Important Result Related to Chi-Square Distribution and Its Statistical Application

LIN Guozhang (061801907)

This report is to explore an important result related to the chi-square distribution. That is, for i.i.d. X_1, ..., $X_n \sim N(\mu, \sigma^2)$, define the sample mean $\overline{X}$ and the sample variance S^2 as

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{1.1}$$

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 \tag{1.2}$$

Then,

$$(n-1)\frac{S^2}{\sigma^2} \sim \chi^2_{n-1} \tag{2}$$

Note that as proved in Nguyen Duc Thanh's report, *Estimating the population variance from a random sample*, the sample mean $\overline{X}$ and the sample variance S^2 are unbiased estimators, i.e.,

$$\mathbb{E}(\overline{X}) = \mu$$
$$\mathbb{E}(S^2) = \sigma^2,$$

That is an important reason why the sample mean and the sample variance are defined as (1.1) and (1.2).

Proof of (2)
To prove (2), a lemma has to be proved. It is as follows.

$\overline{X}$ and S^2 are independent.

We skip the proof of the lemma here. If interested, please refer to page 195 of *Mathematical Statistics and Data Analysis, Third Edition* by John A. Rice.

Now we proceed assuming the lemma is true.
In the report of *Linear Transformation of Normal Random Variables and Its Statistical Application* by Lin Guozhang, standardization of a normal random variable is discussed.
For the proof of (2), one standardizes X_i to obtain

$$\frac{X_i - \mu}{\sigma} \sim N(0,1) \tag{3}$$

In the report of *Chi-square distribution* by Lin Guozhang, it is proved that if independent $X_1, ..., X_n \sim N(0,1)$, then

$$\sum_{i=1}^{n} X_i^2 \sim \chi^2_n \tag{4}$$

By the two results above,

$$\sum_{i=1}^{n}(\frac{X_i-\mu}{\sigma})^2 \sim \chi_n^2 \tag{5}$$

Consider the LHS of (5),

$$\begin{aligned}
\sum_{i=1}^{n}(\frac{X_i-\mu}{\sigma})^2 &= \frac{1}{\sigma^2}\sum_{i=1}^{n}[(X_i-\overline{X})+(\overline{X}-\mu)]^2 \\
&= \frac{1}{\sigma^2}[\sum_{i=1}^{n}(X_i-\overline{X})^2+2\sum_{i=1}^{n}(X_i-\overline{X})(\overline{X}-\mu)+\sum_{i=1}^{n}(\overline{X}-\mu)^2] \\
&= \frac{1}{\sigma^2}[\sum_{i=1}^{n}(X_i-\overline{X})^2+2(n\overline{X}-n\overline{X})(\overline{X}-\mu)+\sum_{i=1}^{n}(\overline{X}-\mu)^2] \\
&= \frac{1}{\sigma^2}[\sum_{i=1}^{n}(X_i-\overline{X})^2+0+\sum_{i=1}^{n}(\overline{X}-\mu)^2] \\
&= \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\overline{X})^2+\frac{1}{\sigma^2}n(\overline{X}-\mu)^2 \\
&= \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\overline{X})^2+(\frac{\overline{X}-\mu}{\sigma/\sqrt{n}})^2
\end{aligned}$$

Thus,

$$\sum_{i=1}^{n}(\frac{X_i-\mu}{\sigma})^2 = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\overline{X})^2+(\frac{\overline{X}-\mu}{\sigma/\sqrt{n}})^2 \tag{6}$$

Define

$$W := \sum_{i=1}^{n}(\frac{X_i-\mu}{\sigma})^2, \text{ the LHS of (6)},$$

$$U := \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\overline{X})^2, \text{ the 1st term of the RHS of (6)},$$

$$V := (\frac{\overline{X}-\mu}{\sigma/\sqrt{n}})^2, \text{ the 2nd term of the RHS of (6)}.$$

Then,

$$W = U + V$$

By (5), $W \sim \chi_n^2$.
As proved in *Linear Transformation of Normal Random Variables and Its Statistical Application*, $\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$.
Therefore, according to *Chi-square distribution*, $V \sim \chi_1^2$.
By comparing (1.2) with U, one can see U is a function of S^2.
By comparing (1.1) with V, one can see V is a function of $\overline{X}$.
By the lemma stating that $\overline{X}$ and S^2 are independent, one finds that U and V are independent.
As proved in *Properties of the moment generating function* by Lin Guozhang, if X and Y are independent, then $M_{X+Y}(t) = M_X(t)M_Y(t)$, which is true on the common interval where both mgf's exist.
Thus,

$$M_W(t) = M_U(t)M_V(t) \tag{7}$$

As proved in *On the gamma distribution* by Matsumoto Kosuke and Nguyen Duc Thanh, the mgf of a random variable $X \sim \Gamma(w(>0), \lambda(>0))$ is

$$M_X(t) = (\frac{\lambda}{\lambda-t})^w, \text{ for } t < \lambda.$$

Since χ^2_n is equivalent to $\Gamma(\frac{n}{2}, \frac{1}{2})$, the mgf of a random variable $X \sim \chi^2_n$ is

$$M_X(t) = (\frac{1/2}{1/2 - t})^{n/2}, \text{ for } t < \frac{1}{2}. \tag{8}$$

Since $W \sim \chi^2_n$ and $V \sim \chi^2_1$, by (7) and (8),

$$(\frac{1/2}{1/2 - t})^{n/2} = M_U(t)(\frac{1/2}{1/2 - t})^{1/2}$$
$$M_U(t) = (\frac{1/2}{1/2 - t})^{(n-1)/2} \tag{9}$$

By comparing (9) with (8) and the uniqueness theorem of mgf,

$$U \sim \chi^2_{n-1} \tag{10}$$

Recall the definition of U and (1.2)

$$U := \frac{1}{\sigma^2} \sum_{i=1}^{n} (X_i - \overline{X})^2$$
$$= (n-1)\frac{S^2}{\sigma^2} \tag{11}$$

Thus, by (10) and (11), (2) is proved.

Q.E.D.

Application
This example is adapted from Example 11 on page 127 of *Biostatistical Analysis, International Edition* by Jerrold H. Zar.

A manufacturer is interested in whether the variability in the dissolving times of a drug is greater than 1.5 sec^2. Conduct n independent experiments of dissolving the drug in gastric juice. Let random variable X_i denote the dissolving time in the i-th experiment, where $i = 1, ..., n$. Assume $X_1, ..., X_n$ are i.i.d. and normal.

The measurement results are as follows.
Dissolving time (in sec) of a drug in gastric juice: 42.7, 43.4, 44.6, 45.1, 45.6, 45.9, 46.8, 47.6.
There are 8 data. Therefore, $n = 8$. The realization of S^2 is 2.6898 sec^2.

Perform hypothesis testing.

Null hypothesis (assumed to be true in the testing):

$$H_0 : \sigma^2 \leq 1.5 \text{ sec}^2.$$

Alternative hypothesis (all the other possibilities) :

$$H_a : \sigma^2 > 1.5 \text{ sec}^2.$$

Set the confidence level as 0.05.

$$U := (n-1)\frac{S^2}{\sigma^2}$$
$$= (8-1)\frac{S^2}{\sigma^2}$$
$$= 7\frac{S^2}{\sigma^2} \sim \chi^2_7$$

As discussed in the report of *Linear Transformation of Normal Random Variables and Its Statistical Application*, the criterion to reject the null hypothesis is

$$\text{p-value} \leq \text{confidence level}$$

In essence, p-value is the probability of obtaining a value that is equally or more extreme than its realization.
The null hypothesis is the popular variance $\sigma^2 \leq 1.5$ sec^2, which is assumed to be true in this testing.
Therefore, the larger value of the sample variance S^2 one obtains, the more extreme it is, and it provides stronger evidence against the null hypothesis.
Introduce S^2_{lim}. As long as one obtains a realization larger (more extreme) than S^2_{lim}, one then rejects the null hypothesis.
Then,

$$\mathbb{P}(S^2 \geq S^2_{\text{lim}}) = 0.05$$

Since U is an increasing function of S^2,

$$\mathbb{P}(U \geq U_{\text{lim}}) = 0.05$$

Referring to the chi-square table (URL in the section of Reference), one has

$$U_{\text{lim}} = 14.06$$

By (11) and $n = 8$, $\sigma^2 = 1.5$ sec^2 (the most conservative value in the null hypothesis),

$$S_{\text{lim}} = 3.013 \text{ sec}^2.$$

As mentioned one page before, the realization of S^2 is 2.6989 sec^2, smaller (less extreme) than S^2_{lim}. Therefore, the null hypothesis is accepted.

References

- *Probability, an introduction* from Grimmett and Welsh
- Lecture notes for SML: Probability by Richard, S.
- *Estimating the population variance from a random sample* by Nguyen Duc Thanh.
- *Properties of the moment generating function* by Lin Guozhang.
- *Mathematical Statistics and Data Analysis, Third Edition* by John A. Rice
- *Linear Transformation of Normal Random Variables and Its Statistical Application* by Lin Guozhang
- *Chi-square distribution* by Lin Guozhang
- *On the gamma distribution* by Matsumoto Kosuke and Nguyen Duc Thanh
- *Biostatistical Analysis, International Edition* by Jerrold H. Zar
- Chi-square table: `https://www.statisticshowto.com/tables/chi-squared-table-right-tail/`