Chi-squared distributions and testing for population variance.

- Suppose that Z_1, …, Z_m are IID $N(0,1)$. Then the distribution of $Z_1^2 + \cdots + Z_m^2$ is called the chi-squared distribution (χ^2 distribution, 卡方分配) with m degrees of freedom, denoted by $\chi^2(m)$.
- Relavent R commands
  - `dchisq(x, m)`: the PDF for χ^2(`m`) evaluated at `x`.
  - `pchisq(x, m)`: the CDF for χ^2(`m`) evaluated at `x`.
  - `qchisq(q, m)`: the `q` quantile for χ^2(`m`).
  - `rchisq(n, m)`: generating `n` IID random variables from χ^2(`m`).
- Example 1. Find the mean and the median of $\chi^2(10)$.

  Sol. Let Z_1, …, Z_{10} be IID $N(0,1)$ random variables. Then the mean of $\chi^2(10)$ is $E(Z_1^2 + \cdots + Z_{10}^2) = 10E(Z_1^2) = 10$. The median of $\chi^2(10)$ is

```
qchisq(0.5, 10) #9.341818
```

  Note that the mean > the median, so the $\chi^2(10)$ distriubtion is positively skewed (or right skewed).
- Definition 1. A distribution is positively skewed (or right skewed) if its mean is greater than its median.
- Definition 2. A distribution is negatively skewed (or left skewed) if its mean is less than its median.
- Example 2. Generate 10000 independent random sums, where each sum is of the form $Z_1^2 + \cdots + Z_9^2$, and Z_1, …, Z_9 are IID $N(0,1)$ random variables. Plot the normalized histogram and add the $\chi^2(9)$ PDF.

```
n.tr <- 10^5
ans <- rep(0, n.tr)
for (i in 1:n.tr){
  z <- rnorm(9)
  ans[i] <- sum(z^2)
}
hist(ans, nclass="scott", freq=FALSE)

dchisq9 <- function(x){ dchisq(x, df=9)}
curve(dchisq9, 0, 40, add=TRUE, col="red")
```

- Suppose that $(X_1, \cdots, X_n)$ is a random sample from $N(\mu, \sigma^2)$. Let

$$S = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}}$$

be the sample standard deviation. Then

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

- Example 3. Generate 10000 random numbers from the distribution of $9S^2/4$, where S is the sample standard deviation of a random sample of size 10 from $N(0,4)$. Plot the normalized histogram for the 10000 generated values and add the $\chi^2(9)$ PDF.

```
x <- 1:10000
for (i in 1:10000){
 S <- sd(rnorm(10, sd=2))
 x[i] <- 9*S^2/4
}
hist(x, nclass="scott", freq=FALSE)

f <- function(x){ dchisq(x, 9) }
curve(f,add=TRUE, col="red")
```

- Suppose that $(X_1, \cdots, X_n)$ is a random sample from $N(\mu, \sigma^2)$. Let S be the sample standard deviation. Consider the testing problem

$$H_0 : \sigma \leq \sigma_0 \text{ v.s. } H_1 : \sigma > \sigma_0.$$

Consider the test that rejects H_0 at level a whenever

$$\frac{(n-1)S^2}{\sigma_0^2} > k_{a,n-1},$$

where $k_{a,n-1}$ is the $(1-a)$ quantile for $\chi^2(n-1)$. Then the test is of size a. The p-value for the test is

$$P\left(\chi^2(n-1) > \text{ observed } \frac{(n-1)S^2}{\sigma_0^2}\right).$$

- Example 4. (modified from the desk production example in the text).
  - weekly desk production $\sim N(\mu, \sigma^2)$
  - data: the numbers of desks produced per week in last 30 weeks
  - sample standard deviation is 14.3.

  Questions

  - Can we conclude that $\sigma > 11$ at the 0.05 significance level?
  - Can we conclude that $\sigma > 11$ at the 0.01 significance level?

  Sol. Consider the testing problem

$$H_0 : \sigma \leq 11 \text{ versus } H_1 : \sigma > 11$$

and the test that rejects H_0 at level a whenever

$$\frac{(30-1)S^2}{11^2} > k_{a,30-1}.$$

The observed test statistic $(30-1)S^2/(11^2)$ is $29(14.3)^2/(11^2) = 49.01$, so the p-value is $P(\chi^2(29) > 49.01)$, which can be found using the R command

```
1-pchisq(49.01, 29)
```

The p-value is approximately 0.01152017, so we can conclude that $\sigma > 11$ at the 0.05 significance level but not at the 0.01 level.

We can also reach the same conclusion by finding $k_{0.05,29} = 42.557$ and $k_{0.01,29} = 49.588$ using the table "Quantiles for χ^2 distributions on the course web site.

- Let S be the sample standard deviation of a random sample of size n from $N(\mu, \sigma^2)$. Given $\alpha \in (0, 0.5)$, to construct a level α test for testing

$$H_0 : \sigma = \sigma_0 \text{ v.s. } \sigma \neq \sigma_0,$$

choose $\theta \in (0,1)$ and consider the test that rejects $H_0 : \sigma = \sigma_0$ if and only if

$$\sigma_0^2 \notin \left[\frac{(n-1)S^2}{k_{\alpha\theta,n-1}}, \frac{(n-1)S^2}{k_{1-\alpha(1-\theta),n-1}}\right],$$

then the test is of size α. Moreover,

$$\left[\frac{(n-1)S^2}{k_{\alpha\theta,n-1}}, \frac{(n-1)S^2}{k_{1-\alpha(1-\theta),n-1}}\right]$$

is a $(1-\alpha)$ confidence interval of σ^2.

- Example 5. Let S be the sample standard deviation of a random sample of size 10 from $N(\mu, \sigma^2)$. Note that running the R command

```
1/qchisq(c(0.02, 0.04, 0.97, 0.99),df=9)
```

gives

```
[1] 0.39488565 0.32209537 0.05411377 0.04615528
```

and running the R command

```
1/qchisq(c(0.02, 0.04, 0.97, 0.99),df=10)
```

gives

```
[1] 0.32689872 0.27052314 0.05019599 0.04308627
```

(a) Give a 95% C.I. for σ^2.

(b) Suppose that we observe $S = 10$. Compute the observed 95% C.I. for σ^2 using the C.I. from Part (a).

(c) Propose a level 0.05 test for testing

$$H_0 : \sigma = 3 \text{ v.s. } H_1 : \sigma \neq 3$$

based on S. Suppose that we observed $S = 10$. Can we conclude that $\sigma \neq 3$ at level 0.05 based on the proposed test?

Sol.

(a) Note that for $\theta \in (0, 1)$,

$$\left[\frac{9S^2}{k_{0.05\theta,9}}, \frac{9S^2}{k_{1-0.05(1-\theta),9}}\right] \tag{1}$$

is a 95% C.I. for σ^2. Take θ so that $0.05\theta = 0.01$, then $\theta = 1/5$, $0.05(1 - \theta) = 0.04$,

$$k_{0.05\theta,9} = k_{0.01,9} = 1/0.04615528,$$

and

$$k_{1-0.05(1-\theta),9} = k_{1-0.04,9} = 1/0.32209537$$

The 95% C.I. in (1) is

$$\left[0.04615528 \cdot 9S^2, 0.32209537 \cdot 9S^2\right] = [0.4153975S^2, 2.898858S^2]. \tag{2}$$

(b) Since the observed S is 10, the observed 95% C.I. for σ^2 in (2) is

$$[0.4153975S^2, 2.898858S^2]\big|_{S=10} = [41.53975, 289.8858]$$

(c) Consider the test that rejects $H_0 : \sigma = 3$ if and only if

$$3^2 \notin \left.\left[\frac{9S^2}{k_{0.05\theta,9}}, \frac{9S^2}{k_{1-0.05(1-\theta),9}}\right]\right|_{\theta=1/5},$$

then the test is of size 0.05 (and hence of level 0.05). From the calculation in Part (a),

$$\left.\left[\frac{9S^2}{k_{0.05\theta,9}}, \frac{9S^2}{k_{1-0.05(1-\theta),9}}\right]\right|_{\theta=1/5} = [0.4153975S^2, 2.898858S^2],$$

so the test rejects $H_0 : \sigma = 3$ if and only if

$$9 \notin [0.4153975S^2, 2.898858S^2].$$

When the observed $S = 10$, the interval

$$[0.4153975S^2, 2.898858S^2]\big|_{S=10} = [41.53975, 289.8858],$$

which does not contain 9. Therefore, we can conclude that $\sigma \neq 3$ at level 0.05 based on the proposed test.