The Dutch new herring scandals

Richard D. Gill

This version: 29 January 2020

Warning: I must declare a conflict of interest

https://en.wikipedia.org/wiki/Soused_herring
https://www.economist.com/europe/2017/11/23/netherlands-fishmongers-accuse-herring-tasters-of-erring
https://sites.google.com/site/benvollaard/home
https://www.volkskrant.nl/nieuws-achtergrond/geen-blubberige-haring-en-oliebollen-als-schoenzolen-meer-ad-stopt-met-alle-smaaktesten~ba44b87b/

Netherlands fishmongers accuse herring-tasters of erring

Can the Dutch still trust their herring-tasters?

Print edition | Europe
Nov 23rd 2017
| AMSTERDAM

HERRING (genus *Clupea*, with four species found in the Baltic and North Seas) have been vital to northern Europe's economy since the Middle Ages, when fishermen worked out how to preserve them in brine. Every north European country maintains that there is a right way to eat the fish, but they differ as to what it is. In Sweden Baltic *surströmming* are fermented until slightly rancid. In Denmark the *sill* are pickled, or cooked and eaten in long strips. In the Netherlands *haring* must be lightly salted for preservation but otherwise raw, dipped in minced onion and accompanied with a pickle. No food is more loved.

So the Dutch were shocked when accusations surfaced in November that there was something rotten about the national herring test. The test, sponsored by the *Algemeen Dagblad*, a newspaper, is carried out by two expert tasters, who each year rate the herring at over a hundred shops and stands across the country. Ben Vollaard, an economist at Tilburg University, was surprised when his respected local fishmonger scored zero. The merchant told Mr Vollaard that one judge routinely tipped the scales, giving higher scores to stores that get their fish from the Atlantic Group, a distributor in Scheveningen. The judge happened to be a consultant for Atlantic, giving courses on how to slice and serve herring.

"I saw how much damage a low rating could do. The judges act like God," says Mr Vollaard, who specialises in using statistics to detect crime. He decided to run the numbers. The ratings include objective criteria, like weight and fattiness, and subjective ones such as taste and appearance. The economist contacted 85% of the shops surveyed in the past two years and asked who their distributors were. He found that whereas the overall average score was 5.5, the average for those supplied by Atlantic was 8.7. The extra boost for the Atlantic stores came mainly from the subjective scores.

Mr Vollaard's study has blown the lid off the sealed world of Dutch herring. Fishmongers who long suspected the judge of bias towards Atlantic now say the test is rotten. Two who received low ratings have vowed to sue the *Algemeen Dagblad* for defamation.

The judge and Atlantic say they have been smeared, and that the statistical evidence is a red herring. They say Mr Vollaard's figures are off, and that their high scores are due to their superior fish. But the charges of *belangenverstrengeling* (conflict of interest) have left the test's reputation for impartiality gutted.

This article appeared in the Europe section of the print edition under the headline "Failing the smell test"
Print edition | Europe
Nov 23rd 2017
| AMSTERDAM

Over Ons

Haring en zeevis(groot)handel Atlantic

Abdullah (Appie) en Umut Tagi zijn samen met hun neef Ali Bagcan Haring en Zeevis(groot)handel Atlantic gestart vanuit hun passie voor goede vis. Atlantic importeert, exporteert, be- en verwerkt en levert haring en verse en diepgevroren (zee)vis(producten) aan groot- en detailhandel en horeca door heel Nederland en daarbuiten. Daarnaast werven, selecteren en detacheren we ervaren en hooggekwalificeerd personeel en leiden we ondernemers en hun medewerkers in de visbranche op in het aloude visambacht.

De groep bestaat uit enkele groothandelsvestigingen, een grootschalige verwerkingslokatie, een groeiend aantal visdetailhandelsvestigingen zijnde winkels en marktverkoopplaatsen en een gerenommeerd opleiding- en detacheringsinstituut.

CentER

TILBURG UNIVERSITY

Economics Ranking

According to the Tilburg University Economics Ranking we are ranked:

Europe:

1. London School of Economics and Political Sciences
2. University of Oxford
3. Tilburg University

World:

19. University of Maryland
20. University of California, San Diego
21. Tilburg University
22. University College London

Office hours: by appointment

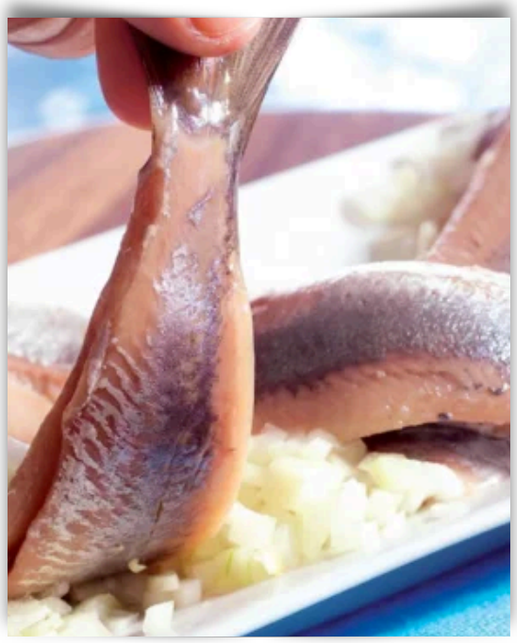
Ben Vollaard (Ph.D. RAND Graduate School) is associate professor at the Economics Department of Tilburg University, where he has been on the faculty since 2008. As of 2019, he leads the Fraud Detection and Prevention Lab. In 2014, he won the Tilburg School of Economics and Management Research Valorization Award. He is also recipient of a TiU Alumni Fund grant in 2016.

Past positions include graduate fellow at the RAND Corporation, reporter at NRC Handelsblad and research economist at the Dutch equivalent of the Council of Economic Advisors (CPB).

His research revolves around the question how to prevent illegal behavior, corporate fraud in particular, in a (cost)effective manner. When studying this question, he focuses on how situational factors affect behavior. For instance, an

Vollaard's work on Dutch new herring

AD Haringtest geeft viszaken in Rotterdam e.o. meer punten (juli 2017). *NEMO Kennislink*, *Vismagazine*.



Belangenverstrengeling bij AD haringtest: testpanel heeft alle schijn tegen, zie persbericht (plus link naar PDF) (nov 2017). *EenVandaag*, *DenHaagFM*, *The Economist*, *Jinek*. Related Stats Assignment by a teacher at a college in NY. Tot slot: *de Volkskrant* nu het doek is gevallen (juni 2018).

Vollaard's work

(on Dutch new herring)

- July 2017: 1st report and press-release
  - Unfair advantage for region close to Rotterdam
  - Dummy variable "> 30 Km" significant
- November 2017: 2nd report and press-release
  - Unfair advantage for clients (retailers) of wholesale company "Atlantic"
  - Dummy variable *not* significant
  - Two "subjective" variables responsible for nearly 50% of the variation
  - Two "objective" variables responsible for most of rest
  - Three further "objective" variables unimportant
  - Subjectively judged "ripeness" is important, but objective "microbiological test" unimportant

Latest news

- LOWI is ready (Advice nr. 23, 2019)
    - "The LOWI would applaud it if a scientific debate could take place between all parties involved, and experts”.
    - "It is clear that mistakes were made in the press release which presented the results of this research to the public"
- Data and analysis programs have been released by Ben Vollaard (see his personal webpage)
- These are a *Stata* data set and a *Stata* script. Fortunately *R* can nowadays input Stata data files,“.dta” extension, and the Stata script, a “.do” file, is a plain text file

NAME _____ 6-DIGIT CODE 100113

Questions (3)-(7). Are herring quality scores being rigged? The Dutch people love to eat herring and take its quality very seriously. A national board regularly samples and scores the quality of herring vendors. However, there have been concerns whether the ratings have been unduly favoring suppliers who are linked to a certain distributor. Professor Vollaard of Tilburg University investigated the matter by collecting the scores obtained by 293 vendors, 21 linked to the distributor, and 272 not linked. Assume that the population variances are approximately equal.

	~~Areas sampled~~	Mean score (range: 1-10)	Sample standard deviation
Linked	21	8.75	2.03
Not linked	272	5.71	2.97

Data from AD Herring Tests 2016 and 2017 (?)

NAME _____ **6-DIGIT CODE 100113**

Questions (3)-(7). Are herring quality scores being rigged? The Dutch people love to eat herring and take its quality very seriously. A national board regularly samples and scores the quality of herring vendors. However, there have been concerns whether the ratings have been unduly favoring suppliers who are linked to a certain distributor. Professor Vollaard of Tilburg University investigated the matter by collecting the scores obtained by 293 vendors, 21 linked to the distributor, and 272 not linked. Assume that the population variances are approximately equal.

	Areas sampled	Mean score (range: 1-10)	Sample standard deviation
Linked	21	8.75	2.03
Not linked	272	5.71	2.97

(3) State the hypotheses to test whether there is a difference in the scores of vendors who are linked to the distributor and those who are not.

A. $H_0: \mu_{Linked} - \mu_{Not} \leq 0, H_a: \mu_{Linked} - \mu_{Not} > 0$
B. $H_0: \mu_{Linked} - \mu_{Not} \neq 0, H_a: \mu_{Linked} - \mu_{Not} = 0$
C. $H_0: \mu_{Linked} - \mu_{Not} \geq 0, H_a: \mu_{Linked} - \mu_{Not} < 0$
D. $H_0: \mu_{Linked} - \mu_{Not} = 0, H_a: \mu_{Linked} - \mu_{Not} \neq 0$
E. None of the above is correct.

(4) The standard error of the difference between the two sample means is closest to:

A. 0.149
B. 0.386
C. 0.436
D. 0.660
E. None of the above is correct.

(5) At 99% confidence, the decision rule is:

A. Reject H_0 if $z > 2.575$ or $z < -2.575$
B. Reject H_0 if $t > 2.316$ *or* $t < 2.316$
C. Reject H_0 if $t > 2.617$ *or* $t < -2.617$
D. Reject H_0 if $z > 2.325$ or $z < -2.325$
E. None of the above is correct.

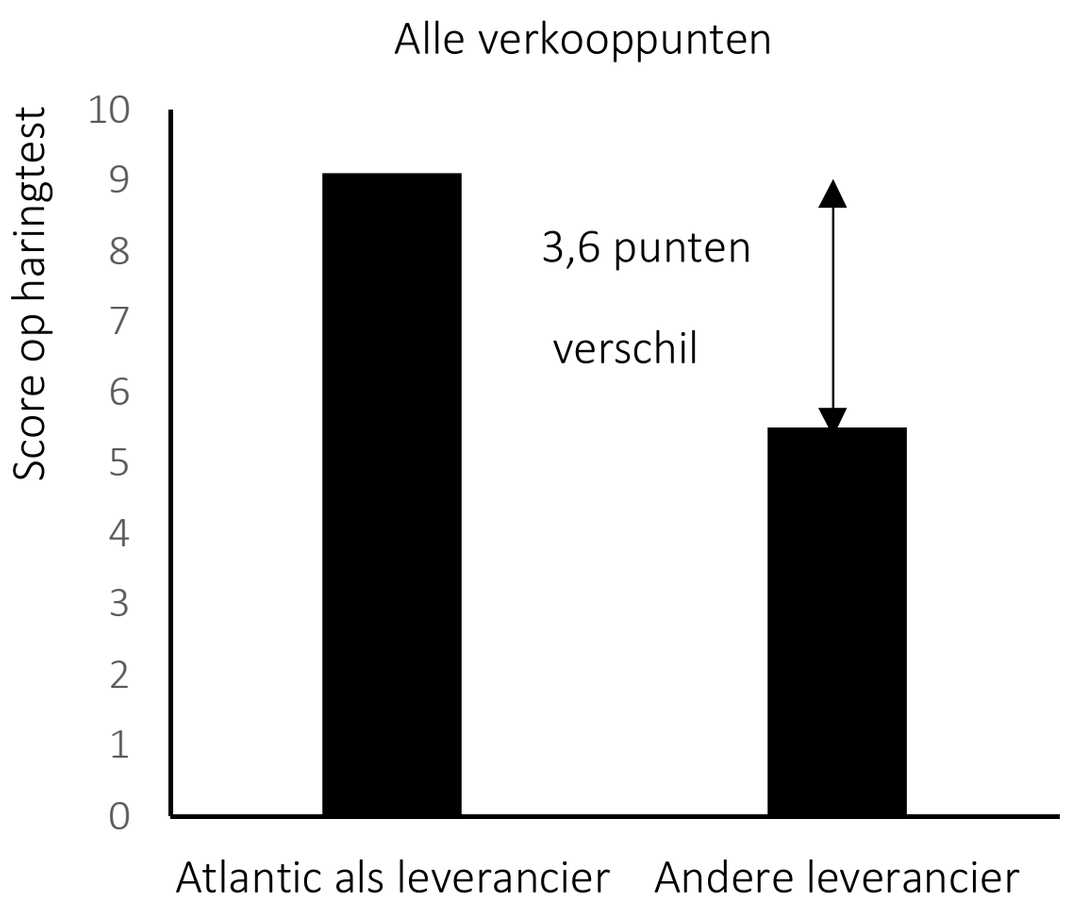
(6) The calculated test statistic is closest to:

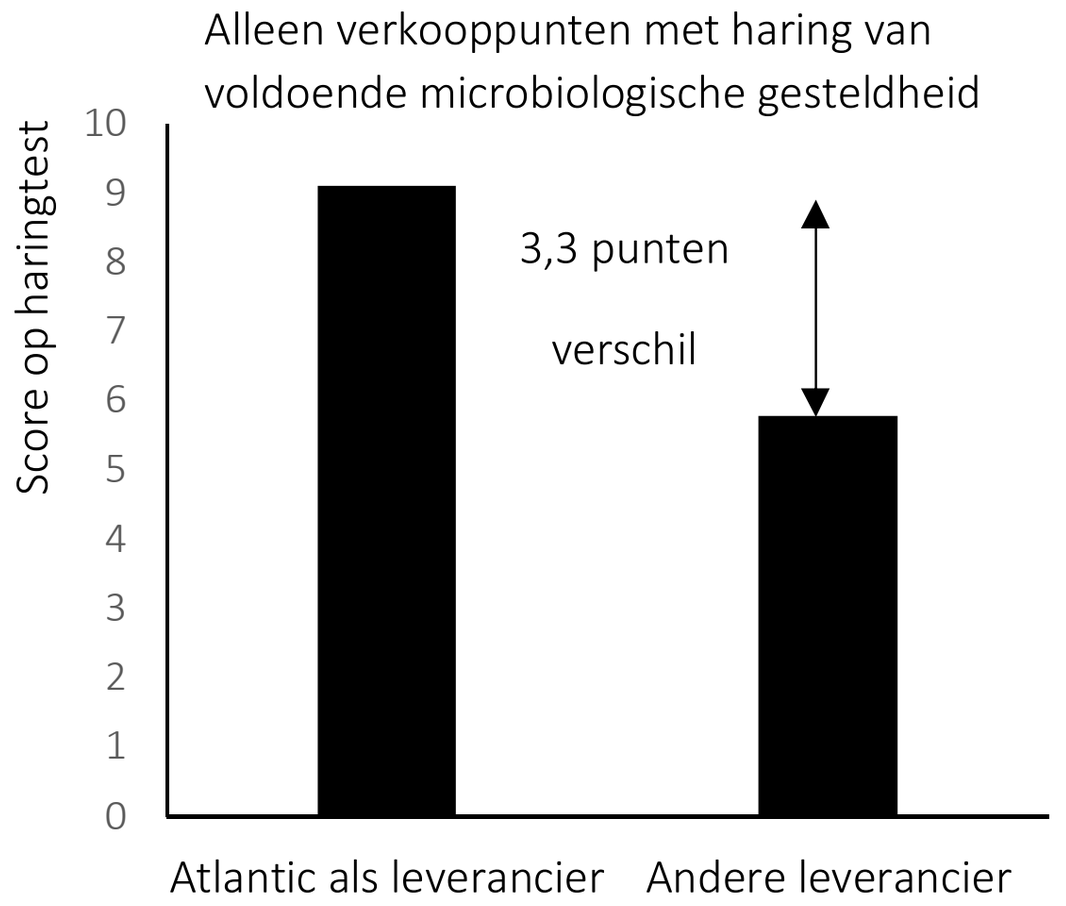
A. -4.34
B. -2.13
C. 2.27
D. 4.60
E. 7.87

(7) At 99% confidence, is there sufficient evidence to conclude that herring vendors linked to the distributor obtain higher test scores?

A. Yes, because the calculated test statistic is greater than the critical value.
B. No, because the calculated test statistic is less than the critical value.
C. Yes, because the calculated test statistic is less than the critical value.
D. No, because the calculated test statistic is greater than the critical value.
E. None of the above is correct.

Figuur 1. *Gemiddelde score van verkooppunten op de haringtest 2016/17, naar leverancier*





n1 + n2 = 292

n1 + n2 = 292 - 37

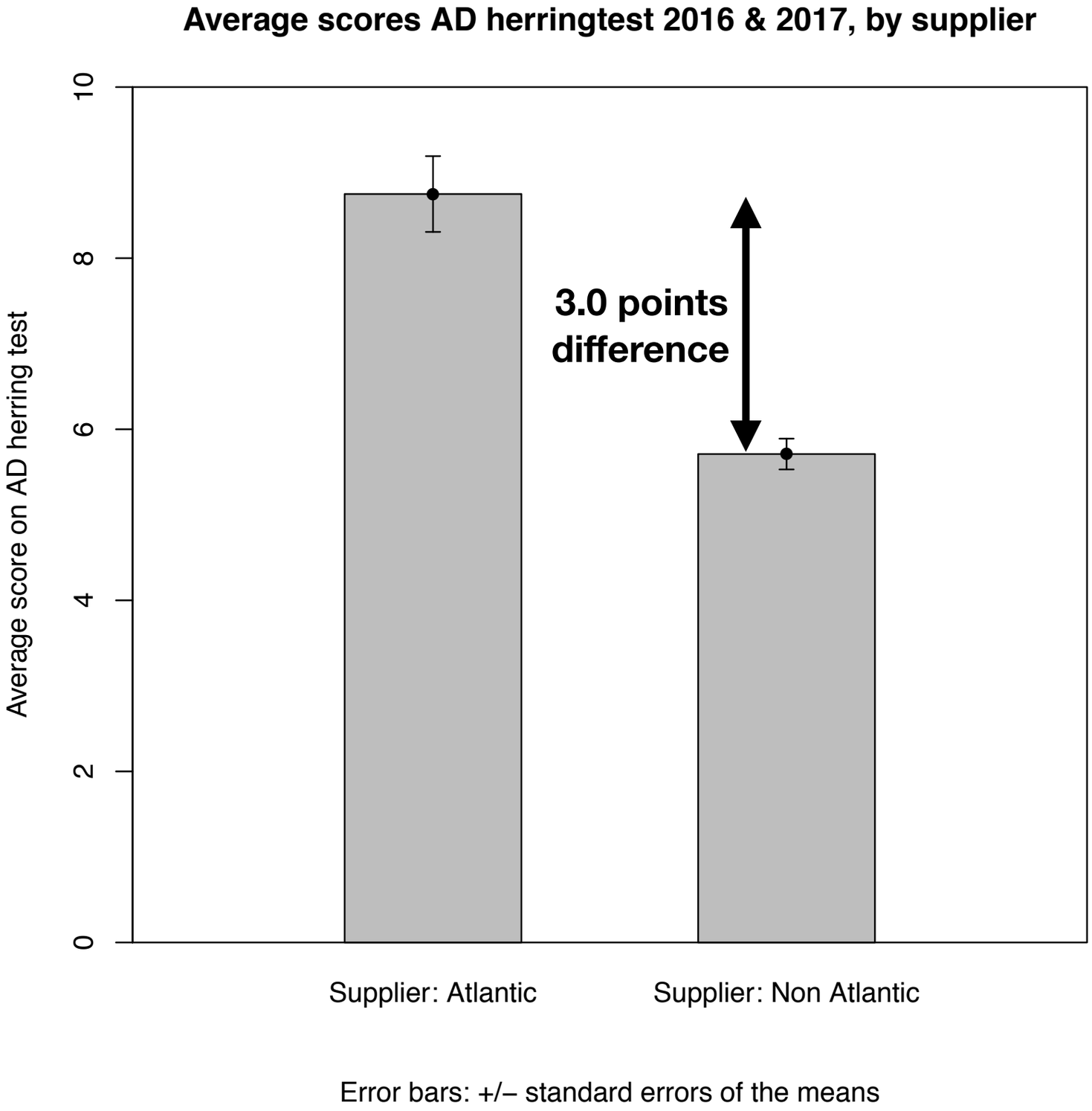
From report 2

```
> n1 <- 21
> n2 <- 272
> xbar1 <- 8.75
> xbar2 <- 5.71
> sdev1 <- 2.03
> sdev2 <- 2.97
> diff <- xbar1 - xbar2
> diff
[1] 3.04
> t <- diff /
+           sqrt((1/n1 + 1/n2) *
+           (((n1 - 1) * sdev1^2 + (n2 - 1) * sdev2^2)/(n1 + n2 - 2)))
> t
[1] 4.60446
> pnorm(t, lower.tail = FALSE)
[1] 2.06769e-06
> pt(t, df = n1 + n2 - 2, lower.tail = FALSE)
[1] 3.093617e-06
```

Data: Stats Assignment @City College of New York.

n1 + n2 = 293

Data: Stats Assignment @City College of New York



p << 0.01

```r
# Simple Bar Plot with error bars

library(Hmisc)
n1 <- 21
n2 <- 272
xbar1 <- 8.75
xbar2 <- 5.71
sdev1 <- 2.03
sdev2 <- 2.97
means <- c(xbar1, xbar2)
se <- c(sdev1/sqrt(n1), sdev2/sqrt(n2))
names <- c("Supplier: Atlantic", "Supplier: Non Atlantic")
lower <- means - se
upper <- means + se
bp = barplot(means, ylim=c(0,10), names.arg = names,
    ylab = "Average score on AD herring test",
    xlab = " ", xpd=T, width=c(0.2,0.2), xlim=c(0,1),
    space=c(1,1),
    main = "Average scores AD herringtest 2016 & 2017,\
                        by supplier",
    sub= "Error bars: +/- standard errors of the means")
errbar(bp, means, upper, lower, add=T)
box(bty="O")
```

eindcijfer = a. gewicht + b. temperatuur + c. vet + d. vers + e. microbiologische gesteldheid + f. rijping + g. schoonmaken + h. regio Rotterdam + i. top10 + j. indicator 2017 + constante + restterm

Tabel A1. Verklaring van het eindcijfer in de haringtest 2016 en 2017

Te verklaren: eindcijfer haringtest (0 tot 10)	
Gewicht (gr)	0,04 (0,01)***
Temperatuur	
beneden 7°C	referentiecategorie
tussen 7 en 10 °C	-0,58 (0,19)***
boven 10 °C	-1,71 (0,22)***
Vetpercentage	
beneden 10%	referentiecategorie
tussen 10 en 14%	0,18 (0,19)
boven 14%	0,66 (0,25)***
Vers van het mes	
niet	referentiecategorie
wel	1,78 (0,20)***
Microbiologische gesteldheid	
(zeer) goed	referentiecategorie
voldoende	-0,14 (0,31)
slecht	-0,50 (0,44)
waarschuwingsfase	-0,14 (0,28)
afgekeurd	-2,45 (0,68)***
Rijping	
licht	referentiecategorie
gemiddeld	-0,36 (0,33)
sterk	-1,87 (0,38)***
bedorven	-4,59 (0,49)***
Schoonmaken	
zeer goed	referentiecategorie
goed	-0,82 (0,21)***
Matig	-1,56 (0,22) ***
Slecht	-2,64 (0,43)***
Regio	
binnen straal van 30 km rond R'dam	referentiecategorie
meer dan 30 km van R'dam	-0,30 (0,17)*
Top 10	
top 10	referentiecategorie
buiten top 10	-0.96 (0,32)***
Indicator 2017	
haringtest 2016	referentiecategorie
haringtest 2017	0,21 (0,17)
Constante	3,72 (0,75)***
Aantal waarnemingen	292
Verklaarde variantie (R^2)	0,84

Noot: Tussen haakjes staan de standaardfouten rond de geschatte coëfficiënten. *** statistisch significant op het 1 procent-niveau; ** op het 5 procent-niveau en * op het 10-procent niveau.

Tabel A1. Verklaring van het eindcijfer in de haringtest 2016 en 2017

Te verklaren: eindcijfer haringtest (0 tot 10)	
Gewicht (gr)	0,04 (0,01)***
Temperatuur	
beneden 7^{o}C	referentiecategorie
tussen 7 en 10 ^{o}C	-0,58 (0,19)***
boven 10 ^{o}C	-1,71 (0,22)***
Vetpercentage	
beneden 10%	referentiecategorie
tussen 10 en 14%	0,18 (0,19)
boven 14%	0,66 (0,25)***
Vers van het mes	
niet	referentiecategorie
wel	1,78 (0,20)***
Microbiologische gesteldheid	
(zeer) goed	referentiecategorie
voldoende	-0,14 (0,31)
slecht	-0,50 (0,44)
waarschuwingsfase	-0,14 (0,28)
afgekeurd	-2,45 (0,68)***
Rijping	
licht	referentiecategorie
gemiddeld	-0,36 (0,33)
sterk	-1,87 (0,38)***
bedorven	-4,59 (0,49)***

Rijping	
licht	referentiecategorie
gemiddeld	-0,36 (0,33)
sterk	-1,87 (0,38)***
bedorven	-4,59 (0,49)***
Schoonmaken	
zeer goed	referentiecategorie
goed	-0,82 (0,21)***
Matig	-1,56 (0,22) ***
Slecht	-2,64 (0,43)***
Regio	
binnen straal van 30 km rond R'dam	referentiecategorie
meer dan 30 km van R'dam	-0,30 (0,17)*
Top 10	
top 10	referentiecategorie
buiten top 10	-0.96 (0,32)***
Indicator 2017	
haringtest 2016	referentiecategorie
haringtest 2017	0,21 (0,17)
Constante	3,72 (0,75)***
Aantal waarnemingen	292
Verklaarde variantie (R^2)	0,84

Noot: Tussen haakjes staan de standaardfouten rond de geschatte coëfficiënten. *** statistisch significant op het procent-niveau; ** op het 5 procent-niveau en * op het 10-procent niveau.

Tabel 1. ***Waarom verkooppunten die niet Atlantic als leverancier hebben lager scoren op de haringtest***

Verklaringsfactor	Verschil tussen verkooppunten met en zonder Atlantic als leverancier	Eenheid	Effect op eindcijfer (0-10)
Gewicht	-2.17	gram	-0.09
Microbiologische gesteldheid			
(zeer) goed	referentiecategorie		
Voldoende	0.07	aandeel	-0.01
Slecht	0.03	aandeel	-0.02
waarschuwingsfase	0.09	aandeel	-0.01
Afgekeurd	0.02	aandeel	-0.04
Gezamenlijk			-0.07
Vetpercentage			
beneden 10%	referentiecategorie		
tussen de 10 en 14%	-0.21	aandeel	-0.04
boven 14%	-0.02	aandeel	-0.01
Gezamenlijk			-0.05
Temperatuur			
beneden 7^{o}C	referentiecategorie		
tussen 7 en 10^{o}C	0.29	aandeel	-0.17
boven 10^{o}C	0.21	aandeel	-0.36
Gezamenlijk			-0.52
Vers van het mes			
niet	referentiecategorie		
wel	-0.29	aandeel	-0.51
Schoonmaken			
zeer goed	referentiecategorie		
goed	0.27	aandeel	-0.22
matig	0.25	aandeel	-0.40
slecht	0.04	aandeel	-0.12
gezamenlijk			-0.73
Rijping			
licht	referentiecategorie		
gemiddeld	-0.24	aandeel	0.09
sterk	0.34	aandeel	-0.63
bedorven	0.06	aandeel	-0.25
gezamenlijk			-0.80
Regio			
binnen straal van 30 km rond R'dam	referentiecategorie		
meer dan 30 km van R'dam	0.60	aandeel	-0.18
Top 10			
top 10	referentiecategorie		

Vollaard's work

(on Dutch new herring)

- July 2017: 1st report and press-release
  - Unfair advantage for region close to Rotterdam
  - Dummy variable "> 30 Km" significant
- November 2017: 2nd report and press-release
  - Unfair advantage for clients (retailers) of wholesale company "Atlantic"
  - Dummy variable *not* significant !!!
  - Two "subjective" variables responsible for nearly 50% of the variation
  - Two "objective" variables responsible for most of rest
  - Three further "objective" variables unimportant
  - Subjectively judged "ripeness" is important, but objective "microbiological test" unimportant

Does analysis tell us anything we didn't know before?

Here: focus on Report 2

- "Atlantic" outlets tend to get high scores
- High scores are associated with the "subjective" factors "how the fish is cleaned" and "ripeness"
- "Temperature" and "freshly cleaned" are also important

NB: temperature < 5 deg is a legal requirement

Obvious Criticism

- Many outlets are in both years' tests
- "Top 10" as an explanatory variable???
- Paucity of information, unavailability of data
- Naivety of "model"
- Author pays lip-service to "correlation is not causation" but makes it clear that he believes that his analysis strengthens the accusations of bias
- One outlet wrongly classified, but analysis has not been corrected
- The data has *not* been publicly released; request to receive data has been refused

Not so obvious Criticism

- The data is not a random sample!!!
- Outlets (and their clients) “nominate themselves” for inclusion
- Badly scoring outlets refuse to participate in future years
- We *have* to model the selection mechanism if want to draw *causal* conclusions

https://www.npo3.nl/de-rekenkamer/11-12-2014/KN_1663955

KORT PROGRAMMA'S NIEUWS FILM & SERIE 3LAB CLUB HUB SKAMNL

DE REKENKAMER

DE REKENKAMER

Do 11 dec 2014 20:26 - 20:51
Wat kost een oliebol?

Wat kost het en wat krijg je ervoor? KRO De Rekenkamer loodst de kijker door het getallenoerwoud zodat we antwoord krijgen op de vraag: wat kost dat nou eigenlijk? En wat krijg ik er (niet) voor? Niemand weet meer wat iets écht kost.

Scientific Integrity?

Scientific competence?

- Some quotes from report 2
  - Intro: “Centraal staat de volgende vraag: heeft het belangenconflict binnen het testpanel invloed op hun oordeel?”
  - Conclusion: “Het grote verschil in gemiddeld eindcijfer tussen verkooppunten met en zonder Atlantic als leverancier is voor bijna de helft te verklaren uit een subjectieve inschatting door het testteam hoe goed de haring is schoongemaakt (zeer goed/goed/matig/slecht) en van de mate van rijping van de haring (licht/ gemiddeld/sterk/bedorven). In mindere mate zijn de gemeten temperatuur van de geserveerde haring, of de haring vers van het mes is, en plaatsing in de top 10 van belang voor het verklaren van het verschil in oordeel. Onbelangrijk blijken verschillen in het gewicht, de microbiologische gesteldheid en het vetpercentage van de haring – alle objectief meetbare factoren.”
  - Final summary: “Deze punten vormen geen direct bewijs van het bevoordelen van vishandels met de bewuste leverancier in de AD haringtest, maar het testteam heeft op basis van dit onderzoek wel alle schijn tegen.

“Het testteam heeft op basis van dit onderzoek wel alle schijn tegen”

“On the basis of this research, all appearances are against the test-team”

The intended reader is the Dutch public.

What is the intended take-home message to the intended reader?

Conclusions and final thoughts

- Thoughts on
  - scientific integrity vs. scientific competence
  - systems for investigating scientific integrity
  - systems for funding of research and education
- Recommendations to the statistical community
- An anecdote
- My recommendations on where to eat Dutch new herring
- Why couldn’t the data be released three years ago?