ST 516: Foundations of Data Analytics

The role of independence

Independence

If knowledge about one observation allows us to make a better guess about another observation, there is a lack of independence, or dependence.

If there is dependence, the t-based methods give misleading results - and getting more data won't help!

In the paired case, we require the pairs to be independent.

In the two sample case, we require the observations are independent from other observations in the same sample, and independent from observations in the other sample.

Independence is a common assumption in more complicated methods. It is generally crucial for the validity of the results.

Types of dependence

Two common types of dependence:

Cluster effects

There is some kind of subgroup within samples, and subjects in the same subgroup are more similar.

For example, 50 animals may have been collected from 10 litters and then randomly assigned to one of two treatment groups. Since animals from the same litter may tend to be more similar in their responses (because of their genetic similarities) than animals from different litters, it is likely independence is lacking.

Types of dependence

Proximity effects

Measurements are made in time (or space) and observations made close in time (or space) are more similar. We often call dependence due to closeness in time, temporal correlation, and dependence due to closeness in space, spatial correlation.

For example, you might record prices for a single product at two different websites over 100 days. Since prices on days close to one another are more similar than prices on days far apart, independence is lacking.

Detecting dependence

Generally, if you can specify the kind of correlation you expect, there are methods designed for detecting it. (Not really useful if you don't know what kind of correlation there might be!)

For this class, we will focus on careful reading of the study design to detect possible sources of dependence.

Examples

Researchers interested in investigating the effect of indoor air pollution on respiratory health randomly select houses in a particular city. The houses are monitored for nitrogen dioxide and classified as being either high or low on the nitrogen dioxide scale. Each member of the household is measured for their respiratory health in terms of breathing capacity. The data set consists of these measures of respiratory health for all individuals for houses with low nitrogen dioxide and all individuals for houses with high nitrogen dioxide.

Examples

Researchers interested in investigating the effect of indoor air pollution on respiratory health randomly select houses in a particular city. The houses are monitored for nitrogen dioxide and classified as being either high or low on the nitrogen dioxide scale. Each member of the household is measured for their respiratory health in terms of breathing capacity. The data set consists of these measures of respiratory health for all individuals for houses with low nitrogen dioxide and all individuals for houses with high nitrogen dioxide.

Cluster effect: individuals from the same household are likely to be more similar that individual from different households due to shared living situation and possible relatedness.

Examples

Researchers are interested in a new fertilizer for tomatoes. They divide a growing bed of tomatoes in half lengthways and apply the new fertilizer to one half and their current fertilizer to the other half. At the end of the season, for each plant, they measure the yield of tomatoes. The data set consists of yields for each plant that had the new fertilizer and each plant that had the old fertilizer.

Examples

Researchers are interested in a new fertilizer for tomatoes. They divide a growing bed of tomatoes in half lengthways and apply the new fertilizer to one half and their current fertilizer to the other half. At the end of the season, for each plant, they measure the yield of tomatoes. The data set consists of yields for each plant that had the new fertilizer and each plant that had the old fertilizer.

Proximty effect: Plants nearby in the growing bed are likely to have more similar yields than plants far from each other in the growing bed, due to sharing similar environmental effects like soil quality, sunlight/shading from other plants, proximity to watering etc.

Positive correlation makes point estimates more variable

Positive correlation between data points results in our usual standard error formulas *underestimating the uncertainty* in our point estimate.

Why? Because each observation is related to the others, they don't contribute the same amount of new information as an independent observation would. Having correlated observations, is like having a smaller sample of uncorrelated observations.

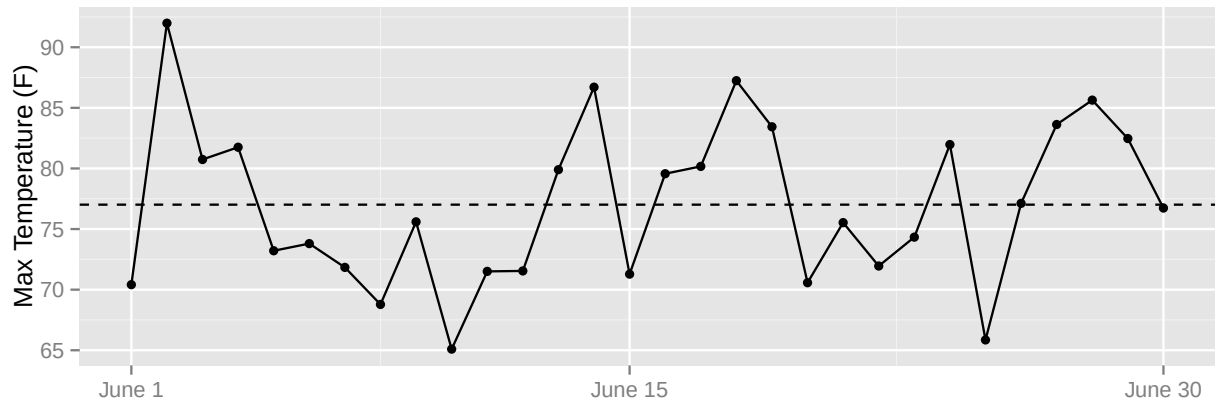
Or in other words, the sampling distribution for our point estimate from a correlated sample has a larger spread than we would estimate assuming an uncorrelated sample.

Without correction, we will declare significance too often, and our confidence intervals will be too narrow.

Example 1: Temporal correlation

Imagine we record the temperature every day in June this year. We want to use these 30 measurements to estimate the mean temperature in June.

Here's what those 30 measurements might look like (for the OSU campus in Corvallis)



I've added a horizontal line at the sample mean. Notice that when the temperature is above the sample mean on one day, it is often followed by another temperature above the mean. This is a classic sign of temporal correlation.

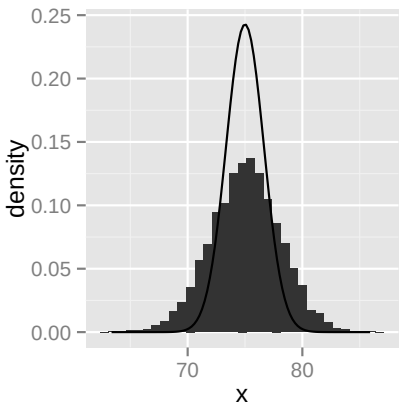
Example 1: Temporal correlation

Let's look at the sampling distribution of the sample mean in this case.

I repeated many times:

- simulate a correlated series of 30 temperatures with population mean 75 and standard deviation 9
- find the sample mean

The histogram of 5000 sample means from the 5000 simulated Junes is shown below along with a Normal distribution with the mean and variance predicted by the central limit theorem (i.e $N(75, \frac{9^2}{n})$).



Example 1: Temporal correlation

The center of the sampling distribution is the population mean: the sample mean is still unbiased.

However, the spread of the sampling distribution is much larger than that predicted by the central limit theorem.

If we use the usual standard error, based on $\sigma/\sqrt{n}$ we will drastically underestimate the uncertainty in our point estimate.

In fact, in this example, 95% CIs based on the usual one sample t procedure, fail to cover the true mean in 29.5% of repeated samples.

Example 2: Paired data

When two samples arise from paired data, there is often dependence between observations that are paired.

For example, in the Schizophrenia case study, if the twin unaffected has a large brain volume, their affected twin also tends to have a large brain volume.

The presence of the dependence violates the assumption of between sample independence for the two sample t procedure.

Last week you demonstrated what happens if we incorrectly analyse paired data using a two sample procedure.

Summary

Be on the look out for sources of dependence.

Presence of dependence will generally invalidate the results from a method that assumes independence.

Getting more data *will not* help!