Packet 4: Interval Estimation

Learning Objects: Learn how to construct and interpret 95% confidence intervals.

midterm I | midterm II

Text book sections: 7.1, 7.2, 7.3, 7.4.

Motivating Example: Wall Street analyst told his manager that the stock price will increase $22.00 tomorrow.
Manager: Are you sure?
Analyst: No.
Manager: What is the probability of that?

It is wise to provide an interval estimate that is associated with some confidence.

e.g. with 95% confidence, increase by \$18.00 ~ \$26.00

Chap 7.1 Confidence Interval for Means

Q: How to find an interval $[a, b]$, such that with 95% confidence $[a, b]$ covers the true value of the unknown parameter θ?

Confidence interval for μ when $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$, σ^2 is known.

In Chap 5.5, we already know that

data, unknown parameter

sample mean $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$, and thus $Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$.

standarization $Y \sim N(a, b)$ $\frac{Y-a}{\sqrt{b}} \sim N(0,1)$ does not depend on μ

In statistics, a *pivotal quantity* is a function of observations and unobservable parameters such that the function's probability distribution does not depend on the unknown parameters.

$Z = h(X, \mu, \sigma^2)$ in this problem is a pivotal quantity. $Z \sim N(0, 1)$ does not depend on the values of the parameters of interest.

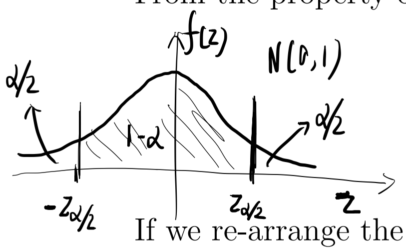
Construct $(1-\alpha)\times 100\%$ confidence interval
(C.I.)

Pivotal quantities are fundamental to the construction of confidence intervals and test statistics.

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

Select the appropriate quantiles of the distribution of the pivotal quantity to build the bounds of the confidence interval; in general, we pull off the $(\alpha/2)^{\text{th}}$ and $(1-\alpha/2)^{\text{th}}$ quantiles of the distribution of the pivotal quantity, where α is some small error rate (usually $\alpha = 0.05$).

From the property of the standard Normal distribution, we know that



interval for pivotal quantity Z

$$P(-1.96 < Z < 1.96) = 0.95.$$

$Z_{\alpha/2}$ is the upper $\alpha/2$ percentile for $Z \sim N(0,1)$

$Z_{0.025} = 1.96$ from normal distn table

If we re-arrange the inequalities inside the above probability in terms of μ,

$$P\left(-1.96 < \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < 1.96\right) = 0.95$$

$$P\left(-1.96\frac{\sigma}{\sqrt{n}} < \bar{X}-\mu < 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95$$

upper bound lower bound

$\mu < \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}$ $\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} < \mu$

$$P\left(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95$$

$\left(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right)$ is the 95% C.I. for μ

In general $P\left(-Z_{\alpha/2} < Z < Z_{\alpha/2}\right) = 1-\alpha$

$\left(\bar{X} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$ is the $100(1-\alpha)\%$ C.I. for μ

known variance

Example 7.1-1 Let X be the length of life of a light bulb $\sim N(\mu, 36^2)$. A random sample of $n = 27$ is tested until they burn out, yielding a sample mean $\bar{x} = 1478$. Find a 95% confidence interval for μ.

$n=27$ $\quad \bar{X} \pm 1.96 \frac{\sigma}{\sqrt{n}}$ $\quad \left(1478 - 1.96 \times \frac{36}{\sqrt{27}},\ 1478 + 1.96 \times \frac{36}{\sqrt{27}}\right)$

$(1464.42,\ 1491.58)$

For a given value of $\alpha \in (0, 1)$ (typically $\alpha = 0.05$), an $100(1-\alpha)\%$ confidence interval of an estimator for some parameter θ is an interval $[l(X),\ u(X)]$ such that

$$P(l(X) < \boxed{\theta} < u(X)) = 1-\alpha$$

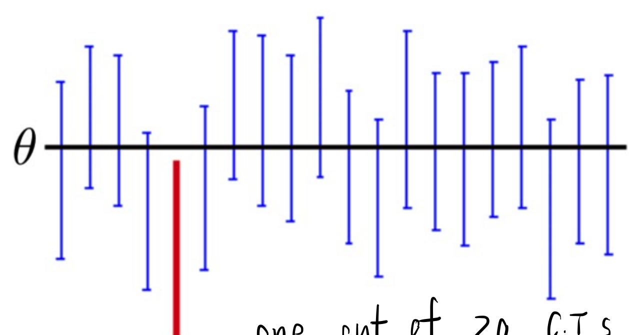
statistic, random

unknown, fixed, no randomness in θ

Note that it is the interval that is random, not the parameter θ. The randomness comes from which data we observe, e.g., the realization of $\bar{X}$, i.e., $\bar{x}$, depends on which data we observe.

After the data are observed ($X = x$), we end up with a realization of this interval $(l(x),\ u(x))$, which is one out of infinitely many possible intervals. Since it either covers θ or does not, we cannot say that θ is in $(l(x),\ u(x))$ with 95% chance.

Instead, it actually means that if the experiment of interest (or random sampling) is repeated 100 times and we compute 100 confidence intervals from the resulting 100 datasets, then 95% confidence intervals out of 100 are expected to contain the unknown true parameter θ (i.e., we would expect 95% of the intervals to cover the true parameter we are estimating).



fixed value

one out of 20 C.I.s does not cover true θ

Broadly speaking, there are three approaches to calculate confidence intervals; (i) exact method based on the analytical solution to the sampling distribution, (ii) approximation based on large sample theory, such as CLT and MLE, and finally (iii) numerical methods such as bootstrapping and other simulation based approaches.

Confidence interval for μ when $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ with σ unknown.

Previously, we use the following pivotal quantity to derive the confidence interval:

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

What if we do not know the value of σ? *A principle in statistics is to replace any unknown parameter with a good estimator.*

A good estimator for σ^2 is the sample variance:

Chap 5.5 Theorem 5.5.2 $$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2. \sim \chi^2(n-1)$$

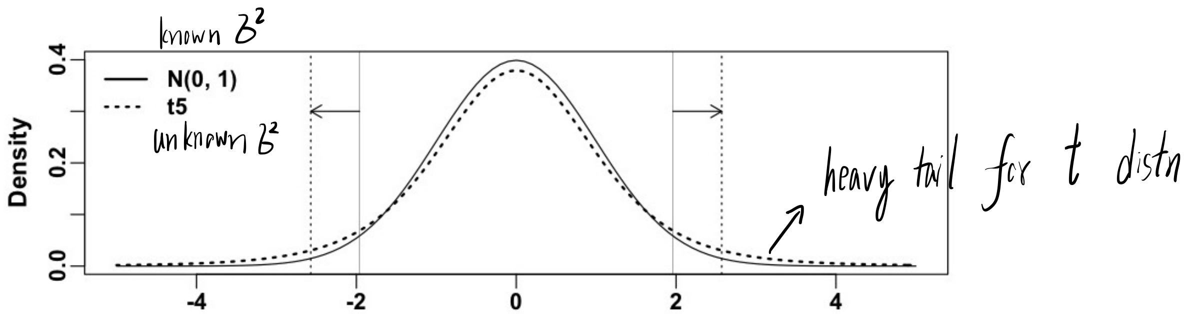
If we replace σ^2 with S^2, the distribution of the following statistic is

unknown parameter

Theorem 5.5.3 $$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}.$$ does not depend on μ.

functions of data

The t_ν-distribution is heavy-tailed. For example, if $T \sim t_5$ $(n = 6)$,



$$P(-1.96 < Z < 1.96) = P(-2.57 < T < 2.57) = 0.95.$$

T has a wider interval, due to the heavy tails because of the additional uncertainty of not knowing σ^2

$100(1-\alpha)\%$ C.I. for μ when σ^2 is unknown

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

More examples from textbook

7.1-3 7.1-4 7.1-6

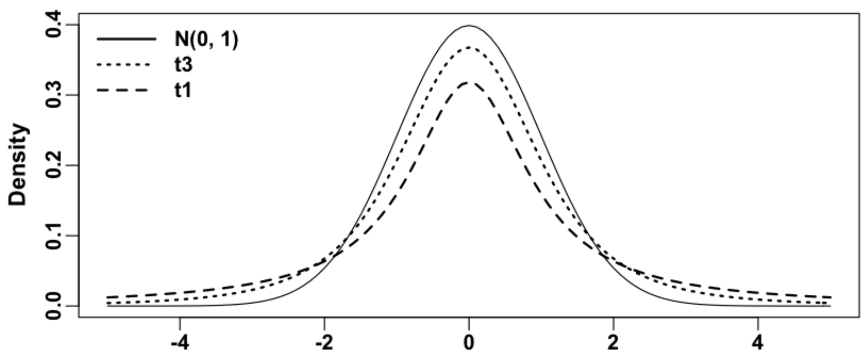
Example 7.1-5 The amount of butterfat in milk production (pounds) of a typical cow during 30 days, take a random sample of size 20 from $N(\mu, \sigma^2)$, both μ and σ^2 are unknown. Suppose $\bar{X} = 507.5$, $s = 89.75$. Find the 90% C.I. for μ.

$n=20 \quad 1-\alpha = 0.9 \quad \alpha = 0.1 \quad t_{\alpha/2}(n-1) = t_{0.05}(19) = 1.729$

$\left(507.5 - 1.729 \frac{89.75}{\sqrt{20}}, \ 507.5 + 1.729 \frac{89.75}{\sqrt{20}} \right)$

$(507.5 - 34.7, \ 507.5 + 34.7)$

As the sample size n increases, the t_{n-1} distribution approaches the standard Normal distribution.



Thus, if n is large (typically $n > 30$), $t_{n-1} \dot\sim N(0, 1)$, and a 95% confidence interval for μ becomes close to