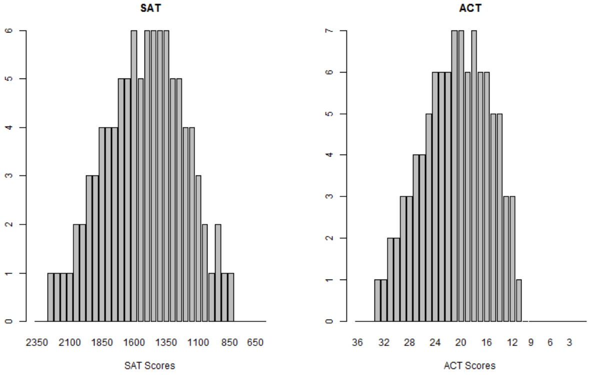
STAT 311: Z-scores and the Empirical Rule

Y. Samuel Wang

Summer 2016

SAT vs ACT scores



Percentiles

If you scored in the 95% on the SAT, what does that mean?

Percentiles

If you scored in the 95% on the SAT, what does that mean?

- The k^{th} percentile of the distribution is the value which has k% of the data at or below it
- Special example is the median (50th percentile), Q1 (25 percentile) or Q3 (75 percentile)

Terminology

For a set of data $\{x_1, x_2, \ldots x_n\}$

- The **rank** (r(x)) is the number of data points in the set that are less than or equal to that data points (including itself).
  - Smallest value has rank of 1
  - Largest value has rank of N
- The **percentile** is the rank divided by the number of observations.

$$\text{percentile} = \frac{r(x)}{N}$$

- **Quantiles** are the values which divide the dataset into equal portions.
  - Percentiles are 100-quantiles
  - Quartiles are 4-quantiles

Minor Details

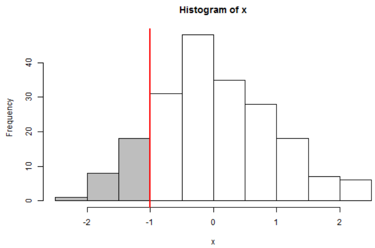
The smallest observation will have a percentile of $\frac{1}{N}$ while the largest observation will have a percentile of $\frac{N}{N} = 1$.

To make things "symmetric" `R` uses a slightly different formulate

$$\text{percentile} = \frac{r(x) - 1}{N - 1}$$

Cumulative Distribution Function

- **Cumulative Distribution Function**, or CDF returns what portion of the distribution is below a certain value
- Often denoted by F(x)
- It is like an "integral of a histogram" or the area under the curve

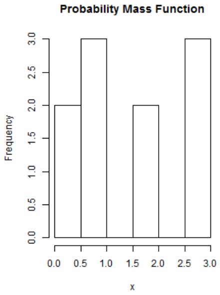


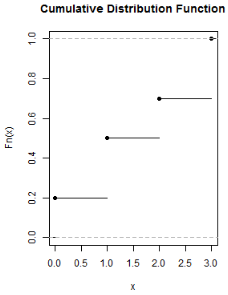
F(-1) is the area of the shaded region

Cumulative Distribution Function

Suppose I have the following data-

Value	0	1	2	3
Count	2	3	2	3

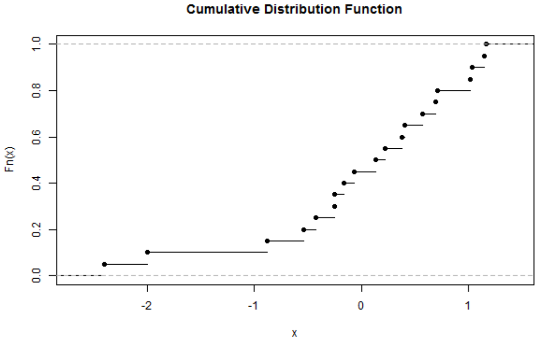




CDF vs Percentile

The CDF and Percentiles are inverse operations.
For a set of data

- CDF: Starts with an observed value, returns a proportion
- Percentile: Starts with a proportion, returns an "observed value"



Z-scores

Given a data set, we can transform each observation in the following way

$$z_i = \frac{x_i - \bar{x}}{s_x} \tag{1}$$

This is called a "z - score"

Mean of the new distribution

$$y_i = x_i - \bar{x}$$

The mean of Y is-

Mean of the new distribution

$$y_i = x_i - \bar{x}$$

The mean of Y is-

$$\bar{y} = \frac{1}{N}\sum_i (y_i) = \frac{1}{N}\sum_i (x_i - \bar{x}) = \frac{1}{N}\sum_i x_i - \frac{1}{N}\sum_i \bar{x} = \bar{x} - \bar{x} \quad (2)$$

$$= 0$$

Proof of transformation

$$z_i = \frac{x_i - \bar{X}}{s_x}$$

The standard deviation of z is-

Proof of transformation

$$z_i = \frac{x_i - \bar{X}}{s_x}$$

The standard deviation of z is-

$$\begin{aligned}
s_z &= \sqrt{\frac{1}{N-1}\sum_i (z_i - \bar{z})^2} = \sqrt{\frac{1}{N-1}\sum_i (z_i)^2} \\
&= \sqrt{\frac{1}{N-1}\sum_i \left(\frac{(x_i - \bar{x})}{s_x}\right)^2} \\
&= \frac{1}{s_x}\sqrt{\frac{1}{N-1}\sum_i (x_i - \bar{x})^2} \\
&= \frac{1}{s_x} s_x = 1
\end{aligned} \tag{3}$$

Why is this useful?

- The z-score has “standardized” each observation
- This takes away the effect of units, and simply tells us how an observation compared to other

Empirical Rule

When the data is normally distributed (bell shaped) with mean=0 and sd = 1.

- Roughly 68% of the data lies between -1 and 1
- Roughly 95% of the data lies between -2 and 2
- Roughly 99.7% of the data lies between -3 and 3

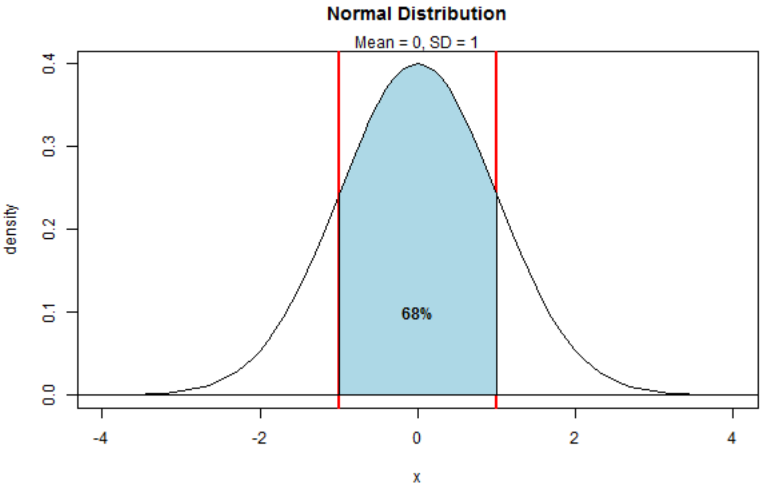
Empirical Rule

When the data is normally distributed (bell shaped) with mean=0 and sd = 1.

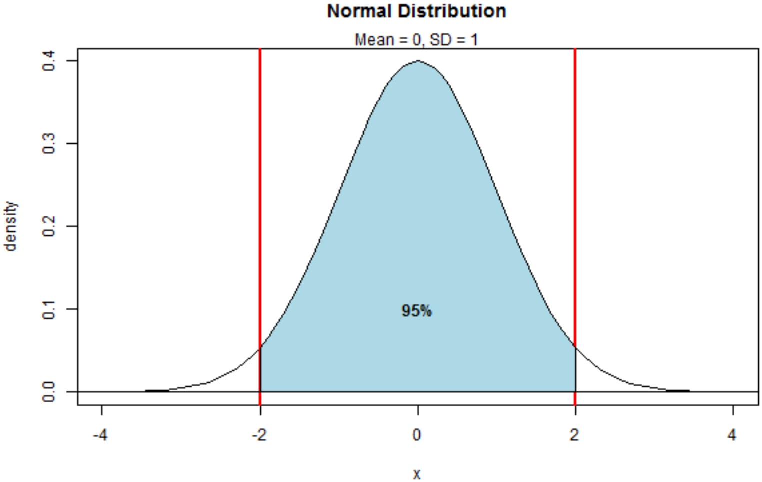
- Roughly 68% of the data lies between -1 and 1
- Roughly 95% of the data lies between -2 and 2
- Roughly 99.7% of the data lies between -3 and 3

Only if the data is roughly normal!

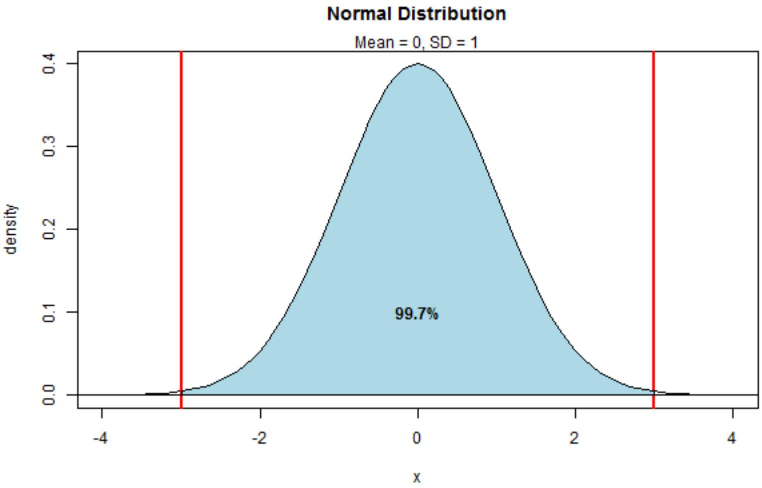
Empirical Rule



Empirical Rule

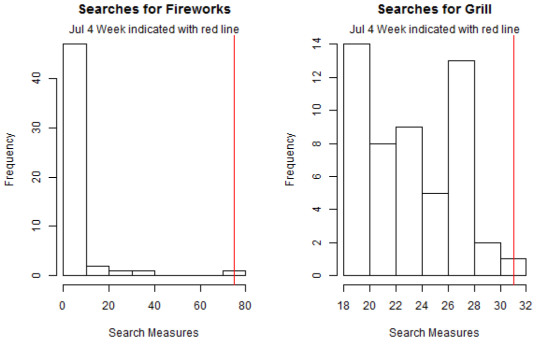


Empirical Rule



Fireworks vs Grills

Data from Google Trends-



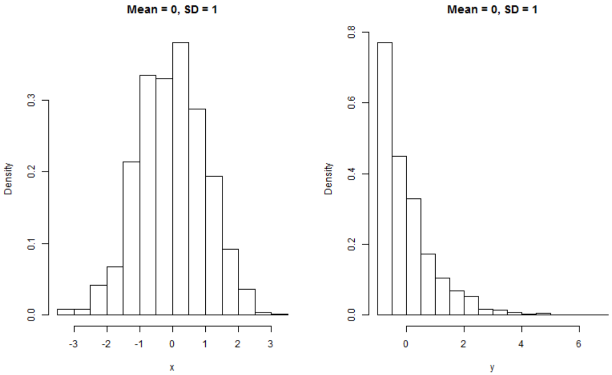
	Fireworks	Grill
Mean	5.35	23.69
SD	11.44	3.49
Jul 4 Obs	75.00	31.00

QQ Plots

How can we tell if a distribution is like another distribution? What if we compare the mean and standard deviation?

QQ Plots

How can we tell if a distribution is like another distribution? What if we compare the mean and standard deviation?



So maybe compare 5 number summaries...

QQ Plots

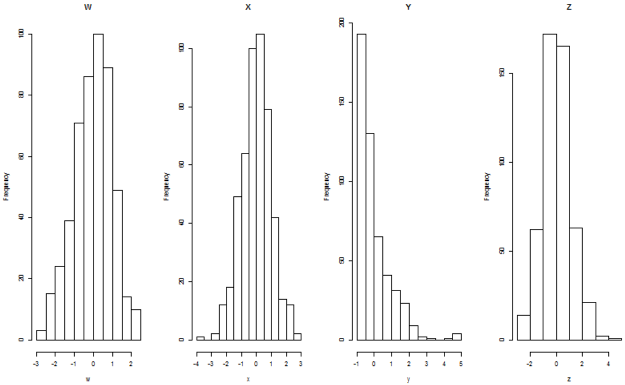
But why stop at 5 numbers, why not compare as many points as we have? We can compare each percentile (which we've observed) against each other using a QQ-plot.

QQ Plots

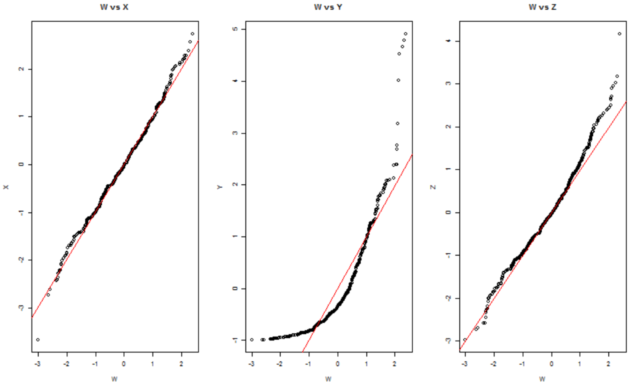
But why stop at 5 numbers, why not compare as many points as we have? We can compare each percentile (which we've observed) against each other using a QQ-plot.

- X-axis: Sorted values from first distribution
- Y-axis: Sorted values from second distribution
- If the two distributions are the same, we would expect the plot to have an intercept close to 0 and slope close to 1

QQ Plots



QQ Plots



More QQ plots

Plotting QQ-Plots of Z scores can be useful to compare the general shape (after spread and mean have been accounted for)