Statistics Notes

Contents

Chapter 1

A word about these notes

These notes are sort of for you, but mostly for me. They give me the basis for a presentation in class, and they should give you a reminder of what happened during that presentation. They aren't a textbook. When there are useful readings, I'll point you to them. In the meantime, take notes! Don't expect to rely a hundred percent on reading what I put in these course notes.

Chapter 2

Statistical reasoning

How do you know how tall you are? You measured yourself, or someone else did. How might you know how tall your neighbour is? You could ask them, and they would tell you something about what happened the last time they were measured.

How tall are we?

Hmm.

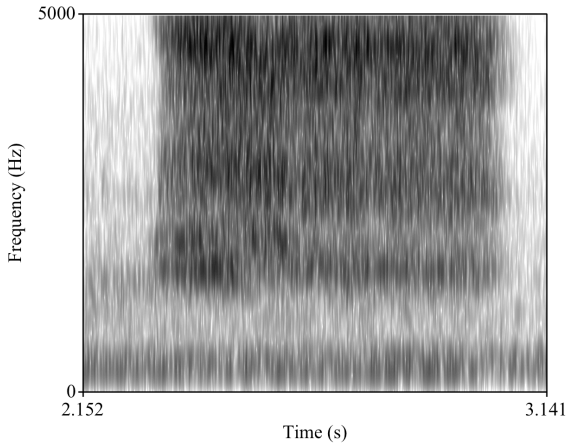
Here's an important linguistic fact: "people" (not just me) find it harder to perceive the differences between "sounds" if those sounds aren't used in their respective native language (than if they are).

Hmm.

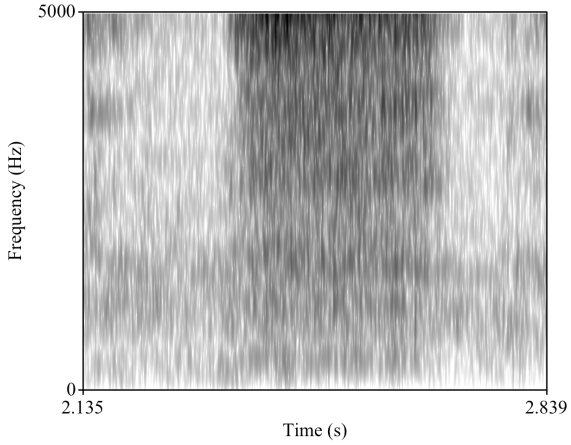
2.1 What is statistics?

Statistics is not math. And it is not about counting things. And it does not involve reasoning that requires you to be extremely clever (except for some points where even statisticians can't agree). Statistics is about reasoning. It's about reasoning through questions like the ones we just saw: we want to answer a question about **one** thing, or make a comparison between **two** things: is A bigger than B? Is A like B? But, manifestly, when we look at thing A and thing B, we find a whole set of things, slightly different from one another. Statistics is about reasoning on those days when we realize that we're not reasoning about things we can observe directly (like "people"), but, rather, that we're reasoning about **underlying patterns.**

In science (language science included), that's almost always. And it's not just because we want to know about how things work. It's also because, even for simple things that we might think we can just observe directly, there is usually some kind of "noise" to filter out. I'll give you an example. This is a spectrogram of me saying **[ʃ]**:



Now. It so happens that, for fricative sounds like **[ʃ]** and **[s]**, the **spectral centre of gravity** is an important acoustic property. If you looked at a spectrogram of a **[s]**, it would have a higher spectral centre of gravity: roughly speaking, it corresponds to the frequency at which the large black noise band that you can see has the most energy. In fact, it's not quite that. It's a mean frequency value, whereas what I just said would correspond to a modal frequency value. But modal values are easier to understand on a graph, and this isn't a phonetics course.



Now, the spectral centre of gravity, this important acoustic measurement, is one single measurement. It's simple to calculate (take a Fourier transform, take the average of the frequency bins, weighted by their magnitude). And it's deterministic. If you give me a sound file, my algorithm for calculating the spectral centre of gravity will give the same number each and every time, no matter how many times I run it.

And yet, this number still poses a statistical problem. I had the window open when I made these recordings, and there was a lot of traffic noise. That traffic noise is in the recording. But I'm not interested in the traffic noise. I'm interested in what came out of my mouth. But I don't have direct access to that. I have access to a measurement of what came out of my mouth, plus some other signal that's highly variable and won't be exactly the same between two measurements.

What that means is that, even if I was only interested in the properties of "the **[ʃ]** that came out of my mouth at exactly 3:13 PM on Friday," that itself is not a fixed, directly measurable thing. My measurement is subject to variability, in the sense that the measurements that I have access to **would have been** different if, for example, I had had the window closed. What I want, in some sense, is really the "underlying pattern": the signal that was produced in the noise.

We can take this a step further. Suppose that I want to learn something about "my **[ʃ]** sound." As a native speaker of a human language, I have some internally consistent way of producing **[ʃ]** sounds. It's similar, although not exactly identical, to the way of producing **[ʃ]** sounds that other English speakers will have internalized during early language acquistion. But

Sentence	Rating	Sentence	Rating
Sentence G	100	Sentence U	1
Sentence G	100	Sentence U	68
Sentence G	100	Sentence U	42
Sentence G	100	Sentence U	17
Sentence G	97	Sentence U	3
Sentence G	100	Sentence U	21
Sentence G	100	Sentence U	100
Sentence G	92	Sentence U	21
Sentence G	100	Sentence U	1
Sentence G	100	Sentence U	19
Sentence G	100	Sentence U	11
Sentence G	100	Sentence U	9
Sentence G	100	Sentence U	1
Sentence G	100	Sentence U	7
Sentence G	100	Sentence U	16

it varies, of course. Even if I say **[ʃ]** in exactly the same phonetic context twice, the position of my articulators will vary slightly for mechanical reasons. And it might even be the case that the position of my articulators will vary for cognitive reasons—I don't think we know this for sure, but it's entirely possible that part of the process of language acquisition is not only to learn to reproduce the sounds that you hear in your native language, but also *the variability in their production.* That is to say, I might unconsciously randomly vary my **[ʃ]** sounds to match up with the random variability that other people around me make. I don't know if we've ever tested this systematically, but this would be coherent with my experience in acoustic phonetics.

That means that, when I talk about "my **[ʃ]** sound," I'm not talking about one measurable token. I'm talking about "my way of producing **[ʃ]** sounds," which, for mechanical and maybe cognitive reasons, has inherent variability. And if you wanted to make a statement like, "my **[ʃ]** sound has a lower centre of gravity than my **[s]** sound," you'd have to have some way of interpreting that statement. Again, it would be some kind of "underlying **[ʃ]**," a pattern of some kind amongst the noise.

Instead of trying to go into mental rotations to try and think a priori about how we would make such comparisons amongst abstract objects to which we don't have access, we're going to be practical. We're going to start by looking at what **kind** of questions we're going to want to ask, habitually. Today we're going to start with the absolute most basic: I've got two groups of observations. Behind them, two "underlying patterns"—maybe. Or maybe only one? Maybe, as far as I can tell from these observations, the two groups are of a piece?

2.2 Grammaticality judgments

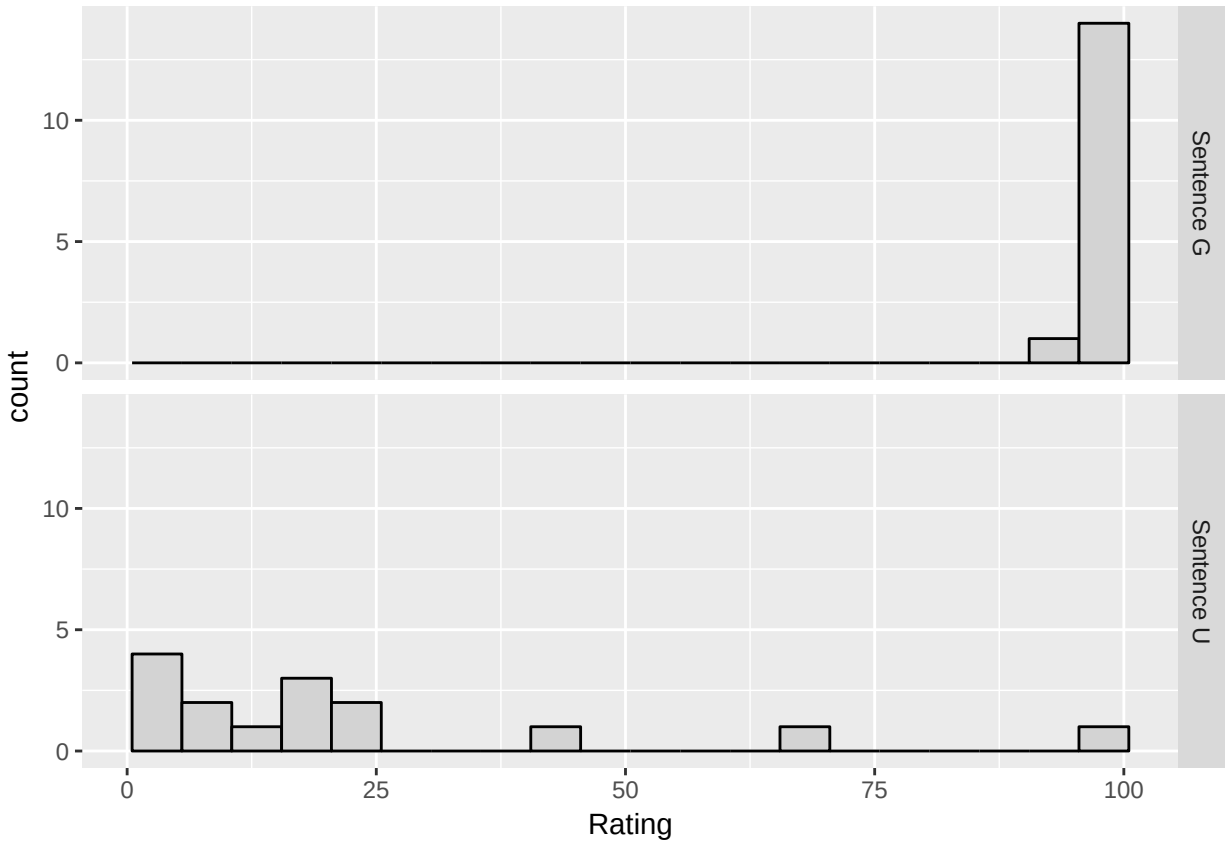
Have a look at the following data set:

These are ratings (from one to a hundred), assigned by a number of native speakers of English, to two utterances:

1. **Sentence G:** Evan's idea is brilliant.
2. **Sentence U:** Evan's the idea is brilliant.

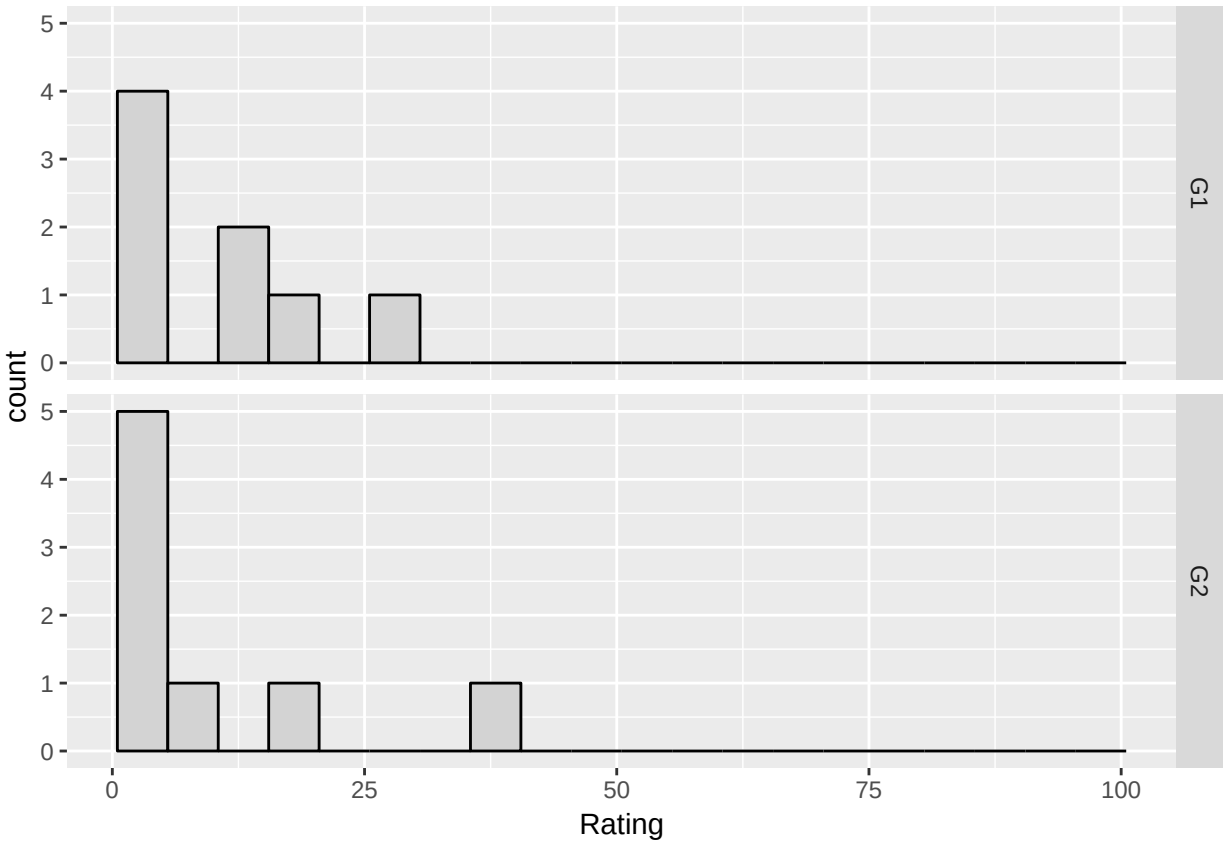
As you can imagine, the "G" stands for "Grammatical," and the "U" for "Ungrammatical." But of course, we don't know this going in. It's an empirical fact: "people" "generally" "like" Sentence G more than Sentence U—if, for example, you ask them to rate the two on a scale of 1 to 100. So let's see that first hand.

Draw a histogram for the ratings for each of the two sentences (one histogram for each sentence). A histogram—that is, a graph showing, for the numerical ratings from one to one hundred (on the x axis), how often each rating is observed (on the y axis). Or perhaps, to make things fit better on your page, each of the following ranges of ratings: one to five; six to ten; eleven to fifteen; sixteen to twenty; …; ninety to ninety-five; ninety-six to one hundred. If so, you should find that your histograms look like this:



So now I want to ask: do people ("people") tend ("tend") to give different kinds of ratings to sentence G than to sentence U? Two questions, in backwards order. First: why do you think that? And second, what does that mean?

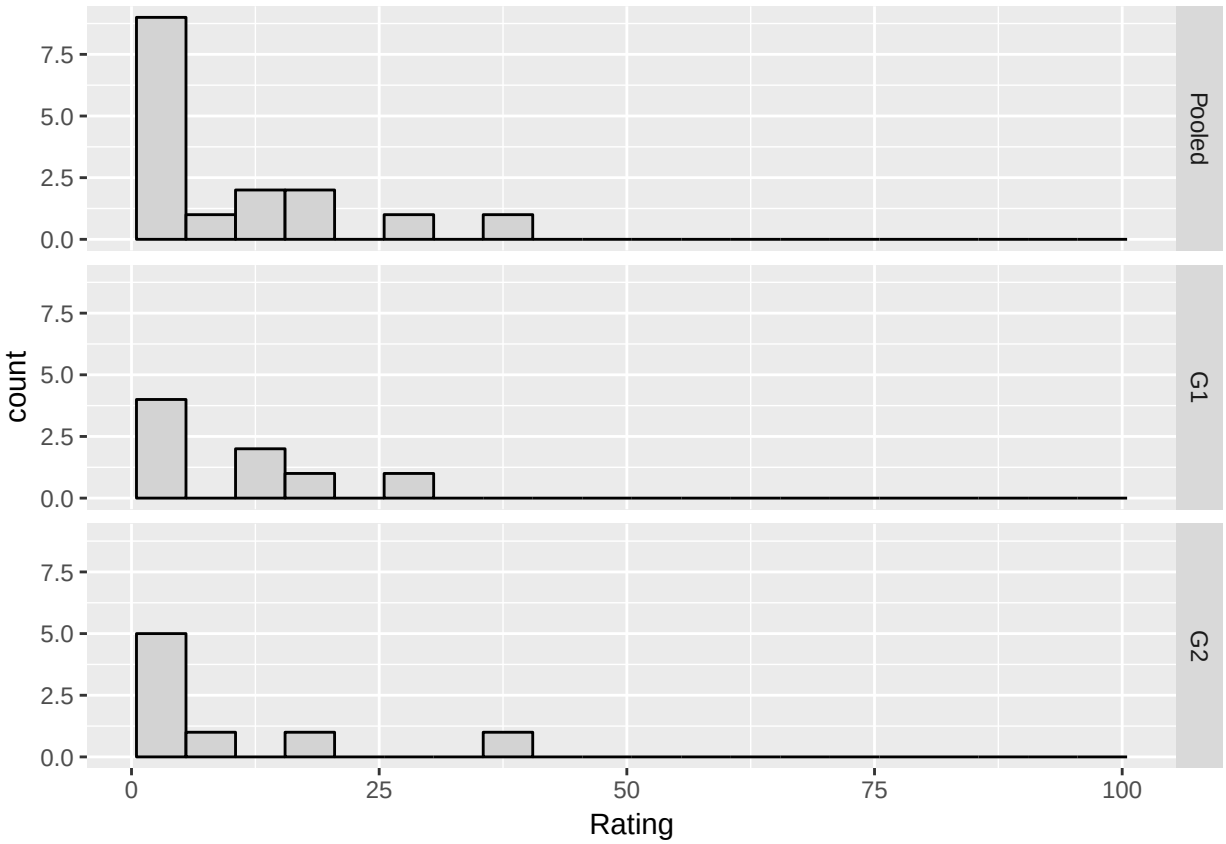
To stimulate your thinking, I'm going to show you a pair of histograms for two sets of data that I'm pretty sure **do** represent "the same kind" of ratings, because I pretty much made them up.



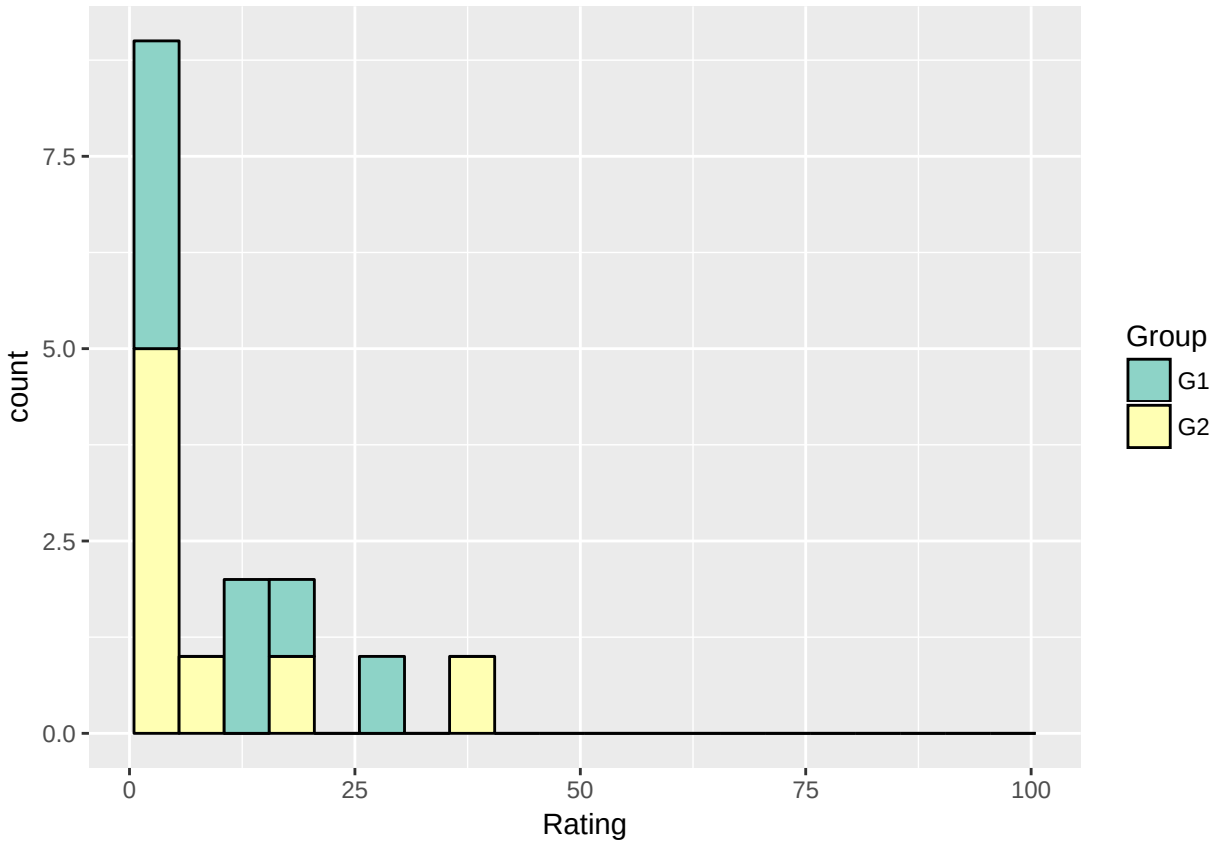
Actually, I didn't make the numbers up, but I made the groups up. These are actually all ratings for the same sentence (*The wedding was beautiful was claimed that by everyone*), and I just split them up at random. So by definition they're of the same kind. There might be interesting differences between certain sub-groups of them, but the point is if I were to look at this graph, and answer the question, are people behaving differently on the top graph than on the bottom graph, if I said yes, I would have to be wrong. Unless by random chance my random numbers have stumbled across some other, unknown difference between certain participants, which is of course always possible.

So: we have one pair of graphs that we think probably represent "two different things" and one pair of graphs that we are pretty sure, unless we had very bad luck, represent "the same thing twice." What's the difference? The difference is, obviously, that in one pair the two graphs look visually similar and in the other one they look visually different. And now if we ask, what do these graphs represent, we'd remember that they represent the number of times we saw each of the possible ratings, grouped in ranges of five. What's the rule? When we have the "same thing" or "the same pattern" underlying two groups of observations, what we mean is that we expect to see similar observations, about the same number of times each (more on this in a minute). When we have two different things, "two patterns," what it means is that we don't expect to see similar observations with similar frequency.

So for these two groups of observations I just gave you, we can observe that both of the graphs are more or less similar to the graph that I get when I pool all of the observations together, like this.

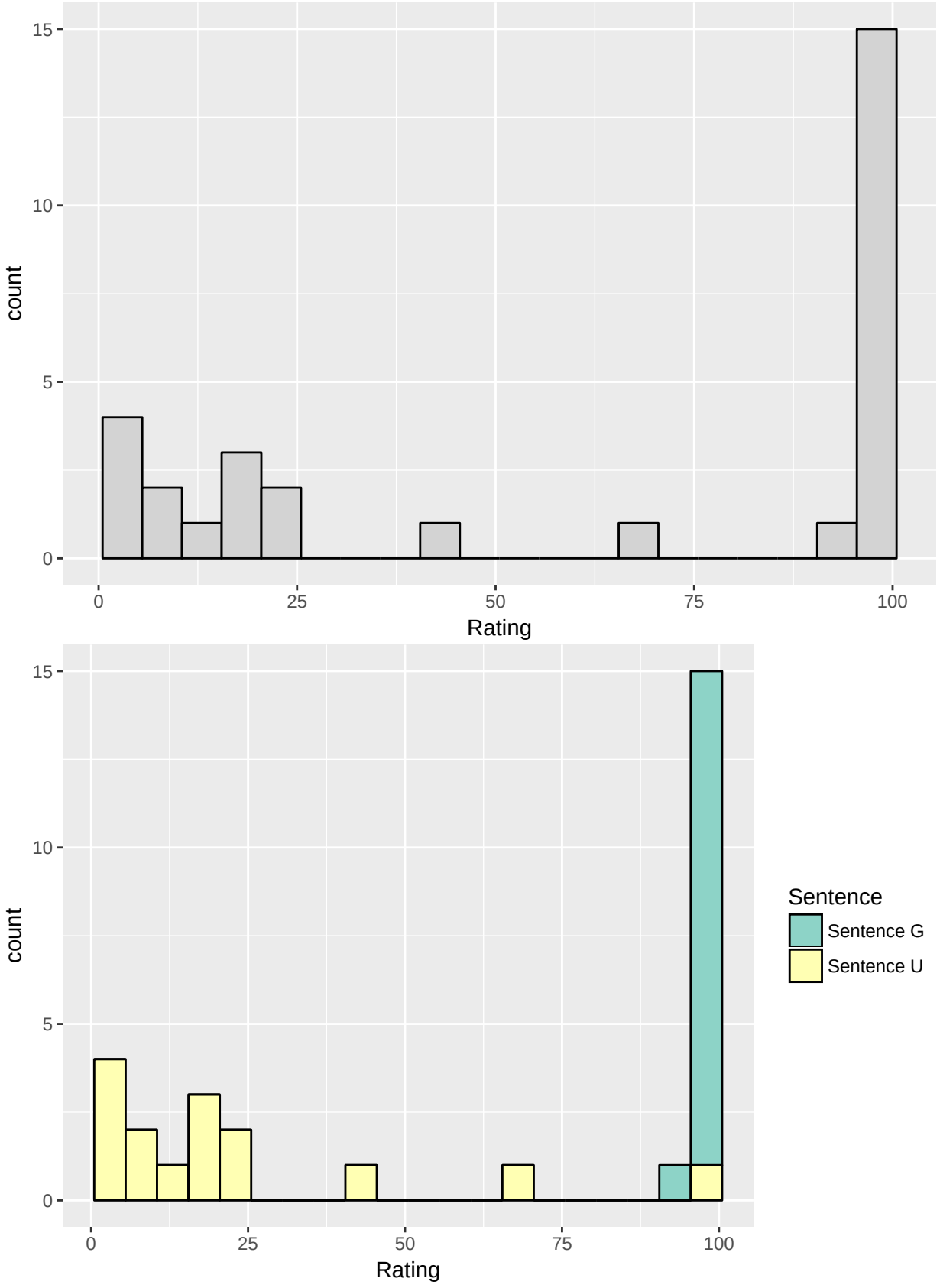


With one caveat, which is that, as you can see now that I've also put the two original graphs right underneath, if I pool all the observations I get bigger counts. We'll come back to that in a minute. For now I'd just like you to ask yourself why you'd *expect* to see the same shape in the graph if you pooled all the observations together, under the assumption that we're dealing with the same thing twice. Now if you know a little bit, you might be tempted to answer, well, because "the same thing twice" means that we have two groups of observations that come from the same distribution, which means the graphs have to have (roughly) the same shape, because the observations have to have (roughly) the same relative frequencies. But there's a simpler answer here, which is why I didn't talk about relative frequencies or distributions yet, and that is that, because I had the same number of observations in each group (eight), and because each graph has about the same shape, if I stack the bars in the two graphs on top of each other, it's going to be roughly double the height.

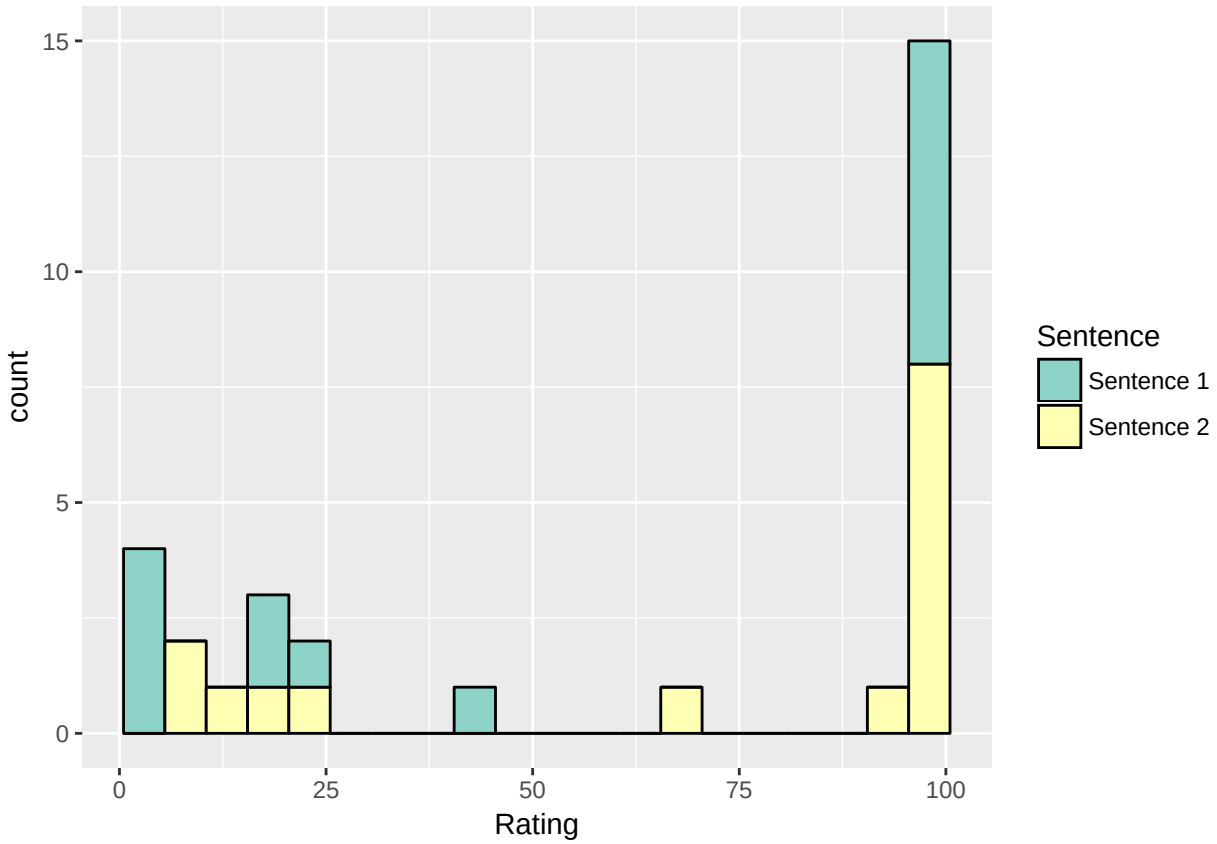


We'll come back to this, but for now let's just see where we are. We started off asking stupid questions like "how tall are we?" And we said that, even though this doesn't make any sense, this is the kind of question we're going to be asking all the time in science, or actually any time we look at data that's subject to variability: "are cats smaller than dogs?" And so we better know how to deal with this sort of question. And so I gave you an example: sentence A was rated higher than sentence B; versus sentence A wasn't rated higher than sentence B, and you knew it wasn't rated higher because it was actually just sentence A twice. And we said, as a first guess, that we'd expect that histograms would have the same, or different, shapes, even if the data weren't exactly the same, because if we took ten ratings for each sentence then we'd expect that they'd be roughly the same, roughly as often. And so maybe in fact what we're really saying is that the **variability** between the ratings of the sentences is the same, in some meaningful sense of "same." And then I showed you that a way to think about "what we would expect if the two groups of observations were really the same" is to put all the observations in one big group and look at the shape of that graph.

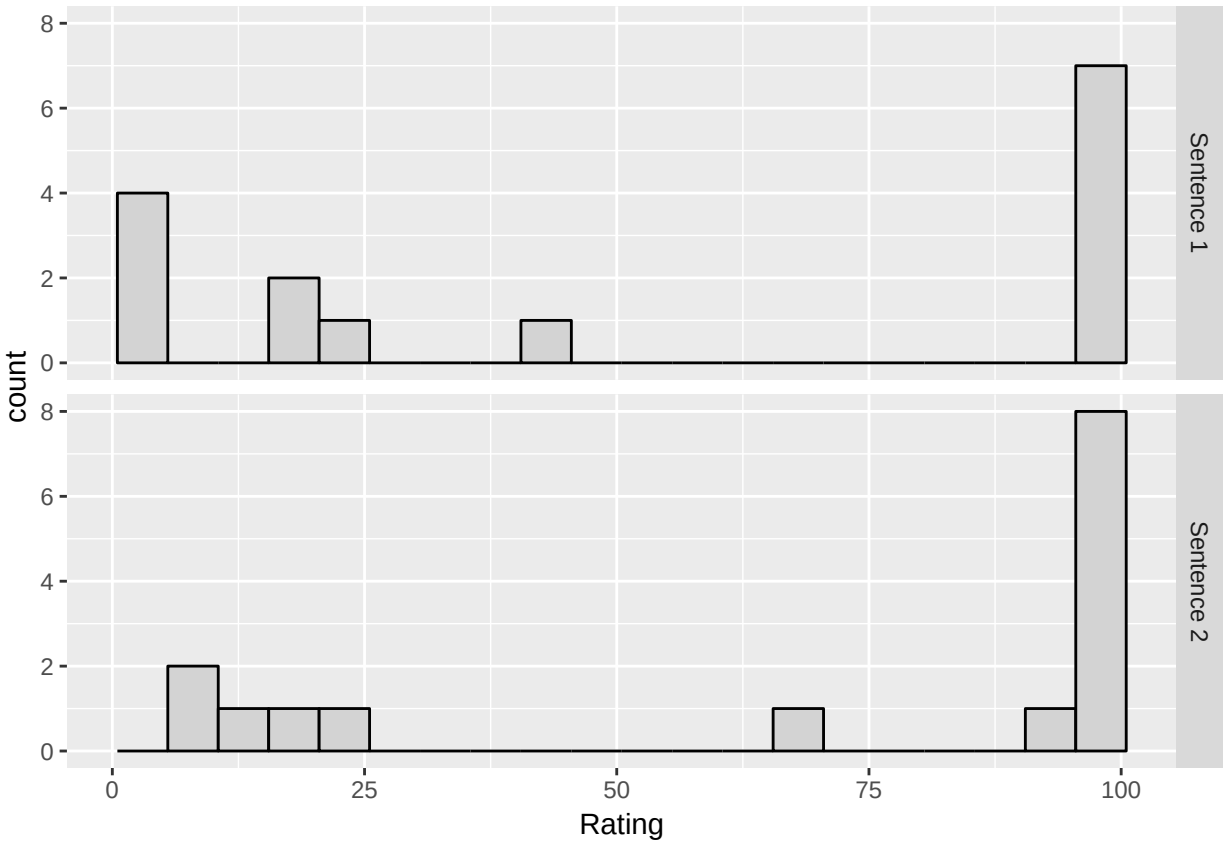
So let's do that for our original data. First, let's look at the actual breakdown, in colours.



Now let's imagine that the observations really were "all the same." Here's what we would expect:



Here's what I've done: I've taken the data, and I've again made fake groups, of the same size as in the original data (15 each). Except that, because the groups are meaningless, we don't expect there to be any systematic difference in the kind of responses I get, and there isn't really, not in any obvious visual way.



Now that doesn't mean it doesn't look like there's very big differences among the responses. You can see that there are. In this thought experiment, there are some extremely high ratings, and some extremely low ratings. But in this thought experiment, that's not because of the sentence we gave people. Maybe it's because of something else. Or maybe it's because that's just what people do when you ask them to rate a sentence: every so often they mess with you. But the two sentences look about the same. They don't have *the same ratings* in the sense of the same number all the time, or even the same set of numbers. But they have similar frequencies for each possible rating.

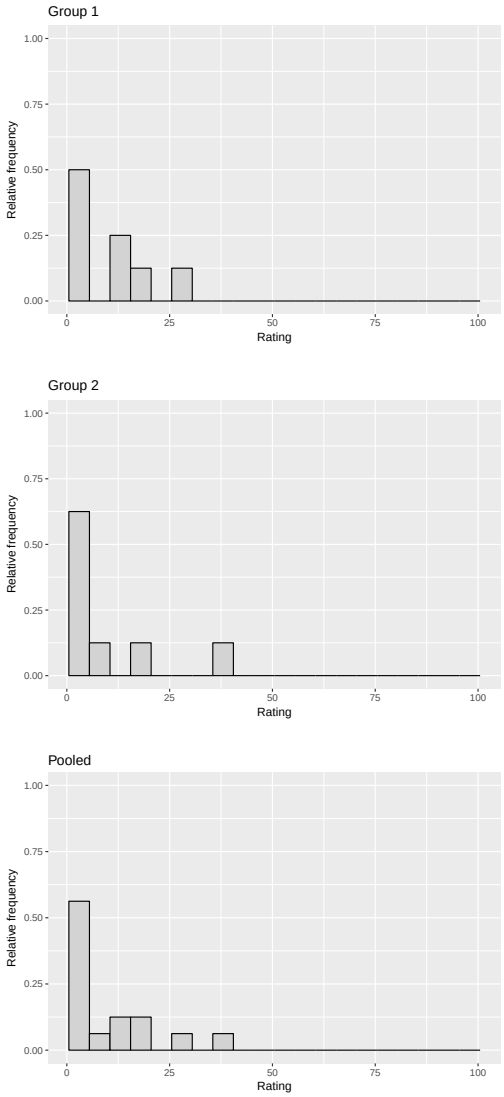
The question, "are cats bigger than dogs," which we might think makes no sense, we can't quite address, but we can at least address a related question, which is, "are cats differen from dogs with regard to their size." That question we can now make sense of, and we can now translate in the following way: do we think that, if we take a group of cats, and we take an equally-sized group of dogs, and we measure their sizes (let's say the length from nose to tail), that across the two groups of observations, we're going to get roughly similar numbers roughly as many times?

2.3 A potentially useful comment

It might be useful to think about this. What if there was no variability? Well if there was no variability—if we could guarantee there was no variability—then everything would be easy. I wouldn't have to graph anything. I'd just look at the numbers I got for sentence A—let's say they're all 100, but I wouldn't have to **check** that they're all 100 because I just told you I knew going in that they were all going to be the same—I'd just have to look at one of them. And then I'd look at my ratings for sentence B, and I'd see that the rating is 35, for example, and I'd be able to conclude that they're different. Why? Because 35 is different from 100. And if they were the same, I'd be able to conclude that the two sentences yield the same ratings, because, say, 100 is the same as 100. But once I introduce the **possibility** of variability—even if there's no actual variability in my data set—then I have to ask a different question. Do I **think** these two groups are different? Or do I think that I simply sometimes get 35 and sometimes get 100 and it has nothing to do with the sentence that I gave them? That would predict that it was just random luck that I didn't come up with the number 35 for sentence A and 100 for sentence B.

2.4 Counts and relative frequencies

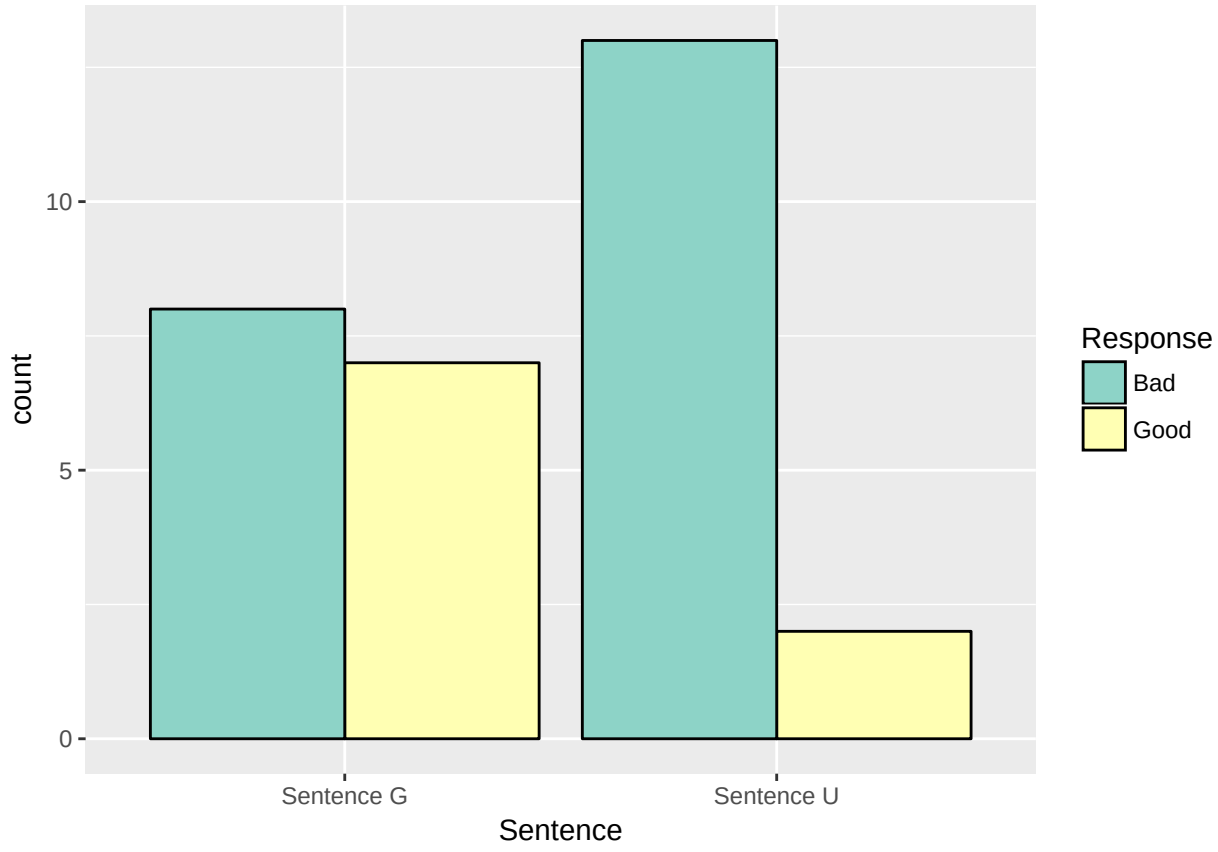
Now, up to now, we've basically only been making comparisons between histograms that have the same number of observations. At one point, though, I did suggest that maybe you could visually assess whether you had two groups of similar observations, or two groups of dissimilar observations, by checking to see whether the two histograms both look like the single histogram that you'd obtain if you pooled everything together. The fact is, that it's a bit annoying even if we're just making visual comparisons between graphs that we have to compare the shape while abstracting away from the size. So let's just abstract away from the size. We said that if we started off with eight observations, and the shape was the same in a group of sixteen observations, then basically the only difference in the graph would be that the bars would be double the size. And vice versa, if we started off with eight and then went to four, the bars ought to be half the size while the shape remains the same. (Of course, again, we were assuming that, since there's variability, the shape of the graph won't be **exactly** the same, but at any rate, it will be **roughly** the same.) Well if we want to have a standardized **shape without size** graph, we can go all the way down to dividing the size of the bars by eight, or in general, the total number of observations. And then we know that we're only looking at the shape, and the shape, in this case, simply represents how **relatively frequent** each of the bins of observations is.



This is useful for many things, as we'll see. For now, it's useful just for reasoning about what the "no difference" distribution looks like.

2.5 Categorical data

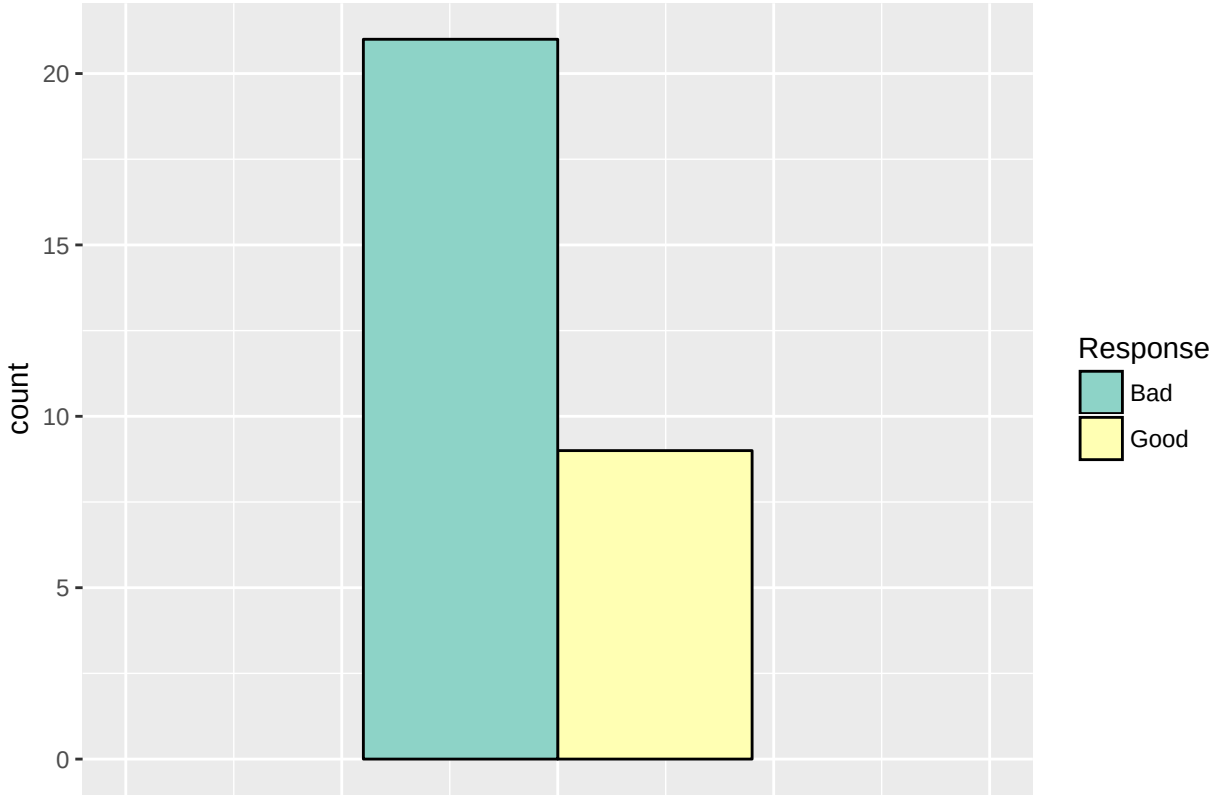
Now, ratings are numbers. But the exact same type of reasoning applies to data that isn't numerical. Here we have a different pair of sentences, a little bit more ambiguous, it seems, and this time the question was just whether they were good or bad. There was no numerical rating.



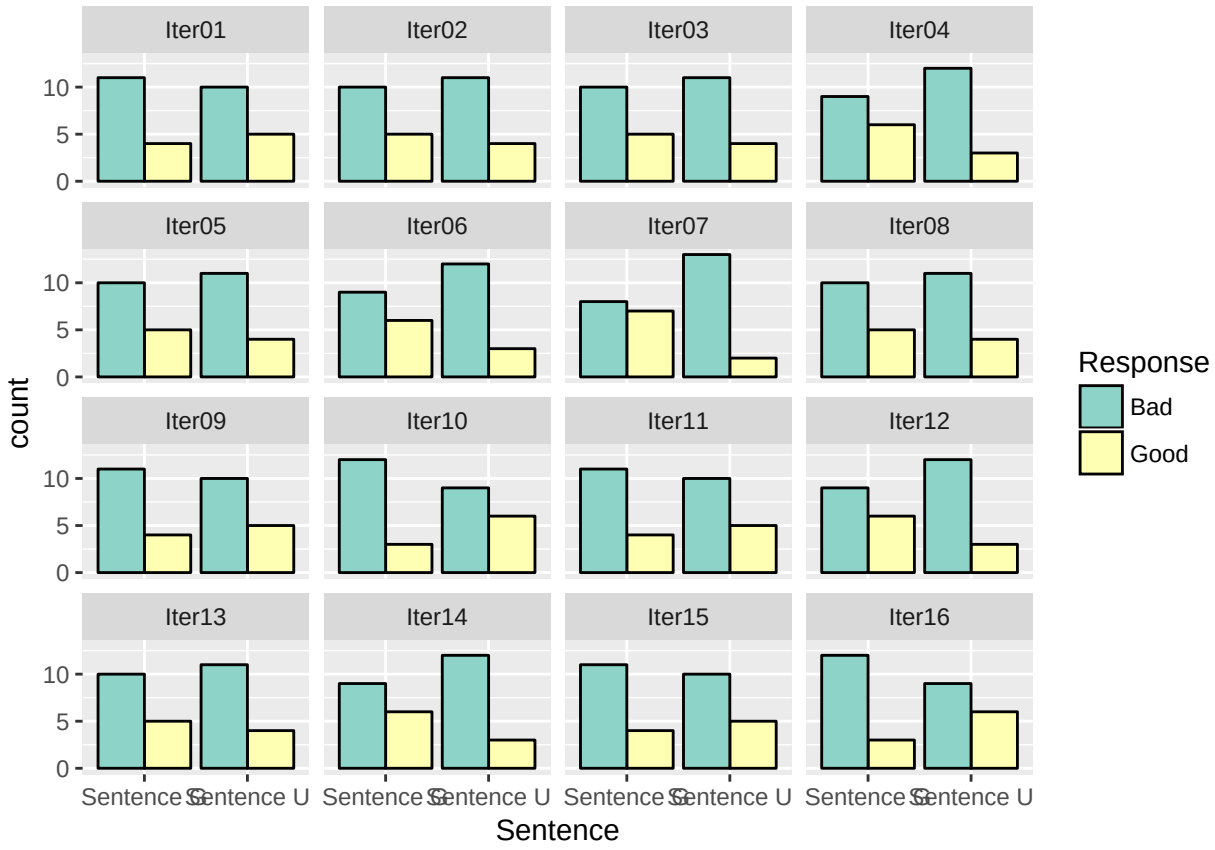
Now, let's do the same thing. We've got two hypotheses:

- **Hypothesis A**: Sentence G and Sentence U are the same in terms of their binary responses.
- **Hypothesis B**: Sentence G and Sentence U are different in terms of their binary responses.

First, let's explain these two hypotheses. Based on what we said before, we'd say that what Hypothesis A means is that we expect the response "Bad" roughly as often for Sentence G as for Sentence U, if we take the same number of data points for each. (And thus we'd expect the response "Good" roughly as often, because there were only two options.) There's a fuzzy part in there, which is the idea of "roughly as often." But let's stick with that for the moment, and envision what that looks like.



We have 15 responses in each group. If they were "the same," then we'd expect "roughly" 10 or 11 "bad" responses (half of 21, the total number of "bad" responses) and the remainder ("roughly" 4 or 5) would be "good." Now, how sure am I really that those first two graphs don't "roughly" look like that? They sort of do. Let's do what we did before, and let's make up some random groupings. Remember that this represents the case where we think that the two sentences ought to look "roughly" the same.



So: does our original data look like this? A little. Not a lot, but a little. In fact, to my surprise, Iteration 7 is exactly like the original data. Intuitively, Hypothesis B seems more plausible than Hypothesis A to me. But these little experiments leave some doubt in my mind. That's my conclusion.

2.6 Distributions and samples

What is all this "roughly" business? Let's look over this last step. I wanted to imagine what I would find if the two kinds of sentences really were the same. So I said, if they're the same, then any sample that I draw should have roughly the same proportions of "Bad" and "Good" responses. Then, the proportions of Bad versus Good would be roughly the same as what I observed overall, as well. Now, in order to give concrete examples, I had to do a thing called "sampling", which means I told the computer, generate some fake data "as if X were true." But of course I had to tell it what X is. What I told it was, act as if the reality is, that the proportions in the overall sample are correct: 21 Bad, 9 Good responses. And now, randomly divide those responses into groups of 15, and see what happens.

Now, that might be on the right track. Or this specific data set might not be entirely representative of what happens when people are asked about these sentences; they might be a little weird—after all, I only tried thirty times. In which case, I might not be drawing a conclusion that's entirely representative of what I ultimately want to know about. But I'm still drawing an important conclusion: that maybe these observations **could have arisen** under the circumstance where the two sentences "behave the same," even based on my limited knowledge.

So, when we said "roughly": we meant two things, actually. We meant, first of all, that of course, since there's variability, we don't expect exactly the same (or necessarily different) **numbers** of observations in each bin, under either of the two hypotheses. We ask whether each bin is **consistent** with some imagined world in which the two sets of observations were both generated by the same **underlying process** (or not), not whether the number of observations is identical. And when we were referring to the idea that the two sets of observations would be "roughly" the same as the shape of the graph we got if we pooled all the data together, we recognized that, that too is just a finite sample reflecting an incomplete knowledge.

In order to compensate for this "roughly"—the inherent variability we have in a finite sample—what we're going to do in this class is find out sensible ways of asking the question "how roughly." But for today, we're just going to stick with the graphs

Here's one thing, and maybe the only thing, that's abstract and a bit mysterious: the underlying distribution. A distribution is a big list of numbers scoring how relatively likely each of the possible outcomes is: for example, good or bad, or each of the numbers between 0 and 100. Now, by convention, but also for some very practical reasons, we state distributions as non-negative numbers that we require to add up to one, once we've exhausted every logical possibility. So, if the two logical possibilities are good and bad, then we need to have numbers like, say, 0.8 Good and 0.2 Bad. Now the numbers are easy to interpret, superficially: if that's the underlying distribution, then, if I look at any observation, eight out of ten times, it would be Good. But the tricky part is, what is this "eight out of ten times"? It's not "eight out of ten actual observations". There is an entire branch of philosophy dedicated to debating this question, and probabilists, physicists and statisticians do not have their collective minds made up. An easy way to think of it is this: whatever the underlying process is, it makes it so that there's a "propensity" weighting in favour of "Good" observations by about a factor of four. We can say that "Good" is about four times more probable, or four times more likely, than "Bad." And that **exact** propensity will yield **roughly** four times more Good than Bad responses.