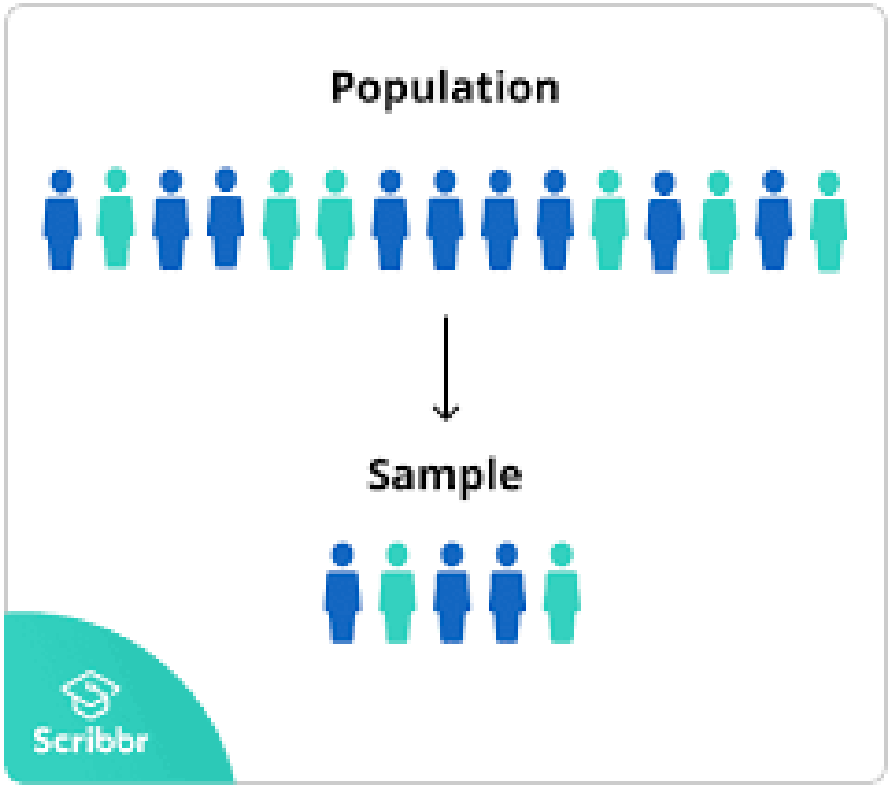
Engineering Statistics



Random Sampling

Dr. Vvn Weian Chao (趙韋安)
https://ce.nctu.edu.tw/member/teachers/23
Department of Civil Engineering, National Yang Ming Chiao Tung University, Taiwan

Purpose

描述統計

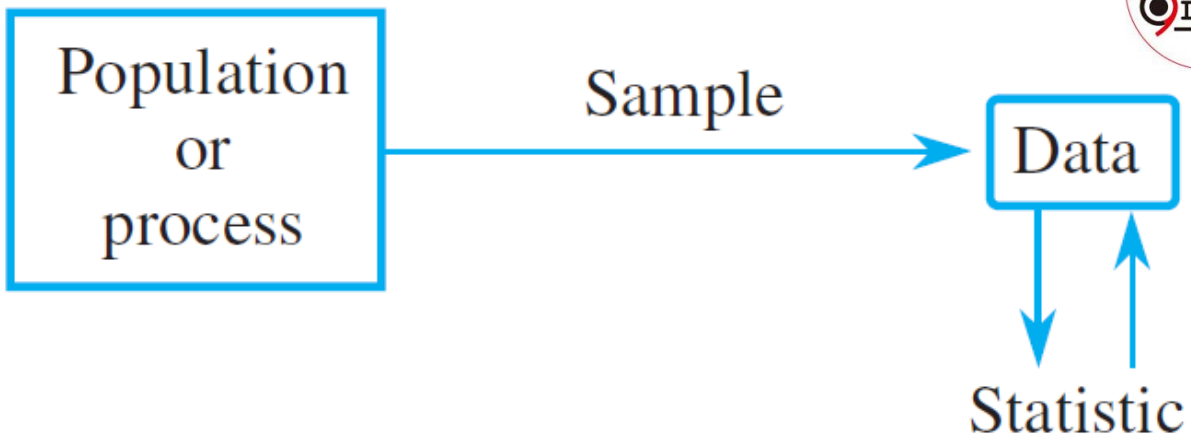
進入

推論統計

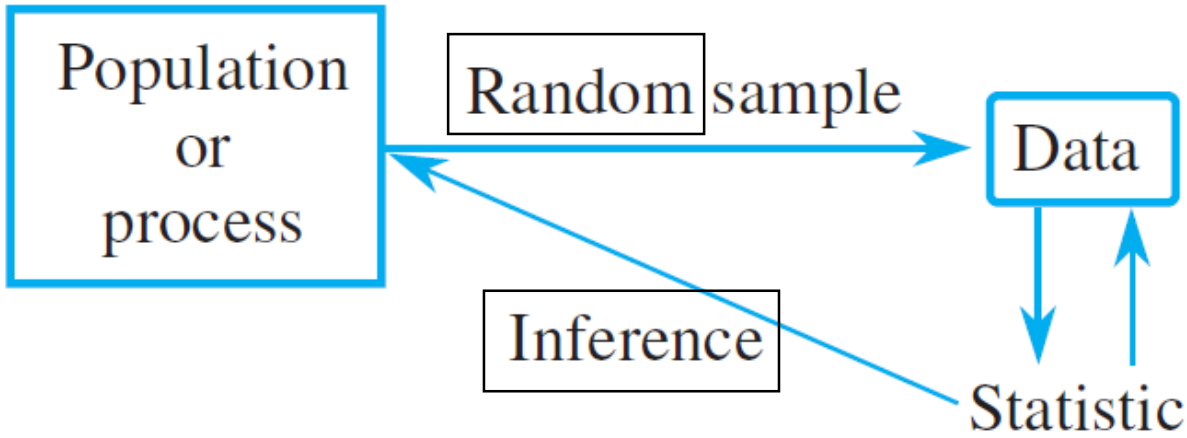
Random Sampling

Probability

Purpose


Population or process
Sample
Data
Statistic

(a)


Population or process
Random sample
Data
Inference
Statistic

(b)

Goal

To draw conclusions about larger entity

Data from sampling

Random

Stratified

Advantage

Savings in time and money

Advantage

Testing done on samples is often more reliable than testing on entire populations

隨著時間疲憊

Data from sampling

To reduce or eliminate bias

To make precise statements

Random number generator

replacement

Without replacement

Key parameter

The sample size used In Random sampling

取樣的數量n與本身的分析預算及準確率需求有關

R: Sampling

```
sample(
x,
size,
replace,
prob).
```

R: Sampling-
same random data

set.seed().

R: Sampling

generate random float number

runif(

n,min,max

).

TRY it in R

R: Sampling

R_sampling_a.R

Data from sampling

Random

Stratified

Advantage

Stratified sampling will generally produce estimates that are more **precise**

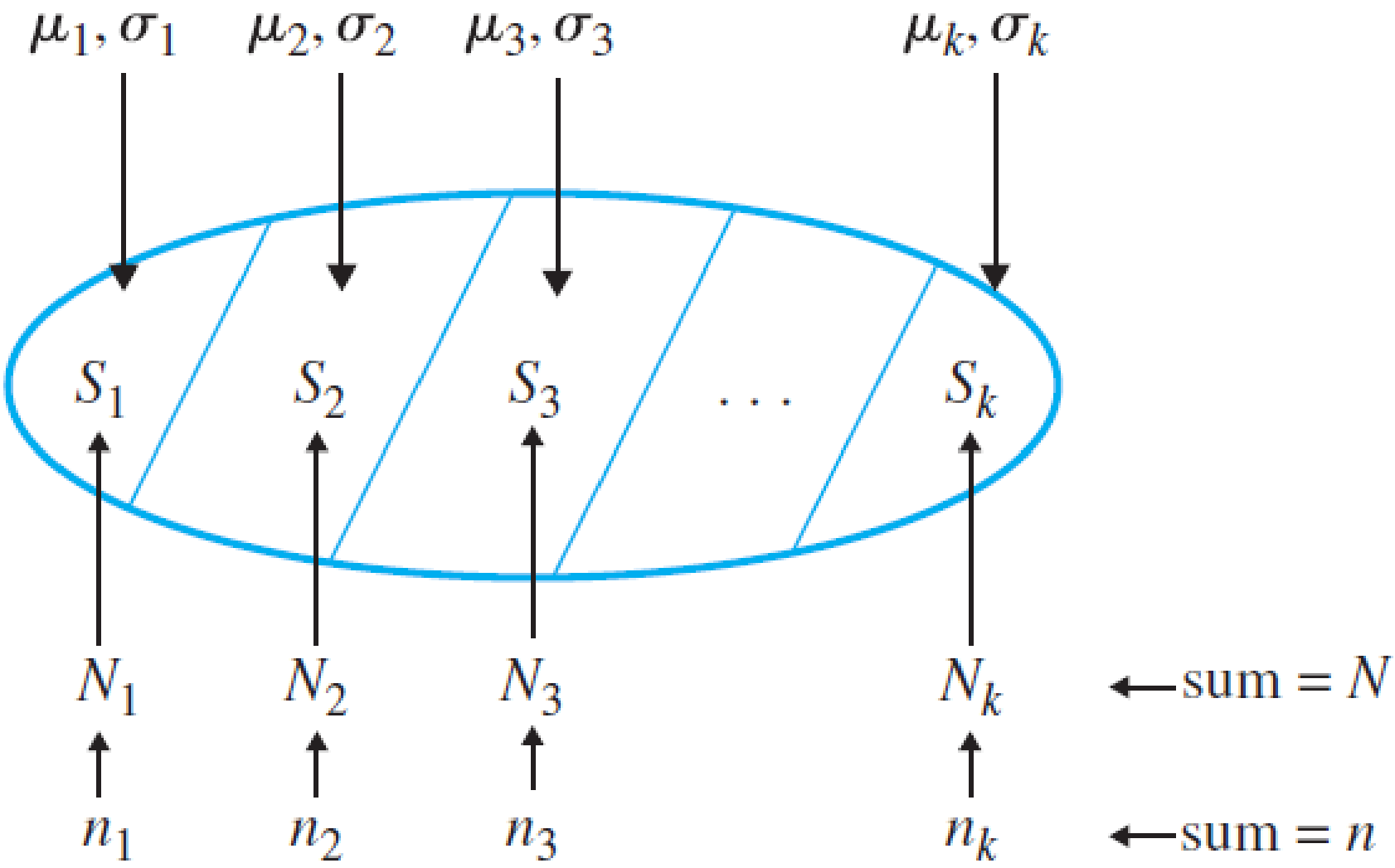
Stratified Sampling

先將母體分類成幾個群組➔決定各群組的**隨機抽樣個數**後再進行抽樣。

General Rules for Choosing Strata

- Decide on a response variable *y* that is of interest.
- Divide the entire population into nonoverlapping groups (i.e., strata) $S_1, S_2, \ldots, S_k$, each of which is as *homogeneous* as possible.
- Decide on the sample sizes $n_1, n_2, \ldots, n_k$ to select from the *k* strata.
- Use SRS to obtain a sample from each stratum.

Stratified Sampling



Estimating a Population Mean

- Given the w_i's, the N_i's, the σ_i's, a confidence level of **95%**, and B, it can be shown that the minimum necessary sample n for estimating the population mean μ to within a margin of error of $\pm B$ is

$$n = \frac{\sum_{i=1}^{k} \frac{N_i^2 \sigma_i^2}{w_i}}{N^2 \left(\frac{B}{1.96} \right)^2 + \sum_{i=1}^{k} N_i \sigma_i^2}$$

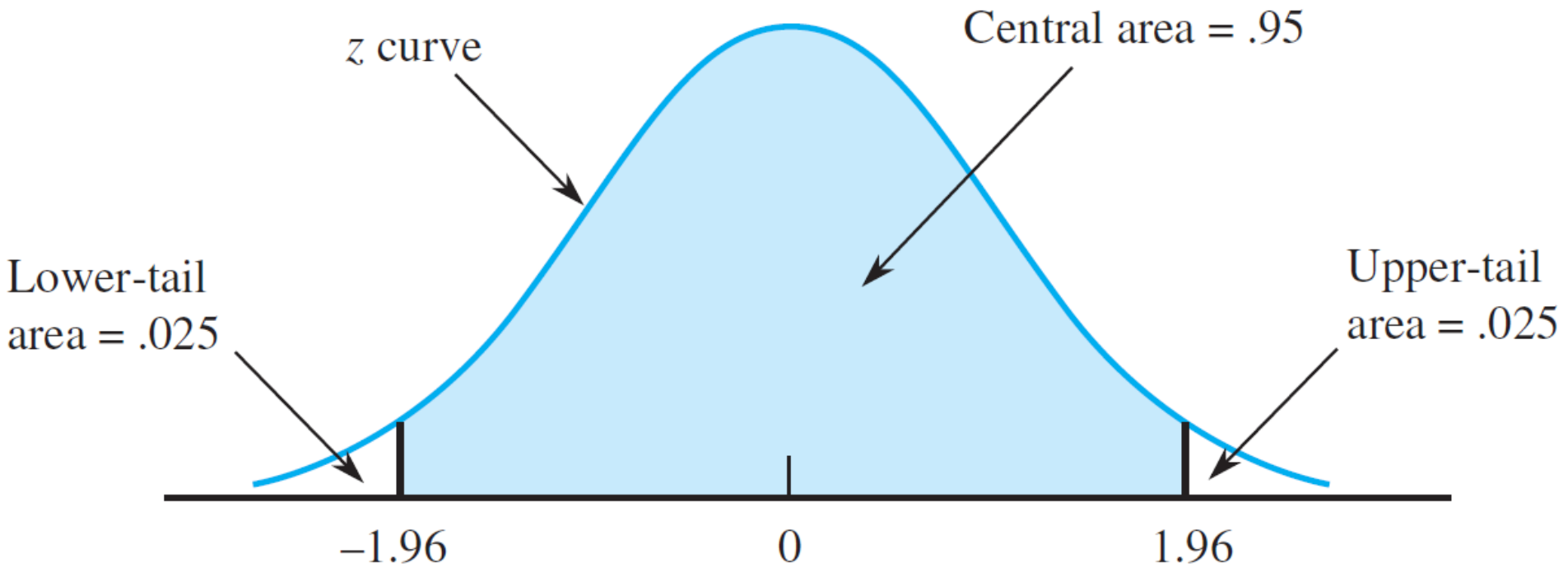
- For stratified sampling, the standard error is approximated by

$$s_{\text{str}} \approx \sqrt{\frac{1}{N^2} \sum_{i=1}^{k} N_i^2 \left(\frac{s_i^2}{n_i} \right) \left(\frac{N_i - n_i}{N_i - 1} \right)}$$

Where is the sample variance of the observations from stratum .

設定個母體權重、數量、母體標準差、信心水準及容許誤差➔決定**最少需要的總抽樣樣本數量，才能估計母體平均**

A confidence interval for μ with CL 95%



$$P(-1.96 \le \frac{\bar{x} - \mu}{s/\sqrt{n}} \le 1.96) = 0.95$$

Estimating a Population Mean

若將容許的誤差B變小，則需要增加抽樣樣本個數n。

1.96對應95%的信心水準(基於標準化常態分布)，**若要提高信心水準，亦需要增加n**。

$$n = \frac{\sum_{i=1}^{k} \frac{N_i^2 \sigma_i^2}{w_i}}{N^2 \left(\frac{B}{1.96} \right)^2 + \sum_{i=1}^{k} N_i \sigma_i^2}$$

Estimating a Population Mean: Example

問題: 自然環境與人為種植對於水質的影響。其中發現與特定作物的覆蓋面積有關。

Example 4.12

Since 1991 the USGS (U.S. Geological Survey) has conducted the National Water Quality Assessment Program (NAWQA), whose purpose is to study natural and human factors that affect water quality. One important measurement that NAWQA produces is an estimate of the percentages of a region covered by various crop types. In one study ("Validation of National Land-Cover Characteristics Data for Regional Water Quality Assessment," *Geocarto International*, Dec. vol. 10, no. 4 1995: 69–80) of the percentages of a region covered by corn crops, a region was divided into the following strata: A (irrigated crops), B (small grains and mixed crops), C (grasslands and small crops), D (wooded areas and crops), E (grasslands), and F (woods and pastures).

The region under study is first divided into smaller regions called *quadrats*, each with an area of 1 km^2. These subregions are then assigned to the various strata categories. Suppose that data from previous studies is used to obtain estimates of the standard deviations σ_i of the percentages of corn crops within each stratum and that this information is collected in the following table:

Estimating a Population Mean: Example

Stratum (S_i)	Stratum size (N_i)	Standard deviation (σ_i)
A	500	.2
B	300	.2
C	100	.3
D	50	.4
E	50	.6
F	200	.8

Since aerial photographs are used to estimate the percentage of corn coverage at a given site, the unit sampling costs will be about the same for each 1 km^2 subregion, so the Neyman allocation can be used. If we specify a [redacted]

$$n = \frac{\left[\sum_{i=1}^{k} N_i\sigma_i\right]^2}{N^2\left(\frac{B}{1.645}\right)^2 + \sum_{i=1}^{k} N_i\sigma_i^2}$$

$$= \frac{[500(.2) + 300(.2) + 100(.3) + 50(.4) + 50(.6) + 200(.8)]^2}{1200^2\left(\frac{0.10}{1.645}\right)^2 + [500(.2^2) + 300(.2^2) + \cdots + 200(.8^2)]}$$

$$= 109.68 \approx 110 \quad \text{(rounding to the nearest integer).}$$

抽樣樣本數至少需要110

Estimating a Population Mean: Example

估計各分類下所需的抽樣樣本個數

Stratum	N_i	σ_i	$N_i\sigma_i$	$n_i = n(N_i\sigma_i / \sum_{i=1}^{k} N_i\sigma_i)$
A	500	.2	100	$n_1 = 110(100/410) = 26.8 \approx 27$
B	300	.2	60	$n_2 = 110(60/410) = 16.1 \approx 16$
C	100	.3	40	$n_3 = 110(40/410) = 10.7 \approx 11$
D	50	.4	20	$n_4 = 110(20/410) = 5.4 \approx 5$
E	50	.6	30	$n_5 = 110(30/410) = 8.0 \approx 8$
F	200	.8	160	$n_6 = 110(160/410) = 42.8 \approx 43$

Estimating a Population Mean: Example

計算分類抽樣平均值及標準差

Stratum	n_i	N_i	$\bar{x}_i$	s_i
A	27	500	.52	.18
B	16	300	.22	.23
C	11	100	.02	.35
D	5	50	.06	.45
E	8	50	.01	.64
F	43	200	.67	.78

$$\bar{x}_{str} = .52(500/1200) + .22(300/1200) + \cdots + .67(200/1200)$$
$$= .39 \text{ (or, 39\%)}$$

39%是受到人為種植影響

- The estimated standard deviation that accompanies this estimate is $s_{str} \approx .03$ (or, 3%).

$$s_{str} \approx \sqrt{\frac{1}{N^2}\sum_{i=1}^{k} N_i^2 \left(\frac{s_i^2}{n_i}\right)\left(\frac{N_i - n_i}{N_i - 1}\right)}$$

Sampling distributions

How sampling distribution are used

General Properties of sampling distribution

How Sampling Distributions Are Used

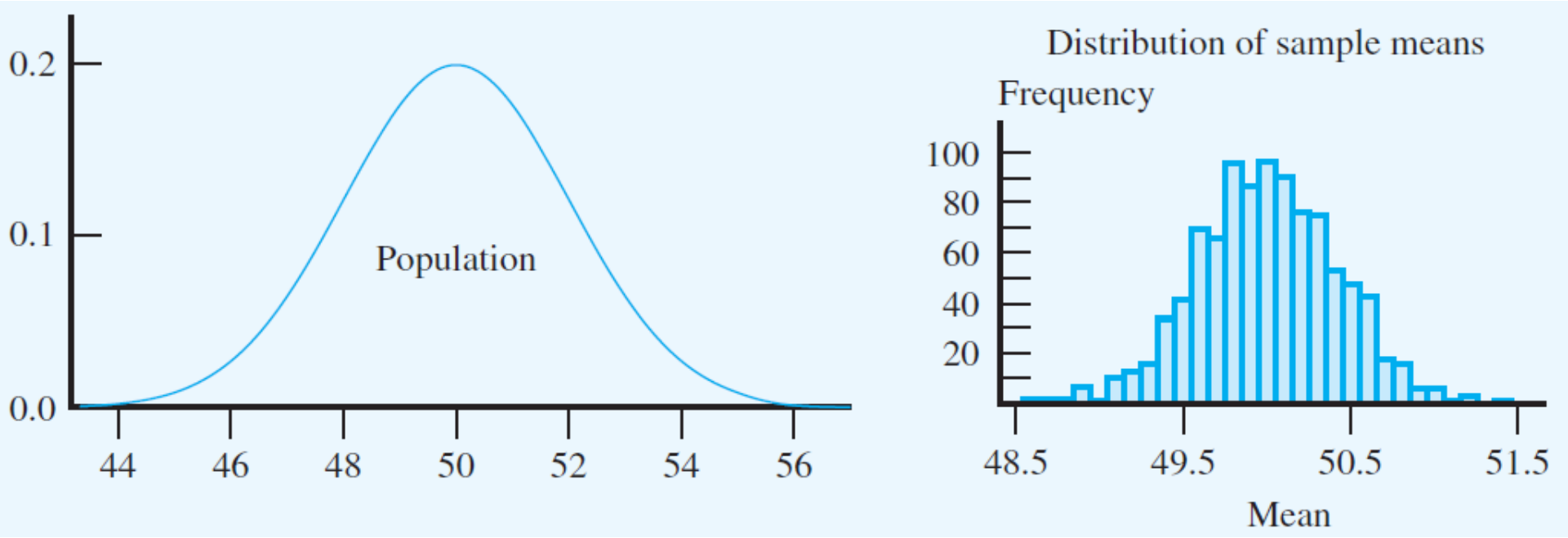
-怎樣的樣本分佈是適合的?

-一般樣本分佈主要是來表達: 從**母體抽樣**中的**樣本**可用於代表的統計意義(特徵值)

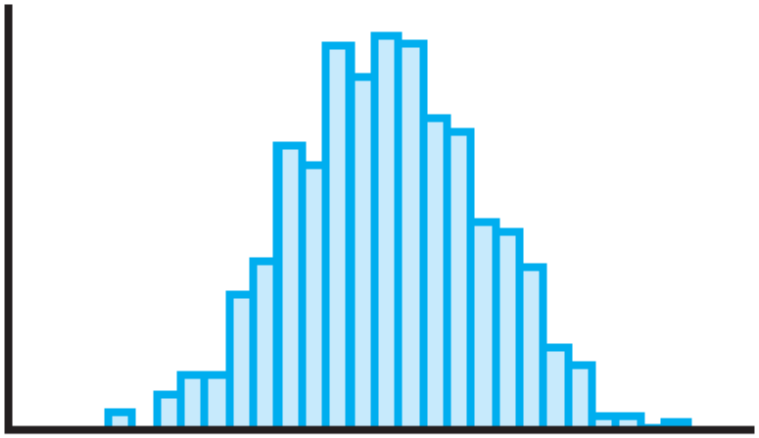
- To approximate the sampling distribution of a statistic, we **repeatedly select** a large number of random samples of **size *n*** from a given population.
- We calculate the value of the statistic for each sample and form a **histogram** of the results.
- We get an approximate picture of the sampling distribution of the statistic.

How Sampling Distributions Are Used

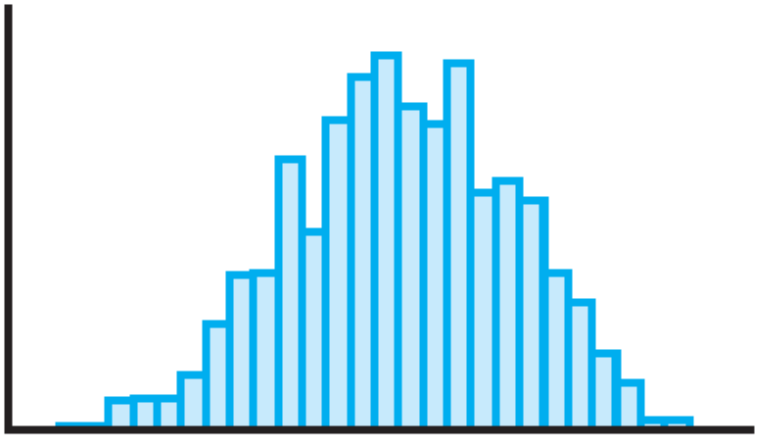
-從平均值為50標準差為2常態分佈(normal distribution)抽樣25個樣本，並計算其樣本平均數、中位數、樣本標準差、變異數。上述動作進行1000次。



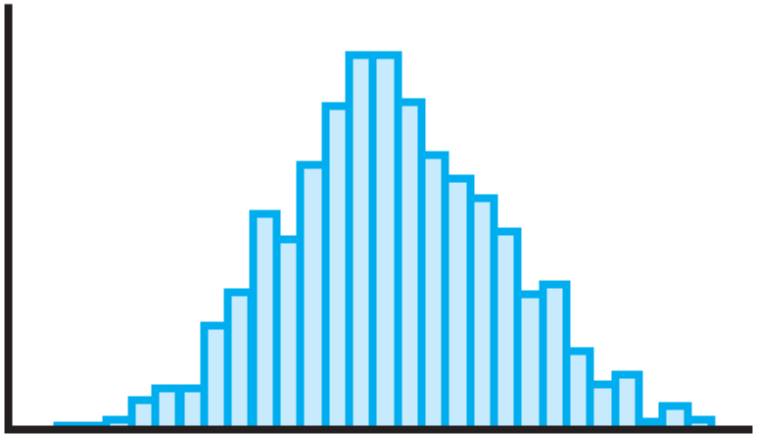
How Sampling Distributions Are Used



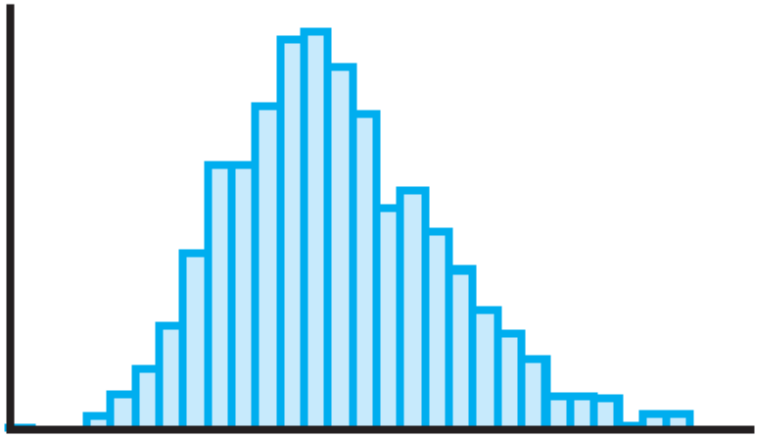
Mean



Median



Standard deviation



Variance

How Sampling Distributions Are Used

1. The sampling distribution of a statistic often tends to be *centered* at the value of the population parameter estimated by the statistic.
2. The spread of the sampling distributions of many statistics tends to grow smaller as the sample size *n* increases.
3. As the **sample size increases**, sampling distributions of many statistics become more and **more bell-shaped** (more and more like normal distributions).

How Sampling Distributions Are Used

-樣本平均值等於母體平均值

-樣本標準差小於母體標準差: 大部份數值落在平均值±1

Population parameter	Actual value	Sample mean of sampling distribution	Sample standard deviation of sampling distribution
			.418
Median	50	49.982	.515
			.2853
Variance, σ^2	4	4.0139	1.1528

-用途: 在母體平均值未知的情況下，我們可以透過抽樣分析求得的平均值來代表可能的母體平均值。

General Properties of Sampling Dist.

-其樣本分佈特徵為何?

-若觀察道樣本抽樣分佈有集中趨勢，代表此母體的統計估計是沒有偏差的。

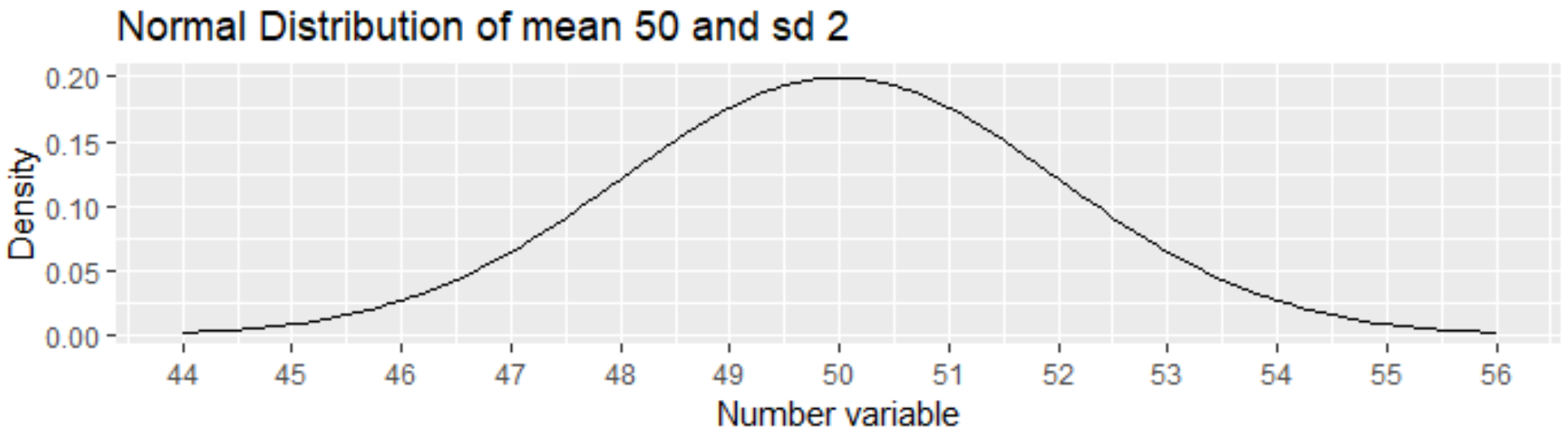
- With reference to Table 5.1, there is a close similarity between the population parameters and the means of the sampling distributions.
- As such, the ***center*** (i.e., the mean) may coincide with the corresponding population parameter.
- When this happens, the statistic is said to be **unbiased**, or that it is an **unbiased estimator** of the population parameter.

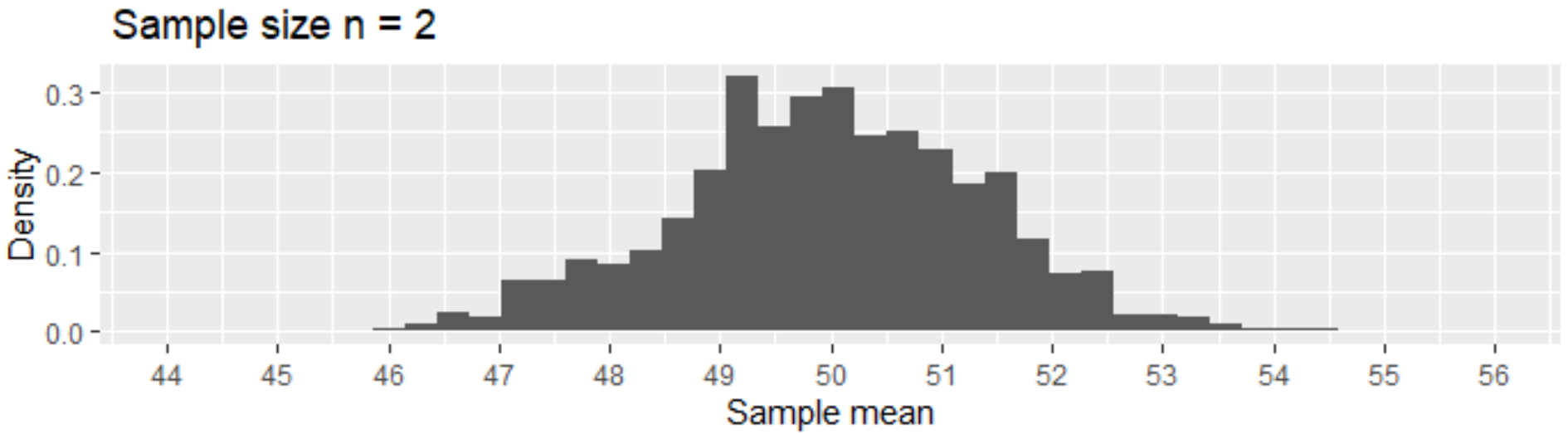
TRY it in R

R: Sampling
For loop in R

```
For(i in 1:1000)
{
}
```

R: Sampling


Normal Distribution of mean 50 and sd 2
Density
0.20
0.15
0.10
0.05
0.00
44
45
46
47
48
49
50
51
52
53
54
55
56
Number variable


Sample size n = 2
Density
0.3
0.2
0.1
0.0
44
45
46
47
48
49
50
51
52
53
54
55
56
Sample mean

R: Sampling

R_sampling_b.R

Describing Sampling Distributions

- Sampling Distribution of $\bar{x}$
- Sampling from a Normal Population
- The Central Limit Theorem
- Sampling Distribution of the Sample Proportion

Sampling Distribution of $\bar{x}$

-樣本平均值是否足夠推估母體平均值?

-定義樣本平均值的標準誤差(**Standard Error**)

$$\mu_{\bar{x}} = \mu \qquad \text{and} \qquad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

- Standard/Sampling Error:

標準誤差用於衡量樣本統計量的離散程度。

在**參數估計**與**假設檢定**中，它是用來衡量樣本統計量與母體參數之間差距的一個重要尺度。

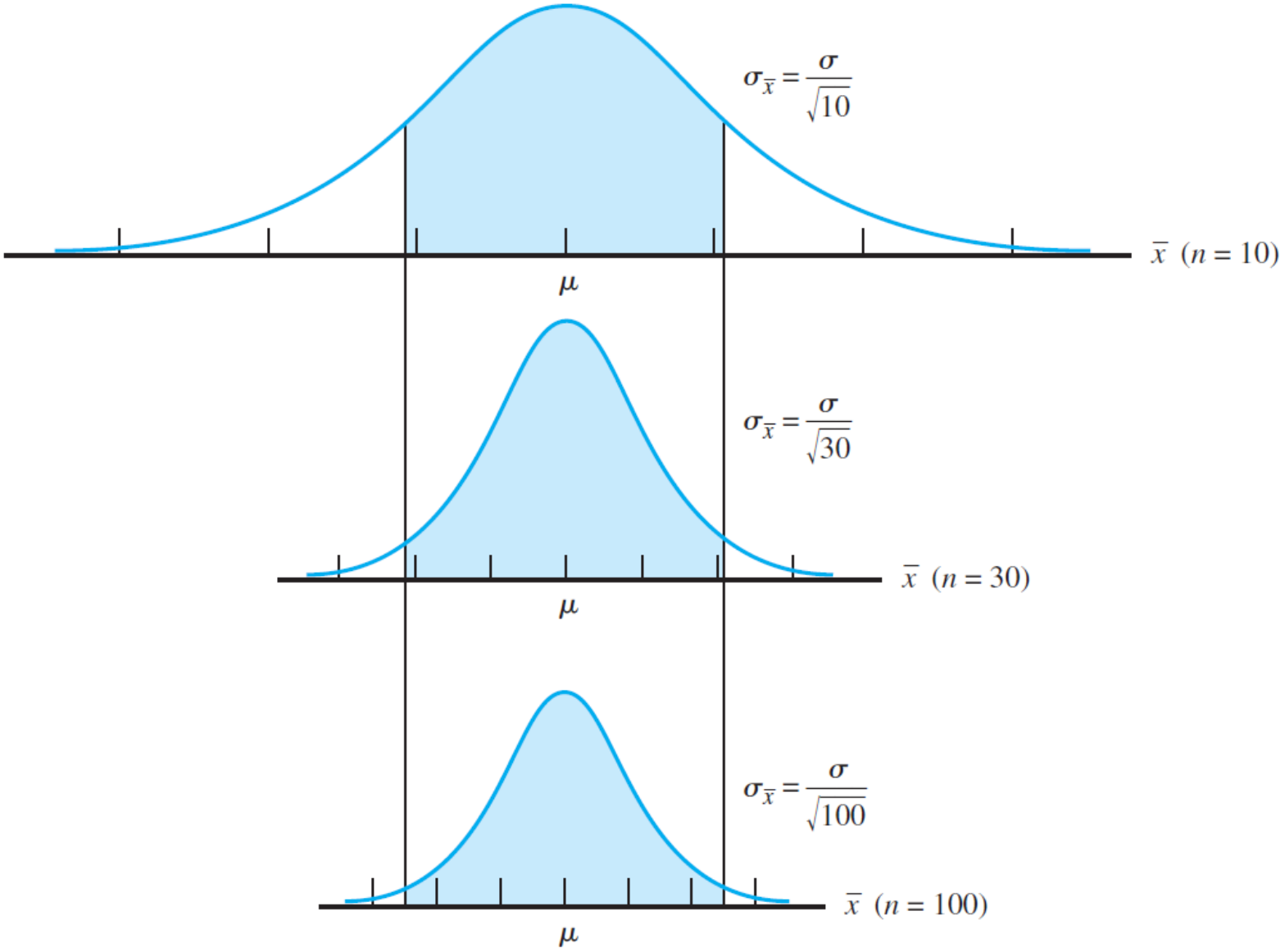
Sampling from a Normal Population

-如果母體本身符合常態分佈，則無論樣本數量n大小，樣本抽樣分佈應該都遵循常態分佈

Sampling Distribution of $\bar{x}$ (Normal Population)

- When a population distribution is normal, the sampling distribution of $\bar{x}$ is also normal, regardless of the size of the sample.
- With the normal distribution, probabilities of events involving $\bar{x}$ reduce to straightforward calculations.

Sampling from a Normal Population



Sampling from a Normal Population

-隨機抽樣樣本數n=5，當探討樣本平均數與實際母體平均數的差異落在2 cm的機率問題。在正規化為標準常態分佈，母體平均值會被移除。n增加，機率變大。

Example 5.18

Physical characteristics of manufactured products are often well described by normal distributions. Suppose, for example, that we want to evaluate the length (in cm) of certain parts in a production process based on the information in a random sample of five such parts. The parts are required to have a nominal length of 20 cm; past experience with this process indicates that the standard deviation is known to be $\sigma = 1.8$ cm. If we assume that the lengths can be described by a normal distribution, what is the probability that the mean of this sample will be within 2 mm of the current process mean μ? That is, what is the probability that $\bar{x}$ will lie between $\mu - 2$ and $\mu + 2$?

The solution to this type of problem lies in recognizing that the sampling distribution of $\bar{x}$ is normal with a mean of $\mu_{\bar{x}} = \mu$ and standard error of $\sigma_{\bar{x}} = \sigma/\sqrt{n} = 1.8/\sqrt{5} = .805$. To find the probability $P(\mu - 2 < \bar{x} < \mu + 2)$, we standardize, making sure to use the mean and standard error of $\bar{x}$ while doing this:

$$P(\mu - 2 < \bar{x} < \mu + 2) = P\left(\frac{\mu - 2 - \mu}{\frac{\sigma}{\sqrt{n}}} < z < \frac{\mu + 2 - \mu}{\frac{\sigma}{\sqrt{n}}}\right)$$

$$= P\left(\frac{-2}{.805} < z < \frac{2}{.805}\right) = .9868$$

That is, there is a 98.68% chance that the mean of a random sample of size $n = 5$ will be within 2 units of the population mean μ. Notice how the unknown mean μ cancels itself during the standardization. In other words, we do not need to know (or assume) a value for μ. Instead, when we select our sample of five parts, we can be relatively confident that the sample mean will be no farther than 2 cm from the true (unknown) process mean.

$$\mu_{\bar{x}} = \mu; \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

The Central Limit Theorem

-無論母體分佈為何，當抽樣樣本數n夠大時，抽樣樣本平均值分佈會趨近於常態分佈。

- The sampling distribution of $\bar{x}$ can be approximated by a normal distribution when the **sample size *n* is sufficiently large**, irrespective of the shape of the population distribution.
- The larger the value of *n*, the better the approximation.

The Central Limit Theorem

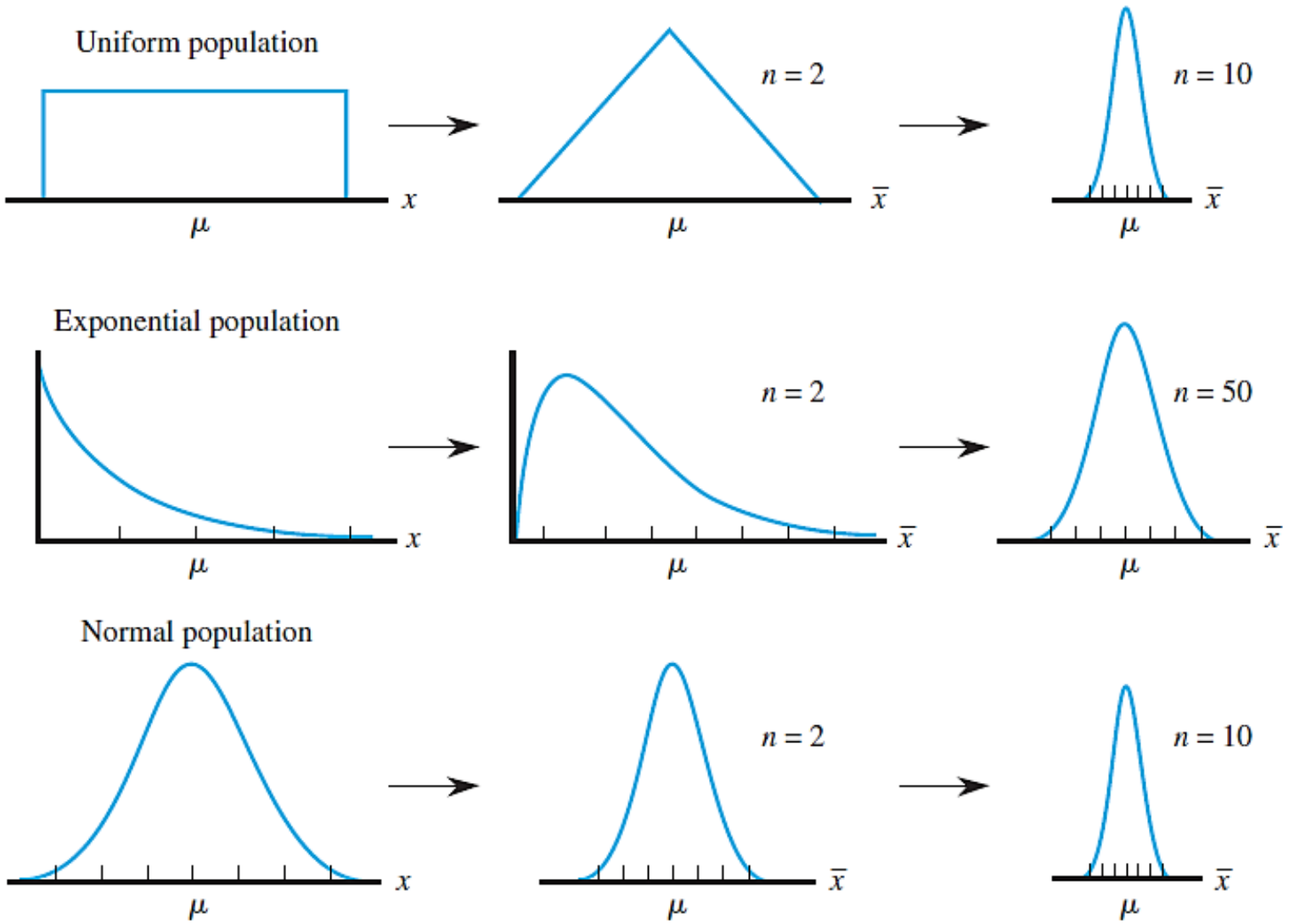


圖 5.20　中央極限定理：當樣本數 $\bar{x}$ 增加，n 的抽樣分配會逼近某一種常態分配。

The Central Limit Theorem

-大致上**樣本數大於30**，就可以滿足中央極值定理

-如果本身母體分佈對稱性差，則需要更大的抽樣樣本數量

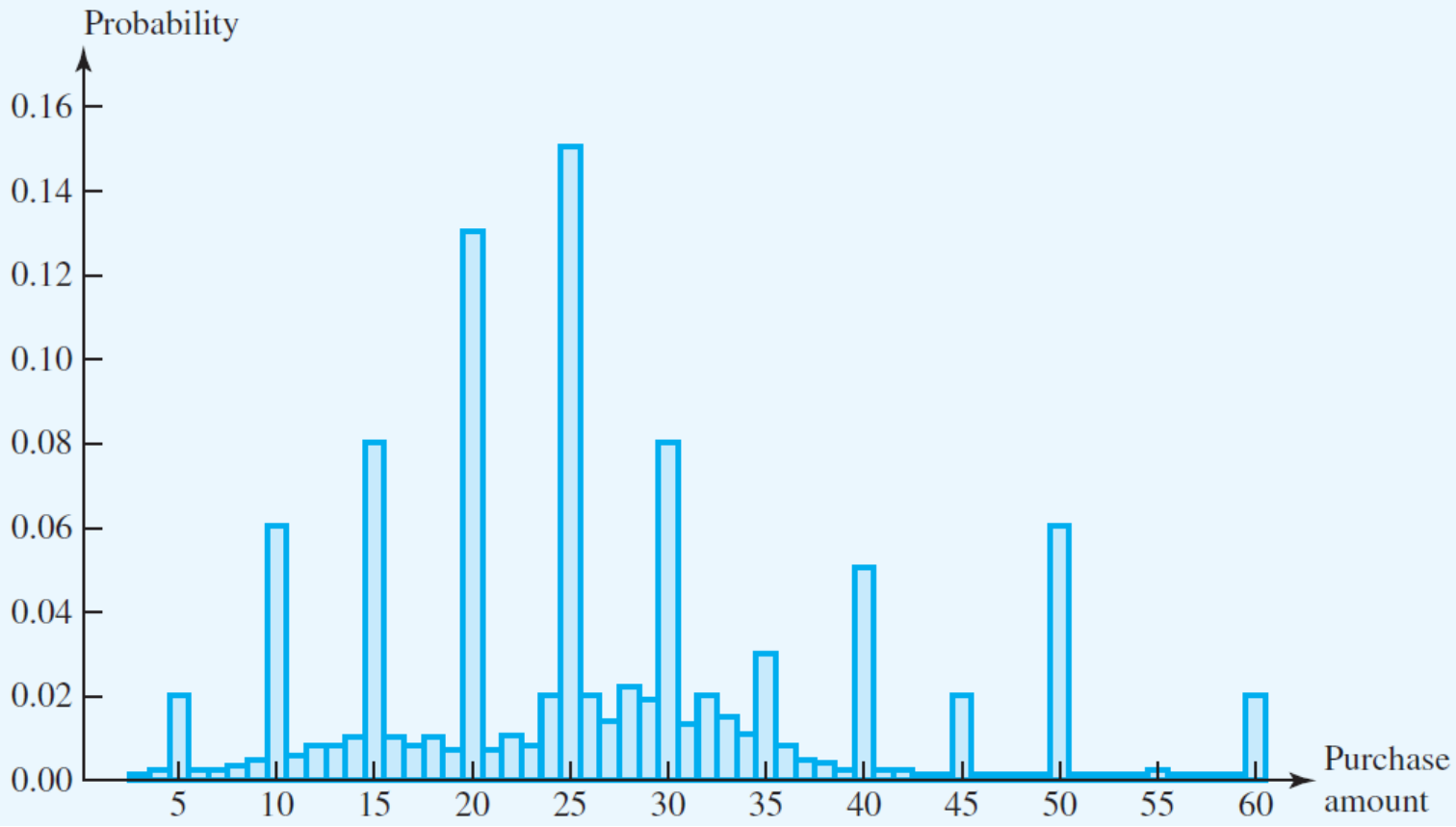
- is commonly used as a rough guide for what constitutes a "large enough" sample size for invoking the Central Limit Theorem.
- However, there are cases where smaller values of will suffice, as well as cases where larger sample sizes are needed.
- As a rule, *the less symmetric a population is, the larger the sample size will have to be to ensure normality of* .

The Central Limit Theorem

-汽油購買數量的統計直方圖。

-抽樣樣本數n=15，進行1000次，並於每次計算平均值。

-觀察平均值分佈情形。



The Central Limit Theorem

and calculate the value of the sample mean $\bar{x}$ for each one. Figure 5.22 is a histogram of the resulting 1000 values; this is the approximate sampling distribution of $\bar{x}$ under the specified circumstances. This distribution is clearly approximately normal even though the sample size is not very large. A normal quantile plot based on the 1000 $\bar{x}$ values exhibits a very prominent linear pattern.

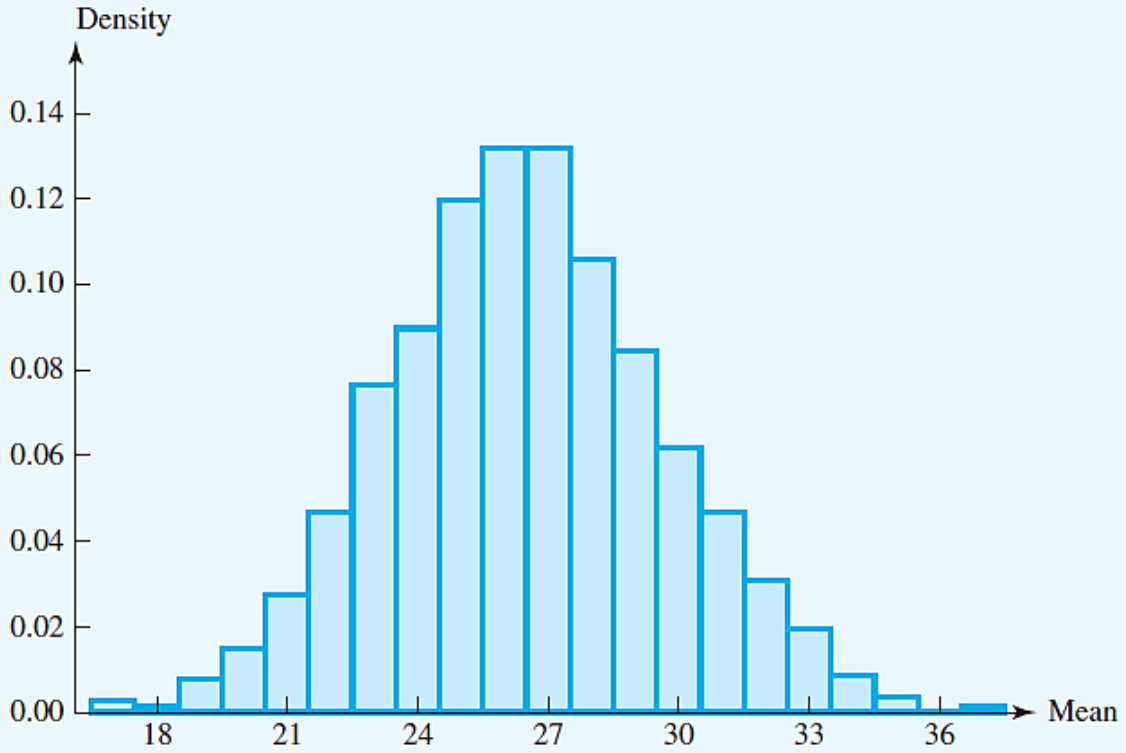


圖 5.22　顯示於圖 5.21 之母體且 $n = 15$ 的時候，樣本平均購買量的近似抽樣分配。

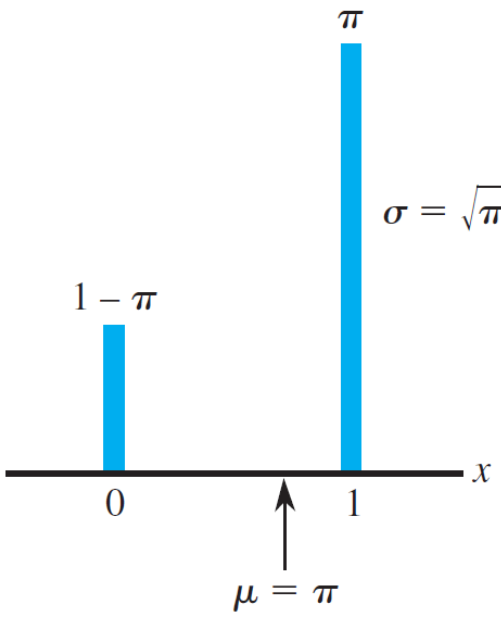
Sampling Distribution of the Sample Proportion

-樣本比例(sample proportion): 為樣本中某種元素的數量除以樣本大小。

-以樣本比例為隨機變數，其機率分佈即為樣本比例的抽樣分佈。

-滿足樣本抽樣分佈為常態分佈條件

nπ ≧ 5 and n(1- π) ≧ 5



$$\mu_p = \pi \qquad \text{and} \qquad \sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}}$$

Sampling Distribution of the Sample Proportion

-一個製成的過程中，產品非符合標準的機率有5%。該公司將每日抽樣100個樣本數來檢驗是否符合5%。如果有一天檢驗結果顯示為12%，則該如何解釋? **其機率相當低**

Example 5.21
管制圖

p 圖

Control charts are graphs that monitor the movements in a sample statistic (such as $\bar{x}$ or p) in periodic samples taken from an ongoing process. Using the sampling distribution of the statistic as a yardstick, values of the statistic "too far" away from the center of the sampling distribution are taken to be signals of possible problems with the process. For example, a p chart is often used to monitor the proportion of nonconforming products in a manufacturing process. Using past data from [redacted] value of π is selected as being representative of the long-run behavior [redacted] Suppose, for example, that a certain process constantly generates an ave[redacted] 5% nonconforming products and that samples of size 100 are taken each day to test whether the 5% nonconformance rate has changed. On one particular day, 12 nonconforming products appear in the sample. How do we interpret this information?

Assuming that the process is behaving as it has in the past, we set $\pi = .05$. For this value of π, $n\pi = 100(.05) = 5$ and $n(1-\pi) = 100(.95) = 95$, so the condition for applying the normal approximation is met. Furthermore, the mean and standard deviation of the sampling distribution of p can be calculated:

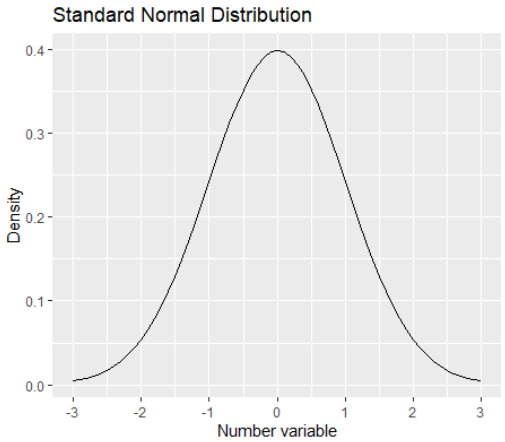
$$\mu_p = .05 \quad \text{and} \quad \sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}} = \sqrt{\frac{(.05)(1-.05)}{100}} = .0218$$

[redacted]

probability of occurring *if* the process is running as usual. Our conclusion is that it is more likely that something has caused an increase in the process nonconformance rate.

常態分布近似

$$\mu_p = \pi \quad \text{and} \quad \sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}}$$



TRY it in R

R: Sampling
Data sorting
dplyr.
select
filter

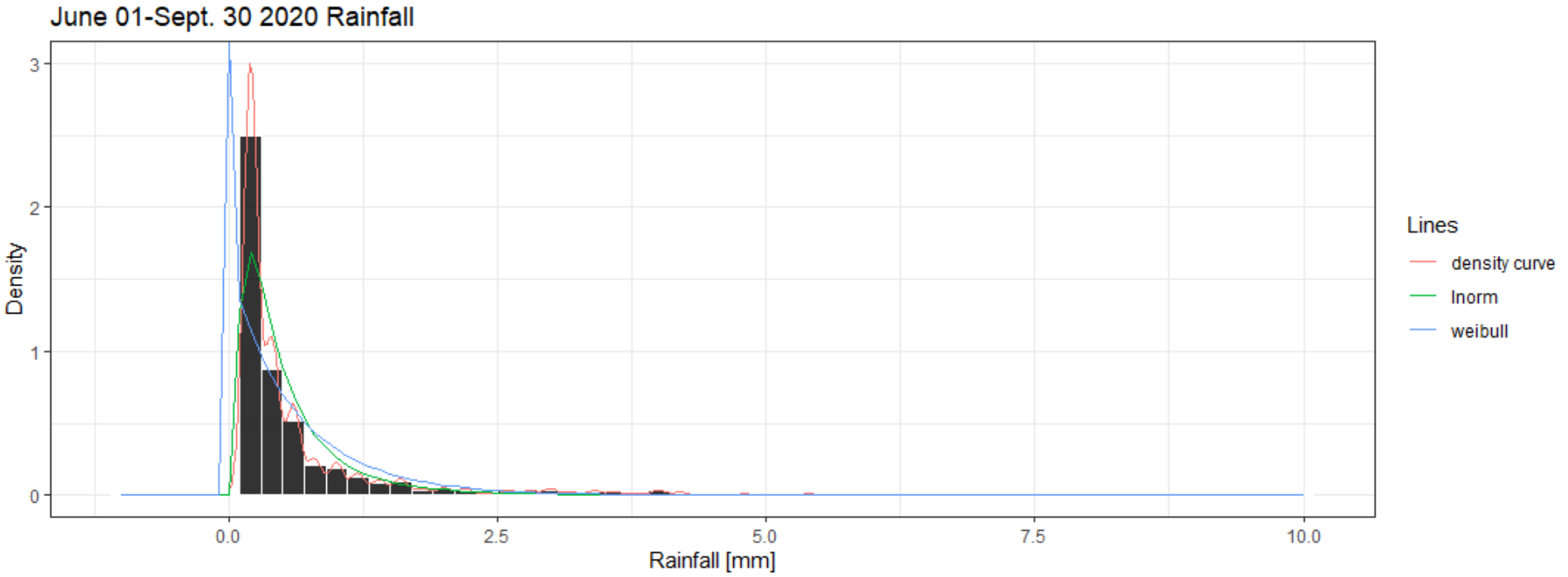
R: dply

If you meet trouble,

Error in select(., ..) : unused argument.

Please restarting your R session: Ctrl + Shift + F10 and running your code again

R: Sampling

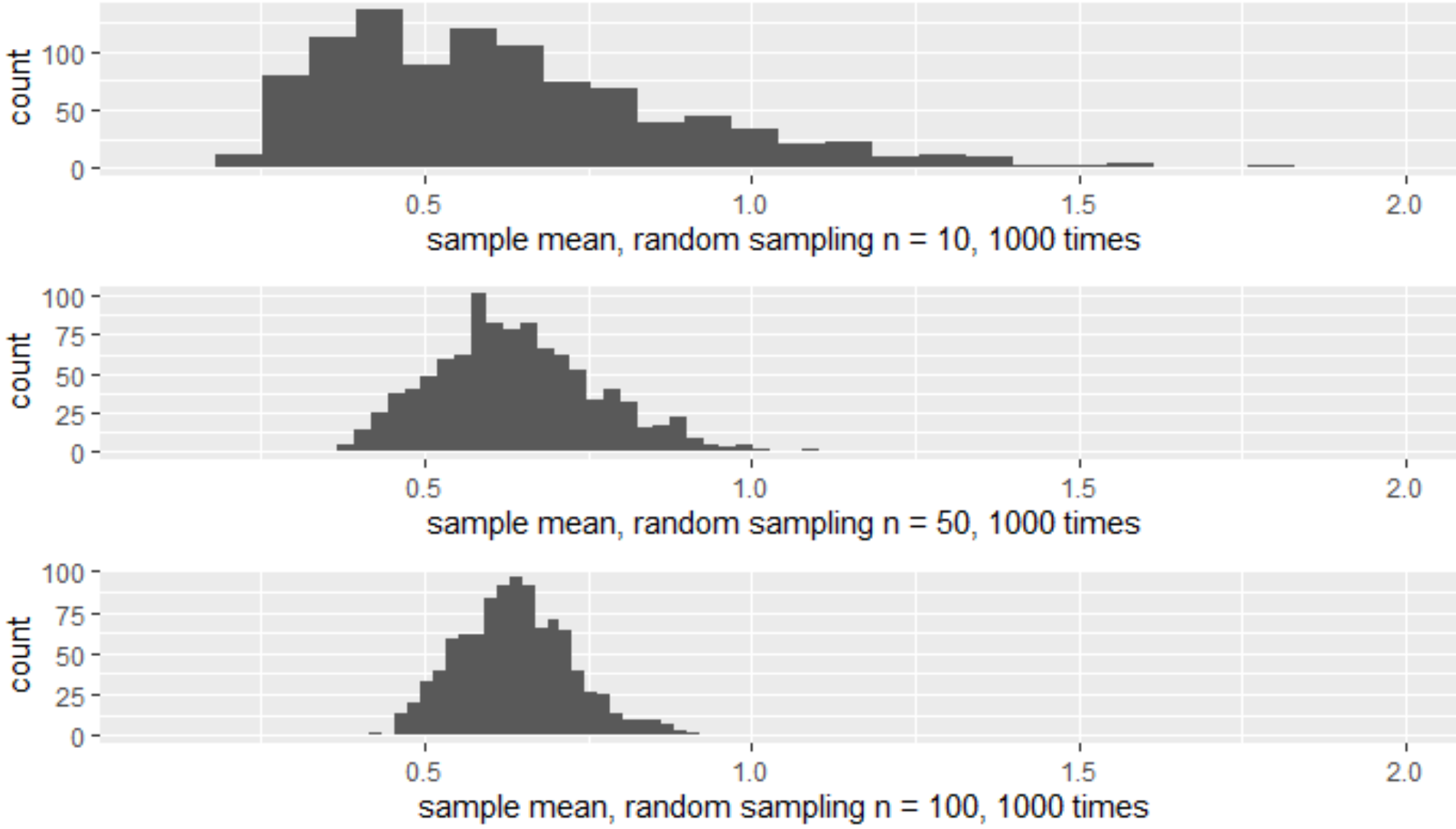

June 01-Sept. 30 2020 Rainfall
Density
3
2
1
0
0.0
2.5
5.0
7.5
10.0
Rainfall [mm]
Lines
density curve
lnorm
weibull

R: Sampling

R_sampling_c1.R

R: Sampling

分布中心應接近0.644



R: Sampling

R_sampling_c2.R

課堂練習: 學號-姓名-ch7-Sampling.R

Reading temperature data in “weatherdata.xlsx”:
試著回答以下問題:

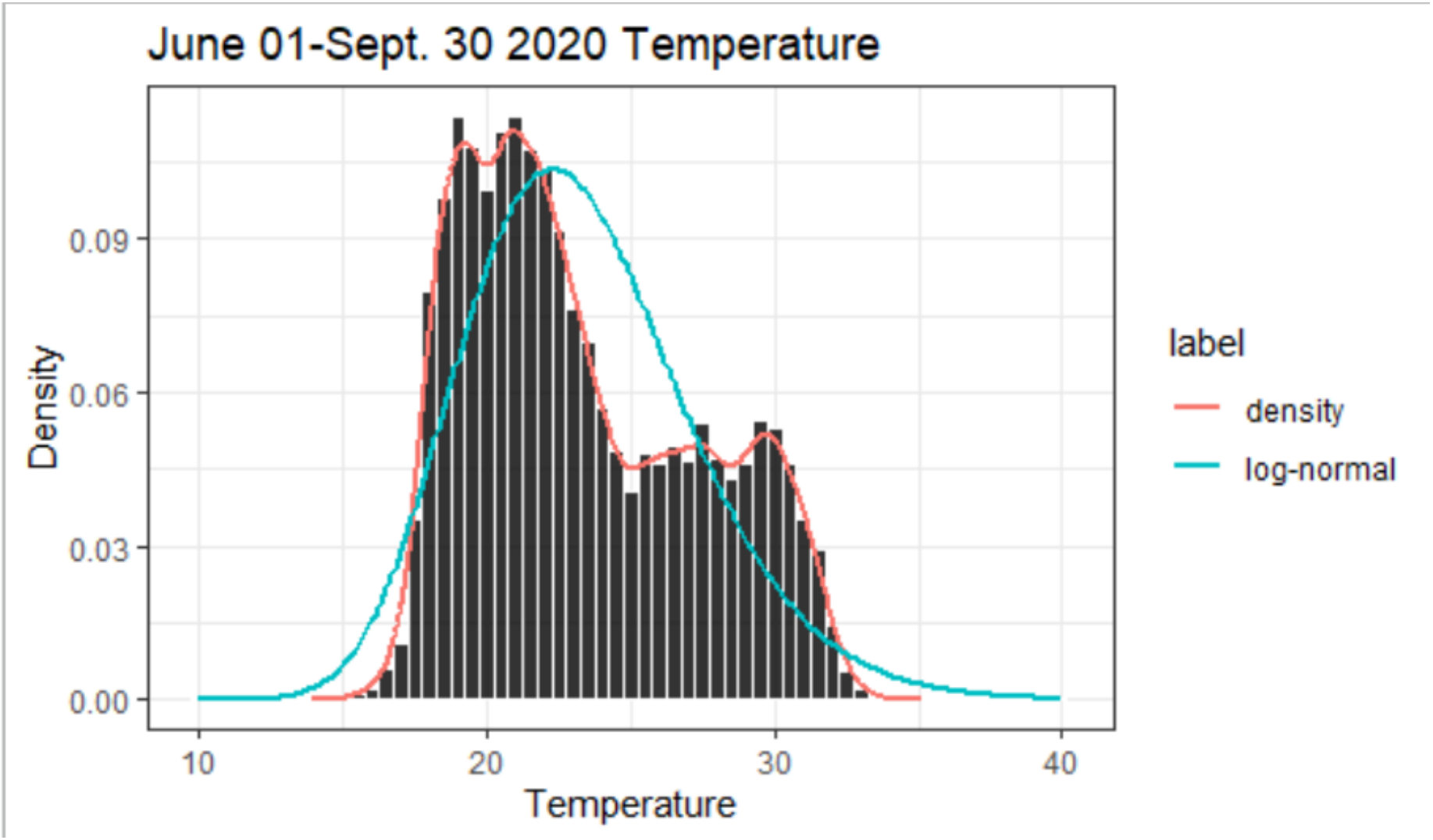
(1) Fitting the data with “log-Normal” distribution and then comparing it with density curve. Make a description on the plotting result.
(2) Based on the temperature dataset, please sampling 1,000 times with sample sizes of 10, 50, 100, and see what’s happened and make a comment.

課堂練習: 學號-姓名-ch7-Sampling.R

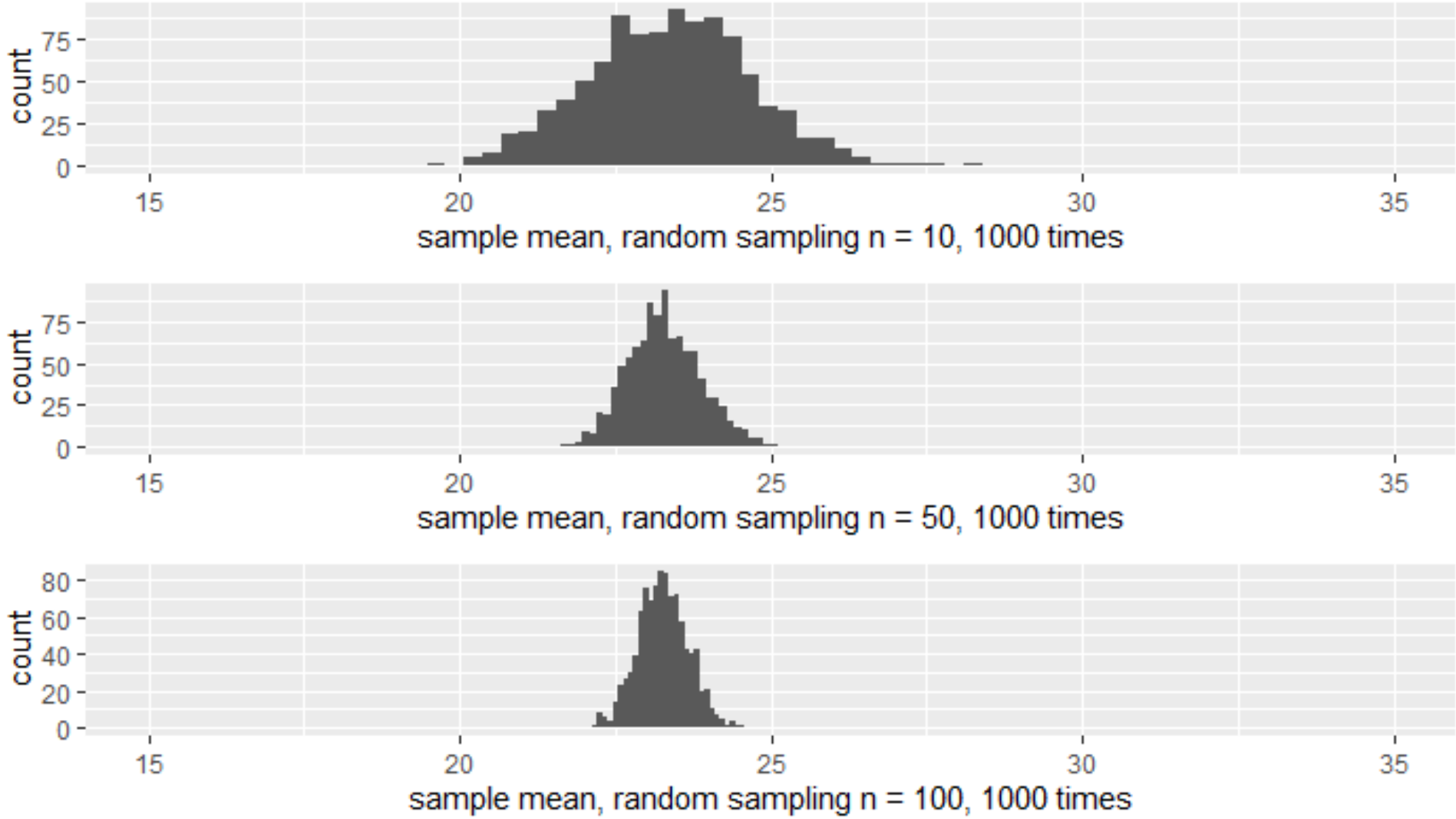
Hint:

```
library(dplyr)
d.temp <- d %>%
  select(Temperature)   # select the Temperature data
str(d.temp)
d.temp <- as.data.frame(d.temp)
```

課堂練習: 學號-姓名-ch7-Sampling.R

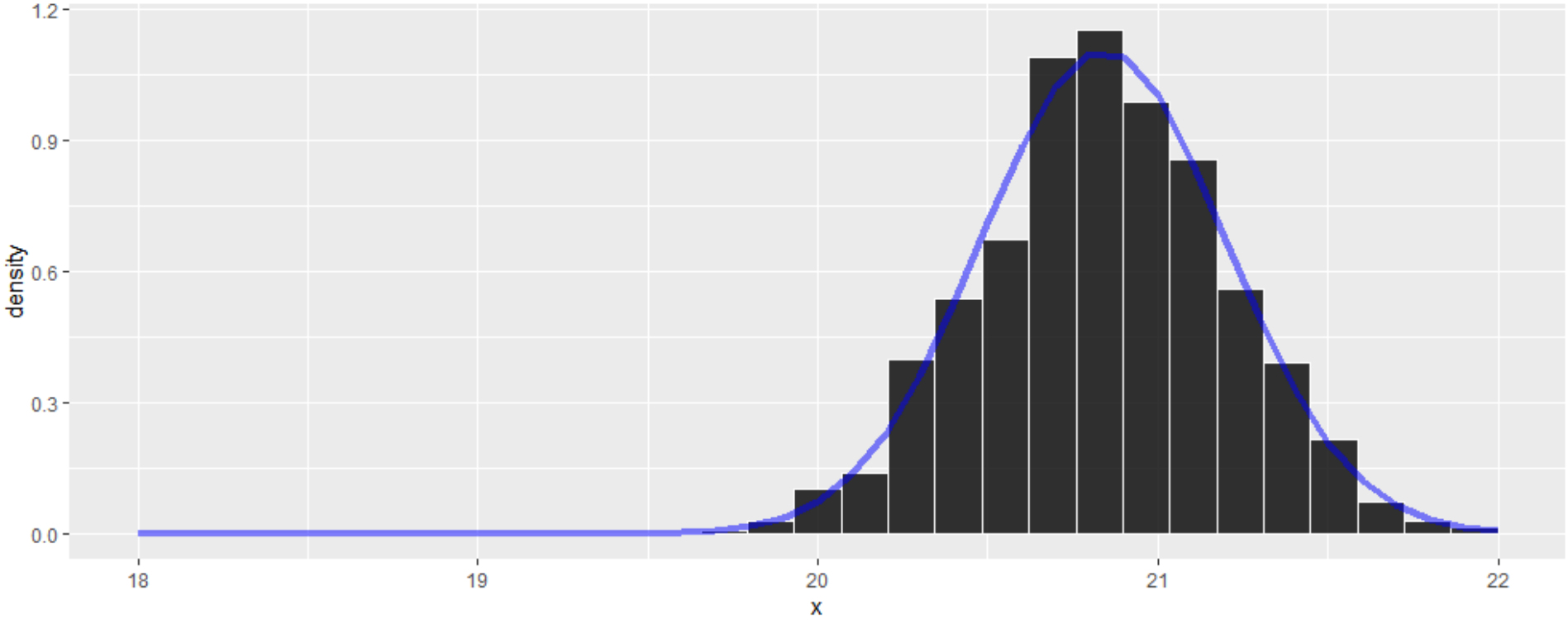


課堂練習: 學號-姓名-ch7-Sampling.R



R: Sampling

額外練習思考:
如何?
只分析溫度小於25度且晚上時間區段(hour>17)



R: Sampling

Hint:

```
library(dplyr)
d.t <- d %>%
   select(Date, Temperature) %>%
   filter(Temperature <= 25.0)

library(lubridate)
d.t$Date <- mdy_hms(d.t$Date)

d.t.sort <- d.t %>%
   filter(hour(d.t$Date) > 17 )
```