Robust Estimators

How should outliers from a sample be analyzed?

- Remove the outliers from the sample since by definition they are not representative of the underlying population?

Example: Patients with atrial fibrillation costs in a given year: ~$10,000 but a few ~$1 million.

- Remove outliers and analyze them separately. The outliers may be more interesting than the rest of the sample.

- Leave the cases with outliers in the data set, especially in situations where many variables are measured on the same experimental unit.

For example, X1, X2, X3 are measured on the same person. X1 is considered an outlier for a particular person. If the person is removed from the sample what about X2 and X3?

- Use estimators of population parameters which are not sensitive to the assumptions made about the data - called Robust Estimators.

Two well-known Robust Estimators of central location include:

- Trimmed mean
- Winsorized mean

Material below from Peter Flom Independent stat consultant for researchers
New York NY

What is the trimmed mean?

The trimmed mean is the mean with some extreme values taken out.; for example, in the 10% trimmed mean the largest and smallest 10% of the values are removed and then the mean is taken on the remaining 80%. This strikes some people as odd, almost as "cheating". But these same people are often perfectly comfortable reporting the median, which is simply the 50% trimmed mean. This is strange. People are fine with chucking all the data, but not with chucking some of it. Still, it is true that the median and median are very often reported, while other trimmed means are not.

What is the Winsorized mean?

The Winsorized mean is similar to the trimmed mean, except that rather than deleting the extreme values, they are set equal to the next largest (or smallest) value.

Example of the trimmed mean

Suppose you are taking an introductory statistics class. The professor decides to collect data on the heights of all the students in the class. However, she is unaware that the coach of the men's basketball team has recommended the class to his team. Each student writes his or her height on a card and passes it in; each student also writes an M or F for male and female.

The results?

For the men: 70, 72, 74, 68, 69, 87, 82, 73, 67, 70, and for the women: 62, 67, 66, 61, 69, 70, 69, 68, 62, 63.

He first takes the two averages. For the men, the mean is 73.2 inches, or 6 feet 1.2 inches. For the women, he gets 65.7 inches, or 5 feet 5.7 inches. The professor knows that the average height of men is not close to 6'1″. She looks at the values and sees some are odd and decides to calculate the 20% trimmed means. That is, she removes the 2 tallest and 2 shortest men and the 2 tallest and 2 shortest women; then she calculates the mean on the remaining 6. It is easier to do this if you first sort the data (e.g. for the men, change:

70, 72, 74, 68, 69, 87, 82, 73, 67, 70

to

67, 68, 69, 70, 70, 72, 73, 74, 82, 87

and then remove the two smallest (67 and 68) and two largest (82 and 87).

The results? For the men, she gets 71.3 inches (much less than 73.2) and for the women she gets 65.8 inches (almost exactly 65.7). Then she uses that information to investigate and writes a note to the coach thanking him for sending students to her class.

Example of the Winsorized mean

To calculate the Winosrized mean we would again order the data but then change the extreme values; for the men:

67, 68, 69, 70, 70, 72, 73, 74, 82, 87

becomes

69, 69, 69, 70, 70, 72, 73, 74, 74, 74

The 20% Winsorized mean is 71.4 (very close to the trimmed mean, which is common).

What about the median?

She decides to also show the median height for each group. The median is defined as the value that splits the group in two: Half are higher, half lower. For the men it is 71.0 inches and for the women it is 66.5 inches. These values are, in this example, quite close to the 20% trimmed means.

How to calculate these numbers?

With only 10 students in each group, it is easy to do the calculations by hand. But if the groups were larger, it would be a pain. Fortunately, statistical software packages exist to do the work for you. In `R` you can do all the above calculations with the following code:

```
men <- c(70, 72, 74, 68, 69, 87, 82, 73, 67, 70) #Data
mean(men) #Mean
mean(men, trim = .2) #Trimmed mean
women <- c(62, 67, 66, 61, 69, 70, 69, 68, 62, 63)
mean(women)
mean(women, trim = .2)
quantile(men) #Quantiles, including the median (50% quantile)
quantile(women)
```

In `R` the `psych` package has winsor and winsor.means functions; these do Winsorization slightly differently; using the values at the exact quintiles (that is, here, the 80%tile and 20%tile) which is easy for computers but not that easy in hand calculation. The result for the men is 71.68.

go to winsor_trim.R

Example of winsorizing and trimming

The UNIVARIATE Procedure
Variable: diameter

Moments			
N	10	Sum Weights	10
Mean	0.0538	Sum Observations	0.538
Std Deviation	0.05542322	Variance	0.00307173
Skewness	2.43269607	Kurtosis	6.3598223
Uncorrected SS	0.05659	Corrected SS	0.0276456
Coeff Variation	103.017138	Std Error Mean	0.01752636

Basic Statistical Measures			
Location		Variability	
Mean	0.053800	Std Deviation	0.05542
Median	0.032000	Variance	0.00307
Mode	0.031000	Range	0.18300
		Interquartile Range	0.04700

Basic Confidence Limits Assuming Normality			
Parameter	Estimate	95% Confidence Limits	
Mean	0.05380	0.01415	0.09345
Std Deviation	0.05542	0.03812	0.10118
Variance	0.00307	0.00145	0.01024

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	3.069662	Pr > |t|	0.0134
Sign	M	5	Pr >= |M|	0.0020
Signed Rank	S	27.5	Pr >= |S|	0.0020

Trimmed Means								
Percent Trimmed in Tail	Number Trimmed in Tail	Trimmed Mean	Std Error Trimmed Mean	95% Confidence Limits		DF	t for H0: Mu0=0.00	Pr > |t|
10.00	1	0.040125	0.009814	0.016919	0.063331	7	4.088554	0.0046
20.00	2	0.037333	0.011231	0.008463	0.066203	5	3.324159	0.0209

Example of winsorizing and trimming

The UNIVARIATE Procedure
Variable: diameter

Winsorized Means								
Percent Winsorized in Tail	Number Winsorized in Tail	Winsorized Mean	Std Error Winsorized Mean	95% Confidence Limits		DF	t for H0: Mu0=0.00	Pr > |t|
10.00	1	0.041800	0.009953	0.018264	0.065336	7	4.199656	0.0040
20.00	2	0.041000	0.011672	0.010997	0.071003	5	3.512829	0.0170

Robust Measures of Scale		
Measure	Value	Estimate of Sigma
Interquartile Range	0.047000	0.034841
Gini's Mean Difference	0.051689	0.045808
MAD	0.011500	0.017050
Sn	0.017889	0.017889
Qn	0.028885	0.020931

Quantiles (Definition 5)	
Quantile	Estimate
100% Max	0.2000
99%	0.2000
95%	0.2000
90%	0.1395
75% Q3	0.0700
50% Median	0.0320
25% Q1	0.0230
10%	0.0175
5%	0.0170
1%	0.0170
0% Min	0.0170

Example of winsorizing and trimming

The UNIVARIATE Procedure
Variable: diameter

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
0.017	1	0.033	6
0.018	2	0.036	7
0.023	3	0.070	8
0.031	5	0.079	9
0.031	4	0.200	10

proportion trimmed	95% CI	Length of interval
.0	(.01414,.09345)	.07930
.1	(.016919,.063331)	.046412
.2	(.008463,.066203)	.057740

adaptive estimator

proportion winsorized	95% CI	Length of interval
.0	(.01414,.09345)	.07930
.1	(.018264,.065336)	.047072
.2	(.010997,.071003)	.060006