# Introduction to Statistical Method

## Simultaneous Estimation of the Mean and Variance

### Chi Random Variable

Consider a problem:

- $z = (z_1, \cdots, z_n) \in \mathbb{R}^n$ randomly chosen.
- the value $z_k$ is determined by random variable $Z_k$, following a standard normal distribution.
- think about distribution about $\chi_n := \sqrt{\sum_{i=1}^n Z_i^2}$
- $\chi_n$ is a chi random variable, follows chi distribution with $n$ degree of freedom.

#### Cumulative Distribution Function

$$F_{\chi_n}(y) = P[\chi_n \le y] = P[\chi_n^2 \le y^2] = P[\sum_{i=1}^n Z_i^2 \le y^2] = \int_{\sum_{k=1}^n z_k^2 \le y^2} f_{Z_1 \dots Z_n}(z_1, \cdots, z_n) dz_1 \dots dz_n$$

Since they are $Z_1, \dots Z_n$ that $n$ independent standard variables, then we see the joint density:

$$f_{Z_1 \dots Z_n}(z_1, \dots z_n) = \frac{1}{(2\pi)^{n/2}} e^{-\sum_{k=1}^n z_k^2/2}$$

Thus $F_{\chi_n}(y) = \int_{\sum_{k=1}^n z_k^2 \le y^2} (2\pi)^{-n/2} e^{-\sum_{k=1}^n z_k^2/2} dz_1 \dots dz_n$

Apply polar coordinate with $(r, \theta_1, \dots, \theta_0)$ with $r > 0$, $0 < \theta_{n-1} < 2\pi$ and $-\pi/2 < \theta_k < \pi/2$ for $k = 1, \dots n-2$:

$$\begin{aligned} x_1 &= r \sin\theta_1 \\ x_2 &= r \cos\theta_1 \sin\theta_2 \\ x_3 &= r \cos\theta_1 \cos\theta_2 \sin\theta_3 \\ &\vdots \\ x_{n-1} &= r \cos\theta_1 \cos\theta_2 \dots \cos\theta_{n-2} \sin\theta_{n-1} \\ x_n &= r \cos\theta_1 \cos\theta_2 \dots \cos\theta_{n-2} \cos\theta_{n-1} \end{aligned}$$

Then the integral become: $$\begin{aligned} F_{\chi_n}(y) = &\int_0^{2\pi} \int_{-\pi/2}^{\pi/2} \cdots \int_{-\pi/2}^{\pi/2} \int_0^y (2\pi)^{-n/2} e^{-r^2/2} r^{n-1} \\ &\times D(\theta_1, \dots, \theta_{n-1}) \, dr \, d\theta_1 \dots d\theta_{n-2} \, d\theta_{n-1} \end{aligned}$$

Since $D(\theta_1, \dots \theta_{n-1})$ is independent of $r$, $C_n = (2\pi)^{-n/2} \int_0^{2\pi} \int_{-\pi/2}^{\pi/2} \cdots \int_{-\pi/2}^{\pi/2} D(\theta_1, \dots \theta_{n-1}) d\theta_1 \cdots d\theta_{n-1}$

we have $F_{\chi_n}(y) = C_n \int_0^y e^{-r^2/2} r^{n-1} dr$ .

Obviously, the $1 = \lim_{y \to \infty} C_n \int_0^\infty e^{-r^2/2} r^{n-1} dr = C_n \Gamma(\frac{n}{2}) 2^{n/2-1}$

Thus $C_n = (\Gamma(\frac{n}{2}) 2^{n/2-1})^{-1}$ and $f_{\chi_n}(y) = \frac{2}{\Gamma(\frac{n}{2}) 2^{n/2}} y^{n-1} e^{-y^2/2}$

## Chi-Squared Distribution

we hence derive from the $F_{\chi_n^2} = (\Gamma(\frac{n}{2})2^{n/2-1})^{-1} \int_0^{\sqrt{y}} e^{-r^2/2} r^{n-1} dr$

$$f_{\chi_n^2} = F'_{\chi_n^2}(y) = (\Gamma(\frac{n}{2})2^{n/2-1})^{-1} e^{-y/2} \sqrt{y}^{n-1} \cdot \frac{d}{dy}\sqrt{y} = \frac{1}{2^{n/2}\Gamma(\frac{n}{2})} y^{n/2-1} e^{-y/2}$$

## Sum of Independent Chi-Squared Variables

Given $\chi_m^2 = \sum_{i=1}^{m} X_i^2$ and $\chi_n^2 = \sum_{j=1}^{n} Y_j^2$ , then $\chi_{m+n}^2 = \chi_m^2 + \chi_n^2 = \sum_{i=1}^{m} X_i^2 + \sum_{j=1}^{n} Y_j^2$

It follows a chi-squared distribution, but with $m + n$ degree of freedom.

It extends to multi-addition case, trivial.

# Joint Sampling of Mean and Variance

In the previous chapter, we were able to analyze the sample mean, and also its distribution, under the assumption of known variance.

If variance $\sigma^2 = E[(X-\mu)^2]$ is unknown, then we must first see $S^2 = \frac{1}{n-1}\sum_{k=1}^{n}(X_k - \overline{X})^2$

So we are using the random sample $X_1, \dots X_n$ to get $\overline{X}$ and $S^2$ at same time.

So we are getting the joint distribution of $\overline{X}$ and $S^2$.

## Theorem

### Predicate

- $X_1, \dots X_n$ $n \geq 2$ be a random sample of size $n$.
- Normal distribution with $\mu$ and variance $\sigma^2$.

### Content

- The sample mean $\overline{X}$ is independent of the sample variance $S^2$.
- $\overline{X}$ is normally distributed with mean $\mu$ and $\sigma^2/n$.
- $(n-1)S^2/\sigma^2$ is chi-squared distributed with $n-1$ degree of freedom.

# Helmert Transformation

The Helmert transformation is a very special kind of orthogonal transformation from a set of $n \geq 2$ normal random variables $X_1, \dots X_n$ to a new set of random variables $Y_1, \dots Y_n$.

$$\begin{aligned}
Y_1 &= \frac{1}{\sqrt{n}}(X_1 + \cdots + X_n) \\
Y_2 &= \frac{1}{\sqrt{2}}(X_1 - X_2) \\
Y_3 &= \frac{1}{\sqrt{6}}(X_1 + X_2 - 2X_3) \\
&\vdots \\
Y_n &= \frac{1}{\sqrt{n(n-1)}}(X_1 + X_2 + \cdots + X_{n-1} - (n-1)X_n)
\end{aligned}$$

$$\begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \cdots & \frac{1}{\sqrt{n}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 & \cdots & 0 \\ \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & -\frac{2}{\sqrt{6}} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{1}{\sqrt{n(n-1)}} & \frac{1}{\sqrt{n(n-1)}} & \frac{1}{\sqrt{n(n-1)}} & \cdots & -\frac{n-1}{\sqrt{n(n-1)}} \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ \vdots \\ X_n \end{pmatrix}$$

The matrix $A$ is orthonormal since $A^{-1} = A^T$. This implies $|\det A| = 1$.

$$\sum_{i=1}^{n} y_i^2 = \langle y, y \rangle = \langle Ax, Ax \rangle = (Ax)^T(Ax) = x^T A^T Ax = \langle A^T Ax, x \rangle = \langle x, x \rangle = \sum_{i=1}^{n} x_i^2$$

Thus the joint distribution:

$$\begin{aligned} f_{X_1 \cdots X_n}(x_1, \dots, x_n) &= \prod_{i=1}^{n} (2\pi)^{-1/2} \sigma^{-1} e^{-\frac{1}{2\sigma^2}(x_i - \mu)^2} \\ &= (2\pi)^{-n/2} \sigma^{-n} e^{-\frac{1}{2\sigma^2} \sum\limits_{i=1}^{n} (x_i^2 - 2\mu x_i + \mu^2)} \\ &= (2\pi)^{-n/2} \sigma^{-n} e^{-\frac{1}{2\sigma^2} \left( \sum\limits_{i=1}^{n} x_i^2 - 2\mu \sum\limits_{i=1}^{n} x_i + n\mu^2 \right)} \end{aligned}$$

Apply back into the $y_n$,

$$\begin{aligned} & f_{Y_1 \cdots Y_n}(y_1, \dots, y_n) \\ &= f_{Y_1 \cdots Y_n}(\mathbf{y}) = f_{X_1 \cdots X_n}(\mathbf{x})_{\mathbf{x} = A^{-1}\mathbf{y}} \cdot \underbrace{|\det DA^{-1}(\mathbf{y})|}_{=1} \\ &= (2\pi)^{-n/2} \sigma^{-n} e^{-\frac{1}{2\sigma^2} \left( \sum\limits_{i=1}^{n} y_i^2 - 2\mu \sqrt{n} y_1 + n\mu^2 \right)} \\ &= (2\pi)^{-n/2} \sigma^{-n} e^{-\frac{1}{2\sigma^2} \left( \sum\limits_{i=2}^{n} y_i^2 + (y_1 - \sqrt{n}\mu)^2 \right)} \\ &= (2\pi)^{-1/2} \sigma^{-1} e^{-\frac{1}{2\sigma^2}(y_1 - \sqrt{n}\mu)^2} \prod_{i=2}^{n} (2\pi)^{-1/2} \sigma^{-1} e^{-\frac{1}{2\sigma^2} y_i^2} \end{aligned}$$

Then $f_{Y_1}(y_1) = (2\pi)^{-1/2} \sigma^{-1} e^{-\frac{1}{2\sigma^2}(y_1 - \sqrt{n}\mu)^2}$ and $f_{Y_i}(y_i) = (2\pi)^{-1/2} \sigma^{-1} e^{-\frac{1}{2\sigma^2} y_i^2}$ for $2 \le i \le n$.

$f_{Y_1 \dots Y_n}(y_1, \dots y_n) = f_{Y_1}(y_1) \cdot f_{Y_2}(y_2) \dots f_{Y_n}(y_n)$

So $Y_1$ is normally distributed with mean $\sqrt{n}\mu$ and variance $\sigma^2$, while $Y_2 \dots Y_n$ are having mean 0 and variance $\sigma^2$.

## Proof for Previous Theorem

So $\overline{X} = n^{-1/2} Y_1$ and

$$\begin{aligned} (n-1)S^2 &= \sum_{k=1}^{n} (X_k - \overline{X})^2 = \sum_{k=1}^{n} X_k^2 - 2\sum_{k=1}^{n} X_k \overline{X} + n\overline{X}^2 \\ &= \sum_{k=1}^{n} X_k^2 - n\overline{X}^2 = \sum_{k=1}^{n} Y_k^2 - Y_1^2 = \sum_{k=2}^{n} Y_k^2 \end{aligned}$$

Since $\overline{X} = n^{-1/2} Y_1$ and $f_{Y_1}(y_1) = (2\pi)^{-1/2} \sigma^{-1} e^{-\frac{1}{2\sigma^2}(y_1 - \sqrt{n}\mu)^2}$, then according to the rule in `ve401 note 3 page 6` that we get $f_{\overline{X}}(x) = (2\pi)^{-1/2} \sigma^{-1} e^{-\frac{1}{2\sigma^2}(\sqrt{n}x - \sqrt{n}\mu)^2} \sqrt{n}$

So the $\overline{X}$ is normally distributed with mean $\mu$ and variance $\sigma^2/n$.

$(n-1)S^2/\sigma^2 = \frac{1}{\sigma^2} \sum_{k=2}^{n} Y_k^2 = \sum_{k=2}^{n} (\frac{Y_k}{\sigma})^2$ is a chi-squared distribution with $n-1$ freedom.

## Independence of Sample Mean and Sample Variance in more General Form

The converse result for the previous theorem is also true:

$X_1 \dots X_n$, with $n \ge 2$ be independent identical distributed random variables.

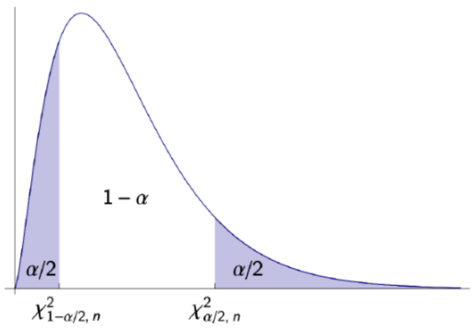
If $\overline{X}$ and $S^2$ are independent, then $X_k$ with $k = 1 \dots n$ follows normal distribution.

This means that the independence of $\overline{X}$ and $S^2$ is a characteristic property of the normal distribution. If in a given situation we assume that $\overline{X}$ and $S^2$ are independently distributed we essentially assuming that the population is normally distributed.

## Interval Estimation of Variability

We let $0 < \alpha < 1$ and we define $\chi^2_{1-\alpha/2,n} \le \chi^2_{\alpha/2,n} \in \mathbb{R}$

We use the previous theorem to find a confidential interval for the variance based on sample variance $S^2$.



Thus we get $1-\alpha = P[\chi^2_{1-\alpha/2,n-1} \le \dfrac{(n-1)S^2}{\sigma^2} \le \chi^2_{\alpha/2,n-1}]$

Then $1-\alpha = P[\dfrac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}} \le \sigma^2 \le \dfrac{(n-1)S^2}{\chi^2_{1-\alpha/2,n-1}}]$, which is the $100(1-\alpha)\%$ confidence interval for $\sigma^2$.

## Interval Estimation for Mean in Variance Unknown

If we know the variance, then $Z = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$.

Our main goal is to derive a general formula for a confidence interval on the mean when the value of $\sigma$ is not known and must be estimated.

So we should know the distribution of $\frac{\overline{X}-\mu}{S/\sqrt{n}}$.

## T-distribution

- $Z$ is a standard normal variable.
- $\chi^2_\gamma$ be an independent chi-squared random variable with $\gamma$ degree of freedom.
- $T_\gamma = \frac{Z}{\sqrt{\chi^2_\gamma/\gamma}}$ is a T-distribution with $\gamma$ degree of freedom.

The density for a T distribution with $\gamma$ degrees of freedom is given by:

$f_{T_\gamma}(t) = \dfrac{\Gamma((\gamma+1)/2)}{\Gamma(\gamma/2)\sqrt{\pi\gamma}}(1+\dfrac{t^2}{\gamma})^{-\frac{\gamma+1}{2}}$ check page 332 to 335

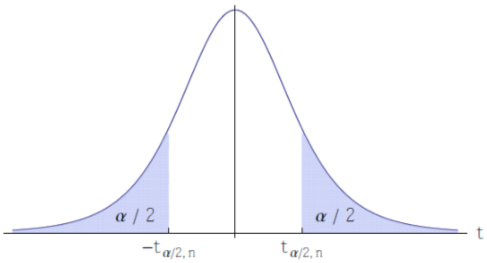
So $X_1 \dots X_n$ be a random sample from a normal distribution with mean $\mu$ and variance $\sigma^2$.

The random variable $T_{n-1} = \frac{\overline{X}-\mu}{S/\sqrt{n}}$ follows T distribution with $n-1$ degree of freedom.

(Come from the $(\overline{X}-\mu)/(\sigma/\sqrt{n})$ as standard normal and $(n-1)S^2/\sigma^2$ as chi-squared with $n-1$ degree of freedom, put them into $T_\gamma$ )

## Interval Estimation of Mean with Variance Unknown

We define $t_{\alpha/2,n} \geq 0$ by $\int_{t_{\alpha/2,n}}^{\infty} f_{T_n}(t)dt = \alpha/2$



### Theorem

$X_1, \dots X_n$ be a random sample of size $n$ from a normal distribution with mean $\mu$ and variance $\sigma^2$.

The $100(1-\alpha)\%$ confidential interval on $\mu$ is given by $\overline{X} \pm t_{\alpha/2,n-1}S/\sqrt{n}$

## Tolerance Limits

A tolerance interval determined from a sample of size $n$ consists of two numbers $L1, L2$, called tolerance limits.

So it goes like $(1-\alpha) \cdot 100\%$ certainty at least $\delta \cdot 100\%$ of population lies between $L1, L2$.

Since we know the fact that $95\%$ of the population lies in the interval $\mu \pm 1.96\sigma$.

The $\overline{X} \pm 1.96S$ will not always cover $95\%$ of the population since it is a random interval.

### Theorem: Two-sided Tolerance Limits

$X$ is normally distributed random variable with $\overline{X}$ and $S^2$ from a sample of size $n$.

$\exists K = K(n, \alpha, \delta)S$ covers at least $\delta \cdot 100\%$ of population with $(1-\alpha) \cdot 100\%$ confidence.

So the $K$ is a two-sided tolerance limit.