NHH

INSTITUTT FOR FORETAKSØKONOMI

DEPARTMENT OF BUSINESS AND MANAGEMENT SCIENCE

FOR 12 2018

ISSN: 1500-4066
September 2018

Discussion paper

Sample statistics as convincing evidence: A tax fraud case

BY
Jostein Lillestøl

Norges Handelshøyskole

NORWEGIAN SCHOOL OF ECONOMICS

Sample statistics as convincing evidence: A tax fraud case

by[1]

Jostein Lillestøl

Department of Business and Management Science
Norwegian School of Economics (NHH), Bergen, Norway

June 19. 2018

Summary

This report deals with the analysis of data used by tax officers to support their claim of tax fraud at a pizzeria. The possibilities of embezzlement under study are overreporting of take-away sales and underreporting of cash payments. Several modelling approaches are explored, ranging from simple well-known methods to presumably more precise tools. More specifically, we contrast common methods based on normal assumptions and models based on Gamma-assumptions. For the latter, both maximum likelihood and Bayesian approaches are covered. Several criteria for the choice of method in practice are discussed, among them, how easy the method is to understand, justify and communicate to the parties. Some dilemmas present itself: the choice of statistical method, its role in building the evidence, the choice of risk factor, the application of legal principles like "clear and convincing evidence" and "beyond reasonable doubt". The insights gained may be useful for both tax officers and defenders of the taxpayer, as well as for expert witnesses.

[1] This work is part of project at Norwegian Centre for Taxation (NoCeT) at NHH.

1. The context

Mr. Way runs a small pizzeria, here named Take-A-Way pizzeria, established by him in 2012, and run all by himself. Being an immigrant, coming from a country with less strict business regulations, he faced many challenges. He had to handle everything from procurement, food preparation, serving and accounting. He faced tough competition and, and had to work long hours most evenings, to pick up occasional customers. With no accounting experience, Mr. Way felt that the reporting for tax purposes was a hassle.

At the end of the second year, the tax office was dissatisfied with his reporting, and asked for detailed accounts of the daily sales, which he could not provide in full for 2012 and 2013. After some guidance, he felt that he was on track, and from now on, he filed his income statements on time and tried to keep backing documentation, as best he could. However, during the year 2014, the tax office became suspicious to the truthfulness of his income statements. At the end of the year, the trouble was mounting for Mr. Way, and he faced action from the tax office, based on a report claiming lousy bookkeeping and misreported sales. Their action included reassessment of the tax for the prior years 2012 and 2013 and, in the worst case, a penalty tax on top of this.

The basis the claim from the tax authorities was essentially

A. Missing documentation according to tax law (cash register rolls etc.)
B. Mismatch between reported procurement and sales (beverages, flour etc.)
C. Direct observations at Take-A-Way pizzeria made by tax officers in 2014
    - Claimed mismatch between observed and reported sales

Item A mainly affected the credibility of the taxpayer, for the years 2012-2013, and partly 2014. So did item B, since efforts by Mr. Way to explain the mismatches were dismissed by the tax officers, and he had insufficient documentation to support himself. Some irregularities in the years prior to 2014 were admitted, and corrections accepted. Item C was the main basis for reassessment of the tax was sales amounts and type of sales, inferred from direct observations at the pizzeria on several occasions in 2014. The findings were then applied to the reassessment of the tax for 2012 and 2013, the argument being that, given the irregularities found for A and B for these years, one could assign little credibility to the numbers related to C, as well. Item C leads to statistical issues as well as legal issues, related to the combination and weighing of evidence. The tax office needs practicable procedures, but may be challenged by statisticians as well as lawyers. We will see how tax officers may face some of these challenges, in light of current tax regulations.

The main opportunities for embezzlement, possible to uncover by C, are as follows:

In this business, some sales are “For Here” and some sales are “To Go”, also named “Take-Away”. The former includes serving, and incurs a higher value added tax (VAT) than the latter. In this case, the value added tax is 25% for “For Here” sales and 15% for “To Go” sales. The value-added tax assigned for a specific year is based on the total sales for the year and the share of each of the two types of sales, calculated as an average of daily shares over the

year. This may offer an opportunity to misreport the shares and pocket the difference. In addition, the customers may pay cash or by card. Cash payment may offer the opportunity not to record the sale at all. It should be noted that the use of cash was on steadily decline during these years, which may add uncertainty to actual customer behavior.

In general, let us see how the tax officers may proceed to uncover possible tax evasion of this kind, and subsequently prepare a case against the taxable. Roughly, the tax officers follow a four-stage process, which can end at any of the stages

1. Examine reporting from taxable -> Suspicion raised
2. Uncover and document credibility -> Go forward, if not credible
3. Uncover and document size of evasion -> Reassessment of taxes
4. Judge graveness -> Impose penalty

The legal framework is the Taxes Management Act of 2016, in operation from January 2017. The law states the obligation of the taxable to give correct and complete information in the tax return (§8). The tax authority may change the tax determination whenever the determination is wrong (§12-1), and discretionary assessment of the factual basis can be made, whenever the taxable own statements do not provide a proper basis for the tax assignment (§12-2 (1)). The discretionary assessment shall appear correct, in view of the available information (§12-2 (2)). Penalty tax is given to a taxable who has given wrong or incomplete information leading to tax advantages (§14-3). If the act is willful or grossly negligent, the reaction is sharpened (§14-6).

The Taxes Management Act is general, and additional guidance is needed for practical use. In particular, this is crucial for the last two stages above, when decisions made by the tax authority may be appealed and possibly lead to court proceedings. Some issues are clarified in regulations made to accompany the law, and they may be modified and supplemented over time. In some cases, advice may come from the preparatory material to the law, and from recent Supreme Court rulings based on the law.

It may be useful to refer to some common law principles[2]:

I. Proven “by a preponderance of the evidence”, that is “more probable than not”[3]
II. Proven “by clear and convincing evidence”, that is “clearly more probable”
III. Proven “beyond reasonable doubt”, that is “almost sure”

At the first stages, principle I may be sufficient to go ahead, while principle II is required to justify discretionary reassessment. This assessment shall establish the factual basis, and have to stay close to principle I, which corresponds to “the most likely state”. To impose penalty tax, a proof like II is necessary and sufficient, but the sharpened penalty requires principle III. Further details on these issues are in Appendix B.

[2] The Norwegian phrasing is: I=“Alminnelig sannsynlighetsovervekt” and II=”Klar sannsynlighetsovervekt”.
[3] Another notion for this in English is “on the balance of probabilities”.

2. The uncovering

The case of Take-A-Way involves all stages 1-4, and the use of direct observations as part evidence (item C above) and basis for reassessment of taxes. The observations were taken at the pizzeria by tax officers, pretending to be customers. On a visit in February 2014, they observed some irregularities. Among others, they paid cash, but the reported total cash payment for the day was less than their single payment. This raised their suspicion against the reporting practice, and they decided on a revisit later on to gather data. This happened at three selected dates, two in May and one in June 2014. On these occasions they observed a higher fraction of cash payments than was reported for the two prior years (which were less than 10%), and a lower fraction of take-away sales than was reported the prior years (which were more than 95%). The results from the three visits as given in the tax examiner report are in Table 1, together with the average of the three numbers, and a downgraded average to be explained later. In addition, we list the yearly average, as reported at the end of the year 2014. [4] The report states that, in order to have practicable routines, it is justified to base the reassessment on average considerations. No measure of variation is given.

Date	#Sales	Sales NOK	Cash%	ForHere%
14.05.2014	10	2502	38.7	26.7
25.05.2014	34	7276	26.7	15.0
18.06.2014	23	4557	15.3	49.9
Average of sample	22	4778	26.9	30.5
Downgraded Average	-	-	21.0	25.0
Reported Average	-	5279	12.7	10.3

Table 1. Reported condensed findings from three visits

The sample averages based on the three observations are considerably higher than the averages reported by Mr. Way the preceding years, and at the end of the year 2014. The latter are unknown at the time of the visit. The tax officers may of course possess some background knowledge of typical shares of the different types of sales in this business. During this period, the use of cash was steadily declining, but supposedly this was taken into account in their judgement. The numbers reported at the end of the year 2014 turned out somewhat less alarming than those of the preceding two years. This may have a variety of possible explanations in addition to pure chance: change in customer behavior and/or improved reporting practice, for example due to better knowledge or being more on alert.

It is clear that a sample introduce randomness, and that the result may have been entirely different, if some other dates were chosen. However, the report states that a sample of n=3 is representative for drawing conclusions about the true yearly average, but makes no efforts to judge the uncertainty in probability terms. However, the tax examiners are seemingly aware of the risk of misjudgment, since they downgrade their numbers, the cash-share from 26.9% to 21.0% and the share consumed at the premises from 30.5% to 25.0%.

[4] The transactions made by the tax officers themselves are taken out.

No principle is stated for the size of downgrading, and it is at best discretionary based on some additional insight or established practice.

3. The data on record

In order to understand the role of sample variation we may look at the record of daily data. We have three variables Sales (in NOK), ForHere%, and Cash%, from which we also have the variables ToGo% and Card%. This data are available from end of day reports (Z-reports), for each year of operation 2012-2014, but we focus on the last year 2014. The Z-report may also provide more detailed sales data, e.g. sales for beverages and food separately.

A time series plot of consecutive daily sales (the variable Sales) is given in Figure 1. The plot shows that the daily sales vary considerably. This may be due to both seasonality and random variation. We may have seasonality over the week and over the year. There may be random variation of different kinds, due to weather conditions a specific day or pure randomness. In addition, the sales may be affected by special events that bring more or less people downtown. Promotion campaigns are also a possibility.

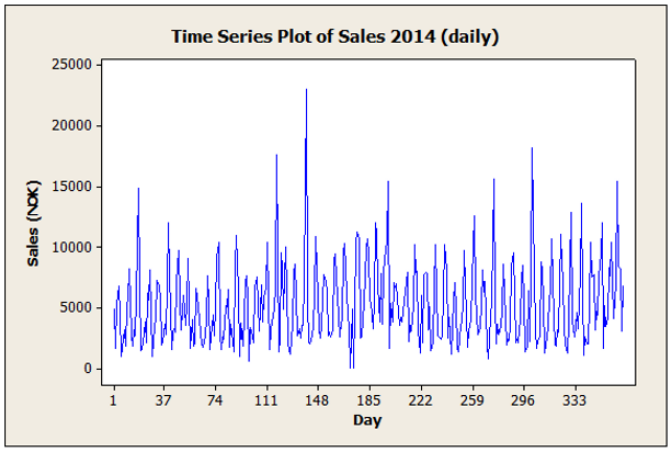

Figure 1. Time series plot of daily sales (NOK) in 2014

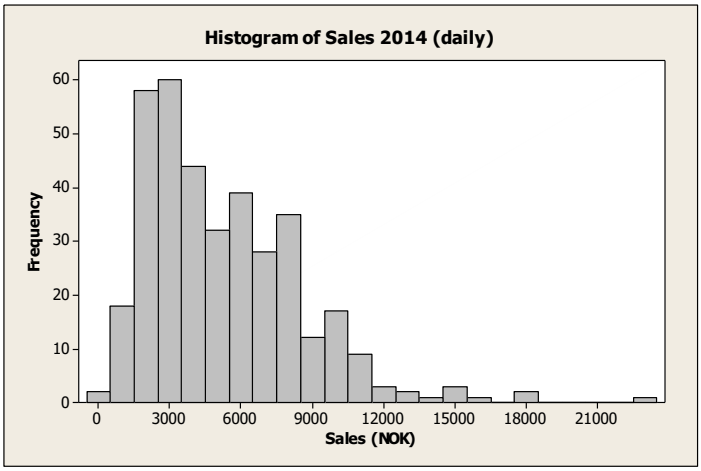

Figure 2. Histogram of daily sales (NOK) in 2014

A histogram of the distribution is given in Figure 2, showing a non-normal distribution, skewed to the right.

Remark. One may ask which theoretical distribution that will fit to this data. So-called probability plots show the lognormal distribution and the Gamma distribution both fit fairly well, the lognormal slightly better than the other for the lower tail, and Gamma slightly better than the other for the upper (right) tail. In our context, the latter may have priority.

The area graphs in Figure 3 and Figure 4 show respectively the shares (ForHere%, ToGo%) and (Cash%, Card%), for consecutive days in year 2014. We see that the vast majority are take-away customers and the vast majority are card-paying customers. However, we see occasional days when more people eat the meal at the premises. There are also occasional days where a lot more people pay cash than usual. Some of the high peaks in the graphs may typically be due to some special cause that day, and Mr. Way may perhaps be able to give a reasonable explanation. On the other hand, the tax officers may find the occasional peaks suspicious, and attribute that to irregularities in reporting. For claiming that, they may have prior knowledge of the customer behavior in this business from elsewhere.

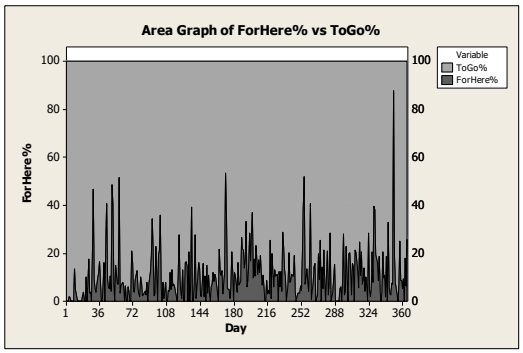

Figure 3. Area graph of ForHere% vs ToGo%

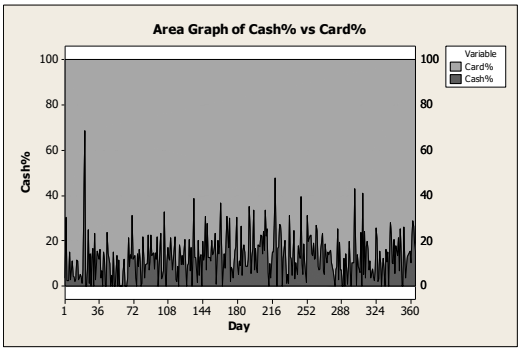

Figure 4. Area graph of Cash% vs Cards%

Histograms for the variables Cash% and ForHere% (or Card% and ToGo%) show that their distributions are skew to the right as well. Taken as shares between zero and one, they are candidates to fit by a Beta-distribution. The fit is reasonable good, better for ForHere% than for Cash%. The existence of a few zero's reduces the fit, and there are more Cash% zero's than ForHere% zero's. In itself, these zero's may in itself raise suspicion among tax officers.

The descriptive characteristics of the variables Sales, Cash% and ForHere% for the year 2014 (No. of Observations, Mean, Standard Error of the Mean, Standard Deviation, Minimum, Lower Quartile, Median, Upper Quartile, Maximum) are as follows:

Variable	Count	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3	Maximum
Sales	361	5279	171	3258	624	2728	4600	7156	23040
Cash%	361	12.70	0.60	11.49	0.00	4.98	11.63	17.29	100.00
ForHere%	361	10.27	0.57	10.97	0.00	2.95	7.11	14.49	97.57

The tax officers may not believe that the records represent the true population of sales. We will therefore examine what can be inferred about the imagined "true" population from the sample. In the following, we will forget about the detailed account of daily sales and

modes of sales exposed above, which may not have been available in the first place, and just relate to the sampled data and the summary reported by the taxpayer, as given in Table 1.

At this point, it is worthwhile to recall the context, where the sampled result may be used at different stages in the preparation of a case against the taxable. The stages required different strengths of evidence, and it is therefore necessary to examine the inferences made from the sample for different guarantee levels, ranging from 50% to 99% or beyond. However, for comparison purposes we stick to the 95% guarantee, frequently used in other legal settings, to represent "Proven by clear and convincing evidence". It may be argued, that a 75% guarantee is enough for this, but a cautionary attitude is recommended. More on this in the discussion section at the end.

We will focus on two modes of statistical analysis. The first, in the next section, makes use of simple and well-known statistical theory and calculations available in standard statistical software, in Excel as well. The second mode of analysis, in the subsequent sections, is based on more advanced statistical theory. Both makes use us assumptions that may be challenged. In practice, the choice has to take into account the ease of communication and fairness with respect to the required ability to understand and challenge the claims based on statistical reasoning. This goes both ways, the tax officer and taxpayer or his/her representative.

4. The Normal analysis

The relevant descriptive statistics for the sample of n=3 turned out to be

Variable	Count	Mean	SE Mean	StDev
Sales	3	4778	1383	2395
Cash%	3	26.9	6.8	11.7
ForHere%	3	30.5	10.3	17.8

First, we note that the standard error of the mean of daily sales are SEMean=1383, so that the reported yearly mean 5279 is well within the range of variation around the observed mean of 4778. In this respect, there is no reason to believe that the three days chosen for examination are special.

Second, we note that the standard error of the mean SEMean for Cash% is 6.8%. This subtracted from the mean 26.9% gives 21.1%, close to the downgraded value in the examiner report. Maybe this is their downgrading principle. However, we see that the standard error of the mean SEMean for ForHere% is 10.3%, which should then lead to the downgrading from 30.5% to about 20% instead of the actual 25%

Question: Does the performed downgrading compensate for the sample uncertainty?

In the current context, we may want a confidence interval for the true means in the population of N=361 opening days, for each of the variables Cash% and ForHere%. This is an interval covering the true mean with a given probability, named confidence coefficient (or confidence level). The interval may be taken as the set of plausible values of the parameter

in question (here true yearly mean), given the data at hand. Typical choice of confidence level in a legal context is 95%, see DeGroot et. al. (1986.

Confidence intervals in the current context are typically constructed by

$$\text{Sample mean} \pm k \cdot \text{Standard Error of Mean}$$

where k is a risk factor, determined according to the desired confidence level. As a crude guidance, k=1 may be taken as 68% guarantee, and k=2 taken as 95% guarantee. Thus the downgrading made by the tax officers, taking a cash share of 21.0% still plausible, given the data, apparently corresponds to an (approximate) 68% guarantee. This consideration is based on the assumption of a large random sample from a population of normally distributed values. Neither of these assumptions are valid here: The sample is small and the population is fairly skewed. The most critical one is the small sample, and we address this first. A confidence interval calculation valid for a small, but still a normal population, is the so-called t-interval, which is available in statistical software, and in Excel as well.[5] With 68% guarantee, we have to increase the safety factor from k=1 to k=1.31, in order to compensate for the fact that the sample is as small as n=3. With 95% guarantee we have to increase the safety factor from k=2 to k=4.3. Taking the sample size n=3 into account, the downgrading of the average Cash% to 21% made by the tax-officers, would correspond to the lower end of a confidence interval with confidence level not 68%, as indicated in the crude analysis, but as little as 53%.

If we want to be fairly sure to pick up the true averages for Cash% and ForHere% in the confidence interval, we may require a confidence level equal to 90%. We then get

Variable	n	Mean	90% Conf.Int
Cash%	3	26.9	(7.2; 46.6)
ForHere%	3	30.5	(0.6; 60.5)

We see that the reported yearly averages 12.6% for Cash% and 10.3% for ForHere% are both is within their respective confidence interval. We can therefore state : With confidence 90%, there is no (statistical) reason to claim that the reported yearly average of Cash% and ForHere% are underreported. On the other hand, if only 50% confidence is required, then the confidence intervals are given by

Variable	n	Mean	50% Conf.Int
Cash%	3	26.9	(21.4; 32.4)
ForHere%	3	30.5	(22.3; 38.9)

We see that the lower confidence limit for the yearly mean of Cash% is close to the downgraded value given by the tax officers.

Although these calculations are standard, also available in Excel, they are based on the assumption that the observations are independent normal with common expectation (the true yearly mean) and common variance (the true yearly variance). Independence may seem reasonable, but note that the sampling is not random from the 361 opening days of the year.

[5] Theory says that with n=3, the judgement should be based on the t-distribution with n-1=2 degrees of freedom.

In fact, the days were selected and in some sense, taken as representative. This may induce dependence. Normality is questionable (see below), common expectation questionable (may vary over the week or over the seasons), common variance questionable (vary with number of daily sales, which itself depends on the weekday). Such violations of assumptions may be used to dismiss the calculations altogether, but the question is: Do they matter much? Typically not, if one accepts that the calculations are crude. Both theory and practice tells that the confidence intervals are likely to become somewhat wider if some of these assumptions are violated. In order to be fair, such knowledge on top of knowledge of standard statistical methods are required. The model uncertainty may then be accounted for, by adopting a slightly higher safety margin.

Some may say that, with a sample of n=3, the confidence intervals get too wide to be useful at all, when a reasonable high confidence guarantee is required. For confidence level 95%, common in legal settings, the interval gets even wider than the one above for 90%, and at the low end goes below zero. This exposes the fact that the method of analysis does not recognize that the shares are numbers between 0% and 100%, and thus non-negative. This may be made up for, by a transformation of data. For original data x, transform x by $y=\log(x/(100-x))$, and calculate confidence limits based on the transformed data, and back-transform these by $x=100\cdot e^y/(1+e^y)$. This brings the data closer to being normal and therefore closer to exact confidence guarantees. The result from this alternative calculation is provided here, for some different choices of confidence level.

Confidence level	Conf. Int. Cash%	Conf.Int. ForHere%
50%	(20.5, 31.8)	(21.0, 37.6)
80%	(14.9, 40.7)	(13.4, 51.0)
90%	(10.7, 49.9)	(8.5, 63.4)
95%	(6.8, 62.2)	(4.4, 78.0)

Table 2. Confidence intervals for different confidence levels (transformed data)

We see that the tax-officers stipulations of the yearly average of Cash% to 21.0 only corresponds to an approximate 50% guarantee, while the reported average of 12.6% is not unreasonable if a 90% guarantee is required. We see that the downgrading for ForHere% to 25.0 is not sufficient to give a 50% guarantee.

Preliminary conclusion: The discretionary judgment of the tax authority concerning Cash% is incompatible with the statistical judgment, unless a fairly high risk of wrongdoing is taken.

We may ask: Does the phrase that corresponds to “proved on the balance of probabilities” really mean application of confidence level 50%?

Remark. The discussion above is based on the common two-sided confidence intervals. It may be argued, in the given context, that the upper confidence limit will never be used, and that only a one-sided guarantee is needed. If a one-sided interval is established by using the

lower confidence limit of the two-sided interval with 100% as upper limit, the guarantees are as follows:

Two-sided guarantee	0%	50%	80%	90%	95%	99%
One-sided guarantee	50%	75%	90%	95%	97.5%	99.5%

However, in practice it may cause confusion to depart from the standard confidence idea.

We may now compare three modes of reporting the uncertainty by confidence intervals:

(i) Z-intervals, based on large sample ideas (e.g. use 68%-95% rule)
(ii) T-intervals, based on small sample theory
(iii) T-intervals, based on small sample theory and transformation of data

Here (i) -> (ii) -> (iii) represents increased trustworthiness with respect to the attached confidence. Going from (i) to (ii) widens the confidence intervals and gives favor to the taxpayer. Going from (ii) to (iii) narrows the confidence interval and gives favor to the tax-officer.

Some prefer to cast the problem as one of hypothesis testing. The natural null-hypothesis is the true yearly average is as reported average by the taxpayer, with the alternative that the true yearly average is in the direction of the claim by the tax office. This way, the null-hypothesis is rejected when there is sufficient evidence from the data in support of the alternative. For Cash% the reported yearly share was 12.7%. The standard T-test for Cash% then gave the following result:

```
T-Test of mu = 12.7 vs > 12.7
Variable Count   Mean   SE Mean     T     P
Cash%        3   26.9       6.8 2.10 0.085
```

With a P-value of 8.5%, the null-hypothesis is not rejected at the 5% significance level. This is essentially equivalent to the calculation of a one-sided 95% confidence interval, which gives the lower limit 7.2%, and the reported 12.7% is well above that. This lower limit may also be obtained from lower limit of the two-sided 90% confidence interval, as seen early on in this section. The corresponding T-test for ForHere% where the reported yearly share was reported to be 10.3%, turned out as follows, with the same conclusion.

```
T-Test of mu = 10.3 vs > 10.3
Variable  Count   Mean   SE Mean     T      P
ForHere%      3   30.5      10.2  1.97  0.094
```

The support for claiming underreported ForHere% is even slightly weaker, despite the fact that the discrepancy between reported and observed is larger, due to a much larger standard error, i.e. larger sample variation.

Remarks. In practice, there are widespread misunderstandings about the probability guarantees for both confidence intervals and hypothesis testing. The confidence level is the probability that the confidence interval capture the unknown. It is formally wrong to say that it is the probability that the unknown lies in the interval, but may be seen as just a semantic difference causing no harm. Significance level is the probability of rejecting the null-hypothesis when it is true. P-value is the probability of getting a T at the level of the one observed or beyond (i.e. in favor of the alternative), calculated under the assumption that the null hypothesis is true. Neither of these are really probabilities about the null-hypothesis or alternative itself, e.g. being true or wrong. Such misunderstandings are more grave, and some statisticians prefer to stay away from hypothesis testing altogether, in order to avoid them. However, such direct probabilities are attainable within a Bayesian context, a theme in a later section.

The statistical analysis in this section is based on the fraction of sales in monetary terms (Card vs Cash and For Here vs To Go). When this is computed for each of three days, we have n=3 observations that allow the evaluation of uncertainty. However, these uncertainties are large and it is felt that more definitive results may be hidden in the data. Are there some alternative professional statistical judgement available? This is the theme of the next section.

5. The Gamma-Beta analysis

Instead of looking at the fraction of sales for each type in each of the three sampled days, we may examine the overall fraction for the three days and support it by some theory. These fractions are given in the following table

	Cash%	ForHere%
Overall Sales three days	24.5	28.8

Note that they are different from the average of the fractions for each of the three days, since the number of sales each day are different. In a sense, the calculation here gives equal weight to each of the sales, instead of equal weight to each of the days. Now we have just n=1 observation for each of the variables. Why that we can judge any uncertainty at all? Here statistical theory may come to help!

Sales are non-negative and may possibly fit to a Gamma-distribution. Gamma-distributions are determined by two parameters, a shape-parameter and a scale-parameter. A sum of independent Gamma-distributed variables, with the same scale parameter, is itself Gamma-distributed with shape parameter equal to the sum of the individual shape-parameters, and keeping the scale-parameter fixed. Details are given in Appendix A.

With this fulfilled, we have two sums of sales S_1 and S_2, say for Cash or Cards respectively, which are independent and distributed Gamma($shape_1$, scale) and Gamma($shape_2$, scale)

respectively. Then the ratio $S_1/(S_1 + S_2)$ is distributed Beta($shape_1$, $shape_2$). More properties may be found in Johnson et.al. (1994) for Gamma, and in Johnson et.al. (1995) for Beta.

To support the analysis one may check whether the Gamma-assumption is reasonable. This will typically be the case, and more so than the normal assumption in the standard analysis in the preceding section. The added requirement of identical scale parameter may be questionable. It may change somewhat over the week and over the year, but this may be a minor problem. More crucial is the assumption that the scale-parameter is the same for both types of sales, say card or cash. Let us look into this with our data.

5.1 Gamma-Beta analysis: Common shape and scale model

The following graphs show the histogram and fitted Gamma-distribution by maximum likelihood, for all sales during the three sampled days.

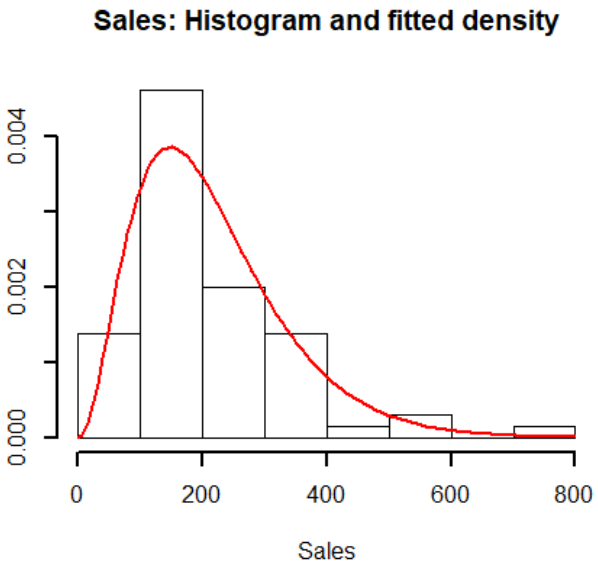

Figure 5. Histogram and fitted Gamma-density for Sales (all)

We see a good fit to the Gamma-distribution, and so far this is promising. A probability plot may also be used as support for this. Let us outline a crude analysis without any worries about the specific assumptions above. The maximum likelihood estimates of the Gamma-parameters based on all sales are as follows[6]:

Sales	Estimate	Std. Error
Shape	3.09	0.50
Scale	0.0145	0.0026

[6] Distribution fitting is performed in R using the program fitdistrplus, see Delignette-Muller and Dutang (2015).

Now assume that all 67 sales are drawn from this distribution, regardless of type of sales. The numbers of each type are given by

Cash: 18	For Here: 22
Card: 49	To Go: 45
All: 67	All: 67

The random variable related to Cash% may the by written as a ratio as follows

$$\frac{\text{Sum of the 18 Cashvariables}}{\text{Sum of the 18 Cashvariables} + \text{Sum of the 49 Cardvariables}}$$

From the theory outlined above, we get that Cash% will have a Beta-distribution with parameters

$$(18 \cdot 3.09, 49 \cdot 3.09) = (55, 151)$$

As a check of consistency the expectation of this distribution is

$$\frac{55}{55 + 151} = 3.09 \cdot \frac{18}{18 + 49} = 0.267$$

which is close to the result 26.9% obtained for the standard averaging method in section 3.

Similarly, the random variable related to ForHere% may the by written as a ratio as follows

$$\frac{\text{Sum of the 22 ForHere variables}}{\text{Sum of the 22 ForHere variables} + \text{Sum of the 45 ToGo variables}}$$

and that ForHere% will have a Beta-distribution with parameters

$$(22 \cdot 3.09, 45 \cdot 3.09) = (68, 138)$$

As a check of consistency, the expectation of this distribution is

$$\frac{68}{68 + 138} = 3.09 \cdot \frac{22}{22 + 45} = 0.330$$

which is somewhat higher than the result 30.5% obtained for the standard averaging method in section 3.

The estimated Beta probability densities of Cash% and ForHere% are as follows:

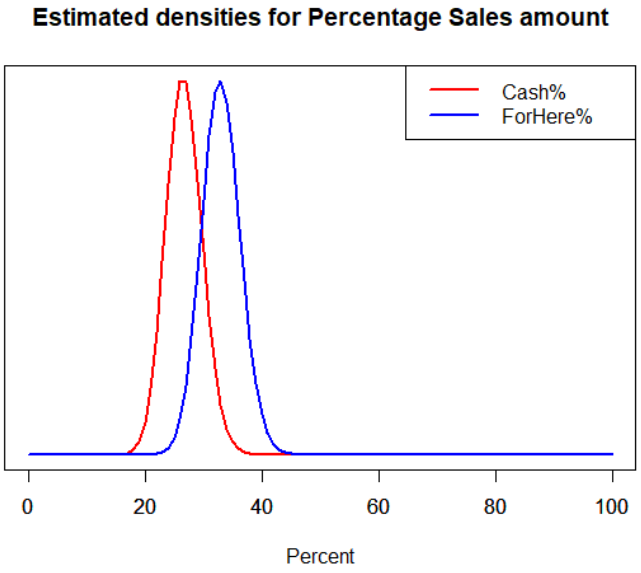

Figure 6. Estimated densities for percentage sales (common shape and scale model)

Quantiles of these distributions are given in the following table:

Quantile	1%	5%	10%	20%	30%	40%	50%
Cash%	19.9	21.8	22.8	24.1	25.0	25.8	26.6
ForHere%	25.7	27.7	28.9	30.2	31.3	32.1	33.0

Table 3. Quantiles of fitted distributions (common shape and scale models)

Probabilities of not exceeding specific percentage levels are given for each of Cash% and ForHere% in the following tables:

%level	20.0	21.0	22.0	23.0	24.0	25.0	26.0
P(Cash%<%level)	0.01	0.03	0.06	0.11	0.19	0.30	0.42

Prob < %level	25.0	26.0	27.0	28.0	29.0	30.0	31.0	32.0
P(ForHere%<%lev)	0.005	0.01	0.03	0.06	0.11	0.18	0.27	0.38

Table4. Non-exceedance probabilities (common shape and scale models)

5.2 Gamma-Beta analysis: Different shape and common scale model

The analysis above is based on the assumption that all sales amounts come from a common Gamma-distribution. This can be weakened by assuming Gamma-distributions with common scale, but possibly different shape parameters in the two groups. However, common scale is crucial, in order for the ratio to be Beta-distributed and independent of the scale. We will now examine the assumption. In the case that the assumptions are violated, we will examine the consequences, and try to answer questions like: Do we gain anything by refining the model, or can we stick with the simple one for practical application?

The descriptive statistics of the 67 sales amounts subset according to type of sale are as follows:

Variable	Count	Mean	SE Mean	StDev
Card	49	225.1	15.9	110.5
Cash	18	191.3	41.6	176.5
ToGo	45	227.0	19.3	129.1
ForHere	22	183.3	29.1	136.6

We see that on average the Card payments are higher than the Cash payments. Likewise, the ToGo payments are on average higher than the ForHere payments. However, none of the two differences are statistically significant.

The following graphs show the histogram and fitted Gamma-distributions by maximum likelihood for Card payments and the Cash payments respectively.

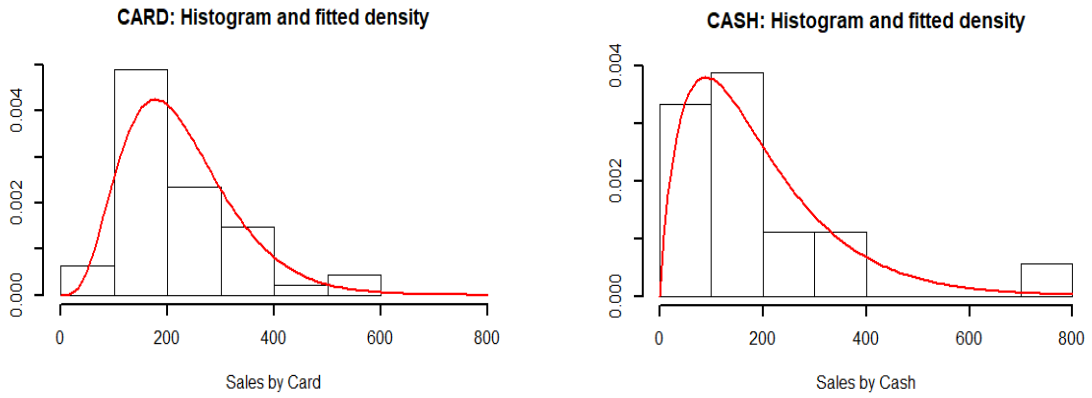

Figure 7. Histograms and fitted densities (common scale models)

We see a reasonably good fit, but the shapes are somewhat different, typically due to less small sales paid by cash. Whether the scale (or rate) is different is hard to tell from the graphs. The respective parameter estimates are

Card	Estimate	Std. Error	Cash	Estimate	Std. Error
Shape	4.72	0.93	Shape	1.86	0.56
Scale	0.0210	0.0043	Scale	0.0097	0.0033

From this, we see that the scale parameters are different as well. In fact, this is no surprise, since the expectation of a Gamma variate is Shape/Scale, and the expected value of sold amount is not that different for the payments by card and cash. So, if the shape is different, the scale have to be different as well.

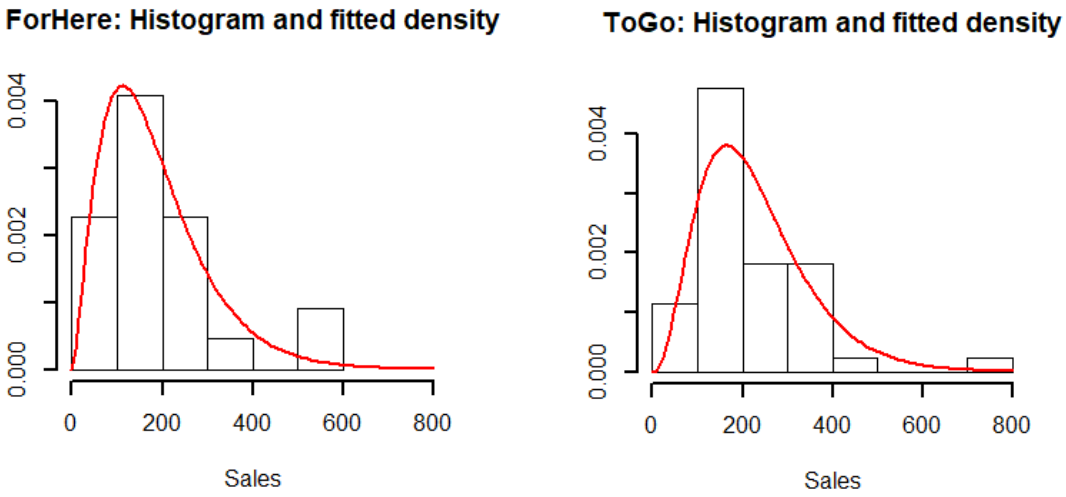

Figure 8. Histograms and fitted densities (common scale models)

The parameter estimates are:

ForHere	Estimate	Std. Error
Shape	2.57	0.72
Scale	0.0140	0.0043

ToGo	Estimate	Std. Error
Shape	3.62	0.73
Scale	0.0159	0.0034

The shape parameters of the ForHere-sales and the ToGo sales differs slightly, while the scale parameters turned out very close. Therefore, a Gamma-Beta model may be justified. An improvement of the crude calculation above may take advantage of this, by re-estimating the model by maximum likelihood with a common scale-parameter. For convenience, we may just re-estimate the shapes, by stipulating the scale at the intermediate value 0.0150. Then the shape estimates are changed to 2.72 and 3.43 respectively. Using this in our calculation, we get that ForHere% will have a Beta-distribution with parameters

$$(22 \cdot 2.72, 45 \cdot 3.43) = (60, 154)$$

The expectation of this distribution is

$$\frac{60}{60 + 154} = 0.280$$

If we, not that well justified, do the same for Card-sales and Cash-sales and stipulate a common scale at the intermediate value 0.0015, the shape-parameter for Card becomes 3.51 instead of 4.72, and for Cash 2.62 instead of 1.86. We then get the Beta-parameters

$$(18 \cdot 2.62, 49 \cdot 3.51) = (47, 172)$$

The expectation of this distribution is

$$\frac{47}{47+172} = 0.215$$

Quantiles of the Cash% and ForHere% distribution are now given by

Quantile	1%	5%	10%	20%	30%	40%	50%
Cash%	15.4	17.1	18.0	19.1	20.0	20.7	21.4
ForHere%	21.2	23.1	24.2	25.4	26.8	27.2	28.0

Table 5. Quantiles of fitted distributions (common scale models)

We see that the expectation and the quantiles are about 5% lower than the ones obtained above, for a model assuming that the Gamma-distribution of sales amount is common to both payment groups. Therefore, the more realistic Gamma-Beta analysis will be more "friendly" to the taxpayer.

5.3 Gamma-Beta analysis: Different shape and scale model

For the situation with different scale-parameters the ratio will have a more complicated distribution than the common Beta-distribution. The probability density is given in Appendix A. Required quantiles may be calculated as tail areas under the density curve. While such calculations for the Beta-distribution are available directly in statistical software like R, these integral calculations may be performed in R and stored as a function with input the four estimated Gamma-parameters and the numbers of each type of sales. This gave the following quantiles:

Quantile	1%	5%	10%	20%	30%	40%	50%
Cash%	16.4	18.4	19.5	20.9	21.9	22.8	23.6
ForHere%	21.3	23.3	24.3	25.6	26.6	27.4	28.2

Table 6. Quantiles of fitted distributions (different shape and scale models)

We see that the low quantiles for Cash% are about 1% higher than those for the common scale case, while the ForHere% quantiles, as expected, are almost identical with the ones above.

An alternative, which circumvents the complicated theory, is to get at the underlying distribution by so-called resampling. This is to simulate new observations from the estimated Gamma-distribution, and then compute the ratio. Repeat this many times, say 10 000, and the empirical distribution of the 10 000 ratios obtained this way, provides a picture of the uncertainties in the original ratio-estimate. This matches well with the theory, as the 1%-quantiles obtained this way turned out to be 16.7% and 21.1% respectively.

Remark. One may ask how to do hypothesis testing within the Gamma-Beta model. The null hypothesis would then be that the expectation of the Beta-distribution for the ratio R is as reported by the taxpayer. This implies a restriction on the underlying Gamma-parameters. In principle, the likelihood-function may be maximized under this restriction, and then give rise to a likelihood ratio test. This becomes too complicated to deserve attention here. However, a way of calculation likelihood-ratios is indicated in Appendix A.

Will the above model take into account all sources of statistical uncertainty? Maybe not! We have taken the number of sales of each type as given. Taken over a given sampling period, here three days, they are random. The total number of sales is random as well, but may be taken as a fixed number n, as if we have determined at the outset to observe n sales. We may then imagine that for each sale, there are probabilities (1-p,p) for the sale being of the specific type, e.g. (Card, Cash). We may then take the assumption of constant p with independence from sale to sale. This will add one parameter to the models above. The extended model may then be estimated by maximum likelihood, as above. The conditional distribution of the ratio, given the number of sales of each type, will be Beta-distributed, but the unconditional distribution becomes complicated. Approximations based on Taylor expansions and asymptotic theory exist, but are complicated and impractical. We will therefore not follow this path here, but we will see in the next section that the extended model may be implemented with less effort within a Bayesian context.

6. Bayesian Gamma-analysis

Various Bayesian formulations assuming the sales being independent Gamma-variates may be tried. As benchmark, we take the full model with different shape and scale-parameters for each type of sale, and with type (e.g. Card/Cash) determined by independent Bernoulli-variables with constant probabilities (1-p, p). This gives a five-parameter model. Alternatives may be the reduced model with common scale (which may be justified for ToGo/ForHere) and an extended model, where the probabilities are dependent on the sales amount expressed by a regression term. We will not pursue the latter.

The Bayesian analysis requires specification of prior distributions for each of the model parameters. In the current context, it is natural to assume that we have no prior knowledge. This is implemented by choosing a joint prior, determined by a non-informative prior for each parameter, combined with independence. Given the data, the joint posterior distribution is determined by Bayes law, which in this case will lead to integral formulas. In practice, the solution may be obtained numerically by so-called Markov Chain Monte Carlo techniques, which amounts to artificial creation of samples from the joint posterior distribution of the parameters. This is implemented in standard Bayesian software. We have used OpenBUGS, within the statistical environment R, and implemented in R2OpenBUGS, see Sturz et.al (2005). The call on BUGS within R is as follows:

```
my.sim <- bugs(data, inits, model.file="BUGSmodel-full.r",
parameters=c( "prob2", "shape1", "shape2", "scale1", "scale2"),
  n.chains=1, n.iter=20000)
```

Here the data is a list with three items (n=no. of observations, here 67, y=sales amount in NOK , x=type of sale, e.g. 0=Card, 1=Cash). The BUGS-call asks for a model file, and the parameters to be returned from MCMC simulation of one chain. The number of iterations is specified to 20 000, from which the 10 000 last ones are returned for further processing. After executing the command, the results are stored in the object my.sim which, among others, contains the 10 000 simulated values of the five parameters, in theory sampled from the joint posterior distribution of the five parameters.

The model specification for the full model is as follows:

```
model
{
 for( i in 1 : n ) {
  y[i] ~ dgamma(a[i],b[i])
  x[i] ~ dbern(prob2)
  a[i] <- (shape2-shape1)*x[i]+shape1
  b[i] <- (scale2-scale1)*x[i]+scale1
 }
 shape1 ~ dgamma(0.001,0.001)
 shape2 ~ dgamma(0.001,0.001)
 scale1 ~ dgamma(0.001,0.001)
 scale2 ~ dgamma(0.001,0.001)
 prob2 ~ dbeta(1,1)
      }
```

Here y[i] represents sale no. i, assumed to be Gamma-distributed, and x[i] is a Bernoulli-variable, being 1 with probability prob2 and 0 otherwise. This will assign the Gamma-parameters for two types of sales (say 1=Card, 2=Cash) accordingly, i.e. (shape1, scale1) with probability prob1=1-prob2, and (shape2, scale2) with probability prob2. We want to assign non-informative priors to the five parameters, and note that BUGS requires proper priors. Since the Gamma-parameters are non-negative and unbounded, Gamma-distributions with large variance will serve the purpose. A common choice is Gamma(0.001, 0.001), which has expectation 1 and variance 1000. For the Bernoulli-parameter prob2 is chosen a uniform prior over the [0,1]-interval, here represented by a Beta(1,1)-distribution. Note that our choice here will not twist the result, as long as the non-informative aspect is taken care of.

For the full model, we get the mean and standard deviation of each posterior:

Card/Cash	Mean	Std. dev
prob2	0.283	0.0548
shape1/	4.65	0.94
shape2	1.79	0.56
scale1/	0.0207	0.0044
scale2	0.0094	0.0034

ToGo/ForHere	Mean	Std. dev
prob2	0.338	0.057
shape1/	3.54	0.74
shape2	2.48	0.68
scale1/	0.0156	0.0035
scale2	0.0136	0.0041

From this an ad hoc measure of expected share of cash may be computed by

$$R = \frac{0.283 \cdot (\frac{1.79}{0.0094})}{0.283 \cdot \left(\frac{1.79}{0.0094}\right) + 0.717 \cdot (\frac{4.65}{0.0207})} = 0.251$$

This is obtained by inserting the averages of the simulated parameter values into a formula expressed by the Gamma-expectations of each type of sales and their expected share. The corresponding percent ForHere sales are 29.1%. Alternatively, these calculations may be done in reverse order, compute the ratio for each of the 10 000 simulated parameter values in the estimation process itself and then average. This will give 25.5% instead of the 25.1%., and 29.4% instead of 29.1%, i.e. only slightly higher. The latter method may be preferred, since the 10 000 simulations at the same time provides insight to the posterior distribution of the ratio, not just its mean. Here follows the R-program for doing this:

```
n=67 # no. of observations
R=rep(0, 10000) # Initialize vector of shares of type 2
for (i in 1:10000) {
x=rbinom(n,1, prob2)  # Generate type of sale (0=type 1, 1=type 2)
s1=sum((1-x)*rgamma(n,shape1, scale1)) # Sum of sales of type 1
s2=sum(x*rgamma(n,shape2, scale2)) # Sum of sales of type 2
R[i]=s2/(s1+s2) # Compute share of type 2 sales
}
```

The posterior distributions of Cash% and ForHere% turned out as follows:

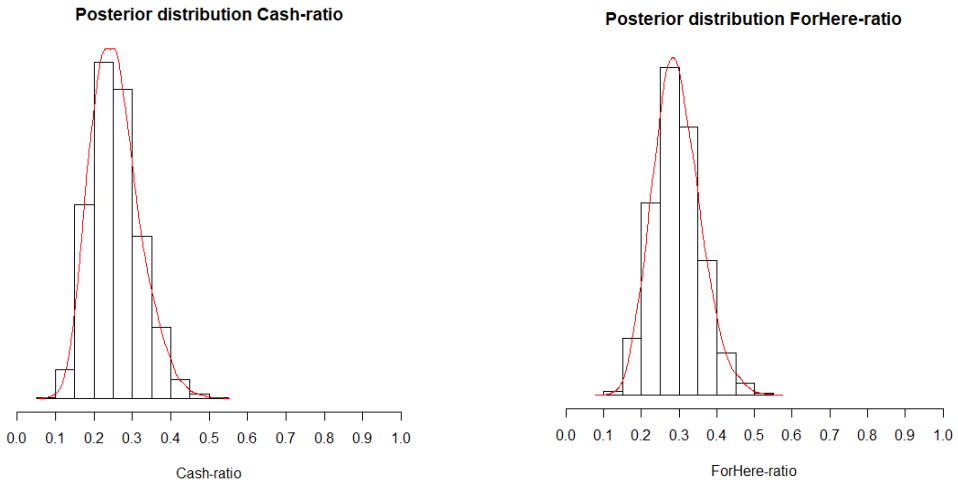

Figure 9. Posterior distributions of ratios (Bayesian full model)

Quantiles of the distributions are given by

Quantile	1%	5%	10%	20%	30%	40%	50%
Cash%	13.3	16.2	17.8	20.0	21.8	23.4	25.0
ForHere%	16.6	19.7	21.7	24.1	26.0	27.5	29.0

Table 7. Quantiles of fitted distributions (Bayesian full model)

It may be of interest to compare 1% quantile obtained by various Gamma-models, including the Bayes-models of this section, and the non-Bayesian ML-models of the preceding section. We believe that non-Bayesian models, which includes the extra probability parameter, will perform much like the Bayesian ones. As said before, such models will be complicated and impractical. However, a possibility is to estimate the probability and the Gamma-parameters separately, and from this, simulate a sample sales of the actual size (here 67) and then compute the shares of each type. This may be repeated many times in order to get at the probability distribution of the ratio. For the ML-estimated models which do not lead to Beta-distributed ratio (marked by *), we have performed simulations with 10 000 repeats. The comparisons may be read from the following table, where no probability means that the observed numbers of each sales type are taken as fixed (non-random).

Type of Gamma Model	Q1%-Cash%	Q1%-ForHere%
Bayes – Full	13.3%	16.6%
Bayes – Common scale	12.1%	16.4%
Bayes – Full – No probability	17.3%	22.1%
Bayes – Common scale – No prob.	16.2%	21.8%
* ML – Full	11.2%	14.8%
* ML – Common scale	10.2%	14.9%
ML – Full – No probability	16.4%	21.2%
ML – Common scale – No prob.	15.4%	21.3%
ML – Common both – No prob.	19.9%	25.7%

* By simulations

Table 8. The 1%-quantiles of fitted distributions for different models

We see that the Bayes-models with random payment type give a considerably lower 1% quantile than the other Bayes-models. Choosing one of these two models corresponds to allow for additional uncertainty, which is both reasonable and favorable to the taxpayer in general. We notice that the quantiles are very close for ForHere%, where the common scale assumption was better justified. For Cash% the 1% quantile is lowest when we assume a common scale. This corresponds to the fact that common scale is not that well justified, and that the solution have to allow for the uncertainty thus created. The choice of this model for Cash% would therefore give extra favor to the taxpayer, undeserved in light of the data. From the “proven beyond reasonable doubt” point of view, it seems that the model should include the extra probability parameter, in order to represent the existing uncertainties and, at the same time, give reasonable favor to the taxpayer.

For the ML-models, we see that the results compare well with the corresponding Bayesian ones, and with a similar pattern between them. In the last line, we see the first simple Gamma-Beta analysis of section 5, which gives the most favor to the taxpayer, but not so much as the Normal analysis of section 4.

Which one of the Gamma-type models should be favored? This may be a compromise between ease of understanding, ease of calculation, ease of communication and reasonable favor to the taxpayer. We deal with some of these aspects in the next section.

7. Discussion

What conclusions on the truthfulness of the taxpayer reporting can we draw from an analysis based on Gamma-type models, Bayesian or not? Recall that the yearly averages reported by the taxpayer were Cash%=12.7 and ForHere%=10.3. The precise judgment on how extreme the numbers are, will depend on the choice of model. In practice, a statement of level may be sufficient, and will not leave an impression of undeserved high accuracy. We see that reported Cash% is about the level of 1% quantile, while reported ForHere% is more extreme, in fact down to the level of 0.1% quantile. This may justify the following:

Conclusion: The data sampled by the tax office provides strong support for their case, and cannot be dismissed for statistical reasons as part evidence.

Let us now discuss some of the general issues raised, but left unanswered, in previous sections. First, some words about the reporting of significance of the sampled data. Contrast the following statements:

A: "Given the available data, the probability is less than p% that the true yearly average is as reported or less".

B: "Given the reported yearly average is true, the probability of getting the sampled average of that size or less Is p%"

Here A is typically a Bayesian statement, while B is a non-Bayesian statement, like the reporting of P-values in hypothesis testing. As mentioned earlier there are some obstacles for treating the Gamma-type model within the hypothesis testing context, and for non-Bayesians we have instead estimated the distribution of the ratios R and taken that as given. This way, statement A may be justified in this case as well, without doing any harm (except to the purist statistician).

We may also use statements like the following:

C: "The reported yearly average is outside the range of plausible values, which is down to x with guarantee p%"

In the Bayesian case, the range is a so-called credibility interval, with probability guarantee attached to the unknown true average. In the non-Bayesian case, the above statement usually refers to a confidence interval, with guarantee on the probability that the interval capture the unknown true average. Statement C may serve both purposes without causing any harm. However, our ranges for the non-Bayesian Gamma-type models are not confidence intervals in the common sense.

Another way of reporting results is as follows:

D: “In light of the sampled data, our claim of true cash-ratio r_1 is x times more likely than your claim r_0”.

In the current case the taxpayer claimed r_0=0.127, while the sample gave r_1=0.245. For the Gamma-Beta model (with common scale), we offer a formula for computation of the, likelihood-ratio LR in Appendix A. This gave LR=220, an impressive number in favor of the tax office. The question remains, whether this calculation will be understood and carry weight as evidence.

Then to the choice of model. There is a difference in interpretation between the two approaches, Bayesian and non-Bayesian. The Bayesian approach provides direct probability statements, while the ML-approach provides confidence statements, which are often misunderstood in practice. The direct probability statements given in section 4 are in a sense fake. We estimate the probability distribution, and then take this as given to calculate the probabilities. This is not a serious objection. One may argue that, given that Bayesian thinking, modelling and calculation now have become commonplace, there are no major reason to stay away from it. Once we have taken the step into Gamma-models, the full Bayesian model is no more complex than the simpler one. It does not add extra restrictions to be argued about, and it gives some favor to the taxpayer. The corresponding full ML-model does not have “exact” theory, but requires some simulations on top of the estimation, easily implemented in Excel. Seemingly, it goes too far in giving favor to the taxpayer. It is usually a virtue in statistical modelling to be parsimonious, that is, add no more than the necessary parameters to the model.[7] This may be seen as an argument for assuming common scale and/or leaving out the type of sale probability, taken the numbers as given. However, the reported cash ratio may be too low for two main reasons: Amounts are downgraded or some payments not reported at all. The latter is the most probable reason. The observation that the cash payments are on average slightly lower than the card payments, is probably just common customer behavior. This is an argument for keeping the type of payment probability as a model parameter to account for this uncertainty.

In summary: The preferred model is the full Bayesian model, and if justified, simplified to common scale.

The next issue is the choice of risk levels. First, the meaning of risk will depend on the context. In the non-Bayesian case, we may take confidence levels 95% or more and P-values 5% or less, in line with common practice of reporting in forensic statistics. In the Bayeasian case, one may refer directly to the probability that the true value is as stated by the taxpayer or beyond on the unlikely side. Then it is a question about how far out in the posterior distribution one should go to claim or act as if it is a misstatement. Then the 5% quantile may be a reasonable choice. However, the tax-officer meet the choice repeatedly in several contexts, which requires different choices. Refer back to section 2, with the outlined four-stage process and three principles for strength of evidence: I=“Proven by a preponderance

[7] This argument will carry more weight when the number of observations is small.

of the evidence/on the balance of probabilities" (at least 51%), II= "Proven by clear and convincing evidence" (say at least 75%), III= "Proven beyond reasonable doubt" (say at least 99%). This means that the common way of statistical reporting, with 95% guarantees, will fulfill strength II, but not strength III. It is therefore sufficient for deciding on discretionary assessment to take place, which will typically be supported by other evidence as well. It is also sufficient for deciding regular penalty tax, but not sharpened penalty tax. There may be good reasons for being cautionary, and go beyond 75% to claim "Clear and convincing evidence ". This may compensate for model uncertainty, i.e. if the solution is derived under standard assumptions, possibly not fully met by the data, and which will require statistical expertise to clarify. The tax authority may face the risk that the statistical evidence may be rejected altogether. An interesting question is whether a raise of probability guarantee would help to prevent this. For more insight, see Appendix C.

The law profession has been hesitant to pinpoint required strength of evidence in terms of probabilities. However, it seems that "proven beyond reasonable doubt" is achieved in court at a lower level than 99%. In a Norwegian study more than 1100 judges (professional and lay) were asked about the likelihood of guilt, and their vote guilty or not ,in a fictitious rape case (Magnussen et. al. , 2014). It turned out that, among those who judged 90% chance of guilty, the majority would vote for guilty. In the US, Gastwirth (1992) refers to an informal survey made by a judge, asking 12 colleagues to pinpoint probabilities. The result was somewhat disturbing, as some of them were down to 60% for "clear and convincing evidence" and down to 80% for "proven beyond reasonable doubt".

Given that discretionary assessment is demonstrated to be legitimate, the assessment itself shall turn out the facts, as best possible. This means that the tax-office may settle on an average as an effort to estimate the true yearly average. It is not given that this estimate need to be a sample average. In fact, the median is in many cases a better choice[8]. In the case of Gamma-model and the approach taken in this work, the median (as 50% quantile) is recommended, also because it fits well with the use of quantiles in general. In a sense, this is an application of the principle "on the balance of probabilities", conditional on the loss of credibility. If the tax-officers want to be cautionary, for instance due to model uncertainty, they my settle at a lower quantile.

In the case of Take-A-Way, the tax officers downgraded the three sampled daily averages in favor of the taxpayer. Finding that the downgraded results were still away from the reported figures, the credibility of the taxpayer was taken as non-existent. Given the loss of credibility, they were entitled to reassess taxes for the preceding years by average calculations. The tax assigned is the most likely, given the sampled data. Their evidence was sufficient to impose penalty tax, but not sharpened penalty.

Note that "proven by preponderance of evidence" is the requirement in civil court cases, and refers to the final conclusion taken by the judge, after taking into account all the evidence, statistical and other. In some cases, the statistical analysis of the kind above will

[8] Note that the most likely outcome from a distribution is the mode, but for our models, the mean, median and the mode will not differ much.

be part of the total evidence. With several statistical statements, based on evidence judged separately and independently, we cannot just work "on the balance of probabilities" (more probable than not) for each of them. Assume that we claim a taxpayer misstatement on two issues, based on independent pieces of evidence with attached 60% confidence on each (40% risk). Then there will be just 36% confidence (64% risk) attached to the joint claim of misstatement. To give a specific statistical evidence weight, it is therefore required to aim at a low risk level, more in in line with "proven beyond reasonable doubt". As a practical rule of thumb, it seems reasonable to require risk levels of 5% or less for individual statements.

The Gamma-Beta model with common scale in section 5 offers additional opportunities for the tax office for analysis, admittedly speculative. Suppose that the low ratio of amounts paid cash is altogether due to unreported cash sales. If we assume that all sales paid by cards are reported, we may ask what is the number of cash sales x that will match the reported cash ratio of 12.7%. As yardstick, it is convenient to take the 49 card payments actually observed. We then get

$$\frac{x \cdot 2.72}{x \cdot 2.72 + 49 \cdot 3.43} = 0.127 \ \rightarrow x = 9$$

This contrasts the 18 cash payments actually observed, and it may be tempting to conclude that 50% of the actual cash sales are not reported. Although this is based on the model with common scale, an alternative argument in terms of the observed average amounts paid for the two types of payment, will give similar result. It may be argued, that direct extrapolation based on averages is, in a sense, on the balance of probabilities. However, both modes of reasoning is not sufficient as part evidence, and should probably be refuted altogether.

The tax office may possess prior knowledge of the fast food business, which may have raised the suspicion in the first place. This could hardly be used as evidence, unless data is relevant, reliable and convincing. A possibility is to extract cash-ratios from a sufficiently large sample of firms in the relevant segment, and from this establish a distribution of cash-ratios. One may then judge whether the specific pizza-place in question is extreme at the low end of the distribution. A complicating factor is that this sample may contain some tax evaders as well. In fact, such sample statistics was gathered by a regional tax office, and used for proofing in a recent case brought to court.[9] The sample was restricted to firms within a chain and not run by the owner, believing that they have no incitement to misreport the cash-ratios. It is of some interest to look into the actual numbers.

	Sample of firms			Taxpayer report	Tax office reassessment
Year	Sample size	Minimum	Order of magnitude		
2009	242	22.2	[26, 40]	9.7	22.7
2010	316	17.1	[22, 31]	11.4	20.5
2011	316	17.3	[20, 30]	8.5	17.9

Table 9. Cash-ratio information from nationwide sample

[9] Oslo Tingrett: TOSLO-2014-193468-UTV-2017.1347

We see that the discretionary assessment is close to the low end of the distribution of the sampled values. However, the way of adjustment from the minimum value, is not stated. The notion “order of magnitude” is not given a precise meaning, but apparently mean the main range of the distribution. The figures are consistent the 90% range: [5% quantile, 95% quantile], and a cautionary assessment set at the 1% quantile. If so, this should be made explicit in the verdict.

The attorney of the taxable firm argued that the sample taken was arbitrary and had no value as evidence. The lower court agreed on this, but the case ended in Supreme Court, where the sample evidence was accepted. More on this in Appendix B.

Acknowledgement

This work has benefitted from input by Certified Internal Auditor Are Lysholm and a meeting with tax office staff Torkel Fure, Jostein Rene, and Anne May Melsom. They are not, in any way, responsible for viewpoints expressed in the paper.

8. References

Aitken, C., Roberts, P. and Jackson, G. (2010) Fundamentals of Probability and Statistical Evidence in Criminal Proceedings. Guidance for Judges, Lawyers, Forensic Scientists and Expert Witnesses, Practitioners Guide No 1. Prepared under the auspices of Royal Statistical Society's Working Group on Statistics and the Law.

Bright, J.C., Kadane, J.B. and Nagin, D.S. (1988) Statistical Sampling in Tax Audits, Law and Social Inquiry, 13, 305-333.

COIC & RSS (2017) Statistics and Probability for Advocates: Understanding the Use of Statistical Evidence in Courts and Tribunals. The Council of Inns of Court (COIC) in cooperation with Royal Statistical Society (RSS).

DeGroot, M.H., Fienberg, S.E. and Kadane, S. P. (eds.) (1986) Statistics and the Law, New York: John Wiley.

Delignette-Muller M. L. and Dutang, C. (2015) fitdistrplus: An R Package for Fitting Distributions, Journal of Statistical Software, 64(4), 1-34.

Eide, E. (2000) Oversikt over litteratur om svart arbeid og skatteunndragelser, Rapport fra prosjektet "Svart økonomi i Norge", finansiert av Skattedirektoratet, ISBN 82-7988-019-4.

Eide, E. (2006) Bevisvurdering – Usikkerhet og Sannsynlighet, Oslo: Cappelen Damm.

Fenton, N. (2011) Science and law: Improve statistics in court, Nature, Vol.479, p.36-37.

Fienberg, S.E. (ed.) (1989) The Evolving Role of Statistical Assessments as Evidence in Courts, New York: Springer Verlag.

Finkelstein, M.O. and Levin, B. (1990) Statistics for Lawyers, New York: Spreinger-Verlag.

Gastwirth, J.L. (1992) Statistical Reasoning in the Legal Setting, The American Statistician, Vol.46(1), p.55-69.

Johnson, N.L., Kotz, S. and Balaskrishnan, N. (1994) Continuous Univariate Distributions, vol. 1, 2 ed., New York: John Wiley.

Johnson, N.L., Kotz, S. and Balakrishnan, N. (1995) Continuous Univariate Distributions, vol. 2, 2 ed., New York: John Wiley.

Magnussen, S. et.al. (2014) "Utover enhver rimelig tvil"? En kvantitativ studie av sikkerhet i bevisvurdering i straffesaker hos norske fagdommere og lekommere, Tidsskrift for rettsvitenskap 127, p. 347-365.

Meier, P. (1986) Damned liars and expert witnesses, Journal of the American Statistical Association, 81, 269-276.

Melsom, A. M. (2017) Sannsynlighetsberegninger- et nyttig kontrollverktøy, Skattedirektoratet: Analysenytt 02/2017. 7-11.

Sturtz, S., Ligges, U. and Gelman, A. (2005) R2WinBUGS: A Package for Running WinBUGS from R, Journal of Statistical Software, 12(3), 1-16.

Tiller, P.A. and Green, E. (eds.) (1988) Probability and Inference in the Law of Evidence: The Use and Limits of Bayesianism, Dordrecht: Kluwer Academic Publishers.

Aalen O. O. (2007) Statistical thinking in criminal cases. In Brantzæg & Eskeland: Rettsmedisinsk sakkyndighet i fortid, nåtid og fremtid, Oslo: Cappelen.

Appendix A: Distribution-theory

$S \sim Gamma(\alpha, \beta)$ denotes a sale amount S with a Gamma-distribution determined by parameters $(\text{shape} = \alpha, \text{scale} = \beta)$. The probability density is

$$f(s) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} \cdot s^{\alpha-1} e^{-\beta s} \; ; \; s \geq 0$$

with expectation $E(S) = \alpha/\beta$ and variance $V(S) = \alpha/\beta^2$. Here $\Gamma(\cdot)$ is the Gamma-function.

Let $S_i, i = 1,2,3, \ldots$ be consecutive sales. Assume they are independent with distribution

$$S_i \sim Gamma(\alpha_i, \beta) \text{ i=1,2,3,.. .}$$

i.e. with common scale-parameter β, but possible differing shape-parameters α_i.
Then the distribution of the sum

(1) $$S = S_1 + S_2 + \cdots + S_n \sim Gamma(\alpha_1 + \alpha_2 + \cdots + \alpha_n, \beta)$$

$R \sim Beta(\alpha_1, \alpha_2)$ denotes a ratio with a Beta- distribution with parameters determined by $(\text{shape1} = \alpha_1, \text{shape2} = \alpha_2)$. The probability density is

$$f(r) = \frac{r^{\alpha_1 - 1}(1-r)^{\alpha_2 - 1}}{B(\alpha_1, \alpha_2)} \; ; 0 \leq r \leq 1$$

with expectation $E(R) = \alpha_1/(\alpha_1 + \alpha_2)$. Here $B(\cdot\,,\,\cdot)$ is the Beta-function.

Let S_1 and S_2 be independent and distributed $S_i \sim Gamma(\alpha_i, \beta)$ i=1,2.

Then the distribution of the ratio

(2) $$R = \frac{S_1}{S_1 + S_2} \sim Beta(\alpha_1, \alpha_2)$$

This may be applied by letting S_1 and S_2 be the sum of sales of type 1 and 2 respectively, and let (α_1, α_2) be the sums of the respective α-parameters determined by use of (1) on each type.

In the case $S_i \sim Gamma(\alpha_i, \beta_i), \; i = 1,2$ with differing scale parameter, the ratio is related to the so-called Beta-prime distribution, and it follows that the probability density of R is

$$f(r) = \frac{r^{\alpha_1 - 1}(1-r)^{\alpha_2 - 1}}{B(\alpha_1, \alpha_2)} \cdot \frac{{\beta_1}^{\alpha_1} {\beta_2}^{\alpha_2}}{(r\beta_1 + (1-r)\beta_2)^{\alpha_1 + \alpha_2}}; \; 0 \leq r \leq 1$$

Here the first factor is ordinary Beta-density, which obtains in the case of $\beta_1 = \beta_2$, when the second factor is one.

For the Gamma-Beta model the likelihood ratio for a specific value r_1 of the true ratio, compared with another value r_0, may be computed by (with "hats" denoting the parameter estimates)

$$LR(r_1 : r_0) = \frac{r_1^{\hat{\alpha}_1 - 1}(1-r_1)^{\hat{\alpha}_2 - 1}}{r_0^{\alpha_1 - 1}(1-r_0)^{\alpha_2 - 1}} = \left(\frac{r_1}{r_0}\right)^{\hat{\alpha}_1 - 1} \left(\frac{1-r_1}{1-r_0}\right)^{\hat{\alpha}_2 - 1}$$

The corresponding likelihood ratio for the Gamma-Beta-prime model follows accordingly.

Appendix B: A Supreme Court ruling on sample evidence

The Norwegian Taxes Management Act of 2016 or its regulations (by 2018) do not spell out the evidence requirements for discretionary assessment and penalty tax. However, a Supreme Court ruling of 2017, related to a restaurant and its cash sales, provides the basis for current practice. In fact, there is no real change from the prior tax laws (the Tax Assessment Act and the VAT Act), and a similar Supreme Court ruling of 2008[10]. The Act adheres to the European Convention on Human rights (6.2): "Everyone charged with a criminal offence shall be presumed innocent until proved guilty according to law".

The case before the Supreme Court in 2017 was an appeal by the State on a verdict in lower courts, in favor of the plaintiff (i.e. the restaurant). The main evidence was the data described in the end of the last section. The issues before the Supreme Court were:

(i) Was a discretionary assessment of the tax office legitimate at all?
(ii) Was the actual discretionary assessment legitimate and defensible?
(iii) Was the assigned penalty tax legitimate?

With respect to (i), the first voting judge remarked that discretionary assessment is entitled only when the violations of laws and accounting regulations in total provides an insufficient basis for the taxation. The judge accepted the sampled evidence as part evidence, contrary to the decision in lower court. In total, a discretionary assessment was found legitimate.

With respect to (ii), the judge stated that the tax office obligation is to fixate the factual basis for the taxation by a free judgment of all available evidence, and get as close to the correct basis as possible, and so that the result does not appear arbitrary or apparently unreasonable. Within these limits, the tax office will have a considerable freedom, and the court limited opportunity to overrule the actual assessment by the tax office. In this case, the judge found the tax office assessment based on the nationwide sample of restaurant was appropriate, based on solid data, and performed with caution. The assessment neither arbitrary nor unreasonable, with a fixation of cash-ratio like the lowest one in the sample. Remark. Here it would have been more clarifying if the Supreme Court had a case where the tax office had settled on the mean or median of the sample. Taking the law literally, this should be ok'd as well. Looking at some lower court verdicts, it seems that "unreasonable" is given a connotation of safeguarding against a felt burden on the taxpayer.

With respect to (iii), the penalty tax, clear and convincing evidence is required, for the assignment itself, and its size. The latter means that, in some cases, it has to be derived from a lower amount than the reassessed basis for the ordinary tax. In the current case, the tax-office had already accounted for this, by assigning cash-ratios at the low end of the nationwide sampled distribution. With some doubt, the first voting judge agreed with this.

In summary, the first voting judge had answered yes on all three questions (i), (ii) and (iii) above, and he voted for exonerate the State from the charges made the plaintiff. The four other judges agreed, in essence. The Supreme Court therefor ruled in favor of the State.

[10] Norwegian Supreme Court: HR-2017-967-A, HR-2008-01861-S

Appendix C: On the use of statistical models in court

Spiegelhalter:

"In general, I don't feel statistical evidence is handled well by the courts. They like either incontrovertible numerical "facts", or overall expert opinions. But statisticians deal with a delicate combination of data and judgment that often gives rise to "rough" numbers, and these don't seem to fit well with the legal profession".

In the 1980's statisticians in the US and UK became increasingly aware of misuse of statistics in courts with grave consequences, and initiatives were taken. In the US, a panel involving statisticians, social scientists and lawyer produced a report Fienberg (1989) with cases and recommendations for the US. At the same time, academic papers and books on statistics and the law started to emerge, e.g. DeGroot et.al. (1986), Meier (1986), Tiller and Green (1988), Gastwirth (1992), Finkelstein and Levin (1990). Tax fraud and statistics are hardly a theme, see however Bright et.al. (1988).

In the UK, the Royal Statistical Society initiated, in cooperation with the legal profession, the project "Communicating and Interpreting Statistical Evidence in The Administration of Criminal Justice". Out of this came four guides for judges, lawyers, forensic scientists and expert witnesses, the first is Aitken et. al. (2010). An abbreviated version is COIC & RSS (2017), where reference to all four guides. See also Fenton (2011).

In Norway, the acceptance of statistical evidence has been less controversial than in the US and UK. Occasionally, Norwegian statisticians have served as expert witnesses, but contributions to the academic literature is rare (see however Aalen, 2007), and there has been no national initiatives. The book by Eide (2016) gives a broad account on probability evidence, favoring the Bayesian view. Wide references are given to differing views in the law profession, in Norwegian as well as international literature.

The law profession have increasingly realized that issues of strength of evidence is probabilistic and statistical. This has also influenced the teaching of law students. In the future, the tax-offices and the courts may be better prepared to handle cases in this area.

A general issue is how advanced statistical methods could be used in the preparation of a case that eventually may end up in court. There is a risk that too advanced statistics will be dismissed as evidence. Norway has been ahead of many other countries in accepting statistical arguments in court.

From a District Court verdict (TOSLO-2014-193468):

"The discretionary assessment is based on an expedient and acceptable method. The court remarks that the assessment could, in principle, be more advanced, by econometric modelling and analysis, replacing simple average calculations. However, more advanced methods cannot represent the minimal requirement. At the same time the court also realize the value of more sophisticated statistical models".