

Classroom Simulation: Indications of Outliers in Boxplots of Normal Data

JSM, Seattle, August 6, 2006

Jacob B. Colvin
jbcolvin@fastmail.fm

Bruce E. Trumbo
bruce.trumbo@csueastbay.edu

Eric A. Suess
eric.suess@csueastbay.edu

Department of Statistics
California State University, East Bay
Hayward, CA 94542, USA

What is an Outlier?

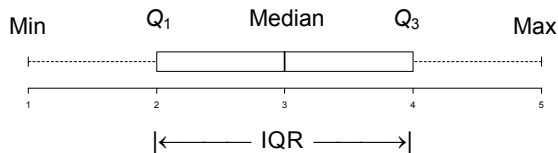
General: An observation in an especially extreme location:

- Far away from most other data points.
- Far away from any other data point.
- Having a disproportional effect on minimum variance estimates.
- In regression: Located more than 2 SD from the mean, or having a standardized residual greater than 2.

Observations That May Be Indicated as "Outliers"

- Naturally occurring extreme values in heavy tailed distributions.
- Observations caused by outside influences mixed with a predominant distribution:
 - Human errors.
 - Equipment failures.
 - Signal against heavy noise background.

Skeletal Boxplot

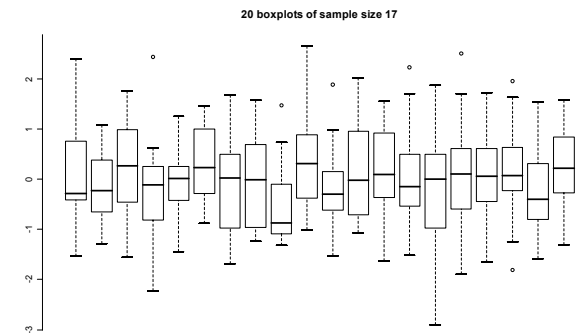


- Graphical representation based on the 5-number summary: percentiles Min=0, $Q_1=25$, Med=50, $Q_3=75$, and Max=100.
- The Interquartile Range: $IQR = Q_3 - Q_1$.
- A boxplot provides no information about sample size n .

Boxplot Indications of Outliers

- A Lower Fence = $Q_1 - K \times IQR$.
- An Upper Fence = $Q_3 + K \times IQR$.
- Observations beyond fences are indications of "outliers":
 - "Possible outliers" if $K = 1.5$,
 - "Probable outliers" if $K = 3.0$,
 - **Alternate criterion considered here is $K = 2.25$.**

Frequent Outlier Indications for Normal Data ($K = 1.5$)



Measuring How Often Normal Observations Are Indicated as Outliers in Boxplots

α_S : Fraction of **Samples** with indications.

α_E : **Expected number** of outlier indications per sample.

For $K = 1.5, 2.25$ or 3 :

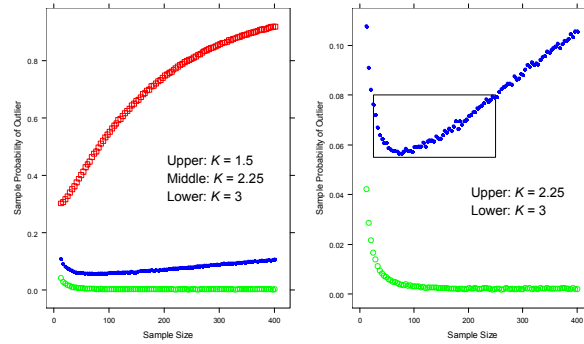
α_S and α_E vary greatly with n :

If $n \rightarrow \infty$, then $\alpha_S \rightarrow 1$ and $\alpha_E \rightarrow \infty$.

We investigate samples of moderate size.

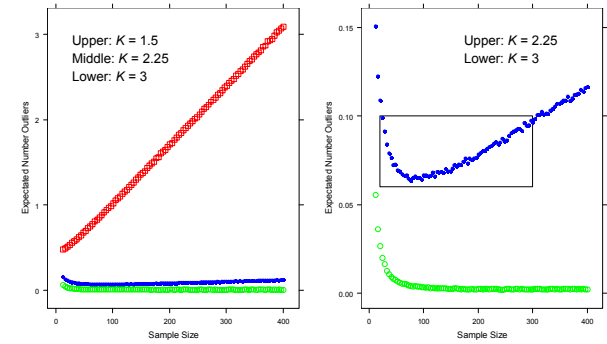
Fraction α_S of Samples with Indications

For $K = 2.25$ and $25 \leq n \leq 250$: $5.5\% < \alpha_S < 8\%$



Avg. No. α_E of Indications Per Sample

For $K = 2.25$ and $20 \leq n \leq 300$: $.06 < \alpha_E < .1$



Another Criterion (Davies et al.): Probability α_I that an Individual Observation is Indicated as an Outlier

Advantage: If $n \rightarrow \infty$, then α_I converges to a limit < 1 , which depends on K .

Limits:

For $K=1.5$: Limit is 0.006976603
For $K=2.25$: Limit is 0.000207510
For $K=3.0$: Limit is 0.000002342

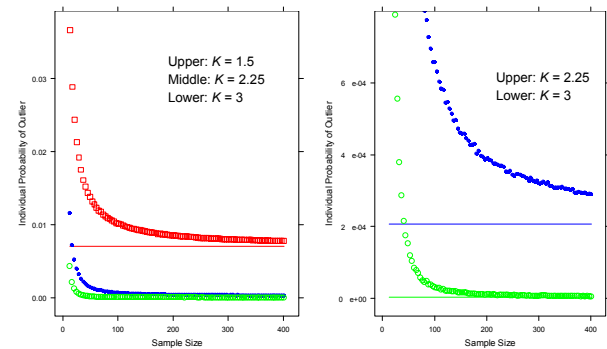
Disadvantage: This criterion may not be useful to practitioners viewing box plots.

Limit of α_I Depends on Quartiles of the Standard Normal Distribution

R code

```
> q1 = qnorm(.25)
> q3 = qnorm(.75)
> iqr = q3 - q1
> k = c(1.5, 2.25, 3)
> 2 * pnorm(q1 - k*iqr)
[1] 6.976603e-03 2.075102e-04 2.341942e-06
```

Probability α_I that an Individual Observation is Indicated as an Outlier



R Code for Simulations of α_S , α_E , α_I

```
# create obs for alpha_E, alpha_I, and alpha_S given K
sim_boxplots=function(reps=100000,k=c(1.5,2.25,3),nobs=17){
  bp=matrix(nrow=reps,ncol=length(k))
  for(i in 1:reps){ # for each simulation
    x=rnorm(nobs) # simulation sample
    for(j in 1:length(k)) # for each value of K
      bp[i,j]=length(boxplot(x, range=k[j],
        plot=FALSE)$out) # number of outliers
  }
  list( # expected number of outliers
    alpha_E=apply(bp,2,mean),
    # probability sample contains an outlier
    alpha_S=apply(bp>0,2,mean),
    # expected number of outliers per observation, or
    # probability observations is an outlier
    alpha_I= apply(bp,2,mean)/nobs) }
```

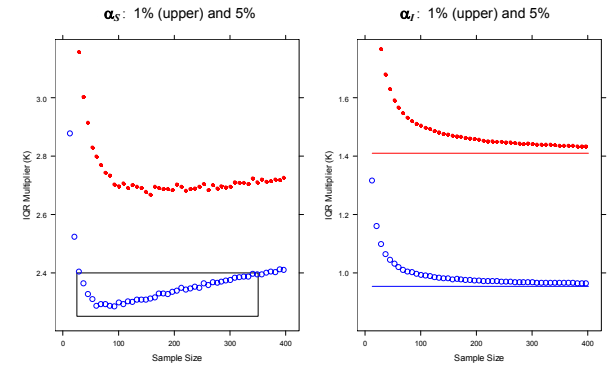
Varying K to Control the Probability of Outlier Indications

- Goal: Hold probabilities of outlier indications constant across samples of different sizes by selecting appropriate values of K_n .
- Values found by simulation.
- We show results for both α_S and α_I .

Conclusion: To hold $\alpha_S = 5\%$ in samples of moderate size, the required values are between $K = 2.2$ and $K = 2.4$.

Values of K for α_S and α_I at 1% and 5%

Solid lines for α_I based on distance between 1st and 3rd quartiles of $N(0,1)$



Comments on the Scope of Our Simulation Study

- Samples sizes less than $n = 13$ were avoided:
 - Sample IQR is too erratic,
 - Dot plots or stripcharts are often better.
- Sample sizes greater than several hundred were explored, but not taken into account in our conclusions:
 - Histograms are often better.

References on Related Simulation Studies

- Davies, Laurie, and Gather, Ursula: "The Identification of Multiple Outliers," *JASA*, Vol. 88, No. 423 (Sept 1993), pages 782-92. Robustness approach.
- Frigge, Michael; Hoaglin, David C.; Iglewicz, Boris: "Some implementations of the boxplot," *The American Statistician*, Vol. 43 (1989), No. 1, 50-54. Suggestions on computer implementations.
- Hoaglin, David C.; Iglewicz, Boris; Tukey, John W.: "Performance of some resistant rules for outlier labeling," *JASA*, Vol. 81 (1986), No. 396, 991-999. Early simulation study to which ours is somewhat similar.
- Hoaglin, David C.; Iglewicz, Boris: "Fine-tuning some resistant rules for outlier labeling," *JASA*, Vol. 82 (1987), No. 400, 1147-1149. Continuation of previous reference.
- Iglewicz, Boris; Hoaglin, David C.: "Use of boxplots for process evaluation," *Journal of Quality Technology*, Vol. 19 (1987), No. 4, 180-190. Applications to control charts.

[We thank Paul Velleman and David Hoaglin for suggesting the last four references.]

General References

- Moore, David S.: *The Basic Practice of Statistics*, 3rd ed., Freeman (2004). Chapter 16 has an excellent discussion of outliers and the robustness of t-procedures.
- Trumbo, Bruce E.: *Learning Statistics with Real Data*, Duxbury (2002). Unit 1 has several examples of outliers in real data arising through various mechanisms.
- Tukey, John W.: *Exploratory Data Analysis*, Addison-Wesley (1977). One of the earlier published discussions of box plots (there called "box-and-whisker plots") appears in Chapter 1.
- www.sci.csueastbay.edu/~btrumbo/JSM2006/Outliers has R code for all figures in this presentation.