Nonparametric Regression: Goodness of Fit Tests

Paweł Polak

March 1, 2016

STAT W4413: Nonparametric Statistics - Lecture 10

Goodness of Fit Test

- Let $x_1, x_2, \ldots, x_n$ be drawn iid from a certain distribution $F(x)$.
- We would like to test the null hypothesis

$$H_0 : F(x) = F_0(x) \text{ vs. } H_1 : F(x) \neq F_0(x)$$

- E.g., we would like to know if our data is drawn from standard normal distribution or not.
- In this lecture we cover several different approaches, including
  - χ^2-test,
  - Kolmogorov-Smirnov test, and
  - Anderson-Darling test.

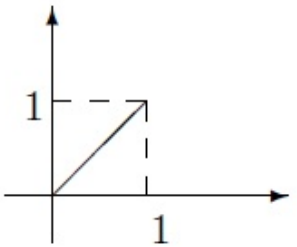
Let us first study some visual inspection tools that provide qualitative information on the correctness of H_0.

Visual Analysis of Goodness of fit

- Visual inspection tools are probably the first type of tool that we employ in applications.
- While they are usually incapable of providing accurate/definitive answers, they provide good first guesses and intuitions.
- Using visual inspection tools in statistics was popularized by Tukey in his "exploratory data analysis" book.
- We will study two visual inspection tools for "goodness of the fit": pp-plot and qq-plot and will discuss their advantages and disadvantages.

pp-plot

Suppose that we know the actual distribution function F. The simplest visual idea is to compare $F(t)$ with $F_0(t)$ on the same graph. Suppose the following graph in which each point is $(F_0(t), F(t))$ for a given $t \in \mathbb{R}$.



- The set of points $(F_0(t), F(t))(t \in \mathbb{R})$ is going to be a curve in the pp-diagram.
- However, if the two distributions are equal, then we see the diagonal line with 45-degree that passes through the origin.
- Any deviation from this straight line is an indication of the fact that our null hypothesis is not correct and shall be rejected.
- Needless to say that in applications $F(t)$ is not known and therefore we cannot plot $F(t)$ versus $F_0(t)$.

pp-plot

To address this issue we replace the distribution function F with its estimate $\hat{F}_n$ given by

$$\hat{F}_n(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(x_i \leq t).$$

- Since $\hat{F}_n(t)$ is an estimate of $F(t)$ and is different from $F(t)$, the graph we observe in pp-plane is not a line.
- But if F is the same as F_0 and the number of samples are large enough then the plot is going to be close to a line.
- In practice we usually accept the null hypothesis if the graph is close enough to the 45-degree line.
- As is clear from this discussion, accepting or rejecting the null hypothesis is based on the personal judgement rather than an accurate test. We will get back to this point later in these slides.

qq-plot

We are again interested in testing the null hypothesis $F(t) = F_0(t)$ for every t.

- We assume for now that $F(t)$ is given but slightly modify the approach from pp-plot.
- Suppose that both F and F_0 are strictly increasing.
- For every $\alpha \in [0, 1]$ we calculate $t_0(\alpha) = F_0^{-1}(\alpha)$ and $t(\alpha) = F^{-1}(\alpha)$ and we plot $t_0(\alpha)$ versus $t(\alpha)$ for $\alpha \in [0, 1]$.
- Clearly, if $F = F_0$ then $t_0(\alpha) = t(\alpha)$ and therefore we must see a 45-degree line again.
- As mentioned before, in applications F is not known and therefore we will use empirical distribution function $\hat{F}_n(x)$ that we discussed before.

qq-plot

- There is only one subtlety in here. Note that usually given α the value of $\hat{F}^{-1}(\alpha)$ is not uniquely specified, since in some regions $\hat{F}_n$ is flat.
- In most cases as long as you choose one of the points that give you the value of α you are interested in, qq-plot works well.
- Most packages have their own conventions of choosing the points that correspond to $\hat{F}^{-1}(\alpha)$.

qq-plot vs. pp-plot?

Let's start with an example.

- We have a data $x_1, x_2, \ldots, x_n$ that we believe is Gaussian.
- But we don't know the mean and the variance. Can visual inspection tools help us check such composite null hypotheses as well?
- It turns out that qq-plot can potentially address such problems more efficiently.
- Since we do not know the exact value of μ and σ, we set μ to zero and $\sigma^2 = 1$ to obtain $F_0 = N(0, 1)$. (Even though we know that the mean is not necessarily zero and the variance is not necessarily equal to one). In the qq-plot we should determine: $(F_0^{-1}(\alpha), F^{-1}(\alpha))$.

qq-plot vs. pp-plot?

- Define

$$\begin{aligned} t_0(\alpha) &\triangleq F_0^{-1}(\alpha), \\ t(\alpha) &\triangleq F^{-1}(\alpha). \end{aligned} \tag{1}$$

- We then have

$$\begin{aligned} &\mathbb{P}\{N(\mu,\sigma^2) < t(\alpha)\} = \alpha \Rightarrow \mathbb{P}\{\sigma(N(0,1)) + \mu < t(\alpha)\} = \alpha \\ &\Rightarrow \mathbb{P}\{N(0,1) < \frac{t(\alpha)-\mu}{\sigma}\} = \alpha \Rightarrow \frac{t(\alpha)-\mu}{\sigma} = F_0^{-1}(\alpha) = t_0(\alpha) \\ &\Rightarrow t(\alpha) = \sigma t_0(\alpha) + \mu \end{aligned} \tag{2}$$

In other words, if our data is $\mathcal{N}(\mu,\sigma)$ then we will still see a line in the qq-plot plane. Slope and intercept of the line provide information on the mean and variance of the distribution. It is pretty straightforward to extend this result to the following theorem. This is left as a homework for you.

qq-plot vs. pp-plot?

Theorem

Consider two random variables X, Y that satisfy $Y = \alpha X + \beta$. The qq-plot of the actual distribution of Y versus the actual distribution of X is a line.

- We don't need to worry about the shift and scale of the random variables in qq-plot.
- This is definitely not the case in pp-plot.
- Convince yourself by an example of $X \sim Unif(0, 1)$ and $Y = 2X$.

qq-plot vs. pp-plot?

- Note that in case you would like to use pp-plot for such composite hypotheses, you should first estimate the unknown parameters (by a method such as maximum likelihood), then convert the composite null hypothesis to a simple one and then use pp-plot for this new distribution.

- For instance, in the Gaussian case, we first estimate $\hat{\mu}, \hat{\sigma}$ and then we change the null hypothesis to $N(\hat{\mu}, \hat{\sigma}^2)$.

- Therefore we change the composite null hypothesis to a simple null hypothesis given by $H_0 : F = N(\hat{\mu}, \hat{\sigma}^2)$.

Flaws of visual inspection tools

Remark: The visual inspection tools are very useful in providing some intuition on the behavior of the data. However, the decision we make based on these tools are to a great extent subjective. Such approaches are not acceptable in many applications. Therefore we would like to have some more objective tools.

This will be discussed in the next slides. We will cover three different tests that are the most popular for this purpose.

Pearson's χ^2 goodness of the fit test

First we study a very special form of random variables known as *categorical* or *finite-valued* random variables.

Definition

A random variable $X : \Omega \to \mathbb{R}$ (Ω denotes the sample space) is called categorical if and only if the range of X is a finite set. In other words, the random variable can be written as $X : \Omega \to \{\alpha_0, \ldots, \alpha_k\}$.

Well-known examples of such random variables are customer ratings, when people rate products and companies. For instance, the ratings of a movie in Netflix is a number in $\{1, 2, 3, 4, 5\}$. The ratings in Amazon or Yelp have the same flavor.

Pearson's χ^2 goodness of the fit test

In case of categorical random variables it is usually more convenient to work with *probability mass functions*.

For goodness of fit tests of categorical random variables we can still work with probability mass functions.

Suppose that we believe

$$\mathbb{P}\{X = \alpha_0\} = \pi_0, \mathbb{P}\{X = \alpha_1\} = \pi_1, \ldots, \mathbb{P}\{X = \alpha_k\} = \pi_k,$$

and we would like to test the hypothesis

$$H_0 = \begin{cases} P(X = \alpha_0) = \pi_0 \\ \vdots \\ P(X = \alpha_k) = \pi_k \end{cases} \quad \text{vs} \quad H_1 : \begin{cases} \text{at least one of the probabilities do not match} \\ i.e.\ \exists i : P(X = \alpha_i) \neq \pi_i \end{cases}$$

Pearson's χ^2 goodness of the fit test

Toward this goal we first estimate

$$p_i \triangleq \mathbb{P}(X = \alpha_i).$$

Note that

$$p_i = \mathbb{E}[\mathbb{I}(X = \alpha_i)].$$

Why?

Replacing $\mathbb{E}$ with the empirical average we obtain the following simple estimate for p_i:

$$\hat{p}_i = \frac{1}{n} \sum_{j=1}^{n} \mathbb{I}(X_j = \alpha_i).$$

Pearson's χ^2 goodness of the fit test

Based on these $\hat{p}_i$ we can construct different goodness of the fit tests.

We should only define a distance between the empirical estimates and the probability mass function under the null hypothesis.

One such a test that has optimality properties is the *Pearson's* χ^2 test.

Definition

The Pearson's χ^2 statistic for the goodness of the fit test is defined as

$$Q_n \triangleq \sum_{i=0}^{k} \frac{(\hat{p}_i - \pi_i)^2}{\pi_i}$$

Pearson's χ^2 goodness of the fit test

Note that the subscript n represents the number of samples that are used in estimating $\hat{p}_i$. Based on Pearson's

$$\text{Reject } H_0 \text{ if } Q_n > \kappa; \text{ otherwise accept } H_0.$$

Remark It is clear that if we define another statistic

$$R = \sum_{i=1}^{k} (\hat{p}_i - \pi_i)^2,$$

then this may lead to another goodness of the fit test.

But note that such statistic usually ignores mismatch of the events with low probability. In order to give more weight to the low probability events we have divided each difference $(\hat{p}_i - \pi_i)^2$ by its probability under null π_i.

As we will see later, another important reason for doing so is the nice limiting distribution of Q_n.

Asymptotic analysis of χ^2 test

- Our next step is to calculate the probability of the Type I error and calculate κ in terms of the significance level.
- Unfortunately the exact characterization of the distribution of Q_n is complicated.
- Therefore, we characterize it in the asymptotic setting $n \to \infty$.
- This asymptotic argument works well if the sample size is larger than 30 (and k is 5 or 6).
- But in many cases it works well even if the sample size is smaller. We will see a few examples in the homeworks.

Theorem

Under the null hypothesis H_0, $nQ_n \xrightarrow{d} \chi^2(k)$.

Heuristic argument

Note that the discussion we provide here is hand-waving, misses lots of details, and should not be considered as a proof.[1]

We have

$$\hat{p}_i - \pi_i = \frac{1}{n}\sum_{j=1}^{n} \mathbb{I}(x_j = \alpha_i) - \pi_i = \frac{1}{n}\sum_{j=1}^{n} (\mathbb{I}(x_j = \alpha_i) - \mathbb{E}[\mathbb{I}(x_j = \alpha_i)]).$$

So we expect that $\sqrt{n}(\hat{p}_i - \pi_i)$ converges to a Gaussian in distribution. Clearly, this holds for every i.

By CLT

[1] We will prove the result very carefully later in the course

Heuristic argument

- Since Q_n is sum of squares of Gaussian random variables, we expect Q to converge to a χ^2 distribution[2].
- If we accept the conjecture that the distribution is χ^2, then we should characterize the degrees of freedom of the χ^2.
- Our first guess would be $k+1$ since we have added squares of $k+1$ Gaussian random variables.
- However, if we inspect the problem more carefully, we realize that the random variables are dependent. In fact, one can easily prove that

$$\sum_{i=0}^{k}(\hat{p}_i - \pi_i) = 0.$$

- Clearly, if the random variables were independent the degrees of freedom of χ^2 would have been $k+1$.
- But the above constraint reduces the degrees of freedom by one.
- So, we guess the asymptotic distribution of nQn to be χ^2 with k degrees of freedom.
- Note that our argument is very hand-waving:
  - It does not keep into account the differences between the variances of the terms $\hat{p}_i - \pi_i$.
  - Also, it assumes that a linear constraint reduces the degrees of freedom of the χ^2 by 1.
- However, such argument can usually provide you a good first guess for the limiting distribution of different statistics. The next step is of course to prove such a guess.

[2] Note that this is the first place we have been sloppy. There are several reasons for that. First, the random variables are not independent. Second, we don't know if they are jointly Gaussian as well.

Proof of Theorem

- The heuristics on the last slide,provide CLT type argument for individual $\hat{p}_i - \pi_i$.
- However, to obtain the distribution of Q_n which is a function of $\hat{p}_0 - p_0, \hat{p}_1 - p_1, \ldots, \hat{p}_k - p_k$, we need to *characterize the joint distribution* of

$$\hat{p}_0 - p_0, \hat{p}_1 - p_1, \ldots, \hat{p}_k - p_k.$$

- In fact, it is important to capture the dependencies among random variables as well.
- Toward this goal, we should put all $\hat{p}_i - \pi_i$ in a vector and use the CLT for random vectors.
- Therefore, our first guess would be to form the vector $[\hat{p}_0 - \pi_0, \ldots, \hat{p}_k - \pi_k]$ and use CLT for this vector.
- However, there is a problem in doing so. As we discussed before, we have

$$(\hat{p}_0 - \pi_0) + (\hat{p}_1 - \pi_1) + \ldots + (\hat{p}_k - \pi_k) = 0.$$

- This means that the covariance matrix of the joint distribution of $[\hat{p}_0 - \pi_0, \ldots, \hat{p}_k - \pi_k]$ is going to be singular (not invertible). You will prove this in the next Homework.
- This implies that we will have some issues in applying the central limit theorem for which we required $\sum$ to be positive definite (so that we can define the final Gaussian distribution).
- To resolve this issue we just drop one of the elements of the vector, e.g. $\hat{p}_0 - \pi_0$, and work with

$$\mathbf{p}^n = [\hat{p}_1 - \pi_1, \hat{p}_2 - \pi_2, \ldots, \hat{p}_k - \pi_k]^T. \tag{3}$$

Note that the superscript n in p_n denotes the number of samples that have been used for estimating the empirical probabilities. Note that $\sqrt{n}\mathbf{p}^n$ converges to a multivariate normal in distribution. To see why, first define the following vector,

$$\mathbf{b}^{n,j} = \begin{bmatrix} \mathbb{I}(x_j = \alpha_1) - \pi_1 \\ \mathbb{I}(x_j = \alpha_2) - \pi_2 \\ \vdots \\ \mathbb{I}(x_j = \alpha_k) - \pi_k \end{bmatrix}. \tag{4}$$

It is straightforward to confirm that

$$\mathbf{p}^n = \frac{1}{n}\sum_{j=1}^{n} \mathbf{b}^{n,j}. \qquad \sim\mathbf{MVN}$$

Since $\mathbf{b}^{n,j}$'s are independent for different values of j, we can see that $\sqrt{n}\mathbf{p}^n$ converges to a Gaussian in distribution (by multivariate CLT). The mean and covariance matrix of the limiting distribution can be characterized by

$$\begin{aligned} \mu &= \mathbb{E}(\mathbf{b}^{n,j}), \\ \Sigma &= \mathbb{E}(\mathbf{b}^{n,j}(\mathbf{b}^{n,j})^T) \end{aligned} \tag{5}$$

Proof of Theorem

Note that under the null hypothesis we have $\mathbb{E}[\mathbb{I}(x_j = \alpha_i)] = \pi_i$. Therefore $\mu = 0$. To characterize the covariance matrix we consider the following two cases:

Case I- $\ell \neq k$:

Bi-linearity and Independence

$$\begin{aligned} \Sigma_{\ell,k} &= cov(\mathbb{I}(x_1 = \alpha_\ell) - \pi_\ell, \mathbb{I}(x_1 = \alpha_k) - \pi_k) \\ &= \mathbb{E}((\mathbb{I}(x_1 = \alpha_\ell)\mathbb{I}(x_1 = \alpha_k)) - \pi_\ell \pi_k \\ &= -\pi_\ell \pi_k. \end{aligned} \tag{6}$$

Case II- $\ell = k$:

$$\begin{aligned} \Sigma_{\ell\ell} &= \mathbb{E}(\mathbb{I}(x_1 = \alpha_\ell)\mathbb{I}(x_1 = \alpha_k)) - \pi_\ell^2 \\ &= \mathbb{E}(\mathbb{I}(x_1 = \alpha_\ell)) - \pi_\ell^2 = \pi_\ell - \pi_\ell^2. \end{aligned} \tag{7}$$

Proof of Theorem

Combining (6) and (7) we obtain:

$$\Sigma = \begin{bmatrix} \pi_1(1-\pi_1) & -\pi_1\pi_2 & \dots & -\pi_1\pi_k \\ -\pi_1\pi_2 & \pi_2(1-\pi_2) & \dots & -\pi_2\pi_k \\ \vdots & \vdots & \dots & \vdots \\ -\pi_k\pi_1 & -\pi_k\pi_2 & \dots & \pi_k(1-\pi_k) \end{bmatrix}$$

You will prove in the next Homework that this matrix is invertible and its inverse is given by

$$\Sigma^{-1} = \begin{bmatrix} \frac{1}{\pi_0} + \frac{1}{\pi_1} & \frac{1}{\pi_0} & \dots & \frac{1}{\pi_0} \\ \frac{1}{\pi_0} & \frac{1}{\pi_0} + \frac{1}{\pi_2} & \dots & \frac{1}{\pi_0} \\ \vdots & \vdots & \dots & \vdots \\ \frac{1}{\pi_0} & \frac{1}{\pi_0} & \dots & \frac{1}{\pi_0} + \frac{1}{\pi_k} \end{bmatrix}.$$

So far we have proved that $\sqrt{n}\mathbf{p}^n \xrightarrow{d} N(0, \Sigma)$.

Proof of Theorem

We now define a new random vector that is more closely related to the χ^2 statistic Q_n. Define

$$\mathbf{Y}^n \triangleq \begin{bmatrix} Y_1 \\ \vdots \\ Y_k \end{bmatrix} \triangleq \Sigma^{-\frac{1}{2}} \sqrt{n} \begin{bmatrix} \hat{p}_1 - \pi_1 \\ \vdots \\ \hat{p}_k - \pi_k \end{bmatrix}.$$

It is straightforward to see that

$$\mathbf{Y}^n \xrightarrow{d} N(0, I),$$

where I is the identity matrix. By employing the continuous mapping theorem we obtain

$$\sum_{i=1}^{k} (Y_i)^2 \xrightarrow{D} \chi^2_k.$$

The last step of the proof is to show that the χ^2 statistic nQ_n is the same as $\sum_{i=1}^{k} Y_i^2$.

Proof of Theorem

We show this in the following way:

$$\begin{aligned}
\sum_{i=1}^{k} Y_i^2 &= \begin{bmatrix} Y_1 \\ \vdots \\ Y_k \end{bmatrix}^T \begin{bmatrix} Y_1 \\ \vdots \\ Y_k \end{bmatrix} = n \begin{bmatrix} \hat{p}_1 - \pi_1 \\ \vdots \\ \hat{p}_k - \pi_k \end{bmatrix}^T \Sigma^{-1} \begin{bmatrix} \hat{p}_1 - \pi_1 \\ \vdots \\ \hat{p}_k - \pi_k \end{bmatrix} \\
&= n \begin{bmatrix} \hat{p}_1 - \pi_1 \\ \vdots \\ \hat{p}_k - \pi_k \end{bmatrix}^T \begin{bmatrix} \vdots \\ \sum_{j=1}^{k} \Sigma_{ij}^{-1} (\hat{p}_j - \pi_j) \\ \vdots \end{bmatrix} \\
&\overset{(a)}{=} n \begin{bmatrix} \hat{p}_1 - \pi_1 \\ \vdots \\ \hat{p}_k - \pi_k \end{bmatrix}^T \dots
\end{aligned}$$

To obtain Equality (a) we have used (8).

Proof of Theorem

$$= \begin{bmatrix} \frac{1}{\pi_1}(\hat{p}_1 - \pi_1) + \frac{1}{\pi_0}\sum_{j=1}^{k}(\hat{p}_j - \pi_j) \\ \vdots \\ \frac{1}{\pi_i}(\hat{p}_i - \pi_i) + \frac{1}{\pi_0}\sum_{j=1}^{k}(\hat{p}_j - \pi_j) \end{bmatrix} \tag{8}$$

-(p0-π0)=sum((i=1:k),(pj-πj))

$$\overset{(b)}{=} n \begin{bmatrix} \hat{p}_1 - \pi_1 \\ \vdots \\ \hat{p}_k - \pi_k \end{bmatrix}^T \begin{bmatrix} \frac{1}{\pi_1}(\hat{p}_1 - \pi_1) - \frac{\hat{p}_0 - \pi_0}{\pi_0} \\ \vdots \\ \frac{1}{\pi_i}(\hat{p}_i - \pi_i) - \frac{\hat{p}_0 - \pi_0}{\pi_0} \\ \vdots \\ \frac{1}{\pi_k}(\hat{p}_k - \pi_k) - \frac{\hat{p}_0 - \pi_0}{\pi_0} \end{bmatrix}$$

$$= n\sum_{j=1}^{k}\frac{(\hat{p}_j - \pi_j)^2}{\pi_j} - \frac{(\hat{p}_0 - \pi_0)}{\pi_0}\sum_{j=1}^{k}(\hat{p}_j - \pi_j) \overset{(c)}{=} nQ_n$$

Equalities (b) and (c) are due to the fact that $\sum_{j=0}^{k}(\hat{p}_j - \pi_j) = 0$.
Therefore

$$nQ_n \xrightarrow{d} \chi^2(k).$$

χ^2 test for composite null hypothesis

So far, we have considered "simple null hypothesis" where under H_0, the exact probabilities are known.

However, in many applications this is not the case.

Consider the following Netflix example. In Netflix people rate movies with one of the numbers in $\{1, 2, 3, 4, 5\}$. 5 means that they like a movie very much. and "1" means that they do not like it at all.

Usually if people like a movie they give 4 or 5 to it, otherwise they will give 1,2, or 3.

Suppose that we observe some ratings of the movie "Crash" and we represent these numbers with $X_1, X_2, \ldots, X_n$. Our null hypothesis is that 90 percent of the people like this movie.

More formally we would like to test:

$$H_0 : \begin{cases} P(X = 4) + P(X = 5) = 0.9, \\ P(X = 1) + P(X = 2) + P(X = 3) = 0.1. \end{cases}$$

As you can see there is more than one probability distribution in the null. For instance,

- $(P(X = 1), P(X = 2), P(X = 3), P(X = 4), P(X = 5)) = (0.02, 0.02, 0.06, 0.45, 0.45)$ is in the null and
- so are the distribution $(P(X = 1), P(X = 2), P(X = 3), P(X = 4), P(X = 5)) = (0.03, 0.03, 0.04, 0.4, 0.5)$ and many others.

χ^2 test for composite null hypothesis

The situation in which there is more than one probability distribution in the null is called the *composite null hypothesis*.

We first formally state the problem of testing "goodness of fit" under the composite null hypothesis for categorical random variables and then describe how the χ^2 test should be modified to address such situations.

We observe n iid samples $X_1, X_2, \ldots, X_n$ of categorical random variable $X_i : \Omega \to \{\alpha_0, \alpha_1, \ldots, \alpha_k\}$. We would like to test

$$H_0 : \begin{cases} P(X = \alpha_0) = g_0(\theta_1, \theta_2, \ldots, \theta_\ell) \\ \vdots \\ P(X = \alpha_k) = g_k(\theta_1, \theta_2, \ldots, \theta_\ell) \end{cases} \quad \text{versus } H_1 : \text{at least one of the equalities in } H_0 \text{ is violated.} \tag{9}$$

As $\theta_1, \theta_2, \ldots, \theta_\ell$ take different values, the null distribution takes on different forms.

χ^2 test for composite null hypothesis

Remark Note that many of the composite null problems are not immediately in the above standard form. However, they can be converted to the form described in (9). For instance consider the Netflix example described above. Define, $\theta_1 = P(X = 1), \theta_2 = P(X = 2), \theta_3 = P(X = 4)$. Then we can convert the null hypothesis to

$$H_0 : \begin{cases} P(X = 1) = \theta_1, \\ P(X = 2) = \theta_2, \\ P(X = 3) = 0.1 - \theta_1 - \theta_2, \\ P(X = 4) = \theta_3, \\ P(X = 5) = 0.9 - \theta_3, \end{cases}$$

which has the form described in (9).

χ^2 test

Let's define the χ^2 test the same way we did before and see if it works. Define,

$$Q(\theta_1, \theta_2, \ldots, \theta_\ell) = \sum_{j=0}^{k} \frac{(\hat{p}_j - g_j(\theta_1, \ldots, \theta_\ell))^2}{g_j(\theta_1, \ldots, \theta_\ell)}.$$

The first issue is that Q is no more a number. It is now a function of the parameters $\theta_1, \theta_2, \ldots, \theta_\ell$. Therefore, the question is how we can obtain a statistic from this $Q(\theta_1, \theta_2, \ldots, \theta_\ell)$ that can help us evaluate the null hypothesis. Since we know that under null, for at least one value of $\theta_1, \theta_2, \ldots, \theta_\ell$, $g_j(\theta_1, \theta_2, \ldots, \theta_\ell) = p_j$ for every j, we can consider

$$Q^{M\chi} = \min_{\theta_1, \ldots, \theta_\ell} Q(\theta_1, \theta_2, \ldots, \theta_\ell),$$

as our statistic.

The other approach would be to first estimate $\theta_1, \theta_2, \ldots, \theta_\ell$ by maximum likelihood principle, and then calculate

$$Q^{ML} = Q(\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_\ell),$$

where $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_\ell$ are the maximum likelihood estimates of $\theta_1, \theta_2, \ldots, \theta_\ell$. It turns out that in the asymptotic settings these two methods are equivalent and lead to the same statistics. The explanation of this fact is beyond the scope of this course.

χ^2 test

The final question that we should answer is how can we characterize the probability of Type I error.

We can prove that under the null hypothesis

$$H_0 : nQ^{M\chi} \xrightarrow{d} \chi^2(k-\ell).$$

The more accurate statement of this result is the following theorem.

Theorem

Consider the testing problem (9). Let $D = D(\theta)$ be the $k \times \ell$ matrix with (i,j) entry $\frac{\partial g_i(\theta)}{\partial \theta_j}$.

If rank$(D) = \ell$, then $nQ^{M\chi} \xrightarrow{d} \chi^2(k-\ell)$

Note that we have also assumed the asymptotic normality of MLE.[3]

[3] If you do not know what "asymptotic normality of MLE" means you can skip this statement. Again the proof is beyond the scope of this course.

χ^2 test: example

Example (Gibbons and Chakraborti)

A quality control engineer has taken 50 samples of size 13 from a production process. The numbers of defectives for these samples are recorded below. Test the null hypothesis at significance level $\alpha = 0.05$. The number of defectives follows the binomial distribution.

Number of defects	Number of samples
0	10
1	24
2	10
3	4
4	1
5 or more	1

χ^2 test: example

As you can see this problem is slightly tricky. The reason is that our last category includes all the packages that have 5 or more defects in them.

Let X_i be the random variable that shows the number of defects in the ith sample.

The correct way to approach this problem is to first define a new random variable Y_i in the following way: $Y_i = X_i$ if $X_i < 5$ and $Y_i = 5$ if $X_i \geq 5$. Note that the probability mass function of Y_i satisfies the following:

$$\begin{aligned} \mathbb{P}(Y_i = k) &= \binom{13}{k} \beta^k (1-\beta)^{13-k}, \quad k < 5, \\ \mathbb{P}(Y_i = 5) &= \sum_{i=5}^{13} \binom{13}{i} \beta^i (1-\beta)^{13-i}, \end{aligned} \tag{10}$$

where β is the probability of each unit being a defect.

χ^2 test

The next step is to find MLE of β. Since the calculation of MLE is complicated and requires iterative algorithms such as Newton method (that we have not seen in this class), we will use an approximation that simplifies the calculations dramatically. If β is small, then $P(Y_i = 5)$ can be well approximated with $\binom{13}{5}\beta^5(1-\beta)^{13-5}$. Therefore, I assume that

$$\mathbb{P}(Y_i = 5) = \binom{13}{5}\beta^5(1-\beta)^{13-5}.$$

Note that since the number of defects is most of the time less than 5, we expect β to be very small. We have,

$$p(Y_1 = k_1, Y_2 = k_2, \ldots, Y_{50} = k_{50}) = \binom{n}{k_1}\binom{n}{k_2}\ldots\binom{n}{k_{50}}\beta^{k_1+k_2+\ldots+k_{50}}(1-\beta)^{(13-k_1)+(13-k_2)+\ldots+(13-k_{50})}.$$

Note that $k_1 + k_2 + \ldots + k_{50}$ is the total number of defects which is equal to $24 \times 1 + 10 \times 2 + 4 \times 3 + 4 \times 1 + 5 \times 1 = 65$. Therefore, the likelihood can be simplified to

$$p(Y_1 = k_1, Y_2 = k_2, \ldots, Y_{50} = k_{50}) = \binom{n}{k_1}\binom{n}{k_2}\ldots\binom{n}{k_{50}}\beta^{65}(1-\beta)^{650-65}.$$

χ^2 test

We take the logarithm of the likelihood and differentiate it with respect to β to obtain the maximizing value of β:

$$\frac{d}{d\beta} \log p(X_1 = k_1, X_2 = k_2, \ldots, X_{50} = k_{50}) = 0 \Rightarrow \frac{65}{\beta} = \frac{650 - 65}{1 - \beta} \Rightarrow \hat{\beta} = 0.1.$$

It turns out that if we do the calculations exactly, the correct MLE is $\hat{\beta} = 0.105$. So, our approximation is acceptable here. The following table summarizes the empirical probabilities with the ones that are coming from Binomial.

number of defects	$\hat{p}$(from observations)	CDF	$\hat{\pi}$(from binomial)
0	.2	.2	0.2542
1	0.48	.68	0.3671
2	.2	.88	0.2448
3	.08	.96	0.0997
4	.02	.98	0.0277
5 or more	.02	1	.0065

Therefore, $Q = \sum_{i=0}^{5} \frac{(\hat{p}_i - \pi_i)}{\pi_i} = 0.0885$ We know that the asymptotic distribution of nQ is $\chi^2(4)$. The significance level of 0.05 gives us threshold 9.48. If we compare $nQ = 4.425$, we conclude that we should accept the null hypothesis.

Goodness of fit tests: noncategorical random variables

Now we would like to study the goodness of the fit tests for real-valued (noncategorical) random variables.

The formal statement of the null and alternate hypotheses. Let $X_1, X_2, \ldots, X_n$ be iid drawn from a continuous distribution F.

We would like to test

$$H_0 : F(t) = F_0(t) \ \forall t \in \mathbb{R} \quad \text{versus} \quad H_1 : F(t) \neq F_0(t) \text{ for at least a } t \in \mathbb{R}.$$

Application of χ^2

The first idea to address this issue is to use the χ^2 test we introduced for categorial random variables.

Toward this goal, we first partition $\mathbb{R}$ into some intervals and then based on these intervals we define a categorical random variable Y which is a function of X.

We finally translate the null hypothesis on X to a null hypothesis on Y and use χ^2 test. Next example will clarify this procedure.

Example

We have observed a data set 0.26, 0.34, 0.22, 0.21, 0.06, 0.57, 0.04, 0.07, 0.43, 0.31, 0.22, 0.15, 0.8, 0.36, 0.59, 0.58, 0.18, 0.2, 0.57, 0.39, 0.17, 0.34, 0.30, 0.98, 0.67. We would like to test

$$H_0 : F = \text{Unif}[0, 1] \text{ versus } H_1 : F \neq \text{Unif}[0, 1]$$

Suppose $F = \text{Unif}[0, 1]$.

If we break the [0,1] interval into four intervals,

$$A_0 = [0, 0.25], A_1 = [0.25, 0.5], A_2 = [0.5, 0.75], A_3 = [0.75, 1],$$

then we expect

$$\mathbb{P}\{X_i \in A_0\} = \mathbb{P}\{X_i \in A_1\} = \mathbb{P}\{X_i \in A_2\} = \mathbb{P}\{X_i \in A_3\}.$$

Application of χ^2

Define the random variable $Y_i = j$ if $X_i \in A_j$. This random variable is clearly categorical. Under H_0, it is clear that

$$P(Y_i = 0) = \frac{1}{4}, P(Y_i = 1) = \frac{1}{4}, P(Y_i = 2) = \frac{1}{4}, P(Y_i = 3) = \frac{1}{4}, \tag{11}$$

Therefore, we can do the χ^2 test for the following null hypothesis:

$$H_0' : \begin{cases} \mathbb{P}\{Y_i = 0\} = \frac{1}{4} \\ \mathbb{P}\{Y_i = 1\} = \frac{1}{4} \\ \mathbb{P}\{Y_i = 2\} = \frac{1}{4} \\ \mathbb{P}\{Y_i = 3\} = \frac{1}{4} \end{cases} \quad \text{versus} \quad H_1' : \begin{cases} \text{At least one} \\ \text{of them is} \\ \text{not } \frac{1}{4} \end{cases}$$

If our data rejects H_0'. then it will reject H_0 as well. Otherwise, it is inconclusive. Clearly this approach suffers from several issues:

Application of χ^2

1. Even if H_0' is accepted by this χ^2 test we cannot be sure whether H_0 is true or not. Because there are many different distributions that are not uniform and still give us $\mathbb{P}\{X_i \in A_0\} = \mathbb{P}\{X_i \in A_1\} = \mathbb{P}\{X_i \in A_2\} = \mathbb{P}\{X_i \in A_3\} = 1/4$.

2. There is no good systematic way to determine the number of intervals and their locations.

For these reasons we would like to come up with new and more efficient approaches.