Associations Between a Categorical and Quantitative Variable

Friendship analytics

In advertising (and many other things), the physical appearance of a spokesperson, actor, actress, etc., matters.

- If selling beauty, style, fashion products, person should be attractive.
- If selling insurance, person should look trustworthy and authoritative.
- If selling household products, person should look relatable and likely to actually use the product.

It is imperative to figure out "what matters" when determining how people perceive someone. Take the case of finding factors associated with someone's "friendship potential".

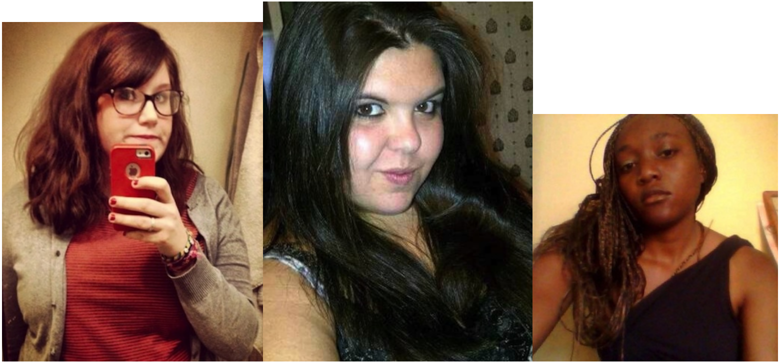
- hair color, eye color
- smile, glasses
- weight, complexion

Friendship survey

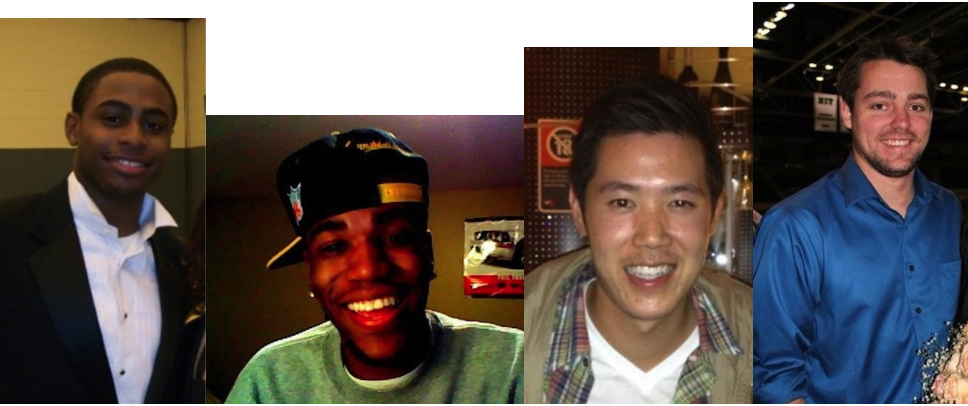
You were asked to rate, on a scale of 1 (low) to 5 (high), how likely it is that you could be friends with 70 people. You were also asked to rate a few characteristics of people that may influence scores (nerdiness, professionalism, etc.). So how can we tell what factors matter?

Girls with most friendship potential

Girls with least friendship potential



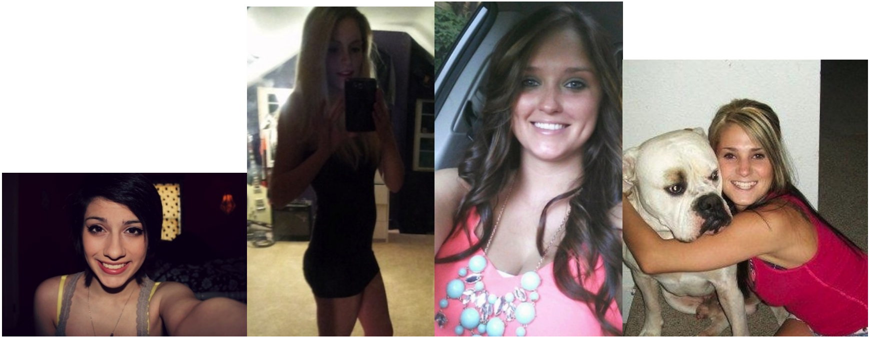
Guys with most friendship potential



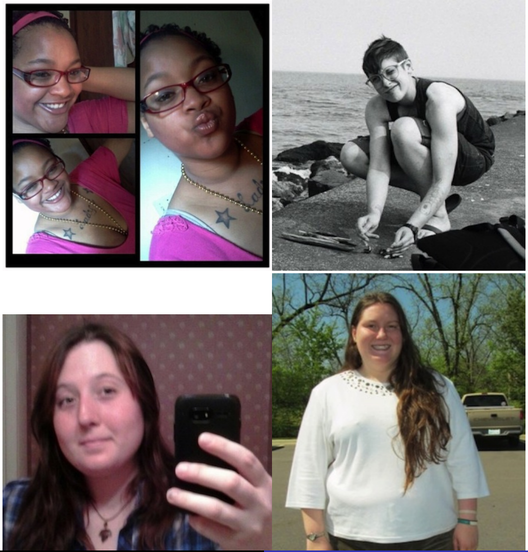
Guys with least friendship potential

Most attractive girls

Classes in the past rated attractiveness instead of friendship potential.

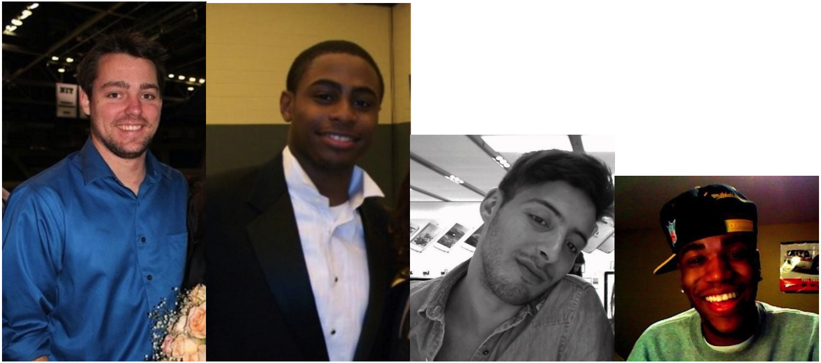


Least attractive girls

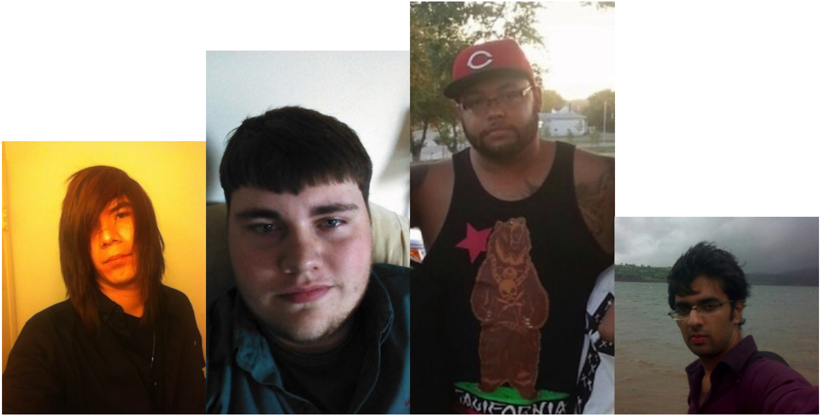


Most attractive guys

Past classes rated attractiveness instead of friendship potential. Notice that 3 of the 4 top guys are the same!



Least attractive guys



Formal Definition of Association

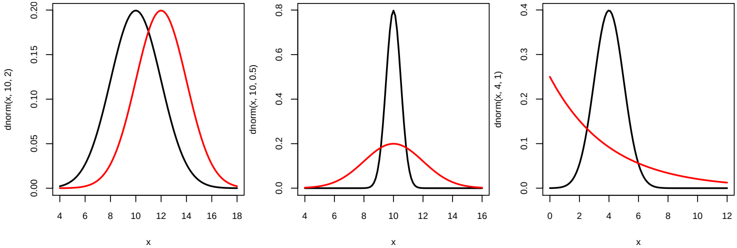
Recall that an association is a general term to describe a relationship between two variables.

> Two quantities are associated when, for whatever reason, knowing the value of one quantity tells you *something* about (i.e. narrows down) the possible values of the other.

In this discussion, we will always assign the role of x to be the categorical variable and y to be the quantitative variable. An association exists between y and x if the distribution of y varies between levels of x.

General Illustration of Association

When y and x have an association, the distribution of y is somehow different between the levels of x.



Left: the two groups differ in terms of their average values of y. Middle and right: the two groups have the same average, but overall distributions of y are different.

Practical definition of association

Checking whether the *distribution* of y is the same for each level of x is a hard problem. Since people are usually more interested in whether the *typical value* of y is different between levels, we will compare only the **average** or **median** values (whichever is appropriate for the distribution).

- Is the *average* friendship potential different between smilers/non-smilers?
- Is the *median* donation amount among all majors the same?

Practical definition of association

In this course, we will only **test** whether the difference in *typical* values of y (rather than the whole of its distribution) between levels of x is statistically significant.

Side-by-side boxplots or histograms give an informal assessment of whether the *distribution* of y as a whole varies between levels of x, but we will not formally test if this difference is significant.

Which typical value?

What number best summarizes the typical value of y?

Use average (mean) when: distributions are roughly symmetric with no extreme outliers.

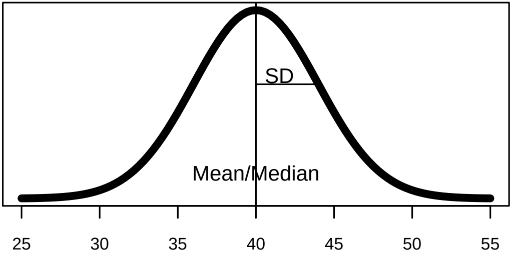
Use median when: distributions are noticeably skewed and/or have extreme outliers.

The resemblance of the distribution of y to a Normal distribution provides an excellent reference point for deciding which of these quantities to use.

Normal distribution

The **Normal distribution** is the standard "bell-curve" that provides an approximate shape for many distributions. It is symmetric and unimodal.

- The mean/median (typical value) are both located at the peak of the distribution.
- The standard deviation or SD (typical difference between values and the mean) is the "half-width" of the curve.



Comparing to a Normal distribution

If the Normal distribution provides a good description of the shape of the distribution, then the average/mean provides a good summary of the typical value of the distribution. If not, the median provides a better summary.

The histogram provides an informal assessment of how much it looks like a Normal distribution, but the **QQ plot** is specialized to do this.

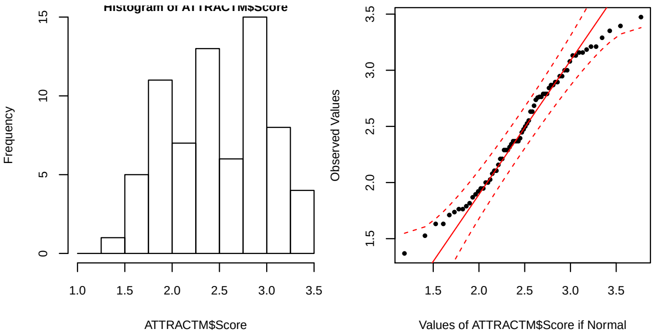
QQ plot overview

Our version of the QQ plot (and there are others) will compared the observed values in the data to the values we would have expected them to be had been generated at random from a Normal distribution.

If the observed and expected values match up well, then the distribution is well-described by a Normal distribution, and the average/mean provides a useful summary for the typical value. If not, the median provides a better summary.

QQ plot example1 (very close to Normal)

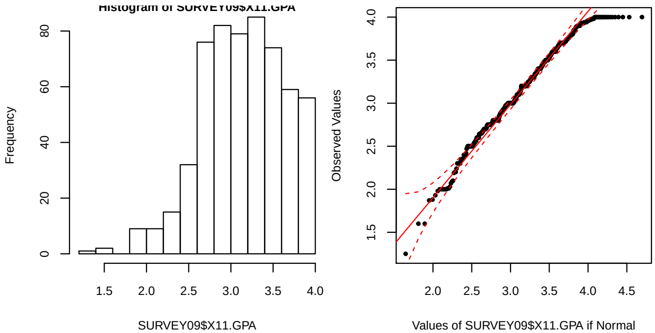
```
hist(ATTRACTM$Score,breaks=seq(1,3.5,.25)); qq(ATTRACTM$Score)
```



About as close to a Normal distribution you'll find, though the histogram doesn't have an "obvious" bell-shape.

QQ plot example2 (small but systematic difference from Normal)

```
hist(SURVEY09$X11.GPA); qq(SURVEY09$X11.GPA)
```



Decent match except for at large GPAs. Some systematic departure from Normality.