MH1820 Week 10

Interval Estimator

An estimator $\widehat{\theta}$ only provides a single "best guess" for θ ("point estimator"), based on a random sample.

Bias and standard error measure average precision of $\widehat{\theta}$. Both types of information, the best guess and average precision, can be combined into a **confidence interval** ("**interval estimation**").

In almost all applications of parameter estimation, confidence intervals are used (point estimation is not enough).

unknown but not random

More precisely, we want to find values $\widehat{\theta}_L$ and $\widehat{\theta}_U$ such that

$$\mathbb{P}(\widehat{\theta}_L < \theta < \widehat{\theta}_U) = 1 - \alpha, \quad \text{where } 0 < \alpha < 1.$$

estimated from sample.

- The interval $[\widehat{\theta}_L, \ \widehat{\theta}_U]$ computed from the selected sample is called a **confidence interval** for θ.

- The fraction $1 - \alpha$ is called the **confidence level**. Some common values of α are 0.01, 0.05, 0.025.

- The endpoints $\widehat{\theta}_L$, $\widehat{\theta}_U$ are called the **lower and upper confidence limits**.

Strategy to Construct Confidence Intervals

- $X_1, \ldots, X_n$ i.i.d with some distribution depending on an unknown parameter θ.
- Goal: Find $100(1-\alpha)\%$ confidence interval for θ.
- Idea: Find a statistic Q involving $X_1, \ldots, X_n$ and θ such that the distribution of Q is known.
- Find a, b such that $\mathbb{P}(a < Q < b) = 1 - \alpha$.
- Transform $a < Q < b$ to an equivalent condition $\theta_L < \theta < \theta_U$. Then $[\theta_L, \theta_U]$ is the required confidence interval.

E.g:

Population

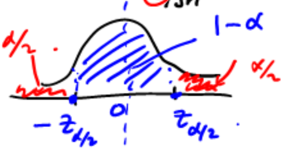
$\theta = \mu$ $N(\mu, \sigma^2)$ $\to X_1, X_2, \ldots, X_n$

unknown

n = sample size.

σ^2 known

Chose $Q = \dfrac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$



$P(-z_{\alpha/2} < Q < z_{\alpha/2}) = 1 - \alpha$ ✓

σ^2 unknown.

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

as proxy to σ^2

$Q = \dfrac{\bar{X} - \mu}{S/\sqrt{n}}$ ✓

Transform back to $\theta = \mu$

$$-z_{\alpha/2} < Q = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < z_{\alpha/2}$$

$$\underbrace{\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}}$$

this is the $(1-\alpha)$ confidence interval for $\theta = \mu$.

$n \geq 30$ (large sample size)

$n < 30$ (small sample size).

$Q \sim N(0,1)$

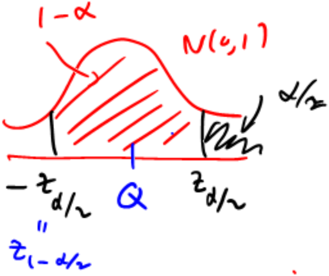
$Q \sim t(n-1)$

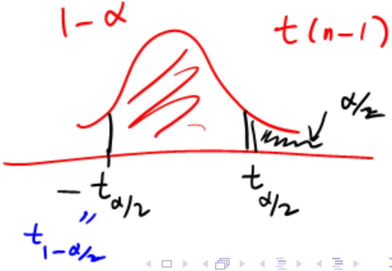
t-distribution w/ degree freedom $n-1$

μ unknown.

Summary: $100(1-\alpha)\%$ Confidence Interval or μ of $N(\mu, \sigma^2)$:

Case	Statistic Q	Dist. of Q	CI
Known σ^2	$\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$	$N(0,1)$	$\left[\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}},\ \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$
Unknown σ^2 ($n \geq 30$)	$\frac{\overline{X}-\mu}{S/\sqrt{n}}$	$\approx N(0,1)$	$\left[\overline{X} - z_{\alpha/2}\frac{S}{\sqrt{n}},\ \overline{X} + z_{\alpha/2}\frac{S}{\sqrt{n}}\right]$
Unknown σ^2 ($n < 30$)	$\frac{\overline{X}-\mu}{S/\sqrt{n}}$	t-distribution degree freedom $n-1$	$\left[\overline{X} - t_{\alpha/2}(n-1)\frac{S}{\sqrt{n}},\ \overline{X} + t_{\alpha/2}(n-1)\frac{S}{\sqrt{n}}\right]$





Theorem 6 (Confidence Interval for μ of normal distribution with known σ^2)

Let $X_1, \ldots, X_n$ i.i.d $\sim N(\mu, \sigma^2)$. The $100(1-\alpha)\%$ confidence interval for μ is given by

$$\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

Recall z_α is the upper $100(1-\alpha)\%$ point of the standard normal, i.e. $\mathbb{P}(\phi > z_\alpha) = \alpha$.

Justification:

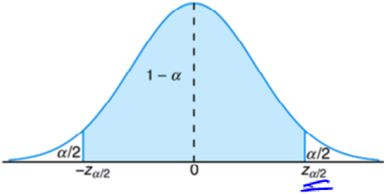
- Select statistic $Q = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$ the standardized sample mean. We know that
$$Q \sim N(0,1).$$

- Choose $a = -z_{\alpha/2}$ and $b = z_{\alpha/2}$. Then
$$\mathbb{P}\left(-z_{\alpha/2} < \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} < z_{\alpha/2}\right) = 1-\alpha.$$

- Rearranging the terms, the interval $-z_{\alpha/2} < \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} < z_{\alpha/2}$ is equivalent to
$$\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

Some commonly used z_α:

α	z_α
0.05	1.645
0.025	1.96
0.005	2.575



$n = 36$

$\bar{x} = 2.6$

Example 7

The average zinc concentration recovered from a sample of measurements taken from 36 different locations in a river is found to be 2.6 g/ml.

Find the 95% **confidence interval** for the mean concentration in the river based on the data above. Assume the population of zinc concentration is normally distributed with standard deviation of 0.3 g/ml.

$1 - \alpha = 0.95$

$\alpha = 0.05$.

$\sigma = 0.3$ (known)

Let $X \sim N(\mu, \sigma^2 = 0.3^2)$ be the random variable for the population zinc concentration in the river.

We have $1 - \alpha = 0.95 \implies \alpha = 0.05$. Thus, the 95% confidence interval for μ is

$$\boxed{\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}}$$

$$2.6 - z_{0.025}\frac{0.3}{\sqrt{36}} < \mu < 2.6 + z_{0.025}\frac{0.3}{\sqrt{36}}$$

$$\boxed{2.5 < \mu < 2.7.}$$

Suppose we repeat for different samples, we may get different intervals, say

[2.4, 2.6], [2.49, 2.9], [2, 2.4], [2.1, 3], . . .

Remarks.

- Ideally, we prefer short interval of high level of confidence.
- Confidence interval computed from a given set of observations either contain the true value or it does not.
- The level of confidence is about proportions of samples of a given size that may be expected to contain the true value.
- That is, for a 95% level of confidence, if many samples of a given size are collected and the confidence intervals are computed, about 95% of these intervals would contain the true value.

Theorem 8 (Confidence Interval for μ of normal distribution with unknown σ^2, $n \geq 30$)

Let $X_1, \ldots, X_n$ i.i.d $\sim N(\mu, \sigma^2)$, $n \geq 30$, and σ^2 is unknown. The $100(1-\alpha)\%$ confidence interval for μ is given by

$$\overline{X} - \underline{\underline{z_{\alpha/2}}}\frac{S}{\sqrt{n}} < \mu < \overline{X} + \underline{\underline{z_{\alpha/2}}}\frac{S}{\sqrt{n}},$$

where S^2 is the sample variance.

When population variance σ^2 is unknown, sample variance is sufficiently good estimator of σ^2 for $n \geq 30$.

For small sample size $n < 30$, the value of sample variance S^2 fluctuate considerably from sample to sample.
To deal with inference on μ, we consider the following statistic

$$Q = \frac{\overline{X} - \mu}{S/\sqrt{n}}$$

The distribution of Q is a t-**distribution with degree of freedom** $n - 1$, under the assumption that the population is (approximately) normal.

The upper $100(1-\alpha)\%$ point of a t-distribution with degree r, denoted by $t_\alpha(r)$, is defined similatly. That is

$$\mathbb{P}(Q > t_\alpha(r)) = \alpha.$$

The values of $t_\alpha(r)$ can be found in the Table (NTULearn $->$ Content $->$ TABLES.pdf).

Using a similar argument, we have the following:

Theorem 9 (Confidence Interval for μ of normal distribution with unknown σ^2, $n < 30$)

Let $X_1, \ldots, X_n$ i.i.d $\sim N(\mu, \sigma^2)$, $n < 30$, and σ^2 is unknown. The $100(1-\alpha)\%$ confidence interval for μ is given by

$$\overline{X} - t_{\alpha/2}(n-1)\frac{S}{\sqrt{n}} < \mu < \overline{X} + t_{\alpha/2}(n-1)\frac{S}{\sqrt{n}},$$

where S^2 is the sample variance.

Example 10

The contents of seven similar containers of sulphuric acid are 9.8, 10.2, 10.4, 9.8, 10.0, 10.2, and 9.6 liters. Find a 95% confidence interval for the mean contents μ of population, assuming a normal distribution.

$\alpha = 0.05$

$n = 7$

$t_{\alpha/2}(n-1) = t_{0.025}(6)$

It is a small sample with $n = 7$, with $\overline{x} = 10$, $s = 0.283$. Also, $\alpha = 0.05$ and $t_{0.025}(6) = 2.447$. A 95% confidence interval for μ is

$$\overline{x} - t_{\alpha}(n-1)\frac{s}{\sqrt{n}} < \mu < \overline{x} + t_{\alpha}(n-1)\frac{s}{\sqrt{n}}$$

$$10 - 2.447\frac{0.283}{\sqrt{7}} < \mu < 10 - 2.447\frac{0.283}{\sqrt{7}}$$

$$\boxed{9.74 < \mu < 10.26.}$$

MH1820 Week 11

Confidence Interval for Variance

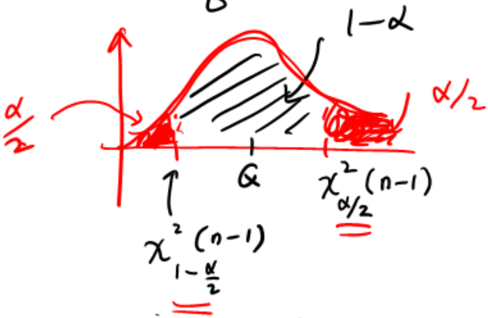
Suppose $X_1, \ldots, X_n$ i.i.d $\sim N(\mu, \sigma^2)$. Notice that

$$
\begin{aligned}
S^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 \\
\frac{(n-1)S^2}{\sigma^2} &= \sum_{i=1}^{n}\left(\frac{X_i - \overline{X}}{\sigma}\right)^2
\end{aligned}
$$

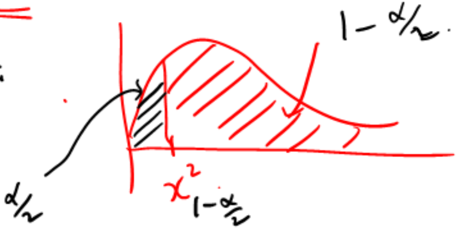
- Recall that $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$ (chi-square distribution with degree of freedom $n-1$) (Week 9).
- We can use this to construct confidence intervals.

$$Q = \frac{(n-1)s^2}{\sigma^2} \sim \chi^2(n-1)$$

Given α:



Recall:



Q =

$$\mathbb{P}\left(\chi^2_{1-\alpha/2}(n-1) < \frac{(n-1)S^2}{\sigma^2} < \chi^2_{\alpha/2}(n-1)\right) = 1 - \alpha.$$

Here, $\mathbb{P}(X > \chi^2_\alpha(r)) = \alpha$, i.e. $\chi^2_\alpha(r)$ is the upper $100(1-\alpha)\%$ point.

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

$$\frac{(n-1)S^2}{\chi^2_{\alpha/2}(n-1)} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{1-\alpha/2}(n-1)}$$

Theorem 1 (Confidence Interval for σ^2 of normal distribution)

Let $X_1, \ldots, X_n$ i.i.d $\sim N(\mu, \sigma^2)$. The $100(1-\alpha)\%$ confidence interval for σ^2 is given by

$$\frac{(n-1)S^2}{\chi^2_{\alpha/2}(n-1)} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{1-\alpha/2}(n-1)},$$

where S^2 is the sample variance.

Example 2

The following are the weights, in decagrams, of 10 packages of grass seed distributed by a certain company:

$$46.4, 46.1, 45.8, 47.0, 46.1, 45.9, 45.8, 46.9, 45.2, 46.0.$$

Find a 95% confidence interval for the variance of the weights of all such packages of grass seed distributed by this company, assuming a normal distribution.

$\alpha = 0.05.$

$n = 10$

Solution. $\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$

$$s^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right) = \frac{1}{9}\left(\sum_{i=1}^{9} x_i^2 - 10 \cdot 46.12^2\right) \approx 0.286$$

For 95% confidence interval, we have $\alpha = 0.05$. From the χ^2-table, with degree of freedom $n-1=9$, we have $\chi^2_{0.025}(9) = 19.02$, $\chi^2_{0.975}(9) = 2.700$.

Therefore, a 95% confidence interval for σ^2 is

$$\boxed{0.135 = \frac{(10-1)(0.286)}{19.02} < \sigma^2 < \frac{(10-1)(0.286)}{2.7} = 0.953.}$$

□

Purpose and Rationale of Hypothesis Tests

Hypothesis Testing.

A **statistical hypothesis** is an assertion or conjecture concerning one or more populations. For example:

- An engineer claims that the fraction of defective in a process is 0.10.
- A manufacturer claims that the average saturated fat content in a certain rice cereal does not exceed 1.5 grams per serving.
- A project manager claims that the abrasive wear of Material A exceeds that of Material B by 2. more units

The aim of hypothesis tests is to **decide**, based on the given observations, whether to accept or reject the claim.

Before we begin, consider a criminal trial by jury:

- A jury must decide between two hypotheses.
  - The **null hypothesis** H_0: The defendant is innocent.
  - The **alternative hypothesis** H_1: The defendant is guilty.
- The jury does not know which hypothesis is true. They must make a decision on the basis of the evidence presented.

There are two possible decisions.

- Convicting the defendant is called **rejecting the null hypothesis** in favor of the alternative hypothesis. That is, the jury is saying that there is enough evidence to conclude that the defendant is guilty (the alternative hypothesis).

- If the jury acquits it is stating that there is **not enough evidence to support the alternative hypothesis**. Notice that the jury is not saying that the defendant is innocent, only that there is not enough evidence to support the alternative hypothesis.

Choosing H_0:

Given observations $x_1, \ldots, x_n$, the purpose of a hypothesis test is to determine whether a certain "**interesting effect**" exists.

evidence against null Hypothesis.

- H_0 should specify a distribution that is reasonable as a population distribution for the observations under the assumption that **no effect** exists.

- Rejecting H_0 means that the observations provide significant evidence for the effect.

- Not rejecting H_0 means that the observations do not contain significant evidence for the effect.

Procedue for Hypothesis Testing:

- Given are observations $x_1, \ldots, x_n$.
- Formulate **null hypothesis** H_0 describing the population distribution from which observations were drawn.
- Choose significance level α (often $\alpha = 0.05$) $1-\alpha =$ confidence level
- Choose test statistic $T(X_1, \ldots, X_n)$ that contains information on the parameters involved in H_0 and whose distribution is known under H_0.
- Assuming H_0, compute probability (p-value) to observe $t = T(x_1, \ldots, x_n)$ or something "**at least as extreme as t" (in the direction of rejection of H_0)**.
- If the p-value is smaller than α, reject null hypothesis.

small p-value means the probability of observing such evidence (t) is small if H_0 is true.

But α = minimum such probability I can accept.

Meaning of "at least as extreme"

- Suppose the statistic $T(X_1, \ldots, X_n)$ is used to test H_0.
- Let $t = T(x_1, \ldots, x_n)$ be the obbserved value of T.
- Let $\mathbb{E}[T]$ be the expectation of T under the assumption that H_0 is true. Often deviation from $\mathbb{E}[T]$ is viewed as evidence against H_0.
- **"at least as extreme as t" (in the direction of rejection of H_0)** means
  - $T \geq t$ (one-sided test)
  - $T \leq t$ (one-sided test)
  - $|T - \mathbb{E}[T]| \geq |t - \mathbb{E}[T]|$ (two-sided test)

$T - \cancel{\mathbb{E}[T]} \geq t - \cancel{\mathbb{E}[T]}$

$T - \cancel{\mathbb{E}[T]} \leq t - \cancel{\mathbb{E}[T]}$

- The direction of rejection is determined by the alternative hypothesis H_1.

p-value

- p-value is the probability to observe t or something "at least as extreme as t" assuming H_0 is true.
- If p-value is small, it means that chances of observing what we have observed (assuming H_0 is true) is small.

$$\Longrightarrow \text{ the smaller the } p\text{-value, the less we should believe in } H_0.$$

- The significance level α is the minimum value of this probability that we are willing to accept before perfoming the test.

$$\Longrightarrow \begin{cases} \text{Reject } H_0 & \text{if } p\text{-value} < \alpha \\ \text{Do not reject } H_0 & \text{otherwise.} \end{cases}$$

Examples of Hypothesis Testing

$n = 100$

Example 3

A random sample of 100 recorded deaths in the US during the past year showed that an average life span of 71.8 years. Assuming a population standard deviation of 8.9 years, does this seem to indicate that the mean life span today is more than 70 years?

$\sigma = 8.9$

Perform a test with $\alpha = 0.05$ as the significance level.

Solution.

Note: population σ^2 is given. By Central Limit Theorem, the sample mean $\overline{X}$, with $n = 100$, is approximately normal. In particular, the statistic

$$T = \frac{X - \mu}{\sigma/\sqrt{n}} \sim N(0, 1).$$

Let μ be the population mean.

- Null Hypothesis H_0: $\mu = 70$ (years)
- Alternative Hypothesis H_1: $\mu > 70$ (years)
- Set $\alpha = 0.05$
- Choose statistic $T = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$

The Alternative Hypothesis H_1: $\mu > 70$ suggests that we do a one-sided test with p-value $\mathbb{P}(T \geq t)$.

Compute p-value based on data:

assume H_0 is true.
$\mu = 70$

$$t = \frac{\overline{x} - \mu}{\sigma/\sqrt{n}} = \frac{71.8 - 70}{8.9/\sqrt{100}} = 2.02$$

$T \sim N(0,1)$

$$\begin{aligned} p\text{-value} &= \mathbb{P}(T \geq t) \\ &= \mathbb{P}(T \geq 2.02) \\ &= 1 - \Phi(2.02) = 1 - 0.9783 = 0.0217 < \alpha. = 0.05 \end{aligned}$$

Decision: Reject H_0 (since the p-value is less than $\alpha = 0.05$).