

# STAT2006 Basic Concepts in Statistics and Probability II

## Tutorial 8 Confidence Interval for Variance and Proportion

Chan Chun Ho, Benjamin<sup>\*†</sup>

March 12, 2018

### Abstract

It aims to introduce concepts of confidence interval for variance and proportion. Some exercises are provided for students to practice. Some materials do credit to former TAs while some are extracted from textbooks “Probability and Statistical Inference” (Hogg and Tanis) used in STAT2001/2006.

## Notations and Definitions

- Set of real numbers:  $\mathbb{R} = (-\infty, \infty)$ ; Closed interval:  $[a, b] = \{x \in \mathbb{R} \mid a \leq x \leq b\}$
- Is defined to be:  $\triangleq$ ; Expectation:  $E(\cdot)$ ; Sample mean:  $\bar{X}$ ; Population mean:  $\mu_X$
- Covariance:  $Cov(\cdot)$ ; Variance:  $Var(\cdot)$ ; Sample variance:  $S_X^2$ ; Population variance:  $\sigma_X^2$
- Probability distribution: A tilde ( $\sim$ ) means “has the probability distribution of”.
- Random sample: Each random variable has the same probability distribution and all are mutually independent.
- Sample size:  $n$ ; Significance level:  $\alpha$ ; Confidence coefficient:  $1 - \alpha$
- Normal distribution:  $N(\mu, \sigma^2)$ , where  $\mu$  is mean and  $\sigma^2$  is variance.
- $t$  distribution:  $t(r)$ ; Chi-squared distribution:  $\chi^2(r)$ , where  $r$  is degrees of freedom.
- $F$  distribution:  $F(r_1, r_2)$ , where  $r_1$  and  $r_2$  are degrees of freedom.
- Normal cutoff: Select  $z_{\alpha/2}$  so that  $P(Z \geq z_{\alpha/2}) = \alpha/2$ , where  $Z \sim N(0, 1)$ .
- $t$  cutoff: Select  $t_{\alpha/2}(r)$  so that  $P[T \geq t_{\alpha/2}(r)] = \alpha/2$ , where  $T \sim t(r)$ .
- $\chi^2$  cutoff: Select  $\chi_{\alpha/2}^2(r)$  and  $\chi_{1-\alpha/2}^2(r)$  so that  $P[X \geq \chi_{\alpha/2}^2(r)] = \alpha/2$ ,  $P[X \geq \chi_{1-\alpha/2}^2(r)] = 1 - \alpha/2$  and  $P[X \leq \chi_{1-\alpha/2}^2(r)] = \alpha/2$ , where  $X \sim \chi^2(r)$ .
- $F$  cutoff: Select  $F_{\alpha/2}(r_1, r_2)$  so that  $P[F \geq F_{\alpha/2}(r_1, r_2)] = \alpha/2$ , where  $F \sim F(r_1, r_2)$ .

---

<sup>\*</sup>For enquiry, please email to 1155049861@link.cuhk.edu.hk.

<sup>†</sup>Personal profile: [www.linkedin.com/in/benjamin-chan-chun-ho](http://www.linkedin.com/in/benjamin-chan-chun-ho)

# 1 Introduction

- The objective is to find confidence intervals for the variance of a normal distribution and for the ratio of the variances of two normal distributions. The variance of a normal distribution is sometimes called the scale parameter.

## 1.1 Confidence Interval for a Variance (Under Normality)

- The confidence interval for the variance  $\sigma^2$  is based on the sample variance

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

- From Tutorial 0: Review of Normal Distribution, if  $X_1, X_2, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} N(\mu_X, \sigma_X^2)$ ,

$$\frac{(n-1)S_X^2}{\sigma_X^2} \sim \chi^2(n-1). \quad (1)$$

In fact, it is a pivot mentioned in Tutorial 7: Confidence Intervals for Means since its distribution does not depend on  $\sigma_X^2$ . Hence use it to find a confidence interval for  $\sigma_X^2$ .

---

**Theorem 1.1.** Let  $X_1, X_2, \dots, X_n$  be a random sample from  $N(\mu_X, \sigma_X^2)$ . Assume that  $\sigma_X^2$  is unknown. The random interval

$$\left[ \frac{(n-1)S_X^2}{\chi_{\alpha/2}^2(n-1)}, \frac{(n-1)S_X^2}{\chi_{1-\alpha/2}^2(n-1)} \right]$$

is a  $100(1-\alpha)\%$  confidence interval for the unknown variance  $\sigma_X^2$ .<sup>1</sup>

*Proof.* Under normal distribution, from (1), we aim to select constants  $a$  and  $b$  such that

$$P\left(a \leq \frac{(n-1)S_X^2}{\sigma_X^2} \leq b\right) = 1 - \alpha.$$

One way is to set  $a = \chi_{1-\alpha/2}^2(n-1)$  and  $b = \chi_{\alpha/2}^2(n-1)$ , i.e.

$$P\left(\chi_{1-\alpha/2}^2(n-1) \leq \frac{(n-1)S_X^2}{\sigma_X^2} \leq \chi_{\alpha/2}^2(n-1)\right) = 1 - \alpha.$$

That is, we select  $a$  and  $b$  so that the probabilities in the two tails are equal. Then, solving the inequalities, we have

---

<sup>1</sup>“Probability and Statistical Inference” 8th ed. (Hogg and Tanis) Ch6 p.314

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

Thus, the probability that the random interval

$$\left[ \frac{(n-1)S_X^2}{\chi_{\alpha/2}^2(n-1)}, \frac{(n-1)S_X^2}{\chi_{1-\alpha/2}^2(n-1)} \right]$$

contains the unknown  $\sigma^2$  is  $1 - \alpha$ . □

---

**Corollary 1.2.** Let  $X_1, X_2, \dots, X_n$  be a random sample from  $N(\mu_X, \sigma_X^2)$ . Assume that  $\sigma_X^2$  is unknown. The random interval

$$\left[ \sqrt{\frac{(n-1)S_X^2}{\chi_{\alpha/2}^2(n-1)}}, \sqrt{\frac{(n-1)S_X^2}{\chi_{1-\alpha/2}^2(n-1)}} \right] = \left[ \sqrt{\frac{n-1}{\chi_{\alpha/2}^2(n-1)}} S_X, \sqrt{\frac{n-1}{\chi_{1-\alpha/2}^2(n-1)}} S_X \right]$$

is a  $100(1 - \alpha)\%$  confidence interval for the unknown standard deviation  $\sigma_X$ .

*Proof.* Since the function  $f(x) = \sqrt{x}$  is a strictly increasing function, i.e.

$$0 < a < b \iff 0 < \sqrt{a} < \sqrt{b}.$$

Hence by taking square root of the confidence limits in Theorem 1.1, we can form a confidence interval for  $\sigma_X$ . □

- 
- Once the values of  $X_1, X_2, \dots, X_n$  are observed to be  $x_1, x_2, \dots, x_n$  and  $s_X^2$  is computed. The interval  $[(n-1)s_X^2/\chi_{\alpha/2}^2(n-1), (n-1)s_X^2/\chi_{1-\alpha/2}^2(n-1)]$  is a  $100(1 - \alpha)\%$  confidence interval for  $\sigma_X^2$ .

## 1.2 Confidence Interval for Ratio of Variances (Under Normality)

- There are occasions when it is of interest to compare the variances of two normal distributions. As a result, a confidence interval for  $\sigma_X^2/\sigma_Y^2$  is useful. The idea is to make use of the two sample variances  $S_X^2$  and  $S_Y^2$ .

---

**Definition 1.1.** If  $U \sim \chi^2(r_1)$ ,  $V \sim \chi^2(r_2)$ ,  $U$  and  $V$  are independent, then  $\frac{U/r_1}{V/r_2} \sim F(r_1, r_2)$ .

---

---

**Theorem 1.3.** Let  $X_1, X_2, \dots, X_n$  be a random sample from  $N(\mu_X, \sigma_X^2)$  and  $Y_1, Y_2, \dots, Y_m$  be another random sample from  $N(\mu_Y, \sigma_Y^2)$ , and the two samples are independent. Assume that  $\sigma_X^2$  and  $\sigma_Y^2$  are unknown. The random interval

$$\left[ \frac{1}{F_{\alpha/2}(n-1, m-1)} \frac{S_X^2}{S_Y^2}, F_{\alpha/2}(m-1, n-1) \frac{S_X^2}{S_Y^2} \right]$$

is a  $100(1 - \alpha)\%$  confidence interval for the unknown ratio of variances  $\sigma_X^2/\sigma_Y^2$ .<sup>2</sup>

*Proof.* Under normal distributions, from (1), we know that

$$\frac{(n-1)S_X^2}{\sigma_X^2} \sim \chi^2(n-1) \quad \text{and} \quad \frac{(m-1)S_Y^2}{\sigma_Y^2} \sim \chi^2(m-1).$$

Moreover, since the two samples are independent, the two chi-squared variables are also independent. Note that

$$F \triangleq \frac{\frac{S_Y^2}{\sigma_Y^2}}{\frac{S_X^2}{\sigma_X^2}} = \frac{\left[ \frac{(m-1)S_Y^2}{\sigma_Y^2} \right] / (m-1)}{\left[ \frac{(n-1)S_X^2}{\sigma_X^2} \right] / (n-1)}.$$

By Definition 1.1,  $F \sim F(m-1, n-1)$ . It is a pivot as before since its distribution does not depend on  $\sigma_X^2$  and  $\sigma_Y^2$ . Hence use it to find a confidence interval for  $\sigma_X^2/\sigma_Y^2$ .

Under normal distributions, we aim to select constants  $c$  and  $d$  such that

$$\begin{aligned} 1 - \alpha &= P\left(c \leq \frac{S_Y^2/\sigma_Y^2}{S_X^2/\sigma_X^2} \leq d\right) \\ &= P\left(c \frac{S_X^2}{S_Y^2} \leq \frac{\sigma_X^2}{\sigma_Y^2} \leq d \frac{S_X^2}{S_Y^2}\right). \end{aligned}$$

One way is to set  $c = F_{1-\alpha/2}(m-1, n-1)$  and  $d = F_{\alpha/2}(m-1, n-1)$ . Note that

$$P\left(\frac{U/r_1}{V/r_2} \geq F_\alpha(r_1, r_2)\right) = \alpha \iff P\left(\frac{V/r_2}{U/r_1} \leq \frac{1}{F_\alpha(r_1, r_2)}\right) = \alpha \iff P\left(\frac{V/r_2}{U/r_1} \geq \frac{1}{F_\alpha(r_1, r_2)}\right) = 1 - \alpha.$$

It implies that

$$F_{1-\alpha}(r_2, r_1) = \frac{1}{F_\alpha(r_1, r_2)}.$$

---

<sup>2</sup>“Probability and Statistical Inference” 8th ed. (Hogg and Tanis) Ch6 p.315-316

Because of the limitations of  $F$  table, we generally let

$$c = F_{1-\alpha/2}(m-1, n-1) = \frac{1}{F_{\alpha/2}(n-1, m-1)} \quad \text{and} \quad d = F_{\alpha/2}(m-1, n-1).$$

Therefore, the random interval

$$\left[ \frac{1}{F_{\alpha/2}(n-1, m-1)} \frac{S_X^2}{S_Y^2}, F_{\alpha/2}(m-1, n-1) \frac{S_X^2}{S_Y^2} \right]$$

is a  $100(1-\alpha)\%$  confidence interval for the unknown ratio of variances  $\sigma_X^2/\sigma_Y^2$ .  $\square$

---

**Corollary 1.4.** Let  $X_1, X_2, \dots, X_n$  be a random sample from  $N(\mu_X, \sigma_X^2)$  and  $Y_1, Y_2, \dots, Y_m$  be another random sample from  $N(\mu_Y, \sigma_Y^2)$ , and the two samples are independent. Assume that  $\sigma_X^2$  and  $\sigma_Y^2$  are unknown. The random interval

$$\left[ \sqrt{\frac{1}{F_{\alpha/2}(n-1, m-1)} \frac{S_X^2}{S_Y^2}}, \sqrt{F_{\alpha/2}(m-1, n-1) \frac{S_X^2}{S_Y^2}} \right]$$

is a  $100(1-\alpha)\%$  confidence interval for the unknown ratio of standard deviations  $\sigma_X/\sigma_Y$ .

*Proof.* Since the function  $f(x) = \sqrt{x}$  is a strictly increasing function, i.e.

$$0 < a < b \iff 0 < \sqrt{a} < \sqrt{b}.$$

Hence by taking square root of the confidence limits in Theorem 1.3, we can form a confidence interval for  $\sigma_X/\sigma_Y$ .  $\square$

---

## 2 Choosing the Quantiles

- The probabilities in the two tails are generally selected to be equal in the two-sided confidence interval. In other words, the confidence interval is usually constructed by choosing the  $1-\alpha/2$  and  $\alpha/2$  quantiles.
- Such choice seems arbitrary, but it is actually optimal (minimum length) for the unimodal symmetric distributions like normal or  $t$  distribution. (See Exercise 2.)
- In general it is not optimal for skewed distributions like chi-squared or  $F$  distribution.

## 3 Confidence Intervals for Proportions

### 3.1 Confidence Intervals for $p$

- In general, when observing  $n$  Bernoulli trials with probability of success  $p$  on each trial. Let  $Y$  be the frequency of success out of the  $n$  observations. Under the assumptions of independence and constant probability  $p$ ,  $Y \sim \text{binomial}(n, p)$ . Thus, the problem is to determine the accuracy of the relative frequency  $Y/n$  as an estimator of  $p$ , and to find a confidence interval for  $p$  based on  $Y/n$ .
- Let  $Y \sim \text{binomial}(n, p)$ . From Tutorial 0: Review of Selected Discrete Distributions, we know that  $E(Y) = np$  and  $\text{Var}(Y) = np(1 - p)$ .
- In fact,  $Y/n$  is an unbiased point estimator for  $p$  since

$$E\left(\frac{Y}{n}\right) = \frac{1}{n}E(Y) = \frac{1}{n} \cdot np = p.$$

- By the Central Limit Theorem (CLT),

$$\frac{Y - np}{\sqrt{np(1 - p)}} = \frac{(Y/n) - p}{\sqrt{p(1 - p)/n}} \rightarrow N(0, 1) \quad \text{as } n \rightarrow \infty.$$

Here it has an approximate  $N(0, 1)$ , provided that  $n$  is large enough.

- Then  $100(1 - \alpha)\%$  two-sided confidence interval for  $p$  can be constructed from

$$P\left(-z_{\alpha/2} \leq \frac{(Y/n) - p}{\sqrt{p(1 - p)/n}} \leq z_{\alpha/2}\right) \approx 1 - \alpha. \quad (2)$$

We would then obtain

$$P\left(\frac{Y}{n} - z_{\alpha/2}\sqrt{\frac{p(1 - p)}{n}} \leq p \leq \frac{Y}{n} + z_{\alpha/2}\sqrt{\frac{p(1 - p)}{n}}\right) \approx 1 - \alpha. \quad (3)$$

- Unfortunately, the unknown parameter  $p$  appears in the endpoints of the inequality. There are two ways out of the dilemma.<sup>3</sup>

---

<sup>3</sup>“Probability and Statistical Inference” 8th ed. (Hogg and Tanis) Ch6 p.319-321

- **Method 1:** Replacing  $p$  with  $Y/n$  in  $p(1-p)/n$  in the endpoints in (3). An approximate  $100(1 - \alpha)\%$  confidence interval for  $p$  is

$$\left[ \frac{Y}{n} - z_{\alpha/2} \sqrt{\frac{(Y/n)(1 - Y/n)}{n}}, \frac{Y}{n} + z_{\alpha/2} \sqrt{\frac{(Y/n)(1 - Y/n)}{n}} \right].$$

- **Method 2:** Solving for  $p$  in the inequality in (2), i.e.

$$\frac{|Y/n - p|}{\sqrt{p(1 - p)/n}} \leq z_{\alpha/2}$$

or by taking square on both sides,

$$\frac{(Y/n - p)^2}{p(1 - p)/n} \leq z_{\alpha/2}^2 \iff H(p) = \left( \frac{Y}{n} - p \right)^2 - \frac{z_{\alpha/2}^2 p(1 - p)}{n} \leq 0.$$

Expanding the square and collecting the terms, we get

$$H(p) = \left( 1 + \frac{z_{\alpha/2}^2}{n} \right) p^2 - \left( 2 \frac{Y}{n} + \frac{z_{\alpha/2}^2}{n} \right) p + \left( \frac{Y}{n} \right)^2.$$

Note that  $H(p)$  is a quadratic expression in  $p$ . Thus, we can find those values of  $p$  for which  $H(p) \leq 0$  by finding the two zeros of  $H(p)$ . By the quadratic formula, the zeros of  $H(p)$  are, after simplification and letting  $\hat{p} = Y/n$ ,

$$\frac{\hat{p} + z_{\alpha/2}^2/(2n) \pm z_{\alpha/2} \sqrt{\hat{p}(1 - \hat{p})/n + z_{\alpha/2}^2/(4n^2)}}{1 + z_{\alpha/2}^2/n}.$$

An approximate  $100(1 - \alpha)\%$  confidence interval for  $p$  is

$$\left[ \frac{\hat{p} + z_{\alpha/2}^2/(2n) - z_{\alpha/2} \sqrt{\hat{p}(1 - \hat{p})/n + z_{\alpha/2}^2/(4n^2)}}{1 + z_{\alpha/2}^2/n}, \frac{\hat{p} + z_{\alpha/2}^2/(2n) + z_{\alpha/2} \sqrt{\hat{p}(1 - \hat{p})/n + z_{\alpha/2}^2/(4n^2)}}{1 + z_{\alpha/2}^2/n} \right],$$

where

$$\hat{p} \triangleq \frac{Y}{n}.$$

- If  $n$  is large,  $z_{\alpha/2}^2/(2n)$ ,  $z_{\alpha/2}^2/(4n^2)$  and  $z_{\alpha/2}^2/n$  are small. Thus, the two confidence intervals are approximately equal when  $n$  is large.

### 3.2 One-sided Confidence Intervals for $p$

- One-sided confidence intervals are sometimes appropriate for  $p$ . For example, we may be interested in an upper bound on the proportion of defectives in manufacturing some item. Or we may be interested in a lower bound on the proportion of voters who favor a particular candidate.<sup>4</sup>
- A one-sided confidence interval for  $p$  is

$$\left[0, \frac{Y}{n} + z_\alpha \sqrt{\frac{(Y/n)(1 - Y/n)}{n}}\right],$$

which provides an upper bound for  $p$ .

- Another one-sided confidence interval for  $p$  is

$$\left[\frac{Y}{n} - z_\alpha \sqrt{\frac{(Y/n)(1 - Y/n)}{n}}, 1\right],$$

which provides a lower bound for  $p$ .

### 3.3 Confidence Intervals for Proportion Difference $p_1 - p_2$

- Let  $Y_1 \sim \text{binomial}(n_1, p_1)$  and  $Y_2 \sim \text{binomial}(n_2, p_2)$ . Assume that  $Y_1$  and  $Y_2$  are independent.
- Since the independent random variables  $Y_1/n_1$  and  $Y_2/n_2$  have respective mean  $p_1$  and  $p_2$  and variances  $p_1(1 - p_1)/n_1$  and  $p_2(1 - p_2)/n_2$ . The difference  $Y_1/n_1 - Y_2/n_2$  has mean and variance as

$$E\left(\frac{Y_1}{n_1} - \frac{Y_2}{n_2}\right) = p_1 - p_2$$

and

$$Var\left(\frac{Y_1}{n_1} - \frac{Y_2}{n_2}\right) = \frac{p_1(1 - p_1)}{n_1} + \frac{p_2(1 - p_2)}{n_2}.$$

Recall that the variances are added to get the variance of a difference of two independent random variables.

- Moreover, the fact that  $Y_1/n_1$  and  $Y_2/n_2$  have approximate normal distributions would suggest that the difference  $Y_1/n_1 - Y_2/n_2$  would have an approximate normal distribution.

---

<sup>4</sup>“Probability and Statistical Inference” 8th ed. (Hogg and Tanis) Ch6 p.322



- By the Central Limit Theorem (CLT),

$$\frac{(Y_1/n_1 - Y_2/n_2) - (p_1 - p_2)}{\sqrt{p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2}} \rightarrow N(0, 1) \quad \text{as } n_1, n_2 \rightarrow \infty.$$

Here it has an approximate  $N(0, 1)$ , provided that  $n_1$  and  $n_2$  are large enough.

- For large enough  $n_1$  and  $n_2$ , we replace  $p_1$  and  $p_2$  in the above denominator by  $Y_1/n_1$  and  $Y_2/n_2$ , respectively to have

$$P\left[-z_{\alpha/2} \leq \frac{(Y_1/n_1 - Y_2/n_2) - (p_1 - p_2)}{\sqrt{(Y_1/n_1)(1-Y_1/n_1)/n_1 + (Y_2/n_2)(1-Y_2/n_2)/n_2}} \leq z_{\alpha/2}\right] \approx 1 - \alpha.$$

- An approximate  $100(1 - \alpha)\%$  confidence interval for  $p_1 - p_2$  is

$$\left[\frac{Y_1}{n_1} - \frac{Y_2}{n_2} - z_{\alpha/2}S, \frac{Y_1}{n_1} - \frac{Y_2}{n_2} + z_{\alpha/2}S\right],$$

where  $S = \sqrt{\frac{(Y_1/n_1)(1-Y_1/n_1)}{n_1} + \frac{(Y_2/n_2)(1-Y_2/n_2)}{n_2}}.$

## 4 Exercises

**Exercise 1.** (a) If  $X_i \sim \text{Gamma}(\alpha_i, \theta)$ ,  $i = 1, 2, \dots, n$  and  $X_i$ 's are independent, what is the distribution of  $\sum_{i=1}^n X_i$ ?

(b) Let  $W_1, \dots, W_n$  be a random sample from  $\exp(\theta) = \text{Gamma}(1, \theta)$ . Then the MLE of the scale parameter  $\theta$  is  $\bar{W}$ . Find the distribution of  $\bar{W}$ . Hence construct a  $100(1 - \alpha)\%$  confidence interval for  $\theta$  based on the quantiles of a chi-squared distribution.

(c) Construct a confidence interval for  $\theta_1/\theta_2$  if there are two independent exponential samples.

---

**Exercise 2.** Let  $X_1, X_2, \dots, X_n \sim N(\mu, \sigma^2)$  and  $X_i$ 's are independent. Both of the parameters are unknown. Then  $\sqrt{n} \frac{\bar{X} - \mu}{S} \sim t(n-1)$  and we can use its quantiles  $a, b$  which satisfy

$$G(b) - G(a) = \Pr\left\{a \leq \sqrt{n} \frac{\bar{X} - \mu}{S} \leq b\right\} = 1 - \alpha$$

to construct a  $100(1 - \alpha)\%$  confidence interval for  $\mu$  where  $G$  is the CDF of  $t(n-1)$ . The pdf of  $t(n-1)$  is given by  $g(t)$ .

- (a) Construct the confidence interval and express its length  $k$  in terms of  $n, S, a, b$ .
- (b) Find the pair of  $(a, b)$  that minimizes the length. (Hints: use Lagrange optimization.)

---

**Exercise 3.** A proportion  $p$  that many public opinion polls estimate is the number of Americans who would say yes to the question, “If something were to happen to the President of the United States, do you think that the Vice President would be qualified to take over as President?” In one such random sample of 1022 adults, 388 said yes.

- (a) On the basis of the given data, find a point estimate of  $p$ .
- (b) Find an approximate 90% two-sided confidence interval for  $p$ .