Tutorial: Poisson distribution in Stata

Using Stata to calculate Poisson probabilities

Suppose X is a random variable that follows a Poisson distribution; X is a count of breast cancer cases.

When $X \sim$ Poisson(m),

`poissonp(m,k)`	returns the probability of observing floor(k) successes
`poisson(m,k)`	returns the probability of observing floor(k) or fewer successes
`poissontail(m,k)`	returns the probability of observing floor(k) or more successes

Example: Ecological Cancer Study

In the United States, the National Cancer Institute (NCI) tracks cancer incidence through the Surveillance Epidemiology and End Results (SEER) database. At various SEER sites, incident cases of cancer, cancer type, and location are tracked. Using data from SEER, epidemiologists can monitor patterns in disease risk and find factors, such as socioeconomic status, that are correlated with disease.

For instance, Los Angeles County is divided into 2,056 census tracts in the 2000 census. Using the SEER database, we can estimate the number of expected breast cancer cases in each census tract, based on breast cancer incidence rates in California and the age distribution within each tract (see standardization lectures). Then, we can compare the number of observed cases in each census tract to the expected, to determine if census tracts have more cases of cancer than expected. We can then try to correlate excess breast cancer cases with other area-level variable, in an ecological study.

Below, we have data on breast cancer incidence for the African-American female population in a census tract in LA County. We choose to model the observed number of breast cancer cases in a census tract using the Poisson distribution, with mean equal to the expected number of breast cancer cases in the census tract.

Age group	Observed	Population	Cancer rate (per 1,000 p-y)	Expected
15-24	0	188	0.008	0.001
25-34	0	163	0.200	0.033
35-44	0	216	0.875	0.189
45-54	0	157	1.868	0.293
55-64	0	137	2.633	0.361
65-74	0	151	3.165	0.478
75-84	0	121	3.452	0.418
84+	0	57	3.313	0.189
Total	0	1,190	1.648	1.962

Table 1: Census tract 1

1. What is the expected number of women with breast cancer in the census tract 1?

   1.962

2. What is the typical departure of the number of women with breast cancer from this mean number?

$$\begin{aligned} sd(X) &= \sqrt{var(X)} \\ &= \sqrt{\mu} \\ &= \sqrt{1.962} \\ &= 1.400714 \end{aligned}$$

3. Does the Poisson distribution provide an appropriate model?

   Count data, so Poisson distribution seems reasonable. Difficult to assess any more information about model fit without data on many census tracts.

4. What is the probability that exactly 0 women develop breast cancer in census tract 1? (use the formula)

$$\frac{e^{-1.962}1.962^0}{0!} = e^{-1.962} = 0.1406$$

Consider another census tract, with a similar total African-American female population to the previous, but with 5 observed breast cancer cases.

Age group	Observed	Population	Cancer rate (per 1,000 p-y)	Expected
15-24	0	187	0.008	0.001
25-34	0	187	0.200	0.037
35-44	1	218	0.875	0.191
45-54	0	193	1.868	0.361
55-64	1	175	2.633	0.461
65-74	1	141	3.165	0.446
75-84	2	66	3.452	0.228
84+	0	17	3.313	0.056
Total	5	1,184	1.504	1.781

Table 2: Census Tract 2.

5. What is the probability that exactly 5 women have breast cancer in census tract 2?

```
. di poissonp(1.781, 5)
.02515706
```

6. What is the probability that at least 5 women have breast cancer in census tract 2?

```
. di poissontail(1.781, 5)
.03504886
```

Alternatively, we could use the `poisson` command to calculate this probability, since $P(X \geq 5) = 1 - P(X \leq 4)$.

```
 di 1 - poisson(1.781, 4)
.03504886
```

Takeaway: Census tracts 1 and 2 have similar population sizes and consequently similar expected breast cancer case counts. However, in census tract 1, we observe no cases; in census tract 2, we observe 5 cases. Using the Poisson distribution, we can calculate the probability of observing case counts as extreme as 0 or 5 in these tracts.

Remember that there are about 2,000 total tracts, so we expect to see some extreme observations. We could also incorporate ecological covariates into our analysis, such as median household income or land-use data, to try to explain some of the differences between observed and expected breast cancer rates.