Analysing foodborne outbreaks in the USA

Project for Design of Experiments – by Renske Bouma

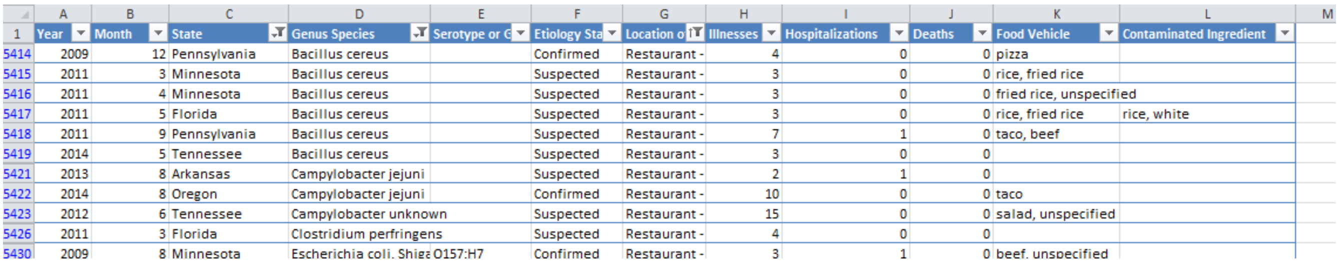
Introduction

Food is vital to life, but can also cause illness or even death. Food can namely be a carrier of dangerous micro-organisms, which then will result in foodborne disease. According to the World Health Organisation (WHO) a foodborne disease is defined as: 'Any disease of an infectious or toxic nature caused by, or thought to be caused by, the consumption of food or water'. This definition also includes diseases caused by non-microbial substances, like harmful pesticides or processing chemicals. Most common are however the illnesses caused by micro-organisms and their toxins (Adams & Moss 2008) and these will be the focus of this report. A foodborne outbreak is defined as: 'An incident in which two or more persons experience a similar illness resulting from the ingestion of a common food' (CDC 2000). Often enough, the outbreaks are larger than two illnesses. In this report foodborne outbreaks in the USA are investigated for the dependence of average outbreak size on location of preparation of the food vehicle, the micro-organism that caused the disease and the state it occurred in.

The database

The database I use is put together by the Centers for Disease Control and Prevention (CDC). This is the organisation in the USA that is working towards a better public health. To know how to do this, the CDC needs to know where the problems lay and therefore it monitors the prevalence of diseases, like foodborne diseases. It created the FOOD tool, the Foodborne Outbreak Online Database (CDC 2015), in which all reported cases of foodborne outbreaks that were reported to the CDC since 1998 are included. I downloaded an extensive excel file from their website to use for the statistical analysis. The CDC warns that the database is not final, reports can still be changed when new information is gathered. The database I used was lastly updated on 16 October 2015. The newer reports could therefore be reflecting the true outbreak less than the older reports, which could lead to systematic errors. However, the database does not contain reports newer than 2014, so also the newest outbreaks had almost a year to be fully reported.

The database includes the following information from every outbreak (see Figure 1) : the year and month it occurred, the state, the specie/ species that (probably) caused the disease, the serotype (if known) of the micro-organism, the etiology status (confirmed or only suspected origin), the location(s) of preparation of the infected food, the resulting illnesses, hospitalizations and deaths, the food vehicle and the contaminated ingredient in this food item. To simplify I only use year, state, genus, location of preparation and resulting illnesses in my analysis. These are the factors that can be grouped most easily in groups that are still big enough for analysis and that seem the most interesting to me.



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Year	Month	State	Genus Species	Serotype or G	Etiology Sta	Location o	Illnesses	Hospitalizations	Deaths	Food Vehicle	Contaminated Ingredient	
5414	2009	12	Pennsylvania	Bacillus cereus		Confirmed	Restaurant -	4	0	0	pizza		
5415	2011	3	Minnesota	Bacillus cereus		Suspected	Restaurant -	3	0	0	rice, fried rice		
5416	2011	4	Minnesota	Bacillus cereus		Suspected	Restaurant -	3	0	0	fried rice, unspecified		
5417	2011	5	Florida	Bacillus cereus		Suspected	Restaurant -	3	0	0	rice, fried rice	rice, white	
5418	2011	9	Pennsylvania	Bacillus cereus		Suspected	Restaurant -	7	1	0	taco, beef		
5419	2014	5	Tennessee	Bacillus cereus		Suspected	Restaurant -	3	0	0			
5421	2013	8	Arkansas	Campylobacter jejuni		Suspected	Restaurant -	2	1	0			
5422	2014	8	Oregon	Campylobacter jejuni		Confirmed	Restaurant -	10	0	0	taco		
5423	2012	6	Tennessee	Campylobacter unknown		Suspected	Restaurant -	15	0	0	salad, unspecified		
5426	2011	3	Florida	Clostridium perfringens		Suspected	Restaurant -	4	0	0			
5430	2009	8	Minnesota	Escherichia coli, Shiga	O157:H7	Confirmed	Restaurant -	3	1	0	beef, unspecified		

Figure 1 the original database

Hypothesis

I would like to know where a mistake causes the most illnesses. Does a mistake by a caterer cause more illnesses on average than a mistake at a banquet? Does an outbreak at a restaurant cause more illnesses than one at health care? Next to that, I am curious if the other factors, genus and state, play a role as well. Which genus causes the most illnesses per outbreak? Are there differences between states in how big the outbreaks are?

Experimental design

I want to know whether or not the differences in amount of illness per outbreak between different locations of preparations, genera and states are significant or not. Is it simply because of chance that they look different or is it likely that there is a real difference? To know this I will analyse the variance of the data with an one-way balanced ANOVA in R. Before I can do this, I have to structure the data.

Structuring of the data

To get a clear result I removed all the data-points with multiple possible species, multiple location of preparations and the multistate outbreaks. I grouped the different species of the most common genera as displayed in Table 1 and omitted all the other data-points from less common genera. I also grouped different location of preparations as displayed in Table 2 and left out all other data-points from less common locations (like camps and festivals). None of the groups has less than 100 data-points, which I believe gives a good reliability.

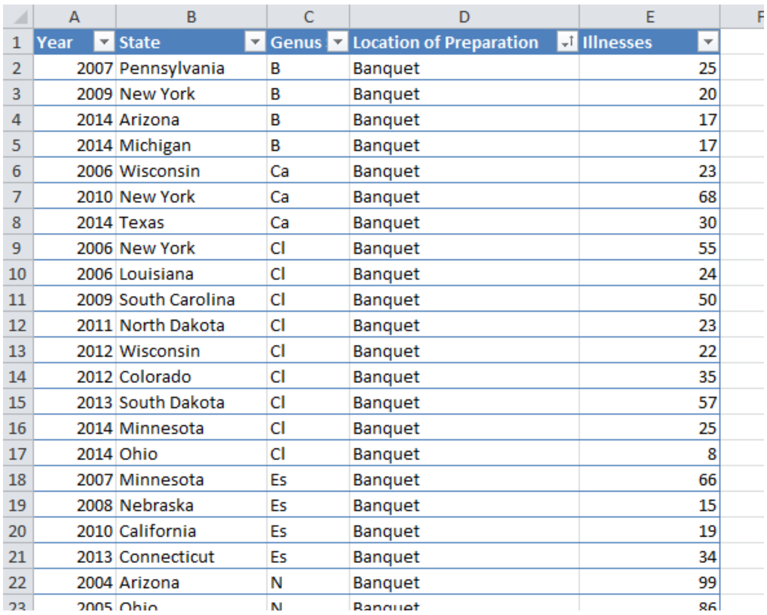
Table 1 grouping of different species in their respective genus

group	contains	data-points
Bacillus (B)	*B. cereus, B. other, B. unknown*	246
Campylobacter (Ca)	*C. jejuni, C. coli, C. fetus, C. other, C unknown*	185
Clostridium (Cl)	*C. perfringens, C. botulinum*	547
Escherichia (Es)	*E. coli*, enteroaggregative, *E. coli* enteropathogenic, *E. coli* other, *E. coli* shiga toxin-producing	222
Norovirus (N)	*Norovirus, Norovirus Genogroup 1, Norovirus Genogroup 2, Norovirus unknown*	3729
Salmonella (Sa)	*Salmonella, S. enterica, S. other, S. unknown,*	1335
Shigella (Sh)	*Shigella, S. boydii, S. dysenteriae, S flexneri, S. sonnei, S. unknown*	112
Staphylococcus (St)	*S. aureus, S. other, S. unknown*	415

Table 2 grouping of different locations of preparation

group	contains	data-points
Banquet	Banquet facility (food prepared and served on-site)	210
Caterer	Caterer (food prepared off-site from where served), Caterer; unknown, Caterer; other	687
Health care	Hospital, Long-term care/nursing home/assisted living facility, long-term care..; Hospital, long-term care..; Other	162
Private home	Private home/residence	861
Restaurant	Restaurant- “Fast food” (drive up service or pay at counter), Restaurant- other or unknown type, Restaurant – other or unknown type; Other, Restaurant – Sit-down dining, Restaurant – sit-down dining; Other	4871

After this structuring and ‘cleaning’, the data-set looked like shown in Figure 2. The dataset still contained more than 6500 data-points.



	A	B	C	D	E	F
1	Year	State	Genus	Location of Preparation	Illnesses	
2	2007	Pennsylvania	B	Banquet	25	
3	2009	New York	B	Banquet	20	
4	2014	Arizona	B	Banquet	17	
5	2014	Michigan	B	Banquet	17	
6	2006	Wisconsin	Ca	Banquet	23	
7	2010	New York	Ca	Banquet	68	
8	2014	Texas	Ca	Banquet	30	
9	2006	New York	Cl	Banquet	55	
10	2006	Louisiana	Cl	Banquet	24	
11	2009	South Carolina	Cl	Banquet	50	
12	2011	North Dakota	Cl	Banquet	23	
13	2012	Wisconsin	Cl	Banquet	22	
14	2012	Colorado	Cl	Banquet	35	
15	2013	South Dakota	Cl	Banquet	57	
16	2014	Minnesota	Cl	Banquet	25	
17	2014	Ohio	Cl	Banquet	8	
18	2007	Minnesota	Es	Banquet	66	
19	2008	Nebraska	Es	Banquet	15	
20	2010	California	Es	Banquet	19	
21	2013	Connecticut	Es	Banquet	34	
22	2004	Arizona	N	Banquet	99	
23	2005	Ohio	N	Banquet	86	

Figure 2 the structured data-set in Excel

First look at the data

Before starting to go into the statistics, I have a look at the raw data. The mean of all illnesses from all outbreaks is 20.33. Plotting the data gives an idea about the range the amount of illnesses per outbreak can be in. In Figure 3 the outbreak size (the amount of illnesses per outbreak) for the different locations is given. It can be seen that most of the outbreaks result in less than 100 illnesses, but there are many exceptions. Five outbreaks were relatively extreme and resulted in more than 600 reported ill people. Most outliers are seen at the restaurant, but no conclusions can be drawn from this, as the amount of data-points are not equal for all locations. The data-set contains by far the most points for restaurants, which can explain the bigger range of points within this group. When only looking at the boxplots, there seems to be a difference between the first three locations and the last two.

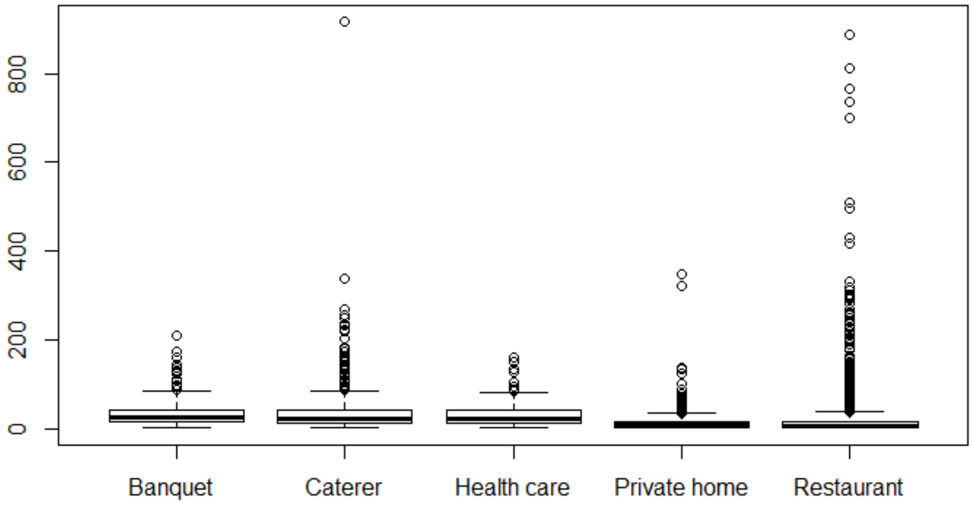


Figure 3 boxplot of outbreak size versus location of preparation

The boxplots per genus is given in Figure 4. Of course most of the outbreaks are again below 100 illnesses, as the same data is plotted as before, just in a different grouping. Interesting to see is where the outliers are. *Escherichia* seems at first sight to be quite dangerous, because of the high outliers, but the boxplot itself is quite small. *Norovirus* has an exceptional amount of outliers, but just as for the restaurants, no conclusion can be drawn from this. The *Norovirus* group contains the most data-points from all groupings, so this could explain the higher amount of outliers. When just looking at the boxplot, it seems that *Clostridium*, *Norovirus* and *Shigella* are resulting in the more illnesses than the other genera. We will see if this difference is seen as well from the ANOVA.

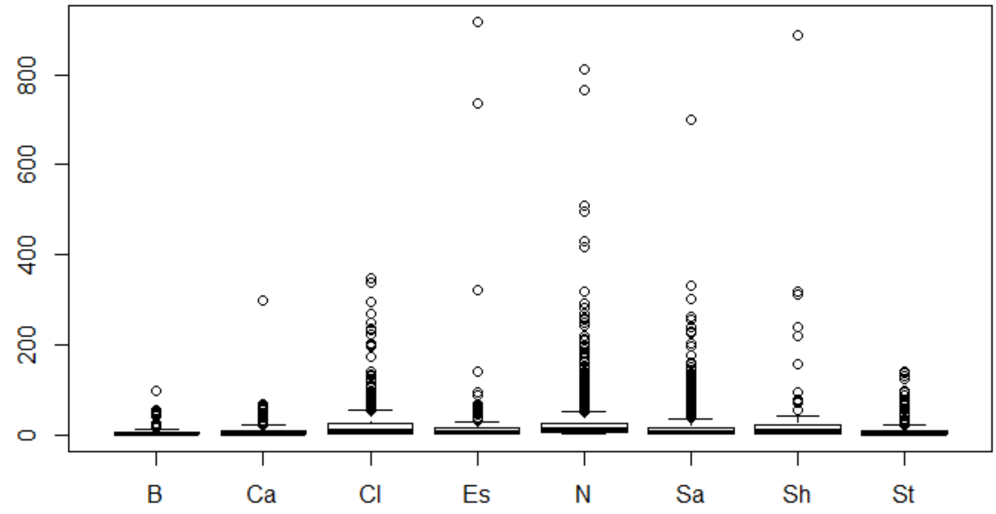


Figure 4 boxplot of outbreak size versus genus

The boxplot per state is given in Figure 5. It is clear that there are differences between the states. We will find out later whether or not these differences are significant or not.

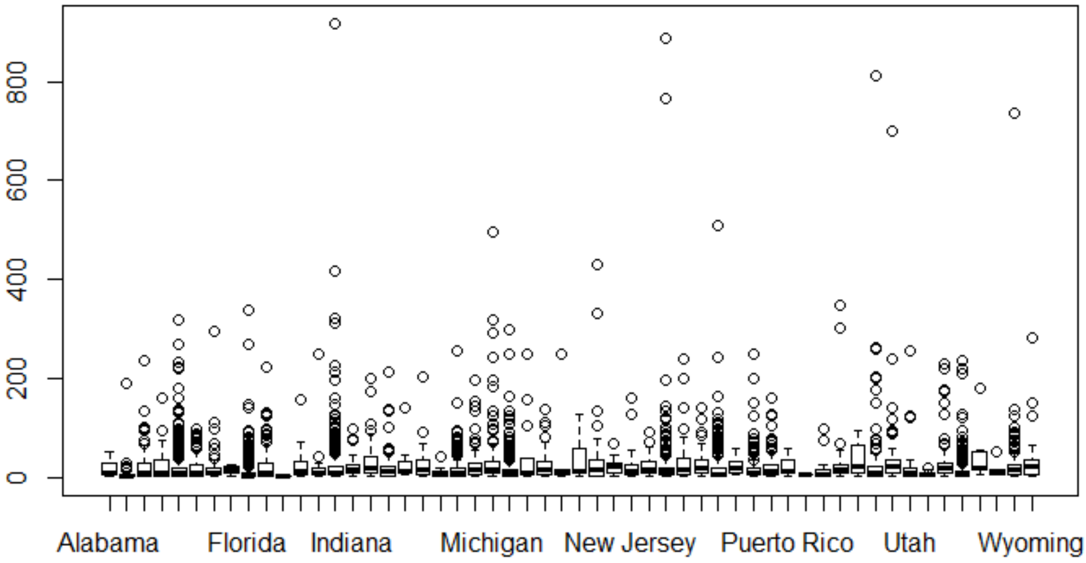


Figure 5 boxplot of outbreak size versus state

Hypothesis testing

My main interest is whether or not the location of preparation of the food that causes the outbreak has a significant influence on the outbreak size. So I perform an ANOVA in R on the location. This is the output:

```
               Df   Sum Sq Mean Sq F value Pr(>F)    
Location        4   333252   83313   54.51 <2e-16 ***
Residuals    6786 10371990    1528                   
---
Signif. codes:  
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The result shows that at least between two locations there is a significant difference in average outbreak size. This difference is so extreme that the probability that this would occur under the null hypothesis (there is no difference) is less than 0.1%.

Validation of assumptions

The interpretation of the ANOVA is only valid when the assumptions are acceptable. The result of ANOVA is meaningful when it can be assumed that the residues are normally, independently and identically distributed (NIID). Normality can be checked by making a Q-Q plot. When the residues are more or less normally distributed a straight line is observed. In Figure 6 it can be seen that this is not the case. A non-parametic test should be used or the data should be transformed so that the normality assumption becomes valid.

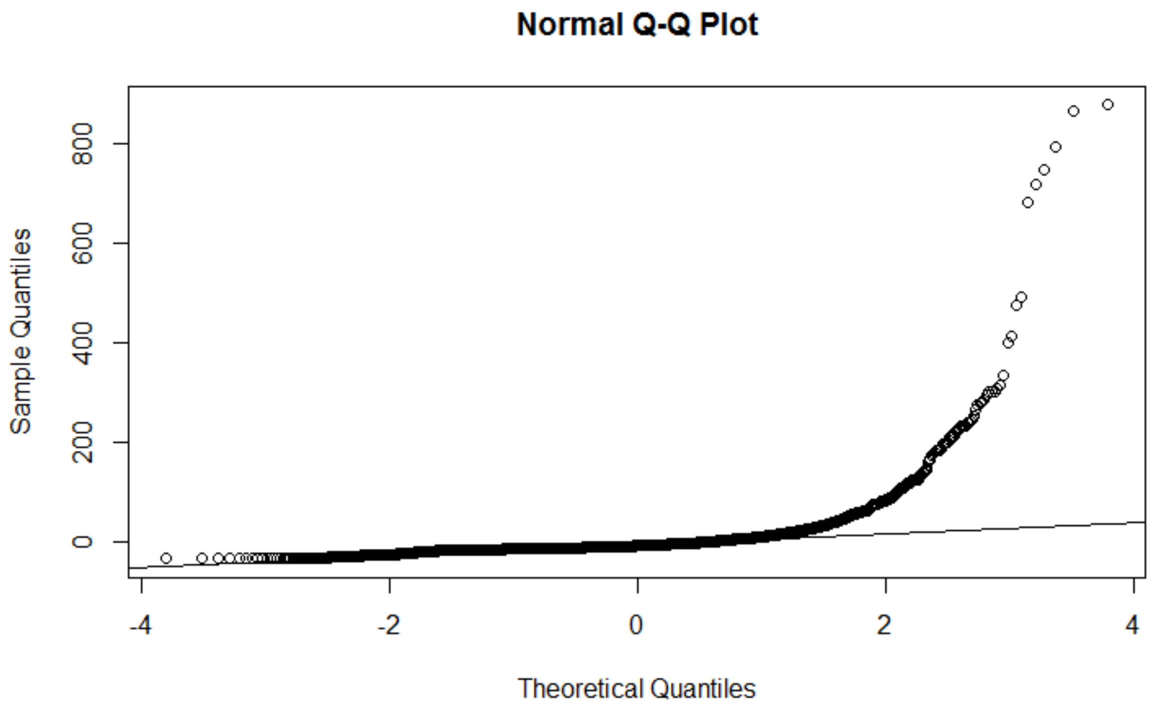


Figure 6 Q-Q plot of residues from ANOVA of location

Hypothesis testing

In the book (Box et al. 2005) it was suggested that taking the logarithm of the output (here outbreak size) can stabilize the variance. I perform the ANOVA with the log-transformed data. This is the output from R:

```
              Df Sum Sq Mean Sq F value Pr(>F)    
Location       4    846  211.52   209.2 <2e-16 ***
Residuals   6786   6861    1.01                   
---
Signif. codes:  
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The result is the same as before: there is a very significance difference. Let us see if this time we can trust the result.

Validation of assumptions

Again, I make a Q-Q plot to check if the residues are normally distributed. As can be seen in Figure 7 the dots are more or less on one line. Normality can therefore be assumed.

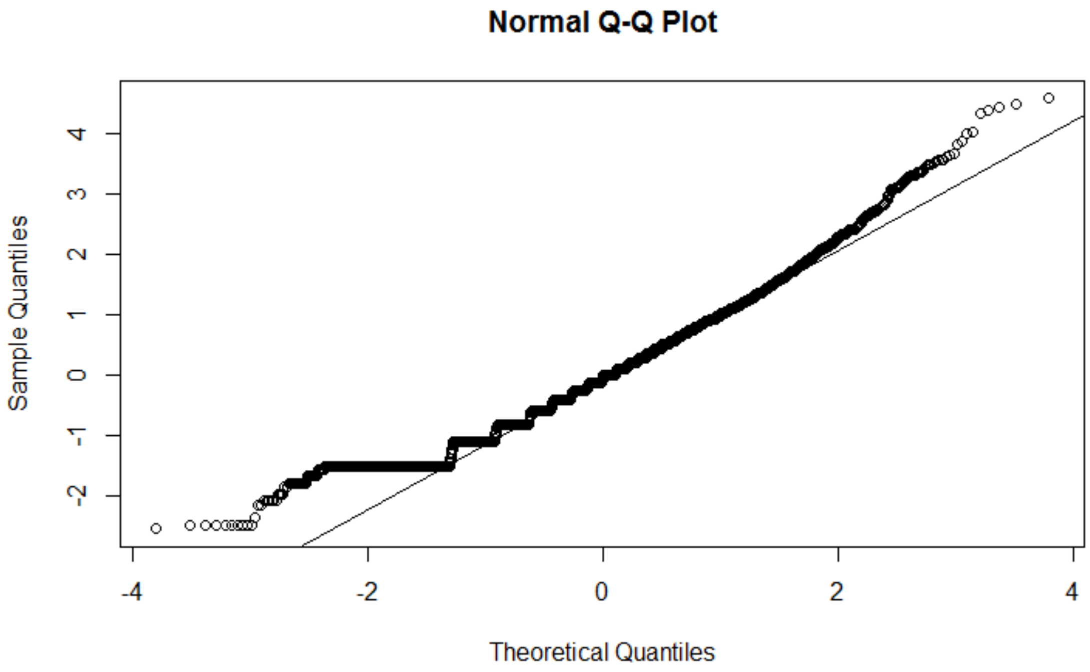


Figure 7 Q-Q plot of residues of the ANOVA of location after transformation

The residues should also be approximately independently and identically distributed. In Figure 8 it can be seen that the residues are spread more or less the same over the years. There is no trend, so probably there is no autocorrelation between the data-points. This means the assumption of independence can be made.

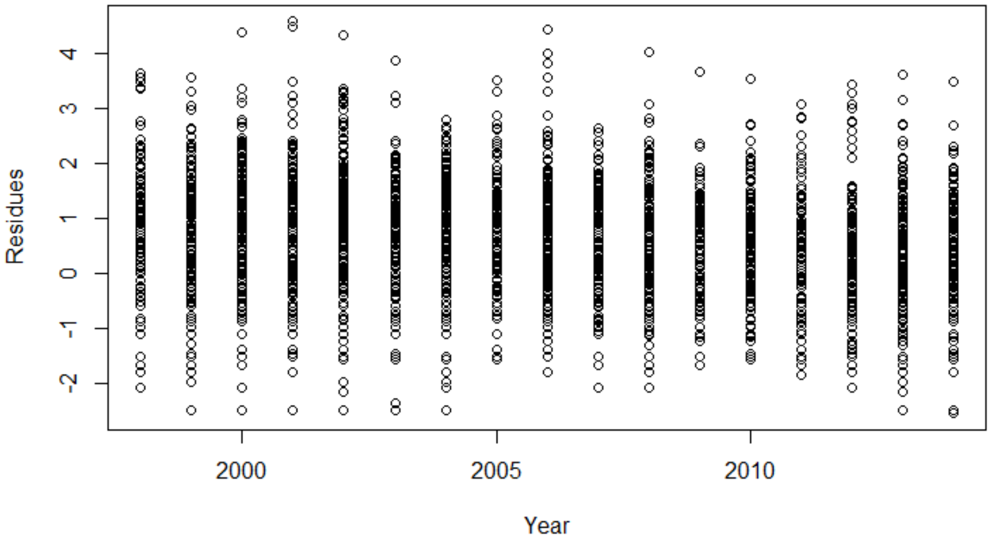


Figure 8 plot of residues in time order

In Figure 9 the residues per expected value are plotted. Curious enough, only four lines appear, while I am researching five locations. Apparently two locations have an expected value very close to each other. The spread of residues should be approximately the same per expected value to be able to assume identical distribution. The spread is not identical here, but it is close enough.

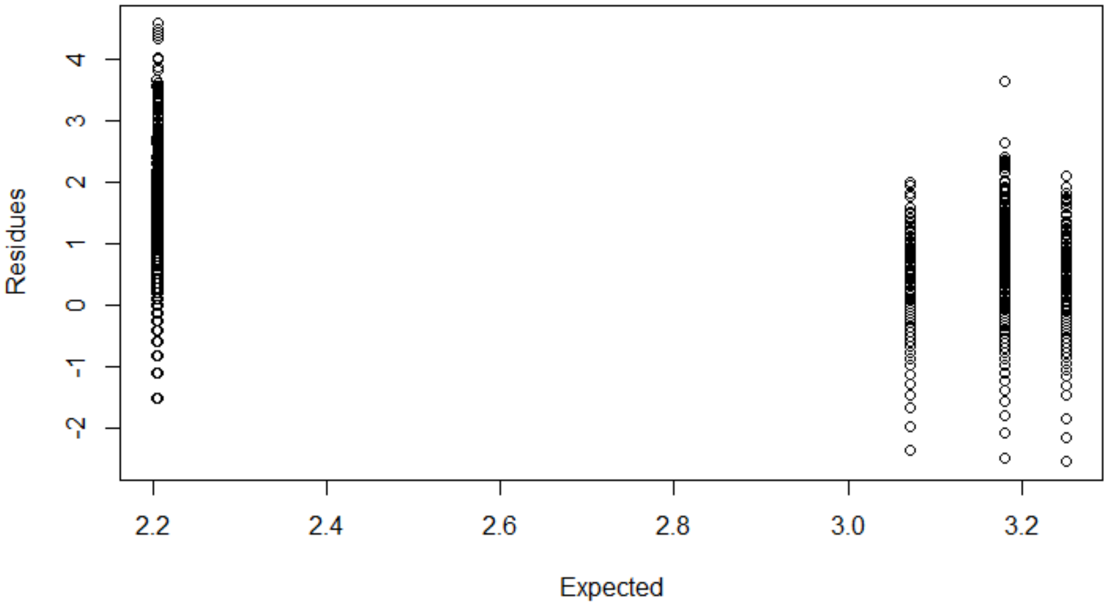


Figure 9 plot of residues versus expected value

Hypothesis testing

Now I know I found an useful transformation of my data I do some more hypothesis testing with the transformed data. I already saw a significant difference between different location of preparations. Now I am also curious if there are differences between genera. In the ANOVA table below it can be seen that the different genera do not result in the same average outbreak size.

```
Df Sum Sq Mean Sq F value Pr(>F)
Genus          7    648   92.50   88.88 <2e-16 ***
Residuals   6783   7060    1.04
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

I also would like to know if there are significant differences between states. From the ANOVA table below it is clear that there are very significant differences between states.

```
             Df Sum Sq Mean Sq F value Pr(>F)
State         53    828  15.620    15.3 <2e-16 ***
Residuals   6737   6879   1.021
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The plots to check the assumptions for these residuals are left out for space reasons, but can be found in the appendix. They give no reason to doubt any of the assumptions.

Blocking?

We have seen that there are differences in average outbreak size between outbreaks in different states, from different locations of preparation and by different genera. Combining the effects in one ANOVA so that the residues are explained by all factors would reduce the unexplained remaining residues. This would increase the sensitivity of the ANOVA test and with that the significance of the results. However, the results are already as significant as it gets, so blocking is not necessary.

Graphical ANOVAs

To see where the differences exactly lie I perform a graphical ANOVA for the locations, genera and states. To know which point in the graph correspondents to what I also show the averages in a table. Next to this, by looking at the real averages (and not the transformed ones) in these tables we can see whether or not a significant difference is also an interesting one. With such a large data-set a difference is already quite fast a significant one, but if the difference is small, this might not be very interesting.

Locations

The graphical ANOVA for the different locations of preparation can be seen in Figure 10 and the averages are given in Table 3.

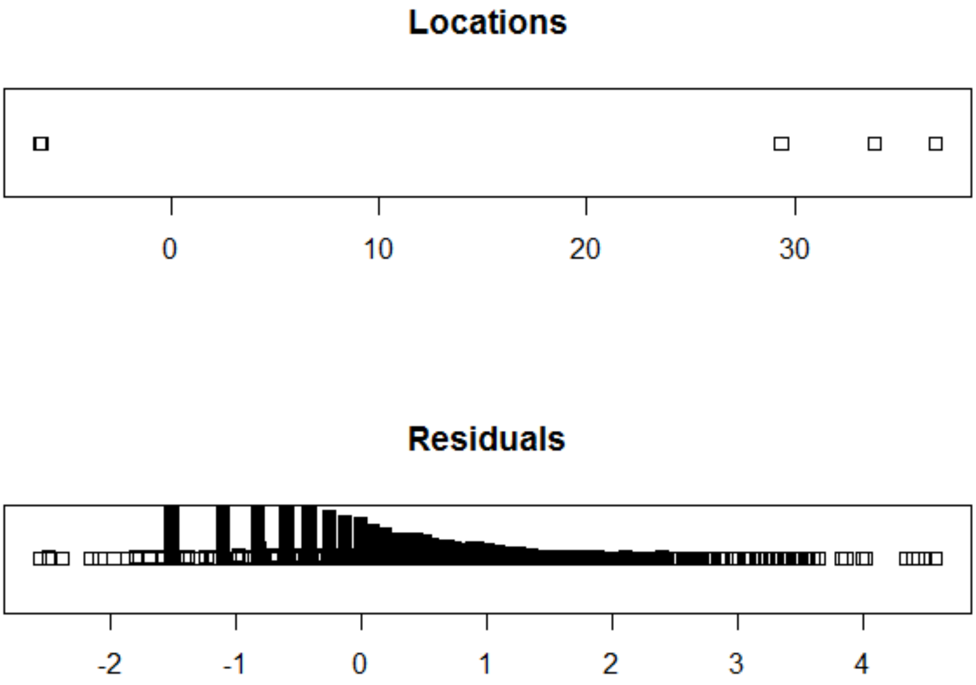


Figure 10 Graphical ANOVA for locations

Table 3 average outbreak size per location of preparation

Location of preparation	Average outbreak size
Banquet	36.30
Caterer	37.37
Health care	31.68
Private home	14.31
Restaurant	17.93

The residuals are spread from -3 to 5 while the difference between locations is as big as 40. It can therefore be seen clearly from this graphical ANOVA that there is a significant difference between some locations. The low point in the graph is actually both the point for private home and the restaurant. They are so close to each other that they cannot be seen as individual points. Thus they are virtually the same. They are however very different from the other locations: banquet, caterer and health care. This makes sense as at a banquet and a caterer food is provided to big groups of people. At health care the amount of people eating the food is not as big as for caterer or banquet, but as the people eating the food are fragile, they are more prone to get sick from a food contamination.

Genera

The graphical ANOVA for the different genera causing the illnesses can be seen in Figure 11Figure 10 and the averages are given in Table 4. The order of average outbreak size from smallest to largest is given to make it easier to see what the different genera have as effect on the average outbreak size.

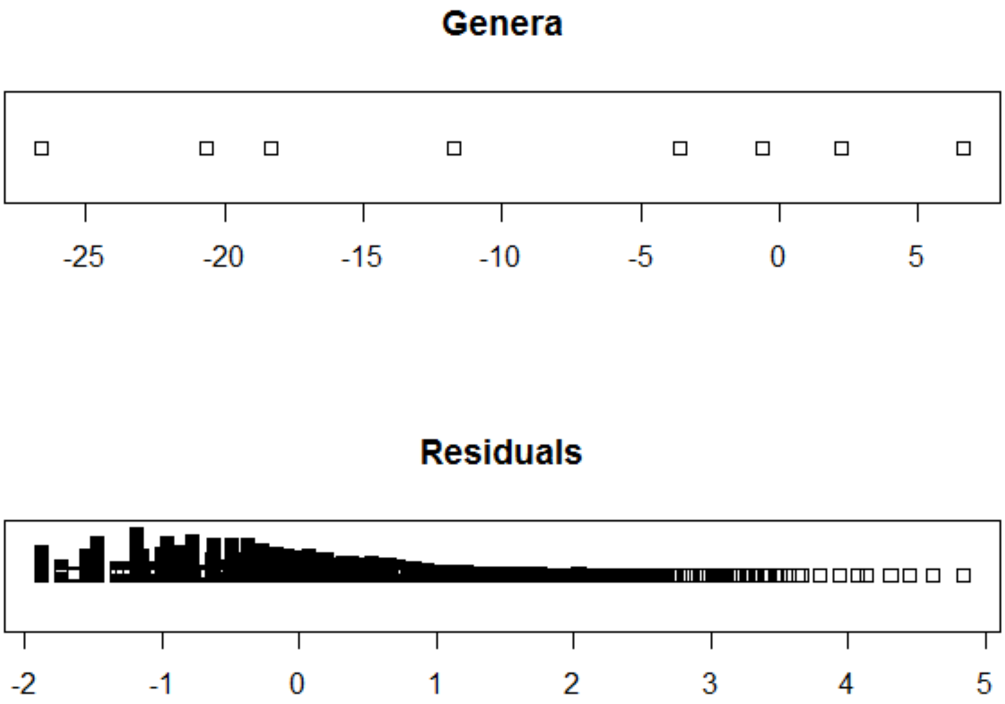


Figure 11 graphical ANOVA for genera

Table 4 average outbreak size per genus causing it

Genus	Average outbreak size	Order
Bacillus (B)	7.34	1
Campylobacter (Ca)	10.79	2
Clostridium (Cl)	24.54	6
Escherichia (Es)	21.44	5
Norovirus (N)	22.34	7
Salmonella (Sa)	18.23	4
Shigella (Sh)	33.39	8
Staphylococcus (St)	11.38	3

Bacillus outbreaks result on average in the least illnesses while *Shigella* outbreaks cause the most. The big average outbreak size by *Shigella* could be explained by the fact that the infectious dose is low. In other words, a few cells are already enough to cause a disease. Commonly, Shigella is spread person-to-person, but when food is prepared by personnel that caries the bacterium, the food can get infected (Adams & Moss 2008). In such situations often many people get ill as one sick person in the personnel of for example a catering company handle the food for many. Also, people that get ill from *Shigella* are likely to report themselves, as the symptoms of shigellosis are often quite extreme and need medical attention.

Bacillus is a very common food pathogen. It forms spores by which it can survive harsh conditions. Different species of *Bacillus* form different enterotoxins which can result in two different illnesses: diarrhoeal and emetic syndrome. Both illnesses are often over in less than 24 hours and the symptoms are in most cases quite normal, like vomiting and diarrhoea (Adams & Moss 2008). As most people do not report these kinds of symptoms if they are over in a day, it is very likely that the amount of illnesses of an outbreak of a *bacillus* specie are under reported. It is however also possible that outbreaks are really smaller as individual products can be the source of an outbreak (instead of personnel handling food for many). For example, *B. cereus* can survive in pasteurized milk, but will only grow and produce toxins when stored at too high temperature(Adams & Moss 2008). In this way it is possible that only one package becomes unsafe. Solely the people eating from that one package then get sick (like a family). This results in smaller outbreaks.

States

The graphical ANOVA for the different states where the outbreaks occurred can be seen in Figure 12Figure 10 and the averages are given in Table 5 on the next page. As there are many states, the most interesting ones are highlighted.

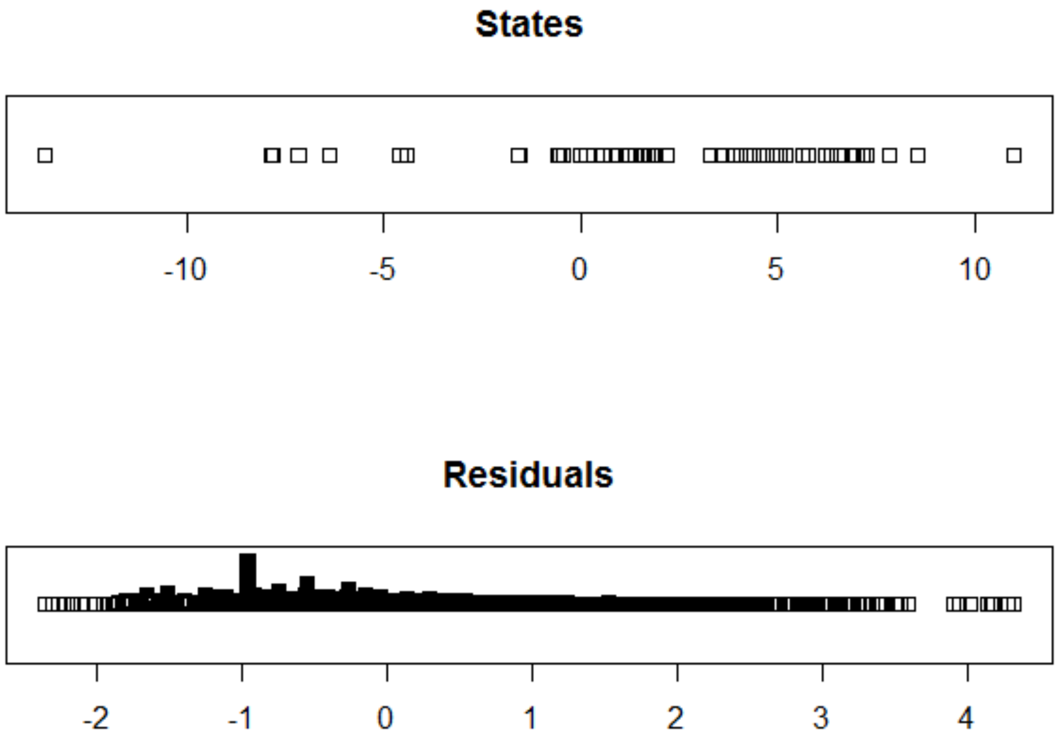


Figure 12 graphical ANOVA for states

Table 5 Average outbreak size per state it occurred in

State	Average outbreak size	State	Average outbreak size
Alabama	18.25	Nebraska	32.93
Alaska	12.78	***Nevada***	***51.45***
Arizona	24.33	New Hampshire	21.74
Arkansas	28.27	New Jersey	21.02
California	17.58	New Mexico	24.89
Colorado	18.46	New York	25.37
Connecticut	15.89	North Carolina	31.21
Delaware	17.25	North Dakota	32.06
Florida	10.87	Ohio	18.71
Georgia	24.60	Oklahoma	23.62
Guam	***3.50***	Oregon	16.63
Hawaii	22.69	Pennsylvania	20.22
Idaho	19.71	Puerto Rico	22.18
Illinois	24.98	***Republic of Palau***	***6.00***
Indiana	21.63	Rhode Island	17.53
Iowa	31.23	South Carolina	29.57
Kansas	20.51	South Dakota	36.83
Kentucky	30.45	Tennessee	29.62
Louisiana	31.25	Texas	47.88
Maine	10.19	Utah	33.97
Maryland	18.49	***Vermont***	***7.40***
Massachusetts	27.59	Virginia	28.89
Michigan	35.71	Washington	13.68
Minnesota	15.90	Washington DC	48.86
Mississippi	**50.62**	West Virginia	15.63
Missouri	28.08	Wisconsin	24.08
Montana	57.40	Wyoming	44.09

Two striking averages are those for Guam and the Republic of Palau, with only 3.5 and 6 illnesses respectively on average per outbreak. A quick google explains why. Guam is a little island far east of the Philippines and is inhabited by less than 200.000 people (Wikipedia 2016a). Palau is another island relatively close to Guam and has even less people: about 25.000 (Wikipedia 2016c). The small amount of people and the distance between the mainland and the islands probably explain the small outbreaks. As it is a colony of the USA with quite a different culture, the eagerness to report is likely to be smaller. Also there are simply less people to get ill.

The mainland state, Vermont, becomes then interesting with an average outbreak size of only 7.4. Vermont has a good reputation when it comes to public health. Vermont got first rank for health outcomes in the USA in 2010. From 2000 to 2008 Vermont was ranked as the healthiest place to live seven out of eight times (Wikipedia 2016d). The low average outbreak size fits in this picture.

Nevada and Mississippi are on the complete other side of the range, with 51.5 and 50.6 illnesses respectively on average per outbreak. Mississippi is infamous for its health care. It was given the lowest rank for health care among all the American states by the Commonwealth Fund (Wikipedia 2016b). The large average outbreak size in Nevada might be caused by the popularity of Las Vegas. Massive scaled buffets are very common in Las Vegas, which means that if there is an outbreak, many (tourists) will get sick at once.

Conclusion and discussion

It can be concluded that foodborne outbreaks differ in size depending on location of preparation of the food, micro-organism causing the illness and the state it occurs in. Size in this case refers to the reported amount of illnesses. The question is however how realistic the reports reflect the real outbreak sizes. Probably, all outbreaks are underreported, but some might be more under reported than others, which is problematic as this might create significant differences where actually there are none. Next to this, I want to mention that a bigger outbreak size does not directly say something about the seriousness of the outbreak. A big outbreak could mean that 50 people had to vomit ones, and a small one could mean that 10 people died. This report is simply and only about the amount of people affected per outbreak.

Another remark I would like to make is the fact that I omitted data because of ambiguity. It could be that a specific genus is hard to distinguish, but actually creates big outbreaks. This would not be seen in this analysis because all data-points with multiple possible micro-organisms causing it, were not taken into account. The omitting of data does have an advantage considering the conclusions about location of preparation. The fact that unsafe food was prepared somewhere does not say that something went wrong in that particular location. It might be that at the factory, or at the farm, or during transport something went wrong by which unsafe food was created. However, if food becomes unsafe in an early step in the production, it is likely to end up in different places. By omitting the data-points with multiple locations, it is more probable that it was actually a mistake at that location causing the disease in the remaining data-points. With this it is possible to make more reliable conclusions about where a mistake results in the biggest foodborne outbreak.

References

Adams, M.R. & Moss, M.O., 2008. *Food Microbiology* 3th ed., The Royal Society of Chemistry.

Box, G.E.P., Hunter, J.S. & Hunter, W.G., 2005. *Statistics for Experimenters* 2nd ed., Wiley-Interscience.

CDC, 2000. Appendix B Guidelines for Confirmation of Foodborne-Disease Outbreaks. Available at: http://www.cdc.gov/mmwr/preview/mmwrhtml/ss4901a3.htm [Accessed May 23, 2016].

CDC, 2015. Foodborne Outbreak Online Database (FOOD Tool). Available at: http://wwwn.cdc.gov/foodborneoutbreaks/ [Accessed May 20, 2016].

Wikipedia, 2016a. Guam. Available at: https://en.wikipedia.org/wiki/Guam [Accessed May 20, 2016].

Wikipedia, 2016b. Mississippi - health. Available at: https://en.wikipedia.org/wiki/Mississippi#Health [Accessed May 20, 2016].

Wikipedia, 2016c. Palau. Available at: https://en.wikipedia.org/wiki/Palau [Accessed May 20, 2016].

Wikipedia, 2016d. Vermont - Public health. Available at: https://en.wikipedia.org/wiki/Vermont#Public_health [Accessed May 20, 2016].

Appendix

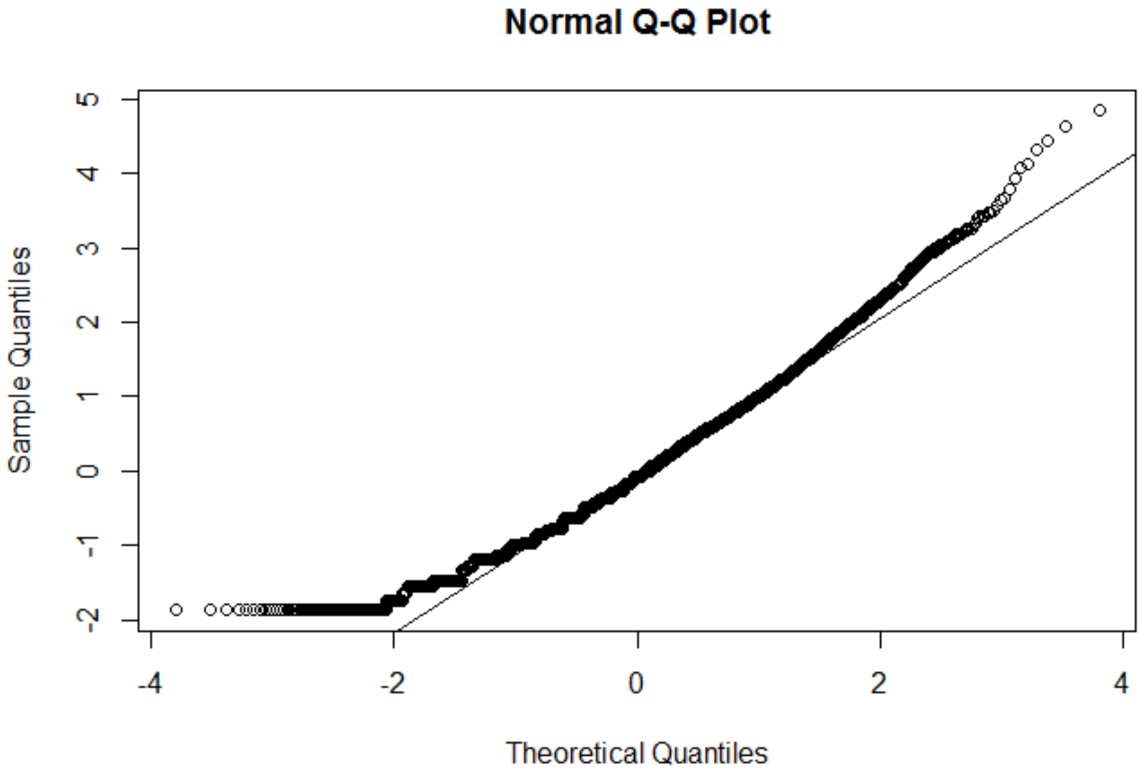
The code

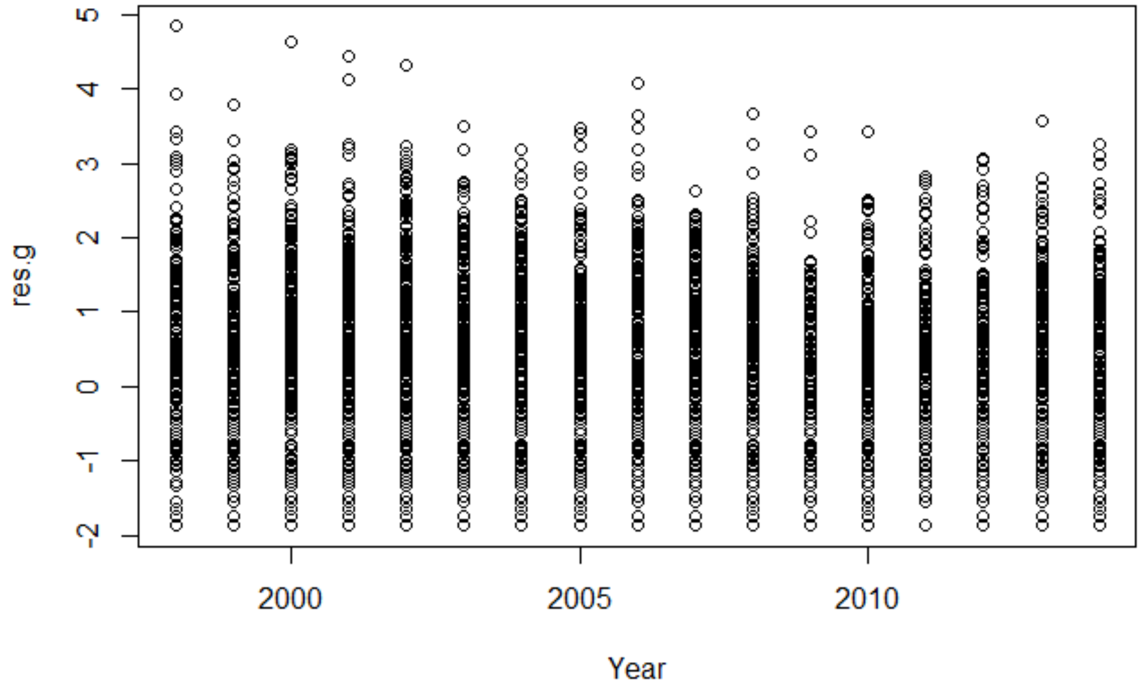
```
setwd("~/R/working directory")

Data=read.table("project data.txt", header=T, "\t")
attach(Data)
mean(Illnesses)

plot(Location, Illnesses)
plot(Genus,Illnesses)
plot(State, Illnesses)

##ANOVA##
#influence location on amount of Illnesses
r.l=aov(Illnesses ~ Location)
summary(r.l)
# location matters, noise did not prevent result, blocking necessary?

#assumptions#
res.l=resid(r.l)
qqnorm(res.l) #normality data --> not normal!
qqline(res.l)
plot(x=Year, y=res.l) #independence --> no trend
plot(fitted(r.l),res.l) #equal variance --> not really!
#interpretation ANOVA is questionable#
#!need of data transformation or other test!#

#data transformation: log
r.tl=aov(log(Illnesses) ~ Location)
summary(r.tl)
res.tl=resid(r.tl)
qqnorm(res.tl)
qqline(res.tl)
plot(Year,res.tl)
plot(fitted(r.tl),res.tl)
#Better! Still very significant influence of location of preparation

#influence genus on amount of Illnesses
r.g=aov(log(Illnesses) ~ Genus)
summary(r.g)
res.g=resid(r.g)
qqnorm(res.g)
qqline(res.g)
plot(Year,res.g)
plot(fitted(r.g),res.g)
#Assumptions are alright, genus matters very significantly!

#influence state on amount of Illnesses
r.s=aov(log(Illnesses) ~ State)
summary(r.s)
```

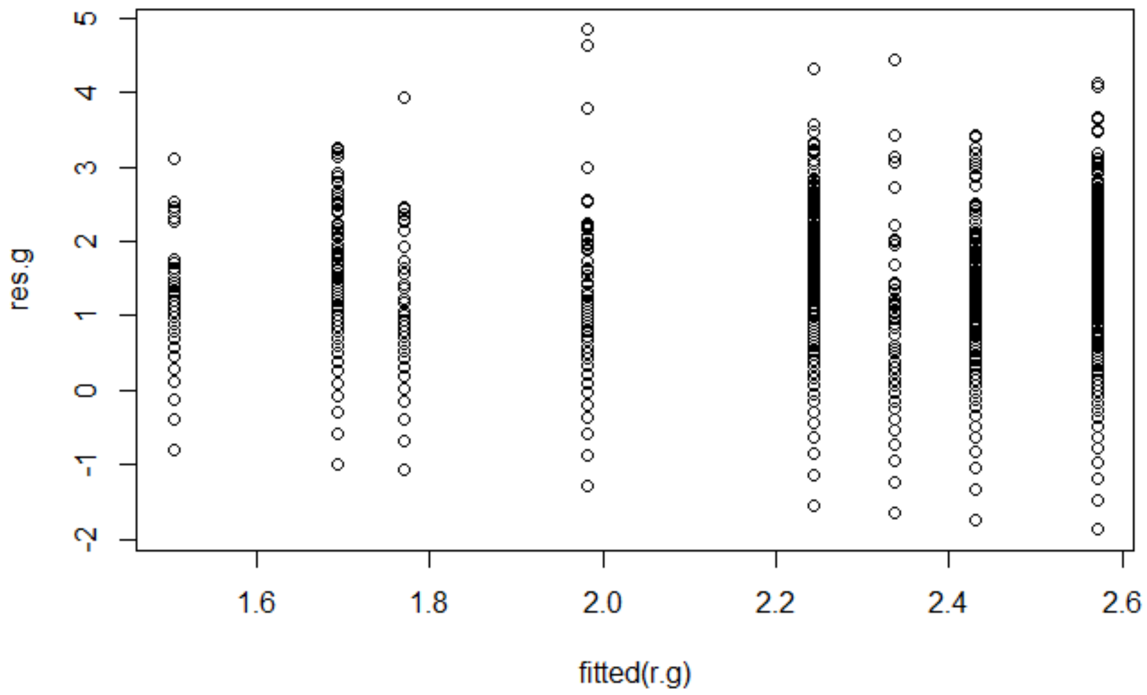
```
res.s=resid(r.s)
qqnorm(res.s)
qqline(res.s)
plot(Year,res.s)
plot(fitted(r.s),res.s)
#Assumptions are alright, State matters also very significantly!

#blocking is possible, but necessary? Test is already very sign.
#now I want to know where the differences are!

##Graphical ANOVA##
raw.total=c(Data$Illnesses)
total=log(raw.total)
ga=mean(total) #grand average
par(mfrow=2:1)

#Locations
meanL=aggregate(x=log(Illnesses), by=list(Location=Location), mean)
devL=meanL$x-ga
stripchart(sqrt(6786/4)*devL, main="Locations")
stripchart(res.tl, main="Residuals", method="stack", offset=0.005)

#Genus
meanG=aggregate(x=log(Illnesses), by=list(Genus=Genus), mean)
devG=meanG$x-ga
stripchart(sqrt(6783/7)*devG, main="Genera")
stripchart(res.g, main="Residuals", method="stack", offset=0.005)

#State
meanS=aggregate(x=log(Illnesses), by=list(State=State), mean)
devS=meanS$x-ga
stripchart(sqrt(6737/53)*devS, main="States")
stripchart(res.s, main="Residuals", method="stack", offset=0.005)
```

Assumptions

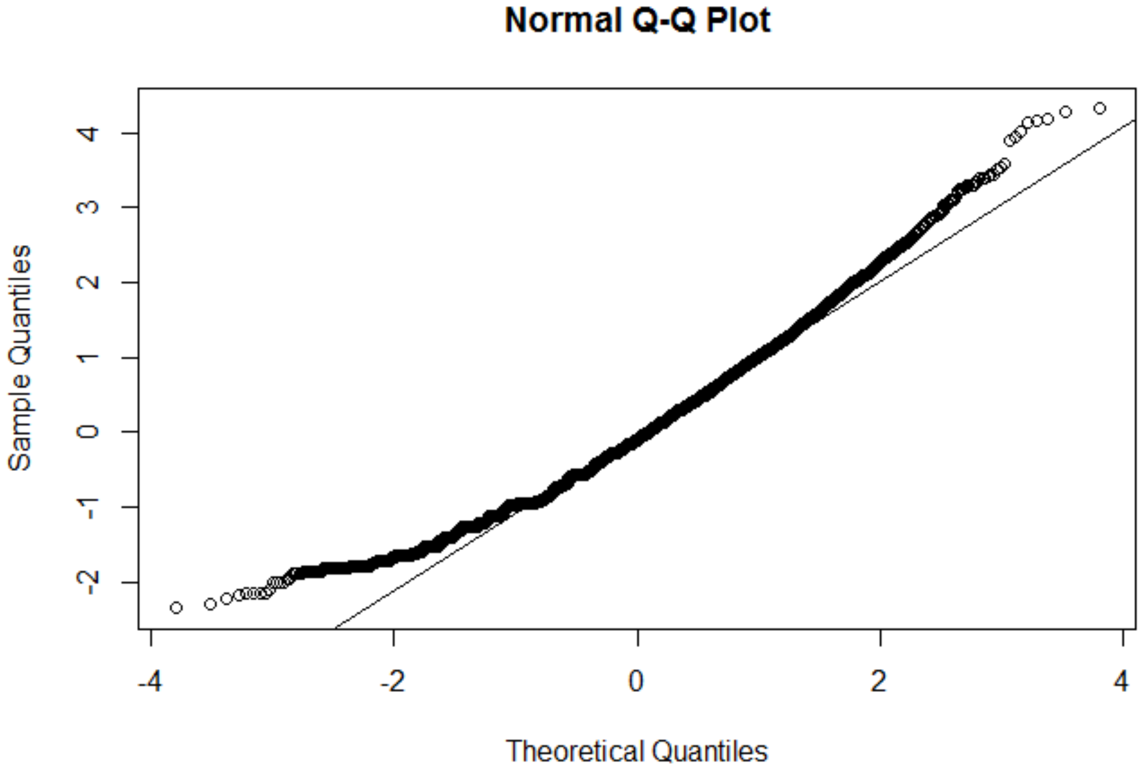
Genera

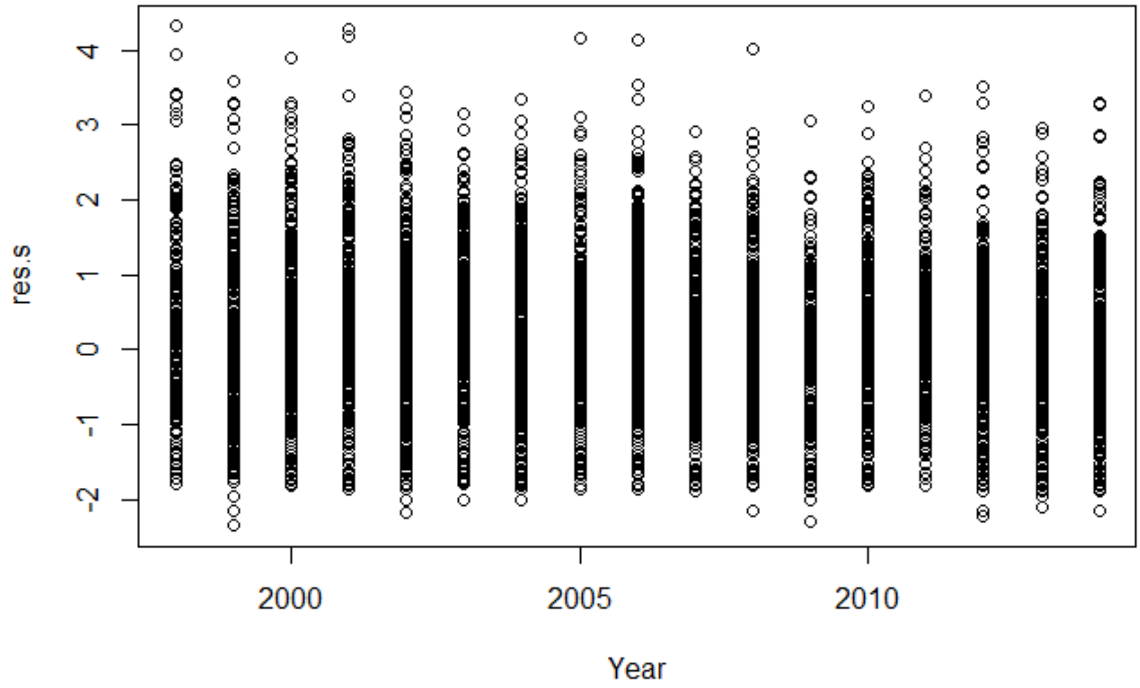


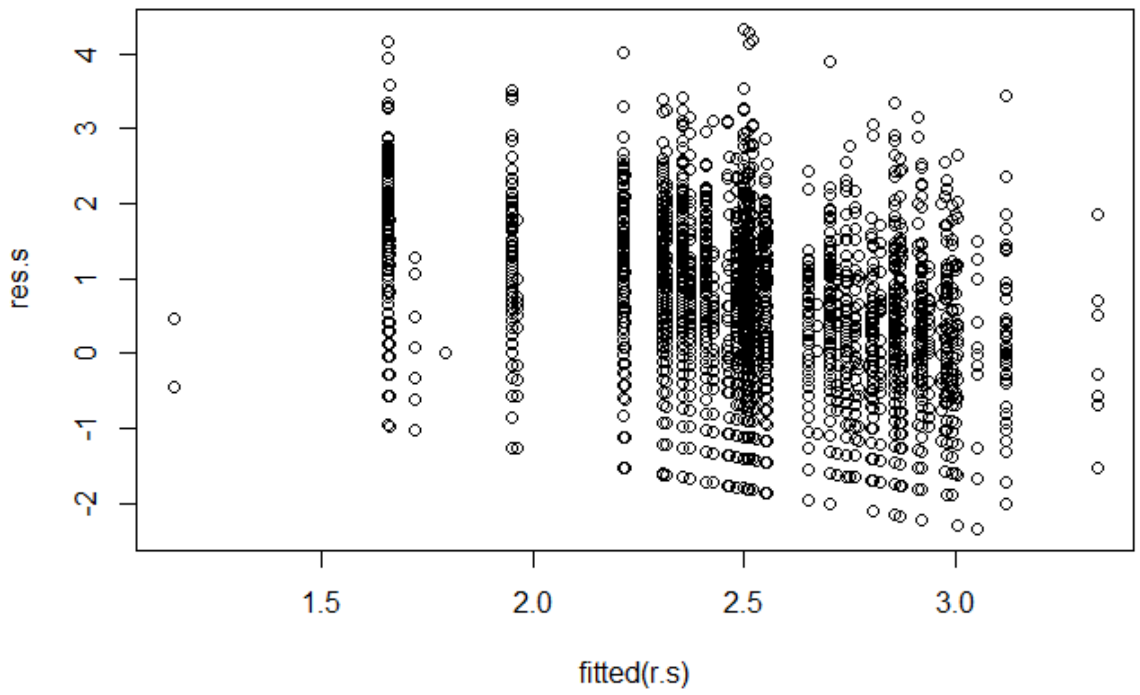



res.g
5
4
3
2
1
0
-1
-2
1.6
1.8
2.0
2.2
2.4
2.6
fitted(r.g)

States


Normal Q-Q Plot
Sample Quantiles
-2
-1
0
1
2
3
4
-4
-2
0
2
4
Theoretical Quantiles


res.s
-2
-1
0
1
2
3
4
2000
2005
2010
Year


res.s
-2
-1
0
1
2
3
4
1.5
2.0
2.5
3.0
fitted(r.s)