# STAT2006 Basic Concepts in Statistics and Probability II

# Tutorial 7 Confidence Intervals for Mean

Benjamin Chun Ho Chan*†‡§

January 4, 2019

**Abstract**

It aims to introduce concepts of confidence intervals for mean and difference of two means. 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* written by Hogg and Tanis and used in STAT2001/2006, and *Statistical Inference* by Casella and Berger and used in STAT4003.

# 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.
- Parameter(s): $\theta$ denotes population characteristic(s) that can be set to different values to produce different probability distributions.
- 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
- 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 $T \sim t(r)$.

*For enquiry, please email to 1155049861@link.cuhk.edu.hk.
†Personal profile: www.linkedin.com/in/benjamin-chan-chun-ho
‡GitHub repository: https://github.com/BenjaminChanChunHo/CUHK-STAT-or-RMSC-Tutorial-Note
§RPubs: http://rpubs.com/Benjamin_Chan_Chun_Ho

# 1 Introduction of Interval Estimation

- In Tutorial 5: Point Estimation, point estimation of a parameter $\theta$ is a guess of a single value as the value of $\theta$. If $\theta$ is real-valued, interval estimation can be discussed.

---

**Definition 1.1.** An **interval estimate** of a real-valued parameter $\theta$ is any pair of functions, $L(x_1, \ldots, x_n)$ and $U(x_1, \ldots, x_n)$, of a sample that satisfy $L(\boldsymbol{x}) \leq U(\boldsymbol{x})$ for all $\boldsymbol{x} \in \mathcal{X}$. If $\boldsymbol{X} = \boldsymbol{x}$ is observed, the inference $L(\boldsymbol{x}) \leq \theta \leq U(\boldsymbol{x})$ is made. The random interval $[L(\boldsymbol{X}), U(\boldsymbol{X})]$ is called an **interval estimator**.[1]

---

- Here $\boldsymbol{X} = (X_1, \ldots, X_n)$ denotes the random sample and $\boldsymbol{x} = (x_1, \ldots, x_n)$ denotes the realized values.
- We can also consider one-sided interval estimates. For instance, if $L(\boldsymbol{x}) = -\infty$, then we have the one-sided interval $(-\infty, U(\boldsymbol{x})]$ and the assertion is that "$\theta \leq U(\boldsymbol{x})$" with no mention of a lower bound. We could similarly take $U(\boldsymbol{x}) = \infty$ and have a one-sided interval $[L(\boldsymbol{x}), \infty)$.
- Although Definition 1.1 mentions a closed interval $[L(\boldsymbol{x}), U(\boldsymbol{x})]$, it will sometimes be more natural to use an open interval $(L(\boldsymbol{x}), U(\boldsymbol{x}))$ or even a half-open and half-closed interval.
- The rationale behind interval estimation is that by giving up some precision in our estimate of $\theta$, we have gained some confidence, or assurance, that our assertion is correct. The purpose of using an interval estimator rather than a point estimator is to have some guarantee of capturing the parameter of interest.

---

**Definition 1.2.** For an interval estimator $[L(\boldsymbol{X}), U(\boldsymbol{X})]$ of a parameter $\theta$, the **coverage probability** of $[L(\boldsymbol{X}), U(\boldsymbol{X})]$ is the probability that the random interval $[L(\boldsymbol{X}), U(\boldsymbol{X})]$ covers the true parameter, $\theta$. In symbols, it is denoted by $P_\theta(\theta \in [L(\boldsymbol{X}), U(\boldsymbol{X})])$.[2]

---

**Definition 1.3.** For an interval estimator $[L(\boldsymbol{X}), U(\boldsymbol{X})]$ of a parameter $\theta$, the **confidence coefficient** of $[L(\boldsymbol{X}), U(\boldsymbol{X})]$ is the infimum of the coverage probabilities $\inf_\theta P_\theta(\theta \in [L(\boldsymbol{X}), U(\boldsymbol{X})])$.

---

- It is important to keep in mind that the interval is the random quantity, not the parameter (from a frequentist point of view). The probability statements $P_\theta(\theta \in [L(\boldsymbol{X}), U(\boldsymbol{X})])$ refer to $\boldsymbol{X}$, not $\theta$. Alternatively we can write $P_\theta(L(\boldsymbol{X}) \leq \theta, U(\boldsymbol{X}) \geq \theta)$, a statement about a random $\boldsymbol{X}$.

[1] "Statistical Inference" 2nd ed. (Casella and Berger) Ch9 p.417
[2] "Statistical Inference" 2nd ed. (Casella and Berger) Ch9 p.418

# 2 Confidence Interval for Means

- To consider the closeness of $\bar{X}$ (the unbiased estimator of $\mu$) to the unknown mean $\mu$, we need to use the distribution of $\bar{X}$ to construct a confidence interval for the unknown parameter $\mu$.

## 2.1 When $\sigma^2$ is Known under Normal Assumption

**Theorem 2.1.** Let $X_1, X_2, \ldots, X_n$ be a random sample from $N(\mu, \sigma^2)$. Assume that $\mu$ is unknown and $\sigma^2$ is known. The random interval

$$\left[\bar{X} - z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right), \bar{X} + z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)\right]$$

is a $100(1-\alpha)\%$ confidence interval for the unknown mean $\mu$.[3]

*Proof.* First of all, note that $\bar{X} \sim N(\mu, \sigma^2/n)$ and hence

$$P\left(-z_{\alpha/2} \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}\right) = 1 - \alpha.$$

The following inequalities are equivalent:

$$\begin{aligned}
-z_{\alpha/2} &\leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}, \\
-z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) &\leq \bar{X} - \mu \leq z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right), \\
-\bar{X} - z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) &\leq \quad -\mu \quad \leq -\bar{X} + z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right), \\
\bar{X} + z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) &\geq \quad\quad \mu \quad\quad \geq \bar{X} - z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right).
\end{aligned}$$

Thus, since the probability of the first of these is $1-\alpha$, the probability of the last must also be $1-\alpha$, because the latter is true if and only if the former is true. That is, we have

$$P\left[\bar{X} - z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) \leq \mu \leq \bar{X} + z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)\right] = 1 - \alpha.$$

So the probability that the random interval

$$\left[\bar{X} - z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right), \bar{X} + z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)\right]$$

includes the unknown mean $\mu$ is $1-\alpha$. □

- The number $100(1-\alpha)\%$, or equivalently, $1-\alpha$, is called the confidence coefficient.
- Once the sample is observed and the sample mean computed to equal $\bar{x}$, the interval $[\bar{x} - z_{\alpha/2}(\sigma/\sqrt{n}), \bar{x} + z_{\alpha/2}(\sigma/\sqrt{n})]$ becomes known. For a particular sample, the interval either does or does not contain the mean $\mu$.

[3] "Probability and Statistical Inference" 8th ed. (Hogg and Tanis) Ch6 p.295-296

- If many such intervals are calculated, the probability that the random interval covers $\mu$ is $1-\alpha$. In other words, the probability is only applicable before the sample is drawn.
- The confidence interval for $\mu$ is centered at the point estimate $\bar{X}$. Note that as $n$ increases, $z_{\alpha/2}(\sigma/\sqrt{n})$ decreases, resulting in a shorter confidence interval with the same confidence coefficient $1-\alpha$. A shorter confidence interval indicates that we have more credence in $\bar{X}$ as an estimate of $\mu$.

## 2.2 When $\sigma^2$ is Known under Any Distribution (Large n)

- When we cannot assume that the sample comes from normal distribution, provided that sample size $n$ is large enough, we can use the central limit theorem to obtain an approximate confidence interval for $\mu$.

---

**Theorem 2.2.** Let $X_1, X_2, \ldots, X_n$ be a random sample without any distributional assumption. Assume that $\mu$ is unknown and $\sigma^2$ is known. If $n$ is large enough, the random interval

$$\left[\bar{X} - z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right), \bar{X} + z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)\right]$$

is an approximate $100(1-\alpha)\%$ confidence interval for the unknown mean $\mu$.[4]

*Proof.* By the central limit theorem, provided that $n$ is large enough, the ratio $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ has the approximate $N(0,1)$ even when the underlying distribution is not normal. Note that

$$P\left(-z_{\alpha/2} \leq \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}\right) \approx 1-\alpha,$$

and hence the probability that the random interval

$$\left[\bar{X} - z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right), \bar{X} + z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)\right]$$

includes the unknown mean $\mu$ is approximately $1-\alpha$. □

---

- The closeness of the approximate probability $1-\alpha$ to the exact probability depends on both the underlying distribution and the sample size. An $n$ of at least 30 is usually adequate.

## 2.3 When $\sigma^2$ is Unknown under Any Distribution (Large n)

- If $\sigma^2$ is unknown and the sample size $n$ is large (like 30 or greater), we shall use the fact that the ratio $\frac{\bar{X}-\mu}{S/\sqrt{n}}$ has an approximate $N(0,1)$. It is true whether or not the underlying distribution is normal.

---

**Corollary 2.3.** Let $X_1, X_2, \ldots, X_n$ be a random sample without any distributional assumption. Assume that $\mu$ is unknown and $\sigma^2$ is unknown. If $n$ is large enough, the random interval

$$\left[\bar{X} - z_{\alpha/2}\left(\frac{S}{\sqrt{n}}\right), \bar{X} + z_{\alpha/2}\left(\frac{S}{\sqrt{n}}\right)\right]$$

is an approximate $100(1-\alpha)\%$ confidence interval for the unknown mean $\mu$.

---

[4] "Probability and Statistical Inference" 8th ed. (Hogg and Tanis) Ch6 p.297-298

## 2.4 When $\sigma^2$ is Unknown under Normal Assumption (Small n)

- In many applications, the sample sizes are small and we do not know the value of $\sigma^2$.
- If the random sample arises from a normal distribution, we use the fact that

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

has a $t$ distribution with $r = n - 1$ degrees of freedom, where $S^2$ is the sample variance.

---

**Theorem 2.4.** Let $X_1, X_2, \ldots, X_n$ be a random sample from $N(\mu, \sigma^2)$. Assume that $\mu$ is unknown and $\sigma^2$ is unknown. The random interval

$$\left[\bar{X} - t_{\alpha/2}(n-1)\left(\frac{S}{\sqrt{n}}\right), \bar{X} + t_{\alpha/2}(n-1)\left(\frac{S}{\sqrt{n}}\right)\right]$$

is a $100(1-\alpha)\%$ confidence interval for $\mu$.[5]

*Proof.* First of all, note that

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t(n-1)$$

and hence

$$\begin{aligned}
1 - \alpha &= P\left[-t_{\alpha/2}(n-1) \leq \frac{\bar{X} - \mu}{S/\sqrt{n}} \leq t_{\alpha/2}(n-1)\right] \\
&= P\left[-t_{\alpha/2}(n-1)\left(\frac{S}{\sqrt{n}}\right) \leq \bar{X} - \mu \leq t_{\alpha/2}(n-1)\left(\frac{S}{\sqrt{n}}\right)\right] \\
&= P\left[-\bar{X} - t_{\alpha/2}(n-1)\left(\frac{S}{\sqrt{n}}\right) \leq -\mu \leq -\bar{X} + t_{\alpha/2}(n-1)\left(\frac{S}{\sqrt{n}}\right)\right] \\
&= P\left[\bar{X} + t_{\alpha/2}(n-1)\left(\frac{S}{\sqrt{n}}\right) \geq \mu \geq \bar{X} - t_{\alpha/2}(n-1)\left(\frac{S}{\sqrt{n}}\right)\right]
\end{aligned}$$

The random sample provides $\bar{X}$ and $S^2$. Therefore, the random interval

$$\left[\bar{X} - t_{\alpha/2}(n-1)\left(\frac{S}{\sqrt{n}}\right), \bar{X} + t_{\alpha/2}(n-1)\left(\frac{S}{\sqrt{n}}\right)\right]$$

is a $100(1-\alpha)\%$ confidence interval for $\mu$. □

---

- Note that the $t$-distribution is symmetric and bell-shaped, like the normal distribution, but has heavier tails. Thus $t_{\alpha/2}(n-1) > z_{\alpha/2}$. Consequently, we would "expect" the interval $\bar{x} \pm z_{\alpha/2}(\sigma/\sqrt{n})$ to be shorter than the interval $\bar{x} \pm t_{\alpha/2}(n-1)(s/\sqrt{n})$. After all, we have more information, namely, the value of $\sigma$, in constructing the first interval. However, the length of the second interval is very much dependent on the value of $s$. If the observed $s$ is smaller than $\sigma$, a shorter confidence interval could result by the second procedure. But on the average, $\bar{x} \pm z_{\alpha/2}(\sigma/\sqrt{n})$ is the shorter of the two confidence intervals.

[5] "Probability and Statistical Inference" 8th ed. (Hogg and Tanis) Ch6 p.299

## 2.5 When $\sigma^2$ is Unknown under Any Distribution (Small n)

- If we are not able to assume that the underlying distribution is normal, but $\mu$ and $\sigma^2$ are both unknown, the ratio $\frac{\bar{X}-\mu}{S/\sqrt{n}}$ has an approximate $t(n-1)$.

---

**Corollary 2.5.** Let $X_1, X_2, \ldots, X_n$ be a random sample without any distributional assumption. Assume that $\mu$ is unknown and $\sigma^2$ is unknown. If $n$ is not large enough, the random interval

$$\left[\bar{X} - t_{\alpha/2}(n-1)\left(\frac{S}{\sqrt{n}}\right), \bar{X} + t_{\alpha/2}(n-1)\left(\frac{S}{\sqrt{n}}\right)\right]$$

is an approximate $100(1-\alpha)\%$ confidence interval for the unknown mean $\mu$.

---

- Generally, the approximation is quite good for many nonnormal distributions. In particular, it works well if the underlying distribution is symmetric, unimodal, and of the continuous type. However, if the distribution is highly skewed, there is great danger in using that approximation.

## 2.6 One-sided Confidence Interval for Means

- In some cases, we want only a lower (or upper) bound on $\mu$. Say $\bar{X}$ is the mean of a random sample of size $n$ from $N(\mu, \sigma^2)$. Here assume that $\sigma^2$ is known. Then

$$P\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \leq z_\alpha\right) = 1-\alpha,$$

  or equivalently,

$$P\left[\bar{X} - z_\alpha\left(\frac{\sigma}{\sqrt{n}}\right) \leq \mu\right] = 1-\alpha.$$

- Once $\bar{X}$ is observed to be equal to $\bar{x}$, it follows that

$$\left[\bar{x} - z_\alpha\frac{\sigma}{\sqrt{n}}, \infty\right)$$

  is a $100(1-\alpha)\%$ one-sided confidence interval for $\mu$. That is, with the confidence coefficient $1-\alpha$, $\bar{x} - z_\alpha(\sigma/\sqrt{n})$ is a lower bound for $\mu$.

- Similarly,

$$\left(-\infty, \bar{x} + z_\alpha\frac{\sigma}{\sqrt{n}}\right]$$

  is a one-sided confidence interval for $\mu$ and $\bar{x} + z_\alpha(\sigma/\sqrt{n})$ provides an upper bound for $\mu$ with confidence coefficient $1-\alpha$.

# 3 (Optional) Pivotal Quantities

- A random variable whose distribution does not depend on the parameter is a quantity known as a pivotal quantity, or pivot.

---

**Definition 3.1.** A random variable $Q(\boldsymbol{X}, \theta) = Q(X_1, \ldots, X_n, \theta)$ is a pivotal quantity (or pivot) if the distribution of $Q(\boldsymbol{X}, \theta)$ is independent of all parameters. That is, if $\boldsymbol{X} \sim F(\boldsymbol{x}; \theta)$, then $Q(\boldsymbol{X}, \theta)$ has the same distribution for all values of $\theta$.[6]

---

- The function $Q(\boldsymbol{X}, \theta)$ will usually explicitly contain both parameters and statistics, but for any $L(\boldsymbol{X})$ and $U(\boldsymbol{X})$, $P_\theta(L(\boldsymbol{X}) \leq Q(\boldsymbol{X}, \theta) \leq U(\boldsymbol{X}))$ cannot depend on $\theta$. The technique of constructing confidence intervals from pivots relies on being able to find a pivot, and $L(\boldsymbol{X})$ and $U(\boldsymbol{X})$ so that $[L(\boldsymbol{X}), U(\boldsymbol{X})]$ is a confidence interval.

- You may not realize that we have been using this strategy in deriving the above confidence intervals. Notice that, in particular, if $X_1, \ldots, X_n$ is a random sample from $N(\mu, \sigma^2)$, then the $t$ statistic
$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t(n-1)$$
is a pivot because the $t$ distribution does not depend on the parameters $\mu$ and $\sigma^2$. Then
$$P\bigg[ - t_{\alpha/2}(n-1) \leq \frac{\bar{X} - \mu}{S/\sqrt{n}} \leq t_{\alpha/2}(n-1)\bigg]$$
does not depend on $\mu$ and $\sigma^2$. Hence
$$\left[\bar{X} - t_{\alpha/2}(n-1)\left(\frac{S}{\sqrt{n}}\right), \bar{X} + t_{\alpha/2}(n-1)\left(\frac{S}{\sqrt{n}}\right)\right]$$
is a confidence interval for $\mu$.

- If $\sigma^2$ is known, another pivot is
$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1).$$

[6] "Statistical Inference" 2nd ed. (Casella and Berger) Ch9 p.427-428

# 4 Confidence Interval for Difference of Two Means

## 4.1 When $\sigma_X^2$ and $\sigma_Y^2$ are Known under Normality

**Lemma 4.1.** Let $X_1, X_2, \ldots, X_n$ be a random sample from $N(\mu_X, \sigma_X^2)$ and $Y_1, Y_2, \ldots, Y_m$ be another random sample from $N(\mu_Y, \sigma_Y^2)$, and the two samples are independent. Then $\bar{X} \sim N\left(\mu_X, \frac{\sigma_X^2}{n}\right)$ and $\bar{Y} \sim N\left(\mu_Y, \frac{\sigma_Y^2}{m}\right)$. Moreover, $\bar{X} - \bar{Y} \sim N\left(\mu_X - \mu_Y, \frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}\right)$.

*Proof.* First of all, the distributions of $\bar{X}$ and $\bar{Y}$ are trivial. Please refer to other tutorials.

Since the random samples are independent, the respective sample means $\bar{X}$ and $\bar{Y}$ are also independent and hence $Cov(\bar{X}, \bar{Y}) = 0$. Note that the sum (or difference) of independent normal random variables also follows a normal distribution. Then

$$E(\bar{X} - \bar{Y}) = E(\bar{X}) - E(\bar{Y}) = E(X) - E(Y) = \mu_X - \mu_Y$$

and by using $Cov(\bar{X}, \bar{Y}) = 0$, we have

$$Var(\bar{X} - \bar{Y}) = Var(\bar{X}) + Var(\bar{Y}) - 2Cov(\bar{X}, \bar{Y}) = Var(\bar{X}) + Var(\bar{Y}) = \frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}.$$

Therefore, we conclude that

$$\bar{X} - \bar{Y} \sim N\left(\mu_X - \mu_Y, \frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}\right).$$

□

**Corollary 4.2.** Continue with the same setting in Lemma 4.1, we have

$$\frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}} \sim N(0, 1).$$

**Theorem 4.3.** Let $X_1, X_2, \ldots, X_n$ be a random sample from $N(\mu_X, \sigma_X^2)$ and $Y_1, Y_2, \ldots, Y_m$ be another random sample from $N(\mu_Y, \sigma_Y^2)$, and the two samples are independent. Assume that $\mu_X - \mu_Y$ is unknown while $\sigma_X^2$ and $\sigma_Y^2$ are known. The random interval

$$\left[\bar{X} - \bar{Y} - z_{\alpha/2}\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}, \bar{X} - \bar{Y} + z_{\alpha/2}\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}\right]$$

is a $100(1-\alpha)\%$ confidence interval for $\mu_X - \mu_Y$.[7]

*Proof.* By Corollary 4.2, we have

$$\frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{\sqrt{\sigma_X^2/n + \sigma_Y^2/m}} \sim N(0, 1).$$

Hence

$$P\left(-z_{\alpha/2} \leq \frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{\sqrt{\sigma_X^2/n + \sigma_Y^2/m}} \leq z_{\alpha/2}\right) = 1 - \alpha,$$

[7] "Probability and Statistical Inference" 8th ed. (Hogg and Tanis) Ch6 p.303-304

which can be rewritten as

$$P[(\bar{X}-\bar{Y})-z_{\alpha/2}\sigma_W \le \mu_X-\mu_Y \le (\bar{X}-\bar{Y})+z_{\alpha/2}\sigma_W],$$

where

$$\sigma_W=\sqrt{\frac{\sigma_X^2}{n}+\frac{\sigma_Y^2}{m}}$$

is the standard deviation of $\bar{X}-\bar{Y}$. □

---

- Once the experiments have been performed and the means $\bar{x}$ and $\bar{y}$ computed, the interval

$$\left[\bar{x}-\bar{y}-z_{\alpha/2}\sqrt{\frac{\sigma_X^2}{n}+\frac{\sigma_Y^2}{m}},\bar{x}-\bar{y}+z_{\alpha/2}\sqrt{\frac{\sigma_X^2}{n}+\frac{\sigma_Y^2}{m}}\right]$$

  provides a $100(1-\alpha)\%$ confidence interval for $\mu_X-\mu_Y$.

- Note that the interval is centered at the point estimate $\bar{x}-\bar{y}$ of $\mu_X-\mu_Y$.

## 4.2 When $\sigma_X^2$ and $\sigma_Y^2$ are Unknown under Normality (Large n)

- If the sample sizes are large and $\sigma_X^2$ and $\sigma_Y^2$ are unknown, we can replace $\sigma_X^2$ and $\sigma_Y^2$ with $S_X^2$ and $S_Y^2$, where $S_X^2$ and $S_Y^2$ are the respective sample variances.

- The distribution of

$$\frac{(\bar{X}-\bar{Y})-(\mu_X-\mu_Y)}{\sqrt{S_X^2/n+S_Y^2/m}}$$

  has an approximate $N(0,1)$.

---

**Corollary 4.4.** Let $X_1,X_2,\ldots,X_n$ be a random sample from $N(\mu_X,\sigma_X^2)$ and $Y_1,Y_2,\ldots,Y_m$ be another random sample from $N(\mu_Y,\sigma_Y^2)$, and the two samples are independent. Assume that $\mu_X-\mu_Y$ is unknown while $\sigma_X^2$ and $\sigma_Y^2$ are unknown. If $n$ and $m$ are large enough, the random interval

$$\left[\bar{X}-\bar{Y}-z_{\alpha/2}\sqrt{\frac{S_X^2}{n}+\frac{S_Y^2}{m}},\bar{X}-\bar{Y}+z_{\alpha/2}\sqrt{\frac{S_X^2}{n}+\frac{S_Y^2}{m}}\right]$$

is an approximate $100(1-\alpha)\%$ confidence interval for $\mu_X-\mu_Y$.

---

## 4.3 When $\sigma_X^2=\sigma_Y^2=\sigma^2$ is Known under Normality

---

**Corollary 4.5.** Let $X_1,X_2,\ldots,X_n$ be a random sample from $N(\mu_X,\sigma_X^2)$ and $Y_1,Y_2,\ldots,Y_m$ be another random sample from $N(\mu_Y,\sigma_Y^2)$, and the two samples are independent. Assume that $\mu_X-\mu_Y$ is unknown while $\sigma_X^2=\sigma_Y^2=\sigma^2$ is known. The random interval

$$\left[\bar{X}-\bar{Y}-z_{\alpha/2}\sigma\sqrt{\frac{1}{n}+\frac{1}{m}},\bar{X}-\bar{Y}+z_{\alpha/2}\sigma\sqrt{\frac{1}{n}+\frac{1}{m}}\right]$$

is a $100(1-\alpha)\%$ confidence interval for $\mu_X-\mu_Y$.

*Proof.* Replace $\sigma_X^2$ and $\sigma_Y^2$ by $\sigma^2$ in Theorem 4.3. □

---

## 4.4 When $\sigma_X^2 = \sigma_Y^2 = \sigma^2$ is Unknown under Normality (Small n)

- The problem of constructing confidence intervals for the difference of the means of two normal distributions, when $\sigma_X^2$ and $\sigma_Y^2$ are unknown but the sample sizes are small, is difficult.
- If we can assume common, but unknown, variances (say, $\sigma_X^2 = \sigma_Y^2 = \sigma^2$), there is a way out of difficulty.

---

**Definition 4.1.** If $Z \sim N(0,1)$, $W \sim \chi^2(r)$, $Z$ and $W$ are independent, then $\dfrac{Z}{\sqrt{W/r}} \sim t(r)$.

---

**Theorem 4.6.** Let $X_1, X_2, \ldots, X_n$ be a random sample from $N(\mu_X, \sigma_X^2)$ and $Y_1, Y_2, \ldots, Y_m$ be another random sample from $N(\mu_Y, \sigma_Y^2)$, and the two samples are independent. Assume that $\mu_X - \mu_Y$ is unknown while $\sigma_X^2 = \sigma_Y^2 = \sigma^2$ are unknown. If $n$ and $m$ are not large enough, the random interval

$$\left[\bar{X} - \bar{Y} - t_{\alpha/2}(n+m-2)S_p\sqrt{\frac{1}{n}+\frac{1}{m}}, \bar{X} - \bar{Y} + t_{\alpha/2}(n+m-2)S_p\sqrt{\frac{1}{n}+\frac{1}{m}}\right]$$

is a $100(1-\alpha)\%$ confidence interval for $\mu_X - \mu_Y$, where

$$S_p^2 \triangleq \frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2}$$

is the pooled sample variance.[8]

*Proof.* By Corollary 4.2, we know that

$$Z \triangleq \frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{\sqrt{\sigma^2/n + \sigma^2/m}} \sim N(0,1).$$

Moreover, using the results of Tutorial 0: Review of Normal Distribution,

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

Since the random samples are independent, the two chi-squared random variables are also independent. By the results of Tutorial 0: Review of Normal Distribution again, the sum of independent chi-squared random variables follows a chi-squared distribution. Thus

$$U \triangleq \frac{(n-1)S_X^2}{\sigma^2} + \frac{(m-1)S_Y^2}{\sigma^2} \sim \chi^2(n+m-2).$$

For normal distribution, the sample means and sample variances are independent. It implies that $Z$ and $U$ are independent. According to Definition 4.1,

$$T \triangleq \frac{Z}{\sqrt{U/(n+m-2)}} \sim t(n+m-2).$$

[8] "Probability and Statistical Inference" 8th ed. (Hogg and Tanis) Ch6 p.304-305

That is

$$T = \frac{\dfrac{(\bar{X}-\bar{Y})-(\mu_X-\mu_Y)}{\sqrt{\sigma^2/n+\sigma^2/m}}}{\sqrt{\left[\dfrac{(n-1)S_X^2}{\sigma^2}+\dfrac{(m-1)S_Y^2}{\sigma^2}\right]\Big/(n+m-2)}}$$
$$= \frac{(\bar{X}-\bar{Y})-(\mu_X-\mu_Y)}{\sqrt{\left[\dfrac{(n-1)S_X^2+(m-1)S_Y^2}{n+m-2}\right]\left[\dfrac{1}{n}+\dfrac{1}{m}\right]}}$$
$$= \frac{(\bar{X}-\bar{Y})-(\mu_X-\mu_Y)}{S_p\sqrt{\dfrac{1}{n}+\dfrac{1}{m}}}.$$

Therefore, we have

$$P[-t_{\alpha/2}(n+m-2) \le T \le t_{\alpha/2}(n+m-2)] = 1-\alpha.$$

Solving the inequalities for $\mu_X - \mu_Y$ yields

$$P\left(\bar{X}-\bar{Y}-t_{\alpha/2}(n+m-2)S_p\sqrt{\frac{1}{n}+\frac{1}{m}} \le \mu_X-\mu_Y \le \bar{X}-\bar{Y}+t_{\alpha/2}(n+m-2)S_p\sqrt{\frac{1}{n}+\frac{1}{m}}\right),$$

where $S_p^2$ is the pooled estimator of the common variance. Therefore, the random interval

$$\left[\bar{X}-\bar{Y}-t_{\alpha/2}(n+m-2)S_p\sqrt{\frac{1}{n}+\frac{1}{m}}, \bar{X}-\bar{Y}+t_{\alpha/2}(n+m-2)S_p\sqrt{\frac{1}{n}+\frac{1}{m}}\right]$$

is a $100(1-\alpha)\%$ confidence interval for $\mu_X - \mu_Y$. □

---

- It is interesting to consider the two-sample $T$ in more detail. It is[9]

$$T = \frac{(\bar{X}-\bar{Y})-(\mu_X-\mu_Y)}{\sqrt{\left[\dfrac{(n-1)S_X^2+(m-1)S_Y^2}{n+m-2}\right]\left[\dfrac{1}{n}+\dfrac{1}{m}\right]}}$$
$$= \frac{(\bar{X}-\bar{Y})-(\mu_X-\mu_Y)}{\sqrt{\left[\dfrac{(n-1)S_X^2}{nm}+\dfrac{(m-1)S_Y^2}{nm}\right]\left[\dfrac{n+m}{n+m-2}\right]}}.$$

  Since $(n-1)/n \approx 1$, $(m-1)/m \approx 1$, and $(n+m)/(n+m-2) \approx 1$, we have

$$T \approx \frac{(\bar{X}-\bar{Y})-(\mu_X-\mu_Y)}{\sqrt{\dfrac{S_X^2}{m}+\dfrac{S_Y^2}{n}}}.$$

  Note that, in this form, each variance is divided by the wrong sample size! That is, if the sample sizes are large or the variances known, we would like

$$\sqrt{\frac{S_X^2}{n}+\frac{S_Y^2}{m}} \quad \text{or} \quad \sqrt{\frac{\sigma_X^2}{n}+\frac{\sigma_Y^2}{m}}$$

  in the denominator. Thus, using $T$ is particularly bad when the sample sizes and the variances are unequal; hence, caution must be taken in using that $T$ to construct a confidence interval for $\mu_X - \mu_Y$.

[9] "Probability and Statistical Inference" 8th ed. (Hogg and Tanis) Ch6 p.307-308

## 4.5 When X and Y are Dependent

- In some applications, two measurements – say, $X$ and $Y$ – are taken on the same subject. In these cases, $X$ and $Y$ are dependent random variables. Many times these are "before" and "after" measurements, such as weight before and after participating in a diet-and-exercise program.[10]
- To compare the means of $X$ and $Y$, it is not permissible to use the $t$ statistics and confidence intervals that we just developed, because in that situation $X$ and $Y$ are independent.
- If the two samples are dependent pairs, but the difference of the pair $X_i - Y_i$ is normally distributed (e.g. when $(X_i, Y_i)$ jointly follows a bivariate normal distribution), then we construct the confidence interval for $\mu_D \triangleq \mu_X - \mu_Y$ in the same way as the 1 sample case.

---

**Theorem 4.7.** Let $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ be $n$ pairs of dependent measurements. Let $D_i = X_i - Y_i$, $i = 1, 2, \ldots, n$. Suppose that $D_1, D_2, \ldots, D_n$ is a random sample from $N(\mu_D, \sigma_D^2)$, where $\mu_D$ and $\sigma_D^2$ are the mean and variance of each difference. Note that

$$T \triangleq \frac{\bar{D} - \mu_D}{S_D/\sqrt{n}} \sim t(n-1),$$

where $\bar{D}$ and $S_D^2$ are, respectively, the sample mean and sample variance of the $n$ differences. A $100(1-\alpha)\%$ confidence interval for $\mu_D = \mu_X - \mu_Y$ is

$$\left[\bar{D} - t_{\alpha/2}(n-1)\left(\frac{S_D}{\sqrt{n}}\right), \bar{D} + t_{\alpha/2}(n-1)\left(\frac{S_D}{\sqrt{n}}\right)\right].$$

---

[10] "Probability and Statistical Inference" 8th ed. (Hogg and Tanis) Ch6 p.309-310

# 5 Exercises

**Exercise 1.** To determine the effect of 100% nitrate on the growth of pea plants, several specimens were planted and then watered with 100% nitrate every day. At the end of two weeks, the plants were measured. Here are the data on seven of them:

$$17.5,\ 14.5,\ 15.2,\ 14.0,\ 17.3,\ 18.0,\ 13.8$$

Assume that these data are observations from a normal distribution $N(\mu, \sigma^2)$.[11]

(a) Find the value of a point estimate of $\mu$.

**Solution.** The sample is $x_1 = 17.5, \ldots, x_7 = 13.8$, the point estimate of $\mu$ is $\bar{x} = 15.757$.

(b) Find the value of a point estimate of $\sigma$.

**Solution.** The point estimate of $\sigma$ is $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} = 1.792$.

(c) Give the endpoints for a 90% confidence interval for $\mu$.

**Solution.** Since $\sigma^2$ is unknown under Normal assumption and $n$ is small (see Section 2.4), the 90% confidence interval should be

$$\left[\bar{x} - t_{\alpha/2}(n-1)\left(\frac{s}{\sqrt{n}}\right), \bar{x} + t_{\alpha/2}(n-1)\left(\frac{s}{\sqrt{n}}\right)\right].$$

In this case, $\bar{x} = 15.757$, $s = 1.792$, $\alpha = 0.1$, $\alpha/2 = 0.05$, $n = 7$, $t_{\alpha/2}(n-1) = t_{0.05}(6) = 1.943$. Hence we have the 90% confidence interval

$$\left[15.757 - 1.943\left(\frac{1.792}{\sqrt{7}}\right), 15.757 + 1.943\left(\frac{1.792}{\sqrt{7}}\right)\right] = [14.441, 17.073].$$

---

**Exercise 2.** Independent random samples of the heights of adults males living in two countries yielded the following results: $n = 12$, $\bar{x} = 65.7$ inches, $s_x = 4$ inches; and $m = 15, \bar{y} = 68.2$ inches, $s_y = 3$ inches. Find an approximate 98% confidence interval for the difference $\mu_X - \mu_Y$ of the means of the populations of heights. Assume that $\sigma_X^2 = \sigma_Y^2$.[12]

**Solution.** Assume that the underlying distributions of $X$ and $Y$ are normal. Because $\sigma_X^2 = \sigma_Y^2$ is unknown and $n$ is small, we can construct a $100(1-\alpha)\%$ confidence interval for $\mu_X - \mu_Y$ based on the $t$ distribution with $n + m - 2$ degrees of freedom (see Section 4.4):

$$\left[\bar{x} - \bar{y} - t_{\alpha/2}(n+m-2)s_p\sqrt{\frac{1}{n} + \frac{1}{m}}, \bar{x} - \bar{y} + t_{\alpha/2}(n+m-2)s_p\sqrt{\frac{1}{n} + \frac{1}{m}}\right],$$

where

$$s_p^2 \triangleq \frac{(n-1)s_X^2 + (m-1)s_Y^2}{n+m-2}$$

is the pooled sample variance. In this case, $n = 12$, $\bar{x} = 65.7$, $s_x = 4$, $m = 15, \bar{y} = 68.2$, $s_y = 3$, $s_p = 3.476$, $\alpha = 0.02$, $\alpha/2 = 0.01$, $n + m - 2 = 25$, $t_{\alpha/2}(n+m-2) = t_{0.01}(25) = 2.485$. Hence we have the approximate 98% confidence interval

$$\left[65.7 - 68.2 - 2.485(3.476)\sqrt{\frac{1}{12} + \frac{1}{15}}, 65.7 - 68.2 + 2.485(3.476)\sqrt{\frac{1}{12} + \frac{1}{15}}\right] = [-5.845, 0.845].$$

[11] "Probability and Statistical Inference" 8th ed. (Hogg and Tanis) Ch6 p.301 Ex2-3

[12] "Probability and Statistical Inference" 8th ed. (Hogg and Tanis) Ch6 p.311 Ex3-3

**Exercise 3.** [13-14 Final Q2(a), (e) Modified, $\sim 8\%$]
Let $X_1, X_2, \ldots, X_n$ be a random sample from $N(\mu_X, \sigma_X^2)$ and $Y_1, Y_2, \ldots, Y_m$ be another random sample from $N(\mu_Y, \sigma_Y^2)$, and the two samples are independent. Furthermore, $\sigma_X^2 = d\sigma_Y^2$, with $\sigma_Y^2$ is unknown and $d$ is a known constant, construct a confidence interval for $\mu_X - \mu_Y$.[13]

**Solution.** Substituting $\sigma_X^2 = d\sigma_Y^2$ in Corollary 4.2 and Theorem 4.6, we have

$$Z \triangleq \frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{\sqrt{\frac{d\sigma_Y^2}{n} + \frac{\sigma_Y^2}{m}}} \sim N(0,1)$$

$$W \triangleq \frac{(n-1)S_X^2}{d\sigma_Y^2} + \frac{(m-1)S_Y^2}{\sigma_Y^2} \sim \chi^2(n+m-2)$$

and $Z$, $W$ are independent as usual. Therefore,

$$\frac{Z}{\sqrt{\frac{W}{n+m-2}}} = \frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{\sqrt{\frac{d}{n} + \frac{1}{m}}\sqrt{\frac{\frac{n-1}{d}S_X^2 + (m-1)S_Y^2}{n+m-2}}} \sim t(n+m-2)$$

and the unknown parameter $\sigma_Y^2$ is canceled out in the process. As a result,

$$\left[\bar{X} - \bar{Y} - t_{\alpha/2}(n+m-2)S_p\sqrt{\frac{d}{n} + \frac{1}{m}}, \bar{X} - \bar{Y} + t_{\alpha/2}(n+m-2)S_p\sqrt{\frac{d}{n} + \frac{1}{m}}\right],$$

is a two-sided $100(1-\alpha)\%$ confidence interval of $\mu_X - \mu_Y$, where

$$S_p^2 \triangleq \frac{\frac{n-1}{d}S_X^2 + (m-1)S_Y^2}{n+m-2}.$$

[13] "Probability and Statistical Inference" 8th ed. (Hogg and Tanis) Ch6 p.301 Ex3-10