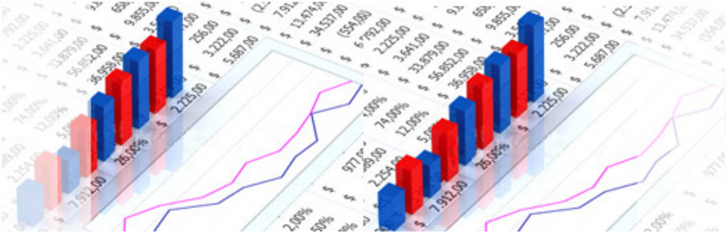
MTMS.01.099 Mathematical Statistics

Lecture 9. Chi-square distribution, t-distribution, Confidence interval for the mean

Tõnu Kollo



Fall 2017

Reminder: Interval Estimation

Let $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ be a sample from a distribution that depends on the unknown parameter θ.

Definition

An interval I_θ that covers θ with probability $1 - \alpha$ is called a **confidence interval** for θ at the *confidence level* $1 - \alpha$.

The left and right endpoints of the interval, $a_1(\boldsymbol{x})$ and $a_2(\boldsymbol{x})$, are called *confidence limits*.

Confidence limits $a_1(\boldsymbol{x})$ and $a_2(\boldsymbol{x})$ are functions of the values in the sample.
Given definition means that if $P(a_1(\boldsymbol{X}) \leq \theta \leq a_2(\boldsymbol{X})) = 1 - \alpha$, where $a_1(\boldsymbol{X})$ and $a_2(\boldsymbol{X})$ are functions of theoretical sample, then

$$I_\theta = (a_1(\boldsymbol{x}), a_2(\boldsymbol{x}))$$

is a confidence interval at the confidence level $1 - \alpha$.

Reminder: Interval Estimation (2)

Requirements for confidence interval:

- as narrow as possible
- confidence level as high as possible (closer to 1)

A very wide confidence interval gives the message that there is a great deal of uncertainty concerning the value of what we are estimating.

If confidence limits $a_1(\boldsymbol{X})$ and $a_2(\boldsymbol{X})$ are finite, the interval is then called **two-sided**, otherwise the interval is called **one-sided** $(-\infty, a_2), (a_1, \infty)$.

In principle, the probability $1 - \alpha$ can be fixed at any value. The classical confidence levels are

$$1 - \alpha = \begin{cases} 0.90 \\ 0.95 \\ 0.99 \end{cases}$$

Remarks

- Confidence interval for any parameter serves to estimate the parameter
- Although the parameter is just a number, the confidence interval is not – it's a much more sophisticated object
- Confidence interval estimate for a parameter has two "aspects" :
  - range of possible values
  - probability

Application to the Normal Distribution: Two important distributions

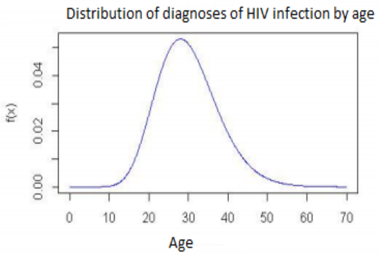
Recall from the idea of General Method (Lecture 8) that the key point when constructing a confidence interval is to find the distribution of a corresponding point estimator.

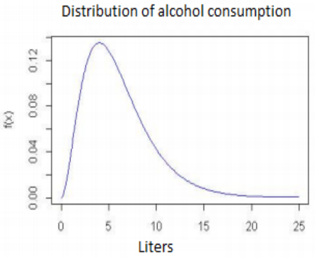
It is very common to construct interval estimates based on data from a normal distribution.

We consider two distributions: **Chi-square** and **t-distribution** (in the book, subsection Auxiliary Tools, p.227) before we go on with interval estimation.

Chi-square distribution (χ^2-distribution)

- χ^2-distribution is one of the most widely used probability distributions in inferential statistics, e.g., in hypothesis testing or in construction of confidence intervals





Chi-square distribution

Definition

A random variable X with a density function of the form

$$kx^{f/2-1}e^{-x/2} \quad (x \geq 0),$$

is said to have a χ^2 *distribution* with f degrees of freedom and k being the normalizing constant

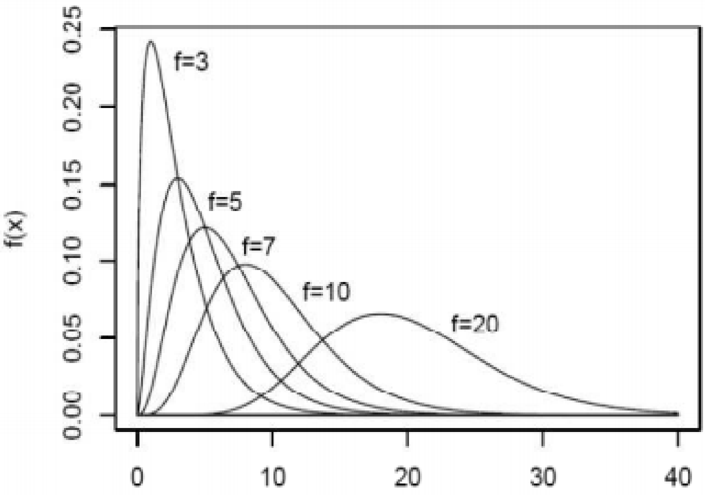
By setting $f = 1, 2, \ldots$ we obtain a whole family of distributions. All distributions in the family are skew, but **the larger is f, the more symmetric** is the distribution.

If $X \sim \chi^2(f)$, then $EX = f$ and $VarX = 2f$.

Chi-square distribution (2)

Remark

If f is large, then $\chi^2(f) \approx N(f, 2f)$.



Chi-square distribution. Relationship to the Gamma Distribution

- The χ^2 distribution is a special case of the gamma distribution $\Gamma(p, a)$ for $p = f/2, a = 2$, that is, $\chi^2 \equiv \Gamma(f/2, 2)$
- The constant k (from definition) is equal to $k = \frac{1}{2^{f/2}\Gamma(f/2)}$, where $\Gamma(\cdot)$ is a Gamma function

$$\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} dt$$

  - $\Gamma(z) = (z-1)\Gamma(z-1), z > 1$
  - $\Gamma(1) = 1$
  - $\Gamma(\frac{1}{2}) = \sqrt{\pi}$

Chi-square distribution

Property

If the random variables $X_1, \ldots, X_n$ are independent and have distributions $\chi^2(f_1), \chi^2(f_2), \ldots, \chi^2(f_n)$, respectively, then

$$Y = \sum_{i=1}^{n} X_i \sim \chi^2(\sum_{i=1}^{n} f_i).$$

Chi-square distribution. Relation to the Normal Distribution

Theorem 1

If $X_i, i = 1, \ldots, n$, are independent random variables and $X_i \sim N(0, 1)$, then

$$\sum_{i=1}^{n} X_i^2 \sim \chi^2(n)$$

and

$$\sum_{i=1}^{n} (X_i - \overline{X})^2 \sim \chi^2(n-1) \quad \left(\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i\right).$$

Hence the theorem states that both the sum of squares of independent standard normal variables and the sum of squares about their arithmetic mean have a χ^2 distribution, but with **different degrees of freedom.**

Chi-square distribution

Theorem 1 can be generalized to

Theorem 1*

If the random variables $X_i, i = 1, \ldots, n$ are independent and $N(\mu, \sigma^2)$, then

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2 \sim \chi^2(n)$$

and

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \overline{X})^2 \sim \chi^2(n-1).$$

Student's t-distribution

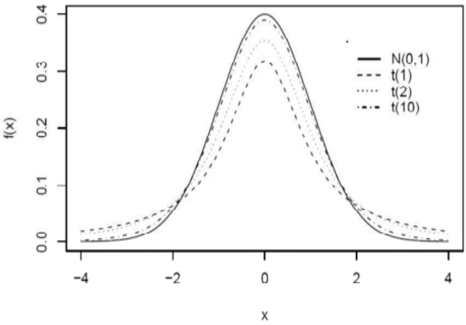
Definition

A random variable X with density function

$$k(1+\frac{x^2}{f})^{-(f+1)/2} \quad ,(-\infty < x < \infty),$$

where $k = \frac{\Gamma(\frac{f+1}{2})}{\sqrt{f\pi}\Gamma(\frac{f}{2})}$, $f \in \{1, 2, \ldots\}$ is said to have a *t-distribution* with f degrees of freedom.

Student's t-distribution (2)



- the distributions for $f = 1, 2, \ldots$ are all symmetric about $x = 0$
- If $X \sim t(f)$, then $EX = 0$. $VarX = f/(f-2)$, only if $f > 2$
- **when f increases to infinity, the t-distribution tends to a standard normal distribution**

Student's t-distribution (3)

Property

If $X \sim N(0,1)$ and $Y \sim \chi^2(f)$, where X and Y are independent, then

$$Z = \sqrt{f}\frac{X}{\sqrt{Y}} \sim t(f)$$

Theorem 2

If $X_i \sim N(\mu, \sigma^2)$ are independent random variables, $i = 1, 2, \ldots, n$, then

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

where $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, $S = \left[\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2\right]^{1/2}$.

A Single Random Sample. Confidence Interval for the **Mean**

Let $x_1, \ldots, x_n$ be a random sample from $N(\mu, \sigma^2)$. We want to construct a confidence interval for the mean μ.
We have the following result

Theorem 3

Let $x_1, \ldots, x_n$ be a random sample from $N(\mu, \sigma^2)$, where μ is unknown. Then

$$I_\mu = \overline{x} \pm \lambda_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \quad \text{if } \ \sigma \ \text{ known}$$

$$I_\mu = \overline{x} \pm t_{\alpha/2}(f) \frac{s}{\sqrt{n}}, \quad \text{if } \ \sigma \ \text{ unknown},$$

where s is a standard deviation of the sample, $\lambda_{\alpha/2}$ and $t_{\alpha/2}(f)$ are $\alpha/2$ complement quantiles of $N(0,1)$ and $t(f)$, $f = n - 1$.

A Single Random Sample. Confidence Interval for the **Mean** (2)

Proof.

If σ is known.

The arithmetic mean $\overline{X}$ is normally distributed,

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim N(\mu, \frac{\sigma^2}{n})$$

It follows that

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

Using the complement quantiles of $N(0, 1)$, the following holds (recall from previous lecture)

$$P(-\lambda_{\alpha/2} < \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} < \lambda_{\alpha/2}) = 1 - \alpha.$$

A Single Random Sample. Confidence Interval for the **Mean** (3)

Now we get

$$\begin{aligned} 1-\alpha &= P(-\lambda_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \overline{X}-\mu < \lambda_{\alpha/2}\frac{\sigma}{\sqrt{n}}) \\ &= P(\overline{X}-\lambda_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \overline{X}+\lambda_{\alpha/2}\frac{\sigma}{\sqrt{n}}) \end{aligned}$$

Replacing $\overline{X} = \overline{x}$, this means that

$$I_\mu = \overline{x} \pm \lambda_{\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

If σ is unknown.
We start again from

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

but since σ here is unknown, we use its estimator S.

A Single Random Sample. Confidence Interval for the **Mean** (4)

$S = \left[\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2\right]^{1/2}$

We know (from Theorem 2) that

$$\frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t(f), \; f = n - 1$$

So using the complement quantiles of $t(f)$, the following holds

$$P(-t_{\alpha/2}(f) < \frac{\overline{X} - \mu}{S/\sqrt{n}} < t_{\alpha/2}(f)) = 1 - \alpha.$$

Now we get

$$\begin{aligned} 1 - \alpha &= P(-t_{\alpha/2}(f)\frac{S}{\sqrt{n}} < \overline{X} - \mu < t_{\alpha/2}(f)\frac{S}{\sqrt{n}}) \\ &= P(\overline{X} - t_{\alpha/2}(f)\frac{S}{\sqrt{n}} < \mu < \overline{X} + t_{\alpha/2}(f)\frac{S}{\sqrt{n}}) \end{aligned}$$

A Single Random Sample. Confidence Interval for the **Mean** (5)

Replacing $\overline{X} = \overline{x}$ and $S = s$, we get

$$I_\mu = \overline{x} \pm t_{\alpha/2}(f)\frac{s}{\sqrt{n}}.$$

Theorem is proved.

Confidence Interval for the Mean

Example

Based on studies in England, it is known that for 13-year-old girls, the mean weight is 101 pounds with a standard deviation of 24.6 pounds. We assume that weights are normally distributed.

The public health officials in Liverpool are interested in the weights of the teens in their town: they suspect that the mean weight of their girls might be different from the given mean weight but assume that the variation is the same.

If they survey a random sample of 150 thirteen-year-old girls and find that their mean weight – an estimate of the population mean weight – is 95 pounds, find 95% confidence interval for the mean?

Confidence Interval for the Mean (2)

Example

We assume the 150 sample values are from a normal distribution, $N(\mu, 24.6^2)$.
Then the sampling distribution of mean weights is $N(\mu, 24.6^2/150)$. Let $\overline{X}$ denote the mean of the 150 weights, so standardizing gives $Z = (\overline{X} - \mu)/(24.6/\sqrt{150}) \sim N(0, 1)$.
For the standard normal random variable Z, we have $P(-1.96 < Z < 1.96) = 0.95$. Thus, we compute

$$\begin{aligned} 0.95 &= P(-1.96 < \frac{\overline{X} - \mu}{24.6/\sqrt{150}} < 1.96) \\ &= P(\overline{X} - 1.96(24.6/\sqrt{150}) < \mu < \overline{X} + 1.96(24.6/\sqrt{150})) \\ &= P(\overline{X} - 3.937 < \mu < \overline{X} + 3.937) \end{aligned}$$

Confidence Interval for the Mean (3)

Example

The random interval $P(\overline{X} - 3.937 < \mu < \overline{X} + 3.937)$ has a probability of 0.95 of covering the mean μ.

Now, once you have drawn your sample, the random variable $\overline{X}$ is replaced by the (observed) sample mean weight of $\overline{x} = 95$ and the interval (91.1, 98.9) is no longer a random interval.

We interpret this interval by stating that we are 95% confident that the population mean weight of 13-year-old girls in Liverpool is between 91.1 and 98.9 pounds.