Statistical Tests and Association Measures for Business Processes.

Sander Leemans, James McGree, Artem Polyvyanyy, Arthur ter Hofstede

IEEE TKDE 2023

Teaching and Research Area of Business Process Management Foundations and Engineering

RWTH AACHEN UNIVERSITY

Frequencies in processes

$$L_1 = [\langle \text{register}, \text{check}, \text{accept} \rangle^{10000}, \langle \text{register}, \text{check}, \text{reject} \rangle^{10000}, \langle \text{register}, \text{accept} \rangle^{1}]$$

Frequencies in processes

$$L_1 = [\langle \text{register}, \text{check}, \text{accept} \rangle^{10000}, \langle \text{register}, \text{check}, \text{reject} \rangle^{10000}, \langle \text{register}, \text{accept} \rangle^{1}]$$

$$L_2 = [\langle \text{register}, \text{check}, \text{accept} \rangle^{9500}, \langle \text{register}, \text{check}, \text{reject} \rangle^{1000} \langle \text{register}, \text{accept} \rangle^{95001}]$$

Frequencies in processes

$$L_1 = [\langle \text{register}, \text{check}, \text{accept} \rangle^{10000}, \langle \text{register}, \text{check}, \text{reject} \rangle^{10000}, \langle \text{register}, \text{accept} \rangle^{1}]$$

L_1

$$L_2 = [\langle \text{register}, \text{check}, \text{accept} \rangle^{9500}, \langle \text{register}, \text{check}, \text{reject} \rangle^{1000} \langle \text{register}, \text{accept} \rangle^{95001}]$$

L_2

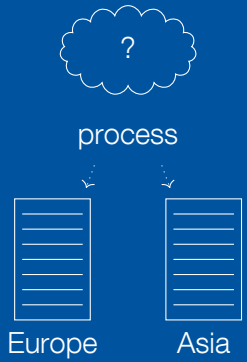
Statistical tests

Statistical tests

Europe

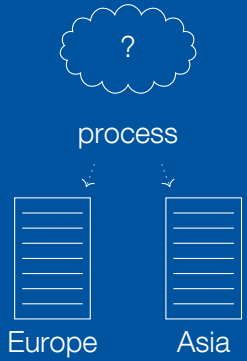
Asia

Statistical tests



Statistical tests

- Were two logs derived from the same underlying process?

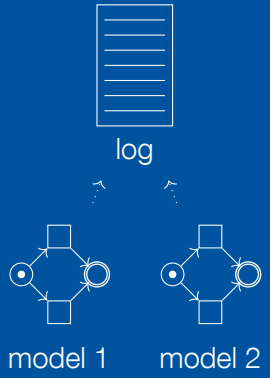


Statistical tests

- Were two logs derived from the same underlying process?
  $\rightarrow$ can we be 95% certain they are different?

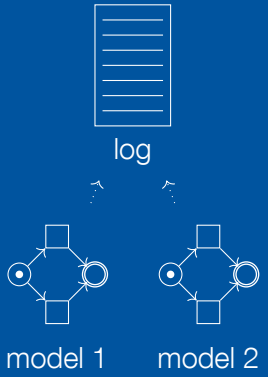
Statistical tests

- Were two logs derived from the same underlying process?
  $\rightarrow$ can we be 95% certain they are different?



Statistical tests

- Were two logs derived from the same underlying process?
  $\rightarrow$ can we be 95% certain they are different?
- Do two process models represent an event log equally well?



Statistical tests

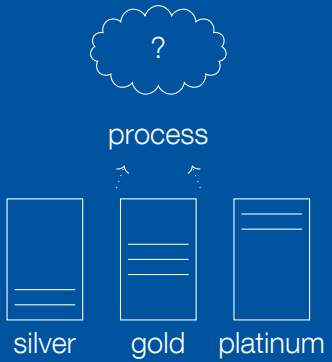
- Were two logs derived from the same underlying process?
  $\rightarrow$ can we be 95% certain they are different?
- Do two process models represent an event log equally well?
  $\rightarrow$ can we be 95% certain one is better?

Statistical tests

- Were two logs derived from the same underlying process?
  $\rightarrow$ can we be 95% certain they are different?
- Do two process models represent an event log equally well?
  $\rightarrow$ can we be 95% certain one is better?

Statistical tests

- Were two logs derived from the same underlying process?
  $\rightarrow$ can we be 95% certain they are different?
- Do two process models represent an event log equally well?
  $\rightarrow$ can we be 95% certain one is better?
- Are all sub-logs derived from identical processes?



Statistical tests

- Were two logs derived from the same underlying process?
  $\rightarrow$ can we be 95% certain they are different?
- Do two process models represent an event log equally well?
  $\rightarrow$ can we be 95% certain one is better?
- Are all sub-logs derived from identical processes?
  $\rightarrow$ can we be 95% certain all are equal?

Stochastic data vs. traces

⟨register, check, accept⟩

Stochastic data vs. traces

⟨register, check, accept⟩

- Numeric
- Ordinal
- Categorical

Stochastic data vs. traces

⟨register, check, accept⟩

- Numeric
- Ordinal
- Categorical

Stochastic data vs. traces

⟨register, check, accept⟩

- Numeric ϟ
- Ordinal ϟ
- Categorical

Stochastic data vs. traces

⟨register, check, accept⟩

- Numeric ϟ
- Ordinal ϟ
- Categorical ϟ

Stochastic data vs. traces

- Numeric ϟ
- Ordinal ϟ
- Categorical ϟ

⟨register, check, accept⟩

- Non-ordinal, distanced

Stochastic data vs. traces

- Numeric
- Ordinal
- Categorical

⟨register, check, accept⟩

- Non-ordinal, distanced

- Infinitely many traces

Stochastic data vs. traces

- Numeric ϟ
- Ordinal ϟ
- Categorical ϟ

⟨register, check, accept⟩

- Non-ordinal, distanced

- Infinitely many traces
- Probability of 0 is acceptable

Stochastic data vs. traces

- Numeric ϟ
- Ordinal ϟ
- Categorical ϟ

⟨register, check, accept⟩

- Non-ordinal, distanced

- Infinitely many traces
- Probability of 0 is acceptable

Log: “I observed ⟨register, check, accept⟩”
Model: “that trace has probability 0”

Stochastic data vs. traces

- Numeric ϟ
- Ordinal ϟ
- Categorical ϟ

⟨register, check, accept⟩

- Non-ordinal, distanced

- Infinitely many traces
- Probability of 0 is acceptable

Log: “I observed ⟨register, check, accept⟩”
Model: “that trace has probability 0”

χ^2 ϟ

Log vs. Categorical Attribute test

Log vs. Categorical Attribute test

Input: event log with categorical trace attribute φ

Log vs. Categorical Attribute test

Input: event log with categorical trace attribute φ
Hypothesis: all sub-logs are derived from identical processes

Log vs. Categorical Attribute test

Input: event log with categorical trace attribute φ
Hypothesis: all sub-logs are derived from identical processes

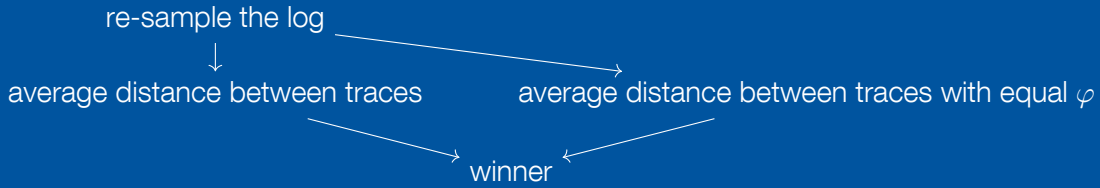
re-sample the log

Log vs. Categorical Attribute test

Input: event log with categorical trace attribute φ
Hypothesis: all sub-logs are derived from identical processes

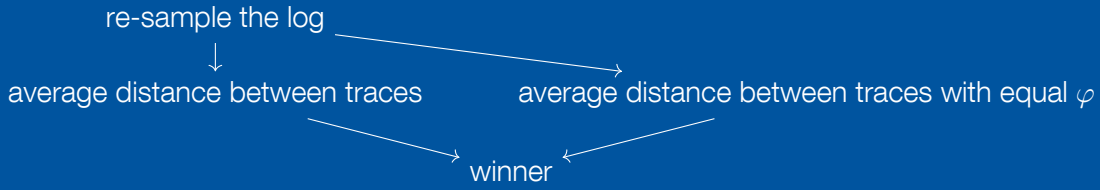
Log vs. Categorical Attribute test

Input: event log with categorical trace attribute φ
Hypothesis: all sub-logs are derived from identical processes



Log vs. Categorical Attribute test

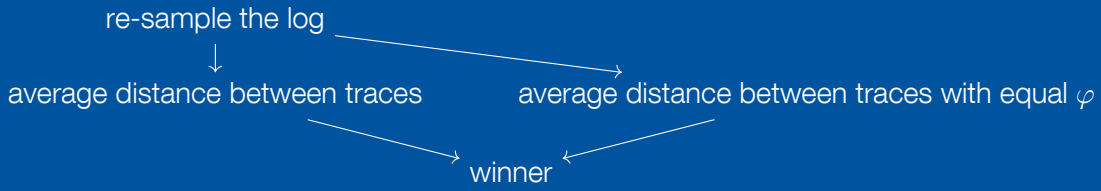
Input: event log with categorical trace attribute φ
Hypothesis: all sub-logs are derived from identical processes



Repeat 500 times

Log vs. Categorical Attribute test

Input: event log with categorical trace attribute φ
Hypothesis: all sub-logs are derived from identical processes



Repeat 500 times
If “with equal φ” wins 95% of the samples, then we have seen enough evidence that considering φ gives information on the process, so we can reject the hypothesis.

Association measure for processes

Road traffic fines log
What is the dependence between the requested loan amount and the process followed?

Association measure for processes

Road traffic fines log
What is the dependence between the requested loan amount and the process followed?

For every pair of traces in L:

Association measure for processes

Road traffic fines log
What is the dependence between the requested loan amount and the process followed?

For every pair of traces in L:
- compute distance between traces
- compute distance between loan amount

Association measure for processes

Road traffic fines log

What is the dependence between the requested loan amount and the process followed?

For every pair of traces in L:

compute distance between traces

compute distance between loan amount

Return correlation between trace distance and loan amount distance

Association measure for processes

Road traffic fines log

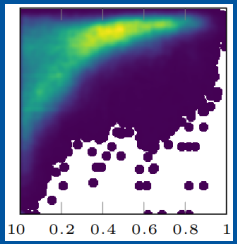
What is the dependence between the requested loan amount and the process followed?

For every pair of traces in L:

compute distance between traces

compute distance between loan amount

Return correlation between trace distance and loan amount distance



An example application

Queensland University of Technology, Brisbane, Australia
PhD journey through milestones

Improving PhD Student Journeys with Process Mining: Insights from a Higher Education Institution. Kanika Goel, Sander J.J. Leemans, Moe T. Wynn, Arthur H.M. ter Hofstede, Janne Barres. International Conference on Business Process Management industry track 2021.

An example application

Queensland University of Technology, Brisbane, Australia
PhD journey through milestones

Improving PhD Student Journeys with Process Mining: Insights from a Higher Education Institution. Kanika Goel, Sander J.J. Leemans, Moe T. Wynn, Arthur H.M. ter Hofstede, Janne Barres. International Conference on Business Process Management industry track 2021.

An example application

Queensland University of Technology, Brisbane, Australia
PhD journey through milestones

No significant difference in journey between male and female candidates*

Improving PhD Student Journeys with Process Mining: Insights from a Higher Education Institution. Kanika Goel, Sander J.J. Leemans, Moe T. Wynn, Arthur H.M. ter Hofstede, Janne Barres. International Conference on Business Process Management industry track 2021.

An example application

Queensland University of Technology, Brisbane, Australia
PhD journey through milestones

No significant difference in journey between male and female candidates*

*in terms of journey steps, trajectories and their frequencies

Improving PhD Student Journeys with Process Mining: Insights from a Higher Education Institution. Kanika Goel, Sander J.J. Leemans, Moe T. Wynn, Arthur H.M. ter Hofstede, Janne Barres. International Conference on Business Process Management industry track 2021.

You have been watching

Sander Leemans
s.leemans@bpm.rwth-aachen.de
https://leemans.ch
http://ebitools.org

RWTH AACHEN UNIVERSITY

You have been watching

- Log vs. log - unknown process test

Sander Leemans
s.leemans@bpm.rwth-aachen.de
https://leemans.ch
http://ebitools.org

Teaching and Research Area of Business Process Management Foundations and Engineering

RWTH AACHEN UNIVERSITY

You have been watching

- Log vs. log - unknown process test
- Process vs. process - log test
- Log vs. categorical attribute test

Sander Leemans
s.leemans@bpm.rwth-aachen.de
https://leemans.ch
http://ebitools.org

RWTH AACHEN UNIVERSITY

You have been watching

- Log vs. log - unknown process test
- Process vs. process - log test
- Log vs. categorical attribute test

- Process vs. attribute association

Future work

Sander Leemans
s.leemans@bpm.rwth-aachen.de
https://leemans.ch
http://ebitools.org

Teaching and Research Area of Business Process Management Foundations and Engineering

RWTH AACHEN UNIVERSITY

You have been watching

- Log vs. log - unknown process test
- Process vs. process - log test
- Log vs. categorical attribute test

- Process vs. attribute association
- Conformance vs. attribute association

Future work

Sander Leemans
s.leemans@bpm.rwth-aachen.de
https://leemans.ch
http://ebitools.org

Teaching and Research Area of Business Process Management Foundations and Engineering
RWTH AACHEN UNIVERSITY

You have been watching

- Log vs. log - unknown process test
- Process vs. process - log test
- Log vs. categorical attribute test

- Process vs. attribute association
- Conformance vs. attribute association

Future work

- Causal inference on processes

Sander Leemans
s.leemans@bpm.rwth-aachen.de
https://leemans.ch
http://ebitools.org

Teaching and Research Area of Business Process Management Foundations and Engineering
RWTH AACHEN UNIVERSITY

You have been watching

- Log vs. log - unknown process test
- Process vs. process - log test
- Log vs. categorical attribute test

- Process vs. attribute association
- Conformance vs. attribute association

Future work

- Causal inference on processes
- Study processual phenomena

Sander Leemans
s.leemans@bpm.rwth-aachen.de
https://leemans.ch
http://ebitools.org

Teaching and Research Area of Business Process Management Foundations and Engineering
RWTH AACHEN UNIVERSITY

You have been watching

- Log vs. log - unknown process test
- Process vs. process - log test
- Log vs. categorical attribute test

- Process vs. attribute association
- Conformance vs. attribute association

Future work

- Causal inference on processes
- Study processual phenomena
- Holy grail: medical trials

Sander Leemans
s.leemans@bpm.rwth-aachen.de
https://leemans.ch
http://ebitools.org

Teaching and Research Area of Business Process Management Foundations and Engineering
RWTH AACHEN UNIVERSITY