Introduction to R　Part3

Satoru UCHIDA (Visiting Student Researcher, UC Berkeley)

http://realize.jounin.jp/R.html

2010/3/10

1. Basic concepts of statistics

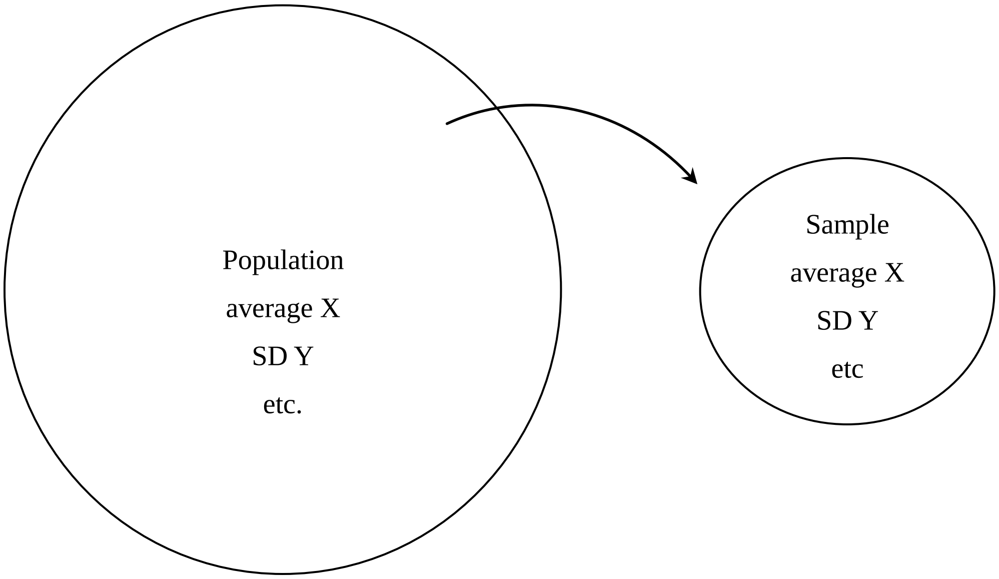
Descriptive statistics

average, sd, median etc.

→Summary of data

Inferential statistics

The aim is to infer the population based on the limited number of samples.



・**point estimation**

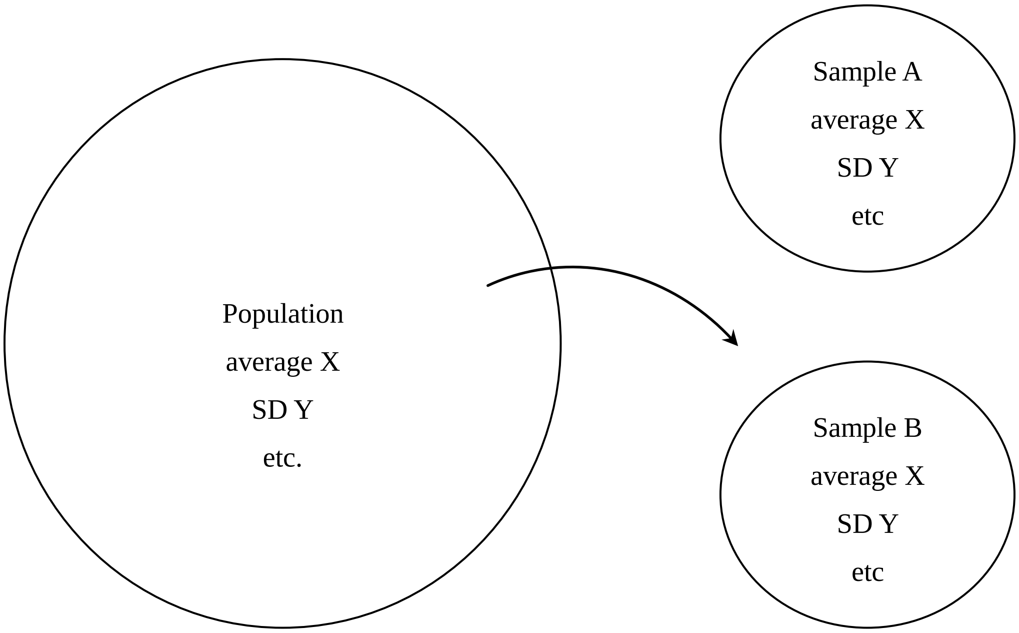
Predict a value (e.g. average) of the population from a sample

Average of a sample =22 ➔Average of the population is 22

・**interval estimation**

Judging from the average of a sample, the average of the population can fall between 20-24.

- **hypothesis testing**



Are there any differences between samples?

If any, how likely is it?

WHAT WE NEED FOR LINGUICTIC ANALYSIS

➔Hypothesis testing

2. Hypothesis testing

First things to remember

Two types of hypothesis

H_0 Null hypothesis (There is no difference between A and B)

H_1 Alternative hypothesis (There is a difference between A and B)

Think Why do we need H_0?

P-value

➔The probability that H_0 holds

Significance level

➔The probability level to reject H_0

P < .05

P<.01

3. How to analyze data

Test for average (t.test (Welch test))

Question: we collected the following data:

coffeeA	score	coffeeB	score
1	70	1	85
2	75	2	80
3	70	3	95
4	85	4	70
5	90	5	80
6	70	6	75
7	80	7	80
8	75	8	90
average	76.88	average	81.88

The average of coffee A is 5 less than that of coffee B. Is this difference significant statistically?

Step1 set up H_0 and H_1

H_0: There is no difference between coffee A and coffee B.

H_1: There is a difference between coffee A and coffee B.

Step2 determine the significance level

This time, let's set $p < .05$

Step3 Conduct an analysis

We use **t.test** for comparing two averages.

```
> coffeeA=c(70,75,70,85,90,70,80,75)
> coffeeB=c(85,80,95,70,80,75,80,90)
> t.test(coffeeA,coffeeB)

            Welch Two Sample t-test

data:    coffeeA and coffeeB
t = -1.2881, df = 13.951, p-value = 0.2187
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  -13.327988      3.327988
sample estimates:
mean of x mean of y
     76.875       81.875
```

Step4 Interpret the result

The p-value (0.22) is larger than 0.05.

There can be 22 cases out of 100 where H_0 holds. (22%)

→We cannot abandon the null hypothesis.

→There is no difference between coffee A and coffee B.

Paired t-test

Suppose that the same person evaluate both coffee and we got the following results:

person	coffeeA	coffeeB	diff
1	90	95	−5
2	75	80	−5
3	75	80	−5
4	75	80	−5
5	80	75	5
6	65	75	−10
7	75	80	−5
8	80	85	−5
average	76.88	81.25	−4.37

Question: Is there any difference between coffee A and coffee B?

Step1 set up H_0 and H_1

H_0: There is no difference between coffee A and coffee B.

H_1: There is a difference between coffee A and coffee B.

Step2 determine the significance level

This time, let’s set p < .05

Step3 Conduct an analysis

```
> coffeeA2=c(90,75,75,75,80,65,75,80)
> coffeeB2=c(95,80,80,80,75,75,80,85)
> t.test(coffeeA2,coffeeB2)

        Welch Two Sample t-test

data:   coffeeA2 and coffeeB2
t = -1.2999, df = 13.878, p-value = 0.2148
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -11.599706    2.849706
sample estimates:
mean of x mean of y
   76.875     81.250
```

```
> t.test(coffeeA2,coffeeB2, paired=TRUE)

        Paired t-test

data:   coffeeA2 and coffeeB2
t = -2.9656, df = 7, p-value = 0.02094
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -7.8633933 -0.8866067
sample estimates:
mean of the differences
                 -4.375
```

TRY Compare these results and notice the difference.

Test for tables (chisq.test)

Question: We got the data as follows:

pet/sex	Male	Female
Dog	16	4
Cat	12	8

Is there any relation between pet and sex?

Step1 set up H_0 and H_1

H_0: There is no relationship between pet and sex.

H_1: Pet and sex is closely related to each other.

Step2 determine the significance level

This time, let's set $p < .05$

Step3 Conduct an analysis

We use **chisq.test** for table

```
#copy the upper data in "table" sheet.
> data=read.delim("clipboard",header=TRUE,row.names=1)
> chisq.test(data, correct=FALSE)
```

```
        Pearson's Chi-squared test

data:   data
X-squared = 1.9048, df = 1, p-value = 0.1675
```

→correct=FALSE (This invalidate the Yate's the correction, which should be applied to 2 * 2 matrix where the data is small (when the expectation of any one of the cells is less than 4).

Step4 Interpret the result

P-value=0.1675 (larger than 0.05)

There can be 17 cases out of 100 where H_0 holds. (17%)

→We cannot abandon the null hypothesis.

→There is no relationship between pet and sex.

TRY Test the other table on the "table" sheet. What do you find?