ANOVA and Chi-Square Tests

John Chandler

You’ve seen a bit about hypothesis testing, where we run a test to see if two means or proportions are different. We’ve also spent time on regression, a technique for modeling the relationship between explanatory (also called “independent”“) variables and a resonse (also called”dependent“) variable. Now we’ll cover a couple of special cases that come up often enough that we have their own techniques.

This document is relatively long, so feel free to tackle it in bite-sized pieces.

Data and EDA

For this exercise we’ll work with data from a Washington credit union. There’s a lot of data in here and we won’t use it all. We’ll read in the data and do some exploratory data analysis.

```
data_dir <- "C:\\Users\\jchan\\Dropbox\\Teaching\\AppliedDataAnalytics\\Data\\"

data_file <- "survey_data.txt"
d <- read_tsv(paste0(data_dir,data_file))

# It's nice to have these levels in a sensible order.
d <- d %>%
  mutate(engagement = factor(engagement,
                             levels=c("Not Engaged",
                                      "Engaged",
                                      "Highly Engaged")))

knitr::kable(head(d[,1:11]))
```

id	age	gender	engagement	mem.edu	zip	channel	progressivism	harm	fair	in.group
346	56	other	Engaged	2	98503	Branch	-0.0833333	1.00	1.00	1.00
348	66	female	Highly Engaged	3	98012	Branch	2.2916667	5.00	4.75	2.25
349	71	male	Not Engaged	7	98506	Branch	1.0000000	3.25	4.25	2.25
352	66	male	Engaged	7	NA	Branch	0.5833333	4.25	4.75	4.75
358	50	male	Highly Engaged	4	98233	Branch	-0.5000000	2.00	4.50	3.50
361	40	female	Engaged	3	98520	Branch	0.5416667	3.75	4.00	3.75

```
knitr::kable(head(d[,12:19]))
```

authority	purity	account.age	region	public.sector	sustainability	localism	pub.greater.priv
1.00	1.25	9.951	Thurston	yes	2.333333	2.75	3
3.25	2.25	8.838	W WA Metro	yes	6.000000	5.75	1
3.25	2.75	5.356	Thurston	yes	5.166667	4.50	4
3.25	3.75	9.192	W WA Metro	no	5.500000	4.50	3
4.50	3.25	11.570	W WA Non Metro	yes	2.833333	4.00	2
3.00	3.25	5.932	W WA Non Metro	yes	3.500000	5.25	3

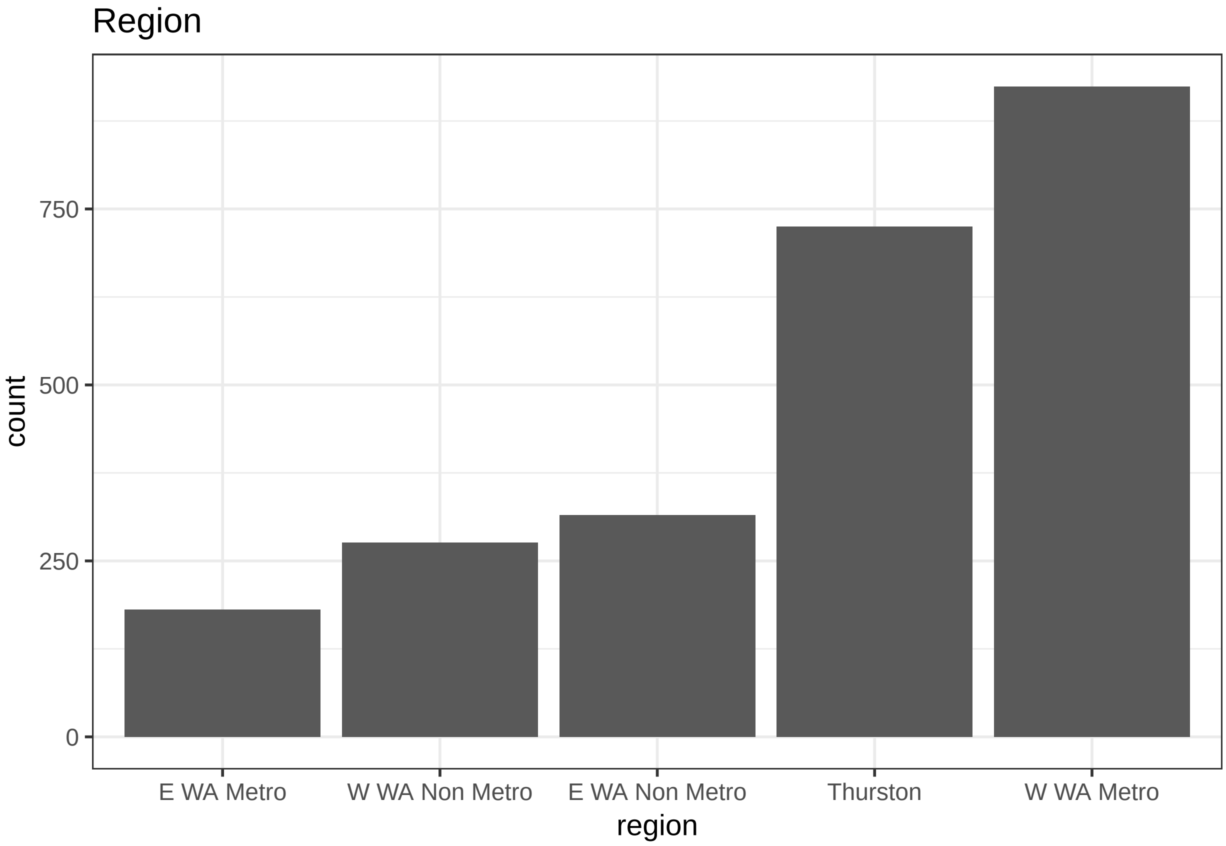
```
knitr::kable(head(d[,20:23]))
```

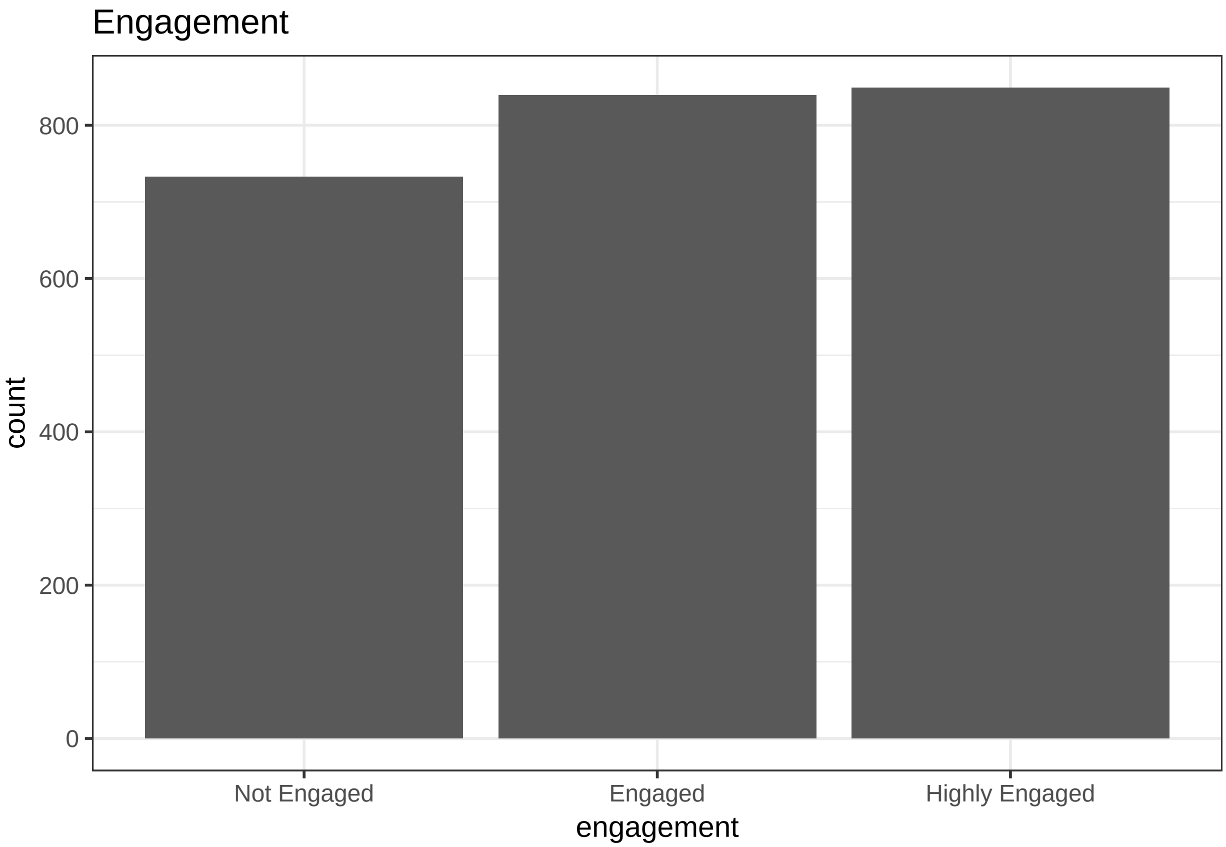
experience.more.important	teachers.underpaid	main.focal.value	support.of.focal.value
3	2	Homelessness	0.000000
4	5	Environment	18.627830
3	6	Hunger/Poverty	28.739986
3	4	Education	18.055372
4	5	Education	4.400921
5	6	Homelessness	28.680137

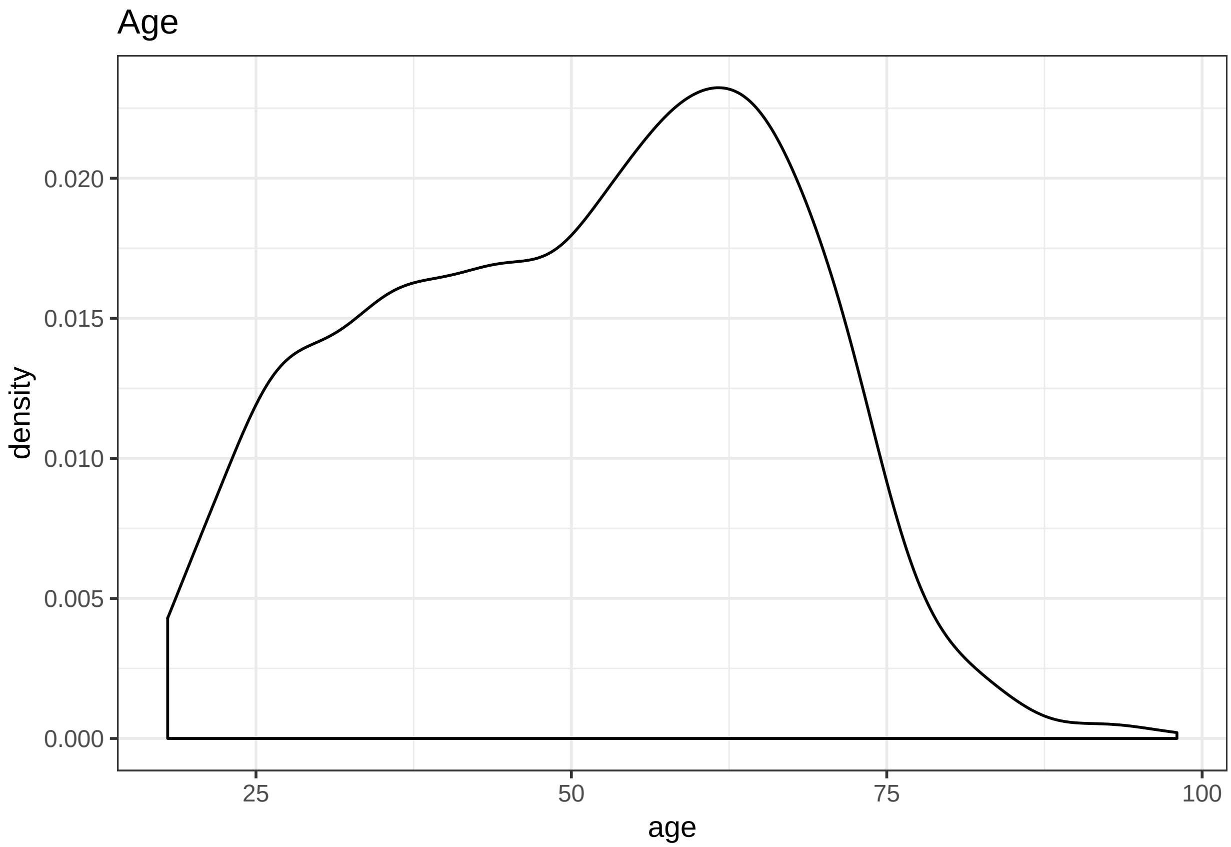
In this tutorial, we'll need the following variables:

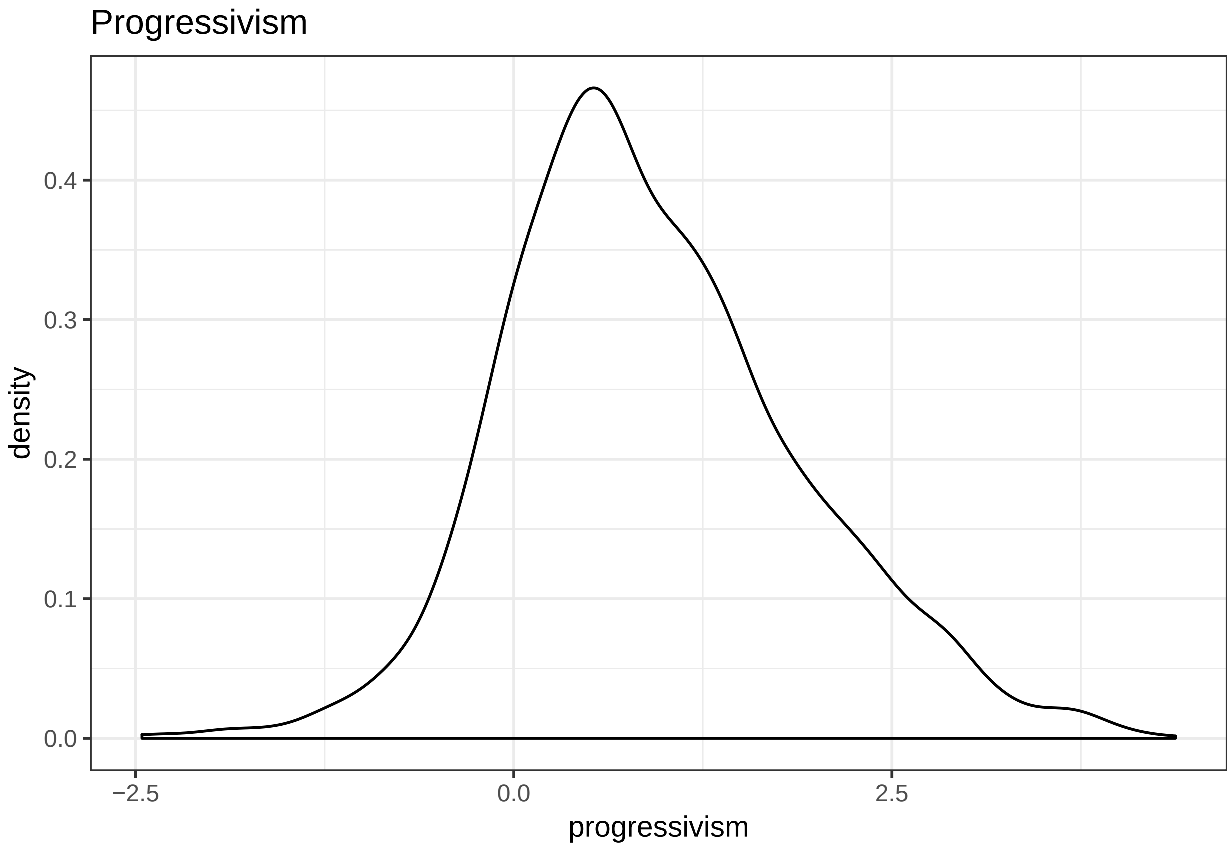
1. *Region*: the region where the credit union (CU) member lives.
2. *Progressivism*: A derived measure from the Moral Foundations Questionnaire. It is the result of the Fair and Harm dimensions minus the Loyalty, Authority and In-group measures. See the appendix for more on this.
3. *Engagement*: A measure determined by the CU to indicate how engaged the member is. Basically a function of number of accounts/products, whether or not they have direct deposit, whether or not they use bill pay, and the number of channels they bank from (online, mobile, branch).
4. *Age*: The member's age. (Confusing, I know.)

Let's do some plotting and descriptive statistics on those. Let's start by just plotting the distributions or counts.




Engagement
count
800
600
400
200
0
Not Engaged
Engaged
Highly Engaged
engagement


Age
density
0.020
0.015
0.010
0.005
0.000
25
50
75
100
age



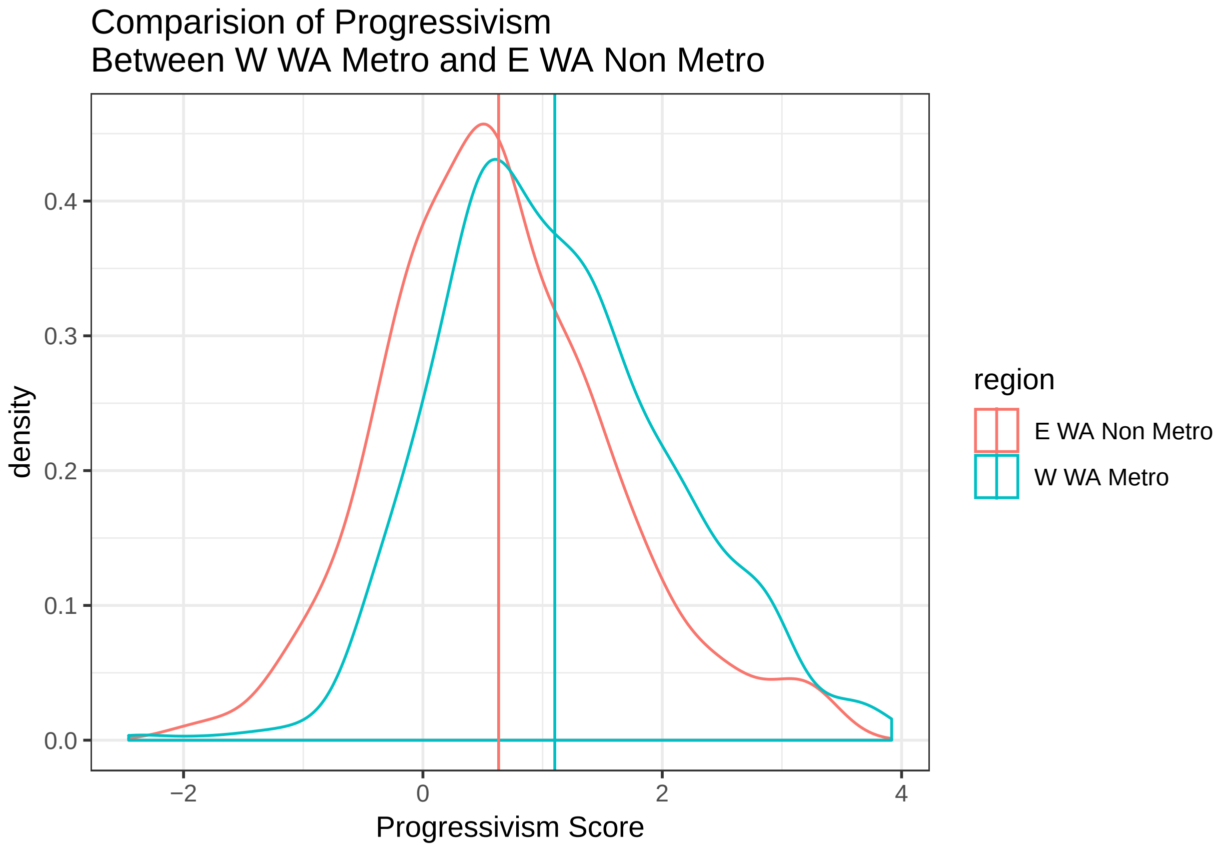
Now some work for you. Edit the code block below to build the following charts:

1. Distribution of age by engagement (use `facet_wrap`).
2. A mosaic plot of enagement by region (use `mosaicplot` which *isn't* in `ggplot2`).
3. Distribution of progressivism by region.
4. Scatterplot of age and progressivism. Add a smoother with `+ stat_smooth(method="gam")` to explore the relationship.
5. Add a layer to the previous plot that colors the plots (and smooth lines) by region.

```
## Your code here.
```

A t-test refresher

We use t-tests when we want to compare two means. For instance, we could compare progressivism between W WA Metro and E WA Non-Metro.

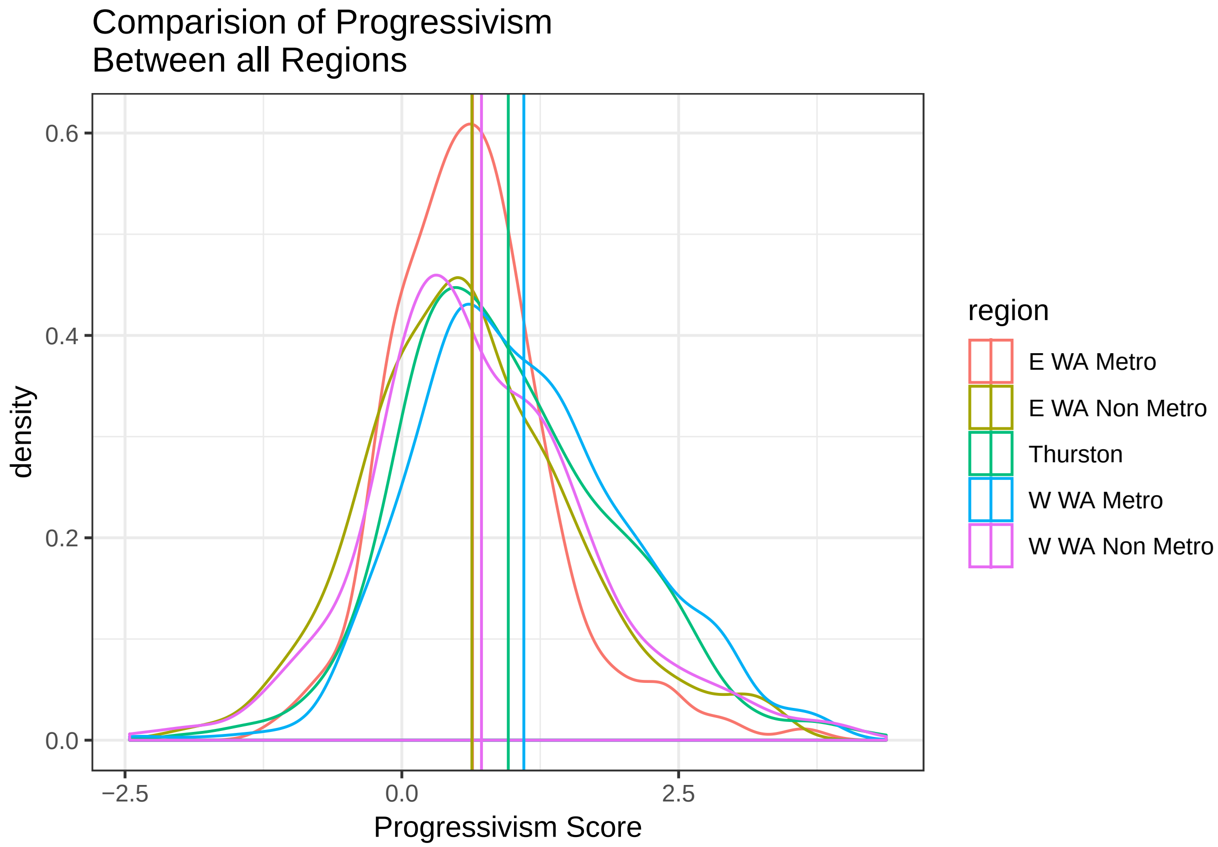


```
##
##  Welch Two Sample t-test
##
## data:  data.for.test %>% filter(region == "W WA Metro") %>% select(progressivism) %>%  and data.for.
## t = 7.4596, df = 548.86, p-value = 3.415e-13
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.3456773 0.5928027
## sample estimates:
## mean of x mean of y
##  1.101912  0.632672
```

The graph makes it look like there is a signficant difference between the means and the t-test confirms it. The t-statistic is 7.5 on on 549 degrees of freedom (which mean this distribution is almost exactly normal). Being seven standard deviations above the mean of a normal gives you a small p-value ($p < 3.4 \cdot 10^{-13}$).

Does this result look surprising? If so, read the next section, *Comparison Between Means*.

Regardless of how that picture looks, there aren't two regions, but five. So the picture becomes more compli-



cated.

Again, there are some clear differences. The two regions in Eastern Washington score lower and their scores are almost identical. Western Washington non-metro closer to Eastern Washington than the high progressivism scores in Thurston county (home of the state capitol Olympia) and Western Washington metro.

How could we verify these differences are not the product of random chance? We could search through all the pairs of regions looking for a significant one. But there are $\frac{5 \cdot 4}{2} = 10$ of those pairs. In this case we might still be okay, but if there were 50 regions, then we'd be making 1225 comparisons and we'd expect to get some false positives. ANOVA solves this problem for us.

Comparison between Means

Glance back at the chart with the two densities. The lines, which represent an estimate of the probability distribution that the values come from. And those distributions overlap substantially. Eastern Washington non-metro ranges from about -2 to 4. Western Washington metro covers the same range, but has a heavier "shoulder", starting around its mean value of 1.

So these distributions overlap quite a bit, yet the p-value was less than $3.4 \cdot 10^{-13}$. What's up with that?

Before we answer that question, let's validate your eyes. When you're looking at how overlapped those distributions are, your brain is implicitly drawing a value from each distribution and seeing which one is bigger. Take a break from reading and go back to that image and try to guess the probability that a point drawn from the blue distribution is larger than one drawn from the red distribution.

When I looked I guessed something like 55%. Let's test it with code.

```
e.wa <- d %>%
  filter(region=="E WA Non Metro") %>%
  select(progressivism) %>%
  sample_n(10000,replace=T)

w.wa <- d %>%
  filter(region=="W WA Metro") %>%
  select(progressivism) %>%
  sample_n(10000,replace=T)

mean(w.wa > e.wa)
```

```
## [1] 0.6321
```

Okay, so that was 0.63, so why is the p-value so low? The answer lies in the fact that we're talking about *samples* from the distribution and summary statistics like the `mean`, taken from that distribution.

To get a clearer sense of what's going on, let's take samples of a variety of sizes and keep track of some summary statistics. We'll do samples in powers of 2: 1, 2, 4, 8, 16, 32, 64, 128, 256. We can keep track of the means as we go up and it'll be interesting to look at the standard deviations of those means. We'll run each experiment 1000 times, but you can make this number smaller if your machine is tired.

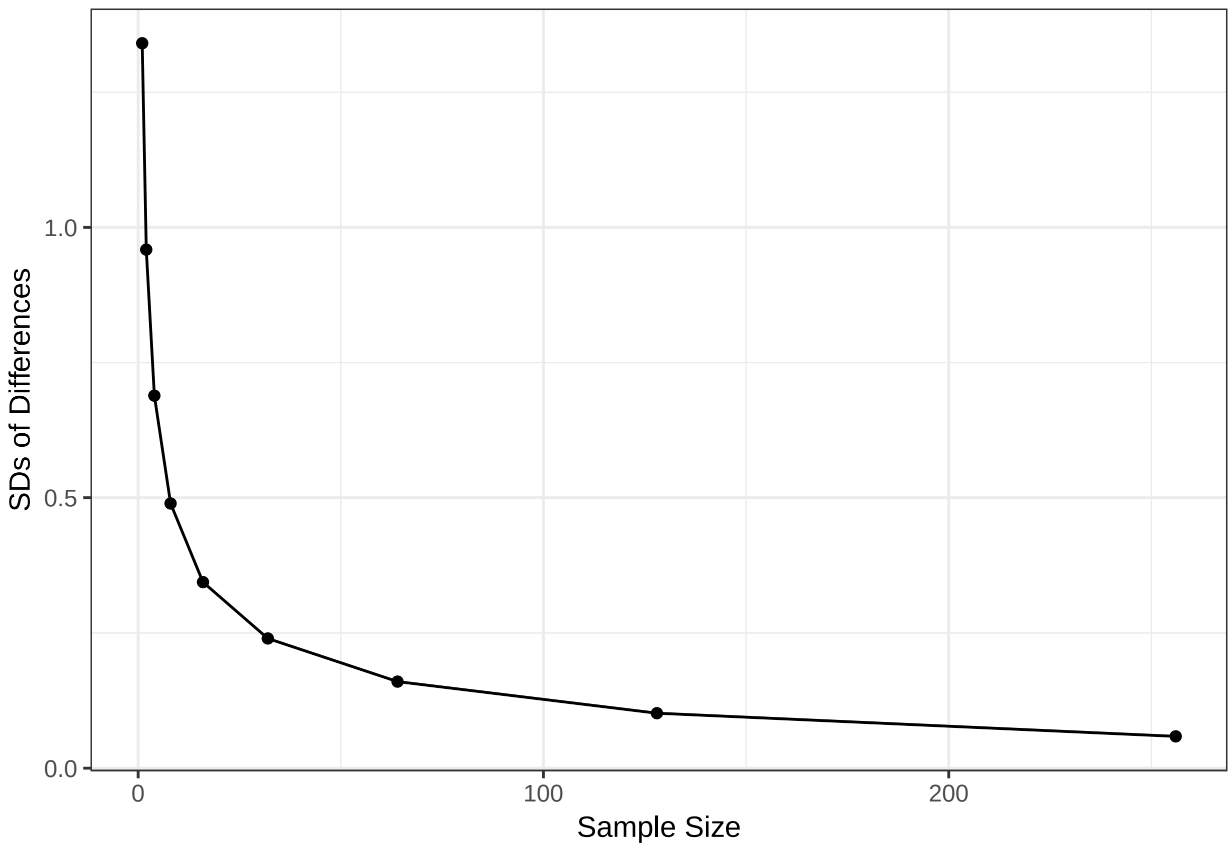
```
# Build a container to hold the results.
results <- data.frame(num_samples=1000,
                      sample_size = 2^(0:8),
                      mean_of_diffs=0.0,
                      sds_of_diffs=0.0,
                      frac_w_gt_e=0.0)

# build the vectors we'll sample from.
e.wa <- d %>%
  filter(region=="E WA Non Metro") %>%
  select(progressivism) %>%
  unlist %>% # this trick gets us out of tibbles
  as.numeric # there might be an easier way.

w.wa <- d %>%
  filter(region=="W WA Metro") %>%
  select(progressivism) %>%
  unlist %>%
  as.numeric

# Loop over all the experiments we want to do.
for (i in 1:nrow(results)) {

  # Grab some data here so we only have to do the
  # df lookup once.
  simulation.size <- results$num_samples[i]
  sample.size <- results$sample_size[i]

  # Set up our storage container
  holder <- data.frame(mean_diff=rep(0.0,simulation.size),
                       w_gt_e = rep(0,simulation.size))

  # Now we run the experiment. There are faster
  # ways to do this, but they get complicated because
  # we need to sample without replacement in each
  # experiment. We're imagining a counterfactual where,
  # rather than having 924 people from W WA and 315
  # people in E WA, we've only sampled `sample_size`
  # people in each place.
  for (j in 1:simulation.size){
    a <- sample(w.wa,size=sample.size,repl=F)
    b <- sample(e.wa,size=sample.size,repl=F)

    a.minus.b <- mean(a) - mean(b)

    holder$mean_diff[j] <- a.minus.b
    holder$w_gt_e[j] <- as.numeric(a.minus.b  > 0)

  }

  results$mean_of_diffs[i] <- mean(holder$mean_diff)
  results$sds_of_diffs[i] <- sd(holder$mean_diff)
  results$frac_w_gt_e[i] <- mean(holder$w_gt_e)

}

knitr::kable(results)
```

num_samples	sample_size	mean_of_diffs	sds_of_diffs	frac_w_gt_e
1000	1	0.4896667	1.3404974	0.632
1000	2	0.4427708	0.9589304	0.672
1000	4	0.5071562	0.6889143	0.766
1000	8	0.4847396	0.4894455	0.835
1000	16	0.4856276	0.3440013	0.919
1000	32	0.4695352	0.2398249	0.982
1000	64	0.4697044	0.1600785	1.000
1000	128	0.4705785	0.1016544	1.000
1000	256	0.4690612	0.0588336	1.000

So, what do we notice from the results table? As the sample size goes up, the mean of the differences stays about the same. In other words, it looks like the mean of 1000 differences doesn't vary too much whether it's a single observation per sample (in which case you've basically just taking a sample of 1000) or 256. But the standard deviations do seem to be dropping. And we seem to be getting the "right" answer more of the time. (I'm assuming that W WA Metro really *is* more progressive than E WA Non Metro. You could make this case based on political voting patterns, or philosophically, by deciding that, for the purposes of our experiments, the survey results *are* the population of interest.) Notice that by $n = 128$, we're not ever getting the wrong answer.

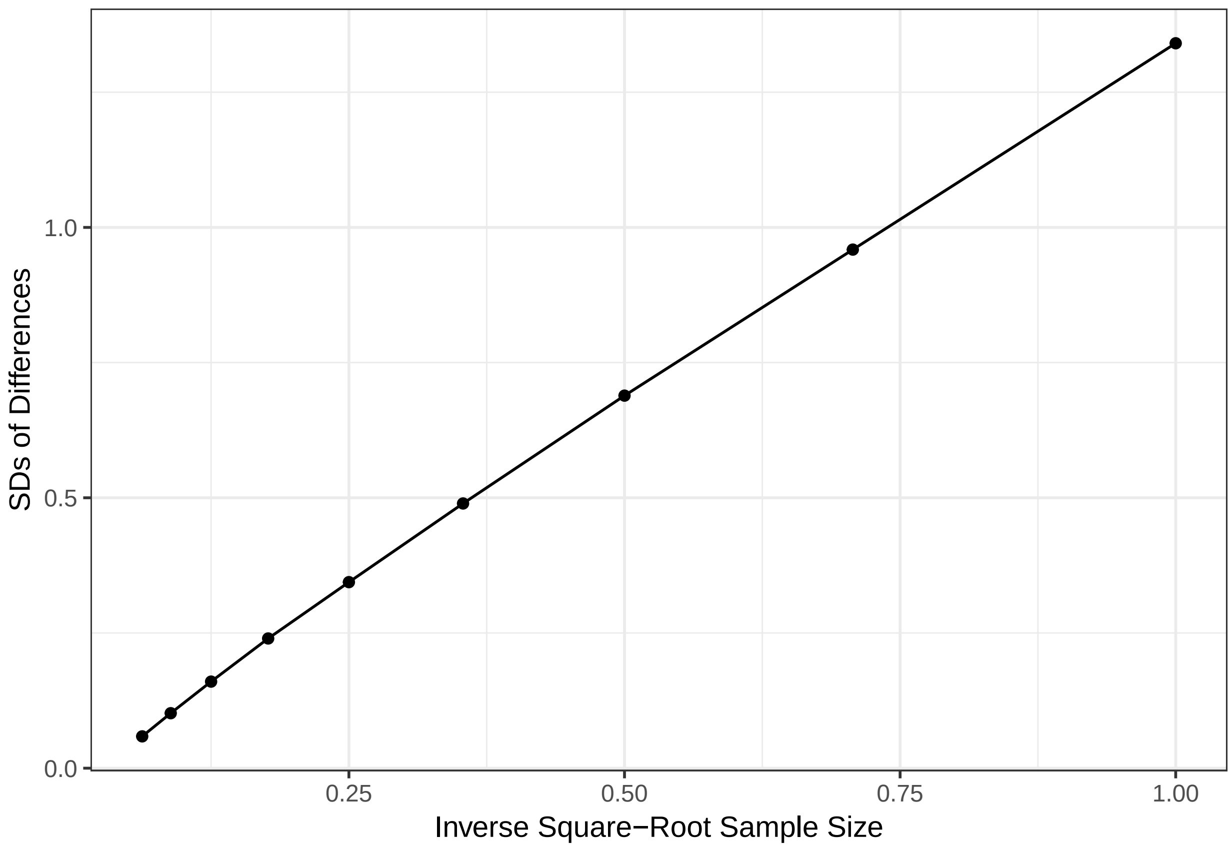
Let's try to figure out what's going on with those standard deviations. First, we can plot them to get a better look.

```
ggplot(results,
       aes(x=sample_size,y=sds_of_diffs)) +
  theme_bw() +
  labs(x="Sample Size",y="SDs of Differences") +
  geom_point() +
  geom_line()
```



Clearly the standard deviation of the mean differences drops off quickly as the sample size increases. It's hard to see the exact functional relationship, although a transformation will help. Since I know the punchline, I'm going to suggest a transformation of $f(n) = \frac{1}{\sqrt{n}}$.

```
ggplot(results,
       aes(x=1/sqrt(sample_size),y=sds_of_diffs)) +
  theme_bw() +
  labs(x="Inverse Square-Root Sample Size",y="SDs of Differences") +
  geom_point() +
  geom_line()
```



The straight line here is no accident. The Central Limit Theorem tells us that, for a random variable X with mean μ and standard deviation σ, the sampling distribution of the mean is the normal distribution. Not only that, but we know the _parameters for the distribution of the mean, $\bar{X}$: $N(\mu, \frac{\sigma}{\sqrt{n}})$.

Why am I going through all this? This property, that the variablity of sample means is smaller than the variability of a population, is basically the foundation of hypothesis testing. So it's worth it to have a good, intuitive grasp of what's happening. As a sample size goes up, small differences in means can start to become quite large.

Let's close with one last simulation to hammer home the point. That value of $p < 3.4 \cdot 10^{-13}$ seemed absurdly small. But, taking samples of 256, which is smaller than the sampling from either region, didn't yield a single example where the mean progressivism from Eastern Washington was larger than that from Western Washington. Let's run a larger simulation to see if we can find one example. We're going to cut some corners here to make this run in a reasonable amount of time, sampling *with* replacement and having 128 credit union members in our samples. This code is set to be cached by default, so if you need to re-execute it you'd have to change the `cache=T` flag to `cache=F`. If you don't want to run it at all, change the `eval=T` to `eval=F`.

```
set.seed(20171119)

simulation.size <- 10^6
sample.size <- 128


e.wa <- d %>%
  filter(region=="E WA Non Metro") %>%
  select(progressivism) %>%
  unlist %>%
  as.numeric

w.wa <- d %>%
  filter(region=="W WA Metro") %>%
  select(progressivism) %>%
  unlist %>%
  as.numeric


# Do sampling in one big step
e.wa <- matrix(sample(e.wa,
                      size=sample.size*simulation.size,
                      repl=T),
               ncol=sample.size)

# And again
w.wa <- matrix(sample(w.wa,
                      size=sample.size*simulation.size,
                      repl=T),
               ncol=sample.size)

# calc the means
results <- cbind(rowMeans(w.wa),
                 rowMeans(e.wa))

# And look for our needle in the haystack
sum(results[,1] < results[,2])
```

```
## [1] 52
```

```
rm(e.wa,w.wa,results)
```

So we got 52 cases where a sample of size 128 from the Eastern Washington Non-metro members had a larger progressivism score than a sample of 128 members from Western Washington Metro. Now maybe that tiny *p*-value doesn't seem that weird.

Categorical comparisons

Imagine now that we'd like to know how engagement varies across region. Here's the contingency table of the values.

	Not Engaged	Engaged	Highly Engaged
E WA Metro	53	52	76
E WA Non Metro	115	113	87
Thurston	171	245	309
W WA Metro	294	341	289
W WA Non Metro	100	88	88

With the ANOVA example, we had some notion of t-tests and then were interested how to expand that to a larger number of comparisons. With this contingency table, we don't really have a "smaller scale" analog of the test we want to run.

It looks like maybe Eastern Washington Metro is more highly engaged and Thurston seems *much* more highly engaged. But regression, and our pairwise hypothesis tests, don't do much for us here. We could theoretically calculate the percentage of one cateogry and then try to model that somehow (repeated tests of proportion?), but we're giving up a lot of power and we're ignoring the other categories.

As we'll see down below, the χ^2 test of association was designed for exactly this problem.

ANOVA

An ANOVA is designed for testing the equality of several means simultaneously. A single *quantitative* response variable is required with one or more *categorical* explanatory variables. So, this is the first thing to take away from ANOVA: > Use ANOVA when you have a continuous response and categorical > explanatory variables.

As I mention in the lecture, there is a *ton* of historical theory surrounding approaches to this problem. For instance, in our case of measuring progressivism across regions, we might propose a model like the following:

$$p_i = \mu + r_i + \epsilon_i$$

where p_i is the progressivism for person i, μ is the grand mean, r_i is the factor that describes region for the ith person and $\epsilon_i \sim N(0, \sigma^2)$ is the error term.

This may look a lot like a regression model–this is no accident. We used to worry a bunch about calculating sums of squares in a bunch of ways. You'd take the overall mean and calculate the errors from that. Then you'd calculate each regional mean and calculate the errors from those. This process yielded estimates for each region and a measure of the importance of the factors.

You live in a beautiful age. Now we just do ANOVA with regression. So our regional comparison can be done like this.

```
reg.lm <- lm(progressivism ~ region, data=d)
anova(reg.lm)
```

```
## Analysis of Variance Table
##
## Response: progressivism
##             Df  Sum Sq Mean Sq F value    Pr(>F)
## region       4   83.32 20.8308  22.698 < 2.2e-16 ***
## Residuals 2416 2217.22  0.9177
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

So we get our ANOVA table using `anova` (we could also use the function `aov` and give it the model formula directly). The interpretation of the table requires a bit of work, but here are the main parts.

1. `Df`: This column tells you how many degrees of freedom are being allocated to each part of the model. In this case there are 4 DFs being used by region (because the number of degrees of freedom used by a categorical variable in a linear model is $n - 1$ when you have n levels).
2. `Sum Sq`: this is the error associated with the level. Here we've got 83 "progressive units" for region and 2416 leftover in the residuals.
3. `Mean Sq`: This is the error explained divided by the degrees of freedom–higher values mean the variables are more predictive.
4. `F Value`: The F value is our test statistic. It's the ratio of the mean squared errors. Larger is better, geneally, and the p-value is dependent on the degrees of freedom and the value.
5. `Pr(>F)`: This is our p-value. Region is massively significant in explaining a piece of progressivism.

We can explore a more complicated model, perhaps controlling for engagement first.

```r
reg.lm.2 <- lm(progressivism ~ region + engagement, data=d)
anova(reg.lm.2)
```

```
## Analysis of Variance Table
##
## Response: progressivism
##              Df  Sum Sq Mean Sq F value    Pr(>F)
## region        4   83.32 20.8308  22.896 < 2.2e-16 ***
## engagement    2   21.00 10.4999  11.541 1.027e-05 ***
## Residuals  2414 2196.23  0.9098
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From this we see that `region` and `engagement` are both significant, with region having a larger impact. We could, now, add age and, since we think there might be an interaction effect between age and region, we'll add that too.

```r
reg.lm.3 <- lm(progressivism ~ region + engagement +
               age + region:age, data=d)
anova(reg.lm.3)
```

```
## Analysis of Variance Table
##
## Response: progressivism
##              Df  Sum Sq Mean Sq F value    Pr(>F)
## region        4   83.32 20.8308 22.8913 < 2.2e-16 ***
## engagement    2   21.00 10.4999 11.5386  1.03e-05 ***
## age           1    3.10  3.0955  3.4017   0.06525 .
## region:age    4    0.98  0.2440  0.2682   0.89855
## Residuals  2409 2192.15  0.9100
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

So age is moderately significant and the interaction term isn't. Let's finish with our final model.

```r
reg.lm.4 <- lm(progressivism ~ region + engagement + age,
               data=d)
anova(reg.lm.4)
```

```
## Analysis of Variance Table
##
## Response: progressivism
##              Df  Sum Sq Mean Sq F value    Pr(>F)
## region        4   83.32 20.8308 22.9191 < 2.2e-16 ***
## engagement    2   21.00 10.4999 11.5526 1.015e-05 ***
## age           1    3.10  3.0955  3.4059   0.06509 .
## Residuals  2413 2193.13  0.9089
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We could use `coef` or `summary` to extract the coefficients, but we've seen the essence of ANOVA. Region matters a lot, even controlling for engagement and age. When you hear about ANOVA, think regression with categorical explanatory variables. You can even throw a continuous parameter in there if you'd like to adjust for it.

χ^2 tests

When all your variables are categorical, we're a bit further away from our regression context. Imagine that someone has asked us to look into the association between region and engagement. If we look again at the cross-tabulation of the factors (called the contingency table) there seems like some association:

	Not Engaged	Engaged	Highly Engaged
E WA Metro	53	52	76
E WA Non Metro	115	113	87
Thurston	171	245	309
W WA Metro	294	341	289
W WA Non Metro	100	88	88

How can we test this association? Let's start with a simpler example, restricting the data to just Thurston and Western Washington Non Metro and removing the "Not Engaged" category.

	Engaged	Highly Engaged
Thurston	245	309
W WA Non Metro	88	88

So Western Washington Non Metro is split pretty evenly, but Thurston has a higher percentage (79%) of Highly Engaged. Note, also, that there are 730 total members in this smaller data set.

If we look at the whole table, we see that 75.9% of the members are in Thurston and, overall, 54.4% are "Highly Engaged". If there were no association between the columns, we might expect $0.759 \cdot 0.544 \cdot 730 = 301.3$ members would be in the cell for "Highly Engaged" *and* Thurston. Instead, we see 288, so we're "off" by 13.3 members.

We can perform this kind of calculation for every cell, resulting in the following table:

```r
mem.count <- nrow(small.d)
region.split <- table(small.d$region)/mem.count
engage.split <- table(small.d$engagement)/mem.count
expected.tbl <- outer(region.split,engage.split)*mem.count

knitr::kable(expected.tbl,caption="Expected values by region and
             engagement, assuming independence.")
```

Table 8: Expected values by region and engagement, assuming independence.

	Engaged	Highly Engaged
Thurston	252.71507	301.28493
W WA Non Metro	80.28493	95.71507

The essence of the χ^2 test hinges on this calculation, the assumption of independence and the calculation of the values we'd expect. These expectations are based on the *marginal* percentages of the categories: the percentages based on the column or row sums.

Before we dig deeper into the test, let's explore the relationship further by building a data set where we randomize the columns and build a table. Let's see what that looks like:

```r
set.seed(20171126)
fake.d <- data.frame(region=small.d$region,
                     engagement=sample(small.d$engagement))

knitr::kable(table(fake.d$engagement,fake.d$region),
             caption="A simulated table using randomization")
```

Table 9: A simulated table using randomization

	Thurston	W WA Non Metro
Engaged	252	81
Highly Engaged	302	95

Looking at this table, we see that the regions have a more similar percentage of high engagement (55% vs 54%). Looking at percentages based on the random allocation is the same as assuming that the two columns are *independent*. This means that knowing `region` doesn't tell you anything about `engagement` and vice versa. In this reduced data set, we have 554 Thurston members and 176 Western Washington Non Metro members. Since the fraction of Thurston members is 0.759 and we have 397 Highly Engaged members, we'd expect that the number of Highly Engaged members who also live in Thurston county would be 301.28. This is the same as the value we calculated above and is pretty close to the randomized value we calculated in our random experiment.

This concept, figuring out the *marginal* percentages and multiplying by the number of observations, is the heart of the χ^2 test. This test assumes independence, calculates the expected values in each cell, and then compares those to what we actually see. If we number the cells $1, \ldots, k$, then the χ^2 test starts by calculating

$$\sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

where O_i is the observed value in cell i and E_i is the expected value in that cell. This sum is normalized by the number of rows and columns. (To be precise, it is divided by $(r-1) \cdot (c-1)$ where r is the number of rows and c is the number of columns. This minus-one construct is done for the exact same reasons we have $l-1$ dummy variables for a categorical variable that has l levels.) This normalized sum has a theoretical distribution. In R we can just call the `chisq.test` function:

```r
chisq.test(x=small.d$engagement,y=small.d$region)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  small.d$engagement and small.d$region
## X-squared = 1.5711, df = 1, p-value = 0.2101
```

This p-value is not significant, so we don't have evidence that the association between these factors violates our assumption of independence.

(Up above we linked to the Wikipedia page on the χ^2 test. Note that this page has *two* major tests with the same name. The first measures fit between a theoretical distribution and a set of data. The second is our test of association. The mechanics of the test are pretty similar–they're based on that observed versus expected technique. It's worth calling attention to just so you know about this confusing name collision.)

Simulating a Simple χ^2 Test

To reinforce the ideas of this, let's simulate a χ^2 test using resampling techniques. The code below goes through the following steps:

1. Determine a statistic that measures the association present in the table.
2. Calculate that statistic on the observed data.
3. Create a container to hold the simulation results.
4. Simulate 1000 replications of the data set and compute the test statistic.

5. Plot the statistics.

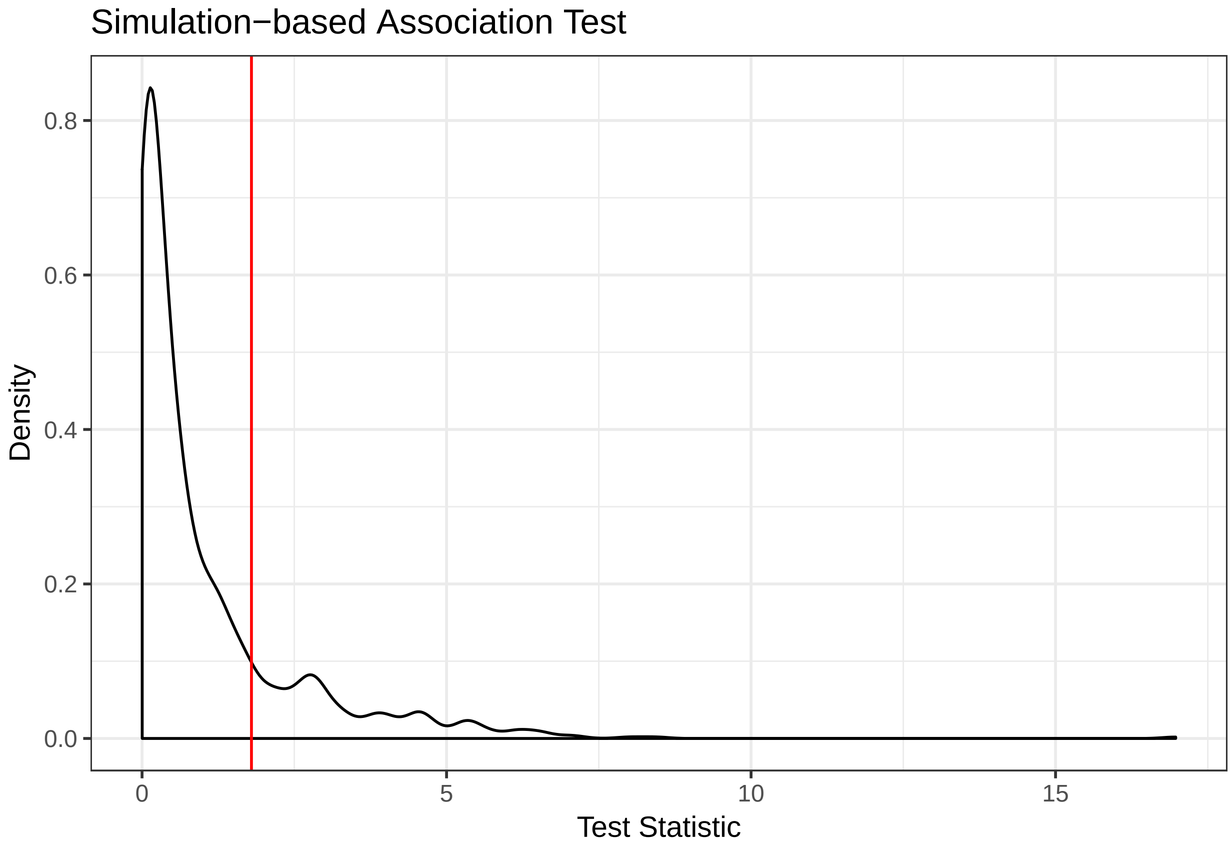
For the statistic that we'll use to summarize how much the table differs from independence, we can follow the example of the χ^2 test and use the sum of the deviations from independence. First, let's calculate the value for our actual data.

```
observed <- as.vector(table(small.d$region,small.d$engagement))
expected <- as.vector(expected.tbl)
test.stat <- sum((observed-expected)^2/expected)
print(test.stat)
```

```
## [1] 1.79635
```

The test statistic is about 1.8—now we will run a simulation to see how extreme that value is.

```
get.stat <- function(tbl,expected) {
  tbl.vec <- as.vector(tbl)
  val <- sum((tbl.vec-expected)^2/expected)

  return(val)
}

results <- data.frame(val=c(test.stat,rep(0,999)))

for (i in 2:nrow(results)){
  fake.d <- data.frame(region=small.d$region,
                     engagement=sample(small.d$engagement))

  results$val[i] <- get.stat(table(fake.d$region,fake.d$engagement),
                             expected.tbl)
}

ggplot(results,
       aes(x=val)) +
  theme_bw() +
  geom_density() +
  geom_vline(aes(xintercept=test.stat),col="red") +
  labs(x="Test Statistic",
       y="Density",
       title="Simulation-based Association Test")
```



```
print(mean(abs(results$val) >= abs(test.stat)))
```

```
## [1] 0.19
```

When we simulate a large number of data sets that obey the assumption of independence, we see that most values fall between 0 and 2.5, though values as large as 15 are possible. Our observed value, 1.8, is larger than about 81% of the simulated values. For the classic, and over-relied-upon, significance of $p < 0.05$, we would need our observed value to be about 3.85. Instead, this p-value, 0.19, is pretty similar to what we get by running `chisq.test`.

A More Complicated Model

Our original question was more complicated than two levels of engagement across two regions. We wanted to know more generally if this table seems to be the result of chance allocation:

	Not Engaged	Engaged	Highly Engaged
E WA Metro	53	52	76
E WA Non Metro	115	113	87
Thurston	171	245	309
W WA Metro	294	341	289
W WA Non Metro	100	88	88

With raw counts it is somewhat difficult to see what's going on. We might want to see the percentage of people in each region that fall into each engagement level. There does seem to be some evidence that Thurston, which has growing numbers as we go up the engagement hierarchy, might be different from the other regions, which are more flat. Eastern Washington Non Metro in particular has a lower fraction of engaged members.

We can test this with a full χ^2 test.

```
chisq.test(x=d$engagement,y=d$region)
```

```
##
##  Pearson's Chi-squared test
##
## data:  d$engagement and d$region
## X-squared = 46.695, df = 8, p-value = 1.753e-07
```

The results are not very descriptive, but the p-value is highly significant. In short, we can state with confidence that engagement varies by region.

Significant p-values are typically the beginning, rather than end, of an association analysis. One natural next step is to look at the residuals from our model. These are the departures from expectation on a cell-by-cell basis:

```
knitr::kable(residuals(chisq.test(x=d$engagement,y=d$region)),
             caption="Residuals for full Region/Engagement Test")
```

Table 11: Residuals for full Region/Engagement Test

	E WA Metro	E WA Non Metro	Thurston	W WA Metro	W WA Non Metro
Not Engaged	-0.2432752	2.0098876	-3.273976	0.851544	1.7980089
Engaged	-1.3542664	0.3671881	-0.394268	1.161634	-0.7820135
Highly Engaged	1.5723127	-2.2325592	3.434040	-1.946007	-0.8932730

We look for values that are large in absolute value. For instance, Thurston shows many more highly-engaged members than we'd expect (and fewer who are not engaged). Western Washington Metro is actually lower on highly-engaged than we'd expect, as is Eastern Washington Non Metro.

Appendix: Full Data Description

A financial institution in Washington has become concerned that their current membership base is not well-aligned with their corporate values. Through that concern they realized that don't actually understand their membership's values very well. They surveyed 2,421 members to shed light on the issue.

The heart of the survey was the Moral Foundations Theory of Jonathan Haidt. Members were surveyed on the Moral Foundations Questionnaire, which you should take so you understand the test. Survey respondents were scored on the five foundations as well as a single-number summary, Progressivism.

The financial institution values Localism, Sustainability, and Education. These aspects of member's values were assessed in the survey as well. Localism and Sustainability used validated scales and thus can be summarized via a single score, where higher values indicate greater support for the values. Education is summarized by the following three questions, which we do not have evidence can be combined into a single score:

- In general, public schools provide a better education than private schools.
- Public school teachers are underpaid.
- Experience is more important than education in determining success in life. These questions were evaluated on a 1 to 6 scale where 1 indicated "Strongly Disagree" and 6 indicated "Strongly Agree".

Finally, we have information on the member that can be used to understand variation in their values.

The data consists of the following columns:

- ID: a unique identifier for the survey respondent.
- age: the age of the respondent.
- gender: gender was evaluated with robust scale and collapsed into male/female/other for those whose gender identity was not male or female.
- engagement: three categories of engagement with the financial institution.
- mem.edu: the self-reported education level of the member with the following scale:
- zip: the member zip code.
- channel: how the member joined the financial institution. Options are “Loan” if they joined via an auto loan, “Branch” if they joined at a branch and other for online or unknown.
- progressivism/harm/fair/in.group/authority/purity: The MFQ results.
- account.age: the age of the member’s account, in years.
- region: The region of Washington the member lives in. May be easier to work with than zip.
- public.sector: has the person ever been a public employee?
- sustainability/localism: Scores on the validated scales. Higher values indicate greater support for the value.
- pub.greater.priv/experience.more.important/teachers.underpaid: The responses to the education questions above.
- main.focal.value: Respondents were asked, “Below is a list of broad areas to which people often dedicate their volunteer or philanthropic efforts. From this list, please select the most important to you. If an area of particular importance is missing, please let us know about it in the space for ‘other.’ ” This column holds the respondents’ answer to that question.
- support.of.focal.value: Respondents were given an opportunity to indicate how they supported their focal value. Those responses were collapsed into a single score, where a higher value indicates more support.