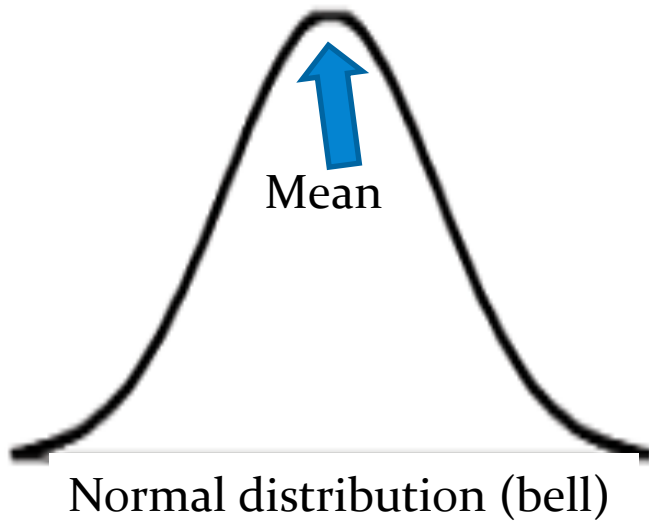


Data Cleaning and spotting outliers with UNIVARIATE

Michael Auld
PhUSE Brighton 2011

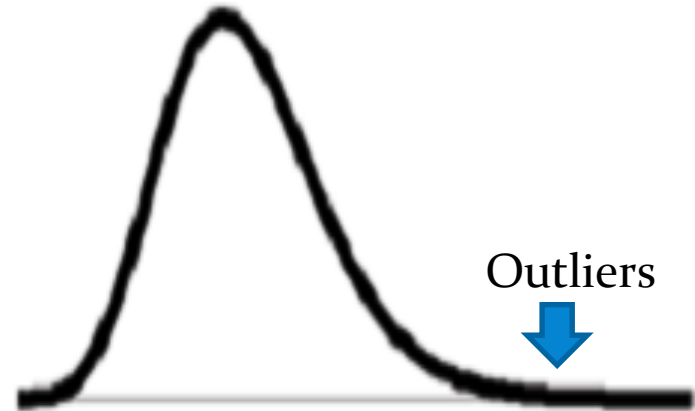
Why outliers are important



- May be an indication of data quality errors
- Good to spot early whatever the reason

- Skewed F-shape curve may reveal bias in the population
- May indicate power of trial isn't strong enough – expensive fail

Asymmetric curve (F-shape)



UNIVARIATE ODS tables

Output Added:

Name: Moments

Label: Moments

Output Added:

Name: BasicMeasures

Label: Basic Measures of Location and Variability

Output Added:

Name: TestsForLocation

Label: Tests For Location

Output Added:

Name: Quantiles

Label: Quantiles

Output Added:

Name: ExtremeObs

Label: Extreme Observations

Output Added:

Name: MissingValues

Label: Missing Values

Extreme values

```
ODS SELECT ExtremeObs;
PROC UNIVARIATE DATA=sds.lb NEXTROBS=10;
  CLASS lbtest;
  ID usubjid;
  VAR lbstresn;
RUN;
```

The UNIVARIATE Procedure
Variable: lbstresn
LBTEST = ABS. NEUTRO.COUNT

Extreme Observations

	Highest-----	-----Lowest-----		-----	
Obs	Value	subjid	Obs	Value	subjid
	0.00000	0074-0018	459425	2730	0067-0017
412339	0.00000	0053-0008	311137	2920	0067-0017
412471	0.00000	0053-0008	311125	2920	0067-0017
412472	0.00036	0100-0012	607593	3200	0067-0017
412498	0.00215	0033-0013	188278	3200	0067-0017
412499	0.01000	0048-0019	279017	3500	0067-0017
412525	0.01000	0048-0019	279016	3500	0067-0017
412526	0.01500	0084-0008	511085	3680	0067-0017
412432	0.01900	0064-0013	397497	3680	0067-0017
412433	0.02000	0048-0019	279007	5330	0059-0005
352791					

The SAS System

The UNIVARIATE Procedure
Variable: lbstresn
LBTEST = ALBUMIN

Extreme Observations

Quantiles analysis

Note the extreme gap between 100th and 99th percentiles

The SAS System

The UNIVARIATE Procedure
Variable: lbstresn
LBTEST = ABS. NEUTRO.COUNT

Quantiles (Definition 5)

Quantile	Estimate
100% Max	5330.000
99%	10.380
95%	6.424
90%	5.350
75% Q3	3.960
50% Median	2.910
25% Q1	2.120
10%	1.512
5%	1.110
1%	0.450
0% Min	0.000

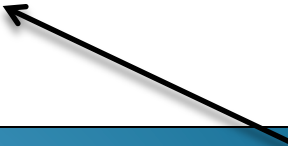


The SAS System

The UNIVARIATE Procedure
Variable: lbstresn
LBTEST = ALBUMIN

Quantiles (Definition 5)

Quantile	Estimate
100% Max	470.000
99%	61.700
95%	53.000
90%	49.000
75% Q3	46.000
50% Median	43.000
25% Q1	39.550
10%	36.300
5%	34.000
1%	26.000
0% Min	0.029



Large gap also observed between 0th and 1st percentiles

Method and Rationale

- Determine programmatically what made those observations stand-out from the crowd
- The answer is context – the distance between that 95th and the 100th percentile when compared with the others
- Why not project back from the 5th and forward from the 95th to determine the expected values at 0th and 100th (the min and max)

The SAS code bit

```
PROC UNIVARIATE DATA=sds.lb NOPRINT;  
  CLASS lbcat lbtest;  
  VAR lbstresn;  
  OUTPUT OUT=mydata PCTLPTS=5 95 MIN=min MAX=max PCTLPRE=p;  
RUN;  
  
DATA nthdegree;  
  SET mydata(WHERE=(NOT MISSING(max)));  
  pn = (p95 - p5)/90;  
  p0 = MAX(p5 - (5*pn), min);  
  p100 = MIN(p95 + (5*pn), max);  
RUN;
```

The SAS code bit

	LAB Test or Examination Name	the largest value, lbstresn	the smallest value, lbstresn	the 5.0000 percentile, lbstresn	the 95.0000 percentile, lbstresn	pn	p0	p100
1	ABS. BAND COUNT	0	0	0	0	0	0	0
2	ABS. BASO.COUNT	20	0	0	0.1	0.0011111111	0	0.1055555556
3	ABS. EOS. COUNT	200	0	0	0.33	0.0036666667	0	0.3483333333
4	ABS. LYMPH. COUNT	2390	0.06	0.7	2.52	0.0202222222	0.5988888889	2.6211111111
5	ABS. MONO. COUNT	800	0	0.14	0.981	0.0093444444	0.0932777778	1.0277222222
6	ABS. NEUTRO.COUNT	5330	0	1.11	6.424	0.0590444444	0.8147777778	6.7192222222
7	ALBUMIN	470	0.029	34	53	0.2111111111	32.9444444444	54.0555555556
8	ALKALINE PHOSPHATASE	4490	1.99	57	427	4.1111111111	36.4444444444	447.5555555556
9	ALT (SGPT)	679	0.28	10.5	80	0.7722222222	6.6388888889	83.8611111111
10	APTT	40.6	20	20	40.6	0.2288888889	20	40.6
11	AST (SGOT)	683	0.27	15	80	0.7222222222	11.3888888889	83.6111111111
12	BANDS	819	0	0	10	0.1111111111	0	10.5555555556
13	BASOPHILS	13.4	0	0	1.64	0.0182222222	0	1.7311111111
14	BILIRUBIN	0	0	0	0	0	0	0
15	BUN	6.4	6.4	6.4	6.4	0	6.4	6.4
16	CALCIUM	214	0.3175	2.07	2.66	0.0065555556	2.0372222222	2.6927777778
17	CHLORIDE	250	10.5	96	110	0.1555555556	95.2222222222	110.77777778
18	CREATININE	8309.6	12.25	46	101	0.6111111111	42.9444444444	104.05555556
19	EOSINOPHILS	74	0	0	6	0.0666666667	0	6.3333333333
20	GGT	884	14	16	794	8.6444444444	14	837.22222222
21	GLUCOSE	38.14	2.1	4.22	8.51	0.0476666667	3.9816666667	8.7483333333
		426	0.198	0.362	42.6	0.4693111111	0.198	44.9465555556
		12300	3.8	100	144	0.4888888889	97.5555555556	146.44444444
		47000	0.56	146	845	7.7666666667	107.16666667	883.83333333
		2368	0	14	50	0.4	12	52
		121	0.09	0.6724	1.05	0.0041955556	0.6514222222	1.0709777778
		444	0	2	15	0.1444444444	1.2777777778	15.7222222222
		8775	0	36	76.8	0.4533333333	33.7333333333	79.0666666667
		132	0.16	0.87	1.568	0.0077555556	0.8312222222	1.6067777778
		161000	0.139	139	403	2.9333333333	124.33333333	417.66666667
		18	1.2	3.57	5.01	0.016	3.49	5.09
		500	0	0	0.3	0.0033333333	0	0.3166666667
		51.5	11	11	51.5	0.45	11	51.5
		60	25.1	25.1	60	0.3877777778	25.1	60

```

PROC UNIVARIATE DATA=sds.lb NOPRINT;
  CLASS lbcat lbtest;
  VAR lbstresn;
  OUTPUT OUT=mydata PCTLPTS=5 95 MIN=min MAX=max
  PCTLPRE=p;
RUN;

DATA nthdegree;
  SET mydata (WHERE= (NOT MISSING (max)) );
  pn = (p95 - p5)/90;
  p0 = MAX(p5 - (5*pn), min);
  p100 = MIN(p95 + (5*pn), max);
RUN;

```

Select the extreme obs and raise queries

```
PROC SQL NOPRINT;  
  CREATE TABLE lab_outliers as  
  SELECT lb.*  
         ,extreme.min  
         ,extreme.p0  
         ,extreme.p5  
         ,extreme.p95  
         ,extreme.p100  
         ,extreme.max  
  FROM nthdegree AS extreme LEFT JOIN sds.lb  
  ON lb.lbcat EQ extreme.lbcat  
  AND lb.lbtest EQ extreme.lbtest  
  AND ((extreme.min <= lb.lbstresn < extreme.p0)  
  OR (extreme.p100 < lb.lbstresn <= extreme.max))  
  ORDER BY usubjid, lbcat, lbtest, visitnum  
  ;  
QUIT;
```

Further reading...

Cody's Data Cleaning Techniques using SAS
(Ron Cody, SAS Press Series 2008)

BASE SAS Procedures Guide , SAS Publishing

Contact...

michael_auld@eisai.net

Questions...

