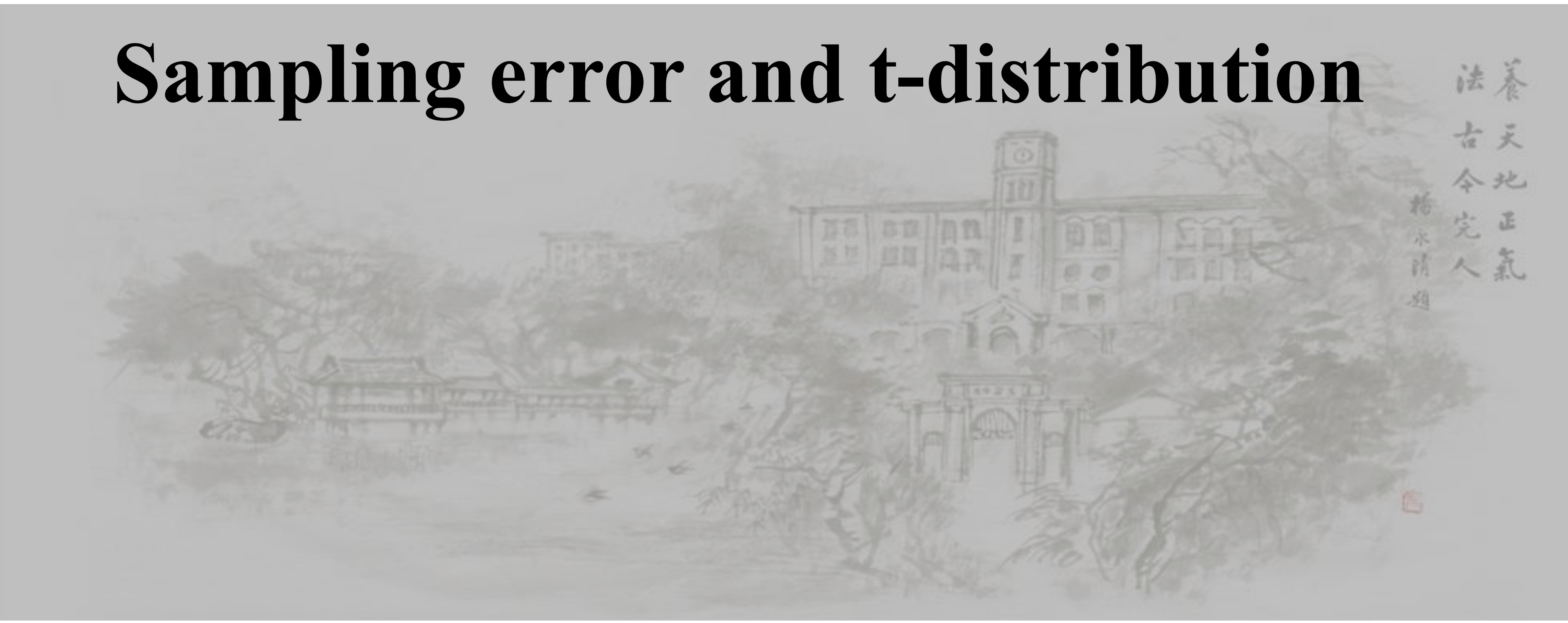
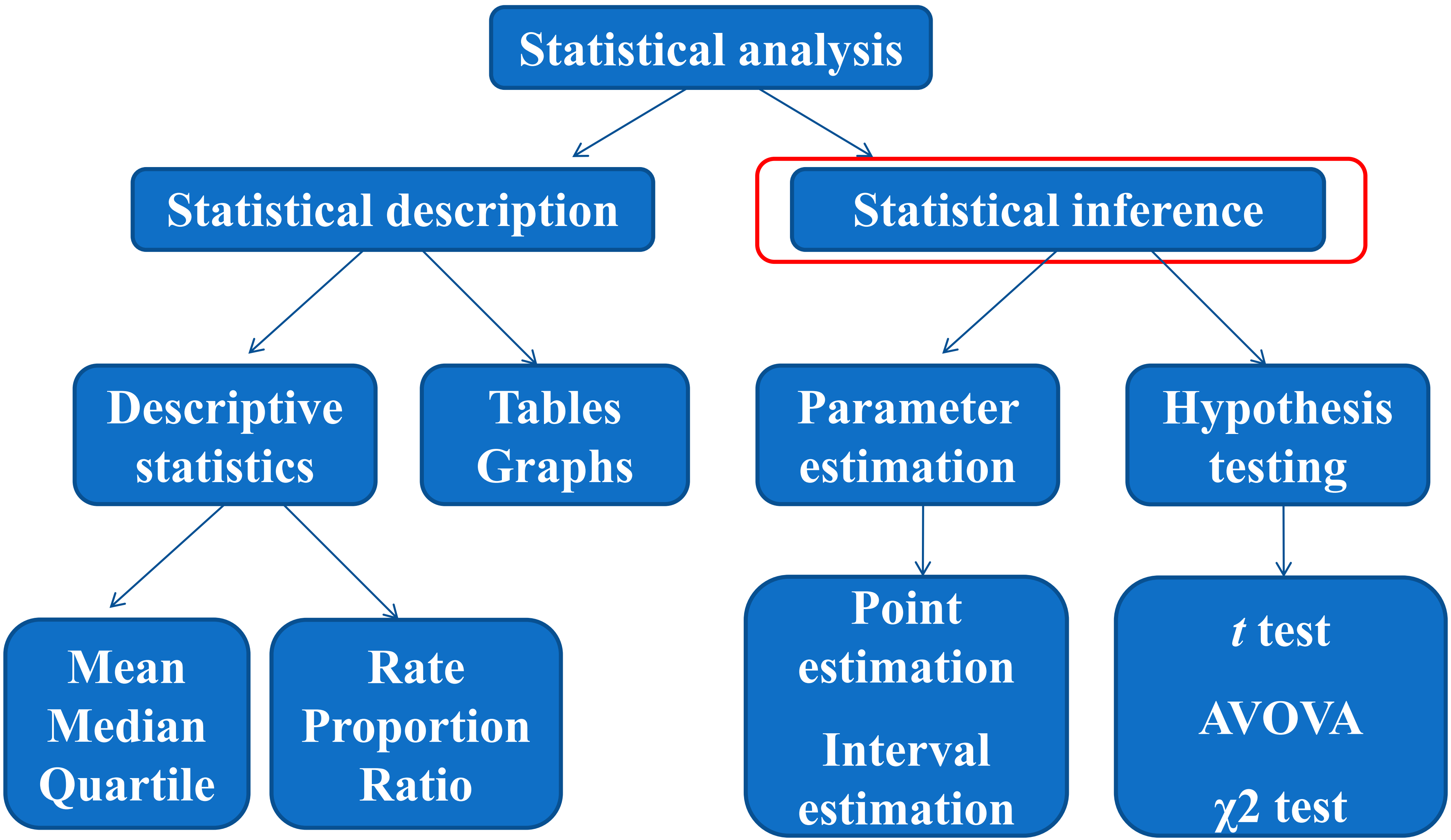
Sampling error and t-distribution



Procedures of Statistical Analysis



What is sampling distribution?

Distribution of sample statistics

Main contents

1
- Sampling distribution of sample means (sampling error)

2
- Sampling distribution of sample rates

3
- *t* distribution

1. Sampling distribution of sample means

Sampling study

parameter?

μ、σ、π

statistic

$\bar{x}$、s、p

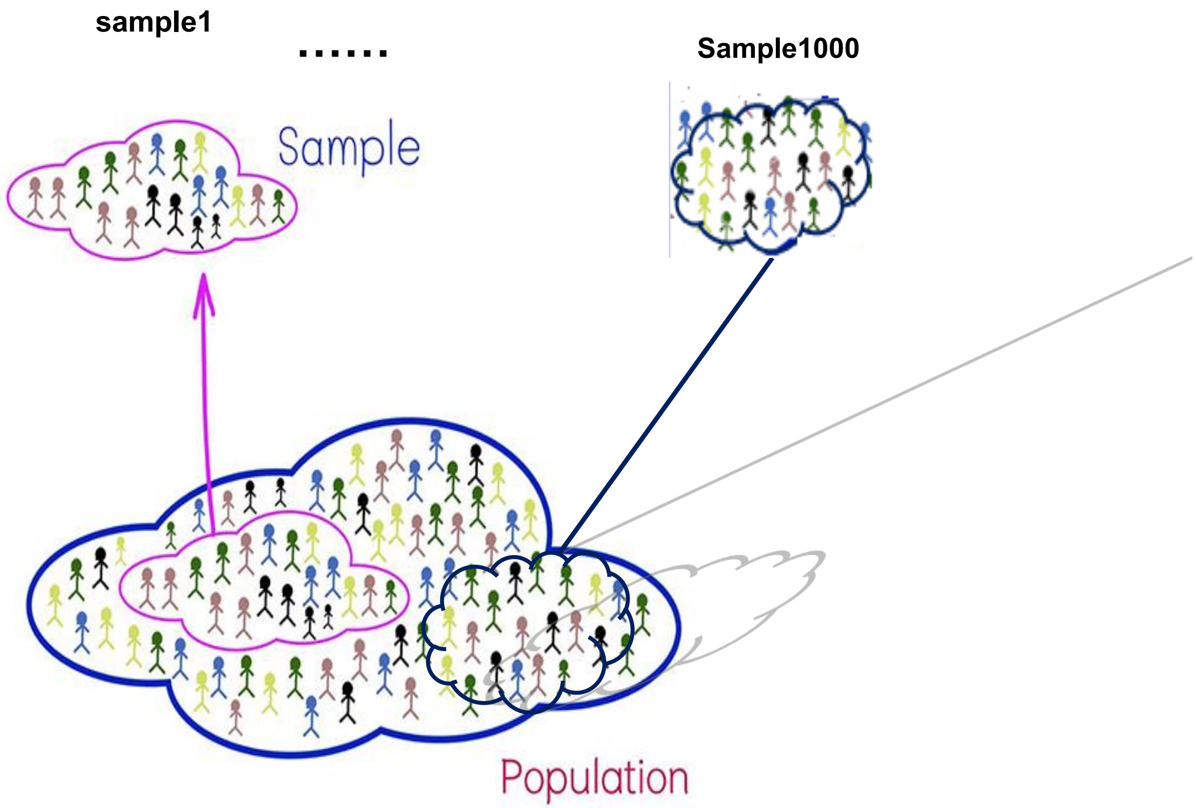
population

Randomly

inference

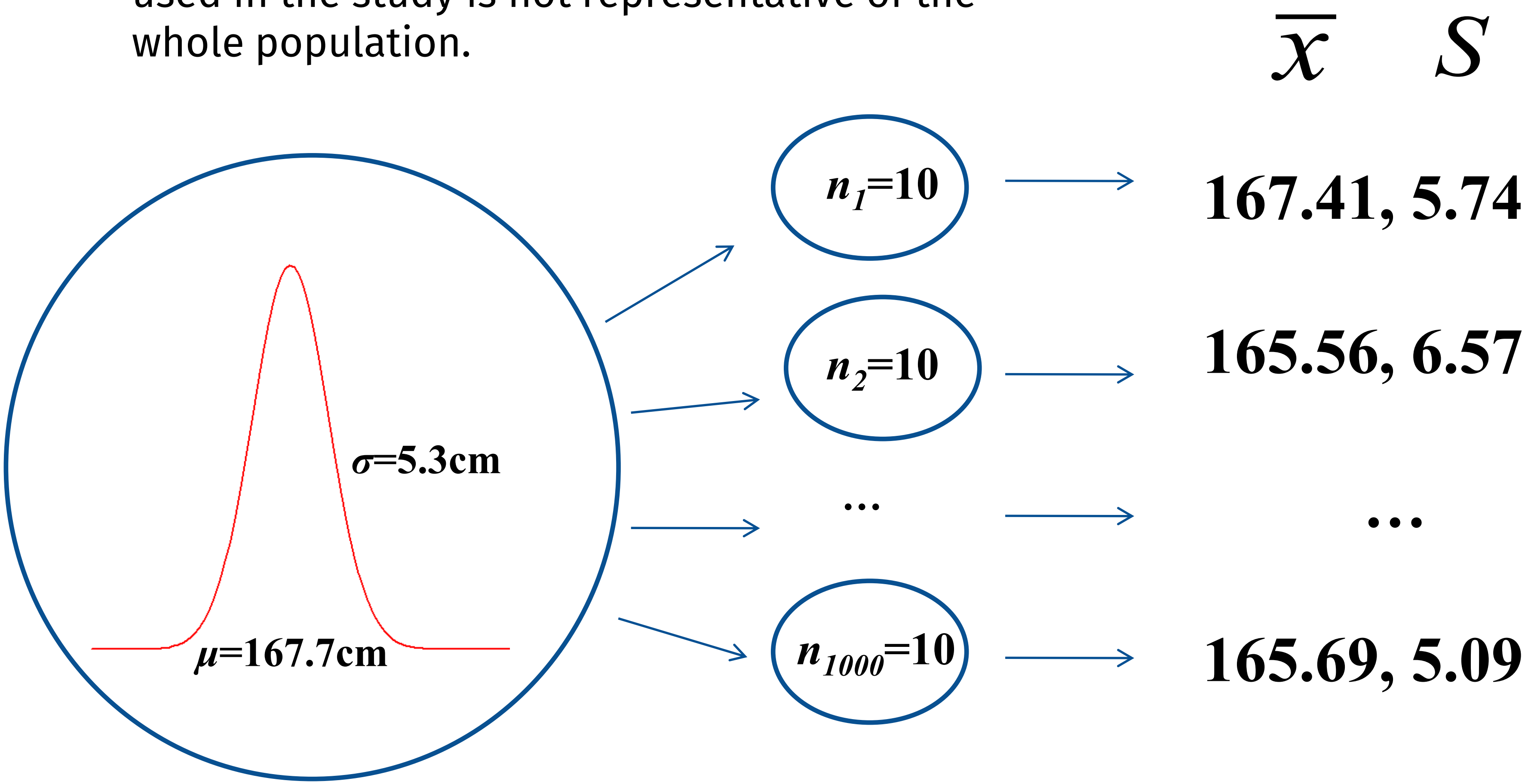
sample

Sampling error

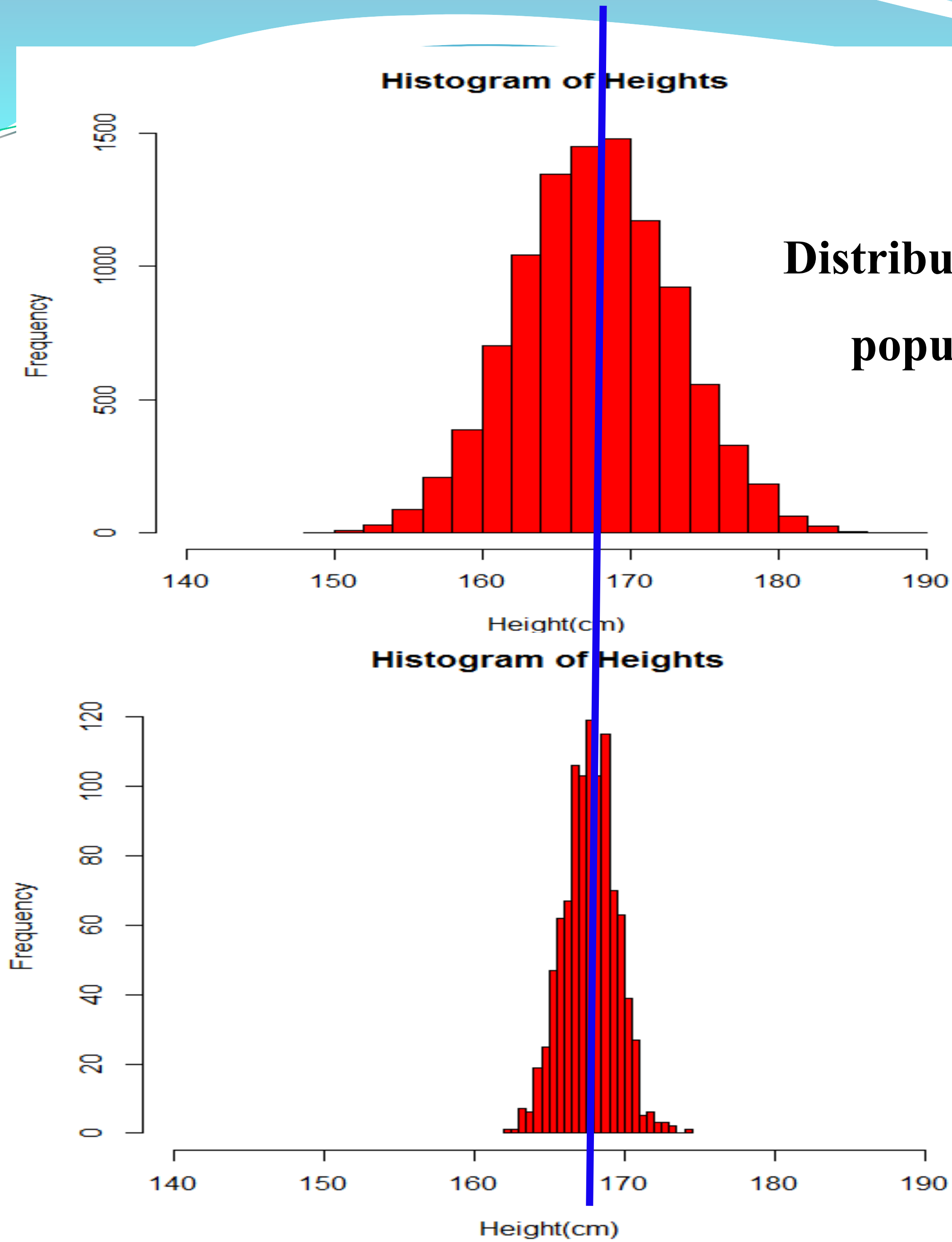


Female height(cm): X~N(167,5.3)

A sampling error occurs when the sample used in the study is not representative of the whole population.



$$\bar{\bar{x}} = 167.5 \qquad S_{\bar{x}} = 1.67$$



Distribution of observation values from the population with μ=167.7cm, σ=5.3cm

Sampling error:

$\overline{x1} \neq \overline{x2}$

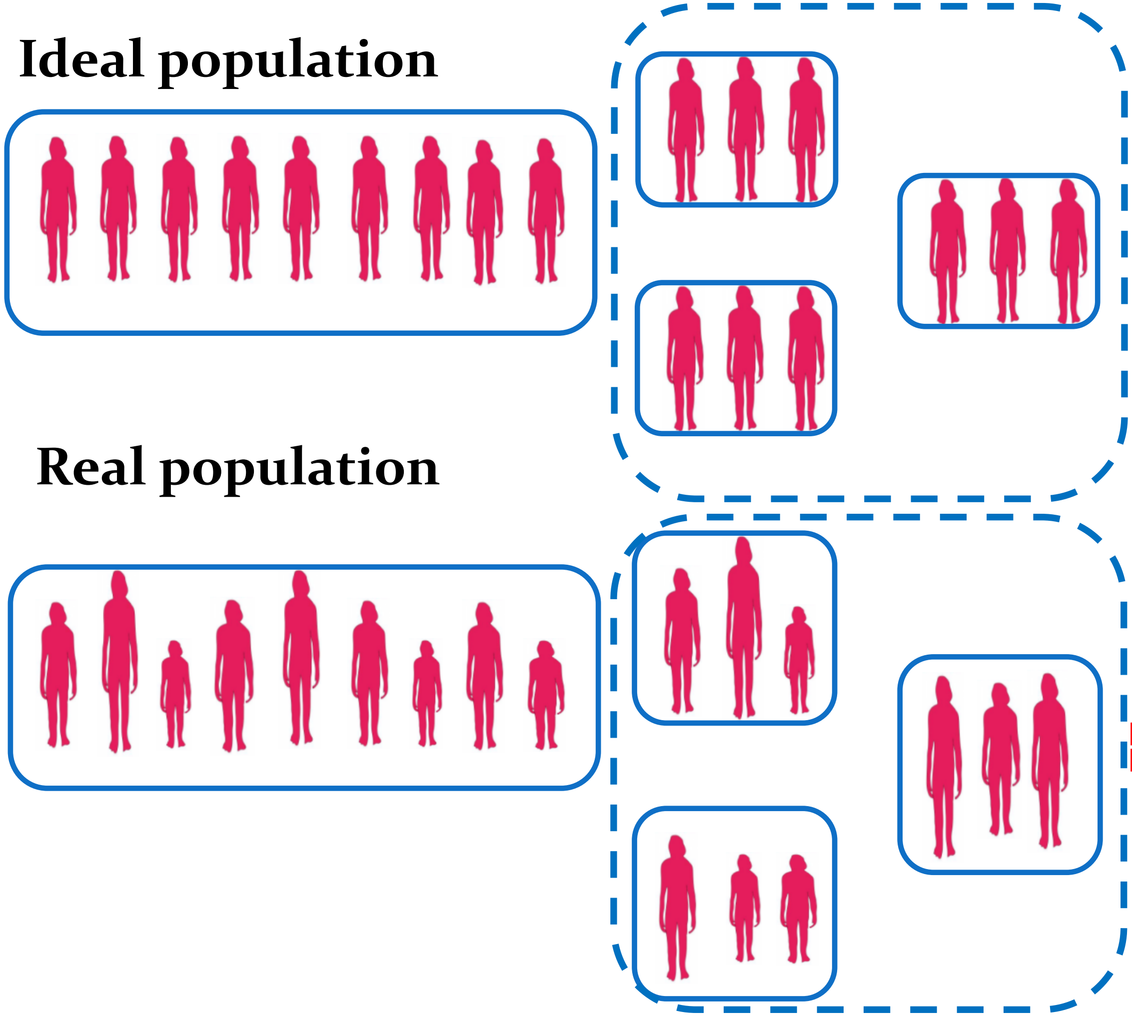
$\overline{x} \neq \mu$

Distribution of sample means

Characteristics of the distribution of sample means:

- ✓ The sample mean dose not necessarily equals the population mean μ ;
- ✓ There are differences among sample means;
- ✓ The sample means are in the normal distribution;
- ✓ The mean of sample means is equal to μ;
- ✓ The deviation of sample mean is <u>smaller</u> than that of the original variable .

sampling error can't be deleted



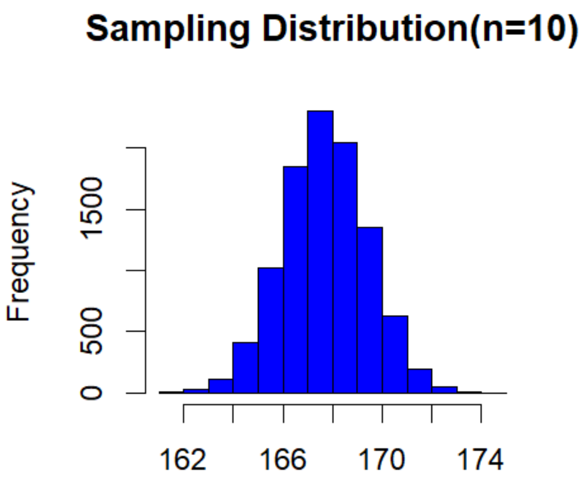
①random selection

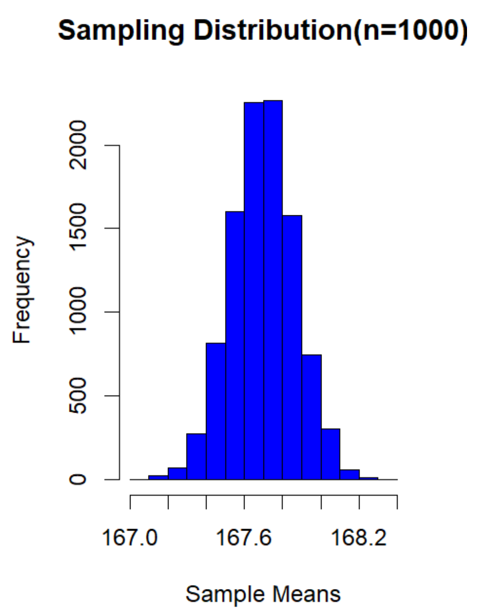
+

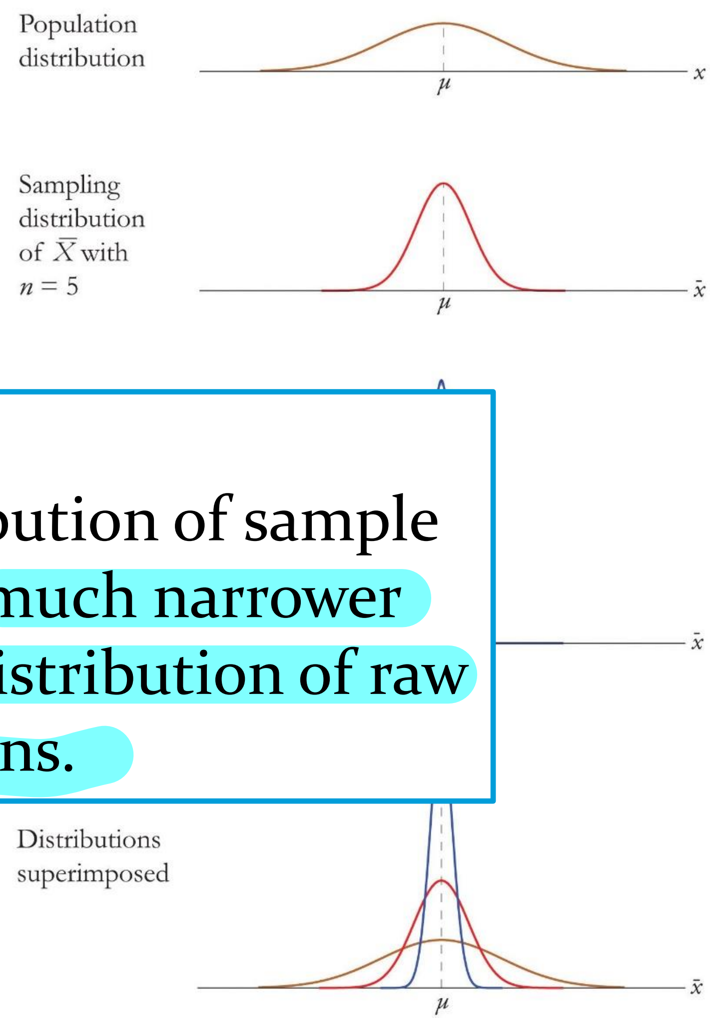
②variation

= Sampling error

The rule of sampling error







$E(\overline{X})=\mu$;
The distribution of sample means, is much narrower than the distribution of raw observations.

Prove

	Mean	Variance
Adding: $T = X + Y$	$\mu_T = \mu_X + \mu_Y$	$\sigma_T^2 = \sigma_X^2 + \sigma_Y^2$
Subtracting: $D = X - Y$	$\mu_D = \mu_X - \mu_Y$	$\sigma_D^2 = \sigma_X^2 + \sigma_Y^2$

$$\mu_{\bar{X}} = \mu_{\frac{(x1+x2+\cdots xn)}{n}}$$

$$= \frac{1}{n}(\mu_{x1} + \mu_{x2} + \cdots .. + \mu_{xn}) = \frac{1}{n}(\mu+\mu+.....+\mu) = \mu$$

Prove

if $X \sim N\left(\mu_1, \sigma_1^2\right)\ Y \sim N\left(\mu_2, \sigma_2^2\right)$

$$\sigma^2_{\bar{x}} = \sigma^2_{\frac{(x1+x2+...+xn)}{n}}$$

$$= \left(\frac{1}{n}\right)^2 \left(\sigma^2_{x1} + \sigma^2_{x2} + ... + \sigma^2_{xn}\right)$$

$$= \left(\frac{1}{n}\right)^2 \left(\sigma^2 + \sigma^2 + ..\right)$$

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

Standard Error of mean（SEM）

For a random variable x (μ,σ^2), the mean of sample mean is still μ, and the standard error of sample mean is

$$\sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}}$$

For a sample to estimate

$$s_{\overline{x}} = \frac{s}{\sqrt{n}}$$

Question:

What are the differences between standard deviation and standard error of mean?

Differences between standard deviation and standard error of mean

- Standard error of mean reflects the variation of sample means and indicates the sampling error; standard deviation reflects the variation of individuals.
- The sign for standard error of mean is: $\sigma_{\bar{X}}$, $S_{\bar{X}}$ and the sign for standard deviation is σ, S.
- Standard error of mean can be calculated by $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$, $S_{\bar{X}} = \frac{S}{\sqrt{n}}$
- Standard error of mean can be decreased by increasing the sample size, but standard deviation can not be controlled.

Definition of Sampling Error

Due to the individual variations and random sampling, it results in:

- **differences between sample statistics and population parameter;**
- **differences between sample statistic and sample statistic.**

Definition of Standard Error（SE）

It is the standard deviation of sample statistics.

- ✓ **Reflect the dispersion of sample statistics**
- ✓ **Reflect the magnitude of sampling error**

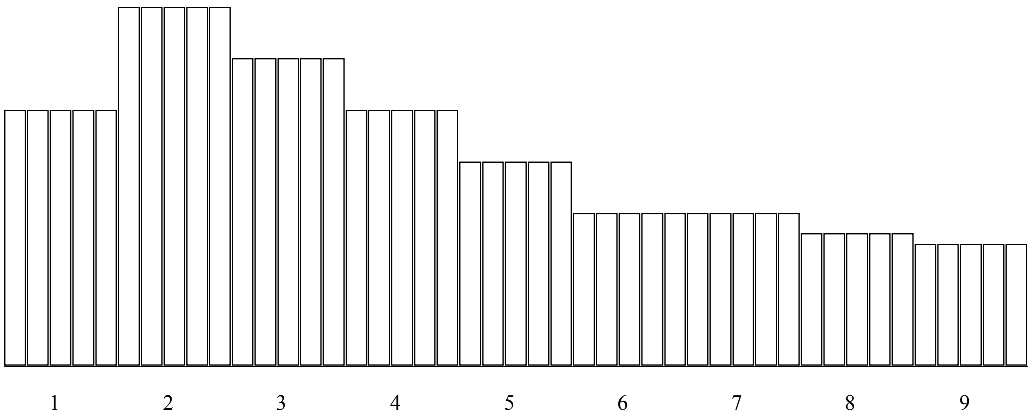
Definition of Standard Error of Mean（SEM）

It is the standard deviation of the sampling distribution of sample means.

- ✓**Reflect the dispersion of sample means**
- ✓**Reflect the magnitude of sampling error**

$$\sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}} \qquad s_{\overline{x}} = \frac{s}{\sqrt{n}}$$

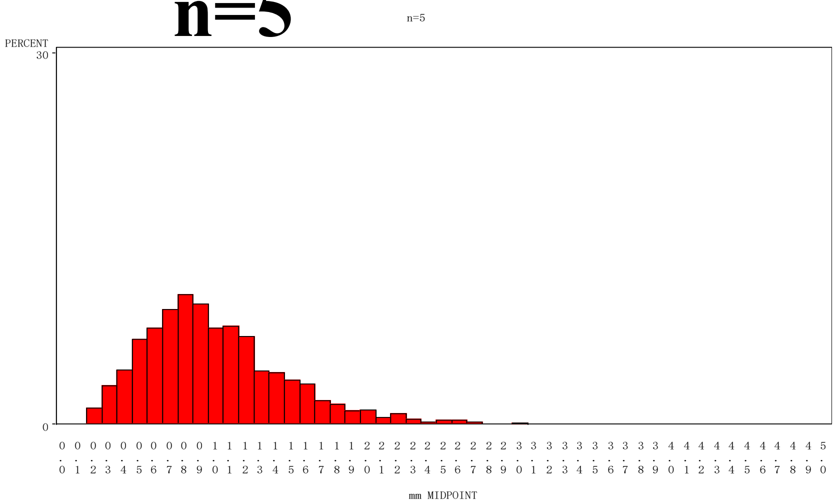
The population is not the normal distribution？



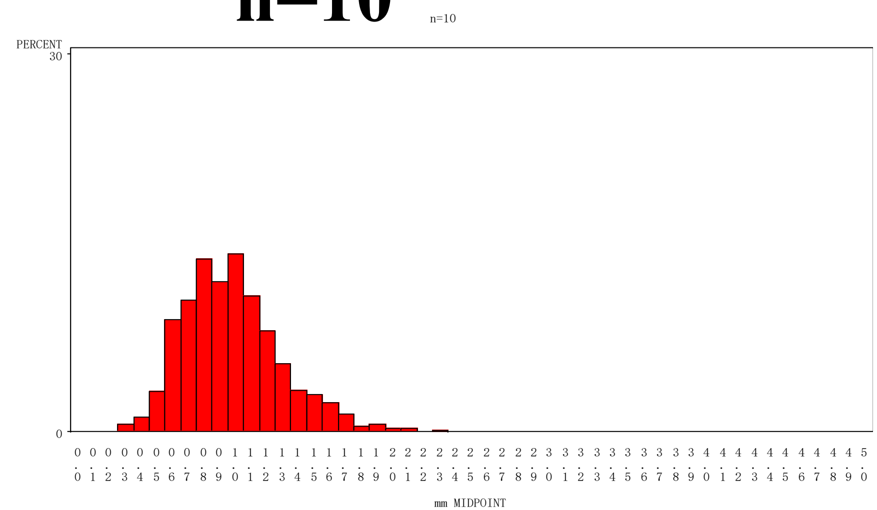
(a)

Computer simulation results

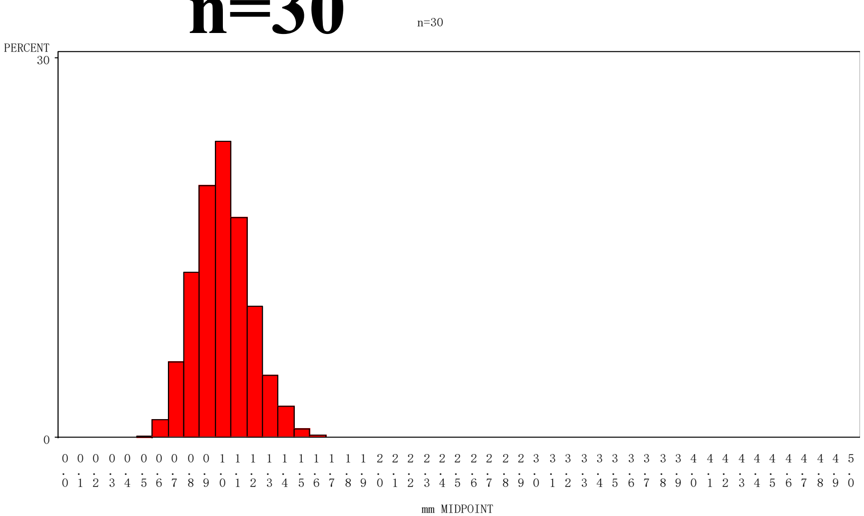
n=5



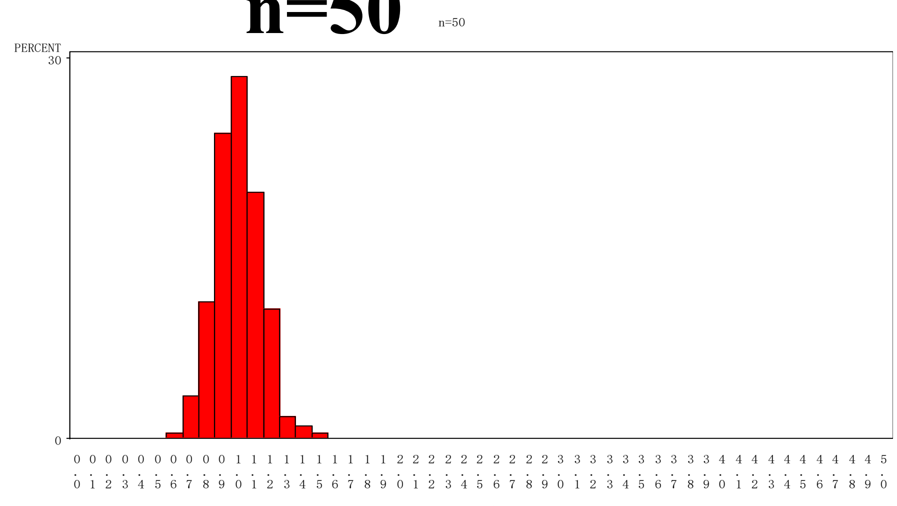
n=10



n=30



n=50



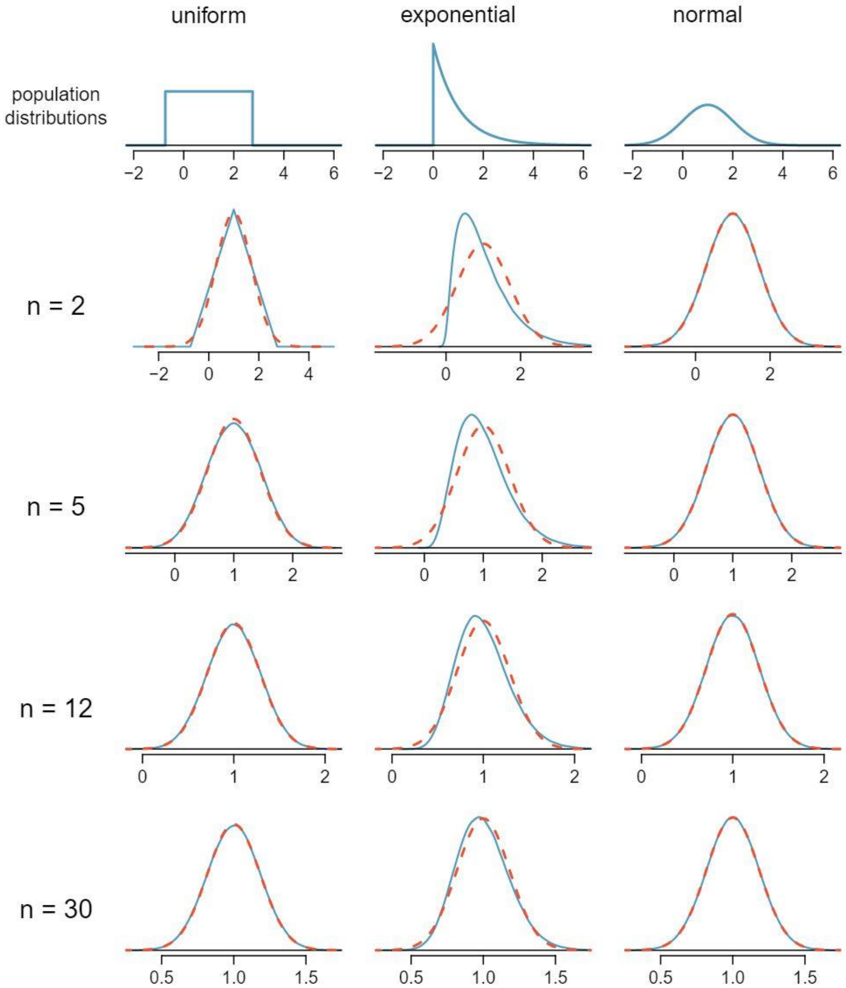
Sampling distribution of sample means

1. $\mu_{\bar{X}} = \mu_X$

2. $\sigma_{\bar{X}} = \dfrac{\sigma}{\sqrt{n}}$

3. *If X is* normal, $\bar{X}$ is normal. If X is nonnormal, $\bar{X}$ is approximately normally distributed for sufficiently large sample size.

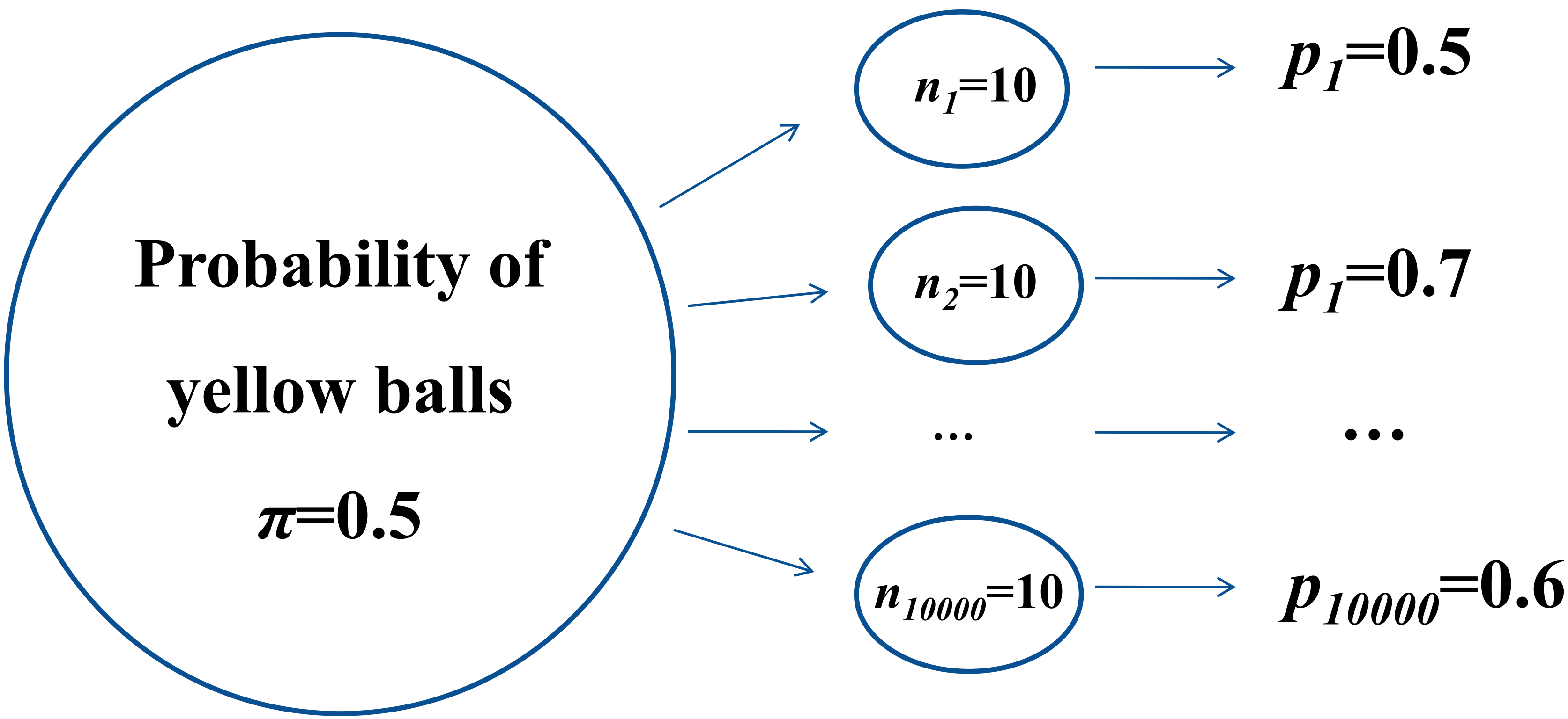
Central Limit Theorem



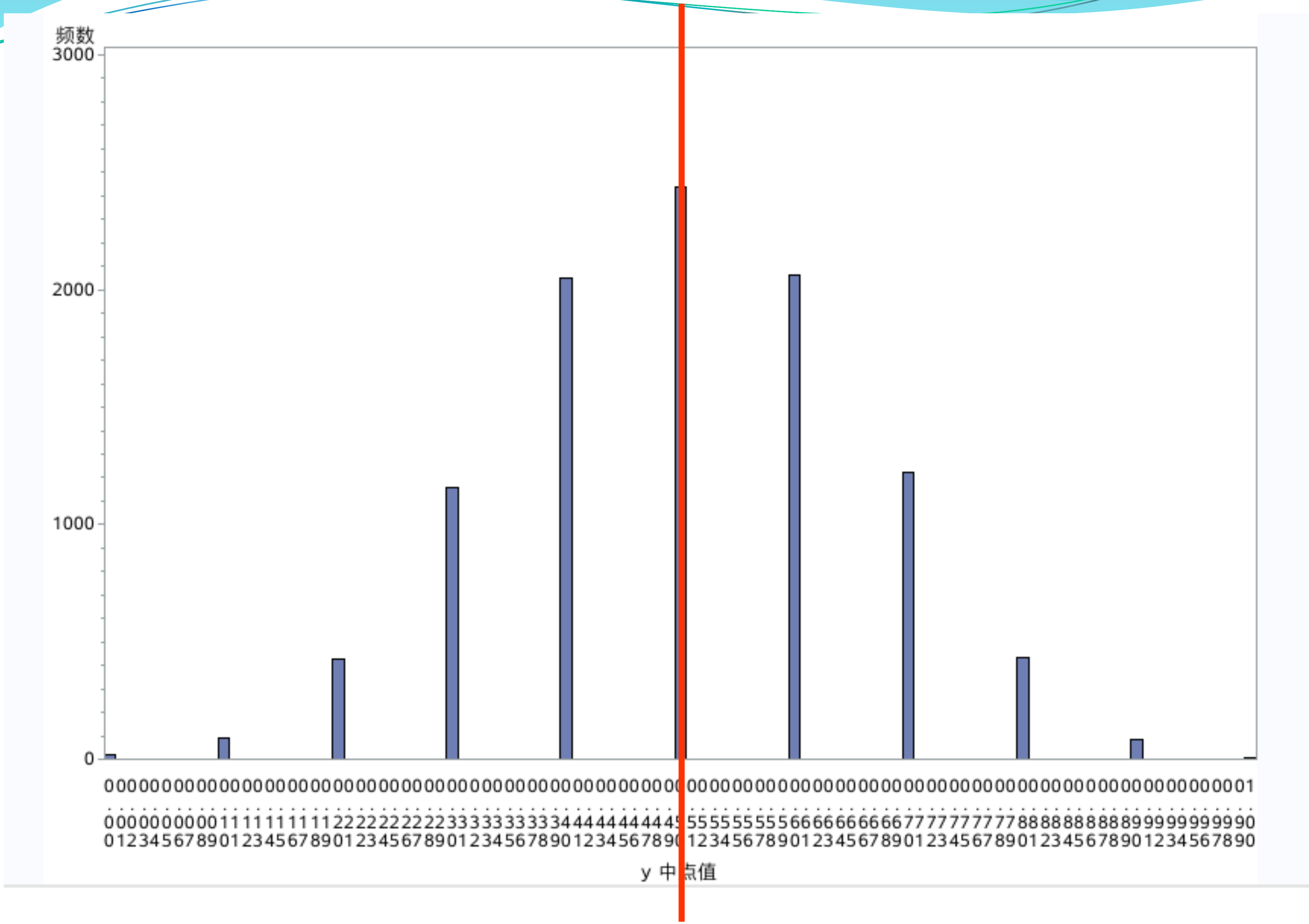
n is large enough
n>30

2. Sampling distribution of sample rates

Example 2



Probability of yellow balls
π=0.5
n_1=10
n_2=10
…
n_{10000}=10
p_1=0.5
p_1=0.7
…
p_{10000}=0.6


频数
3000
2000
1000
0
y 中点值

Sampling distribution of sample rates

- The mean of sample rates is equal to the population rate:

$$\mu_p = \pi$$

- The population standard deviation of sample rates (namely SER) is:

$$\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}}$$

- In practice, the population rate is unknown, so the sample rate is often used instead:

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

Definition of Standard Error of Rate (SER)

It is the standard deviation of the sampling distribution of sample rates.

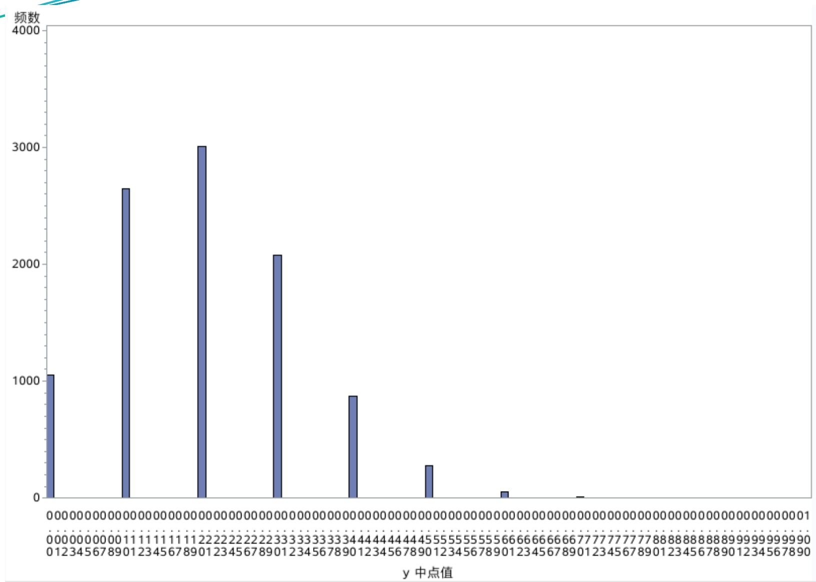
- ✓ **Reflect the dispersion of sample rates**
- ✓ **Reflect the sampling error of sample rates**

The population rate is not equal to 0.5 ($\pi \neq 0.5$)?

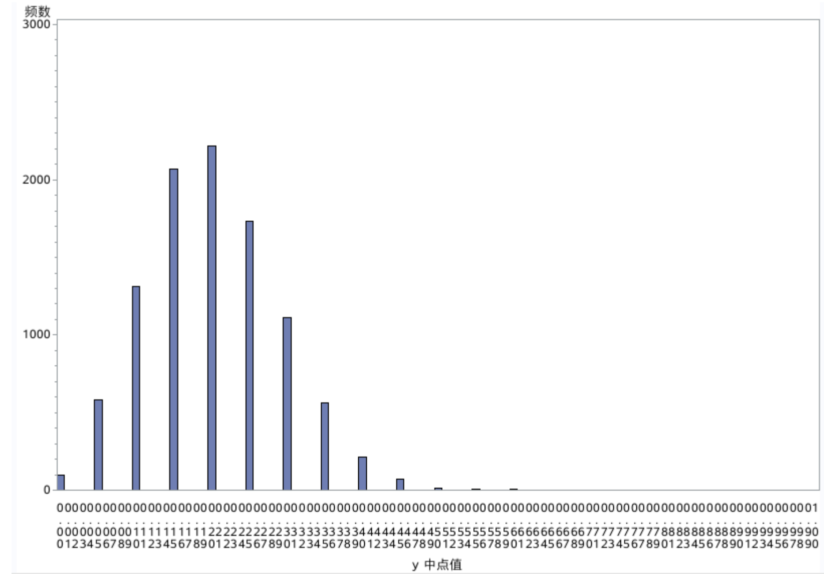
$\pi = 0.8$ *or* $\pi = 0.2$

What is the sampling distribution of sample rates in these cases?

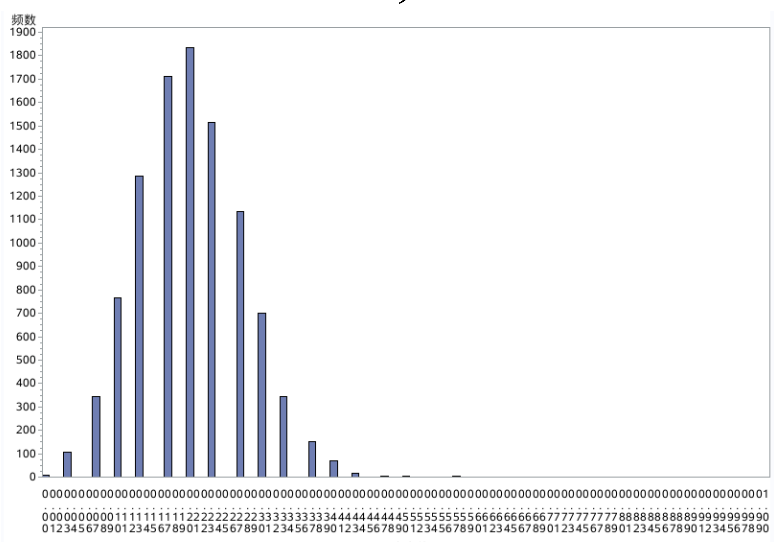
π=0.2, n=10



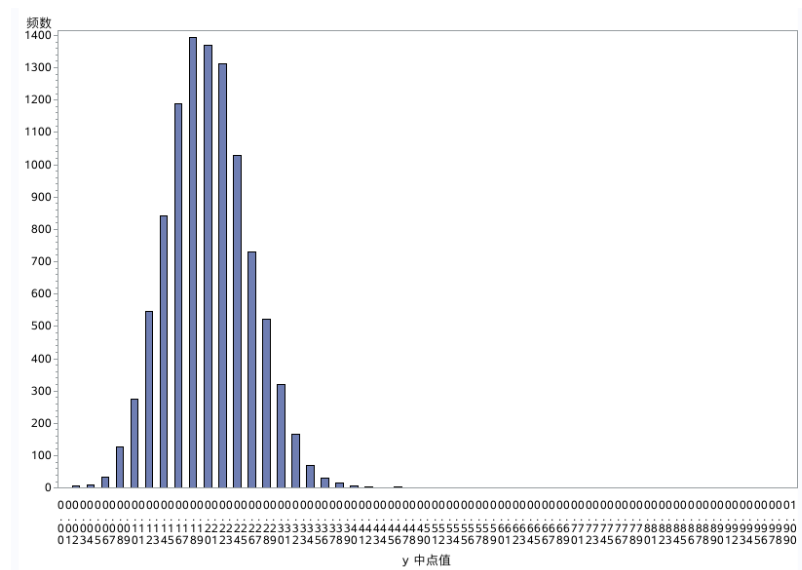
π=0.2, n=20



π=0.2, n=30



π=0.2, n=50



Sampling distribution of sample rates

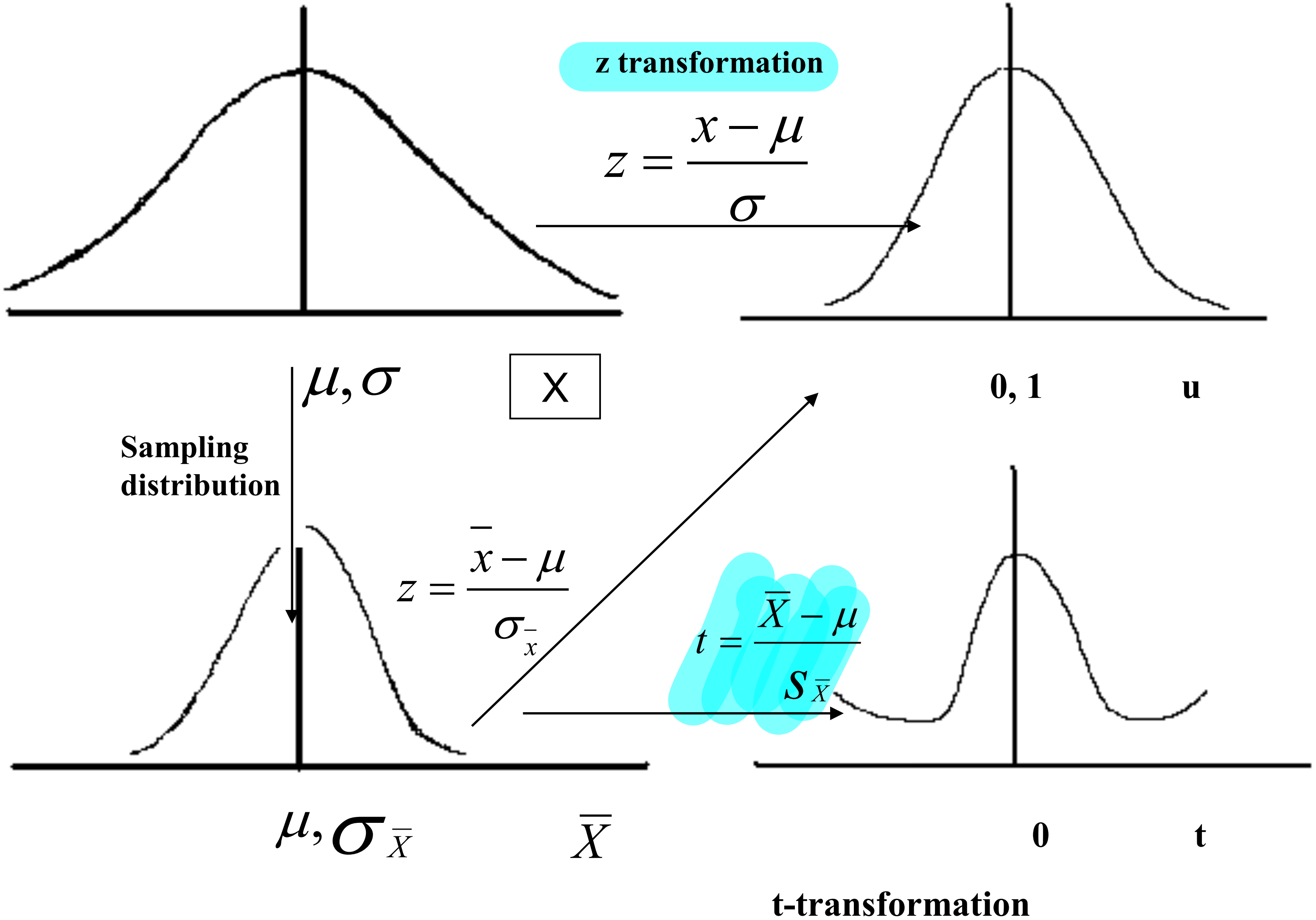
1. $\mu_p = \pi$

2. $\sigma_p = \sqrt{\dfrac{\pi(1-\pi)}{n}}$

3. *If* π *is* **0.5,** p **is normal. If** π **is not 0.5,**
p is approximately normally distributed
when $n\pi \geq 5,\ n(1-\pi) \geq 5.$

3. *t* distribution

t-transformation



Who introduced *t* distribution?

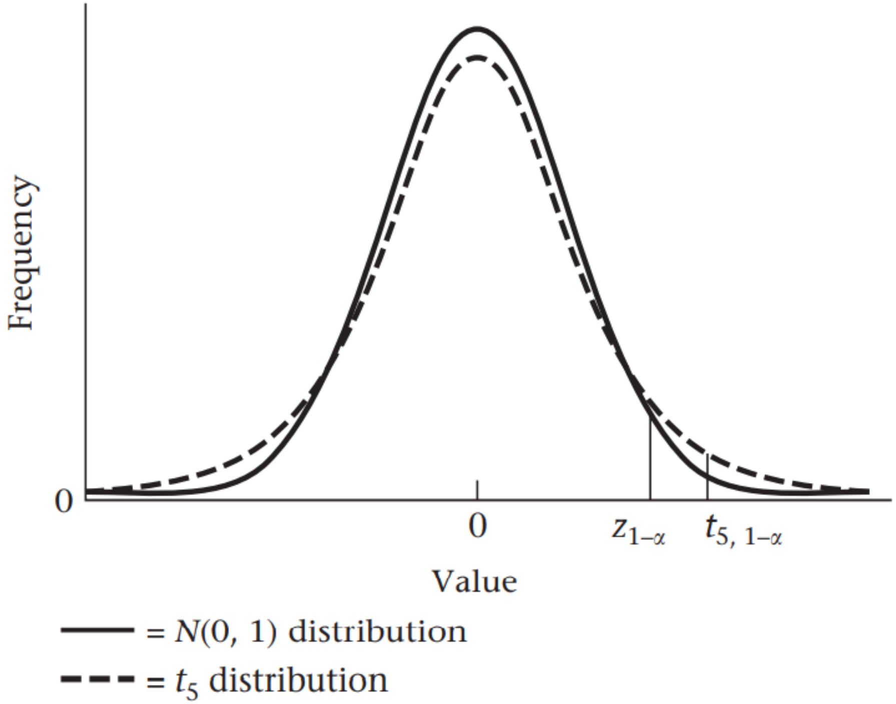
beer brewery in Gunnies

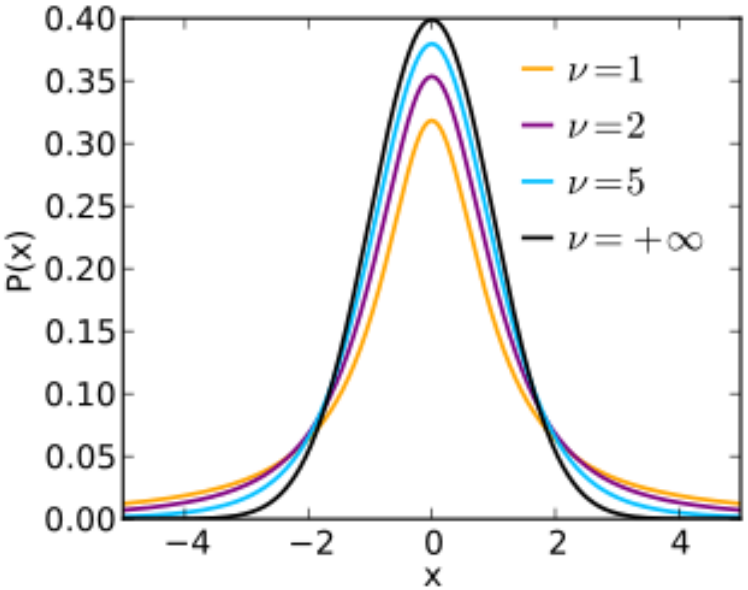
t **distribution, was discovered by *W. S. Gossett [1876-1937].***

$$f(t)=\frac{\Gamma\left[(\nu+1)/2\right]}{\sqrt{\pi\nu}\,\Gamma(\nu/2)}\left(1+\frac{t^2}{\nu}\right)^{-\frac{(\nu+1)}{2}}$$

Student's t-distribution

Student *t* distribution





Comparison of the 97.5th percentile of the *t* distribution and the normal distribution

d	$t_{d,.975}$	$z_{.975}$	d	$t_{d,.975}$	$z_{.975}$
4	2.776	1.960	60	2.000	1.960
9	2.262	1.960	∞	1.960	1.960
29	2.045	1.960			

Characteristics of *t* distribution curve

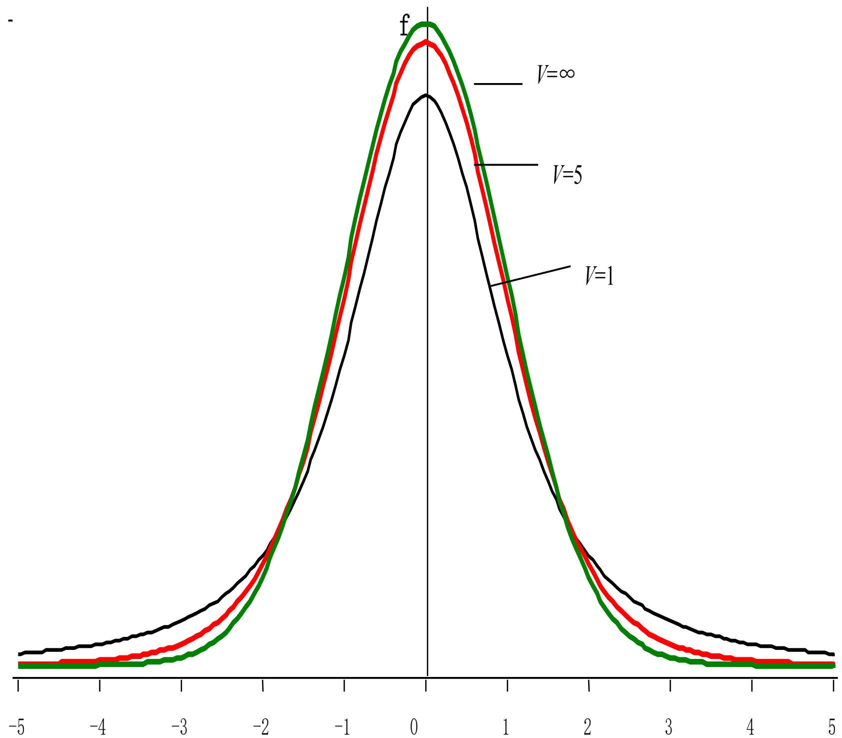
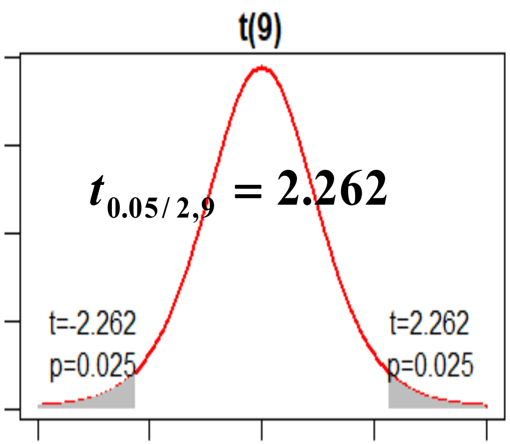


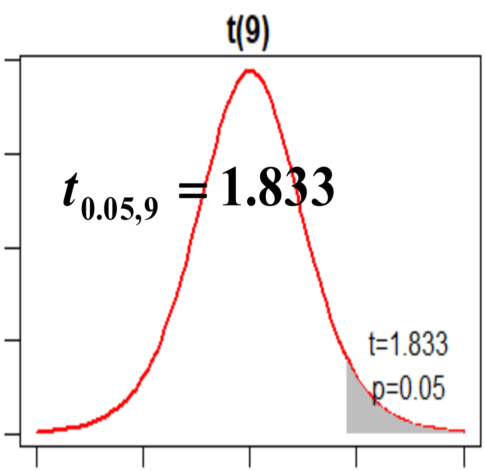
Figure 4-3 t-distribution graphs in different df

- It is symmetrical by y axis and has one apex;
- Only one parameter, ν(degree of freedom, ν=n-1) determines the shape of t distribution.
- The total area under the t distribution equals to 1.
- When ν approaches ∞, t distribution approaches standard normal distribution

t statistic table

df	*P*				
v	1-tail:	0.05	0.025	0.005	0.0005
	2-tail:	0.1	0.05	0.01	0.001
1		6.314	12.706	63.657	636.619
2		2.920	4.303	9.925	31.599
3		2.353	3.182	5.841	12.924
……					
9		1.833	2.262	3.250	4.781
……					
∞		1.645	1.960	2.576	3.291





Summary

✓ **Sampling distribution of sample means**

1. $\mu_{\bar{X}} = \mu_X$

2. $\sigma_{\bar{X}} = \dfrac{\sigma}{\sqrt{n}}$

3. **If X is normal, $\bar{X}$ is normal. If X is nonnormal, $\bar{X}$ is approximately normally distributed for sufficiently large sample size.**

Summary

✓ ***t* distribution**

$$Z = \frac{\overline{X} - \mu}{\sigma_{\overline{X}}}$$

$$\downarrow$$

$$t = \frac{\overline{X} - \mu}{S_{\overline{X}}}$$

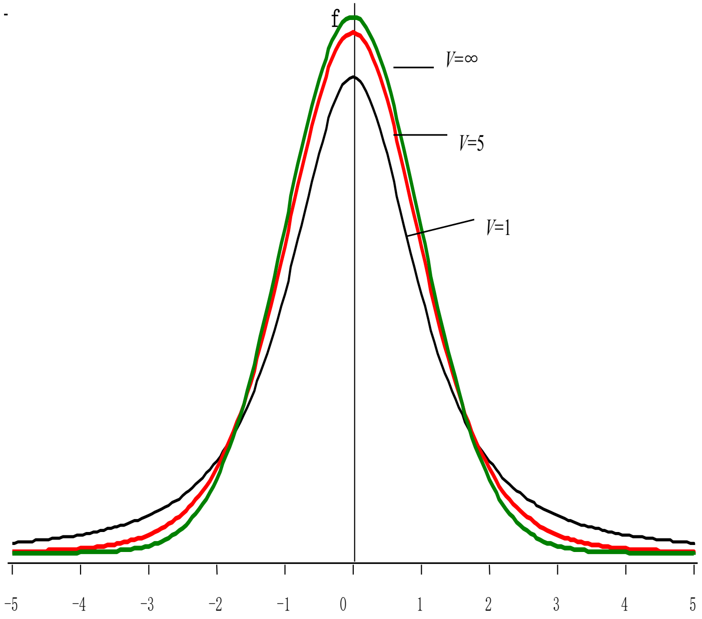
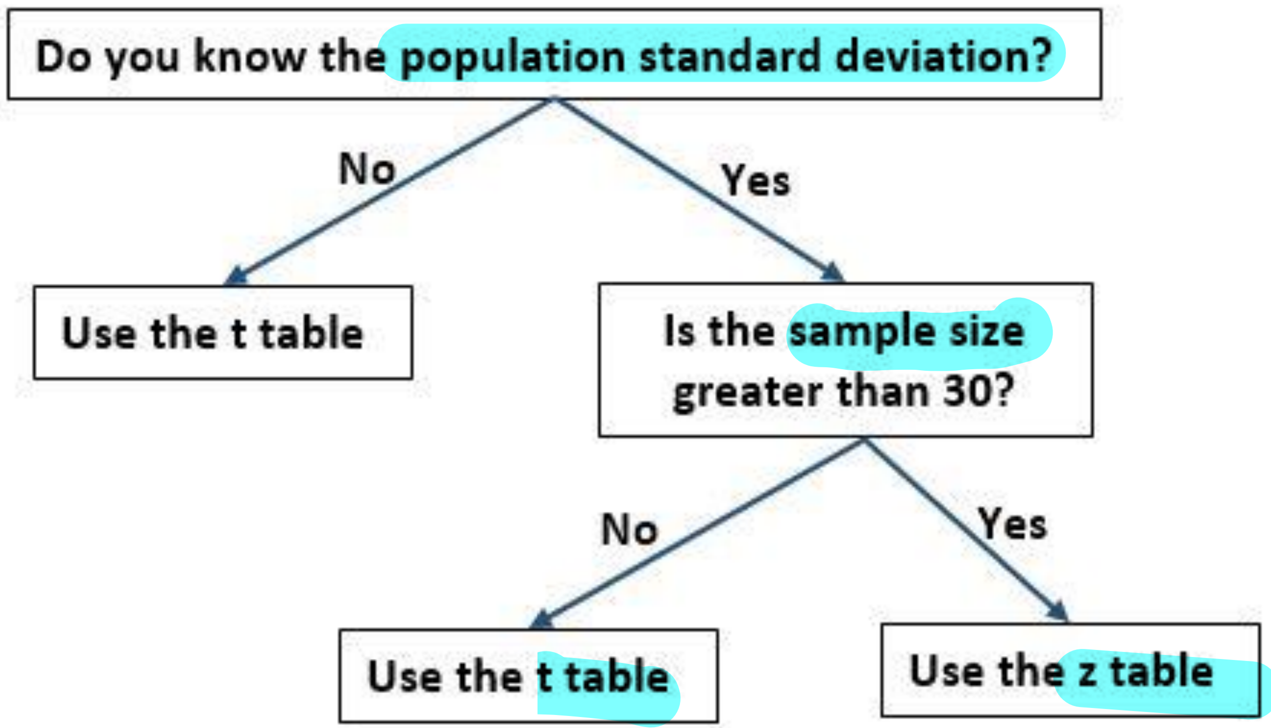


Figure 4-3 t-distribution graphs in different df



Do you know the population standard deviation?
No
Yes
Use the t table
Is the sample size greater than 30?
No
Yes
Use the t table
Use the z table

Thank you!