Introduction to Statistics

Preliminary Lecture Notes

Adolfo J. Rumbos

November 28, 2008

Contents

Chapter 1

Preface

Statistics may be defined as the art of making decisions based on incomplete information or data. The process of making those decisions is known as *statistical inference.* The emphasis in this course will be on statistical inference; in particular, *estimation* and *hypothesis testing.* There are other aspects surrounding this process, such as sampling and exploratory data analysis, which we will also touch upon in this course. We will also emphasize the acquisition of statistical reasoning skills. Thus, the main thrust of the course will not be the mere application of formulae to bodies of data, but the reasoning processes that accompany the involvement of statistics at the various stages of real world statistical applications: from the formulation of questions and hypotheses, experimental design, sampling and data collection, data analysis, to the presentation of results. We will use a combination of lectures, discussions, activities and problem sets which expose the students to the various issues that arise in statistical investigations.

Chapter 2

Introduction to Statistical Inference

Statistical inference refers to the process of going from information gained by studying a portion of a **population**, called a **sample**, to new knowledge about the population. There are two types of statistical inferences that we will look at in this course: **hypothesis testing**, or tests of **significance**, and **estimation**. We begin with a type of significance testing known as a **permutation** or **randomization** test. We illustrate this procedure by analyzing the data set presented in Activity #1: *An Age discrimination Case?*

2.1 Activity #1: An Age Discrimination Case?

The data set presented on page 9 is Display 1.1 on page 5 of *Statistics in Action* by Watkins, Cobb and Scheaffer [GWCS04]. The data were provided by the Envelope Division of the Westvaco Corporation (a pulp and paper company originally based in West Virginia and which was purchased by the Meade Corporation in January of 2002) to the lawyers of Robert Martin. Martin was laid off by Westvaco in 1991 when the company decided to downsize. The company went through 5 rounds of layoffs. Martin was laid off in the second round and he was 54 years old then. He claimed that he had been laid off because of his age. He sued the company later that year alleging age discrimination. Your goal for this activity is to determine whether the data provided show that the suit had any merit.

The rows in the table represent the 50 employees in the envelope division of Westavaco before the layoffs (Robert Martin is in row 44). The Pay–column shows an "H" if the worker is payed hourly, and an "S" if the worker had a salary. The last column gives the age of each worker as of the first of January 1991 (shortly before the layoffs). The "RIF" column shows a "0" if the worker was not laid off during downsizing, a "1" is the worker was laid off in the first round of downsizing, a "2" if laid off in the second round, and so on.

For this activity, you will work in groups. Each team will study the data provided in search for patterns, if any, that might reflect age discrimination on the part of Westvaco management. Based on the analysis that your team makes of the data, your team will take a position on the whether there was age discrimination, or whether the data do not present enough evidence to support Martin's claim.

It was mentioned in the Preface that Statistics may be defined as the art of making decisions based on incomplete information. The decision we are asked to make in Activity #1 is whether Westvaco did discriminate based on age when laying off workers during its downsizing period in 1991. The information provided by the data set is incomplete because it does not state what criteria the company used to select the workers for layoffs. A quick examination of the RIF and Age columns in the table may lead us to believe that, on average, workers that were laid off were older than those that were not. For instance, a quick calculation[1] of the average age for workers on rows with RIF bigger than or equal to 1, i.e., workers that were laid off in one of the rounds, yields

$$\text{mean age laid off workers} = 50;$$

while the average age of those that were kept is

$$\text{mean age employed} = 46.$$

The decision we need to make then is whether a difference of 4 in the mean age of those workers who were fired versus those who remained is sufficient evidence for us to conclude that the company discriminated against older workers.

Example 2.1.1 (A Permutation Test) In this example we perform a **permutation test** for the Westvaco data for the hourly workers involved in the second round of layoffs. The question at hand is to decide whether Wastvaco based its selection of the three hourly workers that were laid off in that round on the workers' age. The ages of the 10 workers are

$$25, 38, 56, 48, \underline{55}, \underline{64}, \underline{55}, 55, 33, 35,$$

the underlined values being the ages of the 3 workers that were laid off in that round. The average age of the three laid off workers is obtained in R by typing

```
mean(c(55,64,55))
```

or 58. Note that the R function `c()` concatenates values in parentheses to form a vector or one–dimensional array.

Observe that, by contrast, the corresponding average age of the workers that were not laid off in the second round was

```
mean(c(25,38,56,48,55,33,35))
```

[1] See Appendix A for a discussion of how these calculations were done using R

Row	Job Title	Pay	Birth		Hire		RIF	Age 1/1/91
			Mo	Yr	Mo	Yr		
1	Engineering Clerk	H	9	66	7	89	0	25
2	Engineering Tech II	H	4	53	8	78	0	38
3	Engineering Tech II	H	10	35	7	65	0	56
4	Secretary to Engin Manag	H	2	43	9	66	0	48
5	Engineering Tech II	H	8	38	9	74	1	53
6	Engineering Tech II	H	8	36	3	60	1	55
7	Engineering Tech II	H	1	32	2	63	1	59
8	Parts Crib Attendant	H	11	69	10	89	1	22
9	Engineering Tech II	H	5	36	4	77	2	55
10	Engineering Tech II	H	8	27	12	51	2	64
11	Technical Secretary	H	5	36	11	73	2	55
12	Engineering Tech II	H	2	36	4	62	3	55
13	Engineering Tech II	H	9	58	11	76	4	33
14	Engineering Tech II	H	7	56	5	77	4	35
15	Customer Serv Engineer	S	4	30	9	66	0	61
16	Customer Serv Engr Assoc	S	2	62	5	88	0	29
17	Design Engineer	S	12	43	9	67	0	48
18	Design Engineer	S	3	37	6	74	0	54
19	Design Engineer	S	3	36	2	78	0	55
20	Design Engineer	S	1	31	3	67	0	60
21	Engineering Assistant	S	6	60	7	86	0	31
22	Engineering Associate	S	2	57	4	85	0	34
23	Engineering Manager	S	2	32	11	63	0	59
24	Machine Designer	S	9	59	3	90	0	32
25	Packaging Engineer	S	3	38	11	83	0	53
26	Prod Spec—Printing	S	12	44	11	74	0	47
27	Proj Eng—Elec	S	9	43	4	71	0	48
28	Project Engineer	S	7	49	9	73	0	42
29	Project Engineer	S	8	43	4	64	0	48
30	Project Engineer	S	6	34	8	81	0	57
31	Supv Engineering Serv	S	4	54	6	72	0	37
32	Supv Machine Shop	S	11	37	3	64	0	54
33	Chemist	S	8	22	4	54	1	69
34	Design Engineer	S	9	38	12	87	1	53
35	Engineering Associate	S	2	61	9	85	1	30
36	Machine Designer	S	2	39	4	85	1	52
37	Machine Parts Cont—Supv	S	10	28	8	53	1	63
38	Prod Specialist	S	9	27	10	43	1	64
39	Project Engineer	S	7	25	9	59	1	66
40	Chemist	S	12	30	10	52	2	61
41	Design Engineer	S	4	60	5	89	2	31
42	Electrical Engineer	S	11	49	3	86	2	42
43	Machine Designer	S	3	35	12	68	2	56
44	Machine Parts Cont Coor	S	9	37	10	67	2	54
45	VH Prod Specialist	S	5	35	9	55	2	56
46	Printing Coordinator	S	2	41	1	62	3	50
47	Prod Dev Engineer	S	6	59	11	85	3	32
48	Prod Specialist	S	7	32	1	55	4	59
49	VH Prod Specialist	S	3	42	4	62	4	49
50	Engineering Associate	S	8	68	5	89	5	23

Display 1.1 The data in *Martin v. Westvaco.*

Source: *Martin v. Envelope Division of Westvaco Corp.*, CA No. 92-03121-MAP, 850 Fed. Supp. 83 (1994).

or 41. There is, hence, quite a dramatic difference in mean ages of the two groups.

Suppose, for the sake of argument, that the company truly did not use age as a criterion for selection. There might have been other criteria that the company used, but the data set provided does not indicate other options. In the absence of more information, if the company did not take into account age in the selection process, then we may assume that choice of the three workers was done purely at random regarding age. What is the chance that a random selection of three workers out of the 10 will yield a sample with a mean of 58 or higher? To estimate this chance, assuming the selection is done at random (meaning that each individual in the group of 10 has the same chance of being selected as any other worker), we may replicate the process of selecting the 3 worker many times. We compute the average of the selected ages each time. The proportion of times that the average is 58, or higher, gives us an estimate of the likelihood, or **probability**, that the company would have selected for layoff three workers whose average is 58 or higher.

Since there are only 10 values that we would like to randomize, or permute, the simulation of the random selection process can be done easily by hand. For instance, we can write the 10 ages in cards and, after shuffling, randomly select three of them and record the ages. These simulations were done in class by 12 groups, each one generating 10 averages for their outcomes. The simulation results can be stored in an R vector called `Xbar`. A histogram of the generated values is shown in Figure 2.1.1. We would like to compute the proportion of

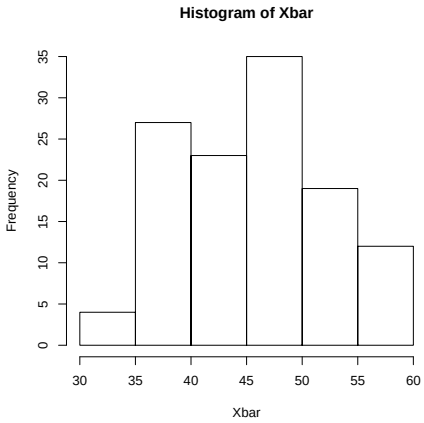

Figure 2.1.1: Sampling Distribution for Class Simulations in Activity #1

times that the average age in the samples is 58 or higher. In order to do this based on the histogram in Figure 2.1.1, it will helpful to refine the "breaks" in the histogram. The `hist()` function in R has an option that allows us to set the brakes for the bins in the histogram. This can be done by defining a vector

continuing the breaks; for instance,

```
b <- seq(30,60,by=1)
```

generates a sequence of values from 30 to 60 in steps of 1. We can then generate a new histogram of the sampling distribution in the simulations by typing

```
hist(Xbar, breaks = b)
```

The new histogram is shown in Figure 2.1.2 We can then see in Figure 2.1.2 that

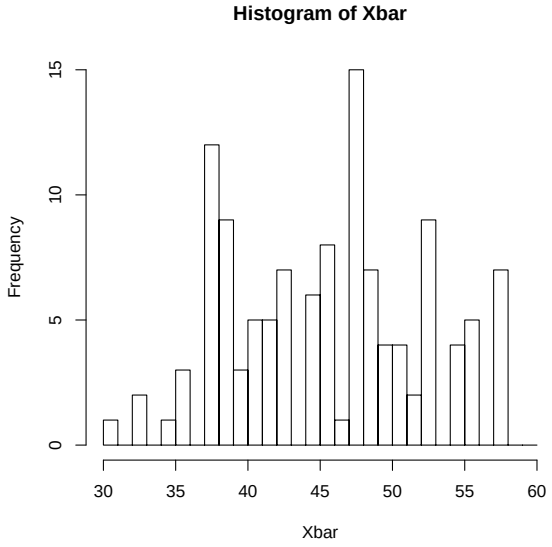

Figure 2.1.2: Sampling Distribution for Class Simulations in Activity #1

7 of the 120 random selections of three ages yielded an average of 58 or higher. This give a proportion of about 0.0583 or 5.83%. Thus, there is about a 5.83% chance that, if the company did a random selection of hourly workers in the second round for layoffs, an average of age of 58 or higher would have turned up. Some people might not consider this to be a low probability. However, in the following section we will see how to refine the simulation procedure by performing thousands of repetitions using R. This will yield a better estimate for that probability and we will see in a later section that it is in fact 5%. Thus, assuming that age had nothing to do with the selection of workers to be laid off in Westvaco's second round of layoffs in 1991, it is highly unlikely that the company selected the three workers that it actually fired. Thus, we can argue that age might have played a prominent role in the decision process.

2.2 Permutation Tests

The simulations that we performed in class for Activity #1 dealing with the 10 hourly workers at Westvaco involved in the second round of layoffs can be easily automated in R by running a script of commands stored in a .R extension file. The code is given in Appendix A.4. The key to the code is the `sample()` function in R. If the ages of the 10 workers are stored in a vector called `hourly2`, typing

```
s <- sample(hourly2, 3, replace = F)
```

selects a random sample of size 3, without replacement, from the 10 ages. Typing `mean(s)` then gives the mean age of the three workers which were selected. We can repeat the sampling procedures as many times as we wish by using the `for (i in 1:NRep)` loop structure in R that will set up a loop running from `NRep` repetitions. The code in Appendix A.4 yields a vector `Xbar` containing the means of `NRep` random samples of size 3 drawn from ages of the 10 hourly paid workers involved in the second round of layoffs at Westvaco in 1991. Figure 2.2.3 shows the histogram for `NRep`= 1000 repetitions. 2.1.1.

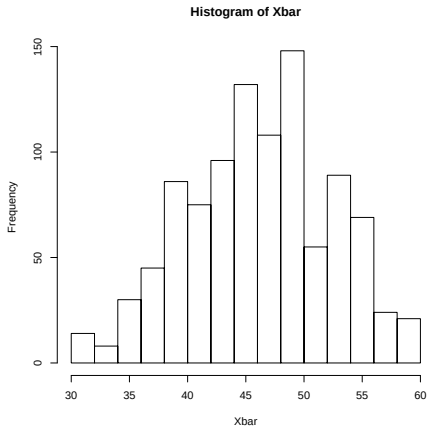

Figure 2.2.3: Sampling Distribution for Mean Age in 1000 Simulations

We can use the histogram in Figure 2.2.3 to estimate the proportion of samples that gave a mean age of 58 or higher as we did in Example 2.1.1. However, we can also use the programming capabilities in R to define a function that would compute for us give `Xbar` and the observed value of 58. The code defining the `pHat()` function can be found in Appendix A.5. An R script containing the code, called `pHatFunction.R`, may be downloaded from the course webpage. After sourcing this script file in R, we can type

```
p_hat <- pHat(Xbar,58)
```

which yields 0.045. This is an estimate of the probability that, assuming that the company selected 3 workers from the 10 involved in the second round of layoffs purely at random, the average age would be 58 or above. Thus, there is less than 5% chance that the company would have selected three workers whose age averages 58 or higher had the selection been purely random. The portion of the data dealing with hourly workers involved in the second rounds of layoffs leads us to believe that the company may have considered the age of the workers when deciding who to lay off.

The procedure used above in the analysis of the data for the 10 hourly workers involved in the second rounds of layoffs at Wesvaco in 1991 is an example of a **Test of Significance** known as a **Permutation Test**. We present here the general outline of the procedure so we can apply it in similar situations.

1. **Begin with a Question**

   We always begin by formulating a question we are trying to answer, or a problem we are trying to solve. In the examples we have just done, the question at hand is: *Did Westvaco discriminate against older workers?*

2. **Look at the Data**

   Do a quick analysis of the data (usually a exploratory analysis) to see if the data show trends that are relevant in answering the question. In Example 2.1.1, for instance, we looked that the 10 hourly paid workers involved in the second round of layoffs and found that the average age of the three laid off workers was 58, while that for the workers that remained was 41. This seems to suggest that the company was biased towards workers older than 55. This analysis also gives as the **test statistics** that we will be using in the significance test; namely, the average age of the workers sleeted for laid off.

3. **Set up a Null Hypothesis or Model**

   Based on the exploratory analysis of the data done previously, we surmise that, perhaps, the company discriminated based on age. Do the data actually support that claim? Suppose that, in fact, the company did not consider age when selecting workers for layoff. We may assume, for instance, that the company did the selection at random. What are the chances that we would see a selection of three workers whose average age is 58 or higher? If the chances are small (say 5%, or 1%), then it is highly unlikely that the company selected the workers that it did, if the selection was done purely at random. Hence, the data would not lend credence to the statement that the selection was done purely at random. Thus, the company has some explaining to do.

   The statement that the "selection was done at random" is an example of a **null hypothesis** or a **null model**. Usually denoted by H_o, a null hypothesis provide a model which can be used to estimate the chances that the test statistic will attain the observed value or more extreme values

under repeated occurrences of an experiment. In this case, the random experiment consists of selected three ages out of the ten at random.

4. **Estimate the p–value**

   The p–value is the likelihood, or probability, that, under the assumption that the null hypothesis is true, the test statistic will take on the observed value or more extreme values. A small p–value is evidence that the data do not support the null hypothesis, and therefore we have reason to believe that the selection was not done purely at random. We then say that the data are **statistically significant**. A large p–value, on the other hand, is indication that our claim that the company discriminated against older workers is not supported by the data. The reasoning here is that, even if the selection had been done at random, observed values as high as then one we saw are likely. Thus, the company could have selected the workers that it did select for layoff even if the selection was done without regard to age.

   The p–value can be estimated by replicating the selection process, under the assumption that the null hypothesis is true, many times and computing the proportion of the number of times that the test statistic is at least as large as the observed value.

5. **Statistical significance**

   The results of our simulations yielded and estimate of the p–value of about 0.045, or less than 5%. We therefore concluded, based on this analysis, that the data for the 10 hourly workers involved in the second rounds of layoffs at Westvaco in 1991 are statistically significant. Therefore, there is evidence that the company discriminated against older workers in that round.

2.3 Activity #2: Comparing two treatments

In the second activity of the course, students are presented with the following hypothetical situation:

Question

Suppose that a new treatment for certain disease has been devised and you claim that the new treatment is better than the existing one. How would you go about supporting your claim?

Discussion

1. Suppose that you also find twenty people that have the disease. Discuss how you would go about designing an experiment that will allow you to answer the question as to which treatment is more effective. Which factors should be considered? What are the variables in question here?

2. Discuss why randomization is important in the experimental design that you set up in 1 above.

3. Suppose that you randomly divide the 20 people into two groups of equal size. Group T receives the new treatment and Group E receives the existing treatment. Of these 20 people, 7 from Group T recover and 5 from Group E recover. What do you conclude about the effectiveness of the two treatments? Do you feel that you have been given enough information to draw an accurate conclusion? Why or why not? (Answer these questions before continuing)

Simulations

You may have concluded that the new treatment is better since it yielded a higher recovery rate. However, this may have occurred simply by chance. In other words, maybe there is really no difference in the effectiveness of the two treatments, and that regardless of which treatment any person in the group got, 12 out of the 20 people would have recovered anyways. Given this, what is the probability that you would observe the results you got regarding the effectiveness of the new treatment? i.e., what is the probability that 7 (or more) out of 10 in the new treatment group recover, given that 12 out of 20 will recover regardless of the treatment?

Your task for this activity is to simulate this experiment using playing cards in order to estimate the probability that, under the assumption that there is no difference between the treatments, 7 or more out of the treatment group will recover. To increase the accuracy of your results, be sure to run the simulation many times. Each team in the class will run simulations, and we will pool the results together.

Conclusions

Based on the results of the simulations, what can you conclude about the effectiveness of the two treatments? Is it possible that the initial findings could simply have resulted by chance? Would you have obtained these same results if there was no difference between the treatments?

We present a statistical significance test of the hypothetical data presented in Activity #2. We follow the procedure outlined in the analysis of the Westvaco data given in the previous secion.

1. **Begin with a Question**

   Is there any difference in the effectiveness of the two treatments? In particular, do the data support the claim that the new treatment is more effective than the existing one?

2. **Look at the Data**

We were given that 7 out of 10 subjects in Group T recovered. This yields a proportion of 7/10, or a 70% recovery rate. On the other hand, the recovery rate for Group E was 50%. The test statistic that we will use for this test is the proportion of subjects in the new treatment group that recover. Denote it by $\overline{p}$.

3. **Set up a Null Hypothesis**

   H_o in this case is the statement that there is no difference between the treatments; that is, 12 subjects out of the the twenty involved in the study would have recovered regardless of the treatment they were given.

4. **Estimate the p–value**

   In this case the p–value is the probability that, under the assumptions that there is no difference between the treatments, 7 or more subjects in the treatment group will recover. Given the information that the data came from a randomized comparative experiment, and the fact that the null hypothesis implies that 12 people of the 20 will recover regardless of which group they are put in, we can set up a replication process using plying cards, for instance. Use a deck of 12 black cards and 8 read cards to model those subject who recover and those who do not, respectively. Shuffling the cards and picking 10 at random will simulation the random selection of subjects to Group T. Count the number of black cards out of the 10. This will be the value of the test statistic for that particular run. Perform many of these replications and determine the proportion of those that showed seven or more recovered in the treatment group. This will yield an estimate of the p–value.

   The class performed 149 replications. Out of those, 49 showed 7 or more recoveries in the treatment group. Thus, 49/149, or about 0.33.

5. **Statistical significance**

   The p–value is about 33%, which is quite big. Hence, the data are not statistically significant. We cannot therefore, based on the data, conclude that the new treatment is better than the old one.

   We will see in the next section that the actual p–value is about 32.5%.

Chapter 3

Introduction to Probability

In the previous chapter we introduced the notion of a p–value as the measure of likelihood, or probability, that, under the assumption that the null hypothesis in a test of significance is true, we would see an observed value for a test statistic, or more extreme values, in repetitions of of a random experiment. We saw how to estimate the p–value by repeating an experiment many times (e.g., repeated sampling in the Westvaco data analysis, or repeated randomization of groups of subjects in a randomized comparative experiment). We then counted the number of times that the simulations yielded an outcome with a test statistic value as high, or higher, than the value observed in the data. The proportion of those times over the total number of replications then yields an estimate for the p–value. The idea behind this approximation is the **frequency interpretation** of probability as a long term ratio of occurrences of a given outcome. In this chapter we give a more formal introduction to the theory of probabilities with the goal of computing p–values, which we estimated through simulations in the previous chapter, through the use of **probability models**. We will also introduce the very important concepts of **random variables** and their **probability distributions**.

3.1 Basic Notions in Probability

3.1.1 Equal likelihood models

A *random experiment* is a process or observation, which can be repeated indefinitely under the same conditions, and whose outcomes cannot be predicted with certainty before the experiment is performed. For instance, tossing a coin is a random experiment with two possible outcomes: heads (H) or tails (T). Before the toss, we cannot predict with certainty which of the two outcomes we will get. We can, however, make some assumptions that will allow us to measure the likelihood of a given event. This measure of likelihood is an example of a **probability function**, P, on the set of outcomes of an experiment. For instance, we may assume that the two outcomes, H or T, are equality likely. It

then follows that

$$P(H) = P(T). \tag{3.1}$$

We assume that P takes on real values between 0 and 1. A value of 1 indicates that the event will certainly happen and a value of 0 means that the event will not happen. Any fraction in between 0 or 1 gives a measure of the likelihood of the event.

If we also assume that each toss of the coin will produce either a head or a tail. We can write this as

$$P(\text{H or T}) = 1, \tag{3.2}$$

indicating that we are absolutely sure that we will get either a head or a tail after the toss. Finally, we can also assume that each toss cannot yield both a head or a tail simultaneously (we say that H and T are **mutually exclusive** events); hence, we can also say that

$$P(\text{H or T}) = P(H) + P(T). \tag{3.3}$$

Combining equation (3.3) with (3.2) and the equal likelihood assumption in (3.1) yields that

$$2P(H) = P(H) + p(H) = P(H) + P(T) = 1,$$

from which we get that

$$P(H) = \frac{1}{2}.$$

Consequently, we also have that $P(T) = 1/2$. This is the probability model for the toss of a fair coin. Each outcome has a 50% chance of occurring.

In general, the equal likelihood model for an experiment with N equally likely and mutually exclusive possible outcomes yields a probability of

$$\frac{1}{N}$$

for each outcome.

Example 3.1.1 (Hourly workers in 2^{nd} round of layoffs at Westvaco) Consider the experiment of selecting at random, and without replacement, three ages from those of the 10 hourly paid workers involved in the second round of layoffs at Westvaco in 1991 (see data set for Activity #1 on page 9). The null hypothesis that the company did the selection purely at random is equivalent to an equal likelihood assumption; that is, each sample of 3 ages out of the 10 ages:

$$25, 38, 56, 48, 55, 64, 55, 55, 33, 35, \tag{3.4}$$

has the same chance of being chosen as any other group of 3. If we can compute the number, N, of all such outcomes, the probability of each outcome would be $1/N$.

To compute the number, N, of all possible samples of 3 ages out of the 10 given in (3.4), we may proceed as follows:

- In each sample of three ages, there are 10 choices for the first one. Once the first choice has been made, since we are sampling without replacement, there are 9 choices for the second are, and 8 choices for the third age. There is then a total of
$$10 \cdot 9 \cdot 8 = 720.$$

- However, we have over counted by quite a bit. For instance, the samples $\{25, 38, 56\}$ and $\{38, 25, 56\}$ are really the same sample, but are counted as two different ones in the previous count. In fact, there are
$$3 \cdot 2 \cdot 1 = 6$$
such repetitions for each sample of three. Thus, the over counting is 6–fold. Hence, the number, N, of all possible samples of size three from the 10 ages in (3.4) is
$$N = \frac{720}{6} = 120.$$

Hence the probability of selecting each sample of size 3 out of the 10 ages in (3.4) is
$$\frac{1}{120} \quad \text{or about } 0.83\%.$$

Observe that the number N computed in the previous example can be written as
$$N = \frac{10 \cdot 9 \cdot 8}{3 \cdot 2 \cdot 1} = \frac{10!}{3! \cdot 7!},$$
where the *factorial* of a positive integer n is defined by
$$n! = n \cdot (n-1) \cdots 2 \cdot 1.$$
For positive integers n and k, with $k \leqslant n$, the expression
$$\frac{n!}{k!(n-k)!},$$
with the understanding that $0! = 1$, counts the number of ways of choosing k objects out of n when the order in which the objects are chosen does not matter. These are also known as **combinations** of k objects out of n. We denote the number of combinations of k objects out of n by
$$\binom{n}{k},$$
read "n choose k," and call it the (n, k)–**binomial coefficient**. In R, the (n, k)–binomial coefficient is computed by typing

```
choose(n,k).
```

Type `choose(10,3)` in R to see that you indeed get 120.

3.1.2 Random Variables

The set of all possible outcomes of a random experiment is called the **sample space** for that experiment. For example, the collection of all samples of 3 ages out of the 10 in (3.4) is the sample space for the experiment consisting of selecting three ages at random and without replacement. A simpler example is provided by flipping a coin three times in a row. The sample space, for this experiment consists of all triples of heads, H, and tails, T:

$$\left.\begin{array}{l} \text{HHH} \\ \text{HHT} \\ \text{HTH} \\ \text{HTT} \\ \text{THH} \\ \text{THT} \\ \text{TTH} \\ \text{TTT} \end{array}\right\} \text{Sample Space} \tag{3.5}$$

Most of the times we are not interested in the actual elements of a sample space. We are more interested in numerical information derived from the samples. In the example of the hourly paid workers at Westvaco, for instance, we were interested in the average of the three ages, the test statistics. A test statistic yields a numerical value from each sample and it therefore defines a real valued function on the sample spaces.

Definition 3.1.2 (Random Variable) A real valued function, X, defined on the sample space of an experiment is called a **random variable**. A random variable may also be defined as a numerical outcome of a random experiment whose value cannot be determined with certainty. However, we can compute the probability that the random variable X takes on a given value or range of values.

Example 3.1.3 Toss a fair coin three times in a row. The sample space for this experiment is displayed in equation (3.5)

By the equal likelihood model implied in the "fair coin" assumption, each element in the sample space has probability 1/8.

Next, define a random variable, X, on the sample space given in (3.5) by

$$X = \text{number of heads in the outcome.}$$

Then, X can take on the values 0, 1, 2 or 3. Using the fact that each element in the sample space has probability 1/8, we can compute the probability that X takes on any of its values. We get

$$P(X = 0) \quad = \quad 1/8$$

since the outcome TTT in the sample space is the only one with no heads in it. Similarly, we get

$$P(X = 1) \quad = \quad 3/8$$

since the event $(X = 1)$ consists of the three outcomes HTT, THT and TTH. Continuing in this fashion, we obtain

$$P(X = k) = \begin{cases} 1/8 & \text{if } k = 0; \\ 3/8 & \text{if } k = 1; \\ 3/8 & \text{if } k = 2; \\ 1/8 & \text{if } k = 3. \end{cases} \tag{3.6}$$

The expression in equation (3.6) gives the **probability distribution** of the random variable X. It is a function which give the probabilities that X takes on a given value or range of values. The graph of this function is shown in Figure 3.1.1. Notice that all the probabilities add up to 1. Observe that X

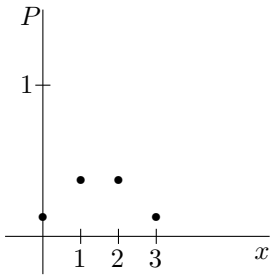

Figure 3.1.1: Probability Distribution of X

can take only a discrete set of values $\{0, 1, 2, 3\}$; that is, X cannot take on values in between those. This makes X into a **discrete random variable**. Random variables can also be **continuous**. For instance, measurements having to do with time, height, weight, pressure are all on done on a continuous scale, meaning that probabilities of ranges of measurements between any two distinct values are non–zero.

Example 3.1.4 Consider the combinations of 3 ages out of the 10 in (3.4). In R, we can find all the elements in this sample space as follows:

- Define a vector, `hourly2`, containing the 10 ages.
- The R statement

  ```
  combn(hourly2,3)
  ```

  produces a two–dimensional array containing combinations of 3 ages out of the 10 ages in `hourly2`. We can store the array in a **matrix** which we denote by `C` by typing

```
C <- combn(hourly2,3)
```

The columns of the matrix C are the combinations. To pick out the first column, for instance, type

```
C[,1].
```

This yields

```
[1] 25 38 56,
```

which are the first three ages in (3.4). The function `t()` in R will transpose a matrix; that is, it yields a matrix whose rows are the columns of the original matrix, and vice versa. Type `t(C)` to get all the 120 combinations of three ages out of the 10 ages of the hourly paid workers involved in the second round of layoffs.

Let $\overline{X}$ denote the mean value of the samples of size 3 drawn from the 10 ages in (3.4). Then, $\overline{X}$ defines a random variable whose values can be stored in a vector `Xbar` by the R commands

```
Xbar <- array(dim = 120)
for (i in 1:120) Xbar[i] <- mean(C[,i])
```

The frequency distribution of the variable $\overline{X}$ is pictured in Figure 3.1.2.

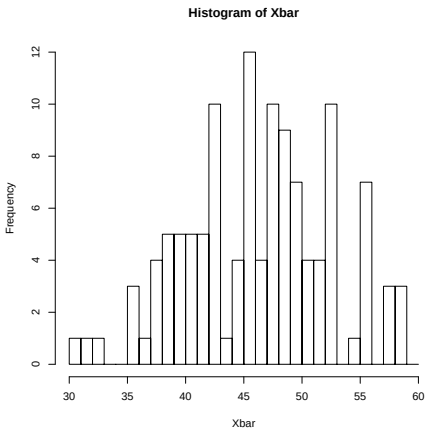

Figure 3.1.2: Frequency Distribution for $\overline{X}$

Using the histogram in Figure 3.1.2 and the fact that each element in the sample space has probability 1/120, we can make the following probability calculations

$$P(30 \leqslant \overline{X} \leqslant 33) = \frac{3}{120} = 0.025,$$

or

$$P(45 \leqslant \overline{X} < 46) = \frac{12}{120} = 0.1,$$

or

$$P(33 < \overline{X} < 35) = 0.$$

The fact that we got a probability of zero for mean ages strictly in between 33 and 35 tells us that $\overline{X}$ behaves more like a discrete distribution, even though we think of ages is a population (measured in time units) as a continuous variable. The reason for the discreteness of the variable in this example has to do with the fact that there are finitely many elements in the sample space.

We can compute the probability of the event $(\overline{X} \geqslant 58)$ using the function `pHat` defined in Appendix A.5 to get that

$$P(\overline{X} \geqslant 58) = 0.05.$$

This is the exact p-value for the Westvaco test of significance that we performed on the portion of the Westvaco data consisting of the ten hourly paid workers involved in the second round of layoffs.

3.2 Cumulative Distribution Function

Sometimes it is convenient to talk about the **cumulative distribution function** of a random variable; especially when dealing with continuous random variables.

Definition 3.2.1 (Cumulative Distribution Function) Given a random variable X, the cumulative distribution function of X, denoted by F_X, is a real valued function defined by

$$F_X(x) = P(X \leqslant x) \quad \text{for all } x \in \mathbf{R}.$$

Example 3.2.2 Let X denote the number of heads in three consecutive tosses of a fair coin. We have seen that X has a probability distribution given by

$$P(X = k) = \begin{cases} 1/8 & \text{if } k = 0; \\ 3/8 & \text{if } k = 1; \\ 3/8 & \text{if } k = 2; \\ 1/8 & \text{if } k = 3. \end{cases} \tag{3.7}$$

We may compute the cumulative distribution function, $F_X(x) = P(X \leqslant x)$, as follows:

First observe that if $x < 0$, then $\mathrm{P}(X \leqslant x) = 0$; thus,

$$F_X(x) = 0 \quad \text{for all } x < 0.$$

Note that $p(x) = 0$ for $0 < x < 1$; it then follows that

$$F_X(x) = \mathrm{P}(X \leqslant x) = \mathrm{P}(X = 0) \quad \text{for } 0 < x < 1.$$

On the other hand, $\mathrm{P}(X \leqslant 1) = \mathrm{P}(X = 0) + \mathrm{P}(X = 1) = 1/8 + 3/8 = 1/2$; thus,

$$F_X(1) = 1/2.$$

Next, since $p(x) = 0$ for all $1 < x < 2$, we also get that

$$F_X(x) = 1/2 \quad \text{for } 1 < x < 2.$$

Continuing in this fashion we obtain the following formula for F_X:

$$F_X(x) = \begin{cases} 0 & \text{if } x < 0, \\ 1/8 & \text{if } 0 \leqslant x < 1, \\ 1/2 & \text{if } 1 \leqslant x < 2, \\ 7/8 & \text{if } 2 \leqslant x < 3, \\ 1 & \text{if } x \geqslant 3. \end{cases}$$

Figure 3.2.3 shows the graph of F_X.

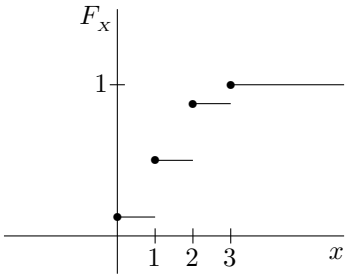

Figure 3.2.3: Cumulative Distribution Function for X

Example 3.2.3 (Cumulative distribution for `Xbar` in Example 3.1.4) In this example we use R to compute and plot the cumulative distribution function for sample mean, `Xbar`, which we computed in Example 3.1.4.

We can modify the code for the `pHat()` function given in Appendix A.5 to define a function, `DumDist()`, that computes the cumulative distribution function for `Xbar`. The code for this function is also given in Appendix A.5.

Assuming that we have a vector `Xbar` of sample means of all possible outcomes of selecting 3 ages at random from the 10 of the hourly workers involved in the second round of layoffs at Westvaco in 1991, we may compute $F_{\overline{x}}(x) =$ CumDist(Xbar,x), for any given value of x. For instance, we may compute

```
CumDist(Xbar,33)= 0.025,
```

which is the probability that the mean age of three workers selected at random out of the 10 is less than or equal to 33.

We can plot the cumulative distribution function of `Xbar` in R as follows:

- First, define a sequence of values of x from 25 to 65, which covers the range of values of `Xbar` as seen in the histogram in Figure 3.1.2. Typing

  ```
  x <- seq(25, 65, length = 1000)
  ```

  will generate the sequence, which has 1000 terms, and store the values in vector called `x`.

- The `plot()` function in R can then be obtain the graph of the cumulative distribution function, $F_{\overline{X}}$ of `Xbar` by typing

  ```
  plot(x, CumDist(Xbar,x), type="p", pch=".")
  ```

  The graph is shown in Figure 3.2.4

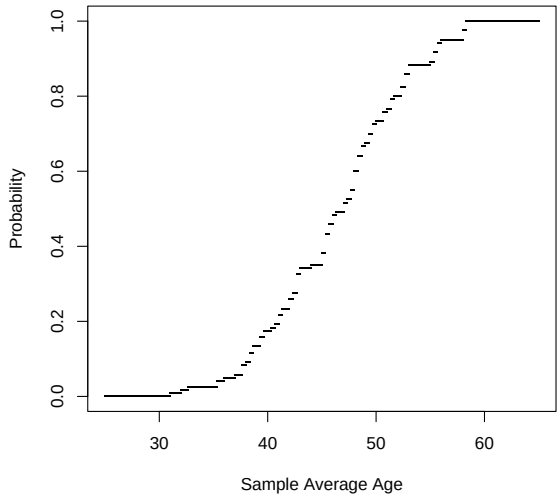

Figure 3.2.4: Cumulative Distribution for $\overline{X}$

The discontinuous nature of the graph of the cumulative distribution of `Xbar` is an indication that `Xbar` is a discrete random variable.

The importance of the cumulative distribution function, F_X, of a random variable X stems from the fact that, if $F_X(x)$ is known for all $x \in \mathbf{R}$, then we can compute the probability of the event $(a < X \leqslant b)$ as follows

$$\mathrm{P}(a < X \leqslant b) = F_X(b) - F_X(a).$$

For example,

```
CumDist(Xbar,35) - CumDist(Xbar,33)
```

yields 0, which shows that $\mathrm{P}(33 < \overline{X} \leqslant 35) = 0$; this is another indication that `Xbar` is a discrete random variable.

3.3 Statistical Independence

Example 3.3.1 Toss a fair coin twice in a row. The sample space for this experiment is the set

$$\{\mathrm{HH}, \mathrm{HT}, \mathrm{TH}, \mathrm{TT}\}.$$

The fairness assumption leads to the equal likelihood probability model:

$$\mathrm{P}(\mathrm{HH}) = \mathrm{P}(\mathrm{HT}) = \mathrm{P}(\mathrm{TH}) = \mathrm{P}(\mathrm{TT}) = \frac{1}{4}.$$

Let A denote the event that head comes up in the first toss and B that of a head coming up in the second toss. Then,

$$A = \{\mathrm{HH}, \mathrm{HT}\} \quad \text{and} \quad B = \{\mathrm{HH}, \mathrm{TH}\},$$

so that

$$\mathrm{P}(A) = \frac{1}{2} \quad \text{and } \mathrm{P}(B) = \frac{1}{2}.$$

The event A and B is the event that a head will come up in both the first and the second toss. We then have that

$$A \text{ and } B = \{\mathrm{HH}\}.$$

Consequently,

$$\mathrm{P}(A \text{ and } B) = \frac{1}{4}.$$

Observe that

$$\mathrm{P}(A \text{ and } B) = \mathrm{P}(A) \cdot \mathrm{P}(B).$$

When this occurs, we say that events A and B are **independent**.

Definition 3.3.2 (Independent Events) Two events, A and B, are said to be independent if

$$\mathrm{P}(A \text{ and } B) = \mathrm{P}(A) \cdot \mathrm{P}(B).$$

In other words, the probability of the joint occurrence of the two events is the product of their probabilities.

Example 3.3.3 Given a deck of 20 cards, 12 of which are black and 8 are red, consider the experiment of drawing two card at random and **without replacement**. Let A denote the event that the first card is black and B the event that the second card is also black. We show in this example that the A and B are not independent.

First, observe that $\mathrm{P}(A) = \dfrac{12}{20}$, or 60%. In the absence of information of what actually occurred in the first draw, we should still get that $\mathrm{P}(B) = 60\%$. On the other hand, we will see that

$$\mathrm{P}(A \text{ and } B) \neq \mathrm{P}(A) \cdot \mathrm{P}(B).$$

To determine $\mathrm{P}(A \text{ and } B)$ we need to compute the proportion of the total number of combinations of two black cards of the 20 which are black; that is,

$$\mathrm{P}(A \text{ and } B) = \frac{\binom{12}{2}}{\binom{20}{2}} = \frac{12 \cdot 11}{20 \cdot 19}.$$

We then see that

$$\mathrm{P}(A \text{ and } B) \neq \mathrm{P}(A) \cdot \mathrm{P}(B).$$

Note that what we do get is

$$\mathrm{P}(A \text{ and } B) = \mathrm{P}(A) \cdot \frac{11}{19}.$$

The fraction $\dfrac{11}{19}$ is the probability that the second card we draw is black, if we know that the first card drawn is black; for in that case there are 11 black cards in the remaining 19. This is known as the **conditional probability of B, given A**.

Definition 3.3.4 (Conditional Probability) Given two events A and B, with $\mathrm{P}(A) > 0$, the conditional probability of B given A, denoted by $\mathrm{P}(B \mid A)$, is defined to be

$$\mathrm{P}(B \mid A) = \frac{\mathrm{P}(A \text{ and } B)}{\mathrm{P}(A)}.$$

In the previous example we have that the probability of a black card in the second draw is black, given that the first draw yielded a black card, is

$$\mathrm{P}(B \mid A) = \frac{\dfrac{12}{20} \cdot \dfrac{11}{19}}{\dfrac{12}{20}} = \frac{11}{19}.$$

Observe that if A and B are independent, and $\mathrm{P}(A) > 0$, then

$$\mathrm{P}(B \mid A) = \frac{\mathrm{P}(A \text{ and } B)}{\mathrm{P}(A)} = \frac{\mathrm{P}(A) \cdot \mathrm{P}(B)}{\mathrm{P}(A)} = \mathrm{P}(B).$$

Thus, information obtained from the occurrence of event A does not affect the probability of B. This is another way to think about statistical independence. For example, the fact that we know that we got a head in the first toss of a coin should not affect the probability that the second toss will yield a head.

3.4 The Hypergeometric Distribution

In this section we compute the exact p–value associated with the the data for randomized comparative experiment described in Activity #2 (*Comparing two Treatments*).

Suppose a deck of 20 cards consists of 12 black cards and 8 red cards. Consider the experiment of selecting 10 cards at random. Let X denote the number of black cards in the random sample. Then, the p–value for the data in Activity #2 is $P(X \geqslant 7)$. In order to compute this probability, we need to determine the probability distribution of the random variable X.

Example 3.4.1 In this example we show how to compute $\mathrm{P}(X = 8)$. We then see how the calculation can be generalized to $\mathrm{P}(X = k)$, for $k = 2, 3, \ldots 10$.

We consider all possible ways of choosing 10 cards out of the 20. The set of all those possible combinations is the sample space. There are $\binom{20}{10}$ elements in the sample space, and they are all equally likely because of the randomization assumption in the experiment. In R, this number is computed to be `choose(20,10)`, or 184, 756. Out of all those combinations, we are interested in the ones that consist of 8 black cards and 2 red ones. There are $\binom{12}{8}$ ways of choosing the black cards. For each one of these, there are $\binom{8}{2}$ choices for the red cards. We then get that there are $\binom{12}{8} \cdot \binom{8}{2}$ random samples of 8 black cards and 2 red cards. Consequently,

$$\mathrm{P}(X = 8) = \frac{\binom{12}{8} \cdot \binom{8}{2}}{\binom{20}{10}} \approx 0.0750.$$

In general, we have that

$$\mathrm{P}(X = k) = \frac{\binom{12}{k} \cdot \binom{8}{10-k}}{\binom{20}{10}} \tag{3.8}$$

for $k = 2, 3, \ldots, 10$ and zero otherwise. This probability distribution is known as the **hypergeometric** distribution, which usually comes up when sampling

without replacement. In R, the `dhyper()` can be used to compute the probabilities in 3.8 as follows

```
dhyper(k,12,8,10).
```

In general, `dhyper(k,b,r,n)` gives the probability of seeing k black objects in a random sample of size n, selected without replacement, from a collection of b black objects and r red objects.

To compute the p–value associated with the data in Activity #2, we compute

$$p\text{–value} = \sum_{k=7}^{10} P(X = k).$$

This calculation can be performed in R as follows

```
pVal <- 0
for (k in 7:10) pVal <- pVal + dhyper(k,12,8,10)
```

This yields 0.3249583. So, the p–value is about 32.5%. In the class simulations we estimated the p–value to be about 33%.

3.5 Expectation of a Random Variable

3.5.1 Activity #3: The Cereal Box Problem

Suppose that your favorite breakfast cereal now includes a prize in each box. There are six possible prizes, and you really want to collect them all. However, you would like to know how many boxes of cereal you will have to eat before you collect all six prizes. Of course, the actual number is going to depend on your luck that one time, but it would be nice to have some idea of how many boxes you should expect to buy, on average.

In today's activity, we will conduct experiments in class to simulate the buying of cereal boxes. Your team will be provided with a die in order to run the simulation. Do several runs, at least 10. We will then pool together the results from each team in order to get a better estimate of the average number of boxes that need to be bought in order to collect all six prizes.

Before you start, discuss in your groups how you are going to run the experiment. Why does it make sense to use a die to run the simulation? What assumption are you making about the probability of finding a certain prize in a given cereal box? Are the assumptions you are making realistic?

Make a guess: how many boxes do you expect to buy before collecting all six prizes?

3.5.2 Definition of Expected Value

In the Cereal Box Problem activity, the random variable in question is the number of cereal boxes that need to be purchased in order to collect all six prizes. We are assuming here that all the prizes have been equally distributed; that is, each prize is in one sixth of all the produced boxes. We also assume that the placement of the prizes in each box is done at random. This justify the use of a balanced six-sided die to do the simulations.

Let Y the number of times the die has to be rolled in order to obtain all six numbers on the faces of the die. Then the values that Y can take are $6, 7, 8, \ldots$ We are not interested in the actually probability distribution for Y. What we would like to know is the following: suppose we run the experiment many times, as was done in class. Sometimes we find it takes 9 rolls to get all six numbers; other times it might take 47; and so on. Suppose we keep track of the values of Y for each trial, and that at the end of the trials we compute the average of those values. What should we expect that average to be in the long run? This is the idea behind the **expected value** of a random variable. We begin by defining the expected value of a discrete random variable with finitely many possible values.

Definition 3.5.1 (Expected Value) Let X be a discrete random variable that can take on the values $x_1, x_2, \ldots, x_N$. Suppose we are given the probability distribution for X:

$$p_X(x_i) = \mathrm{P}(X = x_i) \quad \text{for each } i = 1, 2, \ldots, N.$$

The expected value of X, denoted by $E(X)$, is defined to be

$$E(X) = x_1 p_X(x_1) + x_2 p_X(x_2) + \cdots + x_n p_X(x_N),$$

or

$$E(X) = \sum_{k=1}^{N} x_k p_X(x_k).$$

Thus, $E(X)$ is the weighted average of all the possible values of X, where the weights are given by the probabilities of the values.

Example 3.5.2 Let X denote the number of heads in three consecutive tosses of a fair coin. We have seen that X is a random variable with probability distribution

$$p_X(k) = \begin{cases} 1/8 & \text{if } k = 0; \\ 3/8 & \text{if } k = 1; \\ 3/8 & \text{if } k = 2; \\ 1/8 & \text{if } k = 3. \end{cases}$$

and zero otherwise. We then have that

$$E(X) = 0p_X(0) + 1p_X(1) + 2p_X(2) + 3p_X(3) = \frac{3}{8} + 2\frac{3}{8} + 3\frac{1}{8} = \frac{12}{8} = \frac{3}{2}.$$

Thus, on average, we expect to see 1.5 heads in three consecutive flips of a fair coin.

Remark 3.5.3 If the random variable X is discrete, but takes on infinitely many values $x_1, x_2, x_3, \ldots$, the expected value of X is given by the infinite sum

$$E(X) = \sum_{k=1}^{\infty} x_k p_{_X}(x_k),$$

provided that the infinite sum yields a finite number.

3.5.3 The Law of Large Numbers

Here is an empirical interpretation of the expected value of a random variable X. Recall that X is a numerical outcome of a random experiment. Suppose that the experiment is repeated n times, and that each time we record the values of X:

$$X_1, X_2, \ldots, X_n.$$

We then compute the **sample mean**

$$\overline{X}_n = \frac{X_1 + X_2 + \cdots + X_n}{n}.$$

It is reasonable to assume that as n becomes larger and larger, the sample mean should get closer and closer to the expected value of X. This is the content of a mathematical theorem known as the Law of Large Numbers. This theorem is also the mathematical justification of the frequency interpretation of probability which we have been using in the simulations in this course. For example, flip a fair coin (i.e., $\mathrm{P}(H) = 1/2$) one hundred times. On average, we expect to see 50 heads. This is the reason why in the previous example in which the coin is flipped 3 times we got an expected value of 3/2.

3.5.4 Expected Value for the Cereal Box Problem

Let Y the number of times a balanced die has to be rolled in order to obtain all six numbers on the faces of the die. We can think of Y as simulating the results of an experiment which consists of purchasing cereal boxes until we collect all six prizes. We would like to estimate the expected value of Y. We can use the Law of Large Numbers to get an estimate for $E(Y)$. The class ran 140 simulations of the experiment and the outcomes of the experiments were stored in a vector called `Nboxes`. These are contained in the MS Excel file `CerealBoxProblemAllClassSimulations.xls`, which may be downloaded from `http://pages.pomona.edu/~ajr04747`. The histogram in Figure 3.5.5 shows the frequency distribution of the values in the vecotr `Nboxes`. Observe that the distribution of `Nboxes` is skewed to the right; that is, large values of Y occur less frequently than those around the mean of the distribution, which is

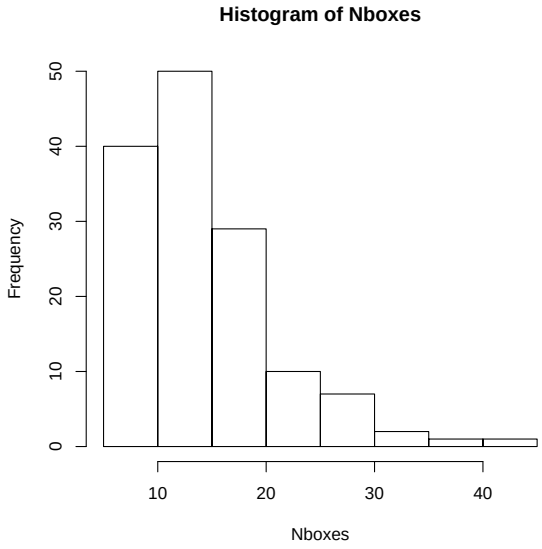

Figure 3.5.5: Cumulative Distribution for $\overline{X}$

14.8. By the Law of Large Numbers, this serves as an estimate for the expected value of Y. We will see shortly how to compute the actual value of $E(Y)$.

Let X_1 denote the number of times it takes to roll any number. Once any number has been rolled, we let X_2 denote the number of rolls of the die that it takes to come up with any number other than the one that has already come up. Similarly, once two distinct numbers have been rolled, let X_3 denote the number of times it takes for a different number to come up. Continuing in this fashion we see that

$$Y = X_1 + X_2 + X_3 + X_4 + X_5 + X_6,$$

so that

$$E(Y) = E(X_1) + E(X_2) + E(X_3) + E(X_4) + E(X_5) + E(X_6). \tag{3.9}$$

It is clear that $E(X_1) = 1$, since any number that comes up is fair game. To compute the other expected values, we may proceed as follows: suppose we want to compute $E(X_5)$, for instance. We have already rolled up four distinct numbers, and so we need to come up up with 2 remaining ones. The probability of rolling up any of the two numbers is

$$p = \frac{2}{6} = \frac{1}{3}.$$

We continue to roll the die until that event with probability $p = 1/3$ occurs. How many tosses of the die would that take on average?

More generally, suppose an experiment has two possible outcomes: one that occurs with probability p, which we call a "success," and and the other, which we call a "failure" occurs with probability $1-p$. We run the experiment many times, assuming that the runs are independent, until a "success" is observed. Let X denote the number of times it takes for the first "success" to occur. X is then a random variable that can take on any of the values

$$1, 2, 3, \ldots$$

We can use the independence assumption to compute the probability distribution for X. Starting with the event $(X = 1)$, a success in the first trial, we see that its probability is p, the probability of success. We then have that

$$p_{_X}(1) = p.$$

The event $(X = 2)$, the first success occurs in the second trial, is the joint occurrence of a failure in the first trial and a succuss in the second one. Since the events are independent, we get that

$$p_{_X}(2) = (1-p)p.$$

Similarly, the event $(X = 3)$ consists of 2 failures followed by 1 success, and therefore

$$p_{_X}(3) = (1-p)^2 p.$$

We can then see that

$$p_{_X}(k) = (1-p)^{k-1} p \quad \text{for } k = 1, 2, 3, \ldots.$$

This is an example of a discrete random variable which can take on infinitely many values. Its distribution is called the **geometric** distribution with parameter p. The expected value of X is then

$$E(X) = \sum_{k=1}^{\infty} (1-p)^{k-1} p.$$

We may compute $E(X)$ in a different and simpler way as follows: If a success occurs in the first trial then only one trial is needed. This occurs with probability p. On the other hand, if a failure occurs, we need to keep going. We then have that

$$E(X) = p \cdot 1 + (1-p) \cdot (E(X) + 1).$$

The second term in the last equation simply states that if a failure occurs in the first trial, then, on average, we need to roll $E(X)$ more times to get a success. Simplifying the previous algebraic expression for $E(X)$ we get that

$$E(X) = p + (1-p)E(X) + 1 - p,$$

from which we get that

$$E(X) = E(X) - pE(X) + 1.$$

Thus,

$$pE(X) = 1,$$

and therefore

$$E(X) = \frac{1}{p}.$$

Hence, if the probability of a success is p, then, on average, we expect to see the first success in $1/p$ trials. Thus, for example,

$$E(X_5) = \frac{1}{\frac{1}{3}} = 3$$

Similarly,

$$E(X_2) = \frac{6}{5},\ E(X_3) = \frac{6}{4},\ E(X_4) = \frac{6}{3},\ \text{ and }\ E(X_6) = 6.$$

Substituting all these values into (3.9), we obtain that

$$E(Y) = 1 + \frac{6}{5} + \frac{3}{2} + 2 + 3 + 6 = 14.7,$$

which shows that our estimate of 14.8 is very close to the actual expected value.

3.6 Variance of a Random Variable

The expected value, $E(X)$, of a random variable, X, gives an indication of where the middle or center of its distribution. We now define a measure of the spread of the distribution from its center.

Definition 3.6.1 (Variance of a Random Variable) Let X be a random variable with expected value $\mu = E(X)$. The **variance** of X, denoted by $\text{Var}(X)$, is defined to be

$$\text{Var}(X) = E[(X - \mu)^2].$$

That is, $\text{Var}(X)$ is the mean square deviation of values of X from $E(X)$.

The positive square root of $\text{Var}(X)$ is called the **standard deviation** of X and is denoted by σ_X.

We can compute the variance of X as follows:

$$\begin{aligned}
\text{Var}(X) &= E[X^2 - 2\mu X + \mu^2] \\
&= E(X^2) - E(2\mu X) + E(\mu^2) \\
&= E(X^2) - 2\mu E(X) + \mu^2 \\
&= E(X^2) - 2\mu^2 + \mu^2 \\
&= E(X^2) - \mu^2.
\end{aligned}$$

Thus,

$$\text{Var}(X) = E(X^2) - [E(X)]^2.$$

In the case in which X is discrete and takes on finitely many values, $x_1, x_2, \ldots, x_N$, we compute $E(X^2)$ as follows

$$E(X^2) = x_1^2 p_X(x_1) + x_2^2 p_X(x_2) + \cdots + x_n^2 p_X(x_N);$$

that is, $E(X^2)$ is the weighted average of the squares of the values of X.

Example 3.6.2 Let X denote the number of heads in three consecutive tosses of a fair coin. Then,

$$E(X) = 0^2 p_X(0) + 1^2 p_X(1) + 2^2 p_X(2) + 3^2 p_X(3) = \frac{3}{8} + 4\frac{3}{8} + 9\frac{1}{8} = \frac{24}{8} = 3.$$

We have seen that $E(X) = 3/2$. Thus,

$$\mathrm{Var}(X) = E(X^2) - [E(X)]^2 = 3 - \left(\frac{3}{2}\right)^2 = \frac{3}{4}.$$

The standard deviation of X then is

$$\sigma_X = \frac{\sqrt{3}}{2}.$$

Chapter 4

Introduction to Estimation

Continuing with our study of statistical inference, we now turn to the problem of estimating parameters from a given population based on information obtained from a sample taken from that population. We begin with the example of estimating the size of a population by means of capture–tag–recapture sampling.

4.1 Activity #4: Estimating the Number of Goldfish in a Lake

Introduction

A lake contains an unknown number of fish. In order to estimate the size of the fish population in the lake, the **capture–tag–recapture** sampling method is used. The nature of this sampling technique involves capturing a sample of size M and tagging the fish. The sample is then released and the fish redistribute themselves throughout the lake. A new sample of size n is then recaptured and the number of tagged fish, t, is recorded. These numbers are then used to estimate the population size.

Description

The goal of this activity is to come up with a good estimate for a population size using the capture–tag–recapture sampling technique described above. We will simulate this sampling method as follows. A bowl containing an unknown number of goldfish crackers will be fish in the lake. We take out 100 (this will be M) of the crackers and replace them with fish crackers of a different color (for example, pretzel fish crackers). This simulates capturing and tagging the fish. Each person/team will recapture a handful and record the sample size n and the number of tagged fish t. Then record this data for each trial.

Discussion

- How would you use the numbers you collected to estimate the size of the population?

- What estimates do the class data yield?
- What assumptions are you making in the process of coming up with an estimate of the fish population size?
- Can you come up with an interval estimate, as opposed to a point estimate, for the population size?
- How confident are you on this interval estimate?

We denote the number of goldfish by N; this is the population parameter we would like to estimate. Assuming the the tagged fish distribute themselves uniformly throughout the lake, then the proportion of tagged fish in the lake is

$$p = \frac{M}{N}.$$

We are also assuming that during the time of sampling no fish are going in or out of the lake. Also, no fish are being born or dying.

The proportion of tagged fish in the sample,

$$\widehat{p} = \frac{t}{n},$$

serves as an estimate for the true proportion p. This is known as a **point estimate**. We then have that

$$\widehat{p} \approx p,$$

or

$$\frac{t}{n} \approx \frac{M}{N}.$$

This yields the estimate

$$N \approx \frac{Mn}{t}$$

for the size of the fish population. We then obtain the **estimator**

$$\widehat{N} = \frac{Mn}{t}$$

for the population size N.

In the class activity we collected samples of various size without replacement and obtained the data shown in Table 4.1.

Notice that the class estimates range all the way from 820 to infinity. There is a lot of variability in the estimates and we do not have an idea yet of the likelihood that the true size of the population lies in that range; in other words, what are the chances that the true population size lies below 820? A single point estimate (in particular, any of the last two ones in the table) is not very useful. We would like develop an estimate for the population size that also give us a measure of the likelihood that the true population parameter lies in a given range of estimates. This is known as a **confidence interval estimate** and we will develop that concept in subsequent sections.

n	t	$\widehat{N}$
41	5	820
27	3	900
63	6	1050
45	4	1125
59	5	1180
51	4	1275
53	4	1325
54	4	1350
183	13	1408
44	3	1467
45	3	1500
138	9	1533
63	4	1575
83	5	1660
50	3	1667
42	2	2100
65	3	2167
87	4	2175
46	2	2300
76	3	2533
50	1	5000
112	2	5600
118	2	5900
57	0	∞
25	0	∞

Table 4.1: Fish–in–the–Lake Activity Class Data

4.2 An interval estimate for the proportion of tagged fish

In this section we develop a confidence interval for the true proportion p of tagged fish in the lake. This will lead to a confidence interval for the population size N.

In order to develop a confidence interval for the proportion, p, of tagged fish in Activity #4, we need to look more carefully into the nature of the sampling that we performed. In the class activity we collected handfuls of fish without replacement. If a sample is of size n, then the probability that the first fish we pick is tagged is p. However, the probability that the second fish we pick is tagged is not going to be p. On the other hand, if we had been sampling with replacement (that is, we pick the fish, note whether it is tagged or not, and put it back in the lake), the probability that a given fish is tagged is p. The distribution of the number of tagged fish in the sample of tagged fish in a sample of size n will then be easier to analyze. We will do this analysis of the sampling with replacement first; this will lead to the study of the **binomial distribution**. Later in this section, we will go back to the sampling without replacement situation which is best treated with the hypergeometric distribution.

4.2.1 The Binomial Distribution

Suppose that we collect a sample of size n of fish from the lake with replacement. Each time we collect a fish, we note whether the fish is tagged or not and we then put the fish back. We define the random variables

$$X_1, X_2, X_3, \ldots, X_n$$

as follows: for each $i = 1, 2, \ldots, n$, $X_i = 1$ if the fish is tagged and $X_i = 0$ if it is not. Then, each X_i is a random variable with distribution

$$p_{X_i}(x) = \begin{cases} 1-p & \text{if } x = 0; \\ p & \text{if } x = 1. \end{cases}$$

for $i = 1, 2, \ldots, n$. Furthermore, the random variables $X_1, X_2, \ldots, X_n$ are independent from one another; in other words, the joint probabilities

$$P(X_{j_1} = x_{j_1}, X_{j_2} = x_{j_2}, \ldots, X_{j_k} = x_{j_k}),$$

for any subset $\{j_1, j_2, \ldots, j_k\}$ of $\{1, 2, 3, \ldots n\}$, are computed by

$$p_{X_{j_1}}(x_{j_1}) \cdot p_{X_{j_2}}(x_{j_2}) \cdots p_{X_{j_k}}(x_{j_k}).$$

Each random variable X_i is called a **Bernoulli Trial** with parameter p. We then have n independent Bernoulli trials.

Then random variable

$$Y_n = X_1 + X_2 + \cdots + X_n$$

then counts the number of tagged fish in a sample of size n. We would like to determine the probability distribution of Y_n. Before do so, we can compute the expected value of Y_n as follows

$$E(Y_n) = E(X_1) + E(X_2) + \cdots + E(X_n)$$

where

$$E(X_i) = 0 \cdot (1-p) + 1 \cdot p = p,$$

for each $i = 1, 2, \ldots, n$, so that

$$E(Y_n) = np.$$

Since the X_i's are independent, it is possible to show that

$$\mathrm{Var}(Y_n) = \mathrm{Var}(X_1) + \mathrm{Var}(X_2) + \cdots + \mathrm{Var}(X_n),$$

where, for each i,

$$\mathrm{Var}(X_i) = E(X_i^2) - [E(X_i)]^2,$$

with

$$E(X_i^2) = 0^2 \cdot (1-p) + 1^2 \cdot p = p,$$

so that

$$\mathrm{Var}(X_i) = p - p^2,$$

or

$$\mathrm{Var}(X_i) = p(1-p).$$

We then have that

$$\mathrm{Var}(Y_n) = np(1-p).$$

We now compute the probability distribution for Y_n. We begin with

$$Y_2 = X_1 + X_2.$$

In this case, observe that Y_2 takes on the values 0, 1 and 2. We then compute

$$\begin{aligned} \mathrm{P}(Y_2 = 0) &= \mathrm{P}(X_1 = 0, X_2 = 0) \\ &= \mathrm{P}(X_1 = 0) \cdot \mathrm{P}(X_2 = 0), \quad \text{by independence,} \\ &= (1-p) \cdot (1-p) \\ &= (1-p)^2. \end{aligned}$$

Next, since the event $(Y_2 = 1)$ consists of the mutually exclusive events

$$(X_1 = 1, X_2 = 0) \text{ and } (X_1 = 0, X_2 = 1),$$

$$\begin{aligned} \mathrm{P}(Y_2 = 1) &= \mathrm{P}(X_1 = 1, X_2 = 0) + \mathrm{P}(X_1 = 0, X_2 = 1) \\ &= \mathrm{P}(X_1 = 1) \cdot \mathrm{P}(X_2 = 0) + \mathrm{P}(X_1 = 0) \cdot \mathrm{P}(X_2 = 1) \\ &= p(1-p) + (1-p)p \\ &= 2p(1-p). \end{aligned}$$

Finally,

$$\begin{aligned} \mathrm{P}(Y_2 = 2) &= \mathrm{P}(X_1 = 1, X_2 = 1) \\ &= \mathrm{P}(X_1 = 1) \cdot \mathrm{P}(X_2 = 1) \\ &= p \cdot p \\ &= p^2. \end{aligned}$$

We then have that the probability distribution of Y_2 is given by

$$p_{Y_2}(y) = \begin{cases} (1-p)^2 & \text{if } y = 0, \\ 2p(1-p) & \text{if } y = 1, \\ p^2 & \text{if } y = 2. \end{cases} \tag{4.1}$$

We shall next consider the case in which we add three mutually independent Bernoulli trials X_1, X_2 and X_3. In this case we write

$$Y_3 = X_1 + X_2 + X_3 = Y_2 + X_3.$$

Then, Y_2 and X_3 are independent. To see why this is so, compute

$$\begin{aligned} \mathrm{P}(Y_2 = y, X_3 = z) &= \mathrm{P}(X_1 + X_2 = y, X_3 = z) \\ &= \mathrm{P}(X_1 = x, X_2 = y - x, X_3 = z) \\ &= \mathrm{P}(X_1 = x) \cdot \mathrm{P}(X_2 = w - x) \cdot \mathrm{P}(X_3 = z), \end{aligned}$$

since the random variables are independent. Consequently, by independence again,

$$\begin{aligned} \mathrm{P}(Y_2 = y, X_3 = z) &= \mathrm{P}(X_1 = x, X_2 = y - x) \cdot \mathrm{P}(X_3 = z) \\ &= \mathrm{P}(X_1 + X_2 = y) \cdot \mathrm{P}(X_3 = z) \\ &= \mathrm{P}(Y_2 = y) \cdot \mathrm{P}(X_3 = z), \end{aligned}$$

which shows the independence of Y_2 and X_3.

To compute the probability distribution of Y_3, first observe that Y_3 takes on the values 0, 1, 2 and 3, and that

$$Y_3 = Y_2 + X_3,$$

where the probability distribution function of Y_2 is given in equation (4.1).

We compute

$$\begin{aligned} \mathrm{P}(Y_3 = 0) &= \mathrm{P}(Y_2 = 0, X_3 = 0) \\ &= \mathrm{P}(Y_2 = 0) \cdot \mathrm{P}(X_3 = 0), \\ &= (1-p)^2 \cdot (1-p) \\ &= (1-p)^3. \end{aligned}$$

Next, since the event $(Y_3 = 1)$ consists of mutually exclusive events

$$(Y_2 = 1, X_3 = 0) \text{ and } (Y_2 = 0, X_3 = 1),$$

$$\begin{aligned} \mathrm{P}(Y_3 = 1) &= \mathrm{P}(Y_2 = 1, X_3 = 0) + \mathrm{P}(Y_2 = 0, X_3 = 1) \\ &= \mathrm{P}(Y_2 = 1) \cdot \mathrm{P}(X_3 = 0) + \mathrm{P}(Y_2 = 0) \cdot \mathrm{P}(X_3 = 1) \\ &= 2p(1-p)(1-p) + (1-p)^2 p \\ &= 3p(1-p)^2. \end{aligned}$$

Similarly,

$$\begin{aligned} \mathrm{P}(Y_3 = 2) &= \mathrm{P}(Y_2 = 2, X_3 = 0) + \mathrm{P}(Y_2 = 1, X_3 = 1) \\ &= \mathrm{P}(Y_2 = 2) \cdot \mathrm{P}(X_3 = 0) + \mathrm{P}(Y_2 = 1) \cdot \mathrm{P}(X_3 = 1) \\ &= p^2(1-p) + 2p(1-p)p \\ &= 3p^2(1-p). \end{aligned}$$

Finally,

$$\begin{aligned} \mathrm{P}(Y_3 = 3) &= \mathrm{P}(Y_2 = 2, X_3 = 1) \\ &= \mathrm{P}(Y_2 = 0) \cdot \mathrm{P}(X_3 = 0) \\ &= p^2 \cdot p \\ &= p^3. \end{aligned}$$

We then have that the probability distribution function of Y_3 is

$$p_{Y_3}(y) = \begin{cases} (1-p)^3 & \text{if } y = 0, \\ 3p(1-p)^2 & \text{if } y = 1, \\ 3p^2(1-p) & \text{if } y = 2 \\ p^3 & \text{if } y = 3. \end{cases}$$

If we go through similar calculations for the case of four mutually independent Bernoulli trials with parameter p, we obtain that for

$$Y_4 = X_1 + X_2 + X_3 + X_4,$$

$$p_{Y_4}(y) = \begin{cases} (1-p)^4 & \text{if } y = 0, \\ 4p(1-p)^3 & \text{if } y = 1, \\ 6p^2(1-p)^2 & \text{if } y = 2 \\ 4p^3(1-p) & \text{if } y = 3 \\ p^4 & \text{if } y = 4. \end{cases}$$

Observe that the terms in the expression for $p_{Y_4}(y)$ are the terms in the expansion of $[(1-p)+p]^4$. In general, the expansion of the n^{th} power of the binomial $a + b$ is given by

$$(a+b)^n = \binom{n}{0} b^n + \binom{n}{1} a b^{n-1} + \cdots \binom{n}{k} a^k b^{n-k} + \cdots \binom{n}{n} a^n,$$

where

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}, \quad k = 0, 1, 2 \ldots, n,$$

are called the **binomial coefficients**. Written in a more compact form,

$$(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k}.$$

Thus, the expansion for $[(1-p)+p]^n$ is

$$[(1-p)+p]^n = \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k}.$$

Observe that $[(1-p)+p]^n$ is also equal to 1. We then have that

$$\sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} = 1,$$

which shows that the terms $\binom{n}{k} p^k (1-p)^{n-k}$ do indeed define the probability distribution of a random variable. This is precisely the distribution for

$$Y_n = X_1 + X_2 + \cdots + X_n,$$

where $X_1, X_2, \ldots, X_n$ are n mutually independent Bernoulli trials with parameter p, for $0 < p < 1$. We therefore have that

$$p_{Y_n}(k) = \binom{n}{k} p^k (1-p)^{n-k} \quad \text{for } k = 0, 1, 2, \ldots, n,$$

and Y_n is said to have a **binomial distribution** with parameters n and p. We write $Y_n \sim B(n, p)$. As we have seen earlier, the expected value and variance of Y_n are

$$E(Y_n) = np \quad \text{and} \quad \text{Var}(Y_n) = np(1-p).$$

Example 4.2.1 Suppose that the true proportion of tagged fish in the lake is $p = 0.06$. If we collect 100 fish with replacement and note the number of fish out of the one hundred that are tagged in a random variable Y, then its distribution is given by

$$p_Y(k) = \binom{100}{k} (0.06)^k (0.94)^{100-k} \quad \text{for } k = 0, 1, 2, \ldots, 100.$$

In R, these probabilities can be computed using the function `dbinom()`. For example, the probability that we will see 5 tagged fish in a sample of size 100 is

```
dbinom(5,100,0.06)
```

or about 16.4%. In general, `dbinom(x,n,p)` gives the probability of x successes in n independent Bernoulli trials with probability of a success p.

A plot of the distribution of Y_n for x between 0 and 20 is shown in Figure 4.2.1. The plot was obtained in R by typing

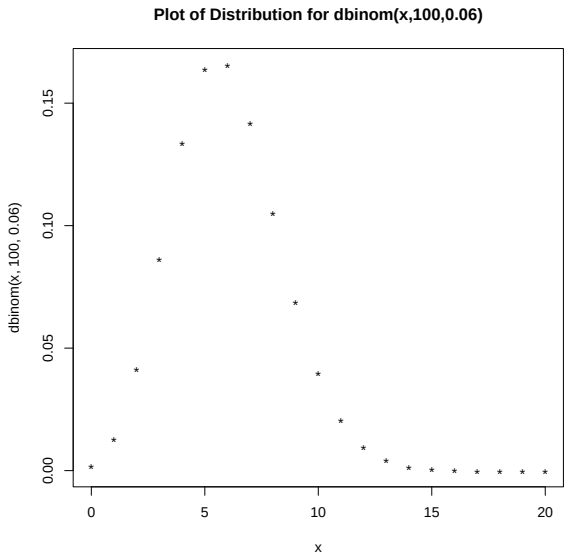

Figure 4.2.1: Probability Distribution of Y_n, for $p = 0.06$ and $n = 100$

```
x <- 0:20
plot(x,dbinom(x,100,0.06),type="p", pch="*")
```

Another way to plot this distribution function makes use of the `barplot()` function in R. Type the R commands

```
x <- 0:20
barplot(dbinom(x,100,0.06), space = c(0,0),names.arg=x,
                                ylab="Probability", xlab="x")
```

to obtain the plot shown in Figure 4.2.2. In the figure, the probability of a given value, x, is to be understood as the area the bar above it. It is therefore assumed that each bar has width 1.

Notice that in both plots in Figures 4.2.1 and 4.2.2 the distribution is centered around the value 6, which is the expected value of Y_n for $n = 100$ and $p = 0.06$. Corresponding plots for Y_n, for $n = 1000$, are shown in Figures 4.2.3 and 4.2.4.

Thus, for a larger value of n we see a very symmetric distribution for Y_n around its expected value of 60. The graphs also reveal an interesting feature of the binomial distribution for large values of n. The distribution function for Y_n "smooths out;" that is, the graphs can be approximated by a smooth curve. This is illustrated in Figure 4.2.5. The points in the graph of p_{Y_n} are indicated by a `+` sign. The smooth curve in the Figure 4.2.5 is very close to the graph

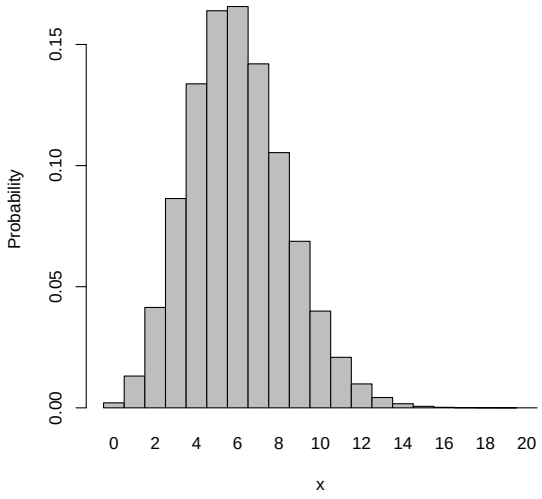

Figure 4.2.2: Probability Distribution of Y_n, for $p = 0.06$ and $n = 100$

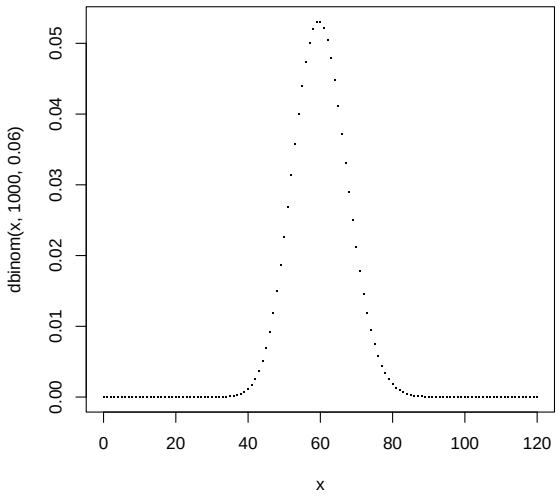

Figure 4.2.3: Probability Distribution of Y_n, for $p = 0.06$ and $n = 1000$

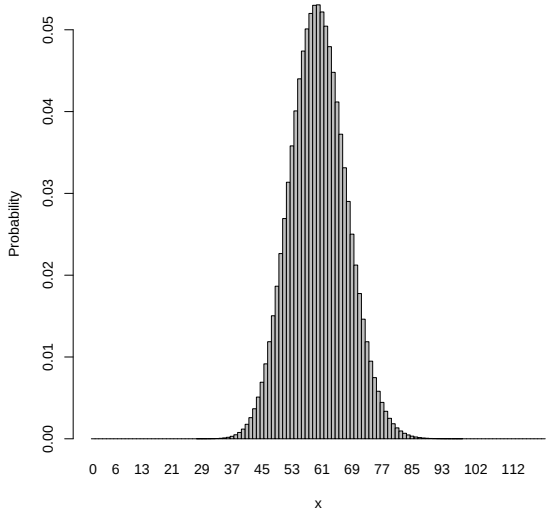

Figure 4.2.4: Probability Distribution of Y_n, for $p = 0.06$ and $n = 1000$

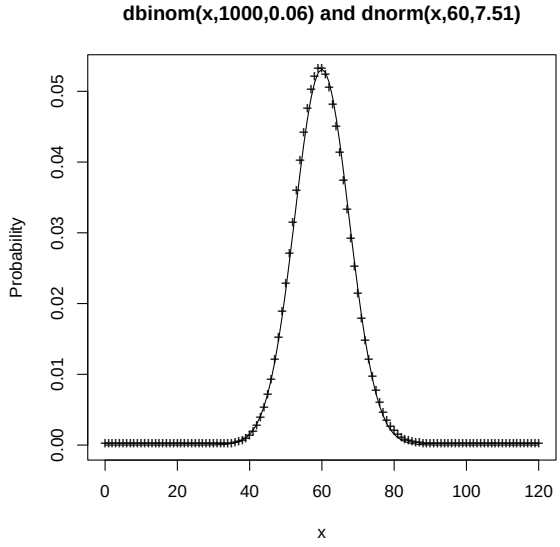

Figure 4.2.5: Probability Distribution of Y_n

of a continuous distribution known as the **normal distribution** with expected value np and variance $np(1-p)$ with $p = 0.06$ and $n = 1000$. This illustrates a very important and surprising theorem in Probability and Statistics: **The Central Limit Theorem**.

4.2.2 The Central Limit Theorem

The smooth curve in Figure 4.2.5 is very close to the graph of the function

$$f(x) = \frac{1}{\sqrt{2\pi}\ \sigma}\ e^{-(x-\mu)^2/2\sigma^2} \quad \text{for } x \in \mathbf{R},$$

where $\mu = 60$ and $\sigma^2 = 56.4$. This function is called the **density function** for a **normally distributed** random variable, X, with parameters μ and σ^2. We write

$$X \sim N(\mu, \sigma^2).$$

It can be shown that

$$\int_{-\infty}^{\infty} f(x)\ \mathrm{d}x = 1.$$

The density function can be used to compute probabilities of events $(a < X \leqslant b)$ for real numbers a and b, with $a < b$ as follows

$$P(a < X \leqslant b) = \int_a^b f(x)\ \mathrm{d}x.$$

We then get that the cumulative distribution function of X is given by

$$F_X(x) = P(X \leqslant x) = \int_{-\infty}^{x} f(t)\ \mathrm{d}t \quad \text{for all } x \in \mathbf{R}.$$

In R, the cumulative distribution function for a $N(\mu, \sigma^2)$ random variable may be estimated by the `pnorm()` function as follows

$$F_X(x) = \texttt{pnorm}(x, \mu, \sigma).$$

Note that we input the standard deviation, σ, and not variance, σ^2. A graph of F_X may be obtained in R by typing

```
curve(pnorm(x,60,7.51),0,120)
```

The graph of F_X for $\mu = 60$ and $\sigma^2 = 56.4$ is shown in Figure 4.2.6.

Recall that once we know the cumulative distribution function, we can compute the probability of events $(a < X \leqslant b)$ by

$$P(a < X \leqslant b) = F_X(b) - F_X(a).$$

For example,

$$P(60 < X \leqslant 67.51) \approx \texttt{pnorm}(67.51, 60, 7.51) - \texttt{pnorm}(60, 60, 7.51) \approx 0.3413$$

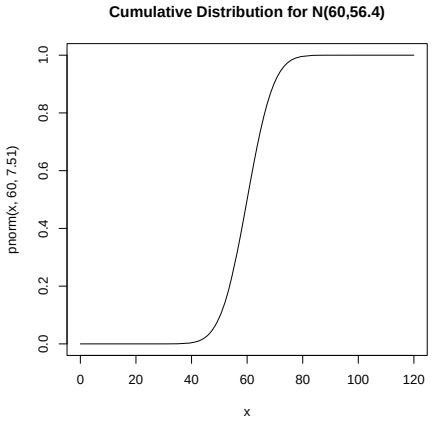

Figure 4.2.6: Cumulative Distribution for $X \sim N(60, 56.4)$

Let

$$X_1, X_2, X_3, \ldots, X_n, \ldots$$

denote a sequence of independent random variables all having the same distribution with expected value μ and variance σ^2; that is,

$$E(X_i) = \mu, \quad \text{and} \quad \text{Var}(X_i) = \sigma^2 \ \text{ for all } i = 1, 2, 3, \ldots.$$

The random variable

$$\overline{X}_n = \frac{X_1 + X_n + \cdots + X_n}{n}$$

is called the sample mean of the random sample

$$X_1, X_2, \ldots, X_n.$$

Note that

$$E(\overline{X}_n) = \mu,$$

since all the X_i's have expected value μ. It is also possible to show, by the independence of the random variables, that

$$\text{Var}(\overline{X}_n) = \frac{\sigma^2}{n}.$$

The Central Limit Theorem states that, for all real numbers a and b with $a < b$,

$$\lim_{n\to\infty} P\left(a < \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \leqslant b\right) = \int_a^b f_Z(z)\, \mathrm{d}z, \tag{4.2}$$

where f_Z is the density function for a normal random variable with expected value 0 and variance 1; that is, Z is a random variable with a probability density

function

$$f_Z(z) = \frac{1}{\sqrt{2\pi}}\, e^{-z^2/2} \quad \text{for } z \in \mathbf{R}.$$

Observe that Equation (4.2) can be written us

$$\lim_{n\to\infty} P\left(a < \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \leqslant b\right) = P(a < Z \leqslant b),$$

which in turn yields the approximation

$$P\left(a < \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \leqslant b\right) \approx P(a < Z \leqslant b) \quad \text{for large } n. \tag{4.3}$$

The probability on the right–hand side of (4.3) is independent of the parameters μ and σ^2 and can be estimated in R using

$$P(a < Z \leqslant b) \approx \texttt{pnorm}(b, 0, 1) - \texttt{pnorm}(a, 0, 1).$$

For example

$$P(-1.96 < Z \leqslant 1.99) \approx \texttt{pnorm}(1.96, 0, 1) - \texttt{pnorm}(-1.99, 0, 1) \approx 0.9500,$$

or about 95%.

Example 4.2.2 (Normal Approximation to the Binomial Distribution) Let $X_1, X_n, \ldots, X, \ldots$ be a sequence of independent Bernoulli trials with parameter p, for $0 < p < 1$. Then,

$$E(X_i) = p, \quad \text{and} \quad \text{Var}(X_i) = p(1-p) \ \text{ for all } i = 1, 2, 3, \ldots$$

According to the Central Limit Theorem (see Equation (4.3)), for large values of n,

$$P\left(a < \frac{\overline{X}_n - p}{\sqrt{p(1-p)}/\sqrt{n}} \leqslant b\right) \approx P(a < Z \leqslant b). \tag{4.4}$$

Observe that

$$\frac{\overline{X}_n - p}{\sqrt{p(1-p)}/\sqrt{n}} = \frac{n\overline{X}_n - np}{\sqrt{np(1-p)}},$$

where $n\overline{X}_n = Y_n \sim B(n, p)$. It then follows from (4.4) that

$$P\left(a < \frac{Y_n - np}{\sqrt{np(1-p)}} \leqslant b\right) \approx P(a < Z \leqslant b) \quad \text{for large values of } n, \tag{4.5}$$

where $Y_n \sim B(n, p)$. In other words, if we look at the distribution of the values of deviations of the values of $Y_n \sim B(n, p)$ from the mean or center of the distribution, μ, scaled by the standard deviation, $\sqrt{np(1-p)}$, for large values of

n, then distribution is very closed to that of a $N(0,1)$ random variable. Loosely speaking, the distribution of the random variables

$$\frac{Y_n - np}{\sqrt{np(1-p)}}$$

for large values of n is close to the distribution of $Z \sim N(0,1)$. Thus, the distribution of Y_n for large values of n is close to the distribution of

$$np + \sqrt{np(1-p)}Z$$

which has a normal distribution with mean np and variance $np(1-p)$. This is illustrated in Figure 4.2.7.

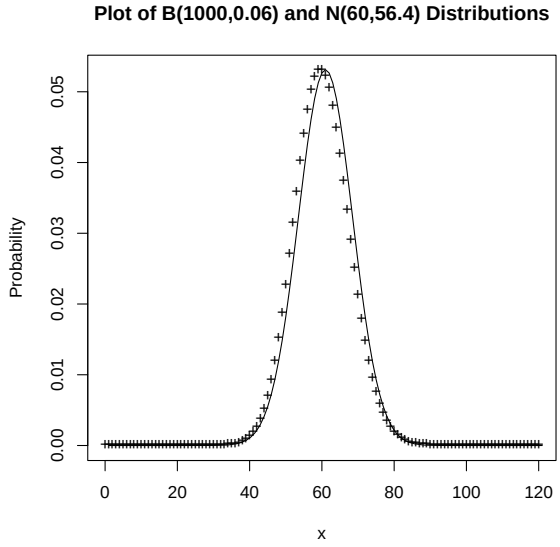

Figure 4.2.7: Probability Distribution of $Y_n \sim B(1000, 0.06)$ and Normal Approximation

4.2.3 Confidence Interval for Proportions

Consider the problem of estimating the number of fish in a fish population of size N by the capture–tag–recapture sampling method. Suppose that we have captured and tagged M fish and return them to the lake. Then, the proportion of tagged fish in the lake is

$$p = \frac{M}{N}.$$

Suppose that we collect a sample of size n of fish from the lake with replacement. Each time we collect a fish, we note whether the fish is tagged or not and we

then put the fish back. We define the random variables

$$X_1, X_2, X_3, \ldots, X_n, \ldots$$

as follows: for each $i = 1, 2, 3 \ldots$, $X_i = 1$ if the fish is tagged and $X_i = 0$ if it is not. Then, the sample proportion

$$\widehat{p}_n = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

is a point estimate for p. We would like to obtain an interval estimate for p of the form

$$(\widehat{p}_n - \delta, \widehat{p}_n + \delta)$$

where δ is a positive number which is chosen so that

$$P(\widehat{p}_n - \delta < p < \widehat{p}_n + \delta) \approx 0.95,$$

for instance, or any other pre–assigned large probability.

By the Central Limit Theorem (see Equation (4.4)), for large samples of size n,

$$P\left(a < \frac{\widehat{p}_n - p}{\sqrt{p(1-p)}/\sqrt{n}} \leqslant b\right) \approx P(a < Z \leqslant b). \tag{4.6}$$

for any real numbers a and b with $a < b$.

Observe that if

$$\widehat{p}_n - \delta < p < \widehat{p}_n + \delta,$$

then

$$-\delta < p - \widehat{p}_n < \delta,$$

or

$$-\delta < \widehat{p}_n - p < \delta.$$

Dividing the last inequality by $\sqrt{p(1-p)}/\sqrt{n}$ yields

$$-\frac{\delta}{\sqrt{p(1-p)}/\sqrt{n}} < \frac{\widehat{p}_n - p}{\sqrt{p(1-p)}/\sqrt{n}} < \frac{\delta}{\sqrt{p(1-p)}/\sqrt{n}}.$$

We then have by (4.6) that, for large values of n,

$$\begin{aligned} P(\widehat{p}_n - \delta < p < \widehat{p}_n + \delta) &= P\left(-\frac{\delta}{\sqrt{p(1-p)}/\sqrt{n}} < \frac{\widehat{p}_n - p}{\sqrt{p(1-p)}/\sqrt{n}} < \frac{\delta}{\sqrt{p(1-p)}/\sqrt{n}}\right) \\ &\approx P\left(-\frac{\delta}{\sqrt{p(1-p)}/\sqrt{n}} < Z < \frac{\delta}{\sqrt{p(1-p)}/\sqrt{n}}\right), \end{aligned}$$

where $Z \sim N(0, 1)$. We now use the Law of Large Numbers to get a further approximation

$$P(\widehat{p}_n - \delta < p < \widehat{p}_n + \delta) \approx P\left(-\frac{\delta}{\sqrt{\widehat{p}_n(1-\widehat{p}_n)}/\sqrt{n}} < Z < \frac{\delta}{\sqrt{\widehat{p}_n(1-\widehat{p}_n)}/\sqrt{n}}\right)$$

Thus, to get

$$P(\widehat{p}_n - \delta < p < \widehat{p}_n + \delta) \approx 0.95,$$

we may choose δ so that

$$\frac{\delta}{\sqrt{\widehat{p}_n(1-\widehat{p}_n)}/\sqrt{n}} = 1.96.$$

We then get that

$$\delta = 1.96\frac{\sqrt{\widehat{p}_n(1-\widehat{p}_n)}}{\sqrt{n}}.$$

This yields an **approximate** 95% **confidence interval** for p:

$$\left(\widehat{p}_n - 1.96\frac{\sqrt{\widehat{p}_n(1-\widehat{p}_n)}}{\sqrt{n}}, \widehat{p}_n + 1.96\frac{\sqrt{\widehat{p}_n(1-\widehat{p}_n)}}{\sqrt{n}}\right).$$

Example 4.2.3 In the fish–in–the–lake activity in class, we collected a sample of size $n = 100$ with replacement and counted $t = 5$ tagged fish. We then have that

$$\widehat{p}_n = \frac{5}{100} = 0.05$$

is a point estimate for the true, unknown, proportion p of tagged fish in the lake. A 95% confidence interval for the true proportion p is then

$$\left(\widehat{p}_n - 1.96\frac{\sqrt{\widehat{p}_n(1-\widehat{p}_n)}}{\sqrt{n}}, \widehat{p}_n + 1.96\frac{\sqrt{\widehat{p}_n(1-\widehat{p}_n)}}{\sqrt{n}}\right),$$

where

$$1.96\frac{\sqrt{\widehat{p}_n(1-\widehat{p}_n)}}{\sqrt{n}} \approx 0.043.$$

Thus, the 95% confidence interval for p is $(0.05-0.043, 0.05+0.043)$ or $(0.007, 0.093)$.

Definition 4.2.4 (Confidence Interval[1]) A level C confidence interval for a parameter is an interval computed from sample data by a method that has a probability C of producing an interval containing the true value of the population parameter.

In the previous example we produced the interval estimate $(0.007, 0.093)$ for the proportion of tagged fish in the fish–in–the–lake activity based on a sample size 100 in which there were 5 tagged fish. The interval was obtained by a procedure with an approximate probability of 95% of producing an interval that contains the true population proportion. In other words, on average, 95% of the intervals computed by the procedure will contain the true population proportion.

[1]This definition was taken from page 359 in Moore, McCabe abd Graig, *Introduction to the Practice of Statistics,* Sixth Edition.

The 95% confidence interval for the proportion of tagged fish in the lake produced in the previous example allows us to obtain a corresponding 95% confidence interval for the population size N. We do this by using the relation

$$N = \frac{M}{p},$$

where $M = 100$ in our case.

From the inequality

$$0.007 < p < 0.093,$$

we obtain

$$\frac{1}{0.007} > \frac{1}{p} > \frac{1}{0.093} \quad \text{or} \quad \frac{1}{0.093} < \frac{1}{p} > \frac{1}{0.007}.$$

Multiplying by $M = 100$ we then obtain that

$$1075 < N < 14,286.$$

Thus, a 95% confidence interval for the number of fish in the lake is $(1075, 14286)$.

To obtain a 90% confidence interval for the proportion of tagged fish in the lake, we use

$$\left(\widehat{p}_n - z_* \frac{\sqrt{\widehat{p}_n(1-\widehat{p}_n)}}{\sqrt{n}}, \widehat{p}_n + z_* \frac{\sqrt{\widehat{p}_n(1-\widehat{p}_n)}}{\sqrt{n}}\right),$$

where z_* is a positive number with the property that

$$\mathrm{P}(-z_* < Z \leqslant z_*) \approx 0.9,$$

where $Z \sim N(0,1)$. Using R we find

$$P(-1.6449 < Z \leqslant 1.6449) \approx \texttt{pnorm}(1.6449, 0, 1) - \texttt{pnorm}(-1.6449, 0, 1) \approx 0.9000.$$

Thus, we can take z_* to be about 1.6449 or 1.645. We then obtain that the 90% confidence interval for the true proportion of tagged fish in the lake is

$$\left(\widehat{p}_n - 1.645 \frac{\sqrt{\widehat{p}_n(1-\widehat{p}_n)}}{\sqrt{n}}, \widehat{p}_n + 1.645 \frac{\sqrt{\widehat{p}_n(1-\widehat{p}_n)}}{\sqrt{n}}\right),$$

where $\widehat{p}_n = 0.05$ and $n = 100$, which yields

$$(0.05 - 0.036, 0.05 + 0.036) \quad \text{or} \quad (0.014, 0.086).$$

This yields the following 90% confidence interval for the true population size in the fish–in–the–lake activity:

$$1163 < N < 7143.$$

Thus, the length of the interval decreases as the confidence level decreases. Another way to decrease the length of the interval is to increase the sample size.

4.3 Sampling from a uniform distribution

The Central Limit Theorem states that if

$$X_1, X_2, \ldots, X_n, \ldots$$

are independent observations from distribution with mean μ and variance σ^2, then, for large values of n, the sample mean

$$\overline{X}_n = \frac{X_1 + X_n + \cdots + X_n}{n}$$

has a distribution which is approximately normal with mean μ and variance $\dfrac{\sigma^2}{n}$; that is, approximately,

$$\overline{X}_n \sim N(\mu, \sigma^2/n) \quad \text{for large values of } n.$$

As another illustration of the Central Limit Theorem, we present here an example in which we sample from a continuous distribution whose probability density function is

$$f(x) = \begin{cases} 1 & \text{if } 0 < x < 1; \\ 0 & \text{otherwise.} \end{cases}$$

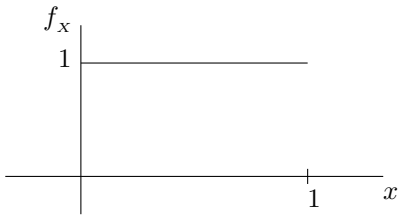

Figure 4.3.8: Probability Density Function for $X \sim U(0, 1)$

Note that

$$\int_{-\infty}^{\infty} f(x)\ \mathrm{d}x = \int_0^1 1\ \mathrm{d}x = 1;$$

so that f is indeed the density function of a continuous random variable, X. Thus, to compute the probability of the event $a < X \leqslant b$ we evaluate

$$\int_a^b f(x)\ \mathrm{d}x.$$

For example, to determine the probability that X lies between 0.1 and 0.47, we compute

$$P(0.1 < X \leqslant 0.47) = \int_{0.1}^{0.47} f(x)\ \mathrm{d}x = 0.37,$$

or 37%.

If X is a random variable having the probability density f, we say that X has a **uniform distribution** over the interval $(0, 1)$. We write $X \sim U(0, 1)$ and denote the density function of X by f_X (see Figure 4.3.8). The expected value of X is computed as follows

$$E(X) = \int_{-\infty}^{\infty} x f_X(x)\, \mathrm{d}x = \int_0^1 x\, \mathrm{d}x = \frac{1}{2}.$$

Similarly,

$$E(X^2) = \int_{-\infty}^{\infty} x^2 f_X(x)\, \mathrm{d}x = \int_0^1 x^2\, \mathrm{d}x = \frac{1}{3}.$$

It then follows that the variance of X is given by

$$\mathrm{Var}(X) = E(X^2) - [E(X)]^2 = \frac{1}{3} - \frac{1}{4} = \frac{1}{12}.$$

We then have that the mean of X is $\mu = 1/2$ and its variance is $\sigma^2 = 1/12$.

Let

$$X_1, X_2, \ldots, X_n, \ldots$$

be independent random variables all having a $U(0, 1)$ distribution. These corresponds to selecting numbers from the interval $(0, 1)$ at random. In R this can be done using the `runif()` function. Typing

```
runif(n,0,1)
```

yields a sample of n random numbers between 0 and 1. We can simulate collecting many random samples of size n and computing the mean of each sample. A histogram of the samples will then give us an idea of what the sampling distribution for the sample mean,

$$\overline{X}_n = \frac{X_1 + X_n + \cdots + X_n}{n},$$

looks like. The following R code generates 10000 of those samples of size $n = 100$, computes their means and stores them in a vector called `Xbar`.

```
n <- 100
Nrep <- 10000    # number of repetitions
Xbar <-  mean(runif(n,0,1))
                 # sets initial value of Xbar
L <- Nrep -1     # we need Nrep - 1 more repetitions
for (i in 1:L)  # Sets up a loop from 1 to L
{
Xbar <- c(Xbar, mean(runif(n,0,1)))
}
```

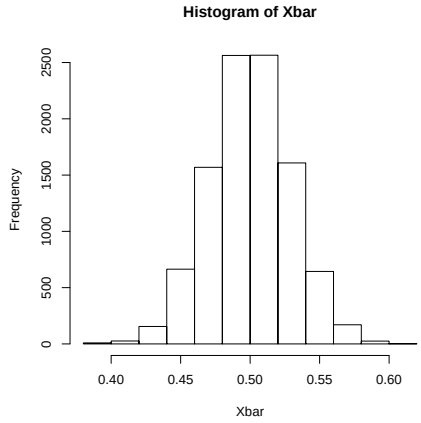

Figure 4.3.9: Histogram of `Xbar`

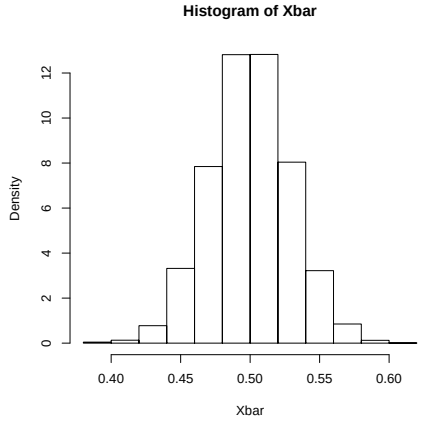

Figure 4.3.10: Density histogram of `Xbar`

The histogram of `Xbar` is shown in Figure 4.3.9. Typing `hist(Xbar,freq=FALSE)` in R yields the "density" histogram for `Xbar` shown in Figure 4.3.10.

The density histogram in Figure 4.3.10 has the property that the area of the bars correspond to the probability that `Xbar` takes on values lying at the base of the bar. In this sense (i.e., in the sense of areas of the bars in the histogram), the picture in Figure 4.3.10 gives the probability distribution of `Xbar`.

To plot the normal approximation density given by the Central Limit Theorem, we need to plot the density function

$$f(x) = \frac{1}{\sqrt{2\pi}\ \sigma}\ e^{-(x-\mu)^2/2\sigma^2} \quad \text{for } x \in \mathbf{R},$$

where $\mu = 0.5$ and $\sigma^2 = \dfrac{1}{1200}$. To superimpose the graph of f on the density plot in Figure 4.3.10, type the following commands in R:

```
x <- seq( 0.38, 0.62, length = 1000)
lines(x,dnorm(x,0.5,0.0288675))
```

where 0.0288675 is (approximately) the standard deviation of $\overline{X}_n$, for $n = 100$. The graph is shown in Figure 4.3.11.

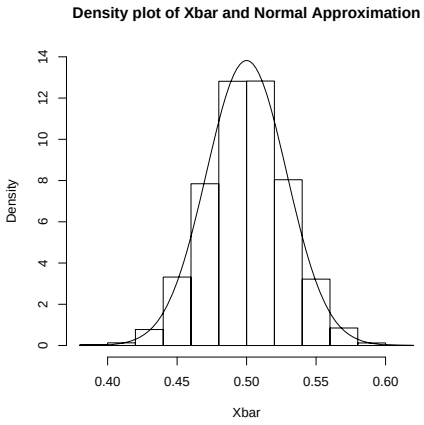

Figure 4.3.11: Density histogram of `Xbar` and Approximating Normal Density

Chapter 5

Goodness of Fit

When estimating parameters for a population based on sample data, we usually make assumptions as to what the distribution of a certain random variable is. For example, in the fish–in–the–lake activity, we assumed that the distribution of the number of tagged fish in a sample of size n followed a binomial distribution with parameters n and p, where p is the population parameter we are trying to estimate (i.e., the true proportion of tagged fish in the lake). Based on that assumption and using the Law of Large Numbers and the Central Limit Theorem, we were able to obtain a confidence interval estimate for p.

Sometimes, we are interested in knowing to what degree the distribution of a random variable follows a some theoretical distribution. For example, is the distribution of tagged fish in a sample truly binomial? Or, going further into the sampling procedure, is it true that the distribution of the outcomes of a single draw of a fish is truly a Bernoulli trial with parameter p?

There is a procedure known as a **goodness of fit test** that allows us to determine how close (or how far) a given distribution is from a hypothetical one based on information contained in a sample. We illustrate the use of this procedure in the following class activity.

5.1 Activity #5: How Typical are our Households' Ages?

Introduction[1]

Consider a sample made up of the people in the households of each pearson in our class. Would their ages be representative of the ages of all households in the United States? Probably not. After all, it would not be a random sample. But how unrepresentative would the sample be with regard to ages of people in the United States?

[1] Adapted from *Activity–Based Statistics,* Second Edition, by Scheaffer, Watkins, Witmer and Gnanadesikan. Key College Publishing, 2004

Question

The latest census in the United States yielded the following age distribution[2]:

Age Range	Proportion in US Population
0–18	0.26
18–64	0.62
64+	0.12

Table 5.1: US 2000 Census: Household Age Distribution

Is the age distribution of the people from the households in the class typical of that of all residents in the United States?

Discussion

1. Would you expect the age distribution from the households in your class to be typical of the age distribution of all US residents? Why or why not?

2. If there is any discrepancy between the two distributions, how would you measure that discrepancy?

Data Collection

In the following table record the number of people in households from this class for each age range. In the last column, assuming that the class distribution will follow the national distribution, record the number of people you would expect to see in each age range.

Age Range	Number Observed in Households from this Class	Expected Number in the Class
0–18		
18–64		
64+		
Total		

Table 5.2: Class Household Age Data

Analysis

[2]Source: `http://www.census.gov/prod/2001pubs/c2kbr01-12.pdf`

If we call the observed numbers in the second column O_1, O_2 and O_3, for each age range respectively, and E_1, E_2 and E_3 the corresponding expected values, then one way to measure the discrepancy between the two distributions is by computing the statistic

$$\frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} + \frac{(O_3 - E_3)^2}{E_3}.$$

This is usually denoted by χ^2 and is known as the *Chi–squared* statistic.

Is the χ^2 statistic you computed large enough to convince you that the ages of the households from this class are not similar to those of a typical random sample of US residents?

Simulations

Simulate drawing random samples of the ages from the population of US residents to see how often we get a value of the χ^2 statistic that is as large or larger than the one from the class.

Discuss how you would use R to run this simulation.

Conclusion

Based on your simulations, estimate the probability that a random sample of US residents (of the same size as the sample in the class) would have a χ^2 statistic as large or larger than the one obtained in class.

Is the distribution of ages of the households from the class similar to those of a typical random sample of US residents? What do you conclude?

5.2 The Chi–Squared Distance

The class data, as well as the expected values based on the USA age distribution shown in Table 5.1 and the sample size, are shown in Table 5.3.

Age Range	Number Observed in Households from this Class	Expected Number in the Class
0–18	24	36
18–64	110	87
64+	6	17
Total	140	140

Table 5.3: Class Household Age Data

The entries in the second column were obtained by multiplying the proportions in the USA Census data distribution by the sample size, which is 140. For

instance, the first entry in the second column is $(0.26) \cdot (140) = 36.4$, which we rounded to 36. In R, these calculations can be performed as follows:

First, define a vector of postulated proportions based on the USA 2000 Census distribution

```
Prop <- c(0.26,0.62,0.12)
```

Then, the expected values are obtained by typing

```
Exp <- round(140*Prop)
```

Typing `Exp` in R then yields

```
[1] 36 87 17
```

which are the first three entries in the third column in Table 5.3.

The Chi-Squared statistic

$$X^2 = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} + \frac{(O_3 - E_3)^2}{E_3},$$

gives a measure of how far the observed counts in each category are from the expected values predicted by the assumption that the class sample is representative of the USA population with regards to age distribution. This assumption constitutes our null hypothesis or null model; that is, the household members of the people in this class is a random sample from a populations having a distribution for categories Cat_1, Cat_2 and Cat_3 given by the proportions p_1, p_2 and p_3:

$$\text{H}_o\colon p_1 = 0.26,\ p_2 = 0.62,\ p_3 = 0.12.$$

In R, the Chi–Squared statistic may be computed as follows:

First, define a vector of observed values

```
Obs <- c(24,110,6)
```

Then,

```
ChiSqr <- sum((Obs-Exp)^2/Exp)
```

This yields a value of about 17.20. What are the chances that, if we sample repeatedly from the USA population, each time collecting 140 ages and computing the Chi–Squared statistic, we would obtain a value of 17.20 or higher? The answer to this question is the p–value for the test. We will first estimate the p–value by simulating the sampling procedure many times and computing the proportion of times that the Chi–Squared statistic is 17.2 or higher.

To perform the simulation, we will use the `runif()` function in R to generate samples of size $n = 140$ of values between 0 and 1; typing `sample <- runif(n,0,1)`, where $n = 140$, will generate a vector `sample1` of 140 values between 0 and 1. We will then count how many of the values in `sample1` are between 0 and $p_1 = 0.26$ and put the result into a category which we will call `SimObs1`. This will simulate the number of people in the sample with ages between 0 and 18. The following code in R will do this for us:

```
n <- 140
Prop <- c(0.26,0.62,0.12)

sample <- runif(n,0,1)

SimObs1 <- 0         # Sets up counter
for (i in 1:n)      # Sets up loop
{
    SimObs1 <- SimObs1 + (sample[i]>=0 & sample[i]<=Prop[1])
                  # adds 1 to counter1 if 0<= value <= Prop[1]
}
```

For the particular random sample that we obtained, this code yields a count of 34.

We can do something similar for the category of ages between 18 and 64. This time we use the code

```
SimObs2 <- 0         # Sets up counter
for (i in 1:n)      # Sets up loop
{
    SimObs2 <- SimObs2 + (sample[i]> Prop[1] & sample[i]<=Prop[1]+Prop[2])
                  # adds 1 to counter1 if p1<= value <= p1+p2
}
```

where we have used the values of `n`, `Prop` and `sample` defined previously.

This yields a count of 89. We don't need to perform the simulation for the third category since we know that it will be $n - 34 - 89$ or 17. We therefore obtain the simulated result shown in Table 5.4. The corresponding Chi–Squared

Age Range	Simulated value	Expected Value
0–18	34	36
18–64	89	87
64+	17	17
Total	140	140

Table 5.4: Age distribution for one simulated sample

statistic may be obtained by typing:

```
SimObs <- c(34,89,17)
ChiSqr1 <- sum((SimObs-Exp)^2/Exp)
```

which yields a Chi–Squared value of about 0.16. We see the value is much lower than the observed one. If we repeat the simulated sampling procedure many times, we can get an estimate of the probability that the Chi–Squared statistic is as high or higher than the observed value of 17.2.

We can automate the sampling procedure in R by generating many samples, computing the Chi–squared statistic for each sample and storing them in a vector, which we shall also call `ChiSqr`. The distribution of this vector will approximate the distribution of the Chi–Squared statistic under the assumption that the samples are randomly selected from the US population. A code performing this simulation is given in the Appendix A.4.

A histogram of `ChiSqr` is shown in Figure 5.2.1.

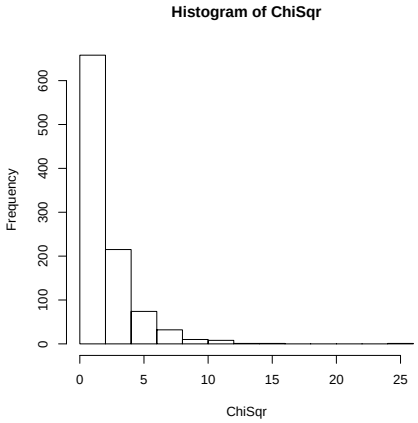

Figure 5.2.1: Histogram of `ChiSqr`

In order to estimate the p–value for the test of hypothesis that the age distribution in our class follows the distribution of the entire US population, we may use the `pHat()` function that we have defined previously in the course. We obtain

$$p\text{–value} \approx \texttt{pHat(ChiSqr,17.2)} \approx 0.001$$

or 0.1%. Thus, it is highly unlikely that the age distribution of our class sample is truly a random sample of the entire US population.

5.3 The Chi–Squared Distribution

Figure 5.3.2 shows that density distribution of the `ChiSqr` vector that resulted from a simulation like the one described in the previous section. Typing the commands

```
x <- seq(0,25,length =1000)
lines(x,dchisq(x, 2),col="blue")
```

superposes a density curve on the plot in 5.3.2. The output is shown in Figure 5.3.3. The curve in Figure 5.3.3 is the graph of the function given by

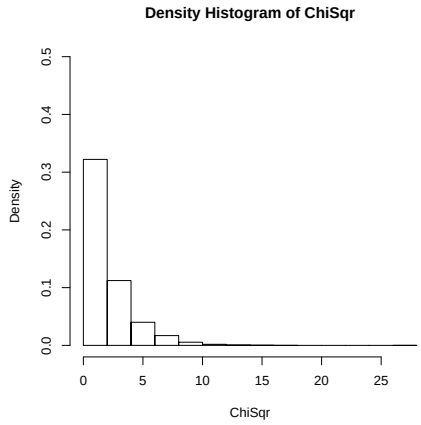

Figure 5.3.2: Density Histogram of `ChiSqr`

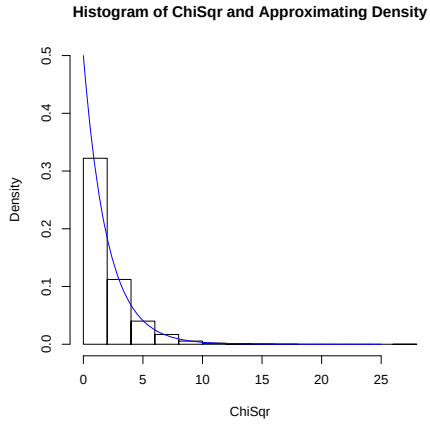

Figure 5.3.3: χ^2_2 Density Curve

$$f(x) = \begin{cases} \frac{1}{2}e^{-x/2} & \text{if } x > 0; \\ \\ 0 & \text{otherwise.} \end{cases}$$

This is the probability density function for what is known as a Chi–Squared random variable, X, with 2 **degrees of freedom**; we write $X \sim \chi^2_2$. In R, the density function f is given, approximately, by

`dchisq(x,2)` for all x.

Figure 5.3.3 illustrates the fact that, for large sample sizes, the density of χ^2_2 distribution approximates the distribution of the vector `ChiSqr` obtained in the simulations of the age distribution activity.

In a goodness of fit test, the degrees of freedoms parameter has to do with the number of categories that we are trying to fit in a model. If k denotes the number of categories, them the number of degrees of freedom is $k-1$. Thus, for a goodness of fit test with k categories, for large sample sizes, the distribution of `ChiSqr` is approximated by a χ^2_{k-1} distribution. Figures 5.3.4 and 5.3.5 we plot the density function for χ^2_1 and χ^2_2 random variables, respectively. The

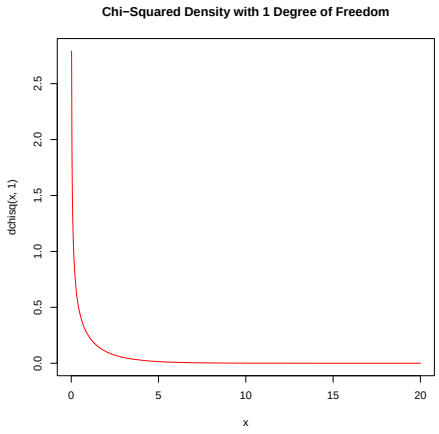

Figure 5.3.4: χ^2_1 Density Curve

corresponding density functions are

$$f_{\chi_1}(x) = \begin{cases} \frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{x}}e^{-x/2} & \text{if } x > 0; \\ \\ 0 & \text{otherwise;} \end{cases}$$

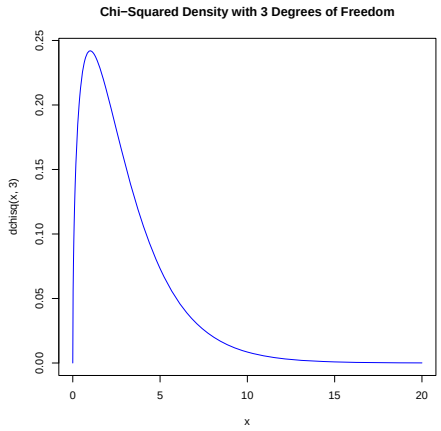

Figure 5.3.5: χ^2_3 Density Curve

$$f_{\chi_3}(x) = \begin{cases} \dfrac{1}{\sqrt{2\pi}}\sqrt{x}\ e^{-x/2} & \text{if } x > 0; \\ \\ 0 & \text{otherwise;} \end{cases}$$

respectively.

The formula for the probability density function of χ^2_4 random variable is

$$f_{\chi_4}(x) = \begin{cases} \dfrac{1}{4}x\ e^{-x/2} & \text{if } x > 0; \\ \\ 0 & \text{otherwise;} \end{cases}$$

and its graph is shown in Figure 5.3.6.

The χ^2_{k-1} distribution can be used to estimate the p–value for the case in which the sample size is so large, and the proportions postulated in the null hypothesis ar such that the **expected values are at least** 5. For example, in the age–distribution class activity, with $X^2 \sim \chi^2_2$, we have that

$$p\text{–value} = P(X^2 \geqslant 17.2) \approx P(\chi^2_2 \geqslant 17.2).$$

Since we know the probability density function, $f(x)$, for a χ^2_2 random variable, we can compute the last probability by integration

$$\begin{aligned} P(\chi^2_2 \geqslant 17.2) &= \int_{17.2}^{\infty} \frac{1}{2} e^{-x/2} \mathrm{d}x \\ &= \left[-e^{-x/2}\right]_{17.2}^{\infty} \\ &= e^{-8.6} \approx 1.84 \times 10^{-4} \end{aligned}$$

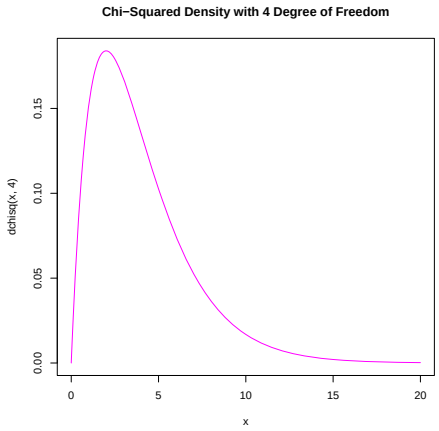

Figure 5.3.6: χ^2_4 Density Curve

which, again, is a very small probability.

In R, we can estimate the p–value using the `pchisq()` function as follows

$$p\text{–value} \approx \texttt{1 - pchisq(17.2,2)} \approx 1.84 \times 10^{-4}.$$

5.4 Randomization for Chi–Squared Tests

In this section we present a more efficient code in R to perform the goodness of fit test that we did in Section 5.2 for the class households age distribution data. It is based on the `table()` function in R, which provides a table of counts for categorical data. The code has the added advantage that can be easily modified for other tests involving the Chi–Squared statistics.

To see how the `table()` function works, consider the situation described in Section 5.2 in which there are three categories of ages which may be labeled "1" for ages between 0 and 18, "2" for those between 19 and 64, and "3" for 65 or above. The proportions of those categories in the entire US population are 26 : 62 : 12, which may be reduced to 13 : 31 : 6. We define a vector made up of the digits 1, 2 and 3 in proportions 13 : 31 : 6, respectively, as follows:

```
Cat1 <- rep(1,13)
Cat2 <- rep(2,31)
Cat3 <- rep(3,6)
categories <- c(Cat1,Cat2,Cat3)
```

This yields the vector whose proportions of 1s ,2s and 3s are 0.26, 0.62 and 0.12, respectively. The same effect can be achieved by simply typing

```
categories <- rep(c(1,2,3),c(13,31,6))
```

We then use the `sample()` function in R to obtain a random sample of size 140, with replacement, from the vector `categories`:

```
s <- sample(categories,140,replace = TRUE)
```

Typing `table(s)` yields the following out put in R:

```
 s
 1  2  3
30 91 19
```

showing that there are 30 values in category 1, 91 in category 2, and 19 in category 3. The values in this table can then be used to compute the Chi–Square statistics as follows:

First, compute the expected values

```
Exp <-  140*c(13,31,6)/50
```

Then, store the simulated counts into a vector called `Sim`:

```
Sim <-  table(s)
```

Finally, compute the Chi–Square statistic by typing

```
ChiSqr <- sum((Sim-Exp)^2/Exp)
```

This process can be repeated many times, `Nrep`, to obtain a vector called `ChiSqr` containing the Chi–Squared values for each of the runs. The distribution of `ChiSqr` will approximate the sampling distribution of the Chi–Squared distance under the assumption that sampling is done from a distribution having the proportions postulated in the null model. Here is the code that will generate `ChiSqr`:

```
Nrep <- 10000
n <- 140
Prop <- c(0.26, 0.62,0.12)
Exp <- round(n*Prop)
    # Computes expected values based on Prop and n
categories <- rep(c(1,2,3),c(13,31,6))
ChiSqr <- array(dim=Nrep) # sets up array ChiSqr
for (i in 1:Nrep)         # sets up loop to generate ChiSqr
{
    s <- sample(categories,n,replace = TRUE)
        # generates sample of size n from categories
    Sim <-  table(s) # Tabulates counts in each group
    ChiSqr[i] <- sum((Sim-Exp)^2/Exp)
        # Computes Chi-Squared Statistic
}
```

The histogram of `ChiSqr` is very similar to the one in Figure ChiSqrSimAgeDistHist1. We can estimate the p–value associated with the test using the `pHat` function:

$$p\text{–value} \approx \texttt{pHat(ChiSqr,17.2)} \approx 0.0001$$

or about 0.01%; a very small p–value telling us that the evidence is very strong against the null hypothesis that the age distribution in the class sample follows that of the entire US population. Hence, the class sample is not representative of the entire use population regarding age distribution.

5.5 More Examples of Randomization for Chi–Squared Tests

Example 5.5.1 *(Vehicle collisions and cell phones)*[3] In study done by Donald A. Redelmeier, M.D., and Robert J. Tibshirani, Ph.D., in 1979, which appeared in the New England Journal of Medicine (*Association between Cellular–Telephone Calls and Motor Vehicle Collisions,* Volume 336, No. 7, pp. 453-458) the following data relating to number of collisions and days of the week were presented:

```
  Day  Count
1 Sun    20
2 Mon   133
3 Tue   126
4 Wed   159
5 Thu   136
6 Fri   113
7 Sat    12
```

We would like to answer the question: Is the number of collisions associated with the day of the week? The data suggest that collisions are more frequent during weekdays. Is the evidence suggested by the data statistically significant? We will answer these questions by doing a goodness of fit test to the null model that collations are equally likely to occur on any day of the week:

$$\mathrm{H}_o : p_i = \frac{1}{7} \quad \text{for all } i = 1, 2, \ldots, 7.$$

We define a vector in R, called `Obs`, which has observed counts for each day of the week. Set `k=length(Obs)`, the number of categories (which in this case is 7). The proportions dictated by the null hypothesis can then be defined in R as follows

```
Prop <- rep(1/k,k)
```

[3]Adapted from Example 9.14 in Moore, McCabe and Graig, *Introduction to the Practice of Statistics,* Sixth Edition, p. 545

that is, `1/k` repeated `k` times. Set `n <- sum(Obs)`; then, `n`, is the sample size. We can then define the vector of expected values by

```
Exp <- n*Prop
```

Next, we select a sample of size `n` with replacement from the set of categories $\{1, 2, 3, 4, 5, 6, 7\}$ by using

```
days <- 1:k
s <- sample(days,n,replace = TRUE)
```

This simulates collecting a sample with information on which days collisions occurred. We then use the `table()` function in R to count how many accidents occurred in each day. For example, we might get

```
> table(s)
s
  1   2   3   4   5   6   7
103  86  98 105 115 104  88
```

This is a possible distribution of collisions in week under the null hypothesis. We store this distribution in a vector called `Sim` by typing

```
Sim <-  table(s)
```

We can then compute the Chi–Squared value for this particular simulation by typing

```
sum((Sim-Exp)^2/Exp)
```

and obtain, for this particular case, 6.197425 or about 6.2.

This process can be automated in R and repeated as many times as needed. Here is the code that runs the simulations

```
# Simulations for Example 9.14 in the text
Nrep <- 10000
n <- sum(Obs)
k <- length(Obs)
Prop <- rep(1/k,k)  # Null Model:
                    #collisions are equally likely in each day
Exp <- n*Prop     # Computes expected values based on Prop and n
days <- 1:k       # Labels of groups
ChiSqr <- array(dim=Nrep) # sets up array ChiSqr
for (i in 1:Nrep)         # sets up loop to compute ChiSqr
{
    s <- sample(days,n,replace = TRUE) # randomization sample
    Sim <-  table(s) # Counts in each group
    ChiSqr[i] <- sum((Sim-Exp)^2/Exp) # Computes Chi-Squared Statistic
}
```

This yields a vector `ChiSqr` with the simulated Chi–Squared values.

The Chi–Squared statistic for the observed values is

```
Xsqr <- sum((Obs-Exp)^2/Exp)
```

or about 208.85

The p–value for the goodness of fit test can then be estimated using the `pHat` function:

```
pHat(ChiSqr,Xsqr)
```

which yields 0. Hence, the data are significant and conclude that the number of collisions is associated with the day of the week. In fact, collisions are more frequent during week days.

We next perform the test week days. For this part of the test, we change the observations vector, `Obs`, to contain the counts only for Monday through Friday. This is achieved in R by typing

```
Obs <- Obs[2:6]
```

For this set of observations, the Chi–Squared statistic is 8.494753 or about 8.49. The corresponding p–value is then about 0.0761, which is bigger than the 5% threshold for significance. We therefore conclude that the weekdays data do not provide evidence in support of differences due to day of the week.

Example 5.5.2 (Challenger Disaster Data) On January 28, 1986, the space shuttle Challenger exploded shortly after take-off and disintegrated over the Atlantic Ocean. It was later discovered that the explosion was due to fuel leakage caused by an O–ring seal failure. The night before, engineers at the company which built the shuttle warned NASA that the launch should be postponed due to predicted low temperatures. Low temperatures had been linked to O–ring seals failures. The data in Table 5.5 are taken from a graph in Richard P. Feynman's autobiographical book: *What do you care what other people think?* (New York: W. W. Norton, 1988). Feynman was a member of the commission that investigated the accident.

Launch Temperature																				
Below 65°F	1	1	1	3																
Above 65°F	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	2

Table 5.5: Number of O–ring failures in 24 space shuttle flights prior to 1/28/1986

The question of interest here is: Are O–ring seal failures associated with low temperatures.

In order to analyze the data using a Chi–Square test, it is helpful to reorganize the information into a table like that shown in Table 5.6.

Failures $\geqslant 1$? \ Temperature	Low	High	Total
Yes	4	3	7
No	0	17	17
Total	4	20	24

Table 5.6: Observed Values

The null model in this situation is that there is no association between the number of failures in a launch and the temperature in the day of the lauch. Thus, failures occur at random regardless of temperature.

To find expected counts based on the null hypothesis, we may proceed as follows:

Let X denote the number of failures that may occur in a day with low temperature. The possible values for X are 0, 1, 2, 3 and 4. To find the probability distribution for X, we compute

$$P(X=0) = \frac{\binom{17}{4}}{\binom{24}{4}} \approx 0.2239789;$$

$$P(X=1) = \frac{\binom{7}{1}\binom{17}{3}}{\binom{24}{4}} \approx 0.4479578;$$

$$P(X=2) = \frac{\binom{7}{2}\binom{17}{2}}{\binom{24}{4}} \approx 0.2687747;$$

$$P(X=3) = \frac{\binom{7}{3}\binom{17}{1}}{\binom{24}{4}} \approx 0.05599473;$$

$$P(X=4) = \frac{\binom{7}{4}\binom{17}{0}}{\binom{24}{4}} \approx 0.003293808;$$

where we have used the R command `dhyper(x,7,17,4)` for each $x = 0, 1, 2, 3, 4$, respectively. The expected value of X is then

$$E(X) = \sum_{k=0}^{4} k \cdot P(X=k)$$

which may be computed in R as follows

```
k <- 0:4
sum(k*dhyper(k,7,17,4))
```

This yields about 1.167. Thus, the entry in the "Yes" row and column and "Low" column of the expected values table in Table 5.7 is 1.167. We can fill out the rest of the entries by considering the marginal totals.

Failures $\geqslant 1$? \ Temperature	Low	High	Total
Yes	1.167	5.833	7
No	2.833	14.167	17
Total	4	20	24

Table 5.7: Expected Values

The Chi–Squared statistic in this case is then 11.65283.

To estimate the p–value of the test, we first generate samples of size 4 from a vector containing all the values in Table 5.5 and count how many values in the sample are 1 or greater. Alternatively, we may define the vector

```
launches <- rep(0:1,c(17,7))
```

which has seventeen 0s and seven 1s. Next, we sample

```
low <- sample(launches,4,replace=F)
```

to simulate four days with low temperature. Then `sum(low)` will yield the number of days with low temperature in which there was at least one O–ring failure. This will be the first entry in a vector that we will call `Sim` of entries in table of simulated values. The other entries are obtained as follows

```
Sim <- array(dim=4)
Sim[1] <- sum(low)
Sim[2] <- 7 -Sim[1]
Sim[3] <- 4 -Sim[1]
Sim[4] <- 13 + Sim[1]
```

Then, Chi–Squared values for each simulation are given by

```
sum((Sim-Exp)^2/Exp)
```

and these can be stored in vector called `ChiSqr`. The R code performing these simulations follows:

```
# Simulations for problem Challenger disaster data
Obs <- c(4,3,0,17)
launches <- rep(0:1,c(17,7))
Nrep <- 10000
```

```
Exp <-  c(1.167,5.833,2.833,14.167)
ChiSqr <- array(dim=Nrep)
for (i in 1:Nrep)
{
    Sim <- c(0,0,0,0)
    low <- sample(launches,4,replace=F)
            Sim[1] <- sum(low)
            Sim[2] <- 7 -Sim[1]
            Sim[3] <- 4 -Sim[1]
            Sim[4] <- 13 + Sim[1]
        ChiSqr[i] <- sum((Sim-Exp)^2/Exp)
}
```

The Chi–Squared test statistic for this test is 11.65283 and the corresponding p–value based on the simulations is 0.0024, or 0.24%. Consequently, the incidence of O–ring failures is associated with low temperatures.

Chapter 6

Association Between Categorical Variables

6.1 Two–Way Tables

Table 5.6, which we obtained in the analysis of the Challenger disaster data in Example 5.5.2 on page 72, is an example of a two–way table:

Failures $\geqslant 1$? \ Temperature	Low	High	Total
Yes	4	3	7
No	0	17	17
Total	4	20	24

Table 6.1: Two–Way Table for Challenger Disaster Data

Table 6.1 displays a relationship between temperature on the day of a launch and the incidence of O–ring seal failures. Both variables are displayed as categorical variables: temperature is either low or high (depending on whether it is below 65°F or above, respectively), and the incidence of O–ring failures is coded as "Yes" if there was at least one failure on the day of the launch, or "No" if there were no failures in that day. The main information contained in Table 6.1 are the counts of launches with more than one O–ring failure incident, or no incidents, as a function of the temperature (either low or high).

The question of interest in the analysis of two–way tables is whether there is an association between the variables, or whether they are independent. In Table 6.1 we see that the proportion of O–ring failure incidents is higher in low temperature days than in those with high temperatures. So, low temperatures seem to have a bearing on incidence of O–ring failures. To determine whether the difference in proportions is significant, we can perform a test in which the

null hypothesis is

$$\mathrm{H}_o\text{: there is no association between the two variables}$$

or, equivalently,

$$\mathrm{H}_o\text{: the two variables are independent}$$

In order to understand the null hypothesis for this test of significance, we need to recall the concept of independent random variables. We will also need to study joint probability distribution as well as marginal distributions.

6.2 Joint Distributions

In order to deal with probabilities for events involving more than one random variable, we need to define their **joint distribution**. In this section we only study the case of two discrete random variables X and Y.

Definition 6.2.1 (Joint probability distribution) Given discrete random variables X and Y, the **joint probability** of X and Y is defined by

$$p_{(X,Y)}(x,y) = \mathrm{P}(X = x, Y = y).$$

Example 6.2.2 Suppose three chips are randomly selected from a bowl containing 3 red, 4 white, and 5 blue chips.

Let X denote the number of red chips chosen, and Y be the number of white chips chosen. We would like to compute the joint probability function of X and Y; that is,

$$\mathrm{P}(X = x, Y = y),$$

were x and y range over 0, 1, 2 and 3.

For instance,

$$\begin{aligned}
\mathrm{P}(X = 0, Y = 0) &= \frac{\binom{5}{3}}{\binom{12}{3}} = \frac{10}{220} = \frac{1}{22}, \\
\mathrm{P}(X = 0, Y = 1) &= \frac{\binom{4}{1}\cdot\binom{5}{2}}{\binom{12}{3}} = \frac{40}{220} = \frac{2}{11}, \\
\mathrm{P}(X = 1, Y = 1) &= \frac{\binom{3}{1}\cdot\binom{4}{1}\cdot\binom{5}{1}}{\binom{12}{3}} = \frac{60}{220} = \frac{3}{11},
\end{aligned}$$

and so on for all 16 of the joint probabilities. These probabilities are more easily expressed in tabular form:

Notice that the probability distributions for the individual random variables X and Y can be obtained as follows:

$$p_{_X}(i) = \sum_{j=0}^{3} P(i,j) \leftarrow \text{ adding up } i^{th} \text{ row}$$

$X\backslash Y$	0	1	2	3	Row Sums
0	1/22	2/11	3/22	2/110	21/55
1	3/22	3/11	9/110	0	27/55
2	3/42	3/55	0	0	27/220
3	1/220	0	0	0	1/220
Column Sums	14/55	28/55	12/55	1/55	1

Table 6.2: Joint Probability Distribution for X and Y, $p_{(X,Y)}$

for $i = 1, 2, 3$, and

$$p_{Y}(j) = \sum_{i=0}^{3} P(i,j) \leftarrow \text{ adding up } j^{th} \text{ column}$$

for $j = 1, 2, 3$.

These are expressed in Table 6.2 as "row sums" and "column sums," respectively, on the "margins" of the table. For this reason p_X and p_Y are usually called **marginal distributions**.

Definition 6.2.3 (Independent Random Variables) The discrete random variable X and Y are said to be independent if and only if

$$p_{(X,Y)}(x,y) = p_X(x) \cdot p_Y(y)$$

for all values of x and y.

Observe that, in the previous example (see Table 6.2),

$$0 = \mathrm{P}(X = 3, Y = 1) \neq p_X(3) \cdot p_Y(2) = \frac{1}{220} \cdot \frac{28}{55},$$

and therefore X and Y are not independent.

6.3 Test of Independence

We now begin the test of the null hypothesis that temperature on the day of the launch and incidence of O–ring failures are independent based on the data in Table 6.1. Using the marginal totals in the table, we can obtain the marginal probabilities shown in Table 6.3. We have denoted temperature by X and incidence of O–ring failure by Y.

Note that we are displaying X and Y as categorical variables with values for X of "Low" or "High," and for Y of "Cat 1" for 1 or more failures, or "Cat 2" for no failures.

Assuming the null hypothesis that X and Y are independent, we obtain the joint probability distribution function by using the formula

$$p_{(X,Y)}(x,y) = p_X(x) \cdot p_Y(y)$$

Y \ X	Low	High	p_Y
Cat 1			7/24
Cat 2			17/24
p_X	1/6	5/6	1

Table 6.3: Marginal Distributions for Challenger Disaster Data

Y \ X	Low	High	p_Y
Cat 1	7/144	35/144	7/24
Cat 2	17/144	85/144	17/24
p_X	1/6	5/6	1

Table 6.4: Joint Distribution of X and Y

for all values of x and y. The results are shown in Table 6.4.

The joint probabilities in Table 6.4 can now be used to obtain the expected values, under the assumption that H_o is true, by multiplying by 24, the sample size. The expected values are shown in Table 6.5. Note that we could have

Y \ X	Low	High	Row Totals
Cat 1	7/6	35/6	7
Cat 2	17/6	85/6	17
Column Totals	4	20	24

Table 6.5: Expected Values

also obtained the values in Table 6.5 by multiplying the row and column totals corresponding to each cell and the dividing by the grand total. For example, the entry in the Cat 1/Low cell is

$$\frac{7 \cdot 4}{24} = \frac{7}{6}.$$

Once we have the expected values predicted by the null model, we can compute the Chi–Squared distance

$$X^2 = \sum \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}$$

from the observed data to the model.

In R, we can automate the calculations of expected values and Chi–Square distance my making use of the `matrix()` function and matrix operations in R. We start out by entering the observed data in Table 6.1 as a 2×2 matrix as follows

```
Obs <- matrix(c(4,3,0,17),nrow=2,byrow=T)
```

The outcome of this is `Obs` as a 2×2 matrix

```
> Obs
      [,1] [,2]
[1,]    4    3
[2,]    0   17
```

which is essentially the two–way table in Table 6.1. Typing `dim(Obs)` in R yields

```
> dim(Obs)
[1] 2 2
```

which gives the dimensions of the matrix `Obs` as a vector; the first entry of the vector gives the number of rows and the second the number of columns. We could have also entered the matrix `Obs` by specifying its entries in the vector `c(4,3,0,17)` and indicating its dimensions:

```
Obs <- matrix(c(4,3,0,17), 2,2,byrow=T)
```

The logical constant `byrow` has to be set to be TRUE since R enters matrices column by column by default.

Typing `rowSums(Obs)` in R yields the row totals in Table 6.1 as a vector. If we want to get a column vector; that is, a 2×1 matrix, we may use the `matrix()` function as follows

```
R <- matrix(rowSums(Obs))
```

This yields the column vector `R`

```
> R
     [,1]
[1,]    7
[2,]   17
```

To obtain a column vector of the column totals in Table 6.1, we type

```
C <- matrix(colSums(Obs),1,2)
```

indicating that we want a 1×2 matrix, or a row vector. The outcome of this is

```
> C
      [,1] [,2]
[1,]    4   20
```

The symbol `%*%` in R denotes the **matrix product**. Typing `R %*% C`, for instance, yields

```
> R %*% C
     [,1] [,2]
[1,]   28  140
[2,]   68  340
```

which are the products of the row totals and column totals. Dividing by `sum(Obs)`, the total sum of all the counts, yields the matrix of a expected values

```
Exp <- (R %*% C)/sum(Obs)
```

that is,

```
> Exp
         [,1]      [,2]
[1,] 1.166667  5.833333
[2,] 2.833333 14.166667
```

These calculations can be automated into a function that we may call `expected` which takes on a matrix of counts and yields the matrix of expected values based on the null hypothesis that the variables displayed in the two–way table are independent. Here is the code for the `expected()` function:

```
# Computing expected values for 2-way tables
# Enter M: the matrix of counts in 2-way table
expected <- function(M)
{
    R <- matrix(rowSums(M))
    C <- matrix(colSums(M),1,dim(M)[2])
    E <-  (R %*% C)/sum(M)
    return(E)
}
```

Typing `Exp <- expected(Obs)` then yields the matrix, `Exp`, of expected values. Once we have the expected values, we can compute the Chi–Squared distance as usual

```
Xsqr <- sum((Obs - Exp)^2/Exp)
```

or about 11.65714.

Next, we estimate the probability that the Chi–Squared distance is `Xsqr` or higher under the assumption that the null hypothesis is true; in other words,

the p–value for the test of independence. To do so, we may proceed as we did in Example 5.5.2 by performing many simulations of 4 launches at low temperature out of the 24, and computing the proportion of times that the Chi–Squared distance is as high or higher than the observed value. However, there is a better way to do this that makes use of the matrix operations and the `table()` function in R. The code has the advantage that it can be applied to two–way tables of any dimension. This will be done in the next section.

6.4 Permutation Test of Independence for Two-Way Tables

We present here a variation of the randomization procedure used in the analysis of the Westvaco data which can be used in the analysis of two-way tables. We will first apply it to the specific case of the Challenger disaster data and then write a code that can be applied to any two–way table. The idea is to look at all possible permutations of the vector `RowGroups` defined by

```
RowGroups <- rep(1:2,c(7,17))
```

`RowGroups` consists of 7 ones, corresponding to the group of launches with one or more failure, and 17 twos, corresponding to the launches with no failures. Typing

```
perms <- sample(RowGroups, length(RowGroups ))
```

yields a sample without replacement of length the size of `RowGroups` from the vector `RowGroups`; for example,

```
> perms
 [1] 1 1 2 2 2 1 2 1 2 2 2 1 1 2 2 2 2 2 2 1 2 2 2 2
```

This is an example of a **permutation** of the vector `RowGroups`. If we now group this permutation according to 4 days of low temperature and 20 days of high temperature, then we simulate selecting 4 days at random from the `RowGroups` vector to be four launches at low temperature. The random selection is in accord with the null hypothesis of no association between temperature and O–ring failure. The simulations can be accomplished in R by means of the `table()` function as follows:

First, type

```
ColGroups <- rep(1:2,c(4,20))
```

corresponding to a grouping according to low temperature (category "1"), or high temperature (category "2"). Typing

```
Sim <- table(perms,ColGroups)
```

yields

```
> Sim
     ColGroups
perms  1  2
    1  2  5
    2  2 15
```

This is a tabulation of how many 1s in the permutation sample fall in column–group 1 or column–group 2 in the first row of counts, and how many 2s fall in the column–group 1 or column–group 2 in the second row. To see more closely how this works, write the permutation sample above the `ColGroups` vector

```
 [1] 1 1 2 2 2 1 2 1 2 2 2 1 1 2 2 2 2 2 2 1 2 2 2 2
 [1] 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
```

and note that two 1s in the `perms` vector match with two 1s in the `ColGroups` vector while five 1s in the `perms` vector match with five 2s in the `ColGroups` vector.

The table in `Sim` corresponds to a simulated table of counts under the assumption that the hull hypothesis of independence is true. This is shown in Table 6.6.

$Y \setminus X$	Low	High	Row Totals
Cat 1	2	5	7
Cat 2	2	15	17
Column Totals	4	20	24

Table 6.6: Simulated Values

Typing

```
sum((Sim - Exp)^2/Exp)
```

computes the Chi–Squared statistic associated with this simulated sample. In this particular example we get 1.008403. This process can be repeated many times storing the Chi–Squared values in a vector called `ChiSqr` whose distribution approximates the sampling distribution of the Chi–Squared statistic. The code given below automates this process by defining a function `IndepTest` which takes on a matrix of observed values and the number of repetitions. The code assumes that the `expected()` function has already been defined.

```
# Permutation Test of Independence for 2-way tables
#
# Enter a matrix, M, of observed counts in 2-way table
# and the number of repetitions, N
# This code assumes that the expected() function has
# been defined.
#
IndepTest <- function(M,N)
    {
    Exp <- expected(M) # computes expected values
    CS <- array(dim=N) # sets up array dimension N
    RowGroups <- rep(1:dim(M)[1],rowSums(M))
        # sets up row groups
    L <- length(RowGroups)
    ColGroups <- rep(1:dim(M)[2],colSums(M))
        # sets up column groups
    for (k in 1:N)
    {
        perms <- sample(RowGroups,L)
        Sim <- table(perms,ColGroups)
        CS[k] <- sum((Sim-Exp)^2/Exp)
}
return(CS)
}
```

Typing `ChiSqr <- IndepTest(Obs,10000)` produces an array of $10,000$ values of the Chi–Squared statistic resulting from a permutation test of independence for the two–way table counts contained in

```
Obs <- matrix(c(4,3,0,17), 2,2,byrow=T)
```

The histogram of `ChiSqr` is shown in Figure 6.4.1.

We can now use the `pHat()` function to estimate the p–value for the test of independence:

$$p\text{–value} \approx \texttt{pHat(ChiSqr,Xsqr)} \approx 0.0029$$

or about 0.29%. We can therefore reject the null hypothesis of no association and conclude that the evidence is statistically significance in support of the assertion that the incidence of O–ring failures is associated with low temperatures.

6.5 More Examples of Permutation Tests for Independence

In this section we present more examples in which we use the `IndepTest()` function defined in the previous section to test the independence of categorical

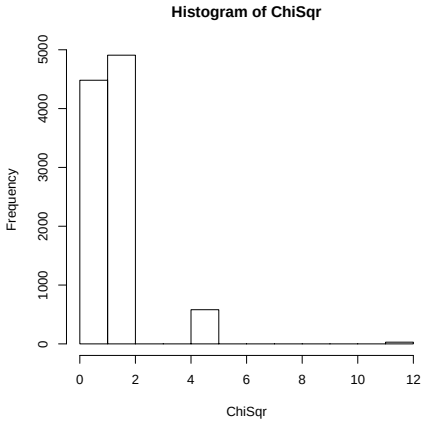

Figure 6.4.1: Histogram of `ChiSqr`

variables whose counts are displayed in two–way tables.

Example 6.5.1 A review of voter registration records in a small town yielded the following table of the number of males and females registered as Democrat, Republican, or some other affiliation:

Affiliation \ Gender	Male	Female
Democrat	300	600
Republican	500	300
Other	200	100

Table 6.7: Data for Example 6.5.1

Perform a permutation test to determine whether or not there is association between gender and political party affiliation for the voters in the town.

We first enter the counts in Table 6.7 as a 3×2 matrix as follows

```
Obs <- matrix(c(300,600,500,300,200,100), 3,2,byrow=T)
```

This yields

```
> Obs
     [,1] [,2]
[1,]  300  600
[2,]  500  300
[3,]  200  100
```

We can then type `ChiSqr <- IndepTest(Obs,10000)` to obtain `ChiSqr` as before. The histogram of `ChiSqr` is shown in Figure 6.5.2.

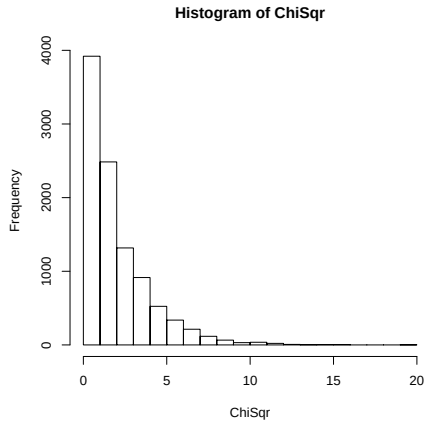

Figure 6.5.2: Histogram of `ChiSqr`

To estimate the p–value based on the distribution of `ChiSqr`, we first compute the expected values under the assumption of no association between gender and political affiliation. Using the the `expected()` function defined previously we obtain

```
> Exp <- expected(Obs)
      [,1] [,2]
[1,]  450  450
[2,]  400  400
[3,]  150  150
```

Knowing the expected values, we compute the Chi–Squared value, `Xsqr`, for the data. In this case we obtain that `Xsqr` is about 183.3.

Next, use the `pHat()` function to estimate the p–value for the test of independence:

$$p\text{–value} \approx \texttt{pHat(ChiSqr,Xsqr)} \approx 0.$$

Thus, the data in Table 6.7 are statistically significant and therefore we have strong evidence to reject the null hypothesis of independence and conclude that there is a strong association between gender and party affiliation.

Remark 6.5.2 Figure shows a density histogram of `ChiSqr` for the previous example as well as the graph of the density function for a χ^2_2 distribution. Figure 6.5.3.

The graph in Figure 6.5.3 shows that the distribution for `ChiSqr` in Example 6.5.1 can be approximated by that of a χ^2 random variable with 2 degrees of

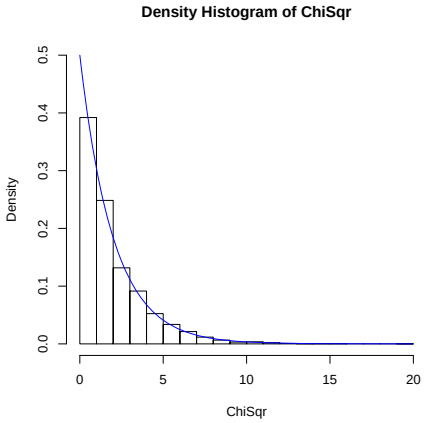

Figure 6.5.3: Density Histogram of `ChiSqr` and χ^2_2 Density

freedom. In general, for large total counts in an $r \times c$ two–way table, and expected values 5 or bigger, the distribution for the Chi–Squared distance can be approximated by a χ^2 distribution with

$$\text{df} = (r-1)(c-1)$$

degrees of freedom, where r denotes the number of rows and c the number of columns in the table. The p–value for the test of independence in the previous example can then be approximated by

$$p\text{–value} \approx \texttt{1 - pchisq(Xsqr,2)} \approx 0.$$

Example 6.5.3 A specific type of electronic sensor has been experiencing failures. The manufacturer has studied 200 of these devices and has determined the type of failure (A, B, C) that occurred and the location on the sensor where the failure happened (External, Internal). The following table summarizes the findings:

Location \ Type	A	B	C	Total
Internal	50	16	31	97
External	61	26	16	103
Total	111	42	47	200

Table 6.8: Data for Problem 6.5.3

Test the null hypothesis is that there is no association between Type of Failure and Location of Failure.

We proceed as in the previous example. This time the observed values are

```
Obs <- matrix(c(50,16,31,61,26,16), 2,3,byrow=T)
```

The outcome of this is the 2×3 matrix

```
> Obs
      [,1] [,2] [,3]
[1,]   50   16   31
[2,]   61   26   16
```

The expected values, assuming that there is no association between the type of failure and location of failure, are obtained by typing `Exp <- expected(Obs)`, where the `expected()` function has been defined previously. We obtain

```
> Exp
       [,1]  [,2]   [,3]
[1,] 53.835 20.37 22.795
[2,] 57.165 21.63 24.205
```

The Chi–Squared distance is then obtained from

```
Xsqr <- sum((Obs-Exp)^2/Exp)}.
```

This yields a value of about 8.085554.

To estimate the p–value for the test, we run the `IndepTest()` function to obtain an array `ChiSqr` by typing

```
ChiSqr <- IndepTest(Obs,10000)
```

The p–value can then be estimated by

$$p\text{–value} \approx \texttt{pHat(ChiSqr,Xsqr)} \approx 0.0185.$$

Hence, the p–value is about 1.85%. Thus, we can reject the null hypothesis at the 5% significance level. We can then say that there is strong evidence to suggest that there is an association between the type of failure and the location.

Appendix A

Calculations and Analysis Using R

A.1 Introduction to R

In this section we present an introduction to R, an integrated suite of programs that are useful in analyzing data and producing graphical displays.

R is free and may be downloaded from

`http://cran.stat.ucla.edu/`

Follow the instructions to download according to the computer platform that you have. It is available for Windows, MacOS X and Linux.

R is being developed by the *R Project for Statistical Computing* whose website is at `http://www.r-project.org/`. In that website you will find links to resources to learn about R; in particular, there is an introduction to R in the `Manuals` link.

In this section, we introduce R in the context of analyzing the data in Activity #1: *An Age Discrimination Case?* The data consist of information on 50 employees at the envelope division of the Westvaco corporation in 1991 (see Display 1.1 on page 9). The data are also provided in an MS Excel file Westvaco.xls which may be downloaded from

`http://pages.pomona.edu/~ajr04747/`

We first see how to enter Westvaco data in the MS Excel file into R. When dealing with small data sets like this one, the simplest thing to do is to read the data from the "clipboard." This is done by first selecting the data in the MS Excel file and copying into the clipboard. It is convenient to have the data in columns where the cell at the top in each column contains the name of the variable that the column's values describe (in other words, a column heading).

From the command line console window in R type

```
WestvacoD <- read.delim("clipboard")
```

This will read the data in the clipboard in a `dataframe` called `WestvacoD`. Typing `WestvacoD` in the command line at the console will display the table of values with variable names at the top of each column.

Alternatively, we can place the contents of the clipboard into a text file and then read the data in the text file from the R console. We do this as follows: After you copy the data to the clipboard in Excel, open a text editor (not a word processor) and paste the data into a text file that you may name `myTable.txt`. Then, in the console window of R, type:

```
myDataFrame <- read.table("myTable.txt")
```

(make sure that the data file is in the directory in which you are working when you use this command). This should load the into the dataframe `myDataFrame`.

A.2 Exploratory Data Analysis of Westvaco Data

We can extract that data contained in the RIF and Age columns, respectively, by using the `$` operator as follows

```
RIF <- WestvacoD$RIF
```

and

```
Age <- WestvacoD$Age
```

We now have the ages of all the workers in the Envelope Division at Westvaco at the beginning of 1991 and their corresponding job status in the data vectors `RIF` and `Age`, respectively. Both variables contain information about each individual in the population of workers in the envelope division at Westvaco. `RIF` puts the workers into categories named 0, 1, 2, etc. depending on whether the workers were not laid off in the first, second, etc. rounds of layoffs. Thus, `RIF` is an example of a **categorical** variable. `Age` is an example of a **quantitative** variable.

Quantitative variables are usually the results of counts or measurements. We can get a picture of the values of `Age` by using a **histogram**, or a display of all values that a quantitative variable can take and their frequencies. In R, a histogram of `Age` is obtained by typing:

```
hist(Age)
```

R then displays the histogram of the variable `Age` in a graphics window, which can be printed or saved as a postcript file by right-clicking on the display and making a selection. The histogram is shown in Figure A.2.1.

The histogram in Figure A.2.1 displays the **distribution** of the variable `Age`; that is, its values and how often they occur. The distribution of of a quantitative variable can also be described numerically. Typing

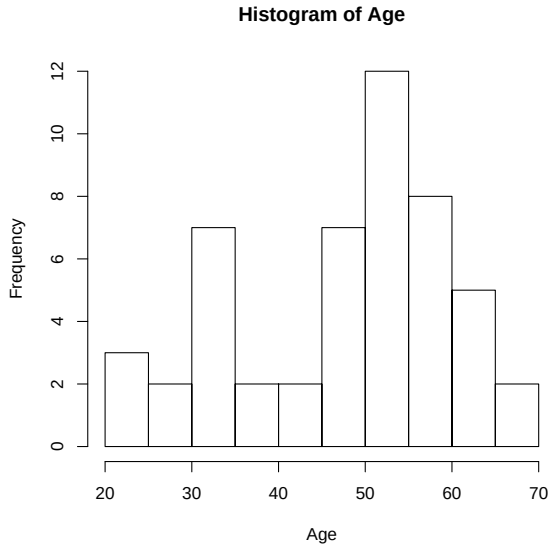

Figure A.2.1: Age Distribution at Westvaco Envelope Division in 1991

```
summary(Age)
```

in R yields

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  22.00   37.25   53.00   48.24   56.00   69.00
```

displaying the minimum, first quartile, median, mean, third quartile and maximum values of the variable `Age`. The mean is the average age of all the workers in the division and can also be obtained in R by typing

```
mean(Age)
```

The median is an age value for which half of the ages are above it and the other half below. Both mean and median are referred to as **measures of centrality**. They provide an idea of where the middle or center of the distribution is located.

The first quartile is a value for which one quarter of the values, or 25%, are below it and 75% ar above it. Similarly, 75% of the values are below the third quartile and 25% are above it.

The values minimum, maximum, mean, median and quartiles are examples of **parameters**, when referring to a **population**, or **statistics**, when derived from **sample** data. In this particular case, the fifty workers in the envelope division of Westvaco can be thought of as the population of workers in the Envelops Division of Westvaco at the beginning of 1991, or a sample (although not a random or representative one) of the entire population of workers in the company.

The five statistics (or parameters) minimum, first quartile, median, third quartile and maximum are usually referred to as the **five-number summary** for the given variable. These can be used to provide another picture of the distribution known as a **box plot**. We can obtain a box plot of `Age` in R by typing

```
boxplot(Age, ylab='Age')
```

This will yield the graphics shown in Figure A.2.2. The box in Figure A.2.2

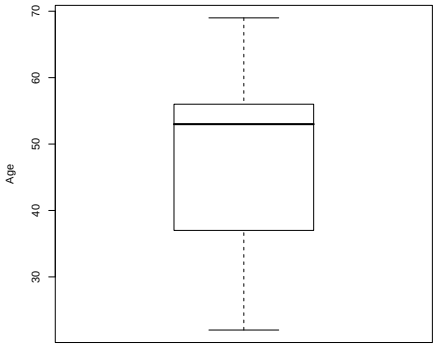

Figure A.2.2: Age Distribution at Westvaco Envelope Division in 1991

gives the boundaries of where 50% of the ages lie. The horizontal line in the box gives the median of the distribution. The two smaller horizontal lies joined to the box by the dashed vertical lines give the extreme values of the distribution.

A.3 Operations on vectors and programming

In this section we show how to use R to pick out particular entries in the `Age` vector based on the status given in the `RIF` variable. In particular, we would like to define a new variable which contains the ages of those workers that were fired by Westvaco in any of the 5 rounds of layoffs; that is, those workers for which the `RIF` column as an entry bigger than or equal to 1. In order to do this, we use the `if ... then ... else` logical structure in R as well as the looping capability of the `for` function. Here is the code we will use

```
laidoff <- array(dim = 0)        # define data structure
L <- length(Age)                 # length of Age vector, or 50
for (i in 1:L)                   # sets up for loop to length of Age
{                                # begins loop
if (RIF[i]>=1)                   # Checks if worker has been laid off
```

```
laidoff <- c(laidoff,Age[i])    # append age if laid off
}                               # end for loop
```

This code may be put into a script in a .R file and sourced after the vectors `RIF` and `Age` have been detached from the dataframe containing the Westvaco data.

We start out by defining the name of the variable that we would like to output: `laidoff`. At the end of the script `laidoff` will be a vector containing the ages of all the workers in the Envelope Division of Westvaco that were laid off in all five rounds of layoffs. The command `array(dim =0)` essentially gives an empty array. The function `length(Age)` gives the number of entries in the Age vector (which we happen to know is 50). The command `for (i in 1:L)` sets up a loop indexed by `i` running from 1 to 50. Typing `Age[i]` in R will yield the ith entry in the vector `Age`. For instance, `Age[44]` will give the age of Robert Martin, who was 54 years old at the beginning of 1991. Typing `RIF[44]` yields 2, which says that Martin was laid off in the second round.

The expression `RIF[i]>=1` is either TRUE or FALSE depending on the value of `i`. For instance, typing `RIF[44]>=1` yields a TRUE value; while typing `RIF[1]>=1` yields a FALSE value. On the other hand, `RIF[44]==0` will be FALSE, while `RIF[1]==0` is TRUE.

As an aside, type `0 + RIF[44]>=1` to get 1 and `0 + RIF[1]>=1` to get 0. In general, `n+(RIF[i]>=1)`, will add 1 to `n` if `RIF[i]>=1` is TRUE, or 0 if it is FALSE. This structure allows one to count the number of times that a certain condition holds true.

The structure

`if` (*condition*) ⟨*action*⟩

will perform the specified action if the condition holds true; and it will not do anything if the condition if false. On the other hand,

`if` (*condition*) ⟨*action 1*⟩ `else` ⟨*action 2*⟩

will perform action 1 if the condition is TRUE, or action 2 if the condition is FALSE.

Thus, the R command

```
if (RIF[i]>=1) laidoff <- c(laidoff,Age[i])
```

will append the age of the ith worker only if the worker has been laid off.

In a similar way, we can pick out the ages of the workers that kept their jobs in the Envelope Division at Westvaco in 1991 by running the R script

```
kept <- array(dim = 0)
L <- length(Age)
for (i in 1:L)
{
if (RIF[i]==0) kept <- c(kept,Age[i])
}
```

This script will yield the vector `kept` which contains the ages of those workers who kept their jobs.

Once the vectors `laidoff` and `kept` we can compute the average age of those workers that were laid off with that of the workers that remained. We can also plot the distribution of the two variables using the `boxplot` function as follows

```
boxplot(laidoff,kept, names=c("Laid off", "Kept"), ylab="Age")
```

The output of this function is shown in Figure A.3.3. The figure gives a strong

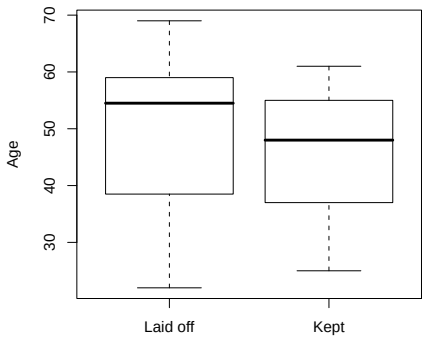

Figure A.3.3: Comparing distribution of ages of laid off versus kept workers

graphical indication of an age bias against older workers by Westvaco.

A.4 Automating simulations with R

In this section we present some codes in R that can be used to automate simulations. We begin with an example which performs permutations on a subset of the Westvaco data, computes means of random samples and stores them in a vector called `Xbar`.

```
# Author: Adolfo J. Rumbos
# Date: 9/3/2008
#
# This R code performs a permutation test on a subset of
# the Westvaco data found on page 5 of Statistics in Action
# by Cobb, Sheaffer and Watkins.
#
#
# The subset consists of the 10 hourly workers involved
# the second of the five rounds of layoffs.  Three of the
```

```
# ten workers were laid off in that round.
#
# The 10 hourly workers involved in the 2nd round of layoffs
# are entered into a vector called hourly2
#

hourly2 <- c(25,38,56,48,55,64,55,55,33,35)

Nsamples <- 1000    # Nsamples is the number of samples drawn

#
# Nsamples random samples of size 3, without replacement,
# in a vector from hourly2 and their averages are computed
# and stored denoted Xbar
#
Xbar <-  mean(sample(hourly2,3,replace=F))
            # sets initial value of x_bar

l <- Nsamples -1  # we need Nsamples - 1 more samples

for (i in 1:l) # Sets up a loop from 1 to Nsamples

{           # Begins body of loop

Xbar <- c(Xbar, mean(sample(hourly2,3,replace=F)))
            # creates vector of average ages of
            # laid off workers.
```

The following code performs random sampling and computes Chi–Squared statistics for a goodness of fit test. The sampled values are taken from a uniform distribution over the interval $(0, 1)$ by means of the `runif()` function in R. The code computes Chi–Squared values for each run of the simulation and stores them in a vector named `ShiSqr`.

```
# Author: Adolfo J. Rumbos
# Date: November 12, 2008

# Simulating sampling from a population with age distribution
#
# Category 1 (Ages 0-18)  : 0.26
# Category 2 (Ages 18-64) : 0.62
# Category 3 (Ages 64+  ) : 0.12
# Samples of size n are collected from a uniform distribution
# over the interval (0,1). Values between 0 and 0.26 are placed
# into category 1; values between 0.26 and 0.88 into Category 2, etc.
# Counts for each category are placed into the vector SimObs.
# These are then used to compute Chi-Squared statistics and placed into
```

```
    # a vector ChiSqr.  The process is repeated Nrep times.

    n <- 140                  # sample size
    Prop <- c(0.26,0.62,0.12) # Proportions postulated by null model
    Exp <- round(n*Prop)      # Computes expected values

    Nrep <- 10000             # Number of repetitions
    ChiSqr <- array(dim=Nrep)
    for (k in 1:Nrep)
    {
    SimObs <- c(0,0,0)        # Initializes counts
    s <- runif(n,0,1)         # Generates n randoms numbers between 0 and 1

    for (i in 1:n) # Sets up loop for obtaining counts in first category
    {
        SimObs[1] <-SimObs[1] + (s[i]>=0 & s[i]<=Prop[1])
                    # adds 1 to first counter
    }
    for (i in 1:n)            # Sets up loop for counts in second category
    {
        SimObs[2] <-SimObs[2] + (s[i]>Prop[1] & s[i]<=Prop[1]+Prop[2])
                              # adds 1 to second counter
    }
    SimObs[3] <- n-SimObs[1]-SimObs[2]
    ChiSqr[k] <- sum((SimObs-Exp)^2/Exp) # Computes X^2
}
```

A.5 Defining functions in R

The goal of this section is to show how the `function()` function in R can be used to define a function, `pHat`, which takes in `Xbar`, the vector of mean ages of the samples resulting from simulations obtained in the previous section, and the observed value of 58, and yields the proportion of values in the `Xbar` vector which are 58 or higher. This yields an estimate for the p–value in the randomization test that we performed on the 10 hourly workers involved in the second rounds of layoffs at Westvaco in Activity #1.

Here is the code:

```
# Author: Adolfo J. Rumbos
# Date: September 5, 2008

# Randomization Test for Westvato data (continued)
#
# Defining pHat()
#
# This script is to be run after the Simulation2roundHourlyWorkers.R has been
```

```
# run.  It assumes that the Xbar vector has been obtained.
# This program defines a function, pHat(), which
# computes the proportion, p_Hat, of samples in the simulation which yield a
# mean greater than or equal to an observed value, obs. This yields
# an estimate of the probability that, under the assumption that Westvaco's
# choice of the three hourly workers to be laid off was entirely random
# (i.e., each worker, regardless of her or his age, had an equal likelihood
# of being chosen).
#
# The function takes on the two arguments: X and obs, where
# X is a vector that contains values for a test statistic obtined from
# many repetions of a sampling, and obs is the  value of the statistic
# observed in the data.
# The result is the proportion of samples that yield a test statistic
# equal to, or larger than, the observed value
#
#
pHat <- function(X,obs)  # Sets up to define the function Phat
{
L<-length(X)             # Computes length of X

NYes <- 0                # Sets up counter of "yeses"

for (i in 1:L)           # Sets up loop through vector X

{ NYes <-NYes +(X[i]>=obs) }
                         # adds 1 to counter if X[i]>= obs

result <- NYes/L         # Computes proportion of NYes in L

return(result)           # Returns estimate of p_hat
}
```

The structure `function()` takes any number of objects (arrays, vectors, matrices, constants, etc.) and returns another another object, in this case it returns a value. Typing

```
p_hat <- pHat(Xbar,58)
```

stores in `p_hat` the proportion of times that the values in the vector `Xbar` are at least 58. Any vector and any observed value can be input into the `pHat()` function in place of `Xbar` and 58, respectively.

The definition of `pHat()` given above has three parts:

- The line

```
pHat <- function(X,obs)
```

established the name of the function and the names of the objects the function will be taking in.

- A code for calculations involving the objects input to functions is placed in braces. The result of the calculations is placed in an object called `result`
- The result of the calculations is then returned by the statement

```
return(result)
```

A modification of the code for `pHat()` yields a code for function `CumDist()` which computes the cumulative distribution function for a random variable based on its set of values on a sample space given in a vector `X`. Here is the modification

```
# Author: Adolfo J. Rumbos
# Date: September 23, 2008

# Defining a cumulative distribution function CumDist()
#
# The function takes on the two arguments: X and v, where
# X is a vector that contains values for a test statistic obtained from
# the sample space of an experiment, and v is a value.
# The result is the proportion of samples that yield a value less than
# or equal to v
#
CumDist <- function(X,v) # Sets up to define the function Phat
{
L<-length(X) # Computes length of X
NYes <- 0           # Sets up counter of "yeses"
for (i in 1:L)      # Sets up loop through vector X
{ NYes <- NYes +(X[i]<=v)   # adds 1 to counter
if Xbar[i]<= v } result <- NYes/L # Computes proportion of NYes in L
return(result)      # Returns cumulative distribution value
}
```

Bibliography

[GWCS04] A. E. Watkins, G. W Cobb and R. L. Scheaffer. *Statistics in Action*. Key College Publishing, 2004.