Comparing Typical Values Across Groups

36-313, Fall 2021

16 September 2021 (Lecture 6)

Contents

We will now begin comparing groups to each other. The methods are the same whether we're comparing two totally distinct populations, or two sub-sets of the same population. I will assuming that we are comparing a single quantitative, numerical characteristic. To keep everything suitably abstract and generic, I'll call the groups A and B; the values of the characteristic for group A will be X, and values for group B will be Y.

By "typical values", what I have in mind is some sort of one-number summary of central tendency of a distribution, or of a sample from a distribution. We've looked at a bunch of these: the median, the mean, the mean of the log, etc. The methods are pretty much the same regardless of *which* notion of typical value we're using, so I'll keep this pretty generic.

The Setting and the Two Fundamental Problems

We have a sample from group A, say $x_1, x_2, \ldots x_{n_A}$. We plug these sample values in to the appropriate formula and get a certain typical value m_A. We also have a sample from group B, $y_1, y_2, \ldots y_{n_B}$, leading to typical value m_B. I need a symbol for the difference $m_A - m_B$, so I'll call it v. When I need to distinguish this observed v from some other hypothetical or imaginary or possible difference, I'll write it v_{obs}.

Here, as a running example, is a sample data set for group A

observation	value
x_1	5
x_2	7
x_3	8
x_4	6

and for group B:

observation	value
y_1	2
y_2	10
y_3	4

If we use the median as our typical value, $m_A = 7$, $m_B = 4$, and $v_{obs} = 3$.

Your own data sets will be fundamentally like this, but larger.

I can now pretty much guarantee you, without looking at your data or knowing anything about your problem at all, that $m_A \neq m_B$. Congratulations, your data shows inequality between the two groups!

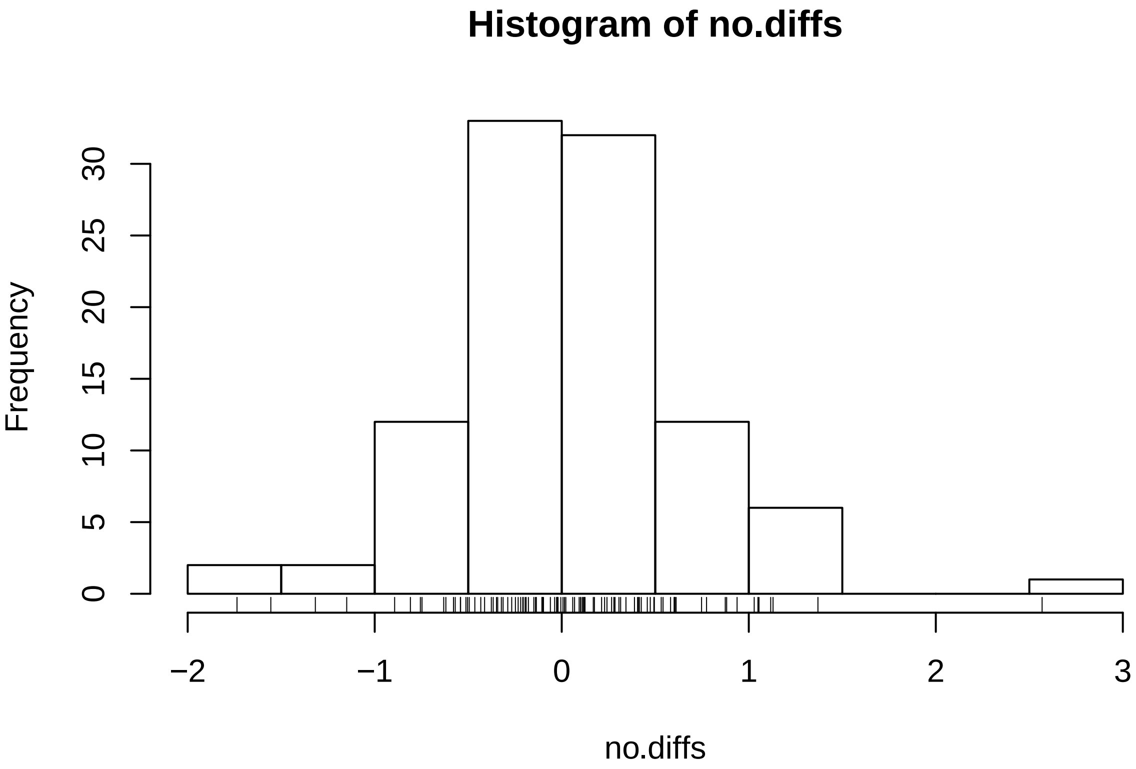
The reason I can make that guarantee is that if I take two samples from the *same* distribution, they will almost always have different means, different medians, etc., unless the population distribution is truly weird.

Here, for instance, I have drawn 10 values from a standard log-normal $\mathcal{LN}(0, 1)$ distribution, and then another 10 values from the same standard log-normal, and calculated the difference in medians:

```
median(rlnorm(10)) - median(rlnorm(10))
```

```
## [1] 0.1583089
```

Lo! It's not 0. This histogram shows what happens when I repeat that exercise 100 times, with little tick marks on the horizontal axis showing the actual values:



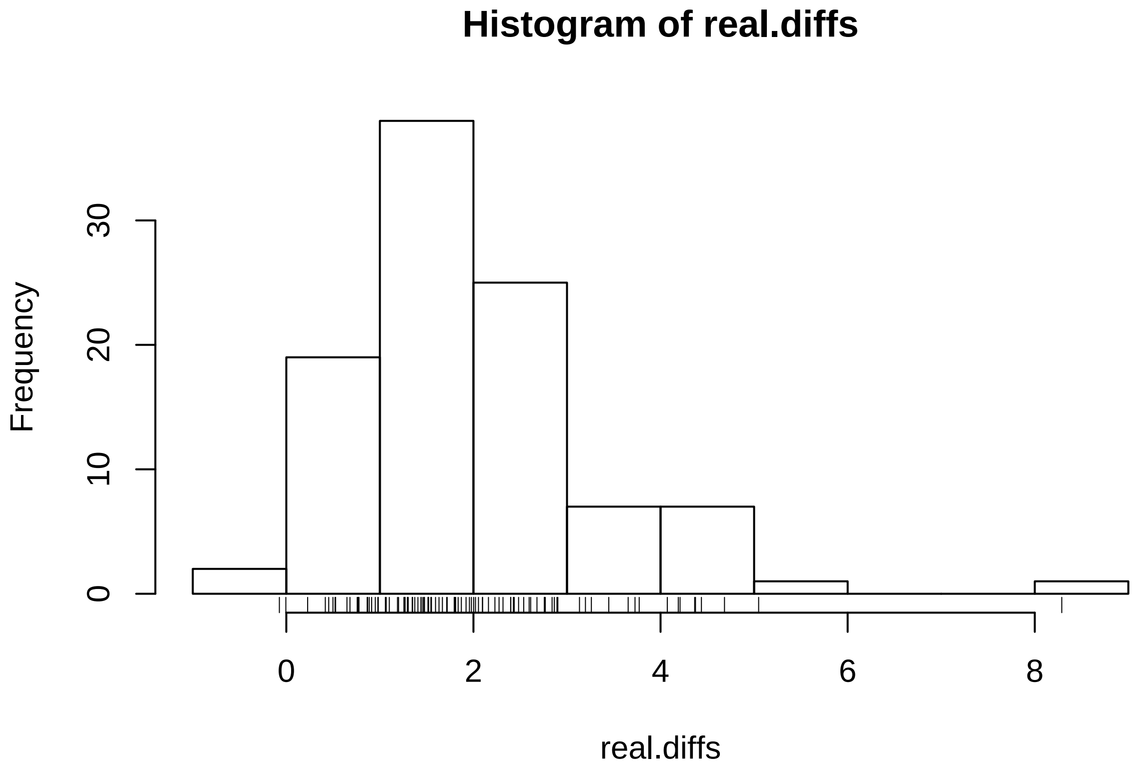
This gives us our first fundamental question in comparing typical values:

> Is the observed difference real, or just sampling noise?

That's still a bit vague. Here is something which is more precise, and potentially calculable:

> How likely is it that we'd see as big a difference as we do, if the two groups really had the same typical value, and the apparent difference is just sampling noise?

If there really are differences, sampling noise doesn't go away. Here one of the groups gets drawn from a $\mathcal{LN}(1,1)$ distribution, while the other is drawn from a $\mathcal{LN}(0,1)$ distribution. At the level of the whole distribution, the difference in medians is $e^1 - e^0 = 1.72$. But if we only draw 10 samples from each group, we're not going to be able to see that very precisely:



Our second question or problem is this:

> How can we use our data to attach some measure of uncertainty or margin of error to our estimate of the difference?

There are a couple of ways of making this more precise, and we'll look at two of them later.

Testing for no difference at all

First, let's think about how we'd test the idea that maybe there's really no difference at all between group A and group B, so X and Y follow exactly the same distribution. This is a fairly extreme statistical hypothesis, it's rare to find two actual social groups people care about where this is true of anything[1], but it can form a useful baseline, and the trick involved is very pretty.

Start by supposing that the two groups really do have exactly the same distribution. Then we might as well merge the two data sets. Here is an example of doing so, using our running example:

value	group
5	A
7	A
8	A
6	A
2	B
10	B
4	B

[1] The technique we're about to discuss actually originated in the analysis of experiments, specifically in the analysis of agricultural experiments, where, at least in the early days, it *wasn't* so strange to imagine that lots of treatments actually did exactly nothing.

I have added an extra column here which records which group each value was originally sampled from. We've **pooled** the data from the two groups, because doing so gives us a better idea of the common underlying distribution, assuming, of course, that there is *a* common underlying distribution.

Now here is the trick. If there really is just one distribution here, then there's no relationship between the observed values and the group labels. So we can **simulate** drawing two samples from this common distribution, of sizes n_A and n_B, calculating the typical values for those samples, and looking at the differences. This gives us an idea of whether the difference we *observed* is big or small compared to what we'd expect from pure sampling noise.

The way we do the simulation is to take the entries in the "group" column and shuffle them at random into a new order. (Back in the day, some statisticians really did write the group labels on cards and then shuffle the cards.) Now split the data into two parts again, based on the new, shuffled group labels.

value	group	shuffled group
5	A	A
7	A	B
8	A	B
6	A	B
2	B	A
10	B	A
4	B	A

This gives us two new typical values, for the "groups" after shuffling, say m^*_A and m^*_B, in this case $m^*_A = 4.5$ and $m^*_B = 7$. The apparent difference after shuffling is $v^* = -2.5$.

Now, there is nothing particularly special about this one shuffling of the group labels, but it is representative of the kind of difference we can get just by sampling from the pooled data. To get a picture of the *distribution* of differences that can arise from sampling, we repeat this many times, with a different shuffling of the group labels each time, getting many values of v^*. We record all the v^*s, and then we see how extreme v_{obs} is compared to this distribution of sampling differences.

Since "shuffling the group labels at random" is a mouthful, and makes us sound like gamblers, we instead talk about random permutations. The procedure for a **permutation test** is then as follows.

0. Start with data $x_1, \ldots x_{n_A}$ and $y_1, \ldots y_{n_B}$, calculate typical values m_A and m_B and difference $v_{obs} = m_A - m_B$.
1. Repeat b times:
    a. Randomly permute the group labels
    b. Use the new group labels to divide the data into $x^*_1, \ldots x^*_{n_A}$ and $y^*_1, \ldots y^*_{n_B}$
    c. Calculate m^*_A and m^*_B and $v^* = m^*_A - m^*_B$
    d. Record v^*
2. Find the number m of v^* values where $|v^*| \geq |v_{obs}|$.
3. Return m/b as the (approximate) p-value[2].

Step (1) is doing many random permutations, and seeing how big a difference each one generates between the groups. By construction, in step 1, both groups are sampled from the same distribution, so the difference between the samples is purely due to sampling noise. In step 2, we ask how many of those simulated differences are *at least as big* as the difference we observed[3]. Step (3) summarizes the test in the form of an approximate p-value.

I say an "approximate" p-value, because there are two sources of error in this calculation.

[2] Notice that this can give 0 if $|v_{obs}|$ is greater than all the simulated $|v^*|$. It can be a bit embarrassing to report a p-value of 0, so some people prefer to use $\frac{m+1}{b+1}$. If b is at all substantial, this will make hardly any difference, except when m is very small.

[3] Strictly speaking, this is a two-sided permutation test. This is usually what you want if you're testing the idea that there's absolutely no difference between the groups. You can work out how a one-sided test would differ.

1. We *only* have a limited set of samples, n_A of them from the first group and n_B from the second. If those initial samples had been different, we'd have seen a somewhat different distribution for v^*.
2. We have done only b random permutations.

The first source of error is limited information. We can only make it smaller by getting more data. The second source of error is computational: we decided to do only b permutations. We can make the computational error as small as we like by making b larger. So how large should we make it?

How many permutations, i.e., how big should you make b?

Ideally, we'd consider every possible permutation of the group labels exactly once. The difficulty is that there are $\binom{n_A+n_B}{n_A}$ possible permutations, and this grows *very* rapidly. Even for our little example, $\binom{7}{4} = 35$, which is tedious but could be ground through. If we made both samples ten times larger, $\binom{70}{40} = 6 \times 10^{19}$. Even if each permutation took the computer a millionth of a second, we'd be looking at about 2 million years.

Now, if we really needed all possible permutations, we'd say that the permutation test is a cute idea but just not practical. The trick that makes it work is that we *don't* need all possible permutations, we can randomly generate a limited number of them, b. This is like drawing a random sample from the distribution of permutations, and we know that random samples quickly become representative of the whole distribution.

We're specifically interested in using the random permutations to approximate a p-value. A reasonable guess, based on general statistical intuition, is that the approximation error will be inversely proportional to the square root of the number of permutations, $\propto 1/\sqrt{b}$. (This is actually true, though it's a bit more involved to show than is useful here.) This suggests that there should be diminishing returns to running more and more permutations. Going from 10 permutations to 100 shrinks our error by a factor of 10, but going from 100 permutations to 200 shrinks it only by another factor of $\sqrt{2} \approx 1.4$. But every permutation takes just as long to run as the one before. So we're paying the same cost in computer time but getting less and less extra accuracy for our pains.

The *optimal* value of b would depend on how much we value accuracy versus how much we value computer time. It's rarely worth *human* time to be precise about this. As a rule of thumb, though, a few hundred or a few thousand permutations is typically plenty. Unless you have an unusual need for precision, or each permutation is really slow, $b = 5000$ is a nice round number[4].

Some code

In R, there is a handy and simple function to permute a vector:

```r
sample(t)
```

returns a random permutation of the vector `t`. I've used it to generate the random shuffling of the group labels above.

Let's build a simple function that we do a permutation test on two vectors.

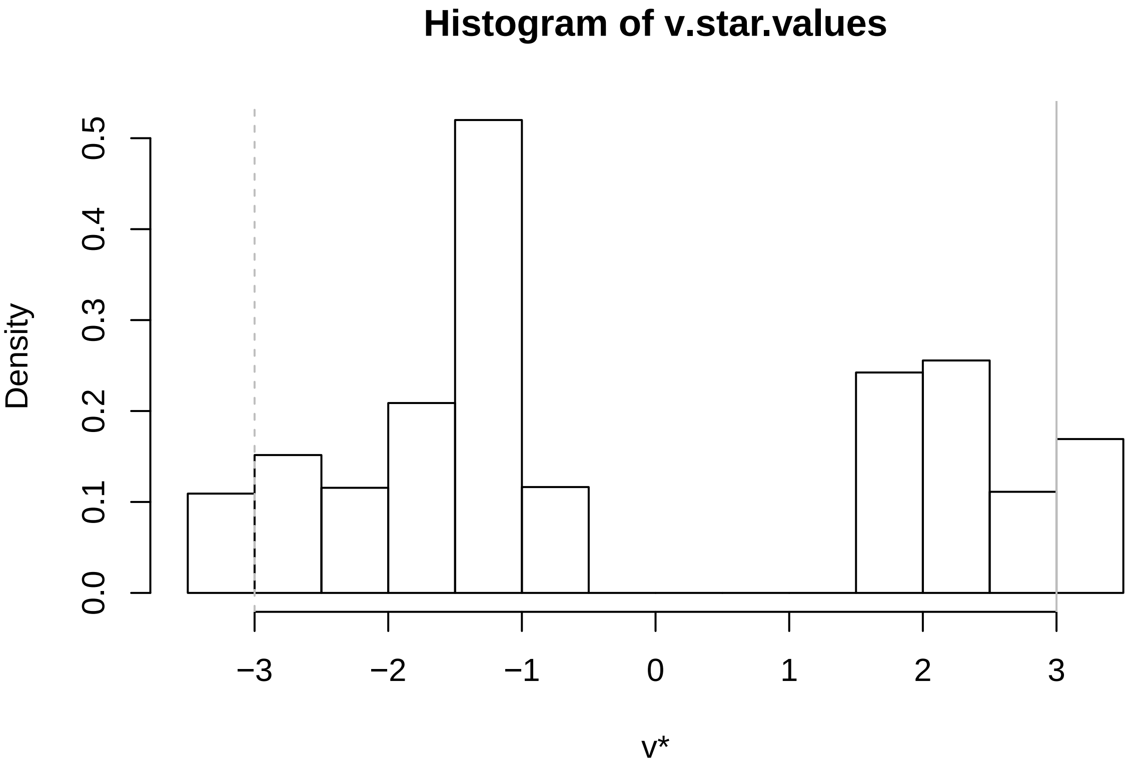
```r
permutation.test <- function(x, y, n.perm = 5000, fn = mean) {
    pooled <- c(x, y)
    labels <- c(rep("A", times = length(x)), rep("B", times = length(y)))
    v.star <- replicate(n.perm, {
        shuffle <- sample(labels)
        fn(pooled[shuffle == "A"]) - fn(pooled[shuffle == "B"])
    })
    v.obs <- fn(x) - fn(y)
    p.value <- mean(abs(v.star) >= abs(v.obs))
```

[4] Or 4999, if you're using the add-one trick from a previous footnote.

```
    return(list(p.value = p.value, draws = v.star))
}
```

The `replicate()` function takes two arguments. The first is the number of times to repeatedly do something. The second argument is an R command, or block of R commands, which it will repeat. This is most useful when the commands do something random. Here, the random thing is permuting the group labels, using `sample()`.

When I try this out on the running-example data, I get the following histogram of v^* values.



I've added vertical lines at v_{obs} (solid) and at $-v_{obs}$ (dashed). The p-value corresponds to the area under the histogram *outside* of those boundaries. In this case, that's 0.1948. This means that even if these two data sets really came from exactly the same distribution, we've still got about 1 chance in 5 of producing at least as big a difference between two samples of that size.

Interpretation of the p-value

The interpretation of the p-value is asymmetric[5]. A small p-value is more or less evidence *against* the two distributions being equal. The difference between the two samples is so big that it's really unlikely sheer sampling noise could have made it.

On the other hand, a large p-value isn't, on its own, evidence *for* the two distributions being exactly the same. It does tell us that it's quite likely for sampling noise to give us that big a difference (or bigger). But maybe *something else* could also give us that big a difference! In particular, maybe the two distributions are very similar but not, exactly, identical…

This suggests that we might well want to know *how big* a difference between the distributions is compatible with our data. That is the topic of the next section.

[5] If you find the logic here intriguing, I strongly recommend the works of Deborah Mayo, particular Mayo (1996) and Mayo and Cox (2006).

Avoiding the permutation test

Suppose we use the mean as our typical value. Then the central limit theorem tells us the following: as $n_A \rightarrow \infty$, if $\mathrm{Var}\left(X\right) < \infty$, then

$$m_A \rightsquigarrow \mathcal{N}(\mathbb{E}\left(X\right), \mathrm{Var}\left(X\right)/n_A) \tag{1}$$

Similarly

$$m_B \rightsquigarrow \mathcal{N}(\mathbb{E}\left(Y\right), \mathrm{Var}\left(Y\right)/n_A) \tag{2}$$

Since we're assuming independent samples from the two groups,

$$m_A - m_B \rightsquigarrow \mathcal{N}(\mathbb{E}\left(X\right) - \mathbb{E}\left(Y\right), \mathrm{Var}\left(X\right)/n_A + \mathrm{Var}\left(Y\right)/n_A) \tag{3}$$

We could use this to calculate a p-value *without* having to go through the elaborate ritual of permutation testing; we'd just need to get estimates of the two variances, say $\hat{\sigma}^2_X$ and $\hat{\sigma}^2_Y$, and then compare $m_A - m_B$ to an $\mathcal{N}(0, \hat{\sigma^2}_X/n_A + \hat{\sigma}^2_Y/n_B)$ distribution. Equivalently, we compare

$$\frac{m_A - m_B}{\sqrt{\hat{\sigma^2}_X/n_A + \hat{\sigma}^2_Y/n_B}} \tag{4}$$

to a standard Gaussian $\mathcal{N}(0, 1)$ distribution. This is a ***z*-test** for the difference in means being 0.

One advantage of the z-test is that it does *not* presume the two groups have exactly the same distribution, just that they have the same mean[6].

There are several reasons why I have emphasized the permutation test over the z-test.

1. The permutation test works in exactly the same way if we employ a different notion of "typical value" than just the mean, e.g., the median. Lots of these other notions of typical value also have central limit theorems (the sample median does!), but with much more complicated variances that you need to work out in each case.
2. More seriously, unless both groups follow a Gaussian distribution to start with, the central limit theorem is an asymptotic approximation, getting better and better as n_A and n_B both tend toward infinity. But the less Gaussian the distributions are to start with, the slower the convergence in the CLT. It is particularly slow for heavy-tailed distributions[7]. But, as you're sick of hearing by now, heavy-tailed distributions are really common whenever we study social inequality.

The Bootstrap, Resampling, and Uncertainty Quantification

We saw above that sampling noise can give rise to a whole distribution of apparent differences between two groups, even when they follow the same distribution. We also saw, in the first section, that when there *are* differences between the two groups, sampling noise can make it hard to discern just how big the difference is. The observed difference between the samples, v_{obs}, is a **point estimate** of the true difference between the populations. We would like to say something about how **precise** or **uncertain** that estimate is.

The ideal would be to repeat our data-generating process (experiment, survey, ...) many times, re-calculating v each time, and so build up a picture of the **sampling distribution** of V. The standard deviation of the sampling distribution would be the **standard error** of V. Knowing the percentiles of the sampling distribution would let us find confidence intervals for the true difference, i.e., give **interval estimates**.

[6] If we're *also* willing to believe that the two groups might have the same variance, we should pool the data to estimate one variance, say $\hat{\sigma}^2$, and the formulas simplify a little.

[7] A result called the Berry-Essen theorem gives a bound on how far the CDF of the sample mean can be from the CDF of a Gaussian. This involves the third absolute moment of the distribution after centering, $\mathbb{E}\left(|X - ExpectX|^3\right)$, more exactly the ratio $\mathbb{E}\left(|X|^3\right)/\mathrm{Var}\left(X\right)^{3/2}$. For heavy-tailed distributions, this ratio can be extremely large, implying very slow convergence of sample means to Gaussians. See Complementary Problem 3.

The only drawback is that we can almost never repeat our data-generating process exactly[8]. What we can do, however, is to use the data to **simulate** repeatedly drawing samples and calculating the difference between them. This procedure is called **resampling** (for reasons that'll be clear in a moment) or **bootstrapping** or **the bootstrap** (for reasons I'll explain later).

Here's how to do resampling. We treat our sample from group A as though it were the entirely population, and draw another sample from it, of exactly the same size, with replacement. (That is, we **resample** group A.) We do the same thing to group B. This gives us new values $x^*_1, \ldots x^*_{n_A}$ and $y^*_1, \ldots y^*_{n_B}$. We use these to calculate new values m^*_A and m^*_B, and the difference $v^* = m^*_A - m^*_B$. We repeat this many times, and build up the distribution of v^* in this way.

So far, this sounds very much like the permutation test. The small but crucial difference is that we *never* swap values between the two groups. Our x^*s all come from resampling the original xs, and our y^*s all come from resampling the original ys. The x^*s don't have *exactly* the same distribution as the xs, but the distribution is close, and the difference between them will be like the difference between a sample and the larger population, because we are, after all, drawing a sample.

Having obtained a lot of v^* values by resampling in this way, there are two common ways to boil the summary down into a measure of uncertainty, or margin of uncertainty, we can apply to v_{obs}:

1. The **bootstrap standard error** in v_{obs} is just the standard deviation of the v^*s.
2. The **bootstrap confidence interval** is a little more involved. If we want a confidence level of $1 - \alpha$, we find the $\alpha/2$ quantile of the v^*s, say $v^*_{\alpha/2}$, and likewise $v^*_{1-\alpha/2}$, and report $(v^*_{\alpha/2}, v^*_{1-\alpha/2})$ as the confidence interval.

There are a *lot* of variations, extensions and refinements to this basic scheme, but this is the core of it.

Some code

The `boot` package offers a powerful and flexible set of functions for doing bootstrapping, even for quite complicated situations, but it will be character-building, and informative, to roll our own.

```
bootstrap.comparison <- function(x, y, n.boot = 5000, fn = mean, level = 0.95) {
    v.obs <- fn(x) - fn(y)
    v.star <- replicate(n.boot, fn(resample(x)) - fn(resample(y)))
    std.err <- sd(v.star)
    lower <- quantile(v.star, (1 - level)/2)
    upper <- quantile(v.star, 1 - (1 - level)/2)
    return(list(uncertainty = c(point.estimate = v.obs, std.err = std.err, lower.ci = lower,
        upper.ci = upper), draws = v.star))
}
```

If you just try to type this in and run it, you'll get an error. That's because the code calls a helper function, `resample()`, which isn't part of base R. But I can define that easily enough[9]:

```
resample <- function(x) {
    sample(x, size = length(x), replace = TRUE)
}
```

Let's try this out:

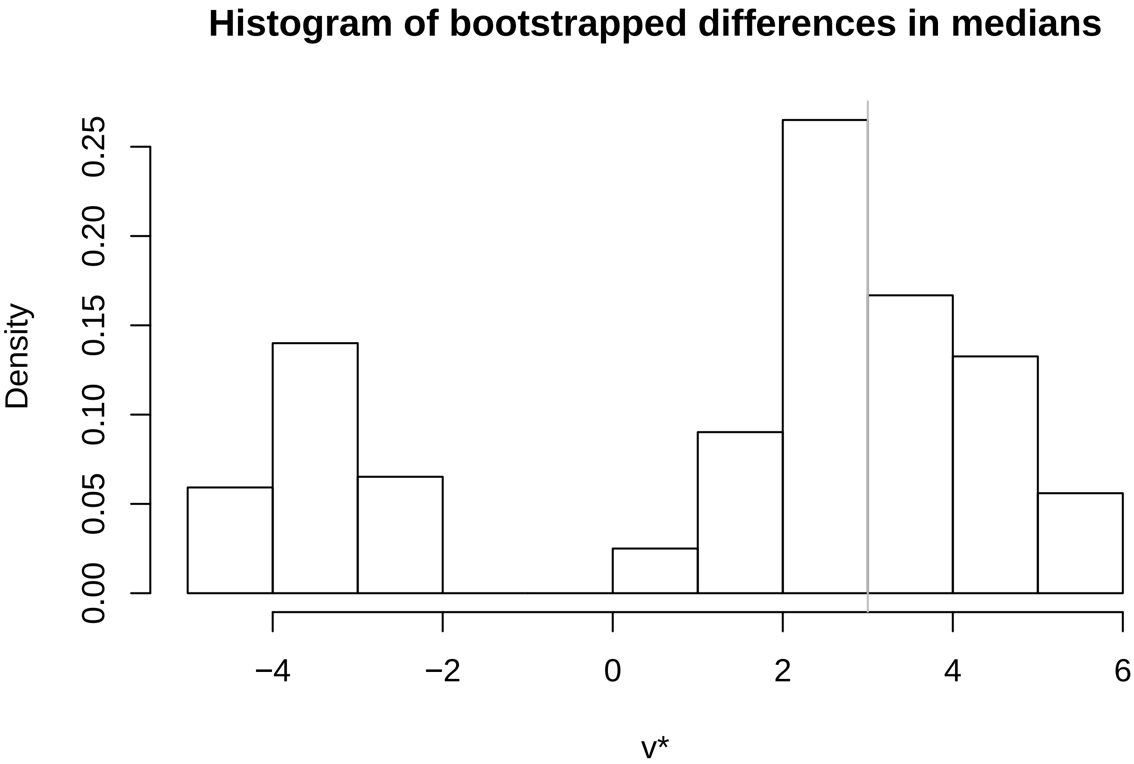
```
bc.AB <- bootstrap.comparison(samples.from.A, samples.from.B, fn = median)
bc.AB$uncertainty
```

[8] And, even if we could, we'd certainly be tempted to combine the data sets to improve our estimate of the difference, but then we're right back where we started when it comes to uncertainty.

[9] In some languages, I'd have to define `resample()` earlier in my code than `bootstrap.comparison()`. R doesn't care what order you define functions in, so long as everything's been defined by the time you *run* a function.

```
## point.estimate        std.err  lower.ci.2.5% upper.ci.97.5%
##       3.000000       3.114717      -4.000000       5.500000
```

We can also plot the histogram of the bootstrap draws, getting something like this:



(Here I have added a vertical grey line at the observed difference v_{obs}.)

What we learn from this is that there is a *lot* of uncertainty about the difference in medians. If we insist on 95% confidence, all we can say is that it's somewhere between NA and NA, which is not very precise. When the permutation test didn't reject the hypothesis of no difference, we couldn't really tell if that was because we had enough information to measure the difference, and could tell that it was small, or if on the contrary we had so little information we can't say squat about the difference. The bootstrap is telling us that we're in the we-have-no-idea situation. With only seven data points, this shouldn't be a surprise!

Why does the bootstrap work?

This section is a little more theoretical, but still hand-wavy.

Let's imagine re-sampling from *one* sample first. We've got a collection of data points $x_1, \ldots x_n$. When we re-sample, we chose each of these x_is with probability $1/n$. This probability distribution is called the **empirical distribution** of the sample. The CDF of the empirical distribution, the **empirical CDF**, is written $\hat{F}_n$. A very important result in statistics[10] tells us that the empirical distribution approaches the true distribution as the sample grows. Specifically, if we write F for the true CDF,

$$\max_x |\hat{F}_n(x) - F(x)| \rightarrow 0 \tag{5}$$

This suggests that sampling from the empirical distribution should come closer and closer to sampling from the true distribution. In particular, if we're interested in quantities which change smoothly in response to

[10] The **Glivenko-Cantelli theorem**. Pitman (1979) calls it "the fundamental theorem of statistics".

small changes in the distribution, like means, medians, means of logs, etc.[11], then sampling from the empirical distribution should be an increasingly good approximation to sampling from the true distribution.

In our setting, of caring out comparisons, we're applying re-sampling from both group A and group B, so we're using the empirical distribution of group A to approximate its true distribution, and like for group B.

Alternatives to re-sampling

The argument in the last few paragraphs suggests that re-sampling isn't the only way to bootstrap. Instead of sampling from the empirical distribution, we could sample from *any* distribution which was a good approximation to the true distribution. The empirical distribution converges on the truth (under a very wide range of circumstances), and it's computationally easy to sample from. But we could also use more elaborate models to estimate the true distribution, and, if we can sample from them, we could use them to bootstrap. This is the **model-based bootstrap**, as opposed to the **re-sampling bootstrap** I've gone over above[12].

The advantage of the model-based bootstrap is (usually) that *if* we've chosen the right model, it gives us a better approximation to the true distribution than does re-sampling. (The empirical distribution converges on the truth, but sometimes it converges very slowly.) The disadvantage is that if our model is wrong, no amount of data will correct the mistaken modeling assumptions, and we'll make systematic errors.

The name "bootstrap"

"Pulling yourself up by your bootstraps" is a proverbially impossible feat; you can't do it[13]. We're using a single sample to say what the sampling distribution looks like. This *sounds* similarly impossible. That is why the inventor of the technique, Bradley Efron, chosen it, as a kind of self-deprecating joke (Efron 1979).

On confidence intervals

A confidence interval offer a probabilistic guarantee about the parameter we're estimating. When we build a $1 - \alpha$ interval, one of three things must be true:

1. The true value of the parameter is in the interval, *or*
2. We're very unlucky, and something whose probability[14] is $\leq \alpha$ happened, *or*
3. The model we're using to calculate probabilities is wrong.

When we lean on a confidence interval to support an argument about what the parameter is, we're putting a lot of reliance on the 2nd item, that we'd have to be really unlucky to be wrong. This shines an uncomfortable light on the conventional use of 95% confidence intervals, since that gives us an error rate of 5%. In more human terms, that's one working day every month. This might make you prefer a higher confidence level and a lower error rate, but that comes at a cost, namely a wider confidence interval (unless we get more data).

[11] These quantities are known as "smooth functionals". ("Functional" is an old term of a function of a function, in this case a function of the distribution function.) The classic example of a quantity which is *not* a smooth functional is the range of a distribution. If you have a distribution where *most* of the probability is on the interval $[-1, 1]$, but there's *some* probability, no matter how small, of being at either -10^{26} or 10^{26}, then the range is $[-10^{26}, 10^{26}]$. The bootstrap does poorly at handling non-smooth functionals. More precise statements involve working out derivatives and calculus for functionals; that's beyond the scope of this class, but see Davison and Hinkley (1997) if you're interested.

[12] Some people call model-based and re-sampling bootstraps "the parametric bootstrap" and "the nonparametric bootstrap", respectively. I avoid these phrase, for a number of reasons. The biggest is that I think "model-based" and "re-sampling" are more descriptive and transparent. The other is that there are such things as nonparametric models, which sometimes get used to bootstrap.

[13] It appears to derive from an 18th century satirical novel, *The Adventures of Baron Munchhausen*, in which it is just one of the absurd and impossible things the title character boasts of doing.

[14] More precisely, something whose probability is $\leq \alpha$ no matter what the true value of the parameter is.

Confidence intervals and hypothesis tests

Suppose we have a way to test whether or not θ takes on a specific value, say θ_0, and that this test gives us a p-value, say $p(\theta_0)$ for that particular value. One way to build a $1-\alpha$ confidence interval[15] is to collect the parameter values we can't reject at level α:

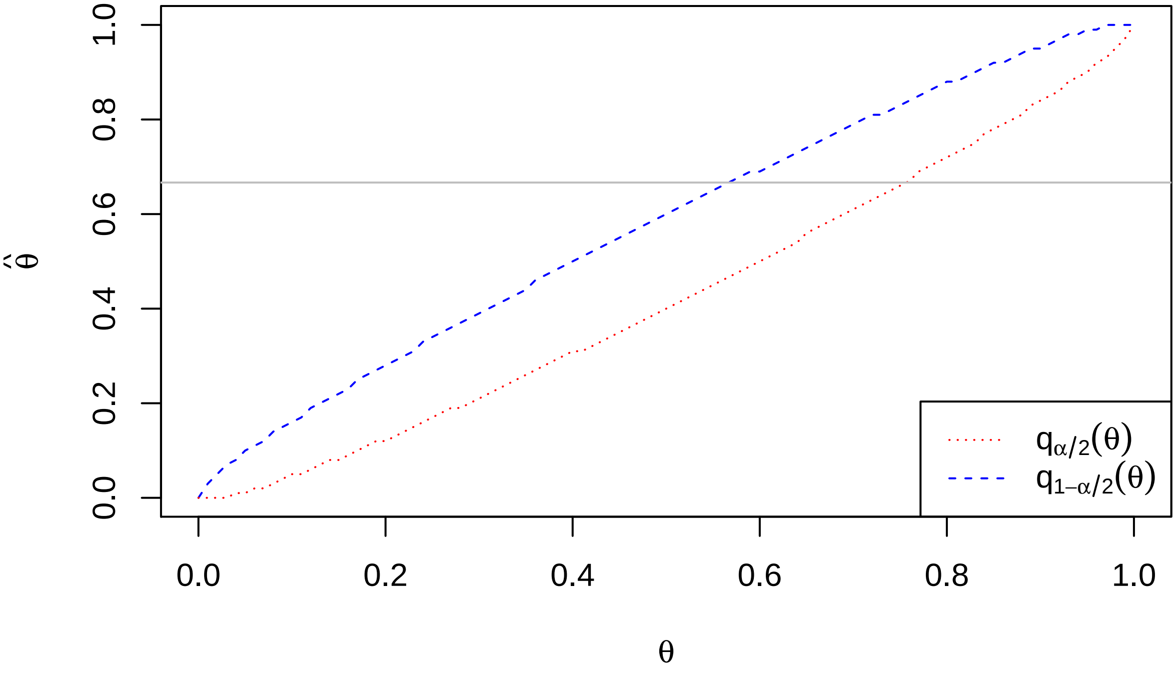
$$\{\theta : p(\theta_0) > \alpha\} \tag{6}$$

We call this **inverting the test**. In fact, *every* confidence interval corresponds to inverting *some* hypothesis test, even if we don't seem to be testing a hypothesis in constructing the interval.

One way to test whether θ takes a specific value, θ_0, is to see whether θ_0 is included in a $1-\alpha$ confidence interval for θ. If so, the hypothesis $\theta = \theta_0$ passes a test of "size" (false positive rate) α. If we want a p-value, we keep adjusting α until θ_0 is just at the edge of the interval.

A little bit more on sampling distributions, hypothesis tests and confidence intervals

Suppose we're trying to estimate some parameter or function of the data-generating distribution θ. (For us, this is the difference in typical values between groups.) Each estimator, say $\hat{\theta}$, will have its own sampling distribution, with a pdf $f(\hat{\theta};\theta)$. This notation indicates that the sampling distribution of $\hat{\theta}$ varies with the true parameter value θ. (It will also vary with the sample size, but I'm leaving that out for right now[16].) This distribution will have some quantiles, and in particular there will be quantiles $q_{\alpha/2}(\theta)$ and $q_{1-\alpha/2}(\theta)$. (Again, notice that these will vary with θ.) The next figure shows a cartoon of what this might look like.



We could use this to test whether θ equals any particular value θ_0. Find $q_{\alpha/2}(\theta_0)$ and $q_{1-\alpha/2}(\theta_0)$. Then collect your data and calculate your estimate $\hat{\theta}$. Reject $\theta = \theta_0$ if, and only if, $\hat{\theta}$ is outside the interval $[q_{\alpha/2}(\theta_0), q_{1-\alpha/2}(\theta_0)]$. Observe that the false-rejection rate (or **size**) of this test is, by construction, exactly α.

[15] Strictly speaking, the next equation defines a set which may or may not be an interval — but for most tests it *is* an interval.

[16] The sampling distribution might also involve other parameters besides θ, which, in this context, would be called **nuisance parameters**. (If we're trying to estimate the population mean, the population variance is a nuisance parameter.) There are various ways of handling this, including transforming the data so the nuisance parameter doesn't matter, or doesn't matter so much. (This is what "studentizing" the data does for the mean.) I'm not going over these complications here because this isn't a theory-of-statistics course.

The corresponding confidence interval consists of all the values of θ which this test does *not* reject. Thus if the $\hat{\theta}$ we estimate from the data is given by the grey horizontal line in the figure, the confidence interval will run from the point where it the line meets the upper blue dashed curve to where it meets the lower red dotted curve. Now one of three things has to be true:

1. The true parameter θ_0 is in that interval; *or*
2. We were very unlucky and event of probability $\leq \alpha$ happened; *or*
3. The model we used to calculate the sampling distribution is wrong.

Of course doing calculations and making plots like this above can get very tedious, and it'd be nicer if there was a way of *directly* calculating the confidence interval without going through it. But any such procedure is logically equivalent to what I've just sketched.

So why do we just read off the quantiles of the bootstrap distribution?

When we resample, we're getting an approximation to the sampling distribution at $\theta = \hat{\theta}$. You might worry that we need to somehow do a modified bootstrap where we impose different values of θ and find the sampling interval for each one, or something complicated like that. I have instead told you to *just* use the quantiles of the bootstrap distribution. This is what's sometimes called a **bootstrap percentile interval**. It is in fact somewhat wrong — the probability that the CI contains, or **covers**, the true parameter isn't, in fact, exactly $1 - \alpha$. The **coverage** does approach the stated (or **nominal**) level as the number of data points grows, but not as quickly as we might want. Statisticians have put a lot of effort over the years into creating modifications of the percentile interval with more accurate coverage, and while this is an important topic, it's a bit beyond the scope of this class. Davison and Hinkley (1997) gives a thorough review of these methods and the associated theory, while Hesterberg (2014) is more tutorial, and focuses explicitly on comparing group means.

Asymptotic approximations

If we're interested in the difference between population means, and we use the sample means of the two groups to estimate that difference, and neither group has a distribution that's too heavy tailed, then

$$V \rightsquigarrow \mathcal{N}(\mathbb{E}\left(X\right) - \mathbb{E}\left(Y\right), \mathrm{Var}\left(X\right)/n_A + \mathrm{Var}\left(Y\right)/n_B) \tag{7}$$

A large-sample approximate confidence interval for the difference in population means is therefore

$$\left[v_{obs} - z_{\alpha/2}\sqrt{\hat{\sigma}^2_X/n_A + \hat{\sigma}^2_Y/n_B}, v_{obs} + z_{1-\alpha/2}\sqrt{\hat{\sigma}^2_X/n_A + \hat{\sigma}^2_Y/n_B}\right] \tag{8}$$

where z_p is the p quantile of the standard Gaussian distribution. Other ways of measuring typical values lead to other, similar central limit theorems, usually with even more complicated expressions for the variance. You should think of these expressions as shot-cuts which let you avoid having to use the bootstrap, *when the assumptions hold.* Hesterberg (2014) has some detailed numerical examples which show that when the data distribution is highly skewed, these Gaussian approximations don't become accurate until both n_A and n_B are very, very large. And we've seen, at length, that variables link income and wealth are usually very, very skewed. (See Complementary Problem 3.)

Complementary Problems

As usual, these are optional problems to think through or practice on, not to turn in.

0. Throughout, I've said that we're interested in $v = m_A - m_B$, the difference in typical values. What, if anything, would have to change if we were interested in the ratio m_A/m_B?

1. Suppose that we consider larger and larger samples from our two groups, but the two samples are always in the sample proportion to each other, so that $n_A = p(n_A + n_B)$ even as $n_A + n_B = n$ grows. Use Stirling's formula[17] to approximate the number of permutations in terms of n and p.
2. Find and read Arthur C. Clarke's short story "The Nine Billion Names of God" (from 1953; dated language is dated). Explain its relevance to permutation testing and bootstrapping.
3. *Convergence of sample means to Gaussians* Write code to draw a sample of 10 values from a standard log-normal distribution and calculate the sample mean. Run this for 1000 replicates and verify that the sample mean does *not* have a Gaussian distribution (e.g., using a Q-Q plot). Now increase the sample size to 30; you should see that it's closer to Gaussian but not a lot. How big to you have to make the sample size to get close to Gaussian?

References

Davison, A. C., and D. V. Hinkley. 1997. *Bootstrap Methods and Their Applications.* Cambridge, England: Cambridge University Press. https://doi.org/10.1017/CBO9780511802843.

Efron, Bradley. 1979. "Bootstrap Methods: Another Look at the Jackknife." *Annals of Statistics* 7:1–26. https://doi.org/10.1214/aos/1176344552.

Hesterberg, Tim. 2014. "What Teachers Should Know About the Bootstrap: Resampling in the Undergraduate Statistics Curriculum." E-print, arxiv:1411.5279. https://arxiv.org/abs/1411.5279.

Mayo, Deborah G. 1996. *Error and the Growth of Experimental Knowledge.* Chicago: University of Chicago Press.

Mayo, Deborah G., and D. R. Cox. 2006. "Frequentist Statistics as a Theory of Inductive Inference." In *Optimality: The Second Erich L. Lehmann Symposium*, edited by Javier Rojo, 77–97. Bethesda, Maryland: Institute of Mathematical Statistics. http://arxiv.org/abs/math.ST/0610846.

Pitman, E. J. G. 1979. *Some Basic Theory for Statistical Inference.* London: Chapman; Hall.

[17] $\log{(n!)} \approx n \log{n}$. (There are also small-order terms, but they don't matter for this problem.)