36-315: Statistical Graphics and Visualization

Handout 2
Date: January 15, 2003

Census tract - A homogenous neighborhood of approximately 3000 people. For each tract, we have information on population density, ethnicity, ages, incomes, family size, house prices, etc. CMU lies in tract 42003-1401.98.

Histogram - A graphic summary of variation in a set of data. The pictorial nature of the histogram lets people see patterns that are difficult to detect in a simple table of numbers.

Mixture distribution - A sum of simple, smooth distributions (such as normal), plus a few isolated points (exceptional values or "outliers"). This is a convenient (an often correct) mental model of how data varies. The distributions being added are called *modes* (even when their peaks are not visible).

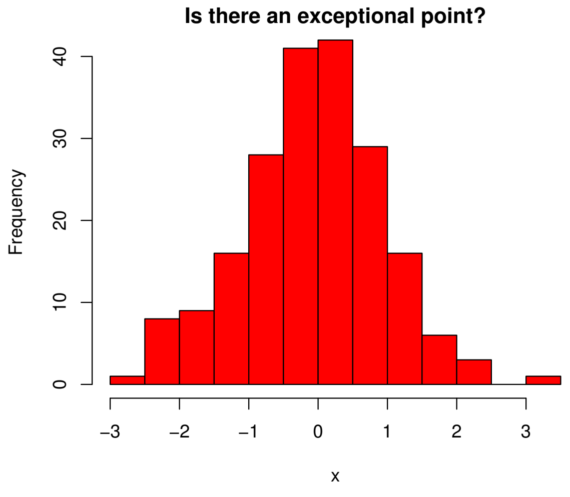
How to identify a mixture - Look for local peaks, or abrupt changes in the slope of the density (forming a "knee" or "corner"). Often easier after a transformation.

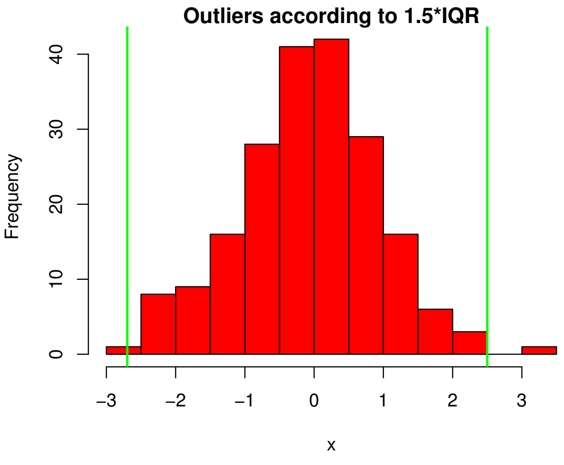
How to identify exceptional values - Look for observations that have low probability under any of the modes. Often easier after a transformation.

Common mistakes:

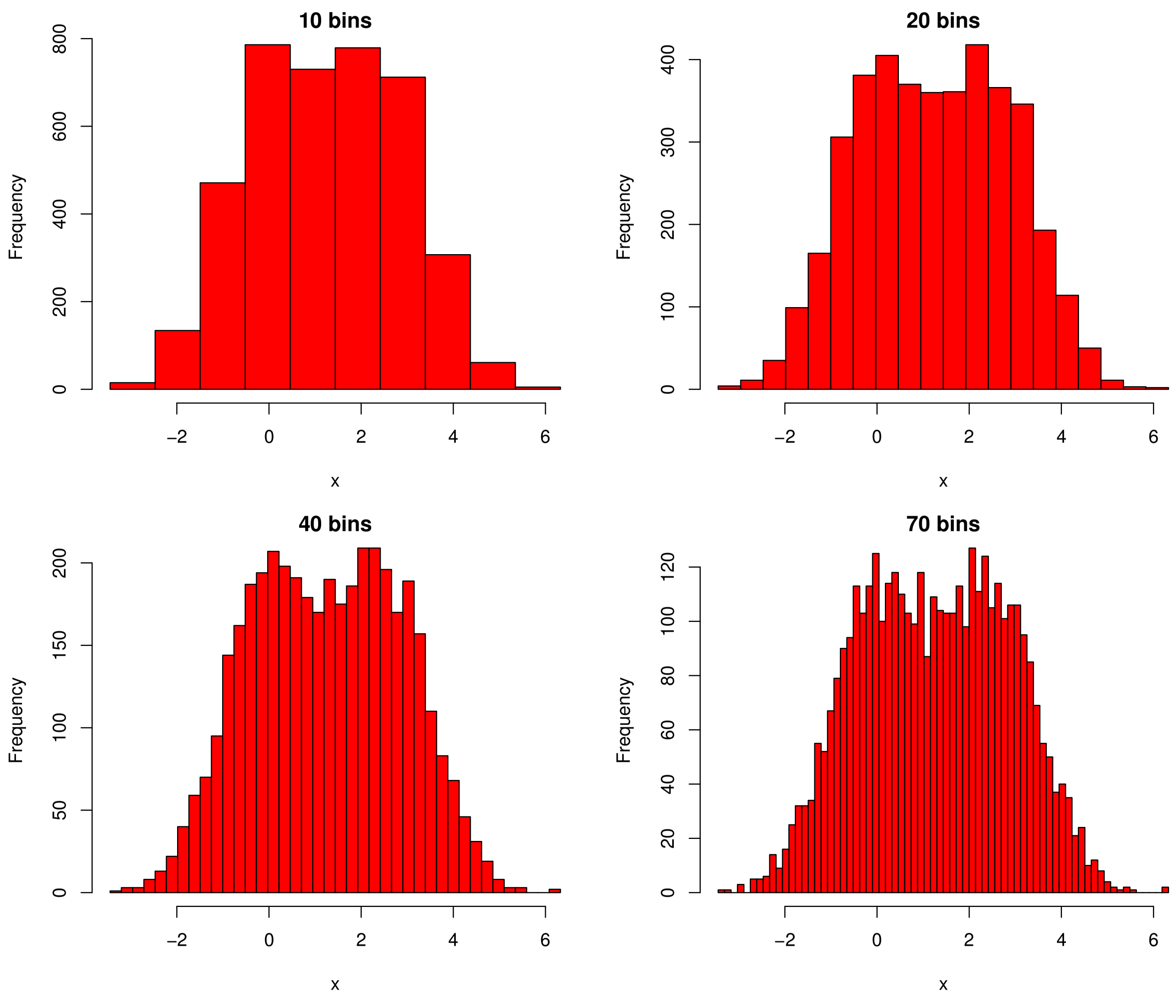
- Too few bins - crude, not enough detail - hides information
- Too many bins - too noisy - gives false information
- Axis range is too wide - hides detail
- Failure to remove skew by transformation - hides modes and outliers

Histograms are very good at identifying outliers (much better than the traditional "1.5 times IQR" rule).


Is there an exceptional point?

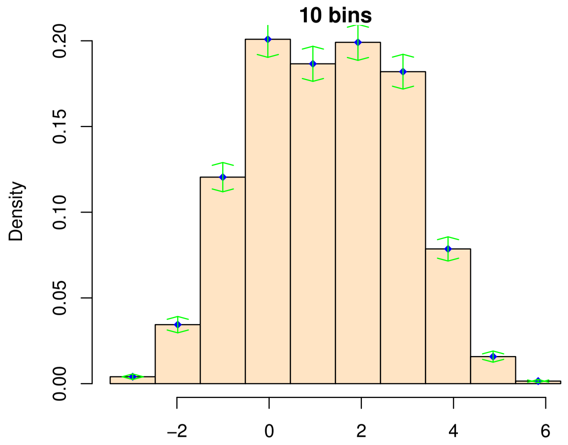

Outliers according to 1.5*IQR

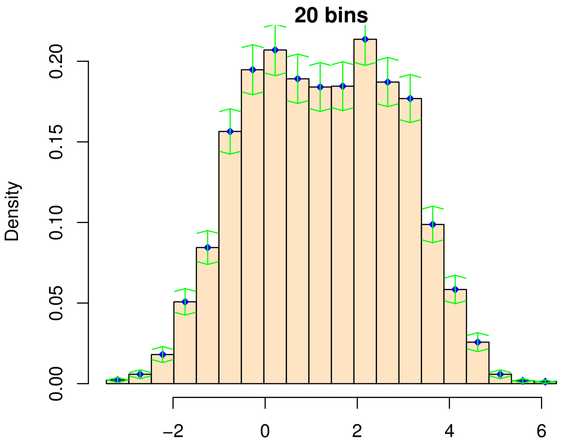
Varying the number of bins:

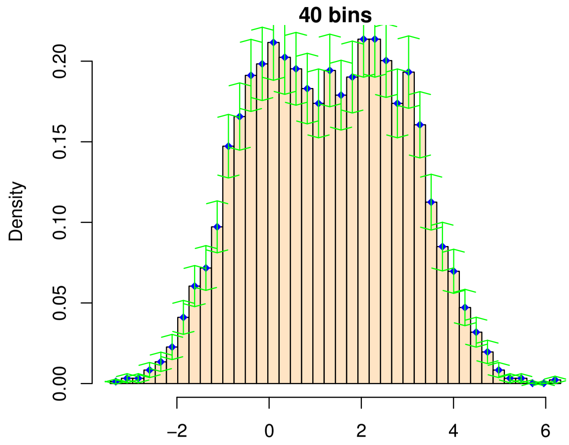


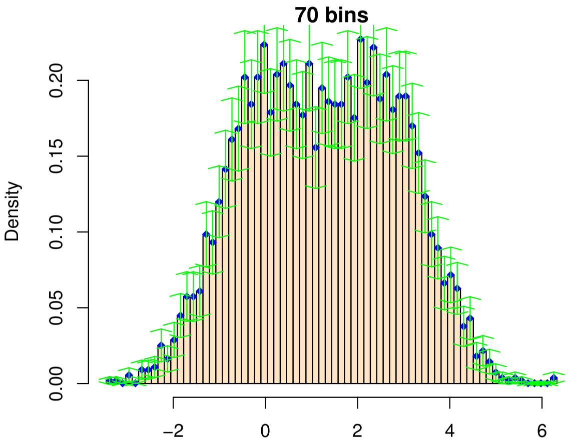
Which histogram is closest to the truth?

With error bars:


10 bins


20 bins


40 bins


70 bins