Math 3070 § 1. **Munitions Example: Negative Binomial** Name: Example
Treibergs **to Predict Number of Accidents** July 17, 2011

This problem comes from Bulmer, *Principles of Statistics,* Dover, 1979, which is a republication of the 1967 second edition published by Oliver & Boyd.

The Poisson Random Variable describes the number of occurances of rare events in a period of time, but it sometimes fails to account best for an observed variable. The data on point comes from a 1920 study of Greenwood & Yule. They counted the number of accidents in five weeks that occurred to a group of women working on high explosives shells in a munitions factory during the 1914–1918 War. One could regard an accident as a rare event and as such would be governed by Poisson's law. However Poisson's law does not fit the data very well; there are too many women in the extreme groups. Indeed the variance is .69 but the mean is .47, which should be the same for a Poisson variable.

Let $X_1, X_2, ..., X_n$ are observations of the number of accidents. If these were a random sample from Pois($\lambda$), then since $\mathrm{E}(X) = \mathrm{V}(X) = \lambda$, an estimator for $\lambda$ and is the sample mean.

$$\bar{X} = S^2 = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

The sample mean is the maximum likelihood estimator $\hat{\lambda}$. We use this to compute the Poisson approximation.

Greenwood & Yule explained the inaccuracy by supposing that the women were not homogeneous, but differed in their accident proneness. This would spread out the distribution of accident numbers. Let us suppose that the accident rate is itself a random variable $\Lambda$. Then for any individual, the number of accidents follows a Poisson distribution with rate constant $\Lambda$. We may say that the conditional distribution of the number of accidents $X$ for given $\Lambda = \lambda$ is

$$p(x|\lambda) = \frac{e^{-\lambda}\lambda^x}{x!}, \qquad \text{for } x = 0, 1, 2, \ldots$$

Also, suppose that there is a constant $a > 0$ such that the rate satisfies a $\chi^2$-distribution with $f > 0$ degrees of freedom, where $f$ is not necessarily an integer,

$$Y = a\Lambda \sim \chi^2(f).$$

The pdf of $Y$ is

$$f_Y(y) = \frac{1}{2^{f/2}\Gamma(f/2)} y^{f/2-1} e^{-y/2},$$

so that the pdf for $\Lambda$ is

$$f_\Lambda(\lambda) = \frac{a}{2^{f/2}\Gamma(f/2)} (a\lambda)^{f/2-1} e^{-a\lambda/2}.$$

The population distribution of $X$ is thus the marginal pmf for $x = 0, 1, 2, \ldots$,

$$\begin{aligned} p_X(x) &= \int_0^\infty f(x,\lambda)\, d\lambda \\ &= \int_0^\infty p(x|\lambda)\, f_\Lambda(\lambda)\, d\lambda \\ &= \int_0^\infty \frac{e^{-\lambda}\lambda^x}{x!} \cdot \frac{a}{2^{f/2}\Gamma(f/2)} (a\lambda)^{f/2-1} e^{-a\lambda/2}\, d\lambda \\ &= \frac{a^{f/2}}{2^{f/2}\Gamma(f/2)\, x!} \int_0^\infty \lambda^{f/2+x-1} e^{-(a/2+1)\lambda}\, d\lambda. \end{aligned}$$

Using the gamma distribution whose pdf for $y > 0$ is

$$f(y;\alpha,\beta) = \frac{1}{\beta^\alpha \Gamma(\alpha)} y^{\alpha-1} e^{-y/\beta},$$

with $\alpha = f/2 + x$ and $1/\beta = a/2 + 1$, and the fact that its total probability is one,

$$\begin{aligned} p_X(x) &= \frac{a^{f/2}\Gamma(f/2+x)}{2^{f/2}\Gamma(f/2)\,x!}(a/2+1)^{-f/2-x} \\ &= \frac{\Gamma(f/2+x)}{\Gamma(f/2)x!}\left(\frac{a}{a+2}\right)^{f/2}\left(\frac{2}{a+2}\right)^{x} \\ &= \frac{\Gamma(r+x)}{\Gamma(r)x!}\,p^r\,(1-p)^x \\ &= \mathrm{nb}(x;r,p) \end{aligned}$$

is negative binomial with parameters $r = f/2$ and $p = a/(a+2)$. In case $r$ is an integer, $\Gamma(r) = (r-1)!$ and the coefficient equals the usual $\binom{x+r-1}{r-1}$.

Let's compute the mean and variance of a negative binomial variable. Observe that since $\Gamma(z+1) = z\Gamma(z)$, we have for $x = 0, 1, 2, \ldots$,

$$\begin{aligned} &\frac{\Gamma(r+x)}{\Gamma(r)x!} = \frac{(r+x-1)}{x}\frac{(r+x-2)}{x-1}\cdots\frac{r}{1} = \binom{r+x-1}{x} \\ &= (-1)^n\binom{-r}{x} = (-1)^n\frac{(-r)}{x}\frac{(-r-1)}{x-1}\cdots\frac{-r-x+1}{1} \end{aligned}$$

Remembering the negative binomial formula from Calculus II, we see that, first of all, it is a pmf

$$\sum_{x=0}^{\infty} p_X(x) = \sum_{x=0}^{\infty}\binom{-r}{x}p^r(-q)^x = p^r(1-q)^{-r} = 1.$$

Differentiating with respect to $q$ gives the first moment

$$r(1-q)^{-r-1} = \frac{d}{d\,q}(1-q)^{-r} = -\sum_{x=0}^{\infty}\binom{-r}{x}x(-q)^{x-1} = \frac{1}{q}\sum_{x=0}^{\infty}\binom{-r}{x}x(-q)^x.$$

so that

$$\mathrm{E}(X) = \sum_{x=0}^{\infty} x\binom{-r}{x}p^r(-q)^x = qr(1-q)^{-r-1}p^r = \frac{rq}{p}.$$

Differentiating again,

$$r(1-q)^{-r-1}+r(r+1)q(1-q)^{-r-2} = \frac{d}{d\,q}rq(1-q)^{-r-1} = -\sum_{x=0}^{\infty}\binom{-r}{x}x^2(-q)^{x-1} = \frac{1}{q}\sum_{x=0}^{\infty}\binom{-r}{x}x^2(-q)^x.$$

so that

$$\mathrm{E}(X^2) = \sum_{x=0}^{\infty} x^2\binom{-r}{x}p^r(-q)^x = qr(1-q)^{-r-1}p^r + r(r+1)q^2(1-q)^{-r-2}p^r = \frac{qr}{p} + \frac{r(r+1)q^2}{p^2}$$

It follows that

$$\mathrm{V}(X) = \mathrm{E}(X^2) - \mathrm{E}(X)^2 = \frac{qr}{p} + \frac{r(r+1)q^2}{p^2} - \frac{r^2q^2}{p^2} = \frac{qr\,[p+q]}{p^2} = \frac{qr}{p^2}.$$

We have shown that if $X \sim \text{nb}(r, p)$, then

$$\text{E}(X) = \frac{r(1-p)}{p}; \qquad \text{V}(X) = \frac{r(1-p)}{p^2}.$$

We fit the best negative binomial distribution using the method of moments. In other words, we equate the theoretical first and second moments to the empirical ones, which is equivalent to equating the mean and variance of the theoretical population to the sample mean and variance, and solve for the parameters. Equating to $\bar{X}$ and $s^2$, we find

$$p = \frac{\bar{X}}{s^2}; \qquad r = \frac{\bar{X}^2}{s^2 - \bar{X}}.$$

Using these parameter values in the negative binomial seems to fit the data better.

---

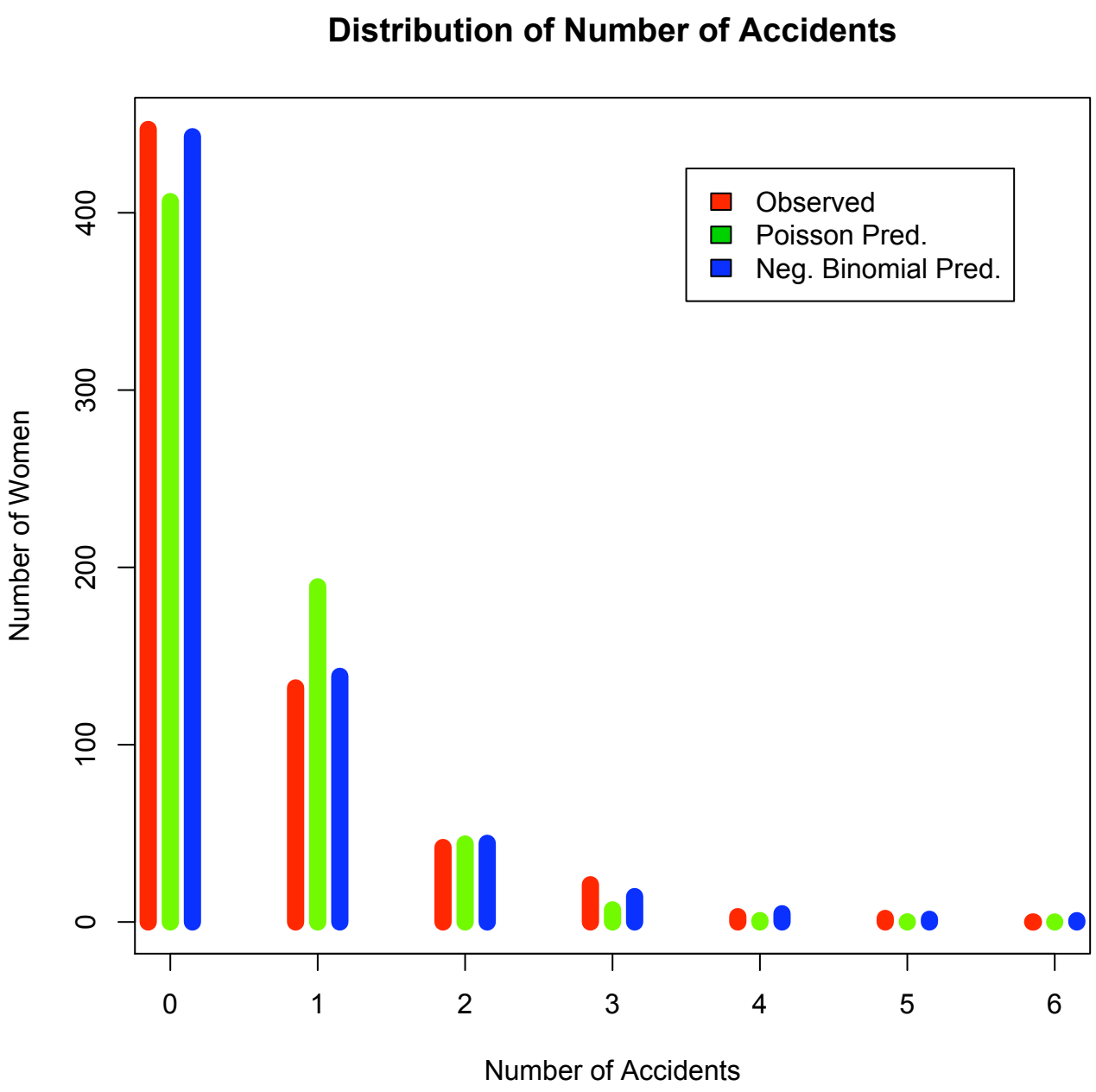
**R Session:**

---

```
R version 2.10.1 (2009-12-14)
Copyright (C) 2009 The R Foundation for Statistical Computing
ISBN 3-900051-07-0

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 1.31 (5538) powerpc-apple-darwin8.11.1]

[Workspace restored from /Users/andrejstreibergs/.RData]

> ###### ENTER ACCIDENT DATA FOR WOMEN AT THE MUNITIONS PLANT ##################
> #  acc  records the freq for each number of accidents
> acc <- scan()
1: 447 132 42 21 3 2 0
8:
Read 7 items

> # noacc records the corresponding number of accidents ("6" means 6 or more)
> noacc <- 0:6
> # The total number of women is  n
> n <- sum(acc); n
[1] 647
```

```
> # The average number of accidents per woman
> xbar <- sum(acc*noacc)/n; xbar
[1] 0.4652241

> # The variance of the number of accidents
> s2 <- (sum(noacc^2*acc)-xbar^2*n)/(n-1); s2
[1] 0.6919002
>
> ########## METHOD OF MOMENTS ESTIMATOR FOR PARAMETERS ########################
> pp <- xbar/s2
> rr <- xbar^2/(s2-xbar)
>
> # Set up vector of probabilities for neg. binomial.
> probnb <- dnbinom(0:6,rr,pp)
> replace probnb[7] by total of six or more
> # lower.tail=F means probability is given for X>5
> probnb[7] <- pnbinom(5,rr,pp,lower.tail=F)
> n
[1] 647
> Distribute the total of 647 women among their number of accidents.
> # enb   is expected no accidents using neg. binom.
> enb <- probnb*n
>
> ########## COMPARE TO BEST POISSON ESTIMATE ###################################
> # Set up vector of probabilities for Poisson.
> probpois <- dpois(0:6,xbar)
> probpois[7] <- ppois(5,xbar,lower.tail=F)
> Distribute the total of 647 women among their number of accidents.
> # epois   is expected no accidents using Poisson.
> epois <- probpois*n
>
> ################## BUILD TABLE TO COMPARE OBS. TO EXPECTED ##################
> m <- matrix( c(acc, sum(acc), round(epois, 3), sum(round(epois, 3)),
+ round(enb, 3),  sum(round(enb, 3))), ncol = 3)
> colnames(m) <- c("Observed","   Poisson","  Neg. Binom.")
> rownames(m) <- c(0:5, "Over 5", "Sum")
>
> m
       Observed    Poisson   Neg. Binom.
0           447    406.312       442.907
1           132    189.026       138.546
2            42     43.970        44.364
3            21      6.819        14.315
4             3      0.793         4.637
5             2      0.074         1.505
Over 5        0      0.006         0.726
Sum         647    647.000       647.000
```

```
> ############ PLOT COMPARATIVE HISTOGRAMS ######################################
>
> clr<-rainbow(15)
> plot(c(0:6,0:6,0:6),c(acc, epois,enb), type="n",
+ main = "Distribution of Number of Accidents", ylab = "Number of Women",
+ xlab = "Number of Accidents")
> points((0:6)-.15,acc, type = "h", col = clr[1], lwd = 10)
> points((0:6),epois, type = "h", col = clr[5], lwd = 10)
> points((0:6)+.15,enb, type = "h", col = clr[11], lwd = 10)
> legend(3.5, 425, legend = c("Observed", "Poisson Pred.",
+ "Neg. Binomial Pred."), fill = c(2,3,4))
> # M3074Munitions1.pdf
```


Distribution of Number of Accidents
Observed
Poisson Pred.
Neg. Binomial Pred.
Number of Women
0
100
200
300
400
0
1
2
3
4
5
6
Number of Accidents

```
> ################# DO CHI-SQUARED TESTS ON FIT ################################
> # Lump 4 or over into single bin
> acc1 <- c(acc[1:4],acc[5]+acc[6]+acc[7])
> acc1
[1] 447 132  42  21   5

> probpois1 <- c(probpois[1:4],probpois[5]+probpois[6]+probpois[7])
> sum(probpois1)
[1] 1
> probnb1 <- c(probnb[1:4],probnb[5]+probnb[6]+probnb[7])
> sum(probnb1)
[1] 1
> chisq.test(acc1,p=probpois1)

Chi-squared test for given probabilities

data:  acc1
X-squared = 70.3724, df = 4, p-value = 1.894e-14

Warning message:
In chisq.test(acc1, p = probpois1) :
  Chi-squared approximation may be incorrect
> # Since one parameter was estimated, the critical .05 chisq and p-value are
> qchisq(.05,3,lower.tail=F)
[1] 7.814728
> pchisq(70.3724,3,lower.tail=F)
[1] 3.552333e-15
>
>
> chisq.test(acc1,p=probnb1)

Chi-squared test for given probabilities

data:  acc1
X-squared = 4.1026, df = 4, p-value = 0.3923

> # Since two parameter were estimated, the critical .05 chisq and p-value are
> qchisq(.05,2,lower.tail=F)
[1] 5.991465
> pchisq(4.1026,2,lower.tail=F)
[1] 0.1285677


> # It seems that we cannot reject H0: distribution is neg. binom,
> # but we do reject H0: distribution is Poisson. However, the test is not
> # valid: the parameters for nb were not MLE estimators for binned data (for
> # that matter neither was the parameter for Poisson!)
```