REJECTING BAD DATA

In this unit of the course we investigate a statistical procedure for eliminating bad data from a sample. The premise is that while sampling data, some error or mistake caused a measurement to be outside the expected bounds based on the other data.

Before throwing out unwanted data, we should consider the following:

> If there was some clear measurement problem, the affected data should be eliminated. For example, if you are recording the speed of a vehicle using a radar gun, and find that the battery charge was low, you should eliminate all values measured using the low battery *even if the measurements seem to be OK.*
>
> Sometimes the ***outlier*** data points are the ones you really want! For example, the temperature in your F-18 engine exhaust suddenly rises. Do you throw out the reading because it is "out of norm"?
>
> Any rejection procedure may eliminate "good" data.
>
> Any rejection procedure may keep "bad" data.

For this course we use the *modified Thompson's-tau technique*, which is the method recommended in *Measurement Uncertainty*, Part I, ASME PTC 19.1-1998.

We start by calculating the mean and standard deviation of the sample, $\bar{x}$ and S. We then look for outliers. An outlier will be the largest or smallest data value (it would either be the first or last data point if the samples were placed in increasing order). We determine which (lowest or highest) point is the biggest outlier by calculating the deviation for each point as:

$$d_i = \left| x_i - \bar{x} \right|$$

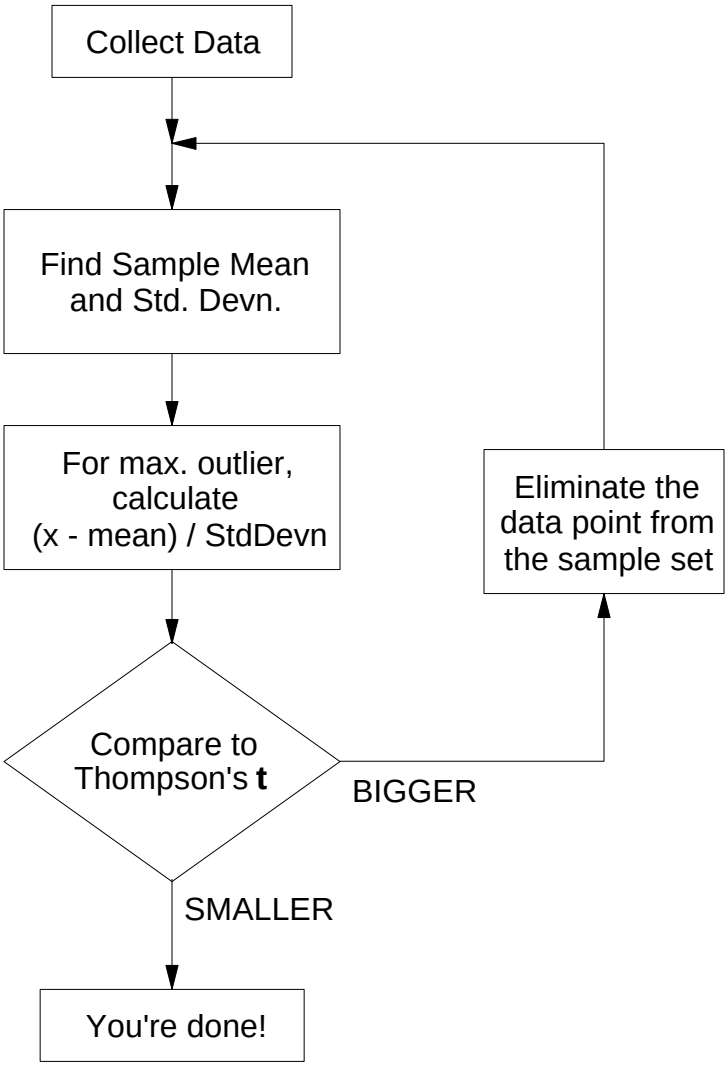
We then standardize the deviations as:

$$t_i = \frac{d_i}{S} = \frac{\left| x_i - \bar{x} \right|}{S}$$

Finally, we compare the largest standardized outlier with tabulated values of Thompson's-τ. If the standardized outlier is larger than Thompson's-τ, we eliminate the associated data point and start the calculation process again.

NOTE: <u>You can only eliminate a single data point at a time.</u> Even if several data points have standardized deviations greater than Thompson's-τ, you can still only eliminate the singe data point with the largest standardized deviation.

The following flow chart shows the entire process:



The values of Thompson's-τ are shown in the following table. The Thompson's-τ value depends on the number of samples, *n*, in the data set. As you eliminate bad data points, the sample size will reduce by one each time.

Values of Thompson's t

n	*t*	*n*	*t*
3	1.150	22	1.893
4	1.393	23	1.896
5	1.572	24	1.899
6	1.656	25	1.902
7	1.711	26	1.904
8	1.749	27	1.906
9	1.777	28	1.908
10	1.798	29	1.910
11	1.815	30	1.911
12	1.829	31	1.913
13	1.840	32	1.914
14	1.849	33	1.916
15	1.858	34	1.917
16	1.865	35	1.919
17	1.871	36	1.920
18	1.876	37	1.921
19	1.881	38	1.922
20	1.885	39	1.923
21	1.889	40	1.924

EXAMPLE: While a vehicle was traveling at an essentially constant speed, a radar gun was used to measure the following speeds (mph):

58, 60, 61, 62, 61, 57, 52, 67, 65, 54, 61, 68, 58, 50, 53, 64, 61, 55, 55,
56, 61, 58, 67, 53, 62, 68, 57, 63, 72, 58

Should you reject any measurements from your analysis? If so, which ones?

<u>STEP 1</u>: (30 samples)

$\bar{x} = 59.90$

$S = 5.352$

$t_{30} = 1.911$

Min. outlier is 50, gives $t_{MIN} = \dfrac{|x_{MIN} - \bar{x}|}{S} = 1.850$

Max. outlier is 72, gives $t_{MAX} = \dfrac{|x_{MAX} - \bar{x}|}{S} = 2.261$ reject this one since 2.261 > 1.911

<u>STEP 2</u>: (29 samples)

$\bar{x} = 59.483$

$S = 4.925$

$t_{29} = 1.910$

Min. outlier is 50, gives $t_{MIN} = \dfrac{|x_{MIN} - \bar{x}|}{S} = 1.925$ reject this one since 1.925 > 1.910

Max. outlier is 68, gives $t_{MAX} = \dfrac{|x_{MAX} - \bar{x}|}{S} = 1.729$

<u>STEP 3</u>: (28 samples)

$\bar{x} = 59.821$

$S = 4.659$

$t_{28} = 1.908$

Min. outlier is 52, gives $t_{MIN} = \dfrac{|x_{MIN} - \bar{x}|}{S} = 1.679$

Max. outlier is 68, gives $t_{MAX} = \dfrac{|x_{MAX} - \bar{x}|}{S} = 1.755$

1.755 < 1.908 so don't reject any more

FINAL REDUCED DATA SET IS:

58, 60, 61, 62, 61, 57, 52, 67, 65, 54, 61, 68, 58, 53, 64, 61,
55, 55, 56, 61, 58, 67, 53, 62, 68, 57, 63, 58

PROBLEMS

Reject-1:

You are to determine the average length of midshipmen's shoes. Do this as follows:

> Measure the length of at least 15 shoes. You can measure different types of shoes (e.g., sandals, leathers) but only one shoe per midshipman. Measure and record the length to an accuracy of 1/10-inch.
>
> Use the Thompson's-τ procedure to determine if any of the measurements should be eliminated.

Report the following:

> Raw data, to include the name of each midshipman, type of shoe and its length.
>
> The mean and standard deviation of the raw data set.
>
> The Thompson's-τ calculations. Clearly indicate any measurements you eliminate.
>
> The final data set, and its mean and standard deviation
>
> A comment about whether you think any "good" data were eliminated or if any "bad" data were maintained in the data set.