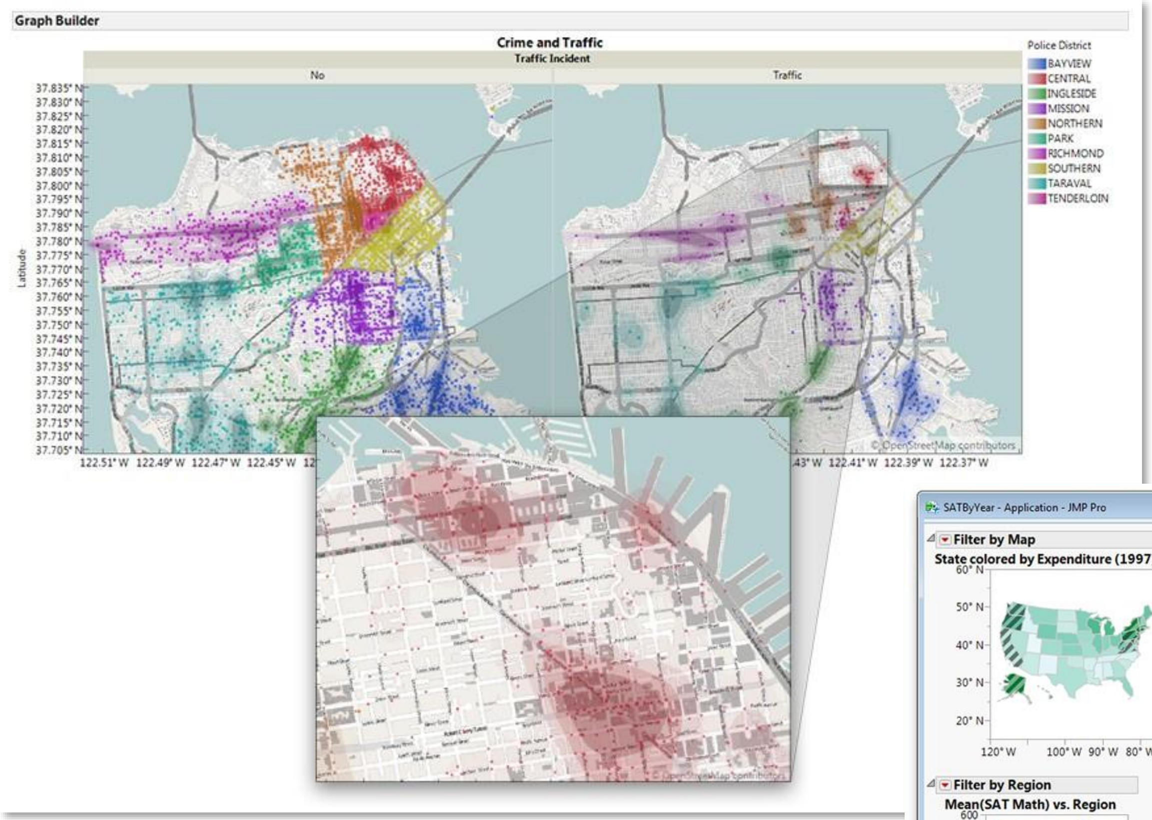


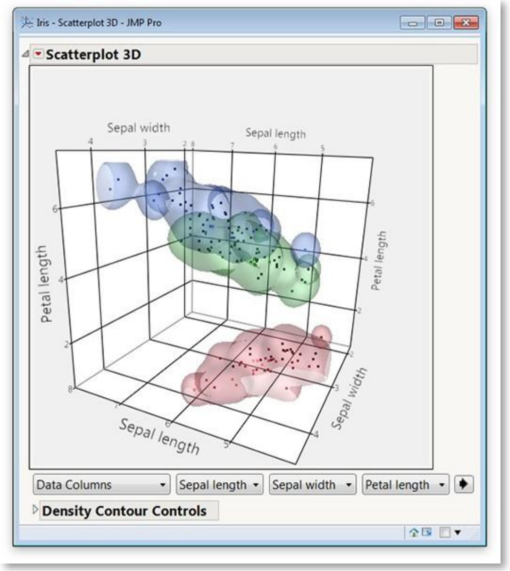


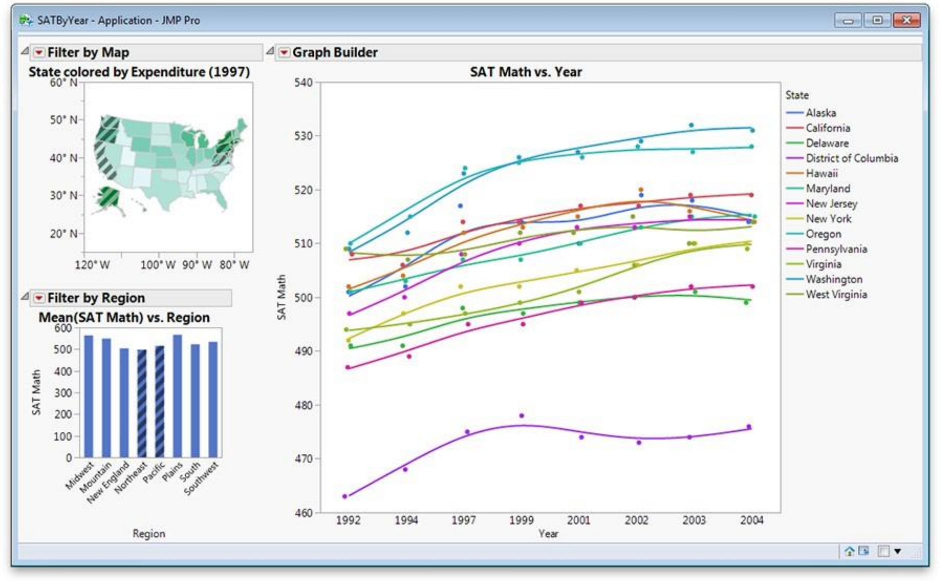


Software & Example

• Graph & Analysis

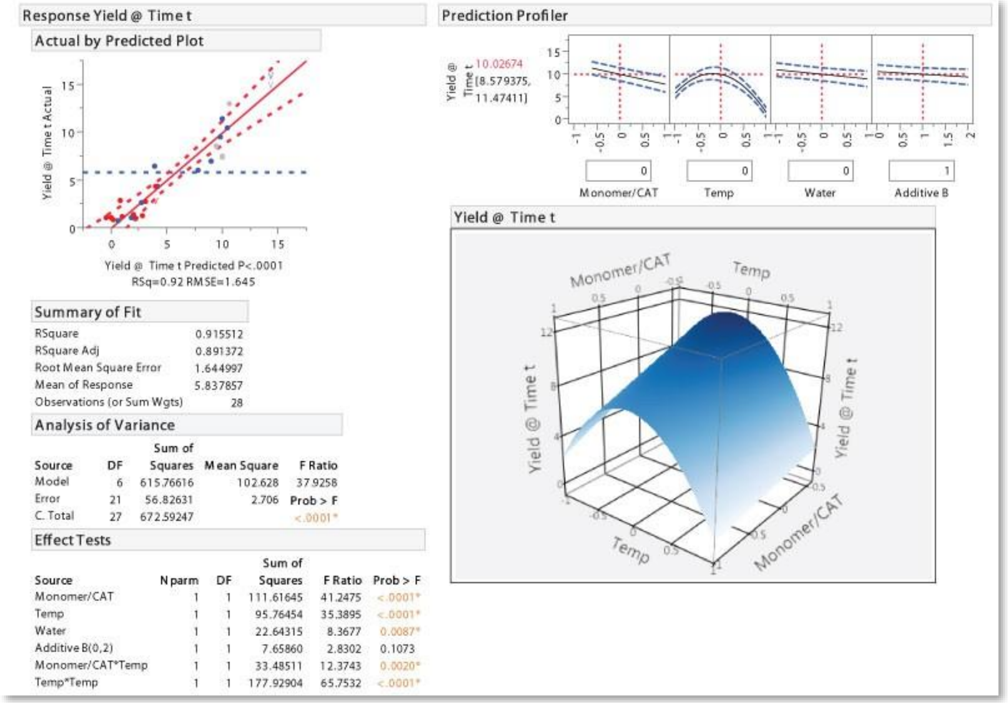


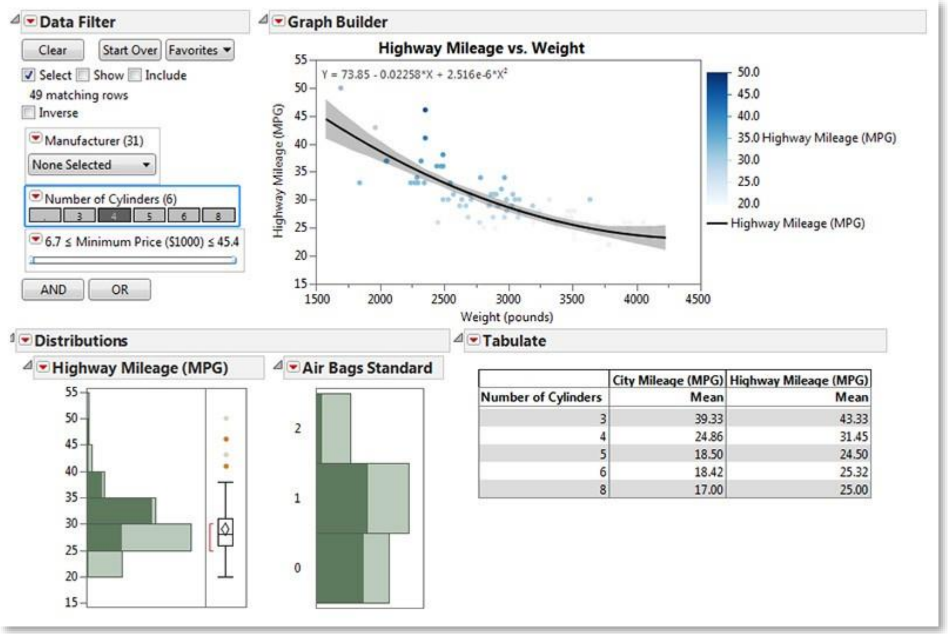


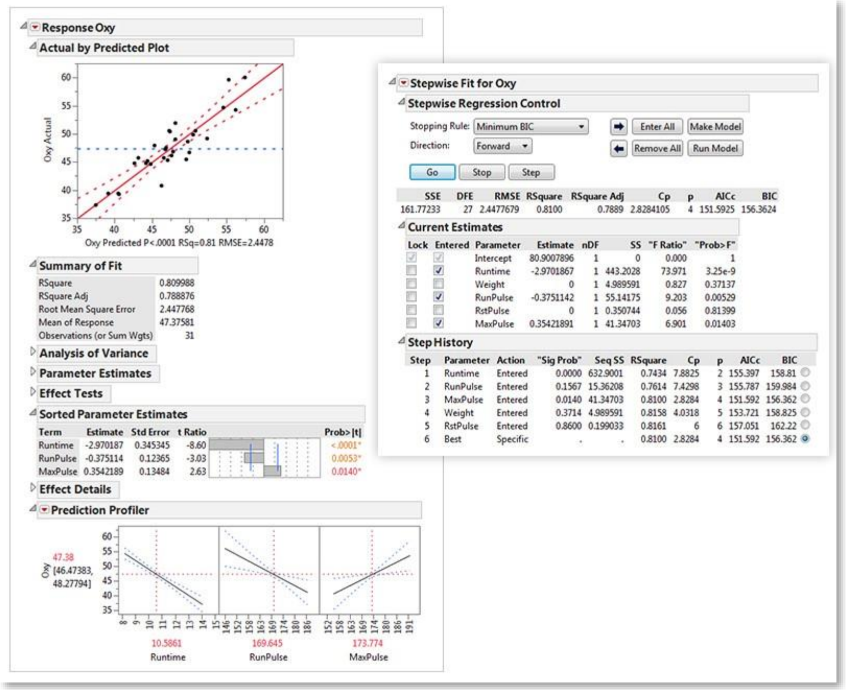


Software & Example

• **Graph & Analysis**







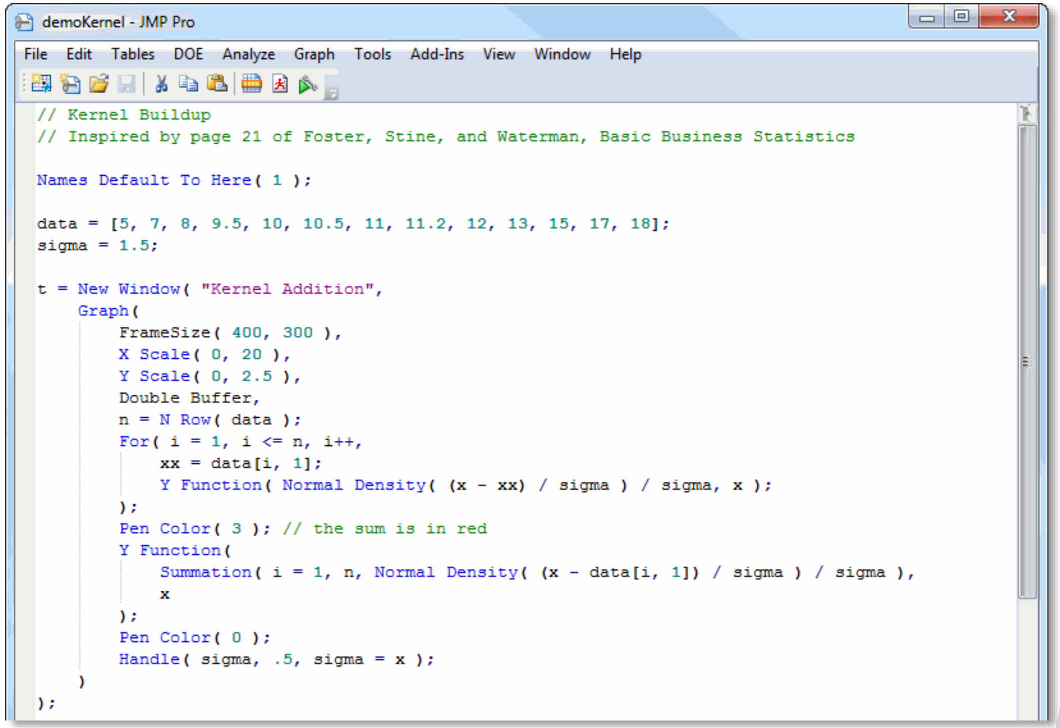
Software & Example

- **Script**

```
Distribution(
Nominal Distribution( Column( :sex ) ),
Continuous Distribution( Column( :height ) ),
Continuous Distribution( Column( :weight ) )
);
```

```
// Expression 1
sum=0; for(i=1,i<=10,i++,sum+=i;show(i,sum))

// Expression 2
Sum = 0;
For( i = 1, i <= 10, i++,
Sum += i;
Show( i, Sum );
);
```

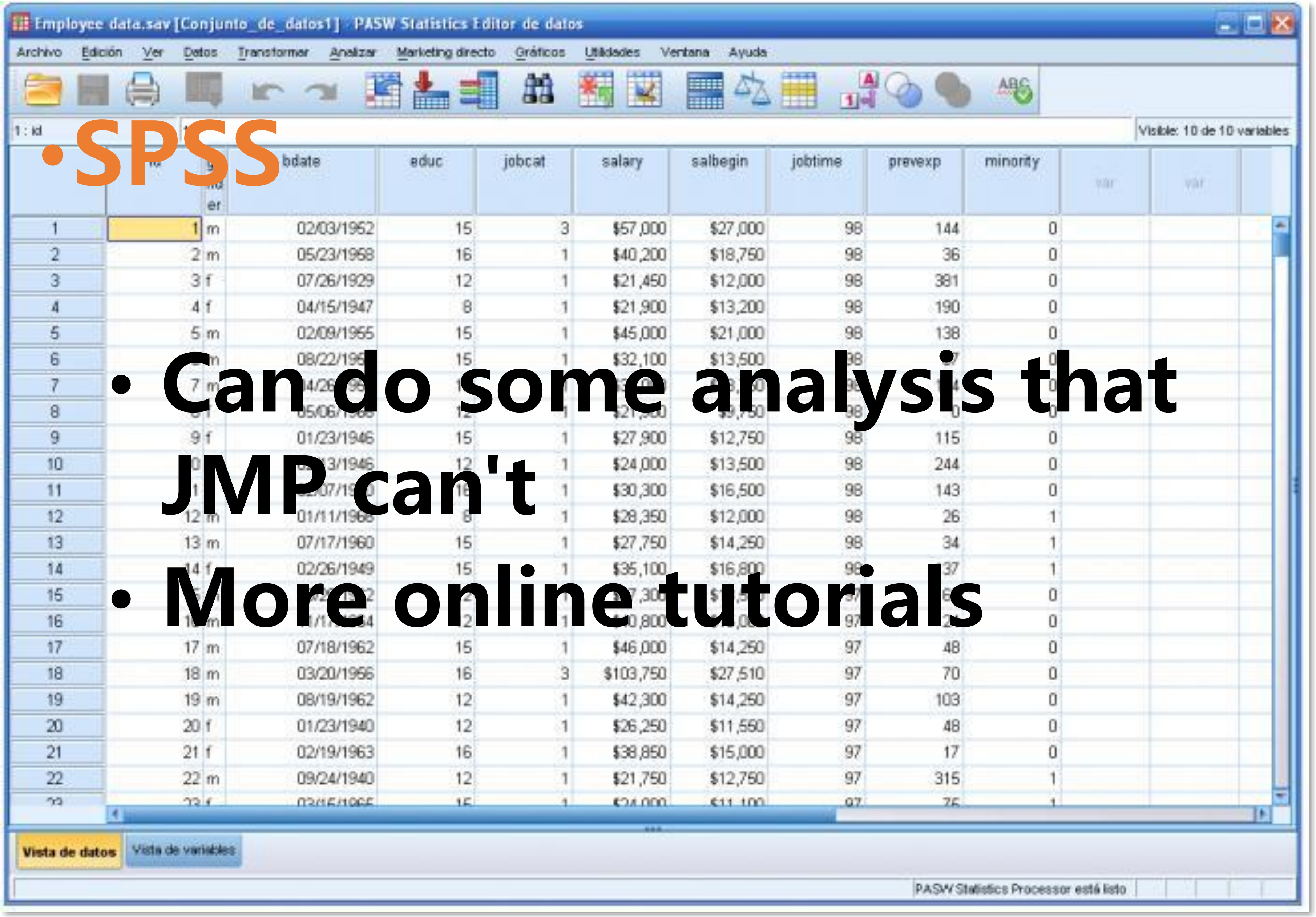
```
Graph Box( title("My Line Graph"),
Frame Size( 300, 500 ),
Marker( Marker State( 3 ), [11 44 77], [75 25 50] );
Pen Color( "Blue" );
Line( [10 30 70], [88 22 44] ));
```




```
// Kernel Buildup
// Inspired by page 21 of Foster, Stine, and Waterman, Basic Business Statistics

Names Default To Here( 1 );

data = [5, 7, 8, 9.5, 10, 10.5, 11, 11.2, 12, 13, 15, 17, 18];
sigma = 1.5;

t = New Window( "Kernel Addition",
    Graph(
        FrameSize( 400, 300 ),
        X Scale( 0, 20 ),
        Y Scale( 0, 2.5 ),
        Double Buffer,
        n = N Row( data );
        For( i = 1, i <= n, i++,
            xx = data[i, 1];
            Y Function( Normal Density( (x - xx) / sigma ) / sigma, x );
        );
        Pen Color( 3 ); // the sum is in red
        Y Function(
            Summation( i = 1, n, Normal Density( (x - data[i, 1]) / sigma ) / sigma ),
            x
        );
        Pen Color( 0 );
        Handle( sigma, .5, sigma = x );
    )
);
```

Software & Example

SPSS



- Can do some analysis that JMP can't
- More online tutorials

http://www-03.ibm.com/software/products/en/spss-stats-premium

Software & Example

Parametric tests

- Student's *t* test
- Paired-samples *t* test
- One-way ANOVA
- N-way ANOVA
- Repeated measures ANOVA
- MANOVA
- Tukey-Kramer HSD test

Categories, counts and proportion tests

- One-sample Pearson Chi-Square (χ^2) test
- Two-sample Pearson Chi –Square (χ^2) test
- Fisher's exact test

Nonparametric tests

- Mann-Whitney *U* test
- Wilcoxon rank-sum test
- Wilcoxon signed-rank test
- Kruskal-Wallis test
- Friedman test

Auxiliary tests

- Shapiro-Wilk *W* test
- Kolmogorov-Smirnov *D* test
- Mauchly's sphericity test
- Levene's test

Wobbrock, J. O. Practical statistics for human-computer interaction. In *Annual Workshop of the HCI Consortium (HCIC'11).*

Software & Example

Parametric tests

- Student's *t* test (P15)
- Paired-samples *t* test
- One-way ANOVA (P19)
- N-way ANOVA (P21)
- Repeated measures ANOVA (P33)
- MANOVA
- Tukey-Kramer HSD test

Categories, counts and proportion tests

- One-sample Pearson Chi-Square (χ^2) test
- Two-sample Pearson Chi –Square (χ^2) test
- Fisher's exact test

Nonparametric tests

- Mann-Whitney *U* test
- Wilcoxon rank-sum test
- Wilcoxon signed-rank test
- Kruskal-Wallis test
- Friedman test (P84)

Auxiliary tests

- Shapiro-Wilk *W* test (P42)
- Kolmogorov-Smirnov *D* test (P43)
- Mauchly's sphericity test
- Levene's test (P88)

Wobbrock, J. O. Practical statistics for human-computer interaction. In *Annual Workshop of the HCI Consortium (HCIC'11).*

Software & Example

1/ Statistics Overview

2/ Statistical Hypothesis Testing

3/ Tests of Significance

4/ Software & Example

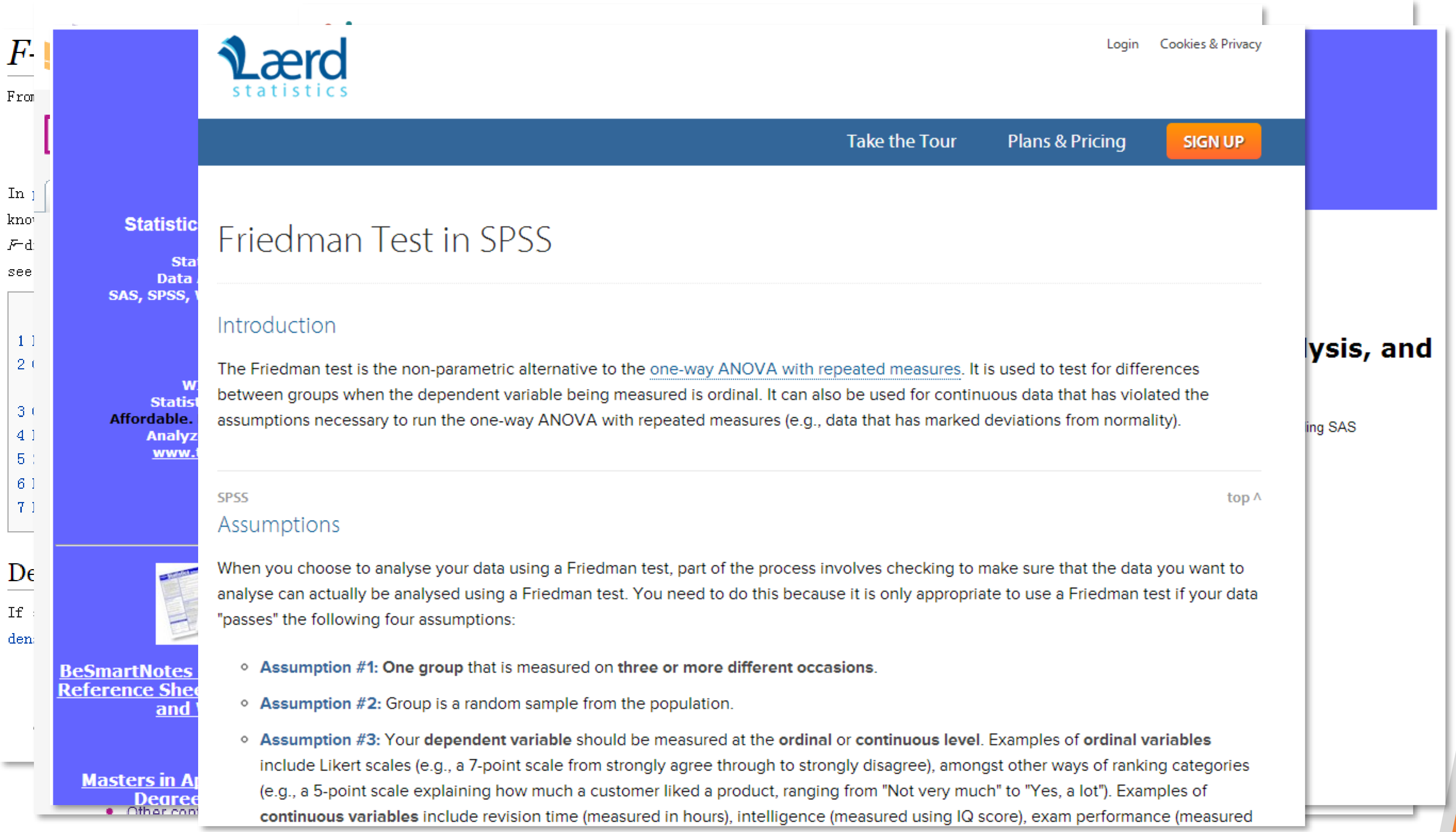
5/ References

Practical Statistics for Human-Computer Interaction

• References

- Website
  - Wikipedia
  - Graphpad Statistics Guide
  - Statistics Tutorials for Statistical Data Analysis: SAS, SPSS, WINKS, R, Excel
  - Statistical Computing
  - Laerd Statistics
- Book
  - Practical Statistics for Human-Computer Interaction
  - Pattern Recognition and Machine Learning
- Software
  - SAS JMP
  - SPSS

- **Website**



Login Cookies & Privacy

Laerd statistics

Take the Tour Plans & Pricing SIGN UP

Friedman Test in SPSS

Introduction

The Friedman test is the non-parametric alternative to the one-way ANOVA with repeated measures. It is used to test for differences between groups when the dependent variable being measured is ordinal. It can also be used for continuous data that has violated the assumptions necessary to run the one-way ANOVA with repeated measures (e.g., data that has marked deviations from normality).

SPSS top ^

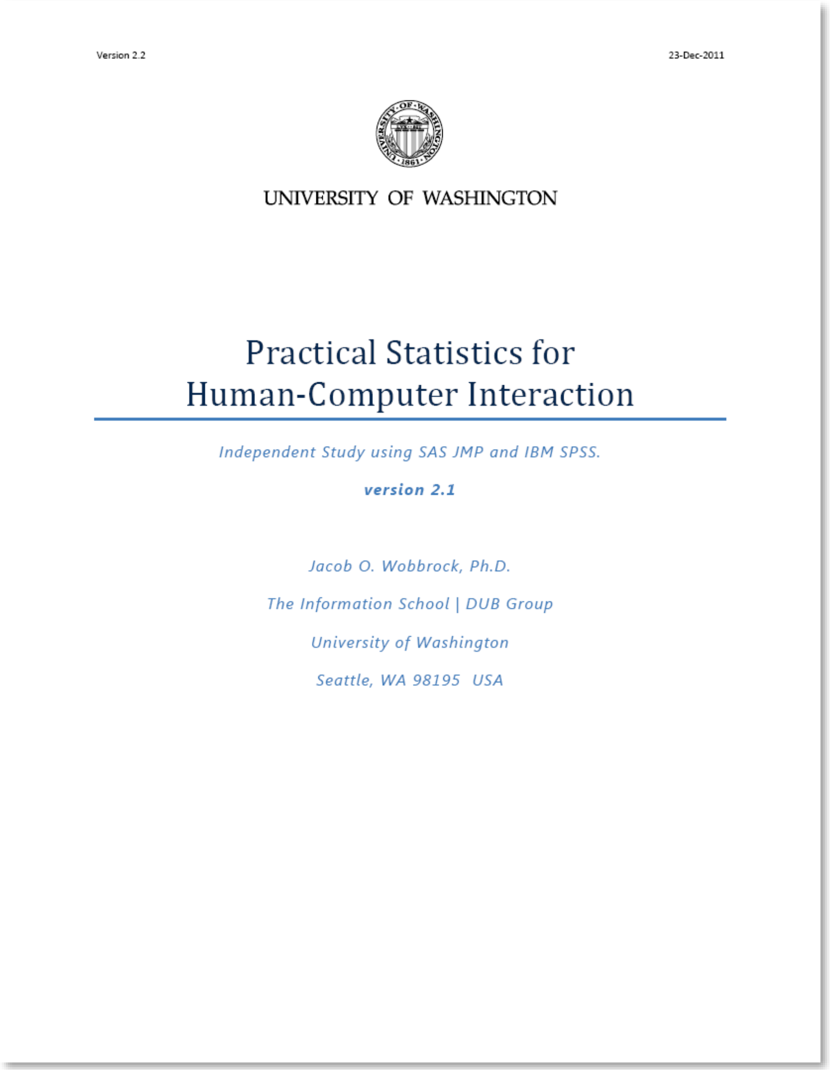
Assumptions

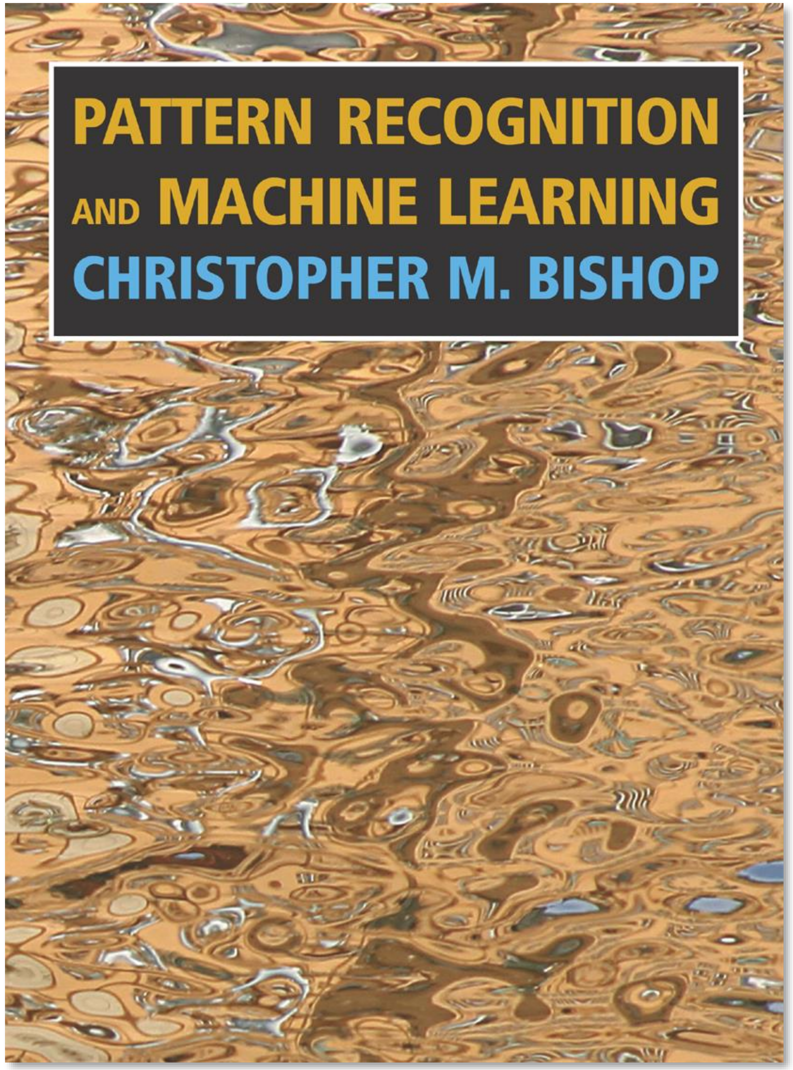
When you choose to analyse your data using a Friedman test, part of the process involves checking to make sure that the data you want to analyse can actually be analysed using a Friedman test. You need to do this because it is only appropriate to use a Friedman test if your data "passes" the following four assumptions:

- **Assumption #1:** **One group** that is measured on **three or more different occasions**.
- **Assumption #2:** Group is a random sample from the population.
- **Assumption #3:** Your **dependent variable** should be measured at the **ordinal** or **continuous level**. Examples of **ordinal variables** include Likert scales (e.g., a 7-point scale from strongly agree through to strongly disagree), amongst other ways of ranking categories (e.g., a 5-point scale explaining how much a customer liked a product, ranging from "Not very much" to "Yes, a lot"). Examples of **continuous variables** include revision time (measured in hours), intelligence (measured using IQ score), exam performance (measured

References

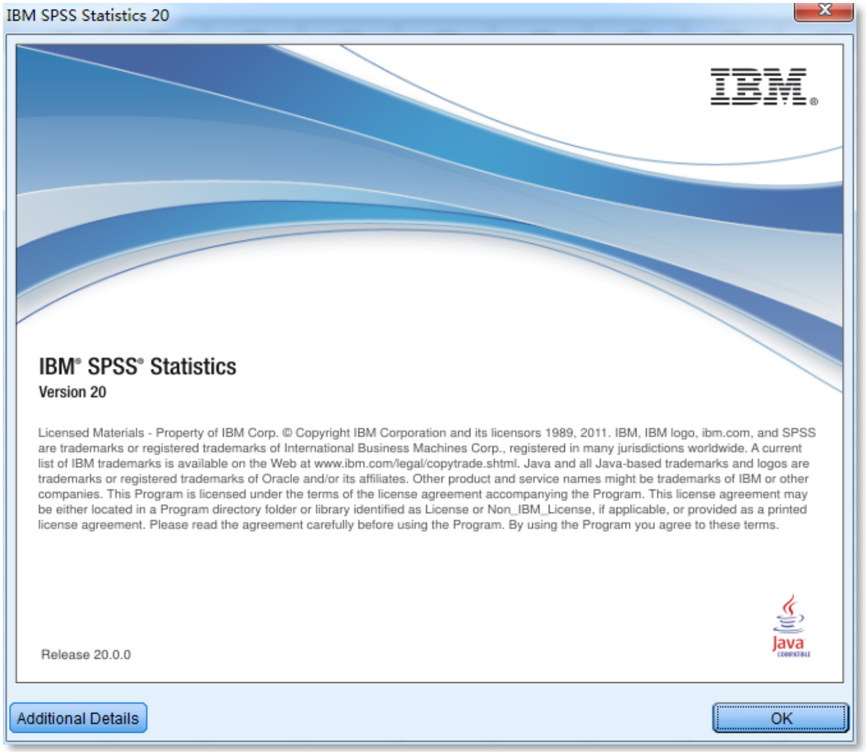
- Book





References

- Software



References

Thanks