Problem Set 6 Solution

17.872 TA Jiyoon Kim
November 17, 2003

Bulmer Excercise 6.5

(a) In the region of England, the variance will be obtained by dividing the sum of mean deviation squares. This will give $\sigma^2 = 2.1 \times 10^{-6}$. However, in theoretical mean, you need to use the formula of $p(1-p)/n$, which is $.514 \cdot (1 - .514)/100000 = 2.5 \times 10^(-6)$.
(b) In the same way, we can obtain the variance of regions of Dorset and it is 3.6×10^{-3}. A theoretical variance is $p(1-p)/n = .502 \cdot (1 - .502)/200 = 1.2 \times 10^{-3}$. As you've seen, the difference between variance of the sample from the theoretical variance is not much of difference in the region of England. Although it is still very small number, the difference in variances is much larger in the case of regions of Dorset.

Bulmer Problem 6.8

We've been through this pretty thoroughly in the recitation, but here it is again. In this question, the important difference between the distribution given here and binomial is **replacement**.

Instead of N which is a total number of balls in the urn, I use T. And, n is the number of trials, that is drawing.

$$\begin{aligned}
E(Z_i) &= p \cdot 1 + (1-p) \cdot 0 = p \\
&= \frac{R}{T} \\
Var(Z_i) &= \sum (z_i - E(Z_i))^2 \cdot p \\
&= (1-p)^2 \cdot p + (0-p)^2 \cdot (1-p) \\
&= p - p^2 = p(1-p) \\
&= \frac{R \cdot T - R}{T \cdot T}
\end{aligned}$$

$$\begin{aligned}
Cov(Z_i, Z_j) \quad &= E[(Z_i - E(Z_i))(Z_j - E(Z_j))] \\
&= E(Z_i \cdot Z_j) - E(Z_i)E(Z_j) \\
&= \frac{R(R-1)}{T(T-1)} - \frac{R \cdot R}{T \cdot T} \\
&= \frac{R^2 - RT}{T^2(T-1)} = -\frac{R(R-T)}{T^2(T-1)} \\
&= -p(1-p) \cdot \frac{1}{T-1} = -\frac{pq}{T-1}
\end{aligned}$$

where $q = 1 - p$

$E(Z_i \cdot Z_j)$ is basically joint probability which is given in the question, $\frac{R(R-1)}{T(T-1)}$.

Now, we need to calculate the expected value and variance of hypergeometric distribution. Since $Z = Z_1 + Z_2 + \cdots Z_n$, Z is $\sum_i^n Z_i$. Therefore,

$$\begin{aligned}
E(Z) \quad &= E(\textstyle\sum_i^n Z_i) = \sum_i^n E(Z_i) \\
&= \textstyle\sum_i^n p = np
\end{aligned}$$

This is as same as the expected value of binomial distribution. However, the variance shows difference due to the covariance term. Remember, binomial was independent n trials because there was replacement. That is, the previous result does not influence the result of now. Covariance of each trial until n, therefore, is zero. On the other hand, hypergeometric does not allow replacement. Therefore, previous result matters, which makes covariance of each trial non-zero.

$$\begin{aligned}
Var(Z) \quad &= Var(\textstyle\sum_i^n Z_i) \\
&= Var(Z_1 + Z_2 + \cdots + Z_n) \\
&= Var(Z_1) + Var(Z_2) + \cdots Var(Z_n) \\
&\quad +2Cov(Z_1, Z_2) + \cdots + 2Cov(Z_{n-1}, Z_n) \\
&= \textstyle\sum_i^n Var(Z_i) + 2\sum_i \sum_{j, i \neq j} Cov(Z_i, Z_j)
\end{aligned}$$

The question is how many covariance terms there are in the variance equation. This is like choosing 2 Z variables out of n Z variables. Since covariance term is a kind of pair or match of two Z variables and there are n Zs in total, you can simply use combination rule. (why? no order is considered. $Z_1 \cdot Z_2$ is as same as $Z_2 \cdot Z_1$.)

$$\begin{aligned} \begin{pmatrix} n \\ 2 \end{pmatrix} &= \frac{n!}{2!(n-2)!} \\ &= \frac{n(n-1)\cdot(n-2)!}{2\cdot(n-2)!} \\ &= \frac{n(n-1)}{2} \end{aligned}$$

Once you plug all the results in the equation,

$$\begin{aligned} Var(Z) &= \textstyle\sum_i^n Var(Z_i) + 2\sum_i \sum_{j,i\neq j} Cov(Z_i, Z_j) \\ &= npq + 2\cdot\frac{n(n-1)}{2}\cdot Cov(Z_i, Z_j) \\ &= npq + n(n-1)\cdot[-\frac{pq}{T-1}] \\ &= npq[1-\frac{n-1}{T-1}] \end{aligned}$$

Hypergeometric matters when there is relatively small size of population. If the sample size is very small compared to the population size, the second term will go to zero since a huge T makes the denominator also very big, therefore makes the whole second term as near to zero. In this case, there is no big difference between hypergeometric and binomial distributions.

Let's apply these results to the previous question in the book.

T is 10, and $R = 4$ and $B = 6$. We are drawing 3 balls, thus, $n = 3$. Drawing a red ball is considered as success. The expected value and variance of binomial is $E(X) = np = 3 \times .4 = .12$ and $Var(X) = npq = 3 \times .4 \times .6 = .72$, respectively. Hypergeometric expected value is as same as binomial, .12. However, the variance is $Var(X) = npq(1 - \frac{n-1}{T-1} = .72(1 - \frac{2}{9}) = .56$.

Bulmer Exercise 8.1

You can do it step by step. The question is asking you the probability s is less than 8.8. If you write this down, it will be $Pr(s < 8.8)$. $\mu = 100$ and $\sigma = 16$. When you have σ or σ^2, you have to consider taking χ^2 distribution to test.

$$\begin{aligned} Pr(s < 8.8) &= Pr\left(\frac{s^2(n-1)}{\sigma^2} < \frac{8.8^2(n-1)}{\sigma^2}\right) \\ &= Pr\left(\chi^2 < \frac{696.96}{16^2}\right) \\ &= Pr\left(\chi^2 < 2.72\right) \end{aligned}$$

If you look up the end of a stat book, there is χ^2 distribution table. There, please find the probability corresponding to χ^2. The book wrote the probability of exceeding a particular χ^2, but we want the probability which is below the χ^2. The table tells us the probability of a particular χ^2 is exceeding 2.72 with 9 degrees of freedom is about 97.5 %. Therefore, the probability that s is less than 8.8 is 100 - 97.5 = 2.5 %, which is .025 in proportional term.

Bulmer Problem 8.3

First, rather easier way to solve this question is the second method, using moment generating function. There is an important property of moment generating function. Please remember that

$$\begin{aligned} M_x(t) &= E(e^{tx}) = \int e^{tx} f(x)dx \quad \text{if continuous} \\ &= \sum e^{tx} p(x) \quad \text{if discrete} \\ M_{x+y}(t) &= E(e^{t(x+y)}) = E(e^{tx}e^{ty}) = M_x(t)\dot{M}_y(t) \end{aligned}$$

Chi-squared distribution's moment generating function is $(1 - 2t)^{-\frac{k}{2}}$ when it has k degrees of freedom. Therefore, if $U = Y_1 + Y_2$ and both Y_1 and Y_2 are chi-squared distributed with f_1 and f_2, Y_1 and Y_2's moment generating functions are $(1-2t)^{-\frac{f_1}{2}}$ and $(1-2t)^{-\frac{f_2}{2}}$, respectively. Then, using the property of moment generating function,

$$\begin{aligned} M_{y_1+y_2}(t) &= E(e^{t(y_1+y_2)}) = E(e^{ty_1}e^{ty_2}) = M_{y_1}(t)M_{y_2}(t) \\ &= (1-2t)^{-\frac{f_1}{2}}(1-2t)^{-\frac{f_2}{2}} = (1-2t)^{-\frac{f_1+f_2}{2}} \end{aligned}$$

The last line is a new moment generating function of U, and as its moment generating function shows, it is a chi-squared distributed with f_1+f_2 degrees of freedom.

Proving this in direct way is actually much harder and needs higher mathematical skill. First, go back to the Bulmer's 3.5. The formula you see is called *convolution integral.*

$$h(u) = \int f(u-y)g(y)dy$$

For convenience reason, let's note that $Y \sim \chi^2_{f_1}$ and $X \sim \chi^2_{f_2}$, and $U = X+Y$. The method using convolution integral needs to insert density functions. χ^2 density function for y is

$$\frac{1}{2^{\frac{f_1}{2}}\Gamma(\frac{f_1}{2})}y^{\frac{f_1}{2}-1}e^{-\frac{y}{2}}$$

Then, density function for $u-y$ is

$$\frac{1}{2^{\frac{f_2}{2}}\Gamma(\frac{f_2}{2})}(u-y)^{\frac{f_2}{2}-1}e^{-\frac{u-y}{2}}$$

Therefore, the $h(u)$ can be written as

$$h(u) = \int_0^u \left[\frac{1}{2^{\frac{f_1}{2}}\Gamma(\frac{f_1}{2})}y^{\frac{f_1}{2}-1}e^{-\frac{y}{2}}\right]\left[\frac{1}{2^{\frac{f_2}{2}}\Gamma(\frac{f_2}{2})}(u-y)^{\frac{f_2}{2}-1}e^{-\frac{u-y}{2}}\right]dy$$

$$= \frac{1}{2^{\frac{f_1+f_2}{2}}\Gamma(\frac{f_1}{2})\Gamma(\frac{f_2}{2})}e^{-\frac{u}{2}}\int_0^u y^{\frac{f_1}{2}-1}(u-y)^{\frac{f_2}{2}-1}dy$$

Please note that the boundary of u is given between 0 and some u. Since it is a distribution of the sum of two chi-squared distributions, it cannot be less than 0.

Let's do some substitution to solve the problem. Let $y = zu$ and z is any variable that makes this equation work. Then, $dy = u \cdot dz$, and we will insert this result into the integral equation. Since we are substituting dz for dy, we also need to change the range of the integration. When a previous integration was from 0 to u over dy, a new integration is from 0 to 1. Now, replacing y with zu and dy with udz yields

$$\int_0^1 (zu)^{\frac{f_1}{2}-1}(u-zu)^{\frac{f_2}{2}-1}udz$$

$$= \int_0^1 (z)^{\frac{f_1}{2}-1}(u)^{\frac{f_1}{2}-1}(u)^{\frac{f_2}{2}-1}(1-z)^{\frac{f_2}{2}-1}udz$$

$$= \int_0^1 (u)^{\frac{f_1+f_2}{2}-1}(z)^{\frac{f_1}{2}-1}(1-z)^{\frac{f_2}{2}-1}dz$$

$$= (u)^{\frac{f_1+f_2}{2}-1}\int_0^1 (z)^{\frac{f_1}{2}-1}(1-z)^{\frac{f_2}{2}-1}dz$$

(The above result only shows the inside calculation of the integral.)
By definition, the integral part is a beta distribution density function, which is

$$\int_0^1 (z)^{\frac{f_1}{2}-1}(1-z)^{\frac{f_2}{2}-1}dz = \frac{\Gamma(\frac{f_1}{2})\Gamma(\frac{f_2}{2})}{\Gamma(\frac{f_1+f_2}{2})}$$

Inserting the result into an original equation produces

$$h(u) = \frac{1}{2^{\frac{f_1+f_2}{2}}\Gamma(\frac{f_1}{2})\Gamma(\frac{f_2}{2})}e^{-\frac{u}{2}}u^{\frac{f_1+f_2}{2}-1}\frac{\Gamma(\frac{f_1}{2})\Gamma(\frac{f_2}{2})}{\Gamma(\frac{f_1+f_2}{2})}$$

$$= \frac{1}{2^{\frac{f_1+f_2}{2}}}\frac{1}{\Gamma(\frac{f_1+f_2}{2})}e^{-\frac{u}{2}}u^{\frac{f_1+f_2}{2}-1} \quad \sim \chi^2_{f_1+f_2}$$

As shown in the last line, u is chi-squared distributed (since it has a chi-square density function) with $f_1 + f_2$ degrees of freedom.

Polling Question
I took four polling reports from CNN, ABC, CBS and NBC.

	rating	n
CNN	.5	1004
ABC	.57	1023
CBS	.49	1177
NBC	.51	1003

Confidence interval can be obtained by calculating standard deviation of each poll. For instance, the variance of ABC poll is $\frac{.57\cdot(1-.57)}{1023} = .00024$. Standard deviation is, therefore, $\sqrt{.00024} = .01548$. 95 % confidence interval is $.57 \pm 1.96 \cdot .01548$. They are .6003 and .5397.

Mean Squared Error is obtained by calculating variance across polls. The average of five polls' approval rating is .5175 and the variance based on this average is $\sum_{i=1}^{4} \frac{(s_i - .5175)^2}{3} = .00129$, which is MSE.

If you consider that these 4 polls are using the same technique and random sampling, and there is no bias, you can calculate the variance of total by combining all 4 polls. For example, if there is .5 of approval rating in CNN poll, there were 502 people who approved of Bush. And, 583, 577 and 512 people approved of Bush in each poll in the order. Since the total number of samples of four polls is 4207, the approval rating of total is $2173/4207 = .5166$. The variance, therefore, is $\frac{.5166(1-.5166)}{4207} = .000059$. Compared with the previous result, it says that there is a bias in MSE calculation because it is much larger than the variance.