The **sensR** package: Difference and similarity testing

Per B Brockhoff

DTU Compute
Section for Statistics
Technical University of Denmark
perbb@dtu.dk

August 17 2015

DTU Compute
Department of Applied Mathematics and Computer Science

Main purposes of **sensR**

- Statistical tests of sensory discrimation and similarity data
- Power and sample size computations for discrimination and similarity tests
- Thurstonian analyses via d' (d-prime) estimation
- Improved confidence intervals via profile likelihood methods
- Linking "one sample at a time" Thurstonian analysis to more generic statistical analysis (regression and anova)

Main functions in **sensR**

d', CI, tests	Power & Sample size	Transformation	Illustration	d' comparisons	Miscellaneous
discrim	discrimPwr	rescale	plot	dprime_compare	findcr
AnotA	d.primePwr	psyfun	ROC	dprime_test	clm2twoAC
samediff	discrimSS	psyinv	AUC	dprime_table	SDT
twoAC	d.primeSS	psyderiv		posthoc	discrimR
betabin	twoACpwr	pd2pc			discrimSim
glm		pc2pd			samdiffSim
clm					
clmm					

The **sensR** package

Official site:
`www.cran.r-project.org/packages=sensR`

- Reference manual (51 pages)
- News / change-log
- Vignettes:
  - Statistical methodology for sensory discrimination tests and its implementation in **sensR**
  - Examples for papers

Supporting package: **ordinal** for ratings data, A-not A with sureness, 2-AC etc.

Obtaining **sensR**

1. Download and install **R** from `http://www.cran.r-project.org/`
2. Open **R** and run `install.packages("sensR")`
3. Tell **R** that you want to use it with `library(sensR)`
4. Analyze your data — try for example `discrim(10, 15, method="triangle")`

Citation

```
citation("sensR")

##
## To cite the sensR-package in publications use:
##
##   Christensen, R. H. B. & P. B. Brockhoff (2015). sensR - An
##   R-package for sensory discrimination. R package version 1.4-5.
##   http://www.cran.r-project.org/package=sensR/.
##
## A BibTeX entry for LaTeX users is
##
##   @Misc{,
##     title = {sensR---An R-package for sensory discrimination},
##     author = {R. H. B. Christensen and P. B. Brockhoff},
##     year = {2015},
##     note = {R package version 1.4-5. http://www.cran.r-project.org/package=sensR/},
##   }
```


Protocols supported in **sensR**

Discrimination	d' estimation	Difference test	Similarity test	Power	Sample size	Simulation	Likelihood CI	Replicated	Regression analysis
Duo-Trio, Triangle, Tetrad	✓	✓	✓	✓	✓	✓	✓	✓	✓
2-AFC, 3-AFC	✓	✓	✓	✓	✓	✓	✓	✓	✓
A-not A	✓	✓	✓				✓	✓	✓
Same-Different	✓	✓	(✓)	✓		✓	✓		
2-AC	✓	✓	✓	✓			✓	✓	✓
A-not A w. Sureness	✓	✓	(✓)				✓	✓	✓

Publications related to **sensR** (and **ordinal**)

- Brockhoff , P. B., & Christensen, R. H. B. (2010). Thurstonian models for sensory discrimination tests as generalized linear models. Food Quality and Preference, 21, 330–338.
- Christensen, R. H. B., & Brockhoff, P. B. (2009). Estimation and Inference in the Same Different Test. Food Quality and Preference, 20, 514–524.
- Christensen, R. H. B., Cleaver, G., & Brockhoff, P. B. (2011). Statistical and Thurstonian models for the A-not A protocol with and without sureness. Food Quality and Preference, 22, 542-549.
- Christensen, R. H. B., Lee, H-S. & Brockhoff , P. B. (2012). Estimation of the Thurstonian Model for the 2-AC Protocol. Food Quality and Preference, 24, 119-128.
- Christensen, R. H. B., & Brockhoff, P. B. (2013). Analysis of sensory ratings data with cumulative link models. J. Soc. Fr. Stat. & Rev. Stat. App., 154(3), 58-79.
- Christensen, R.H.B., Ennis, J.M., Ennis, D.M. & Brockhoff, P.B.(2014). Paired preference data with a no-preference option - Statistical tests for comparison with placebo data. Food Quality and Preference, 32, 48-55.
- John M Ennis, Rune HB Christensen (2014). Precision of measurement in Tetrad testing. Food Quality and Preference, Vol 32, 98-106.
- Ennis, J.M. & Christensen, R.H.B.(2015). A Thurstonian comparison of the Tetrad and Degree of Difference tests. Food Quality and Preference, Vol 40, 263-269.
- T. Næs, P.B. Brockhoff and O. Tomic, (2010). Statistics for Sensory and Consumer Science, John Wiley & Sons, Chapter 7.

Discrimination testing - what can we use it for?

Sensory discrimination testing is used to

- Show difference or similarity between products
- Detect unwanted product changes in product development
- Guide ingredient substitution (e.g. health initiatives)
- Detect production stability
- Measure consumer sensitivity
- Substantiate claims ("75% of consumers can tell the difference")
- Dispute claims ("our product is at least as good as yours")

Outline

Example 1: Ingredient substitution

In a calorie reduced yoghurt natural sweetener from from the *Stevia* plant is used. Unfortunately, consumers complained about a bitter after taste. The company decided to change to a different supplier of Stevia sweetener and now wants to know if consumers can actually tell the difference between the current production and the new yoghurt.

- It is decided to use a Triangle test with 100 consumers.
- Each subject is presented with 3 samples, two of which are equal
- Question: which sample is most different from the two others?

X Y X

Example 1: Ingredient substitution

The experiment was performed:

- Sensory discrimination protocol: Triangle
- Sample size: $N = 100$, no. correct answers: $X = 45$

Do consumers perceive the current and new yoghurts as *significantly different*?

- What are high and low values of X?
- What is the alternative hypothesis, H_A?
- What is the null hypothesis, H_0?

What can we do to test our hypotheses?

- How many correct answers would we expect if H_0 is true and $p_c = 1/3$?
- What would an extreme observation be if H_0 is true?
- How extreme is 45 correct out of 100 if products are not different?

What is the expected proportion of correct answers?

Which range of values of p_c are plausible given 45 correct answers out of 100?

- We don't know the true p_c
- We can estimate p_c: $\hat{p}_c = X/N = 0.45$
- Could the true p_c be 0.40?
- Could the true p_c be 0.33?
- Could the true p_c be 0.90?
- Which values for the true p_c are in fact plausible?

We need a *confidence interval*!

Conclusions from Example 1

- p_c: The *expected* proportion of correct answers
- $\hat{p}_c$: The *observed* proportion of correct answers
- H_0: Products are *not* different, $p_c = 1/3$
- H_A: Products are different, $p_c > 1/3$
- $x_c = 45$ is the critical value: $Pr(X \geq 45) \leq 0.05$ if products are not different
- p-value: The probability of observing 45 correct answers or more if the products are not different (H_0 is true) is $p = 0.01$
- Plausible values of p_c are $(0.35;\ 0.55)$ based on a 95% confidence interval

Duo-Trio, Triangle, Tetrad, 2-AFC, and 3-AFC

Duo-Trio

- 1 reference sample X_R and 2 test samples X and Y
- *Which sample is most similar to the reference?*

Triangle:

- 3 samples (X, X', Y)
- *which sample is most different from the two others?*

Tetrad:

- 4 samples (X, X', Y, Y')
- *Group the samples based on similarity?*

2-AFC:

- 2 samples X and Y
- *which sample has the strongest/weakest sensory magnitude?*

3-AFC:

- 3 samples (X, X', Y)
- *which sample has the strongest/weakest sensory magnitude?*

Duo-Trio, Triangle, Tetrad, 2-AFC, and 3-AFC

Protocol	Samples	Alternatives	p_0	Type
Duo-Trio	3	2	$1/2$	Unspecified
2-AFC	2	2	$1/2$	Specified
Triangle	3	3	$1/3$	Unspecified
Tetrad	4	3	$1/3$	Unspecified
3-AFC	3	3	$1/3$	Specified

Table: Four protocols with binomial answers.

- p_0: The probability of a correct answer by guessing
- Unspecified: General product differences
- Specified: Attribute specific differences (sweetness, saltiness, etc.)

Which protocol should I choose?

Different criteria:

- Specified or unspecified problem?
- No. samples to taste in each trial
- Statistical power

Recommendations:

- The Triangle test is often preferred over the Duo-Trio test due to its higher power.
- The Tetrad test is more powerful than the Triangle test, but requires that 4 instead of 3 samples be assessed in each trial.
- The 2-AFC test is often preferred over the 3-AFC test because it has similar power, but fewer samples in each trial.

Which proportion of subjects can actually tell the difference between the products?

Assume two kinds of subjects: *Discriminators* and *Guessers*

- Discriminators can tell the difference and *always* choose the right sample.
- Guessers can only guess and will *sometimes* choose the right sample.

We denote the *proportion discriminators* with p_d.

Is this model realistic?

Example 2: Proportion discriminators in a Triangle test

If $N = 100$ in a Triangle test and 10% can tell the difference, how many correct samples can I expect?

Proportion discriminators and proportion correct

From p_d to p_c:

Tetrad, Triangle, 3-AFC: $p_c = \frac{1}{3} + p_d \frac{2}{3}$ Duo-Trio, 2-AFC: $p_c = \frac{1}{2} + p_d \frac{1}{2}$

From p_c to p_d:

Tetrad, Triangle, 3-AFC: $p_d = \frac{p_c - 1/3}{2/3}$ Duo-Trio, 2-AFC: $p_d = \frac{p_c - 1/2}{1/2}$

Example 3: What is the proportion of discriminators?

What is the proportion of discriminators in the sweetener example?

Examples in R

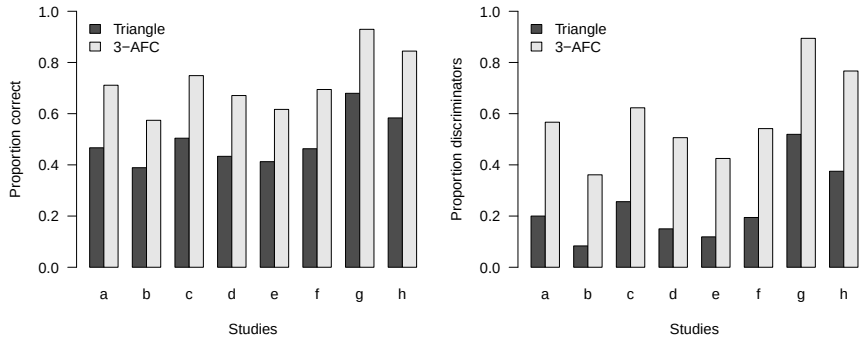
What proportion of correct answers should we expect if we had used the Duo-Trio protocol instead?

Gridgeman's paradox

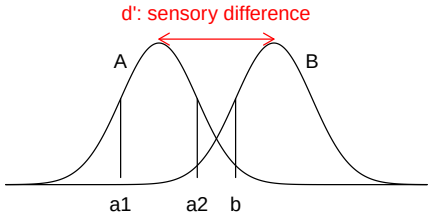
Study	Product	no. tests	Triangle	3-AFC
Byer & Abrams (1953)	Bitter solutions	45	21	32
Stillman (1993)	Onion dip	108	42	62
Tedje et al. (1994)	Salt solutions	720	363	539
		240	104	161
		240	99	148
Masuoka et al. (1995)	Beer	108	50	75
Delwiche & O'Mahony (1996)	Choc. pudding	156	106	145
Rousseau & O'Mahony (1997)	Yoghurt	180	105	152

Why does 3-AFC consistently give more correct answers than Triangle?

Gridgeman's paradox

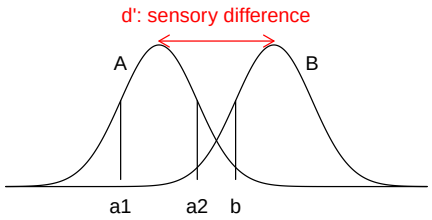


The Thurstonian model, 3-alternatives



- Normal distributions with equal variance
- $d' = (\mu_2 - \mu_1)/\sigma$ — a *signal-to-noise* measurement
- Each decision rule (protocol) leads to a specific relation between d' and p_c
- Is the answer correct for Triangle? for 3-AFC?

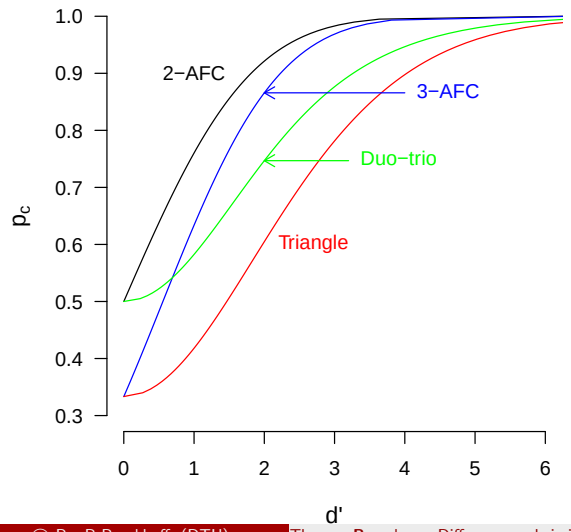
Variability in Thurstonian models



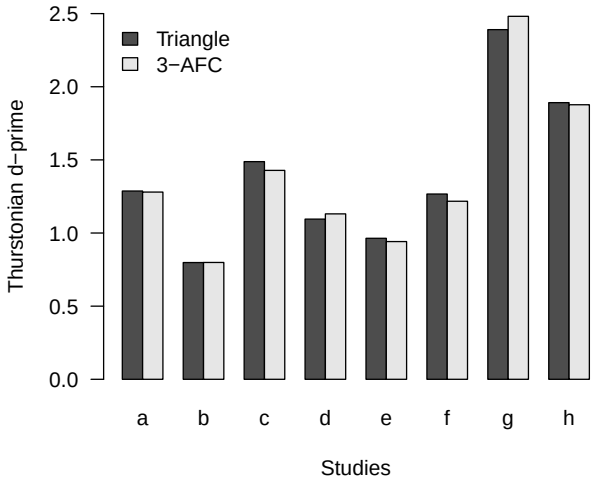
Variability in sensory intensity comes from

- Variability in stimulus (sample-sample variation, temperature, etc.)
- Variability within the subject (fatigue, learning, presentation order)

Psychometric functions

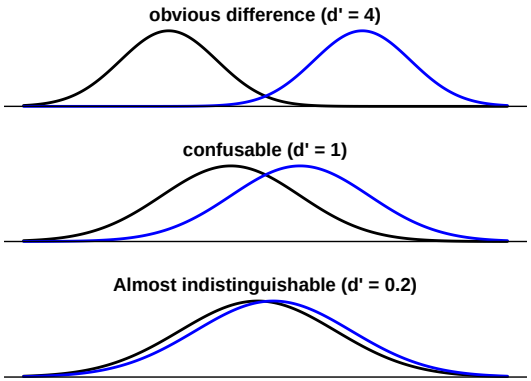


Gridgeman's paradox resolved



Triangle and 3-AFC agree for all studies when d' is used as a measure of sensory difference.

Examples of d-prime



Example 4: d-prime estimation

What is d-prime in the sweetener example?

How low or high could we reasonably expect the true d-prime (δ) to be?

Why plan sensory difference tests?

Because

- Too many tests is expensive and a waste of resources
- Too few tests is useless and a waste of resources

Type I and Type II error rates

	Accept H_0	Reject H_0
H_0 is true	✓	Type I error (α)
H_A is true	Type II error (β)	✓

Five parameters involved when designing a sensory difference test:

- Sample size
- Power ($1-\beta$)
- Significance level (α)
- Sensory difference that you want to detect (d', p_c, or p_d)
- Test protocol (Triangle, Duo-Trio, Tetrad, 2-AFC, 3-AFC)

Usually we want a power between 80% and 90%.

What is the power of a sensory difference test really?

Power is the probability

- of finding a difference (if it is really there)
- of finding a difference if H_A is true
- of rejecting H_0 if H_A is true
- of correctly rejecting H_0
- that the number of correct answers is at least as large as the critical value
- that the p-value from a sensory difference test is less than α

Example 5: Power estimation

What is the power of detecting $d' = 2$ in a Triangle test with $\alpha = 0.05$ and $n = 15$?

- What are the null and alternative hypotheses?
- What is the critical value for the test?

Example 6: Power and sample size estimation

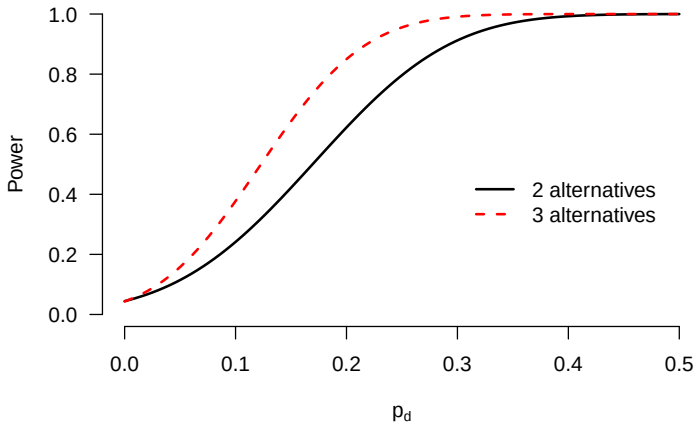
Some new technical equipment has been bought to your company to replace an old machine used in the production of your calorie reduced yoghurt. The producer of the new machine promises that your yoghurt will remain exactly as before, but upon tasting the result, you strongly believe there is a subtle but detectable difference.

To make sure you don't get complaints from your costumers, you decide to test whether consumers can tell the difference between the yoghurts made with the old and the new equipment.

You decide to use the Duo-Trio test and have a budget for 100 consumers.

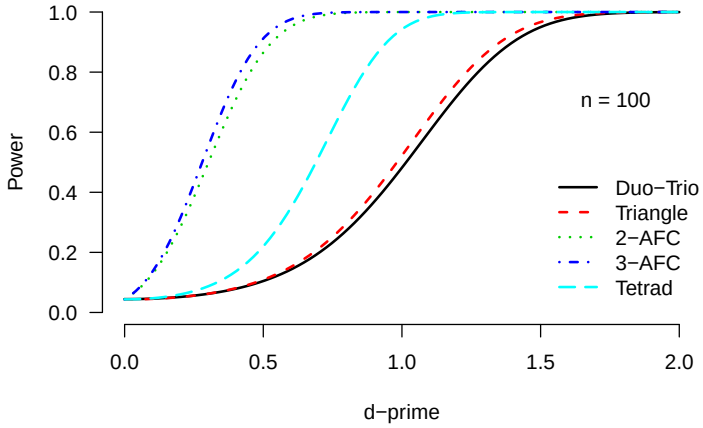
- What is the probability that you can detect a difference of $d' = 1$?
- How many consumers would you need to achieve 80% power?
- If you are convinced that the difference is primarily a difference in smoothness, then what can you do?

Power versus p_d



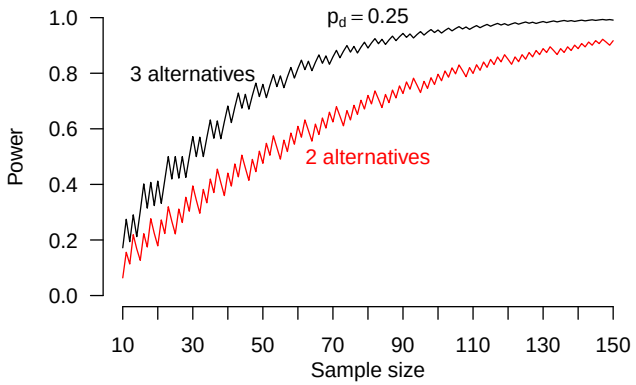
More alternatives means more power

Power versus d-prime



Specified tests are more powerful than unspecified tests

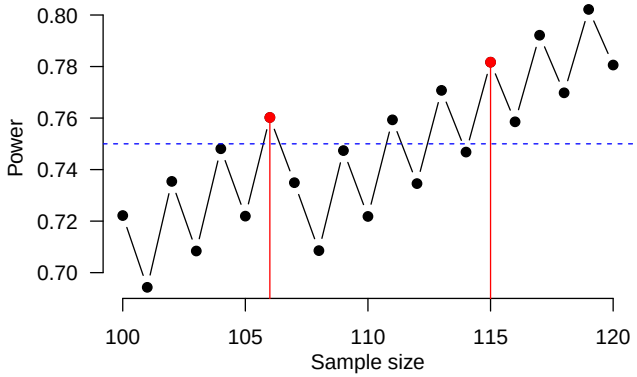
Power versus sample size



- Power is a zig-zag function of sample size.
- This is due to the discreteness of the binomial distribution
- Smoother behavior for 3 alternatives

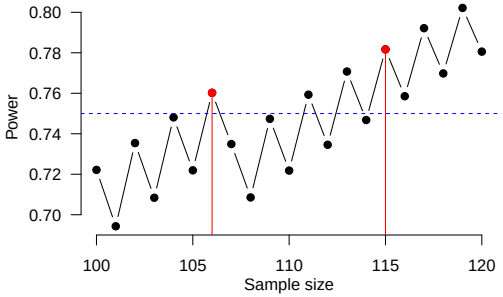
Reject the next consumer?

Power can **decrease** with increasing sample size!



- The **smallest** sample size that gives the required power
- Sample sizes provide guidelines — not conclusive requirements.

First exact and stable exact sample sizes



First exact sample size (conventional):

- The smallest sample size that has the required power

Stable exact:

- The smallest sample size that has the required power *such that no greater sample size has a power smaller than that required*.

Example 7: First-exact and stable-exact sample sizes

The ISO standard for the Triangle test describes that for $p_d = 0.10$, $\alpha = 0.05$ and $\beta = 0.001$, the required (first exact) sample size is 1181.

Lets see if we can replicate this number with sensR.

Protocols supported in **sensR**

Discrimination	d' estimation	Difference test	Similarity test	Power	Sample size	Simulation	Likelihood CI	Replicated	Regression analysis
Duo-Trio, Triangle, Tetrad	✓	✓	✓	✓	✓	✓	✓	✓	✓
2-AFC, 3-AFC	✓	✓	✓	✓	✓	✓	✓	✓	✓
A-not A	✓	✓	✓				✓	✓	✓
Same-Different	✓	✓	(✓)	✓		✓	✓		
2-AC	✓	✓	✓	✓			✓	✓	✓
A-not A w. Sureness	✓	✓	(✓)				✓	✓	✓

Main functions in **sensR**

d', CI, tests	Power & Sample size	Transformation	Illustration	d' comparisons	Miscellaneous
`discrim`	`discrimPwr`	`rescale`	`plot`	`dprime_compare`	`findcr`
`AnotA`	`d.primePwr`	`psyfun`	`ROC`	`dprime_test`	`clm2twoAC`
`samediff`	`discrimSS`	`psyinv`	`AUC`	`dprime_table`	`SDT`
`twoAC`	`d.primeSS`	`psyderiv`		`posthoc`	`discrimR`
`betabin`	`twoACpwr`	`pd2pc`			`discrimSim`
`glm`		`pc2pd`			`samdiffSim`
`clm`					
`clmm`					

Publications related to **sensR** (and **ordinal**)

- Brockhoff , P. B., & Christensen, R. H. B. (2010). Thurstonian models for sensory discrimination tests as generalized linear models. Food Quality and Preference, 21, 330–338.
- Christensen, R. H. B., & Brockhoff, P. B. (2009). Estimation and Inference in the Same Different Test. Food Quality and Preference, 20, 514–524.
- Christensen, R. H. B., Cleaver, G., & Brockhoff, P. B. (2011). Statistical and Thurstonian models for the A-not A protocol with and without sureness. Food Quality and Preference, 22, 542-549.
- Christensen, R. H. B., Lee, H-S. & Brockhoff , P. B. (2012). Estimation of the Thurstonian Model for the 2-AC Protocol. Food Quality and Preference, 24, 119-128.
- Christensen, R. H. B., & Brockhoff, P. B. (2013). Analysis of sensory ratings data with cumulative link models. J. Soc. Fr. Stat. & Rev. Stat. App., 154(3), 58-79.
- Christensen, R.H.B., Ennis, J.M., Ennis, D.M. & Brockhoff, P.B.(2014). Paired preference data with a no-preference option - Statistical tests for comparison with placebo data. Food Quality and Preference, 32, 48-55.
- John M Ennis, Rune HB Christensen (2014). Precision of measurement in Tetrad testing. Food Quality and Preference, Vol 32, 98-106.
- Ennis, J.M. & Christensen, R.H.B.(2015). A Thurstonian comparison of the Tetrad and Degree of Difference tests. Food Quality and Preference, Vol 40, 263-269.
- T. Næs, P.B. Brockhoff and O. Tomic, (2010). Statistics for Sensory and Consumer Science, John Wiley & Sons, Chapter 7.

The Triangle test

- 3 samples (X, X', Y) OR (X, Y, Y')
- 6 presentation orders $XX'Y$, XYX', YXX', $YY'X$, YXY', XYY'
- Unspecified test
- Question: *which sample is most different from the two others?*

The Duo-Trio test

- 1 reference sample X_R and 2 test samples X and Y
- 4 presentation orders X_RXY, X_RYX, Y_RYX, Y_RXY
- Unspecified test
- Question: *Which sample is most similar to the reference?*
- Constant or variable reference?

The 2-AFC test

- 2 samples X and Y
- 2 presentation orders XY and YX
- Specified test
- Question: *which sample has the strongest/weakest sensory mangnitude?*
- Example: Which sample is the sweetest?

The 3-AFC test

- 3 samples (X, X', Y) OR (X, Y, Y')
- 6 presentation orders $XX'Y$, XYX', YXX', $YY'X$, YXY', XYY'
- Specified test
- Question: *which sample has the strongest/weakest sensory mangnitude?*
- Example: Which sample is the sweetest?