

Experiment Planning and Design

Lecture 5: Techniques for Hypothesis Testing

Claus Aranha

caranha@cs.tsukuba.ac.jp

Department of Computer Science

2015-06-02

Outline

Techniques for hypothesis testing in specific situations

- Calculation of α and β
- Power of a test, and sample size
- Testing Variances and Distributions
- Non-parametric testing, Difference testing, and Proportional Testing

Homework Review

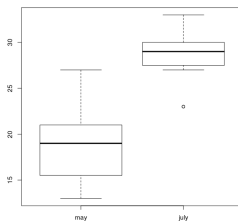
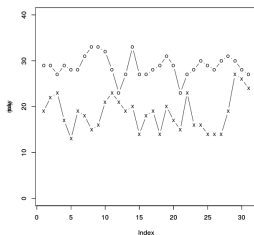
Describe Your Data

- Analyze your data (or sample data) using statistical means
- Means, Variances, Confidence interval
- Draw a simple hypothesis test about your data

Best report (R.Y.)

Comparison of temperature data for months May and July in “air quality” data set.

- 1 Simple analysis of the data with mean, var, max and min (maybe SD would be better?)
- 2 Drawing plots of the two sets of interest:



- 3 A difference is observed. To confirm this difference, a hypothesis test is made.

Best report (R.Y.)

Comparison of temperature data for months May and July in “air quality” data set.

- 1 Kolmogorov-Smirnov test (KS-test) to determine the distribution of the data sets (goodness of fit)

```
> ks.test(may, "pnorm", mean=mean(may), sd=sd(may))  
D = 0.1072, p-value = 0.8683
```

- 2 F-test to compare the variances of the two data sets

```
> var.test(may, july)  
F = 2.4769, num df = 30, denom df = 30, p-value = 0.01538
```

- 3 Because the variances are different, use Welch's t-test.

```
> t.test(may, july, var.equal=F)  
t = -12.66, df = 50.829, p-value < 2.2e-16
```

- 4 Hypothesis test confirms the observation that July is bigger than may.

Common Mistakes: Lack of testing

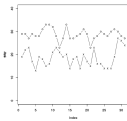
Many of you calculated mean, variance, and plots for two or more variables. Based on these values, you reported some trend.

The figure shows that car 1 is faster than car 2

- Eye observation is good for guessing facts about the data, but is not good enough for confirming these facts.
- Comparing CI intervals is a little better, but not completely – it describes data, but does not infer relationships.
- Ideally, you should perform **hypothesis tests** to confirm the instincts that you obtained from pure observation.

Remember that humans suffer from confirmation bias!

Common Mistakes: Lack of testing

1. Observe Trend in Data	
2. Formulate Hypothesis	Null Hypothesis: means are equal. Alternate Hypothesis: means are different
3. Hypothesis Testing	<code>t.test(may,july,var.equal=F)</code>
4. Make a decision	“We can reject the null hypothesis”

Remember, though: this cannot be purely mechanical! You need to be conscious about what each step means for your experiment!

Bad Mistake: Parroting



The professor wants tests, so I will do... tests... *squawk!*

```
> t.test(car)  
...any result...
```

“The test showed that the average of car is bigger than the average of bicycle!”

- First: **Tests** only exist in the context of **questions**.
It makes no sense to make a test if you do not decide what are you testing beforehand
- Second: `t.test` without parameters test the sample against the null hypothesis “mean is equal to 0”, and the alternative hypothesis “mean is different than zero”
- **Important:** If you do not understand a concept – please ask questions! You will not learn if you don’t ask questions.

Image taken from <http://www.eurosavant.com/2010/02/01/is-that-a-parrot-in-your-pocket-or/>

Homework: Conclusions

- Each of you should have received a **grade** and some **comments** on Manaba.
- These grades **will not** influence your final grade
- However, these grades should give you a good idea of how I will evaluate the final report.

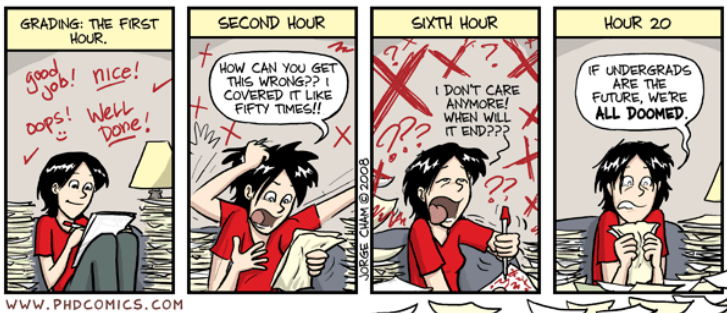


Image from PhdComics: <http://www.phdcomics.com/comics/archive.php?comicid=974>

Review of Last Lecture (1)

Experimental Hypothesis

During an experiment, we **observe** a characteristic of the data. This allows us to formulate a **hypothesis**, which can be tested to confirm or refute our observation.

For example:

- The average weight of the part produced by this machine is 500g
- The air pollution in this city is 15ppm;
- Method A is faster than method B by 10%;
- Participants in a survey like chocolate more than vanilla
- This method has best performance when $\rho = 60$;

Review of Last Lecture (2)

After we decide what is the **effect** that we are **measuring**, we have to formulate a **null hypothesis** and an **alternate hypothesis** that describe our observation.

Null Hypothesis

The null hypothesis usually states that **there is no relationship between the factors we are studying**. Or **any differences that we see are not significant**.

The Null hypothesis indicates the absence of an effect.

Alternate hypothesis

The Alternative hypothesis indicates the presence of **an effect**.

We cannot “prove” the alternate hypothesis. We can show that it is **much more likely** than the null hypothesis (which we call “rejecting the null hypothesis”).

Examples of Null Hypothesis

Factors	Null Hypothesis
Methods (A,B), Speed	The speed of the process is independent of the method
Air Pollution, Cities in region	The air pollution in all cities in this region is the same
Parameter ρ , Efficiency	The method has the same efficiency for any value of ρ
Individual Machine, Part Weight	The average weight of the parts produced by all machines is the same

Of course, determining the right hypothesis is a bit of an art. It requires practice, knowledge of your own research, and reflection.

Review of Last Lecture (3)

Testing the hypothesis

After we define the hypothesis, we will have enough information to run a statistical test. The statistical test will usually get us:

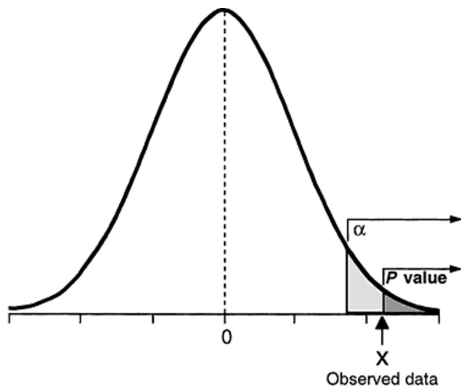
- Estimated value for the parameter of interest;
- Confidence interval for the parameter of interest;
- **p-value** for the alternate hypothesis;
- Other information;

P-value

Probability that, given the null hypothesis, the experiment result would be that much, or higher.

If the **p-value** is low enough (depends on the problem, 0.05 is standard), you can **reject the null hypothesis**

Review of Last Lecture (4)



α value: Parameter of the testing procedure that decides the desired accuracy.

p-value: Probability that, given the null hypothesis, the experiment result would be that much or higher.

Image from Steven N. Goodman, "Towards Evidence Based Medical Statistics"

Example of Hypothesis Testing

Algorithm Analysis

A new algorithm (GA) is used for modeling. We hope that this algorithm produces better models than the (RI) algorithm, as measured by their log-likelihood.

- **Null Hypothesis:** Log likelihood of the model is independent of the method used (for methods GA and RI)
- **Alternate Hypothesis:** Log likelihood of models generated by GA are higher than the value generated by RI models.

Scenario	Log Likelihood		p-value
	RI	GA	
2005	-2263.4	-2253.2 (16.5)	0.01
2006	-2252.28	-2234.72 (14)	0.01
2007	-2113.84	-2108.95 (11.1)	0.03
2008	-2110.79	-2096.75 (11.8)	0.01
2009	-2487.88	-2482.88 (10.3)	0.02
2010	-2132.11	-2099.13 (16.3)	0.01
2011	-20083.09	-19983.73 (144.4)	0.01
2012	-3225.39	-4435.34 (248)	1.00

Can we start with new material?

Types of Errors

In the last class, we defined two types of errors:

- **Type I errors**: rejecting the null hypothesis when it should not be rejected (false positive)
- **Type II errors**: not rejecting the null hypothesis when it should be rejected (false negative)

A type I error depends only on the significance level α , which is a parameter we control;

The probability of a type II error is defined by the *power* of a test, defined by β .

How to control for type II error?

The power of a test is defined by 4 factors

- 1 Actual size of the difference
- 2 Variability of the observations
- 3 Significance level
- 4 Sample size

Minimally interesting effect

One way to estimate a lower bound for the power of a test is by defining a *minimally interesting effect* δ^*

It is essential to have a good understanding of the problem in order to define this value.

Example of power calculation

Suppose that on the green peas example one is really interested in detecting deviations from the nominal value greater than 1%, i.e., $\delta^* = 0.01 * 500 = 5g$. The researcher defines that, for this minimally interesting effect, a test power of 0.85 is desired. The test will again be performed with $\alpha = 0.01$.

The same sample of $n = 10$ packs is used. The estimated standard deviation for this sample is $s = 6.97g$. From this data, we can compute the power of this test as:

```
> s<-sd(sample)
> power.t.test(n=10, delta=5, sd=s, sig.level=0.01,
+             type = "one.sample", alternative = "one.sided")
```

One-sample t test power calculation

```
n = 10
delta = 5
sd = 6.970382
sig.level = 0.01
power = 0.3474724
alternative = one.sided
```

Example of sample size calculation

What is the smallest sample size needed to obtain the desired power of 0.85?

```
> power.t.test(power=0.85, delta=5, sd=s, sig.level=0.01,  
  type = "one.sample", alternative = "one.sided")
```

One-sample t test power calculation

$n = 24.76091$

$\delta = 5$

$sd = 6.970382$

$\text{sig.level} = 0.01$

$\text{power} = 0.85$

$\text{alternative} = \text{one.sided}$

We need at least 25 observations to detect a $-5g$ (1%) or larger deviation on the mean weight of the green peas packages with a power level of 0.85.

Hypothesis Testing Types

- Last week we saw a simple example of statistical testing calculation (t-test)
- The t-test makes many assumptions about the data (normal distribution, known variance, etc)
- To break these assumptions, we need slightly different statistical methods.

Reference: Peter Dalgaard, “Introductory Statistics with R” (Chapters 5,7,8)

One Sample t-test

- Assumes that the data comes from the **normal** distribution $N(\mu, \sigma)$;
- We estimate μ and σ from the mean and standard deviation of the sample;
- We compare the μ of the sample with a single value μ_0 . This is the null hypothesis.
- The t statistic is given by:

$$t = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$$

- The t value is compared to the critical regions in the Normal distribution to define the *p value*;

One Sample t-test – example

```
> install.packages("ISwR")
> library(ISwR)
> attach(intake); pre
[1] 5260 5470 5640 6180 6390 6515 6805 7515 7515 8230 8770

> t.test(pre,mu=7725)
```

One Sample t-test

```
data:  pre
t = -2.8208, df = 10, p-value = 0.01814
alternative hypothesis: true mean is not equal to 7725
95 percent confidence interval:
 5986.348 7520.925
sample estimates:
mean of x
 6753.636
```

One Sample t-test – arguments

- `t.test(data,mu=9000,alternative="greater")`
Does a one-tailed test (h_a is true mean is greater than h_0);
- `t.test(data,mu=9000,alternative="less")`
- `t.test(data,mu=9000,conf.level=0.99)` Changes the required confidence level to 99%;

Assumptions for the one sample t-test

The t-test (and the z-test as well) make a few assumptions about the nature of the data:

- The distribution of the population is approximately normal;
- Independence of Residuals (absence of outside factors);
- Equal variance (in case of multiple samples, see later);

We can test these assumptions in two ways:

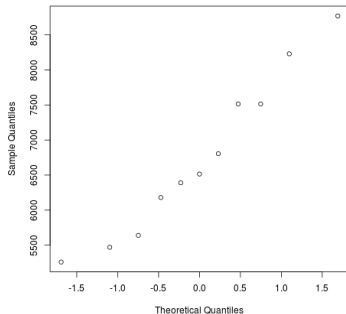
- Graphical/Qualitative Tests;
- Analytical/Quantitative Tests;

Normality Assumption (1)

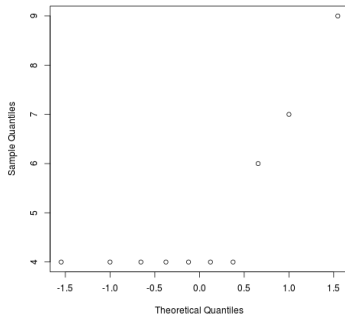
The QQ plot is a great tool to verify if your data follows a certain distribution.

```
> qqnorm(pre)
> rpois(10,5)
[1] 10 10 2 5 5 4 7 7 4 4
> qqnorm(rpois(10,5))
```

Normal Q-Q Plot



Normal Q-Q Plot



Normality Assumption (2)

There is a large number of statistical methods for normality testing. We saw previously the NK test. Another quite used test is the Lilliefors test. Even though the Lilliefors test is possibly the most widely used for normality testing, the Shapiro-Wilk test tends to be more sensitive and has some other interesting properties, so we will be using it throughout the course.

```
> shapiro.test(sample)
```

```
Shapiro-Wilk normality test  
data:  sample  
W = 0.8809, p-value = 0.1335
```

Summary of part I

- Look at your data, and formulate the null and alternate hypothesis;
- The t.test is normally used for comparison of a single sample against pontual values;
- The t.test assumes that
 - the samples are independent,
 - the population is normally distributed,
 - multiple samples have the same variance
- You need to test for each of these assumptions, or risk getting imprecise, or wrong, results.

Any questions?

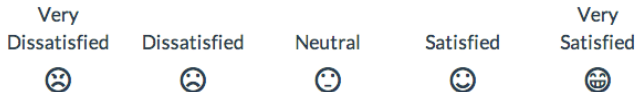
Non-normal distributions

We said that the t-test **assumes that the distribution is normal**. So what do we do if we find out that our distribution is non-normal?

Actually, most cases of non-normality can be solved with judicial application of the **Central Limit Theorem**. Do you still remember it?

Still, for some cases, that is not enough. This is specially true for data that has **no clear scale**. So what do we do?

How did you feel about your most recent experience?



Wilcoxon Signed-Rank Test (1)

Non-parametric Tests

Non-parametric tests ignore the absolute differences between data, and consider relative differences;

For example, the Wilcoxon test calculates the ranks of each sample (best, second, third...). The statistical test is made on these rank values.

Wilcoxon Signed-Rank Test (2)

Consider the following data. We want to compare this data against a h_0 of 7720.

Sample	Difference	Rank
5260	-2460	-8
5470	-2250	-7
5640	-2080	-6
6180	-1540	-5
6390	-1330	-4
6515	-1205	-3
6805	-915	-2
7515	-205	-0.5
7515	-205	-0.5
8230	510	1.0
8770	1050	2.0

- The ranks are the values of each observation against h_0 .
- Two observations with the same value share the same rank.

Wilcox test example:

```
> install.packages("ISwR")
> library(ISwR)
> attach(intake); pre
[1] 5260 5470 5640 6180 6390 6515 6805 7515 7515 8230 8770

> wilcox.test(pre, mu=7725)
```

Wilcoxon signed rank test with continuity correction

```
data: pre
V = 8, p-value = 0.0293
alternative hypothesis: true location is not equal to 7725
```

Warning message:

```
In wilcox.test.default(pre, mu = 7725) :
cannot compute exact p-value with ties
```


Two sample tests

We want to compare two samples coming from different experiments or treatments.

- The mathematics behind two-sample statistical testing is similar to that of the one-sample test.

$$t = \frac{\bar{\mu}_1 - \bar{\mu}_2}{SEDM}$$

- The two-sample tests may be done with the assumption of equal variance, or without it. R by default assumes different variances;
- The wilcox non-parametric test for two samples is treated in a similar manner;

Two sample test – Examples (1)

```
> install.packages("ISwR")
> library(ISwR)
> attach(energy)
> energy
```

	expend	stature
1	9.21	obese
2	7.53	lean
3	7.48	lean
(...)		
20	7.58	lean
21	9.19	obese
22	8.11	lean

Two sample test – Examples (2)

```
> t.test(expend~stature)
```

Welch Two Sample t-test

```
data: expend by stature
```

```
t = -3.8555, df = 15.919, p-value = 0.001411
```

```
alternative hypothesis: true difference in means is not  
                        equal to 0
```

```
95 percent confidence interval:
```

```
-3.459167 -1.004081
```

```
sample estimates:
```

```
mean in group lean mean in group obese  
      8.066154      10.297778
```

Remember: `expend ~ stature` indicates that column “`expend`” is explained by column “`stature`”

Two sample test – Examples (3)

```
> wilcox.test(expend~stature)
```

Wilcoxon rank sum test with continuity correction

```
data: expend by stature
```

```
W = 12, p-value = 0.002122
```

```
alternative hypothesis: true location shift is not equal to 0
```

```
Warning message:
```

```
In wilcox.test.default(x = c(7.53, 7.48, 8.08, 8.09, 10.15, 8.4,  
  cannot compute exact p-value with ties
```

What are proportion tests

In some experiments, the result is not reported as a numerical result, or even as levels. Instead, we have only a “success/failure” measure.

- TSP-like problems;
- Finding Global Optimals;
- Prediction Hit Ratios;
- etc;

In these cases, we want to use [proportion tests](#).

Single Proportion Test

- Test is based on a binomial distribution with size N and probability p ;
- The test statistic for $p = p_0$, where x is the number of successes, becomes:

$$u = \frac{x - Np_0}{\sqrt{Np_0(1 - p_0)}}$$

- u is assumed to have a normal distribution with mean 0 and sd 1;

Single Proportion Test (Example)

In a survey, 39 of 215 randomly selected patients in a hospital have asthma. We test the hypothesis of “the proportion of asthmatic people in the hospital is 0.15”.

```
> library(ISwR)
> prop.test(39,215,.15)

1-sample proportions test with
continuity correction

data: 39 out of 215, null probability 0.15
X-squared = 1.425, df = 1, p-value = 0.2326
alternative hypothesis: true p is not equal to 0.15
95 percent confidence interval:
 0.1335937 0.2408799
sample estimates:
           p
0.1813953
```

Don't confuse the proportion p with the p -value!

The End!

(There is more next class)

Paired Tests (1)



How would you test the differences between two types of shoes (A and B)?

- 1 Give some kids the shoe A, and some kids the shoe B
The personality of the kid becomes a factor – more aggressive kids may wear the shoe faster?
- 2 Give each kid one of each shoe: Left shoe A, right shoe B (or random)
In this way each kid will use both shoes equally.

Can we use this information (same kid used both shoes) to improve our statistical testing?

Image from <http://leagueathletics.com/>

Paired Tests (2)

Interface Testing

Imagine you are comparing two HCI technologies using a survey. You ask each surveyed person to compare two systems using a score.

- Scores given by the same person should show the same biases;
- Thus, you want to compare the scores described by the same voters;
- **Important!** Don't forget to change the order of voting!

Paired T-Test

- The Paired T-test is based on the idea of calculating the difference between the samples.
- This difference is then treated as a one-sample test.

```
> attach(intake)
> pre; post; post-pre
[1] 5260 5470 5640 6180 6390 6515 6805 7515 7515 8230 8770
[1] 3910 4220 3885 5160 5645 4680 5265 5975 6790 6900 7335
[1] -1350 -1250 -1755 -1020 -745 -1835 -1540 -1540 -725 -1330
```

Paired T-Test

- The Paired T-test is based on the idea of calculating the difference between the samples.
- This difference is then treated as a one-sample test.

```
> t.test(pre,post,paired=T)
```

```
Paired t-test
```

```
data:  pre and post
t = 11.9414, df = 10, p-value = 3.059e-07
alternative hypothesis: true difference in means
                        is not equal to 0
95 percent confidence interval:
 1074.072 1566.838
sample estimates:
mean of the differences
      1320.455
```

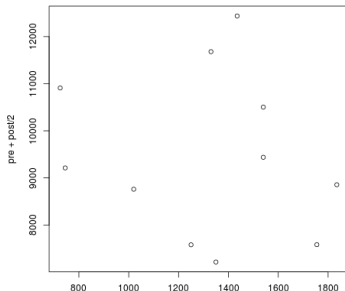
Paired T-Test – Important points 1

- If your data is paired (there is a correspondence between the two samples), you **MUST** use a paired test for correct results;
- Conversely, if your data is **NOT** paired, you should **NOT** use a paired test;

Paired T-Test – Important points 2

- The paired t-test **assumes** that the difference between the samples is **independent of the level**;
- This means that higher values do not have a bigger difference than lower values;
- You can examine this by plotting the difference of the pair against the average of the pair;
- Even if the differences scales with the level, a transformation of the data can be used to fix this problem.

```
plot (pre-post,pre+post/2)
```



Paired Wilcoxon Test

- The basic idea of the Paired Wilcoxon test is similar to the paired T-test;
- The “rank of the differences” is calculated; (so the size of the differences has a smaller effect)

```
> wilcox.test(pre,post,paired=T)
```

```
Wilcoxon signed rank test with continuity correction
```

```
data: pre and post
```

```
V = 66, p-value = 0.00384
```

```
alternative hypothesis: true location shift is  
not equal to 0
```

```
Warning message:
```

```
In wilcox.test.default(pre, post, paired = T) :  
cannot compute exact p-value with ties
```

Analysis of Variance

Introduction

The Analysis of Variance (ANOVA) is a statistical method for testing many samples at the same time.

- We want to find the best values for 4 parameters in our method. Each parameter can be valued between 0.0 and 1.0;
- Can we experiment the parameter values two at a time? Why? Why not?

Analysis of Variance (Model)

Let each observation x_{ij} be the j -th observation in group i . We decompose this observation as:

$$x_{ij} = \bar{x} + (\bar{x}_i - \bar{x}) + (x_{ij} - \bar{x}_i)$$

In other words, each observation is composed by the grand average of all samples, plus the deviation of the group, plus the (independent) error of the sample.

Analysis of Variance (Model)

- The Analysis of Variance starts from the assumption that the variance of all groups is the same;
- If this assumption holds, then we can compare the variance difference between groups to test the null hypothesis that “All groups have the same mean”;

Although it is called Analysis of Variance, and the main calculations are based on the variance of the groups, the final goal is a hypothesis centered on the means of the groups.

Analysis of Variance (Example – 1)

```
> attach(red.cell.folate)
> summary(red.cell.folate)
folate          ventilation
Min.      :206.0    N2O+O2, 24h:8
1st Qu.:249.5    N2O+O2, op  :9
Median :274.0    O2, 24h      :5
Mean     :283.2
3rd Qu.:305.5
Max.     :392.0
```

The red cell folate data set has three different categories: N2O+O2 24h, op, and O2 24h;

Analysis of Variance (Example – 2)

```
> anova(lm(folate~ventilation))
```

Analysis of Variance Table

Response: folate

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
ventilation	2	15516	7757.9	3.7113	0.04359 *
Residuals	19	39716	2090.3		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

- Sum Sq and Mean Sq show the variances calculated for the groups;
- The F statistic shows that the null hypothesis that all groups are the same;

Comparing multiple groups with ANOVA

The F statistic in the Anova test tells us that not all groups are the same. How do we calculate the actual differences between these groups?

```
> summary(lm(folate~ventilation))
Call:
lm(formula = folate ~ ventilation)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-73.625	-35.361	-4.444	35.625	75.375

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)		
(Intercept)	316.62	16.16	19.588	4.65e-14	***	
ventilationN2O+O2,op	-60.18	22.22	-2.709	0.0139	*	
ventilationO2,24h	-38.62	26.06	-1.482	0.1548		

Signif. codes:	0	'***'	0.001	'**'	0.01	'*'
		0.05	'.'	0.1	' '	1

Residual standard error: 45.72 on 19 degrees of freedom

Multiple R-squared: 0.2809, Adjusted R-squared: 0.2052

F-statistic: 3.711 on 2 and 19 DF, p-value: 0.04359

Comparing multiple groups with ANOVA (2)

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	316.62	16.16	19.588	4.65e-14
ventilationN2O+O2, op	-60.18	22.22	-2.709	0.0139
ventilationO2, 24h	-38.62	26.06	-1.482	0.1548

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

- The Intercept is the **mean of the first group** (N2O+O2,24h);
- The Estimates are the **differences** between the corresponding groups and the first;
- The **Test Statistic** (last column), shows the p-value for the alternate hypothesis “The mean of this group is different from the mean of the first group”;
- This version of the test compares only the first group (baseline) against all;

Comparing multiple groups with ANOVA(3)

Comparing All against All

As we increase the number of comparisons, we increase the chance of finding a “significant” result. Therefore, our p-values must be adjusted to compensate for this.

```
> pairwise.t.test(folate,ventilation)
```

Pairwise comparisons using t tests with pooled SD

data: folate and ventilation

N2O+O2,24h N2O+O2,op

N2O+O2,op 0.042 -

O2,24h 0.310 0.408

P value adjustment method: holm

Non-parametric version

The Kruskal-Wallis test is the non-parametric counterpart of the one-way ANOVA:

```
> kruskal.test(folate~ventilation)
```

```
      Kruskal-Wallis rank sum test
```

```
data:  folate by ventilation
```

```
Kruskal-Wallis chi-squared = 4.1852, df = 2, p-value = 0.12
```


Selecting Levels and Parameters

Problem

Your new method has 3 different parameters, that can be set individually. How do you define the values for these parameters during the experiment?

- “Best Guess”
- “Analyse 1 factor at a time”

Other ideas?

Selecting Levels and Parameters

The Latin HyperCube Sampling (LHS)

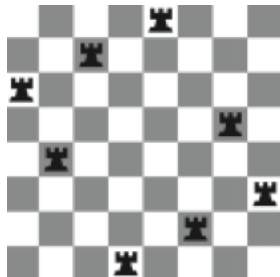
The LHS is a strategy for generating a random, robust parameter set from a large number of dimensions.

- Define the desired number of experimental runs k , and the number of parameter (dimensions) n ;
- For each parameter x_i , divide the range of that parameter into k parts of equal size;
- The range of all n parameters are arranged in a hypercube;
- Choose k sets of parameters so that for each “row” and “column” in the hypercube there is only one sample;

Selecting Levels and Parameters (2)

When to use LHS

- As a initial exploration of the parameter space;
- You are interested in seeing if there is a change in performance, but you don't know where to look;
- You want to show that your model/method is resistant to change in parameters;
- You want to concentrate on one parameter (not included in LHS), but don't care about the others;



When NOT to use LHS

- You want to study the sensibility of one particular parameter;