The geometry of diet: using projections to quantify the similarity between sets of dietary patterns

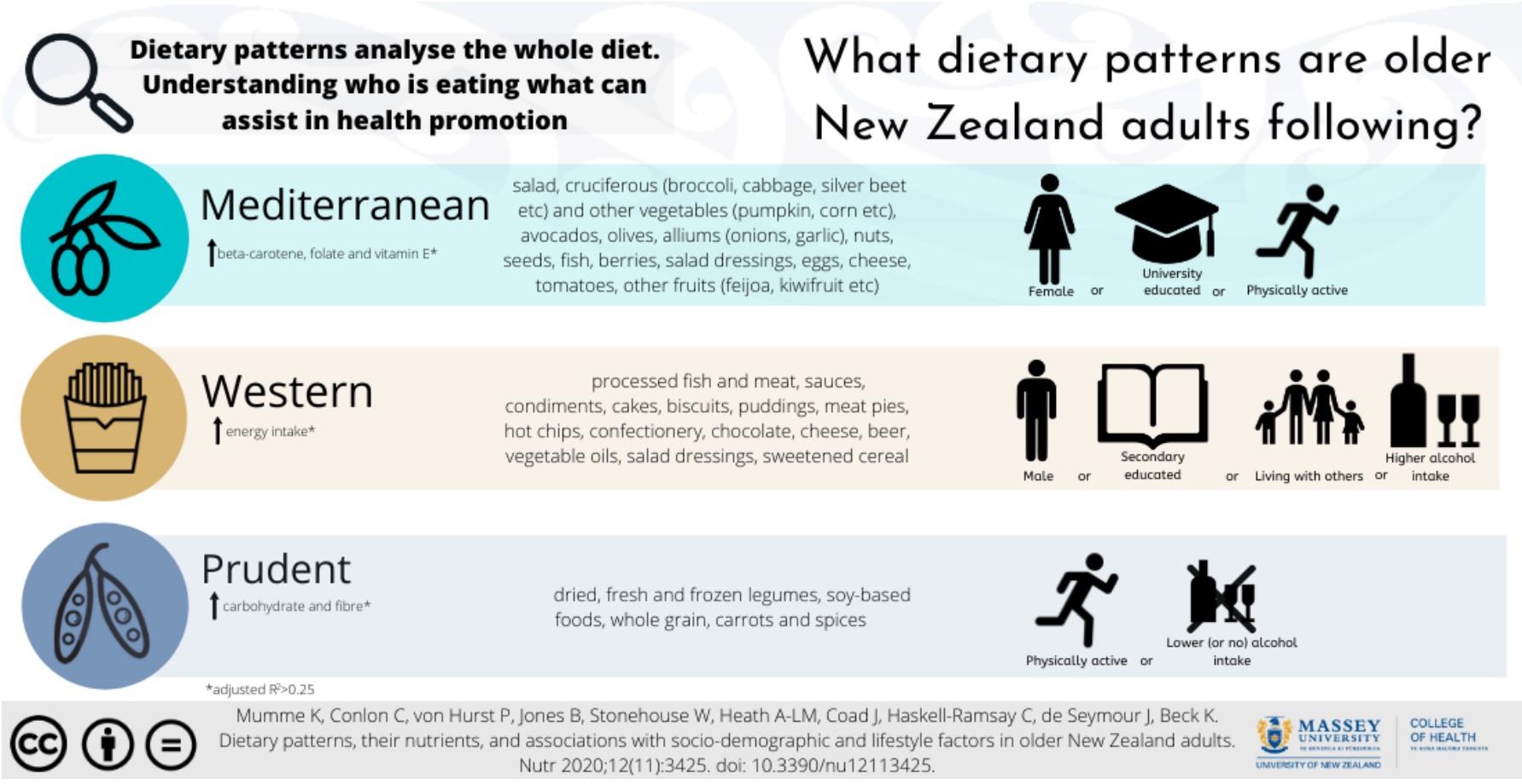
Beatrix Jones*, joint work with Karen Mumme, Shiyu Fan and John Thompson

27 November 2025

Creation of Empirical Dietary Patterns (Typically)

- Food Frequency questionnaire or food diary/recall followed by categorization into food groups.
- PCA/EFA dimension selection. Work with correlation matrix.
- Usually 2-5 components
- Varimax rotation of loadings on the correlation scale
- Use correlations with original food groups to interpret the axes.

Example:REACH study n=371, p=56



Example: REACH study comparisons

"Dietary pattern and education associations . . . found in the current study are consistent with other studies in older adults. 'Mixed', 'fat and meat', 'Western', and 'traditional' dietary patterns have frequently been associated with a lower education [10,12,14,16], while dietary patterns comprising more healthy food groups, such as 'vegetable based', 'fruit and milk', 'plant-based', or 'healthy', are frequently associated with a higher education [9,10,11,14,15,17], although some exceptions have been reported."

Can we be more quantitative than this? Yes. Caveat–we need the other study to use the same variables.

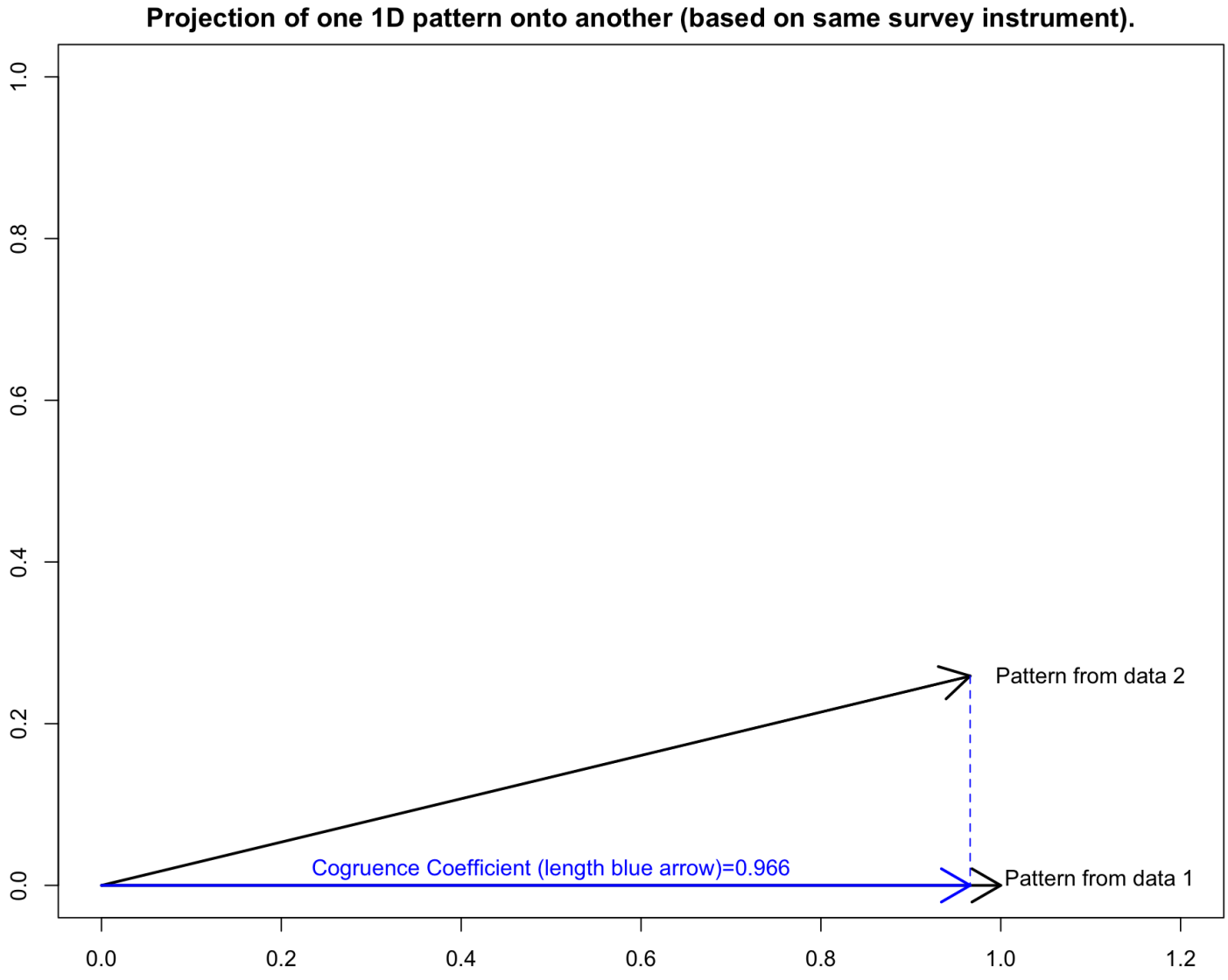
Comparing individual dietary patterns

$$\varphi(\mathbf{x}, \mathbf{y}) = \frac{\sum_{i=1}^{p} x_i y_i}{\sqrt{\sum_{i=1}^{p} x_i^2 \sum_{i=1}^{p} y_i^2}}$$

"Cosine similarity","Tucker's congruence coefficient," "congruence coefficient"

Like the correlation without centering... if we shift all values of $\mathbf{x}$ by a constant, the coefficient changes.

Cosine similarity: Geometry.



Cosine of the angle between the directions.

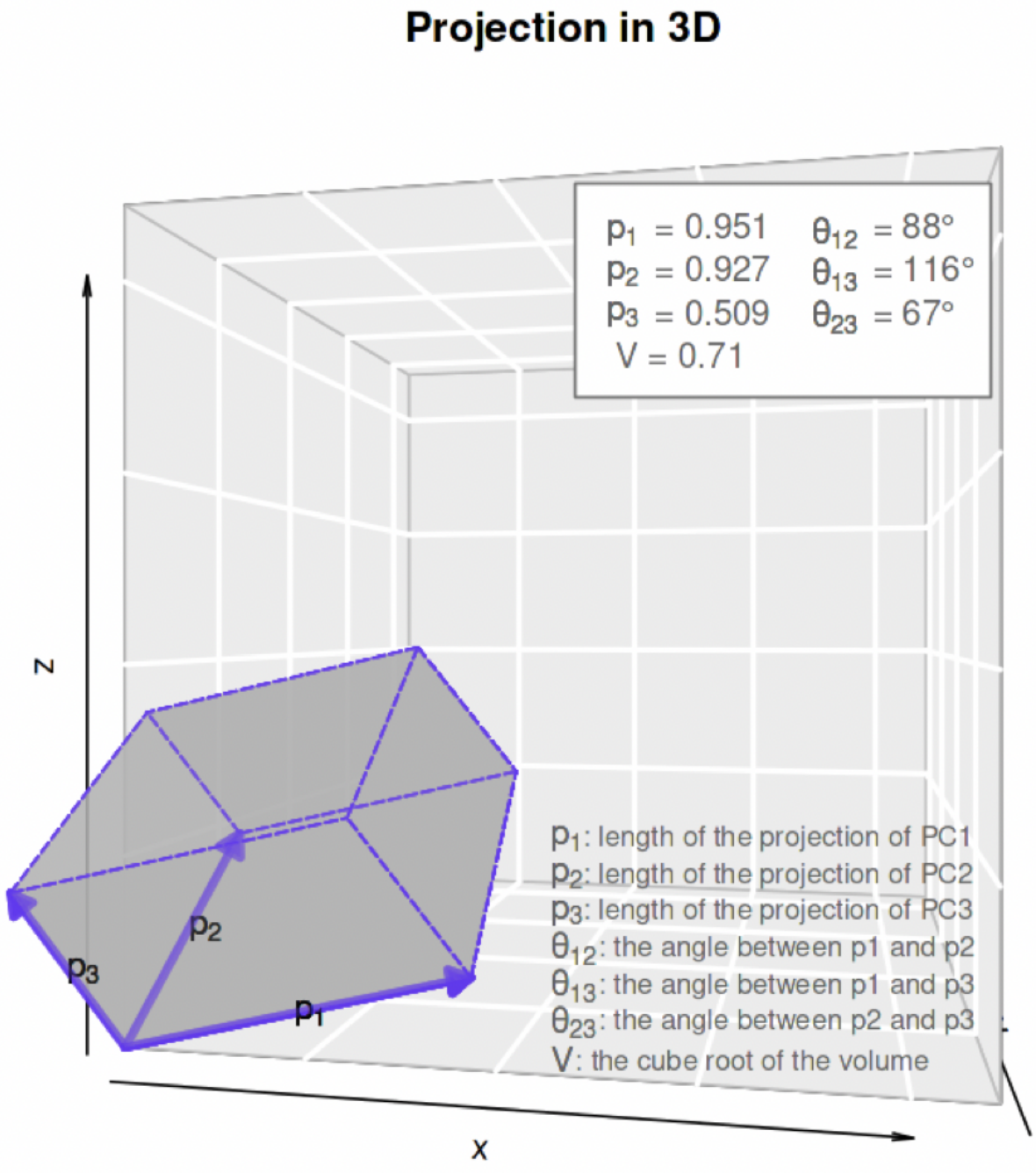
Multivariate Cosine Similarity

We have two, 3-D dietary pattern spaces (eg one from men, one from women)

1. Project a unit cube from one coordinate system onto the other coordinate system.
2. How much did it shrink?
3. Take the cube root to make it more comparable to the 1-D quantities.

Can extend to any number of dimensions

Geometry


Projection in 3D
p1 = 0.951
p2 = 0.927
p3 = 0.509
V = 0.71
θ12 = 88°
θ13 = 116°
θ23 = 67°
z
x
p1
p2
p3
p1: length of the projection of PC1
p2: length of the projection of PC2
p3: length of the projection of PC3
θ12: the angle between p1 and p2
θ13: the angle between p1 and p3
θ23: the angle between p2 and p3
V: the cube root of the volume

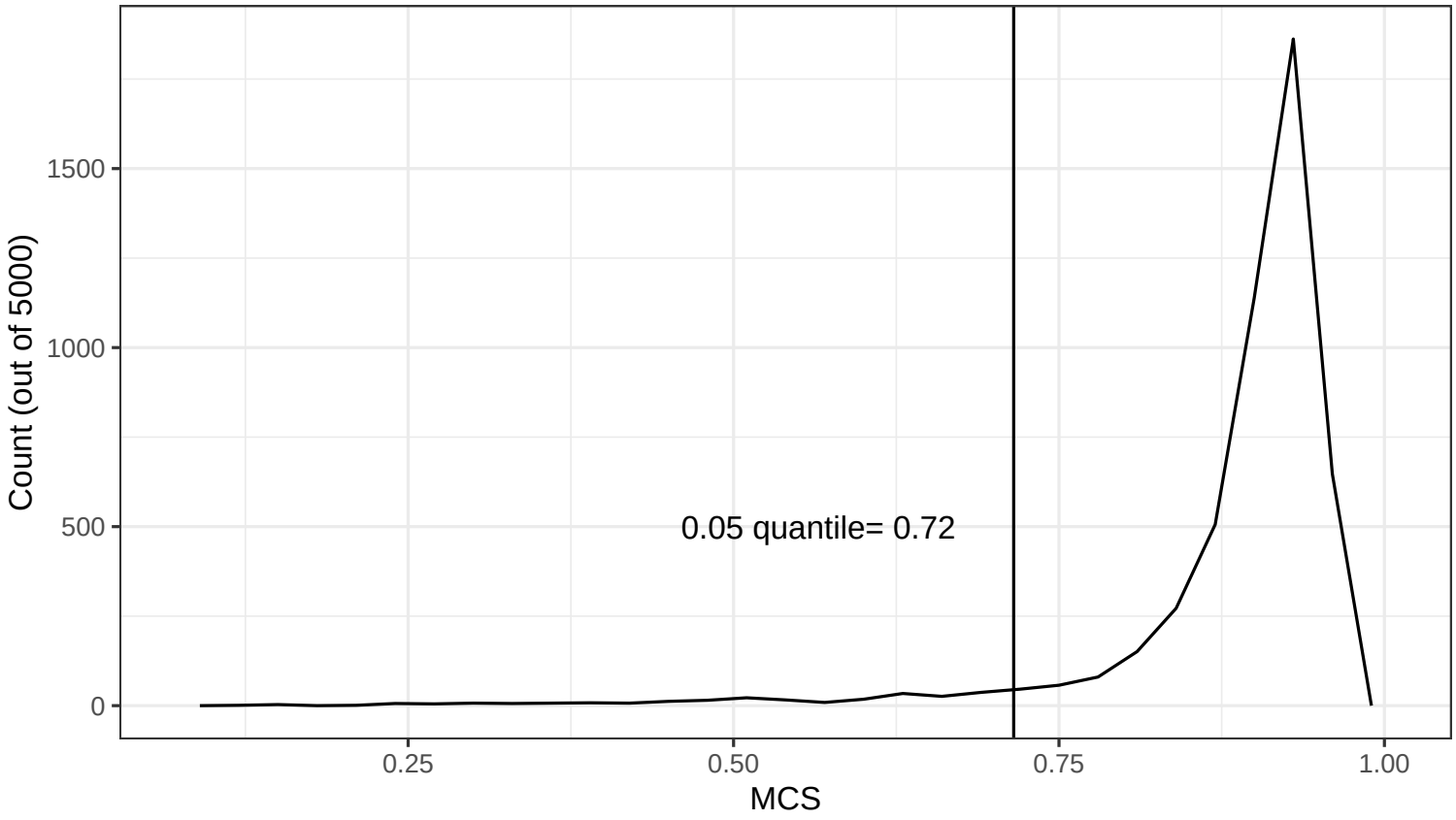
Volume of the projection

Let $\boldsymbol{\beta_{1,1:k}}$ be the vectors of coefficients of the PC's/factors for the population, $\boldsymbol{\beta_{2,1:k}}$ the vectors of coefficients of the PC's/factors for the sample. Both orthonormal sets of vectors.

$$MCS = \sqrt[k]{|\det\left(\boldsymbol{\beta_{1,1:k}}^T \boldsymbol{\beta_{2,1:k}}\right)|}$$

- Invariant to rotation/reflection
- Need to start with an orthonormal basis (for EFA can find with Gram-Schmidt, if necessary)

Bootstrap interval for MCS between first 3 sample and population PCs



But what about. . .

Average congruence after procrustes rotation ("Pro") $\boldsymbol{\beta^*_{1.}}$ = rotated version of $\boldsymbol{\beta_{1,1:k}}$ to best match $\boldsymbol{\beta_{2,1:k}}$

$$\sum_{i=1}^{k} (\boldsymbol{\beta^*_{1i}}^\top \boldsymbol{\beta_{2i}})^2 / k$$

Krzanowski Index, normalized to have maximum 1 ("Krz")

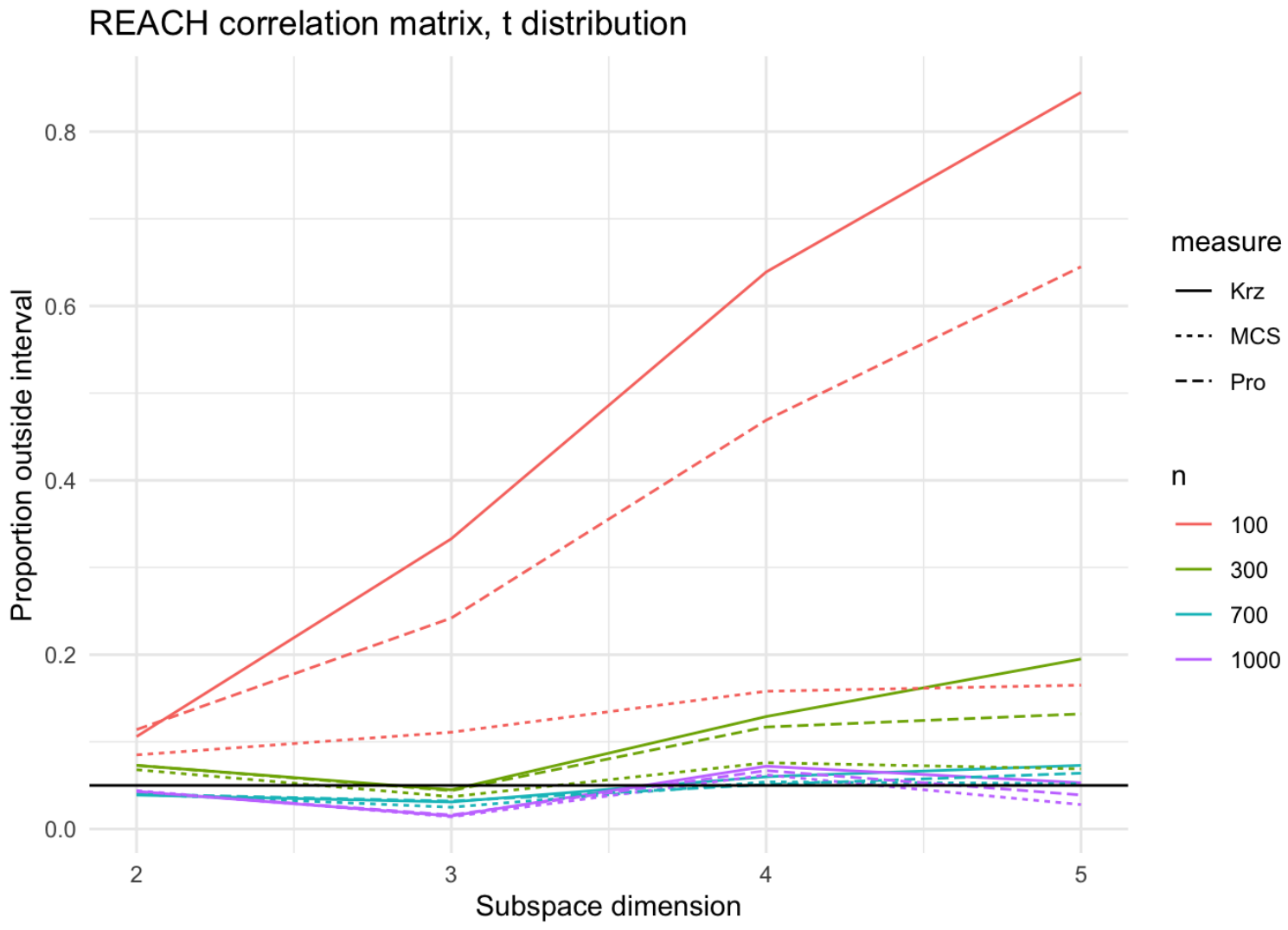
$$\frac{1}{k} \sum_{i=1}^{k} \sum_{j=1}^{k} (\boldsymbol{\beta_{1i}}^\top \boldsymbol{\beta_{2j}})^2$$

Could be used in the same way. . . and how do intervals perform relative to nominal coverage?

Simulation

- Use correlation matrix from REACH as population value.
- Simulate samples using this matrix and the t-5 distribution.
- Sample n =100, 300, 700, 1000
- Look at subspaces of dimension 2, 3, 4, 5
- Nominally 95% intervals; for 1000 replications, in what proportion does the constructed interval exclude the true MCS between population and sample?

Results

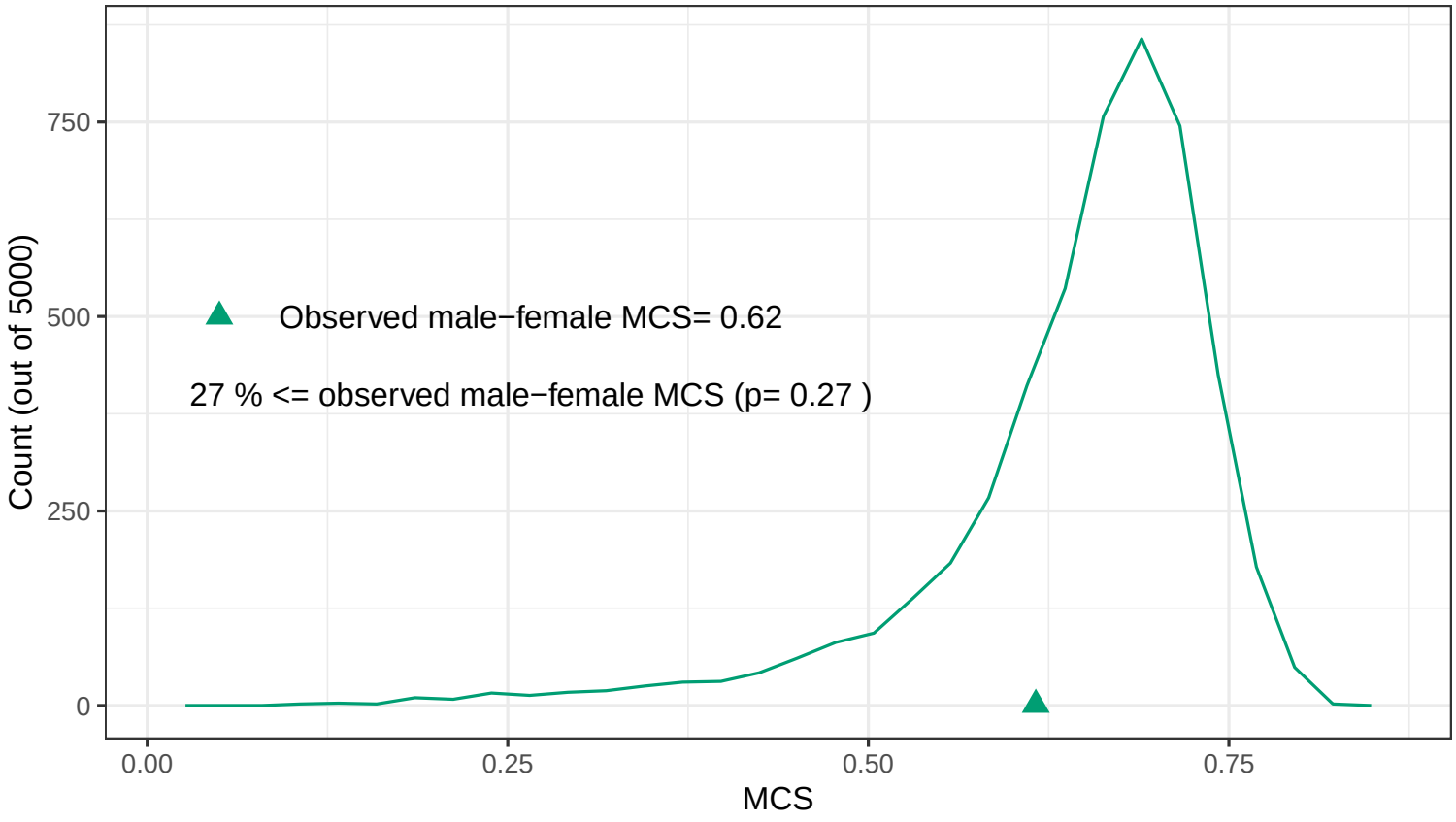

REACH correlation matrix, t distribution
Proportion outside interval
0.0
0.2
0.4
0.6
0.8
2
3
4
5
Subspace dimension
measure
Krz
MCS
Pro
n
100
300
700
1000

Things to note:

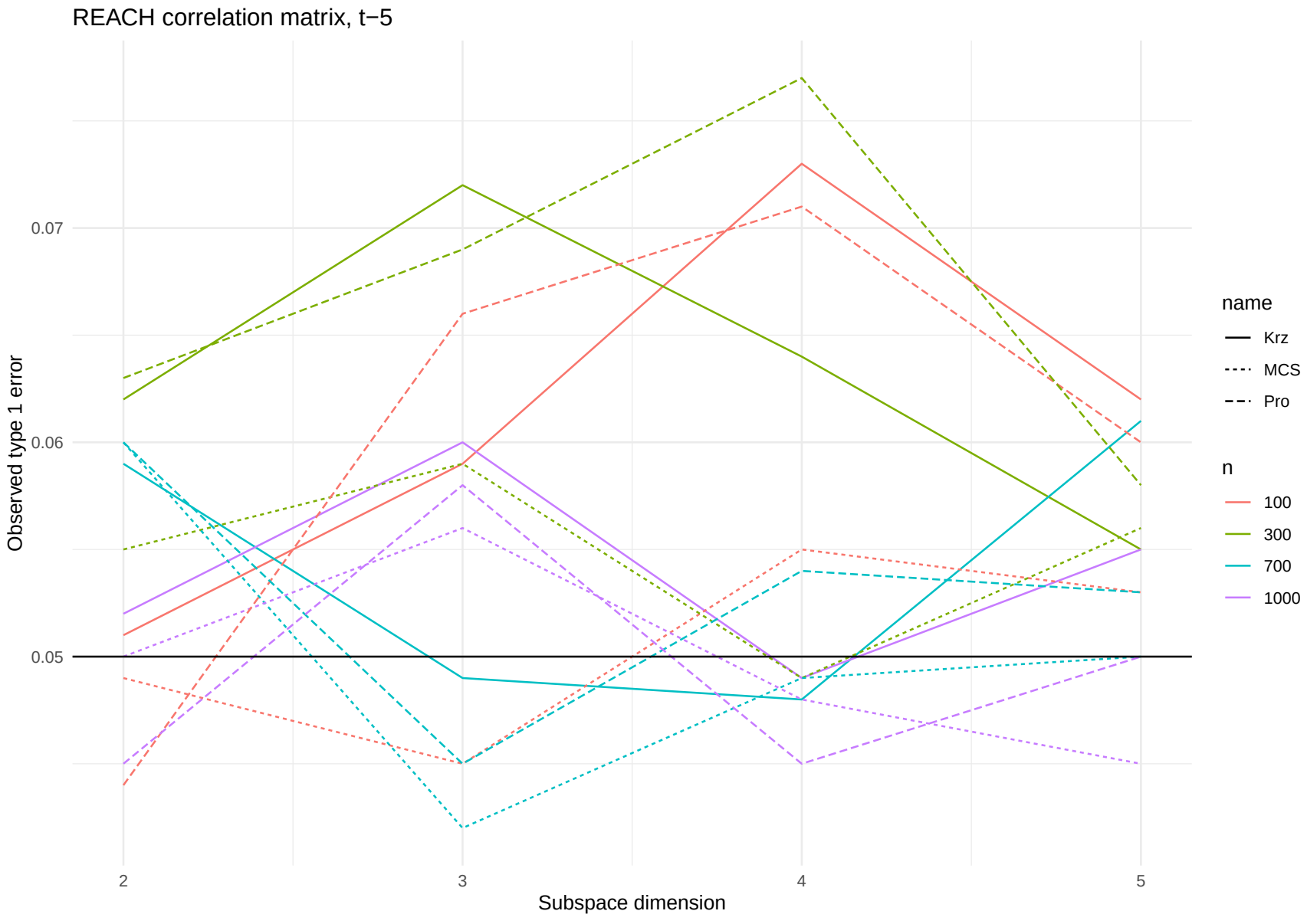
- MCS much better at higher subspace dimension and small-to-moderate sample size.
- Based on other examples, issue not subspace dimension but how distinctive consecutive eigenvalues are.

Example: Permutation test for population differences

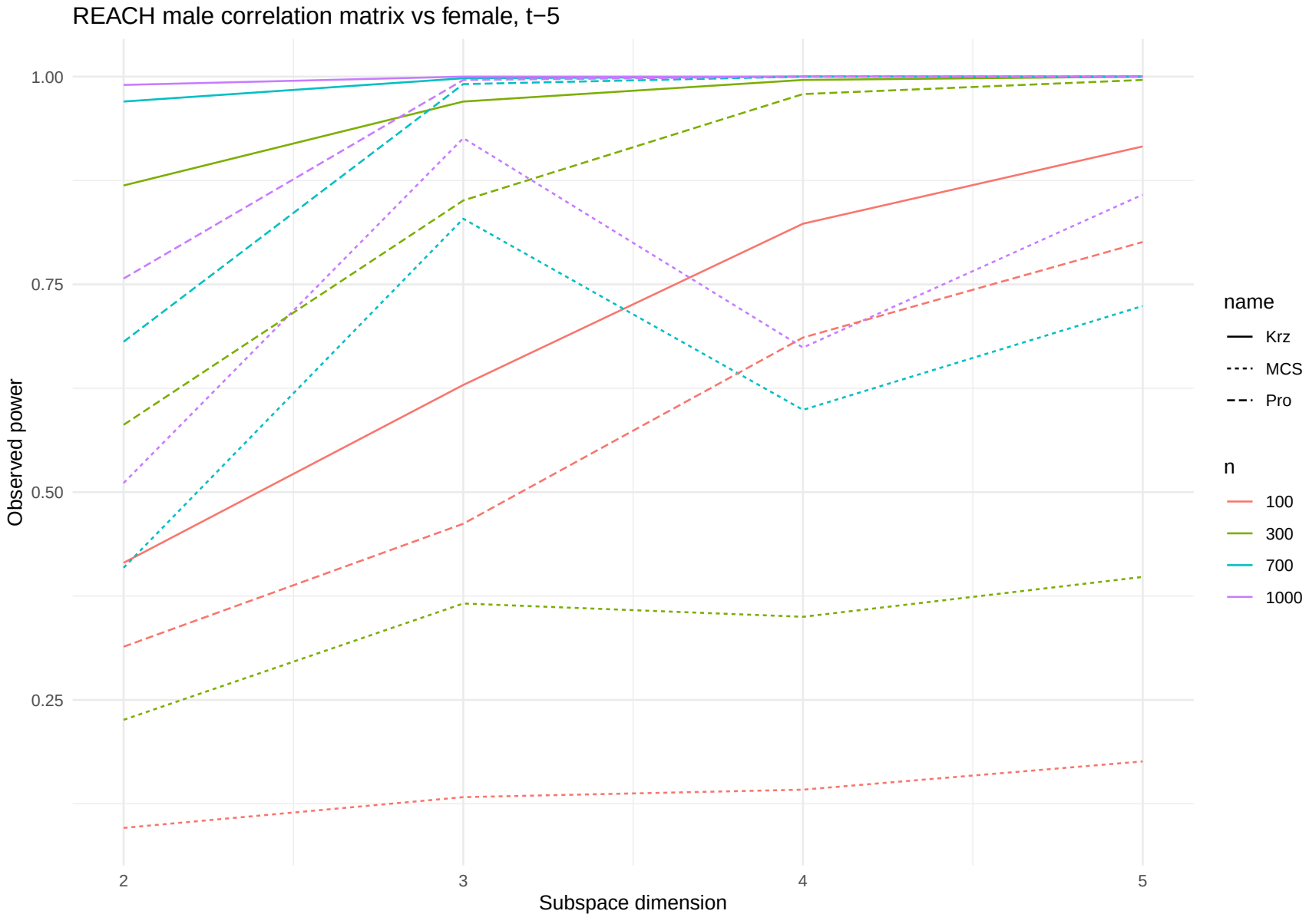
REACH– Compare 131 men to 234 women. Observed MCS is 0.62. Use as test statistic in a permutation test.



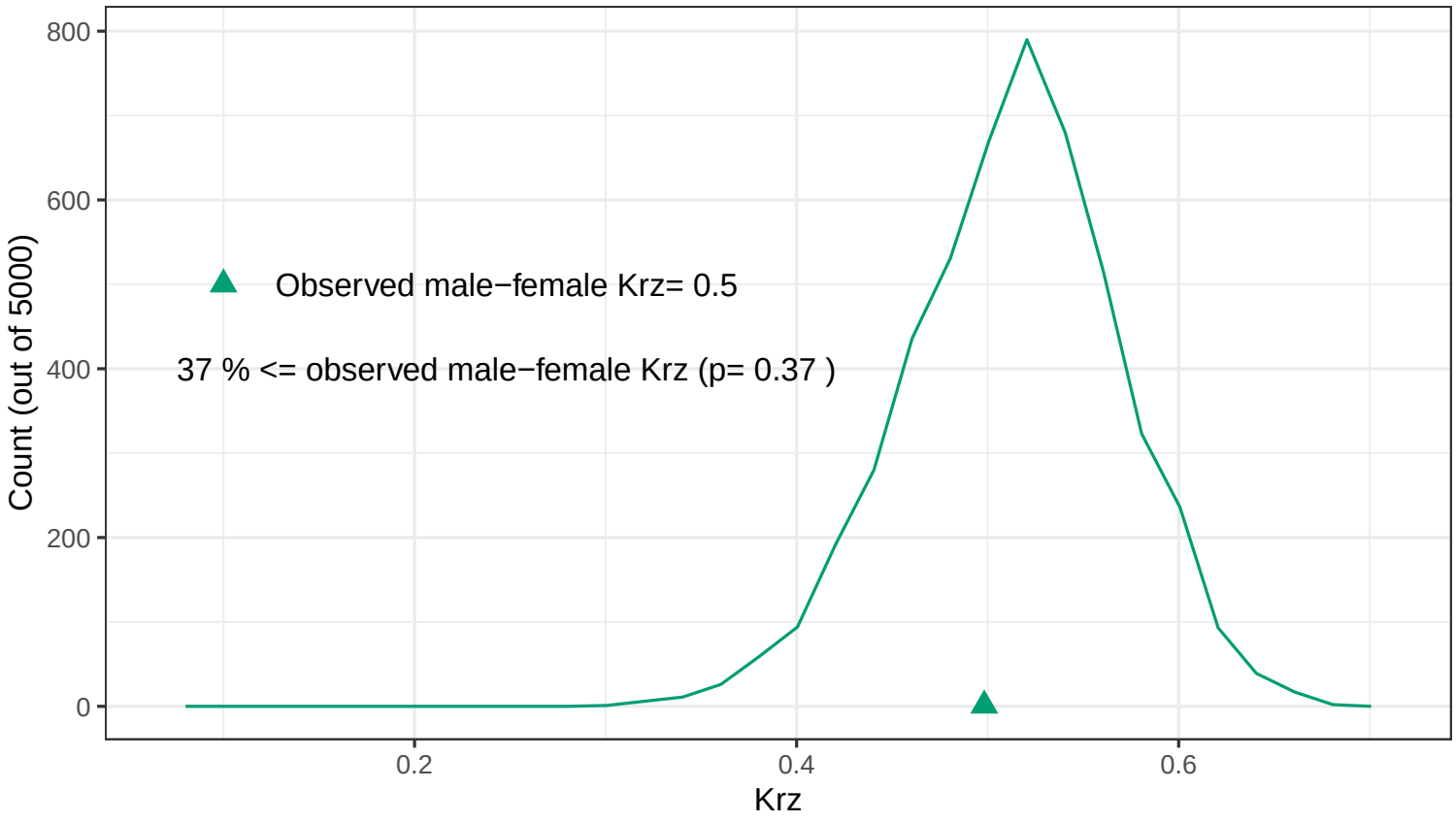
Type 1 error (vs Krz, Pro)



Power (vs Krz, Pro)



So what if we redo the test with Krz?



Other uses

- Are whole population PCs adequate for subgroups?
- Permutation test for paired data (longitudinal questions, reproducibility)
- Could use with supervised dietay patterns (canonical correlation/reduced rank regression/PLS)