Goodness of fit tests

- Suppose that $(X_1, \ldots, X_n)$ is a random sample with distribution D. A goodness of fit testing problem is

$$H_0 : D = D_0 \text{ v.s. } H_1 : D \neq D_0,$$

where D_0 is a given distribution.

- Chi-squared goodness of fit test (卡方適合度檢定). Suppose that $(Y_1, \ldots, Y_n)$ is a random sample such that Y_1 takes values in $\{1, \ldots, k\}$. Let $p_j = P(Y_1 = j)$ for $j \in \{1, \ldots, k\}$. Consider the testing problem

$$H_0 : (p_1, \ldots, p_k) = (p_{0,1}, \ldots, p_{0,k}) \text{ v.s. } H_1 : (p_1, \ldots, p_k) \neq (p_{0,1}, \ldots, p_{0,k}), \tag{1}$$

where $p_{0,1}, \ldots, p_{0,k}$ are given postive numbers such that $\sum_{j=1}^k p_{0,j} = 1$. Let

$$N_j = \sum_{i=1}^n I_{\{j\}}(Y_i)$$

for $j = 1, \ldots, k$. The chi-squared goodness of fit test rejects H_0 at level α if and only if

$$\sum_{j=1}^k \frac{(N_j - np_{0,j})^2}{np_{0,j}} > c_{\alpha,k-1},$$

where $c_{\alpha,k-1}$ is the $(1-\alpha)$ quantile of $\chi^2(k-1)$.

- The chi-squared goodness of fit test is an approximate size α test since

$$\sum_{j=1}^k \frac{(N_j - np_j)^2}{np_j} \xrightarrow{\mathcal{D}} \chi^2(k-1) \text{ as } n \to \infty. \tag{2}$$

The proof of (2) is based on the following facts and the result in Problem 37. The proof is left as an exercise.

Fact 1 *Suppose that Z is a $k \times 1$ random vector such that $Z \sim N(0, \Sigma)$ and $\Sigma^2 = \Sigma$. Then $Z^T Z \sim \chi^2(m)$, where m is the trace of Σ.*

The proof of Fact 1 is given at the end of this handout.

Fact 2 *Suppose that $X_n \xrightarrow{\mathcal{D}} X$ as $n \to \infty$, A is a region such that $P(X \in A) = 1$ and g is continuous on A. Then $g(X_n) \xrightarrow{\mathcal{D}} g(X)$ as $n \to \infty$.*

The proof of Fact 2 is beyond the scope of this course. The special case where X is a constant c can be proved using Fact 2 in the handout "Central Limit Theorem and approximate confidence intervals".

- The quantile of a χ^2 distribution can be obtained using the R command `qchisq`. For example, the R output after running the command `qchisq(0.95, 1:4)` is

```
3.841459 5.991465 7.814728 9.487729
```

  so the 95% quantiles of $\chi^2(m)$ for $m = 1, 2, 3, 4$ are 3.841459, 5.991465, 7.814728, and 9.487729 respectively.

- Example 1. Suppose that $(X_1, \ldots, X_n)$ is a random sample from some distirbution D, where $n = 500$ and X_1 takes values in $\{0, 1, 2\}$. Suppose that we observe 200 0's, 220 1's and 80 2's in the sample. Can we conclude that D is not $Bin(2, 0.4)$ at level 0.05 based on the chi-squared goodness of fit test? The R output after running the command `qchisq(0.95, 1:4)` is

```
3.841459 5.991465 7.814728 9.487729
```

  Sol. Let $q_j = P(Bin(2, 0.4) = j) = C_j^2 0.4^j 0.6^{2-j}$ for $j = 0, 1, 2$, then

$$(q_0, p_1, q_2) = (0.36, 0.48, 0.16).$$

  Let $(N_0, N_1, N_2) = (200, 220, 80)$, then the observed test statistic is

$$\sum_{j=0}^{2} \frac{(N_j - 500q_j)^2}{500q_j} = 3.888889.$$

  The 0.95 quantile of $\chi^2(2)$ is 5.991465, so the observed test statistic does not exceed the 0.95 quantile of $\chi^2(2)$ and we cannot conclude that the distribution D is not $Bin(2, 0.4)$ at level 0.05.

- The chi-squared goodness of fit test can be applied to grouped data to test whether the original data are from a certain distribution.

  Example 2. Suppose that we have summaried annual salary information of 500 employees in a city as follows.

salary range (in 10^4 NTDs)	counts
$(0, 30]$	51
$(30, 60]$	216
$(60, 90]$	120
$(90, 120]$	52
$(120, \infty)$	61

  Suppose that the 500 annual salaries are IID random variables from some distribution D. Can we conclude that D is not $N(65, 36^2)$ at level 0.05 based on the chi-squared goodness of fit test? Note that if we run the following commands

```
a <- c(0, 30, 60, 90, 120)
b <- pnorm(a, mean=65, sd=36)
p <- diff(b)
c(p, 1-sum(p))
```

in R, then the output is

```
0.12997611 0.27929897 0.31152926 0.18041790 0.09877776
```

Also, the R output after running the command `qchisq(0.95, 2:5)` is

```
5.991465  7.814728  9.487729 11.070498
```

Sol. Let

$$(q_1, \ldots, q_5) = (0.12997611, 0.27929897, 0.31152926, 0.18041790, 0.09877776)$$

and

$$(N_1, \ldots, N_5) = (51, 216, 120, 52, 61),$$

then the observed test statistic is

$$\sum_{j=1}^{5} \frac{(N_j - 500q_j)^2}{500q_j} = 71.87922.$$

The 0.95 quantile of $\chi^2(4)$ is 9.487729, so the observed test statistic exceeds the 0.95 quantile of $\chi^2(4)$ and we can conclude that the distribution D is not $N(65, 36^2)$ at level 0.05. Note that we can compute the observed test statistics by running the following R caomands

```
q <- c(0.12997611, 0.27929897, 0.31152926, 0.18041790, 0.09877776)
N <- c(51, 216, 120,  52,  61)
sum( (N-500*q)^2/(500*q) )
```

- Suppose that $(X_1, \ldots, X_n)$ is a random sample with CDF F. The Kolmogorov-Smirnov test can be applied to test

$$H_0 : F = F_0 \text{ v.s. } H_1 : F \neq F_0, \tag{3}$$

where F_0 is a given CDF.

- The Kolmogorov-Smirnov test is based on the statistic

$$\sup_x |\hat{F}(x) - F_0(x)|,$$

where $\hat{F}$ is the empirical cumulative distribution function (empirical CDF) based on the sample $(X_1, \ldots, X_n)$, which is defined by

$$\hat{F}(x) = \frac{1}{n} \sum_{i=1}^{n} I_{(-\infty, x]}(X_i) \text{ for } x \in (-\infty, \infty).$$

- The R command for computing the empirical CDF based on a sample `x` is `ecdf(x)`. For instance, suppose that we generate `x`: a random sample of size 100 from $N(0,1)$ by running the following commands in R:

```
set.seed(1)
x <- rnorm(100)
```

Then we can compute and plot the empirical CDF based on the sample `x` and add the curve of the $N(0,1)$ CDF in the plot for comparison by running the following R commands:

```
x.ecdf <- ecdf(x)
m <- min(x)
M <- max(x)
curve(x.ecdf, m-1, M+1)
curve(pnorm, m-1, M+1 , add=TRUE, col=2)
```

- Suppose that $(X_1, \ldots, X_n)$ is a random sample with CDF F and let $\hat{F}$ be the empircal CDF based on the sample $(X_1, \ldots, X_n)$. When F is strictly increasing and continuous, it can be shown that the distribution of the Kolmogorov-Smirnov test statistic

$$\sup_x |\hat{F}(x) - F(x)|$$

does not depend on F. Let D_0 denote this disitribution, then the Kolmogorov-Smirnov test rejects the H_0 in (3) at level α if and only if

$$\sup_x |\hat{F}(x) - F_0(x)| > c_\alpha, \tag{4}$$

where c_α is the $(1-\alpha)$ quantile of D_0. Here (4) is equivalent to

$$\alpha > \underbrace{P(D_0 > \text{ observed } \sup_x |\hat{F}(x) - F_0(x)|)}_{p\text{-value}}.$$

Note.

  - The Kolmogorov-Smirnov test statistic can be computed based on the equality

$$\sup_x |\hat{F}(x) - F_0(x)| = \max_{i \in \{1,\ldots,n\}} |\underbrace{\hat{F}(X_{(i)})}_{=i/n} - F_0(X_{(i)})|.$$

  - The p-value $P(D_0 > \text{ observed } \sup_x |\hat{F}(x) - F_0(x)|)$ can be obtained using simulated data from D_0.

- The R command for conducting the the Kolmogorov-Smirnov test is

```
ks.test(x, F0)
```

where `x` is the sample and `F0` is the F_0 in (3).

- `ks.test(x,F0)$statistic` gives the observed test statistic $\sup_x |\hat{F}(x) - F_0(x)|$.
- `ks.test(x,F0)$p.value` gives the p-value of the Kolmogorov-Smirnov test.

- The R command for reading data into R is `read.table` if the data set is in a table form stored in a text file. Suppose that the data file is called `data.txt`, the file is in the working directory of R, and the table columns are separated by spaces. Run the R command

```
x <- read.table(file="data.txt")
```

and the data set is read into R. The working directory of R can be found by running the command

```
getwd()
```

in R.

- Example 3. Suppose that $(X_1, \ldots, X_n)$ is a random sample from some distribution D, and the observed sample is stored in the data file

`https://stat.walkup.tw/teaching/math_stat_under/data/example_data.txt`

Determine whether it is reasonable to assume that D is a normal distribution.

Sol. Let $\Phi_{\mu,\sigma}$ be the CDF of $N(\mu, \sigma^2)$ for $\mu \in (-\infty, \infty)$ and $\sigma > 0$. We would like to apply Kolmogorov-Smirnov test to check whether there is a strong evidence against the hypothesis that $D = N(\mu_0, \sigma_0)$, where

$$(\mu_0, \sigma_0) = \arg\min_{(\mu,\sigma)} \left(\sup_x |\hat{F}(x) - \Phi_{\mu,\sigma}(x)| \right).$$

To find (μ_0, σ_0), we will define a function `g` in R such that

$$\mathrm{g}(\mu, \sigma) = \sup_x |\hat{F}(x) - \Phi_{\mu,\sigma}(x)|$$

and then use the R function `optim` to find (μ_0, σ_0): the minimizer of `g`. The R scripts for the above procedure are given below.

```
data <- read.table(file="example_data.txt")  # read the data into R

data <- as.numeric(data[,1])   #take the first column of data
                               #and change its mode to a numeric vector,
                               #and then replace data by this vector.

#define the function g
g <- function(mu, sigma){
  F <- function(x){ pnorm(x, mean=mu, sd= sigma) }
  return(ks.test(data, F)$statistic)
}

#reparametrization for unconstrained optimization
#define a function g1
# g1( c(mu,eta)) = g(mu, sigma)
# eta = log(sigma); sigma = exp(eta)
g1 <- function(mu.eta){
 mu <- mu.eta[1]
 sigma <- exp(mu.eta[2])
 return(g(mu, sigma))
}

#compute initial value of parameters for optimization
mu1 <- mean(data)
sigma1 <- sd(data)
eta1 <- log(sigma1)

#perform optimization for g1
opt <- optim(c(mu1, eta1), g1)

#compute (mu0, sigma0): the minimizer of g
mu0 <- opt$par[1]
eta0 <- opt$par[2]
sigma0 <- exp(eta0)

# perform the Kolmogorov-Smirnov test
F0 <- function(x){ pnorm(x, mean=mu0, sd=sigma0) }
ks.test(data, F0)
```

The R output after running the above scripts is

```
One-sample Kolmogorov-Smirnov test

data:  data
D = 0.0017443, p-value = 0.9212
```

```
alternative hypothesis: two-sided
```

and it shows the p-value for the Kolmogorov-Smirnov test is 0.9212, so we do not have evidence against the hypothesis that $D = N(\mu_0, \sigma_0)$. It is reasonable to assume that D is a normal distribution.

- For the testing problem in (3), the Kolmogorov-Smirnov test statistic can be used to measure the distancs between F and F_0. If we have to choose one of two parametric families as the family for the data distribution, we can use the Kolmogorov-Smirnov test statistic.

- Example 4. Suppose that the data in Example 3 are IID from $N(\mu, \sigma)$ for some (μ, σ) or $U(a, b)$ (the uniform distribution on (a, b)) for some (a, b). Determine the data are from the normal family or the uniform family based on

$$D_j = \min_{F \in \mathcal{F}_j} \sup_x |\hat{F}(x) - F(x)|$$

for $j \in \{1, 2\}$, where $\hat{F}$ is the empirical CDF, $\mathcal{F}_1$ is the collection of CDFs of the distirbutions in the normal family, and $\mathcal{F}_2$ is the collection of CDFs of the distirbutions in the uniform family.

Sol. From the solution to Example 3, we have $D_1 = 0.0017443$. To compute D_2, we will modify the R scripts in Example 3 as follows:

```
data <- read.table(file="example_data.txt")  # read the data into R
data <- as.numeric(data[,1])

#define a function g so that g(a,b) is
#the Kolmogorov-Smirnov test staistic with F = the U(a,b) CDF
g <- function(a,b){
  F <- function(x){ punif(x, min=a, max=b) }
  return(ks.test(data, F)$statistic)
}


#define a function g1
# g1(c(a,eta)) = g(a,b)
# b = a + exp(eta)
g1 <- function(a.eta){
  a <- a.eta[1]
  b <- a+exp(a.eta[2])
  return(g(a,b))
}

#compute initial value of (a, eta) for optimization
a1 <- min(data)
b1 <- max(data)
```

```
eta1 <- log(b1-a1)

#perform optimization for g1
opt <- optim(c(a1, eta1), g1)


#find (a0,b0): the minimizer of g
a0 <- opt$par[1]
b0 <- a0+exp(opt$par[2])

# perform the Kolmogorov-Smirnov test
F0 <- function(x){ punif(x, min=a0, max=b0) }
ks.test(data, F0)
```

The R output after running the above scripts is

```
One-sample Kolmogorov-Smirnov test

data:  data
D = 0.048435, p-value < 2.2e-16
alternative hypothesis: two-sided
```

and it shows that the Kolmogorov-Smirnov test statistic is 0.04835, so $D_2 = 0.04835$. Since $D_1 = 0.0017443 < 0.04835 = D_2$, we choose the normal family as the family of data distribution.

- In Examples 3 and 4, we have computed the "distance" between the empirical CDF to the collection of CDFs of a family, and chosen a family as the family of distributions for the data based on the distance. This approach can be extended to the case where we have more than two families to choose from.

- Proof of Fact 1.

  (i) We will first show that the eigenvalues of Σ are in $\{0, 1\}$. Note that Σ is a $k \times k$ symmetric matrix, so it has k real eigenvalues. Moreover, the condition that $\Sigma^2 = \Sigma$ implies that each eigenvalue of Σ is either 0 or 1. To see this, let λ be an eigenvalue of Σ and v is the corresponding eigenvector. Then by the definition of eigenvalue/eigenvector,

  $$\Sigma^2 v = \Sigma\Sigma v = \Sigma\lambda v = \lambda\Sigma v = \lambda^2 v. \tag{5}$$

  In addition, by the assumption that $\Sigma^2 = \Sigma$, we have

  $$\Sigma^2 v = \Sigma v = \lambda v. \tag{6}$$

  From (5) and (6), we have $\lambda v = \lambda^2 v$. Since v is not a vector of zeros, we must have $\lambda = \lambda^2$, which implies that $\lambda = 0$ or 1.

  (ii) Next, we will show that $Z^T Z$ is equal to sum of squares of IID $N(0, 1)$ random variables, so the distribution of $Z^T Z$ is a χ^2 distribution. To see this, note that the matrix Σ can be decomposed as $\Sigma = PDP^T$, where PP^T is the $k \times k$ identity matrix, and D is a $k \times k$ diagonal matrix whose diagonal elements are the eigenvalues of Σ. Let $U = P^T Z$, then $U \sim N(0, P^T \Sigma P) = N(0, D)$. Since D is a diagonal matrix whose diagonal elements are in $\{0, 1\}$, some components of U are zeros and the other components are IID $N(0, 1)$ random variables. Let r be the number of nonzero components in U, then

  $$r = \text{trace of } D \text{ and } U^T U \sim \chi^2(r).$$

  Therefore,

  $$Z^T Z = (PU)^T PU = U^T U \sim \chi^2(r),$$

  where r is the trace of D. Since

  $$\text{trace of } \Sigma = \text{ trace of } PDP^T = \text{ trace of } DP^T P = \text{ trace of } D,$$

  r is equal to the trace of Σ. The proof of Fact 1 is complete.