ZERO-INFLATED STATISTICAL MODELS FOR ANIMAL BEHAVIOR STUDIES

Jay Harrison

AFB International®

Too many zeroes?

Hypothetical data: Which is the better treatment?

Treatment 1	Treatment 2
0	100
0	120
0	180
200	200

Average: 50

Average without zeroes: 200

Sample size without zeroes: 1

Average: 150

Average without zeroes: 150

Sample size without zeroes: 4

AFB International overview

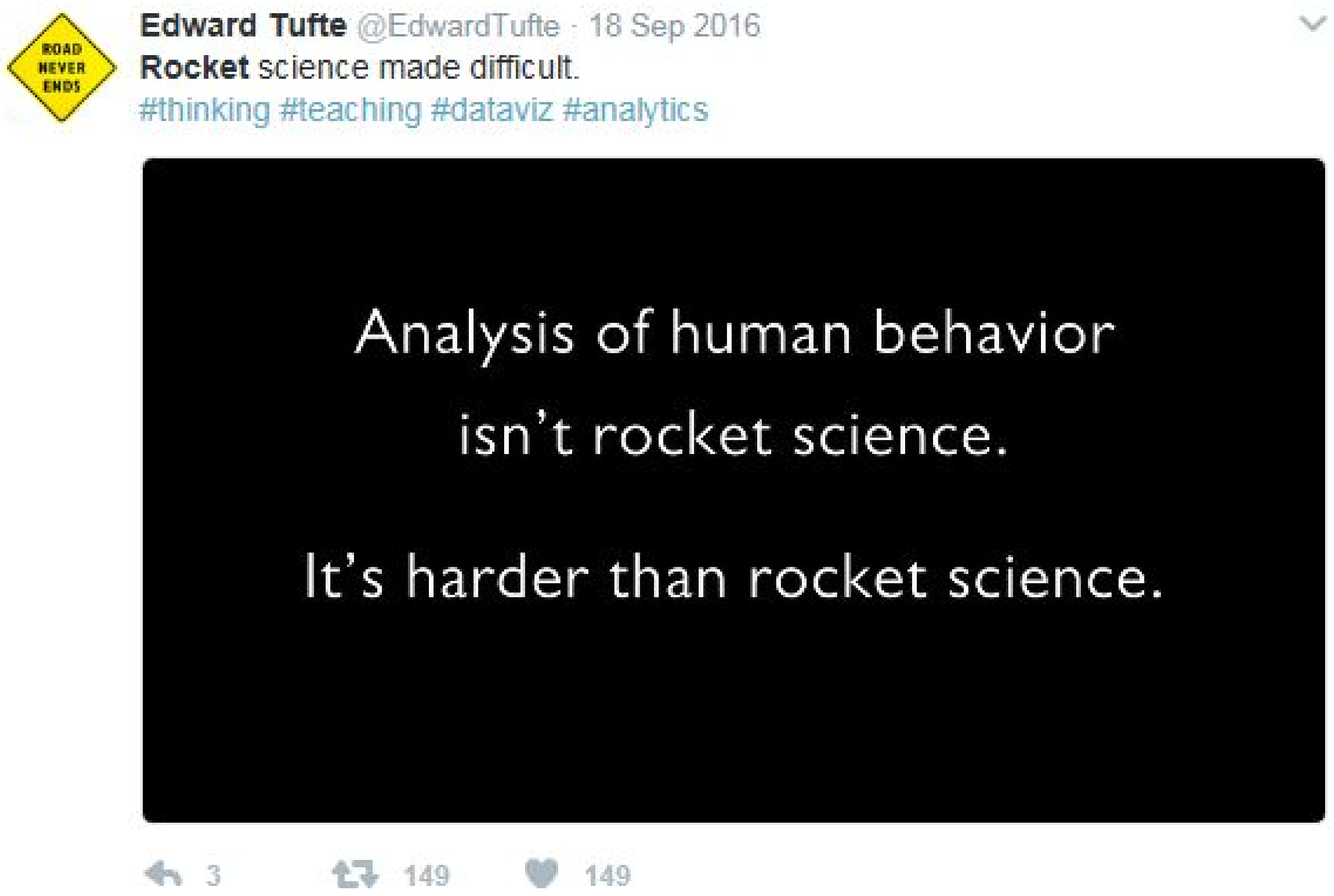
AFB International manufactures palatants for pet food, treats, and toys.

- http://afbinternational.com/research-and-development

- Palatability Assessment Resource Center

- Standard test: two bowls, two days

Which science is hardest?



Edward Tufte @EdwardTufte · 18 Sep 2016

Rocket science made difficult.

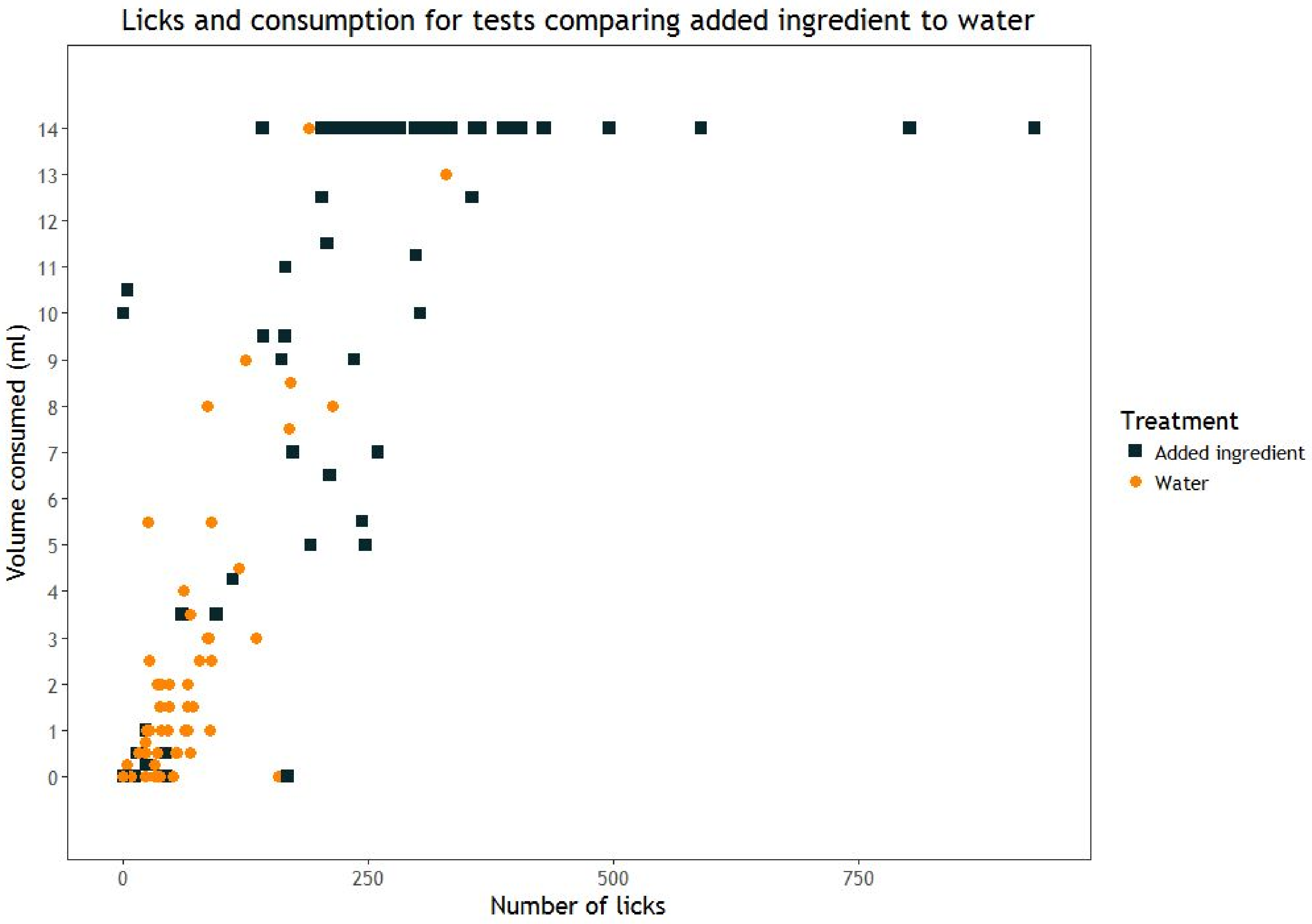
#thinking #teaching #dataviz #analytics

Analysis of human behavior
isn't rocket science.

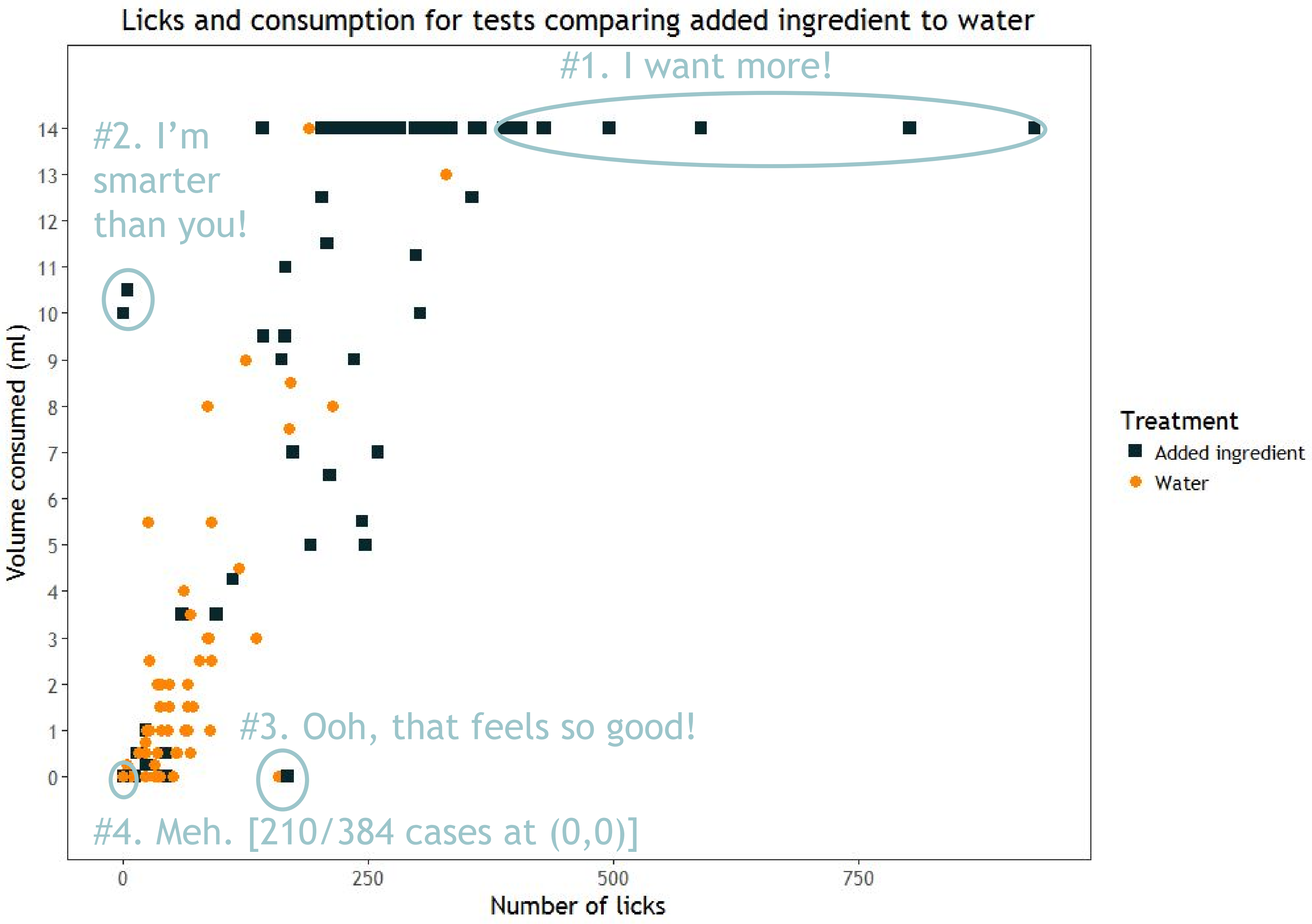
It's harder than rocket science.

3 149 149

Example - Lickometer study



Special cases



Licks and consumption for tests comparing added ingredient to water

#1. I want more!

#2. I'm smarter than you!

#3. Ooh, that feels so good!

#4. Meh. [210/384 cases at (0,0)]

Displaying extra zeroes

1. Use jittering to indicate clusters of points.

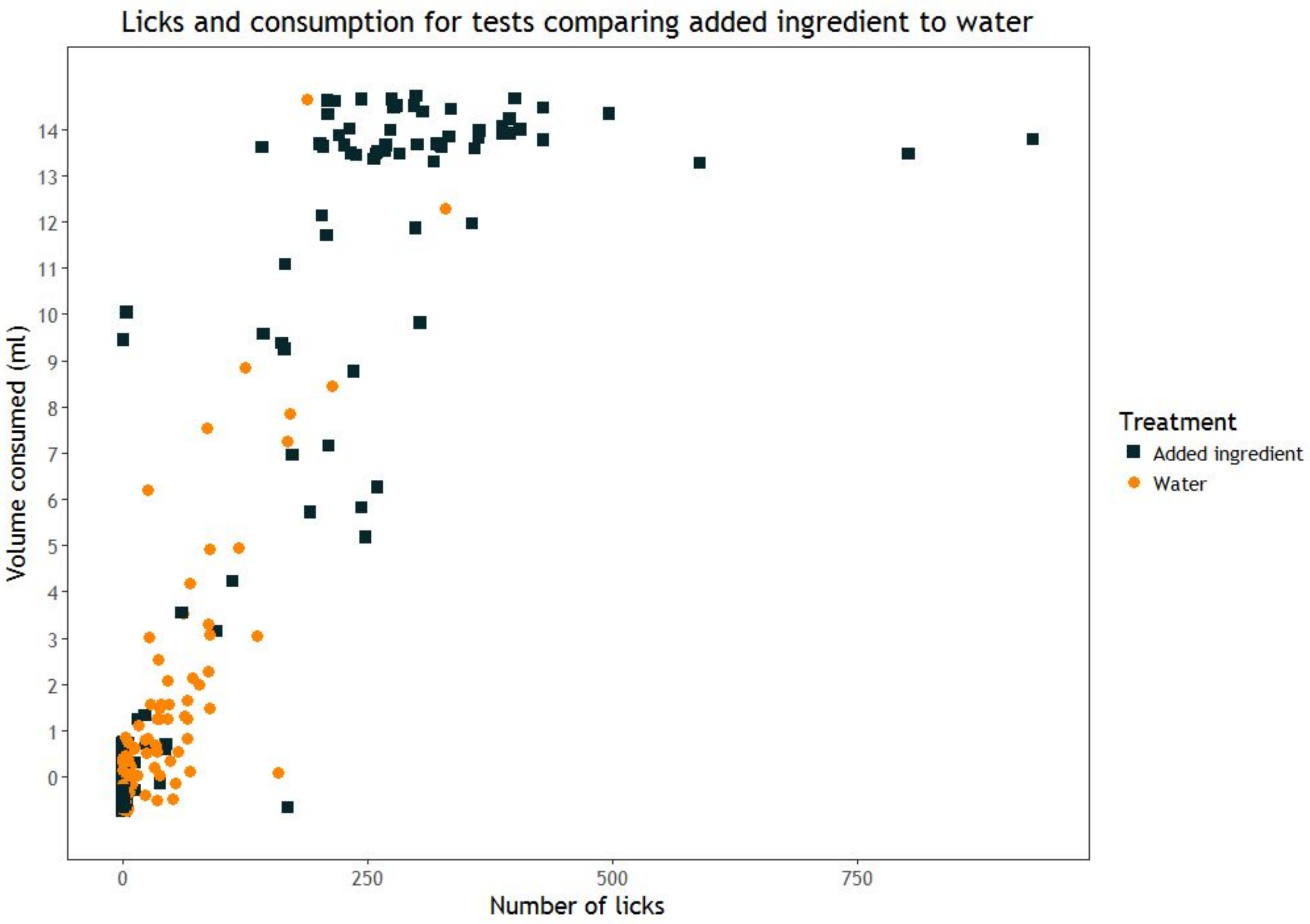
Data frame licks:

Test	Animal	Day	Ration	Licks	Consumed
<chr>	<chr>	<dbl>	<chr>	<dbl>	<dbl>
Test1	Cat1	1	A	236	9
Test1	Cat1	1	B	0	0
Test1	Cat1	2	A	211	6.5
Test1	Cat1	2	B	0	0
Test1	Cat2	1	A	0	0
Test1	Cat2	1	B	0	0
Test1	Cat2	2	A	256	14
Test1	Cat2	2	B	68	0.5

etc.

Displaying extra zeroes

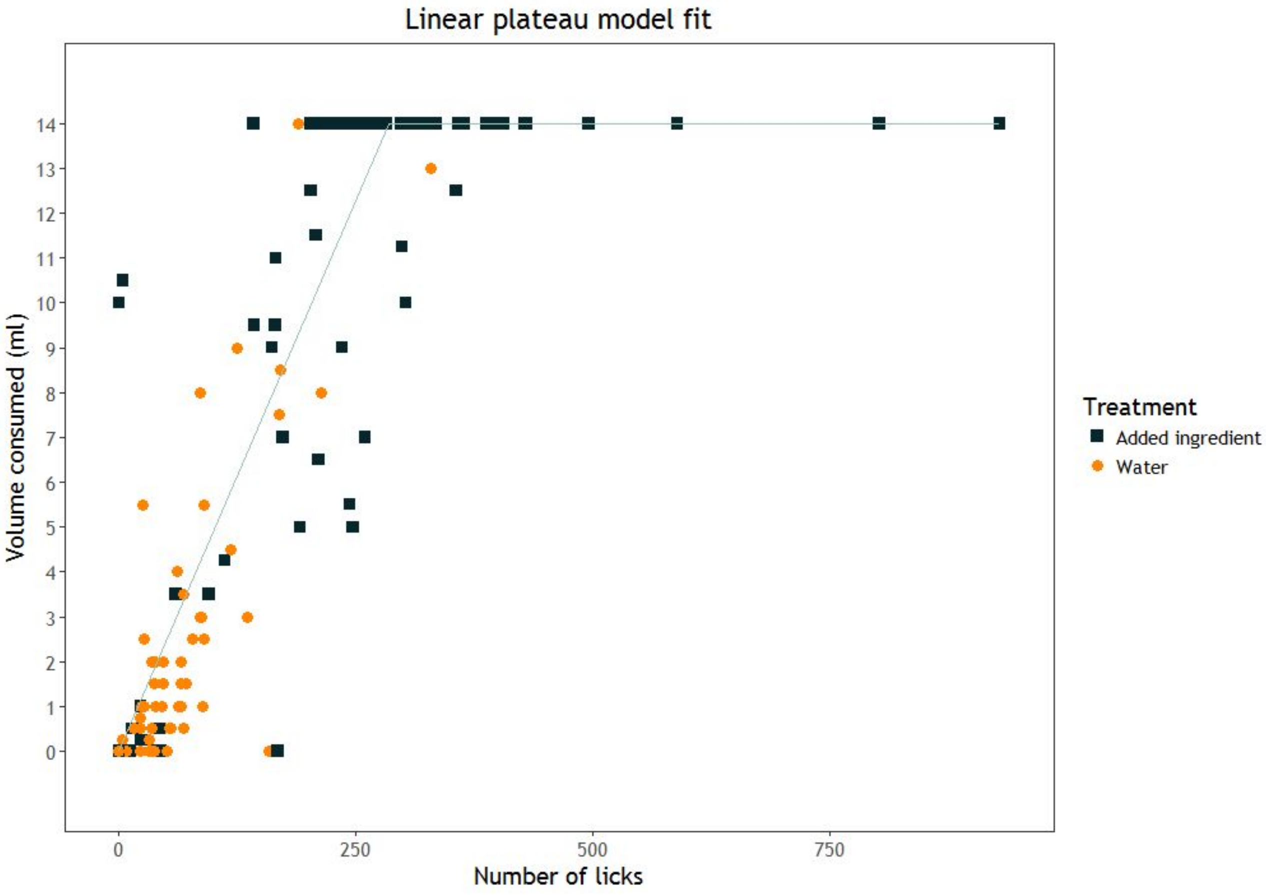
1. Use jittering to indicate clusters of points.



```
set.seed(79311)
library(ggplot2)

plot1 <- ggplot(licks,
aes(x=Licks,
y=Consumed)) +
geom_jitter(size=3,
height=.75,width=.75,
aes(shape=Ration,
colour=Ration))

plot1
```

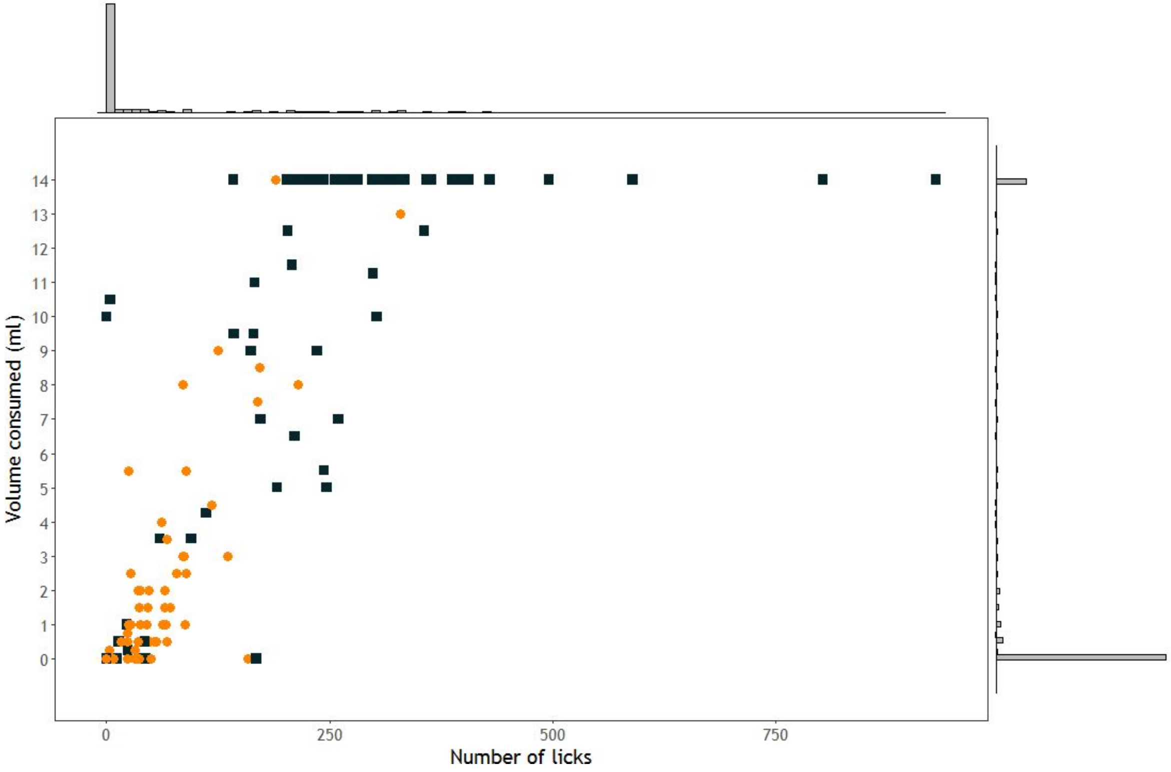
This illustrates density at the expense of accuracy.

2. Use a regression line that can be influenced by the zeroes to indicate the trend.



```
library(stats)

fornls <- licks %>%
select(Licks,
Consumed)

plateau <- nls(Consumed ~
(b*Licks)*(b*Licks<=14) +
14*(b*Licks>14),
data=fornls,
start=list(b=1))
splateau <-
summary(plateau)
joinpoint <- 14/
splateau$parameters[1,1]
```

3. Use marginal histograms in the ggExtra package to indicate large frequencies at (0,0) and at 14 ml.



```
library(ggExtra)

plot1b <- ggplot(licks,
aes(x=Licks,
y=Consumed)) + …

ggMarginal(plot1b,
type="histogram",
margins="both",
bins=100)
```

Reducing excessive zeroes

Can the numbers of zeroes be reduced?

Example: Use two-day feeding totals instead of single days

One or two zeroes do not cause excessive violations of ordinary statistical assumptions.

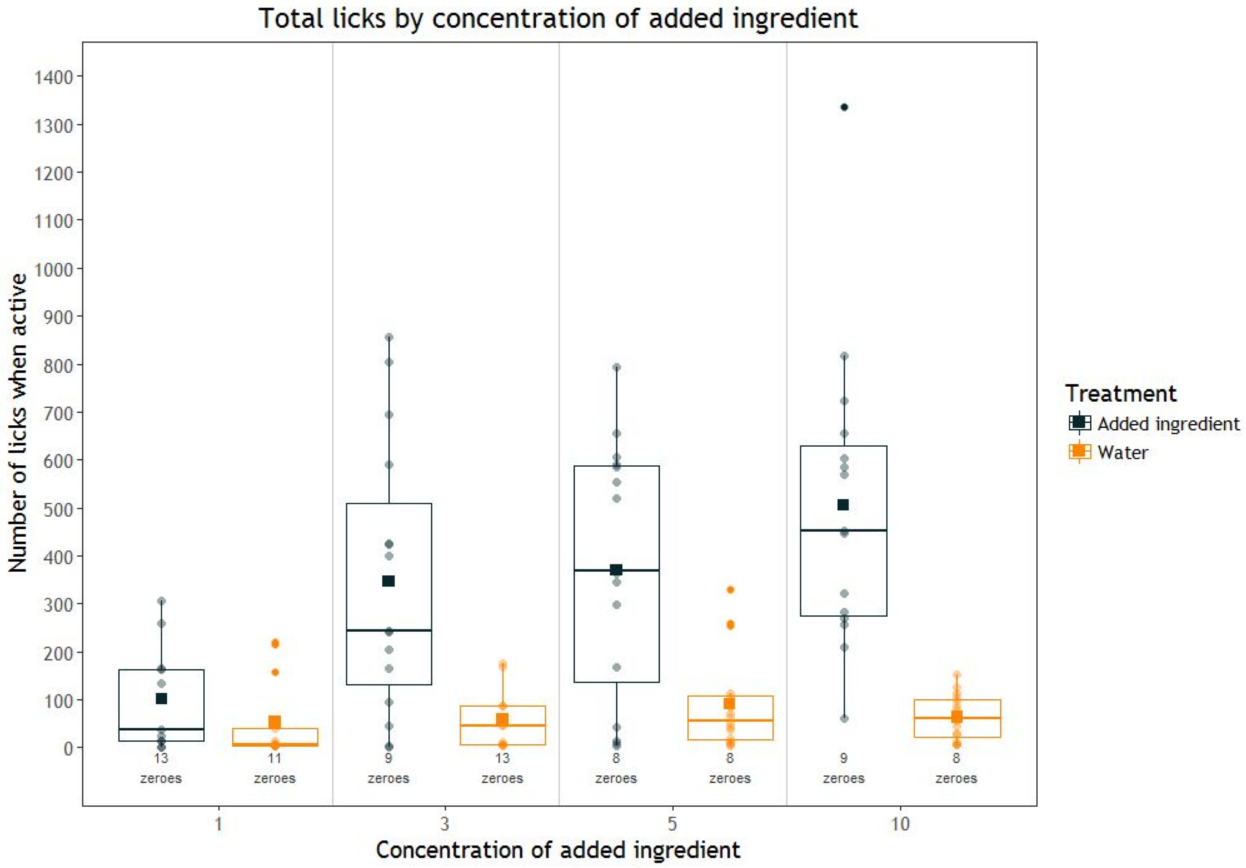
ZINB - Zero-inflated negative binomial model

Data frame licks2 counts the total number of licks over 2 days (Test.Conc identifies the pairing and the concentration)

```
Animal.     Test.Conc Ration Licktotal   xconc
<fctr>        <fctr>  <chr>      <dbl>   <dbl>
 Cat1        Test1.1      A          0     1
 Cat1        Test1.0      B          4     0

 Cat2        Test1.1      A          0     1
 Cat2        Test1.0      B          0     0

 Cat3        Test1.1      A         25     1
 Cat3        Test1.0      B          0     0
 etc.
```

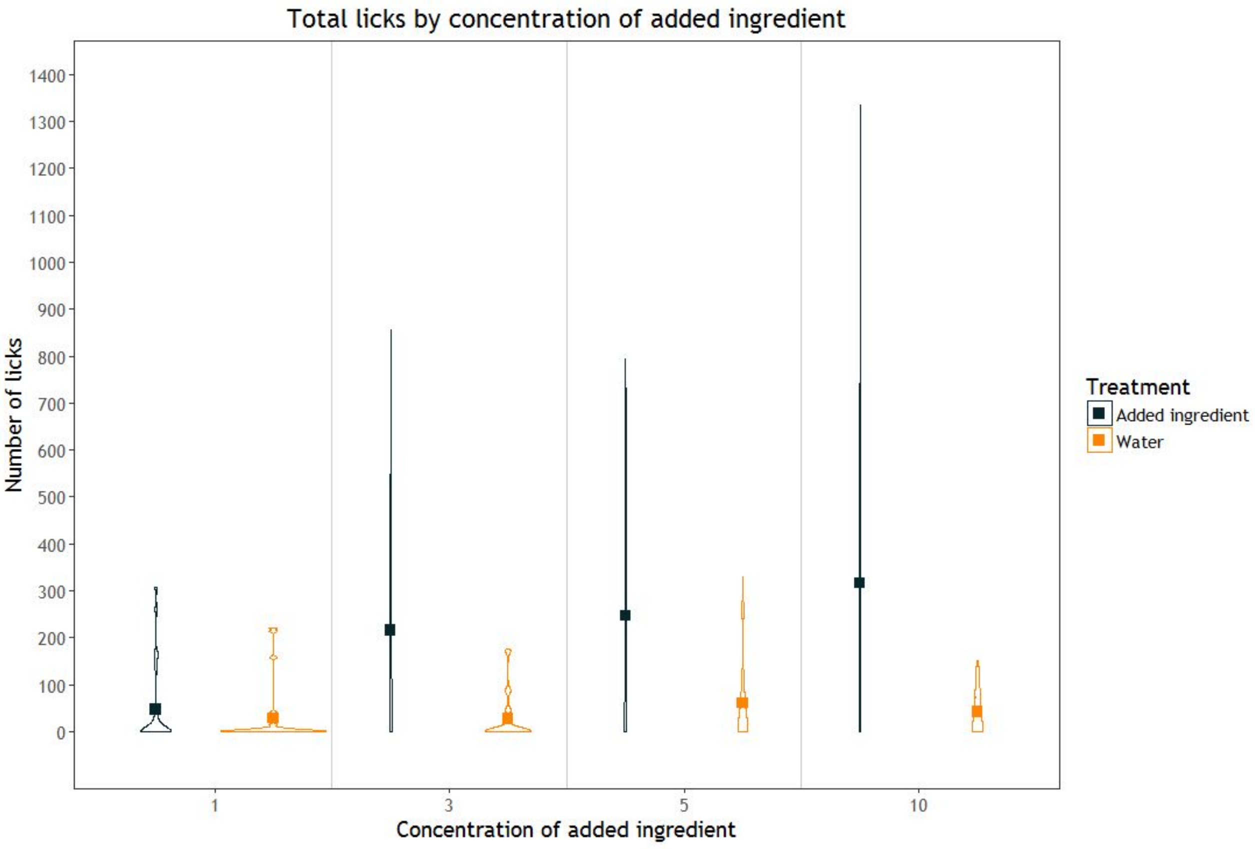
Displaying extra zeroes

4. Consider the total number of licks over 2 exposures to be the primary response of interest and display univariate data with annotated boxplots.



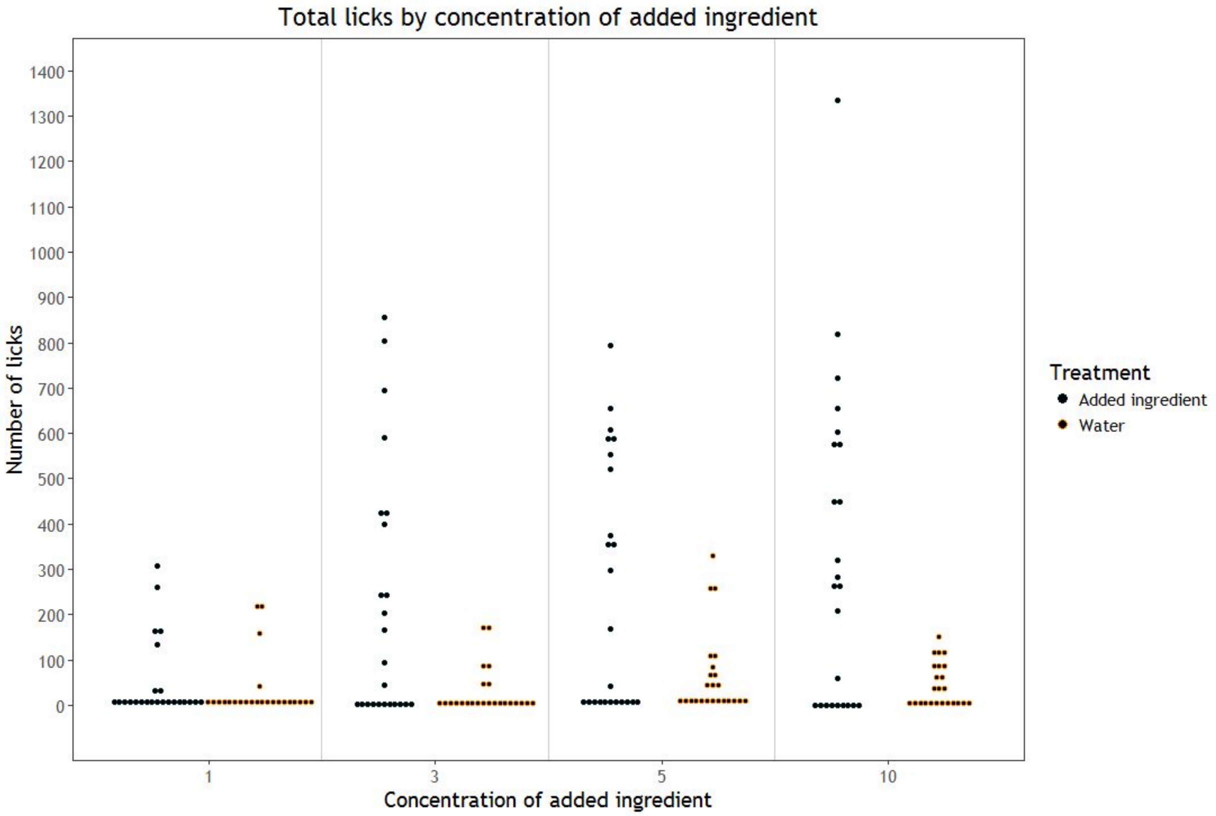
```
plot1 <- ggplot(licks2 %>%
filter(Licktotal>0),
aes(x=xconc,y=Licktotal,
colour=Ration)) +

geom_boxplot(position=
position_dodge(1)) +

geom_point(position=
position_dodge(1)) +

stat_summary(fun.y=mean,
geom="point") +

geom_text(data=lickdata,
aes(x=seq(.75,4.25,by=.5),
y=rep(-40,8)),
label=paste(lickdata$num0s,
"zeroes",sep="\n"))
```

5. Violin plots



```
plot2 <-ggplot(licks2,
aes(x=xconc,y=Licktotal,
colour=Ration)) +

geom_violin(position=
position_dodge(1)) +

stat_summary(fun.y=mean,
geom="point", shape=15,
size=3,position=
position_dodge(width=1))
```

6. Alternative: Stacked dot plots



```
plot3 <-ggplot(licks2,
aes(x=xconc,y=Licktotal,
colour=Ration)) +

geom_dotplot(binaxis='y',
stackdir='center',
dotsize=.6,
binwidth=20,
position="dodge")
```

Tabular summaries with excessive zeroes

A table of descriptive statistics was used to explicitly draw attention to the high numbers of zeroes.

Test	Ration	Concentration of added ingredient	No. of cats	No. of cats with 0 licks	% of cats with 0 licks	No. of cats with >0 licks	Avg. licks (when positive)	Min. no. of licks (when positive)	Median no. of licks (when positive)	Max. no. of licks (when positive)	Total no. of licks for all cats
1	A	1x	24	13	54	11	102	1	38	307	1121
1	B	0	24	11	46	13	53	3	6	221	686
2	A	3x	24	9	38	15	346	1	244	855	5191
2	B	0	24	13	54	11	58	3	45	175	641
3	A	5x	24	8	33	16	370	4	369	793	5923
3	B	0	24	8	33	16	91	3	56	329	1459
4	A	10x	24	9	38	15	506	60	452	1336	7583
4	B	0	24	8	33	16	63	6	62	152	1014

Note that Max > 2(Avg)

Tabular summaries with excessive zeroes

The table was generated using dplyr:

```
lickdata <- licks2 %>%
group_by(Test,Ration,xconc) %>%
summarize(num.cats=n(),num.zeroes=sum(Licktotal==0),
pct.zero=round(100*num.zeroes/num.cats),
num.positive=sum(Licktotal>0),
avg.positive=mean(Licktotal[Licktotal>0]),
min.licks=min(Licktotal[Licktotal>0]),
median.licks=median(Licktotal[Licktotal>0]),
max.licks=max(Licktotal[Licktotal>0]),
total.licks=sum(Licktotal))
```

Modeling data with excessive zeroes

- Hurdle models combine a left-truncated count component with a right-censored hurdle component
- Zeroes can only occur in the count component.

Example: A laboratory instrument reports 0 if the amount of a substance falls below a detection limit.

Modeling data with excessive zeroes

- Zero-inflated models combine a count component and a point mass at zero.
- Zeroes can arise in both situations.

Example: A participant may not be able to complete a task; or, if able, may choose not to do it.

Modeling data with excessive zeroes

Both hurdle and zero-inflated models are provided in the pscl package in R (Zeileis, Kleiber, and Jackman, “Regression Models for Count Data in R,” available at cran.r-project.org/web/packages/pscl/vignettes/countreg.pdf).

Other applications of these models include:

- econometrics
- political science (PSCL @ Stanford)
- agriculture

Models for count portion of a zero-inflated model

1. Poisson
   a) Traditional application: Number of occurrences of an event per unit of time or space
   b) Variance = mean

Models for count portion of a zero-inflated model

2. Negative binomial
    a) Traditional application: Number of binomial trials needed to achieve a specified number of successes
    b) Variance > mean
    c) With two parameters for mean and variance, and multiple equivalent parameterizations for a negative binomial distribution, results are harder to interpret

ZINB - Zero-inflated negative binomial model

Data frame licks2 (Test.Conc identifies the pairing and the concentration)

```
Animal.    Test.Conc Ration Licktotal
<fctr>       <fctr>  <chr>     <dbl>
 Cat1       Test1.1      A         0
 Cat1       Test1.0      B         4

 Cat2       Test1.1      A         0
 Cat2       Test1.0      B         0

 Cat3       Test1.1      A        25
 Cat3       Test1.0      B         0
 etc.
```

ZINB - Zero-inflated negative binomial model

Model in R (zeroinfl is in the pscl library)

Count data

is conditional, based on

Participation

```
zinb0 <- zeroinfl(Licktotal~Test.Conc + Animal.|Test.Conc,
data=licks2, dist="negbin“,link=“logit”)
```

The chance that a cat will “decide to participate” is assumed to be related to the ingredients in the bottles, using (by default) a logit link.

Given that a cat chooses to participate, the mean number of licks is related to an ingredient effect and a cat effect via a log link.

ZINB - Zero-inflated negative binomial model

```
> summary(zinb0)
Pearson residuals:
      Min        1Q    Median        3Q       Max
-1.003604 -0.654252 -0.001159  0.209992  4.304403

Count model coefficients (negbin with log link):
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)       -0.5302     0.6813  -0.778 0.436481
Test.concTest1.0  -0.1333     0.4449  -0.300 0.764509
Test.concTest0.0   0.1343     0.3999   0.336 0.737021
...
Animal.Cat1        4.4654     0.7964   5.607 2.06e-08 ***
Animal.Cat2        4.2511     0.7805   5.447 5.13e-08 ***
...
Log(theta)         0.1848     0.1526   1.211 0.226046

Zero-inflation model coefficients (binomial with logit link):
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)        -0.8903     0.5348  -1.665   0.0959 .
Test.concTest1.0    0.3002     0.7527   0.399   0.6900
Test.concTest1.1  -17.0435  3440.1875  -0.005   0.9960
...
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Theta = 1.2029
Number of iterations in BFGS optimization: 57
Log-likelihood: -710.2 on 40 Df
```

ZINB - Zero-inflated negative binomial model

```
> names(zinb0)
"coefficients"  "residuals"     "fitted.values" "optim"         "method"
"control"       "start"         "weights"       "offset"        "n"
"df.null"       "df.residual"   "terms"         "theta"         "SE.logtheta"
"loglik"        "vcov"          "dist"          "link"          "linkinv"
"converged"     "call"          "formula"       "levels"        "contrasts"
"model"         "y"
```

ZINB - Zero-inflated negative binomial model

```
zinb0 <- zeroinfl(Licktotal~Test.Conc + Animal.|Test.Conc,
data=licks2, dist="negbin“, link=“logit”)
```

```
zinb1<- zeroinfl(Licktotal~Test.Conc + Animal.|1,
data=licks2, dist="negbin“, link=“logit”)
```

```
vuong(zinb0,zinb1) #From package pscl
lrtest(zinb0,zinb1) #From package lmtest
```

If there was not a significant feature (such as color or scent) that attracted the cats to one bottle instead of the other, then the Test.Conc term on the right can be a simple intercept (1).

ZINB - Zero-inflated negative binomial model

```
zinb0 <- zeroinfl(Licktotal~
Test.Conc + Animal.|
Test.Conc,
data=licks2, dist="negbin“, link=“logit”)

zinb2<- zeroinfl(Licktotal~
Test.Conc + Animal.|
Test.Conc + Animal.,
data=licks2, dist="negbin“, link=“logit”)
```

The second model did not converge, because some cats never registered a lick.

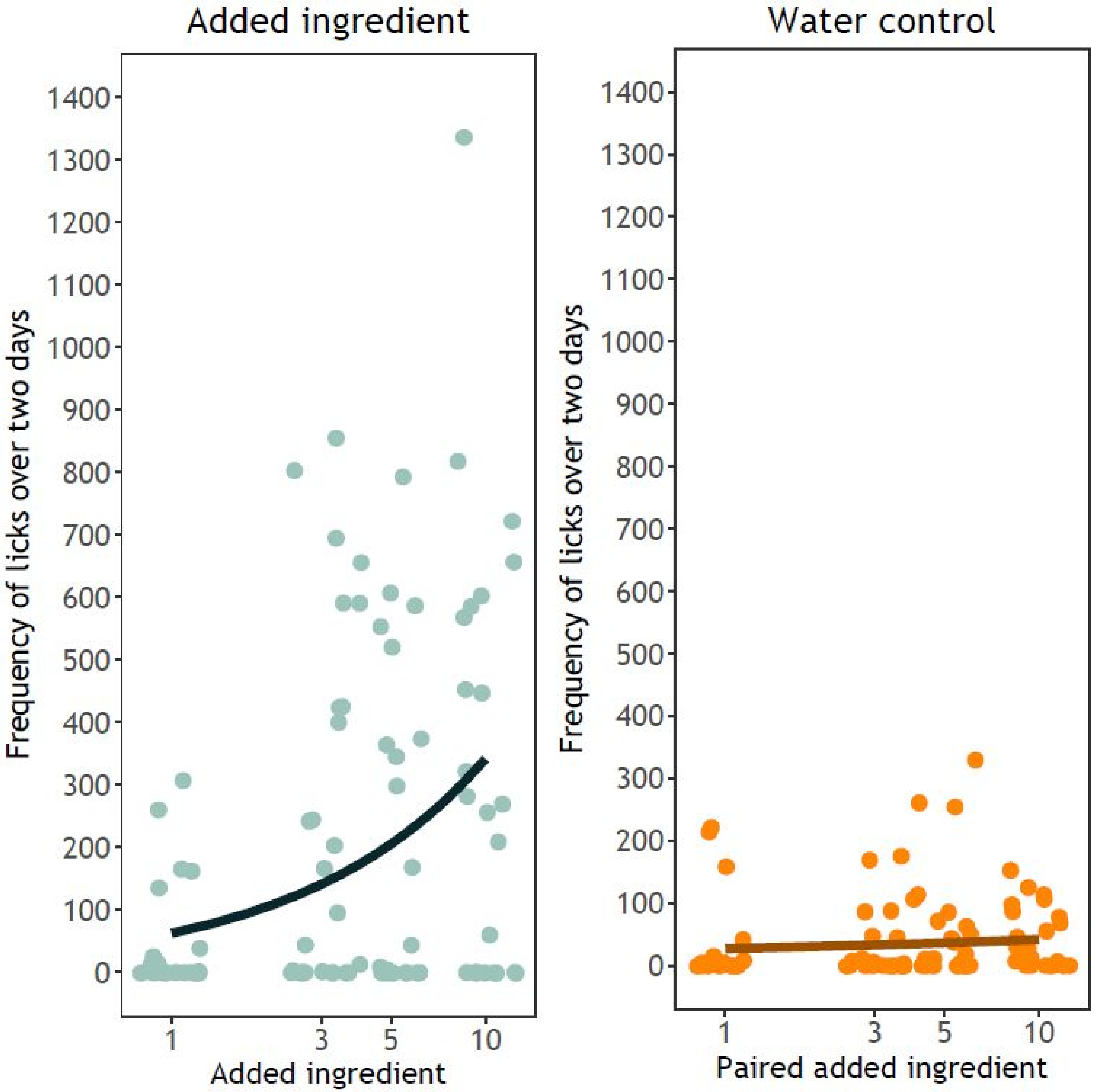
ZINB - Zero-inflated negative binomial model

```
zinb0 <- zeroinfl(Licktotal~Test.Conc +
Animal.|Test.Conc, data=licks2, dist="negbin“,
link=“logit”)
zinb3 <- zeroinfl(Licktotal~1 + controlintercept +
logConc + controlslope + Animal.|1,
data=licks2, dist="negbin“, link=“logit”)
```

Idea: Replace the fixed effect of the different concentrations (1x, 3x, 5x, and 10x) with a slope with covariate values of log(1), log(3), log(5), and log(10) to obtain a dose-response relationship.

ZINB - Zero-inflated negative binomial model

Overlays with regression lines worked with geom_jitter but not with geom_dotplot.

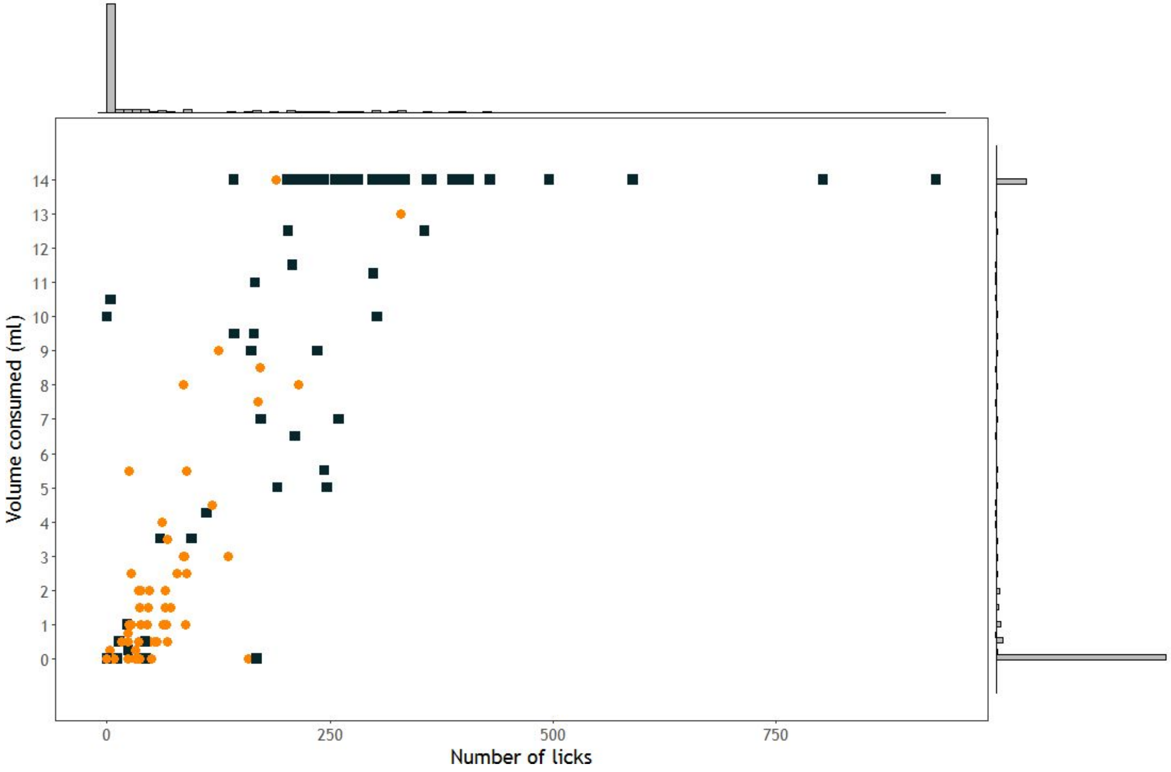


ZINB - Zero-inflated negative binomial model

ANOVA-like comparisons of treatments can be performed using the delta method with the model estimates and their variance-covariance matrix.

R packages msm and car have delta method approximations that require entry of a function to estimate, but no derivatives are needed.

Two inflated points



What if we want to model the volume consumed instead of the number of licks?

There are many zeroes and many ones.

ZOIB - Zero-one inflated beta distribution

First, convert the volume consumed (0 ml to 14 ml) to a fraction consumed (0 to 1).

The R package zoib fits models using a beta distribution with inflation at 0, at 1, or both.

ZOIB - Zero-one inflated beta distribution

- Instead of a “count” side and a “participation” side, the ZOIB model can include up to 5 different sections for the mean and shape of the beta distribution, zero and one inflation, and random effects
- Uses Bayesian estimation with Gibbs sampling via JAGS

Convergence may be difficult.

Kumaraswamy? $f(x|a,b) = abx^{a-1}(1 - x^{a})^{b-1}$

Aroma tests

Cats are given two minutes to explore two perforated cans that contain different aromatic agents.

Responses of interest: Lengths of time that the cat sniffs Aromas A and B.

For some length of time, the cat will not be sniffing either container. This has ranged from 44 seconds (37%) to 120 seconds (100%).

Analogy to industrial mixture experiments

See references from John Cornell

In a 3-component mixture experiment:

- Three components A, B, and C are blended in specified proportions
- These amounts are linearly dependent:

  $A + B + C = 1$
- Additionally, there may be other constraints, e.g.

  $A < .1$ and

  $A + B < .3$
- Goal: Optimize some feature of interest (taste, viscosity)

Ternary plot

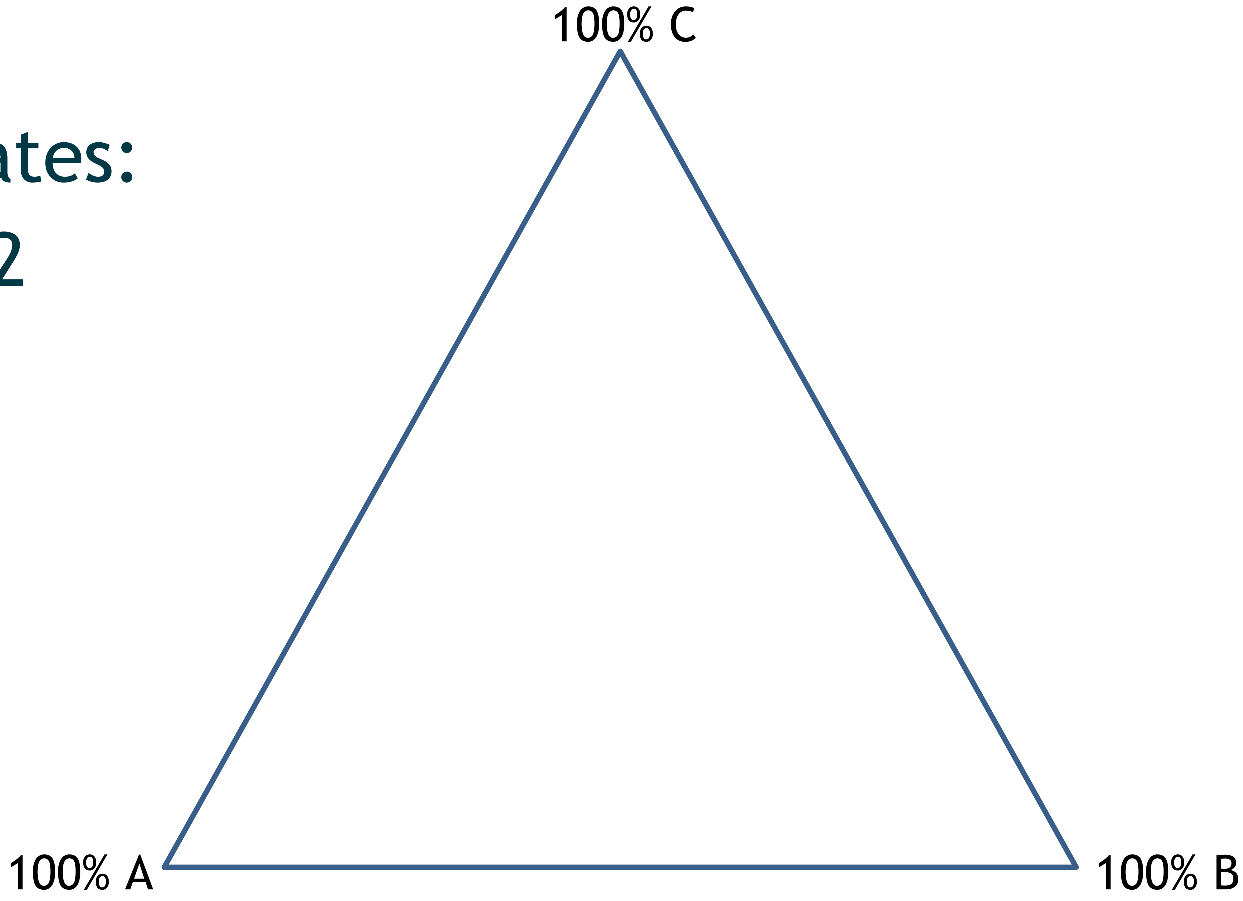
A 3-component mixture can be represented with an equilateral triangle in 2 dimensions, since

$A + B + C = 1$

New coordinates:

$X = (2B + C)/2$

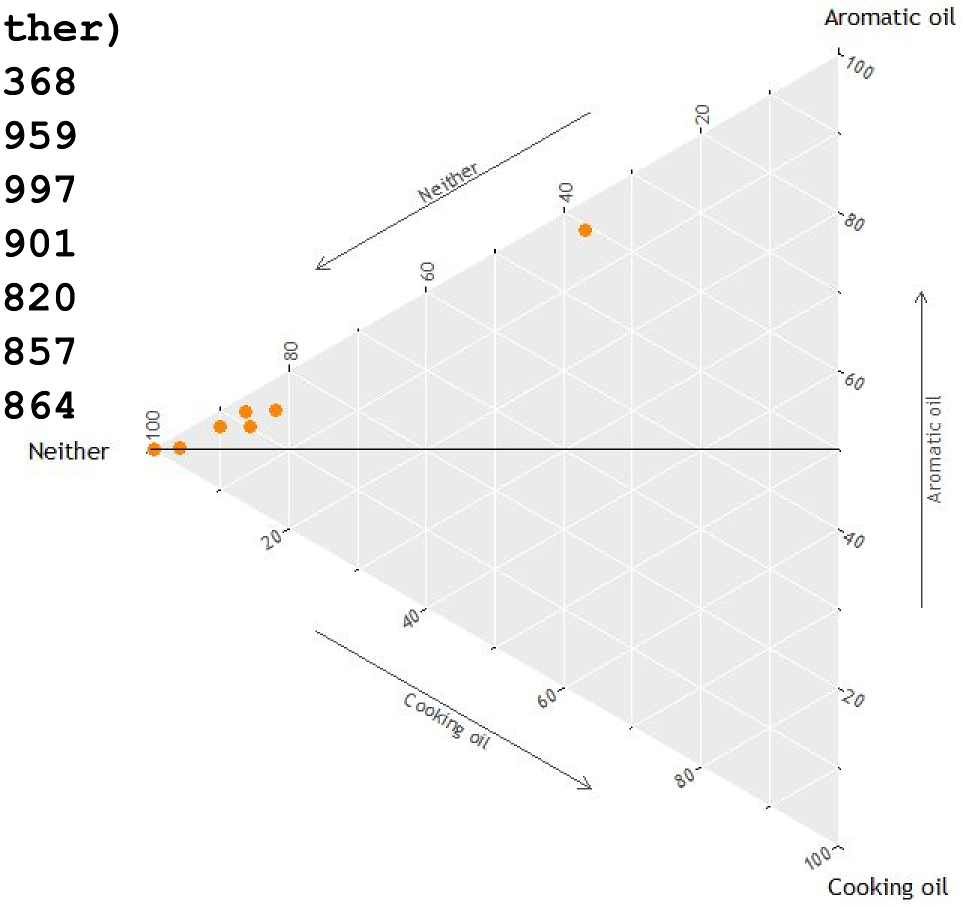
$Y = (C\sqrt{3})/2$



Ternary plot example

Data frame catdata:

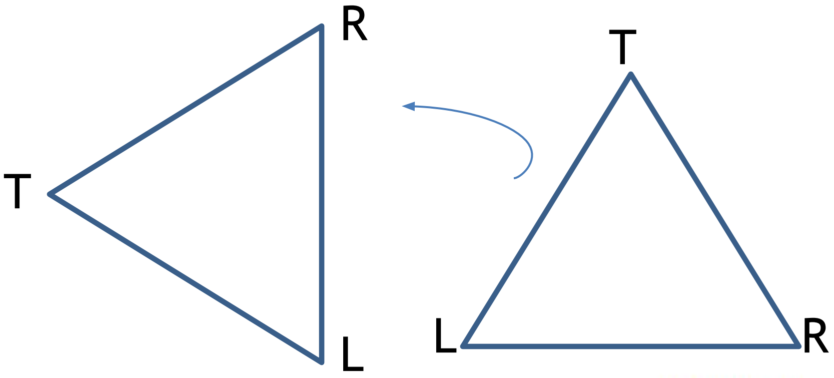
	Cooking	Aromatic	NS
Cat	(oil)	(oil)	(Neither)
1	0.038	0.594	0.368
2	0.018	0.023	0.959
3	0.000	0.003	0.997
4	0.019	0.080	0.901
5	0.039	0.141	0.820
6	0.042	0.100	0.857
7	0.020	0.116	0.864

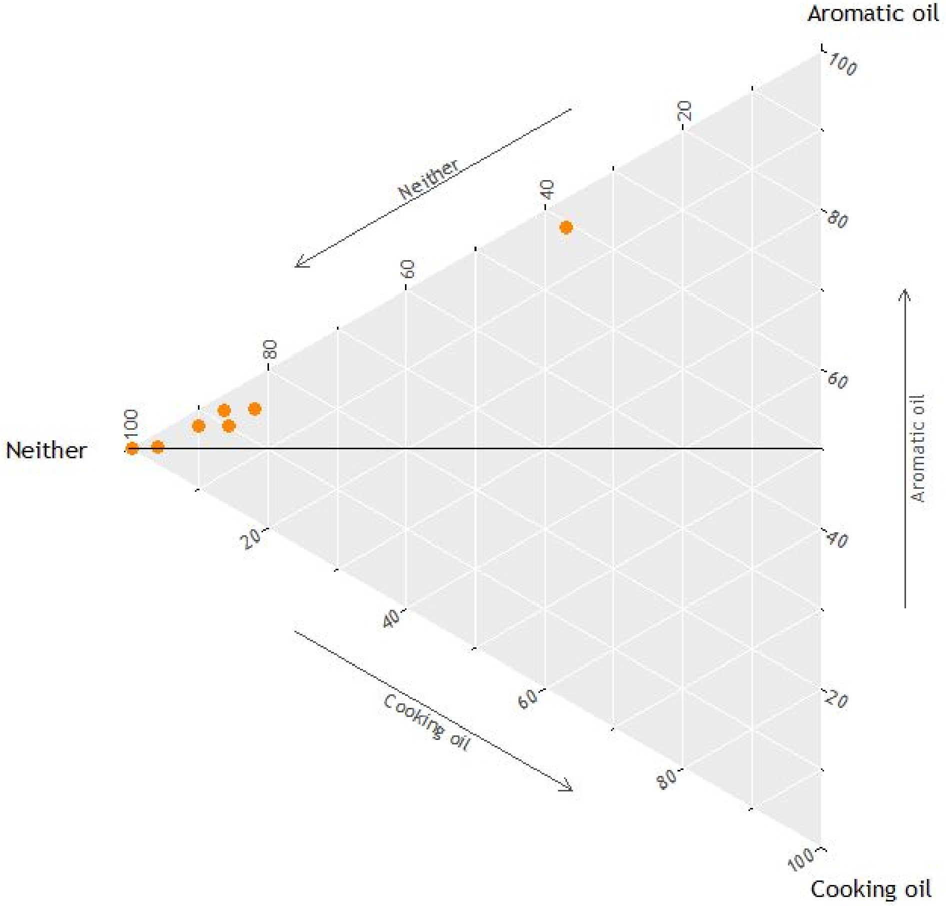


Ternary plot example

The R package ggtern makes ternary plots using ggplot2-like syntax.

```
triplot <- ggtern(data=catdata,
aes(x=Cooking, y=NS, z=Aromatic) +
geom_Tisoprop(value=0.5) +

Tlab("Neither",
labelarrow="Neither") +
Llab("Cooking oil",
labelarrow=“Cooking oil") +
Rlab("Aromatic oil",
labelarrow=“Aromatic oil") +
theme_gray() +
theme_showarrows() +
theme_nomask() +
theme_rotate(degrees=90)
```

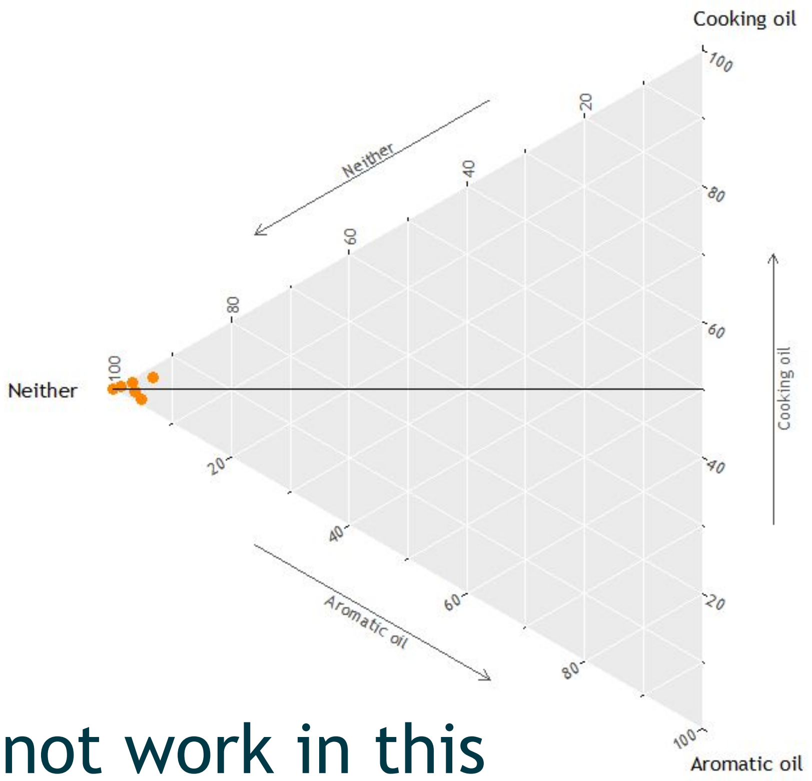




Multiple zeroes in a ternary plot

Data from another trial:

Cat	Cooking oil	Aromatic oil	Neither
1	0.010	0.002	0.987
2	0.027	0.006	0.967
3	0.000	0.000	1.000
4	0.009	0.038	0.952
5	0.000	0.000	1.000
6	0.051	0.015	0.934
7	0.015	0.021	0.965



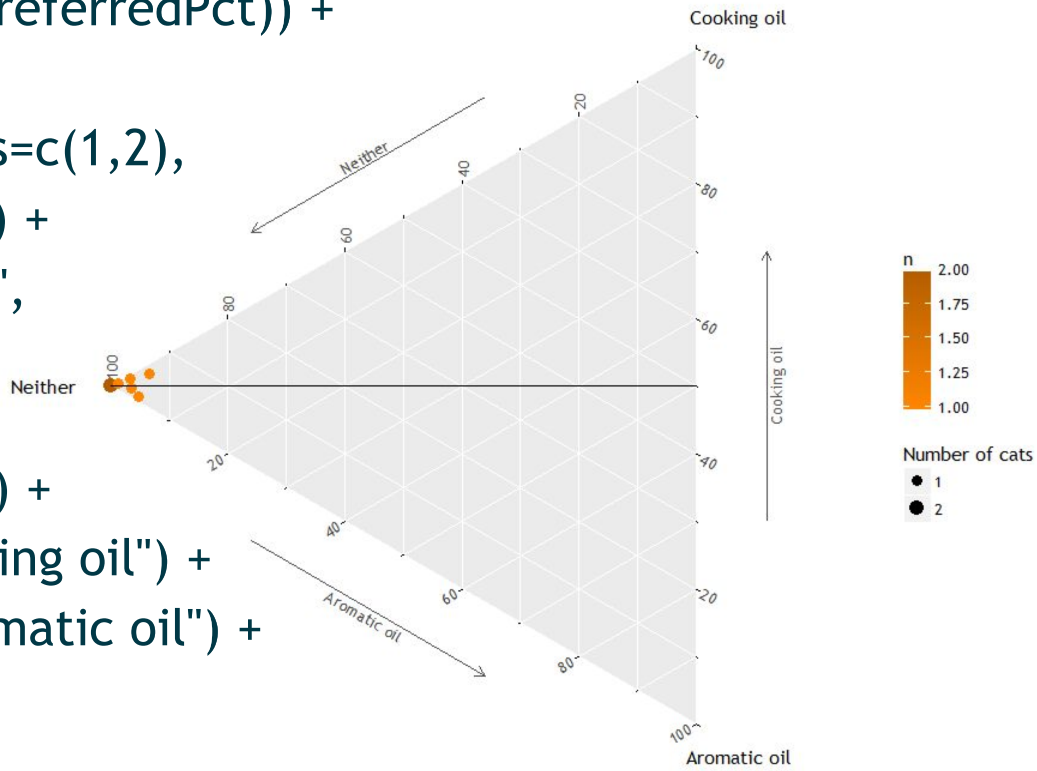
Jittering, violin plots, etc. do not work in this coordinate system

Zooming in the left vertex changes the “room map”

Contour and density plots try to interpolate outside the region where cats actually had data

Multiple zeroes in a ternary plot

```
triplot2 <- ggtern(data=catdata,
aes(x=NotPreferredPct,y=NSPct,z=PreferredPct)) +
geom_count(aes(colour=..n..)) +
scale_size_area(limits=c(1,2),breaks=c(1,2),
name="Number of cats",max_size=4) +
scale_colour_gradient(low="#ff8500",
high="#b35d00") +
geom_Tisoprop(value=0.5) +
Tlab("Neither",labelarrow="Neither") +
Rlab("Cooking oil",labelarrow="Cooking oil") +
Llab("Aromatic oil",labelarrow="Aromatic oil") +
theme_gray(base_size = 12) +
theme_showarrows() +
theme_nomask() +
theme_rotate(degrees=90)
```



Thank you!

Jay Harrison
jharrison@afbinternational.com

AFB International®