Chapter 5
Basic Inferences and Bivariate Association

In this chapter, we begin to use *inferential statistics* and *bivariate statistics*. In Chap. 4, we were content simply to characterize the properties of a single variable in the sample at hand. Usually in Political Science, though, our motivation as researchers will be to argue whether a claim can be generalized. Hence, inferential statistics are designed to draw inferences about a broader population. Further, we frequently want to measure the level of association between variables, and bivariate statistics serve as measures of the degree to which two variables are associated with each other.

In later chapters, linear regression models, generalized linear models, time series models, and other models that we estimate offer the opportunity to draw an inference about a broader population. They also allow us to evaluate bivariate or multivariate relationships among variables. For now, we focus on a handful of stepping stone inferential and bivariate statistics: Tests about means, associations between two categorical variables (via cross-tabulations), and correlations between two continuous variables.

In this chapter, our working example data will be the same data from LaLonde's (1986) analysis of the National Supported Work Demonstration. Information about the features of these data can be reviewed at the start of Chap. 4. In this case, every member of our sample is someone who was long-term unemployed. Hence, as we draw inferences, it would not be fair to try to conclude something about the entire population of the USA in the mid-1970s because these data do not compose a sample of that population. For the purposes of our working example, we will attempt to draw conclusions about the population of long-term unemployed persons in the USA. More information on how this sample was drawn is available in LaLonde (1986),

Electronic supplementary material: The online version of this chapter (doi: 10.1007/978-3-319-23446-5_5) contains supplementary material, which is available to authorized users.

J.E. Monogan III, *Political Analysis Using R*, Use R!,
DOI 10.1007/978-3-319-23446-5_5

and the reader is urged to read about *random sampling* if more information is desired about the theory of statistical inference.

5.1 Significance Tests for Means

Before computing any inferential statistics, we must load LaLonde's data once again. Users who have already installed the `cem` library can simply type:

```
library(cem)
data(LL)
```

Users who did not install `cem` in Chap. 4 will need to type `install.packages("cem")` before the two preceding lines of code will work properly. Once these data are loaded in memory, again by the name `LL`, we can turn to applied analysis[1].

We begin by testing hypotheses about the mean of a population (or multiple populations). We first consider the case where we want to test whether the mean of some population of interest differs from some value of interest. To conduct this significance test, we need: (1) our estimate of the sample mean, (2) the standard error of our mean estimate, and (3) a null and alternative hypothesis. The sample mean is defined earlier in Eq. (4.1), and the standard error of our estimate is simply the standard deviation of the variable [defined in Eq. (4.5)] divided by the square root of our sample size, $s_x/\sqrt{n}$.

When defining our null and alternative hypotheses, we define the null hypothesis based on some value of interest that we would like to rule out as a possible value of the population parameter. Hence, if we say:

$$H_0: \quad \mu = \mu_0$$

This means that our null hypothesis (H_0) is that the population mean (μ) is equal to some numeric value we set (μ_0). Our *research hypothesis* is the alternative hypothesis we would like to reject this null in favor of. We have three choices for potential research hypotheses:

$$H_A: \quad \mu > \mu_0$$
$$H_A: \quad \mu < \mu_0$$
$$H_A: \quad \mu \neq \mu_0$$

The first two are called one-tailed tests and indicate that we believe the population mean should be, respectively, greater than or less than the proposed value μ_0. Most research hypotheses should be considered as one of the one-tailed tests, though occasionally the analyst does not have a strong expectation on whether the mean

[1] As before, these data also are available in comma-separated format in the file named `LL.csv`. This data file can be downloaded from the Dataverse on page vii or the chapter content link on page 63.

should be bigger or smaller. The third alternative listed defines the two-tailed test, which asks whether the mean is simply different from (or not equal to) the value μ_0.

Once we have formulated our hypothesis, we compute a t-ratio as our *test statistic* for the hypothesis. Our test statistic includes the sample mean, standard error, and the population mean defined by the null hypothesis (μ_0). This formula is:

$$t = \frac{\bar{x} - \mu_0}{\text{SE}(\bar{x}|H_0)} = \frac{\bar{x} - \mu_0}{s_x/\sqrt{n}} \tag{5.1}$$

This is distributed Student's-t with $n - 1$ degrees of freedom under the null (and asymptotically normal).[2] Once we have this test statistic, we compute our *p-value* as follows:

$$p - \text{value} = \begin{cases} P(t^* \leq t|H_0) & H_A : \mu < \mu_0 \\ P(t^* \geq t|H_0) & H_A : \mu > \mu_0 \\ P(|t^* - \mu_0| \geq |t - \mu_0||H_0) & H_A : \mu \neq \mu_0 \end{cases}$$

In this case, assume t^* is the actual value of our statistic that we compute. The typical action in this case is to have a pre-defined *confidence level* and decide whether to reject the null hypothesis or not based on whether the p-value indicates that rejection can be done with that level of confidence. For instance, if an analyst was willing to reject a null hypothesis if he or she could do so with 90 % confidence, then if $p < 0.10$, he or she would reject the null and conclude that the research hypothesis is correct. Many users also proceed to report the p-value so that readers can draw conclusions about significance themselves.

R makes all of these calculations very straightforward, doing all of this in a single line of user code. Suppose that we had a hypothesis that, in 1974, the population of long-term unemployed Americans had a lower income than $6,059, a government estimate of the mean income for the overall population of Americans. In this case, our hypothesis is:

$$\begin{aligned} H_0&: \quad \mu = 6059 \\ H_A&: \quad \mu < 6059 \end{aligned}$$

This is a one-tailed test because we do not even entertain the idea that the long-term unemployed could have an on-average higher income than the general population. Rather, we simply ask whether the mean of our population of interest is discernibly lower than $6,059 or not. To test this hypothesis in R, we type:

```
t.test(LL$re74, mu=6059, alternative="less")
```

[2]This statistic has a t distribution because the sample mean has a normally distributed *sampling distribution* and the sample standard error has a χ^2 sampling distribution with $n - 1$ degrees of freedom. The ratio of these two distributions yields a t distribution.

The first argument of the `t.test` lists our variable of interest, `LL$re74`, for which R automatically computes the sample mean and standard error. Second, the `mu=6059` argument lists the value of interest from our null hypothesis. Be sure to include this argument: If you forget, the command will still run assuming you want `mu=0`, which is silly in this case. Finally, we specify our alternative hypothesis as `"less"`. This means we believe the population mean to be less than the null quantity presented. The result of this command prints as:

```
One Sample t-test

data:  LL$re74
t = -10.4889, df = 721, p-value < 2.2e-16
alternative hypothesis: true mean is less than 6059
95 percent confidence interval:
     -Inf 4012.025
sample estimates:
mean of x
 3630.738
```

This presents a long list of information: At the end, it reports the sample mean of 3630.738. Earlier, it shows us the value of our t-ratio is -10.4889, along with the fact that our t distribution has 721 degrees of freedom. As for the p-value, when R prints `p-value < 2.2e-16`, this means that p is so minuscule that it is smaller than R's level of decimal precision, much less any common significance threshold. Hence, we can reject the null hypothesis and conclude that long-term unemployed Americans had a significantly lower income than $6,059 in 1974.

5.1.1 Two-Sample Difference of Means Test, Independent Samples

As an alternative to using one sample to draw an inference about the relevant population mean, we may have *two* samples and want to test whether the two populations means are equal. In this case, if we called the set of observations from one sample x and the observations from the second sample y, then we would formulate our null hypothesis as:

$$H_0: \quad \mu_x = \mu_y$$

Again, we will pair this with one of the three alternative hypotheses:

$$H_A: \quad \mu_x < \mu_y$$

$$H_A: \quad \mu_x > \mu_y$$

$$H_A: \quad \mu_x \neq \mu_y$$

Again, the first two possible alternative hypotheses are one-tailed tests where we have a clear expectation as to which population's mean should be bigger. The third possible alternative simply evaluates whether the means are different. When building our test statistic from this null hypothesis, we rely on the fact that H_0 also implies $\mu_x - \mu_y = 0$. Using this fact, we construct our t-ratio as:

$$t = \frac{(\bar{x} - \bar{y}) - (\mu_x - \mu_y)}{\text{SE}(\bar{x} - \bar{y}|H_0)} \tag{5.2}$$

The last question is how we calculate the standard error. Our calculation depends on whether we are willing to assume that the variance is the same in each population. Under the assumption of unequal variance, we compute the standard error as:

$$\text{SE}(\bar{x} - \bar{y}|H_0) = \sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}} \tag{5.3}$$

Under the assumption of equal variance, we have:

$$\text{SE}(\bar{x} - \bar{y}|H_0) = \sqrt{\frac{(n_x - 1)s_x^2 + (n_y - 1)s_y^2}{n_x + n_y - 2}} \sqrt{\frac{1}{n_x} + \frac{1}{n_y}} \tag{5.4}$$

As an example, we can conduct a test with the last observation of income in the National Supported Work Demonstration, which was measured in 1978. Suppose our hypothesis is that income in 1978 was higher among individuals who received the treatment of participating in the program (y) than it was among those who were control observations and did not get to participate in the program (x). Our hypothesis in this case is:

$$\begin{aligned} H_0&: \quad \mu_x = \mu_y \\ H_A&: \quad \mu_x < \mu_y \end{aligned}$$

Again, this is a one-tailed test because we are not entertaining the idea that the treatment could have reduced long-term income. Rather, the treatment either increased income relative to the control observations, or it had no discernible effect. R allows us to conduct this two-sample t-test using either assumption. The commands for unequal and equal variances, respectively, are:

```
t.test(re78~treated,data=LL,alternative="less",var.equal=F)
t.test(re78~treated,data=LL,alternative="less",var.equal=T)
```

The first argument, `re78~treated`, is in functional notation and indicates that income in 1978 is being separated based on values of the treatment indicator. The `data` option allows us to name the dataset so that we do not have to call it for each variable. When we state `alternative="less"`, we are declaring our alternative hypothesis to mean that the average income for the lower value of

treated (group 0, the control) should be lower than the average for the higher value of **treated** (group 1, the treated group). The only difference between the commands is that the first sets `var.equal=F` so that variances are assumed unequal, and the second sets `var.equal=T` so that variances are assumed equal.

The results print as follows. For the assumption of unequal variances, we see:

```
Welch Two Sample t-test

data:  re78 by treated
t = -1.8154, df = 557.062, p-value = 0.035
alternative hypothesis: true difference in means is less
   than 0
95 percent confidence interval:
      -Inf -81.94117
sample estimates:
mean in group 0 mean in group 1
       5090.048        5976.352
```

Meanwhile for the equal variance assumption, we see:

```
Two Sample t-test

data:  re78 by treated
t = -1.8774, df = 720, p-value = 0.03043
alternative hypothesis: true difference in means is less
  than 0
95 percent confidence interval:
      -Inf -108.7906
sample estimates:
mean in group 0 mean in group 1
       5090.048        5976.352
```

The results are pretty similar, as would be expected. Both report the same estimates of the means for the control group (5090.048) and treated group (5976.352). However, because the standard error is computed differently, we get slightly different t values in each case (and a differently calculated degrees of freedom for the unequal variances). The p-value is slightly larger when assuming unequal variances, but in this case either option yields a similar conclusion. Hence, in either case, we reject the null hypothesis at the 95 % confidence level and conclude that for the treated group income was higher in 1978 than for the control group. It should be noted that one limitation of a test like this is that we have not controlled for any of the other variables known to influence income, and the treatment assignment was not random in this case. Chapters 6–8 offer several examples of methods designed to control statistically for other predictors. Specifically, in Sect. 8.3 we revisit this exact example with a more advanced technique.

5.1.2 Comparing Means with Dependent Samples

A third mean-related test statistic we may want to compute is a difference of means with a dependent sample (e.g., comparing matched samples). In other words, suppose we have a situation in which each observation in sample 1 matches with an observation in sample 2. This could mean we are studying the same person before and after an event, a person doing the same task with different treatments, using a twin study, or using matching methods to pair treated observations to control observations. In this case, we should no longer treat each sample as independent, but compute differences for each pairing and analyze the differences. An easy way to think of this would be to create a new variable, $w_i = x_i - y_i$ where x and y are matched for each case i. In this case, our null hypothesis is $H_0 : \mu_w = 0$, and our alternative hypothesis can be any of the three choices:

$$H_A: \quad \mu_w < 0$$
$$H_A: \quad \mu_w > 0$$
$$H_A: \quad \mu_w \neq 0$$

The test statistic in this case is given by Eq. (5.5), computed for the new variable w.

$$t = \frac{\bar{w} - 0}{\text{SE}(\bar{w}|H_0)} = \frac{\bar{w}}{s_w/\sqrt{n}} \tag{5.5}$$

As can be seen, this is effectively the same test statistic as in Eq. (5.1) with w as the variable of interest and 0 as the null value. The user technically could create the w variable him or herself and then simply apply the code for a single-sample significance test for a mean.

More quickly, though, this procedure could be automated by inserting two separate variables for the linked observations into the t-test command. Suppose, for instance, that we wanted to know if our control observations saw a rise in their income from 1974 to 1978. It is possible that wages may not increase over this time because these numbers are recorded in real terms. However, if wages did increase, then observing how they changed for the control group can serve as a good baseline for comparing change in the treated group's wages in the same time frame. To conduct this paired sample t-test for our control observations, we type:

```
LL.0<-subset(LL,treated==0)
t.test(LL.0$re74,LL.0$re78,paired=T,alternative="less")
```

In the first line we create a subset only of our control observations. In the second line, our first argument is the measure of income in 1974, and the second is income in 1978. Third, we specify the option `paired=T`: This is *critical*, otherwise R will assume each variable forms an independent sample, but in our case this is a paired sample where each individual has been observed twice. (To this end, by typing `paired=F` instead, this gives us the syntax for a two-sample t-test if our separate

samples are in differing columns of data.) Finally, `alternative="less"` means that we expect the mean of the first observation, in 1974, to be lower than the mean of the second, in 1978. Our results are:

```
Paired t-test

data:  LL.0$re74 and LL.0$re78
t = -3.8458, df = 424, p-value = 6.93e-05
alternative hypothesis: true difference in means is
less than 0
95 percent confidence interval:
     -Inf -809.946
sample estimates:
mean of the differences
              -1417.563
```

This output tells us that earnings were on average $1,417.56 lower in 1974 than in 1978. Our t-ratio is $t = -3.8458$, and the corresponding p-value is $p = 0.00007$. Hence at any common confidence threshold, we can reject the null hypothesis and conclude that incomes were higher in 1978 than in 1974 among long-term unemployed who did not receive the treatment.

Just as a final comparison, we could compute the same sort of test on the treated group as follows:

```
LL.1<-subset(LL,treated==1)
t.test(LL.1$re74,LL.1$re78,paired=T,alternative="less")
```

The results are somewhat similar:

```
Paired t-test

data:  LL.1$re74 and LL.1$re78
t = -4.7241, df = 296, p-value = 1.788e-06
alternative hypothesis: true difference in means is
less than 0
95 percent confidence interval:
      -Inf -1565.224
sample estimates:
mean of the differences
              -2405.353
```

We observe a bigger growth of $2,405.35 among the treated observations, and this result is also statistically discernible. This larger growth rate is encouraging for the long-term earnings potential of the program. For bonus points, the reader is encouraged to look up difference-in-differences techniques and consider how they might apply to studies with a design such as this.

5.2 Cross-Tabulations

In situations where we want to analyze the association between two nominal or ordinal variables, a cross-tabulation is often a good tool for inference. A cross tabulation tests the hypothesis that an independent categorical variable affects the conditional distribution of a dependent categorical variable. The researcher asks: Will certain values of the dependent variable be noticeably more or less frequent when moving from one category of an independent variable to another? In assessing this effect, the analyst should always break down relative percentages of categories of the dependent variable within categories of the independent variable. Many people make a mistake by breaking down percentages within categories of the dependent variable; such a mistake prevents a researcher from substantively evaluating the stated hypothesis that the independent variable causes the dependent variable and not the reverse. The results of a cross-tabulation substantively compare the percentages of the same value of the dependent variable across categories of the independent variable.

Some common mistakes to avoid: First, once again, avoid breaking down percentages by the dependent variable. Second, avoid comparing the largest percentage in each category of an independent variable. The hypothesis states that the frequency of the dependent variable will vary by value of the independent variable; it does not argue what value of the dependent variable will be most frequent. Lastly, avoid drawing inferences based on the pure magnitude of percentages; the researcher's task is to look at differences in the distribution. For example, if vote choice is the dependent variable, and 66 % of Republicans support the Democratic candidate, while 94 % of Democrats support the Democratic candidate, the researcher should not focus on majority support from both parties. Instead, the researcher should observe that a 28 percentage point difference implies that partisanship has an important effect on how individuals vote.

Consider two examples from the LaLonde dataset. First, we can simply ask whether being unemployed in 1974 (`u74`) served as a good predictor of being unemployed in 1975 (`u75`). We would have to think that an individual's prior work status shapes current work status. To build a cross-tabulation in R we need to install the `gmodels` package and then load the library. Once we have done this, we can use the `CrossTable` function:

```
install.packages("gmodels")
library(gmodels)
CrossTable(y=LL$u75,x=LL$u74,prop.c=F,prop.t=F,
     prop.chisq=F,chisq=T,format="SPSS")
```

In this code, `y` specifies the column variable, and `x` specifies the row variable. This means our dependent variable makes up the columns and the independent makes up the rows. Because we want the conditional distribution of the dependent variable for each given value of the independent variable, the options `prop.c`, `prop.t`, and `prop.chisq` are all set to `FALSE` (referring to **prop**ortion of the **c**olumn, **t**otal sample, and contribution to the **chisq**uare statistic). This means that each cell

only contains the raw frequency and the row-percentage, which corresponds to the distribution conditional on the independent variable. The option `chisq=T` reports Pearson's Chi-squared (χ^2) test. Under this test, the null hypothesis is that the two variables are independent of each other. The alternative hypothesis is that knowing the value of one variable changes the expected distribution of the other.[3] By setting the `format` option to `SPSS`, rather than `SAS`, we are presented with percentages in our cells, rather than proportions.

The results of this command are printed below:

```
   Cell Contents
|-------------------------|
|                   Count |
|             Row Percent |
|-------------------------|

Total Observations in Table:  722

             | LL$u75
      LL$u74 |         0  |         1  | Row Total  |
-------------|------------|------------|------------|
           0 |       386  |         9  |       395  |
             |   97.722%  |    2.278%  |   54.709%  |
-------------|------------|------------|------------|
           1 |        47  |       280  |       327  |
             |   14.373%  |   85.627%  |   45.291%  |
-------------|------------|------------|------------|
Column Total |       433  |       289  |       722  |
-------------|------------|------------|------------|


Statistics for All Table Factors


Pearson's Chi-squared test
------------------------------------------------------------
Chi^2 =  517.7155     d.f. =  1     p =  1.329138e-114

Pearson's Chi-squared test with Yates' continuity
correction
------------------------------------------------------------
Chi^2 =  514.2493     d.f. =  1     p =  7.545799e-114


        Minimum expected frequency: 130.8906
```

[3] Note that this is a symmetric test of association. The test itself has no notion of which is the dependent or independent variable.

As we can see, among those who were employed in 1974 (**u74**=0), 97.7 % were employed in 1975. Among those who were unemployed in 1974 (**u75**=1), 14.4 % were employed in 1975.[4] This corresponds to an 83.3 percentage point difference between the categories. This vast effect indicates that employment status in 1 year does, in fact, beget employment status in the following year. Further, our test statistic is $\chi^2_{1\text{df}} = 517.7155$ with a minuscule corresponding p-value. Hence, we reject the null hypothesis that employment status in 1974 is independent from employment status in 1975 and conclude that employment status in 1974 conditions the distribution of employment status in 1975.

As a more interesting question, we might ask whether receiving the treatment from the National Supported Work Demonstration shapes employment status in 1975. We would test this hypothesis with the code:

```
CrossTable(y=LL$u75,x=LL$treated,prop.c=F,prop.t=F,
    prop.chisq=F,chisq=T,format="SPSS")
```

The output from this command is:

```
   Cell Contents
|-------------------------|
|                   Count |
|               Row Percent |
|-------------------------|

Total Observations in Table:  722

              | LL$u75
  LL$treated |         0  |         1  | Row Total  |
-------------|------------|------------|------------|
           0 |       247  |       178  |       425  |
             |   58.118%  |   41.882%  |   58.864%  |
-------------|------------|------------|------------|
           1 |       186  |       111  |       297  |
             |   62.626%  |   37.374%  |   41.136%  |
-------------|------------|------------|------------|
Column Total |       433  |       289  |       722  |
-------------|------------|------------|------------|


Statistics for All Table Factors
```

[4]To get more meaningful levels than 0 and 1 in this case, we would need to create copies of the variables **u74** and **u75** that recorded each value as text (e.g., "Unemployed" and "Employed"). The `recode` command from the `car` library offers a straightforward way of doing this, if desired.

```
Pearson's Chi-squared test
-----------------------------------------------------
Chi^2 =  1.480414     d.f. =  1     p =  0.2237097

Pearson's Chi-squared test with Yates' continuity
correction
-----------------------------------------------------
Chi^2 =  1.298555     d.f. =  1     p =  0.2544773


        Minimum expected frequency: 118.8823
```

Substantively, the effects are in the expected direction. Among control observations (**treated**=0), 58.1 % were employed in 1975. Among treated observations (**treated**=1), 62.6 % were employed in 1975. Hence we see a 4.5 percentage point bump in employment among treated observations over control observations. However, our test statistic is $\chi^2_{1\text{df}} = 1.4804$. The corresponding p-value is $p = 0.2237$. Hence, if we set our confidence threshold at 90 % or anything higher, we would fail to reject the null hypothesis and conclude that there was no discernible relationship between the treatment and employment status.

5.3 Correlation Coefficients

As a preview to the next chapter, we conclude our look at bivariate statistics by showing how to calculate a correlation coefficient in R. Correlation coefficients are calculated as a measure of association between two continuous variables. We specifically focus on Pearson's r, the correlation coefficient for a linear relationship. This value shows how well the independent variable linearly predicts the dependent variable. This measure will range between -1 and 1. A correlation coefficient of 0 suggests the absence of any linear relationship between the two variables. (Though, importantly, a nonlinear relationship also might yield $r = 0$ and some erroneous conclusions.) A value of 1 would imply a perfect positive relationship, and a value of -1 would indicate a perfect negative relationship. The square of a Pearson's r (r^2) calculates the amount of variance explained by the predictor.

The formula for a Pearson correlation coefficient is essentially the covariance of two variables, x and y, divided by the standard deviation of each variable:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})}} \tag{5.6}$$

Within R, this quantity is computed with the `cor` command.[5]

Suppose we wanted to evaluate whether the number of years of education served as a good predictor of our first measure of income, in 1974. We could type:

```
cor(LL$education,LL$re74)
cor(LL$education,LL$re74)^2
```

The first line computes the actual correlation coefficient itself. R returns a printout of: `[1] 0.08916458`. Hence, our correlation coefficient is $r = 0.0892$. The second line recalculates the correlation and squares the result for us all at once. This tells us that $r^2 = 0.0080$. The implication of this finding is that by knowing a respondent's number of years of education, we could explain 0.8 % of the variance in 1974 income. On its face, this seems somewhat weak, but as a general word of advice always gauge r^2 (or multiple R^2, in the next chapter) values by comparing them with other findings in the same area. Some sorts of models will routinely explain 90 % of the variance, while others do well to explain 5 % of the variance.

As a final example, we can consider the idea that income begets income. Consider how well income in 1975 correlates with income in 1978. We compute this by typing:

```
cor(LL$re75,LL$re78)
cor(LL$re75,LL$re78)^2
```

The first line returns the correlation coefficient between these two variables, printing: `[1] 0.1548982`. Our estimate of $r = 0.1549$ indicates that high values of income in 1975 do generally correspond to high values of income in 1978. In this case, the second line returns $r^2 = 0.0240$. This means we can explain 2.4 % of the variance of income in 1978 by knowing what someone earned in 1975.

Remember that the graphing tools from Chap. 3 can help us understand our data, including any results that we quantify such as correlation coefficients. If we are wondering why earlier income does not do a better job of predicting later income, we could draw a scatterplot as follows:

```
plot(x=LL$re75,y=LL$re78,xlab="1975 Income",ylab="1978 Income",
     asp=1,xlim=c(0,60000),ylim=c(0,60000),pch=".")
```

Notice that we have used the `asp=1` option to set the **asp**ect ratio of the two axes at 1. This guarantees that the scale of the two axes is held to be the same—which is appropriate since both variables in the figure are measured in inflation-adjusted dollars. The output is reported in Fig. 5.1. As can be seen, many of the observations cluster at zero in one or both of the years, so there is a limited degree to which a linear relationship characterizes these data.

We now have several basic inferences in hand: t-tests on means and χ^2 tests for cross-tabulations. Difference in means tests, cross-tabulations, and correlation

[5]The `cor` command also provides a `method` option for which available arguments are `pearson`, `kendall` (which computes Kendall's τ, a rank correlation), and `spearman` (which computes Spearman's ρ, another rank correlation). Users are encouraged to read about the alternate methods before using them. Here, we focus on the default Pearson method.

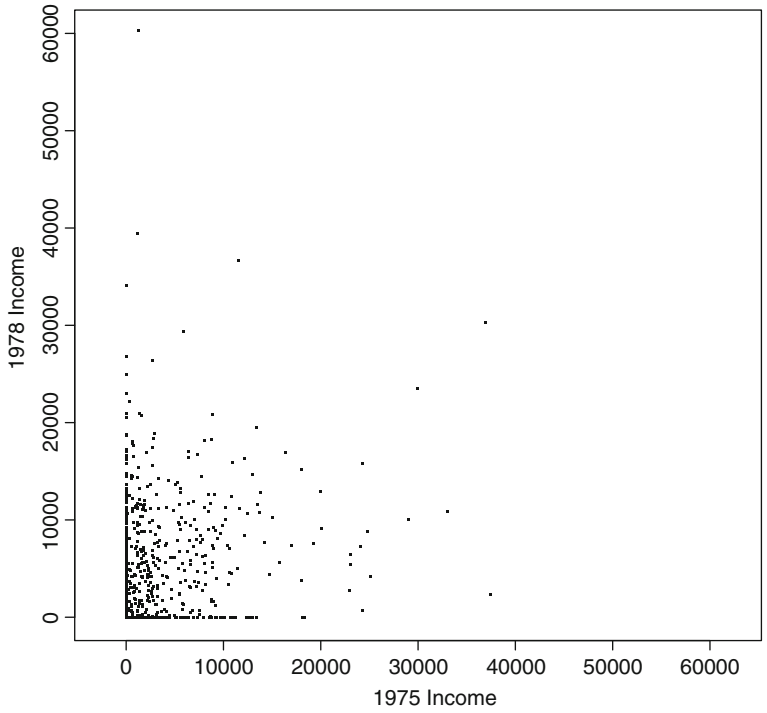

Fig. 5.1 Scatterplot of income in 1975 and 1978 from National Supported Work Demonstration data

coefficients have also given us a good sense of evaluating bivariate relationships. In the next chapter, we turn to multivariate statistics, specifically using linear regression methods. This will build on the linear techniques that correlation coefficients use and allow us to introduce the concept of statistical control.

5.4 Practice Problems

Please load the `foreign` library and download Alvarez et al.'s (2013) data, which are saved in the Stata-formatted file `alpl2013.dta`. This file is available from the Dataverse named on page vii or the chapter content named on page 63. These data are from a field experiment in Salta, Argentina in which some voters cast ballots through e-voting, and others voted in the traditional setting. The variables are: an indictor for whether the voter used e-voting or traditional voting (**EV**), age group (**age_group**), education (**educ**), white collar worker (**white_collar**), not a full time worker (**not_full_time**), male (**male**), a count variable for number of six possible technological devices used (**tech**), an ordinal scale for political knowledge

(**pol_info**), a character vector naming the polling place (**polling_place**), whether the respondent thinks poll workers are qualified (**capable_auth**), whether the voter evaluated the voting experience positively (**eval_voting**), whether the voter evaluated the speed of voting as quick (**speed**), whether the voter is sure his or her vote is being counted (**sure_counted**), whether the voter thought voting was easy (**easy_voting**), whether the voter is confident in ballot secrecy (**conf_secret**), whether the voter thinks Salta's elections are clean (**how_clean**), whether the voter thinks e-voting should replace traditional voting (**agree_evoting**), and whether the voter prefers selecting candidates from different parties electronically (**eselect_cand**).

1. Consider the number of technological devices. Test the hypothesis that the average Salta voter has used more than three of these six devices. (Formally: $H_0 : \mu = 3; H_A : \mu > 3$.)
2. Conduct two independent sample difference of means tests:
   a. Is there any difference between men and women in how many technological devices they have used?
   b. Is there any difference in how positively voters view the voting experience (**eval_voting**) based on whether they used e-voting or traditional voting (**EV**)?
3. Construct two cross-tabulations:
   a. Construct a cross-tabulation where the dependent variable is how positively voters view the voting experience (**eval_voting**) and the independent variable is whether they used e-voting or traditional voting (**EV**). Does the distribution of voting evaluation depend on whether the voter used e-voting? This cross-tabulation addressed the same question as is raised in #2.b. Which approach is more appropriate here?
   b. Construct a cross-tabulation where the dependent variable is how positively voters view the voting experience (**eval_voting**) and the independent variable is the ordinal scale of political knowledge (**pol_info**). Does the distribution of voting evaluation change with the voter's level of political knowledge?
4. Consider the correlation between level of education (**educ**) and political knowledge (**pol_info**):
   a. Compute Pearson's r between these two variables.
   b. Many argue that, with two ordinal variables, a more appropriate correlation measure is Spearman's ρ, which is a rank correlation. Compute ρ and contrast the results from r.