Opinionated Lessons in Statistics

by Bill Press

#26 The Poisson Count Pitfall

The Poisson-count pitfall: $\chi^2 = \sum_i \frac{(x_i - \mu_i)^2}{\mu_i}$ is actually not *Chisquare* !

You can get a statistic that is "accurately" chi-square either by summing (any number of) terms that are accurately squares of Normal t-values,

or by summing a large number of terms that individually have the same mean and variance as normal t^2 values. This uses the CLT, so the exactness of chi-square is no better than its normal approximation.

Compute moments of chi-square with 1 d.f.:

```
In[31]:= py = (1 / (Sqrt[2 Pi y])) Exp[-(1 / 2) y]
```

Out[31]=

$$\frac{e^{-y/2}}{\sqrt{2\pi}\sqrt{y}}$$

```
In[32]:= Integrate[py {1, y, y^2}, {y, 0, Infinity}]
```

Out[32]=

```
{1, 1, 3}
```

$$\text{So,} \quad \mu = 1, \quad \sigma^2 = 3 - 1 = 2$$

$$\text{Hence,} \quad \text{Chisquare}(\nu) \rightarrow \text{Normal}(\nu, \sqrt{2\nu}) \text{ as } \nu \rightarrow \infty$$

If you are going to rely on the CLT and sum up lots of not-exactly-t bins, they must have the expected mean and variance.

Poisson doesn't have. (People often get this wrong!)

```
In[39]:= poi[n_] := Exp[-mu] mu^n / n!

In[48]:= poimean = Sum[ n poi[n], {n, 0, Infinity}]

Out[48]=
         mu
```

OK

```
In[50]:= poivar =
          Simplify[Sum[ n^2 poi[n], {n, 0, Infinity}] -
            poimean^2]

Out[50]=
         mu
```

OK

```
In[51]:= tmean = Sum[ ((n - mu) ^2 / mu) poi[n], {n, 0, Infinity}]

Out[51]=
         1
```

OK

```
         tvar =
          Simplify[
           Sum[ ((n - mu) ^2 / mu) ^2 poi[n], {n, 0, Infinity}] -
            tmean^2]

Out[53]=
         2 + 1/mu
```

Not OK!

Poisson, Pearson chi-square statistic: $\chi^2 = \sum_i \frac{(x_i - \mu_i)^2}{\mu_i}$

We now know that this χ^2 is <u>not</u> Chi-square distributed! Rather, asymptotically,

$$\chi^2 \sim \text{Normal}\left(\nu, 2\nu + \sum_i \mu_i^{-1}\right)$$

What about bins with μ near zero? (Decide in advance!)

I wonder if Modified Neymann, $\chi^2 = \sum_i \frac{(n_i - Np_i)^2}{n_i + \alpha}$ is any closer to true chisquare?

```
poi = @(n,mu) exp(-mu).*mu.^n./factorial(n);
mus = [0.1 0.5 1.0 1.5 2.0 3 5 7 10 20 30];
nsum = 200;
for j=1:numel(mus),
   mu = mus(j);
   pois = poi(0:nsum,mu);
   ts = ((0:nsum)-mu).^2 ./ mu;
   tas = ((0:nsum)-mu).^2 ./ ((0:nsum)+1);
   tmean = sum(ts.*pois);
   tamean = sum(tas.*pois);
   tvar = sum(ts.^2.*pois)-tmean^2;
   tavar = sum(tas.^2.*pois)-tamean^2;
   fprintf(1,'%4.1f %8.5f %8.5f %8.5f
      %8.5f\n',mu,tmean,tamean,tvar,tavar);
end
```

```
 0.1  1.00000  0.05147 12.00000  0.01954
 0.5  1.00000  0.27061  4.00000  0.05447
 1.0  1.00000  0.52848  3.00000  0.25011
 1.5  1.00000  0.73696  2.66667  0.80999
 2.0  1.00000  0.89099  2.50000  1.70583
 3.0  1.00000  1.06780  2.33333  3.85907
 5.0  1.00000  1.15149  2.20000  6.40207
 7.0  1.00000  1.13452  2.14286  6.24527
10.0  1.00000  1.09945  2.10000  4.88251
20.0  1.00000  1.05000  2.05000  3.10583
30.0  1.00000  1.03333  2.03333  2.68296
```

Wow, it's much worse! I never knew that!
Verdict: Don't use Modified Neymann for a goodness-of-fit test unless the number of counts is way large!