

- (1) (*t* Distribution) The random sample in this exercise is an $M \times n$ matrix \mathbf{X} . For this exercise, $M = 1000, n = 20$. The element of each row of \mathbf{X} , x_1, x_2, \dots, x_n , are iid normal distribution with mean $\mu = 10$, variance $\sigma^2 = 9$; there sample mean $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ and sample variance $s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$. Then the number of rows M is the number of independent realization of each random row vector. Additionally, let Z be a standard normal distribution, and $U \sim \chi_n^2$ be the chi-square distribution with degrees of freedom n . Please answer the following questions.

- (a) Here we briefly show why $\frac{\bar{x} - \mu}{s/\sqrt{n}} \sim t_{n-1}$.

If Z and U are independent, then the distribution $Z/\sqrt{U/n}$ is called the *t distribution* with n degrees of freedom with probability density function

$$f_n(t) = \frac{\Gamma[(n+2)/2]}{\sqrt{n\pi}\Gamma(n/2)} \left(1 + \frac{t^2}{n}\right)^{-(n+1)/2},$$

where the *gamma function* $\Gamma(\cdot)$ is defined by

$$\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt, \quad x \in \mathbb{R}.$$

Now we first check that $(n-1)s^2/\sigma^2$ is the chi-square distribution with $n-1$ degrees of freedom. Note that

$$\frac{1}{\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 = \sum_{i=1}^n \left(\frac{x_i - \mu}{\sigma} \right)^2 \sim \chi_n^2.$$

Also,

$$\frac{1}{\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 = \frac{1}{\sigma^2} \sum_{i=1}^n [(x_i - \bar{x}) + (\bar{x} - \mu)]^2.$$

Expanding the square and using the fact that $\sum_{i=1}^n (x_i - \bar{x}) = 0$, we obtain

$$\frac{1}{\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 = \frac{1}{\sigma^2} \sum_{i=1}^n (x_i - \bar{x})^2 + \left(\frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \right)^2 \quad (1)$$

$$\Rightarrow W = \frac{(n-1)s^2}{\sigma^2} + V, \quad (2)$$

, where $W \sim \chi_n^2$, and $V \sim \chi_1^2$.

At this moment, it can be shown geometrically that the random variable \bar{x} and the vector of random variables $(x_1 - \bar{x}, x_2 - \bar{x}, \dots, x_n - \bar{x})$ are independent. Hence, W is independent of V . Consequently, $(n-1)s^2/\sigma^2$ is the chi-square distribution with $n - 1$ degrees of freedom.

Next we show that

$$\frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{\left(\frac{\bar{x} - \mu}{\sigma/\sqrt{n}}\right)}{\sqrt{s^2/\sigma^2}} = \frac{Z}{\sqrt{s^2/\sigma^2}}.$$

Hence it is a t distribution with $n - 1$ degrees of freedom.

Now from the random sample of this exercise, we make the transformation

$$y_m = \frac{\sum_{n=1}^N \mathbf{X}_{m,n}/N - \mu}{s/\sqrt{N}}, \quad m = 1, 2, \dots, M;$$

that is, y_m is the M times realization of \bar{x} . Make the histogram of y_m , $m = 1, 2, \dots, M$ and also plot the pdf curve of the t distribution with degrees of freedom $n-1$ on the same figure (the pdf of t distribution can be plotted by the Matlab function `tpdf`).

- Solution: See Fig. 1.

- (2) (Confidence Interval from Point Estimation) Here we will illustrate the insight behind the *confidence interval*. But before going on, we will first introduce the concept of *estimation*. A *point estimation* is a function $\hat{\theta} = g(x)$ of the observation vector $x = [x_1, \dots, x_n]$. The corresponding random variable $\hat{\theta} = g(\mathbf{x})$ is the *point estimator* of θ . Recall that any function of the sample vector $\mathbf{x} = [\mathbf{x}_1, \dots, \mathbf{x}_n]$ is called a *statistic*. Thus a point estimator is a statistic. Can we draw the near certainty a conclusion about the true value of θ ? We cannot do so if we claim that θ equals its point estimate $\hat{\theta}$ or any other constant. We can, however, conclude with near certainty that θ equals $\hat{\theta}$ within specified tolerance limits. This leads to the concept of *interval estimate*. An *interval estimate* of a parameter θ is an interval (θ_1, θ_2) , the endpoints of which are functions $\theta_1 = g_1(x)$ and $\theta_2 = g_2(x)$ of the observation vector x . The corresponding random interval (θ_1, θ_2) is the *interval estimator* of θ . We shall say that (θ_1, θ_2) is a γ *confidence interval* of θ if

$$\Pr(\theta_1 < \theta < \theta_2) = \gamma = 1 - \alpha.$$

The constant γ is the *confidence coefficient* of the estimate and α is the *confidence (significance) level*. Consequently, a *confidence interval* for a population parameter θ is a random interval, calculated from the sample which

contains θ with some specified probability. For example, a $100(1 - \alpha)\%$ confidence interval for θ is a random interval that contains θ with probability $1 - \alpha$.

- (a) Here we wish to estimate the mean μ of the normal random vector $x = [x_1, \dots, x_n]$. We use the sample mean $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ as the point estimate of μ . Suppose first that the variance σ^2 of x is known. Hence the point estimator \bar{x} of μ is $\mathcal{N}(\mu, \sigma^2/n)$. Denoting by $z(u)$ the point beyond which the standard normal distribution has probability u , we conclude that

$$\Pr \left(-z(\alpha/2) \leq \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \leq z(\alpha/2) \right) \quad (3)$$

$$= \Pr \left(\mu - z(\alpha/2) \frac{\sigma}{\sqrt{n}} \leq \bar{x} \leq \mu + z(\alpha/2) \frac{\sigma}{\sqrt{n}} \right) \quad (4)$$

$$= \Pr \left(\bar{x} - z(\alpha/2) \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z(\alpha/2) \frac{\sigma}{\sqrt{n}} \right) \quad (5)$$

$$= 1 - \frac{\alpha}{2} - \frac{\alpha}{2} = 1 - \alpha. \quad (6)$$

Thus we can state with significance level α that μ is in the confidence interval $\bar{x} \pm z(\alpha)\sigma/\sqrt{n}$.

If σ is unknown, then we cannot use the above derivations (equation (3)-(6)). Since the sample variance $s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$ is an unbiased estimate of σ^2 and it tends to σ^2 as $n \rightarrow \infty$. Hence, for large n we can use the approximation $s \simeq \sigma$ in the above derivations ($\frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$). However, because the estimator \mathbf{s} is also the random variable, the random variable $\frac{\bar{x} - \mu}{s/\sqrt{n}}$ is no longer normal distributed. Fortunately, from question (1) we know that it has a Student t distribution with $n - 1$ degrees of freedom. Denoting by $t_{n-1}(u)$ that point beyond which t_{n-1} has probability u , we conclude that

$$\begin{aligned} & \Pr \left(-t_{n-1}(\alpha/2) \leq \frac{\bar{x} - \mu}{s/\sqrt{n}} \leq t_{n-1}(\alpha/2) \right) \\ &= \Pr \left(\bar{x} - t_{n-1}(\alpha/2) \frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + t_{n-1}(\alpha/2) \frac{s}{\sqrt{n}} \right) \\ &= 1 - \alpha \end{aligned}$$

Note that the random sample x in this exercise has M realizations, please plot the figure of the confidence interval of μ with unknown σ^2 from 1st to 20th realizations with significance level $\alpha = 0.1$ and the true

population value μ . Additionally, calculate the percentage that μ falls in the confidence interval of M realizations. (*Hint: $t_{n-1}(\alpha/2)$ can be calculated by the Matlab function `tinv`*)

- Solution: See Fig. 2. The percentage that μ falls in the confidence interval of M realizations is 0.9980 .
- (b) Now we want to derive the confidence interval of the variance σ^2 of the normal random vector $x = [x_1, \dots, x_n]$. We assume first that the mean μ of x is known and we use the point estimator of σ^2 as

$$\hat{s}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2. \quad (7)$$

The reason that we choose this point estimator is that it is not only the *maximum likelihood estimator*, but also the *consistent estimator*; that is, it can be shown that

$$\begin{aligned} \mathbb{E}(\hat{s}^2) &= \sigma^2; \\ \sigma_{\hat{s}^2}^2 &= \frac{2\sigma^4}{n} \rightarrow 0 \quad \text{as } n \rightarrow \infty. \end{aligned}$$

Now we find the confidence interval of it. Recall that $n\hat{s}^2/\sigma^2$ is a χ_n^2 distribution with degrees of freedom n . To determine the confidence interval, we introduce two constant c_1 and c_2 such that

$$\begin{aligned} \Pr\left(\frac{n\hat{s}}{\sigma^2} \leq c_1\right) &= \alpha/2 \\ \Pr\left(\frac{n\hat{s}}{\sigma^2} \geq c_2\right) &= \alpha/2. \end{aligned}$$

Without loss of generality, we choose $c_1 = \chi_n^2(1 - \alpha/2)$, $c_2 = \chi_n^2(\alpha/2)$ for convenience, which yields

$$\begin{aligned} &\Pr\left(\chi_n^2(1 - \alpha/2) \leq \frac{n\hat{s}^2}{\sigma^2} \leq \chi_n^2(\alpha/2)\right) \\ &= \Pr\left(\frac{n\hat{s}^2}{\chi_n^2(\alpha/2)} \leq \sigma^2 \leq \frac{n\hat{s}^2}{\chi_n^2(1 - \alpha/2)}\right) \\ &= 1 - \alpha. \end{aligned}$$

If μ is unknown, we can only use the sample mean $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ to replace μ in equation (7); that is

$$n\hat{s}^2/\sigma^2 \rightarrow (n-1)s^2/\sigma^2.$$

Noting that $(n-1)s^2/\sigma^2 \sim \chi_{n-1}^2$, hence

$$\begin{aligned} & \Pr(\chi_{n-1}^2(1-\alpha/2) \leq (n-1)s^2/\sigma^2 \leq \chi_{n-1}^2(\alpha/2)) \\ = & \Pr\left(\frac{(n-1)s^2}{\chi_{n-1}^2(\alpha/2)} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi_{n-1}^2(1-\alpha/2)}\right) \\ = & 1 - \alpha. \end{aligned}$$

Therefore, a $100(1-\alpha)\%$ confidence interval for σ^2 is

$$\left(\frac{(n-1)s^2}{\chi_{n-1}^2(\alpha/2)}, \frac{(n-1)s^2}{\chi_{n-1}^2(1-\alpha/2)} \right).$$

Similar with question (2)(b), plot the figure of the confidence interval of σ^2 with unknown μ from 1st to 20th realizations with significance level $\alpha = 0.1$ and the true population value σ^2 . Additionally, calculate the percentage that σ^2 falls in the confidence interval of M realizations. (*Hint*: $\chi_{n-1}^2(\alpha/2)$ can be calculated by the Matlab function `chi2inv`)

- Solution: See Fig. 3. The percentage percentage that σ^2 falls in the confidence interval of M realizations is 0.9100.

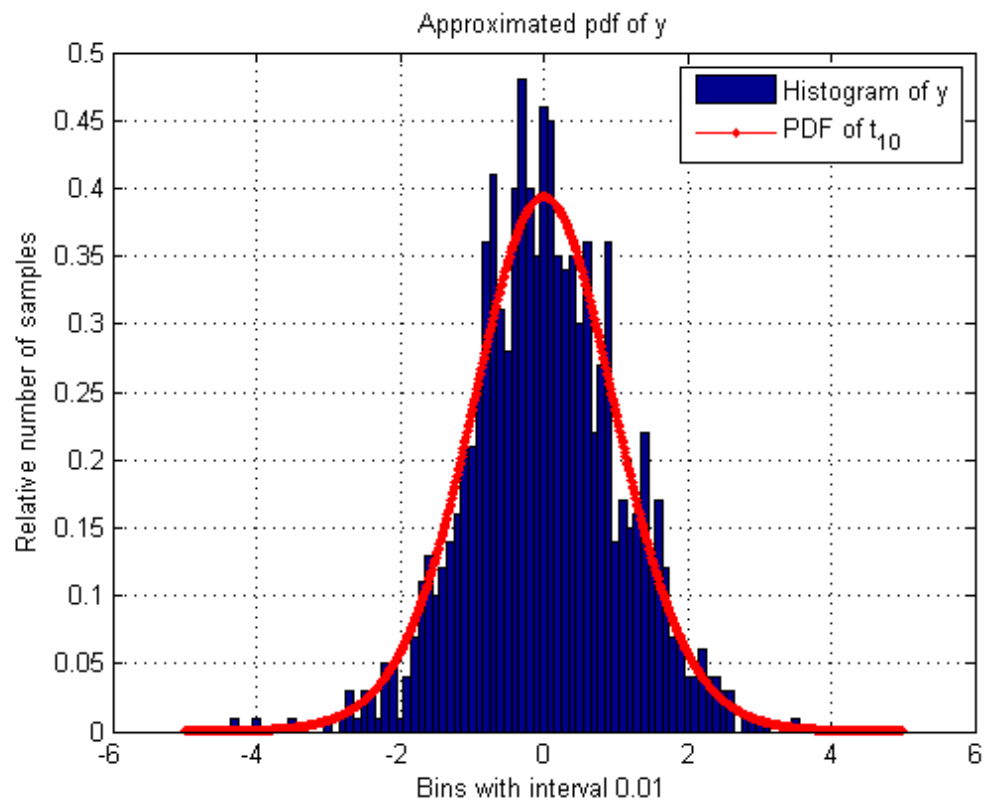


Figure 1: Approximated probability density function of t distribution

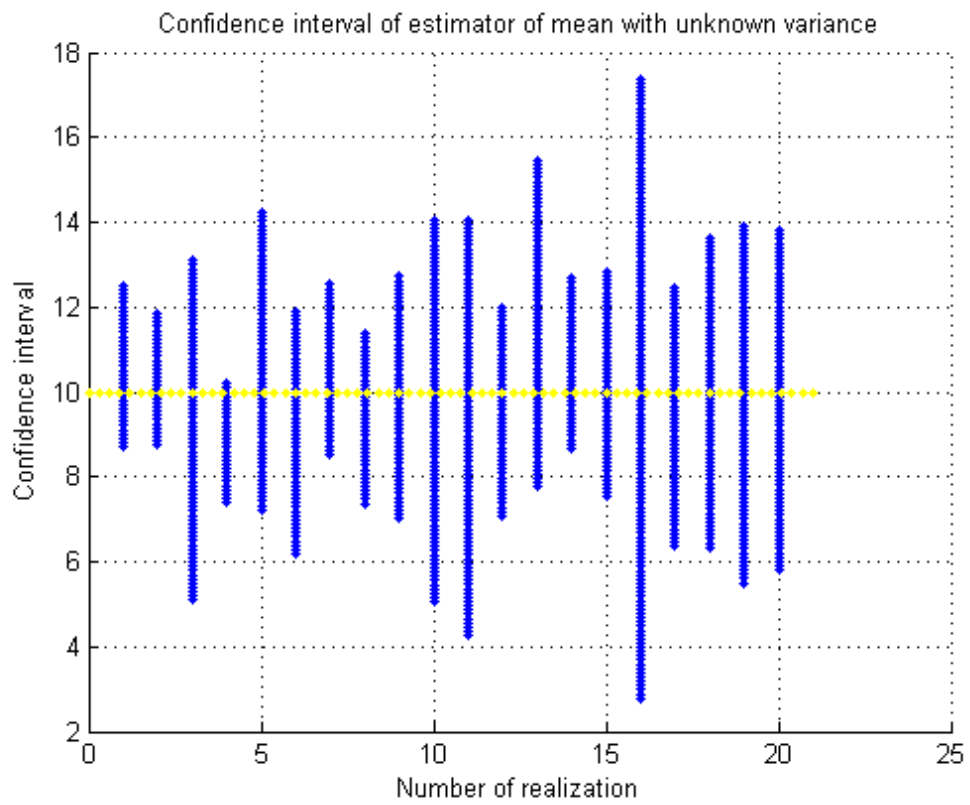


Figure 2:

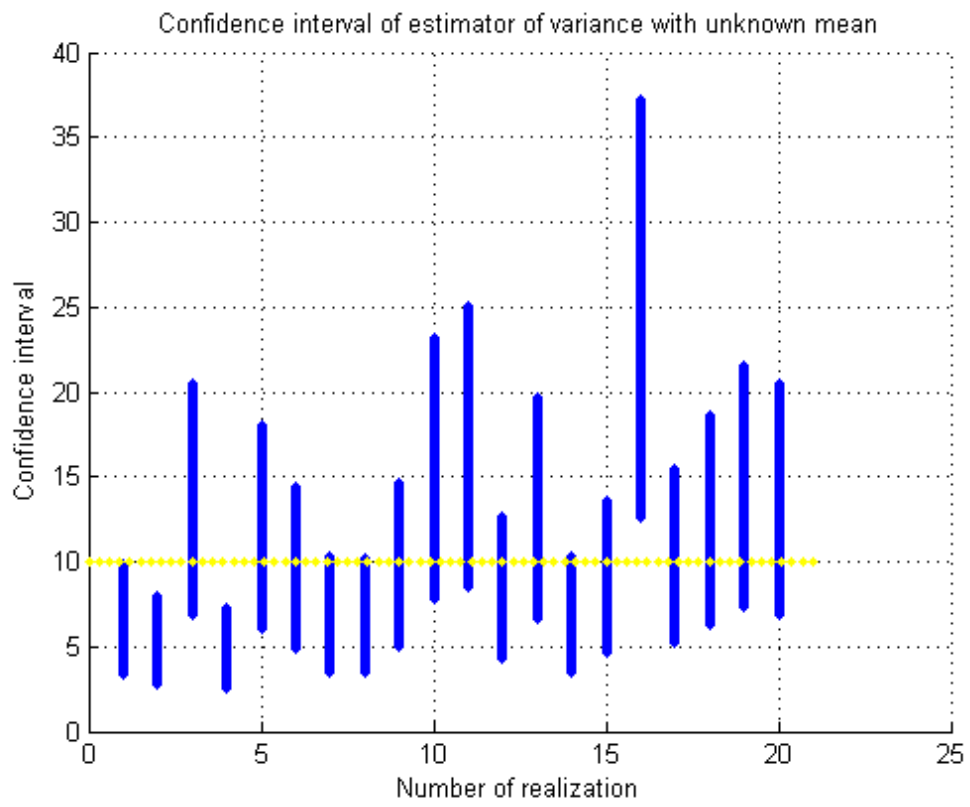


Figure 3: