Southern University of Science And Technology
School of Life Sciences
BIO210 Biostatistics
(Autumn, 2022)

Assignment 5
Due on 2nd Dec, 11 p.m.

1. **One-sided confidence interval 1 (10 points)**: A random sample of size n has been drawn from a population that follows a normal distribution with the standard deviation σ. Follow similar procedures described in **Lecture 19**, construct two one-sided 95% confidence intervals for the population mean, one for a lower bound and one for an upper bound.

We need to find out a and b, such that $P(\mu \geqslant a) = 0.95$ and $P(\mu \leqslant b) = 0.95$. Using the same technique, we start with:

$$
\begin{aligned}
P(z \leqslant Z_{0.05}) = 0.95 \Leftrightarrow P\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \leqslant Z_{0.05}\right) &= 0.95 \\
\Leftrightarrow P\left(-\mu \leqslant Z_{0.05} \cdot \frac{\sigma}{\sqrt{n}} - \bar{X}\right) &= 0.95 \\
\Leftrightarrow P\left(\mu \geqslant \bar{X} - Z_{0.05} \cdot \frac{\sigma}{\sqrt{n}} = 0.95\right)
\end{aligned}
$$

This is the lower bound of the one-sided 95% CI. Similarly, we can use the same logic to get the upper bound. The answers are:

$$
\begin{aligned}
\text{lower bound: } & \left[\bar{X} - Z_{0.05} \cdot \frac{\sigma}{\sqrt{n}}, +\infty\right), \text{ or: } \left[\bar{X} - 1.645 \cdot \frac{\sigma}{\sqrt{n}}, +\infty\right) \\
\text{upper bound: } & \left(-\infty, \bar{X} - Z_{0.95} \cdot \frac{\sigma}{\sqrt{n}}\right], \text{ or: } \left(-\infty, \bar{X} + 1.645 \cdot \frac{\sigma}{\sqrt{n}}\right]
\end{aligned}
$$

2. **One-sided confidence interval 2 (10 points)**: A random sample of size n has been drawn from a population that follows a normal distribution with unknown standard deviation. Construct two one-sided 95% confidence intervals for the population mean, one for a lower bound and one for an upper bound.

Using the same procedures as the previous question:

$$
\begin{aligned}
\text{lower bound: } & \left[\bar{X} - \boldsymbol{t}_{0.05} \cdot \frac{s}{\sqrt{n}}, +\infty\right) \\
\text{upper bound: } & \left(-\infty, \bar{X} - \boldsymbol{t}_{0.95} \cdot \frac{s}{\sqrt{n}}\right], \text{ or: } \left(-\infty, \bar{X} + \boldsymbol{t}_{0.05} \cdot \frac{s}{\sqrt{n}}\right]
\end{aligned}
$$

3. Write out the null (H_0) and the alternative (H_1) hypotheses for the following scenarios:

3.1) **(2.5 points)** An engineering department had 7% of female students. The department had been working hard to increase the percentage of female students. A researcher wanted to test if the female student rate had actually increased, so he obtained a random sample of students to see what proportion of the sample was female.

$\boldsymbol{H_0}$: $\pi \leqslant 7\%$

$\boldsymbol{H_1}$: $\pi > 7\%$

3.2) **(2.5 points)** Li Lei is a college basketball player who has scored 75% of the free-throws he has attempted in his career. He decided to practice a new technique for shooting his free-throws. Li Lei was curious if this new technique produced significantly better or worse results. He tried the new technique and made 70% of 50 attempts.

$\boldsymbol{H_0}$: $\pi = 75\%$

$\boldsymbol{H_1}$: $\pi \neq 75\%$

3.3) **(2.5 points)** A restaurant owner installed a new automated drink machine. The machine is designed to dispense 530 mL of liquid on the medium size setting. The owner suspects that the machine may be dispensing too much in medium drinks. They decide to take a sample of 30 medium drinks to see if the average amount is significantly greater than 530 mL.

$\boldsymbol{H_0}$: $\mu \leqslant 530$ mL

$\boldsymbol{H_1}$: $\mu > 530$ mL

3.4) **(2.5 points)** A healthcare provider saw that 48% of their members received their flu shot in a recent year. The healthcare provider tried a new advertising strategy in the following year, and they took a sample of members to test if the proportion who received their flu shot had changed.

$\boldsymbol{H_0}$: $\pi = 48\%$

$\boldsymbol{H_1}$: $\pi \neq 48\%$

4. Renal Disease: The mean serum-creatinine level measured in 12 patients one day after they received a newly proposed antibiotic was 1.2 mg/dL. Suppose the mean and standard deviation of serum creatinine in the general population are 1.0 and 0.4 mg/dL, respectively, and suppose the new antibiotic does not change the dispersion of the serum-creatinine level. Use a significance level of 0.05, perform a hypothesis testing to see whether the mean serum-creatinine level in this group is different from that of the general population.

4.1) **(2.5 points)** The population represented by these 12 patients is:

(A) All people.

(B) All healthy people

(C) All patients that have taken the new antibiotic

(D) All patients that have NOT taken the new antibiotic

4.2) **(2.5 points)** Write the null and alternative hypotheses for the test.

$H_0 : \mu = 1.0$ mg/dL

$H_0 : \mu \neq 1.0$ mg/dL

4.3) **(2.5 points)** Compute the test statistic.

$$z = \frac{1.2 - 1.0}{0.4/\sqrt{12}} = 1.732$$

4.4) **(2.5 points)** If H_0 were true, what will be the distribution of the test statistic?

If H_0 were true, the test statistic will follow a standard normal distribution $\mathcal{N}(0, 1)$.

4.5) **(2.5 points)** What is the two-sided 95% confidence interval for the patient population mean based on the data from the 12 patients.

$1.2 \pm 1.96 \cdot \frac{0.4}{\sqrt{12}}$, which is [0.974, 1.426]

4.6) **(2.5 points)** What is the $p-$value of the test, and do you reject H_0 ?

$p = 2 \times P(z \geqslant 1.732) = 0.083$. Since $p > 0.05$, we do not reject the null hypothesis.

5. Pulmonary Disease: Suppose the annual incidence of asthma in the general population among children 0–4 years of age is 2.8% for boys and 2% for girls. If 20 cases are observed over 1 year among 500 boys 0–4 years of age with smoking mothers, then test whether there is a significant difference in asthma incidence between this group and the general population.

5.1) **(2.5 points)** The population represented by these 500 boys is:

(A) All children.

(B) All children with smoking mothers

(C) All boys with smoking mothers

(D) All boys 0-4 years of age with smoking mothers

5.2) **(2.5 points)** Write the null and alternative hypotheses for the test.

$H_0 : \pi = 2.8\%$

$H_1 : \pi \neq 2.8\%$

5.3) **(2.5 points)** Compute the test statistic.

$$z = \frac{0.04 - 0.028}{\sqrt{\frac{0.028 \cdot (1 - 0.028)}{500}}} = 1.63$$

5.4) **(2.5 points)** If H_0 were true, what will be the distribution of the test statistic?

If H_0 were true, the test statistic will follow a standard normal distribution $\mathcal{N}(0, 1)$.

5.5) **(2.5 points)** What is the two-sided 95% confidence interval for the true population proportion based on the data from these 500 boys?

$0.04 \pm 1.96 \cdot \sqrt{\frac{0.04 \cdot (1 - 0.04)}{500}}$, which is [0.0228, 0.0572]

5.6) **(2.5 points)** What is the $p-$value of the test, and do you reject H_0 ?

$p = 2 \times P(z \geqslant 1.63) = 0.174$, since $p > 0.05$, we do not reject the null hypothesis.

5.7) **(2.5 points)** Suppose that four cases are observed over 1 year among 300 girls 0–4 years of age with smoking mothers. Can you make a hypothesis testiong using Z-statistics from this data? Why or why not?

No, because $np = 4 < 10$, which violates our assumption. The test statistic does not follow a standard normal distribution.

6. Serum cholesterol level: In the "serum cholesterol level" example during **Lecture 26**, we demonstrated that when taking a sample of size $n = 25$ and setting a significance level $\alpha = 0.05$, to test the null hypothesis ($H_0 : \ \mu \leqslant 180 \ mg/100mL$), we have a power of 0.702 when $\mu_1 = 200$.

6.1) **(2.5 points)** Now, for the same test, suppose we increase the sample size to $n = 50$ and assume the actual mean $\mu_1 = 200$. What is the power of the test?

First, calculate the standard error of the distribution:

$$s.e. = \frac{46}{\sqrt{50}} = 6.505$$

Then, we can get $\bar{x}_{reject}$:

$$\bar{x}_{reject} = \bar{x} + 1.645 \times 6.505 = 190.701$$

Now, we can calculate the power as:

$$\begin{aligned} Power &= P(\text{reject } H_0 \,|\, H_0 \text{ is false}) = P(\bar{x} \geqslant 190.701 \,|\, H_0 \text{ is false}) \\ &= P(\bar{x} \geqslant 190.701 \,|\, \mu_{\bar{x}} = 200, \sigma_{\bar{x}} = 6.505) \\ &= P\left(z \geqslant \frac{190.701 - 200}{6.505}\right) = 92.36\% \end{aligned}$$

6.2) **(2.5 points)** Now, for the same test, suppose the actual mean $\mu_1 = 190$. What will be the required sample size to reach a significance level $\alpha = 0.05$ and a power of 0.702?

The logic is the same as the previous question. First, calculate the standard error of the distribution:

$$s.e. = \frac{46}{\sqrt{n}}$$

Then, we can get $\bar{x}_{reject}$:

$$\bar{x}_{reject} = \bar{x} + 1.645 \times s.e. = 180 + 1.645 \times \frac{46}{\sqrt{n}} = 180 + \frac{75.67}{\sqrt{n}}$$

Now, we let the power be 0.702:

$$\begin{aligned}
Power &= P(\text{reject } H_0 \,|\, H_0 \text{ is false}) = P\left(\bar{x} \geqslant 180 + \frac{75.67}{\sqrt{n}} \,|\, H_0 \text{ is false}\right) \\
&= P\left(\bar{x} \geqslant 180 + \frac{75.67}{\sqrt{n}} \,|\, \mu_{\bar{x}} = 190, \sigma_{\bar{x}} = \frac{46}{\sqrt{n}}\right) \\
&= P\left(z \geqslant \frac{180 + \frac{75.67}{\sqrt{n}} - 190}{46/\sqrt{n}}\right) = 0.702 \\
&\Leftrightarrow \frac{180 + \frac{75.67}{\sqrt{n}} - 190}{46/\sqrt{n}} = -0.53
\end{aligned}$$

Solve the equation, we have $n = 100.1$. Therefore, we need a sample size of $n = 101$.

6.3) **(7.5 points)** Which of the following is true ?

(A) The power of the test will decrease when sample size increases.

(B) When the null hypothesis is false and the difference between μ_0 and μ_1 is small, you need larger sample size to reject the null hypothesis.

(C) When the difference between μ_0 and μ_1 is small, you need smaller sample size to reach the same significance level and power.

(D) With large enough sample size, you can always reject a null hypothesis.

7. Type I & II errors: A Type I error is when we reject a true null hypothesis. Lower values of α makes it harder to reject the null hypothesis, so choosing lower values for α can reduce the probability of a Type I error. The consequence here is that if the null hypothesis is false, it may be more difficult to reject using a low value for α. So using lower values of α can increase the probability of a Type II error. A Type II error is when we fail to reject a false null hypothesis. Higher values of α makes it easier to reject the null hypothesis, so choosing higher values for α can reduce the probability of a Type II error. The consequence here is that if the null hypothesis is true, increasing α makes it more likely that we commit a Type I error (rejecting a true null hypothesis). Now consider the following examples:

7.1) **(5 points)** Employees at a health club do a daily water quality test in the club's swimming pool. If the level of contaminants are too high, then they temporarily close the pool to perform a water treatment. We can state the hypotheses for their test as $\{H_0$: The water quality is acceptable$\}$ $v.s.$$\{H_1$: The water quality is not acceptable$\}$. What would be the consequence of a Type I error in this setting?

(A) The club closes the pool when it needs to be closed.

(B) The club closes the pool when it doesn't need to be closed.

(C) The club doesn't close the pool when it needs to be closed.

(D) None of the above.

7.2) **(5 points)** What would be the consequence of a Type II error in this setting?

(A) The club closes the pool when it needs to be closed.

(B) The club closes the pool when it doesn't need to be closed.

(C) The club doesn't close the pool when it needs to be closed.

(D) None of the above.

7.3) **(7.5 points)** In terms of safety, which error has the more dangerous consequences in this setting?

(A) Type I error.

(B) Type II error.

7.4) **(7.5 points)** What significance level should they use to reduce the probability of the more dangerous error?

(A) 0.01

(B) 0.025

(C) 0.05

(D) 0.10