Section 10

Chi-squared goodness-of-fit test.

Example. Let us start with a Matlab example. Let us generate a vector X of 100 i.i.d. uniform random variables on $[0, 1]$:

```
X=rand(100,1).
```

Parameters $(100, 1)$ here mean that we generate a 100×1 matrix or uniform random variables. Let us test if the vector X comes from distribution $U[0, 1]$ using χ^2 goodness-of-fit test:

```
[H,P,STATS]=chi2gof(X,'cdf',@(z)unifcdf(z,0,1),'edges',0:0.2:1)
```

The output is

```
H = 0, P = 0.0953,
STATS = chi2stat: 7.9000
        df: 4
        edges: [0 0.2 0.4 0.6 0.8 1]
        O: [17 16 24 29 14]
        E: [20 20 20 20 20]
```

We accept null hypothesis $H_0 : \mathbb{P} = U[0, 1]$ at the default level of significance $\alpha = 0.05$ since the p-value 0.0953 is greater than α. The meaning of other parameters will become clear when we explain how this test works. Parameter 'cdf' takes the handle @ to a fully specified c.d.f. For example, to test if the data comes from $N(3, 5)$ we would use '@(z)normcdf(z,3,5)', or to test Poisson distribution $\Pi(4)$ we would use '@(z)poisscdf(z,4).'

It is important to note that when we use chi-squared test to test, for example, the null hypothesis $H_0 : \mathbb{P} = N(1, 2)$, the alternative hypothesis is $H_0 : \mathbb{P} \neq N(1, 2)$. This is different from the setting of t-tests where we would assume that the data comes from normal distribution and test $H_0 : \mu = 1$ vs. $H_0 : \mu \neq 1$.

□

Pearson's theorem.

Chi-squared goodness-of-fit test is based on a probabilistic result that we will prove in this section.

Figure 10.1:

Let us consider r boxes $B_1, \ldots, B_r$ and throw n balls $X_1, \ldots, X_n$ into these boxes independently of each other with probabilities

$$\mathbb{P}(X_i \in B_1) = p_1, \ldots, \mathbb{P}(X_i \in B_r) = p_r,$$

so that

$$p_1 + \ldots + p_r = 1.$$

Let ν_j be a number of balls in the jth box:

$$\nu_j = \#\{\text{balls } X_1, \ldots, X_n \text{ in the box } B_j\} = \sum_{l=1}^{n} I(X_l \in B_j).$$

On average, the number of balls in the jth box will be np_j since

$$\mathbb{E}\nu_j = \sum_{l=1}^{n} \mathbb{E}I(X_l \in B_j) = \sum_{l=1}^{n} \mathbb{P}(X_l \in B_j) = np_j.$$

We can expect that a random variable ν_j should be close to np_j. For example, we can use a Central Limit Theorem to describe precisely how close ν_j is to np_j. The next result tells us how we can describe the closeness of ν_j to np_j simultaneously for all boxes $j \leq r$. The main difficulty in this Thorem comes from the fact that random variables ν_j for $j \leq r$ are not independent because the total number of balls is fixed

$$\nu_1 + \ldots + \nu_r = n.$$

If we know the counts in $n - 1$ boxes we automatically know the count in the last box.

Theorem.(*Pearson*) *We have that the random variable*

$$\sum_{j=1}^{r} \frac{(\nu_j - np_j)^2}{np_j} \to^d \chi^2_{r-1}$$

converges in distribution to χ^2_{r-1}*-distribution with* $(r-1)$ *degrees of freedom.*

Proof. Let us fix a box B_j. The random variables

$$I(X_1 \in B_j), \ldots, I(X_n \in B_j)$$

that indicate whether each observation X_i is in the box B_j or not are i.i.d. with Bernoulli distribution $B(p_j)$ with probability of success

$$\mathbb{E}I(X_1 \in B_j) = \mathbb{P}(X_1 \in B_j) = p_j$$

and variance

$$\mathrm{Var}(I(X_1 \in B_j)) = p_j(1 - p_j).$$

Therefore, by Central Limit Theorem the random variable

$$\begin{aligned} \frac{\nu_j - np_j}{\sqrt{np_j(1-p_j)}} &= \frac{\sum_{l=1}^n I(X_l \in B_j) - np_j}{\sqrt{np_j(1-p_j)}} \\ &= \frac{\sum_{l=1}^n I(X_l \in B_j) - n\mathbb{E}}{\sqrt{n\mathrm{Var}}} \to^d N(0,1) \end{aligned}$$

converges in distribution to $N(0,1)$. Therefore, the random variable

$$\frac{\nu_j - np_j}{\sqrt{np_j}} \to^d \sqrt{1-p_j}N(0,1) = N(0, 1-p_j)$$

converges to normal distribution with variance $1 - p_j$. Let us be a little informal and simply say that

$$\frac{\nu_j - np_j}{\sqrt{np_j}} \to Z_j$$

where random variable $Z_j \sim N(0, 1-p_j)$.

We know that each Z_j has distribution $N(0, 1-p_j)$ but, unfortunately, this does not tell us what the distribution of the sum $\sum Z_j^2$ will be, because as we mentioned above r.v.s ν_j are not independent and their correlation structure will play an important role. To compute the covariance between Z_i and Z_j let us first compute the covariance between

$$\frac{\nu_i - np_i}{\sqrt{np_i}} \text{ and } \frac{\nu_j - np_j}{\sqrt{np_j}}$$

which is equal to

$$\begin{aligned} &\mathbb{E}\frac{\nu_i - np_j}{\sqrt{np_i}}\frac{\nu_j - np_j}{\sqrt{np_j}} = \frac{1}{n\sqrt{p_i p_j}}(\mathbb{E}\nu_i\nu_j - \mathbb{E}\nu_i np_j - \mathbb{E}\nu_j np_i + n^2 p_i p_j) \\ &= \frac{1}{n\sqrt{p_i p_j}}(\mathbb{E}\nu_i\nu_j - np_i np_j - np_j np_i + n^2 p_i p_j) = \frac{1}{n\sqrt{p_i p_j}}(\mathbb{E}\nu_i\nu_j - n^2 p_i p_j). \end{aligned}$$

To compute $\mathbb{E}\nu_i\nu_j$ we will use the fact that one ball cannot be inside two different boxes simultaneously which means that

$$I(X_l \in B_i)I(X_l \in B_j) = 0. \tag{10.0.1}$$

Therefore,

$$\begin{aligned}
\mathbb{E}\nu_i\nu_j &= \mathbb{E}\Big(\sum_{l=1}^{n} I(X_l\in B_i)\Big)\Big(\sum_{l'=1}^{n} I(X_{l'}\in B_j)\Big) = \mathbb{E}\sum_{l,l'} I(X_l\in B_i)I(X_{l'}\in B_j)\\
&= \underbrace{\mathbb{E}\sum_{l=l'} I(X_l\in B_i)I(X_{l'}\in B_j)}_{\text{this equals to 0 by (10.0.1)}} + \mathbb{E}\sum_{l\neq l'} I(X_l\in B_i)I(X_{l'}\in B_j)\\
&= n(n-1)\mathbb{E}I(X_l\in B_j)\mathbb{E}I(X_{l'}\in B_j) = n(n-1)p_ip_j.
\end{aligned}$$

Therefore, the covariance above is equal to

$$\frac{1}{n\sqrt{p_ip_j}}\Big(n(n-1)p_ip_j - n^2p_ip_j\Big) = -\sqrt{p_ip_j}.$$

To summarize, we showed that the random variable

$$\sum_{j=1}^{r}\frac{(\nu_j-np_j)^2}{np_j}\to\sum_{j=1}^{r}Z_j^2.$$

where normal random variables $Z_1,\ldots,Z_n$ satisfy

$$\mathbb{E}Z_i^2 = 1-p_i \text{ and covariance } \mathbb{E}Z_iZ_j = -\sqrt{p_ip_j}.$$

To prove the Theorem it remains to show that this covariance structure of the sequence of (Z_i) implies that their sum of squares has χ^2_{r-1}-distribution. To show this we will find a different representation for $\sum Z_i^2$.

Let $g_1,\ldots,g_r$ be i.i.d. standard normal random variables. Consider two vectors

$$\boldsymbol{g} = (g_1,\ldots,g_r)^T \text{ and } \boldsymbol{p} = (\sqrt{p_1},\ldots,\sqrt{p_r})^T$$

and consider a vector $\boldsymbol{g}-(\boldsymbol{g}\cdot\boldsymbol{p})\boldsymbol{p}$, where $\boldsymbol{g}\cdot\boldsymbol{p} = g_1\sqrt{p_1}+\ldots+g_r\sqrt{p_r}$ is a scalar product of $\boldsymbol{g}$ and $\boldsymbol{p}$. We will first prove that

$$\boldsymbol{g}-(\boldsymbol{g}\cdot\boldsymbol{p})\boldsymbol{p} \text{ has the same joint distribution as } (Z_1,\ldots,Z_r). \tag{10.0.2}$$

To show this let us consider two coordinates of the vector $\boldsymbol{g}-(\boldsymbol{g}\cdot\boldsymbol{p})\boldsymbol{p}$:

$$i^{th}: g_i - \sum_{l=1}^{r} g_l\sqrt{p_l}\sqrt{p_i} \quad \text{and } j^{th}: g_j - \sum_{l=1}^{r} g_l\sqrt{p_l}\sqrt{p_j}$$

and compute their covariance:

$$\begin{aligned}
&\mathbb{E}\Big(g_i-\sum_{l=1}^{r} g_l\sqrt{p_l}\sqrt{p_i}\Big)\Big(g_j-\sum_{l=1}^{r} g_l\sqrt{p_l}\sqrt{p_j}\Big)\\
&= -\sqrt{p_i}\sqrt{p_j} - \sqrt{p_j}\sqrt{p_i} + \sum_{l=1}^{n} p_l\sqrt{p_i}\sqrt{p_j} = -2\sqrt{p_ip_j}+\sqrt{p_ip_j} = -\sqrt{p_ip_j}.
\end{aligned}$$

Similarly, it is easy to compute that

$$\mathbb{E}\Big(g_i - \sum_{l=1}^{r} g_l \sqrt{p_l}\sqrt{p_i}\Big)^2 = 1 - p_i.$$

This proves (10.0.2), which provides us with another way to formulate the convergence, namely, we have

$$\sum_{j=1}^{r}\Big(\frac{\nu_j - np_j}{\sqrt{np_j}}\Big)^2 \to^d |\boldsymbol{g} - (\boldsymbol{g}\cdot\boldsymbol{p})\boldsymbol{p}|^2.$$

But this vector has a simple geometric interpretation. Since vector $\boldsymbol{p}$ is a unit vector:

$$|\boldsymbol{p}|^2 = \sum_{l=1}^{r}(\sqrt{p_i})^2 = \sum_{l=1}^{r} p_i = 1,$$

vector $\boldsymbol{V_1} = (\boldsymbol{p}\cdot\boldsymbol{g})\boldsymbol{p}$ is the projection of vector $\boldsymbol{g}$ on the line along $\boldsymbol{p}$ and, therefore, vector $\boldsymbol{V_2} = \boldsymbol{g} - (\boldsymbol{p}\cdot\boldsymbol{g})\boldsymbol{p}$ will be the projection of $\boldsymbol{g}$ onto the plane orthogonal to $\boldsymbol{p}$, as shown in figure 10.2.

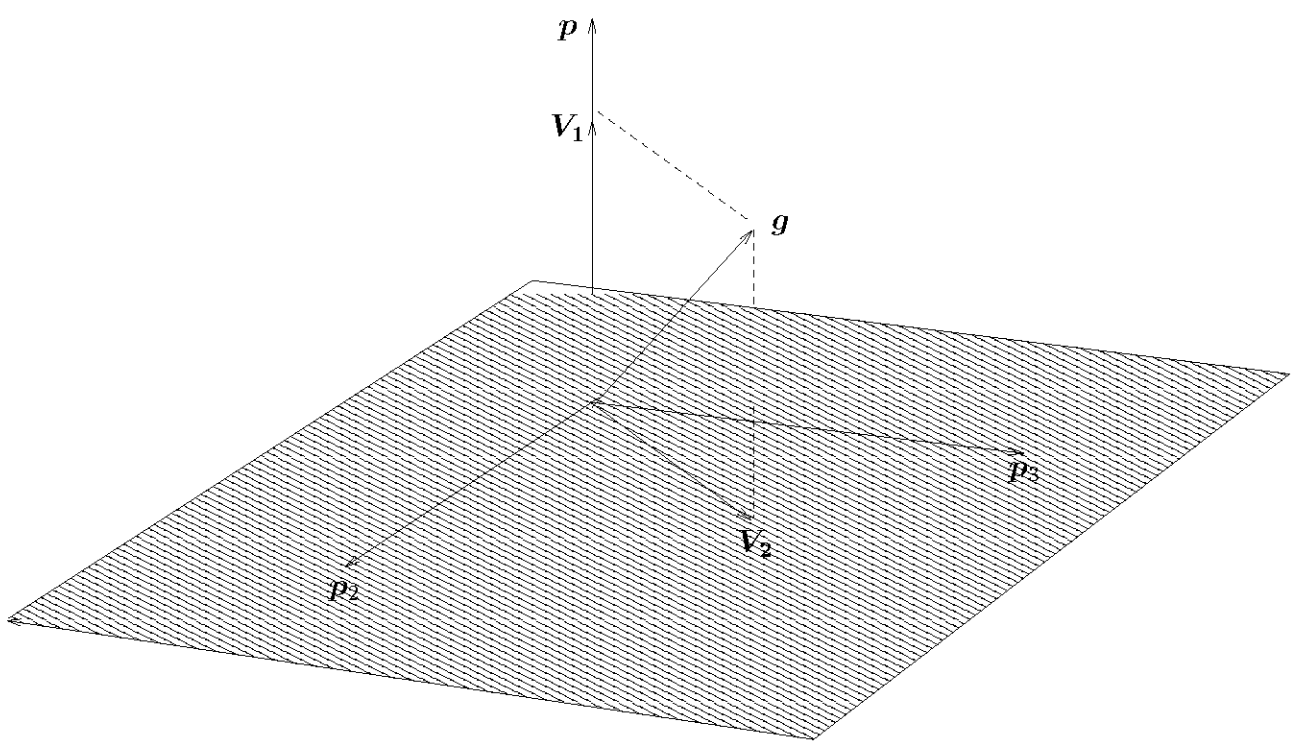

Figure 10.2: New coordinate system.

Let us consider a new orthonormal coordinate system with the first basis vector (first axis) equal to $\boldsymbol{p}$. In this new coordinate system vector $\boldsymbol{g}$ will have coordinates

$$\boldsymbol{g}' = (g_1', \ldots, g_r') = V\boldsymbol{g}$$

obtained from $\boldsymbol{g}$ by orthogonal transformation

$$V = (\boldsymbol{p}, \boldsymbol{p}_2, \ldots, \boldsymbol{p}_r)$$

that maps canonical basis into this new basis. But we proved in Lecure 4 that in that case $g'_1, \ldots, g'_r$ will also be i.i.d. standard normal. From figure 10.2 it is obvious that vector $\boldsymbol{V_2} = \boldsymbol{g} - (\boldsymbol{p}\cdot\boldsymbol{g})\boldsymbol{p}$ in the new coordinate system has coordinates

$$(0, g'_2, \ldots, g'_r)^T$$

and, therefore,

$$|\boldsymbol{V_2}|^2 = |\boldsymbol{g} - (\boldsymbol{p}\cdot\boldsymbol{g})\boldsymbol{p}|^2 = (g'_2)^2 + \ldots + (g'_r)^2.$$

But this last sum, by definition, has χ^2_{r-1} distribution since $g'_2, \cdots, g'_r$ are i.i.d. standard normal. This finishes the proof of Theorem.

□

Chi-squared goodness-of-fit test for simple hypothesis.

Suppose that we observe an i.i.d. sample $X_1, \ldots, X_n$ of random variables that take a finite number of values $B_1, \ldots, B_r$ with unknown probabilities $p_1, \ldots, p_r$. Consider hypotheses

$$\begin{aligned} H_0 &: \quad p_i = p_i^\circ \text{ for all } i = 1, \ldots, r, \\ H_1 &: \quad \text{for some } i, p_i \neq p_i^\circ. \end{aligned}$$

If the null hypothesis H_0 is true then by Pearson's theorem

$$T = \sum_{i=1}^{r} \frac{(\nu_i - np_i^\circ)^2}{np_i^\circ} \to^d \chi^2_{r-1}$$

where $\nu_i = \#\{X_j : X_j = B_i\}$ are the observed counts in each category. On the other hand, if H_1 holds then for some index i, $p_i \neq p_i^\circ$ and the statistics T will behave differently. If p_i is the true probability $\mathbb{P}(X_1 = B_i)$ then by CLT

$$\frac{\nu_i - np_i}{\sqrt{np_i}} \to^d N(0, 1 - p_i).$$

If we rewrite

$$\frac{\nu_i - np_i^\circ}{\sqrt{np_i^\circ}} = \frac{\nu_i - np_i + n(p_i - p_i^\circ)}{\sqrt{np_i^\circ}} = \sqrt{\frac{p_i}{p_i^\circ}} \frac{\nu_i - np_i}{\sqrt{np_i}} + \sqrt{n}\frac{p_i - p_i^\circ}{\sqrt{p_i^\circ}}$$

then the first term converges to $N(0, (1 - p_i)p_i/p_i^\circ)$ and the second term diverges to plus or minus ∞ because $p_i \neq p_i^\circ$. Therefore,

$$\frac{(\nu_i - np_i^\circ)^2}{np_i^\circ} \to +\infty$$

which, obviously, implies that $T \to +\infty$. Therefore, as sample size n increases the distribution of T under null hypothesis H_0 will approach χ^2_{r-1}-distribution and under alternative hypothesis H_1 it will shift to $+\infty$, as shown in figure 10.3.

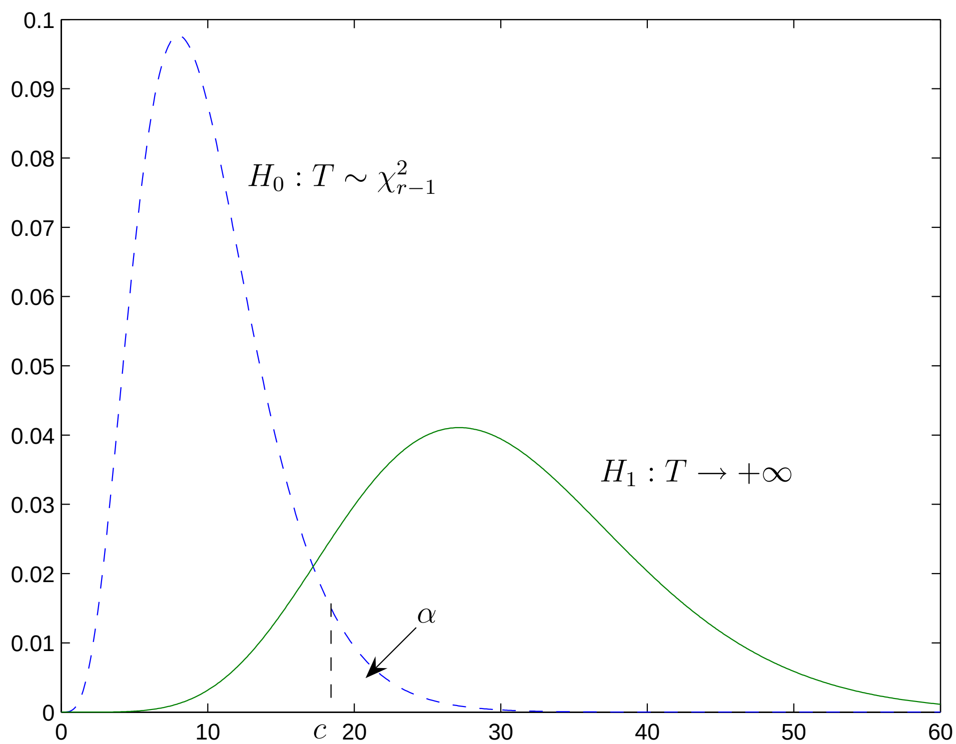

Figure 10.3: Behavior of T under H_0 and H_1.

Therefore, we define the decision rule

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

We choose the threshold c from the condition that the error of type 1 is equal to the level of significance α :

$$\alpha = \mathbb{P}_1(\delta \ne H_1) = \mathbb{P}_1(T > c) \approx \chi^2_{r-1}(c, \infty)$$

since under the null hypothesis the distribution of T is approximated by χ^2_{r-1} distribution. Therefore, we take c such that $\alpha = \chi^2_{r-1}(c, \infty)$. This test δ is called the *chi-squared goodness-of-fit* test.

□

Example. (*Montana outlook poll.*) In a 1992 poll 189 Montana residents were asked (among other things) whether their personal financial status was worse, the same or better than a year ago.

Worse	Same	Better	Total
58	64	67	189

We want to test the hypothesis H_0 that the underlying distribution is uniform, i.e. $p_1 = p_2 = p_3 = 1/3$. Let us take level of significance $\alpha = 0.05$. Then the threshold c in the chi-squared test

$$\delta = \begin{cases} H_0 : & T \le c \\ H_1 : & T > c \end{cases}$$

is found from the condition that $\chi^2_{3-1=2}(c, \infty) = 0.05$ which gives $c = 5.9$. We compute chi-squared statistic

$$T = \frac{(58 - 189/3)^2}{189/3} + \frac{(64 - 189/3)^2}{189/3} + \frac{(67 - 189/3)^2}{189/3} = 0.666 < 5.9$$

which means that we accept H_0 at the level of significance 0.05.

□

Goodness-of-fit for continuous distribution.

Let $X_1, \ldots, X_n$ be an i.i.d. sample from unknown distribution $\mathbb{P}$ and consider the following hypotheses:

$$\begin{cases} H_0 : & \mathbb{P} = \mathbb{P}_0 \\ H_1 : & \mathbb{P} \ne \mathbb{P}_0 \end{cases}$$

for some particular, possibly continuous distribution $\mathbb{P}_0$. To apply the chi-squared test above we will group the values of Xs into a finite number of subsets. To do this, we will split a set of all possible outcomes $\mathcal{X}$ into a finite number of intervals $I_1, \ldots, I_r$ as shown in figure 10.4.

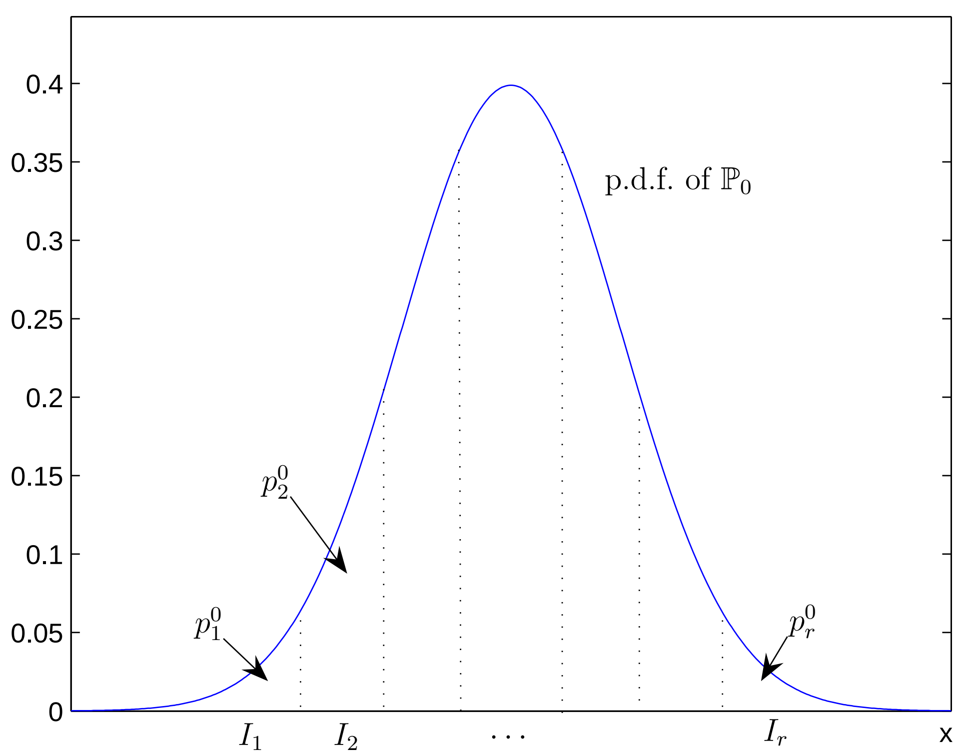

Figure 10.4: Discretizing continuous distribution.

The null hypothesis H_0, of course, implies that for all intervals

$$\mathbb{P}(X \in I_j) = \mathbb{P}_0(X \in I_j) = p_j^0.$$

Therefore, we can do chi-squared test for

$$\begin{aligned} H_0' &: \ \mathbb{P}(X \in I_j) = p_j^0 \text{ for all } j \leq r \\ H_1' &: \ \text{otherwise.} \end{aligned}$$

Asking whether H_0' holds is, of course, a weaker question that asking if H_0 holds, because H_0 implies H_0' but not the other way around. There are many distributions different from $\mathbb{P}$ that have the same probabilities of the intervals $I_1, \ldots, I_r$ as $\mathbb{P}$. On the other hand, if we group into more and more intervals, our discrete approximation of $\mathbb{P}$ will get closer and closer to $\mathbb{P}$, so in some sense H_0' will get 'closer' to H_0. However, we can not split into too many intervals either, because the χ^2_{r-1}-distribution approximation for statistic T in Pearson's theorem is asymptotic. The rule of thumb is to group the data in such a way that the expected count in each interval

$$np_i^0 = n\mathbb{P}_0(X \in I_i) \geq 5$$

is at least 5. (Matlab, for example, will give a warning if this expected number will be less than five in any interval.) One approach could be to split into intervals of equal probabilities $p_i^0 = 1/r$ and choose their number r so that

$$np_i^0 = \frac{n}{r} \geq 5.$$

Example. Let us go back to the example from Lecture 2. Let us generate 100 observations from Beta distribution $B(5, 2)$.

```
X=betarnd(5,2,100,1);
```

Let us fit normal distribution $N(\mu, \sigma^2)$ to this data. The MLE $\hat{\mu}$ and $\hat{\sigma}$ are

```
mean(X) = 0.7421, std(X,1)=0.1392.
```

Note that 'std(X)' in Matlab will produce the square root of unbiased estimator $(n/n-1)\hat{\sigma}^2$. Let us test the hypothesis that the sample has this fitted normal distribution.

```
[H,P,STATS]= chi2gof(X,'cdf',@(z)normcdf(z,0.7421,0.1392))
```

outputs

```
H = 1, P = 0.0041,
STATS = chi2stat: 20.7589
        df: 7
        edges: [1x9 double]
        O: [14 4 11 14 14 16 21 6]
        E: [1x8 double]
```

Our hypothesis was rejected with p-value of 0.0041. Matlab split the real line into 8 intervals of equal probabilities. Notice 'df: 7' - the degrees of freedom $r - 1 = 8 - 1 = 7$.

□