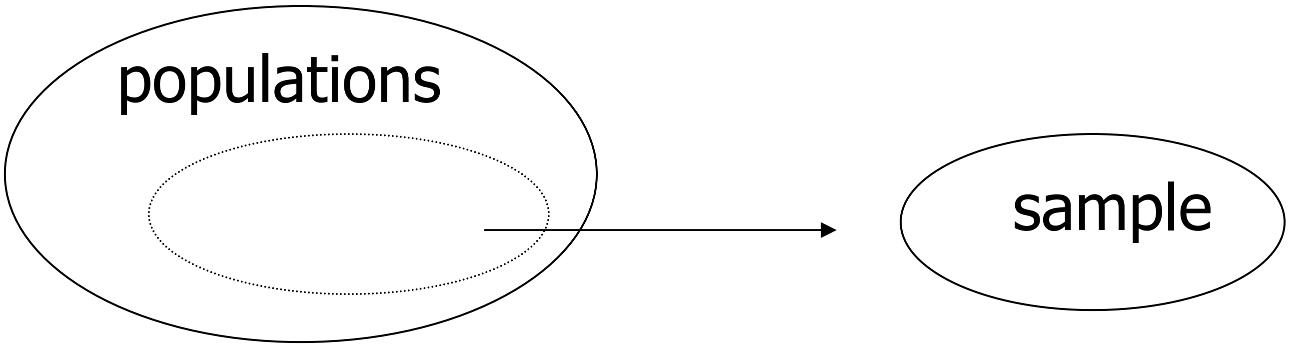
Statistical Methods and Data Analysis

- What is statistics?
  - The science of learning from data.
  - Plays an important role in almost all areas of science, business, industry

- Why study statistics?
  - Interpret the results of sampling (survey or experiment)
  - Evaluate published numerical facts
  - Forecast sales and profit in business

Some definitions

- Populations: the set of all measurements
- Samples: set of measurements selected from the population



Median

- The middle value when the measurements are arranged from lowest to highest.
- 95 86 78 90 62 73 89 92 84 76
- 62 73 76 78 84 86 89 90 92 95
- Median = (84 + 86) / 2 = 85

Mean

- Mean: sum of measurements divided by the total number. (Average, the balancing point of the data set).

$$\bar{y} = \frac{\sum y_i}{n} = \frac{y_1 + y_2 + \ldots y_n}{n}$$

- μ - population mean; $\bar{y}$ - sample mean

Variance

- Variance: the sum of the squared deviations divided by n-1

$$S^2 = \frac{\sum (y - \bar{y})^2}{n - 1}$$

- Measure of variability,
- S^2 represents the sample variance,
- σ^2 (sigma) represents the population variance

Standard Deviation

- Standard deviation (SD): positive square root of the variance

$$S = \sqrt{\frac{\sum (y - \bar{y})^2}{n-1}}$$

- The same units as the original data,
- *S* is sample SD; σ is population SD

Sensory tests:

- 1. Discriminative tests
  - Determine whether a difference exists between samples.
  - Triangle test, Duo-trio test, two-out-of-five test…
- 2. Descriptive tests
  - Determine the nature and intensity of the differences.
  - Scaling methods; descriptive analysis methods.
- 3. Affective tests
  - Based on either a measure of preference (or acceptance) or a measure from which we can determine relative preference.
  - Paired comparison preference test; hedonic test; ranking test.

Discriminative tests

- Triangle test
  - Whether or not a detectable difference exists between two samples.
- Paired comparison test
  - To compare the intensity of some particular characteristics.

Descriptive tests

- Scaling methods:
  - Nine-point scale, seven-point scale.
- A statistical test: T-test, F-test,
  - 1. Research hypothesis, $H_{a,,}$ alternative hypothesis
  - 2. Null hypothesis, H_o
  - 3. Test statistics, T.S.
  - 4. Rejection region, R.R.
  - 5. Check assumption and draw conclusion

Research Hypothesis

- H_0: $\mu > 4$ (null)
- H_a: $\mu \leq 4$ (alternative)

- H_0: $\mu_1 = \mu_{2,}$ there is no difference (null)
- H_a: $\mu_1 \ddagger \mu_2$, there is a significant difference (alternative)

- Type I error: when the null hypothesis is rejected when it is true.
- Type II error: when the null hypothesis is accepted when in fact it is false.

The level of significance (α)

- α: the probability of making a type I error.
  - 0.05 (5%): 1/20 of saying there is a difference when there is no difference
  - 0.01 (1%): 1/100 …..
- β: the probability of making a type II error.

Summary of a statistical test

- One-tailed test: directional difference test.
  - Case 1. H_0: $\mu \le \mu_o$ vs. H_a: $\mu > \mu_o$ (right-tailed test)
  - Case 2. H_0: $\mu \ge \mu_o$ vs. H_a: $\mu < \mu_o$ (left-tailed test)
- Two-tailed test: no expectation about the results.
  - Case 3. H_0: $\mu = \mu_o$ vs. H_a: $\mu \ne \mu_o$ (two-tailed test)

- T.S.:
- *P* : level of significance or *p*-value
- If the *p*-value $\le \alpha$, then reject H_0
- If the *p*-value $> \alpha$, then accept H_0

T-test

- Student’s t distribution: determine whether two samples have the same mean.

$$t = \frac{\bar{y} - u_0}{s / \sqrt{n}}$$

A Paired T-test

- H_0: $\mu_1 = \mu_2$, no difference,
- H_a: μ_1 ‡ μ_2, significant difference
  - If the $t > t_{\alpha}$, then reject H_0
  - If the $t \leq t_{\alpha}$, then accept H_0

- Assumes that the intervals between categories are equal.
- df: degree of freedom, df = n -1

Analysis using t-test

- 1. Calculate average difference:

$$\bar{d} = \bar{A} - \bar{B}$$

- 2.

$$S = \sqrt{\frac{\left\{\sum d^2 - [(\sum d^2)/n\right\}}{n-1}}$$

- 3

$$t = \frac{\bar{d}}{S / \sqrt{n}}$$

Step 1

$$\bar{d} = \bar{A} - \bar{B}$$

$= 4.7 - 3.1 = 1.6$

Step 2

- $\sum d = 32$
- $\sum d^2 = 114$
- $(\sum d)^2 = (32)^2 = 1024$

Step 3

$$S = \sqrt{\frac{\left\{\sum d^2 - [(\sum d^2) / n\right\}}{n-1}}$$

- S = ((114-1024/20)/19)^0.5=1.82

Step 4

- df = n-1 = 20 –1 = 19
- t_α= 2.093
- t = 3.93

T-test

- 1. Specify the null and alternative hypothesis.
- 2. Specify a value for α
- 3. Determine the weight of evidence (t, *p*)

Conclusion

- $t=3.93 > t_{\alpha}=2.093$
- Conclusion: Cheese A was significantly more bitter than cheese B ($P \leq 0.05$)

F-test

- Test of whether two samples are drawn from different populations have the same SD or variance, with specified confidence level. Samples may be of different size.
- The variability of a population is as important as the mean
- As little variation as possible

F-test: comparing two population variances

- H_0: $\sigma_1^2 = \sigma_2^2$
- H_a: $\sigma_1^2 \ \ddagger\ \sigma_2^2$
  - If the $F \geq F_{\alpha,df1,df2}$, then reject H_0, there is significant difference
  - If the $F < F_{\alpha,df1,df2}$, then accept H_0, there is no significant difference

  - If the *p*-value $\leq \alpha$, then reject H_0
  - If the *p*-value $> \alpha$, then accept H_0

ANOVA

- Analysis of variance
  - A statistical test about more than two population means

Assumptions

- Each set has a normal distribution
- The variances of all sets are equal
- All sets are independent random samples

ANOVA

- H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$
- H_a: at least one of the means differs from the rest
  - If the $p \leq \alpha$, then reject H_0, at least one of the means differs from the rest
  - If the $p > \alpha$, then accept H_0, there is no significant difference among four means

ANOVA- single factor analysis

- df: degree of freedom : df=n-1
- SS: sum of squares
- MS: the mean square
- F: the variance ratio
- STAT 412, 512

Example: hardness of four wieners

- $P = 0.00 < \alpha=0.05$, reject H_0
- Conclusion: there was a significant difference in hardness among the four brands of wieners ($p \leq 0.01$).

- ($F = 16.74 > F_{\alpha}=2.95$, reject H_0)

Problem of one-way ANOVA

- Don't know which means differ from each other
- $\mu_1, \mu_2, \mu_3, \mu_4$

ANOVA-two way without replication

- Rows : p=0.22 > α=0.05
- Columns: p=0.00 < α=0.05
- Conclusions: there is a significant difference among the protein content of four types of cereals.

ANOVA- different treatments

- Fisher’s LSD
- Tukey’s test
- Dunnett’s test: compare with a control

Fisher's LSD

- $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$
- H_a: at least one of the means differs from the rest

- If H_0 is rejected, define the least significant difference (LSD)
- $\mu_1 - \mu_2 > LSD$, different means
- $\mu_1 - \mu_2 < LSD$, same means

Minitab

- Stat\ANOVA\One-way\Comparisons\
  - Tukey’s
  - Fisher’s
  - Dunnett’s

SAS

- 6.5 9.2   12.4   4.4
- C   B   A   C