MEDICAL UNIVERSITY OF VIENNA

CeMSIIS

Test statistics for two-group comparisons of zero-inflated intensity values

Andreas Gleiss
Center for Medical Statistics, Informatics, and Intelligent Systems
Medical University of Vienna, Austria

Joint work with:
Georg Heinze (CeMSIIS)
Mohammed Dakna (Mosaiques Diagnostics)
Harald Mischak (Mosaiques Diagnostics)

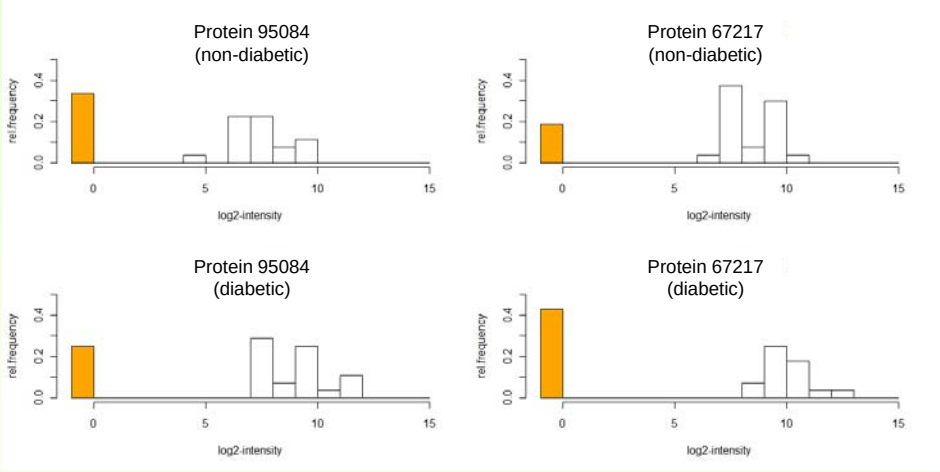
partly supported by EU grant HEALTH-F2-2009-241544

Motivating Example

- patients with chronic kidney disease (Jantos-Siwy et al, 2009)
- urine samples from 28 diabetic CKD
- urine samples from 27 non-diabetic CKD
- 4,290 proteins

- Goal:
  to find proteins differentially expressed in diabetic and non-diabetic patients
  - significance
  - magnitude of difference (shift in location)

Example Proteins



consonant **dissonant**

Example Proteins

- Percentage of **point-mass values (PMVs)**:

	Min.	Q1	Median	Q3	Max.
non-diab. (n=27)	0%	67%	85%	93%	100%
diabetic (n=28)	0%	71%	86%	96%	100%

- Within compounds: PMV proportions mostly similar between groups

- Consonant vs. dissonant:

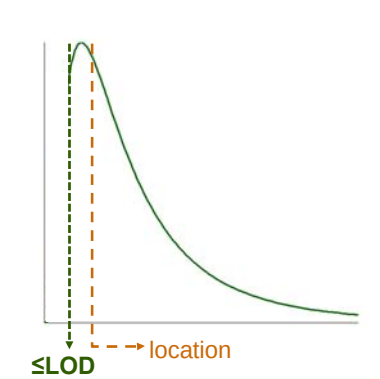
	consonant	equal prop.	dissonant
number of proteins	2634	17	1637

Concepts

- Two alternative models for data distribution:

i. left-censored continuous distribution:

- e.g. log-normal
- censored at limit of detection
- H_0: equality of location parameters
- dissonant group differences only by chance
- e.g., Zhang et al (2009)

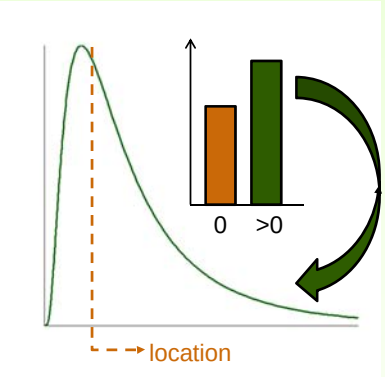


Concepts

- Two alternative models for data distribution:

ii. mixture distribution:

- binomial part = presence of signal
- continuous part = intensity if signal present (log-normal, gamma, …)
- H_0: compound hypothesis!
- consonant as well as dissonant group differences
- e.g., Taylor & Pollard (2009)



Concepts: Comments (1/3)

- Location parameters measure different things!
- Both concepts do not a priori distinguish between
  - „**technical PMVs**“: signal present but not detectable
    - high variance and low mean
    - incorrect detection or misinterpretaion of signal
  - „**biological PMVs**“: no signal present
    - biological phenomenon expressed through more than one compound → bimodal distribution

Concepts: Comments (2/3)

- Taylor & Pollard (2009):
  - „truncated values“ [technical]:
    - below limit of detection (LOD)
    - “lower bound on meaningful signal set by researcher“
  - „true zeros“ [biological]
- Hallstrom (2009):
  - „rancid part of the sample“ [biological]
- Senn et al (2012):
  - „the ghost of departed quantities“ [technical?]

Concepts: Comments (3/3)

- Hard (impossible?) to identify concept that corresponds best to real data
- Choice of concept crucial for results of simulation studies!
- Up to now: simulation studies only based on concept that fits the presented test statistic

Test Statistics: concept (i)

- One-part tests:
  - **Wilcoxon's rank-sum test**
    - Hodges-Lehmann estimator of location shift
  - **T-test** for unequal variances (+ imputation)
    - difference of means
- Truncated tests:
  - **Truncated Wilcoxon-test** (Hallstrom, 2009)
    - apply Wilxocon's rank-sum test to data after removing equal number of point-mass values
    - Hodges-Lehmann estimator after truncation
  - **Truncated T-test**
    - same idea, adapt degrees of freedom
    - difference of means after truncation

Test Statistics: concept (i)

- **Tobit model** (following Zhang et al, 2009)
  - parametric model based on truncated normal distribution
  - maximum likelihood estimation, LR p-value
  - direct estimate for difference of means

Test Statistics: concept (ii)

- Two-part tests
  (Taylor & Pollard, 2009, based on Lachenbruch, 1976):

  X^2 statistic for binomial part
  + (statistic for contin. part)^2
  $\approx$ X^2 statistic with 2 d.f.

  - **Two-part Wilcoxon's rank-sum test**
    - Hodges-Lehmann estimator for continuous parts
  - **Two-part T-test** for unequal variances
    - difference of means for continuous parts
  - **Empirical LR Test** (Taylor & Pollard, 2009)
    - LR = L(p, μ,…) / L(p_0,p_1, μ_0, μ_1,…)
    - empirical distributions of groups plugged into LRT
    - no direct effect estimate available!

Test Statistics: concept (i+ii)

- **Left-inflated mixture (LIM) model**
  - combine Tobit model with two-part approach!
  - Yang & Simpson (2010)

biolocial PMVs — technical PMVs

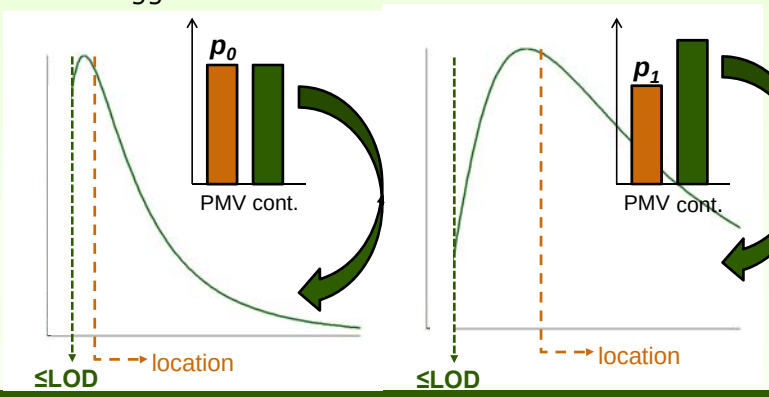
$$f_{LIM}(x_i \mid \mu_0, \sigma_0, p_0; lod) = \begin{cases} p_0 + (1-p_0)\Phi(lod \mid \mu_0, \sigma_0) & \text{for } x_i \text{ a PMV} \\ (1-p_0)\phi(x_i \mid \mu_0, \sigma_0) & \text{otherwise} \end{cases}$$

  - X_i for $i = 1,\ldots, n_0$: log-transformed values of first group
  - p_0 = proportion of biological PMVs
  - μ_0 and σ_0: mean and standard deviation of (underlying but left-censored) normal distribution
  - $\Phi()$ and φ (): normal distribution
  - analogously for second group with same $lod = \log_2(LOD)$

→ likelihood ratio test (under H_0: $\mu_0 = \mu_1$, $p_0 = p_1$, $\sigma_0 \neq \sigma_1$)

Simulations

- **General model**
  - ***p*₀** and ***p*₁** control "biological PMVs"
  - difference in **location** parameters (log-normal)
  - **LOD** triggers "technical PMVs"!



Simulations

- Scenarios:
  - n_0, n_1 = 10, 25
  - shift δ = 0, 1, 2 (= $\mu_1 - \mu_0$ on $\log_2$-scale)
  - p_0, p_1 = 0.0, 0.2, 0.5
  - *lod* = -∞, -0.5, 0.0 on $\log_2$-scale
    corresp. to 0%, 31%, 50% „technical PMVs" with μ = 0
    and 0%, 7%, 16% „technical PMVs" with μ = 1

Simulations

- Scenarios:
  - n_0, n_1 = 10, 25
  - shift δ = 0, 1, 2 (= $\mu_1 - \mu_0$ on $\log_2$-scale)
  - p_0, p_1 = 0.0, 0.2, 0.5
  - *lod* = -∞, **-0.5**, 0.0 on $\log_2$-scale
    corresp. to 0%, **31%**, 50% „technical PMVs" with $\boldsymbol{\mu = 0}$
    and 0%, **7%**, 16% „technical PMVs" with $\boldsymbol{\mu = 1}$
  - simulate 500 independent proteins,
    of which 250 are differentially expressed
  - 500 simulations for each scenario
  - Note: all variances = 1

Simulations (n_0, n_1 = 25)

	δ	p_0	p_1	*lod*	% cons.	% equal p	% diss.
P	**0**	0.2	0.5	-∞	**49.7**	0.9	**49.4**
B1							
B2							
B3							
T							
M1							
M2							
M3							

Simulations (n_0, n_1 = 25)

	δ	p_0	p_1	*lod*	% cons.	% equal p	% diss.
P	**0**	0.2	0.5	-∞	**49.7**	0.9	**49.4**
B1	1	0.5	0.2	-∞	**98.0**	0.9	1.1
B2	1	**0.2**	**0.2**	-∞	**43.1**	14.1	**42.8**
B3	1	0.2	0.5	-∞	1.1	0.9	**98.0**
T							
M1							
M2							
M3							

Simulations (n_0, n_1 = 25)

	δ	p_0	p_1	*lod*	% cons.	% equal p	% diss.
P	**0**	0.2	0.5	-∞	**49.7**	0.9	**49.4**
B1	1	0.5	0.2	-∞	**98.0**	0.9	1.1
B2	1	**0.2**	**0.2**	-∞	**43.1**	14.1	**42.8**
B3	1	0.2	0.5	-∞	1.1	0.9	**98.0**
T	1	**0.0**	**0.0**	-0.5	**97.8**	1.0	1.1
M1							
M2							
M3							

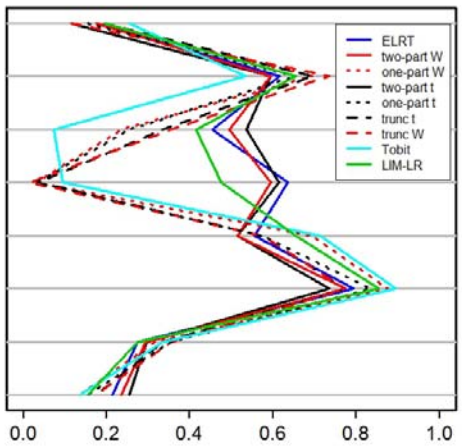
Simulations (n_0, n_1 = 25)

	δ	p_0	p_1	*lod*	% cons.	% equal p	% diss.
P	**0**	0.2	0.5	-∞	**49.7**	0.9	**49.4**
B1	1	0.5	0.2	-∞	**98.0**	0.9	1.1
B2	1	**0.2**	**0.2**	-∞	**43.1**	14.1	**42.8**
B3	1	0.2	0.5	-∞	1.1	0.9	**98.0**
T	1	**0.0**	**0.0**	-0.5	**97.8**	1.0	1.1
M1	1	0.2	0.0	-0.5	**99.0**	0.1	0.9
M2	1	**0.2**	**0.2**	-0.5	**89.4**	4.2	6.4
M3	1	0.0	0.2	-0.5	**60.6**	11.4	**28.1**

Simulations: True Positive Rates

	δ	p_0	p_1	*lod*
P	**0**	0.2	0.5	-∞
B1	1	0.5	0.2	-∞
B2	1	**0.2**	**0.2**	-∞
B3	1	0.2	0.5	-∞
T	1	**0.0**	**0.0**	-0.5
M1	1	0.2	0.0	-0.5
M2	1	**0.2**	**0.2**	-0.5
M3	1	0.0	0.2	-0.5

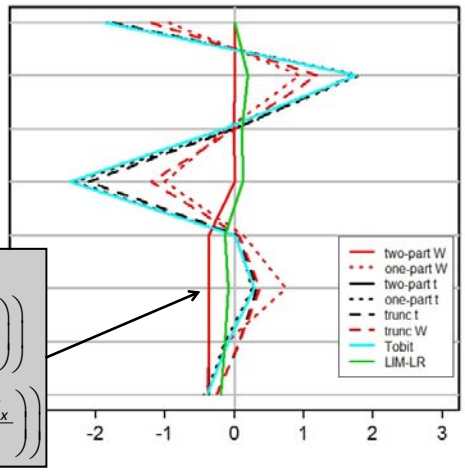
TPR at FPR = 0.005



Simulations: Effect Estimates

	δ	p_0	p_1	*lod*
P	**0**	0.2	0.5	-∞
B1	1	0.5	0.2	-∞
B2	1	**0.2**	**0.2**	-∞
B3	1	0.2	0.5	-∞
T	1	**0.0**	**0.0**	-0.5

Median Bias (DE)



Mean bias for two-part t-Test:

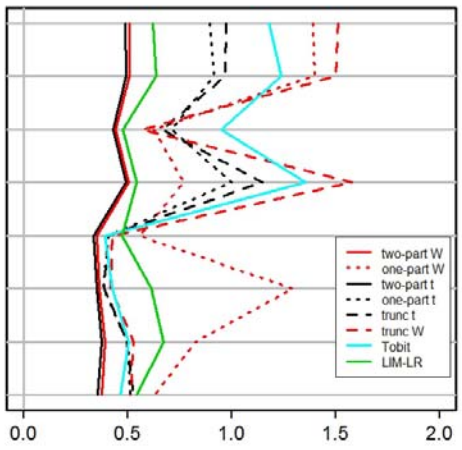
$$\varphi\left(\frac{lod-\mu_y}{\sigma}\right)\Big/\left(1-\Phi\left(\frac{lod-\mu_y}{\sigma}\right)\right)$$

$$-\varphi\left(\frac{lod-\mu_x}{\sigma}\right)\Big/\left(1-\Phi\left(\frac{lod-\mu_x}{\sigma}\right)\right)$$

Simulations: Effect Estimates

	δ	p_0	p_1	*lod*
P	**0**	0.2	0.5	-∞
B1	1	0.5	0.2	-∞
B2	1	**0.2**	**0.2**	-∞
B3	1	0.2	0.5	-∞
T	1	**0.0**	**0.0**	-0.5
M1	1	0.2	0.0	-0.5
M2	1	**0.2**	**0.2**	-0.5
M3	1	0.0	0.2	-0.5

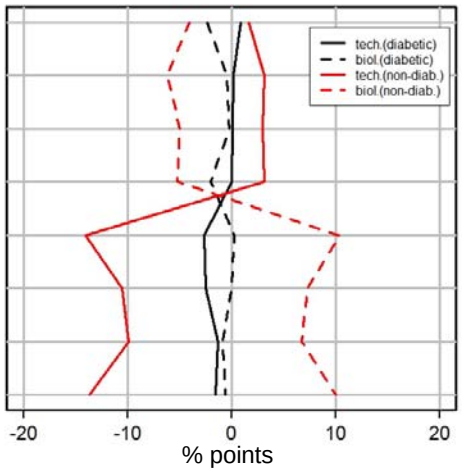
Interquartile Range (DE)



Simulations: PMV prop. for LIM-LR

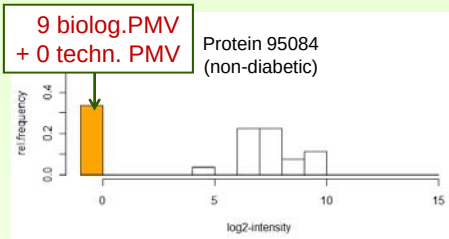
	δ	p_0	p_1	*lod*
P	**0**	0.2	0.5	-∞
B1	1	0.5	0.2	-∞
B2	1	**0.2**	**0.2**	-∞
B3	1	0.2	0.5	-∞
T	1	**0.0**	**0.0**	-0.5
M1	1	0.2	0.0	-0.5
M2	1	**0.2**	**0.2**	-0.5
M3	1	0.0	0.2	-0.5

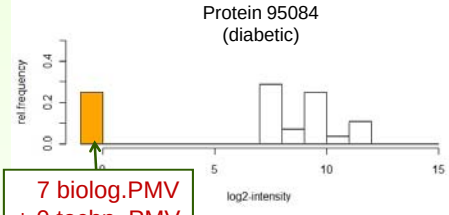
Median Bias (DE)



Application: Example Proteins

9 biolog.PMV
+ 0 techn. PMV



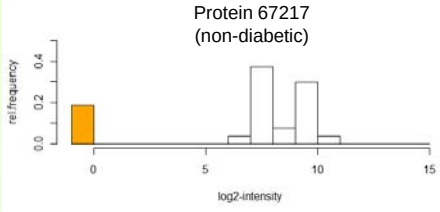


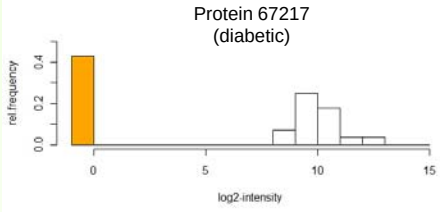
7 biolog.PMV
+ 0 techn. PMV

consonant

Test	p-value	effect estimate
one-part W	0.009	1.62
one-part T	0.227	0.97
truncated W	<0.001	1.93
truncated T	<0.001	2.33
Tobit	0.078	2.05
two-part W	**0.003**	**1.61**
two-part T	0.001	1.73
ELRT	<0.001	1.61
LIM-LR	**0.001**	**1.76**

Application: Example Proteins

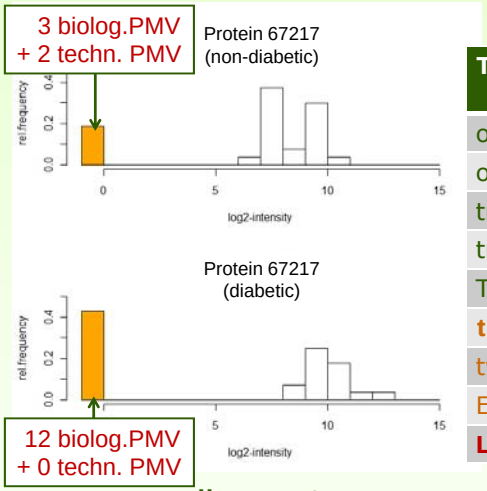




dissonant

Test	p-value	effect estimate
one-part W	0.701	0.00
one-part T	0.880	0.14
truncated W	0.394	0.67
truncated T	0.236	-1.08
Tobit	0.287	-1.41
two-part W	**0.001**	**1.60**
two-part T	<0.001	1.54
ELRT	<0.001	1.60
LIM-LR		

Application: Example Proteins



dissonant

Test	p-value	effect estimate
one-part W	0.701	0.00
one-part T	0.880	0.14
truncated W	0.394	0.67
truncated T	0.236	-1.08
Tobit	0.287	-1.41
two-part W	**0.001**	**1.60**
two-part T	<0.001	1.54
ELRT	<0.001	1.60
LIM-LR	**<0.001**	**1.72**

Conclusions

- strong dependence of results on underlying data-generating mechanism
- One-part Tests (incl. Tobit) distracted by biological PMVs
- **Two-part Wilcoxon test**:
  - „good compromise": high TPRs, low bias
  - if no prior knowledge about technical and biological PMVs
  - no distributional assumptions
- **LIM likelihood ratio test**:
  - generally low bias in all scenarios
  - higher TPRs in technical and consonant mixed scenarios
  - assuming log-normal distribution
  - classification of PMVs (technical vs. biological)

References

- M. Dakna et al: *Adressing the challenge of defining valid proteomic biomarkers and classifiers* (BMC Bioinformatics, 2010)
- A.P. Hallstrom: *A modified Wilcoxon test for non-negative distributions with a clump of zeros* (Statistics in Medicine, 2009)
- J. Jantos-Siwy et al*: Quantitative Urinary Proteome Analysis for Biomarker Evaluation in Chronic Kidney Disease* (J. Proteome Research, 2009)
- P.A. Lachenbruch: *Analysis of data with clumping at zero* (Biometrische Zeitschrift, 1976)
- P.A. Lachenbruch: *Comparison of two-part models with competitors* (Statistics in Medicine, 2001)
- S. Senn et al: *The ghost of departed quantities: approaches to dealing with observations below the limit of quantitation* (Stat. Med., 2012)
- S. Taylor & K. Pollard: *Hypothesis tests for point-mass mixture data with application to ´omics data with many zero values* (Stat. Appl. Genet. Mol. Biol., 2009)
- Y. Yang & D.G. Simpson: *Conditional decomposition diagnostics for regression analysis of zero-inflated and left-censored data* (Stat. Methods in Med. Research, 2012)
- D. Zhang et al: *Nonparametric methods for measurements below detection limit* (Statistics in Medicine, 2009)