As time goes by...

On the performance of significance tests in reaction time experiments

Wolfgang Wiedermann & Bartosz Gula

wolfgang.wiedermann@uni-klu.ac.at
bartosz.gula@uni-klu.ac.at

Department of Psychology
University of Klagenfurt, Austria

8th Alps-Adria Psychology Conference
Ljubljana, October 2 – 4, 2008

Distribution of reaction time (RT) scores

- empirical distributions of reaction times (RTs) are typically *unimodal* and *positively skewed* (e.g. van Zandt, 2002), resembling rather ex-Gaussian, gamma, Wald, or Weibull distributions
- thus, RTs will often violate the assumption of normality made in several parametric tests (e.g. Students't t-test, ANOVAs F-Test)
- What happens, if a parametric test is applied to highly non-normal data?
  1. Previous studies of e.g. Boneau (1960) and Posten (1978) showed that the t-test is quite robust against distributional violations, if sample size is moderately large (e.g. $n \geq 20$).
  2. However, the t-test shows a power disadvantage, compared to nonparametric tests (Zimmerman & Zumbo, 1993).

Distribution of reaction time (RT) scores

- empirical distributions of reaction times (RTs) are typically *unimodal* and *positively skewed* (e.g. van Zandt, 2002), resembling rather ex-Gaussian, gamma, Wald, or Weibull distributions
- thus, RTs will often violate the assumption of normality made in several parametric tests (e.g. Students't t-test, ANOVAs F-Test)
- What happens, if a parametric test is applied to highly non-normal data?
    1. Previous studies of e.g. Boneau (1960) and Posten (1978) showed that the t-test is quite robust against distributional violations, if sample size is moderately large (e.g. $n \geq 20$).
    2. However, the t-test shows a power disadvantage, compared to nonparametric tests (Zimmerman & Zumbo, 1993).

Review of research articles (1)

- How is typically dealt with non-normality of RTs in research practice? Are distributional considerations mentioned?
- Journal of Experimental Psychology: Human Perception and Performance (JEP:HPP): Review of 2000 and 2007 Volumes
- Main coding categories:
  1. Are RTs analyzed?
  2. What kind of distributional considerations/procedures are mentioned, if any?

Review of research articles (2)

Volume	2000	2007	Overall
no. of empirical articles	104	102	206
no. of experiments	385	368	753
no. of experiments analyzing RTs	229 (60%)	230 (63%)	**459 (61%)**
trimming/outlier removal	107 (44%)	114 (47%)	**221 (46%)**
log transformation	5	1	6 (1%)
fitting of RT models	13 (5%)	3 (1%)	16 (3%)
other/special *	23 (10%)	19 (8%)	42 (9%)
not mentioned but parametric tests (t- or F-test) used	93 (39%)	101 (42%)	**194 (41%)**

* winsorized mean (Tukey, 1962); biweight estimates of means & interquartile stretch criterion (Hoaglin, Mosteller, & Tukey, 1983); recursive trimming & non-recursive shifting z-scores (van Selst & Jolicoeur, 1994); vincentized distributional analysis (Ratcliff, 1979)

III. Alternatives

Data Transformation (1)

- transforming raw scores increases the power to detect differences (Doksum & Wong, 1983; Rasmussen & Dunlap, 1991)
- nonlinear transformations can achieve normality by altering the distance between data points
- For RT-measures the log-transformation is considered as an adequate tool to overcome non-normality (Kirk, 1983)

$$x' = \log(x) \qquad \text{or} \qquad x' = \log(x + c)$$

if some sample values are zero.

Data Transformation (2)

Adaptive Transformation (Kirk, 1983)

- This procedure combines the reciprocal, the log, and the square root transformation.
- Decision Rule:
  1. Each transformation is applied on the smallest and largest score within each experimental condition.
  2. Determine the range within each treatment level and compute the ratio of smallest to the largest range.
  3. The transformation generating the smallest ratio is selected.

Data Trimming (1)

- The sample trimmed mean reduces relatively large standard errors and thus represents a more robust measure of location, if samples are heavily skewed.
- Application:
  1. Reorder the sample ascendingly.
  2. Determine the trimming criterion g.
  3. Remove the g-largest and g-smallest values and use the remaining observations for the further analysis (Wilcox, 2005).

Data Trimming (2)

Adaptive Trimming (Leger & Romano, 1990)

- Besides the usage of constant trimming (in terms of percentage of removed observations, in terms of SDs, or in the case of RT measures using fixed time values) it is also possible to determine the trimming proportion empirically.
- To this end, the standard error of the trimmed mean is computed for values like 0, 10%, and 20% and the value producing the smallest standard error is used for trimming.

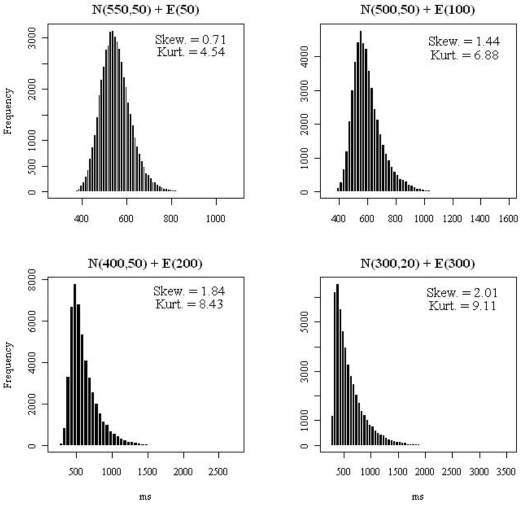
Non-parametric tests

Wilcoxon–Mann–Whitney *U*–test/ Kruskal–Wallis test

- The U – test (as well as the Kruskal Wallis test) gains a power advantage over parametric procedures if the normality assumption is not fulfilled (cf. Zimmerman, 1994; Zimmerman & Zumbo, 1993).
- This can be explained through the conversion of initial scores to ranks, which reduces the distortive influence of extremely deviant scores.

IV. Monte Carlo Simulation

Shape of four selected distributions



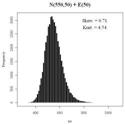
Simulation Procedure

- Samples were evaluated using the two-sample t-test on
  1. raw RT scores
  2. log-transformed RT scores
  3. adaptively transformed RT scores
  4. trimmed RT scores $(2\sigma, 2.5\sigma, 3\sigma)$
  5. adaptively trimmed scores
  6. nonparametric U-Test.
- To evaluate the power of the tests, differences in location were induced by adding constants $(\delta = 0, 1, 2, 3)$ to the *raw values* in one sample.
- The sampling procedure was replicated 50,000 times, all tests were non-directional using $\alpha = 5\%$.

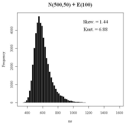
Results:

		RAW	**TRANSFORM**		**TRIMMING**				
δ	**Skewness** (κ)	***t***	***LOG***	***ADAPT***	***2SD***	***2.5SD***	***3SD***	***ADAPT***	***W***
0	**0.71**								
	1.44								
	1.84								
	2.01								
1	**0.71**								
	1.44								
	1.84								
	2.01								
2	**0.71**								
	1.44								
	1.84								
	2.01								
3	**0.71**								
	1.44								
	1.84								
	2.01								

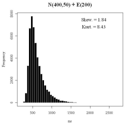
$\kappa_1 = 0.71$



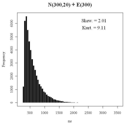
$\kappa_1 = 1.44$



$\kappa_1 = 1.84$



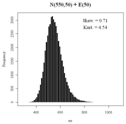
$\kappa_1 = 2.01$



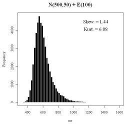
Results: Overall

		RAW	TRANSFORM		TRIMMING				
δ	Skewness (κ)	*t*	*LOG*	*ADAPT*	*2SD*	*2.5SD*	*3SD*	*ADAPT*	*W*
0	**0.71**	4.92	4.88	5.09	4.88	4.80	4.89	4.72	4.82
	1.44	4.90	5.02	5.12	4.96	5.03	5.15	4.61	5.01
	1.84	4.88	5.03	5.02	5.16	5.07	4.99	4.49	5.10
	2.01	4.80	4.95	4.99	5.11	4.93	4.97	4.35	5.02
1	**0.71**	16.44	17.18	**17.78**	15.61	16.64	16.84	16.18	17.13
	1.44	16.53	18.88	19.88	19.07	18.69	18.05	17.89	**20.92**
	1.84	16.57	22.38	23.70	21.13	19.89	18.95	18.73	**25.98**
	2.01	16.16	26.24	26.86	21.88	20.10	18.91	18.39	**31.35**
2	**0.71**	49.94	52.19	**53.16**	47.20	50.62	50.95	49.97	52.64
	1.44	49.82	56.98	58.27	56.52	55.65	53.90	54.73	**62.01**
	1.84	50.27	65.75	64.34	61.96	58.84	56.45	57.51	**71.11**
	2.01	49.91	73.22	67.10	63.72	59.85	57.05	57.35	**76.11**
3	**0.71**	83.80	85.64	**86.23**	80.65	84.05	84.48	83.76	85.68
	1.44	83.52	88.88	89.12	87.41	87.34	86.30	86.81	**90.88**
	1.84	83.29	93.06	91.23	90.77	89.04	87.48	88.03	**94.09**
	2.01	83.86	**95.98**	92.84	91.97	90.03	88.27	88.07	**95.32**

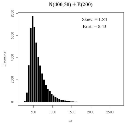
$\kappa_1 = 0.71$



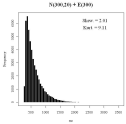
$\kappa_1 = 1.44$



$\kappa_1 = 1.84$



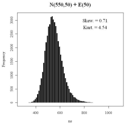
$\kappa_1 = 2.01$



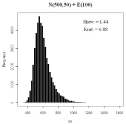
Results: Transformations

		RAW	TRANSFORM		TRIMMING				
δ	**Skewness** (κ)	*t*	*LOG*	*ADAPT*	*2SD*	*2.5SD*	*3SD*	*ADAPT*	*W*
0	**0.71**	4.92	4.88	5.09	4.88	4.80	4.89	4.72	4.82
	1.44	4.90	5.02	5.12	4.96	5.03	5.15	4.61	5.01
	1.84	4.88	5.03	5.02	5.16	5.07	4.99	4.49	5.10
	2.01	4.80	4.95	4.99	5.11	4.93	4.97	4.35	5.02
1	**0.71**	16.44	17.18	**17.78**	15.61	16.64	16.84	16.18	17.13
	1.44	16.53	18.88	**19.88**	19.07	18.69	18.05	17.89	**20.92**
	1.84	16.57	22.38	**23.70**	21.13	19.89	18.95	18.73	**25.98**
	2.01	16.16	26.24	**26.86**	21.88	20.10	18.91	18.39	**31.35**
2	**0.71**	49.94	52.19	**53.16**	47.20	50.62	50.95	49.97	52.64
	1.44	49.82	56.98	**58.27**	56.52	55.65	53.90	54.73	**62.01**
	1.84	50.27	**65.75**	64.34	61.96	58.84	56.45	57.51	**71.11**
	2.01	49.91	**73.22**	67.10	63.72	59.85	57.05	57.35	**76.11**
3	**0.71**	83.80	85.64	**86.23**	80.65	84.05	84.48	83.76	85.68
	1.44	83.52	88.88	**89.12**	87.41	87.34	86.30	86.81	**90.88**
	1.84	83.29	**93.06**	91.23	90.77	89.04	87.48	88.03	**94.09**
	2.01	83.86	**95.98**	92.84	91.97	90.03	88.27	88.07	**95.32**

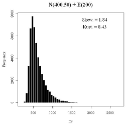
$\kappa_1 = 0.71$



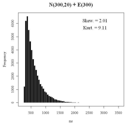
$\kappa_1 = 1.44$



$\kappa_1 = 1.84$



$\kappa_1 = 2.01$



As time goes by…

Wiedermann & Gula

I. Introduction

Reaction time (RT) scores

II. Review

JEP:HPP

III. Alternatives

Transformation

Trimming

Non-parametric tests

IV. Monte Carlo Simulation

Procedure

Results

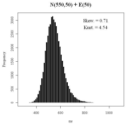
Discussion

V. Conclusions

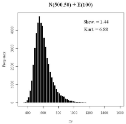
Results: Trimming

		RAW	TRANSFORM		TRIMMING				
δ	**Skewness (κ)**	*t*	*LOG*	*ADAPT*	***2SD***	***2.5SD***	***3SD***	***ADAPT***	*W*
0	**0.71**	4.92	4.88	5.09	4.88	4.80	4.89	4.72	4.82
	1.44	4.90	5.02	5.12	4.96	5.03	5.15	4.61	5.01
	1.84	4.88	5.03	5.02	5.16	5.07	4.99	4.49	5.10
	2.01	4.80	4.95	4.99	5.11	4.93	4.97	4.35	5.02
1	**0.71**	16.44	17.18	**17.78**	15.61	16.64	**16.84**	16.18	17.13
	1.44	16.53	18.88	19.88	**19.07**	18.69	18.05	17.89	**20.92**
	1.84	16.57	22.38	23.70	**21.13**	19.89	18.95	18.73	**25.98**
	2.01	16.16	26.24	26.86	**21.88**	20.10	18.91	18.39	**31.35**
2	**0.71**	49.94	52.19	**53.16**	47.20	50.62	**50.95**	49.97	52.64
	1.44	49.82	56.98	58.27	**56.52**	55.65	53.90	54.73	**62.01**
	1.84	50.27	65.75	64.34	**61.96**	58.84	56.45	57.51	**71.11**
	2.01	49.91	73.22	67.10	**63.72**	59.85	57.05	57.35	**76.11**
3	**0.71**	83.80	85.64	**86.23**	80.65	84.05	**84.48**	83.76	85.68
	1.44	83.52	88.88	89.12	**87.41**	87.34	86.30	86.81	**90.88**
	1.84	83.29	93.06	91.23	**90.77**	89.04	87.48	88.03	**94.09**
	2.01	83.86	**95.98**	92.84	**91.97**	90.03	88.27	88.07	**95.32**

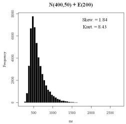
$\kappa_1 = 0.71$



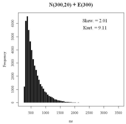
$\kappa_1 = 1.44$



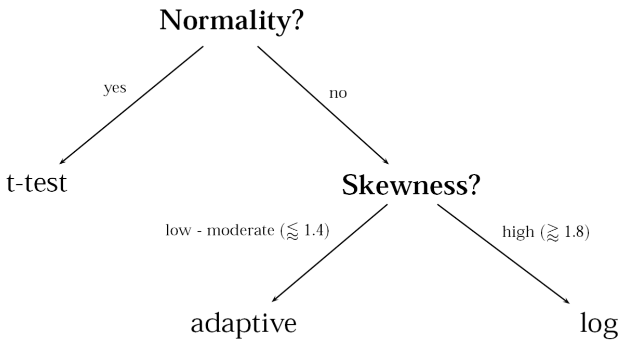
$\kappa_1 = 1.84$



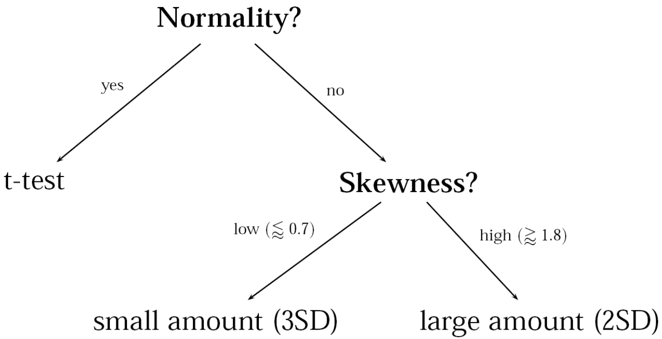
$\kappa_1 = 2.01$



Discussion: Transformation



Discussion: Trimming



Take Home Message (1)

From a *statistical* point of view:

- The Wilcoxon test is robust independent of the degree of skewness and most powerful (also true for nonparametric tests in general?)
- Adaptive trimming is less powerful than constant trimming.
- For highly skewed distributions trimming a large amount is more powerful than trimming a small amount (small n vs. non-normality dilemma)
- Adaptive transformation (Kirk, 1983) outperforms log transformation in case of low-to-moderate skewness.
- In general transformation is slightly more powerful than trimming.
- Among all procedures the t-test on raw scores is least powerful.

Take Home Message (2)

From a *theoretical* point of view:

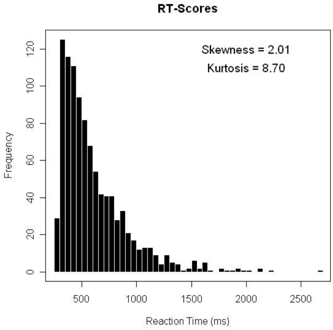
Trimming and transformation cause problems of interpretability (e.g. effect sizes) and methods of RT modeling may be more suitable!

Thank you for your attention.

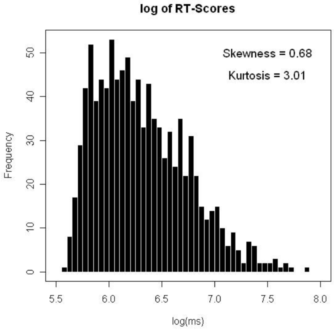
ALPEN-ADRIA
UNIVERSITÄT
KLAGENFURT

Data Transformation (2)

Log – Transformation

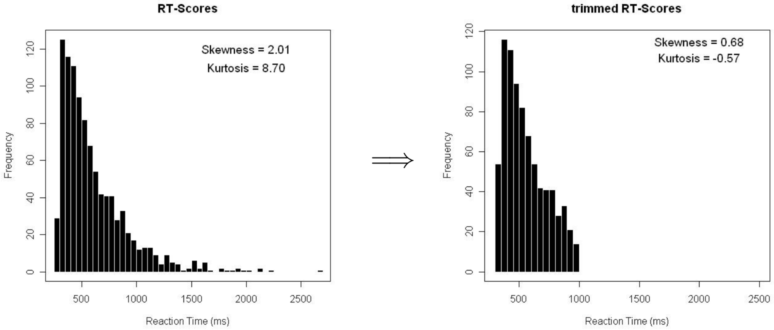


$\Longrightarrow$

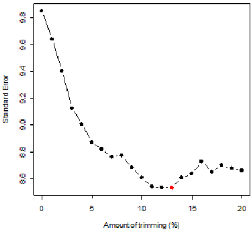


Data Trimming (2)

Amount of trimming: 10%

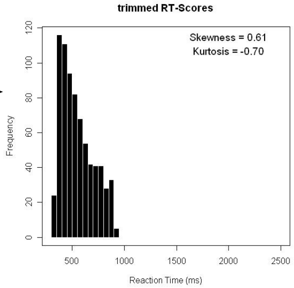


Data Trimming (4)



$\Longrightarrow$

The smallest standard error was found using 13% trimming.



Simulation Procedure

- Miller (1988) defined twelve ex-Gaussian distributions which reflect the shape and range of typically found empirical RTs
- RT scores were simulated using the ex-Gaussian distribution

  $$x = N(\mu, \sigma) + E(\lambda),$$

  where $N(\mu, \sigma)$ is a normal distribution with mean μ and variance σ, and $E(\lambda)$ is an exponential distribution with mean λ.
- Normal deviates were generated using the Ziggurat-method (Marsaglia & Tsang, 2000).
- $E = -\log(u) - 1$, where u denotes a random uniform variable with interval [0,1].