Efficient subjective evaluation: Pair-wise comparisons

Rafal Mantiuk

Bangor University

UK

RIVIC

Research Institute of Visual Computing

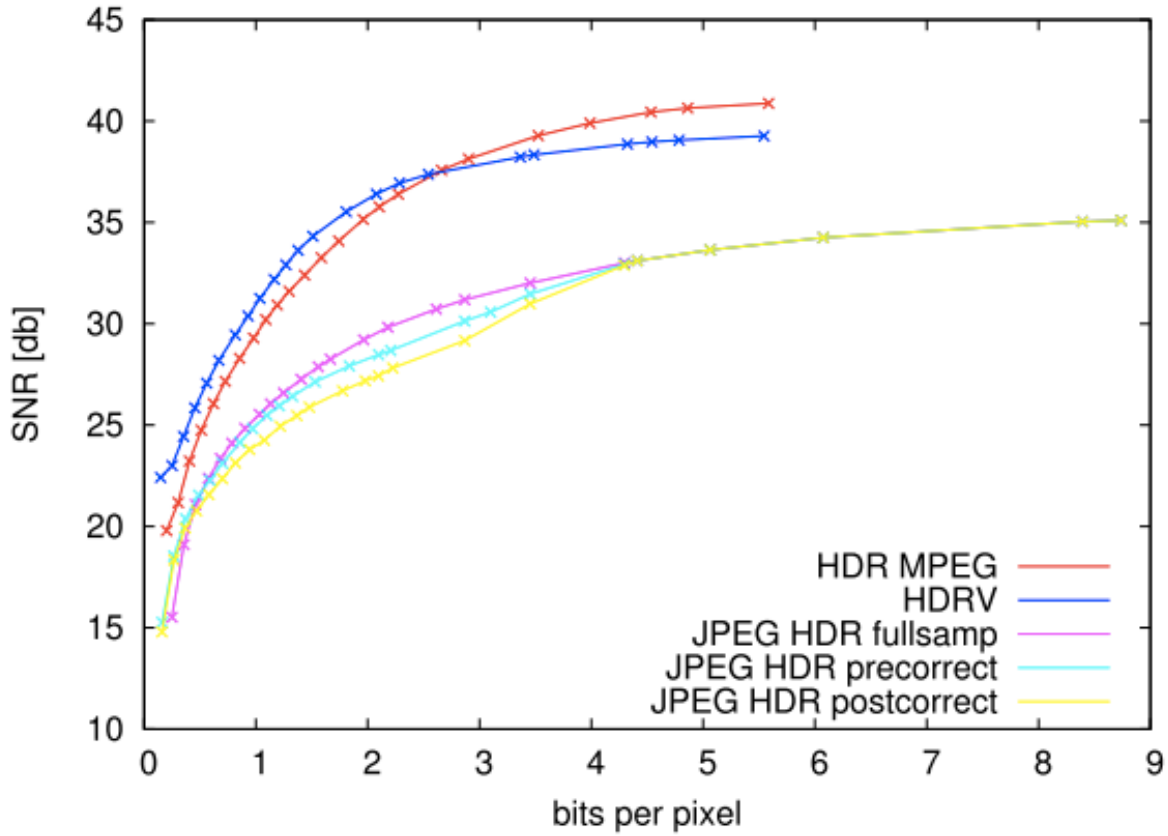
Research Institute of Visual Computing

Outline

- Why do we need quality assessment?
- Quality assessment – overview
- The method of pair-wise comparisons
- Basic statistics – review
- Pair-wise comparison – data analysis
  - Statistical significance
  - Practical significance

The purpose of quality assessment

- To compare algorithms in terms of image or video quality



The purpose of quality assessment

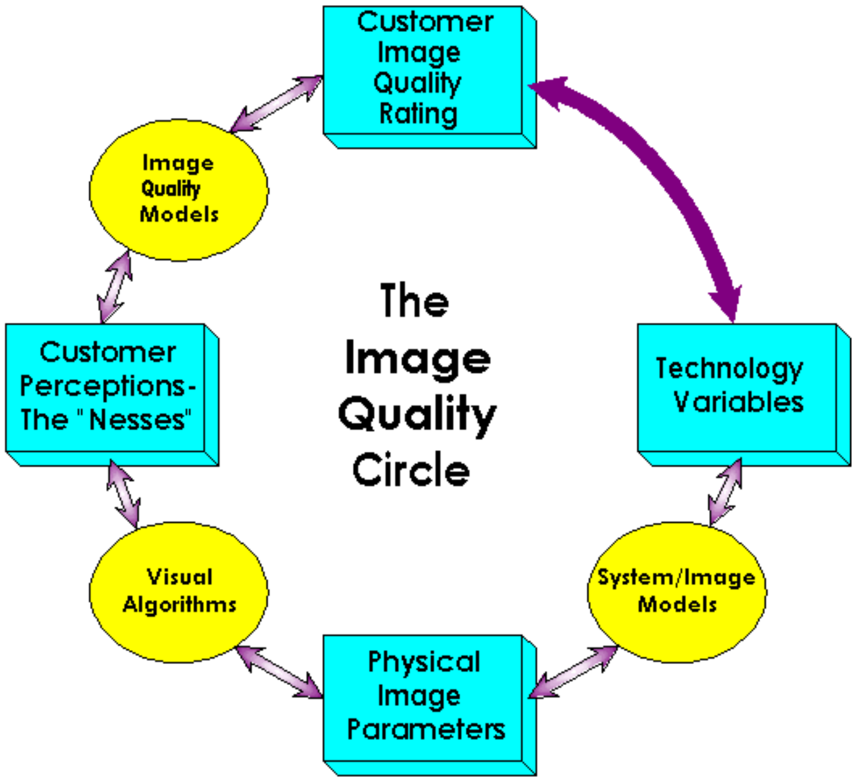
- To provide evidence of improvement over the state-of-the-art

Algorithm A　　Algorithm B　　Algorithm C

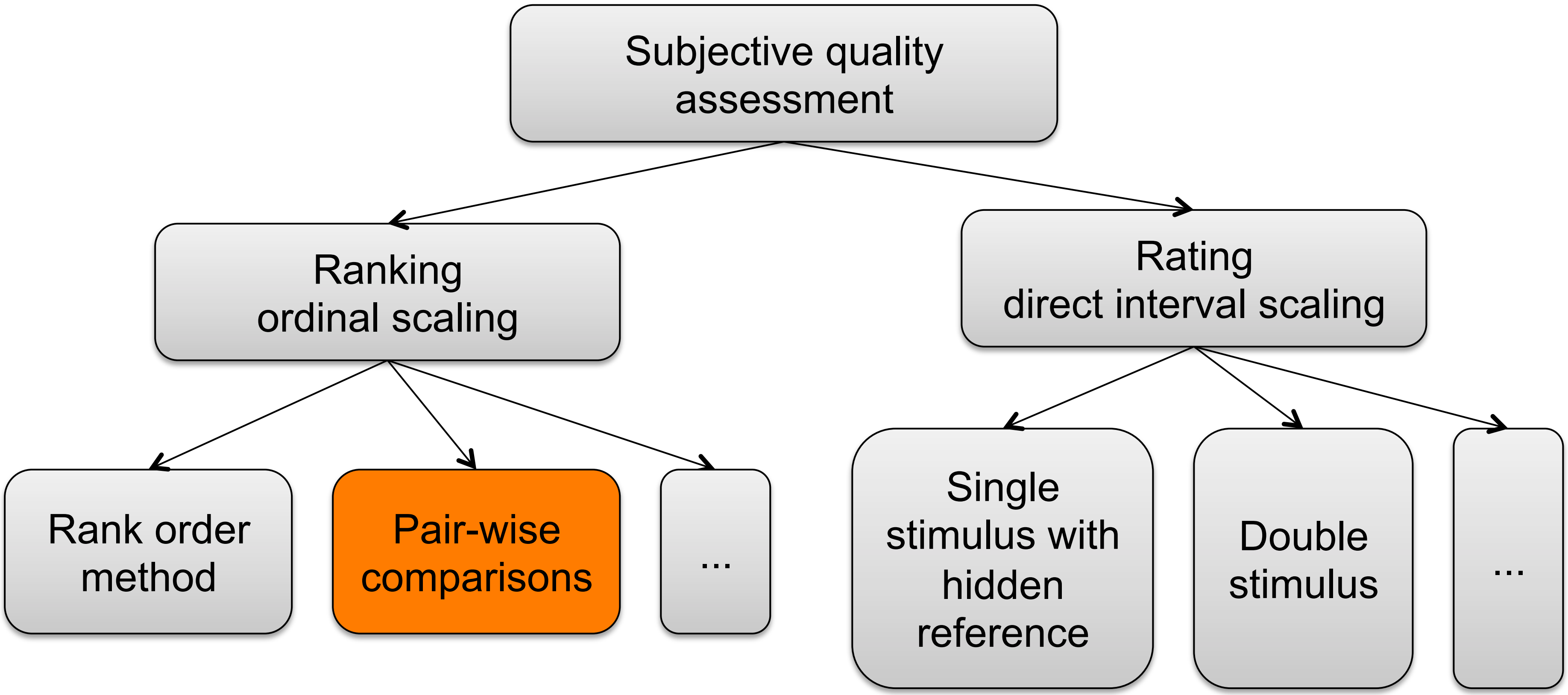
The purpose of quality assessment

- To optimize perceptual quality of a system
  - The best trade-off between cost and quality

- The impact of *technology variables* (resolution, contrast, etc.) on perceived image quality



Subjective quality assessment methods



Pair-wise comparison method

- Example: video quality
- Task: You will see two video sequences one after another. Select the sequence of higher quality.

Comparison matrix

- Results can be stored in a comparison matrix

$$
\begin{array}{cc}
 & \begin{array}{ccc} \text{C1} & \text{C2} & \text{C3} \end{array} \\
C = & \begin{bmatrix} 0 & 3 & 1 \\ 3 & 0 & 2 \\ 5 & 4 & 0 \end{bmatrix} \begin{array}{c} \text{C1} \\ \text{C2} \\ \text{C3} \end{array}
\end{array}
$$

- $C_{ij} = n$ means that
  - condition Cj was preferred over Ci *n* times

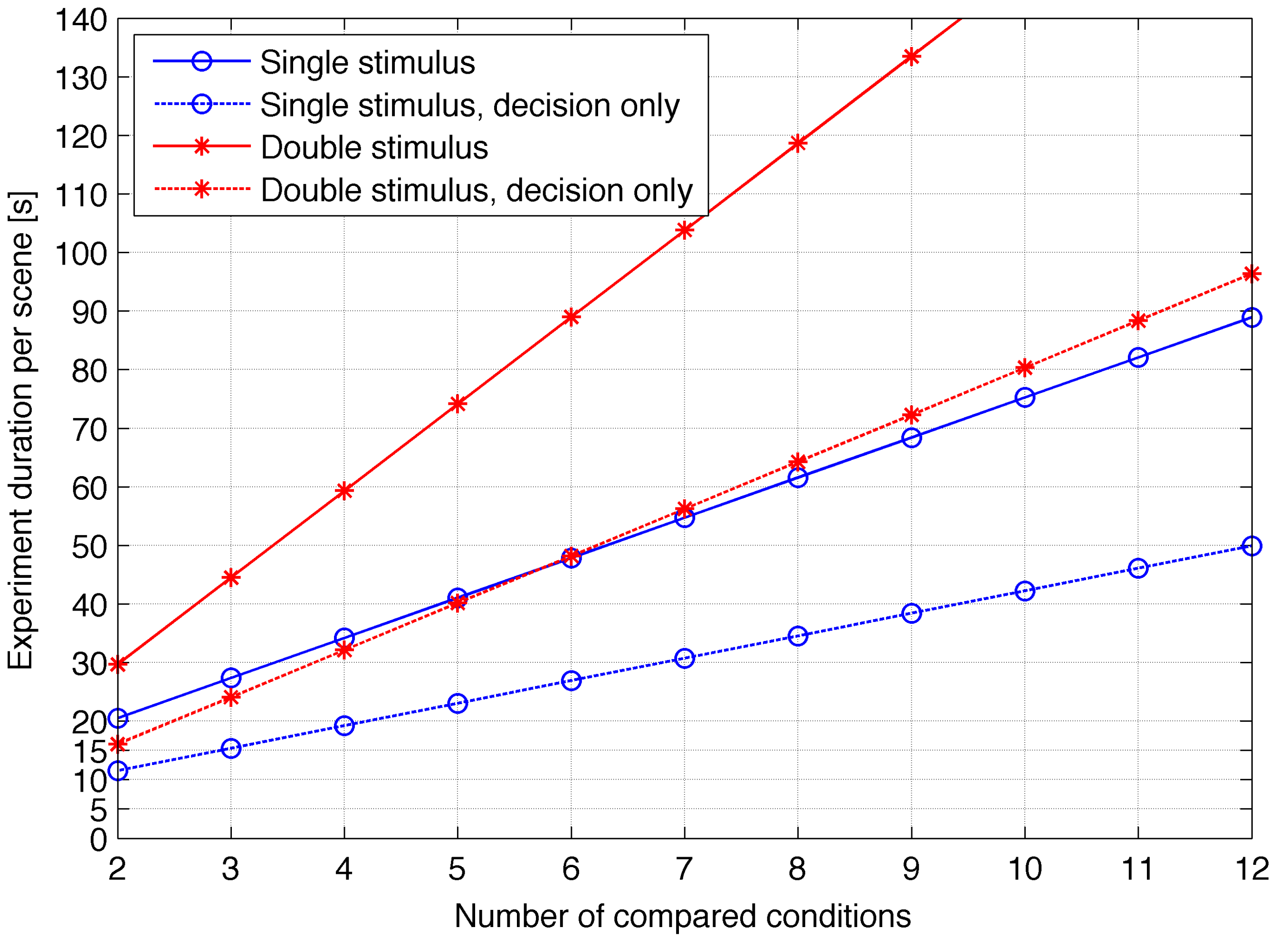
Full and reduced designs

- Full design
  - Compare all pairs of conditions
  - This requires *0.5*n*(n-1)* comparisons for *n* conditions
  - Tedious if *n* is large

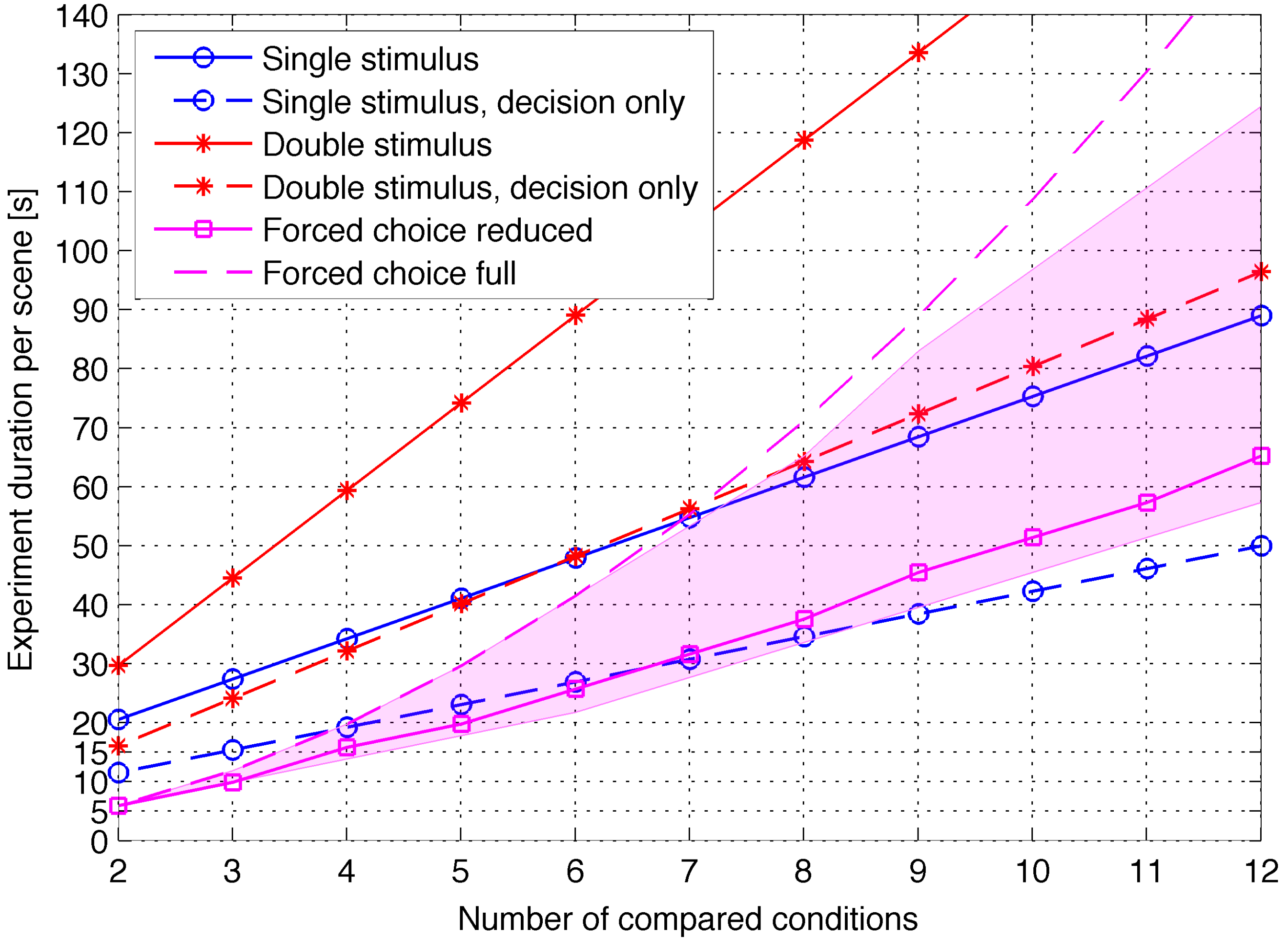
$$C = \begin{array}{cc} & \begin{array}{ccc} C1 & C2 & C3 \end{array} \\ & \begin{bmatrix} 0 & 3 & 1 \\ 3 & 0 & 2 \\ 5 & 4 & 0 \end{bmatrix} \begin{array}{c} C1 \\ C2 \\ C3 \end{array} \end{array}$$

- Reduced design
  - We assume transitivity
    - If C1 > C2 and C2 > C3 then C1 > C3
      - no need to do all comparisons
  - There are numerous “block designs” (before computers)
  - But the task is also a sorting problem
    - The number comparison can be reduced to *n*log(n)* for a “human quick-sort”

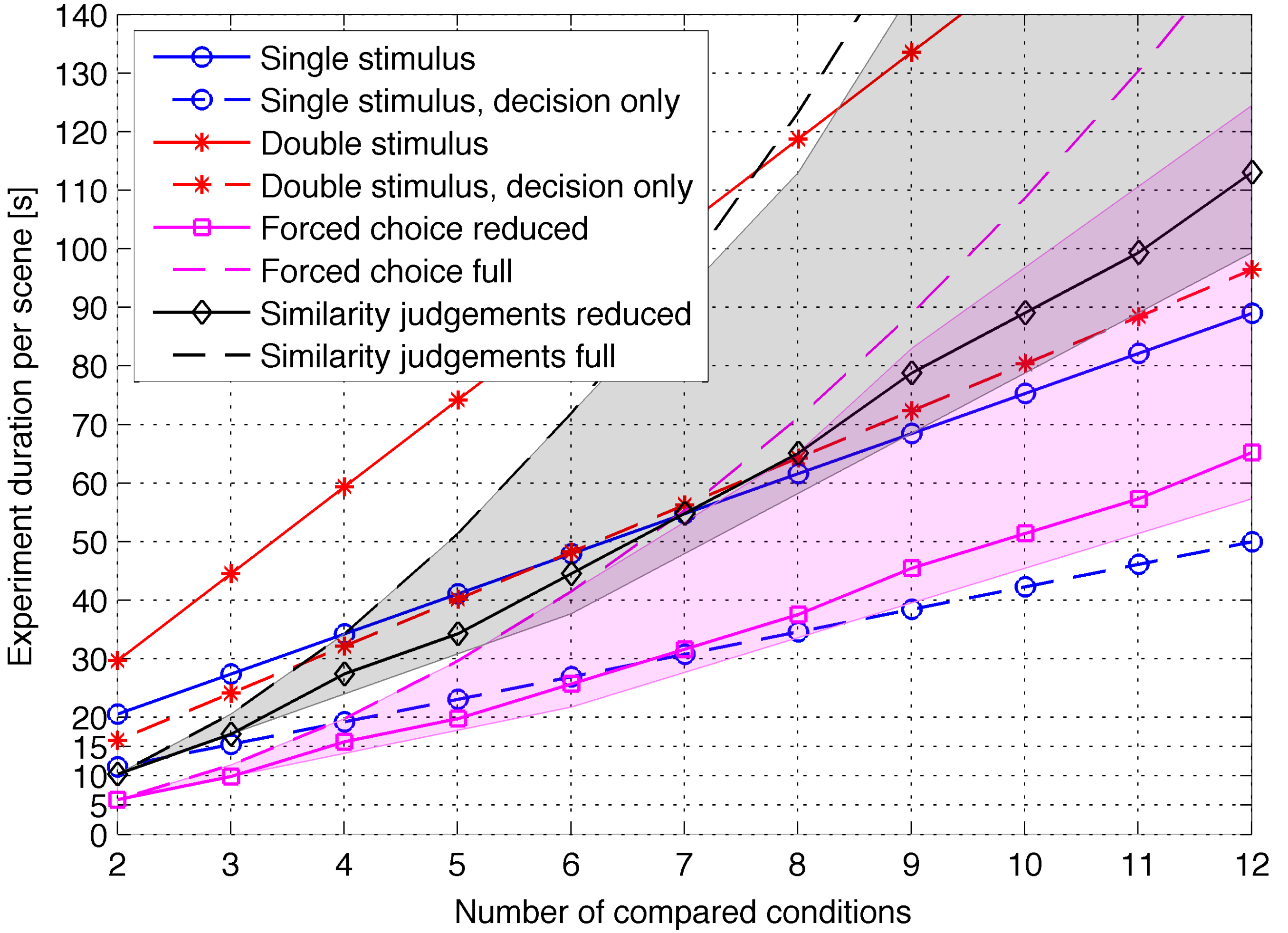
Time efficiency



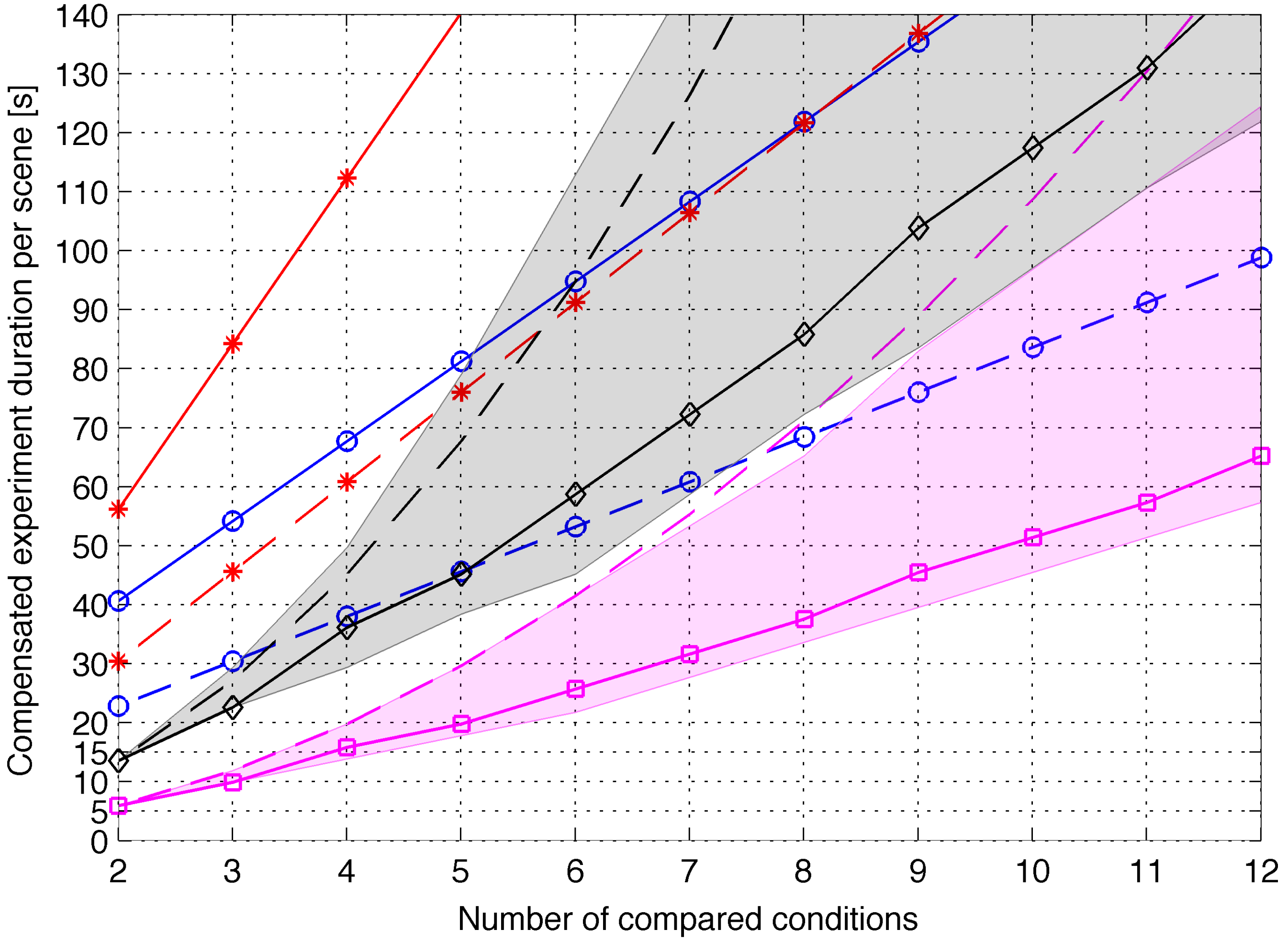
Time efficiency



Time efficiency



Time efficiency – corrected for the effect size



Learning effect

- Participants may
  - change their criteria,
  - become more sensitive (training),
  - become less sensitive (tiredness)
  - in the middle of the experiment
- To control the learning effect
  - Run training session
    - should cover the range of stimuli so that a participant can determine his/her criteria
  - Keep the sessions short (<20 min, <40 min)
  - Pay participants
  - Randomize stimuli (as much as possible)
    - To hide bias in the variance

Data collection

- Typical results file

observer	session	scene	condition_id_1	condition_id_2	select	criterion
jpt	3	window	irawan05	ledda04	1	perceptual
jpt	3	corridor	irawan05	ledda04	1	perceptual
jpt	3	corridor	irawan05	ferwerda96	1	perceptual
jpt	3	park	hateren06	irawan05	1	perceptual
jpt	3	park	hateren06	pattanaik00	0	perceptual
jpt	3	lab	irawan05	pattanaik00	1	perceptual
jpt	3	entrance	pattanaik00	ferwerda96	1	perceptual
jpt	3	window	irawan05	pattanaik00	1	perceptual
jpt	3	corridor	irawan05	benoit09	0	perceptual
jpt	3	entrance	pattanaik00	benoit09	0	perceptual
jpt	3	lab	irawan05	hateren06	1	perceptual

- Store it as CSV (comma-separated values)
- Matlab has great tools to analyze such data
  - Check statistical toolbox, “dataset” class

Experiment considerations

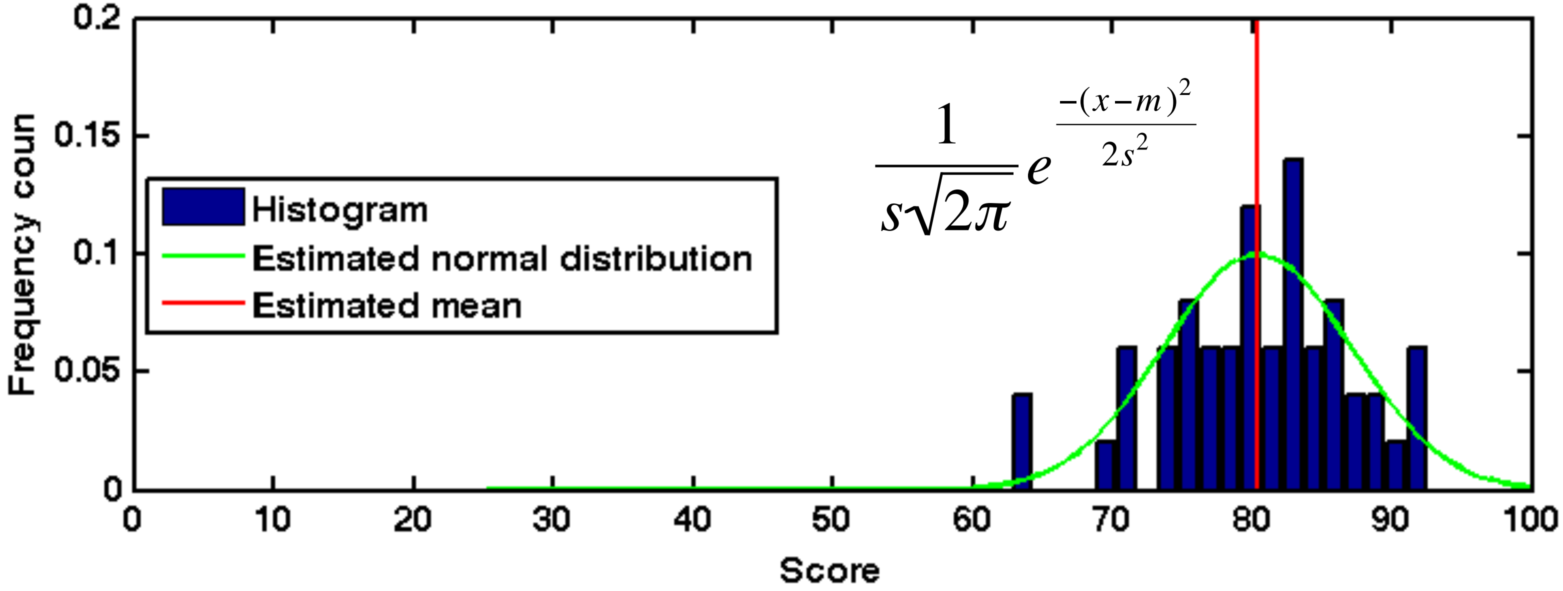
- How many observers?
  - Depends, but between 15 and 30 is usually sufficient
  - Retrospective power analysis can help finding the right number of observers
- Repeated measurements?
  - The same observer completes the experiment more than once
  - Makes the analysis more complicated - better avoided
  - Unless the data are averaged per participant before the analysis
- How many images?
  - It is very difficult to collect a representative sample
    - Some standards recommend using about 100 images – impractical in most case
  - Focus more on difficult / extreme case
  - Avoid averaging results over all images

Outline

- Why do we need quality assessment?
- Quality assessment – overview
- The method of pair-wise comparisons
- Basics of statistics – review
- Pair-wise comparison – data analysis
  - Statistical significance
  - Practical significance

Statistics - review

- Measured values are random variables
  - Example: 30 observers rated an image from 0 to 100



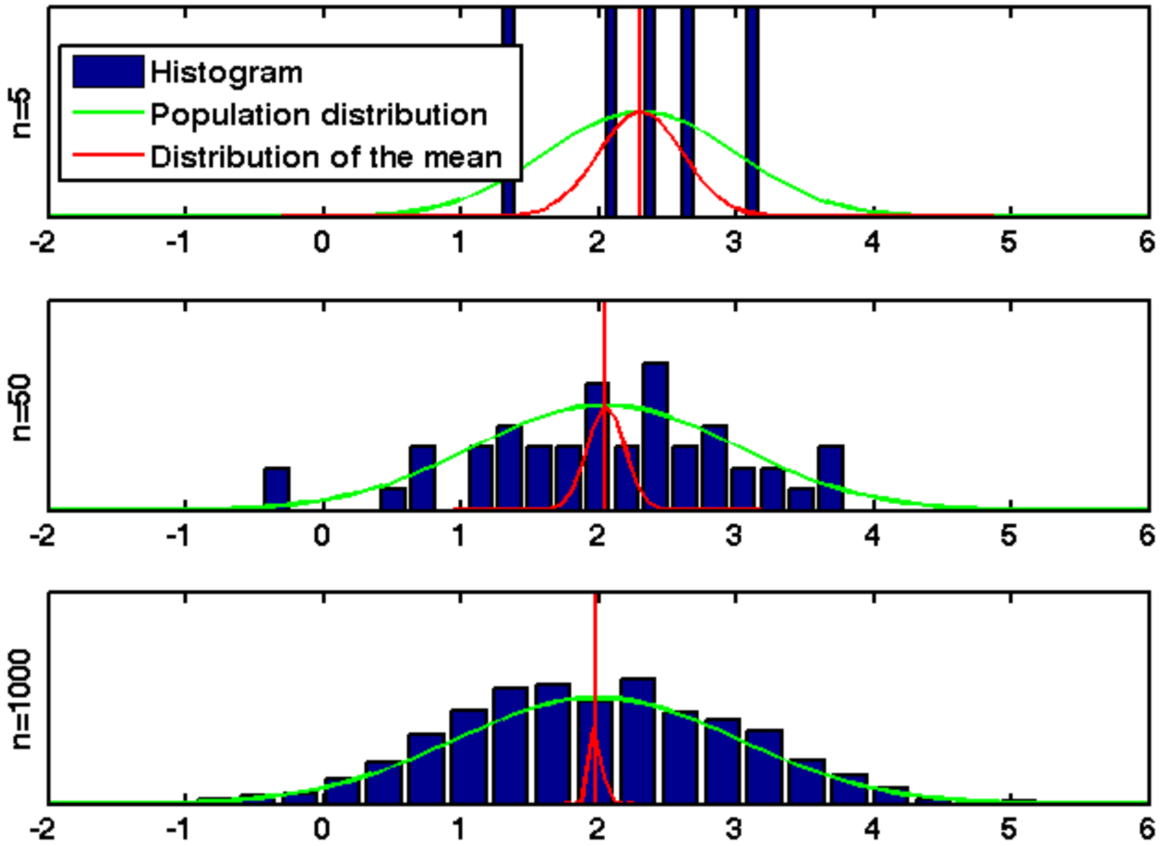
  - Assuming that the measured distribution is normal
    - mean – **estimates** how an average observer rates

Standard deviation and standard error

Mean

$$s = \sqrt{\frac{1}{N} \sum_{i=1..N} (x_i - m)^2}$$

Sample size



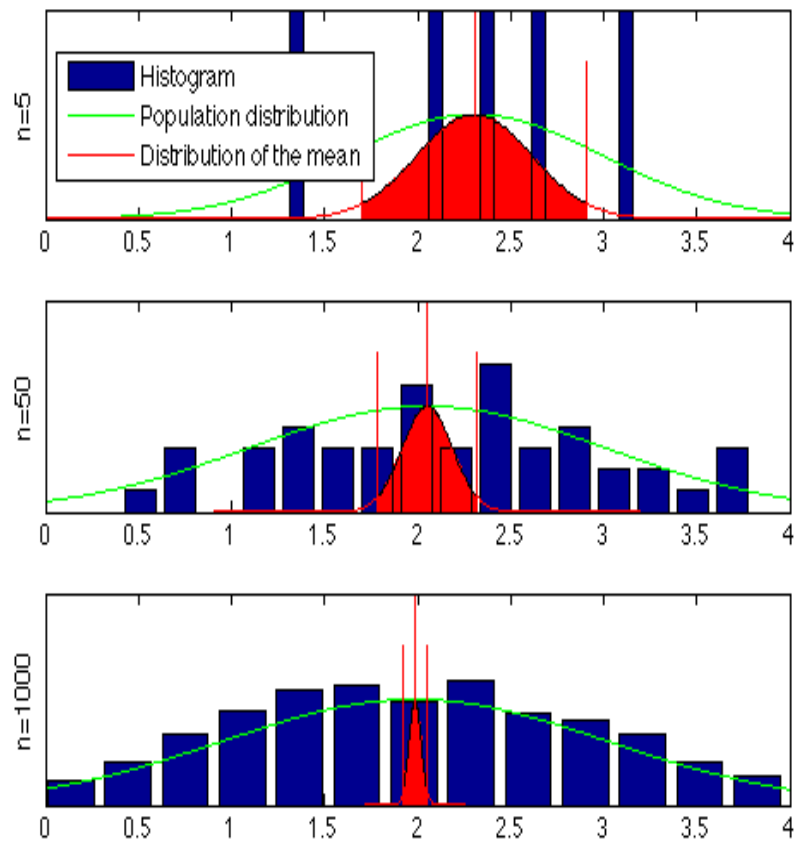
Standard error **of the mean**

$$SE = \frac{s}{\sqrt{N}}$$

Should the results in the paper include error bars for the standard error or for the standard deviation?

Confidence intervals

If the same experiment is repeated with the same number of observers, in 95% of the cases the average value is expected to be within the range of the confidence interval



$$ci = \left[-1.96\cdot SE;\quad 1.96\cdot SE\right]$$

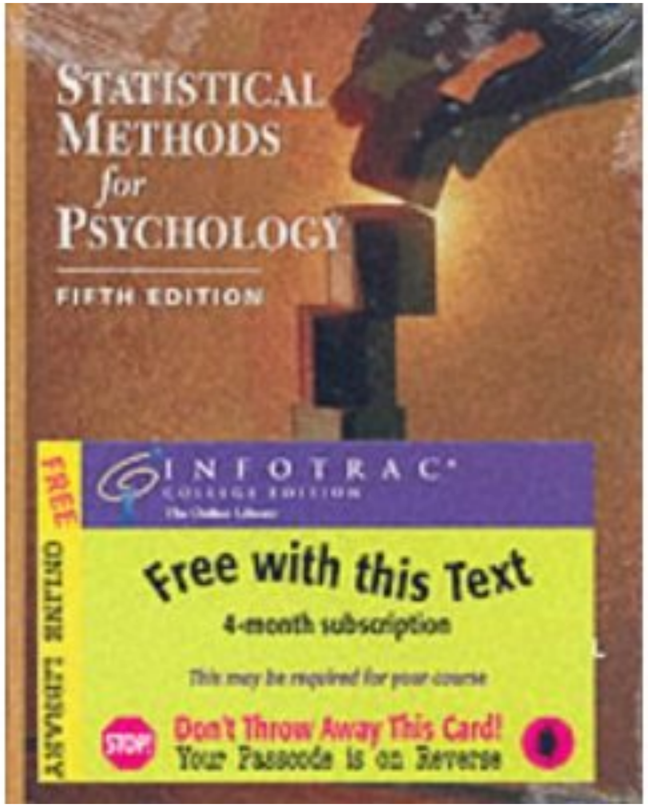
95% confidence interval is the most common choice for the error bars.

Psychometrical scales

- Nominal [red; green; blue]
  - Determination of equality
  - Task: Assign one of the labels to a stimulus
- Ordinal [1^{st}; 2^{nd}; 3^{rd}]
  - Determination of greater or less than
  - Task: Order stimuli according to *ness
- Interval [1; 2.5; 3.2] x better than the reference
  - Determination of differences (distances)
  - Task: Assign score 0-100 to a stimulus
- Ratio [20; 30; 80] points in an absolute scale
  - Determination of equality of ratios (reference “0” is known)

Good reference

- *Statistical Methods for Psychology*
  David. C. Howell



Outline

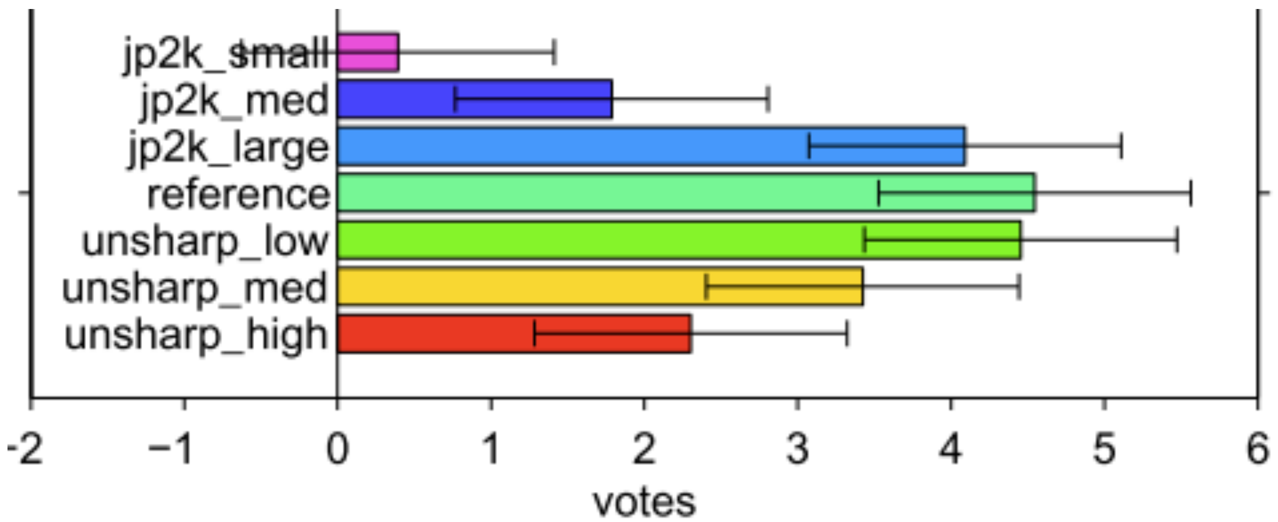
- Why do we need quality assessment?
- Quality assessment – overview
- The method of pair-wise comparisons
- Basic statistics – review
- Pair-wise comparison – data analysis
  - Statistical significance
  - Practical significance

Data analysis

- Statistical significance
  - Whether there is enough **evidence** in the data to say that condition A is better than condition B
  - Involves statistical testing
    - We want to reject H_0 at 0.05 significance level
  - The more samples with have, the more likely we will reject H_0
- Practical significance
  - What percentage of the population will notice that A is different than B

Statistical significance

- Condition *jp2k_large* on average collected less votes than condition *reference* in this experiment
- But would it collect less votes if we run the experiment again with different observers?
- Statistical testing is meant to provide an *evidence* that the difference in votes will be observed in at least 95% of repetitions of the experiment (under certain assumptions)



Statistical significance (with matlab)

- Step 1: Create **per-observer** comparison matrices

$$C = \begin{bmatrix} 0 & 1 & 1 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$$

- Step 2: Sum up columns to computer per observer votes

$$V = \begin{bmatrix} 0 & 2 & 1 \end{bmatrix}$$

Statistical significance

- Step 3: Create data set with the number of votes
  - per scene, observer, condition

observer	scene	condition	votes
AdnanNez	AbruptMotion	Photomatix	3
AdnanNez	AbruptMotion	Photoshop	1
AdnanNez	AbruptMotion	Sen2013	2
AdnanNez	AbruptMotion	Zimmer2011	0
AlmaS	AbruptMotion	Photomatix	2
AlmaS	AbruptMotion	Photoshop	1
AlmaS	AbruptMotion	Sen2013	3
AlmaS	AbruptMotion	Zimmer2011	0
AmarB.	AbruptMotion	Photomatix	3

- Step 4: For each scene, run Kruskall-Wallis test

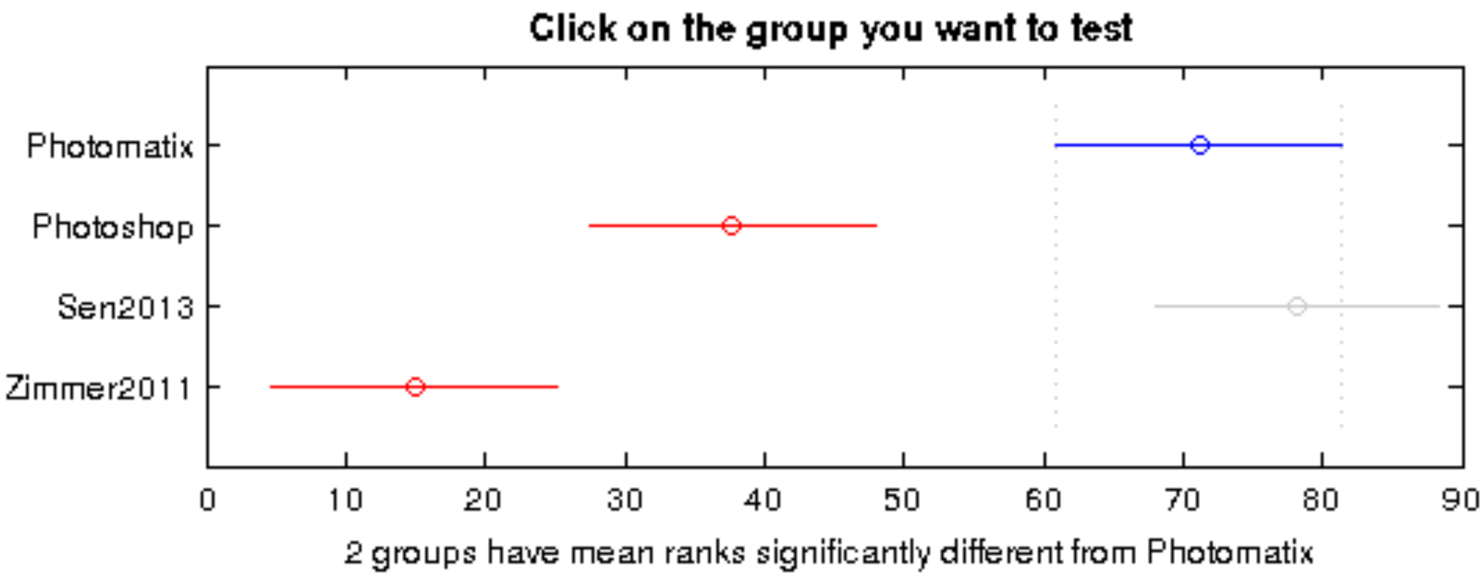
in Matlab

  - D = dataset(‘File’, ‘results.csv’, ‘Delimiter’, ‘,’);
  - Dss = D(strcmp(D.scene, ‘AbruptMotion’), :);
  - [p, t, stats] = kruskalwallis(Dss.votes’, Dss.condition’);

Statistical significance

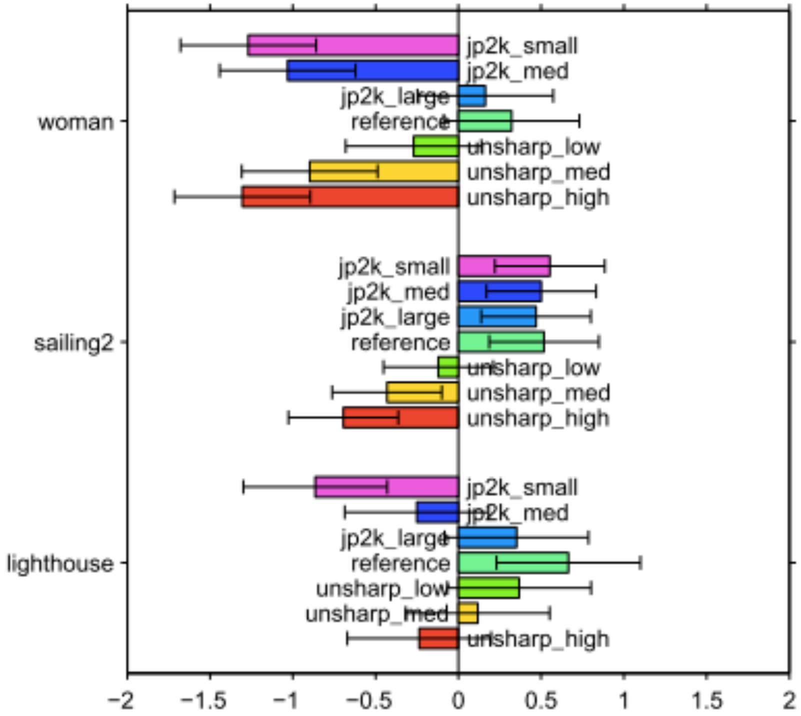
- Step 5: Run multiple-comparison test:
  - multcompare(stats)

in Matlab

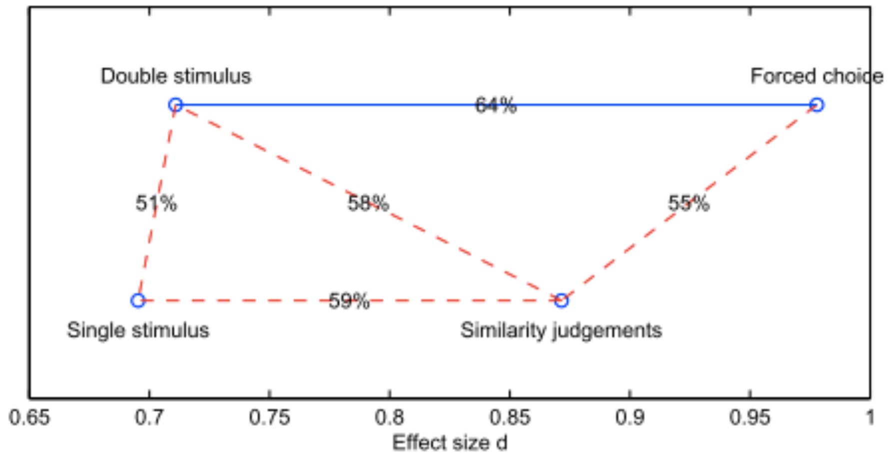


Statistical significance

- Step 6: Report which conditions are statistically significantly different



I	P	A	H	L	B
206	154	142	120	78	20

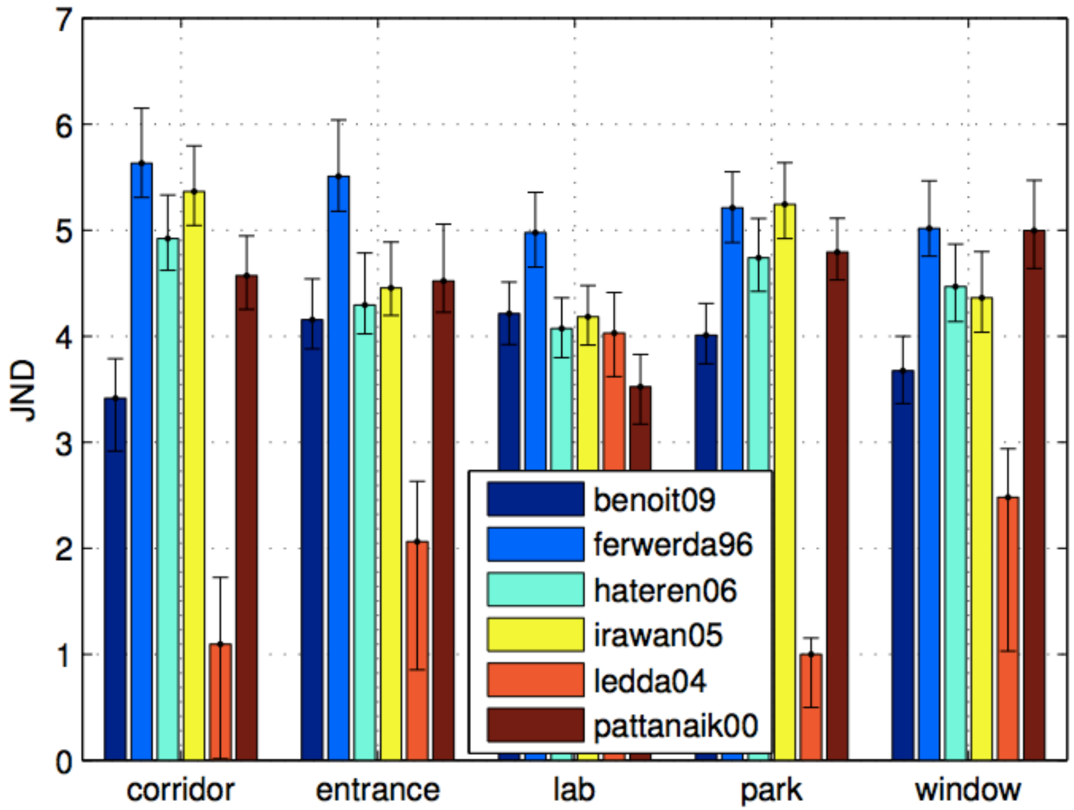


Statistical testing – common misconceptions

- No statistical significance does not mean that the two conditions are the same
  - Statistical test is likely to fail (H_0 cannot be rejected) if there are not enough observers
  - It is a good idea to run retrospective power analysis
- The standard statistical testing does not generalize the results to the entire population of images
  - It only ensures that the results are likely to be the same for different group of observers, but the same images
  - It is very hard to prove that the quality difference generalizes to the entire population of images

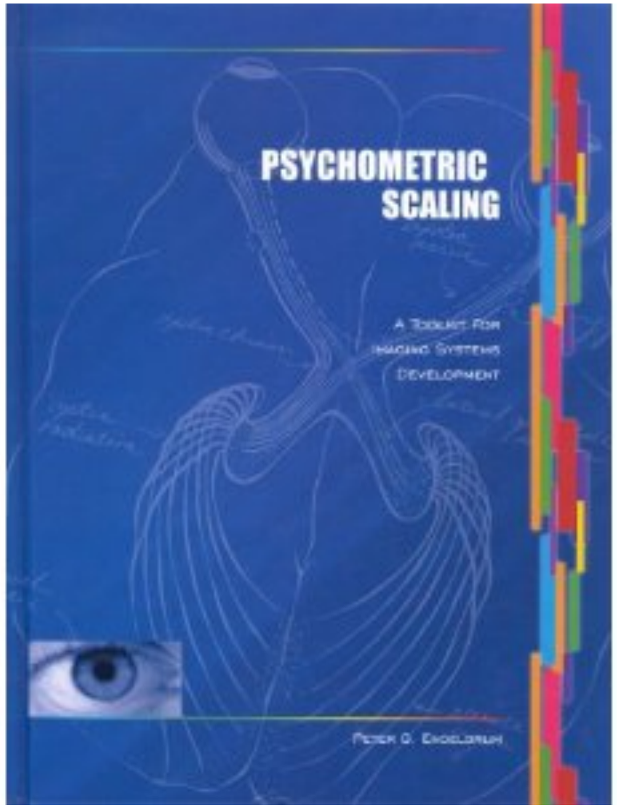
Practical significance - scaling

- Scaling: to map user judgments into meaningful interval scale
- Typically that scale is in just-noticeable-difference units
  - The difference of 1 JND means that 75% of observers would choose one condition over another
  - Useful to show "practical" significance



Data analysis - scaling

- Good reference:
  - Psychometric Scaling: A Toolkit for Imaging Systems Development
  - Peter G. Engeldrum
  - 2000



Data analysis - scaling

- Step 1: Create **per-scene** comparison matrix

$$C = \begin{bmatrix} 0 & 3 & 1 \\ 3 & 0 & 2 \\ 5 & 4 & 0 \end{bmatrix}$$

- Step 2: Change the votes into probabilities

$$P = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & \frac{1}{6} \\ \frac{1}{2} & \frac{1}{2} & \frac{1}{3} \\ \frac{5}{6} & \frac{2}{3} & \frac{1}{2} \end{bmatrix}$$

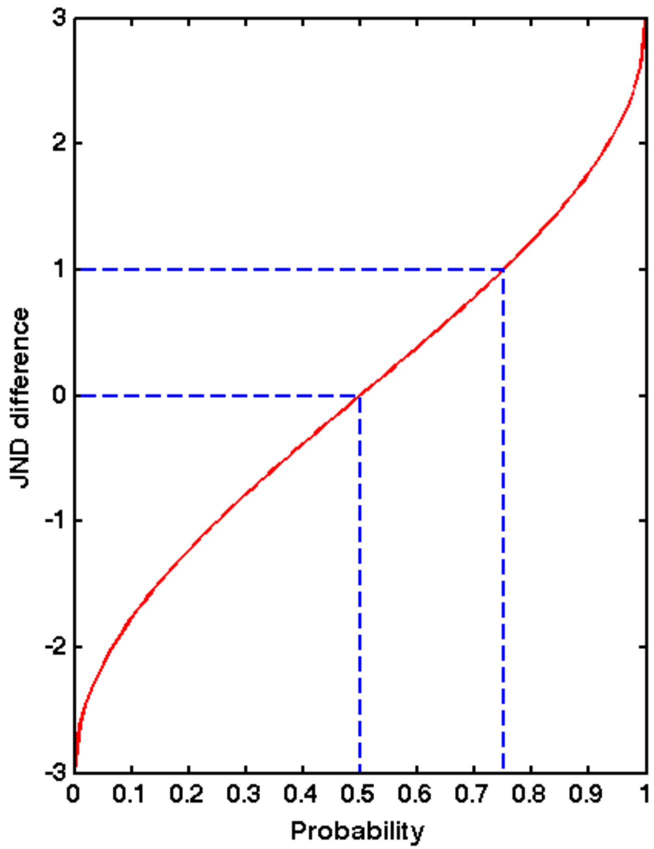
Put 0.5 on the diagonal

Data analysis - scaling

- Step 3: Transform probabilities into JND difference values

$$S_d(P) = \frac{12}{\pi} a\sin\left(\sqrt{P}\right) - 3$$

  - Used instead of the inverse cumulative normal distrib.



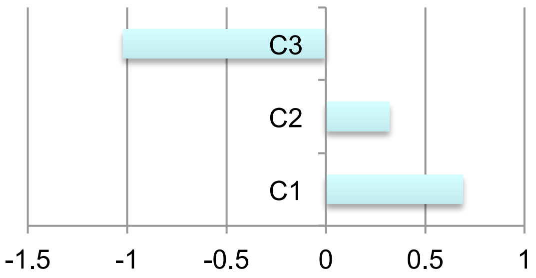
Data analysis - scaling

$$C = \begin{bmatrix} 0 & 3 & 1 \\ 3 & 0 & 2 \\ 5 & 4 & 0 \end{bmatrix}$$

- Step 4: Solve for $S_{1..3}$

$$S_d = \begin{bmatrix} S_1 - S_1 & S_2 - S_1 & S_3 - S_1 \\ S_1 - S_2 & S_2 - S_2 & S_3 - S_2 \\ S_1 - S_3 & S_2 - S_3 & S_3 - S_3 \end{bmatrix}$$

- The least square solution (up to an arbitrary offset) can be found by summing up the 0.5 of the columns



$$S_d = \begin{bmatrix} 0 & 0 & -1.39 \\ 0 & 0 & -0.65 \\ 1.39 & 0.65 & 0 \end{bmatrix}$$

$$S = \begin{bmatrix} 0.69 & 0.32 & -1.02 \end{bmatrix}$$

Problem with scaling

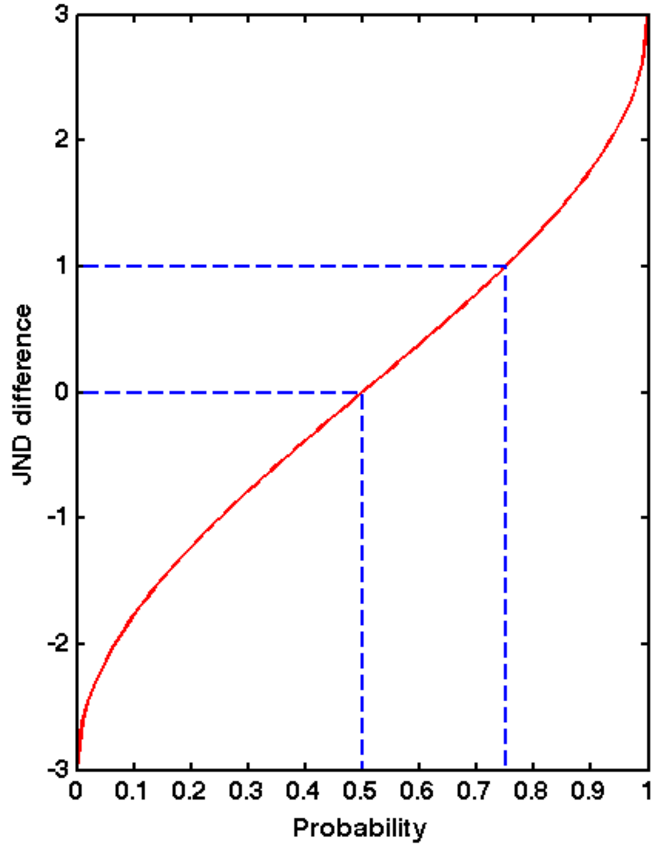
- If the observers are unanimous for any pair, the JND difference is undefined
- The function

$$S_d(P) = \frac{12}{\pi} a\sin\left(\sqrt{P}\right) - 3$$

is a quick fix that limits JND different to <-3;3> range (unlike normal distribution)

- Better solution:

  Silverstein, D., & Farrell, J. (2001). Efficient method for paired comparison. Journal of Electronic Imaging, 10, 394.



Confidence interval for JND scaling

- Can be found by **bootstrapping**
- From the original sample generate 500 (or more) random samples with repetitions
  - Original sample: A, B, C, D, E (letters are any numbers)
  - Random sample 1: A, A, C, D, E
  - Random sample 2: B, C, C, D, D
  - …
- Compute statistics or perform JND scaling on each random sample
- Compute 5th and 95th percentile of the resulting distribution

Summary

- Quality assessment – overview
- Pair-wise comparisons
- Basic statistics – review
- Pair-wise comparison – data analysis
  - Statistical significance
  - Practical significance