Sebaran Percontohan

Sampling Distribution

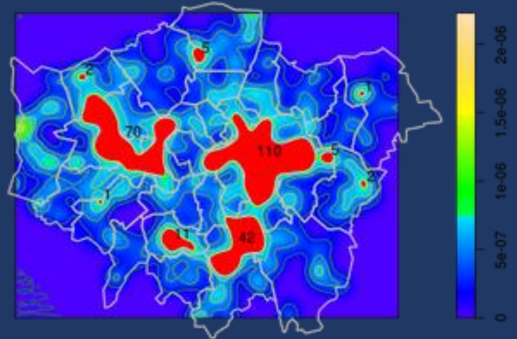
twitter: @kh_notodiputro

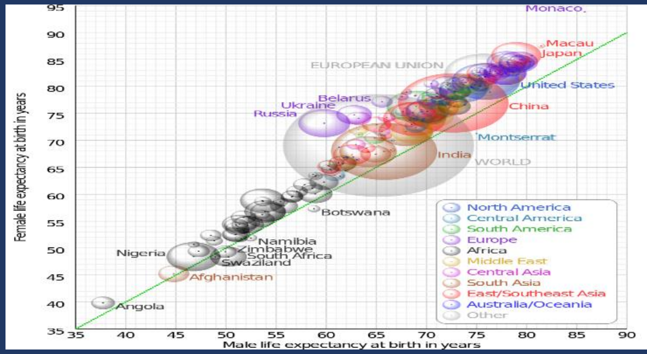
email: khairil@apps.ipb.ac.id

Prof. Dr. Ir. Khairil Anwar Notodiputro, MS

Guru Besar Tetap pada Fakultas Matematika dan Ilmu Pengetahuan Alam
Institut Pertanian Bogor

Kuliah 06





Motivasi

Kita ingin mengetahui pendapat mahasiswa STK tentang program “green campus”, setuju ataukah tidak setuju? ← **Tujuan**

Populasi ke contoh

Pilih beberapa mahasiswa STK yang bisa mewakili seluruh mhs ← **Populasi: seluruh mhs STK; Contoh: sebagian mhs STK**

Tanyakan pendapatnya ← Pengumpulan data

Gunakan metode statistika utk menganalisis pendapat mhs ← Pemodelan dan analisis

Populasi vs Contoh

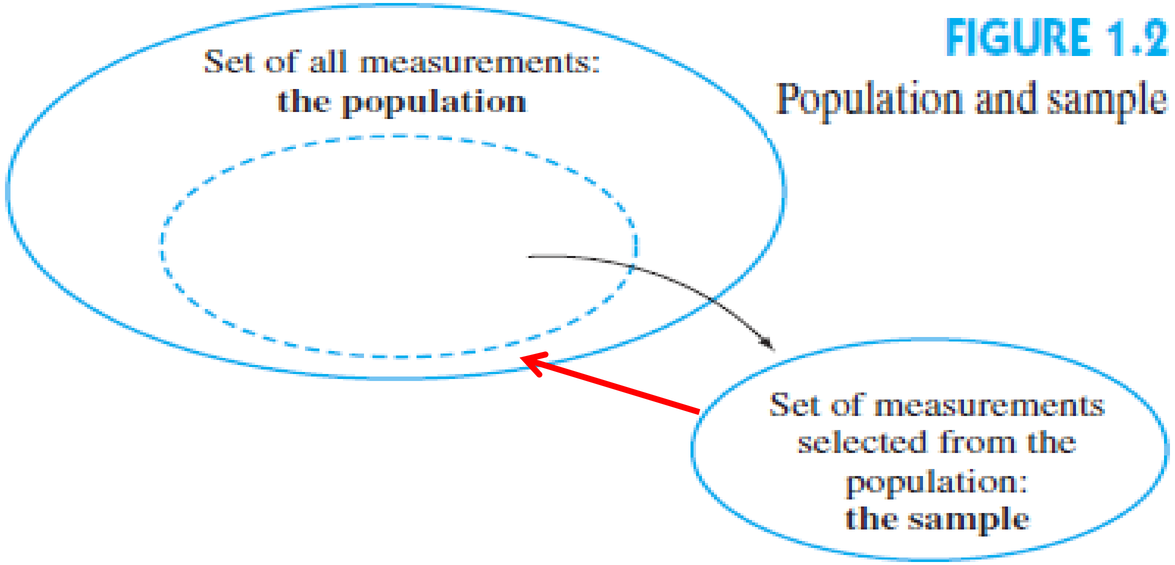


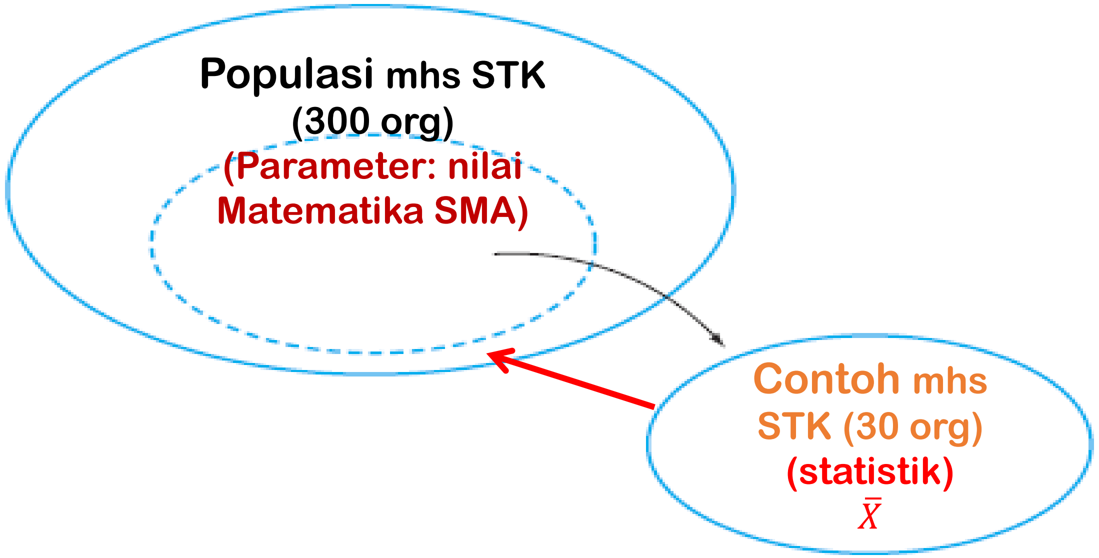
FIGURE 1.2
Population and sample

A **population** is the set of all measurements of interest to the sample collector. (See Figure 1.2.)

A **sample** is any subset of measurements selected from the population. (See Figure 1.2.)

We Observe Samples But Are Interested in Populations

Populasi vs Contoh



- Statistik $\bar{X}$ dihitung dari contoh yang terpilih.
- Jika contoh dipilih secara acak maka $\bar{X}$ merupakan peubah acak.
- Karena $\bar{X}$ peubah acak maka besarnya akan berubah-ubah dari satu contoh ke contoh yang lain.
- Dengan demikian statistik $\bar{X}$ memiliki sebaran peluang.
- Sebaran dari statistik seperti ini yang menjadi fokus bahasan dalam kuliah kali ini.

Sebaran Percontohan? *Sampling Distribution*

Mendenhall :

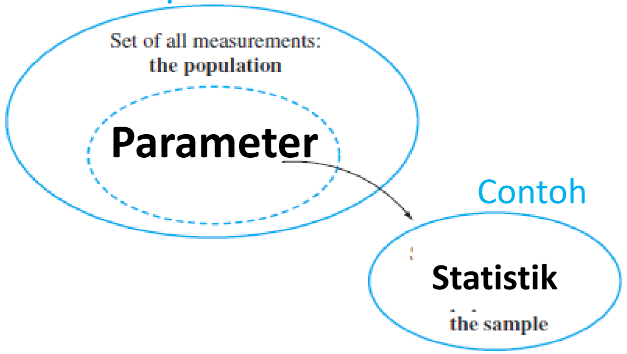
Definition The **sampling distribution of a statistic** is the probability distribution for the possible values of the statistic that results when random samples of size n are repeatedly drawn from the population.

Agresti :

Sampling Distribution

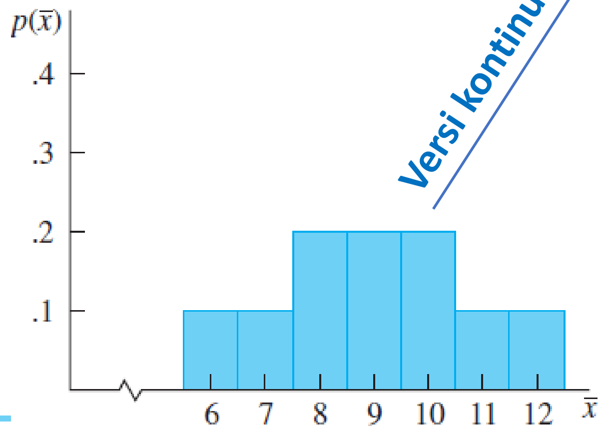
The **sampling distribution** of a statistic is the probability distribution that specifies probabilities for the possible values the statistic can take.

Populasi



Values of $\bar{x}$ and m for Simple Random Sampling when $n = 3$ and $N = 5$

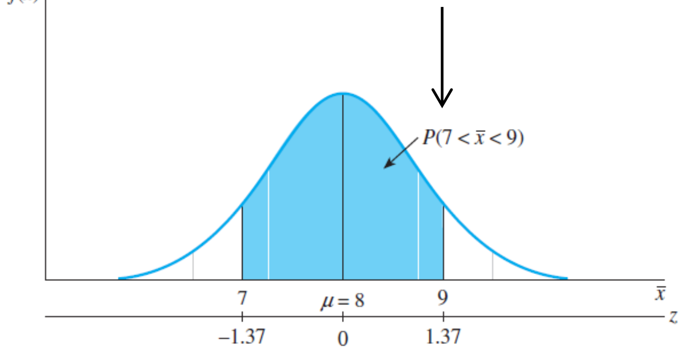
Sample	Sample Values	$\bar{x}$	m
1	3, 6, 9	6	6
2	3, 6, 12	7	6
3	3, 6, 15	8	6
4	3, 9, 12	8	9
5	3, 9, 15	9	9
6	3, 12, 15	10	12
7	6, 9, 12	9	9
8	6, 9, 15	10	9
9	6, 12, 15	11	12
10	9, 12, 15	12	12



Versi kontinu

The probability of interest is

$$P(7 < \bar{x} < 9) = P(-1.37 < z < 1.37)$$
$$= .9147 - .0853 = .8294$$



Sebaran Percontohan

(Sampling Distribution)

Sample	Value of $\bar{x}$	Sample	Value of $\bar{x}$
2, 3	2.5	3, 10	6.5
2, 4	3	3, 11	7
2, 5	3.5	4, 5	4.5
2, 6	4	4, 6	5
2, 7	4.5	4, 7	5.5
2, 8	5	4, 8	6
2, 9	5.5	4, 9	6.5
2, 10	6	4, 10	7
2, 11	6.5	4, 11	7.5
3, 4	3.5	5, 6	5.5
3, 5	4	5, 7	6
3, 6	4.5	5, 8	6.5
3, 7	5	5, 9	7
3, 8	5.5	5, 10	7.5
3, 9	6	5, 11	8

(2, 3, 4, 5, 6, 7, 8, 9, 10, 11)

Parameter
$\mu = 6.5$; $\sigma^2 = 8.25$

Hipotetical Data:
X = Number of household members

Contoh ukuran
n=2 → 45
kombinasi

Sample	Value of $\bar{x}$
6, 7	6.5
6, 8	7
6, 9	7.5
6, 10	8
6, 11	8.5
7, 8	7.5
7, 9	8
7, 10	8.5

Sample	Value of $\bar{x}$
7, 11	9
8, 9	8.5
8, 10	9
8, 11	9.5
9, 10	9.5
9, 11	10
10, 11	10.5

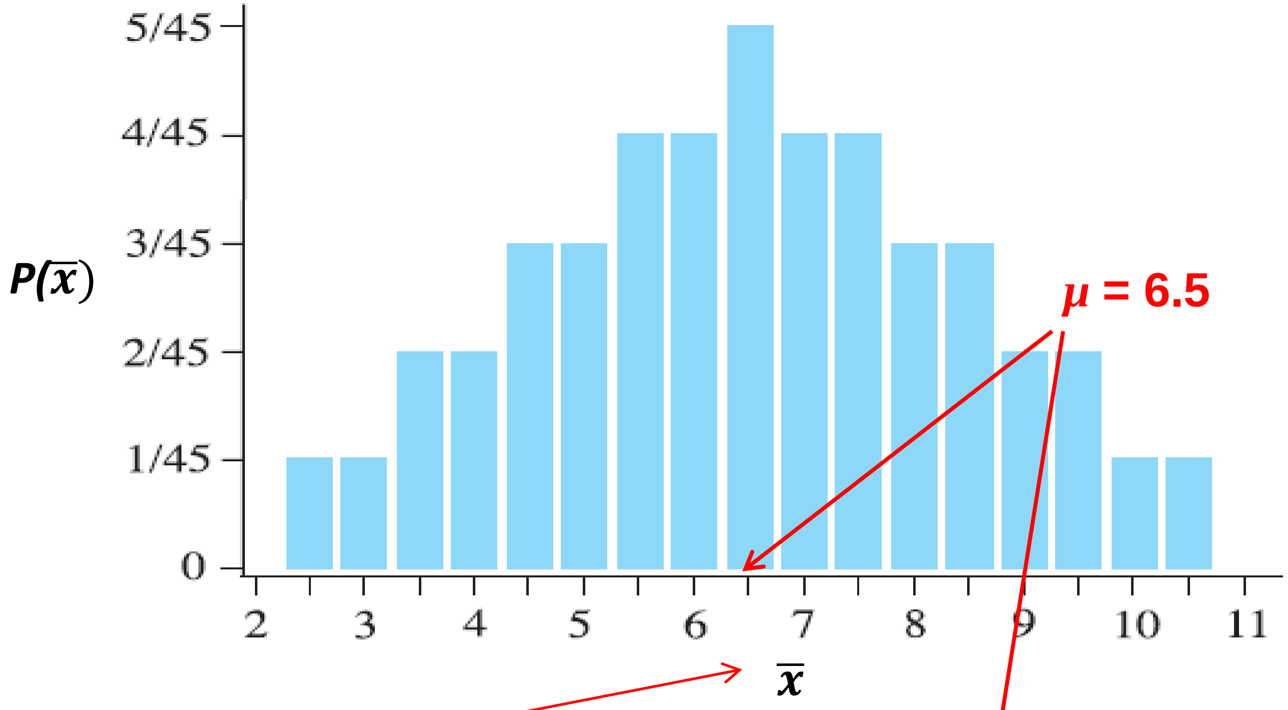
45 kemungkinan

Sebaran Percontohan

(*Sampling Distribution*)

Sampling distribution for $\bar{x}$

$\bar{x}$	$P(\bar{x})$	$\bar{x}$	$P(\bar{x})$
2.5	1/45	7	4/45
3	1/45	7.5	4/45
3.5	2/45	8	3/45
4	2/45	8.5	3/45
4.5	3/45	9	2/45
5	3/45	9.5	2/45
5.5	4/45	10	1/45
6	4/45	10.5	1/45
6.5	5/45		



μ = 6.5

- Sebaran dari $\bar{x}$ cenderung simetrik
- **Nilaitengah dari $\bar{x}$** persis sama dengan μ = 6.5
- Sebaran yang simetrik dikenal sbg sebaran normal

$\mu = E(\bar{x})$
$= 2.5*(1/45)+\ldots+ 10.5*(1/45)$
$= 6.5$

Lebih Jauh ttg Contoh

- Contoh adalah sebagian dari populasi;
- Dari contoh dihasilkan statistik untuk menduga besarnya parameter
- Jika contoh diperoleh secara acak maka:
  - Statistik yang dihasilkan berbeda-beda dari satu contoh ke contoh lainnya → bisa ada sebanyak k statistik: $\bar{X}_1$, $\bar{X}_2$, …, $\bar{X}_k$
  - Rata-rata dari statistik yang dihasilkan akan sama dengan parameternya → jadi $\mu = \frac{\bar{X}_1+\bar{X}_2+\ldots+\bar{X}_k}{k}$
  - Statistik $\bar{X}_k$ adalah peubah acak → jadi $\bar{X}_k$ punya sebaran peluang

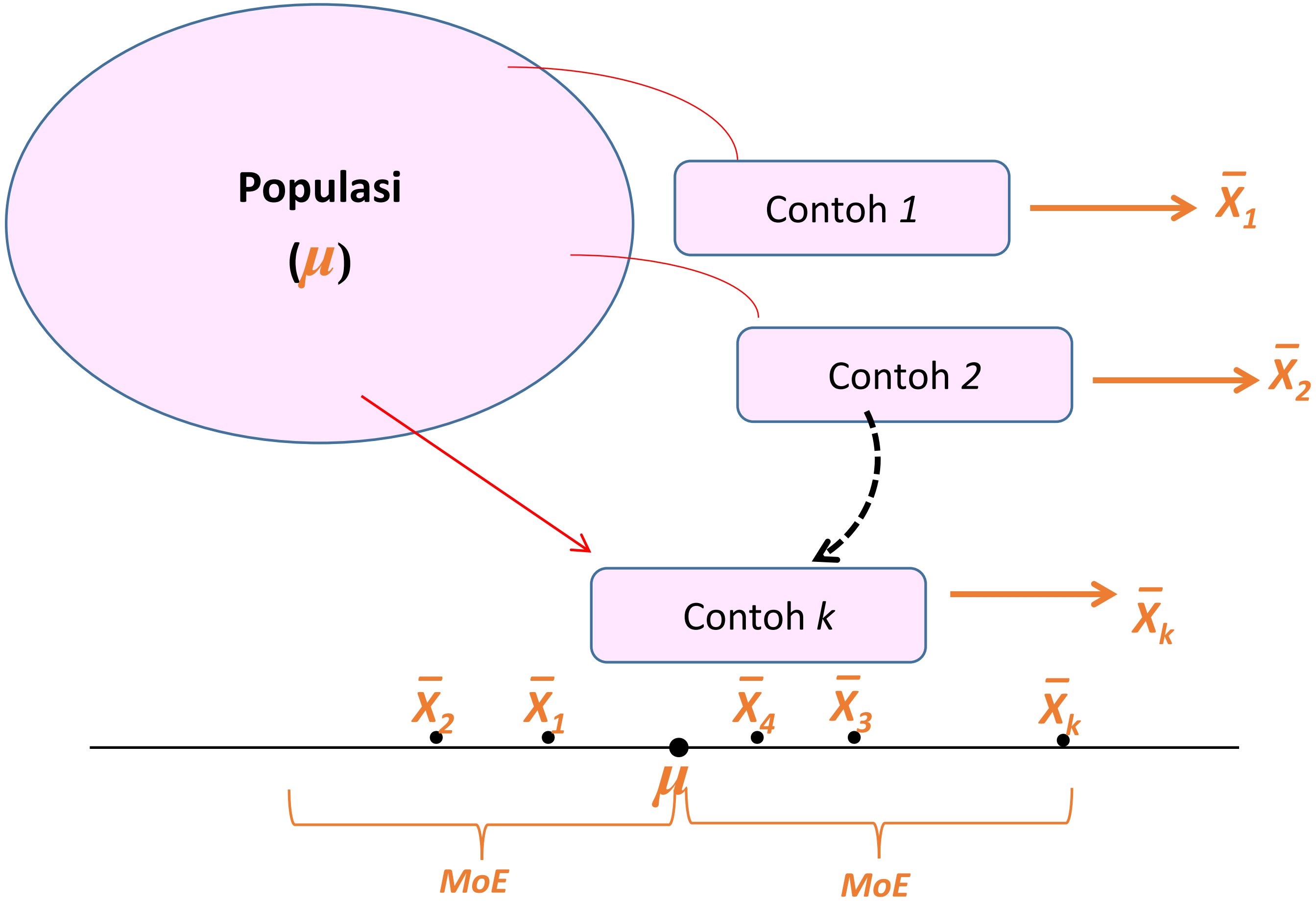
$\bar{X}_k \sim f(.)$

Memiliki sebaran peluang tertentu

Keakuratan Percontohan Acak Sederhana

(*Accuracy of Simple Random Sampling*)

- ***Margin of error*** **(*MoE*) adalah ukuran keragaman hasil dugaan dari satu contoh ke contoh berikutnya.**



MoE = penyimpangan maksimum dari statistik

X rata-rata terhadap parameternya (μ) dari hasil suatu percontohan acak

$\mu \Leftrightarrow \bar{X}_k$

Jika digunakan tingkat kepercayaan 95% maka *MoE* = 1.96*SE($\bar{X}_k$)

Rehat dulu

Sesi 1… beres!!!

Sesi 2…

Dalil Limit Pusat
(*Central Limit Theorem*)

General Properties of the Sampling Distribution of $\bar{x}$

Let $\bar{x}$ denote the mean of the observations in a random sample of size n from a population having mean μ and standard deviation σ. Denote the mean value of the $\bar{x}$ distribution by $\mu_{\bar{x}}$ and the standard deviation of the $\bar{x}$ distribution by $\sigma_{\bar{x}}$. Then the following rules hold:

Rule 1. $\mu_{\bar{x}} = \mu$.

Rule 2. $\sigma_{\bar{x}} = \dfrac{\sigma}{\sqrt{n}}$. This rule is exact if the population is infinite, and is approximately correct if the population is finite and no more than 10% of the population is included in the sample.

Rule 3. When the population distribution is normal, the sampling distribution of $\bar{x}$ is also normal for any sample size n.

Rule 4. (Central Limit Theorem) When n is sufficiently large, the sampling distribution of $\bar{x}$ is well approximated by a normal curve, even when the population distribution is not itself normal.

X_i *indept* $\sim (\mu, \sigma^2) \rightarrow \bar{\bar{X}} \sim N(\mu, \sigma^2/n)$

Dalil Limit Pusat:

Misalkan $X_1, X_2, \ldots, X_n$ adalah contoh acak dari populasi dengan nilai tengah μ dan ragam σ^2 (sebarannya tdk harus normal). Jika

$$\bar{X} = \frac{X_1 + X_2 + \ldots + Xn}{n} \quad \text{dan}$$

n besar (artinya $n \rightarrow \infty$)

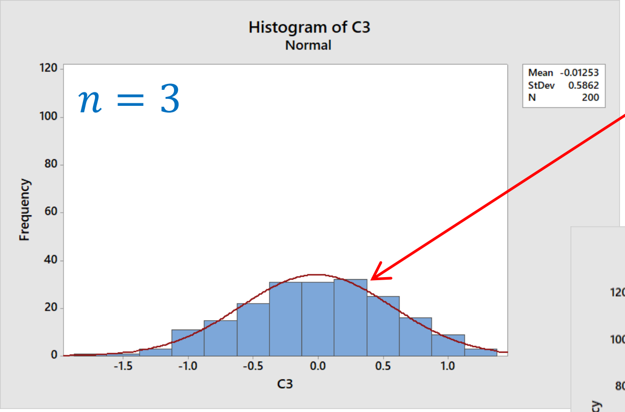
maka $\bar{X}$ akan menyebar NORMAL dengan nilaitengah μ dan ragam σ^2/n.

Dalil limit pusat sangat berguna sebagai dasar atau alasan mengapa kita sering menggunakan sebaran NORMAL dalam inferensi statistika walaupun sebaran datanya TIDAK NORMAL.

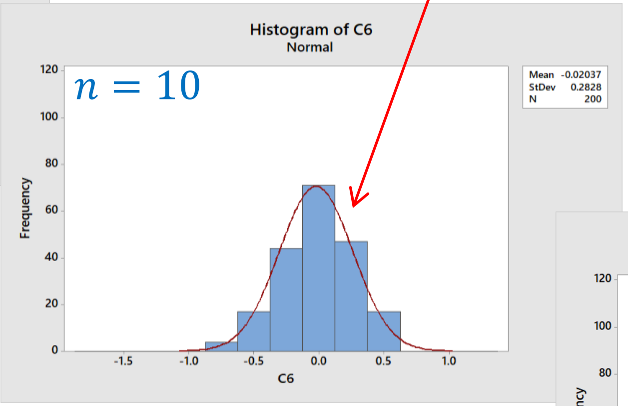
Simulasi Hubungan n dan Ragam
(*Sample size and variance*)

- Mari kita lakukan **simulasi** untuk untuk menghitung rata2 dari peubah acak $X \sim Normal\ (0,1)$ dengan ukuran contoh $n = 3$, $n = 10$, dan $n = 30$ masing masing diulang sebanyak 200 kali.

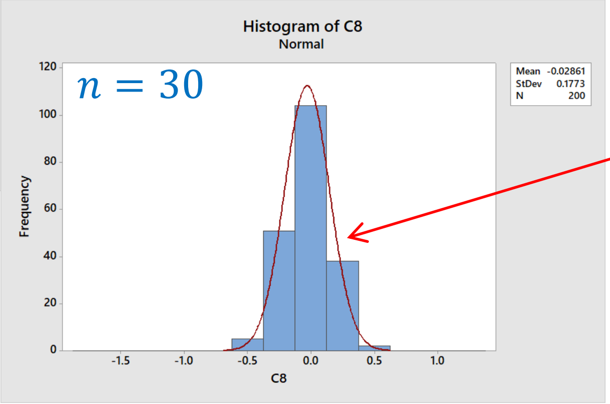
$n = 3$		$n = 10$		$n = 30$	
Ulangan ke	$\bar{X}_i$	Ulangan ke	$\bar{X}_i$	Ulangan ke	$\bar{X}_i$
1	-0.849	1	-0.146	1	-0.059
2	0.707	2	-0.240	2	-0.227
3	0.336	3	-0.002	3	0.013
:	:	:	:	:	:
:	:	:	:	:	:
199	-0.426	199	0.501	199	-0.138
200	1.058	200	-0.244	200	0.188



$s_{\bar{X}} = 0.58$



$s_{\bar{X}} = 0.28$

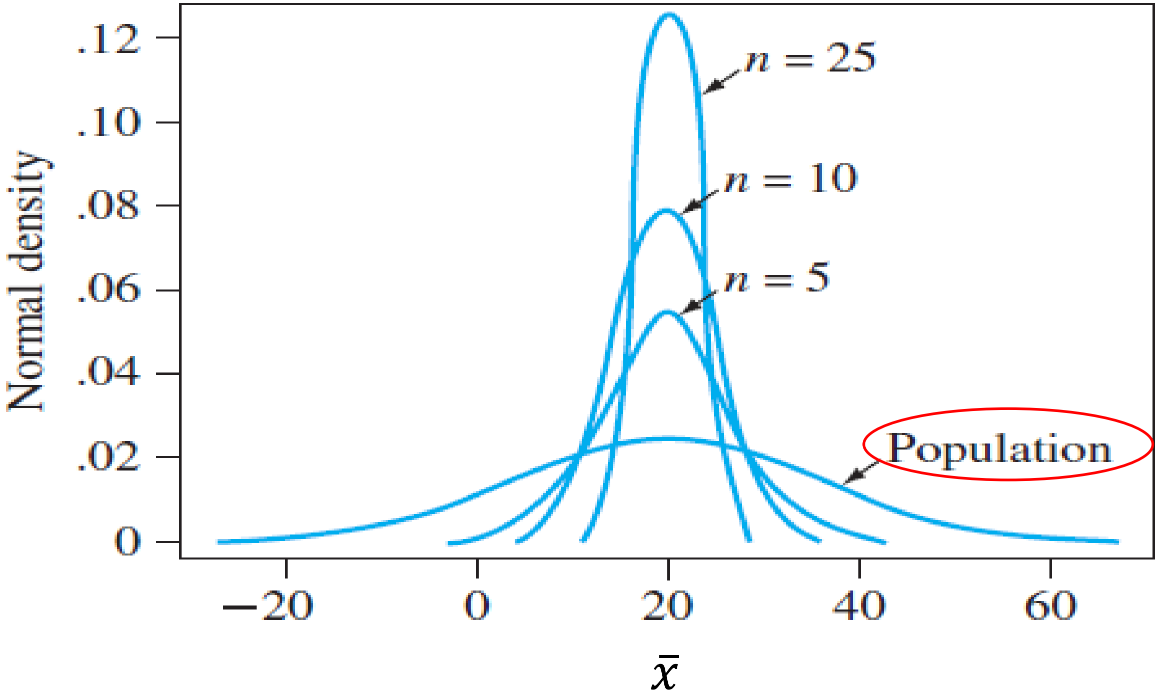


$s_{\bar{X}} = 0.18$

Perhatikan bahwa besarnya $s_{\bar{X}} \approx \frac{\sigma}{\sqrt{n}}$

Hubungan *n* dan Ragam
(*Sample size and variance*)

- **Misal p.a. $X_1, X_2, \ldots, X_n$ merupakan contoh acak dari populasi normal dengan nilai tengah dan ragam tertentu. Maka hubungan antara *n* dg sebaran $\bar{x}$:**

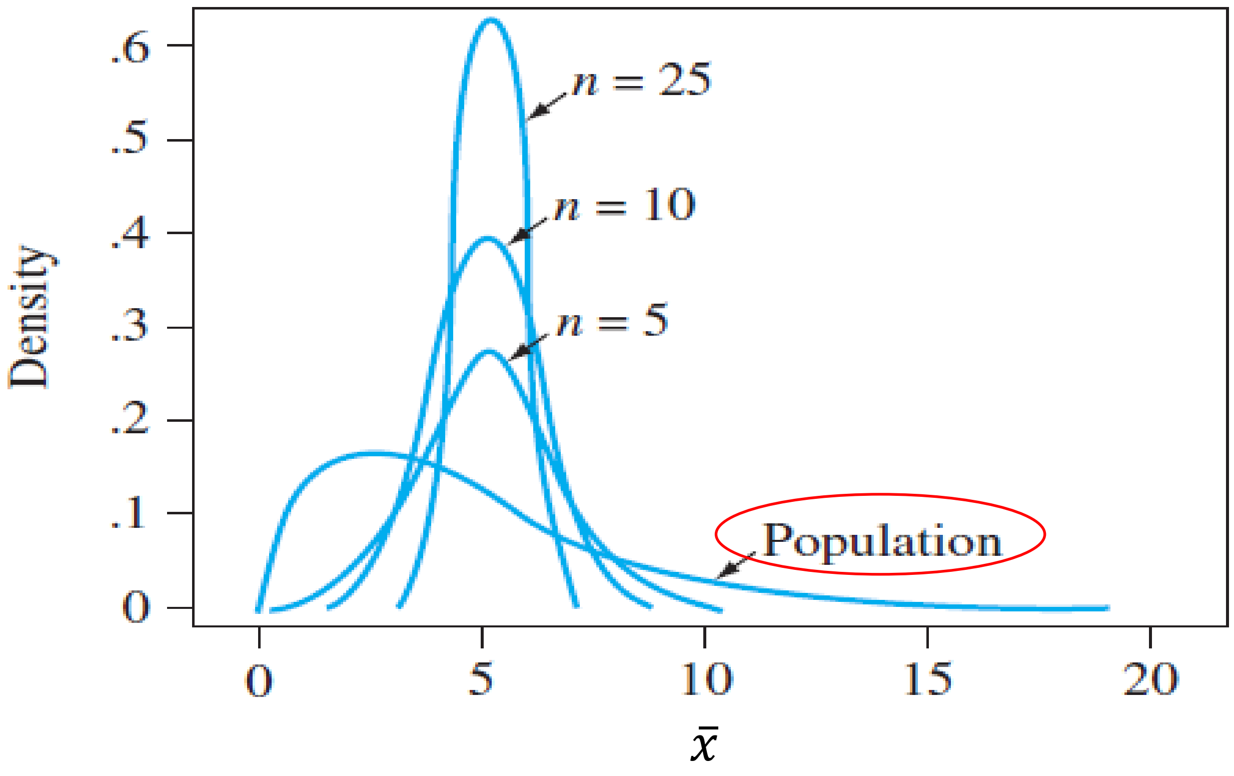


Semakin besar *n*, semakin kecil ragam dari $\bar{x}$ → kurva semakin 'kurus'.

Hubungan *n* dan Ragam

(*Sample size and variance*)

- **Misal p.a. $X_1, X_2, \ldots, X_n$ merupakan contoh acak dari populasi non-simetrik dg nilaitengah dan ragam tertentu. Maka hubungan antara *n* dg sebaran $\bar{x}$**



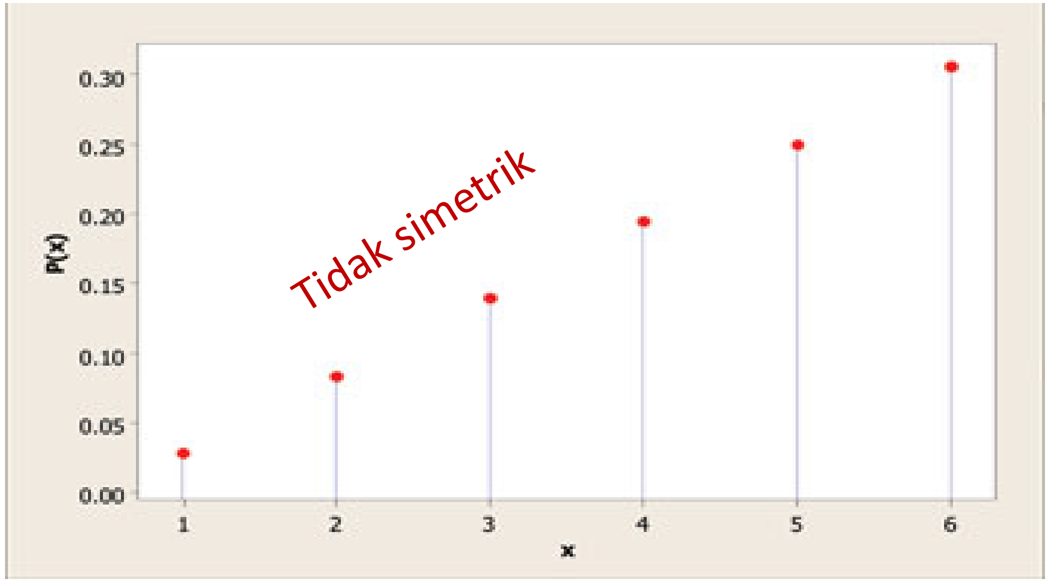
Walaupun sebaran populasi tidak simetrik tetapi sebaran dari rataan $\bar{x}$ simetrik

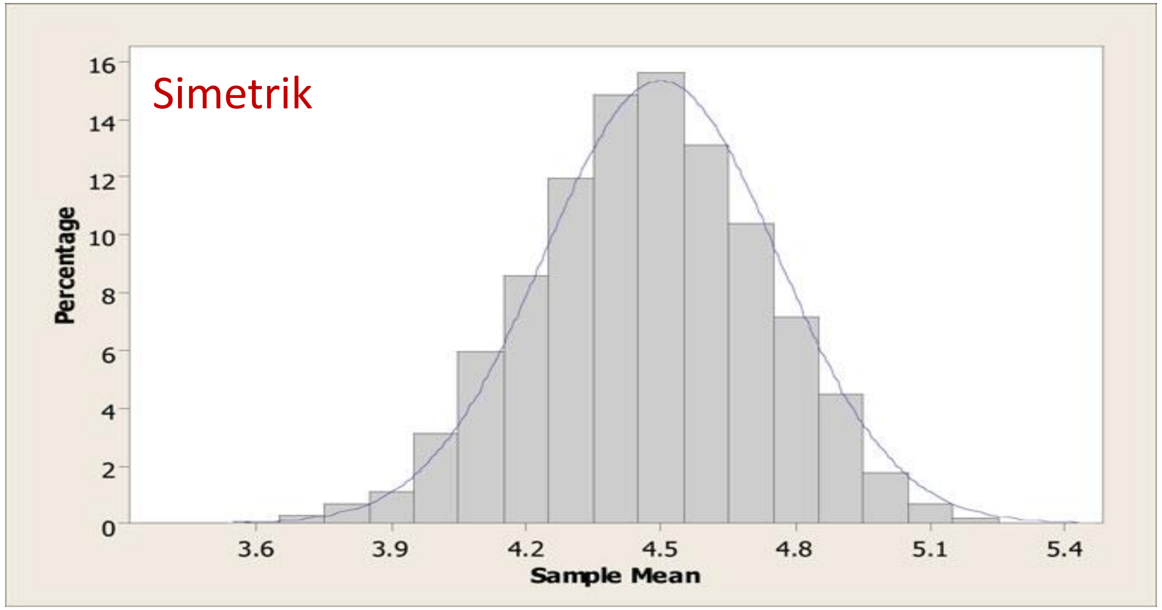
Tidak Simetrik berubah menjadi Simetrik
(From asymmerics to symmetrics)

- **Jika sepasang dadu setimbang dilempar dan diamati angka terbesar yang muncul maka :**

  Jika dadu 1 muncul angka 2, sdgkan dadu 2 muncul 5 → dicatat 5

  Jika dadu 1 muncul angka 4, sdgkan dadu 2 muncul 3 → dicatat 4, dan seterusnya.





- **Sebaran dari rataan angka terbesar jika sepasang dadu dilempar sebanyak 30 kali dan diulang 1000 kali terlihat sebaran rataannya simetrik**

Beberapa Teknik Percontohan

(*Sampling Techniques*)

- Selanjutnya kalau kita menyebut contoh (tanpa embel-embel) maka yang kita maksud adalah contoh yang diperoleh secara acak → contoh acak (*random sample*)
- Ada banyak cara untuk memperoleh contoh acak, yang sangat populer ada 3 (tiga) percontotah (*sampling*):
  - Percontohan acak sederhana (*simple random sampling*)
  - Percontohan acak berstrata (*stratified random samping*)
  - Percontohan acak klaster atau gerombol (*cluster random sampling*)
- Setiap teknik percontohan memiliki keunggulan dan keterbatasan masing-masing.

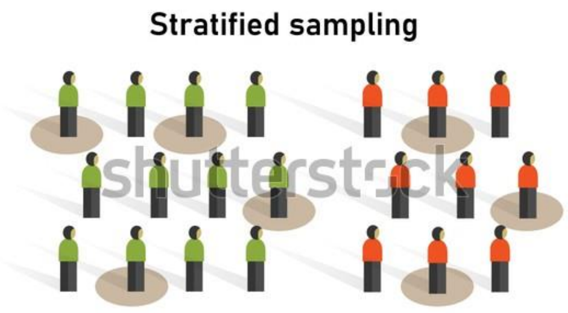
Apa pun teknik percontohannya, maka contoh yang terambil harus memenuhi:

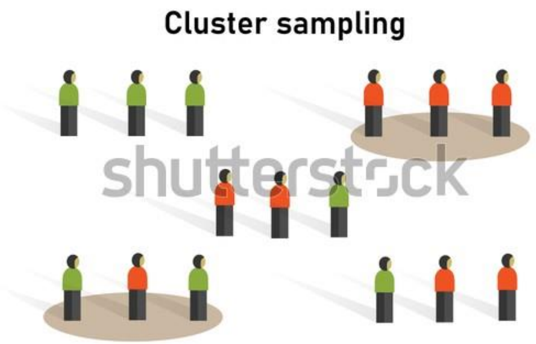
1. Representatif → cerminan populasinya;
2. Cukup jumlahnya → mencapai target *MoE* yang diiinginkan

Beberapa Teknik Percontohan

(*Sampling Techniques*)

- Percotohan acak sederhana (*simple random sampling*):
    - memilih secara acak 30 mhs dari 300 mhs STK, kemudian diamati nilai MAT, FIS, KIM, BIO selama di SMA.
    - setiap mhs berpeluang 0.1 untuk terpilih.
- Percontohan acak berstrata (*stratified random sampling*): mengelompokkan 300 mhs STK menjadi dua kelompok: Jabodetabek 250 mhs dan non-Jabodetabek 50 mhs.
    - dari kelompok pertama dipilih 20-30 mhs dan dari kelompok kedua dipilih 5-10 mhs secara acak.
    - peluang seorang mhs untuk terpilih bisa tidak sama.
- Percontohan klaster (*cluster random sampling*):
    - memilih secara acak 5 provinsi asal mhs, kemudian semua mhs yang berasal dari provinsi terpilih diamati nilai MAT, FIS, KIM, BIO selama di SMA.
    - provinsi dalam hal ini menjadi klaster mahasiswa.





Teknik percontohan mana yang paling mudah dikerjakan?

Contoh Soal

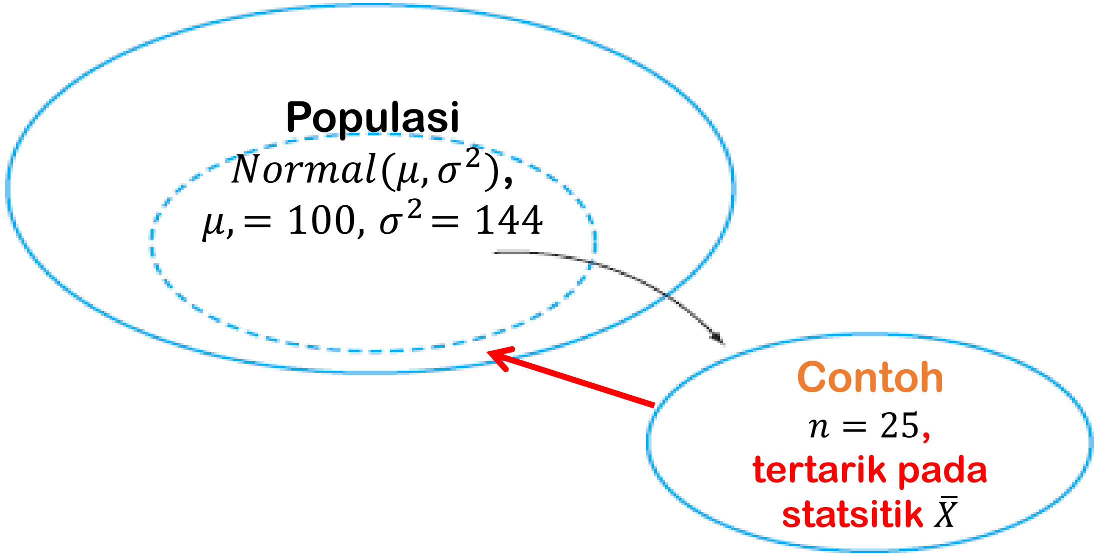
Mendenhall (2013):

7.25 Suppose a random sample of $n = 25$ observations is selected from a population that is normally distributed with mean equal to 106 and standard deviation equal to 12.

a. Give the mean and the standard deviation of the sampling distribution of the sample mean $\bar{x}$.

b. Find the probability that $\bar{x}$ exceeds 110.

c. Find the probability that the sample mean deviates from the population mean $\mu = 106$ by no more than 4.



a. $\bar{X} \sim$ Normal (106, $\frac{144}{25}$), sehingga simpangan baku dari $\bar{X}$ sebesar $\frac{12}{5} = 2.4$

b. P($\bar{X}$ > 110) = P(Z > $\frac{110 - 106}{12/5}$) = P(Z > 1.6667) = ……???

c. P(|$\bar{X}$ - 106| < 4) = P(-4 < ($\bar{X}$ - 106) < 4) = P(102 < $\bar{X}$ < 110)

= P ($\frac{102 - 106}{12/5}$ < Z < $\frac{110 - 106}{12/5}$) = P(-1.6667 < Z < 1.6667) = …….??

Rehat dulu

Sesi 2... beres!!!

Sesi 3...

Sebaran Percontohan utk $\hat{p}$

- Dalam praktek kita ingin menduga proporsi, misal proporsi mahasiswa yang setuju thd suatu kebijakan universitas.
- Parameternya *p* sedangkan penduganya adalah $\hat{p}$ yang dihitung dari hasil survei.
- Besarnya $\hat{p}$ dihitung dari berapa banyaknya mahasiswa yang setuju (*x*) dibagi dengan banyaknya seluruh mahasiswa yang disurvei (*n*).

Jadi $\hat{p} = \frac{x}{n}$

- Ingat kembali bhw X menyebar Binom dengan nilaitengah np dan ragam sebesar $np(1-p)$.

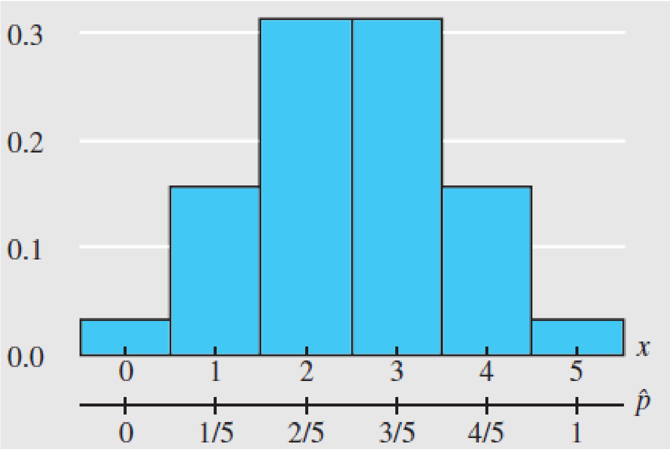
Jika $X \sim Binom\ (n, p)$ maka
$E(X) = np;\ Var(X) = np(1-p)$.

Jika $\hat{p}$ adalah dugaan proporsi sukses maka $\hat{p} = \frac{\sum X}{n}$ sehingga

$$E(\hat{p}) = \frac{np}{n} = p$$

$$Var(\hat{p}) = \frac{np(1-p)}{n^2} = \frac{p(1-p)}{n}$$

- Sebaran percontohan dari $\hat{p}$ simetrik seperti sebaran dari $\bar{X}$ (lihat gambar).



Galat baku dari $\hat{p}$ adalah

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

Sebaran Percontohan utk $\hat{p}$

PROPERTIES OF THE SAMPLING DISTRIBUTION OF THE SAMPLE PROPORTION, $\hat{p}$

- If a random sample of n observations is selected from a binomial population with parameter p, then the sampling distribution of the sample proportion

  $$\hat{p} = \frac{x}{n}$$

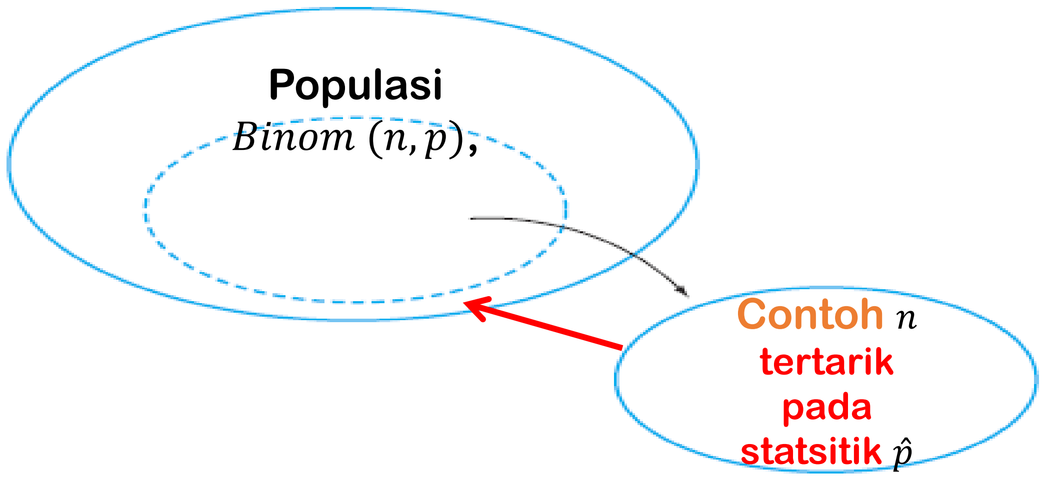
  will have a mean

  $$p$$

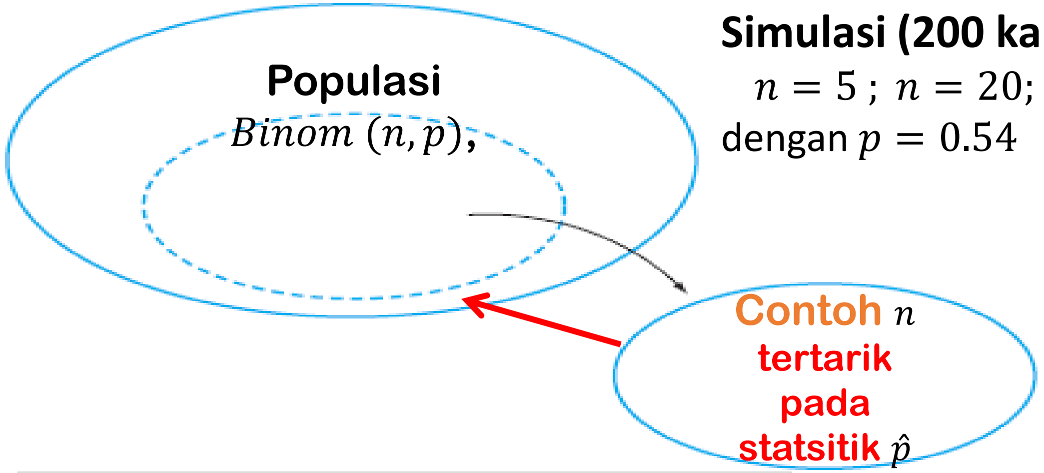
  and a standard deviation

  $$SE(\hat{p}) = \sqrt{\frac{pq}{n}} \qquad \text{where } q = 1 - p$$

- When the sample size n is large, the sampling distribution of $\hat{p}$ can be approximated by a normal distribution. The approximation will be adequate if $np > 5$ and $nq > 5$.

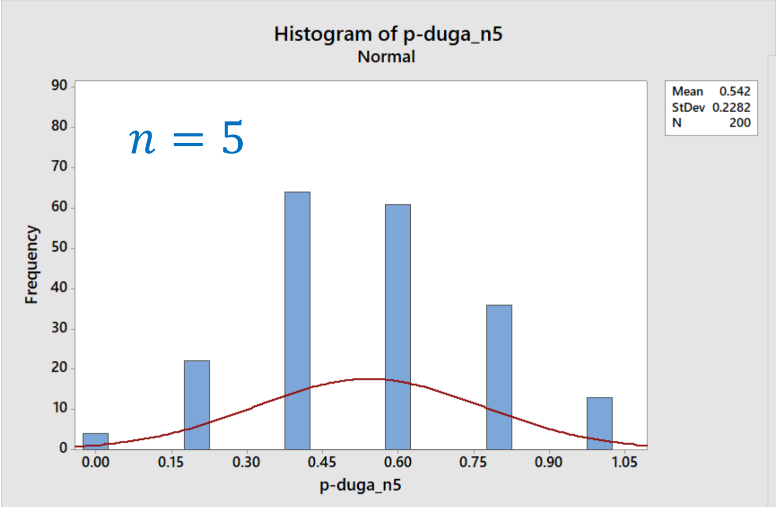


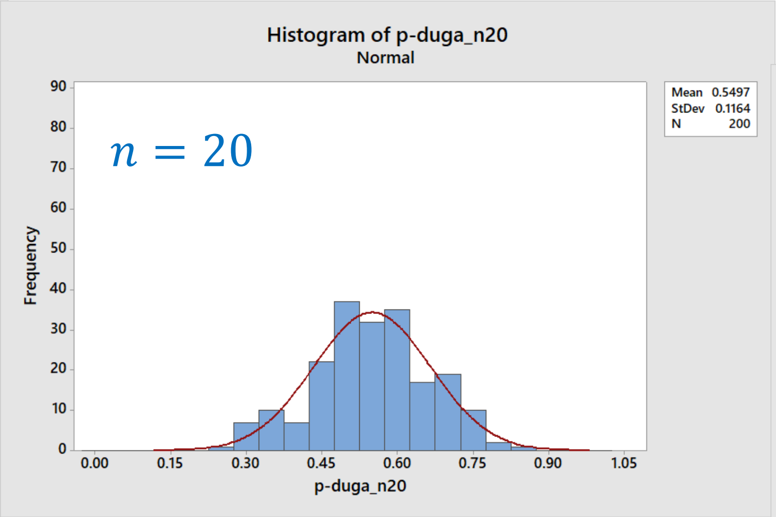
Simulasi Sebaran Percontohan utk $\hat{p}$

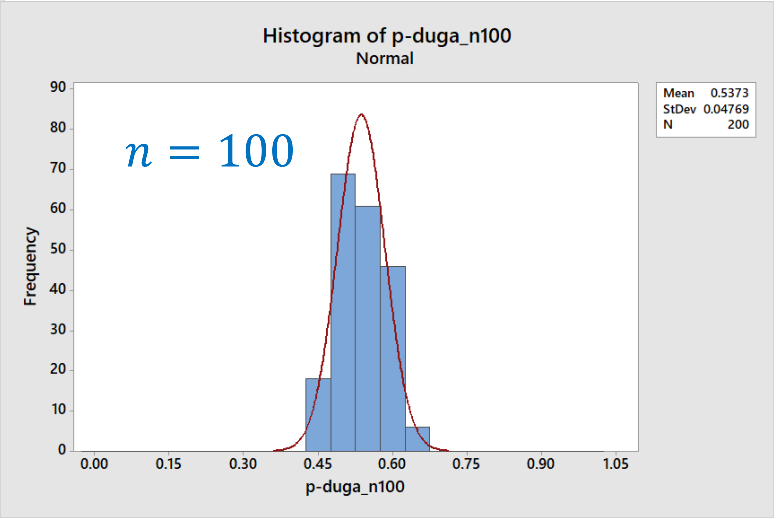


Simulasi (200 kali):
$n = 5$; $n = 20$; $n = 100$
dengan $p = 0.54$

$n = 5$		$n = 20$		$n = 100$	
Ulangan ke	$\hat{p}_i$	Ulangan ke	$\hat{p}_i$	Ulangan ke	$\hat{p}_i$
1	0.4	1	0.70	1	0.50
2	0.2	2	0.70	2	0.57
3	0.8	3	0.60	3	0.61
:	:	:	:	:	:
200	0.40	200	0.50	200	0.53







Perhatikan bahwa galat baku $S_{\hat{p}} \approx \sqrt{\frac{p(1-p)}{n}}$

Ilustrasi

Predicting the Election Outcome

Picture the Scenario

Let's now conduct an analysis that uses the actual exit poll of 3889 voters for the 2010 California gubernatorial election. In that exit poll, 53.1% of the 3889 voters sampled said they voted for Jerry Brown.

Questions to Explore

a. Given that the actual population proportion supporting Brown was 0.538, what are the values of the sample proportion we would expect to observe from random sampling?

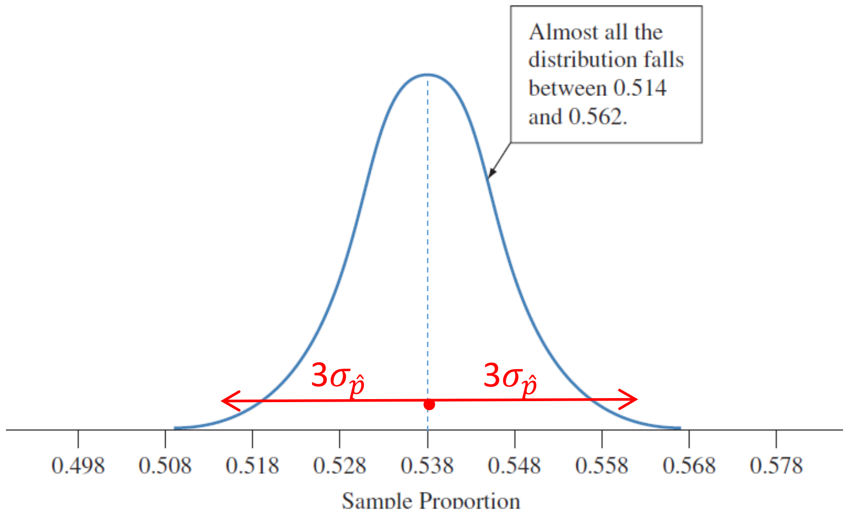
b. Based on the results of the exit poll, would you have been willing to predict Brown as the winner on election night while the votes were still being counted?

Hasil suatu ***exit poll :*** dari 3889 pemilih 53.1% mengatakan telah memilih Jerry Brown.

a. Andai p = proporsi pemilih JB yang sebenarnya adalah 0.538, berapa kisaran nilai-nilai p yang mungkin teramati jika diambil contoh acak?

b. Apakah kita akan memprediksi JB sebagai pemenang dalam pemilihan ini?

a. Perhatikan gambar di samping ini. Jika $p = 0.538$ maka contoh acak akan menghasilkan dugaan p, yaitu $\hat{p}$ yang memiliki sebaran percontohan dengan nilaitengah E($\hat{p}$) = 0.538 dan simpangan baku $\sqrt{\dfrac{(0.538)(0.462)}{3889}}$ = 0.008.



b. Hampir semua $\hat{p}$ yang diperoleh dari contoh acak terletak di antara 0.514 sampai dengan 0.562 (3 kali simpangan baku). Jadi kita yakin JB akan menjadi pemenangnya.

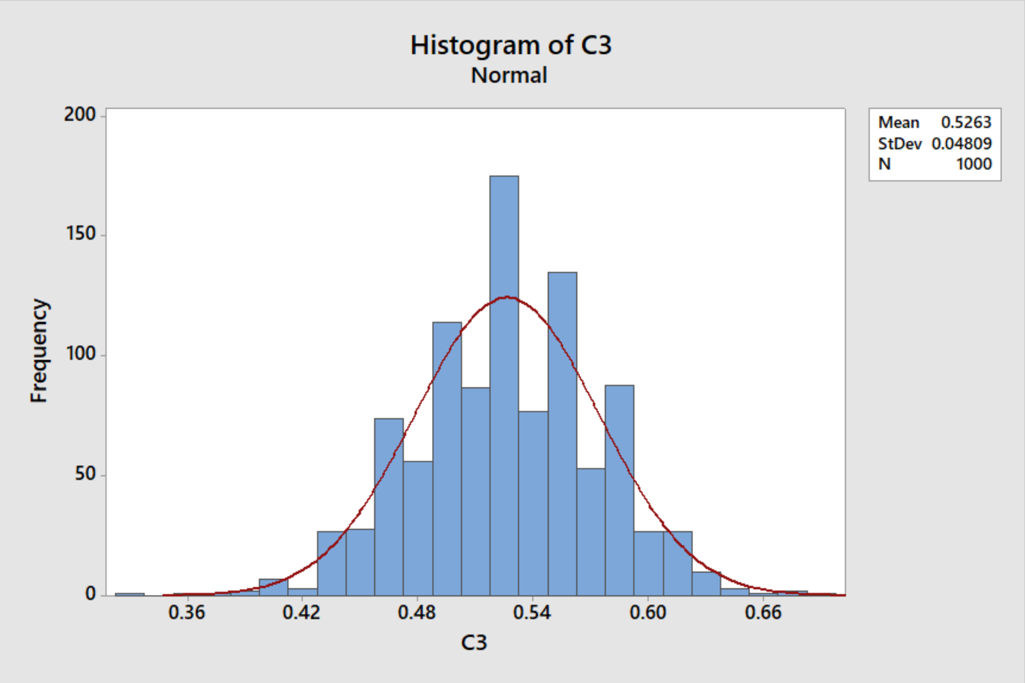
Ilustrasi

SUMMARY: Sampling Distribution of a Sample Proportion

For a random sample of size n from a population with proportion p, the sampling distribution of the sample proportion has

$$\text{mean} = p, \text{ standard deviation} = \sqrt{\frac{p(1-p)}{n}}.$$

If n is sufficiently large so that the expected numbers of outcomes of the two types, np in the category of interest and $n(1-p)$ not in that category, are both at least 15, then the sampling distribution of a sample proportion is approximately normal.



Simulating the exit poll Simulate an exit poll of 100 voters, using the Sampling Distributions applet on the text CD, assuming that the population proportion is 0.53. Refer to Activity 1 for guidance on using the applet.

a. What sample proportion did you get? Why do you not expect to get exactly 0.53?

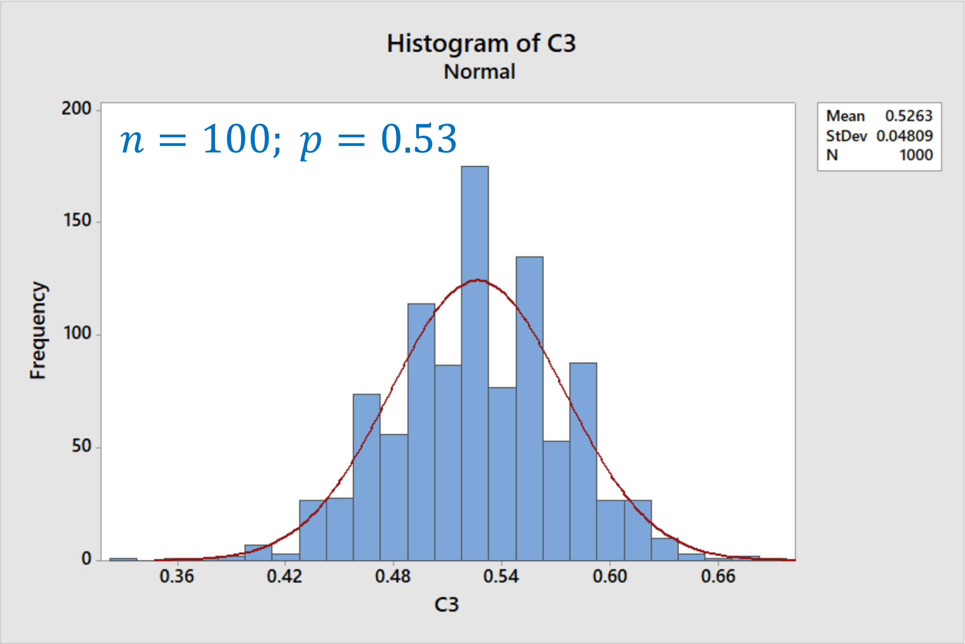
b. Simulate this exit poll 10,000 times (set N = 10,000 on the applet menu). Keep the sample size at $n = 100$ and $p = 0.53$. Describe the graph of the 10,000 sample proportion values.

c. Use a formula from this section to predict the value of the standard deviation of the sample proportions that you generated in part b.

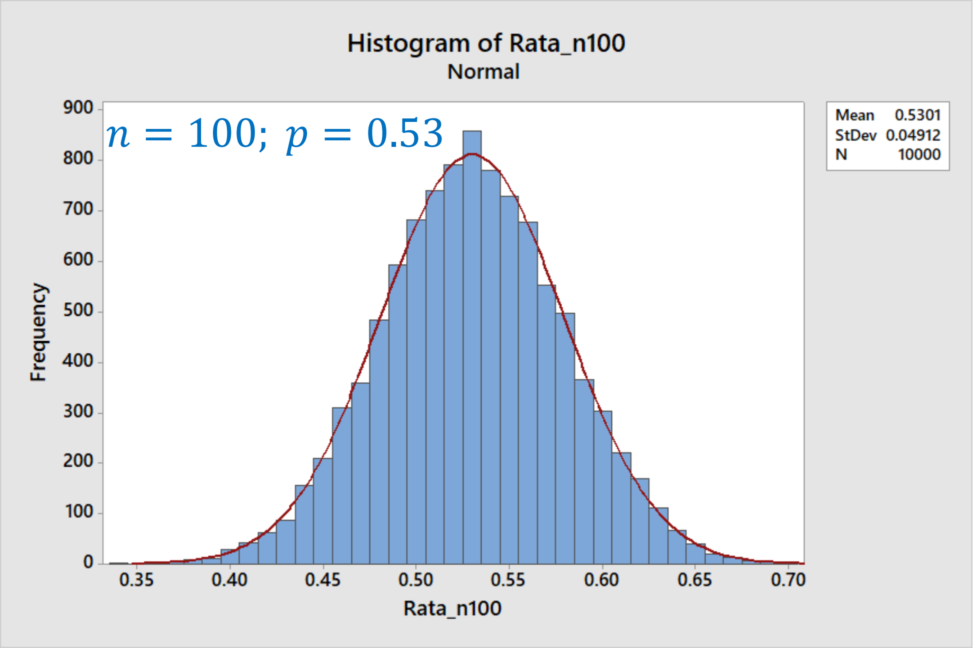
d. Now change the population proportion to 0.7, keeping the sample size $n = 100$. Simulate the exit poll 10,000 times. How would you say the results differ from those in part b?

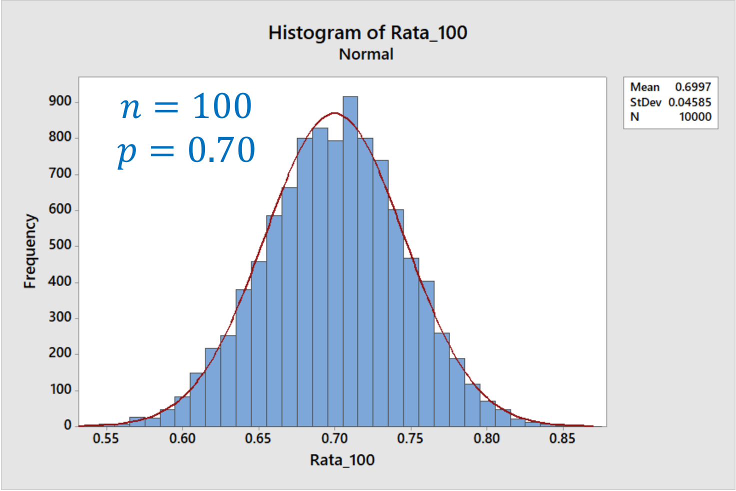
a. Ambil contoh acak (n =100) menghasilkan $\hat{p} = 0.57 \neq 0.53$. Mengapa? Karena yang kita peroleh adalah statistik $\hat{p}$.
b. Lihat grafik di samping ini sebagai hasil simulasi dengan $n = 100$ dan simulasi sebanyak 1000 kali.
c. $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.53(0.47)}{100}} = 0.05$
d. Jika $p = 0.7$ maka akan diperoleh histogram hasil simulasi yang lebih sempit, artinya akan diperoleh $\sigma_{\hat{p}}$ yang lebih kecil.

Ilustrasi



Simulasi 1000 kali





Simulasi 10000 kali

Sebaran Percontohan? *Sampling Distribution*

IPB University — Bogor Indonesia —

Mendenhall :

Definition The **sampling distribution of a statistic** is the probability distribution for the possible values of the statistic that results when random samples of size n are repeatedly drawn from the population.

Agresti :

Sampling Distribution

The **sampling distribution** of a statistic is the probability distribution that specifies probabilities for the possible values the statistic can take.

Contoh sebaran percontohan statistik:

- $\bar{X} \sim$ Normal (μ, σ^2)
- $\hat{p} \sim$ Binom (n, p)
- $\frac{(n-1)S^2}{\sigma^2} \sim$ Khi-kuadrat $(n-1)$
- Dan seterusnya

Sebaran percontohan tidak lain adalah <u>sebaran peluang dari suatu statistik</u>. Misal, sebaran peluang dari $\bar{X}$ atau sebaran peluang dari $\hat{p}$, atau sebaran peluang dari S^2, dan seterusnya.

Diskusikan

Let y be a normal random variable with $\mu = 500$ and $\sigma = 100$. Find the following probabilities:

√

a. $P(500 < y < 665)$
b. $P(y > 665)$
c. $P(304 < y < 665)$
d. k such that $P(500 - k < y < 500 + k) = .60$

√

German mobile study The contingency table shows results from the German study about whether there was an association between mobile phone use and eye cancer (Stang et al., 2001).

a. The study was retrospective. Explain what this means.
b. Explain what is meant by cases and controls in the headings of the table.
c. What proportion had used mobile phones, of those in the study who (i) had eye cancer and (ii) did not have eye cancer?

Eye Cancer and Use of Mobile Phones

Mobile Phones	Cases	Controls
Yes	16	46
No	102	429
Total	**118**	**475**

Diskusikan

√

Aspirin prevents heart attacks? During the 1980s approximately 22,000 physicians over the age of 40 agreed to participate in a long-term study called the Physicians' Health Study. One question investigated was whether aspirin helps to lower the rate of heart attacks. The physicians were randomly assigned to take aspirin or take placebo.

a. Identify the response variable and the explanatory variable.

b. Explain why this is an experiment, and identify the treatments.

c. There are other explanatory variables, such as the amount of exercise a physician got, that we would expect to be associated with the response variable. Explain how such a variable is dealt with by the randomized nature of the experiment.

√

Sample size and margin of error

a. Find the approximate margin of error when $n = 1$.

b. Show the two possible percentage outcomes you can get with a single observation. Explain why the result in part a means that with only a single observation, you have essentially no information about the population percentage.

c. How large a sample size is needed to have a margin of error of about 5% in estimating a population percentage? (*Hint*: Take the formula for the approximate margin of error, and solve for the sample size.)

Diskusikan

A person visits her doctor with concerns about her blood pressure. If the systolic blood pressure exceeds 150, the patient is considered to have high blood pressure and medication may be prescribed. A patient's blood pressure readings often have a considerable variation during a given day. Suppose a patient's systolic blood pressure readings during a given day have a normal distribution with a mean $\mu = 160$ mm mercury and a standard deviation $\sigma = 20$ mm.

a. What is the probability that a single blood pressure measurement will fail to detect that the patient has high blood pressure?

b. If five blood pressure measurements are taken at various times during the day, what is the probability that the average of the five measurements will be less than 150 and hence fail to indicate that the patient has high blood pressure? √

c. How many measurements would be required in a given day so that there is at most 1% probability of failing to detect that the patient has high blood pressure?

√ A random sample of 16 measurements is drawn from a population with a mean of 60 and a standard deviation of 5. Describe the sampling distribution of $\bar{y}$, the sample mean. Within what interval would you expect $\bar{y}$ to lie approximately 95% of the time?



email: khairil@apps.ipb.ac.id

Thank You

twitter: @kh_notodiputro