Laboratory 1b: Goodness of Fit - The χ^2 test

Introduction

Imagine that you have a quantity, x, that you wish to measure. Let us assume that you have taken Lab. 1a to heart and have made N measurements of x and calculated the mean, standard deviation, and error in the mean for your data. As you have decided to measure x because you wish to test some hypothesis. Perhaps you expect x to have a particular value or perhaps you expect x to have a particular standard deviation. The question at hand is this: Does your data support your hypothesis.

To answer this question you will need to assume a theoretical distribution which you believe applies (Gaussian, Poisson, or Binomial). Make this assumption based on your knowledge of the experiment and how that will affect the dispersion of values about the mean. You can then use this theoretical probability distribution, $P(x_j)$, to predict the theoretical frequency,

$$F(x_j) = NP(x_j) \tag{1}$$

for any value x_j. Note that $F(x_j)$ will have the same spread as our data if our assumed distribution was correct.

It is important to recognize that with our finite number of measurements we cannot expect the theoretical frequency, $F(x_j)$, to be *exactly* equal the measured frequency, $f(x_j)$, in any given interval. Rather we would expect $F(x_j) \cong f(x_j)$ and that the deviation between the experimental frequency and assumed theoretical frequency to be statistical in nature. That is to say, we expect $f(x_j) - F(x_j)$ to be given by some statistical error $\sigma(f(x_j))$ associated with measuring $f(x_j)$; i.e.

$$f(x_j) - F(x_j) \cong \sigma\left(f(x_j)\right) \tag{2}$$

The key question is: what are the expected values for $\sigma[f(x_j)]$? To understand this, consider what happens if we were to replicate the experiment. Next time arround the N measurements will be distributed in a slightly different way because of the random nature of errors. A large number, m, of

replicates of the experiment would yield a discrete distribution of possible frequencies for each value x_j.

Example 1—a follow along example:

Consider again the large data set from Lab. 1a involving a block (see http://facultyweb.cortland.edu/douglas.armstead/S15/Intermediate/block.csv), a histogram of the width data appears in Fig. 1. The histogram was produced using hist() [a similar graph can be made with histogram() using type='count' but it doesn't give such a nice bin width in this example].

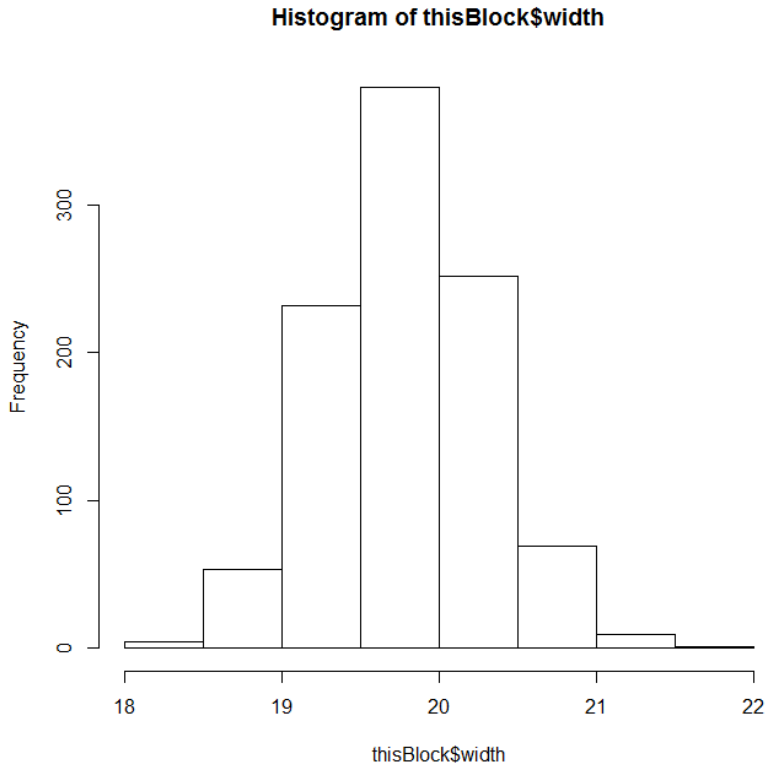

Figure 1: Histogram of block widths from the data_frame thisBlock.

Table 1 is a frequency table, f, for the data and requires more effort to construct[1]. We must create a set of bins, x, and chose their edges. We use the sequence command, seq(), to create the bins and then cut up our data among the bins:

- x<-seq(18, 22, by=0.5)
- f<-table(cut(blockB$width, breaks=x))

[1] A fully functioning script appears at the end of these lab instructions.

Note that f *is* $f(x_j)$.

Interval	Frequency
(18, 18.5]	4
(18.5, 19]	53
(19, 19.5]	232
(19.5, 20]	380
(20, 20.5]	252
(20.5, 21]	69
(21, 21.5]	9
(21.5, 22]	1

Table 1, frequencies, f=f(x_j) for the data set blockB$width.

Because of the measurement process, we expected a Gaussian distribution for the width of the block. Calculating the mean and standard deviation of the data and using it in a Gaussian/normal distribution determines the theoretical observation rate, $F(x_j)$.

- mw=mean(blockB$width)
- sw=sd(blockB$width)
- n<-length(blockB$width)
- px<-pnorm(x, mw, sw)

Note that $px \neq F(x_j)$, rather it is the probability that a measured value is less than x_j. To find F(x_j) we need the new function diff(v,lag) which subtracts one element in the vector v from its neighbor the lag steps away, i.e., x[n+lag]-x[n].

- F<- n*diff(px,1)

Note that F *is* F(x_j).

Motivating χ^2

Recall our plan to replicate the experiment. Consider a specific bin, the one labeled x_j = 19 cm (and centered on 19.25cm). The frequency from table 1 measured for the interval x_j = 19 cm was f(19cm) = 232. If you were to make 12 more replicates of the experiment (each with 100 measurements) there would be a total of m = 13 replicates. Figure 2 is a histogram showing the results of the replicates experiment with the different values that f(19cm) took in the m sets of replicates. Note that the total

number of outcomes f_k(19cm) must equal the total number of replicates, i.e. Σf_k(19cm) = m. The mean frequency is <k>=233.6. It is reasonable to expect that the mean over all the replicates, $F(19cm) \cong 233.6$, is closer to the parent distribution's mean value than any one replicate, including the first with f(19cm)=232.

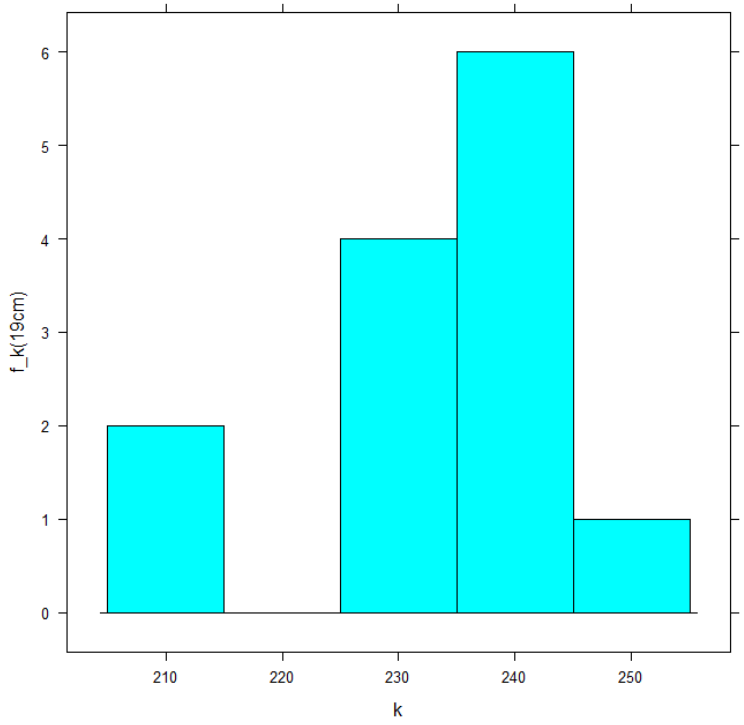

Figure 2: Frequency distribution for 13 replicates for bin 19cm. The mean, <k>=233.6.

An alternative to making replicates of our experiment to gain more information is to use our observations in each of the histogram bins for the data we have, (x_j), and anticipating their distribution. Since $f(x_j)$ is discrete (i.e., each value is an integer) the distribution in Fig. 2 should be a Poisson. The theoretical prediction we made for $f(x_j)$ is $F(x_j)$ and should be a good estimate of the mean of the Poisson distribution. It immediately follows that the standard deviation is $\sqrt{F(x_j)}$. Thus we expect, on average,

$$\left[f(x_j) - F(x_j)\right]^2 \cong \sigma^2\left(f(x_j)\right) = F(x_j) \tag{3}$$

or

$$\frac{\left[f(x_j) - F(x_j)\right]^2}{F(x_j)} \cong 1 \tag{4}$$

We should repeat the process for each of the n intervals x_j. Summing all of these gives the χ^2 value for our single replicate

$$\chi^2 = \sum_{j=1}^{n} \frac{\left[f(x_j) - F(x_j)\right]^2}{F(x_j)} \tag{5}$$

If our assumption about the theoretical distribution is correct we would expect $\chi^2 \cong n$ for our dataset, (recall that n is the number of intervals and N is the number of measurements in a replicate). More precisely we expect $\chi^2 \cong \nu$, where ν is the number of degrees of freedom for the histogram.

- For a Gaussian distribution $\nu = n - 2$. The only way we can estimate the mean and standard deviation of the parent distribution is to assume they have the same value as the sample mean and the sample standard deviation. Both of these assumptions consume degrees of freedom.
- For a Poisson distribution $\nu = n - 1$ since the data is used to determine the sample mean and the standard deviation immediately follows from that.
- For a binomial distribution $\nu = n$ as the data is not needed to determine the distribution.

We can calculate χ^2in R as well as ν, but it is useful to take it one step further and calculate the probability that random fluctuations are responsible for the differences we found between the observed $f(x_j)$ and and theoretical $F(x_j)$ values.

- data.chisq<-sum(((f-F)^2/F))
- df=length(f)-2
- 1-pchisq(data.chisq,df)

The probability that random fluctuations are responsible for the difference between the observed frequencies and Gaussian frequencies is 0.863 or 86.3%.

The reduced χ^2_ν is defined as

$$\chi^2_\nu = \frac{\chi^2}{\nu} \tag{6}$$

We expect $\chi^2 \cong \nu$ for a good fit or $\chi^2_\nu \cong 1$. If $\chi^2_\nu \gg 1$ then the deviations are larger than statistically predicted and we might question the validity of *the choice of distribution*. This corresponds to 1-pchisq()<<1. If $\chi^2_\nu \ll 1$ then the deviations are smaller than statistically predicted and we might wonder if the deviations are indeed statistical, this is something that 1-pchisq() doesn't do a good job of conveying. For the data set blockB $width $\chi^2_\nu = 0.42$ which falls in the category $\chi^2_\nu \cong 1$.

Example 2: A Fair Coin?

Imagine trying to determine if a coin is fair or not. If the coin is fair, then the probability of getting heads

is p=0.5 and the probability of getting tails is q=0.5, if it is not fair p≠0.5 and q≠0.5. Since a coin has only two sides, q=1-p, this will have important consequences for testing our coin. If we toss the coin 100 times, we will expect to get heads 100 × 0.5=50 times. We know, however, that while the probability of getting heads is 0.5, there is a significant chance that we would actually get a few more or a few less than 50 heads in 100 tosses. The question is, how much variation in the number of heads will we need to see before we are confident that someone is trying to cheat us? How much variation from the average do we need to reject the hypothesis that p=0.5? This is where the χ^2 Goodness of Fit test comes in handy. Imagine you now perform the experiment to test the hypothesis that the coin is fair, you toss the coin 100 times and observe that it landed on heads 38 times. From this data, the design of our experiment, and the nature of coins, we are able to determine that the coin must have landed on tails 62 times and we note this in the Table 2.

	Observed	Expected
Heads	38	50
Tails	62	50

Table 2: Observed and expected results of 100 coin tosses.

With this data in our hands, we can compute a χ^2 test statistic and use it to determine the fairness of the coin. That is:

$$\chi^2 = \frac{(38-50)^2}{50} + \frac{(62-50)^2}{50} = 5.76 \tag{7}$$

In order to examine our value in the context of a χ^2 distribution we calculate the total degrees of freedom, ν, by looking at the total number of parameters in our model, 2 (p and q), and subtracting 1 because q is not independent from p since q=1- p:

$$\nu = \text{total number of parameters} - 1 \tag{8}$$

$$\chi^2_\upsilon = \frac{\chi^2}{\upsilon} \tag{9}$$

In our case, $\nu = 2-1 = 1$, and as a result $\chi^2_1 = \chi^2 = 5.76$. We found $\chi^2_\upsilon \gg 1$ and so the theoretical distribution cannot apply to our data. We must reject the hypothesis that the coin is fair.

Third example: A Random Number?

Imagine trying to determine if a random number generator is truly random or whether it has a bias that

makes some numbers come up more often than others. If the random number generator is unbiased, then the probability of getting any particular number should be 1/N where N is the total range of numbers the generator might produce. In this example we will be working with a random number generator that produces integers between 1 and 6 to simulate the dice that you used in the Lab 1. If the generator is unbiased each number should appear with a probability of 1/6 = 0.167. As with the fair coin, however, we know there is a significant chance that we could get a few more or a few less than the average for any finite number of trials. This is where the χ^2 Goodness of Fit test comes into the experiment and will again help us answer the question of whether the variations we observe are consistent with the hypothesis that the probability is 0.167 for all six possible outcomes.

Procedure

Task 1

In Lab. 1a you constructed histograms for several distributions. For each of these distributions, using the above examples as a guide

1. Use the size and range of bins in the histogram as a guide for extracting the observed frequency, $f(x_i)$, from your data. Record this.
2. Calculate the theoretical frequency, $F(x_i)$, for each bin in step 1 using an appropriate theoretical distribution. Record this.
3. From the results of step 1 and 2, calculate the reduced chi squared parameter, χ_υ^2 , and the probability that chance has caused $\chi^2$to be as large as it is. Record these.
4. Comment on the appropriateness of the assumed theoretical distributions.

Again you should think of this as a programming exercise, consolidate the code from the example in script, use the R editor (file, new script) and then run it all. How can you make most of the lines of code stay the same from one distribution to the next?

Report your results for analyzing each of the histograms from part a

- ***blockB width***
- ***table width***
- ***emission data***
- ***radiation data (50, 500, 5000)***
- ***fair and unfair coin data.***

Include a copy of the full script you wrote to do the analysis.

Task 2

Use the runif function in R to create a data set with 2,400 points randomly and uniformly distributed between 0 and 1:

- flips<-runif(2400)

Decide how to chop the interval from (0:1) into 6 equal pieces and identify one piece with each roll outcome 1, 2, 3, 4, 5, and 6. Using the same technique as before:

1. Construct a table of the observed frequencies, $f(x_i)$.
2. Calculate the theoretical frequency for each outcome, $F(x_i)$.
3. Use the results from steps 1 and 2 to calculate the reduced chi squared parameter, χ^2_υ, and the probability that chance has caused $\chi^2$to be as large as it is.
4. Comment on the fairness the random number generator in R, does it appears to have a bias towards any particular number.

Include a copy of the full script you wrote to do the analysis.

WHEN YOU FINISH: LEAVE THINGS AS YOU FOUND THEM!

Script for blockB calculation:

```
require(fastR)

blockB<-read.csv("http://facultyweb.cortland.edu/douglas.armstead/S16/Intermediate/block.csv")

# To find the full range of the data consult summary

summary(blockB$width)

x<-seq(18, 22, by=0.5)

f<-table(cut(blockB$width, breaks=x))

mw=mean(blockB$width)

sw=sd(blockB$width)

n<-length(blockB$width)

px<-pnorm(x, mw, sw)

F<- n*diff(px,1)

data.chisq<-sum(((f-F)^2/F))

df=length(f)-2

pRandomError<-1-pchisq(data.chisq,df)

#Now for the key results

mw

sw

data.chisq

df

pRandomError
```