Inter-rater reliability

How to measure it reliably

Dr. Gordon McDonald

Markers

Student’s assignments

	Alice	Bob
Student 1	Pass	Fail
Student 2	Pass	Pass
Student 3	Fail	Pass
Student 4	Pass	Pass
⋮		

Markers

Student's assignments

	Alice	Bob
Student 1	Pass	Fail
Student 2	Pass	Pass
Student 3	Fail	Pass
Student 4	Pass	Pass
. . .		

Two raters - all of them rated every item

Ratings can be binary or categorical

% agreement

Cohen’s kappa

Fleiss’ kappa

Scott’s Pi

Krippendorff’s alpha

Convert to a contingency table of proportions

Alice's Ratings

Bob's ratings

	Fail = 0	Pass = 1	Bob's marginal
Fail = 0	a	b	p_b
Pass = 1	c	d	$q_b = 1 - p_b$
Alice's Marginal	p_a	$q_a = 1 - p_a$	

$$\textbf{Agreement} = 1 - \frac{\textbf{Observed Disagreement}}{\textbf{Expected Disagreement}}$$

Convert to a contingency table of proportions

Alice's Ratings

Bob's ratings

	Fail = 0	Pass = 1	Bob's marginal
Fail = 0	a	b	p_b
Pass = 1	c	d	$q_b = 1 - p_b$
Alice's Marginal	p_a	$q_a = 1 - p_a$	

fraction of agreement $= 1 - \dfrac{b + c}{1}$

(no measure of what you expect by chance)

Convert to a contingency table of proportions

Alice's Ratings

Bob's ratings

	Fail = 0	Pass = 1	Bob's marginal
Fail = 0	a	b	p_b
Pass = 1	c	d	$q_b = 1 - p_b$
Alice's Marginal	p_a	$q_a = 1 - p_a$	

Cohen's $\kappa = 1 - \dfrac{b + c}{p_a q_b + p_b q_a}$

Convert to a contingency table of proportions

Alice's Ratings

Bob's ratings

	Fail = 0	Pass = 1	Bob's marginal
Fail = 0	a	b	p_b
Pass = 1	c	d	$q_b = 1 - p_b$
Alice's Marginal	p_a	$q_a = 1 - p_a$	

Fraction of 0's

$$p_0 = \frac{p_a + p_b}{2} = \frac{a + b + a + c}{2}$$

Fraction of 1's

$$q_1 = \frac{q_a + q_b}{2} = \frac{c + d + b + d}{2} = 1 - p_0$$

Scott's $\pi = 1 - \dfrac{b + c}{2 \cdot p_0 \cdot q_1}$

Convert to a contingency table of proportions

Alice's Ratings

Bob's ratings

	Fail = 0	Pass = 1	Bob's marginal
Fail = 0	a	b	p_b
Pass = 1	c	d	$q_b = 1 - p_b$
Alice's Marginal	p_a	$q_a = 1 - p_a$	

Fraction of 0's

$$p_0 = \frac{p_a + p_b}{2} = \frac{a + b + a + c}{2}$$

Fraction of 1's

$$q_1 = \frac{q_a + q_b}{2} = \frac{c + d + b + d}{2} = 1 - p_0$$

Krippendorff's $\alpha = 1 - \frac{b + c}{2 \cdot p_0 \cdot q_1} \cdot \frac{n - 1}{n}$

Markers

Student’s assignments

	Alice	Bob
Student 1	Pass	Fail
Student 2	Pass	Pass
Student 3	Fail	Pass
Student 4	Pass	Pass
. . .		

Markers

Student’s assignments

	Alice	Bob
Student 1	Pass	Fail
Student 2	N/A	Pass
Student 3	Fail	Pass
Student 4	Pass	Pass
⋮		

Markers

Student’s assignments

	Alice	Bob	Cathy
Student 1	Pass	Fail	Pass
Student 2	N/A	Pass	Pass
Student 3	Fail	Pass	Fail
Student 4	Pass	Pass	N/A
⋮			

Multiple raters - not all of them rated every item

Ratings can be binary, numeric, ordinal, interval, circular…

Krippendorff's alpha!

(works for the simple cases and simplifies to each of the other indices in the appropriate limits)

Markers

Student's assignments

	Alice	Bob	Cathy
Student 1	Pass	Fail	Pass
Student 2	N/A	Pass	Pass
Student 3	Fail	Pass	Fail
Student 4	Pass	Pass	N/A
⋮			

So how to calculate in general?

$$\alpha = 1 - \frac{\textbf{Observed Disagreement between raters within units}}{\textbf{Expected Disagreement between raters within units}}$$

$$= 1 - \frac{D_o}{D_e}$$

Distance between this pair of marks

$$D_o = \frac{1}{n} \sum_{\textbf{assignments}} \sum_{\textbf{all pairs of marks}} \delta \cdot m \cdot p$$

$\delta =$ **Some appropriate distance metric between pairs of ratings**

So how to calculate in general?

$$\alpha = 1 - \frac{\textbf{Observed Disagreement between raters within units}}{\textbf{Expected Disagreement between raters within units}}$$

$$= 1 - \frac{D_o}{D_e}$$

Distance between this pair of marks

No. of markers for this assignment

$$D_o = \frac{1}{n} \sum_{\textbf{assignments}} \sum_{\textbf{all pairs of marks}} \delta \cdot m \cdot p$$

Something to do with permutations

$D_e =$ **Same thing averaged over how you expect it to come out randomly…**

…whatever, just use the R package *irr*

What does it mean?

$$\alpha = 1 - \frac{\textbf{Observed Disagreement between raters within units}}{\textbf{Expected Disagreement between raters within units}}$$

$$= 1 - \frac{D_o}{D_e}$$

They all disagree on purpose

$\alpha < 0$

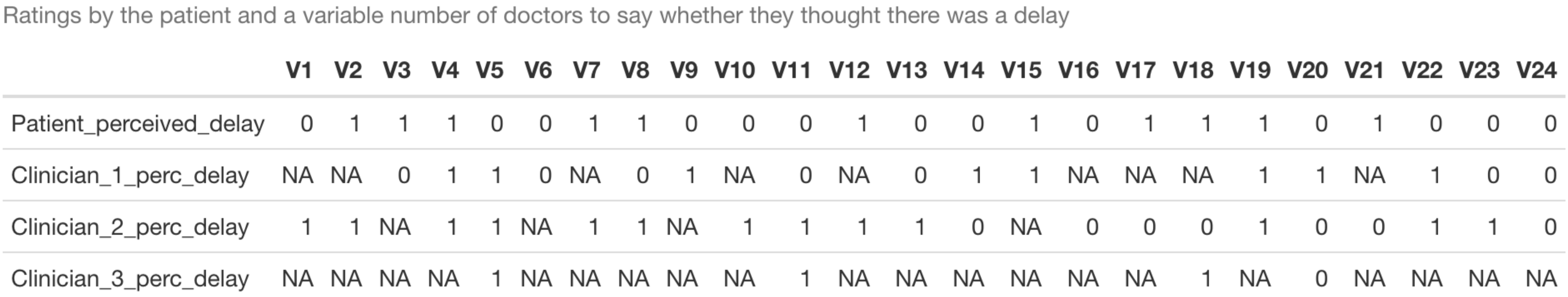
Everybody's guessing and it's all random

$\alpha = 0$

It all perfectly agrees

$\alpha = 1$

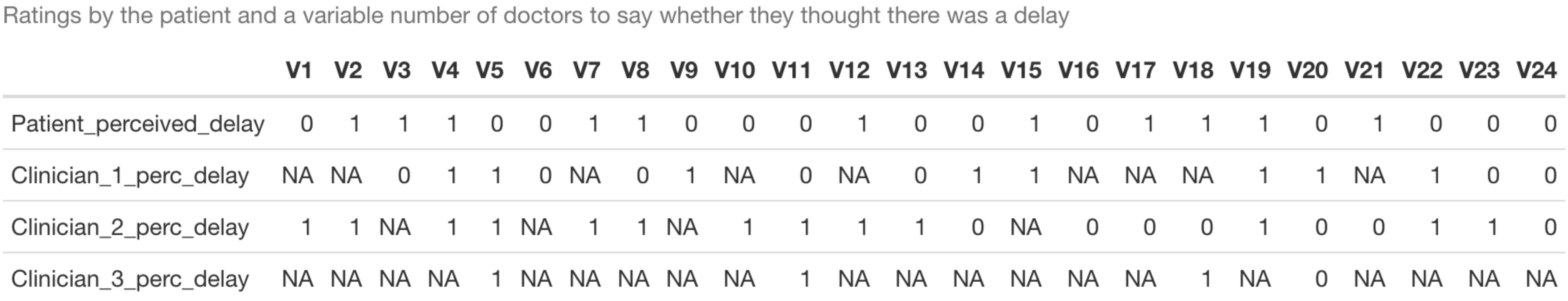
Example - doctors and patients saying whether there was a delay



Ratings by the patient and a variable number of doctors to say whether they thought there was a delay

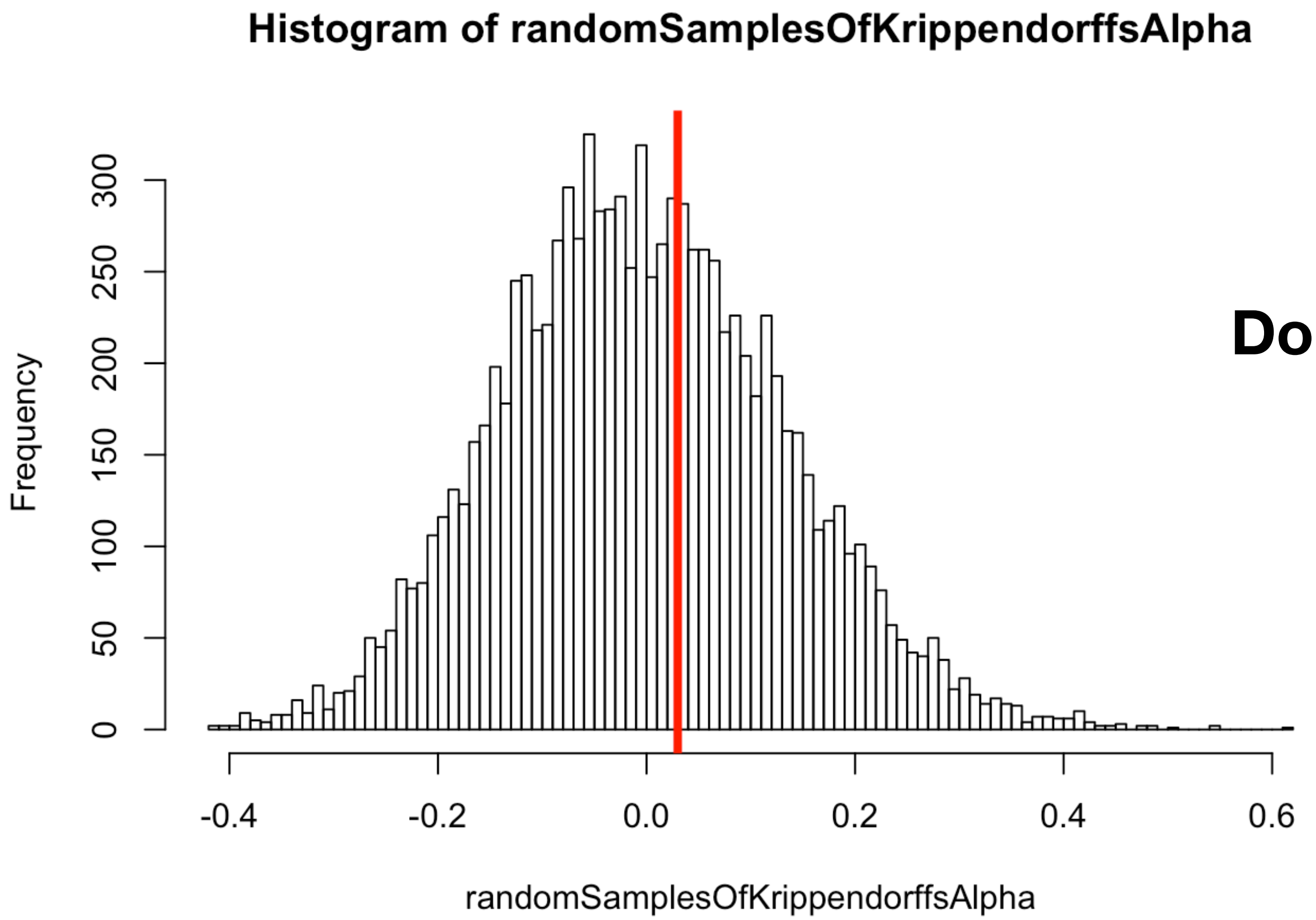
	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14	V15	V16	V17	V18	V19	V20	V21	V22	V23	V24
Patient_perceived_delay	0	1	1	1	0	0	1	1	0	0	0	1	0	0	1	0	1	1	1	0	1	0	0	0
Clinician_1_perc_delay	NA	NA	0	1	1	0	NA	0	1	NA	0	NA	0	1	1	NA	NA	NA	1	1	NA	1	0	0
Clinician_2_perc_delay	1	1	NA	1	1	NA	1	1	NA	1	1	1	1	0	NA	0	0	0	1	0	0	1	1	0
Clinician_3_perc_delay	NA	NA	NA	NA	1	NA	NA	NA	NA	NA	1	NA	NA	NA	NA	NA	NA	1	NA	0	NA	NA	NA	NA

Example - doctors and patients saying whether there was a delay



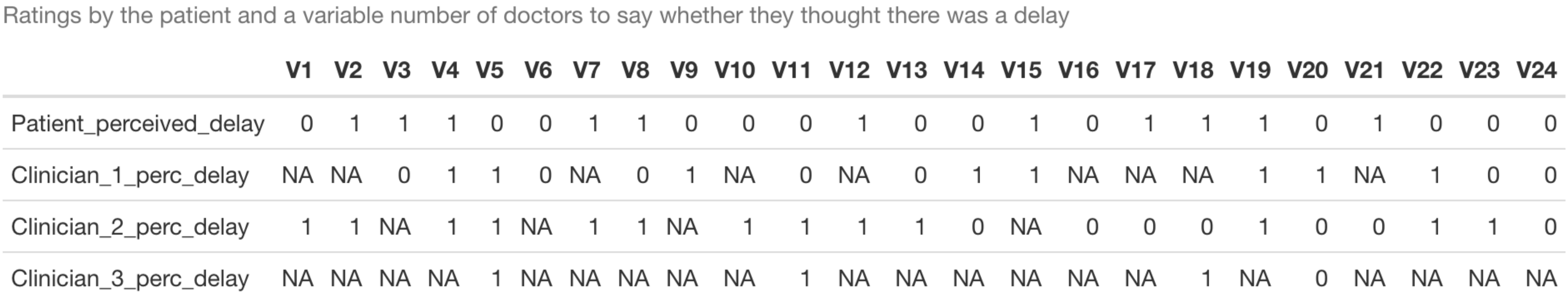
Ratings by the patient and a variable number of doctors to say whether they thought there was a delay

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14	V15	V16	V17	V18	V19	V20	V21	V22	V23	V24
Patient_perceived_delay	0	1	1	1	0	0	1	1	0	0	0	1	0	0	1	0	1	1	1	0	1	0	0	0
Clinician_1_perc_delay	NA	NA	0	1	1	0	NA	0	1	NA	0	NA	0	1	1	NA	NA	NA	1	1	NA	1	0	0
Clinician_2_perc_delay	1	1	NA	1	1	NA	1	1	NA	1	1	1	1	0	NA	0	0	0	1	0	0	1	1	0
Clinician_3_perc_delay	NA	NA	NA	NA	1	NA	NA	NA	NA	NA	1	NA	NA	NA	NA	NA	NA	1	NA	0	NA	NA	NA	NA



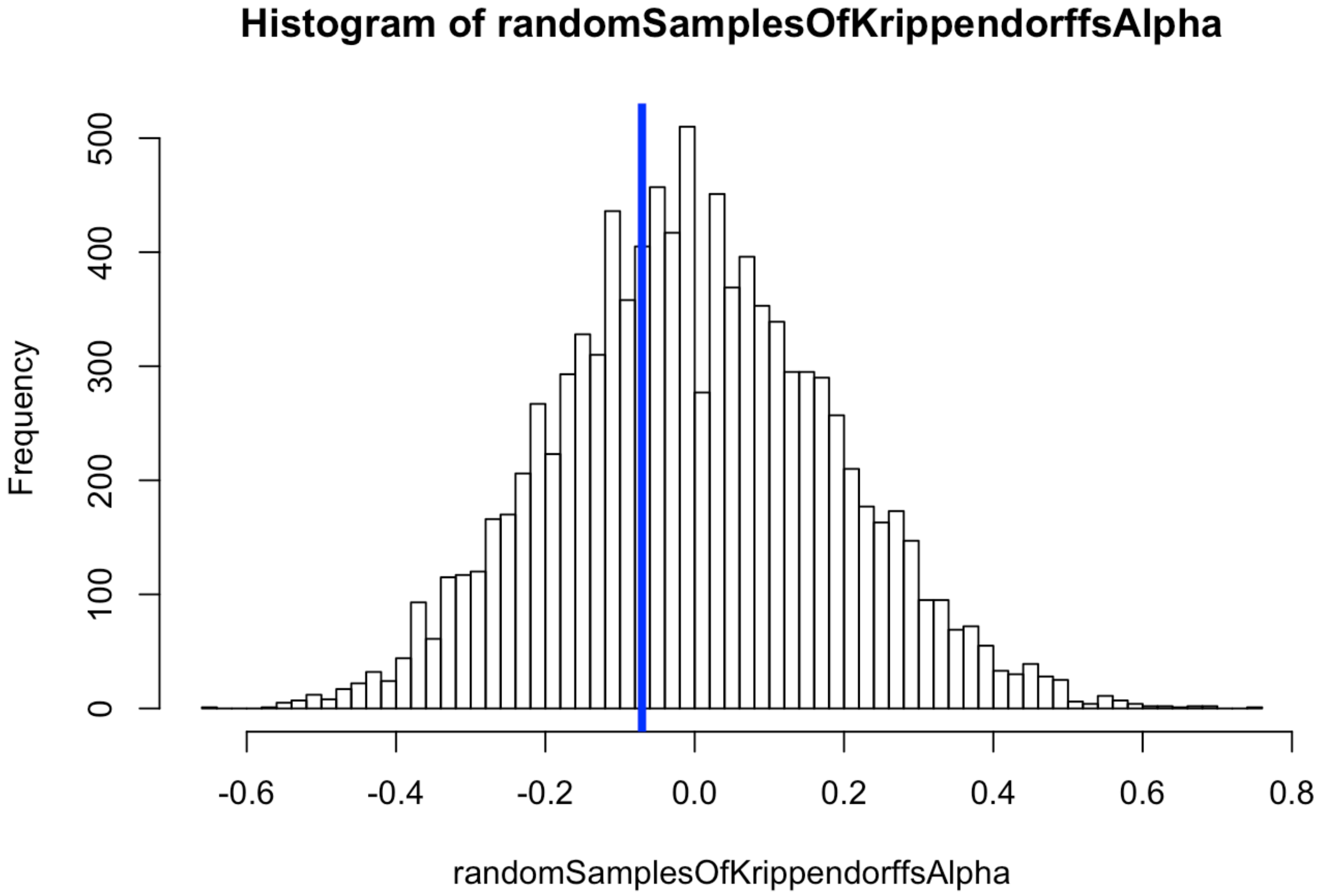
Doctors and patients

Example - doctors and patients saying whether there was a delay



Ratings by the patient and a variable number of doctors to say whether they thought there was a delay

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14	V15	V16	V17	V18	V19	V20	V21	V22	V23	V24
Patient_perceived_delay	0	1	1	1	0	0	1	1	0	0	0	1	0	0	1	0	1	1	1	0	1	0	0	0
Clinician_1_perc_delay	NA	NA	0	1	1	0	NA	0	1	NA	0	NA	0	1	1	NA	NA	NA	1	1	NA	1	0	0
Clinician_2_perc_delay	1	1	NA	1	1	NA	1	1	NA	1	1	1	1	0	NA	0	0	0	1	0	0	1	1	0
Clinician_3_perc_delay	NA	NA	NA	NA	1	NA	NA	NA	NA	NA	1	NA	NA	NA	NA	NA	NA	1	NA	0	NA	NA	NA	NA

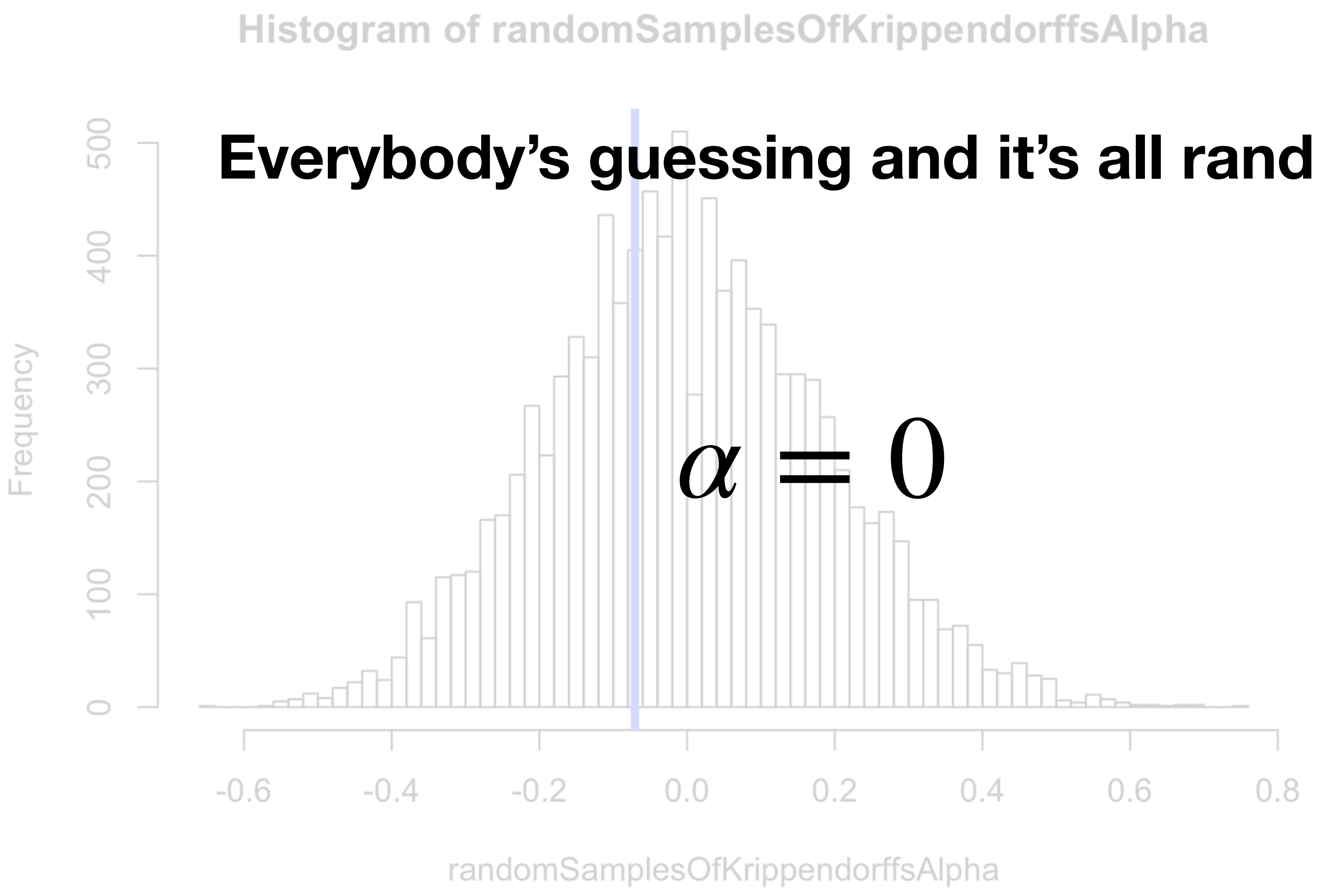


Just doctors

Example - doctors and patients saying whether there was a delay

Ratings by the patient and a variable number of doctors to s [illegible] the [illegible] ere was a delay

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14	V15	V16	V17	V18	V19	V20	V21	V22	V23	V24
Patient_perceived_delay	0	1	1	1	0	0	1	1	[illegible]	[illegible]	[illegible]	[illegible]	0	0	1	0	1	1	1	0	1	0	0	0
Clinician_1_perc_delay	NA	NA	0	1	1	0	NA	0	[illegible]	NA	0	NA	0	1	1	NA	NA	NA	1	1	NA	1	0	0
Clinician_2_perc_delay	1	1	NA	1	1	NA	1	1	NA	1	1	1	1	0	NA	0	0	0	1	0	0	1	1	0
Clinician_3_perc_delay	NA	NA	NA	NA	1	NA	NA	NA	NA	NA	1	NA	NA	NA	NA	NA	NA	1	NA	0	NA	NA	NA	NA



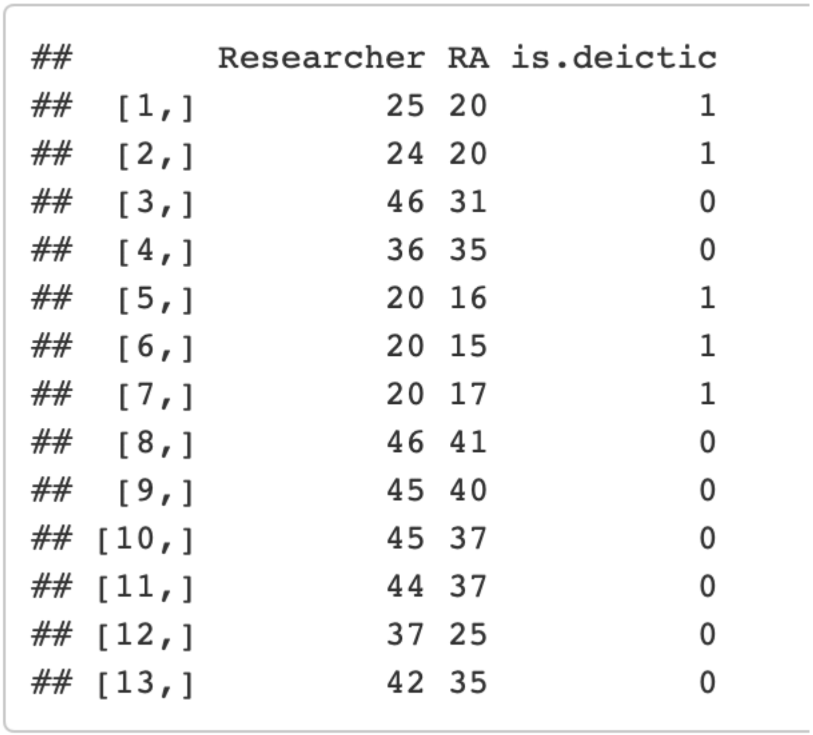
Everybody's guessing and it's all random

$$\alpha = 0$$

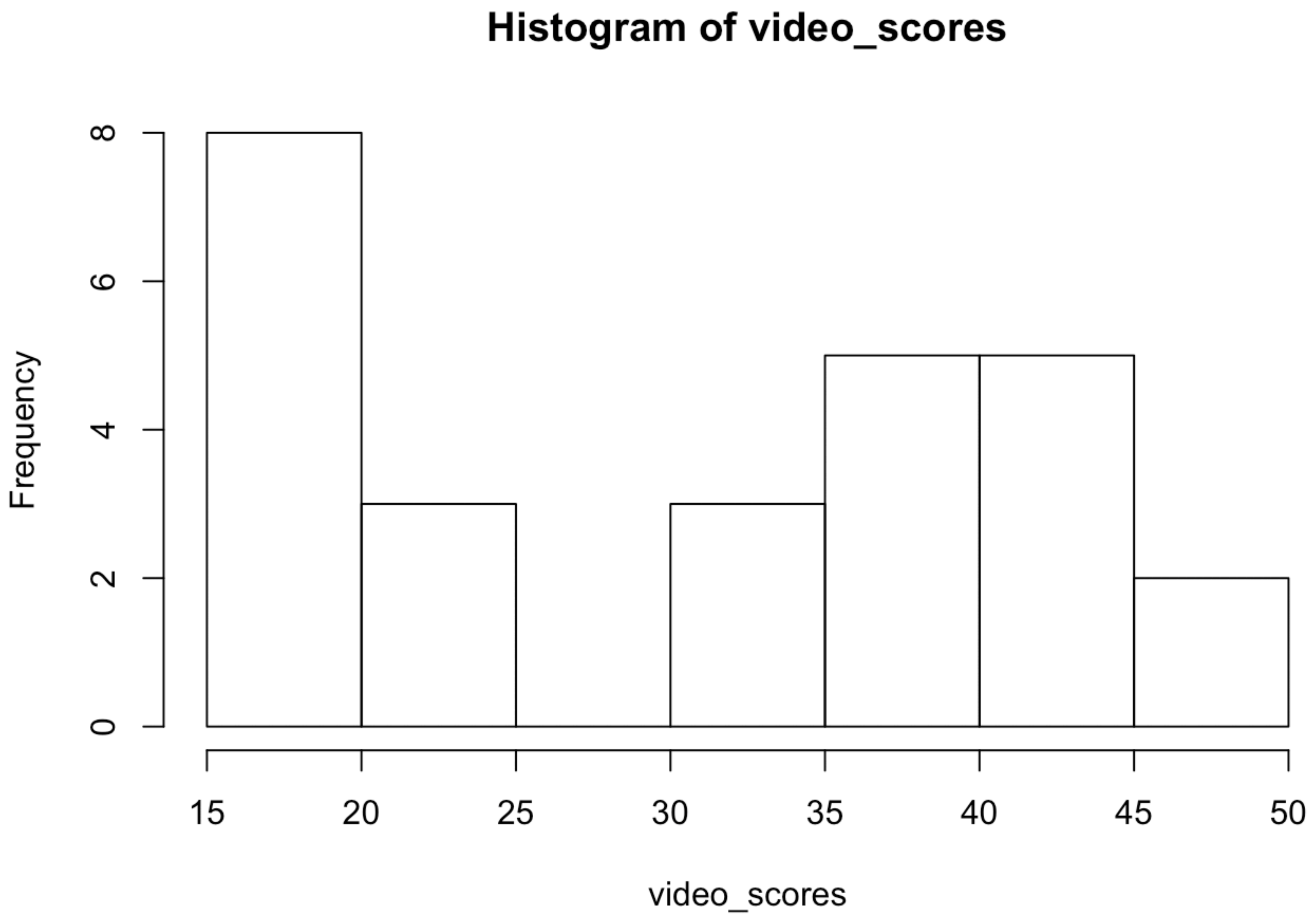
Just doctors

p≈ 0.5

Example - different people counting behaviours in a video

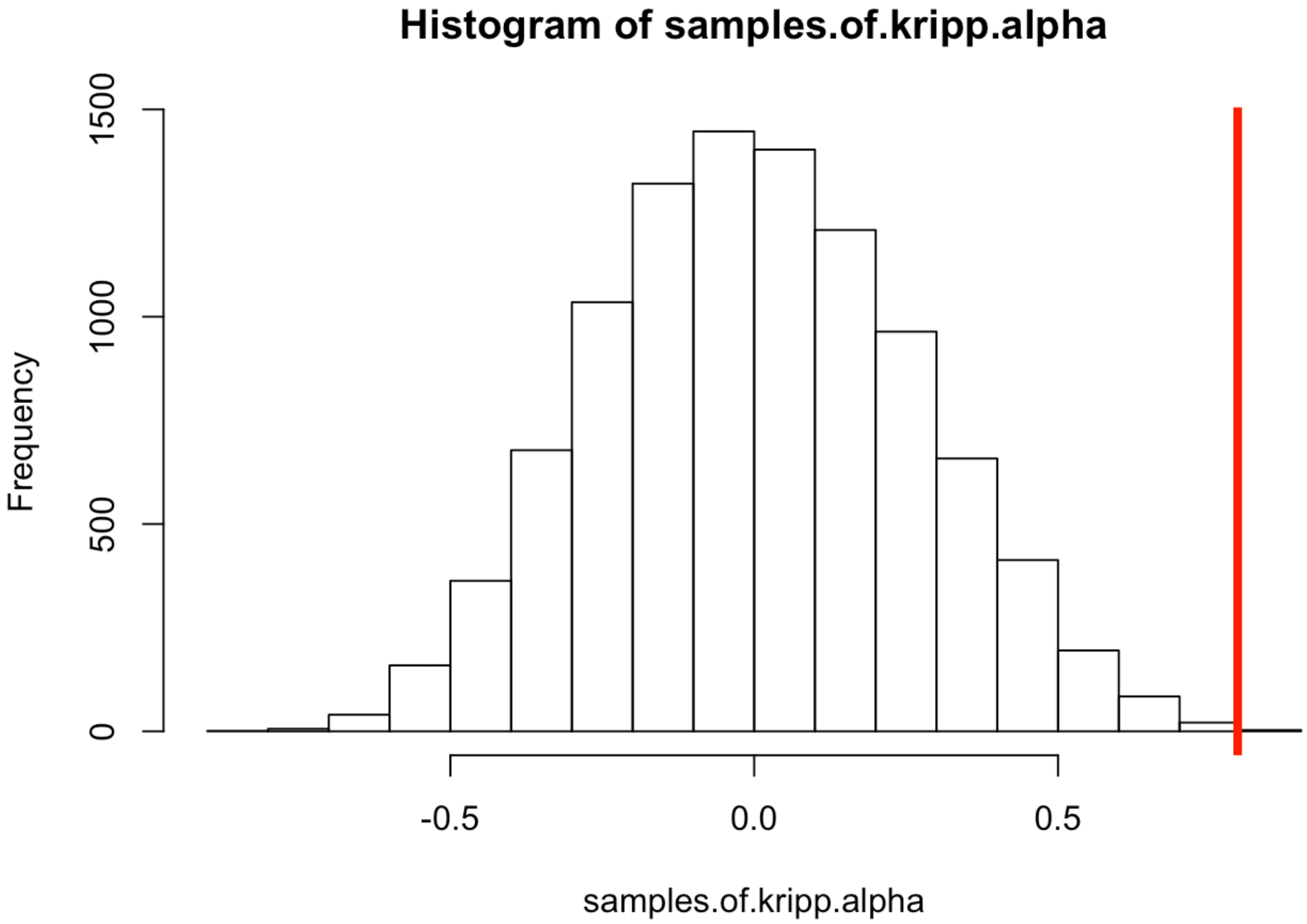


```
##       Researcher RA is.deictic
##  [1,]         25 20          1
##  [2,]         24 20          1
##  [3,]         46 31          0
##  [4,]         36 35          0
##  [5,]         20 16          1
##  [6,]         20 15          1
##  [7,]         20 17          1
##  [8,]         46 41          0
##  [9,]         45 40          0
## [10,]         45 37          0
## [11,]         44 37          0
## [12,]         37 25          0
## [13,]         42 35          0
```



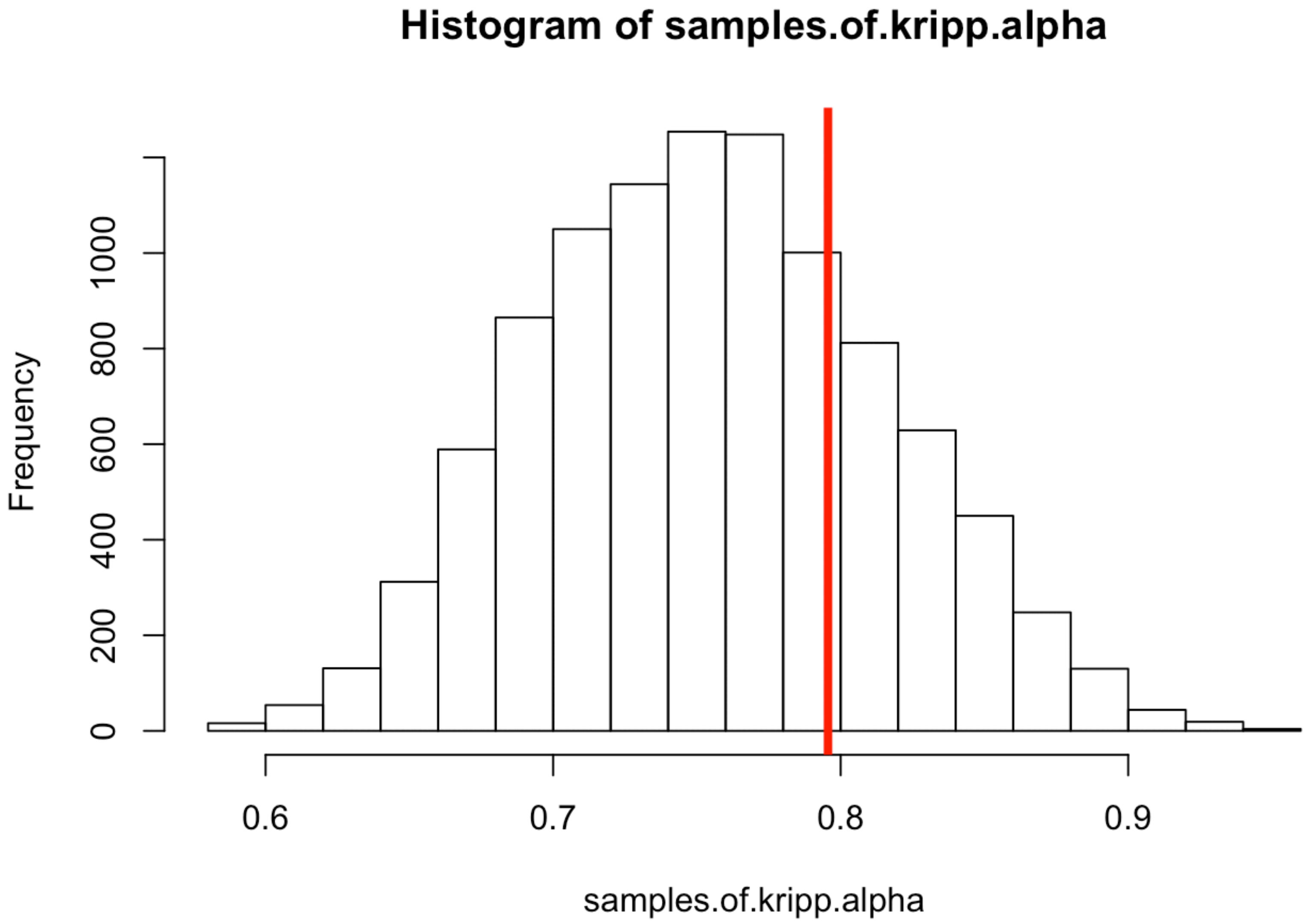
Example - different people counting behaviours in a video

Naive resampling



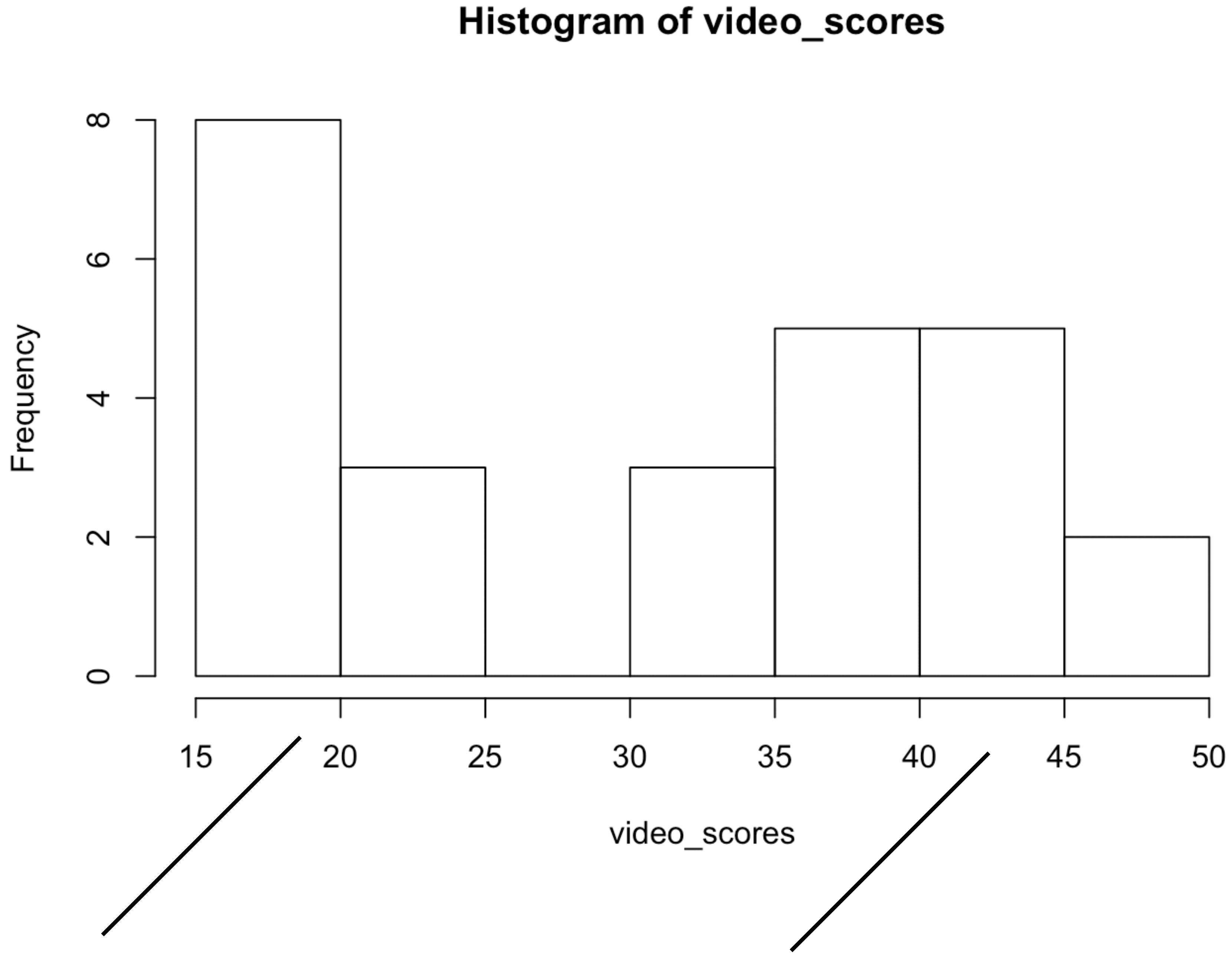
Example - different people counting behaviours in a video

Less naive resampling



Example - different people counting behaviours in a video

Sampling from ideal distributions

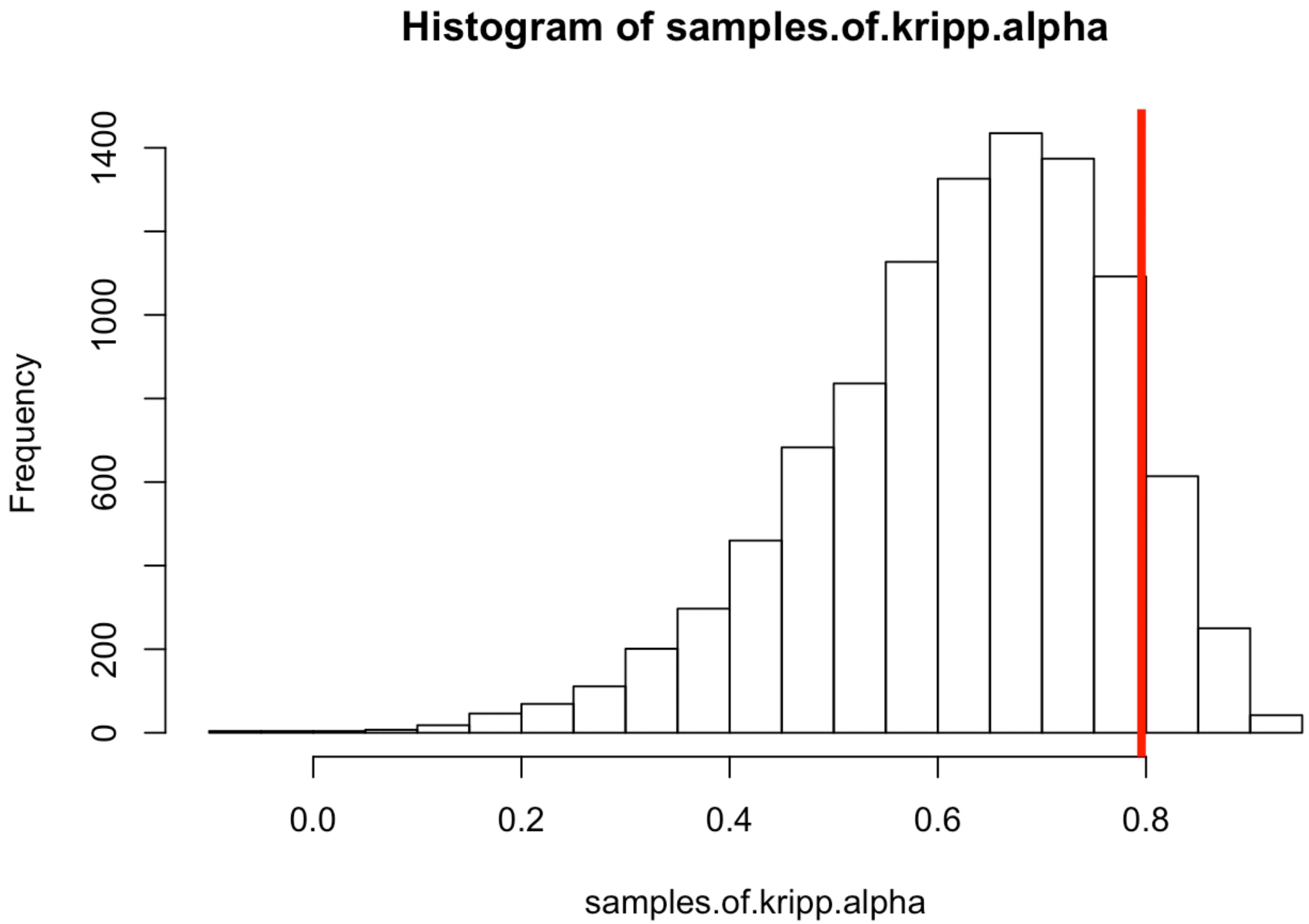


Looks binomial

Looks poisson

Example - different people counting behaviours in a video

Sampling from ideal distributions

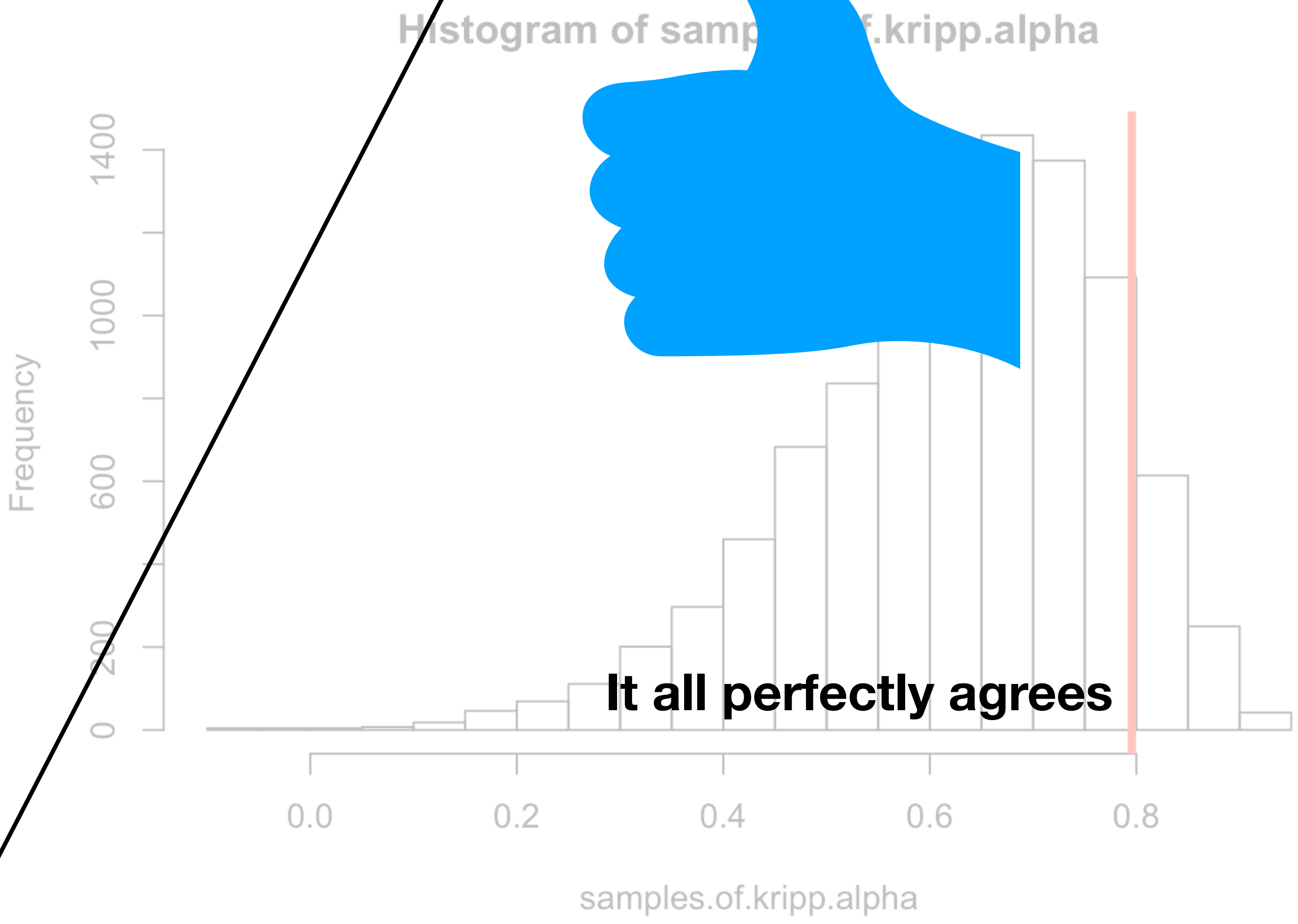


Example - different people counting behaviours in a video

Sampling from ideal distributions

Everybody's guessing and it's all random

$\alpha = 0$



It all perfectly agrees

$$\alpha = 1$$

p=0.09

So in summary, use Krippendorff’s alpha

Pick the right distance metric

Choose an appropriate null hypothesis

Bootstrap if you have enough data to work out confidence intervals or p-values

Otherwise use ideal distributions to sample from

Some references…

https://en.wikipedia.org/wiki/Krippendorff's_alpha

Reliability in Content Analysis: Some Common Misconceptions and Recommendations.

-K. Krippendorff, 2004 University of Pennsylvania Departmental papers,

https://repository.upenn.edu/cgi/viewcontent.cgi?article=1250&context=asc_papers

https://cran.r-project.org/web/packages/irr/index.html