Statistics and Data Analysis

Lecture 9

Lecture by Dr. Zohar Yakhini
Typeset by Steven Karas

2017-01-03
Last edited 20:59:40 2017-01-03

1 Homework comments

HWA1 has been graded, but grades are not yet 100% on Moodle yet. Please write names on the homework (ID numbers are not easy to map back). Homeworks are graded as printed copies. He strongly prefers that submissions include at the least a PDF. Despite stating that he doesn't want code, he wants to see code. Also, in some cases, a proof was necessary, even though some people did not provide them. In the SH/RG rent distribution problem, what happens if we replace IQR with the $Q_{95} - Q_5$?

HW3 will be in pairs. The data will be significantly larger than HW2, but more open ended.

In case of future inquiries, it is recommended to contact him via zohar.yakhini [AT] gmail.com.

2 Agenda

1. Haberman analysis
2. MLE, fitting distributions

3 Student's t-test

The probably of seeing the observed difference in means if two independent and identically distributed samples of sufficient size are drawn from distributions with the same means and well behaved variances (for CLT).

We want to determine if $\mu_A = \mu_B$. If we want to assess if $\mu_A > \mu_B$, then we check $\mu_A^* > \mu_B^*$ (otherwise accepting the null: $\mu_A^* \leq \mu_B^*$) and apply ttest2 and reject the null at $1 - p/2$.

4 Haberman data

The dataset contains morbidity cases from a study that was conducted from 1958-1970 at U of Chicago Billing's Hospital on therapeutic (not prophylactic) surgery for breast cancer.

4.1 Data description

1. Age of patient at time of operation
2. Year of operation - 1900
3. Number of positive axillary nodes detected
4. Survival status – 1 = survived at least 5 years, 2 = died within 5 years

4.2 Analysis

Correlation between year and morbidity. T-test between age of patient sliced by morbidity.

4.3 Matlab Code

On slide 8[1].

Listing 1: Comparing Young vs Old number of LNMs

```
with open('haberman.csv') as file:
  data = [[float(field) for field in line.split(',')]
    for line in file.readlines()]

young = [record for record in data if record[0] < 50]
old = [record for record in data if record[0] >= 60]

young_lnm = [record[2] for record in young]
old_lnm = [record[2] for record in old]

y_tot = len(young_lnm)
o_tot = len(old_lnm)

# Plot them against each other

import numpy
by_lnm, bins = numpy.histogram(young_lnm, 20)
bo_lnm, _ = numpy.histogram(old_lnm, bins)

y_freq = by_lnm / y_tot
o_freq = bo_lnm / o_tot
```

4.4 Fitting to a Poisson distribution

Note that graphing the fitted poissonic distribution shows that it is a very poor fit.

NOTE: the poisson compare function is useful for checking if the fit is a good one, and gives us the error. Dr. Yakhini promised to upload some additional code that explains this.

[1] I've translated this into some roughly equivalent python code

5 Negative Binomial Distribution

The distribution of the needed Bernoulli trials for r successes. The Geometric distribution is a special case where $r = 1$.

For this number to equal y, we should have exactly r-1 successes in first y-1 trials, followed by a success.

$$p(y) = \binom{y-1}{r-1} p^r (1-p)^{y-r} \qquad y = r, r+1, ...$$

$$E(Y) = \frac{r}{p}$$

$$V(Y) = \frac{r(1-p)}{p^2}$$

5.1 Polya

Matlab has a function called PolyaComp which is a generalized negative binomial comparison function.

A useful way to slice the data is also "metastatic" instances vs. non-metastatic instances (where LNM=0).

Listing 2: Comparing Young vs Old morbidity

```
with open('haberman.csv') as file:
  data = [[float(field) for field in line.split(',')]
    for line in file.readlines()]

young = [record for record in data if record[0] < 50]
old = [record for record in data if record[0] >= 60]

young_survival = [record[3] for record in young]
old_survival = [record[3] for record in old]

y_tot = len(young_survival)
o_tot = len(old_survival)

y_sct = sum(1 for r in young_survival if r == 1)
o_sct = sum(1 for r in old_survival if r == 1)

y_survival_rate = y_sct / y_tot
o_survival_rate = o_sct / o_tot
```

5.2 Normal approximation/fit

We want to specify a standard procedure for fitting a normal distribution. In the future, we'll discuss how to evaluate these fittings, and expand it to other distribution families.

Assume we have the observations $x_1, ..., x_n$ and we want to fit the best normal distribution for this data. Let $f(x \mid \mu, \sigma)$ denote the density function for $N(\mu, \sigma^2)$. If the data is $N(\mu, \sigma^2)$ then we define the joint density of the data to be:

$$J(x_1, ..., x_n \,|\, \mu, \sigma) = f(x_1 \,|\, \mu, \sigma) \cdot ... \cdot f(x_n \,|\, \mu, \sigma)$$

We then define the likelihood of μ, σ given the observed data to be:

$$L(\mu, \sigma \,|\, x_1, ..., x_n) = J(x_1, ..., x_n \,|\, \mu, \sigma)$$

We call the μ_0, σ_0 that maximize the above is the maximum likelihood estimator (MLE) for the distribution that generates the observed data.

Example: Coin Tossing The parameter is obviously p. Let's say that the result was 30 heads out of 100 tosses. Thus, for any p, we have that $L(p|D) = Prob(D|p) = \binom{100}{30} p^{30}(1-p)^{70}$. We can take the derivative with respect to p to get:

$$\frac{dL}{dp} = C \cdot [(30p^{29} \cdot (1-p)^{70}) + 70p^{30}(-1)(1-p)^{69}]$$

Setting this to 0, we get that $0 = p^{29}(1-p)^{39}[30 - 100p]$. The edges $p = 0$ and $p = 1$ yield $L = 0$ and MLE is $p = 0.3$.

Gaussian MLE Note, this isn't algebra, but rather a rough explanation that was done on the whiteboard:

$$P(D|\Theta) = \frac{P(D \cap \Theta)}{P(\Theta)}$$

$$L(\Theta|D) = \frac{P(D \cap \Theta)}{P(D)} = \frac{P(\Theta)}{P(D)} \cdot P(D|\Theta)$$

$$L(\mu, \sigma|x_1, ..., x_n) = \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{1}{2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma}\right)$$

$$= \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{1}{2}} \exp\left(-\frac{\sum_{i=1}^{n}(x_i - \sigma)^2}{2\sigma}\right)$$

Now take the log and then the derivative:

$$\log(L(\mu, \sigma|D)) = -\frac{n}{2} - \log(2\pi\sigma^2) - \frac{1}{2\sigma}\left(\sum (x_i - \mu)^2\right)$$

$$\frac{\partial}{\partial \mu} L(\mu, \sigma) = -\frac{1}{2\sigma}\sum_{i=1}^{n}(-2x_i + 2\mu)$$

$$0 = -\sum x_i + n\mu$$

$$\mu = \overline{x_n}$$

5.3 Lognormal fitting

Sometimes lognormal is better than normal.

5.4 Pearson χ^2 test

Tells us which distribution family we should fit. Bin the data - n instances into r bins. Make sure that all bins have at least 5 samples. We observe binned data as O_i - the samples in the i-th bin. Assuming a certain distribution, we obtain the modeled probability of each bin: $p(i)$. We now compute the Pearson χ^2:

$$PQ = \sum_{i=1}^{r} \frac{(O_i - p(i)n)^2}{p(i)n}$$

When r is not too small, we get:

$$\frac{PQ - r}{\sqrt{2r}} \sim N(0,1)$$

5.5 ErrVecs

Note that we are expected to give the error vectors? in the homework?

Also, the notation used is inconsistent, and is sometimes:

$$PQ = \sum_{i=1}^{r} \frac{(\frac{O_i}{n} - p(i))^2}{\frac{p(i)}{n}}$$

6 Next week

1. Compound Poisson
2. Matched t-test
3. Wilcoxon rank sum test