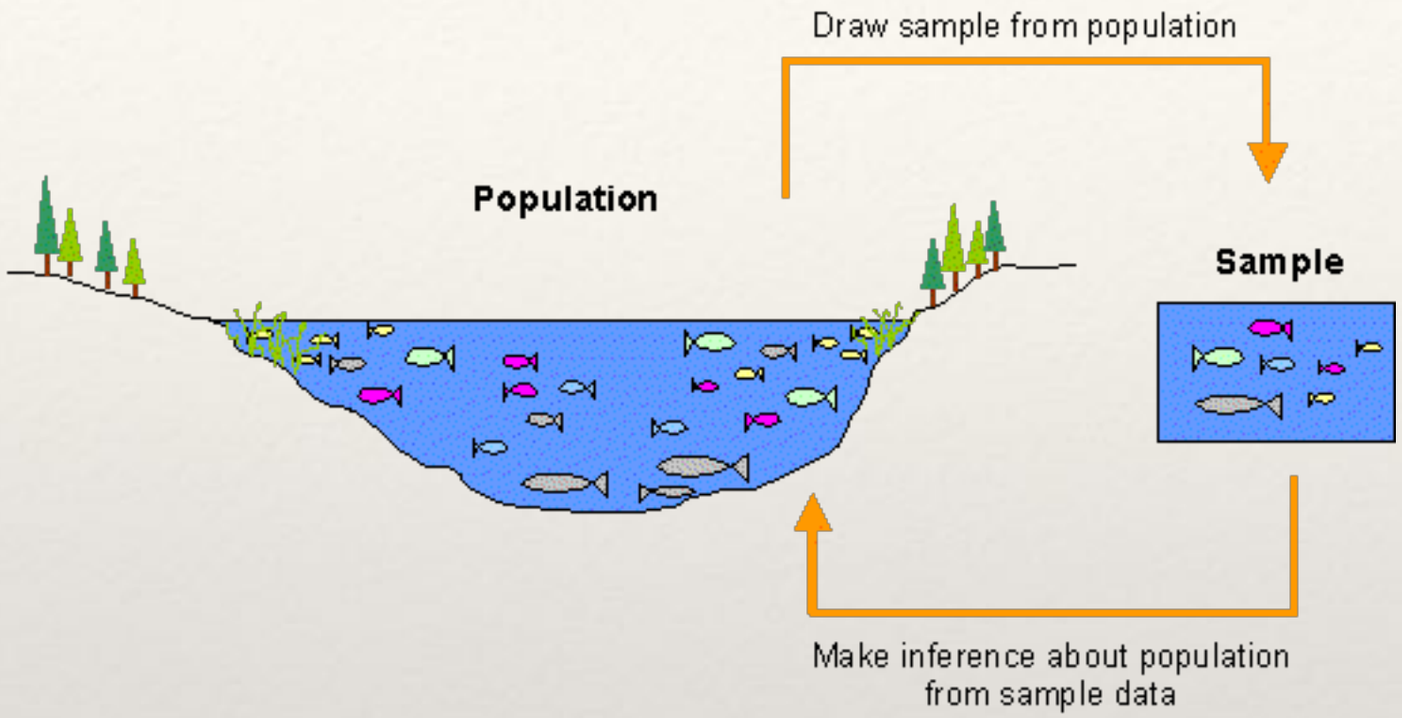
March 5, 2018

CS 361: Probability & Statistics

Samples and populations + hypothesis testing

Review: 3 means, 3 variances



Source: http://labs.geog.uvic.ca/geog226/frLab3.html

Weight of fish in the sample and population both have a mean and variance

The sample mean is a random variable which means it has a distribution, i.e. an expected value (or mean) and a random variable variance.

To analyze whether the sample mean is a good estimator of the population mean, we will have to analyze the expectation and variance of the random variable which describes the randomness according to which we got our sample

Estimating population statistics

We showed that for estimating the population mean, the sample mean was a good single-number estimate.

We also analyzed the distribution of the sample mean as a random variable and showed how we can come up with an **interval estimate** for the population mean from the sample that we called a **confidence interval**

We will usually use the 95% confidence interval in our examples. The confidence interval is wide enough so that 95% of samples drawn according to our sampling assumptions would contain the true population mean within the interval

Confidence intervals for samples

What was nice about the population mean was that our estimate for it—the sample mean—was a normal random variable whose standard deviation—the standard error—was easy to reason about.

This meant that determining a confidence interval involved invoking what we know about how much of a normal dataset is going to be within *k* standard deviations of its mean

The dataset in this case is a bit abstract, it's every possible sample that we could draw and each sample with N points is a single point in this normally distributed dataset, but our conclusions about normally distributed datasets still hold

Confidence intervals for samples

The true value of the population mean is contained in the interval

Sample standard deviation

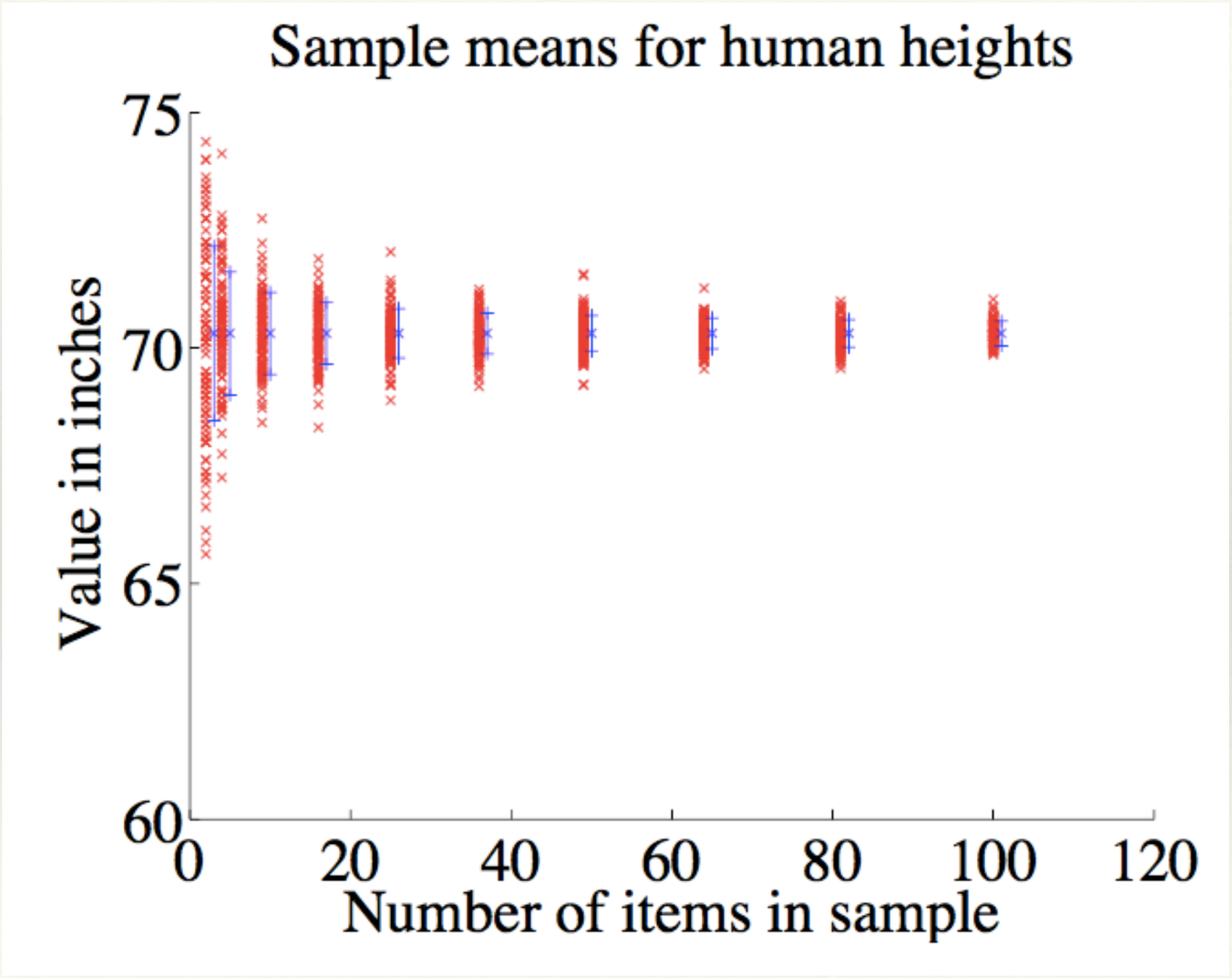
$$X^{(N)} - \frac{sd(\{x\})}{\sqrt{N}} \leq \text{popmean} \leq X^{(N)} + \frac{sd(\{x\})}{\sqrt{N}}$$

in 68% of samples we could have drawn. Or in 95% of samples we have

$$X^{(N)} - 2\frac{sd(\{x\})}{\sqrt{N}} \leq \text{popmean} \leq X^{(N)} + 2\frac{sd(\{x\})}{\sqrt{N}}$$

We call these intervals the 68% and 95% confidence interval, respectively

Simulation



251 human heights in the population and various sample estimates of the mean

popmean and 1 popsd standard error bars in blue. Estimates from simulating the sampling process in red

Caveat

If our sample size N isn't sufficiently large, the sample mean doesn't exactly follow a normal distribution. We will talk a bit later about what distribution we can use to better describe the sample mean for a smaller N

Estimating other quantities

In general, we might consider any **statistic** of a sample and want to come up with an interval of estimates for what would happen if we could measure that statistic in the whole population

Unfortunately it isn't always straightforward to calculate the standard error of our estimate. If we were trying to estimate the median, for example, the standard error of the sample median is a bit more difficult to reason about

The bootstrap

Now if we had multiple samples, we would probably want to combine them into one larger sample in order to estimate some population parameter. But if we had multiple samples we could also compute the statistic for each sample and maybe use the range of values we get to come up with an interval estimate for the statistic of interest

The bootstrap is a process to generate synthetic datasets from a single dataset to do just that.

Suppose we have a sample of N items from a population. We compute a **bootstrap replicate** of the sample by sampling N items from our original sample, uniformly and with replacement. We can do this as many times as we wish.

Since we are sampling with replacement, in general our bootstrap replicates won't just be a copy or a simple permutation of our original data.

Example

Suppose we have the following sample of heights of UIUC students in inches: {60, 58, 54, 62, 64, 73, 65, 62}. We sample from this dataset uniformly and with replacement to get the following bootstrap replicates of our dataset

{64, 60, 65, 54, 65, 54, 64, 64}
{60, 73, 60, 62, 62, 73, 58, 60}
{54, 62, 65, 54, 73, 64, 65, 64}
{64, 58, 60, 60, 60, 54, 65, 62}

Estimating the standard error

If we compute *r* different bootstrap replicates, we can use the following to get a good approximation of the standard deviation of our estimate

Compute S({**x**}_i) — that is the value of the statistic of interest on the i-th replicate, for each replicate. Use all *r* values we get this way to calculate:

$$\overline{S} = \frac{\sum_i S(\{\mathbf{x}\}_i)}{r}$$

Calculate

$$\text{stderr}\left(\{S\}\right) = \sqrt{\frac{\sum_i \left[S(\{\mathbf{x}\}_i) - \overline{S}\right]}{r-1}}$$

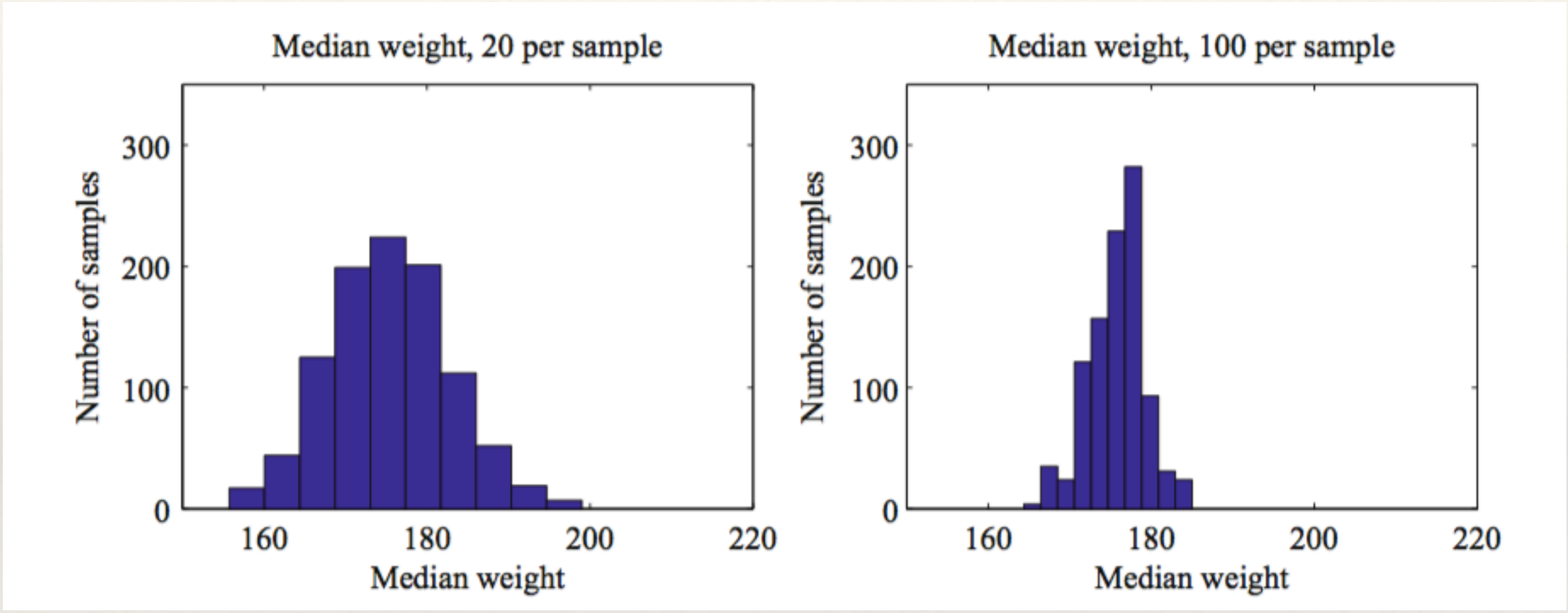
Estimating other quantities

Now we can construct a confidence interval by looking at the value of our statistic for each of the bootstrap replicates and giving a range that contains, say, 95% of the data

Or if we have reason to believe that our estimate follows a normal distribution, we can use the standard error to bound the data. A 95% confidence interval would be given by reporting

$$\overline{S} \pm 2\text{stderr}$$

Example


Median weight, 20 per sample
Number of samples
300
200
100
0
160
180
200
220
Median weight
Median weight, 100 per sample
Number of samples
300
200
100
0
160
180
200
220
Median weight

Hypothesis testing

Hypotheses and evidence

Assume we have a hypothesis. We will be looking at ways to assess how much the data **contradicts** the hypothesis

We aren't interested in how much the data **supports** the hypothesis because in general we don't prove scientific hypotheses true, we just fail to show them to be false. One piece of data can show a hypothesis to be false, a mountain of data still can't show it is true

Examples

Patriot missiles: Pentagon claims 80% accuracy. A reporter saw 13 out of 14 misses

MTG-DAF: we're told the deck has 24 lands and 36 spells. We draw a 7 card hand and get no lands 3 times in a row

Does the population have this mean?

Suppose we believe the average human body temperature is 95 degrees. Let $\overline{T}$ be the random variable evaluated by collecting a random sample of people, measuring their temperatures, and computing the average of these temperatures. This is the sample mean, then. Which means its expected value is the population mean and its standard error is *s*

If our hypothesis is true, then the following is a standard normal random variable

$$G = \frac{(\overline{T} - 95^o)}{s}$$

We can now tell whether the evidence contradicts our hypothesis because we can calculate how unusual the sample we actually observed would be if the hypothesis were true

Does the population have this mean?

Denote the actual value we observe for the random variable $\overline{T}$ as $\overline{t}$ and use it to calculate g

$$g = \frac{(\overline{t} - 95^o)}{s}$$

Now if our hypothesis is true G is a standard normal random variable which means that for 68% of the temperature samples we could have drawn this **test statistic** above, g, would have been between -1 and 1

If we calculate

$$f = \frac{1}{\sqrt{2\pi}} \int_{-|g|}^{|g|} \exp\left(\frac{-u^2}{2}\right) du$$

f is the fraction of samples, assuming our hypothesis is true, which would have had values less extreme than the one we observed

Does the population have this mean?

So a value of g with a large absolute value gives an f very close to 1. Which means that if our hypothesis was true, we have observed a very unusual sample —most samples would have given a value of g closer to 0. If assuming our hypothesis is true makes our data highly unusual, we can say that the data fails to support the hypothesis

Traditionally we look at p=1-f which is called the **p-value**

We reject the hypothesis we are evaluating if the p-value is sufficiently small. The cutoff used in many scientific disciplines is p=0.05 or smaller. A p-value of 0.05 means that you would see evidence this unusual in about one experiment out of twenty if the hypothesis were true

Example: outliers

Assess the hypothesis that the average BMI of humans is 27

We have a dataset of 252 heights/weights. Two are outliers which we discard. We get a mean BMI of 25.3 and a standard deviation of 3.35

Giving a standard error of 0.21 and a test statistic of -8.1 which implies a p-value of almost 0. Which means this is an extremely unusual dataset if the hypothesis is true, so we should reject the hypothesis

If we include the outliers, the mean BMI is 25.9 and the standard deviation is 9.56. This causes us to have a p-value of 0.08 which might make us accept the hypothesis

Summary

If we want to hypothesize about the mean of a population based on a sample, rejecting that hypothesis will occur if the dataset we observe would have needed to be quite unlikely, given the hypothesis

The p-value tells us what proportion of possible samples would have had a less unusual value than the one we observed if the hypothesis were true

Outliers can blow up the standard deviation dramatically and can therefore lower our threshold for non-rejection of a hypothesis

Note the connection to confidence intervals: there we knew the distribution of the sample mean (approximately) and reported an interval where the sample mean would lie in 95% of possible samples. Hypothesis testing asks how large of a confidence interval we would have to draw to enclose the test statistic