STA 138 Discussion 3 Solutions

Fall 2020

Data analysis

The table below contains the days on which 350 randomly sampled births occured.

(a) Is there evidence at significance level 0.05 to conclude that the days of the week differ in the proportions of births that occur in them? H_0: P_1=...=P_7=1/7 the test statistic is chi-square distributed with df=7-1=6

(b) Do any of the days of the week stand out to you as seemingly notable in this respect?

To answer these, we can make a table of observed vs. expected counts:

chisq.test( ) in R

	Observed (Y_i)	Expexted ($n\pi_{j0}$)	$(Y_j - n\pi_j)^2/(n\pi_j)$
Sunday	33.00	50.00	5.78
Monday	41.00	50.00	1.62
Tuesday	63.00	50.00	3.38
Wednesday	63.00	50.00	3.38
Thursday	47.00	50.00	0.18
Friday	56.00	50.00	0.72
Saturday	47.00	50.00	0.18

1-pchisq(15.24, df = 6)

The resulting χ^2 test statistic, obtained by summing over the last column of this table, has value 15.24. From the null distribution (χ^2_6 in this case), then, we get a p-value of 0.0185. As this is less than 0.05, we would reject the null hypothesis here, concluding that there is evidence at this level of non-uniform births over the days of the week.

We can see from the table that the largest single component to the test statistic is from births on Sundays, which are notably less common in our sample than uniformity would suggest.

Exam review

1. A salesperson will cold-call ten phone numbers independently chosen from the phonebook. Suppose that, when called, $\frac{1}{100}$ of the numbers in the phone book will result in a big sale, $\frac{9}{100}$ of the numbers will result in a small sale, and the result will result in no sale whatsoever.

   (a) From the multinomial distribution, we have: Multinomial(10, 0.9, 0.09, 0.01)

   $$\frac{10!}{9!0!1!}\left(\frac{9}{10}\right)^9\left(\frac{9}{100}\right)^1\left(\frac{1}{100}\right)^0 \approx 0.349$$

   (b) Similarly:

   $$\frac{10!}{9!1!0!}\left(\frac{9}{10}\right)^9\left(\frac{9}{100}\right)^0\left(\frac{1}{100}\right)^1 \approx 0.039$$

   (c) Multivariate hypergeometric

   (d) No, because the sample size would be much smaller than the population size, so the multinomial distribution would approximate the corresponding multivariate hypergeometric.

poisson(np)/normal(np, np(1-p)) approximate binomial(n, p)

2. Suppose that the numbers of aphids on a tomato plant are recorded each day. Suppose further that some days are sunny and others are cloudy; and on sunny days, the average number of aphids is smaller than it is on cloudy days. We can think of whether the day is cloudy or not as represented by a random coin flip determining whether a Poisson variable with a smaller or larger parameter is being sampled. The recorded numbers of aphids from day to day would then be overdispersed.

   For Poisson distribution, we would expect EX=VarX, but often find VarX>EX (overdispersion).

The largest k st $2\sum_{i=0}^{k} P(X=i) \le p$

3. Suppose that a Binomial exact test is carried out with $n = 12$ and $\pi_0 = 0.5$.

H_0: \pi=\pi_0, the test statistic is the number of successes in the n=12 trials.

(a) Using the table below, we can compute the probability that the $p-$value is less than or equal to 0.03 to be approximately $2(0.0029 + 0.0002) = 0.0062$.

(b) Using the table below, we can compute the probability that the $p-$value is less than or equal to 0.02 to be approximately $2(0.0029 + 0.0002) = 0.0062$.

(c) Using the table below, we can compute the probability that the $p-$value is less than or equal to 0.01 to be approximately $2(0.0029 + 0.0002) = 0.0062$.

k=2, 0.0384>0.03

x	0	1	2	3	4	5	6	7	8	9	10	11	12
$P_0(X = x)$	0.0002	0.0029	0.0161	0.0537	0.1208	0.1934	0.2256	0.1934	0.1208	0.0537	0.0161	0.0029	0.0002
p^*	0.0005	0.0063	0.0386	0.1460	0.3877	0.7744	1.0000	0.7744	0.3877	0.1460	0.0386	0.0063	0.0005

In general, the p-value is defined as $2\times P(X\ge k)$ if $k>np$, $2\times P(X\le k)$ if $k<np$, 1 if $k=np$

In ideal situations, with a continuous test statistic and no systematic uncertainties, and assuming the null hypothesis H0 is correct, p-values will be uniformly distributed between 0 and 1. In contrast, when the data is discrete rather than continuous (e.g. for a Poisson distribution, where the data values are only integers), the possible p-values are also discrete, are not uniformly spaced in p, and do not have equal weights.

Classical mechanisms leading to overdispersion:

1. *The clustered Poisson process*

Poison of pois son

$$y = z_1 + \cdots + z_N,\quad z_i\ iid,\quad N \sim \text{Poisson},\quad N \text{ independent of } z_i's.$$

Examples:

- Line transect sampling: Z_i = no. of animals, animals sighted in clusters around transect line.
- Insured accidents: N= no of accidents; z_i = damage at i-th accident.

Then, by independence of N and z_i,

$$\begin{aligned}
E(y) &= E(E(y|N)) = E(N(Ez_1)) = (EN)(Ez_1) \\
\text{var}(y) &= E(\text{var}(y|N)) + \text{var}(E(y|N)) \\
&= E(N\text{var}(z_1)) + \text{var}(NEz_1) \\
&= (EN)\text{var}(z_1) + (Ez_1)^2\text{var}(N) \quad \text{VarN=EN} \\
&= (EN)(Ez_1^2) \quad (\text{as } EN = \text{var}(N),\ Ez_1^2 = \text{var}(z_1) + (E(z_1))^2) \\
&> Ey \qquad \boxed{\text{if } Ez_1^2 > Ez_1}
\end{aligned}$$

Binomial(1, p)

2. *Mixture model*

$X \sim \text{Bernoulli}(p)$ then $EX = p$, $VarX = p(1-p)$

$$Y \sim \begin{cases} \text{Poisson}(\lambda_1) & \text{if } X=0 \\ \text{Poisson}(\lambda_2) & \text{if } X=1 \end{cases}$$

(1)
$$\begin{aligned}
EY &= E(E(Y|X)) \\
&= pE(Y|X=1) + (1-p)E(Y|X=0) \\
&= p(\lambda_1 - \lambda_2) + \lambda_2
\end{aligned}$$

Code Appendix

```r
library(xtable)
Obs <- c(33,41,63,63,47,56,47)
Exp <- sum(Obs)*rep(1/7,7)
Mat <- cbind(Obs,Exp,(Obs-Exp)^2/Exp)
  rownames(Mat) <- c("Sunday","Monday","Tuesday","Wednesday","Thursday","Friday","Saturday")
# xtable(Mat)
pmass <- dbinom(0:12,12,0.5)
pval <- c(2*pbinom(0:5,12,0.5),1,2*pbinom(5:0,12,0.5))
# xtable(rbind(0:12,pmass,pval),digits=4)
```

$$(2)\ \mathrm{Var}Y = \overset{\textcircled{1}}{E(\mathrm{Var}(Y|X))} + \overset{\textcircled{2}}{\mathrm{Var}(E(Y|X))}$$

$$\mathrm{Var}(Y|X=1) = \lambda_1 \ ,\ \mathrm{Var}(Y|X=0) = \lambda_2$$

$$E(Y|X=1) = \lambda_1 \ ,\ E(Y|X=0) = \lambda_2$$

$$\therefore \textcircled{1} = p(\lambda_1 - \lambda_2) + \lambda_2$$

$$\textcircled{2} = p(\lambda_1 - (p\lambda_1 + (1-p)\lambda_2))^2 + (1-p)(\lambda_2 - (p\lambda_1 + (1-p)\lambda_2))^2$$

$$= p(1-p)^2(\lambda_1 - \lambda_2)^2 + p^2(1-p)(\lambda_1 - \lambda_2)^2$$

$$= p(1-p)(\lambda_1 - \lambda_2)^2$$

$$\therefore \mathrm{Var}Y = \textcircled{1} + \textcircled{2} = p(\lambda_1 - \lambda_2) + \lambda_2 + p(1-p)(\lambda_1 - \lambda_2)^2$$