Distributions of Sampling Statistics

CSF2600102 – Statistics and Probability

Fakultas Ilmu Komputer
Universitas Indonesia
2014

Credits

These course slides were prepared by Alfan F. Wicaksono. **Suggestions**, **comments**, and **criticism** regarding these slides are welcome. Please kindly send your inquiries to alfan@cs.ui.ac.id.

Technical questions regarding the topic should be directed to current lecturer team members:

- Ika Alfina, S.Kom., M.Kom.
- Prof. T. Basaruddin, Ph.D.
- Alfan F. Wicaksono

The content was based on previous semester's (odd semester 2013/2014) course slides created by **all previous lecturers**.

References

- Introduction to Probability and Statistics for Engineers & Scientists, 4th ed.,
  - Sheldon M. Ross, Elsevier, 2009.
- A First Course in Probability, 8th Edition.
  - Sheldon M. Ross
- Applied Statistics for the Behavioral Sciences, 5th Edition,
  - Hinkle., Wiersma., Jurs., Houghton Mifflin Company, New York, 2003.
- Probability and Statistics for Engineers & Scientists, 4^{th} Edition
  - Anthony J. Hayter, Thomson Higher Education

Outline

- Introduction
- Sample Mean
- Central Limit Theorem
  - Approximate Distribution of The Sample Mean
  - How Large a Sample is Needed ?
- Sample Variance
- Sampling Distribution from Normal Distribution
  - Distribution of Sample Mean (Normal Population)
  - Joint Distribution of Sample Mean & Variance
- Sampling from a Finite Population

To use sample data to make inference about an entire population, it is necessary to assume that there is an underlying (population) probability distribution.

Definition

If $\mathbf{X_1}$, $\mathbf{X_2}$, **...**, $\mathbf{X_n}$ are **indepedent** random variables having a common distribution **F**, then we say that they constitute a ***sample*** (or ***random sample***) from the distribution **F**.

Goal: to make inferences about a (population) distribution F using the samples taken from F.

- **Parametric inference problem**: The form of F is specified up to a set of unknown parameters, e.g.:
  - F is assumed as a normal distribution function having an unknown mean and variance

- **Nonparametric inference problem**: nothing is assumed about the form of F

Sample Mean

Let $\mathbf{X_1, X_2, ..., X_n}$ be a sample of values from a population having *expectation* $\boldsymbol{\mu}$ and *variance* $\boldsymbol{\sigma^2}$.

The sample Mean is:

$$\bar{X} = \frac{X_1 + X_2 + ... + X_n}{n}$$

$\bar{X}$ is also a **random variable** since X_i is a random variable in the sample

Expectation & Variance of Sample Mean

$$E[\bar{X}] = E\left[\frac{X_1 + X_2 + \ldots + X_n}{n}\right]$$

$$= \frac{1}{n}\left(E[X_1] + \ldots + E[X_n]\right)$$

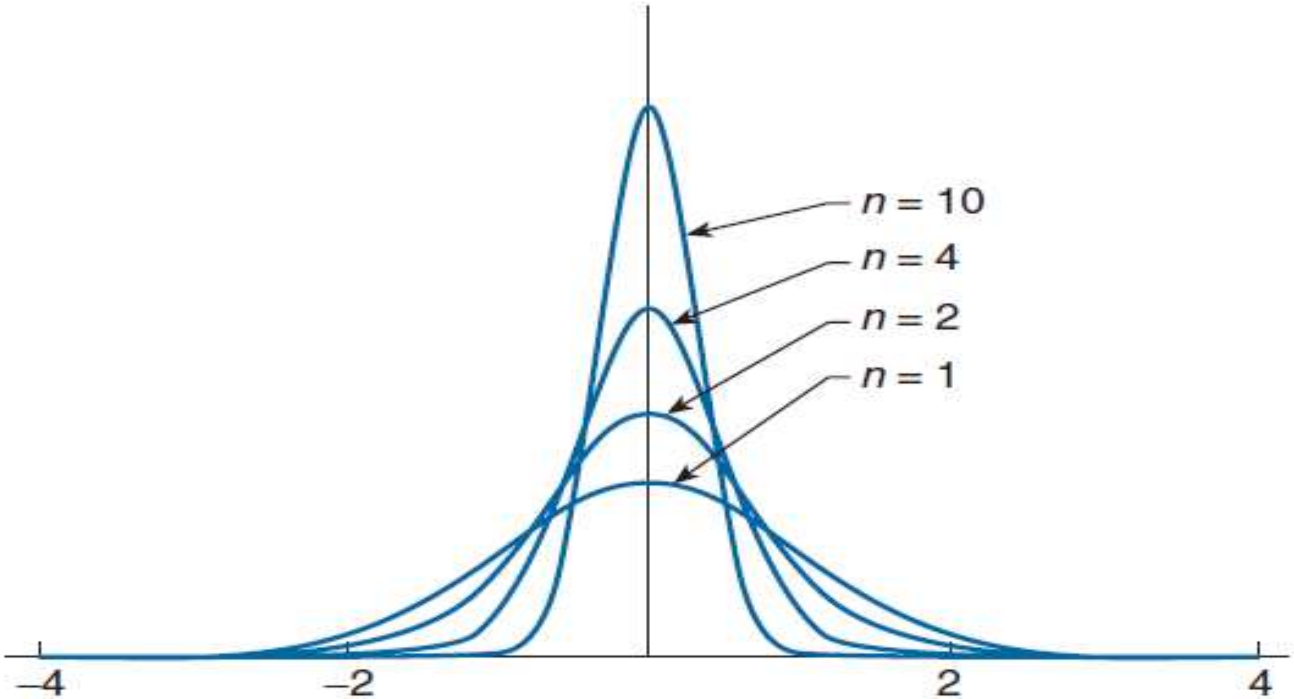
$$= \mu$$

$$Var(\bar{X}) = Var\left(\frac{X_1 + X_2 + \ldots + X_n}{n}\right)$$

$$= \frac{1}{n^2}\left(Var(X_1) + \ldots + Var(X_n)\right)$$ By independence

$$= \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

$\bar{X}$ is also centered about the population mean μ, but its spread becomes more and more reduced as the sample size increases.



Densities of sample means from a standard normal population.

Central Limit Theorem

Let $\mathbf{X_1}$, $\mathbf{X_2}$, **...**, $\mathbf{X_n}$ be a sequence of *independent and identically distributed (i.i.d)* random variables each having mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\sigma^2}$.

$$X_i \overset{iid}{\sim} D(\mu, \sigma^2)$$

Then, for **n large**, the distribution of

$$X_1 + X_2 + \ldots + X_n$$

is **approximately normal** with

$$X_1 + X_2 + \ldots + X_n \sim N(n\mu, n\sigma^2)$$

It follows from Central Limit Theorem that, for **n large**

$$\frac{X_1 + X_2 + \ldots + X_n - n\mu}{\sigma\sqrt{n}} \sim N(0,1)$$

Standard normal random variable

$$P\left(\frac{X_1 + X_2 + \ldots + X_n - n\mu}{\sigma\sqrt{n}} < x\right) = P(Z < x)$$

Example

An insurance company has 25,000 automobile policy holders. If the yearly claim of a policy holder is a random variable with **mean 320** and **standard deviation 540**, approximate the probability that the total yearly claim exceeds 8.3 million.

Let **X** denote the total yearly claim. Number the policy holders, and let $\mathbf{X_i}$ denote the yearly claim of policy holder i.

With n = 25000, we have from the central limit theorem that $X = \sum_{i=1}^{n} X_i$ will have approximately a normal distribution with

$$X = \sum_{i=1}^{n} X_i \sim N\left(\mu, \sigma^2\right) \qquad \begin{aligned} \mu &= 320 \times 25000 = 8 \times 10^6 \\ \sigma &= 540\sqrt{25000} = 8.5381 \times 10^4 \end{aligned}$$

$$\begin{aligned} P\left(X > 8.3 \times 10^6\right) &= P\left(\frac{X - 8 \times 10^6}{8.5381 \times 10^4} > \frac{8.3 \times 10^6 - 8.3 \times 10^6}{8.5381 \times 10^4}\right) \\ &\approx P(Z > 3.51) \\ &\approx 0.00023 \end{aligned}$$

Approximate Distribution of The Sample Mean

Let $\mathbf{X_1, X_2, ..., X_n}$ be a sample of values from a population having *expectation* $\boldsymbol{\mu}$ and *variance* $\boldsymbol{\sigma^2}$.

It follows from central limit theorem that $\bar{X}$ will be approximately normal when sample **size *n* is large**.

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

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

Standard normal distribution

Approximate Distribution of The Sample Mean

Supaya lebih paham ...

$$E[\bar{X}] = \mu$$

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \Rightarrow \quad Var(\bar{X}) = \frac{\sigma^2}{n}$$

$$SD(\bar{X}) = \sqrt{Var(\bar{X})} = \sigma \big/ \sqrt{n}$$

$$\frac{\bar{X} - E[\bar{X}]}{SD(\bar{X})} = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \sim N(0,1)$$

The weights of a population of workers have **mean of 167** and **standard deviation of 27**.

a) If a sample set of 36 workers is chosen, approximate the probability that the sample mean of their weights lies between 163 and 170.

b) Repeat part (a) when the sample is of size 144.

(a) It follows from central limit theorem that

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \qquad \begin{aligned} &\mu = 167 \\ &\sigma/\sqrt{n} = 27/\sqrt{36} = 4.5 \end{aligned}$$

$$\begin{aligned} P(163 < \bar{X} < 170) &= P\left(\frac{163-167}{4.5} < \frac{\bar{X}-167}{4.5} < \frac{170-167}{4.5}\right) \\ &\approx 2P(Z < 0.8889) - 1 \\ &\approx 0.6259 \end{aligned}$$

(b) It follows from central limit theorem that

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \qquad \begin{aligned} &\mu = 167 \\ &\sigma/\sqrt{n} = 27/\sqrt{144} = 2.25 \end{aligned}$$

$$\begin{aligned} P(163 < \bar{X} < 170) &= P\left(\frac{163-167}{2.25} < \frac{\bar{X}-167}{2.25} < \frac{170-167}{2.25}\right) \\ &\approx 2P(Z < 1.7778) - 1 \\ &\approx 0.9246 \end{aligned}$$

How Large a Sample is Needed ?

- A general rule of thumb is that one can be confident of the normal approximation whenever the sample size **n** is at least 30.

- That is, practically speaking, no matter how non normal the underlying population distribution is, the sample mean of a sample of size at least 30 will be approximately normal.

- In most cases, the normal approximation is valid for much smaller sample sizes.

How Large a Sample is Needed ?

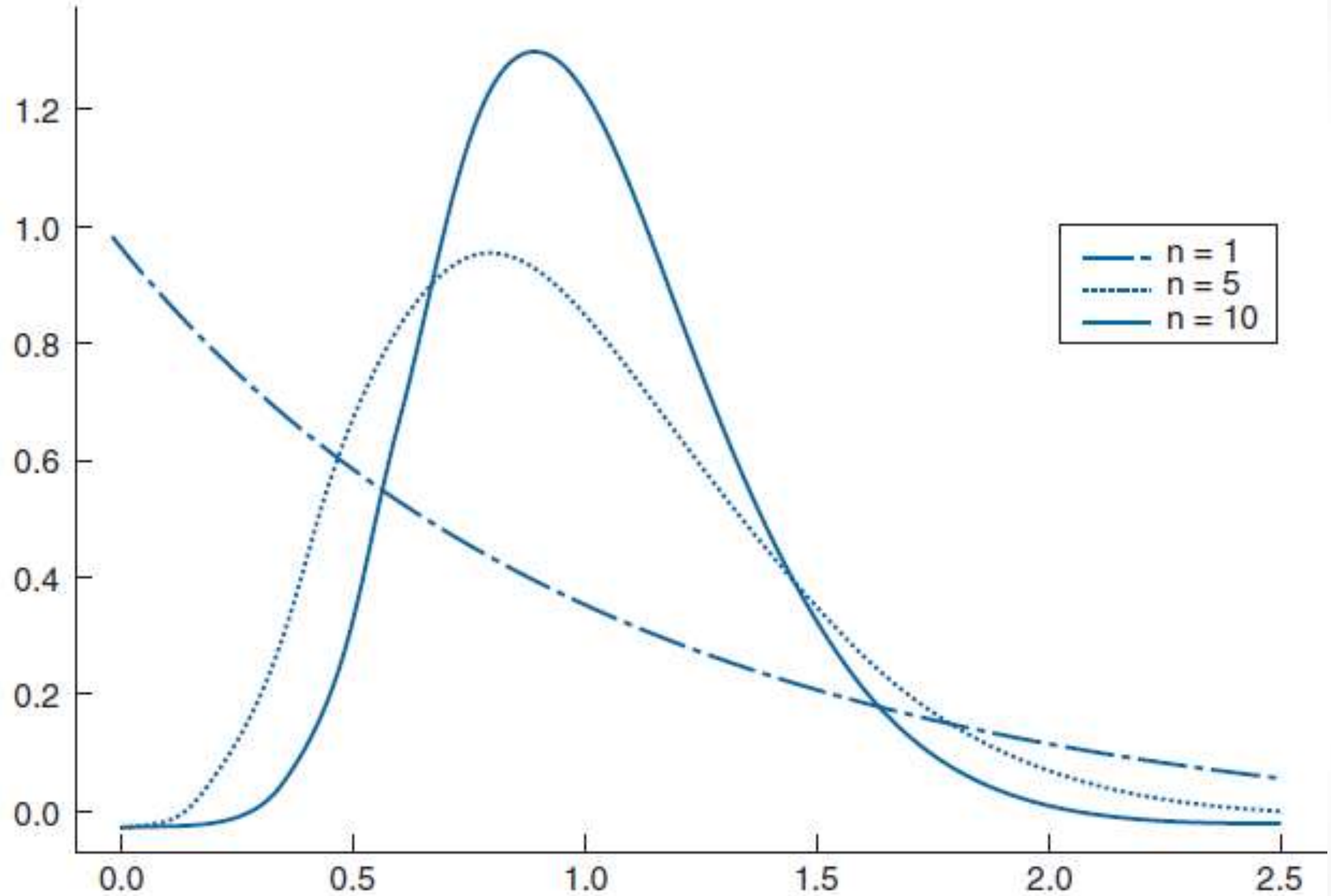


FIGURE 6.4 *Densities of the average of n exponential random variables having mean 1.*

The Sample Variance

Let $\mathbf{X_1}$, $\mathbf{X_2}$, ..., $\mathbf{X_n}$ be a sample of values from a population having *expectation* $\boldsymbol{\mu}$ and *variance* $\boldsymbol{\sigma^2}$.

Recall the definition of **sample variance**:

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

It's clear that sample variance is also a random variable.

The Expectation

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

$$\begin{aligned}(n-1)E[S^2] &= E\left[\sum_{i=1}^{n} X_i^2\right] - nE[\overline{X}^2] \\ &= nE[X_1^2] - nE[\overline{X}^2] \\ &= nVar(X_1) + n(E[X_1])^2 - nVar(\overline{X}) - n(E[\overline{X}])^2 \\ &= n\sigma^2 + n\mu^2 - n(\sigma^2/n) - n\mu^2 \\ &= (n-1)\sigma^2\end{aligned}$$

$$E[S^2] = \sigma^2$$

Sampling Distributions from A Normal Population

Let $\mathbf{X_1}$, $\mathbf{X_2}$, **...**, $\mathbf{X_n}$ be a sample of values from a **NORMAL** population having *expectation* $\boldsymbol{\mu}$ and *variance* $\boldsymbol{\sigma^2}$.

That is, they are **independent** and

$$X_i \sim N\left(\mu, \sigma^2\right)$$

Recall that mean & variance of the sample

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

We want to compute their distributions !

Distribution of The Sample Mean

Since the sum of independent normal random variables is normally distributed, it follows that $\bar{X}$ **is also a normal R.V.**:

$$E[\bar{X}] = \sum_{i=1}^{n} \frac{E[X_i]}{n} = \mu$$

$$Var(\bar{X}) = \frac{1}{n^2} \sum_{i=1}^{n} Var(X_i) = \frac{\sigma^2}{n}$$

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

Joint Distribution of $\bar{X}$ and S^2

Let $\mathbf{X_1}$, $\mathbf{X_2}$, **…**, $\mathbf{X_n}$ be a sample of values from a **NORMAL** population having *expectation* **μ** and *variance* $\boldsymbol{\sigma^2}$.

Then, **$\bar{X}$ and S^2** are **independent random variables** with

(1) $$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

(2) $$(n-1)\frac{S^2}{\sigma^2} \sim \chi^2_{n-1}$$ **being a chi-square with n-1 degrees of freedom**

Corollary

Let $\mathbf{X_1}$, $\mathbf{X_2}$, **...,** $\mathbf{X_n}$ be a sample of values from a **NORMAL** population having *expectation* $\boldsymbol{\mu}$ and *variance* $\boldsymbol{\sigma^2}$.

Then,

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

That is, $\sqrt{n}\,\dfrac{\left(\overline{X}-\mu\right)}{S}$ has a **t-distribution** with **n-1 degrees of freedom**.

Another Proposition ...

Let $\mathbf{X_1}$, $\mathbf{X_2}$, **...**, $\mathbf{X_n}$ be a sample of values from a **NORMAL** population having *expectation* $\boldsymbol{\mu}$ and *variance* $\boldsymbol{\sigma^2}$.

Then,

TheVariance of Sample Variance

$$Var\left(S^2\right) = \frac{2\sigma^4}{(n-1)}$$

Prove this equation for your practice ! ☺

Example

The time it takes a central processing unit to process a certain type of job is **normally** distributed with **mean 20 seconds** and **standard deviation 3 seconds**.

If a sample of 15 such jobs is observed, what is the probability that the sample variance will **exceed 12** ?

Since the sample is of size **n = 15** and **σ^2 = 9**, write

$$P\left(S^2 > 12\right) = P\left((15-1)\frac{S^2}{9} > \frac{(15-1)}{9}.12\right)$$

$$= P\left(\frac{14S^2}{9} > 18.67\right)$$

$$= P\left(\chi^2_{14} > 18.67\right)$$

$$= 1 - P\left(\chi^2_{14} \leq 18.67\right)$$

$$= 1 - 0.8221$$

$$= 0.1779$$

We compute this using Chi-square dist. **calculator**

Sampling from A Finite Population

Consider a population of ***N*** elements, and suppose that ***p*** is the proportion of the population that has a certain characteristic of interest; that is

- ***Np*** elements have this characteristic
- ***N(1-p)*** do not

A sample of size **n** from this population is said to be a ***random sample*** if it is chosen in such a manner that each of the $\binom{N}{n}$ population subsets of size **n** is equally likely to be the sample.

Suppose now that a random sample of size **n** has been chosen from a population of size **N**. For **i = 1, 2, ..., n**, let

$$X_i = \begin{cases} 1 & \text{If the } \mathbf{i^{th}} \text{ member of the sample has the characteristic} \\ 0 & \text{Otherwise} \end{cases}$$

When the population size **N** is large with respect to the sample size **n**, then $\mathbf{X_1}$, $\mathbf{X_2}$, **...**, $\mathbf{X_n}$ are approximately **independent**.

Let $$X = \sum_{i=1}^{n} X_i$$

It follows that **X** can be thought of as representing the total number of success in **n** trials.

Hence, if the $\mathbf{X_i}$ were independent, then **X** would be a **binomial random variable** with parameters **n** and **p**.

$$X \sim B(n, p)$$

$$E[X] = np$$

$$Var(X) = np(1 - p)$$

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

Now, we will suppose that the underlying **population is large in relation to the sample size** and we take the distribution of **X** to be binomial.

Since $\bar{X}$, the **proportion** of the sample that has the characteristics, is equal to ***X/n***, we from the preceding that

$$E[\bar{X}] = E[X/n] = p$$

$$Var(\bar{X}) = \frac{1}{n^2} Var(X) = \frac{p(1-p)}{n}$$

$$SD(\bar{X}) = \sqrt{\frac{p(1-p)}{n}}$$

Example

Suppose that 45 percent of the population favors a certain candidate in an upcoming election. If a random sample of size 200 is chosen, find

- The expected value and standard deviation of the number of members of the sample that favor the candidate
- The probability that more than half the members of the sample favor the candidate

(a) The expected value and standard deviation of the number of members of the sample that favor the candidate

$$X = X_1 + X_2 + \ldots + X_{200}$$

$$E[X] = np = 200(0.45) = 90$$

$$SD(X) = \sqrt{np(1-p)} = \sqrt{200(0.45)(1-0.45)} = 7.0356$$

(b) Since **X** is binomial with **n = 200, p = 0.45**, the solution is

$$P(X \geq 101) = 0.681$$

If we use Normal approximation:

$$P(X \geq 101) = P(X \geq 100.5) \qquad \text{Continuity correction}$$

$$= P\left(\frac{X-90}{7.0356} \geq \frac{100.5-90}{7.0356}\right)$$

$$\approx P(Z \geq 1.4924) \approx 0.0678$$

Jika 10 dadu (fair dice) dilemparkan, hitunglah probabilitas (aproksimasi) bahwa jumlah semua nilai yang didapatkan adalah diantara 30 dan 40 !

Suatu populasi penduduk di kota A mempunyai informasi rataan tinggi badan 167 cm dan standar deviasi 27 cm.

Jika 36 orang dari kota A diambil sebagai sampel, berapa probabilitas bahwa rataan sampel berada diantara 163 cm dan 170 cm ?

Seorang guru dari pengalaman sebelumnya mengetahui bahwa rataan nilai ujian siswa adalah 77 dan standar deviasi 15.

Saat ini, guru tersebut mengajar di dua kelas: kelas A dan kelas B. Kelas A terdiri dari 25 siswa dan kelas B terdiri dari 64 siswa.

- Tentukan probabilitas rataan di kelas A antara 72 dan 82 !
- Ulangi pertanyaan sebelumnya untuk kelas B !
- Tentukan probabilitas bahwa rataan nilai ujian di kelas A lebih tinggi dari rataan di kelas B !

Suatu perusahaan memproduksi bola lampu yang umurnya berdistribusi Normal dengan rataan 800 jam dan simpangan baku 40 jam.

Hitunglah peluang bahwa suatu sampel acak dengan 16 bola lampu akan mempunyai rata-rata umur kurang dari 775 jam !

Suhu suatu logam pada kondisi tertentu diketahui mempunyai distribusi Normal dengan variansi 2.

Jika kemudian suhu logam tersebut diukur lagi sebanyak 5 kali,

- Tentukan probabilitas bahwa variansi sampel kurang dari 3,6 !
- Berapa ukuran sampel yang diperlukan (berapa kali mengukur) agar probabilitas pada kasus **a)** paling sedikit 0,95 ?

Diketahui 45% dari penduduk desa A menyukai caleg Cecep pada pemilu 2004. Jika sampel acak berukuran 200 orang dipilih dari desa A, hitunglah:

- Harapan dan standar deviasi dari banyaknya orang/ penduduk yang suka caleg Cecep pada sampel ?
- Probabilitas bahwa lebih dari separuh anggota sampel suka caleg Cecep ?