Stop Using the Wilcoxon Test:

Myth, Misconception and Misuse in IR Research

Julián Urbano

TU Delft

SIGIR 2026 · July 23rd · Melbourne

We've Been Here Before…

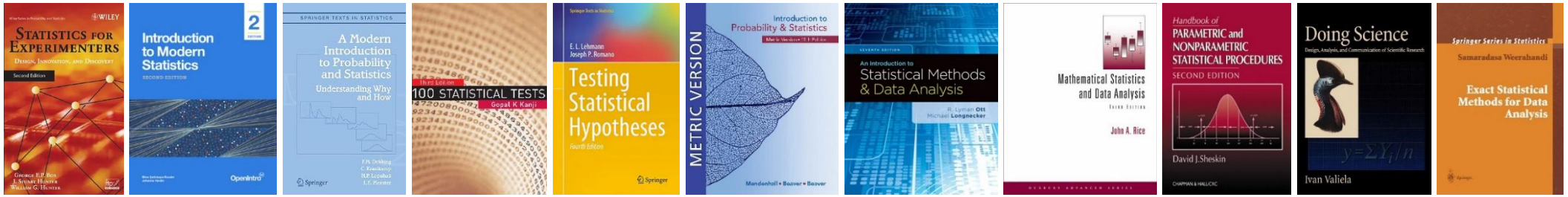
- System X vs system Y, evaluated over n topics
- Compute metric scores $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$
- Test the statistical hypothesis $H_0: \mu_X = \mu_Y$

- Two choices dominate the literature and IR practice (≈90% prevalence)
  - Paired Student's **t-test**: a parametric test assuming that scores are normally distributed
  - **Wilcoxon** signed-rank test: a non-parametric alternative when normality is violated

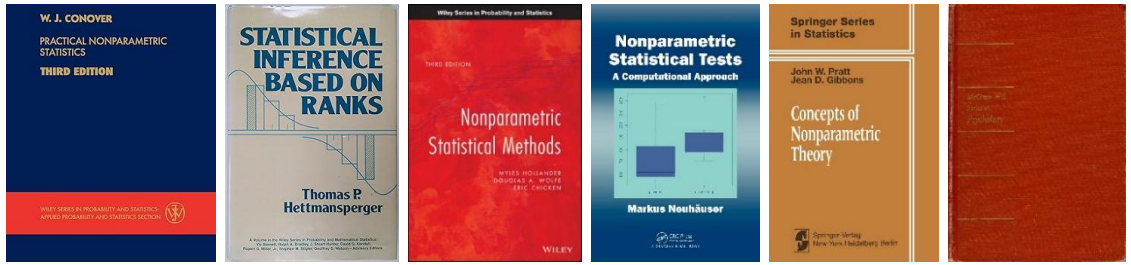
- This narrative is misleading, dangerous … and catastrophic for IR

Review of 25 textbooks

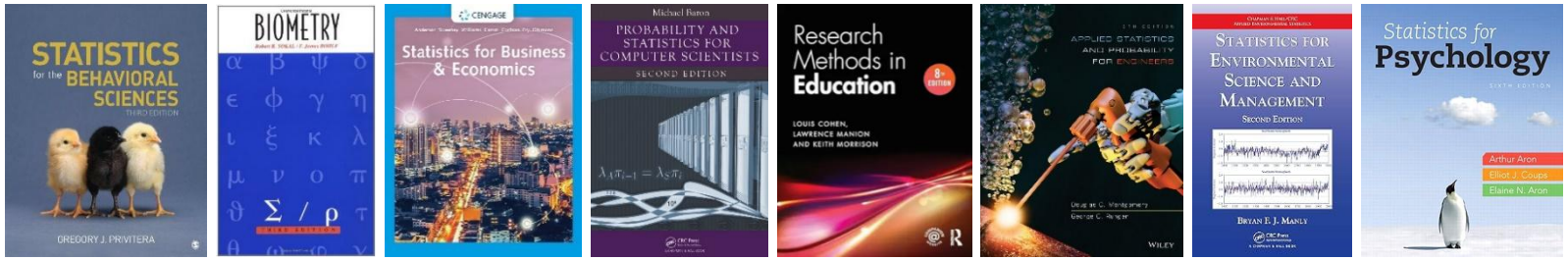
Statistics / Research Methods



Non-parametric Statistics



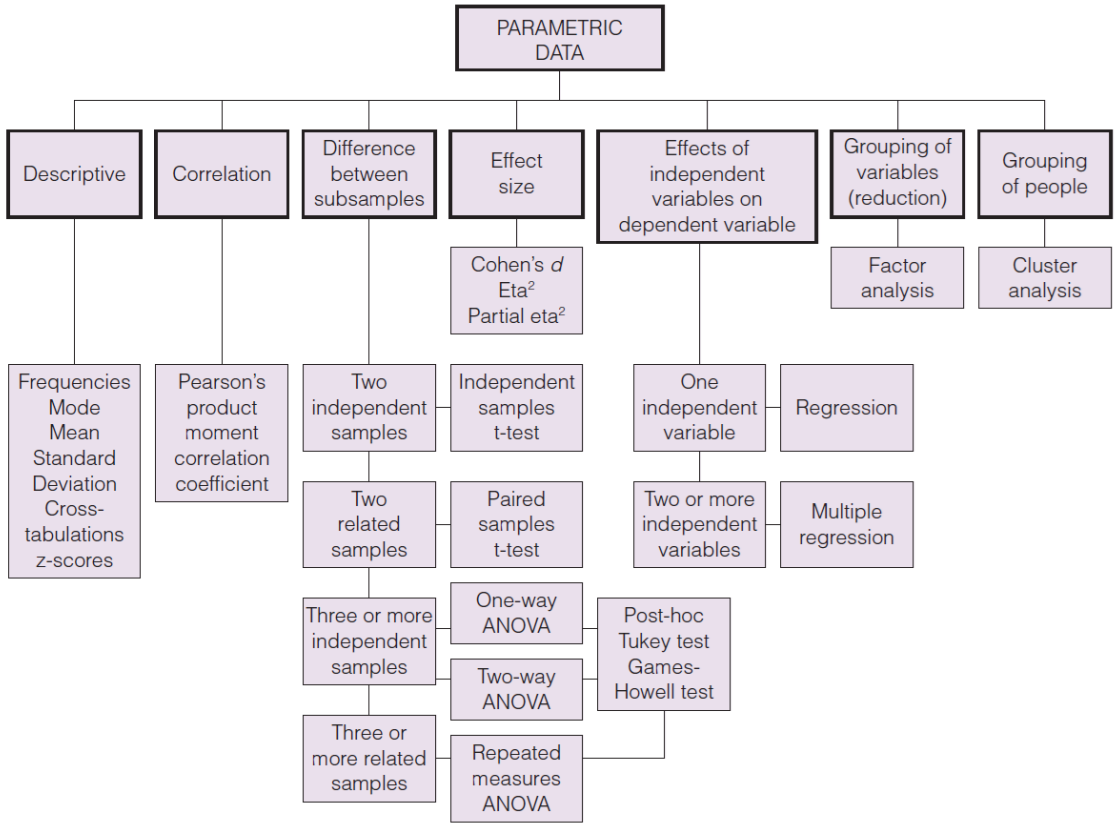
Field-specific Statistics

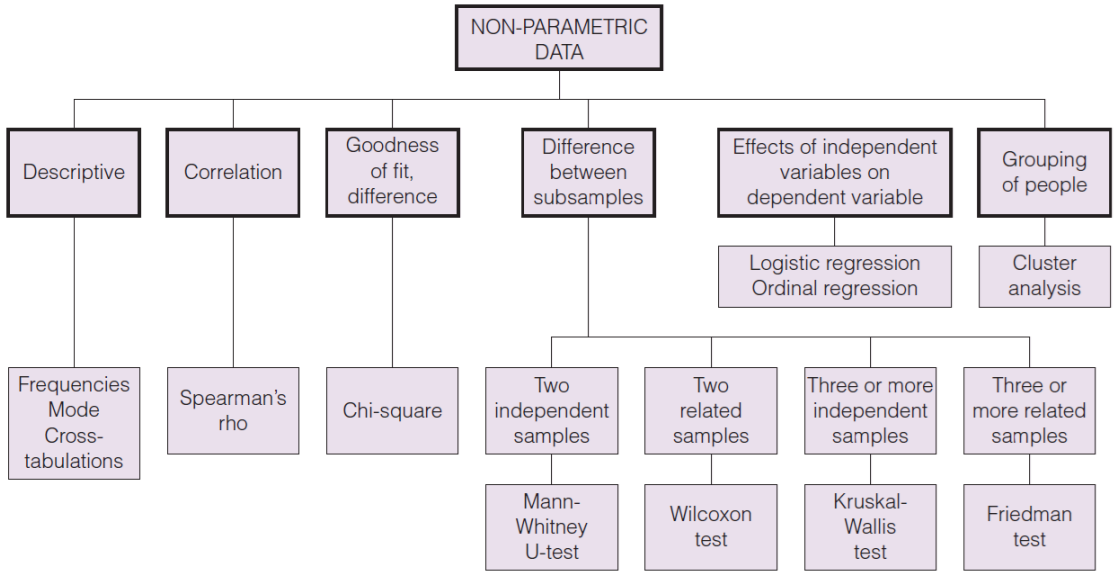


Myth, Misconception & Misue

What Test Should I Use?

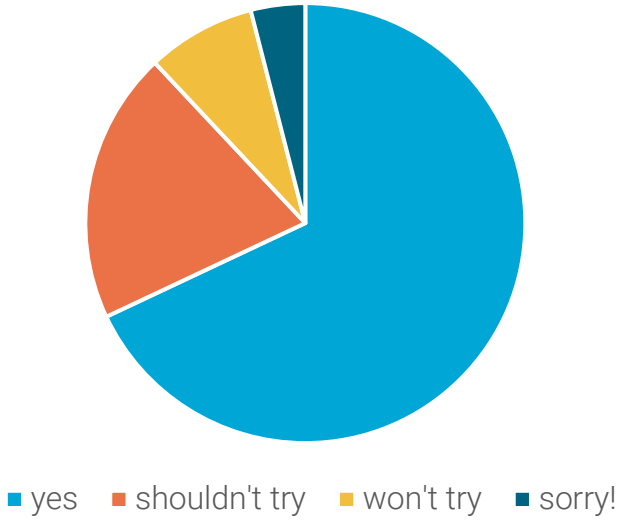
- Roadmaps/decision trees: parametric vs non-parametric



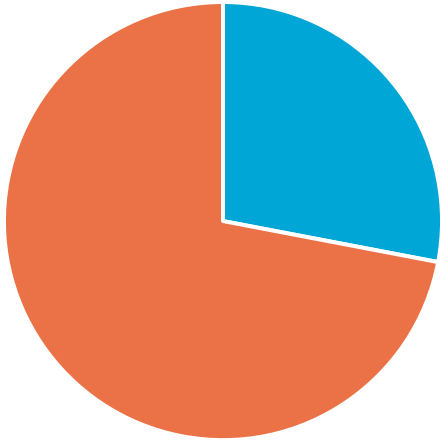


It’s all about Assumptions

Define "non-parametric" methods?



Assumptions…



…about what?

5 give conflicting statements

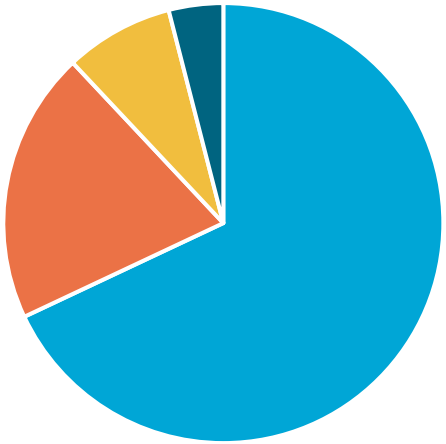
“Do you have normal distributions?”

“Clearly not!”

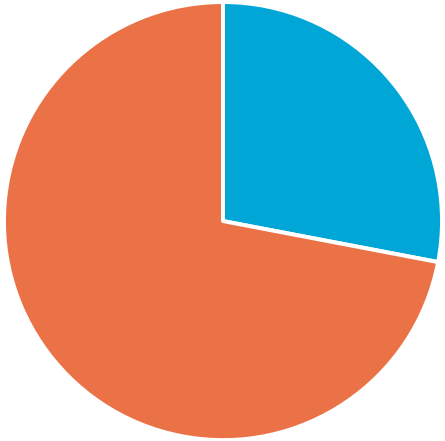
9 say they are **“distribution-free”**

It’s all about Assumptions

Define "non-parametric" methods?



Assumptions…



…about what?

…ements

Non-parametric methods have assumptions of their own…

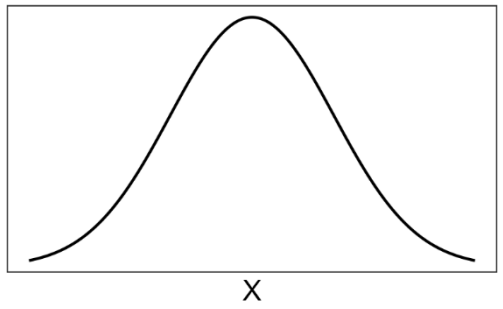
…but are typically neglected

Myth

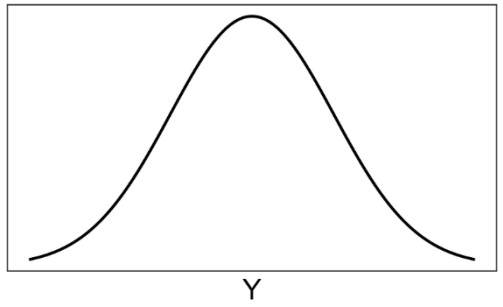
non-parametric methods are “safer”

t-test...as Explained in Books

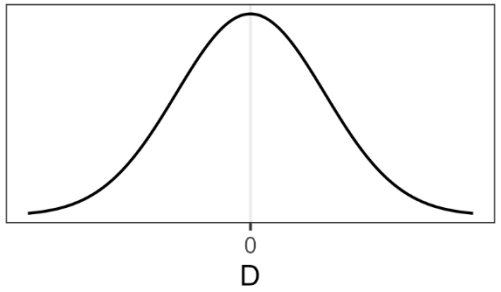
Population: X ~ N



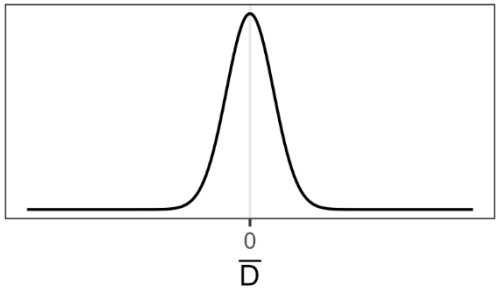
Population: Y ~ N



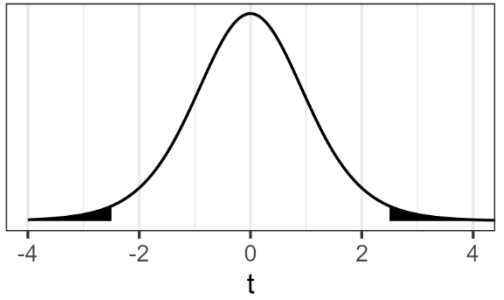
Population: $D \sim N(\mu_D, \sigma_D^2)$



Sample mean: $\overline{D} \sim N(\mu_D, \sigma_D^2/n)$

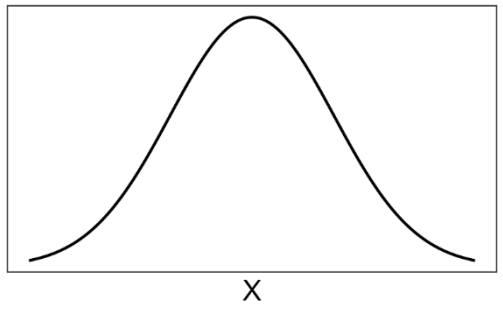


t-score: $t \sim T(n-1) \rightarrow$ p-value

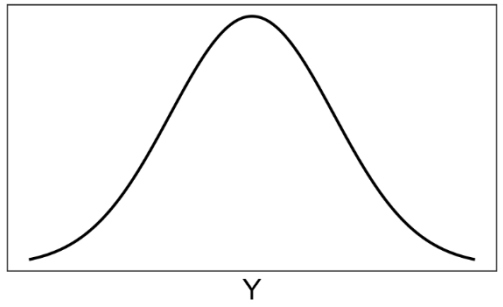


t-test...as Explained in Books

Population: X ~ N



Population: Y ~ N

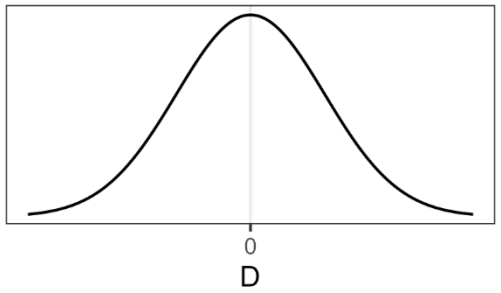


5 books

Which of these should be normal?

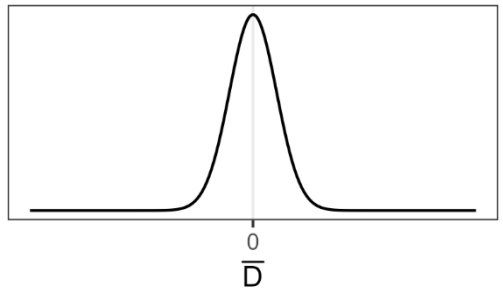
5-6 books unclear

Population: $D \sim N(\mu_D, \sigma_D^2)$

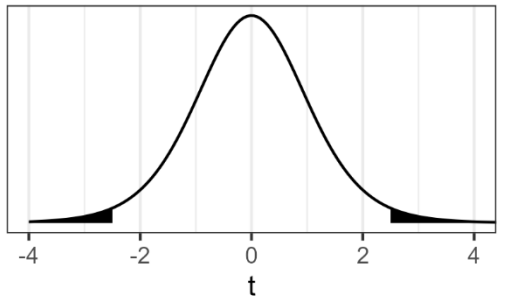


9-13 books

Sample mean: $\overline{D} \sim N(\mu_D, \sigma_D^2/n)$

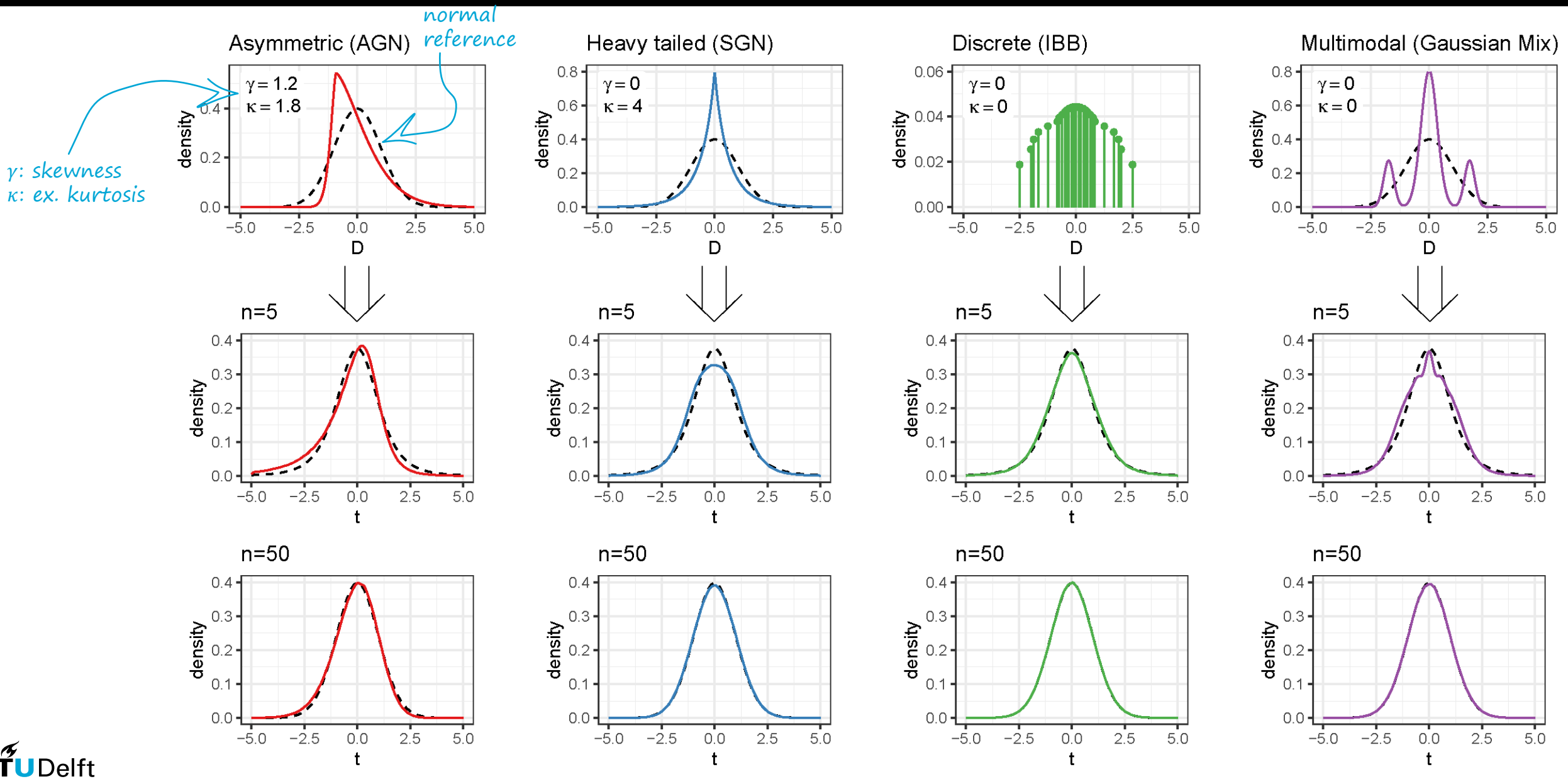


t-score: $t \sim T(n-1) \rightarrow$ p-value

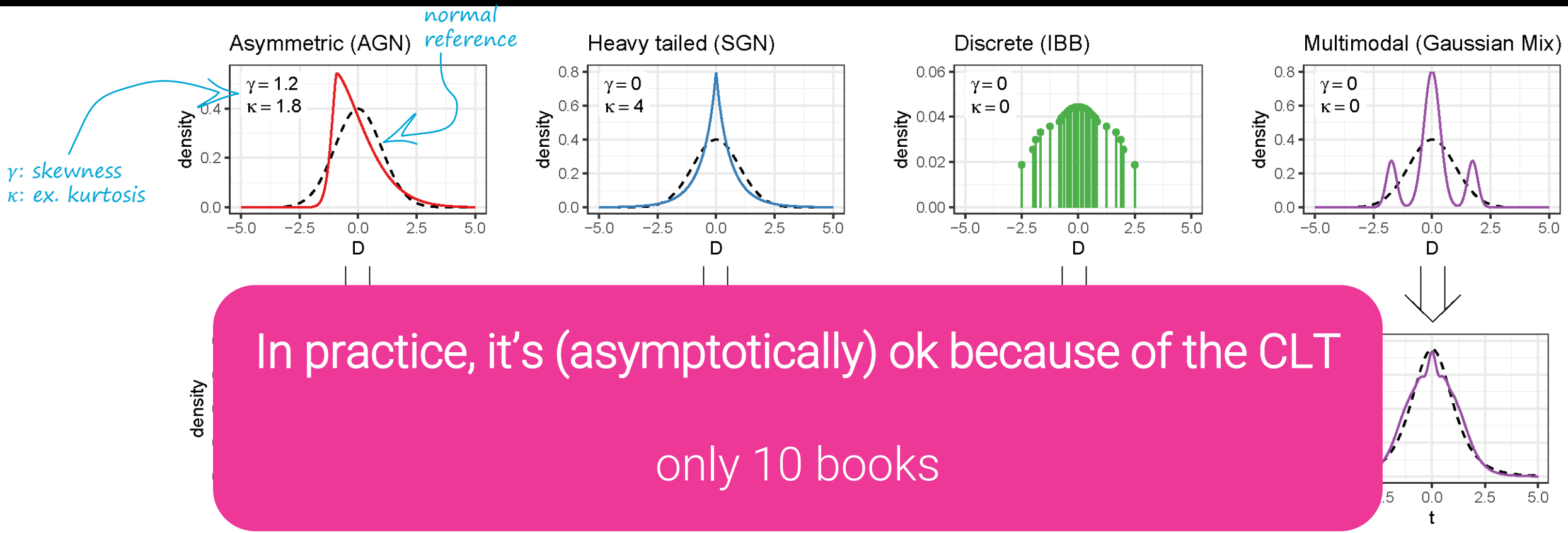


1-2 books

CLT to the Rescue



CLT to the Rescue



In practice, it's (asymptotically) ok because of the CLT

only 10 books

Myth

non-parametric methods are “safer”

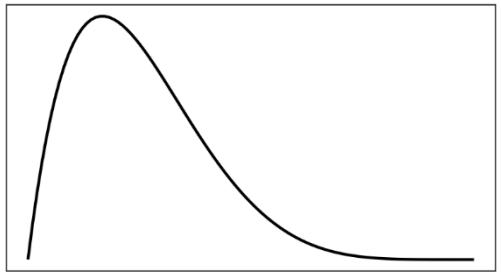
Misconception

t-test requires normally distributed scores

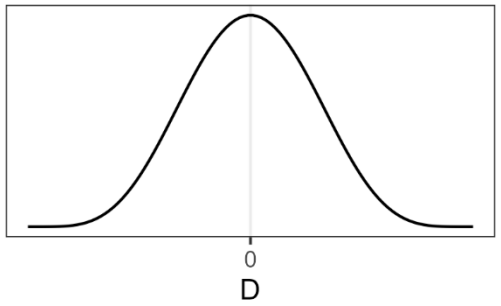
Misuse

Wilcoxon...as Explained in Books

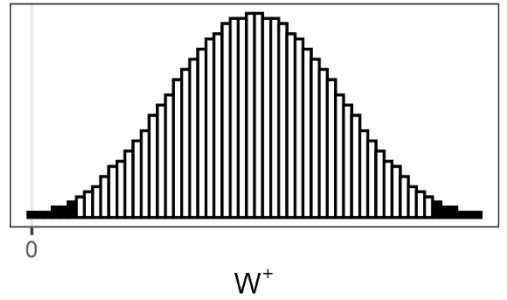
Population: X, Y ~ F



Population: D ~ Symmetric



W^+ ~ Wilcox(n) → p-value

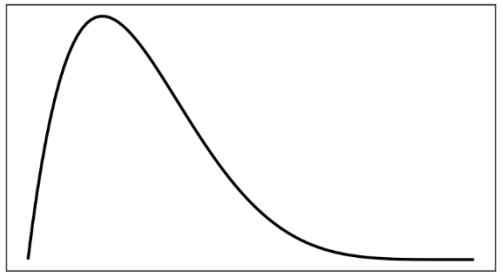


Wilcoxon...as Explained in Books

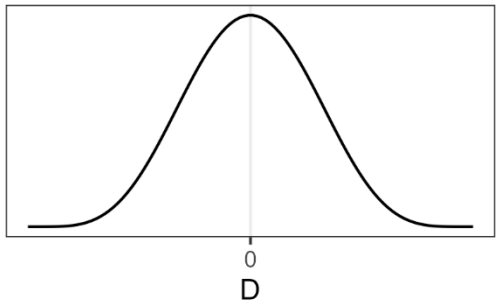
Formally a test for the median because it uses ranks (valid for the mean *only* under symmetry in D)

11-13 books

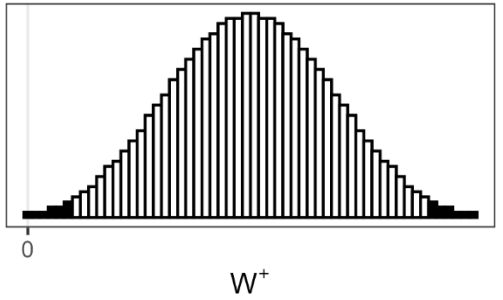
Population: X, Y ~ F



Population: D ~ Symmetric



W^+ ~ Wilcox(n) → p-value

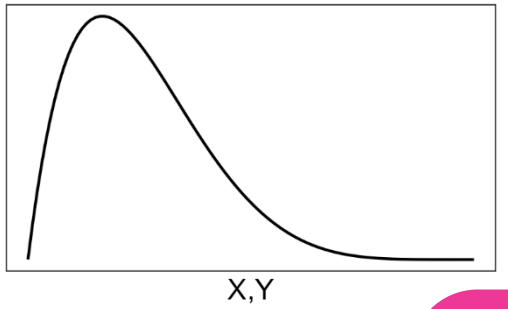


Wilcoxon…as Explained in Books

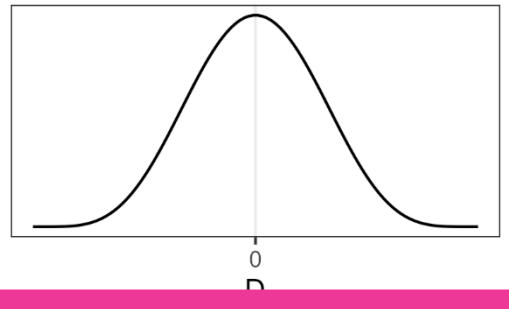
Formally a test for the median because it uses ranks
(valid for the mean *only* under symmetry in D)

11-13 books

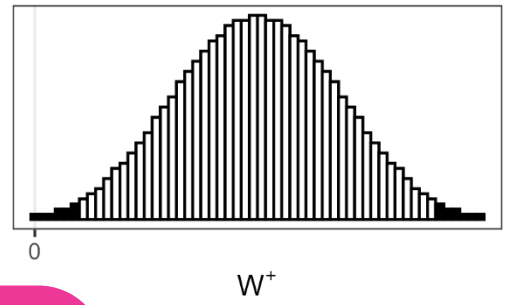
Population: X, Y ~ F



Population: D ~ Symmetric



W^+ ~ Wilcox(n) → p-value



With large n it turns into a test of symmetry…

…and it rejects H_0 for the wrong reason

Myth

non-parametric methods are "safer"

Misconception

t-test requires normally distributed scores

Misuse

rely on Wilcoxon, regardless of symmetry

Impact on IR Research

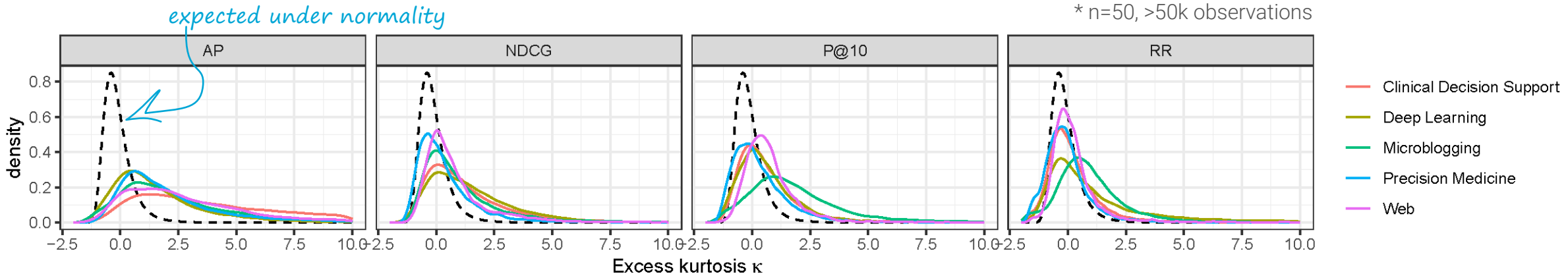
Departures from the Assumptions

- How do progressive departures from normality affect test validity?
  - Along controllable and interpretable dimensions
  - 6 levels: low, medium, high, very high, extremely high, pathologically high

Departures from the Assumptions

- How do progressive departures from normality affect test validity?
  - Along controllable and interpretable dimensions
  - 6 levels: low, medium, high, very high, extremely high, pathologically high

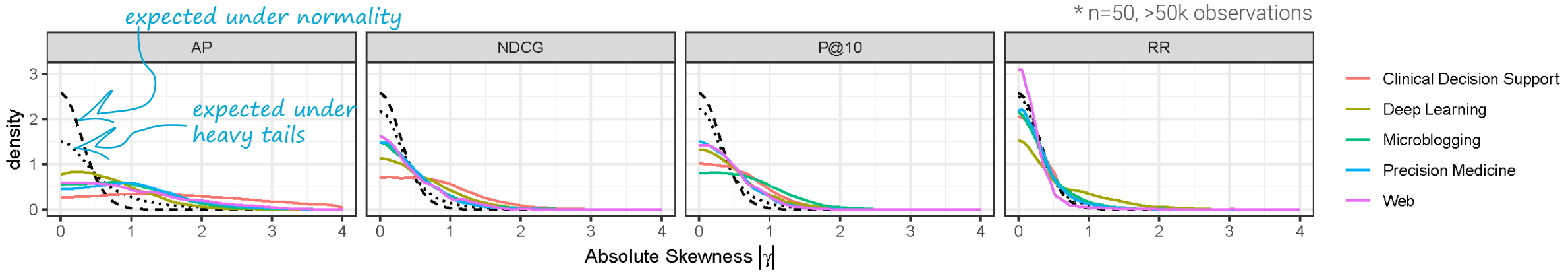
- Tail heaviness



Departures from the Assumptions

- How do progressive departures from normality affect test validity?
  - Along controllable and interpretable dimensions
  - 6 levels: low, medium, high, very high, extremely high, pathologically high

- Tail heaviness
- Asymmetry



Departures from the Assumptions

- How do progressive departures from normality affect test validity?
  - Along controllable and interpretable dimensions
  - 6 levels: low, medium, high, very high, extremely high, pathologically high

- Tail heaviness
- Asymmetry
- Discreteness
  - P@k and RR@k, k=1000, 500, 100, 50, 10, 5
- ~~Multimodality~~

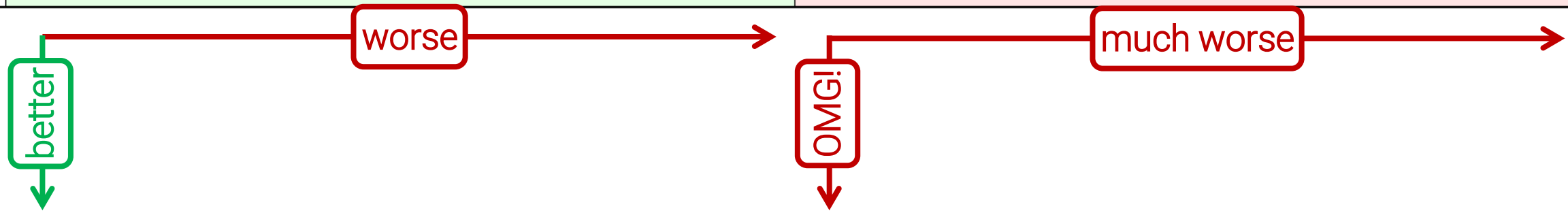
Type I Errors when Departing from Normality

n	*t*-test Low	Med.	High	Very H.	Extr. H.	Patho. H.	Wilcoxon Low	Med.	High	Very H.	Extr. H.	Patho. H.
	Asymmetric											
	$\gamma = 0.25$	$\gamma = 0.5$	$\gamma = 1$	$\gamma = 1.5$	$\gamma = 3$	$\gamma = 5$	$\gamma = 0.25$	$\gamma = 0.5$	$\gamma = 1$	$\gamma = 1.5$	$\gamma = 3$	$\gamma = 5$
5	.052	.056	.070	.087	.136	.189	0	0	0	0	0	0
50	.051	.052	.054	.058	.075	.098	.052	.061	.092	.141	.316	.493
500	.051	.051	.051	.051	.054	.058	.077	.156	.452	.757	.992	1
5000	.049	.050	.050	.050	.051	.051	.321	.841	1	1	1	1
	Heavy tailed											
	$\kappa = .5$	$\kappa = 1.5$	$\kappa = 3$	$\kappa = 5$	$\kappa = 15$	$\kappa = 30$	$\kappa = .5$	$\kappa = 1.5$	$\kappa = 3$	$\kappa = 5$	$\kappa = 15$	$\kappa = 30$
5	.047	.043	.039	.036	.030	.028	0	0	0	0	0	0
50	.050	.050	.049	.049	.046	.045	.049	.049	.049	.049	.049	.049
500	.051	.051	.051	.050	.050	.049	.050	.050	.050	.050	.050	.050
5000	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050
	Light tailed											
	$\kappa = -.2$	$\kappa = -.4$	$\kappa = -.7$	$\kappa = -.9$	$\kappa = -1.1$	$\kappa = -1.2$	$\kappa = -.2$	$\kappa = -.4$	$\kappa = -.7$	$\kappa = -.9$	$\kappa = -1.1$	$\kappa = -1.2$
5	.052	.054	.057	.060	.064	.066	0	0	0	0	0	0
50	.051	.051	.051	.051	.051	.051	.049	.049	.049	.049	.049	.049
500	.051	.051	.050	.051	.050	.050	.050	.050	.050	.050	.050	.050
5000	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050
	Discrete											
	$k = 1000$	$k = 500$	$k = 100$	$k = 50$	$k = 10$	$k = 5$	$k = 1000$	$k = 500$	$k = 100$	$k = 50$	$k = 10$	$k = 5$
5	.043	.043	.042	.039	.037	.044	0	0	0	0	.001	.002
50	.050	.050	.050	.050	.049	.049	.049	.049	.049	.049	.048	.047
500	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050	.050
5000	.050	.050	.050	.050	.050	.050	.051	.050	.050	.050	.050	.050

Nothing relevant here

Type I Errors under Asymmetry

n	*t*-test						Wilcoxon					
	Low	Med.	High	Very H.	Extr. H.	Patho. H.	Low	Med.	High	Very H.	Extr. H.	Patho. H.
	Asymmetric											
	$\gamma = 0.25$	$\gamma = 0.5$	$\gamma = 1$	$\gamma = 1.5$	$\gamma = 3$	$\gamma = 5$	$\gamma = 0.25$	$\gamma = 0.5$	$\gamma = 1$	$\gamma = 1.5$	$\gamma = 3$	$\gamma = 5$
5	.052	.056	.070	.087	.136	.189	0	0	0	0	0	0
50	.051	.052	.054	.058	.075	.098	.052	.061	.092	.141	.316	.493
500	.051	.051	.051	.051	.054	.058	.077	.156	.452	.757	.992	1
5000	.049	.050	.050	.050	.051	.051	.321	.841	1	1	1	1



Wilcoxon fails *systematically* under asymmetry

Large collections only make things worse

Three Perspectives

On Test Optimality

- Two schools to argue for test optimality
  - Orthodox: maximize power *subject to maintaining Type I error rate* at α level
  - Trade-off: maximize power *even if Type I error rate deviates mildly* from the α level

- **Wilcoxon should not even be part of the conversation**
  - It doesn't meaningfully trade power for Type I errors: it gets out of control
  - No multiple-comparisons adjustment can repair it
- Not advocating for t-test (though I really am)
  - Resampling methods probably better idea, but need further exploration

- Avoid goodness-of-fit rabbit holes in recent simulation studies
  - Examine how procedures behave when altering distributional properties

On Importing Results from Other Fields

- Isolated sources from other fields may not reflect the full scope of debate
  - Extensive literature on test robustness is typically overlooked in general summaries
- Even textbooks may be historically contingent or internally contested

- Avoid methodological imports based on single sources
  - Risk of arguments by authority that may be contested at the source field
  - Risk of hard-coding contested views into our practice, CfP, reviewer guidelines, etc.

- e.g. Wilcoxon in the textbooks
  - 50% change of missing the symmetry assumption; not really “safer”
  - Not enough to mention it; the implications and consequences were not sufficiently clarified

On Moving from Testing to Modeling

- We rely too much on contrast-based analysis
  - Isolate a single experimental factor while treating all others as fixed or non-existent
  - Classical pairwise system comparisons and modern ablation studies
- We miss other factors: how much variability, interactions, control

- We've had some work in that direction
  - Generalizability Theory for additional sources of uncertainty
  - Generalized linear models for testing system components
  - Bootstrap ANOVA to account for document collection variation

- Acknowledge and embrace variability, move from isolated hypothesis tests to explicit full-fledged analysis models incorporating experiment structure

To Take Home

Conclusions

- Wilcoxon is definitely *not* a safe choice with standard IR data
  - The *strong* assumption behind the t-test has a *weak* impact
  - The *weak* assumption behind the Wilcoxon test has a *catastrophic* impact

- Whatever the conversation on IR & stats, Wilcoxon shouldn't be a part of it
- Careful with methodological imports from other fields
- Embrace experimental variability; move towards explicit analysis models