Estimation

Antonio Falcó

Unit 6

Question

We believe for a population that DBP $\sim N(\mu, \sigma^2)$. *How can the parameters of the distribution* (μ, σ^2) *be estimated ?*

Procedure

- Assume that the size of the population is very high $\approx \infty$
- Take a sample of size n, namely $x_1, \ldots, x_n$
- Compute

$$\overline{x} = \overline{x}(x_1, \ldots, x_n) = \frac{1}{n} \sum_{i=1}^{n} x_i$$

and

$$s^2 = s^2(x_1, \ldots, x_n) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x}(x_1, \ldots, x_n))^2.$$

- Use $(\overline{x}(x_1, \ldots, x_n), s^2(x_1, \ldots, x_n)) \approx (\mu, \sigma^2)$.

Consider as data population the $\{1, 2, 3, 4\}$ an take samples of size $n = 2$ from this data set. **We assume that** $(x_1, x_2) = (x_2, x_1)$.

Table:

x_1	x_2	$(\overline{x}(x_1, x_2), s^2(x_1, x_2))$
1	2	(1.50,0.50)
1	3	(1.41,2.00)
1	4	(2.50,4.50)
2	3	(2.50,0.50)
2	4	(3.00,2.00)
3	4	(3.50,0.50)

Observe $x_1 = \{1, 2, 3, 4\}$ and $x_2 = \{1, 2, 3, 4\}$

Definition

A *random sample* is a selection of some members of the population such that each member is independly choosen and has know nonzero probability of being selected.

Definition

A *simple random sample* is a random sample in which each group member has the same probability of being selected.

Definition

The *reference, target or study population* is the group we want to study. The random sample is selected from the study population.

Remark

- *From now one, we identify our target with a random variable X defined over our population subject to study.*
- *We assume that the random variable X is completely characterized by a set of parameter values $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_n)$.*

Example

- $X \sim B(n, p)$ here $\boldsymbol{\theta} = (n, p)$.
- $X \sim \mathcal{P}(\mu)$ here $\boldsymbol{\theta} = (\mu)$.
- $X \sim N(\mu, \sigma^2)$ here $\boldsymbol{\theta} = (\mu, \sigma^2)$.

Consider as target $X = \{1, 2, 3, 4\}$ where $Pr(X = x) = 0.25$. Then we have

$$E[X] = \mu = 2.50 \text{ and } Var(X) = \sigma^2 = 1.67$$

x_1	x_2	$(\overline{x}(x_1, x_2), s^2(x_1, x_2))$
1	2	(1.50,0.50)
1	3	(1.41,2.00)
1	4	(2.50,4.50)
2	3	(2.50,0.50)
2	4	(3.00,2.00)
3	4	(3.50,0.50)

The probability of a sample is $1/6$. Then, the average and the standard deviation generate two new random variables:

$$\overline{X} = \{1.50, 1.41, 2.50, 2.50, 3.00, 3.50\}$$

and

$$S^2 = \{0.50, 2.00, 4.50, 0.50, 2.00, 0.50\}.$$

Remark

- $E[\overline{X}] = \frac{1.50+1.41+2.50+2.50+3.00+3.50}{6} = 2.50 = E[X]$
- $E\left[\frac{6}{5}\, S^2\right] = \frac{6}{5}\, \frac{0.50+2.00+4.50+0.50+2.00+0.50}{6} = 1.67 = Var(X)$

Framewok

Consider a target given by the random variable X and consider a simple random sample of size n, namely,

$$X_1, \ldots, X_n.$$

Then we define the *sample mean of size n* as

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

Example

A supervisor has six employees, whose experiences (in terms of years on the job) are

$$2, 4, 6, 6, 7, 8.$$

(a) Construct all possibles samples of four observations.

(b) Compute all sample means.

Solution

SAMPLE	SAMPLE MEAN	SAMPLE	SAMPLE MEAN
2, 4, 6, 6	4.50	2, 6, 7, 8	5.75
2, 4, 6, 7	4.75	2, 6, 7, 8	5.75
2, 4, 6, 8	5.00	4, 6, 6, 7	5.75
2, 4, 6, 7	4.75	4, 6, 6, 8	6.00
2, 4, 6, 8	5.00	4, 6, 7, 8	6.25
2, 4, 7, 8	5.25	4, 6, 7, 8	6.25
2, 6, 6, 7	5.25	6, 6, 7, 8	6.75
2, 6, 6, 8	5.50		

Proposition

Let $X_1, \ldots, X_n$ be a simple random sample from some target X with mean μ. Then the sample mean $\overline{X}$ satisfy

$$E(\overline{X}) = \mu.$$

Example

A supervisor has six employees, whose experiences (in terms of years on the job) are

$$2, 4, 6, 6, 7, 8.$$

Compute the target mean by means the probability distribution of the sample mean.

Estimation Framewok

- To estimate the parameter $\mu = E(X)$, denoted in general by θ, we use the estimator $\overline{X} = \overline{X}(X_1, \ldots, X_n)$, denoted in general by $\widehat{\theta} = \widehat{\theta}(X_1, \ldots, X_n)$.
- How large is the estimation error ? We can compute it by means

$$E((\overline{X} - \mu)^2) = E((\overline{X}(X_1, \ldots, X_n) - \mu)^2) = Var(\overline{X}).$$

- In general, if $E(\widehat{\theta}) = \theta$:

$$E\left(\left(\widehat{\theta} - \theta\right)^2\right) = E\left(\left(\widehat{\theta}(X_1, \ldots, X_n) - \theta\right)^2\right) = Var(\widehat{\theta}).$$

Proposition

Let $X_1, \ldots, X_n$ be a simple random sample from some target X with mean μ and variance σ^2. Then

$$Var(\overline{X}) = \frac{\sigma^2}{n}.$$

Proposition

Let $X_1, \ldots, X_n$ be a simple random sample from some study population X normally distributed with mean μ and variance σ^2. Then $\overline{X}$ is normally distributed with mean μ and variance $\frac{\sigma^2}{n}$, i.e.

$$\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$$

Theorem (Central Limit Theorem)

Let $X_1, \ldots, X_n$ be a sample from a population X with mean μ and standard deviation σ. Then for large n, (say us $n \geq 30$)

$$\overline{X} \approx N\left(\mu, \frac{\sigma^2}{n}\right).$$

where $\approx$ means "approximately distributed". Thus,

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

Interval Estimation

If $Z \sim N(0,1)$ then it is well-known that

$$Pr(-1.96 \leq Z \leq 1.96) \approx 0.95.$$

From the Central Limit Theorem (CLT)

$$Pr\left(-1.96 \leq \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq 1.96\right) \approx 0.95$$

$$Pr\left(-1.96\,\frac{\sigma}{\sqrt{n}} \leq \overline{X} - \mu \leq 1.96\,\frac{\sigma}{\sqrt{n}}\right) \approx 0.95$$

$$Pr\left(\overline{X} - 1.96\,\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + 1.96\,\frac{\sigma}{\sqrt{n}}\right) \approx 0.95$$

Therefore, 95% of all sample means will fall within the interval

$$(\overline{X} - 1.96\frac{\sigma}{\sqrt{n}}, \overline{X} + 1.96\frac{\sigma}{\sqrt{n}})$$

Example

Assume tha we previously know that for a children population the DBP (mm Hg) has a standard deviation $\sigma = 4.57$. From a sample of size $n = 35$ we obtain $\overline{x} = 64.0$. What we can say ? From CLT:

$$\text{“} \qquad Pr\left(64.0 - 1.96\,\frac{4.57}{\sqrt{35}} \leq \mu \leq 64.0 + 1.96\,\frac{4.57}{\sqrt{35}}\right) \approx 0.95. \qquad \text{”}$$

This implies that the real mean μ satisfies from our data that:

$$\text{“} \qquad Pr\,(62.409 \leq \mu \leq 65.591) \approx 0.95. \qquad \text{”}$$

The interval $(62.409, 65.591)$ is the 95%–confidence interval for the mean μ.

Bad News

In real life problems the standard deviation σ of the population X is unknown.

Theorem

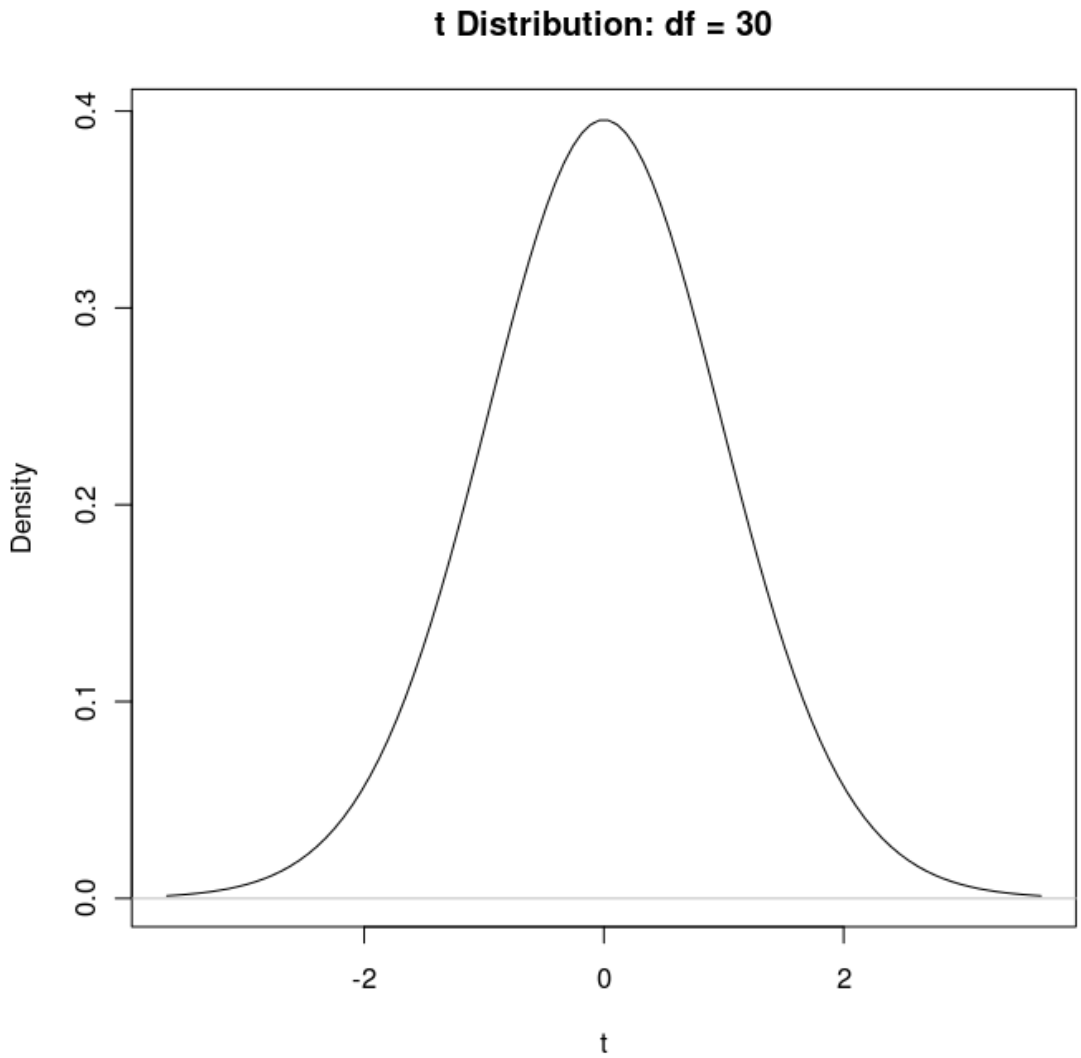
If $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ and are independent, then

$$\frac{\overline{X} - \mu}{\frac{S}{\sqrt{n}}} \sim t_{n-1}.$$

where

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

and t_{n-1} follows a t-distribution with $n-1$ degrees of freedom.


t Distribution: df = 30
Density
0.0
0.1
0.2
0.3
0.4
-2
0
2
t

Example

Assume tha we study for a children population the DBP (mm Hg) From a sample of size $n = 35$ we obtain $\overline{x} = 64.0$ and $s = 4.57$. What we can say ? First,

$$S = s\sqrt{\frac{n}{n-1}}, \text{ thus } S = 4.57\sqrt{\frac{35}{34}} = 4.6367.$$

By using R-commander:

$$Pr(-2.03 \leq t_{34} \leq 2.03) \approx 0.95.$$

$$\text{“} \quad Pr\left(64.0 - 2.03\,\frac{4.6367}{\sqrt{35}} \leq \mu \leq 64.0 + 2.03\,\frac{4.6367}{\sqrt{35}}\right) \approx 0.95. \quad \text{”}$$

This implies that the real mean μ satisfies from our data that:

$$\text{“} \qquad Pr\,(62.486 \leq \mu \leq 65.514) \approx 0.95. \qquad \text{”}$$

For 95% of all random samples, the interval will contain the real mean μ of the population

Argument

$$Pr\left(-2.03 \leq \frac{\overline{X}-\mu}{\frac{S}{\sqrt{35}}} \leq 2.03\right) \approx 0.95$$

$$Pr\left(-2.03\,\frac{S}{\sqrt{35}} \leq \overline{X}-\mu \leq 2.03\,\frac{S}{\sqrt{35}}\right) \approx 0.95$$

$$Pr\left(\overline{X}-2.03\,\frac{S}{\sqrt{35}} \leq \mu \leq \overline{X}+2.03\,\frac{S}{\sqrt{35}}\right) \approx 0.95.$$

Confidence interval for the Mean of a N.D.

Let $t_{n-1,\alpha/2}$ the value obtained from a t-distribution t_{n-1} by

$$Pr(t_{n-1} \leq t_{n-1,\alpha/2}) = 1 - \frac{\alpha}{2}$$

or

$$Pr(-t_{n-1,\alpha/2} \leq t_{n-1} \leq t_{n-1,\alpha/2}) = 1 - \alpha.$$

Then

$$Pr\left(\overline{X} - t_{n-1,\alpha/2}\frac{S}{\sqrt{n}} \leq \mu \leq \overline{X} + t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}\right) = 1 - \alpha.$$

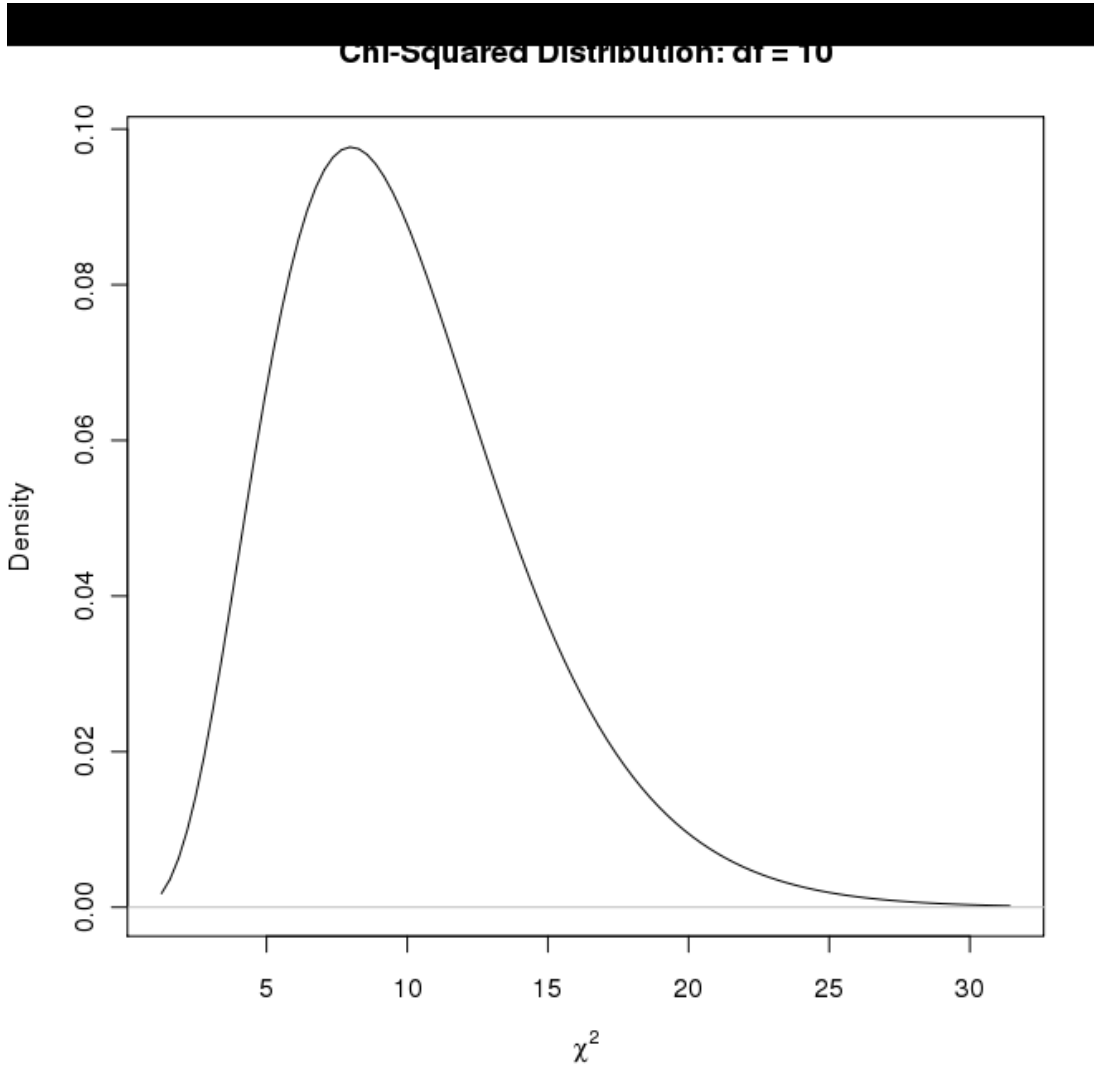
Glass ionomer cement (GI) is one of several new classes of adhesive luting agents recently introduced as an alternative to zinc phosphate cement. GI has been shown to reduce microleakage, increase retention, and improve physical properties, compared with zinc phosphate cements. Different factors might influence the hydraulic response and therefore the film thickness of different classes of adhesive luting agents. Let X be a random variable representing the film thickness of GI. White et al. measured the film thickness (micrometers) of GI with a load of $5\,\mathrm{kg}$ applied vertically to the plates. The sample mean and sample SD of 10 samples are $\overline{X} = 19.9\,\mathrm{kg}$ and $S = 1.3\,\mathrm{kg}$. If X is normally distributed, find a 95% confidence interval for μ.

Definition

Assume that $X_1, \ldots, X_n$ are independent $N(0, 1)$. Then

$$G = \sum_{i=1}^{n} X_i^2 \sim \chi_n^2$$

follows a *Chi-Square distribution with n degrees of freedom (df).*


Chi-Squared Distribution: df = 10
Density
0.00
0.02
0.04
0.06
0.08
0.10
5
10
15
20
25
30
χ^2

Proposition

Assume that $X_1, \ldots, X_n$ *are independent* $N(\mu, \sigma^2)$. *Then it can be shown that*

$$S^2 \sim \frac{\sigma^2}{n-1}\chi^2_{n-1},$$

or equivalently

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}.$$

Procedure

- Fix $\alpha(= 0.10)$ then $1 - \alpha(= 0.90)$.
- Given the sample size $n(= 10)$ compute

$$Pr(z_{\alpha/2} \leq \chi^2_{n-1} \leq z'_{\alpha/2}) = 1 - \alpha.$$

qchisq(c(0.05), df=9, lower.tail=TRUE)
[1] 3.325113
qchisq(c(0.05), df=9, lower.tail=FALSE)
[1] 16.91898
This implies $z_{\alpha/2} = 3.325113$ and $z'_{\alpha/2} = 16.91898$.

- Then

$$Pr\left(3.325113 \leq \frac{9S^2}{\sigma^2} \leq 16.91898\right) = 0.90$$

or

$$Pr\left(\frac{9S^2}{16.91898} \leq \sigma^2 \leq \frac{9S^2}{3.325113}\right) = 0.90$$

Binomial Distribution

Assume $X \sim B(n, p)$. Then consider $\widehat{p} = \frac{X_1 + \ldots + X_n}{n}$ here $X_1 = \{0, 1\}$. Compute $z_{\alpha/2}$ such that

$$Pr(-z_{\alpha/2} \leq N(0,1) \leq z_{\alpha/2}) = 1 - \alpha.$$

Then

$$Pr\left(\widehat{p} - z_{\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} \leq p \leq \widehat{p} + z_{\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}\right) = 1 - \alpha.$$