Section 11

Goodness-of-fit for composite hypotheses.

Example. Let us consider a Matlab example. Let us generate 50 observations from $N(1,2)$:

```
X=normrnd(1,2,50,1);
```

Then, running a chi-squared goodness-of-fit test 'chi2gof'

```
[H,P,STATS]= chi2gof(X)
```

outputs

```
H = 0, P = 0.8793,
STATS = chi2stat: 0.6742
        df: 3
        edges: [-3.7292 -0.9249 0.0099 0.9447 1.8795 2.8142 5.6186]
        O: [8 7 8 8 9 10]
        E: [8.7743 7.0639 8.7464 8.8284 7.2645 9.3226]
```

The test accepts the hypothesis that the data is normal. Notice, however, that something is different. Matlab grouped the data into 6 intervals, so chi-squared test from previous lecture should have $r - 1 = 6 - 1 = 5$ degrees of freedom, but we have 'df: 3'! The difference is that now our hypothesis is not that the data comes from a *particular given* distribution but that the data comes from a *family* of distributions which is called a *composite* hypothesis. Running

```
[H,P,STATS]= chi2gof(X,'cdf',@(z)normcdf(z,mean(X),std(X,1)))
```

would test a simple hypothesis that the data comes from a particular normal distribution $N(\hat{\mu}, \hat{\sigma}^2)$ and the output

```
H = 0, P = 0.9838
STATS =   chi2stat: 0.6842
```

```
df: 5
edges: [-3.7292 -0.9249 0.0099 0.9447 1.8795 2.8142 5.6186]
O: [8 7 8 8 9 10]
E: [8.6525 7.0995 8.8282 8.9127 7.3053 9.2017]
```

has 'df: 5.' However, we **can not** use this test because we estimate the parameters $\hat{\mu}$ and $\hat{\sigma}^2$ of this distribution using the data so this is not a particular given distribution; in fact, this is the distribution that fits the data the best, so the T statistic in Pearson's theorem will behave differently.

$\square$

Let us start with a discrete case when a random variable takes a finite number of values $B_1, \ldots, B_r$ with probabilities

$$p_1 = \mathbb{P}(X = B_1), \ldots, p_r = \mathbb{P}(X = B_r).$$

We would like to test a hypothesis that this distribution comes from a family of distributions $\{\mathbb{P}_\theta : \theta \in \Theta\}$. In other words, if we denote

$$p_j(\theta) = \mathbb{P}_\theta(X = B_j),$$

we want to test

$$\begin{array}{ll} H_0: & p_j = p_j(\theta) \text{ for all } j \leq r \text{ for some } \theta \in \Theta \\ H_1: & \text{otherwise.} \end{array}$$

If we wanted to test H_0 for one particular fixed θ we could use the statistic

$$T = \sum_{j=1}^{r} \frac{(\nu_j - np_j(\theta))^2}{np_j(\theta)},$$

and use a simple chi-squared goodness-of-fit test. The situation now is more complicated because we want to test if $p_j = p_j(\theta), j \leq r$ at least for some $\theta \in \Theta$ which means that we have many candidates for θ. One way to approach this problem is as follows.

(Step 1) Assuming that hypothesis H_0 holds, i.e. $\mathbb{P} = \mathbb{P}_\theta$ for some $\theta \in \Theta$, we can find an estimate θ^* of this unknown θ and then

(Step 2) try to test if, indeed, the distribution $\mathbb{P}$ is equal to $\mathbb{P}_{\theta^*}$ by using the statistics

$$T = \sum_{j=1}^{r} \frac{(\nu_j - np_j(\theta^*))^2}{np_j(\theta^*)}$$

in chi-squared goodness-of-fit test.

This approach looks natural, the only question is what estimate θ^* to use and how the fact that θ^* also depends on the data will affect the convergence of T. It turns out that if we let θ^* be the maximum likelihood estimate, i.e. θ that maximizes the likelihood function

$$\varphi(\theta) = p_1(\theta)^{\nu_1} \ldots p_r(\theta)^{\nu_r}$$

then the statistic

$$T = \sum_{j=1}^{r} \frac{(\nu_j - np_j(\theta^*))^2}{np_j(\theta^*)} \to^d \chi^2_{r-s-1} \tag{11.0.1}$$

converges to χ^2_{r-s-1} distribution with $r - s - 1$ degrees of freedom, where s is the dimension of the parameter set Θ. Of course, here we assume that $s \le r - 2$ so that we have at least one degree of freedom. Very informally, by dimension we understand the number of free parameters that describe the set

$$\Big\{(p_1(\theta), \ldots, p_r(\theta)) : \theta \in \Theta\Big\}.$$

Then the decision rule will be

$$\delta = \begin{cases} H_1 : & T \le c \\ H_2 : & T > c \end{cases}$$

where the threshold c is determined from the condition

$$\mathbb{P}(\delta \ne H_0 | H_0) = \mathbb{P}(T > c | H_0) \approx \chi^2_{r-s-1}(c, +\infty) = \alpha$$

where $\alpha \in [0, 1]$ is the level of sidnificance.

Example 1. Suppose that a gene has two possible alleles A_1 and A_2 and the combinations of these alleles define three genotypes A_1A_1, A_1A_2 and A_2A_2. We want to test a theory that

$$\begin{array}{rl} \text{Probability to pass } A_1 \text{ to a child } & = \theta \\ \text{Probability to pass } A_2 \text{ to a child } & = 1 - \theta \end{array}$$

and that the probabilities of genotypes are given by

$$\begin{aligned} p_1(\theta) &= \mathbb{P}(A_1A_1) = \theta^2 \\ p_2(\theta) &= \mathbb{P}(A_1A_2) = 2\theta(1-\theta) \\ p_3(\theta) &= \mathbb{P}(A_2A_2) = (1-\theta)^2. \end{aligned} \tag{11.0.2}$$

Suppose that given a random sample $X_1, \ldots, X_n$ from the population the counts of each genotype are ν_1, ν_2 and ν_3. To test the theory we want to test the hypothesis

$$\begin{array}{ll} H_0 : & p_1 = p_1(\theta),\ p_2 = p_2(\theta),\ p_3 = p_3(\theta) \text{ for some } \theta \in [0,1] \\ H_1 : & \text{otherwise.} \end{array}$$

First of all, the dimension of the parameter set is $s = 1$ since the distributions are determined by one parameter θ. To find the MLE θ^* we have to maximize the likelihood function

$$p_1(\theta)^{\nu_1} p_2(\theta)^{\nu_2} p_3(\theta)^{\nu_3}$$

or, equivalently, maximize the log-likelihood

$$\begin{aligned} \log p_1(\theta)^{\nu_1} p_2(\theta)^{\nu_2} p_3(\theta)^{\nu_3} &= \nu_1 \log p_1(\theta) + \nu_2 \log p_2(\theta) + \nu_3 \log p_3(\theta) \\ &= \nu_1 \log \theta^2 + \nu_2 \log 2\theta(1-\theta) + \nu_3 \log(1-\theta)^2. \end{aligned}$$

If we compute the critical point by setting the derivative equal to 0, we get

$$\theta^* = \frac{2\nu_1 + \nu_2}{2n}.$$

Therefore, under the null hypothesis H_0 the statistic

$$\begin{aligned} T &= \frac{(\nu_1 - np_1(\theta^*))^2}{np_1(\theta^*)} + \frac{(\nu_2 - np_2(\theta^*))^2}{np_2(\theta^*)} + \frac{(\nu_3 - np_3(\theta^*))^2}{np_3(\theta^*)} \\ &\to^d \chi^2_{r-s-1} = \chi^2_{3-1-1} = \chi^2_1 \end{aligned}$$

converges to χ^2_1-distribution with one degree of freedom. Therefore, in the decision rule

$$\delta = \begin{cases} H_1 : & T \le c \\ H_2 : & T > c \end{cases}$$

threshold c is determined by the condition

$$\mathbb{P}(\delta \ne H_0 | H_0) \approx \chi^2_1(T > c) = \alpha.$$

For example, if $\alpha = 0.05$ then $c = 3.841$.

□

Example 2. A blood type O, A, B, AB is determined by a combination of two alleles out of A, B, O and allele O is dominated by A and B. Suppose that p, q and $r = 1 - p - q$ are the population frequencies of alleles A, B and O correspondingly. If alleles are passed randomly from the parents then the probabilities of blood types will be

Blood type	Allele combinations	Probabilities	Counts
O	OO	r^2	$\nu_1 = 121$
A	AA, AO	$p^2 + 2pr$	$\nu_2 = 120$
B	BB, BO	$q^2 + 2pr$	$\nu_3 = 79$
AB	AB	$2pq$	$\nu_4 = 33$

We would like to test this theory based on the counts of each blood type in a random sample of 353 people. We have four groups and two free parameters p and q, so the chi-squared statistics T under the null hypotheses will have $\chi^2_{4-2-1} = \chi^2_1$ distribution with one degree of freedom. First, we have to find the MLE of parameters p and q. The log likelihood is

$$\begin{aligned} &\nu_1 \log r^2 + \nu_2 \log(p^2 + 2pr) + \nu_3 \log(q^2 + 2qr) + \nu_4 \log(2pq) \\ &= 2\nu_1 \log(1 - p - q) + \nu_2 \log(2p - p^2 - 2pq) + \nu_3 \log(2q - q^2 - 2pq) + \nu_4 \log(2pq). \end{aligned}$$

Unfortunately, if we set the derivatives with respect to p and q equal to zero, we get a system of two equations that is hard to solve explicitly. So instead we can minimize log likelihood numerically to get the MLE $\hat{p} = 0.247$ and $\hat{q} = 0.173$. Plugging these into formulas of blood type probabilities we get the estimated probabilities and estimated counts in each group

	O	A	B	AB
$\hat{p}_i$	0.3364	0.3475	0.2306	0.0855
$n\hat{p}_i$	118.7492	122.6777	81.4050	30.1681

We can now compute chi-squared statistic $T \approx 0.44$ and the p-value $\chi^2_1(T, \infty) = 0.5071$. The data agrees very well with the above theory.

□

We could also use a similar test when the distributions $\mathbb{P}_\theta, \theta \in \Theta$ are not necessarily supported by a finite number of points $B_1, \ldots, B_r$, for example, continuous distributions. In this case if we want to test the hypothesis

$$H_0 : \mathbb{P} = \mathbb{P}_\theta \text{ for some } \theta \in \Theta$$

we can group the data into r intervals $I_1, \ldots, I_r$ and test the hypothesis

$$H_0 : p_j = p_j(\theta) = \mathbb{P}_\theta(X \in I_j) \text{ for all } j \leq r \text{ for some } \theta.$$

For example, if we discretize normal distribution by grouping the data into intervals $I_1, \ldots, I_r$ then the hypothesis will be

$$H_0' : p_j = N(\mu, \sigma^2)(I_j) \text{ for all } j \leq r \text{ for some } (\alpha, \sigma^2).$$

There are two free parameters μ and σ^2 that describe all these probabilities so in this case $s = 2$. Matlab function 'chi2gof' tests for normality by grouping the data and computing statistic T in (11.0.1) - that is why it uses χ^2_{r-s-1} distribution with

$$r - s - 1 = r - 2 - 1 = r - 3$$

degrees of freedom and, thus, 'df: 3' in the example above.

Example. Let us test if the data 'normtemp' from normal body temperature dataset fits normal distribution.

```
[H,P,STATS]= chi2gof(normtemp)
```

gives

```
H = 0, P = 0.0504
STATS =  chi2stat: 9.4682
         df: 4
         edges: [1x8 double]
         O: [13 12 29 27 35 10 4]
         E: [9.9068 16.9874 27.6222 31.1769 24.4270 13.2839 6.5958]
```

and we accept null hypothesis at the default level of significance $\alpha = 0.05$ since p-value $0.0504 > \alpha = 0.05$. We have $r = 7$ groups and, therefore, $r - s - 1 = 7 - 2 - 1 = 4$ degrees of freedom.

□

In the case when the distributions $\mathbb{P}_\theta$ are continuous or, more generally, have infinite number of values that must be grouped in order to use chi-squared test (for example, normal or Poisson distribution), it can be a difficult numerical problem to maximize the "grouped" likelihood function

$$\mathbb{P}_\theta(I_1)^{\nu_1} \cdot \ldots \cdot \mathbb{P}_\theta(I_r)^{\nu_r} \to \max_\theta \to \theta^*.$$

It is tempting to use a usual non-grouped MLE $\hat{\theta}$ of θ instead of the above θ^* because it is often easier to compute, in fact, for many distributions we know explicit formulas for these MLEs. However, if we use $\hat{\theta}$ in the statistic

$$T = \sum_{j=1}^{r} \frac{(\nu_j - np_j(\hat{\theta}))^2}{np_j(\hat{\theta})} \tag{11.0.3}$$

then it will no longer converge to χ^2_{r-s-1} distribution. A famous result in [1] proves that typically this T will converge to a distribution "in between" χ^2_{r-s-1} and χ^2_{r-1}. Intuitively this is easy to understand because θ^* specifically fits the grouped data $\nu_1, \ldots, \nu_r$ so the expected counts

$$np_1(\theta^*), \ldots, np_r(\theta^*)$$

should be a better fit compared to the expected counts

$$np_1(\hat{\theta}), \ldots, np_r(\hat{\theta}).$$

On the other hand, these last expected counts should be a better fit than simply using the true expected counts

$$np_1(\theta_0), \ldots, np_r(\theta_0)$$

since the MLE $\hat{\theta}$ fits the data better than the true distribution. So typically we would expect

$$\sum_{j=1}^{r} \frac{(\nu_j - np_j(\theta^*))^2}{np_j(\theta^*)} \leq \sum_{j=1}^{r} \frac{(\nu_j - np_j(\hat{\theta}))^2}{np_j(\hat{\theta})} \leq \sum_{j=1}^{r} \frac{(\nu_j - np_j(\theta_0))^2}{np_j(\theta_0)}.$$

But the left hand side converges to χ^2_{r-s-1} and the right hand side converges to χ^2_{r-1}. Thus, if the decision rule is based on the statistic (11.0.3):

$$\delta = \begin{cases} H_1 : & T \leq c \\ H_2 : & T > c \end{cases}$$

then the threshold c can be determined conservatively from the tail of χ^2_{r-1} distribution since

$$\mathbb{P}(\delta \neq H_0 | H_0) = \mathbb{P}(T > c) \leq \chi^2_{r-1}(T > c) = \alpha.$$

□

References:

[1] Chernoff, Herman; Lehmann, E. L. (1954) The use of maximum likelihood estimates in χ^2 tests for goodness of fit. *Ann. Math. Statistics* **25**, pp. 579-586.