# Checking Independent Data
# Stat 021 Fall 2020
# Swarthmore College

**Independence Assumption**
The sampled values of your random variable must be independent of each other. Deciding whether or not the independence assumption is plausible depends entirely on your knowledge of the setting and situation.

Even though this assumption isn't something we can usually verify, it is always a good habit to *think about ways in which the data values somehow affect each other.* This is especially true when the data you are collecting is temporally related or spatially related.

When fitting a statistical model, we need to check the independence condition specifically for the *response variable*.

For a sample of *observations of a categorical random variable*, the data is independent if knowing the category one individual belongs to does not tell us anything about the category another individual belongs to. In a chi-square goodness of fit test or when drawing conclusions about population proportions, the response variable is categorical.

For a sample of *observations of a quantitative random variable*, the data is independent if knowing the value for one individual does not tell us anything about the value of another individual. In linear regression, ANOVA models, or when drawing conclusions about population means the response variable is quantitative.

To help us determine the plausibility of an assumption about independence, *in addition to* thinking about the questions above, we should also verify the following two conditions when appropriate.

1. **Simple Random Sample (SRS) Condition**
   If every possible sample of size n has an equal chance of being selected, then the sample is representative of the population because it is a SRS. To check the appropriateness of this assumption we need to think about ways in which the sampling method might have been biased and/or ways in which the sample may not be representative of the population.

   Even if the sample is not exactly a SRS, we can proceed as long as an unbiased *randomization* scheme was used in selecting the sample from the population.

Source: *Stats: Data and Models 2nd Edition.* De Veaux Velleman and Bock. (2008) Pearson Education Inc.

This is most important to check in observational studies because experiments should be designed to incorporate randomization.

2. **10% Condition**
   The problem with sampling (without replacement) from a finite population is that technically the observed values of your random variable are not independent. If you have an infinite (or very large) population however, this is not really an issue. This condition is most important to check when the population itself is small or when conducting inference about population proportions.

**Example 1:** Suppose a cereal manufacturer puts pictures of famous athletes on cards in boxes of cereal, in the hope of increasing sales. The manufacturer announces that in the shipments to Philadelphia's Target supermarkets 20% of the boxes contain a picture of Dwane Wade, 30% a picture of Serena Williams, and the rest a picture of Simone Biles. You randomly select 2 boxes of cereal to buy from the Target closest to you.

The random variable you are modeling is a categorical variable with three levels, each of the famous athletes.

**Q:** Are the boxes you select independent of one another? In other words, if you find a Dwane Wade card in one box, does this have any effect on the probability of finding his card among the other boxes?
**A:** Technically, yes, because if exactly 20% of the boxes have Dwane Wade's card, then when you find one, you've reduced the number of remaining Wade cards. With a few hundred thousand boxes of cereal to choose from however, this difference is negligible. But, if we knew there were 2 Dwane Wade cards hiding in the 10 boxes on this particular shelf, then finding one in the first box you try would clearly change your chances of finding this card in the next box you try.

**Example 2:** Suppose 90 students filter into a 200 seat school auditorium. As these students sit down, you record whether or not each student is left-handed.

The random variable you are modeling is the binary categorical variable, whether or not someone is left-handed.

**Q:** Is student hand dominance independent?
**A:** Mostly likely, yes, the probability of one student being left-handed is not changed by the fact that another student is left or right handed. Hence even if the 90 students are not randomly selected, there is no reason to suspect the data is not independent.

Source: *Stats: Data and Models 2nd Edition.* De Veaux Velleman and Bock. (2008) Pearson Education Inc.

(However, if not randomly selected, then this group of students may not be representative of the population!)
**Example 3:** Someone loads 20 moving boxes onto a service elevator in an apartment complex and you record their weight.

The random variable you are modeling is the quantitative variable, weight.

**Q:** Is the weight of one box going to contain any information on the weight of another box?
**A:** It depends. We were given no information about how the sample of boxes was drawn or what was the population of interest. If the population of interest is a class of students moving onto campus and the boxes are sampled randomly, then the answer is no, and the weights are independent. However if the boxes are not randomly selected, then the answer could be yes because perhaps international students pack lighter than domestic students, for example.

**Example 4:** Suppose you are collecting data on the speed of cars on I-676. You record the speed of 25 cars sampled randomly throughout the same day.

The random variable you are modeling is the quantitative variable, car speed.

**Q:** Is the speed of one car going to affect the speed of another car?
**A:** Although this is a convenience sample, as long as the sampled cars were not driving near each other, then their recorded speeds are plausibly independent.

**Using Assumptions as Disqualifiers**
The reason we check the assumptions for various statistical methods is to know whether we can make a meaningful conclusion from the data. In general, we proceed with our analysis unless there is a serious problem with an assumption. If we find minor issues, the best thing to do is to take note of them in our analysis and express caution about our results.

- If the sample is not a SRS but we believe it is representative of some population, limit our conclusions to the corresponding population.
- If there are outliers, perform the analysis both with and without them.
- If the sample looks bimodal, try to analyse the subgroups separately.

Only when there's a major issue, such as a strongly skewed small sample or an obviously non-representative sample, are we unable to proceed at all.

Source: *Stats: Data and Models 2nd Edition.* De Veaux Velleman and Bock. (2008) Pearson Education Inc.