

3 Randomization Tests

3.1 Introductory Example

- a Hail prevention: (“Grossversuch IV” in central Switz. 1978-83)
 Does spreading of silver iodide into potential hail clouds
 diminish the total hail energy?
 (Simple ideas, only simple combinatorics and prob. needed.)

Target variable: Hail energy, measured for n clouds.

Two groups: treated versus control.

$$Y_i : \text{Hail energy for cloud } i$$

$$G_i = \begin{cases} 1 & \text{if cloud } i \text{ is treated,} \\ 0 & \text{otherwise.} \end{cases}$$

We expect Y_i to be usually smaller for $G_i = 1$ than for $G_i = 0$.

b Observed:

$Y_i = y_i^*$		16672	25	855	0	152	0	46	1219
$G_i = g_i^*$		1	1	0	0	0	1	1	0

g_i^* : random choice of the clouds that are treated.

(In reality there were 216 clouds of which 94 were treated.)

Statistical test! H_0 : no effect.

(\longrightarrow Proof by contradiction!)

t test for independent samples?

We do not like to assume any particular distr. for the Y_i s!!

3.2 Statistical idea

a **Null hypothesis** = Probability model

usually is a distribution of the Y_i . $G_i = g_i^*$ assumed to be given.
 Randomization tests: G_i random, $Y_i = y_i^*$ considered as fixed!
 (Analysis “conditional, given the y_i^* s”.)

If the treatment has no influence on hail energy,
 the same observations y_i^* would result,
 if the treatment had been given by $\underline{g}^{(1)} = [0, 1, 0, 0, 1, 1, 0, 1]$
 or according to any other choice.

Random choice:

Each choice of $n/2 = 4$ elements from $n = 8$ has the same probability

$$p = \binom{8}{4}^{-1} = \frac{1}{70}$$

This determines the null hypothesis.

- b **Test statistic:** designed to assume extreme values when the alternative is true.

Alternative: y_i^* with $g_i^* = 1$ are generally smaller.

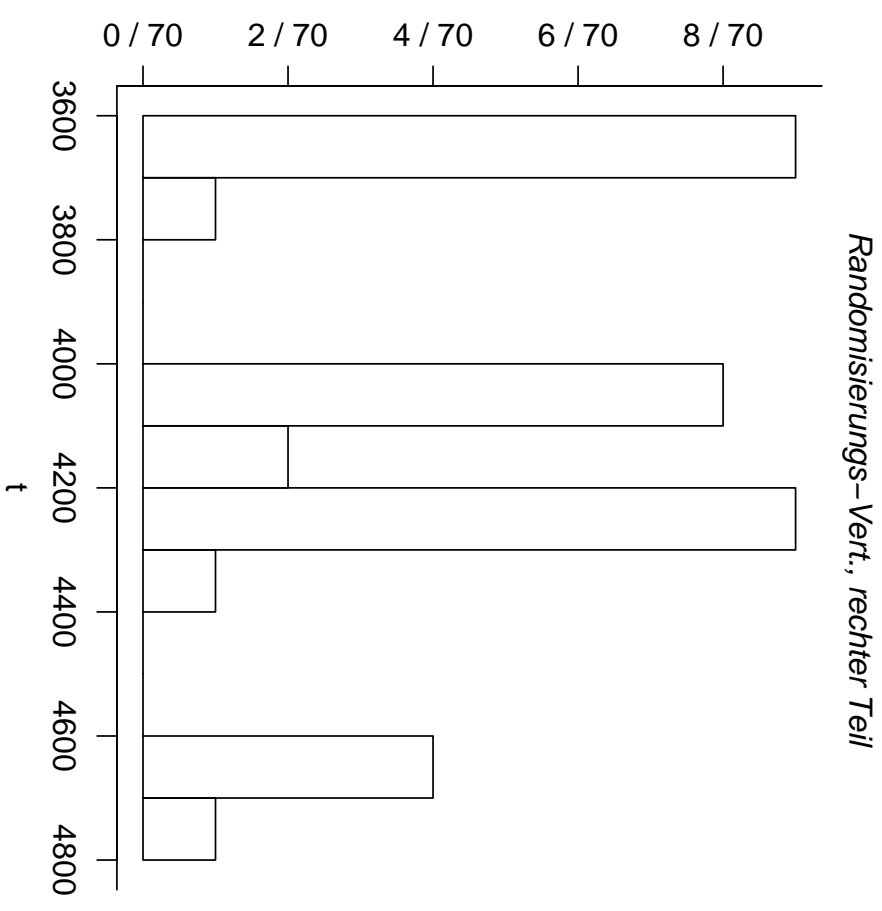
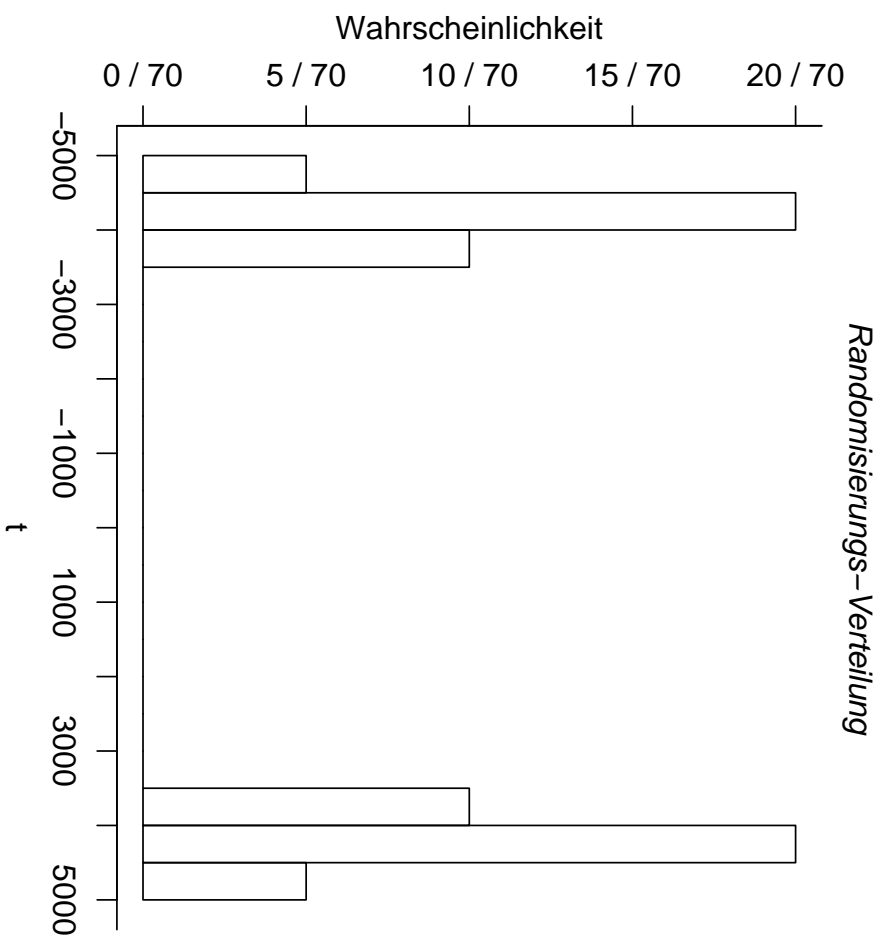
$$T(\underline{g}, \underline{y}^*) = \frac{1}{n/2} \sum_{i:g_i=0} y_i^* - \frac{1}{n/2} \sum_{i:g_i=1} y_i^* = \frac{2}{n} \sum_i y_i^* (1 - 2g_i).$$

- c What is the distribution of T under H_0 ?

y_1^*, \dots, y_n^* given $\rightarrow \leq \binom{n}{n/2}$ possible values for T .

$$P(T(\underline{G}, \underline{y}^*) = t) = \frac{\#\{\underline{g} \mid T(\underline{g}, \underline{y}^*) = t\}}{\binom{n}{n/2}}$$

„randomization distribution“



d **Rejection region:** $\alpha = 5\%$ most extreme values (as precisely as possible).

Example: $\{t \mid t \geq 4643.25\}$ (one sided).

e Experiment:

$$T(\underline{g}^*, \underline{y}^*) = \frac{1}{4}(855 + 0 + 152 + 1219) \\ - \frac{1}{4}(16672 + 25 + 0 + 46) = -3629.25$$

An effect in the wrong direction was observed!

Null hypothesis is not rejected; no effect is demonstrated.

* Assumption of the test: **Independence**

→ Randomization among 76 “potential hail days”

Among these, 33 have been assigned to the treatment.

Number of treated days is random.

→ Analysis conditional on the number of hail days with treatm.

g $\binom{76}{33} = 36 \cdot 10^{20}$ possible choices

→ Simulation of the randomization distribution.

3.3 Tests for the Two Sample Problem

- a Randomization tests are adequate even if the experimental procedure does not contain any randomization.

Assumptions in this case:

- The observations must be equally distributed under H_0 and
- independent

Then, the presupposed probability α of error of the first kind holds precisely.

The randomization tests are in this sense the “gold standard” of statistical tests.

(* Weaker assumption: “Exchangeability”.)

b If the **observations are random**:

Sample $[Y_1, \dots, Y_n]$ \longrightarrow ordered sample $Y_{[1]}, \dots, Y_{[n]}$
or empirical distribution function \hat{F}_n (s. Bootstrap)

Distribution of the test statistic, conditional on \hat{F}_n ,
is the randomization distribution.

Cond. prob. of an error of the first kind, given \hat{F}_n , is α

c **Arbitrary test statistic.**

The difference of means is not robust ...

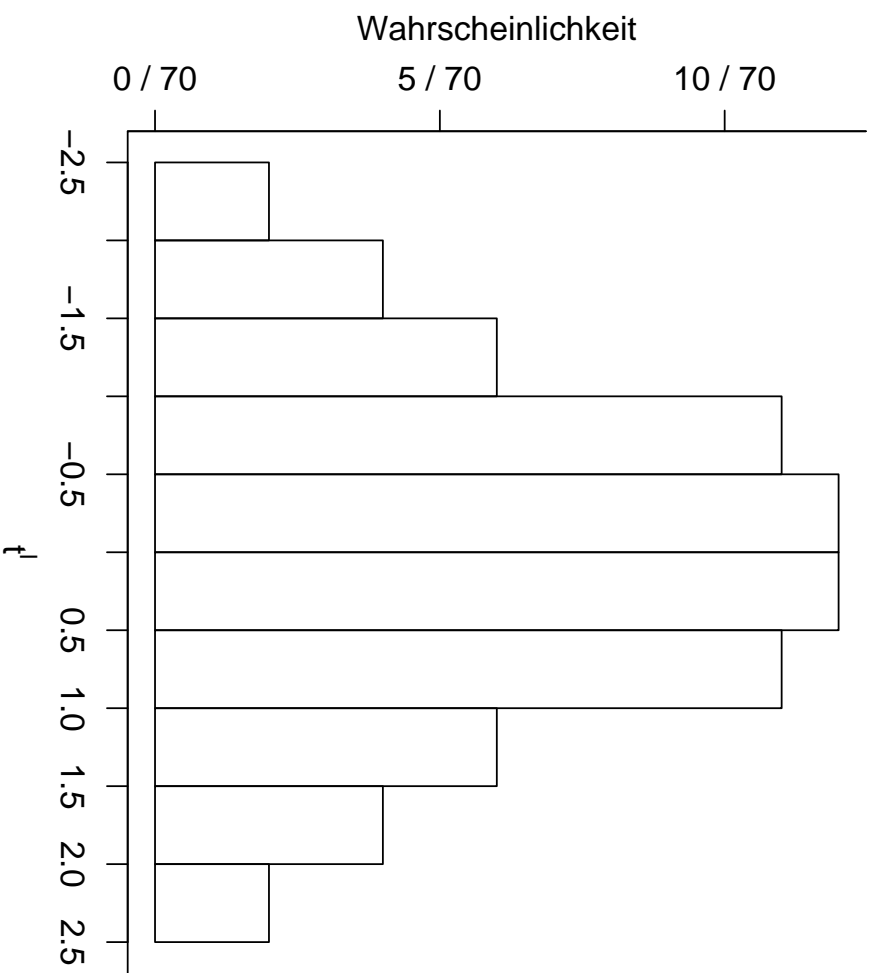
Optimal test statistic? → optimize power for alternative(s)!

Needs **fixed** (family of) distribution(s)

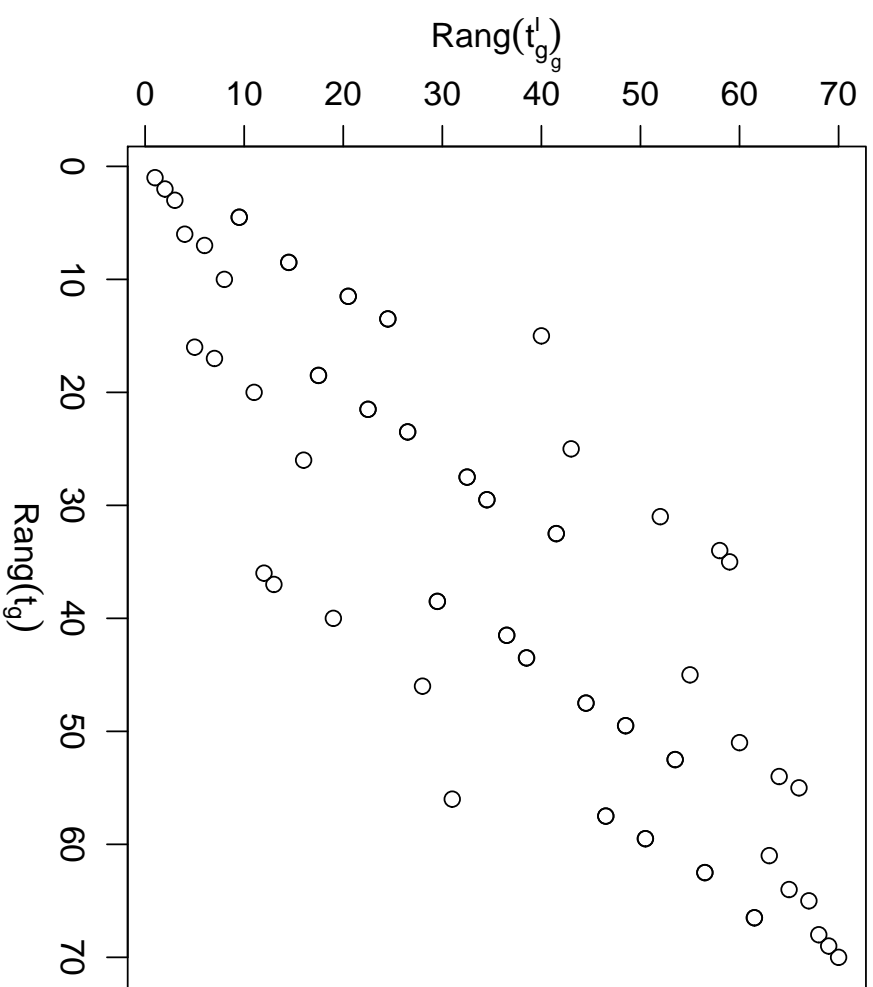
→ optimal test statistic (e.g., likelihood ratio test)

d Example: Log transformation, then difference of means
(preferably robustized)

Rand. Vert. für log. Werte



Vergleich der Test-Statistiken



e **Robustness.** Why should we use a robust test statistic, if the test keeps the level without this “preventive measure”?

f **Rank sum test** of Wilcoxon, Mann and Whitney (U-Test),

$$T\langle \underline{g}, \underline{y} \rangle = \sum_{g_i=1} R_i = \sum_i g_i R_i ,$$

Quite robust \rightarrow First choice for the 2 sample problem

Distribution of the test statistic under H_0 as before.

g^* Hail experiment: Complicated test statistic, two-dimensional
 \rightarrow two dimensional rejection region.

3.4 One Sample and Matched Pairs

a Example Tranquilizer.

Target variable: „Hamilton depression scale factor IV”.
9 patients, before and after taking the tranquilizer

before ($X_i^{(1)}$)	1.83	0.50	1.62	2.48	1.68	1.88	1.55	3.06	1.30
after ($X_i^{(2)}$)	0.878	0.647	0.598	2.05	1.06	1.29	1.06	3.14	1.29
Difference ($-Y_i$)	0.952	-0.147	1.022	0.43	0.62	0.59	0.49	-0.08	0.01

b **Matched Pairs.**

Differences $Y_i = X_i^{(2)} - X_i^{(1)}$

distributed **symmetrically** around 0?

H_0 : For each Y_i , + and – sign is equally probable

$G_i = \text{sign}, |Y_i| = „Y_i”$ of a two sample problem.

For each configuration $\underline{g}^{(\ell)} = [g_1^{(\ell)}, \dots, g_n^{(\ell)}]$

the probability is $= 1/2^n$.

c Fix a test statistic $T(\underline{g}, \underline{z})$

$$g_i = +1 \text{ or } -1, \quad z_i > 0.$$

Randomization distr. $P(T(\underline{G}, \underline{z}) = t) = \#\{g \mid T(\underline{g}, \underline{z}) = t\} / 2^n$

- $T(\underline{g}, \underline{z}) = (1/n) \sum_i g_i z_i = \text{ave}_i \langle y_i \rangle$
corresponds to the t test for matched pairs.

- $T(\underline{g}, \underline{z}) = \#\{i : g_i = 1\}$: sign test.

- $T(\underline{g}, \underline{z}) = \sum_{i:g_i=1} R_i$, R_i : rank of z_i :
signed rank test of Wilcoxon

e Example:

```
> wilcox.test(d.tranquilizer[,1], d.tranquilizer[,2],  
             paired=TRUE)  
      Wilcoxon signed rank test  
data:  d.tranquilizer[, 1] and d.tranquilizer[, 2]  
V = 40, p-value = 0.03906  
alternative hypothesis: true mu is not equal to 0  
barely significant.
```

Beware of before-after comparisons!

Adequate: Comparison with control or cross over experiment

3.5 Estimators and Confidence Intervals

a **Model:** Testing problem was: Is the distr. symmetric around 0?

More general: ... symmetric around μ

$\Leftrightarrow Y_i - \mu$ symmetric around 0.

Test: Test statistic $T\langle \underline{g}, \underline{y} - \mu \underline{1} \rangle$.

Large values indicate deviation from $H_0 : \mu$.

b This yields an **estimator**:

$$\hat{\mu} = \arg \min_{\mu} \langle T\langle \underline{g}, \underline{y} - \mu \underline{1} \rangle \rangle$$

- c Signed rank Test \longrightarrow Hodges-Lehmann estimator.

Form Walsh averages $(X_h + X_i)/2$.

$$\hat{\mu} = \text{med}_{h \leq i} \langle (X_h + X_i)/2 \rangle .$$

Example Tranquilizer: 45 Walsh averages

-0.1470, -0.1135, -0.0800, -0.0685, -0.0350, 0.0100, ..., 1.022

Median $\hat{\mu} = 0.46$

d* Derivation: $X_{[k]}$ k th smallest value.

$$X_{[k]} > 0, Z_{hk} = (X_{[h]} + X_{[k]})/2, h < k$$

$$Z_{hk} < 0, \text{ if } |X_{[h]}| > |X_{[k]}|.$$

$$\#\{Z_{hk} < 0\} = \#\{h \mid |X_{[h]}| < |X_{[k]}|\} = R_{[k]} - 1$$

$$R_{[k]} = \#\{h \mid Z_{hk} > 0, h \leq k\}.$$

$$X_{[k]} < 0 \implies Z_{hk} < 0, \text{ if } h < k.$$

$$T(\underline{g}, \underline{z}) = \sum_{i:g_i=1} R_i = \#\{[h, k] \mid Z_{hk} > 0, h \leq k\}$$

Null hypothesis $\mu = \mu_0$:

$$T(\underline{g}, \underline{z}) = \sum_{i:g_i=1} R_i = \#\{[h, k] \mid Z_{hk} > \mu_0, h \leq k\}$$

Test is least significant if $= \frac{n(n+1)}{2}$

$$\longrightarrow \hat{\mu} = \text{median}\langle Z_{hk} \mid h \leq k \rangle.$$

f **Confidence intervall** for the signed rank test:

Limits of the acceptance interval of T :

$$c \text{ and } c' = n(n+1)/2 + 1 - c$$

Confidence limits = c th and c' th Walsh average.

Example Tranquilizer: $c = 6$, $c' = 40$,

Confidence interval [0.01, 0.786].

h For a general test statistic $T(\underline{G}, \underline{z}^*; \mu)$: Let

$$Q(\beta) = P(T(\underline{G}, \underline{z}^*; \mu) > T(\underline{g}^*, \underline{z}^*; \mu)) - \beta$$

Estimator = solution of $Q(\beta = 0.5) = 0$.

Confidence limits = solution of

$$Q(\beta = 0.025) = 0 \text{ and } Q(\beta = 0.975) = 0.$$

Not difficult!

3.6 More than 2 Samples

a Simple Analysis of Variance

Randomization = assignment of observations to groups.

Number of observations in each group is fixed.

Rank the y_i s among all observations $\longrightarrow R_i$

Average the ranks over groups

$$\bar{R}_h = \text{ave}_{g_i=h} R_i. \quad \mathcal{E}\langle \bar{R}_h \rangle = (n+1)/2.$$

Form weighted mean of squares of deviations

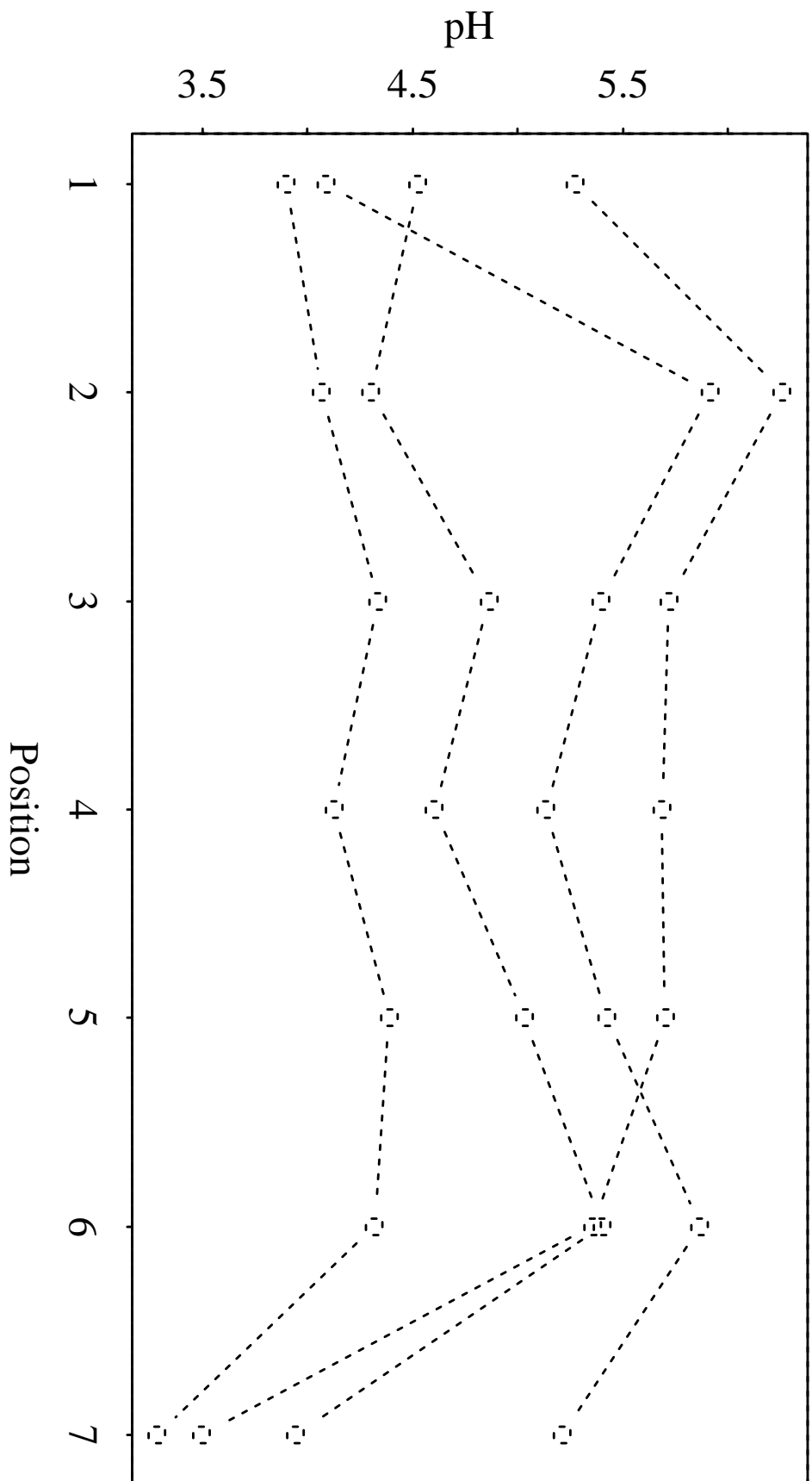
$$T\langle \underline{g}, \underline{y} \rangle = \frac{12}{n(n+1)} \sum_h n_h \left(\bar{R}_h - \frac{n+1}{2} \right)^2$$

Kruskal-Wallis test. 2 samples \longrightarrow U-Test.

- b **Matched Samples** = block design
 n blocks, m treatments.
 Randomization?

- c **Example acidic soils**

Block	Position						
	1	2	3	4	5	6	7
1	4.09	5.91	5.40	5.13	5.43	5.87	5.21
2	3.90	4.07	4.34	4.13	4.39	4.32	3.29
3	5.27	6.26	5.72	5.69	5.70	5.36	3.50
4	4.53	4.30	4.86	4.61	5.03	5.40	3.95



Friedman test. R_{ij} = Rank of observation j in block i .

$\tilde{R}_j = \text{ave}_i \langle R_{ij} \rangle$ average rank of sample j .

$$T = \frac{12n}{m(m+1)} \sum_{j=1}^m (\tilde{R}_j - (m+1)/2)^2.$$

d

	Position						
Block	1	2	3	4	5	6	7
1	1	7	4	2	5	6	3
2	2	3	6	4	7	5	1
3	2	7	6	4	5	3	1
4	3	2	5	4	6	7	1
Summe	8	19	21	14	23	21	6
Mittel	2	4.75	5.25	3.5	5.75	5.25	1.5

```
> friedman.test(t.dt)
      Friedman rank sum test
data:  t.dt
Friedman chi-squared = 14.8, df = 6,
p-value = 0.02199
```

- e Analysis of variance for this example:

```
> summary(aov(pH~trans+pos, data=t.d))
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
trans	1	0.12	0.12	0.17	0.68
pos	1	0.18	0.18	0.27	0.61
Residuals	25	16.57	0.66		

3.7 Correlation and Regression

a **Correlation and simple regression.**

X_i, Y_i (X_i random or fixed)

Null hypothesis: “no relationship”

Randomization = matching = permutation of \underline{Y} .

Probability of each permutation = $1/n! = 1/(n(n-1)\dots 2 \cdot 1)$.

Test statistic:

- simple (Pearson) correlation,
- rank correlation,
- robust estimator of the regression coefficient, ...

b **Multiple Regression:**

Permutation of \underline{Y} for testing the null hypothesis that there is no relationship between **all** explanatory variables and the target variable.

c **Time Series:** Are the observations independent?

Randomization: Permutation.

Test statistic: e.g. first autocorrelation.

d **Multiple Regression: Single coefficient (or several)**

→ no proper randomization model.

* e **Permutations and other randomizations.**

Regression and correlation: permutations.

Two or more samples: subsets (“choices”).

there are many more permutations;

many of them lead to the same partition into groups

→ same randomization distribution.

Hail experiment: Number of potential hail days was random, proportions of treated days also random

randomization distribution: Choices of 33 from 76 days

→ **conditional test**, given the number of treated and control days.

Messages

Randomization Tests

- **Randomization tests keep the level exactly,** without any assumptions on the distribution. (Independence of observations is essentially assumed.)
- The **test statistic may be arbitrarily complicated.** Choose considering (informally) the power. Choose robust test statistic (e.g., based on ranks)!
- Confidence intervals can also be constructed.