

# Resampling: Because P211 is difficult enough without having to learn a bunch of statistics.

R. A. Hullinger  
Indiana University  
P211 - Methods of Experimental Psychology

P211 is not a statistics course. Therefore, we are not going to ask you to learn about various statistical formulas or how to use those mysterious tables in the back of your statistics book. We will not even ask you to memorize much terminology. There will be plenty of time for that in stats classes and in graduate school.

We do, however, need a way to determine if the results of our experiments are “statistically valid.” That is, we need to prepare you to be able to figure out if the difference that you measure between your experimental and control groups is likely to be the result of our experimental manipulation or if the difference could be due to chance. “Resampling” is the technique we will use to accomplish this task.

## What is Resampling and Why Should You Care?

Resampling is a powerful yet simple statistical technique that can be used to generate inferential statistics<sup>1</sup> and test hypotheses when you only have a small amount of data collected and/or the data you have are not suitable for “standard” statistical methods. In our particular case, resampling is handy for at least five reasons:

- **Less Data.** As mentioned above, we won’t need to collect tons of data before we can perform a reasonable analysis. This makes resampling very valuable in situations where it is either expensive, time consuming, or otherwise difficult to collect large quantities of data.
- **Fewer Assumptions.** Traditional statistical tests rely on several assumptions about the data that are collected. If the data violate those assumptions, the traditional tests don’t work well. Resampling will often work even if those assumptions do not hold, so we don’t have to concern ourselves with any of those potential problems.

---

<sup>1</sup>Inferential statistics are statistics about an entire population that are *inferred* from the data collected from a small section of that population. For example if you wanted to estimate the average age of all college freshmen (a population) using the data collected from 100 randomly sampled freshmen at IU (a section of the population), that would be an inferential statistic.

---

For her great work in designing and coding the software described in this document, the author thanks Melissa Troyer. For her great patience and stamina, her crazy grammar skills and her helpful comments while proofreading all 28 drafts of this paper, the author also expresses profound thanks to The Amazing, Charming, Beautiful and Incomparable Reviewer who wishes to remain anonymous.

- It's Intuitive. If we explain it well, resampling should be conceptually easier to understand than traditional statistical methods. If we explain it *really* well, resampling will make understanding those traditional methods much easier when you are exposed to them.
- Fancy Software. We have a really cool piece of easy-to-use software that does the resampling calculations for us and even draws helpful (and animated!) graphs.
- It's what we're going to use in P211. We're convinced that resampling is the way to go, so it's the only method we're going to teach you and it's the method we're going to require you to use.

Having been won over by our cogent arguments, you are now undoubtedly quite excited about this new technique and you are thinking to yourself, "I really wish I knew how resampling worked." Fortunately, we have anticipated your enthusiasm.

## How Resampling Works

The basic idea behind resampling is not complex. However, it can be confusing when you are first exposed to it. As you work your way through this section, you may find yourself wondering why we're doing some of the things we're doing. Don't worry – it should all make sense soon enough.

First, we need to be clear about exactly what sort of data we have and exactly what sort of information we need to extract from those data.

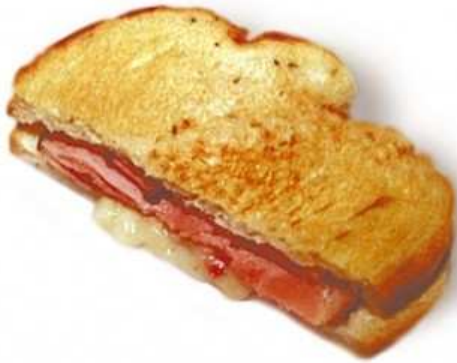
***What do we have?*** After you run your experiment, you will have data from a "representative sample" of some larger population. For the purposes of this guide, we will assume that you have performed a simple within-subjects experiment and you have one set of data values from the control condition and one set of data values from the experimental condition. We will address the differences for other types of experimental designs after we have explained the fundamentals of resampling.

One of the major stumbling blocks with learning statistics is that it often feels very abstract. To avoid that problem, we will illustrate what we're doing with a simple and somewhat far-fetched example. We will test the hypothesis that eating a Dagwood's sandwich instantly increases your IQ.<sup>2</sup> To test this hypothesis we will randomly select 6 IU students to participate in our within-subjects experiment. In the control condition, each participant would be asked to come to the lab before lunch to take a short test to measure his/her IQ. Suppose that in this condition we measure the following IQs: [117, 98, 104, 131, 96, 114], giving an average IQ value of 110. After taking the IQ test, each participant would be given a fresh, delicious Dagwood's sandwich to eat. Finally, each one would be asked to take the IQ test once again (the experimental condition). For the purposes of this example, we will suppose that after eating, the IQs are as follows: [125, 103, 101, 131, 110, 120], giving an average IQ score of 115. On the surface, at least, it looks like eating a Dagwood's sandwich increases IQs by an average of 5 points. (It is worth noting that this simple experiment has a confound. It's possible that most people who take the test a second time do better simply because they are familiar with the test, with or without a sandwich. Since this primer is focused on statistical analysis and not experimental design, we will ignore this explanation. However, if we were really running this experiment, we would want to use a more advanced design to avoid this potential confound.)

---

<sup>2</sup>Sadly, the author is living proof that this hypothesis is deeply flawed.

**What do we want?** What we *really* want is world peace or, failing that, at least one really cool super power, but that's neither here nor there. Statistically what we want is the ability to know if the difference in the data that we collected from our two conditions can be trusted. How sure can we be that our results are due to the experimental manipulation? How likely is it that if we ran the experiment again we would generate results that still support our hypothesis? If we did run the experiment again and got different results, would we have to throw out the hypothesis? We can find these answers with resampling.



*Figure 1.* An old ham sandwich. This is a totally unnecessary figure, but this primer is long and boring and it seems easier to read if it has more pictures in it.

You may have noticed that in the paragraph above we didn't ask simple yes/no questions like "Are we sure that...?" or "Will we get the same results?" That is because we can never be completely certain about results. The very best that we can do is calculate just how likely (60% likely? 85% likely? 99.99% likely?) it is that the observed experimental difference is due to a specific difference between the two conditions – presumably a difference that is a result of our experimental manipulation.

How could results not be due to a specific difference between the two conditions? Well, unfortunately the world is a messy place. Imagine that we gave a group of people an IQ test, then *didn't* give them a Dagwood's sub, and then gave them the IQ test again. It is reasonable to think that not everyone would get exactly the same score on the IQ test the second time. Therefore, even though there wasn't any real difference between the two conditions, we might still see a change in the average IQ scores. How can we distinguish this scenario from the one where the Dagwood's sub caused an increase in IQ?

This, then, is what we want, The Holy Grail of P211 statistics: a method to analyze the data we've collected that will tell us exactly how much we can trust that the difference between the two conditions was a result of our experimental manipulation.

#### ***How can we figure this out?***

If our experimental result was just a fluke, we can reasonably expect to get very different results if we run the experiment again. However, if our hypothesis is correct and the result was due to our experimental manipulation, we should be able to generate reasonably consistent results each time we run the experiment. Therefore, the more consistent our experimental results, the less likely we are to accept dumb luck as an explanation for our data.

Based on this thought process, one very straightforward way to gain trust in our results is to run the experiment again...and again...and again. If we could run the experiment 100 times, and on 96 of those runs we collected data that showed eating a Dagwood's sub leads to an increase in IQ scores, you would be pretty convinced that not only are Dagwood's subs nature's perfect food, but also that they increase intelligence. Unfortunately, you'd also

think that we're nuts for running so many experiments. You'd be right. It is prohibitively expensive and time consuming to run even a simple experiment multiple times, so we need to find another way.

Before we dive into resampling, we need to start with one very important assumption. We will assume that we did a great job of randomly sampling the population when we recruited participants for our experiment. Accepting this assumption means that the variation in our sample of participants is an accurate representation of the variation in the general population. With this assumption in place, we are ready to proceed.

Within-subjects resampling allows us to use the same basic idea of running the experiment multiple times, but it only requires us to run the experiment once. Here's how it works: First, we calculate the difference between each participant's scores for the two conditions to create a list of 6 difference values. For example, the first participant's original IQ score was 117 and his IQ score after eating the sub was 125, a difference of 8 points. We do this for each participant and generate the list of differences: [8, 5, -3, 0, 14, 6].

Next, we run a theoretical duplication of the experiment. Instead of going out and recruiting 6 new individuals to test, we will select 6 individuals from the sample we already have. "Whoa there, hold on one minute!" In the unlikely event that this document has not already lulled you into a stupor, alarm bells should be going off in your head. "How the heck can we re-run the experiment using the data that we've already collected?!" That has to be some sort of cheating, right? No. This is where our initial assumption comes in handy. Remember we stated that we did a great job of randomly sampling the population we're studying. That means that whatever interesting or relevant features are possessed by "Participant 1" are also possessed by many other members of the population that we didn't sample. Similarly, "Participant 2" is a perfect proxy for many other potential participants, too, and so on for all of the participants. So when we use Participant 1's data in our theoretical duplication, don't think of it as using the same person's data again. Think of it as running the experiment again after collecting data from a new participant who happens to be just like Participant 1 in all experimentally-relevant ways.

That idea is bound to take some getting used to. Even after we've explained it, you may think it's a little fishy, but we promise, it's not. What you should realize, though, is that the accuracy of your resampling result is tightly coupled to having an accurate, unbiased, and representative sample. Therefore you need to be thoughtful about how you select your participants.

We shall proceed by selecting six individuals from the existing participants. This, too, might seem problematic. If we only have 6 difference values to start with and we randomly pick 6 scores from that list, we'll always get the same result, right? No matter what order we pick the values from [8, 5, -3, 0, 14, 6], if we pick all 6 of them and then calculate the average value it's always going to be 5. To get around this we pick the scores "with replacement" which means that after we pick a value, we put it back into the group (replace it) so that it could be picked again. This technique means that given our 6 difference scores, when we generate the "new" data over and over again in the resampling process, we might get a list of six scores like this [5, 14, 0, 8, 14, 6]. In this case we randomly selected 14 twice and never selected -3<sup>3</sup>. We then calculate the average of those 6 values (about 8), and store

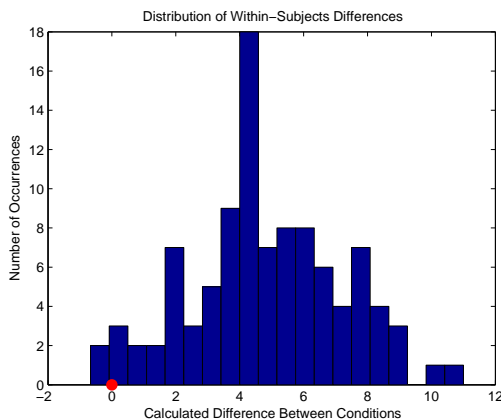
---

<sup>3</sup>Again, this process relies on the assumption that our samples are representative – if we wind up using Participant 4's data twice in one iteration of resampling, it's equivalent to running the experiment after

that.

So far, so good. We’ve essentially run the experiment a second time and gotten results that support our hypothesis (an average IQ increase of 8 points after eating the sub). However, two experiments aren’t much more convincing than just one, so we continue this process for many more iterations, and store the average increase (or decrease) in IQ after each iteration. Suppose we repeated the process of creating these “theoretical experiments” 100 times and then created a histogram of the average change in IQ scores across the experiments. What would that look like?

On many of our experimental iterations when we randomly sampled values from our original data, we would get a few of the values that are larger than 5 and a few of the values that are smaller than 5. Usually, then, we would expect to generate an average difference somewhere around 5. It would be possible, but less likely, to sample only 5s, -3s and 0s from the data, generating a much lower average difference, or to only sample 8s and 14s from the data, giving a very high average difference. It would be even less likely, but possible, to sample data that contained nothing but -3s and 0s, yielding a negative difference between the two conditions. Thinking through the process like this, you should be able to see that the histogram created from the 100 iterations should look something like the one pictured in Figure 2. On the majority of the iterations, the average difference will be around 5, the mean value from the actual experiment.



*Figure 2.* Example histogram showing the distribution of 100 difference calculations after performing 100 within-subjects calculations.

get more extreme. So what? Where does that leave us? That’s a lot of work for a picture, and to be honest, it’s not even a cool picture. In fact it almost looks like the statistics are flipping us off. Stupid statistics. The good news is that the picture, as unexciting as it may be, answers our question. You do remember our question right? Several pages (and perhaps

As expected, though, occasionally the values were randomly selected in a way that gave much higher (or lower) difference scores, creating a few data points on the histogram at more extreme values like 10 or -1. In statistics, the shape of this histogram is called a “normal distribution” or a “bell curve.” As we move away from the center of the distribution it becomes less and less likely that we will encounter any particular value.<sup>4</sup>

OK, we took our experimental data and used it to simulate running the experiment over and over again. This bizarre process allowed us to build a histogram of differences showing a normal distribution of values centered around our experimental mean and falling off quickly as the values

selecting two individuals similar to Participant 4.

<sup>4</sup>Figure 2 doesn’t look like a nice perfect normal distribution because of the random nature of the process used to create it, but as more and more samples are added, say, after 10,000 trials, the histogram would begin to look very much like a bell-shaped normal distribution.

a nap or two) ago, right before things got really messed up and confusing, we asked, “How likely is it that if we ran the experiment again we would generate results that still support our hypothesis?”

The histogram is telling us how likely it is that any particular IQ difference between the two conditions would be seen in our population as we continued to run the experiment. Our hypothesis was that Dagwood’s subs increase IQ. Therefore all of the bars on the histogram that represent results where the IQ difference in the two conditions was greater than zero show us the count of simulated experiments where the data confirmed our hypothesis. Conversely, the portion of the histogram that represents difference values that are less than zero show the iterations where the data did not support our hypothesis. Therefore, we can use the histogram to determine what percentage of the time we would expect to get results that confirm our hypothesis.

Looking at Figure 2 we can see that on 98 of the 100 runs, we collected data that showed a positive increase in IQ after eating a sub. That means that we only generated data that didn’t support the hypothesis 2% of the time. Pretty convincing, and we didn’t have to administer the experiment over and over again.

At a base level, that is resampling. The good news is that aside from collecting the original data and entering it into the computer, the rest of the process is automated by the really cool software we mentioned at the start. You don’t have to do all the sampling and calculating by hand. Actually, that’s *really* good news — we repeated that process 100 times in the example, but when we use resampling to analyze data, we repeat the process 10,000 or more times to ensure accuracy.

### A Very Important Statistical Aside

We know that we promised not to force you to memorize statistical jargon or dive too deeply into mathematical stuff, but there are two key pieces of statistical information that you need to understand. First, we have to clarify one bit of terminology. So far, we’ve been talking about the “percent chance that additional results would confirm your hypothesis.” Because statisticians are curious creatures, they don’t like to report the percent chance that the results *would* confirm the hypothesis, but prefer to talk about the percent chance that the results *would not* support the hypothesis<sup>5</sup>. So instead of saying that there’s a 98% chance that the results would confirm the hypothesis, they report the 2% chance that the results would not provide support.

In statistics, this “probability that future results would not provide support” is reported using a “p-value”. In our example, there was about a 2% chance that we could run this experiment and get results that did not support our hypothesis. So when reporting the statistics, we would say: “After eating a Dagwood’s sub, IQ measures increased by an average of 5 points,  $p < .02$ .” You may have seen this notation in articles or textbooks before, and now you know what it means:  $p < 0.x$  means that there is less than an  $x\%$  probability of generating results that do not support the hypothesis<sup>6</sup>. This interpretation

<sup>5</sup>This is not really because statisticians are curious creatures, but rather, it is based on the philosophy behind this style of statistical testing. If you are interested in such things, you can search for “Null hypothesis significance testing” to learn far more about this topic

<sup>6</sup>This is the interpretation of the p-value for our within-subjects resampling analysis. The interpretation is slightly different for the between-subjects case, as is explained below

also means that the *smaller* the p-value, the *better*, which may not be intuitive. However, as long as you remember that the p-value indicates how likely it is that you could run the experiment and get results that *do not* necessarily support the hypothesis, then it should make sense that as the p-value gets smaller, the support for the hypothesis gets larger.

Second, we need to be clear that the p-value is the primary measure of an experiment's results. It is very possible to run an experiment and record what appears to be a very large numerical difference between two conditions, but statistical analysis may reveal a large p-value indicating that the results are not that strong. Similarly, a very small numerical difference could still yield a small p-value, indicating a strong, statistically significant effect. The explanation of why this is true has to do with certain mathematical properties of your data and is beyond the scope of this primer. What is critical to understand is that a large numerical difference between conditions isn't always good and a small difference doesn't always mean the experiment failed, so you *must* look at the p-value to determine whether or not there was an important effect.

Those two concepts are so important, not only to resampling, but to statistical analysis in general, that they are worth restating:

- The p-value provides a measure of how much you can trust the experimental results, and the *smaller* the p-value, the *stronger* the evidence that the experimental manipulation worked.
- You cannot tell if an experiment “worked” by looking at the size of the difference between your conditions. You *must* use the p-value in order to know if the experiment provided support for the hypothesis in question. A corollary to this is that you *must* report the p-value in your experiments so that others can tell how much faith to put in your results.

Finally, we need to take a moment to think about how to interpret experimental results. As the p-value gets smaller, your belief that an experiment “worked” increases. At some point – and that point is often arbitrarily set to “ $p < 0.05$ ” – psychologists begin to say that the results are “statistically significant” which is our way of saying “we’re pretty sure we’re right.” However, if the p-value is greater than 0.05, we say that the result was “not significant,” but that is not the same as saying “our experiment did not work.” Rather, all “not significant” means is that we cannot rule out the possibility that the difference between the two conditions was due to chance.

### Resampling In a Nutshell

Now that you've made it through the excruciating details, perhaps a quick overview will help solidify the core concepts in your mind. First, you do an excellent job of randomly selecting participants from the population. Next, you run your experiment. Then, enter your data into the resampling applet and let the software take over. The applet executes this process:

1. Use the existing data to run a theoretical duplication of the experiment you performed.
2. Calculate the mean difference between the two conditions in the theoretical experiment.

3. Store that difference for later use.
4. Repeat thousands of times.

After generating thousands of theoretical experimental duplications the software will plot a histogram to show you the expected distribution of results. Finally, based on that histogram, the applet will report what percentage of the time the results did not support the hypothesis you are testing. The lower this percentage, the more confident you can be that your experiment worked.

### A Few More Mostly, Somewhat or at Least Occasionally Important Details

At the beginning of this document we made a decision to simplify the explanation of resampling by sticking with a within-subjects experimental design. The same basic mathematical technique can be used to analyze data from a between-subjects experiment or to check the significance of correlations. To do this, though, the process has to undergo slight modifications. Please only read this section if the basic resampling process is clear to you. If you're still confused, even slightly, stop now. That's right, we're giving you permission to stop reading. If the within-subjects explanation isn't crystal clear, the next sections are only going to confuse you even more. It makes us very sad, but it's true.

#### *Resampling for between-subjects designs*

OK, but don't say we didn't warn you, smarty-pants. When you are working with a between-subjects design, you do not have pairs of scores from each participant. Instead you have one set of scores from a control group and a separate set of scores from a distinct experimental group. In this scenario we will attack the problem in a slightly different way.

First, let's recast the original experiment as a between-subjects design. In this case, we recruit 12 participants. We give 6 participants the IQ test without giving them any food, and record scores of 117, 98, 104, 131, 96, and 114. We give the other 6 participants a fresh and tasty Dagwood's sub, followed by the IQ test, and we record scores of 125, 103, 101, 131, 110, and 120. Just like the within-subjects data, the participants in the Dagwood's condition have, on average, a 5 point higher IQ score.

We will start with the same initial assumption that we have a good random and representative sample of the population in our two groups. We will also start with a second, seemingly odd assumption. We will assume that our experimental manipulation didn't have any effect on our participants<sup>7</sup>. That seems like a messed-up place to start, but it's similar to how we handled the within-subjects analysis. When we worked through the within-subjects process, we finished by calculating how likely we were to get results that did not confirm our hypothesis, and the smaller that number (the p-value), the more trust we had in our results. For between-subjects analysis we calculate how likely it is that our results could come about due to nothing more than chance. If it turns out that our results are very unlikely by chance alone, then we know that it's very likely that our experimental manipulation really did work.

What we're about to do next is going to seem crazy at first, but trust us, it will all work out in the end. We're going to take the data that we collected from everyone in

---

<sup>7</sup>In statistics terminology, this is known as the "Null hypothesis."

the control group and the data we collected from everyone in the experimental group and we're going to put it all into one big pile without any regard for which group it came from. "But," you're saying, "We can't just mix it up like that! Some of that data came from the experimental condition!" You're right, it did. The important realization is that we are working from the assumption that the experimental manipulation had no effect. If it had no effect, then the control group should be just like the experimental group, right? Therefore we can just pile it all together: [120, 117, 103, 98, 104, 131, 101, 96, 114, 125, 131, 110].

Next we're going to randomly draw out as many samples as we had in the experiment's control condition, and put those samples into a new "control" group (since there were 6 participants in the control group, we would draw out 6 samples from the pile). We then take the rest of the samples from the pile and put them into a new "experimental" group. "But there was no experiment!" Right again. We warned you this was going to sound crazy. If the experimental manipulation did not matter, then it wouldn't make any difference if someone was in the control group or the experimental group, so it's safe to (randomly) assign them:

"Control": [120, 131, 101, 114, 96, 117]

"Experimental": [103, 131, 98, 125, 104, 110]

Now that we have new "control" and "experimental" groups, we can do the same calculations that we did in the actual experiment, and we can record the difference between the two groups, which is about -1 point (indicating that the control group now has the slightly higher average IQ). Once we've recorded the difference, we can throw all the data back into one big pile again, shuffle it up, and randomly reassign 6 values to the "control" and "experimental" groups. Once again, we can calculate the difference between the two groups, and record that difference. Next we throw all the data back into one big group, shuffle it ... and so on.

Suppose, then, that we do this over and over and over again, randomly placing some data in the "control" group and the rest of the data in the "experimental" group each time, and dutifully recording the difference between these two groups. If we looked at a histogram showing the distribution of those calculated differences after 100 repetitions of this process, just like we did in the within-subjects analysis, what would the histogram look like?

Recall that in the actual experiment there was a 5 point difference between the two groups. However, if we put all the data together into one big group and then randomly reassign the values to new groups as described above, we would expect each of the new groups to have a few of the high values and a few of the low values, as they did in the first example. Therefore we expect that the difference between the two groups

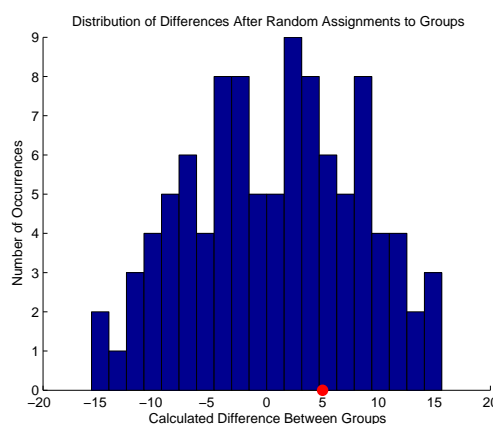


Figure 3. Example histogram showing the distribution of 100 difference calculations after randomly dividing a group of participants into two groups.

will often be pretty close to zero. Since the assignments were random, the only thing that could cause more of the high (or low) values to wind up in one group or the other is chance. Sometimes, though, when we divide up the groups, more of the high values will be placed into one group and more of the low values will be placed in the other group. In that situation there would be a large difference between the two groups, but cases like that will be more rare than the cases where the two groups have about equal averages.

Following that logic, the histogram of differences should look something like Figure 3. This figure shows that on the majority of the trials, the difference between the two groups was around zero. As expected, though, occasionally the values were randomly assigned in a way that gave one group much higher (or lower) scores than the other, so there are a few bars on the histogram that represent large differences between the groups. As in the within-subjects design, the histogram shows a more or less normal distribution, but unlike the within-subjects analysis, this histogram is centered around zero, not around the experimental mean.

So we garbled all of our data together and then randomly divided it into two groups over and over and over again. This process allowed us to build a histogram of differences showing a normal distribution of values centered around zero and falling off quickly as the values get more extreme. As with the within-subjects analysis, this histogram can be used to figure out if we should trust our results. This time, however, we aren't looking at how many results fall on one side of zero or the other. (Since the histogram is centered on zero, we wouldn't get much from that analysis.) Instead, recall that we ran all of these theoretical experiments under the assumption that there was no difference between the two groups. So the histogram we created is telling us how likely it is that any particular IQ difference between two groups has nothing to do with Dagwood's subs, but rather is a result of chance alone. Therefore, what we care about in this case is how often did we get differences that were at least as large as the original experimental difference?

In the experiment, we measured a difference of 5 IQ points between the two groups. Looking at the histogram, we see that it is actually quite likely that chance alone would cause a difference of 5 points or more. In fact, it looks like there were about 58 times out of 100 when a value 5 points or larger was seen. For this analysis, the term "larger" should be interpreted as "having a larger absolute value" so the samples larger than 5, as well as the samples that are less than -5, should all be viewed as "larger". What that means is that according to our analysis, there is a 58% chance (58 out of 100 samples) that our results could have come about by chance alone. That makes it hard to put much faith in the idea that our experimental manipulation (eating a Dagwood's sub) really did raise people's IQs. We could *not* make the claim that eating a Dagwood's sub causes a "statistically significant" increase in IQ.

In the case of between-subjects resampling, the p-value reports what we measured in the paragraph above: the probability that we would expect to see results at least as large as the experimental results if the experimental manipulation had no effect. In this example, we would report " $p < .58$ ." As in the within-subjects design, the smaller the p-value, the stronger our belief that the experiment worked. Again, in psychology, the statistical cut-off is usually set at  $p < .05$ . If there is less than a 5% chance that the results would come about due to chance alone, we claim that the experiment "worked" because we have "statistically significant" results.

If you've made it this far, it may strike you as odd that the data which resulted in statistically significant results when analyzed as a within-subjects measure gave non-significant results in the between-subjects analysis. How can the exact same data yield such different results? The answer lies in the power of within-subjects designs. In the within-subjects design, we were looking at the IQ change for each participant, so their individual differences in IQ don't affect the results. In the between-subjects design we had separate individuals in the two conditions, so it's possible that we selected people with a higher average IQ in the Dagwood's eating condition. Intuitively, then, it makes sense that we would be less convinced by the results of the between-subjects design than the within-subjects design. The techniques used to perform the resampling analysis in each design take these differences into account and present differing levels of trust in the two experiments accordingly.

### *Resampling for correlations*

When working with correlational data you are not trying to determine if there is a difference between two groups (as in the between-subjects designs) or a difference between two measurements from one group (as in the within-subjects designs). Instead, you are trying to determine if there is a systematic relationship between two measurements. For example, you may suppose that there is a relationship between a person's height and his or her shoe size. If you believe that taller people generally have larger feet, then you would say that there is a positive correlation between height and shoe size. If you believed that taller people usually have smaller feet than short people, then you would say that there is a negative correlation between height and shoe size. You would probably also say that it's really easy to tip tall people over, but that's neither here nor there.

Suppose then, that we decide to measure the correlation between height and shoe-size. First, we find a random group of participants and record each participant's height and shoe size. That gives us a set of pairs of height and shoe-size data that may look something like the data in Table 1.

If you were to perform the statistical analysis on this data, you would calculate a "correlation coefficient." The correlation coefficient is a value between -1 and 1 that indicates the strength of the relationship between the two measures. If there was no relationship between the values, then the correlation coefficient would be 0. If there was a "perfect" relationship between the two values, such that if you knew one of the measurements you could exactly predict the other measurement, then the coefficient would be 1 (if as one value increased the other value also increased) or -1 (if as one value increased the other value decreased.) For the data in Table 1, the correlation coefficient is about .87, indicating a very strong correlation between height and shoe size.

How do we know if that correlation coefficient is statistically significant? That is,

Table 1: Hypothetical Shoe Size Data

Participant	Height	Shoe Size
1	5' 10"	11
2	5' 6"	8
3	6' 0"	13
4	5' 1"	6.5
5	5' 11"	9
6	6' 3"	12

given our results from the sample participants, how likely is it that the correlation coefficient would actually be zero?

In this case we perform a similar process to the one used for within-subjects designs. We cannot separate our data into two groups, and a particular participant's height and shoe-size data have to stay together. So we create our "new" set of data by randomly picking 6 pairs of data from the original data set and we pick the data with replacement, just like the within-subjects process. We might get the data from participant 5 in the new data set 3 times, or maybe that data won't be in the new data set at all, but just like the original data, we'll have 6 pairs of measurements to work with.

Once we have our new data, we calculate the correlation coefficient for that data set, and store it off. As with the other resampling processes, we do this thousands of times, storing the correlation coefficient each time. When we are done, we can create a histogram that shows the expected distribution of coefficients from our sample data.

From this point forward, analysis of the correlation is exactly like the within-subjects design analysis. The histogram of coefficients will be centered on the coefficient that we calculated when we ran the actual experiment (in this case, about .39). Therefore, we answer the "how likely is it that the coefficient is actually zero?" question by figuring out where 0 falls on the histogram. If it is near the center of the distribution, then a coefficient of zero is likely, and our results may not be significant. If, however, zero is on one of the extreme tails of the histogram (as it would be with this example), that indicates that only a very small percentage of our simulated experiments resulted in a correlation coefficient of zero. Therefore we may be able to claim a significant result because the likelihood that the actual correlation coefficient of the population is zero is quite small.

Note that like the actual size of the numerical differences between conditions in the between- and within-subjects designs, the magnitude of the correlation coefficient is not of critical importance. It is not safe to assume that just because the coefficient you calculated is large, you must have a significant result. Similarly, just because your results show a very small correlation, that doesn't mean the results are not significant. What is important, from a significance standpoint, is the size of the p-value. If the p-value is very small, then the results are significant regardless of particular value of the correlation itself.

### The Finale

That is resampling in all its glory. While it may initially seem complex and somewhat obtuse, it really is a straightforward process. You use the data you already have to simulate performing your experiment thousands of times and based on those results, you can make an estimate about what percentage of the time you expect to get "interesting" results (e.g. a marked difference between two groups or a non-zero correlation). If the simulations show that you would expect to get "uninteresting" results less than some small percentage of the time (as noted above, psychologists often arbitrarily choose 5% as their cutoff), then you can say that you have significant results.