Statistical Testing and Sample Size Calculation
Nonparametric Tests

06.03.2024

Christina Abele
Center for Thrombosis and Hemostasis
University Medical Center, Mainz
christina.abele@uni-mainz.de

Agenda

- Why do we need nonparametric testing?
- Two independent samples
- More than two independent samples
- Paired samples

Agenda

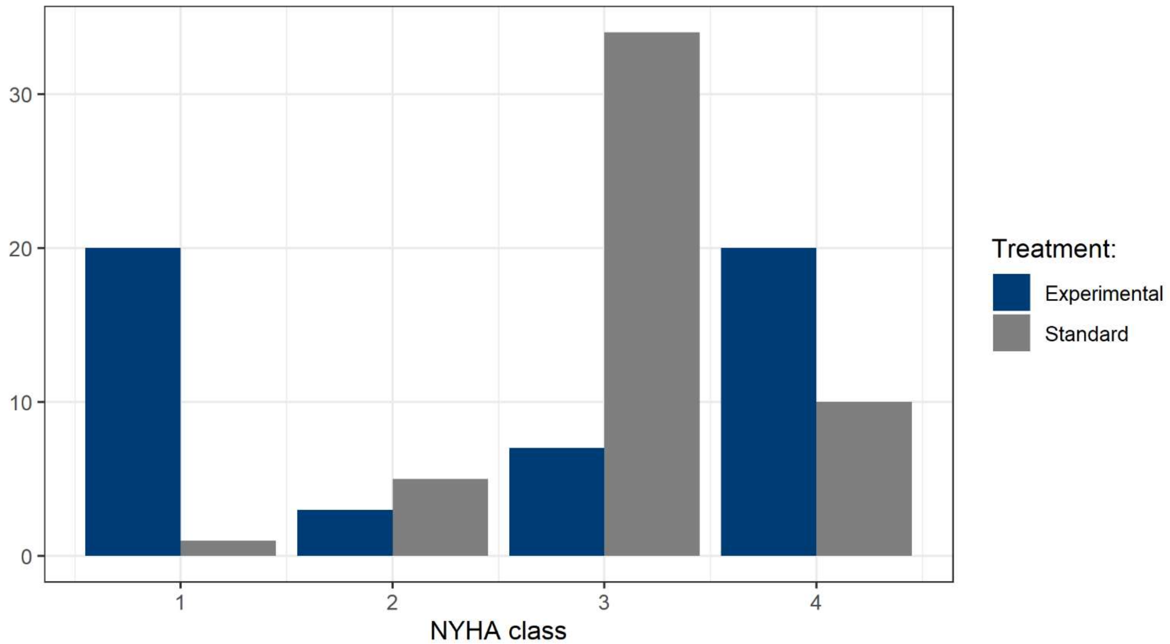
- **Why do we need nonparametric testing?**
- Two independent samples
- More than two independent samples
- Paired samples

Why do we need nonparametric testing?

Experimental vs. standard treatment for PH patients with moderate symptoms of heart failure (NYHA class III):

https://www.pennmedicine.org/updates/blogs/heart-and-vascular-blog/2022/july/heart-failure-classification--stages-of-heart-failure-and-their-treatments

Results after one year of treatment:



Why do we need nonparametric testing?

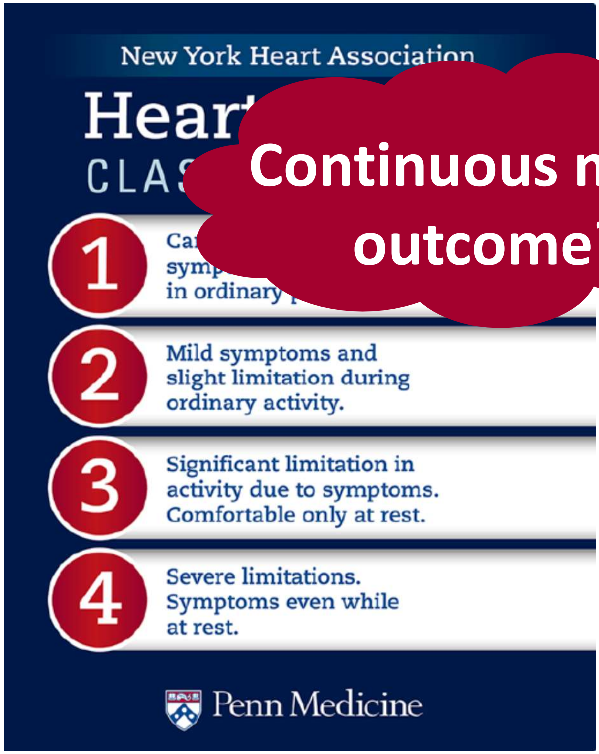
Experimental vs. standard tre… …nts wit… symptoms of heart failure (…

Normally distributed?!

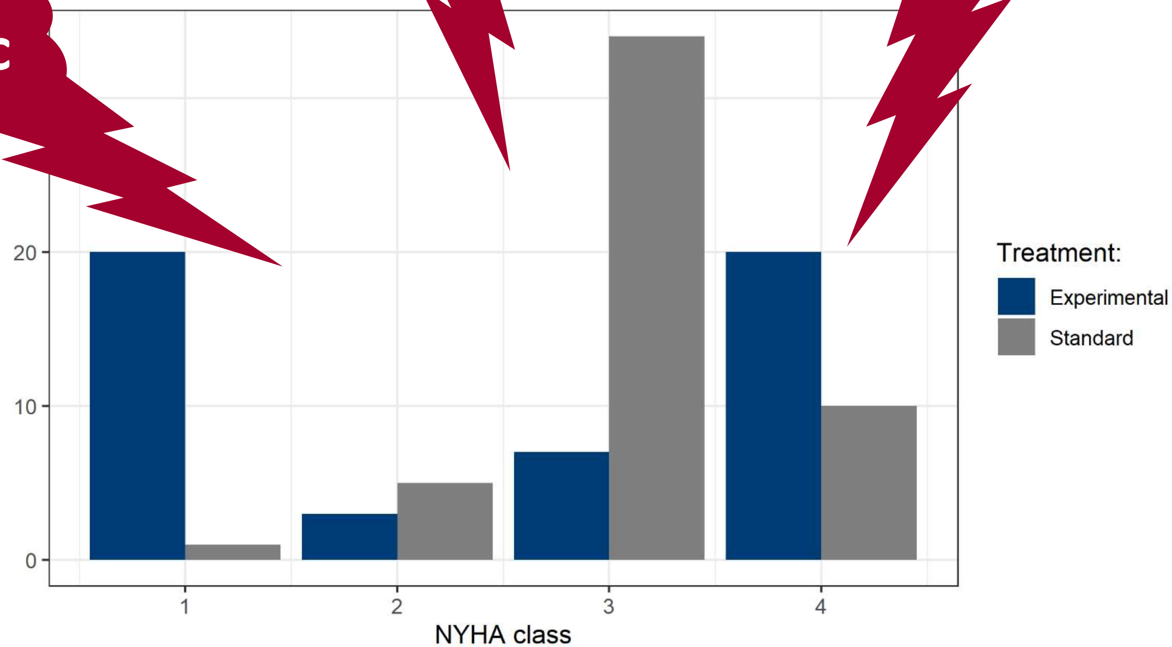
Equal variance?!

Continuous metric outcome?!

Res… …ear of t…



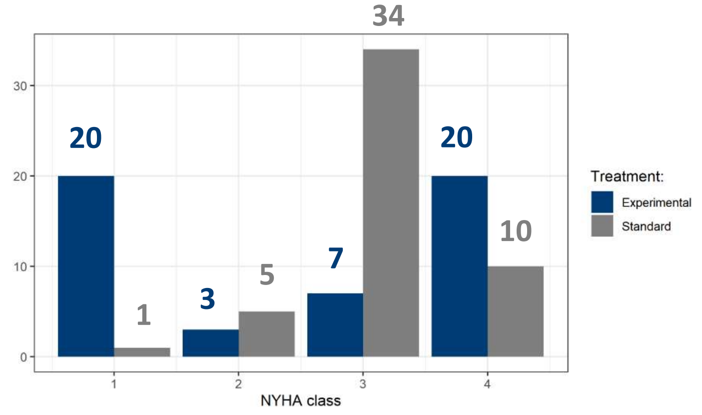
https://www.pennmedicine.org/updates/blogs/heart-and-vascular-blog/2022/july/heart-failure-classification--stages-of-heart-failure-and-their-treatments



Why do we need nonparametric testing?

t-test output:

```
      Welch Two Sample t-test
data:  outcome by treatment
t = -2.4404, df = 68.162, p-value = 0.01728
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.94518092 -0.09481908
sample estimates:
mean in group Experimental     mean in group Standard
                      2.54                       3.06
```



Pairwise comparisons (50·50 = 2 500 possible pairs):

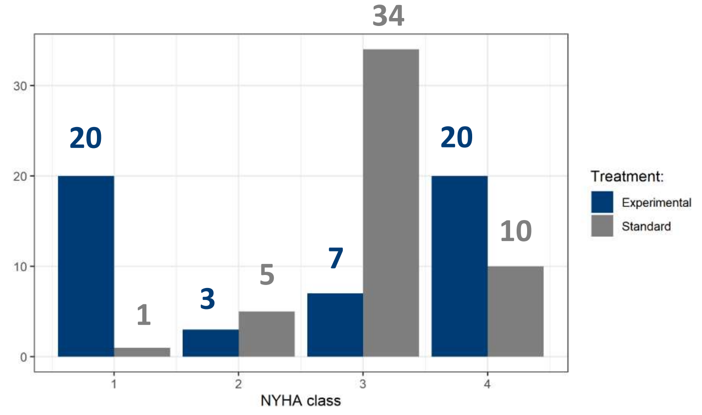
Experimental treatment works

- better in 20·49 + 3·44 + 7·10 = 1 182 cases (47.3%)
- equally well in 20·1 + 3·5 + 7·34 + 20·10 = 473 cases (18.9%)
- worse in 3·1 + 7·6 + 20·40 = 845 cases (33.8%)

Why do we need nonparametric testing?

t-test output:

```
      Welch Two Sample t-test
data:  outcome by treatment
t = -2.4404, df = 68.162, p-value = 0.01728
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.
samp
mean
```



We shouldn´t believe in the apparent superiority the t-test suggests!

Pai… …airs):

Exp…

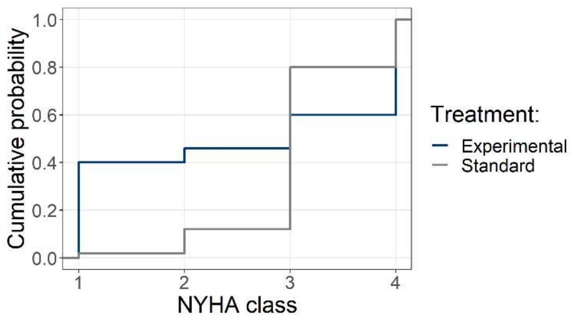
- better in 20·49 + 3·44 + 7·10 = 1 182 cases (47.3%)
- equally well in 20·1 + 3·5 + 7·34 + 20·10 = 473 cases (18.9%)
- worse in 3·1 + 7·6 + 20·40 = 845 cases (33.8%)

Why do we need nonparametric testing?

Alternatives to comparing average outcomes:

- Rank outcomes and compare mean ranks.
- Do pairwise comparisons of the outcomes and count which treatment works better more often.
- Do pairwise comparisons of the outcomes and count which treatment works better more often (count comparisons with equal outcomes for both sides with 50% each).

  The probability of this is called the **relative effect**. We can test if it is significantly different from 0.5.
- Compare cumulative distribution functions.



Agenda

- Why do we need nonparametric testing?
- **Two independent samples**
- More than two independent samples
- Paired samples

Two independent samples

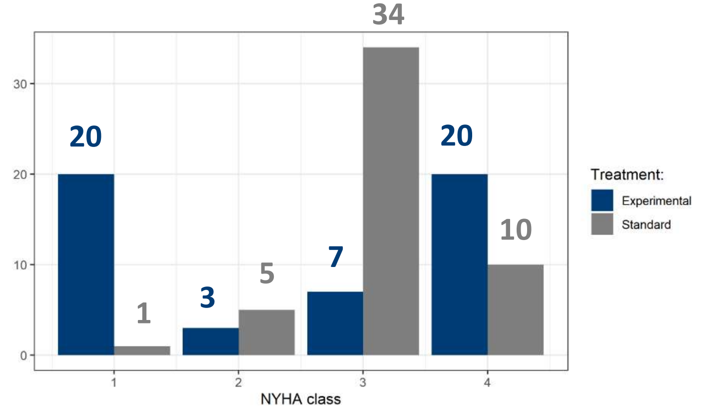
Wilcoxon-Mann-Whitney test (also called WMW test, Wilcoxon rank sum test or (Mann-Whitney-) U test)

- H_0: Same distribution of the outcome in both groups
  $\Rightarrow$ All possibilities of ranking have equal likelihood
- Test statistic for small samples: $T = (\sum R_2 - \sum R_1) - (n_2 - n_1) \cdot \frac{N+1}{2}$
  ($\sum R_{1/2}$: group sums of ranks, $n_{1/2}$: group sizes, $N = n_1 + n_2$)

  Distribution under H_0: Count how many of the N! ranking possibilities lead to a certain value of T $\Rightarrow$ exact p-value
- Large samples (normal approx.): Perform two-sample z- or t-test on ranks
- Exact test recommended if smaller group size is below 15

Two independent samples

Example (continued):



NYHA class	I	II	III	IV
Rank (midranks for ties)	11.0	25.5	50.0	85.5

WMW test in R and output:

```
library(exactRankTests)
wilcox.exact(outcome ~ treatment, data = ds, alternative = "two.sided", exact=FALSE)

	Asymptotic Wilcoxon rank sum test

data:  outcome by treatment
W = 1081.5, p-value = 0.2193
alternative hypothesis: true mu is not equal to 0
```

Two independent samples

Requirements for the WMW test:

- Two independent samples
- At least ordinal data

Very easy to satisfy!!

Approximate total sample size needed for power 1-β (Noether´s formula):

- $N = \frac{1}{12t(1-t)} \left(\frac{u_{1-\alpha/2} + u_{1-\beta}}{\vartheta - 1/2} \right)^2$

  (ϑ: expected relative effect under H_1, $t = n_1/N$: fraction of allocation to the first group, u_p: p-quantile of the standard normal distribution)
- Assumptions: sample large enough, no ties, same variance under H_0 / H_1.
- This is the only simple sample size formula in nonparamteric statistics.

Agenda

- Why do we need nonparametric testing?
- Two independent samples
- **More than two independent samples**
- Paired samples

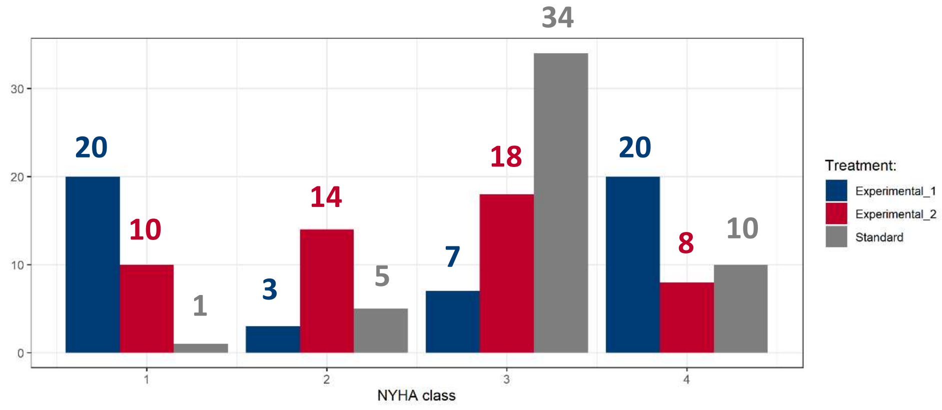
More than two independent samples

Kruskal-Wallis test

- Generalization of WMW test for more groups
- H_0: Same distribution of the outcome in **all** groups
- Test statistic: $T = \sum_{i=1}^{\#groups} \left(\widehat{\vartheta}_i - \frac{1}{2} \right)^2 \sim \chi^2_{\#groups-1}$
  ($\widehat{\vartheta}_i$: relative effect of group i vs. (un)weighted mean distribution)
- ANOVA using (pseudo-)ranks
- Not ideal with very small samples due to normal approximation, but exact version not implemented in R
- Unweighted mean distribution / pseudoranks recommended if group sizes are very different

More than two independent samples

Example (extended):



Kruskal-Wallis test in R and output:

```
library(pseudorank)
kruskal_wallis_test(outcome ~ treatment, data = ds, pseudoranks = FALSE)

Kruskal-Wallis Test

Call:
outcome ~ treatment

Test Statistic:  6.827197
Distribution of Statistic:  Chisq
Degrees of Freedom:  2
unweighted relative Effects / Pseudo-ranks:  FALSE
p-Value:  0.03292251
```

More than two independent samples

Requirements for the Kruskal-Wallis test:

- Independent samples
- At least ordinal data
- Sample size not too small (less than 5 per group)

When the Kruskal-Wallis test is significant ...

... be careful with pairwise comparisons to find out where the difference is:

- Relative effects are not transitive, so you can get „A is better than B, B is better than C, and C is better than A“ (e. g. google „Efron´s dice“).
- So **compare only against reference category** (standard of care, placebo).
- Don´t forget to adjust for multiplicity.

Agenda

- Why do we need nonparametric testing?
- Two independent samples
- More than two independent samples
- **Paired samples**

Paired samples

Typical comparisons: baseline/follow-up, left/right body side, twins, ...

Three different tests are commonly used:

	Sign test	Wilcoxon signed ranks test	Paired ranks test
Quantity analyzed	Direction/sign of difference	sign and rank of difference	difference of ranks
Outcome scale	metric, ordinal, binary	only metric	metric, ordinal, binary
Pairs with zero differences	omitted	omitted	included
Magnitude of difference	irrelevant	relevant	relevant
Power	low	higher	higher
Robustness	high	very sensitive with respect to symmetry of random errors	high

Paired samples

Paired ranks test

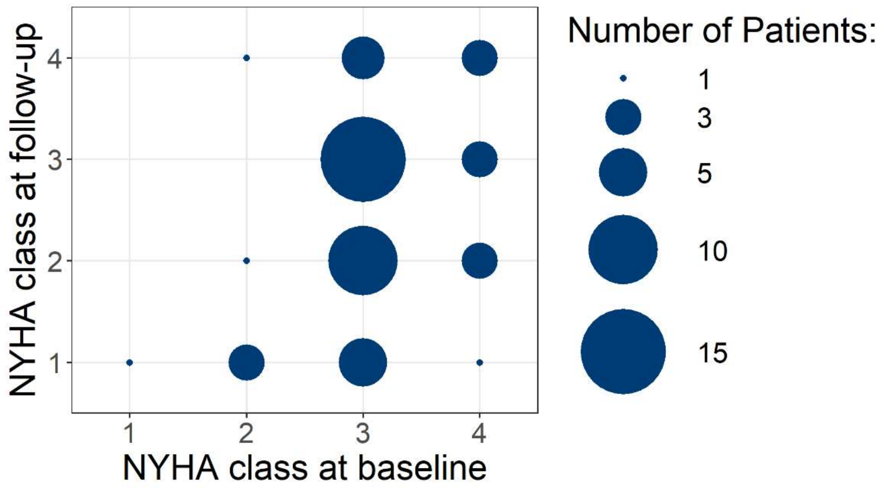
- H_0: Same (patient-dependent) distribution of the outcome in both groups
  $\Rightarrow$ All ranking scenarios where the two ranks related to each patient are the same have equal likelihood
- Test statistic for small samples: $T = \sum R_2 - \sum R_1 = \sum_{i=1}^{n}(R_{2i} - R_{1i})$
  ($\sum R_{1/2}$: group sums of ranks, $R_{1/2i}$: rank of patient i in group 1/2, n: sample size)

  Distribution under H_0: Count how many of the 2^n ranking possibilities lead to a certain value of T $\Rightarrow$ exact p-value
- Large samples (normal approx.): Perform one-sample z- or t-test on ranks
- Exact test recommended if sample size is below 15

Paired samples

Example:

		NYHA class at baseline			
		I	II	III	IV
NYHA class at follow-up	I	1	3	5	1
	II	0	1	10	3
	III	0	0	15	3
	IV	0	1	4	3



Data preparation for the paired ranks test in R:

```
ds <- ds %>%
  mutate(rank = rank(outcome, ties.method = "average"),
         rankx2 = 2 * rank)
```

Paired samples

Exact paired ranks test in R and output:

```
library(exactRankTests)
perm.test(rankx2 ~ time, data = ds, paired=TRUE, alternative = "two.sided", exact = TRUE)

	1-sample Permutation Test

data:  rankx2 by time
T = 2109, p-value = 0.0003987
alternative hypothesis: true mu is not equal to 0
```

Asymptotic paired ranks test in R and output:

```
t.test(rank ~ time, data = ds, paired = TRUE, alternative = "two.sided")

	Paired t-test

data:  rank by time
t = 3.8317, df = 49, p-value = 0.0003629
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  8.027037 25.732963
sample estimates:
mean of the differences
                  16.88
```