ST 516: Foundations of Data Analytics

Other properties of distributions

Other properties of distributions

There are many other properties of distributions that might be of interest beyond the mean and variance.

For any summary of a distribution, remember that there are generally two contexts in which we might use it.

The *sample* summary is a **statistic** we can calculate when we have a sample from the population.

The *population* summary is a **parameter**, or property, of the population, that we could only evaluate if we knew the entire population. We generally think about this as the value the sample summary would approach as we take larger and larger samples.

The median

The median is value for which, the probability of seeing a value below it, is the same as the probability of seeing a value above it.

This is an alternative measure of center to the mean.

The *sample* median is found by finding the middle value in the list of ordered observations (if the number of observations is even, the average of the two central values is often used).

The *population* median is generally defined as the value at which the area under the curve to the right of the value, and the area under the curve to the left of the value is 0.5.

The median can sometimes be more useful than the mean because it is less sensitive to what happens in regions far from the center.

Quantiles/Percentiles

Another name for the median is the 0.5 quantile (or 50th percentile) because the area to the left of the value is 0.5.

Other quantiles can also be defined. For example, the 0.25 quantile, is the value such that the area to the left is 0.25. Or in other words we expect about 25% of values to fall below the 0.25 quantile.

The 0.25, 0.5, and 0.75 quantiles are often referred to as *quart*iles because they divide the possible outcomes into equally likely *quart*ers.

Quantiles/Percentiles

Common quantiles of interest (or reported) include 0.05, 0.25, 0.5, 0.75, and 0.95, but any number between 0 and 1 can define a quantile.

Once again, a *sample* quantile is calculated on a sample, and roughly corresponds to the value such that the corresponding proportion of the data falls below it (although the rules for dealing with what happens when this falls between two observations get more complicated).

The *population* quantile is the number which we would expect our sample quantiles to approach in very large samples.

Interquartile range

The interquartile range (or IQR) is a measure of spread based on quantiles.

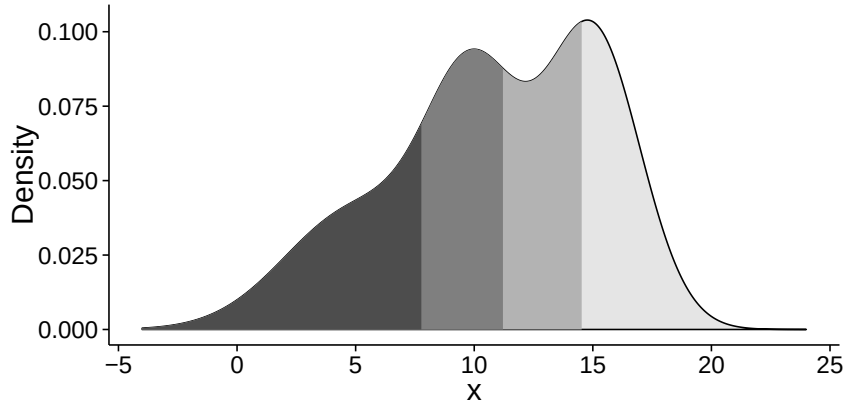
The IQR is the distance between the 0.25 and 0.75 quantile. It measures the extent of the central 50% of the population.

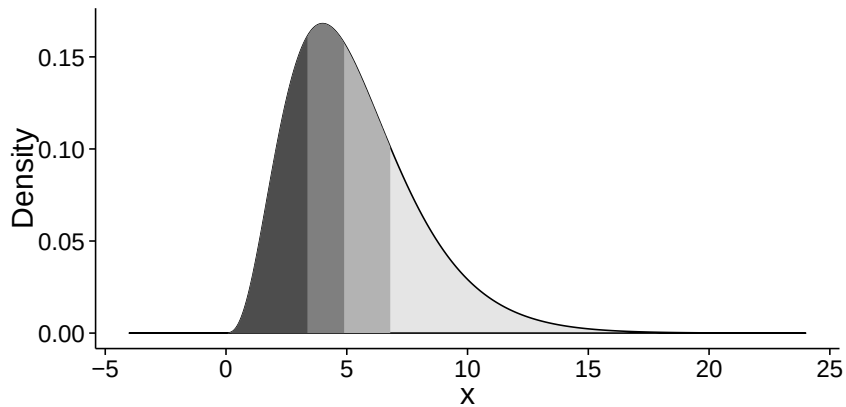
The *sample* IQR is the difference between the *sample* 0.75 quantile and the *sample* 0.25 quantile.

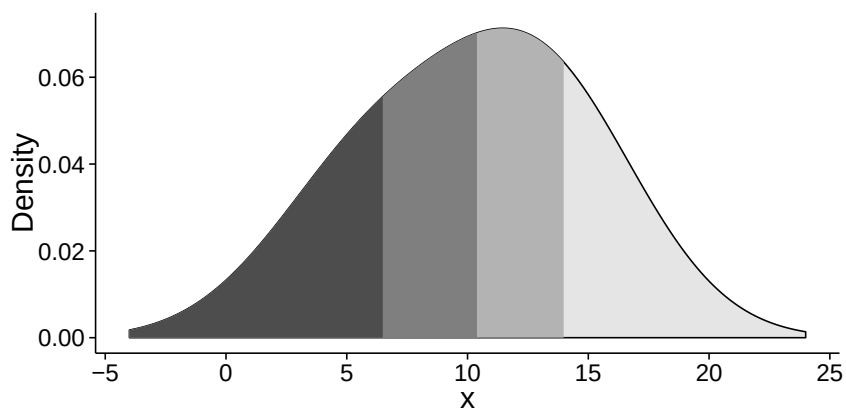
The *population* IQR is the difference between the *population* 0.75 quantile and the *population* 0.25 quantile.

Like the median, the IQR is sometimes used instead of the standard deviation as it less sensitive to what happens in regions far from the center.

Examples


0.100
0.075
0.050
0.025
0.000
Density
−5
0
5
10
15
20
25
X


0.15
0.10
0.05
0.00
Density
−5
0
5
10
15
20
25
X


0.06
0.04
0.02
0.00
Density
−5
0
5
10
15
20
25
X

Mode

Another measure of center is the mode. It has a slightly different meaning between discrete and continuous variables.

For discrete variables, it refers to the value with the highest probability of occuring (or value**s** if more than one value shares the same highest probability of occuring).

For the *sample* this will be the value that was observed most often, in the *population* it is the value with the highest bar.

For continuous variables, it refers to peaks in the distribution curve (*population*) or sample histogram (*sample*). Unlike discrete variables, the mode doesn't refer to only the highest peak.

For continuous variables, we often call a distribution **unimodal** if it has a single peak, and **bimodal** if it has two peaks. More than two peaks? We generally say **multimodal**.

Symmetry and skewness

We say a variables distribution is symmetric, if the shape of the distribution to the right of the center is a mirror image of the distribution to the left of center.

For a symmetric population distribution, the population mean and population median take the same value.

One measure of non-symmetry is **skew**.

We will mostly use skew to describe non-symmetric distibruions in an informal way, but there is a value that can be calculated (for sample or population).

For skewed distributions the mean and median are not equal.

Examples of questions of interest that aren't about means

We'll see a lot of questions of interest that are about population means, but the population mean isn't the only parameter that might be interesting:

- to guarantee a consistent product, a food manufacturing firm might be more interested in the standard deviation of the butter content in their population of cookies
- to make sure they have enough money on hand to pay out claims, an insurance company might be more interested in the 90th percentile of the size of claims in the population of thier clients.