Week 14: Confidence intervals, Hypothesis testing

Armenak Petrosyan

Tuesday class

Confidence intervals

- Let $X_1, \ldots, X_n$ be a random sample of size n.
- From the Law of large numbers, the sample mean is close to the population mean for a large sample size

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \approx \mu.$$

- But how close is $\bar{X}$ to μ?
- Notice that

$$|\bar{X} - \mu| \leq \delta \quad \Leftrightarrow \quad \mu \in [\bar{X} - \delta, \bar{X} + \delta].$$

- Let

$$\alpha = P(|\bar{X} - \mu| > \delta) \text{ or } 1 - \alpha = P(|\bar{X} - \mu| \leq \delta).$$

Definition

$[\bar{X} - \delta, \bar{X} + \delta]$ is called the $100(1 - \alpha)\%$ **confidence interval of the mean** μ.

- In other words, with $1 - \alpha$ probability the unknown true mean lies in the $[\bar{X} - \delta, \bar{X} + \delta]$ interval.
- $1 - \alpha$ is called **confidence coefficient** or **confidence level**.

Given α, $0 < \alpha < 1$, how to find the corresponding confidence interval?

- Assume $X_1, \ldots, X_n$ be a random sample from the $N(\mu, \sigma^2)$ where σ is known but μ is NOT.
- Let $z_{\alpha/2}$ be the value such that, for Z with standard normal distribution,

$$F(z_{\alpha/2}) = P(Z \leq z_{\alpha/2}) = 1 - \alpha/2.$$

- We can find this value for certain α-s from the table in the appendix of the book.
- Then

$$\begin{aligned} P(|Z| \leq z_{\alpha/2}) &= P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = F(z_{\alpha/2}) - F(-z_{\alpha/2}) \\ &= F(z_{\alpha/2}) - (1 - F(z_{\alpha/2})) \\ &= 1 - \frac{\alpha}{2} - (1 - (1 - \frac{\alpha}{2})) = 1 - \alpha. \end{aligned}$$

- $Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{\sqrt{n}(\bar{X} - \mu)}{\sigma}$ is $N(0, 1)$ so

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

- Or equivalently,

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

Normal case with known variance

The formula of the $100(1-\alpha)\%$ confidence interval of the mean when samples are normal and σ is known:

$$\boxed{[\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}]}.$$

- For large n, the confidence interval is smaller: i.e. $\bar{X}$ lies much closer to the true mean with given confidence level.
- With confidence interval for the mean we are estimating an interval where the unknown mean lies with high probability.
- For this reason it falls into **interval estimation** techniques in contrast to point estimation where we output a single value.

General case with known variance and large n

- Assume the distribution from which $X_1, \ldots, X_n$ are sampled is not necessarily normal but σ is known.
- From CLT, for large n, $Z = \frac{\sqrt{n}(\bar{X}-\mu)}{\sigma}$ is only approximately $N(0,1)$ so

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

- Or equivalently,

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

We can use

$$\boxed{[\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}]}.$$

as an estimate of the $100(1-\alpha)\%$ confidence interval of the mean when n is sufficiently large.

General case with unknown variance and large n

- Assume the distribution from which $X_1, \ldots, X_n$ are sampled is not normal and σ is also unknown.
- Let S^2 be the sample variance

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

- For large n, $S^2 \approx \sigma^2$ and $Z = \frac{\sqrt{n}(\bar{X}-\mu)}{S}$ is approximately $N(0,1)$.

We can use

$$\boxed{[\bar{X} - z_{\alpha/2} \frac{S}{\sqrt{n}}, \bar{X} + z_{\alpha/2} \frac{S}{\sqrt{n}}]}.$$

as an estimate of the $100(1-\alpha)\%$ confidence interval of the mean when n is large enough.

Normal case with unknown variance and small n

- Assume the distribution from which $X_1, \ldots, X_n$ are sampled is normal and σ is also unknown.
- For small n (≤ 30), $S^2 \approx \sigma^2$ and $\frac{\sqrt{n}(\bar{X}-\mu)}{S}$ is NOT approximately $N(0,1)$.

Definition

Let $X_1, \ldots, X_n$ be i.i.d. $N(0,1)$. The distribution of the random variable

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

is called **Student's t distribution with $n-1$ degrees of freedom**.

- The pdf of Student's t distribution with $n-1$ degrees of freedom is given by

$$f(t) = \frac{\Gamma(\frac{n}{2})}{\sqrt{(n-1)\pi} \cdot \Gamma(\frac{n-1}{2})} \left(1 + \frac{t^2}{n-1}\right)^{-\frac{n}{2}}.$$

- Let $t_{\alpha/2}$ be the value such that, for T with Student's t distribution with $n-1$ degrees of freedom,
$$F(t_{\alpha/2}) = P(T \leq t_{\alpha/2}) = 1 - \alpha/2.$$

The formula of the $100(1-\alpha)\%$ confidence interval of the mean when samples are normal and σ is unknown:
$$\boxed{[\bar{X} - t_{\alpha/2}\frac{S}{\sqrt{n}}, \bar{X} + t_{\alpha/2}\frac{S}{\sqrt{n}}]}.$$

One-sided confidence intervals

- Let
$$1 - \alpha = P(\mu \geq \bar{X} - \delta).$$
- With probability $1 - \alpha$ we are confident that the unknown $\mu \geq \bar{X} - \delta$.

Definition

We say that $[\bar{X} - \delta, \infty]$ is the one-sided $100(1 - \alpha)\%$ **confidence interval of the mean** μ.

- Note that
$$P(\mu \geq \bar{X} - \delta) = P(\frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \leq \delta \frac{\sqrt{n}}{\sigma}).$$
- Per our definition, for Z which is $N(0,1)$,
$$P(Z \leq z_\alpha) = 1 - \alpha.$$
- When $X_1, \ldots, X_n$ are sampled from normal distribution (or n is large enough), the (corresp. approximate) one-sided $100(1-\alpha)\%$ confidence interval of the mean is at
$$z_\alpha = \frac{\sqrt{n}\delta}{\sigma} \quad \Longrightarrow \quad \delta = z_\alpha \frac{\sigma}{\sqrt{n}}.$$

$$[\bar{X} - z_\alpha \frac{\sigma}{\sqrt{n}}, \infty).$$

- When the variance is unknown and n is large we can use as an estimate of the one-sided confidence interval the

$$[\bar{X} - z_\alpha \frac{S}{\sqrt{n}}, \infty).$$

where S is the sample mean.

- When the variance is unknown, $X_1, \ldots, X_n$ are sampled from a normal distribution and n is small, the one sided confidence interval is given as

$$[\bar{X} - t_\alpha \frac{S}{\sqrt{n}}, \infty).$$

- We can similarly do one-sided confidence intervals for the upper bound of μ as

$$(-\infty, \bar{X} + z_\alpha \frac{\sigma}{\sqrt{n}}].$$

Exercise 1

Problem (7.1-7 in the textbook)

Thirteen tons of cheese, including "22-pound" wheels (label weight), is stored in some old gypsum mines. A random sample of $n = 9$ of these wheels yielded the following weights in pounds:

21.50	18.95	18.55	19.40	19.15
22.35	22.90	22.20	23.10	

Assuming that the distribution of the weights of the wheels of cheese is $N(\mu, \sigma^2)$, find a 95% confidence interval for μ.

Solution

$$\bar{x} = \frac{21.50 + 18.95 + 18.55 + 19.40 + 19.15 + 22.35 + 22.90 + 22.20 + 23.10}{9} = 20.9$$

$$S = \sqrt{\frac{(21.50 - 20.9)^2 + \cdots + (23.10 - 20.9)^2}{8}} \approx 1.86.$$

Solution (cont.)

- *From the table, for* $n-1=8$, $t_{0.025}=2.306$.
- *The confidence interval is*

$$\left[20.9-2.306\cdot\frac{1.86}{3},\ 20.9+2.306\cdot\frac{1.86}{3}\right]\approx[19.47,22.33].$$

Sample size

How large should the n be so that with $100(1-\alpha)\%$ confidence the mean lies in the interval $[\bar{X}-\epsilon, \bar{X}+\epsilon]$.

- ϵ is called **the maximum error of the mean estimate**.
- We can take n large enough so that

$$[\bar{X} - z_{\alpha/2}\frac{S}{\sqrt{n}}, \bar{X} + z_{\alpha/2}\frac{S}{\sqrt{n}}] \subseteq [\bar{X}-\epsilon, \bar{X}+\epsilon].$$

- Or, equivalently,

$$z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \epsilon \quad \Longrightarrow \quad \boxed{n \geq \frac{z_{\alpha/2}^2\sigma^2}{\epsilon^2}}.$$

Thursday class

Hypothesis testing: motivation

- We have a random sample $X_1, \ldots, X_n$ from $N(\mu, \sigma)$ where μ is unknown and σ is known.
- Suppose we want to check if a certain value μ_0 was the true mean.
- If $\mu = \mu_0$ was the true mean then with $1 - \alpha$ probability it lies in the confidence interval

$$[\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}].$$

- We can use this to reject or accept whether μ_0 was the true value.
- What if we have two values μ_0, μ_1 and want to decide which one is the correct mean between the two?
- We can pick the one that is in the interval. But that may be inconclusive as both can be in the interval or both be outside.
- For that reason we give priority to one of them, say μ_0, and call the $\mu = \mu_0$ the **null hypothesis** denoted by H_0.
- We call $\mu = \mu_1$ the **alternative hypothesis** denoted by H_1.
- We want to have a trustworthy way to simultaneously check if the alternative hypothesis is correct and the null hypothesis is wrong before we can accept it.

Another way to think about null and alternative hypothesis is as follows.

- Null hypothesis is your accepted truth (e.g. people cannot fly).
- The alternative hypothesis tries to challenge the null hypothesis. If the alternative hypothesis fails (e.g. a guy with feather wings jumping from a cliff and failed to prove people can fly) that does not mean the null hypothesis was correct, but means we **failed to reject it**.
- Maybe another alternative hypothesis will come around and challenge the null hypothesis better. If the alternative hypothesis is shown to be true (someone has a better idea of how built planes), then the null hypothesis is rejected and the alternative hypothesis is accepted as a new null.

Hypothesis testing: step-by-step

1. We assume we have two hypothesis H_0 and H_1 about the underlying distribution.
2. We sample $X_1 = x_1, \dots, X_n = x_n$.
3. If $(x_1, \dots, x_n) \in C$ then we reject H_0 and accept H_1.
4. Otherwise we reject H_1 and fail to reject H_0.

- The set C is called **critical region**. It must be designed in a way that allows to reject the null hypothesis when H_0 is wrong and also accept the alternative when H_1 is correct.
- This is the region where the alternative hypothesis is accepted.
- In the previous example of choosing between μ_0, μ_1 we can choose the critical region as

$$\mu_0 \notin [\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}] \quad \Leftrightarrow \quad \boxed{|\bar{x} - \mu_0| > z_{\alpha/2}\frac{\sigma}{\sqrt{n}}}.$$

- Here the sample mean is called $teststatistic$ because the critical region is defined using the mean:

$$C = \{(x_1, \dots, x_n) : \ |\bar{x} - \mu_0| > z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\}.$$

- However this is not the best choice for this test as we will see later (it may wrongly reject the H_1 when μ and μ_1 are close to each other).

Errors

What can go wrong?

1. H_0 is rejected when it was in fact true (**Type I error**).
2. H_1 is rejected when it was in fact true which is same as, H_0 is not rejected when it is false (**Type II error**).

Definition

The probability of Type I error is called **significance level** and denoted by α

$$\alpha = P((X_1, \ldots, X_n) \in C | H_0).$$

We also denote the probability of Type II error by β:

$$\beta = P((X_1, \ldots, X_n) \notin C | H_1).$$

- The best scenario is when both α and β are small.
- They depend on the critical region and on the sample size.

Example

- We have $n = 16$ samples from a normal distribution.
- We want to test whether the sampled data is from $N(50, 36)$ or $N(55, 36)$.
- Let
$$C = [(x_1, \dots, x_n) : \bar{x} \geq 53].$$
- Then,
$$\alpha = P(\bar{X} \geq 53 | \mu = 50) = P\left(\frac{\bar{X} - 50}{6/4} \geq \frac{53 - 50}{6/4}\right) = 1 - F(2) = 0.0228.$$
- Similarly,
$$\beta = P(\bar{X} < 53 | \mu = 55) = P\left(\frac{\bar{X} - 55}{6/4} < \frac{53 - 55}{6/4}\right)$$
$$= F\left(-\frac{4}{3}\right) = 1 - F\left(\frac{4}{3}\right) = 1 - 0.9087 = 0.0913.$$
- Turns out, the above C is the best critical region for this problem: for fixed α, it has the smallest β (follows from **Neyman-Pearson lemma**).

Simple null v.s. composite alternative

- The null hypothesis we will test here will have the form
$$H_0 = \{\mu = \mu_0\}.$$
This type of hypothesis are called **simple** because we are testing a single value.
- The alternative hypotheses we will consider have one of these forms
$$H_1 = \{\mu > \mu_0\}$$
$$H_1 = \{\mu < \mu_0\}$$
$$H_1 = \{\mu \neq \mu_0\}$$
These type of hypotheses are called **composite hypotheses** because they are testing for a range of values and not for a single value.

Normal case with known variance

1. $H_0 = \{\mu = \mu_0\}, \quad H_1 = \{\mu > \mu_0\}$. We test with
$$\bar{x} \geq \mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}}.$$
2. $H_0 = \{\mu = \mu_0\}, \quad H_1 = \{\mu < \mu_0\}$. We test with
$$\bar{x} \leq \mu_0 - z_\alpha \frac{\sigma}{\sqrt{n}}.$$
3. $H_0 = \{\mu = \mu_0\}, \quad H_1 = \{\mu \neq \mu_0\}$. We test with
$$|\bar{x} - \mu_0| \geq z_{\alpha/2} \frac{\sigma}{\sqrt{n}}.$$

- We use the same test for general distributions with large n using the CLT.
- These are what's called **uniformly most powerful** critical regions for the above tests.

Normal case with unknown variance and small n

1. $H_0 = \{\mu = \mu_0\}, \quad H_1 = \{\mu > \mu_0\}$. We test with
$$\bar{x} \geq \mu_0 + t_\alpha \frac{s}{\sqrt{n}}$$
(i.e. C is the complement of the one-sided confidence interval)
2. $H_0 = \{\mu = \mu_0\}, \quad H_1 = \{\mu < \mu_0\}$. We test with
$$\bar{x} \leq \mu_0 - t_\alpha \frac{s}{\sqrt{n}}.$$
3. $H_0 = \{\mu = \mu_0\}, \quad H_1 = \{\mu \neq \mu_0\}$. We test with
$$|\bar{x} - \mu_0| \geq t_{\alpha/2} \frac{s}{\sqrt{n}}.$$

Definition

Assume the null hypothesis is true and $X_1 = x_1, \ldots, X_n = x_n$. The largest value of α, for which we will wrongly reject the null hypothesis, given that the null hypothesis is true, is called p-**value**.

- If p-value $> \alpha$, fail to reject the null hypothesis.
- If p-value $\leq \alpha$ we reject the null hypothesis and accept the alternative.

The advantage of using p-values is that you compute it once and know what significance levels you can tolerate.

1. $H_0 = \{\mu = \mu_0\}, \quad H_1 = \{\mu > \mu_0\}$. We test with

$$p\text{-value} = P(\bar{X} \geq \bar{x} | \mu = \mu_0) = P\left(\frac{\sqrt{n}(\bar{X} - \mu_0)}{\sigma} \geq \frac{\sqrt{n}(\bar{x} - \mu_0)}{\sigma}\right).$$

2. $H_0 = \{\mu = \mu_0\}, \quad H_1 = \{\mu < \mu_0\}$. We test with

$$p\text{-value} = P(\bar{X} \leq \bar{x} | \mu = \mu_0) = P\left(\frac{\sqrt{n}(\bar{X} - \mu_0)}{\sigma} \leq \frac{\sqrt{n}(\bar{x} - \mu_0)}{\sigma}\right).$$

3. $H_0 = \{\mu = \mu_0\}, \quad H_1 = \{\mu \neq \mu_0\}$. We test with

$$p\text{-value} = P(|\bar{X} - \mu_0| \geq |\bar{x} - \mu_0| \; | \mu = \mu_0) = P\left(\frac{\sqrt{n}|\bar{X} - \mu_0|}{\sigma} \geq \frac{\sqrt{n}|\bar{x} - \mu_0|}{\sigma}\right).$$

Exercise 2

Problem

Assume that IQ scores for a certain population are approximately $N(\mu, 100)$. To test $H_0 : \mu = 110$ against the one-sided alternative hypothesis $H_1 : \mu > 110$, we take a random sample of size $n = 16$ from this population and observe $\bar{x} = 113.5$.

(a) *Do we accept or reject H_0 at the 5% significance level?*

(b) *Do we accept or reject H_0 at the 10% significance level?*

(c) *What is the p-value of this test?*

Solution

(a) $z_{0.05} = 1.64$. *Therefore*

$$\mu_0 + z_{0.05}\frac{\sigma}{\sqrt{n}} = 110 + 1.64 \cdot \frac{10}{4} = 114.1 > \bar{x}$$

so we fail to reject the null hypothesis.

(b) $z_{0.1} = 1.28$

$$\mu_0 + z_{0.1}\frac{\sigma}{\sqrt{n}} = 110 + 1.28 \cdot \frac{10}{4} = 113.2 < \bar{x}$$

so we reject the null hypothesis and accept the alternative.

Solution (cont.)

(c)

$$p\text{-value} = P\left(\frac{\sqrt{n}(\bar{X} - \mu_0)}{\sigma} \geq \frac{\sqrt{n}(\bar{x} - \mu_0)}{\sigma}\right).$$

Note that

$$\frac{\sqrt{n}(\bar{x} - \mu_0)}{\sigma} = \frac{4(113.5 - 110)}{10} = 1.4.$$

And

$$P(Z \geq 1.4) = 1 - F(1.4) = 1 - 0.9192 = 0.0808.$$

From here also we can conclude that at the significance level $\alpha = 0.05$ we will fail to reject (p-value $> c$) and we will accept the alternative at $\alpha = 0.1$.