Math 3080 § 1. **Pregnancy Example: χ^2 Goodness of Fit Test** Name: Example
Treibergs **Continuous Variable with Known Parameters** March 22, 2014

This **R**© program explores testing for goodness of fit involving a continuous distribution. The essential method is to assign bins, count the occurrences in each bin and test via chi-squared test whether the observed frequencies match the observed ones.

This data was taken from larsen and Marx, *An Introduction to mathematical Statistics and its Applications 4th ed.,* Prentice Hall, Upper Saddle River, NJ, 2006. The duration of pregnancy is thought to be a normal variable with mean $\mu = 266$ days and a standard deviation of $\sigma = 16$ days. The authors say that for the last 70 births at the Davidson County General Hospital in Nashville had durations

251	264	234	283	226	244	269	241	276	274
263	243	254	276	241	232	260	248	284	253
265	235	259	279	256	256	254	256	250	269
240	261	263	262	259	230	268	284	259	261
268	268	264	271	263	259	294	259	263	278
267	293	247	244	250	266	286	263	274	253
281	286	266	249	255	233	245	266	265	264

Accepting that μ and σ are the true parameter values, are the data plausibly from this normal distribution?

To these the hypothesis, we count the number of data points in bins $220 \le y < 230 \le y < 240$, etc., and check whether the observed frequencies are the same as the theoretical frequencies. As the data is rounded to the nearest integer, we set the break points to be $b_1 = 219.5$, $b_2 = 229.5$, and so on. Then count the number of data points in these bins, so f_1 is the number of observations less than b_2, f_i is the number between b_i and b_{i+1} for $i = 1, \ldots, 7$ and f_8 is the number greater than b_7. We compute the theoretical probabilities. If Y denotes a $N(\mu, \sigma)$ variable then $\pi_1 = P(Y < b_2)$, $\pi_2 = P(b_2 < Y < b_3$ and so on.

It turns out that the expected number in each bin $70 * \pi_i$ is less than 5 for some bins. We lump the first three bins together to make up for small expected observations. The last bin has and expected 4.967 observations which we leave unchanged. We end up with six bins. We test the hypotheses

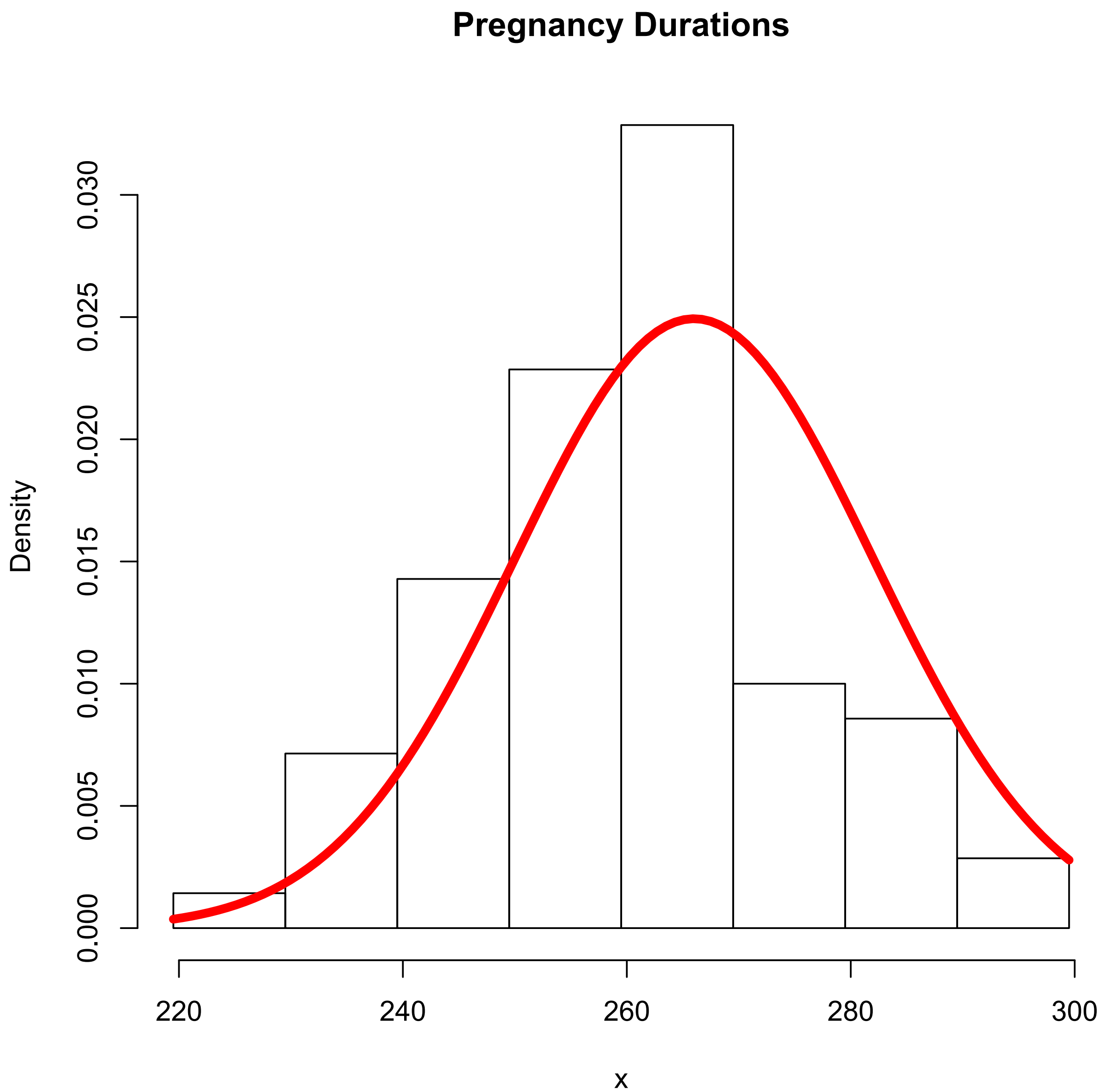
$$\mathcal{H}_0 : p_i = \pi_i \text{ for all } i = 1, \ldots, 6;$$
$$\mathcal{H}_a : p_i \neq \pi_i \text{ for some } i = 1, \ldots, 6.$$

The chi-squared test shows that there is significant evidence (p-nalue is 0.030) that this data does not come from the assumed distribution.

R Session:

```
R version 2.14.0 (2011-10-31)
Copyright (C) 2011 The R Foundation for Statistical Computing
ISBN 3-900051-07-0
Platform: i386-apple-darwin9.8.0/i386 (32-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[R.app GUI 1.42 (5933) i386-apple-darwin9.8.0]

[Workspace restored from /home/1004/ma/treibergs/.RData]
[History restored from /home/1004/ma/treibergs/.Rhistory]

> x=scan()
1: 251 264 234 283 226 244 269 241 276 274
11:  263 243 254 276 241 232 260 248 284 253
21:  265 235 259 279 256 256 254 256 250 269
31:  240 261 263 262 259 230 268 284 259 261
41:  268 268 264 271 263 259 294 259 263 278
51:  267 293 247 244 250 266 286 263 274 253
61:  281 286 266 249 255 233 245 266 265 264
71:
Read 70 items
> max(x); min(x)
[1] 294
[1] 226
> ################  SET BIN BREAKS  ################################
> brk=seq(219.5,299.5,10);brk
[1] 219.5 229.5 239.5 249.5 259.5 269.5 279.5 289.5 299.5
>
> ################ PLOT A HISTOGRAM AND DO FREQUENCY COUNTS  ######$
> h1=hist(x,breaks=brk,right=F,freq=F,main="Pregnancy Durations")
> curve(dnorm(x,266,16),add=T,col=2,lwd=5)
>
> fr=h1$counts; fr
[1]  1  5 10 16 23  7  6  2
>
```

```
> ###########  COMPUTE THE PROBABILITY FOR EACH BIN  #################

> pi=c(pnorm(brk[2],266,16),pnorm(brk[3:8],266,16)-pnorm(brk[2:7],266,16),
     pnorm(brk[8],266,16,lower.tail=F))
>
> ########  TABLE OF OBSERVED VS EXPECTED FREQUENCIES  ################

> rbind(fr,70*pi)
       [,1]     [,2]      [,3]     [,4]     [,5]     [,6]     [,7]     [,8]
fr 1.000000 5.000000 10.000000 16.00000 23.00000  7.00000 6.000000 2.000000
   0.788678 2.629814  7.166334 13.37474 17.10087 14.98125 8.991798 4.966521

> #############  LUMP FIRST THREE BINS TOGETHER  ######################
> fr2=c(fr[1]+fr[2]+fr[3],fr[4:8])
> pi2=c(pi[1]+pi[2]+pi[3],pi[4:8])
> sum(fr2);sum(pi2)
[1] 70
[1] 1
> rbind(fr2,70*pi2)
        [,1]     [,2]     [,3]     [,4]     [,5]     [,6]
fr2 16.00000 16.00000 23.00000  7.00000 6.000000 2.000000
    10.58483 13.37474 17.10087 14.98125 8.991798 4.966521

> ###### CHI-SQ TEST FOR OBSERVED VS THEORETICAL FREQUENCIES  ##########
> chisq.test(fr2,p=pi2)

Chi-squared test for given probabilities

data:  fr2
X-squared = 12.34, df = 5, p-value = 0.03041

Warning message:
In chisq.test(fr2, p = pi2) : Chi-squared approximation may be incorrect
>
```


Pregnancy Durations
Density
0.000
0.005
0.010
0.015
0.020
0.025
0.030
220
240
260
280
300
x