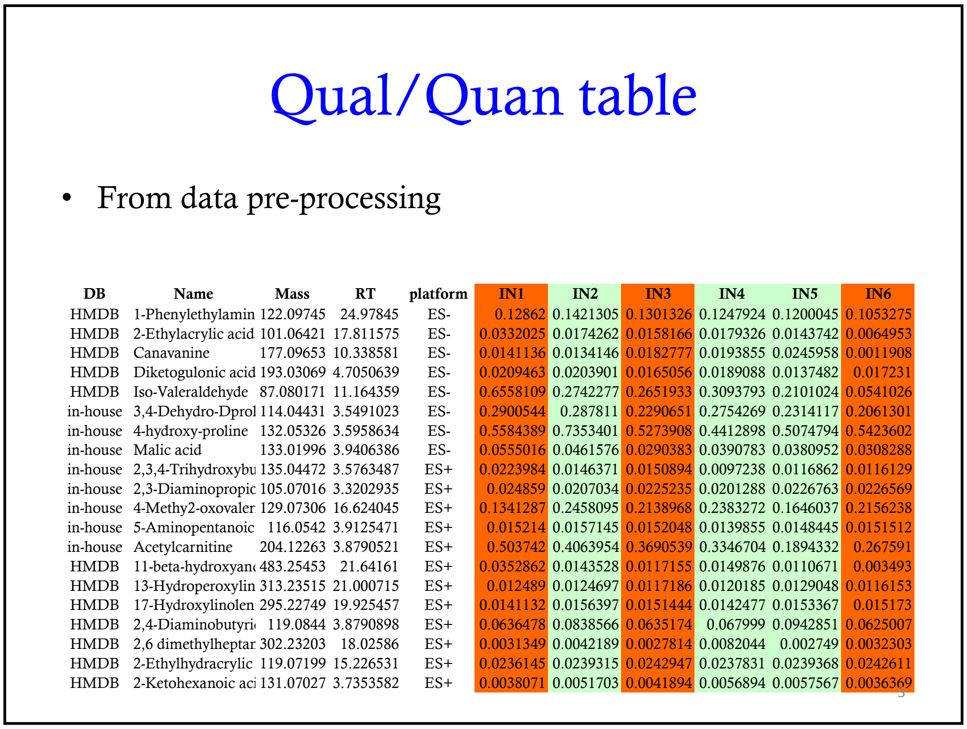
Statistical Analysis of GC- and LC-MS Metabolomics Data

Xiuxia Du
Department of Bioinformatics & Genomics
University of North Carolina at Charlotte

Outline

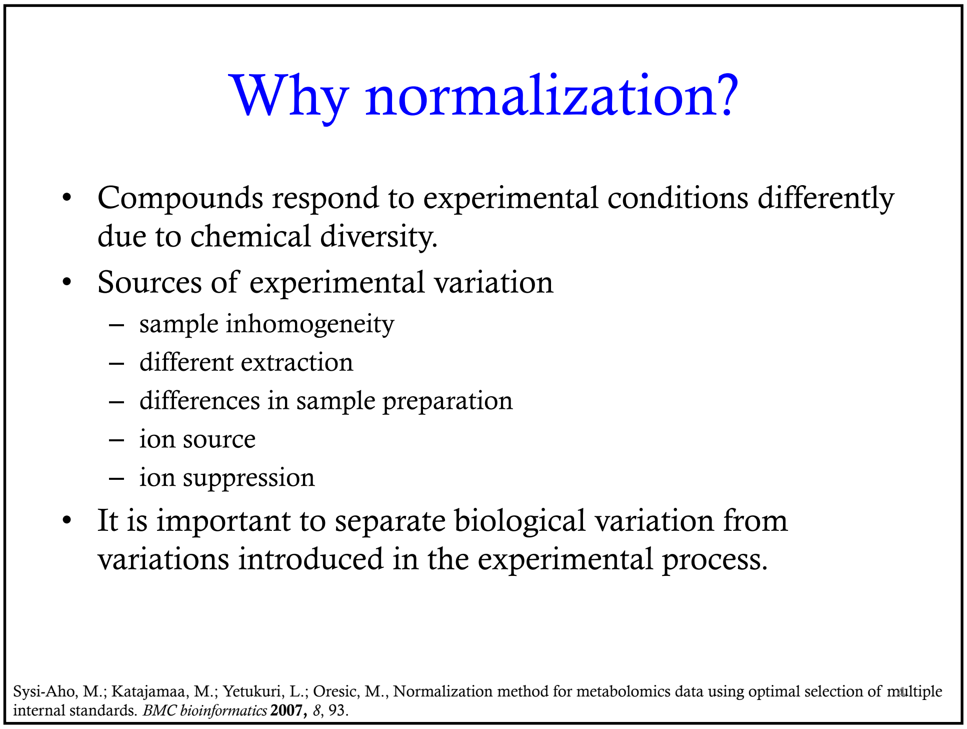
- Introduction
- Data pre-treatment
  1. Normalization
  2. Outlier analysis
  3. Centering, scaling, transformation
- Univariate analysis
  1. Student's *t* test
  2. *p*-value correction
  3. Normality test
  4. Non-parametric test
  5. ANOVA
- Multivariate analysis
  1. Correlation analysis
  2. Clustering
  3. Dimension reduction
  4. Classification
- Software packages



Qual/Quan table

- From data pre-processing

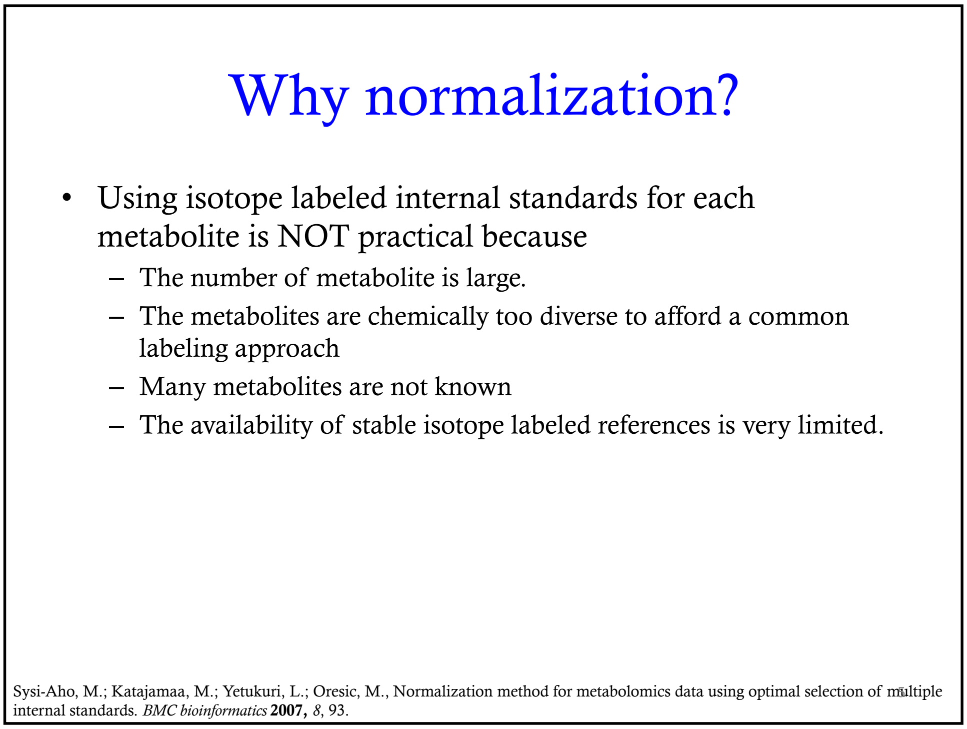
DB	Name	Mass	RT	platform	IN1	IN2	IN3	IN4	IN5	IN6
HMDB	1-Phenylethylamin	122.09745	24.97845	ES-	0.12862	0.1421305	0.1301326	0.1247924	0.1200045	0.1053275
HMDB	2-Ethylacrylic acid	101.06421	17.811575	ES-	0.0332025	0.0174262	0.0158166	0.0179326	0.0143742	0.0064953
HMDB	Canavanine	177.09653	10.338581	ES-	0.0141136	0.0134146	0.0182777	0.0193855	0.0245958	0.0011908
HMDB	Diketogulonic acid	193.03069	4.7050639	ES-	0.0209463	0.0203901	0.0165056	0.0189088	0.0137482	0.017231
HMDB	Iso-Valeraldehyde	87.080171	11.164359	ES-	0.6558109	0.2742277	0.2651933	0.3093793	0.2101024	0.0541026
in-house	3,4-Dehydro-Dprol	114.04431	3.5491023	ES-	0.2900544	0.287811	0.2290651	0.2754269	0.2314117	0.2061301
in-house	4-hydroxy-proline	132.05326	3.5958634	ES-	0.5584389	0.7353401	0.5273908	0.4412898	0.5074794	0.5423602
in-house	Malic acid	133.01996	3.9406386	ES-	0.0555016	0.0461576	0.0290383	0.0390783	0.0380952	0.0308288
in-house	2,3,4-Trihydroxybu	135.04472	3.5763487	ES+	0.0223984	0.0146371	0.0150894	0.0097238	0.0116862	0.0116129
in-house	2,3-Diaminopropic	105.07016	3.3202935	ES+	0.024859	0.0207034	0.0225235	0.0201288	0.0226763	0.0226569
in-house	4-Methy2-oxovaler	129.07306	16.624045	ES+	0.1341287	0.2458095	0.2138968	0.2383272	0.1646037	0.2156238
in-house	5-Aminopentanoic	116.0542	3.9125471	ES+	0.015214	0.0157145	0.0152048	0.0139855	0.0148445	0.0151512
in-house	Acetylcarnitine	204.12263	3.8790521	ES+	0.503742	0.4063954	0.3690539	0.3346704	0.1894332	0.267591
HMDB	11-beta-hydroxyan	483.25453	21.64161	ES+	0.0352862	0.0143528	0.0117155	0.0149876	0.0110671	0.003493
HMDB	13-Hydroperoxylin	313.23515	21.000715	ES+	0.012489	0.0124697	0.0117186	0.0120185	0.0129048	0.0116153
HMDB	17-Hydroxylinolen	295.22749	19.925457	ES+	0.0141132	0.0156397	0.0151444	0.0142477	0.0153367	0.015173
HMDB	2,4-Diaminobutyri	119.0844	3.8790898	ES+	0.0636478	0.0838566	0.0635174	0.067999	0.0942851	0.0625007
HMDB	2,6 dimethylheptar	302.23203	18.02586	ES+	0.0031349	0.0042189	0.0027814	0.0082044	0.002749	0.0032303
HMDB	2-Ethylhydracrylic	119.07199	15.226531	ES+	0.0236145	0.0239315	0.0242947	0.0237831	0.0239368	0.0242611
HMDB	2-Ketohexanoic aci	131.07027	3.7353582	ES+	0.0038071	0.0051703	0.0041894	0.0056894	0.0057567	0.0036369



Why normalization?

- Compounds respond to experimental conditions differently due to chemical diversity.
- Sources of experimental variation
  - sample inhomogeneity
  - different extraction
  - differences in sample preparation
  - ion source
  - ion suppression
- It is important to separate biological variation from variations introduced in the experimental process.

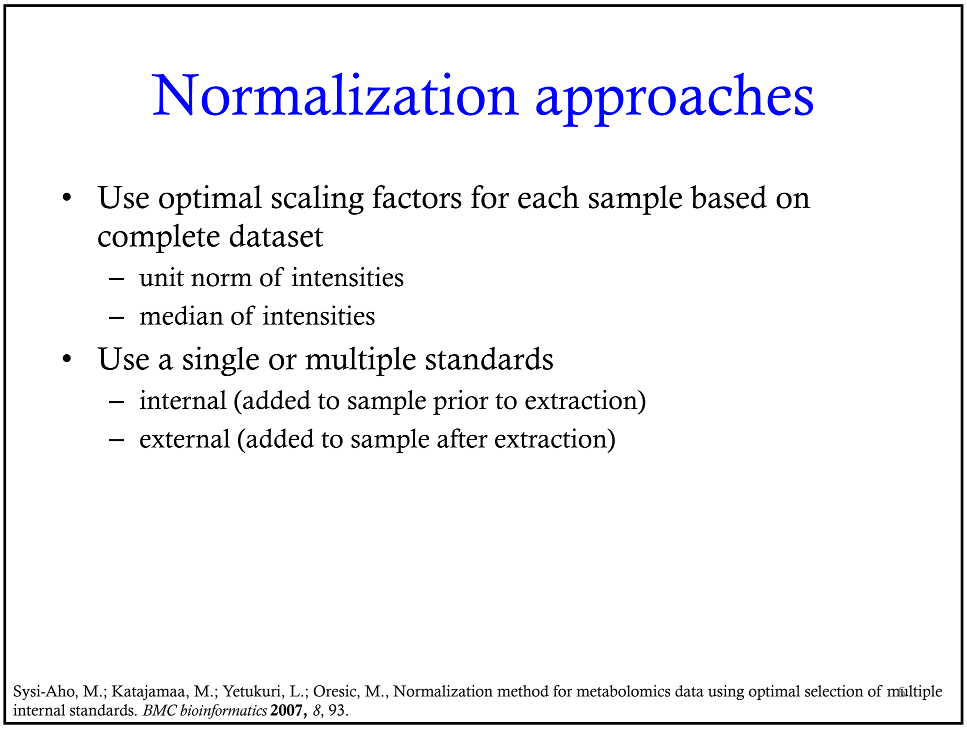
Sysi-Aho, M.; Katajamaa, M.; Yetukuri, L.; Oresic, M., Normalization method for metabolomics data using optimal selection of multiple internal standards. *BMC bioinformatics* **2007**, *8*, 93.



Why normalization?

- Using isotope labeled internal standards for each metabolite is NOT practical because
  - The number of metabolite is large.
  - The metabolites are chemically too diverse to afford a common labeling approach
  - Many metabolites are not known
  - The availability of stable isotope labeled references is very limited.

Sysi-Aho, M.; Katajamaa, M.; Yetukuri, L.; Oresic, M., Normalization method for metabolomics data using optimal selection of multiple internal standards. *BMC bioinformatics* **2007,** *8*, 93.



Normalization approaches

- Use optimal scaling factors for each sample based on complete dataset
  - unit norm of intensities
  - median of intensities
- Use a single or multiple standards
  - internal (added to sample prior to extraction)
  - external (added to sample after extraction)

Sysi-Aho, M.; Katajamaa, M.; Yetukuri, L.; Oresic, M., Normalization method for metabolomics data using optimal selection of multiple internal standards. *BMC bioinformatics* **2007,** *8*, 93.

Normalization approaches

Limitations of approach 1:

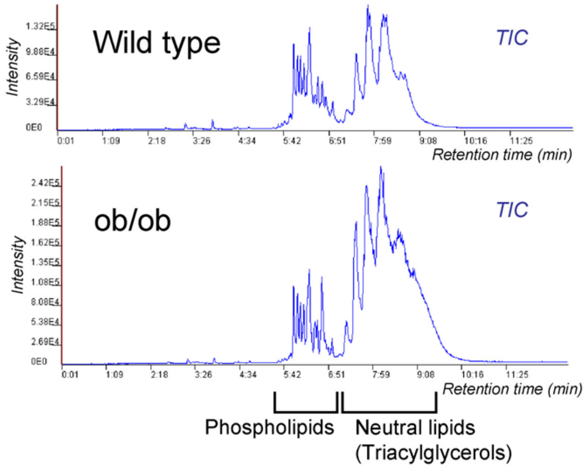
- Suffering from the lack of an absolute concentration reference for different metabolites
- Constraining the data to a specific norm based on total signal affects its covariance structure.
  - Metabolite concentration increase in a specific group of metabolites is not balanced by a decrease of another group.

Limitations of approach 2:

- Assignment of the standards to normalize specific peaks is unclear.

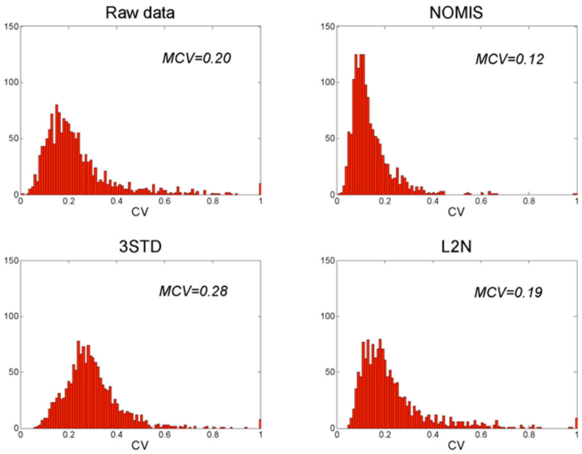
Sysi-Aho, M.; Katajamaa, M.; Yetukuri, L.; Oresic, M., Normalization method for metabolomics data using optimal selection of multiple internal standards. *BMC bioinformatics* **2007**, *8*, 93.

An example



An alternative approach

- Normalization using Optimal selection of Multiple Internal Standards (NOMIS)



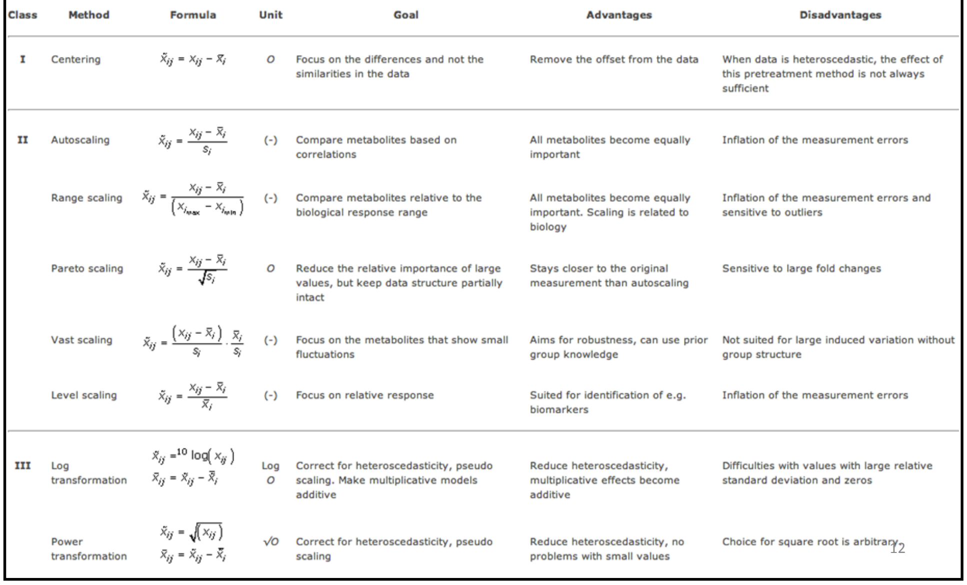
Centering, scaling, transformation

- Induced biological variation through experimental design are what we look for.
- But other factors need to be considered too:
  - Metabolites present in high concentrations are not necessarily more important than those present in low concentrations.
  - Concentrations of metabolites in the central metabolism are generally relatively constant, while the concentrations of metabolites that are present in pathways of the secondary metabolism usually show much larger differences in concentration depending on the environmental conditions.
  - Uninduced biological variation: fluctuations in concentration under identical experimental conditions.
  - Technical variation
  - Heteroscedasticity

Centering, scaling, transformation

- Centering, scaling, and transformation of metabolomics data relate the differences in metabolite concentrations in the different samples to differences in the phenotypes of the cells from which these samples were obtained.

Centering, scaling, transformation



Class	Method	Formula	Unit	Goal	Advantages	Disadvantages
I	Centering	$\tilde{x}_{ij} = x_{ij} - \bar{x}_i$	O	Focus on the differences and not the similarities in the data	Remove the offset from the data	When data is heteroscedastic, the effect of this pretreatment method is not always sufficient
II	Autoscaling	$\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_i}{s_i}$	(-)	Compare metabolites based on correlations	All metabolites become equally important	Inflation of the measurement errors
	Range scaling	$\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_i}{\left(x_{i_{max}} - x_{i_{min}}\right)}$	(-)	Compare metabolites relative to the biological response range	All metabolites become equally important. Scaling is related to biology	Inflation of the measurement errors and sensitive to outliers
	Pareto scaling	$\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_i}{\sqrt{s_i}}$	O	Reduce the relative importance of large values, but keep data structure partially intact	Stays closer to the original measurement than autoscaling	Sensitive to large fold changes
	Vast scaling	$\tilde{x}_{ij} = \frac{\left(x_{ij} - \bar{x}_i\right)}{s_i} \cdot \frac{\bar{x}_i}{s_i}$	(-)	Focus on the metabolites that show small fluctuations	Aims for robustness, can use prior group knowledge	Not suited for large induced variation without group structure
	Level scaling	$\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_i}{\bar{x}_i}$	(-)	Focus on relative response	Suited for identification of e.g. biomarkers	Inflation of the measurement errors
III	Log transformation	$\tilde{x}_{ij} = {}^{10}\log\left(x_{ij}\right)$ $\bar{x}_{ij} = \tilde{x}_{ij} - \bar{\tilde{x}}_i$	Log O	Correct for heteroscedasticity, pseudo scaling. Make multiplicative models additive	Reduce heteroscedasticity, multiplicative effects become additive	Difficulties with values with large relative standard deviation and zeros
	Power transformation	$\tilde{x}_{ij} = \sqrt{\left(x_{ij}\right)}$ $\bar{x}_{ij} = \tilde{x}_{ij} - \bar{\tilde{x}}_i$	√O	Correct for heteroscedasticity, pseudo scaling	Reduce heteroscedasticity, no problems with small values	Choice for square root is arbitrary

Centering

- Converts all the concentrations to fluctuations around zero instead of around the mean of the metabolite concentrations.
- Focus on the fluctuating part of the data.
- Applied in combination with data scaling and transformation.

Scaling

- Divide each variable by a factor
- Different variables have a different scaling factor
- Aim to adjust for the differences in fold differences between the different metabolites.
- Results in the inflation of small values
- Two subclasses
  - Uses a measure of the data dispersion
  - Uses a size measure

Scaling: subclass 1

- Use data dispersion as a scaling factor
  - auto: use the standard deviation as the scaling factor. All the metabolites have a standard deviation of one and therefore the data is analyzed on the basis of correlations instead of covariance.
  - pareto: use the square root of the standard deviation as the scaling factor. Large fold changes are decreased more than small fold changes and thus large fold changes are less dominant compared to clean data.
  - vast: use standard deviation and the coefficient of variation as scaling factors. This results in a higher importance for metabolites with a small relative sd.
  - range: use (max-min) as scaling factors. Sensitive to outliers.

Scaling: subclass 2

- Use average as scaling factors
  - The resulting values are changes in percentages compared to the mean concentration.
  - The median can be used as a more robust alternative.

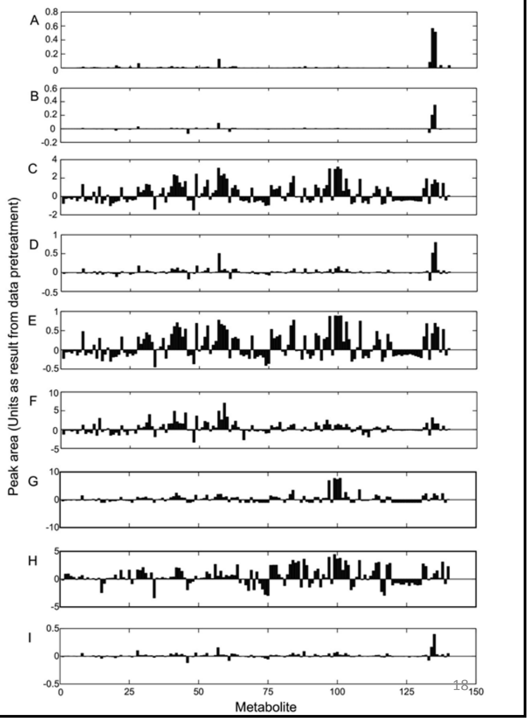
Transformation

- Log and power transformation
- Both reduce large values relatively more than the small values.
- Log transformation
  - pros: removal of heteroscedasticity
  - cons: unable to deal with the value zero.
- Power transformation
  - pros: similar to log transformation
  - cons: not able to make multiplicative effects additive

17

Centering, scaling, transformation

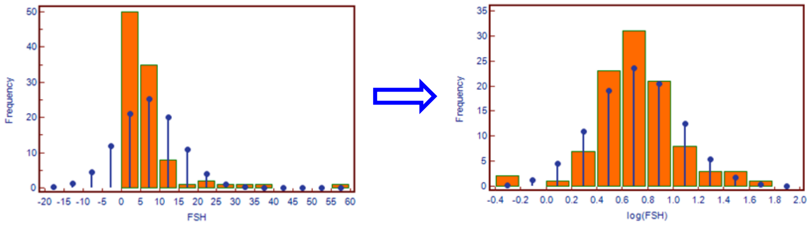
A. original data
B. centering
C. auto
D. pareto
E. range
F. vast
G. level
H. log
I. power



18

Log transformation, again

- Hard to do useful statistical tests with a skewed distribution.

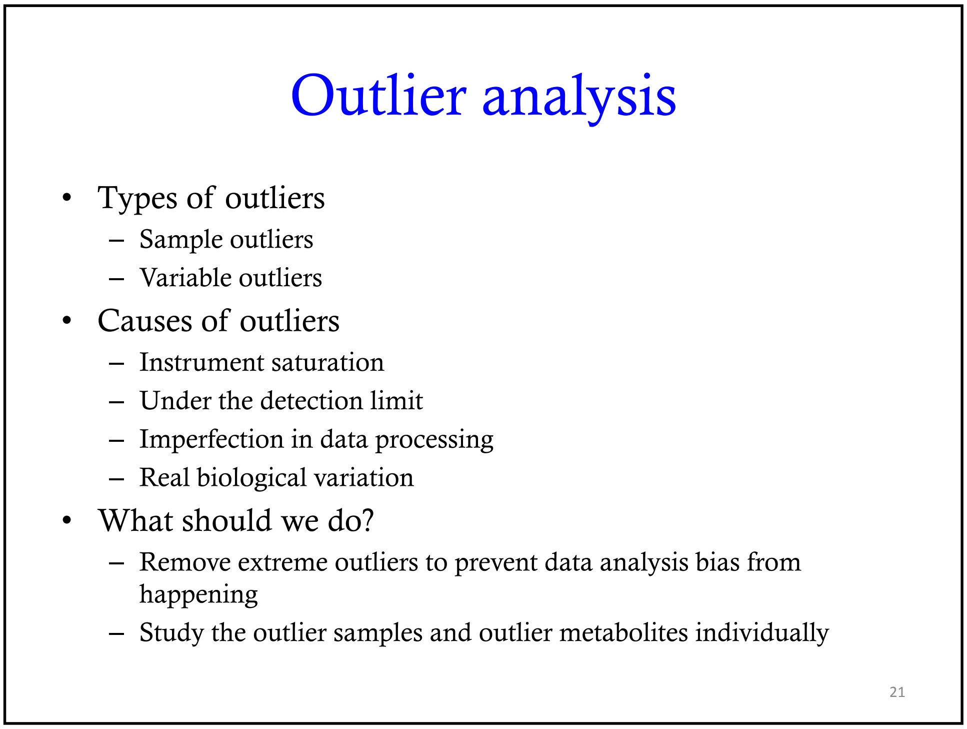


- A skewed distribution or exponentially decaying distribution can be transformed into a Gaussian distribution by applying a log transformation.

http://www.medcalc.org/manual/transforming_data_to_normality.php

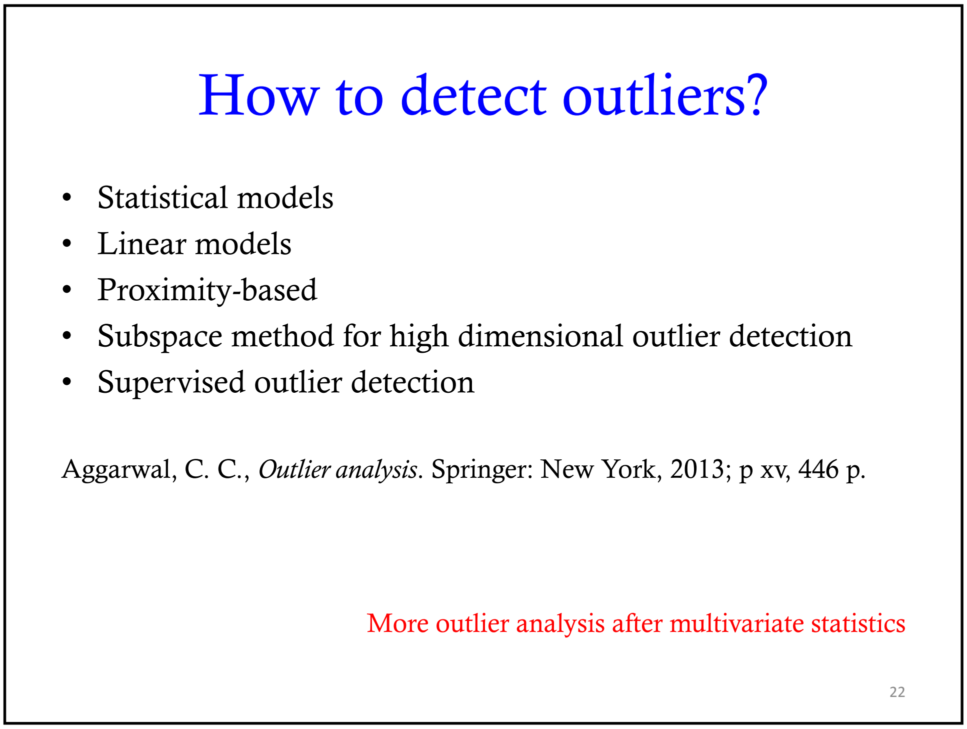
Outlier analysis

DB	Name	Mass	RT	Instrument	CP1	CP4	CP5	CP2	CP3	CP6
NIST	(+)-galactose	217	15.6432	GC	0.07220543	0.06763952	0.05645402	0.19147302	0.89388092	0.52391403
NIST	(+)-Mannose	111	16.1069	GC	0.06917746	0.00706453	0.05991109	0.00102921	0.14233213	0.092648
NIST	1,4-Butanedian	174	15.5695	GC	0.01915597	0.01540993	0.01316491	0.00985555	0.01740541	0.01441497
NIST	1,5-anhydroglu	147	15.6755	GC	0.27335179	0.69264869	1.33827843	0.66726119	1.27042292	0.06215285
HMDB	11-beta-hydrox	483.254531	21.6416101	ES+	0.01971402	0.01280484	0.00774709	0.00515638	0.01824955	0.01379582
HMDB	13-Hydroperox	313.235153	21.0007149	ES+	0.0147208	0.0120644	0.01158581	0.01320837	0.01149144	0.01216372
HMDB	17-Hydroxylino	295.227487	19.9254568	ES+	0.01621269	0.01457397	0.01452179	0.01704355	0.01475745	0.01451786
NIST	1-cyclohexene	127	5.4681	GC	0.05719642	0.02546542	0.07211745	0.04553947	0.03202725	0.04938444
HMDB	1-Phenylethyla	122.097448	24.9784501	ES-	0.14633149	0.12267952	0.10285151	0.13400616	0.14149369	0.10108422
NIST	**Glucose**	**73**	**16.2833**	**GC**	**194.464526**	**193.390828**	**378.660169**	**300.732186**	**539.778597**	**808.889528**
S	2,3,4-Trihydrox	135.044716	3.57634868	ES+	0.02561722	0.01533536	0.0089078	0.01876651	0.01079665	0.0110569
S	2,3-Diaminopro	105.070158	3.32029346	ES+	0.02569076	0.02508627	0.02008465	0.02179467	0.02051684	0.02129889
NIST	2,4-bishydroxyl	103	10.5092	GC	0.02483397	0.01906404	0.01757337	0.01352737	0.02714883	0.01498949
HMDB	2,4-Diaminobut	119.084405	3.87908984	ES+	0.07609924	0.06936676	0.04789434	0.06989164	0.06511184	0.11452725
HMDB	2,6 dimethylhe	302.232031	18.0258597	ES+	0.00828555	0.00897917	0.00535005	0.00325899	0.00279613	0.00174071
S	2-aminobutyric	130	7.16317	GC	0.47301493	0.59640279	0.62730713	0.36893587	0.36989687	0.40698242
HMDB	2-Ethylacrylic a	101.064209	17.8115754	ES-	0.02246093	0.01582154	0.01235352	0.00815676	0.0261188	0.02319674
HMDB	2-Ethylhydracry	119.071994	15.2265313	ES+	0.02424006	0.02386794	0.02395416	0.02393768	0.02427207	0.02398221
NIST	2-Hydroxy-3-m	145	7.07231	GC	0.16934316	0.14489713	0.01160834	0.07573712	0.06076782	0.03866568
S	2-hydroxybutyr	131	6.52167	GC	0.78522263	0.58500452	1.09530082	0.58572461	0.68570527	0.54563315
S	2-Hydroxygluta	198	12.5112	GC	0.00898708	0.01172375	0.01207569	0.0127851	0.00683606	0.00905586
NIST	2-hydroxypyrid	152	5.21487	GC	2.14103338	2.15321383	2.38078173	1.52075446	2.672602	1.83195088
HMDB	2-Ketohexanoic	131.070273	3.73535823	ES+	0.0037897	0.00544538	0.00522184	0.00472749	0.00469586	0.00667645
NIST	2-methy-2-hyd	221	7.67217	GC	0.01402093	0.01259207	0.01729476	0.01095531	0.01533539	0.01034363
HMDB	2-Methylaceto:	117.0538	3.6120234	ES+	0.02579758	0.03159884	0.02355465	0.03002218	0.03539992	0.03038233
HMDB	2-Methylbutyrc	246.169434	19.1986568	ES+	0.00933674	0.00692789	0.0041819	0.00388336	0.01132417	0.01192467
HMDB	2-Octenedioic a	173.079059	15.3173769	ES+	0.00964669	0.00700057	0.0043726	0.00278267	0.0134011	0.01273207
NIST	2-oxo-3-methy	89	6.26107	GC	0.06824451	0.06385102	0.09296196	0.06557308	0.09416076	0.06153601
S	2-oxo-4-methy	71	7.76032	GC	0.07848267	0.06625182	0.09493752	0.0647436	0.1161296	0.07321774



Outlier analysis

- Types of outliers
  - Sample outliers
  - Variable outliers
- Causes of outliers
  - Instrument saturation
  - Under the detection limit
  - Imperfection in data processing
  - Real biological variation
- What should we do?
  - Remove extreme outliers to prevent data analysis bias from happening
  - Study the outlier samples and outlier metabolites individually



How to detect outliers?

- Statistical models
- Linear models
- Proximity-based
- Subspace method for high dimensional outlier detection
- Supervised outlier detection

Aggarwal, C. C., *Outlier analysis*. Springer: New York, 2013; p xv, 446 p.

More outlier analysis after multivariate statistics

Univariate vs. multivariate stats

- In math, univariate statistics include all statistical techniques for analyzing a single variable of interest.
  - **_t_-tests**
  - **ANOVA**
  - multiple regression
- Multivariate statistics includes all statistical techniques for analyzing two or more variables of interest.
- In metabolomics, each metabolite is a variable. Each sample is represented as a vector of many dimensions.

Univariate vs. multivariate stats

An example:

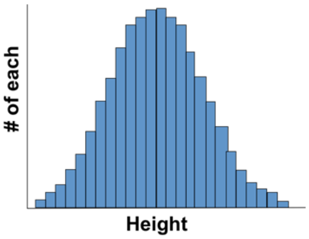
- To quantify the nutritional habits of American women, nutrient intake was measured for a random sample of 1000 women. In a univariate study, we might ask each woman in the survey how much vitamin C they take in on a daily basis.
- In a multivariate study, we might look at not only vitamin C, but calcium, iron, vitamin A as well.

http://sites.stat.psu.edu/~ajw13/stat505/fa06/01_courseintro/WK1_courseoverview.htm

Univariate statistics

- A basic way of presenting univariate data is to create a frequency distribution of the individual cases.

Due to the Central Limit Theorem, many of these frequency distributions can be modeled as a normal/Gaussian distribution.



Gaussian distribution

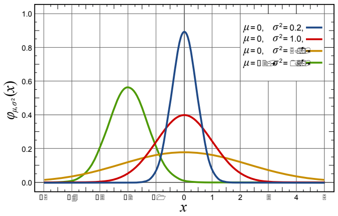
- The total area underneath each density curve is equal to 1.

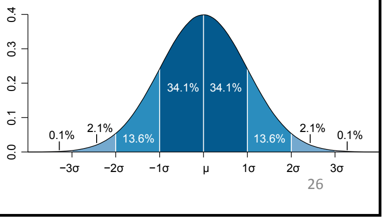
$$\varphi(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\left(\frac{x-\mu}{\sigma}\right)^2}$$

$$\text{mean} = \mu$$

$$\text{variance} = \sigma^2$$

$$\text{standard deviation} = \sigma$$





https://en.wikipedia.org/wiki/Normal_distribution

Sample statistics

Sample mean: $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

Sample variance: $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$

Sample standard deviation: $S = \sqrt{S^2}$

https://en.wikipedia.org/wiki/Normal_distribution

Test of significance

- One-sample *t*-test: is the sample drawn from a known population?

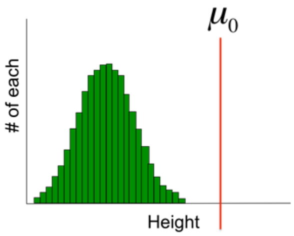
Null hypothesis H_0: $\mu = \mu_0$

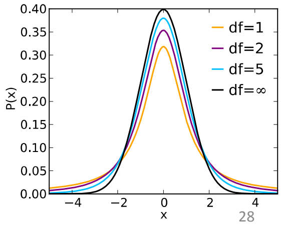
Alternative hypothesis H_1: $\mu < \mu_0$

Test statistic: $t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$

Sample standard deviation: $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}$

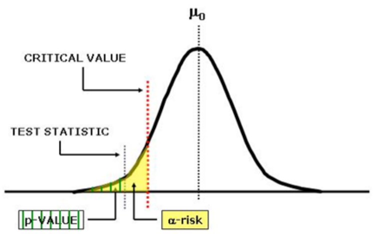
The test statistic *t* follows a student's t distribution. The distribution has *n-1* degrees of freedom.

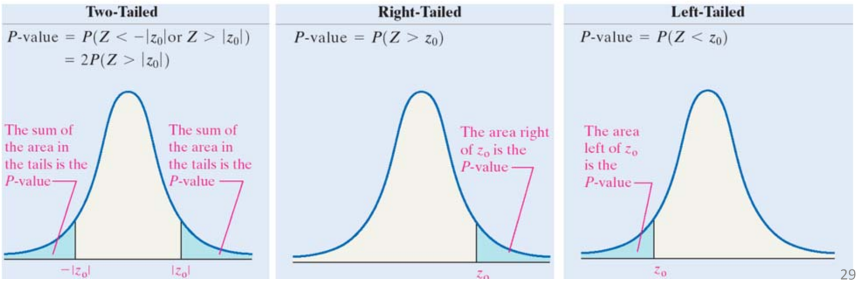




Student’s *t*-test and *p*-value

When the null hypothesis is rejected, the result is said to be statistically significant.




Two-Tailed
P-value $= P(Z < -|z_0| \text{ or } Z > |z_0|)$
$= 2P(Z > |z_0|)$
The sum of the area in the tails is the P-value
The sum of the area in the tails is the P-value
$-|z_0|$
$|z_0|$
Right-Tailed
P-value $= P(Z > z_0)$
The area right of z_0 is the P-value
z_0
Left-Tailed
P-value $= P(Z < z_0)$
The area left of z_0 is the P-value
z_0

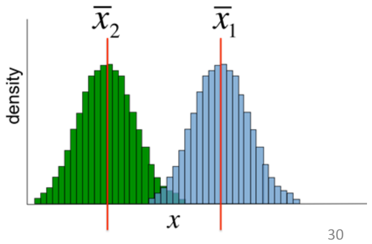
Test of significance

- Two-sample *t*-test: are the two populations different?

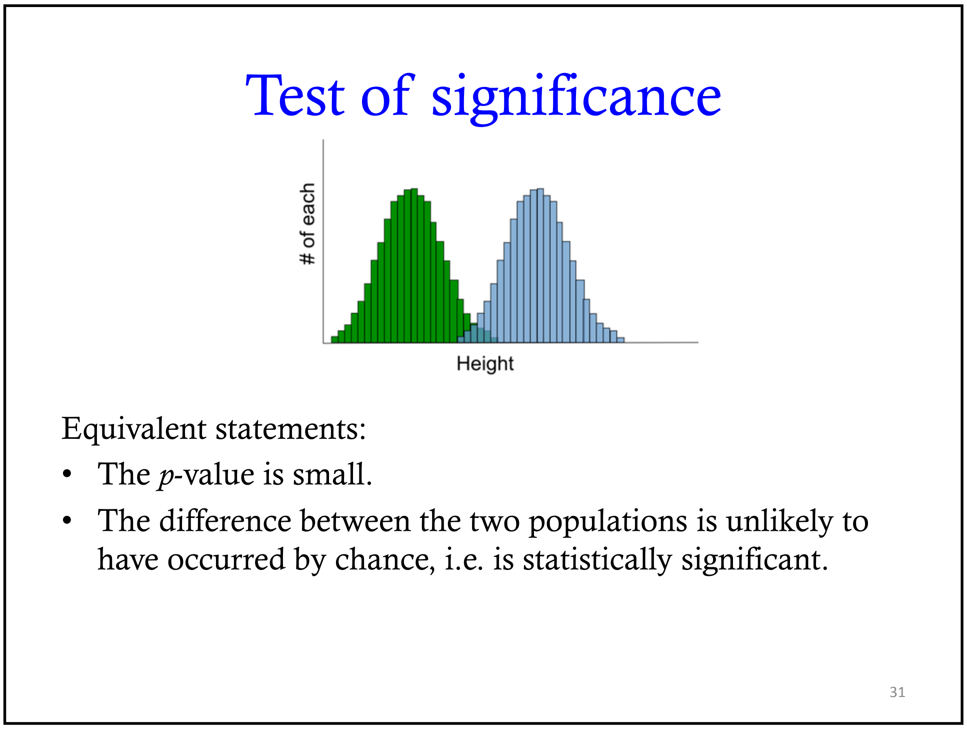
Null hypothesis H_0: $\mu_1 - \mu_2 = 0$

Alternative hypothesis H_1: $\mu_1 - \mu_2 \neq 0$

Test statistic: $t = \dfrac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$



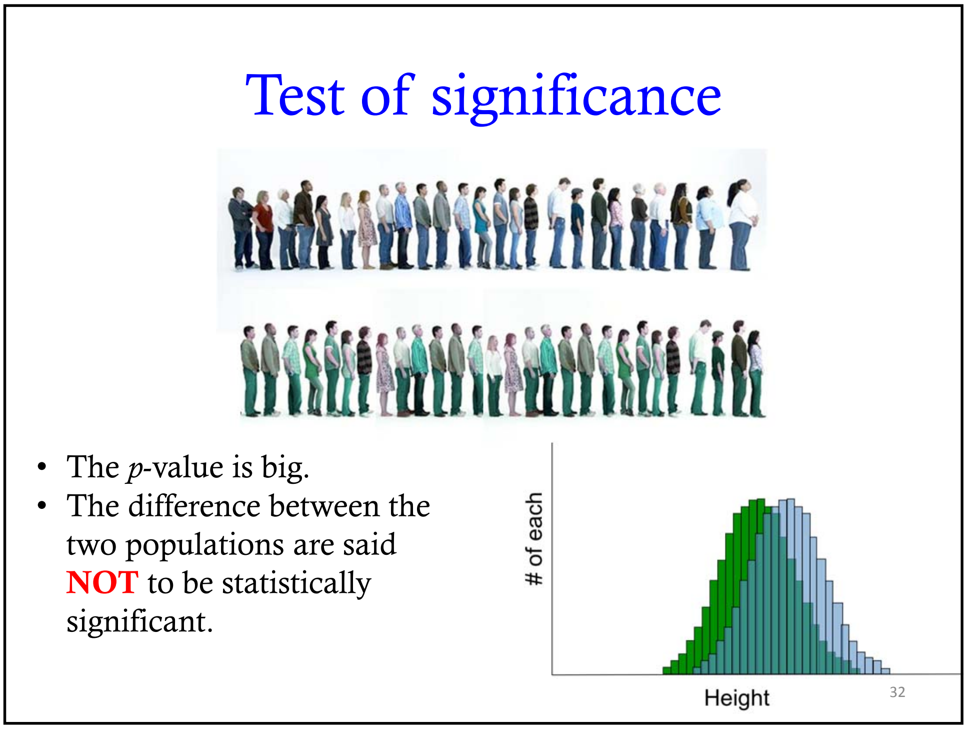
- The two samples should be independent.



Test of significance

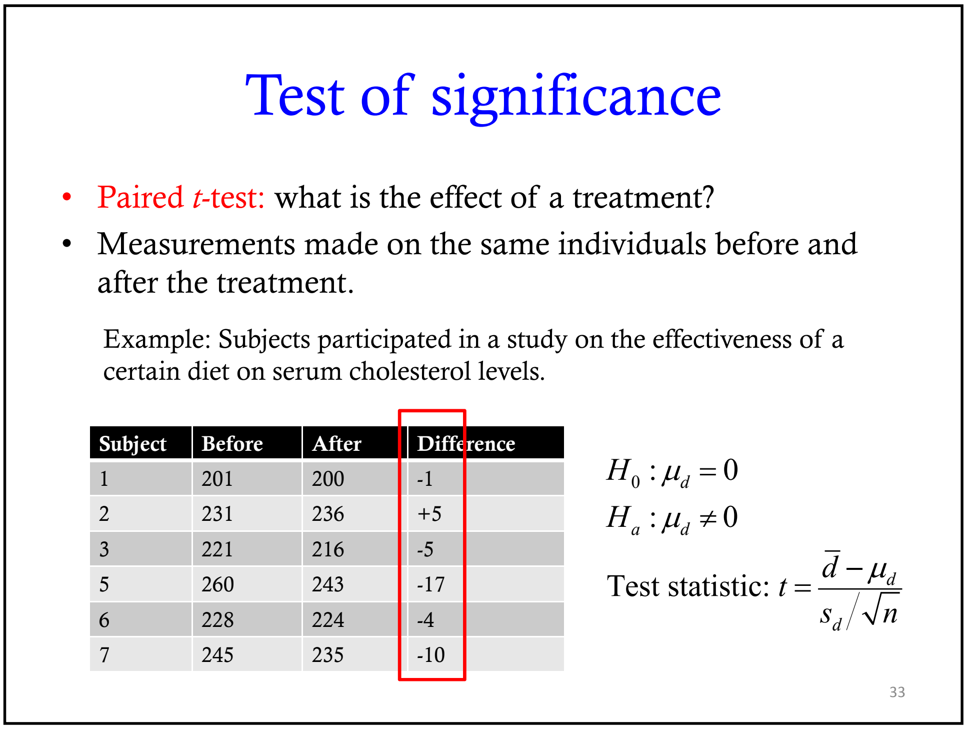
Equivalent statements:

- The *p*-value is small.
- The difference between the two populations is unlikely to have occurred by chance, i.e. is statistically significant.



Test of significance

- The *p*-value is big.
- The difference between the two populations are said **NOT** to be statistically significant.



Test of significance

- Paired t-test: what is the effect of a treatment?
- Measurements made on the same individuals before and after the treatment.

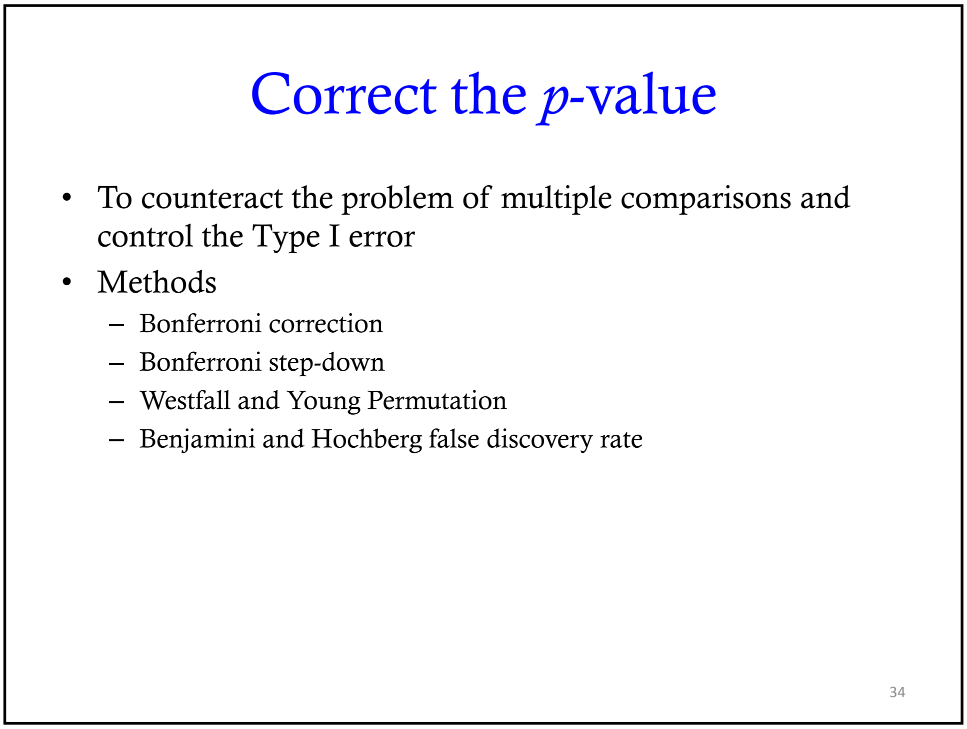
Example: Subjects participated in a study on the effectiveness of a certain diet on serum cholesterol levels.

Subject	Before	After	Difference
1	201	200	-1
2	231	236	+5
3	221	216	-5
5	260	243	-17
6	228	224	-4
7	245	235	-10

$$H_0 : \mu_d = 0$$

$$H_a : \mu_d \neq 0$$

Test statistic: $t = \dfrac{\bar{d} - \mu_d}{s_d \big/ \sqrt{n}}$



Correct the p-value

- To counteract the problem of multiple comparisons and control the Type I error
- Methods
  - Bonferroni correction
  - Bonferroni step-down
  - Westfall and Young Permutation
  - Benjamini and Hochberg false discovery rate

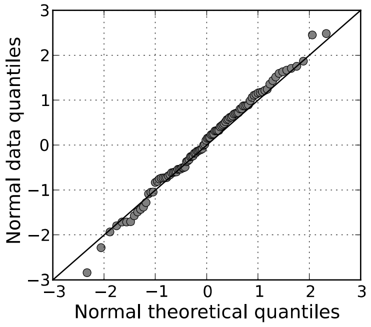
Normality test

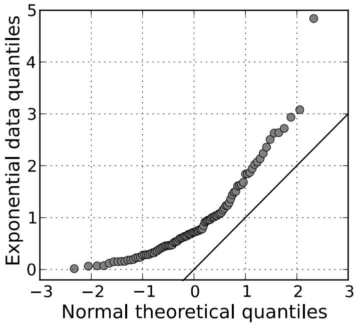
- Assumptions of a *t*-test
  - Each of the two populations being compared should follow a Gaussian distribution.
  - The samples are randomly drawn without any selection bias.

- Normality test
  - **Quantile-quantile plot (QQ plot)**
  - **Shapiro-Wilk**
  - Kolmogorov-Smirnov
  - Pearson’s chi-squared test

Normality test

- QQ plot: compare two probability distributions by plotting their quantiles against each other





http://en.wikipedia.org/wiki/Quantile-quantile_plot

Normality test

- Shapiro-Wilk test: tests the null hypothesis that a sample came from a Gaussian distributed population
- Test statistic

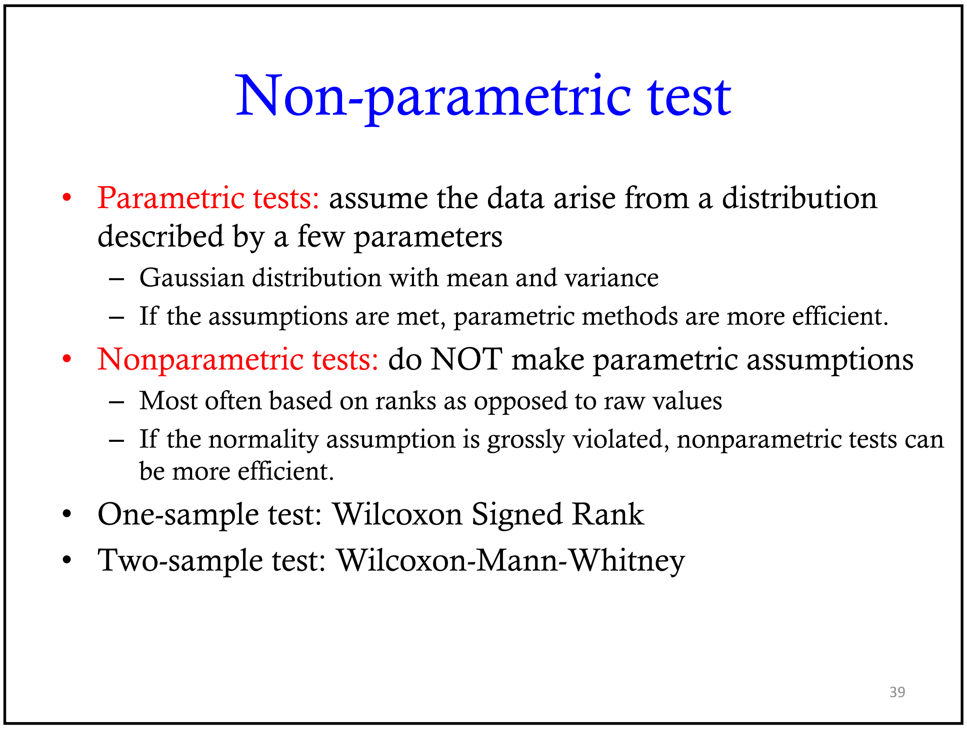
$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2}$$

- $x_{(i)}$ is the i-th smallest number in the list.
- The constants a_i are given by
- $\left(a_1, \cdots, a_n\right) = \dfrac{m^T V^{-1}}{\left(m^T V^{-1} V^{-1} m\right)^{1/2}}$
- m=$\left(m_1, \cdots, m_n\right)^T$ are the expected values of the order statistics of i.i.d. random variables sampled from the standard normal distributions.
- V is the covariance matrix of those order statistics.

http://en.wikipedia.org/wiki/Shapiro-Wilk_test

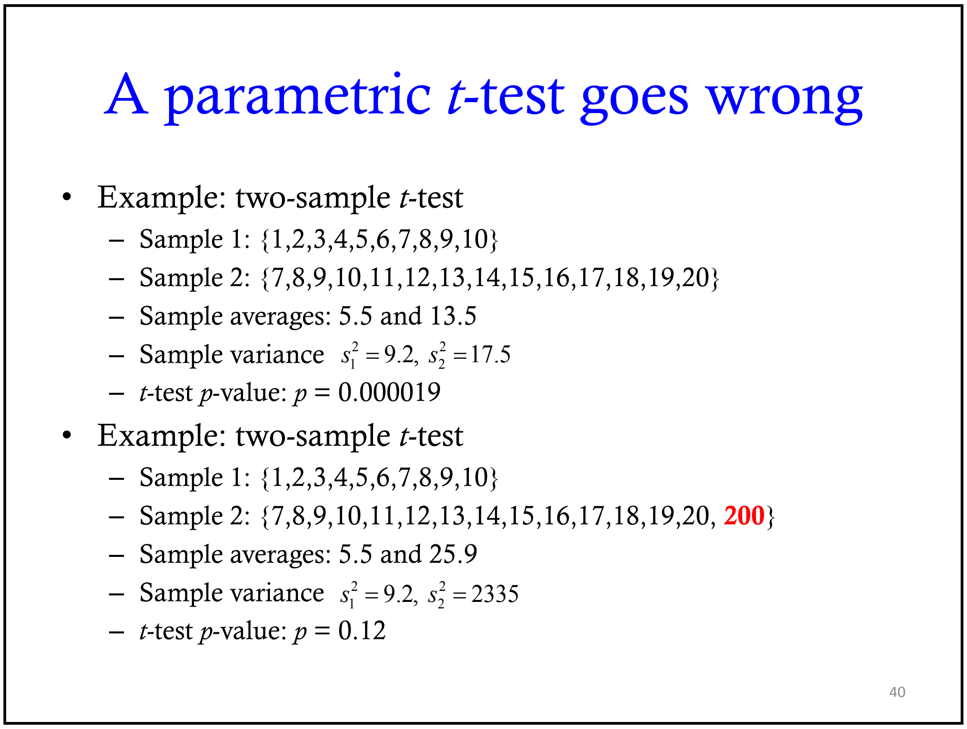
t-test and normality assumption

- Based on the Central Limit Theorem, the distribution of a sample mean approaches Gaussian when the sample size is sufficiently large.
  - Rule of thump: n > 30
- What if the normality assumption is violated?
  - Use non-parametric methods



Non-parametric test

- Parametric tests: assume the data arise from a distribution described by a few parameters
  - Gaussian distribution with mean and variance
  - If the assumptions are met, parametric methods are more efficient.
- Nonparametric tests: do NOT make parametric assumptions
  - Most often based on ranks as opposed to raw values
  - If the normality assumption is grossly violated, nonparametric tests can be more efficient.
- One-sample test: Wilcoxon Signed Rank
- Two-sample test: Wilcoxon-Mann-Whitney



A parametric *t*-test goes wrong

- Example: two-sample *t*-test
  - Sample 1: {1,2,3,4,5,6,7,8,9,10}
  - Sample 2: {7,8,9,10,11,12,13,14,15,16,17,18,19,20}
  - Sample averages: 5.5 and 13.5
  - Sample variance $s_1^2 = 9.2,\ s_2^2 = 17.5$
  - *t*-test *p*-value: $p = 0.000019$
- Example: two-sample *t*-test
  - Sample 1: {1,2,3,4,5,6,7,8,9,10}
  - Sample 2: {7,8,9,10,11,12,13,14,15,16,17,18,19,20, **200**}
  - Sample averages: 5.5 and 25.9
  - Sample variance $s_1^2 = 9.2,\ s_2^2 = 2335$
  - *t*-test *p*-value: $p = 0.12$

Non-parametric tests

- Many non-parametric methods convert raw values to ranks and then analyze ranks.
- In case of ties, mid-ranks are used.
  - Raw data: {105, 120, 120, 121}
  - Ranks: {1, 2.5, 2.5, 4}

Parametric test	non-parametric counterpart
one-sample *t* test	Wilcoxon signed-rank
two-sample *t* test	Wincoxon two-sample rank-sum
k-sample ANOVA	Kruskal-Wallis
Pearson *r*	Spearman ρ

ANOVA

- Compare the means of three or more populations
- A generalization of the two-sample *t*-test
- Use the *F*-distribution to test for significance
- Factor: an independent treatment variable whose settings (values) are controlled and varied by the experimenter
- Level: the intensity setting of a factor
- 1-way ANOVA
  - Only one factor is considered.

    Null hypothesis: There is NO difference in the population means of the different levels of the only factor.

    Alternative hypothesis: The means are not the same, i.e. at least one pair of means is different.

2-way ANOVA

- Two factors, A and B, are considered.
- Possible hypotheses
  - Case 1

    Null: There is no difference in the means of factor A.

    Alternative: The means are not equal.
  - Case 2

    Null: There is no difference in means of factor B.

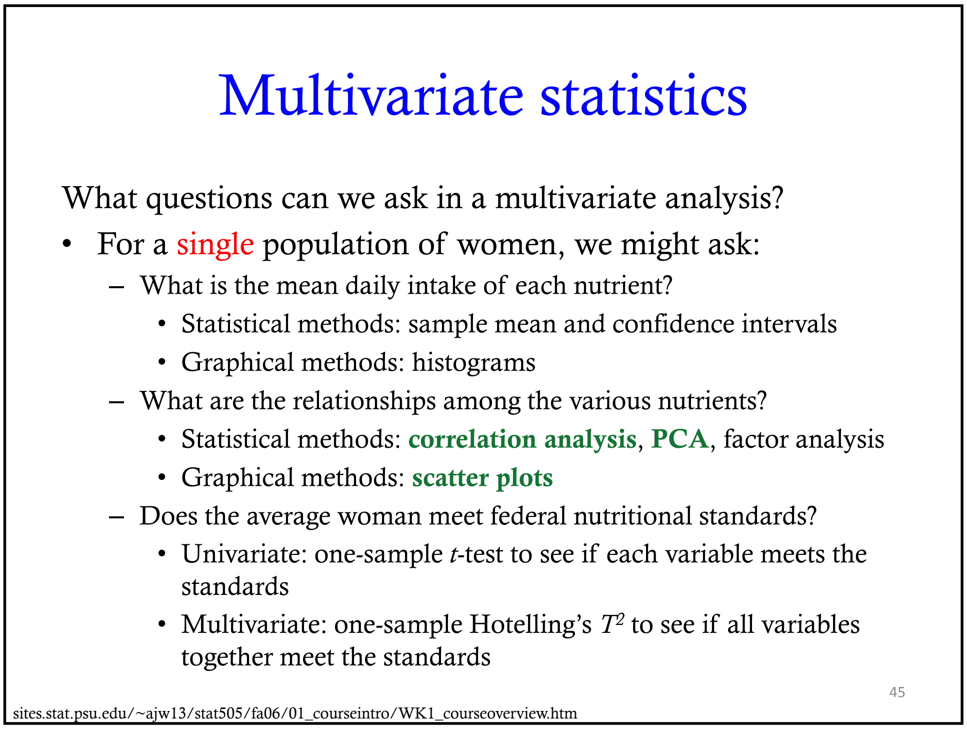
    Alternative: The means are not equal.
  - Case 3

    Null: There is no interaction between factors A and B.

    Alternative: There is an interaction between factors A and B.

3-way ANOVA

- Three factors (A, B, and C) are considered.
- The main effects are factors A, B, and C.
- The two-factor interactions are: AB, AC, and BC.
- There is also a three-factor interaction: ABC.
- For each of the seven cases
  - Null: There is no difference in means.
  - Alternative: The means are not equal.

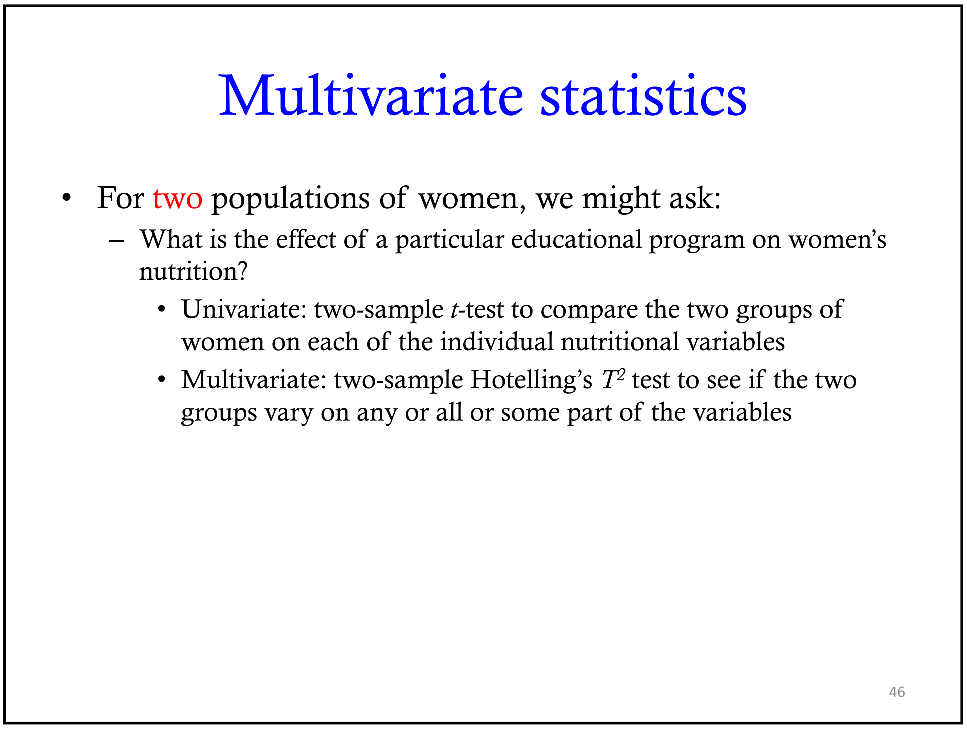


Multivariate statistics

What questions can we ask in a multivariate analysis?

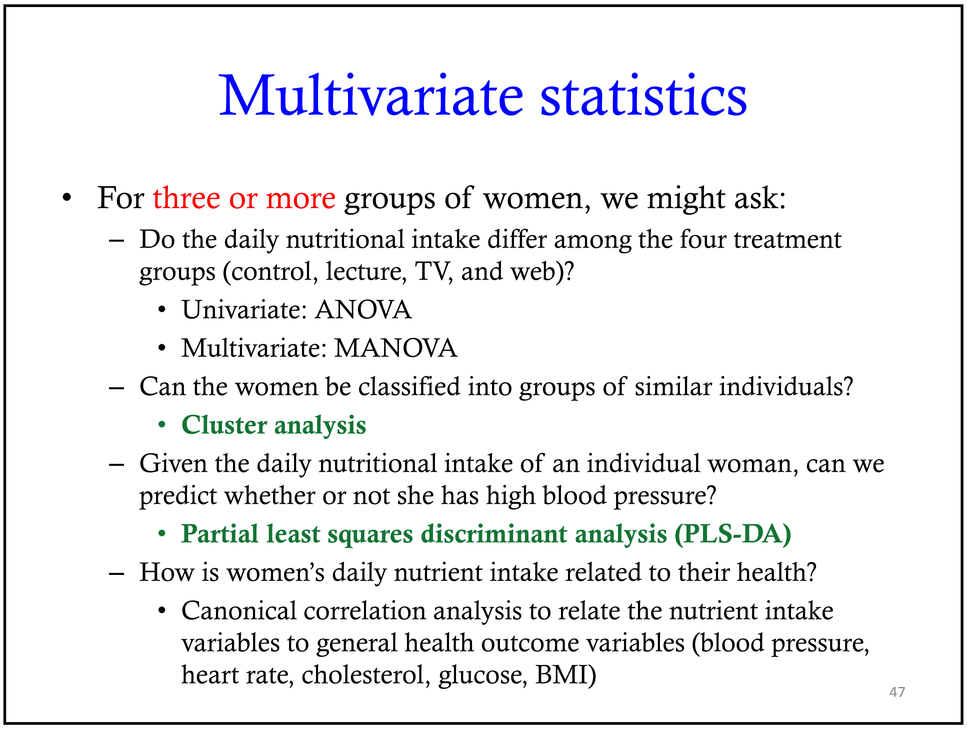
- For a single population of women, we might ask:
  - What is the mean daily intake of each nutrient?
    - Statistical methods: sample mean and confidence intervals
    - Graphical methods: histograms
  - What are the relationships among the various nutrients?
    - Statistical methods: **correlation analysis**, **PCA**, factor analysis
    - Graphical methods: **scatter plots**
  - Does the average woman meet federal nutritional standards?
    - Univariate: one-sample *t*-test to see if each variable meets the standards
    - Multivariate: one-sample Hotelling's T^2 to see if all variables together meet the standards

sites.stat.psu.edu/~ajw13/stat505/fa06/01_courseintro/WK1_courseoverview.htm



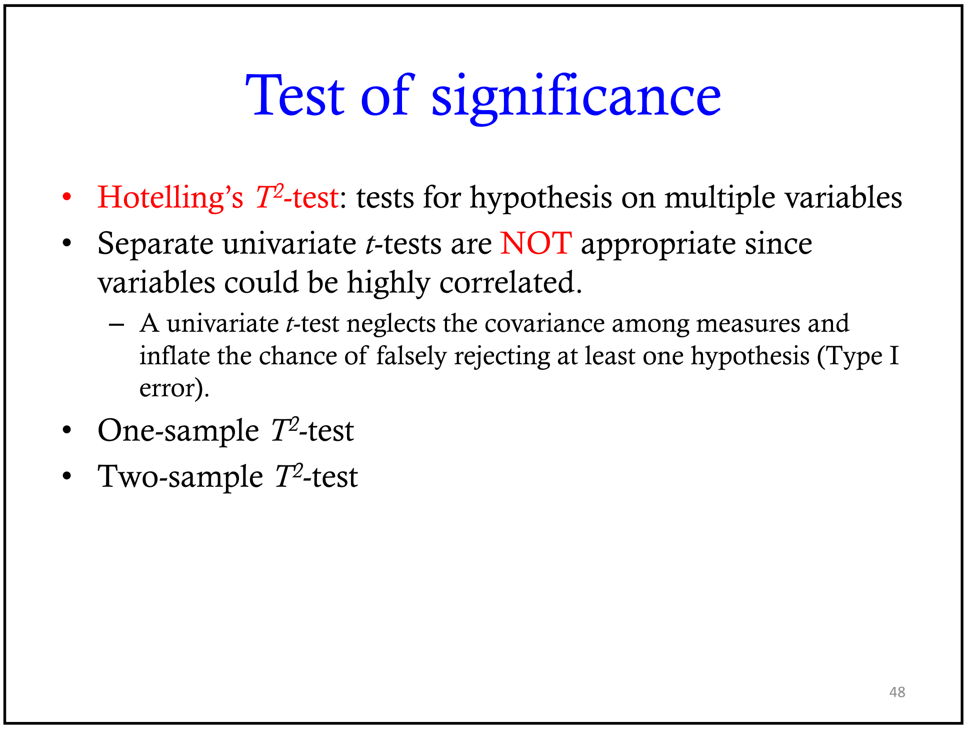
Multivariate statistics

- For two populations of women, we might ask:
  - What is the effect of a particular educational program on women's nutrition?
    - Univariate: two-sample *t*-test to compare the two groups of women on each of the individual nutritional variables
    - Multivariate: two-sample Hotelling's T^2 test to see if the two groups vary on any or all or some part of the variables



Multivariate statistics

- For three or more groups of women, we might ask:
  - Do the daily nutritional intake differ among the four treatment groups (control, lecture, TV, and web)?
    - Univariate: ANOVA
    - Multivariate: MANOVA
  - Can the women be classified into groups of similar individuals?
    - **Cluster analysis**
  - Given the daily nutritional intake of an individual woman, can we predict whether or not she has high blood pressure?
    - **Partial least squares discriminant analysis (PLS-DA)**
  - How is women's daily nutrient intake related to their health?
    - Canonical correlation analysis to relate the nutrient intake variables to general health outcome variables (blood pressure, heart rate, cholesterol, glucose, BMI)



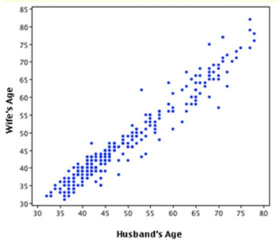
Test of significance

- Hotelling's T^2-test: tests for hypothesis on multiple variables
- Separate univariate t-tests are NOT appropriate since variables could be highly correlated.
  - A univariate t-test neglects the covariance among measures and inflate the chance of falsely rejecting at least one hypothesis (Type I error).
- One-sample T^2-test
- Two-sample T^2-test

Correlation

- Graphical method: scatter plot
- Pearson product-moment correlation coefficient

$$\rho = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$



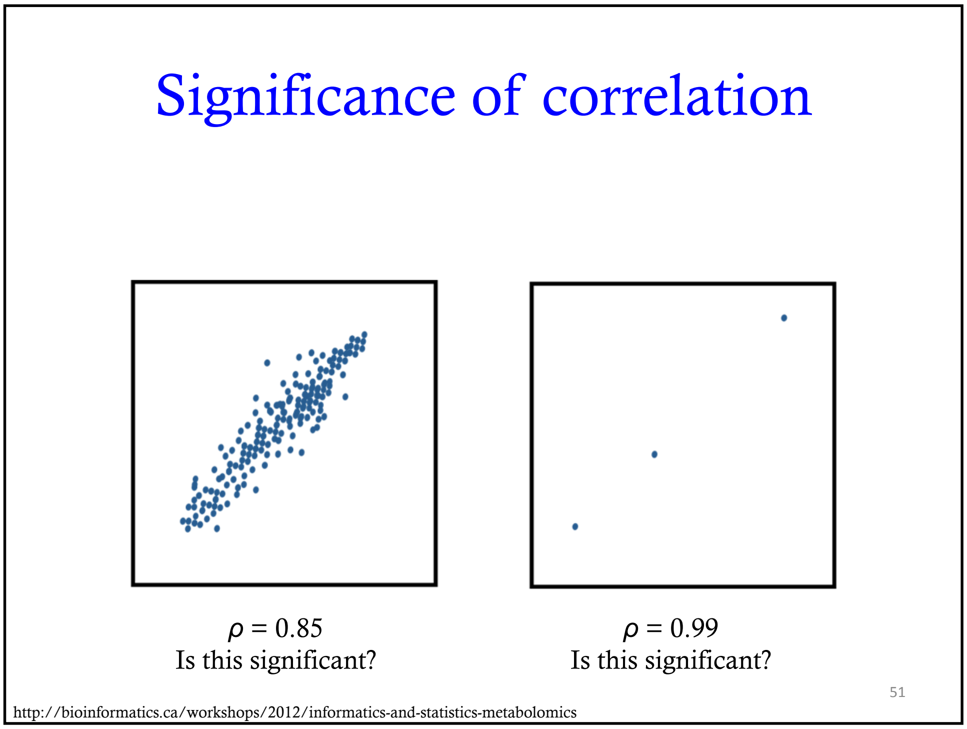
- ρ measures the linear relationship between variable x and y.
- ρ and R^2 (coefficient of determination)

49

Correlation

1	0.8	0.4	0	-0.4	-0.8	-1
1	1	1		-1	-1	-1
0	0	0	0	0	0	0

50



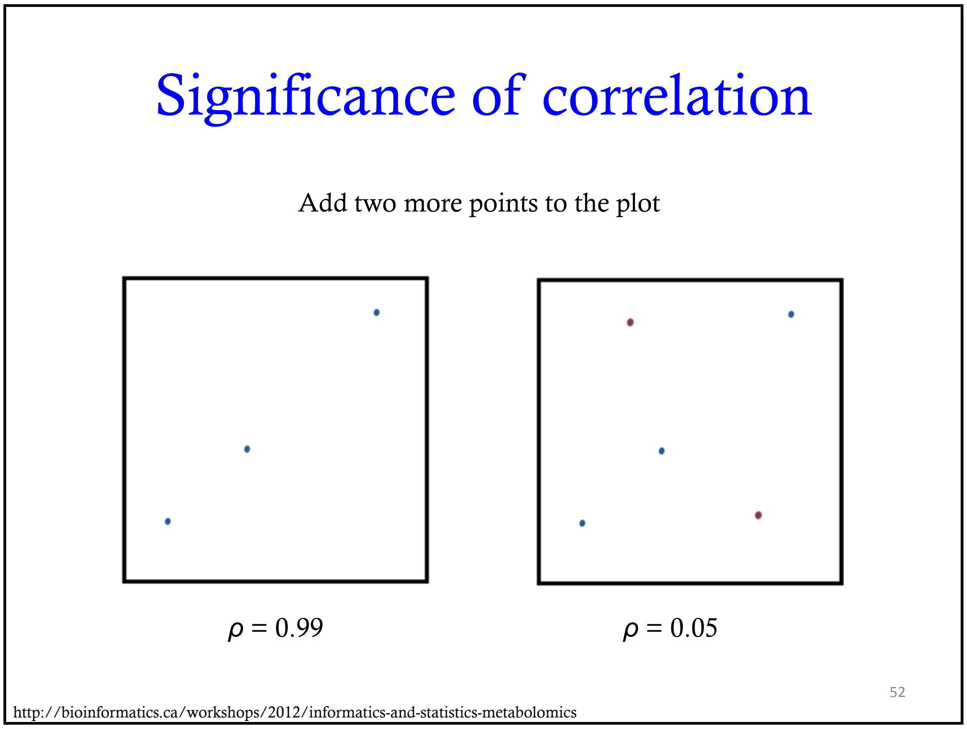
Significance of correlation

$\rho = 0.85$
Is this significant?

$\rho = 0.99$
Is this significant?

51

http://bioinformatics.ca/workshops/2012/informatics-and-statistics-metabolomics



Significance of correlation

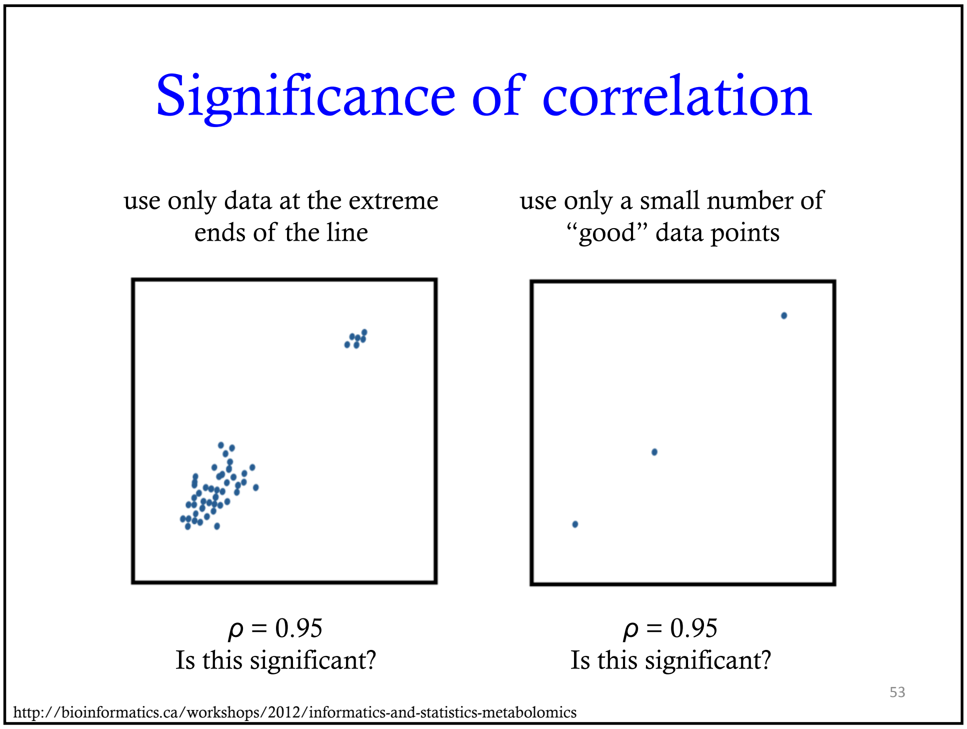
Add two more points to the plot

$\rho = 0.99$

$\rho = 0.05$

52

http://bioinformatics.ca/workshops/2012/informatics-and-statistics-metabolomics



Significance of correlation

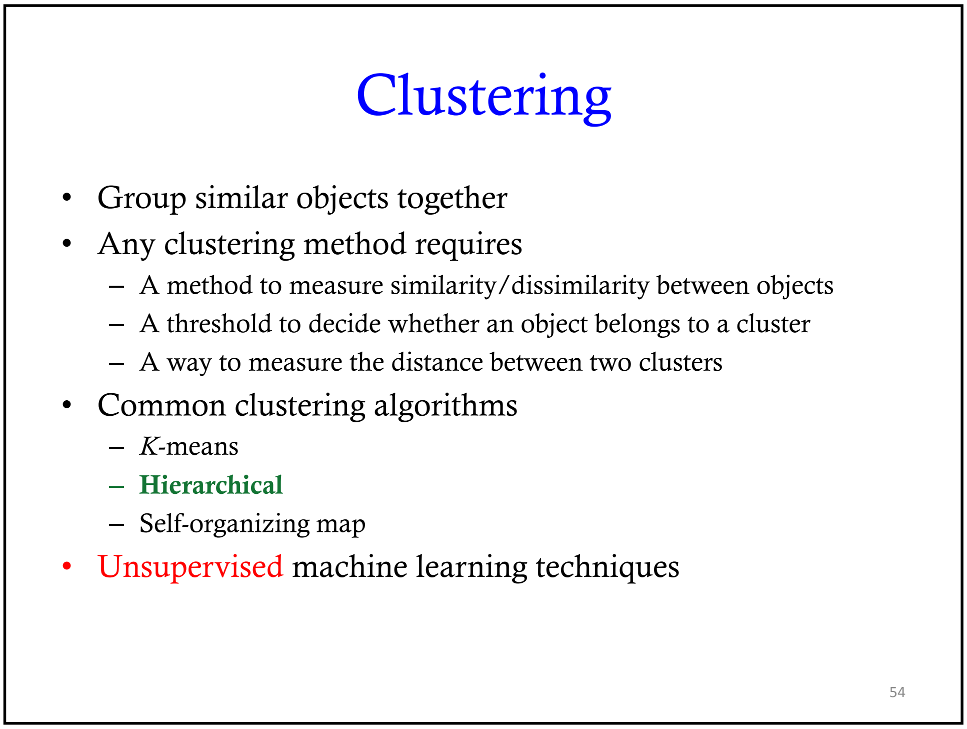
use only data at the extreme ends of the line

use only a small number of "good" data points

$\rho = 0.95$
Is this significant?

$\rho = 0.95$
Is this significant?

http://bioinformatics.ca/workshops/2012/informatics-and-statistics-metabolomics

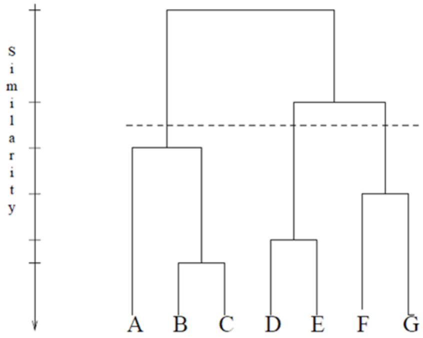


Clustering

- Group similar objects together
- Any clustering method requires
  - A method to measure similarity/dissimilarity between objects
  - A threshold to decide whether an object belongs to a cluster
  - A way to measure the distance between two clusters
- Common clustering algorithms
  - *K*-means
  - **Hierarchical**
  - Self-organizing map
- Unsupervised machine learning techniques

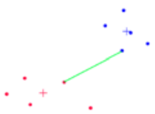
Hierarchical clustering

1. Find the two closest objects and merge them into a cluster
2. Find and merge the next two closest objects (or an object and a cluster, or two clusters)
3. Repeat step 2 until all objects have been clustered

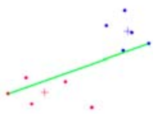


Hierarchical clustering

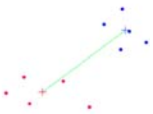
- Methods to measure similarity between objects
  - Euclidean, Manhattan
  - Pearson correlation
  - Cosine similarity
- Linkage: ways to measure the distance between two clusters



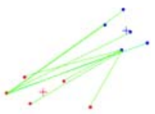
single



complete

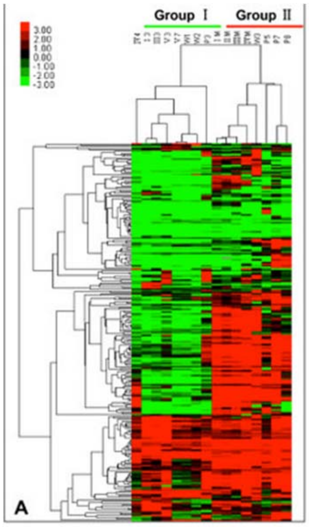


centroid



average

Hierarchical clustering

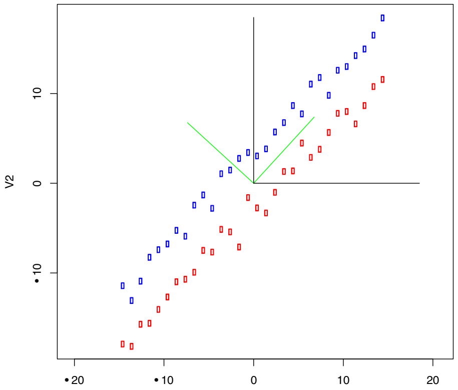


Dimension reduction

- Why to reduce dimensions?
  - Help visualize and interpret dependencies among sets of variables
- Methods
  - **PCA**
  - Single value decomposition (SVD)
- PCA
  - Transform correlated variables into uncorrelated variables
  - Order the uncorrelated variables by the amount of variance they explain in the data
  - Discard low-variance variables

PCA

- Unsupervised
- Note of caution: dimension of maximum variance is not necessarily the dimension of maximum separation.



Classification

- Use a training set of correctly-identified observations to build a predictive model
- Predict to which of a set of categories a new observation belongs
- Supervised machine learning
- Methods
  - Linear discriminant analysis
  - Support vector machine (SVM)
  - Artificial neural network (ANN)
  - *k*-nearest neighbor
  - Random forest
  - **Partial least squares discriminant anlaysis**

PLS-DA

- The predictive model describes the relationship between the dependent and independent variables.
- Interpretation of the model
  - R^2X and R^2Y
    - fraction of the variance that the model explains in the independent and dependent variables
    - Range: 0-1
  - Q^2Y
    - measure of the predictive accuracy of the model
    - usually estimated by cross validation or permutation testing
    - Range: 0-1
    - > 0.5 is considered good while > 0.9 is outstanding

Note of caution

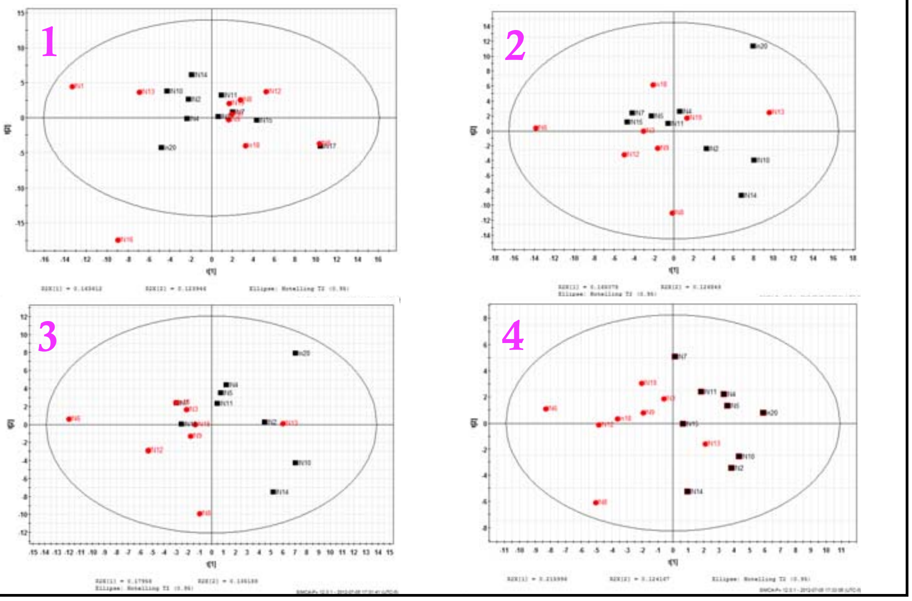
- Supervised classification methods are powerful.
- BUT, they can overfit your data, severely.

Do NOT skip the clustering step.
Do assess the significance of the predictive model.

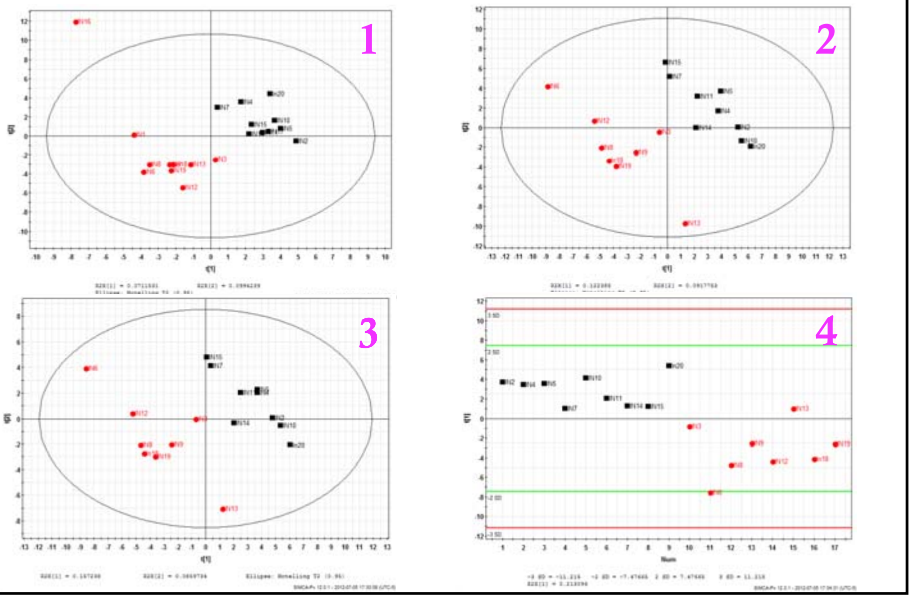
Outlier analysis, again

Model	Description	PCA		PLS-DA			
		# of PCs	R^2X	# of PCs	R^2X	R^2Y	Q^2Y
Model 1	20 samples with all 211 Variables	5	0.54	3	0.26	0.97	-0.07
Model 2	211 variables after excluding samples 1, 16 and 17 from model 1	4	0.50	3	0.30	0.98	0.15
Model 3	137 variables after excluding variables with VIP < 0.5 in model 2	5	0.59	3	0.33	0.97	0.35
Model 4	70 variables after excluding variables with VIP < 1 in model 2	5	0.61	1	0.21	0.70	0.49

PCA plots

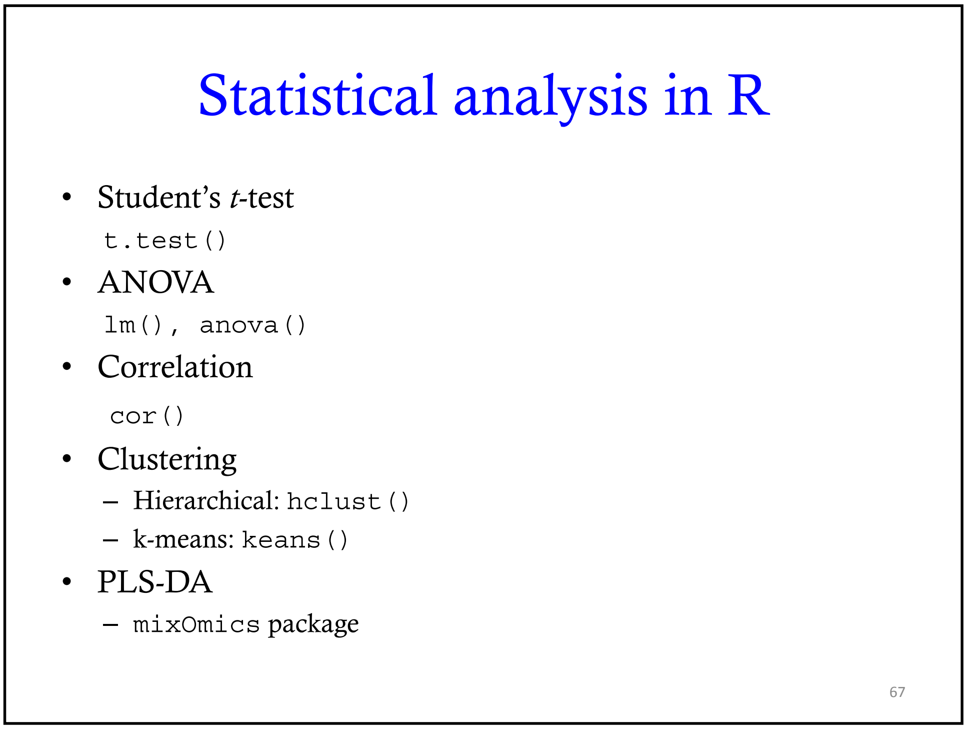


PLS-DA plots



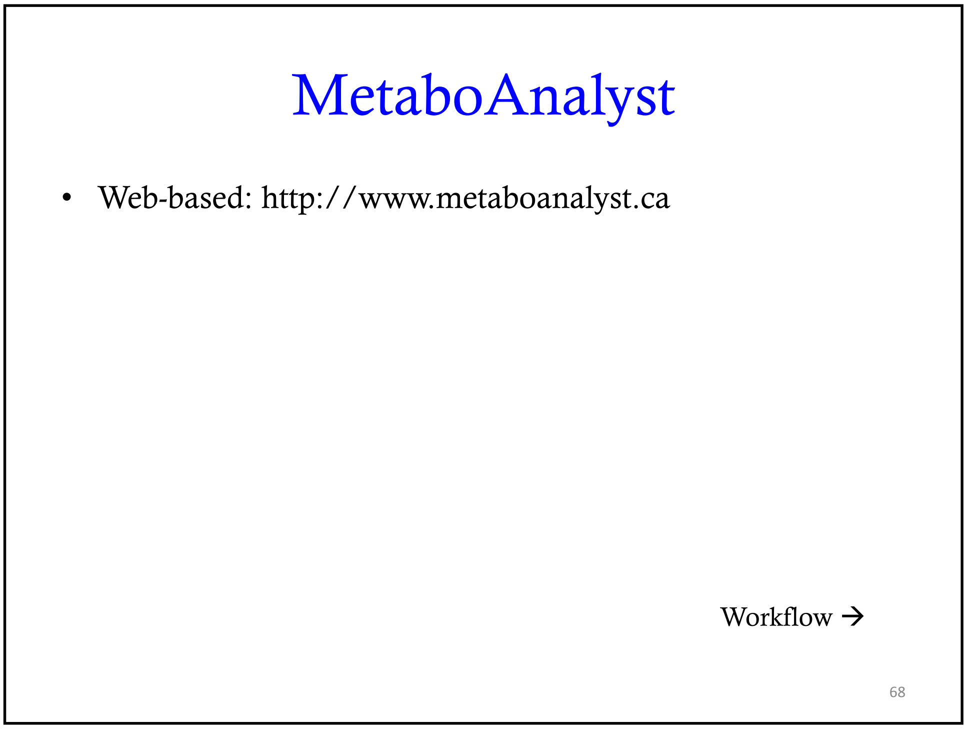
Software packages

- Open source
  - **R**
  - **MetaboAnalyst**
  - MultiExperiment Viewer (MeV)
  - Octave (very similar to MATLAB)
  - Many others ……
- Proprietary
  - SIMPA-P
  - SAS
  - MATLAB
  - Many others ……



Statistical analysis in R

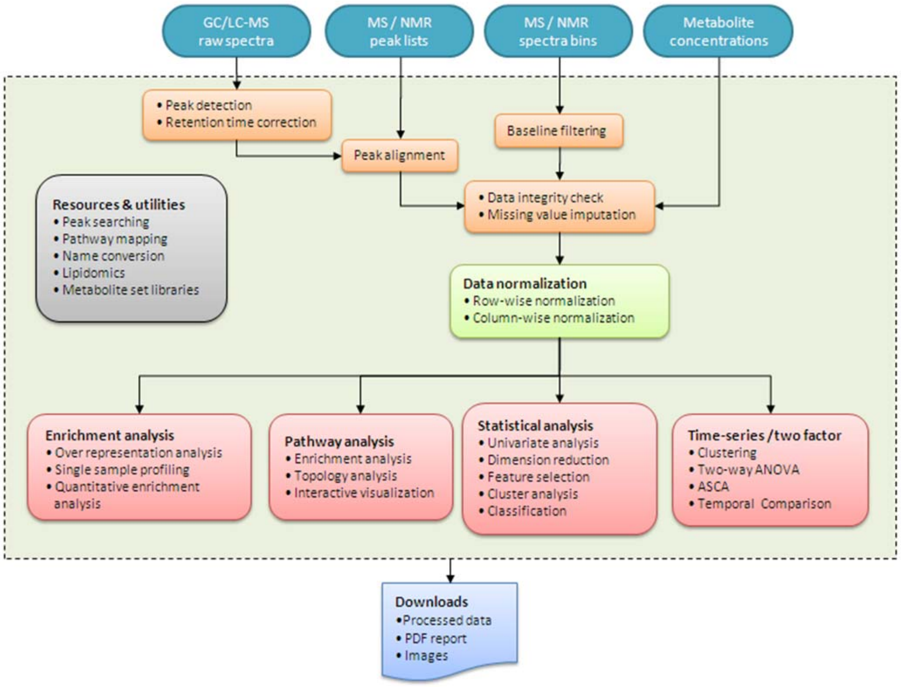
- Student's *t*-test
  `t.test()`
- ANOVA
  `lm(), anova()`
- Correlation
  `cor()`
- Clustering
  - Hierarchical: `hclust()`
  - k-means: `keans()`
- PLS-DA
  - `mixOmics` package



MetaboAnalyst

- Web-based: http://www.metaboanalyst.ca

Workflow →



GC/LC-MS raw spectra

MS / NMR peak lists

MS / NMR spectra bins

Metabolite concentrations

- Peak detection
- Retention time correction

Baseline filtering

Peak alignment

- Data integrity check
- Missing value imputation

Resources & utilities
- Peak searching
- Pathway mapping
- Name conversion
- Lipidomics
- Metabolite set libraries

Data normalization
- Row-wise normalization
- Column-wise normalization

Enrichment analysis
- Over representation analysis
- Single sample profiling
- Quantitative enrichment analysis

Pathway analysis
- Enrichment analysis
- Topology analysis
- Interactive visualization

Statistical analysis
- Univariate analysis
- Dimension reduction
- Feature selection
- Cluster analysis
- Classification

Time-series /two factor
- Clustering
- Two-way ANOVA
- ASCA
- Temporal Comparison

Downloads
- Processed data
- PDF report
- Images

Thank you!

Two-sample *t*-test

- Degree of freedom for a two-sample *t*-test

$$d.f. = \frac{\left(s_1^2 / n_1 + s_2^2 / n_2\right)^2}{\left(s_1^2 / n_1\right)^2 \Big/ \left(n_1 - 1\right) + \left(s_2^2 / n_2\right)^2 \Big/ \left(n_2 - 1\right)}$$

T^2 test

- One-sample T^2 test

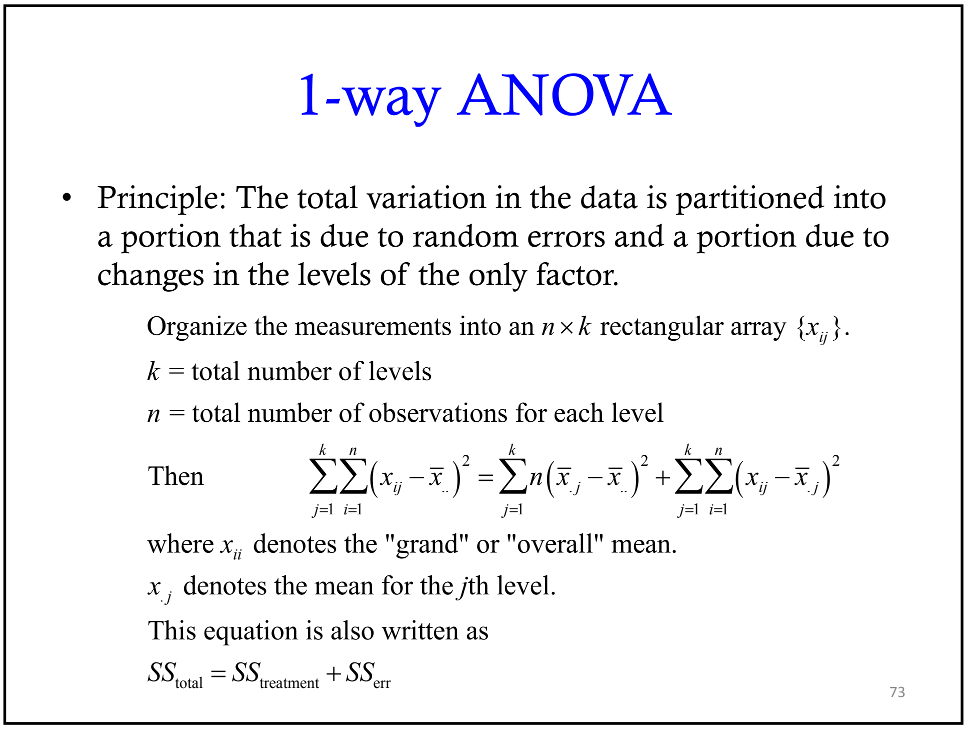
$$T^2 = n\left(\bar{X} - \mu_0\right)^T S^{-1}\left(\bar{X} - \mu_0\right)$$

$\bar{X}$ is the vector of column means

S is the sample covariance matrix

- Two-sample T^2 test

$$T^2 = \frac{n_1 n_2}{n_1 + n_2}\left(\bar{X}_1 - \bar{X}_2\right)^T S_{\text{pooled}}^{-1}\left(\bar{X}_1 - \bar{X}_2\right)$$



1-way ANOVA

- Principle: The total variation in the data is partitioned into a portion that is due to random errors and a portion due to changes in the levels of the only factor.

Organize the measurements into an $n \times k$ rectangular array $\{x_{ij}\}$.

k = total number of levels

n = total number of observations for each level

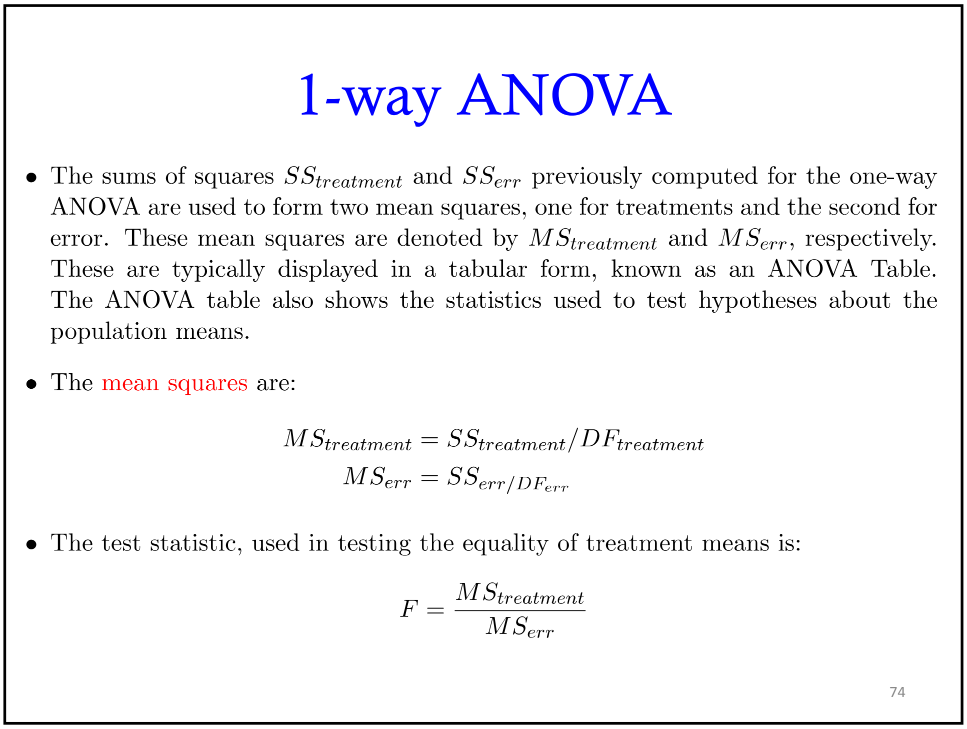
Then $$\sum_{j=1}^{k}\sum_{i=1}^{n}\left(x_{ij}-\bar{x}_{..}\right)^2=\sum_{j=1}^{k}n\left(\bar{x}_{.j}-\bar{x}_{..}\right)^2+\sum_{j=1}^{k}\sum_{i=1}^{n}\left(x_{ij}-\bar{x}_{.j}\right)^2$$

where x_{ii} denotes the "grand" or "overall" mean.

$x_{.j}$ denotes the mean for the jth level.

This equation is also written as

$SS_{\text{total}} = SS_{\text{treatment}} + SS_{\text{err}}$



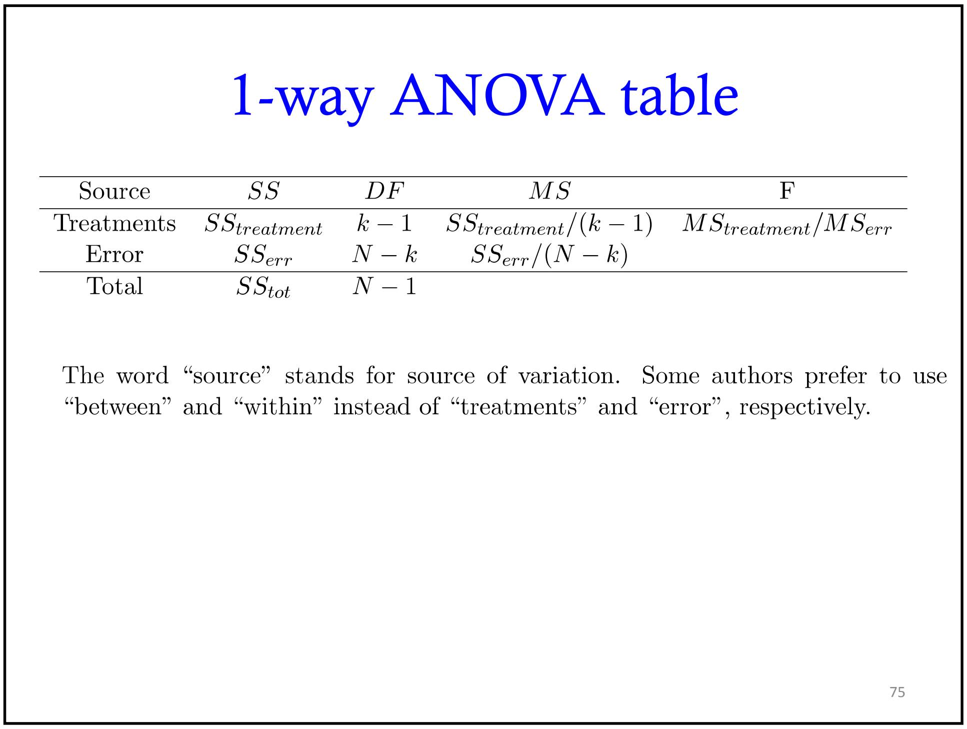
1-way ANOVA

- The sums of squares $SS_{treatment}$ and SS_{err} previously computed for the one-way ANOVA are used to form two mean squares, one for treatments and the second for error. These mean squares are denoted by $MS_{treatment}$ and MS_{err}, respectively. These are typically displayed in a tabular form, known as an ANOVA Table. The ANOVA table also shows the statistics used to test hypotheses about the population means.
- The mean squares are:

$$MS_{treatment} = SS_{treatment}/DF_{treatment}$$
$$MS_{err} = SS_{err/DF_{err}}$$

- The test statistic, used in testing the equality of treatment means is:

$$F = \frac{MS_{treatment}}{MS_{err}}$$



1-way ANOVA table

Source	SS	DF	MS	F
Treatments	$SS_{treatment}$	$k-1$	$SS_{treatment}/(k-1)$	$MS_{treatment}/MS_{err}$
Error	SS_{err}	$N-k$	$SS_{err}/(N-k)$	
Total	SS_{tot}	$N-1$		

The word “source” stands for source of variation. Some authors prefer to use “between” and “within” instead of “treatments” and “error”, respectively.