SMDE FINAL REVIEW

FIRST PART: Probability and Statistics

probability, ANOVA / intro to MANOVA / linear regression / Principal Component Analysis

the p-value is the probability of obtaining a test statistic at least as extreme as the one that was actually observed, assuming that the null hypothesis is true.
"**reject the null hypothesis"** <u>when the p-value</u> turns out to be **less than a certain significance level**, often 0.05[2][3] or 0.01.
Such a result indicates that the observed result would be highly unlikely under the null hypothesis.

* Intro to R

```
options()           <-- set options ( digits, etc),
apropos(table)      <-- see all functions which have the word table in the name
x <-  scan()        <-- handy when you have a column of values, say, stored in a text file, and then you can copy and paste them all at the 1: prompt, and R will store all of the values instantly in the vector x.
seq(from = 2, by = -0.1, length.out = 4)
x[c(1,3,4,8)]
x[-c(1,3,4,8)]
LETTERS and letters vectors
A <- data.frame(v1 = x, v2 = y)   <-- as long as x and y are same length
unique(y[match(as.vector(x), y, 0L)])
prop.table          <-- gives properties of a table variable ( count frequencies, etc )
lots of plots:  eventually look into stemplots(for lists of numbers such as years), indexplots (good for timeseries data!)
```

tips and searchs: http://wiki.r-project.org/rwiki/doku.php and http://www.rseek.org
for graphs: http://addictedtor.free.fr/graphiques/   and http://bm2.genes.nig.ac.jp/RGM2/index.php

- Exercise of Data Input Using R

```
library(DAAG)
with(rainforest, table(complete.cases(root), species))
rainforest[which(rainforest$species == "Acacia mabellae"),]

library(RcmdrPlugin.IPSUR)
data(RcmdrTestDrive)
attach(RcmdrTestDrive)
names(RcmdrTestDrive) # shows names of variables

with(RcmdrTestDrive,table(race))          #make a table of race variable
race
AfricanAmer      Asian   Caucasian    Hispanic      Other
        46          13        73             34          2
dfh = plot(with(RcmdrTestDrive,table(race)))

max(rcd[which(rcd$gender == "Female"),]$salary)             --> 1025.09
sd(rcd[which(rcd$gender == "Female"),]$salary)              --> 130.7053
boxplot(rcd[which(rcd$gender == "Female"),]$salary)

max(rcd[which(rcd$gender == "Male"),]$salary)               --> 1156.16
sd(rcd[which(rcd$gender == "Male"),]$salary)                --> 158.5423
boxplot(rcd[which(rcd$gender == "Male"),]$salary)
```

*   Intro to Probability

Conditional Probablity: P(AlB) = P(A ∩ B) / P(B)
if P(AlB) = P(A) we say that A is independent of B
P(A ∩ B) = P(A) * P(B) if A and B are independent
Bayes Theorem: P(AlB) = P(BlA) * (P(A) / P(B))   <-- gives a relationship to P(AlB) and P(BlA)

random variables:
distribution function: F(x) = P(X < x)
expectation value: E[x] = ∑xi * pi
variance: ∑(xi - u)^2   OR E[X^2] - (E[x])^2
correlation: E[XY] - (E[x] * E[y])

<u>exponential distribution</u>
probabliity density function :   f(x) = λe^(-λx)   <-- curve(exp(1)^(-1.0 * x),xlim=c(0,5))   for λ = 1.0
cumulative distribution func: F(x) = 1 - e^(-λx)   <-- curve(1 - exp(1)^(-1.0 * x),xlim=c(0,5))   for λ = 1.0
E[x] = 1/ λ
V(X) = 1 / λ^2

poisson distribution
probabliity density function :   f(x) = (e^-λ * λ^k) / k!   <-- curve(((exp(1)^(-1)) * ((1)^x))/factorial(x),xlim=c(0,20),ylim=c(0,1))
λ is the expected # of occurrences in an interval
k is the number of occurrences.

-   Exercises of Probability ANOVA SRM using R

Q2: An extremely important concept in queuing theory is the difference between rates and time.
If λ is an arrival rate for the customers of a shop by unit of time, explain why 1/λ is the time between two arrivals.

A2: If λ is an arrival rate of 5 customers per minute, then that means that 1 customer arrives every 20 seconds (1 / 5) so that is the time between arrivals

Q3: Guests arrive following a **Poisson distribution with an average rate of 30 per hour**

a. How many customers arrive per minute?   30 / 1 hr =   30 / 60 minutes = 1 / 2 minutes = .5 customers arriving per minute

b. How many cusomters are expected to arrive within an interval of ten minutes?   .5 customers per minute * 10 minutes = 5 customers.

c. Determine the probability that there are exactly n=0, n=1, n=2 and n=3 arrivals in an interval of 10 minutes? hint use the POISSON (x, average, cumulative, "whether or not" of Excel)

Based on your slide, it seems P(n(T) = k) = (e ^ -λ * λ^k) / k!   where λ is the expected value of occurrences in an interval and k is the number of occurrences.
So for n=0,   P(n = 0) = (e^(-5) * 5^0) /   0!   = .0067
for n=1,   P(n = 1) = (e^(-5) * 5^1)   /   1!   = .03368
for n=2,   P(n = 2) = (e^(-5) * 5^2)   /   2! = .08422
for n=3,   P(n = 3) =   (e^(-5) * 5^3) /   3!   = .1403739

d. What is the probability that there are more than 3 arrivals in an interval of 10 minutes?
This is just 1 minus the summation of P(0) , P(1), P(2), P(3) so 1 - 0.265 = .7349

Q4: Now the service rate is **40 customers per hour** and follows **an exponential distribution**

a. What is the expectation of service time per customer?   2 customers per 3 minutes =   2/3 customers per minute so E[x] = 3/2 minutes per customer

b/c. given that P(0 ≤ x ≤ 1) = e ^ (-μ*a)   - e^(-μ * b)
determine the probability that the service time of a customer is less than or equal to a minute:
P(0 ≤ x ≤ 1) = e ^ (-μ*a)   - e^(-μ * b)   =   e ^ (-3/2 * 0) - e^(-3/2 * 1)   = 0.7768698

d. Calculate the probability that the service time of a customer is between 2 and 5 minutes, less than 4 and more than 3 minutes
P(2 ≤ x ≤ 5) = e ^ (-μ*a)   - e^(-μ * b)   =   e ^ (-3/2 * 2) - e^(-3/2 * 5)   = 0.04923398
P(0 ≤ x ≤ 4) = e ^ (-μ*a)   - e^(-μ * b)   =   e ^ (-3/2 * 0) - e^(-3/2 * 4)   = 0.9975212
P(0 ≤ x ≤ 3) = e ^ (-μ*a)   - e^(-μ * b)   =   e ^ (-3/2 * 0) - e^(-3/2 * 3)   = 0.988891,   1 - 0.988891 = 0.011109

Q5: SKIP

Q6: Suppose a queuing system with two servers, with a time between arrivals following an exponential distribution with an average of 2 hours and a service time that follows an exponential distribution of 2 hours. We know that a customer has arrived at 1:00pm
What is the probability that the number of arrivals between 13:00 and 14:00 is zero? And one? And two or more?

E[x] = 1 customer in 120 minutes   and .5 customers per 60 minutes
P(n(T) = k) = (e ^ -λ * λ^k) / k!
P(n = 0) = (e^(-.5) * .5^0) /   0! = 0.6065307
P(n = 1) = (e^(-.5) * .5^1) /   1! = 0.3032653
Two or more 1 - P(1) + P(0) = .0902

Q7: SKIP

Q8:

P1: makes 25% of chips , 99 % error
P2: makes 75% of chips , 90 % error

what is the probability that a chip selected at random is from P2 if it is error free? use probability tree and bayes
1) P(P2 l -D) = P(P2 && -D) / P(-D)   = 0.75· 0.9 /   0.25· 0.99 + 0.75· 0.9 = .732
2) P(P2 l -D) = P(-D l P2) * P(P2) / P(-D / P1) * P(P1) + P(-D / P2) * P(P2) = .732

1st Exercise: **GOOD FOR DOE ASSIGNMENT 2, PART 1!**

```
library(RcmdrPlugin.IPSUR)
testd = data(RcmdrTestDrive)
attach(testd)
```

does race affect salary? Similar means and variances so no.

```
plot(salary~race,main="Race vs salary",col=heat.colors(2))
tapply(salary,race,summary)
oneway.test(salary~race,data = testd)

        #One-way analysis of means (not assuming equal variances)
        #data:  salary and race
        #F = 0.1132, num df = 4.000, denom df = 7.661, p-value = 0.9741   <--- null hypothesis of means being equal cannot be rejected
#additionally
fit = lm(salary~race)
anova(fit)

        #Analysis of Variance Table
        #Response: salary
        #            Df   Sum Sq        Mean Sq      F value         Pr(>F)
#race              4    13292          3323         0.1469          0.9642
#Residuals       163   3687449        22622
```

t-test if just two

* Random Variables

cumulative distribution function P(x ≤ a) must add to 1. Probability of A and less.   cumulative starts in UPPERCase F(x) , where is probability mass distribution starts in lower p(x)
important distributions:
binomial; geometric; negative binomial; poisson;
E[x] for a discrete (contiunous distribution) = ∑xi * p(x) or (⌠ xi px)
σ^2 = Var(x) = E[(x-μ)^2] = ∑ (xi - μ)^2 * p(x)

* Intro to ANOVA

comparison of two distributions with equal variances:
H0: μA = μB
yA ~ N(μA , σA / √nA)   and   yB ~ N(μB , σB / √nB)
yA - yB ~ N(μA - μB, √ (σA^2)/2 + (σB^2)/2

so for test, (yA - yB - μA - μB) / s * √ 1/n + 1/n > t1-α,n   where n = nA + nB - 2
we reject H0 if this is true

comparison of means in 2 groups with <u>equal variances</u> in R
t.test(formula, dataframe, var.equal=TRUE,alternative) # Normality - Parametric Test
wilcox.test(formula, dataframe) # Non parametric, Wilcoxon test, useful for non normal distributed response data

type 1 error: hypothesis is true and we reject it incorrectly
type 2 error: hypothesis is false and we accept it

i don't get page 12.

comparison of means in 2 groups in R if we can't assume equal variances
t.test(formula, dataframe, var.equal= FALSE,alternative) # Normality - Parametric Test
wilcox.test(formula, dataframe) # by default no equal variance groups is assumed. usefull for non-normal distributed response.

ANOVA - Analysis of Variance
Used with 3 or more groups to test for MEAN DIFFS
We have at least 3 means to test, e.g., H0: μ1 = μ2 = μ3.
Could take them 2 at a time, but really want to test all 3 (or more) at once.
Instead of using a mean difference, we can **use the variance of the group means about the grand mean over all groups**.
Logic is to compare the observed variance among means (observed difference in means in the t-test) to what we would expect to get by chance.

The observations within each sample must be independent -> Durbin Watson test **. .**   dwtest(RegModel.3, alternative = “two.sided”)
The populations from which the samples are selected must be normal. -> Shapiro test **. .**   shapiro.test(residuals(regModel.3))
The populations from which the samples are selected must have equal variances -> Breusch Pagan test   **…**   lmtest::bptest(Regmodel.3)

Comparison of means in k groups with equal variances in R
oneway.test(formula, dataframe, var.equal=TRUE, alternative)   # Normality - Parametric Test
kruskal.test(formula, dataframe) # Non parametric, useful for non normal distributed response data

Variance test in normal 2 groups population:
H0: variance of a = variance of b
Sa^2 / Sb^2 < F na-1, nb-1   , if true, do not reject

var.test(formula, dataframe) # Normality - Parametric Test
fligner.test(formula, dataframe) # Non parametric, useful for non normal distributed response data

Variance test in normal k groups population:
bartlett.test(formula, dataframe) # Normality - Parametric Test
fligner.test(formula, dataframe)   # Non parametric, useful for non normal distributed response data

Example:
mean of A = 25.14,   mean of B = 23.62
sd of A = 1.242,   sd of B = 1.237

if H0: uA = uB,   then: uA - uB / (1.24 * sqrt(1/10 + 1/10)) = 2.74 > t .5, 18 = 1.734   so reject H0
to calculate p-value, P(t18 > 2.74) = .305

SSB = SStreatments is sum of squares between groups.

a = rows
xGM is grand mean
xiM is mean of row i

SSB = sum(i=1 to a) Ni * (xiM - xGM) ^ 2

In ANOVA the variability is estimated by the Mean Square Error, or MSE
The Mean Square Error is a measure of the variability after the group effects have been taken into account (measures variability within group).
MSE = 1 / (N - K) Sum(over i) Sum(over j) (xij - xiM) ^ 2
where xij is the jth observation in the ith group.

We can break the total variance in a study into meaningful pieces that correspond to treatment effects and error. That’s why we call this Analysis of Variance.

Notes on MSE:
If there are only two groups, the MSE is equal to the pooled estimate of variance used in the equal- variance t test.
ANOVA assumes that all the group variances are equal.

Other options should be considered if group variances differ by a factor of 2 or more.

The ANOVA F test is based on the F statistic
F = (SSB / (K –1)) / MSE
where K is the number of groups and N is the total number of observations
Under H0 the F statistic has an “F” distribution, with K-1 and N-K degrees of freedom (N is the total number of observations)

I**MPORTANT** example 60 to 70
1. SSB = ∑ Na(Xa –XG)^2 <-- Xa is row average XG is overall average
2. SSW = ∑ (Xi –Xa)^2 <-- Xi is a column observation for person i , Xa is row average
3. MSE = SSW / (N - K) <-- where K is total groups, N is total observations
4. F = SSB / (K - 1) / MSE

then lookup what F should be. Given K=3 groups, and N=15 total observations
5. use F k-1 , n-k

in example, we calculate 12.5 from 1 - 4, and then we look up 3.89 from F-Table

Review
1. Set alpha (.05).
2. State Null & Alternative
H0: µ1= µ2= µ3
H1: not all µ are =
3. Calculate test statistic: F = 12.5 = (Sa^2 / Sb^2)
4. Determine critical value F.05(2,12) = 3.89
5. Decision rule: If test statistic > critical value, reject H0.
6. Decision: Test is significant (12.5>3.89). ie, rejct H0 so means in population are different.

If the t-test is significant, you have a difference in population means.
If the F-test is significant, you have a difference in population means. But you don’t know where.
With 3 means, could be A=B>C or A>B>C or A>B=C.

ANOVA just says that the means differ, but not which ones. We have to do additional tests to determine.
When are post hoc tests done? As the name implies after an ANOVA
But only after a rejection of the null hypothesis.
Only if there are 3 more treatments; k > 2. If only 2 treatments we can just do a t-test.
Post hoc tests are going to let us go back through our data and compare individual treatments 2 at a time: --> **Bonferroni correction**
see slide 85 of 85 for Least Significant Difference Test

- Lab Session 2 EDA R

```
library(AER)
 data("CPS1985")
 df<-CPS1985
 attach( df )

 # Bivariate analysis: 2 numeric variables
 plot(education,wage,col=as.numeric(ethnicity)+1,main="Wage(Y) vs Education (X) | Race",pch=19)
 legend("topleft",legend=levels(ethnicity),col=2:4,pch=19)

 library(car)
 scatterplot(wage~education|ethnicity,main="Wage(Y) vs Education (X) | Race",smooth=FALSE)

 cor(wage,education,method="spearman")

 # Bivariate analysis: 1 numeric variable and 1 factor - Wage vs Race
 plot(wage~ethnicity,main="Wage(Y) vs Race", col=heat.colors(3),pch=19)
 list<-Boxplot(wage~ethnicity,main="Wage(Y) vs Race", col=heat.colors(3),pch=19)
 df[list,c(1,5)]

 tapply(wage,ethnicity,summary)   # run summary grouping ethnicity columns and getting wage from them

 # Bivariate analysis: 2 factors
 plot(gender~ethnicity,main="Gender(Y) vs Race",col=heat.colors(2))
 ta<-table(gender,ethnicity)
 prop.table(ta,2)

 xtabs(~gender+ethnicity)     #  exact same as table(gender,ethnicity)
```

- Lab Session 34

```
library(car)
data(Duncan)

# Create a new factor: Dicothomy Professional or Not
Duncan$prof<-ifelse(Duncan$type=="prof",1,0)
Duncan$prof<-factor( Duncan$prof, labels=c("NoProf","Prof"))
attach(Duncan)

library(lmtest)
dwtest(prestige~prof)

        Durbin-Watson test
        data:  prestige ~ prof
        DW = 1.2886, p-value = 0.00441
        alternative hypothesis: true autocorrelation is greater than 0

#correlation matrix
cor(Duncan[,c(2:4)],method="spearman")

# Test on means for k=2 groups defined by prof
t.test(prestige~prof, var.equal=TRUE, data=Duncan)
        Two Sample t-test

        data:  prestige by prof
        t = -10.9707, df = 43, p-value = 4.817e-14   <--- low p-value so reject null hypo
        alternative hypothesis: true difference in means is not equal to 0
        mean in group NoProf   mean in group Prof
                 25.85185            80.44444

# Test on variances for k=2 groups defined by prof
var.test(prestige~prof)  # Parametric test: normal data (Y)

# Test on means for k=3 groups defined by type
oneway.test(prestige~type) # Parametric test: normal data (Y)

# Test on variances for k=3 groups defined by type
bartlett.test(prestige~type) # Parametric test: normal data (Y)

library(FactoMineR)
condes(Duncan,4)
catdes(Duncan,1)
```

- SMDE exercises Computational statistical inference -- first page is great for R tests

Normality Test: **shapiro.test()**.
Independence Test of Durbin-Watson: **dwtest(***formula***)**.
Clàssic T-TEST (dicotòmic factor):
 t.test(*formula, dataframe, var.equal=c(TRUE,FALSE),alternative*)
 Non parametric version: ***wilcox.test****(formula, dataframe)*

Library lmtest in R contains most used normality tests.
Use acf() for a more graphic tool. Clàssic T-TEST (dicotòmic factor):

Parametric contrast for the equal mean hypothesis in groups defined by the level of 1 factor:
 ONEWAY – Analysis of Variance for 1 factor: **aov**(*formula, dataframe*) or
 oneway.test(*formula,dataframe,var.equal=c(TRUE,FALSE)). Ex: oneway.test(*Y ~ A)
Non Parametric contrast for the equal mean hypothesis in groups defined by the level of 1 factor:
 ONEWAY – Analysis of Variance for 1 factor: **kruskal.test**(*formula,dataframe,var.equal=c(TRUE,FALSE)). Ex:* kruskal.*test(*Y ~ A)

Correlation test for 2 numeric variables is given in R by:
Parametric version for normal-like variables: **cor**(var1, var2,method="**Pearson**") (default option in R)
Non-parametric version for general variables: **cor**(var1, var2,method="**Spearman**")

Parametric contrasts (assuming normal distribution of Y) for equal dispersion (variance) in groups defined by levels of the studied factor (Y ~ A is the formula parameter):
 Dichotomic Case: **var.test**(*formula,dataframe)*
 Polytomic Case: **bartlett.test**(*formula,dataframe).*
 Breusch Pagan Test: **bptest**(prestige~type) # popular in econometrics
Non Parametric contrasts (normal distribution of Y not required) for equal dispersion (variance) in groups defined by levels of the studied factor (Y ~ A is the formula parameter):
 fligner.test(*formula,dataframe).*
Comparison between individual group means: Provided that F test shows a difference between groups, the question arises of wherein the difference lies.
 Parametric version: **pairwise.t.test**(Y, A) .
 Non-Parametric version: **pairwise.wilcox.test**(Y, A) .

Feature Selection: Let Y be a response numeric variable that has to be described in terms of the rest of variables in data set, either numerical or factors. Which of the variables are associated with response Y?
Profiling: Going a little further, do levels of the factors show mean group values in Y significativelly different to the gross mean?
FactoMineR in R covers Feature Selection and Profiling for target either continuous (condes()) or factors (catdes()). Warning: no missing data should be included as response variable.

R features related to Computational Statistical Inference
Formula equation: Y ~ A+B or Y ~ A*B, where Y is the numeric response variable and A and B are factor (qualitative variables).
plot.factor(formula, dataframe) and plot.design(.) are descriptive tools for graphically assessing how a numeric response variable distributes for each level of considered factors (either dichotomy or polytomic).
Be careful with the default order of factor levels!
 Reorder to simplify interpretation: factor(variable, levels=c(levell1, ..., levelk))
 If factor levels are not meaningful include labels for factor levels: factor(variable, levels=c(level1, ..., levelk),labels=c(name1,…,namek)).

- Lab Session 5
  - anscombe and duncan data (using lm(Column ~ column) and resid() and plotting things
  - SEE **10/1** in smde_r_notes

- SMDE exercises Linear Regression using R
  -slide 4 of 8 in pdf
  - SEE **10/8 notes** in smde_r_notes

```
load("/Users/diego/Documents/UPC-MIRI/semester1/SMDE/labs-and-exercises/Anscombe73raw (1).RData")
cor(anscombe$XC, anscombe$YC)

# calculate the linear correlation coefficient
par(mfrow=c(1,1))
plot(anscombe$XC, anscombe$YC)

# bivariant diagram of original data
anscombe.lmC <- lm(anscombe$YC ~ anscombe$XC, data=anscombe)
summary(anscombe.lmC)

# Calculation of the simple linear model: results of the adjustment
lines(anscombe$XC,anscombe.lmC$fitted.values)
text(x=anscombe$XC,y=anscombe$YC,labels=row.names(anscombe), adj=1)

#Exceeding the Diagonal. Bivariate X vs. Y, straight fit, identifying observations by its id.
par(mfrow=c(2,2))
plot(anscombe.lmC)

# Gràphics of standard diagnosis
par(mfrow=c(1,1))
levC <- hatvalues(anscombe.lmC)
cooC <- cooks.distance(anscombe.lmC)
tresC <- rstudent(anscombe.lmC)
anscombe <- data.frame( anscombe, levC, cooC, tresC )

#Calculate: anchoring factor, dist.cook, Resident Student model C,  keep the columns in the dataframe Anscombe. Then standard plot of residuals vs hii, resid vs Cook,  resid vs. fit,
attributes(anscombe)
plot(anscombe$levC,anscombe$tresC)
text(x=anscombe$levC,y=anscombe$tresC,labels=row.names(anscombe), adj=1)

plot(anscombe$cooC,anscombe$tresC)
text(x=anscombe$cooC,y=anscombe$tresC,labels=row.names(anscombe), adj=1)

plot(anscombe.lmC$fitted.values,anscombe$tresC)
text(x=anscombe.lmC$fitted.values,y=anscombe$tresC,labels=row.names(anscombe ), adj=1)
```

MULTIPLE REGRESSION

$$\boxed{Y_i=\beta_1+\beta_2 X_{2,i}+\cdots+\beta_p X_{p,i}+\varepsilon_i \text{ on } \varepsilon_i\approx N\left(0,\sigma^2\right) \text{independents}}$$

Example: Duncan data on prestige of professions or weight vs height in Davis
Study correlations between numeric variables appearing in the work space.
Explicative variables are **income** and **education**.
Response variable is **prestige**
and we have to propose a multiple regression model to explain the prestige of jobs.

Suggested steps
Correlation matrix in R: cor(duncan1, use="pairwise.complete.obs")
Matrix of 2 by 2 scatterplots.

Forward regression from the nul model with a direction forward option in method step().

```
> duncan1.lm0 <- lm( prestige ~1, data=duncan1)
> summary(duncan1.lm0)
> step(duncan1.lm0, ~income+education, direction="forward", data=duncan1)
```

Backward regression from the model with INCOME+EDUCATION in backward direction option in method step().

```
> duncan1.lm2 <- lm( prestige ~ income+education, data=duncan1)
> summary(duncan1.lm2)
> step(duncan1.lm2, direction="backward",data=duncan1)
```

Use method step(.) in R from the nul model to the maximal model with direction specification "both" (it is the default)

```
> duncan1.lml <- lm( prestige ~income+education, data=duncan1)
> summary(duncan1.lm1)
> duncan1.lm<- step(duncan1.lm1, ~income+education, data=duncan1)
```

Linear correlation between a response variable and explicative variables
might not be significative once some of the explicative variables are already included in the model.

A touch on diagnostics:
* Check outliers in residuals and influent data in the selected model.

* Compute histogram of studentized residuals (rstudent(model)), leverage (hatvalues(model)) and Cook's distance (cooks.distance(model)).
1. R2 and global regression test *H*0:β2 = … = β*p* = 0.
2. Residual analysis:
    - Detection of o*utliers.*
    - Scatterplot of studentitzed residual *vs. Yhat* .
    - Scatterplot of studentitzed residual *vs.* Yhat *vs. Xi* .
    - Detection of *a priori* and *a posterior influent data.*
    - Scatterplot of studentitzed residual *vs. leverage.*
    - Scatterplot of studentitzed residual *vs.* Cook's distance.

Example: weight vs height in Davis

The Davis data frame has 200 rows and 5 columns. The subjects were men and women engaged in regular exercise. There are some missing data.
This data frame contains the following columns:

sex: A factor with levels: F, female; M, male.
weight:Measured weight in kg. ,   height: Measured height in cm.
r_weight : Reported weight in kg.,   r_height : Reported height in cm.

Firstly, we examine the relationship between the reported weight and the actual weight in order to assess how data behaves. Pay attention to outliers.
Secondly, we focus on the classical relationship between weight (Y) and height (X): does a quadratic fit hold? Why?

Suggested steps

- Correlation matrix in R,
- Matrix of 2 by 2 scatterplots.
- Multiple regression weight (Y) vs r_weight (Y). Interpret the regression equation and quality of the fit
- Multiple regression weight (Y) vs height (X). Interpret the regression equation and quality of the fit
- Multiple regression weight (Y) vs poly(height,2) (X). Can you Interpret the regression equation and quality of the fit?

* Intro to General Linear Models (GLM)

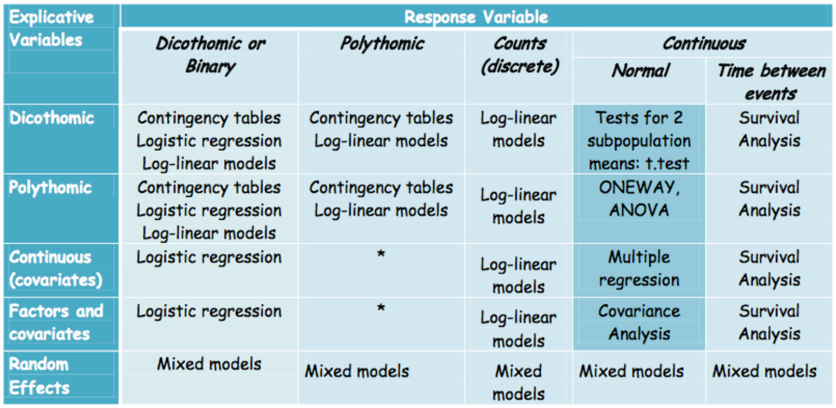
TestScore = B1 + B2(Student-Teacher-Ratio)
The **OLS estimator** minimizes the average squared difference between the actual values of Yi and the prediction (predicted value) based on the estimated line.
returns estimated slope (B2=-2.28) and estimated intercept (B1=698.9)
so estimated regression line = 698.9 - 2.28*STR
then
One of the districts in the data set for which STR = 25 and Test Score = 621
predicted value (using estimated regression line):   = 698.9 – 2.28*25 = 641.9
residual (measured - predicted) :   = 621 – 641.9   = -20.9

The OLS regression line is an estimate, computed using our sample of data; a different sample would have given a different value of β2HAT .

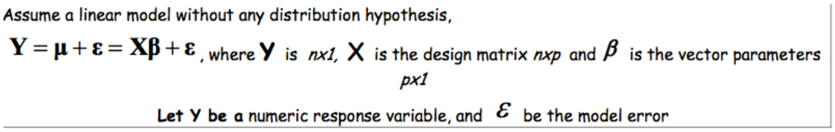
We are going to proceed in four steps:
• The probability framework for linear regression
• Estimation
• HypothesisTesting
• Confidence intervals

A vector-matrix notation for regression elements will be considered since it simplifies the mathematical framework when dealing with several explicative variables (regressors) .

Classification of statistical tools for analysis and modeling



Explicative Variables	Response Variable				
	Dicothomic or Binary	*Polythomic*	*Counts (discrete)*	*Continuous*	
				Normal	*Time between events*
Dicothomic	Contingency tables Logistic regression Log-linear models	Contingency tables Log-linear models	Log-linear models	Tests for 2 subpopulation means: t.test	Survival Analysis
Polythomic	Contingency tables Logistic regression Log-linear models	Contingency tables Log-linear models	Log-linear models	ONEWAY, ANOVA	Survival Analysis
Continuous (covariates)	Logistic regression	*	Log-linear models	Multiple regression	Survival Analysis
Factors and covariates	Logistic regression	*	Log-linear models	Covariance Analysis	Survival Analysis
Random Effects	Mixed models	Mixed models	Mixed models	Mixed models	Mixed models



Assume a linear model without any distribution hypothesis,

$\mathbf{Y} = \boldsymbol{\mu} + \boldsymbol{\varepsilon} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\mathbf{Y}$ is *nx1,* $\mathbf{X}$ is the design matrix *nxp* and β is the vector parameters *px1*

Let Y be a numeric response variable, and ε be the model error

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} == \begin{pmatrix} \underbrace{}_{1} & \overbrace{}^{j=2} & \overbrace{}^{j=3} & \overbrace{}^{\cdots} & \overbrace{}^{j=p} \\ \mathbf{1} & x_{12} & x_{13} & \cdots & x_{1p} \\ \mathbf{1} & x_{22} & x_{23} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathbf{1} & x_{n-1\ldots2} & x_{n-13} & \cdots & x_{n-1p} \\ \mathbf{1} & x_{n2} & x_{n3} & \cdots & x_{np} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_p \end{pmatrix}$$

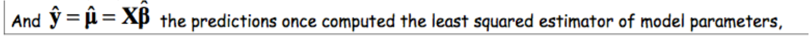
The Ordinary Least Squares estimation of the model parameters β can be written in the general case as,

$$\sum_{i=1\ldots n} \left(Y_k - \mathbf{x}_k^T \beta \right)^2$$

<-- The OLS estimator minimizes the average squared difference between the actual values of Yi and the prediction (predicted value) based on the estimated line.

1. for X, At least as many observations as there are coefficients in the model are needed.
2. The columns of X must not be perfectly linearly related, but even near collinearity can cause statistical problems.

hat - always implies prediction wheras



And $\hat{\mathbf{y}} = \hat{\boldsymbol{\mu}} = \mathbf{X}\hat{\boldsymbol{\beta}}$ the predictions once computed the least squared estimator of model parameters,

slide 15 about properties of H which is the matrix such that Yhat = H * Y (so H transforms Y to predicted values of Y)

Any individual coefficient $\hat{\beta}_j$ is distributed normally with expectation β_j and sampling variance $\mathbf{V}\left(\hat{\beta}_{\mathrm{j}}\right)=\sigma^2\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{T}_{jj}$ and we can test the simple hypothesis (i.e., *make some inference*):

$$H_0: \quad \beta_j=\beta_j^0 \text{ with } Z_0=\frac{\hat{\beta}_j-\beta_j^0}{\sigma\sqrt{\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{T}_{jj}}}\approx N(0,1)$$

But since it does not help so much since β_j and σ^2 are unknown an unbiased estimator of σ^2 is proposed based on the standard error of regression s^2 and to estimate the sample variance of $\hat{\beta}_j$, $\hat{\mathbf{V}}\left(\hat{\beta}_{\mathrm{j}}\right)$.

➡ If the hypothesis hold then the unbiased estimator of σ^2, noted s^2 is efficient (mínimum variance),

$$\boxed{s^2=\frac{\mathbf{e}^{\mathrm{T}}\cdot\mathbf{e}}{n-p}=\frac{\left(\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}}\right)^{\mathrm{T}}\cdot\left(\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}}\right)}{n-p}=\frac{RSS}{n-p}}$$

Residual Sum of Squares

and then

Student t with n-p degrees of freedom (since $\hat{\beta}_j$ and s^2 are independent)

$$t_0=\frac{\beta_j-\beta_j^0}{s\sqrt{\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{T}_{jj}}}\approx Student\ t_{n-p}$$ can be defined and thus $P(H_0)=P(t_{n-p}>t_0)$ computed or a bilateral confidence interval at $100(1-\alpha)\%$ for $\beta_j\in\hat{\beta}_j\pm t_{n-p}^{\alpha/2}SE\left(\hat{\beta}_j\right)$

Inference for Multiple Coefficient will be presented further by F-test

5 HYPOTHESIS TESTS IN MULTIPLE REGRESSION

- If the hypothesis H is true than it can be shown that,

$$F=\frac{\left(RSS_H-RSS\right)/q}{RSS/(n-p)}=\frac{\left(\mathbf{A}\hat{\beta}-\mathbf{c}\right)^{T}\left(\mathbf{A}\left(\mathrm{X}^{T}\mathrm{X}\right)^{-1}\mathbf{A}^{T}\right)^{-1}\left(\mathbf{A}\hat{\beta}-\mathbf{c}\right)}{q\,s^2}\to F_{q,n-p}$$

in R.

Linear.hypothesis() in **car package**: following the previous generic example

```
library(car)
linearHypothesis(model,
    hypothesis.matrix=matrix(c(1,1,0,-4,1,-1,0,0),nrow=2,ncol=4,byrow=TRUE),
    rhs=as.vector(c(2,0)) )
```

Individual confidence interval for β_i in OLS resumes:

$$\boxed{t=\frac{\hat{\beta}_i-\beta_i}{\hat{\sigma}_{\hat{\beta}_i}}\approx t_{n-p} \quad \to \quad \hat{\beta}_i\pm t_{n-p}^{\alpha/2}\hat{\sigma}_{\hat{\beta}_i} \text{ donde } \hat{\sigma}_{\hat{\beta}_i}=s\sqrt{\left(X^TX\right)^{-1}_{ii}} \text{ y } s=\hat{\sigma}=\sqrt{\frac{RSS}{n-p}}}$$

$t_{n-p}^{\alpha/2}$ is the *t de Student* for bilateral confidence interval 1-$\boxed{\alpha}$. Degrees of freedon are (*n-p*) and correspond to the standard error of regression.

Goodness of Fit:

- ***Multiple correlation coefficient R,*** is a goodness of fit measured of a regression model defined as the Pearson correlation coefficient between fitted values $\hat{y}_k$ and observations y_k :

$$R = cor(y, \hat{y})$$

- **The squared of the multiple correlation coefficient R^2 is called the coefficient of determination...**

$$\boxed{R^2 = \frac{\sum_k \left(\hat{y}_k - \bar{\hat{y}}\right)^2}{\sum_k \left(y_k - \bar{y}\right)^2} = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}}$$

1. TSS= $\sum_k (y_k - \bar{y})^2$ where $\bar{y} = \frac{1}{n}\sum_k y_k$ is the mean of the observed response data.
2. ESS= $\sum_k (\hat{y}_k - \bar{y})^2$ and RSS= $\sum_k (y_k - \hat{y}_k)^2$.
3. $\boxed{\textbf{TSS=ESS+RSS}}$, $\sum_k (y_k - \bar{y})^2 = \sum_k (\hat{y}_k - \bar{y})^2 + \sum_k (y_k - \hat{y}_k)^2$.

- The **global test of regression** is a particular case of a multiple contrast of hypothesis where all parameters related to explicative variables are tested to be simultaneously zero.
-

H: $\beta_2 = 0, \ldots, \beta_p = 0$.

$$\boxed{F = \frac{(RSS_H - RSS)/q}{RSS/(n-p)} = \frac{(TSS - RSS)/(p-1)}{RSS/(n-p)} = \frac{ESS/(p-1)}{TSS/(n-p)} = \frac{ESS}{(p-1)s^2} \approx F_{p-1,n-p,}}$$

pg 27 // summary of linear model in R

model validation: Residual Analysis (28 - 34)

Residual analysis constitutes a practical tool for graphically assessing model fitting and satisfaction of optimal hypothesis for OLS estimates:
Residuals are the difference between observed response values and fitted values
scaled residual, ci = ei / s where s is the standard error of regression estimate for the model , ei is the residual of i
standardized residual, di = ci / sqrt(1 - hii) ,

studentized residual, $\boxed{r_i = \frac{e_i}{s_{(i)}\sqrt{1-h_{ii}}}}$ where $s_{(i)}^2 = \frac{(n-p)s^2 - e_i^2/(1-h_{ii})}{n-p-1}$ n - observations and p - groups
Outliers for ri can be detected using t.Student lower and upper bounds for sample size or by univariate descriptive graphical tools as a boxplot.

(34 - 41) anscombe data is a case where there exists **outlier data** which affects the predictor and should be classified as apriori influential (ie, discarded) so as not to be given undue influence.

Multiple regression: we have to think in a cloud of points defined by regressors in **X** (each column in an axis) and center of gravity of those points.

Points **x** ($\mathbf{x} \in \Re^p$) heterogenous regarding the cloud of X points and their center of gravity identify *a priori* influential data.

- The most common measure of **leverage** is the hat - value, h$_i$, the name hat - values results from their calculation based on the fitted values ($\hat{y}_j$): Leverages h_i measure distance from the point $\mathbf{x}_i$ to the center of gravity of the whole set of observation data.
- And thus the average value for the leverage is $\bar{h} = \frac{\sum_i h_{ii}}{n} = \frac{p}{n}$.
- Leverage cut-off: $\boxed{\text{if obs i has } h_{ii} > 2\bar{h} \text{ or } h_{ii} > 3\bar{h} \text{ then is an unusual data}}$.

(40 and 41 one good for outlier, but above my head for the moment and unnecessary for exam)

Best Model Selection (42 - 49)

Model selection should satisfy trade-off between simplicity and goodness of fit, often called **parsimony criteria**.
Available elements to assess the quality of a particular multiple regression **(goodness of fit)** model are: (pg 43) 1. Determination coefficient, *R2* .
2. Stabilibity on the standard error of regression estimate. Estimation of σ^2 by s^2 on underfitting is biased and greater than the true value.
Stability on s^2 confirms or at least points to goodness of fit.
3. Residual analysis.
4. Unusual and influent data analysis
5. calculating Cp, AIC or BIC (in R, AIC(model)) models with lower values are preferred

Stepwise Regression

Backward Elimination is an heuristic strategy to select the best model given a number of regressor and a maximal model built from them.
It is a robust method that supresses non significant terms from the maximal model to the point that all mantained terms are statistically significative and can not be removed. It has been proven to be very effective for polynomial regression.
Stepwise Regression is a strategy that is forward increasing from the starting model, but at each iteration regressor terms are checked for statistical significance.

R software has a sophisticated implementation of these heuristics in the method step(model, target model) based on AIC criteria for model selection at each step.

```
step(duncan1.lm0, ~income+education, direction="forward",data=duncan1)
```

INTRO TO GENERAL LINEAR MODELS (50 - 76)

- Does the relationship between weight on height depend on gender?
- Does profession prestige in Duncan data depend on the type of profession? And after controlling for income and education?
  Both height and prestige are numeric response data and a first random component stated as normal may be assumed leading to OLS estimator.
- Gender is a dicothomic factor (two levels, Male and Female)

- Type of Profession is a polythomic factor consisting in three levels “Blue collar” “White collar” and “professional”.
  How to interpret the R2 ? Exactly as we did in multiple regression
- A high R2 means that the regressors explain the variation in Y.
- A high R2 does not mean that you have eliminated omitted variable bias.
- A high R2 does not mean that the included variables are statistically significant – this must be determined using hypotheses tests.

10 INTRODUCTION TO GENERAL LINEAR MODEL: ONE-WAY ANOVA

The ANOVA model of a factor (generically with I levels)- setting the ideas:

- Formulation and construction of the design matrix for the models of regression,
- Interpretation of its parameters
- Discussion of inference

Group 1	$y_{11}, y_{12}, \cdots, y_{1n_1}$	Mean $\overline{y}_1$
Group 2	$y_{21}, y_{22}, \cdots, y_{2n_2}$	Mean $\overline{y}_2$
...	...	...
Group I	$y_{I1}, y_{I2}, \cdots, y_{In_I}$	Mean $\overline{y}_I$

(1) $Y_{ij} = \mu_i + \varepsilon_{ij}$, I parameters $\boldsymbol{\varepsilon} \approx \mathbf{N}_n\left(\mathbf{0}, \sigma^2 \mathbf{I}\right)$.

(2) $Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$, μ is the overall expected mean and α_i effect for *i* level, I+1 parameters.

The usual null hypothesis is that there are no differences between the expected mean of the groups and can be written according to the formulations as:

(1) $\mathbf{H_0} : \mu_1 = \cdots = \mu_I = \mu$ versus $\mathbf{H_1} : \exists \mu_i = \mu$.

(2) $\mathbf{H_0} : \alpha_1 = \cdots = \alpha_I = 0$ versus $\mathbf{H_1} : \exists \alpha_i \neq 0$.

Prestige of Canadian Occupations in data.frame Prestige in car library for R (Fox and Weisber 2011)

```
library(car)
data(Prestige)
attach(Prestige)
> summary(Prestige)
```

// Default R graphic plot to inspection the relation between a numeric variable (prestige) and a factor (type) works nice

```
> plot(prestige~type, main="prestige vs type",col=3)
```

//Group descriptive statistics and standard procedure for 1 way ANOVA:

```
> tapply(prestige,type,summary)
     $bc
        Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
       17.30   27.10   35.90   35.53   42.60   54.90 
     $prof
        Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
       53.80   61.00   68.40   67.85   72.95   87.20 
     $wc
        Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
       26.50   35.90   41.50   42.24   47.50   67.50 
> oneway.test(prestige~type,var=TRUE)# Corresponds to F-Test
    One-way analysis of means
    F = 109.5916, num df = 2, denom df = 95, p-value < 2.2e-16
> kruskal.test(prestige~type)# Non Parametric version for One-way means test
    Kruskal-Wallis rank sum test Kruskal-Wallis chi-squared = 63.3965, df = 2, p-value = 1.713e-14

> model<-lm(prestige~type, data=Prestige[!is.na(Prestige$type),], contrasts=list(type="contr.treatment"))
> summary(model)
 Call:
 lm(formula = prestige ~ type, data = Prestige[!is.na(Prestige$type), 
     ], contrasts = list(type = "contr.treatment"))

 Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
 (Intercept)   35.527      1.432  24.810  < 2e-16 ***
 typeprof      32.321      2.227  14.511  < 2e-16 ***
 typewc         6.716      2.444   2.748  0.00718 ** 
 ---
 Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

 Residual standard error: 9.499 on 95 degrees of freedom
 Multiple R-squared: 0.6976,     Adjusted R-squared: 0.6913 
 F-statistic: 109.6 on 2 and 95 DF,  p-value: < 2.2e-16
```

$$\boxed{\begin{array}{ll} & i = 1 \equiv bc \quad \hat{y}_{1j} = \overline{y}_1 = \hat{\mu} + \hat{\alpha}_1 = 35.527 + 0 = 35.527 \\ \underbrace{\hat{y}_{ij} = \hat{\mu} + \hat{\alpha}_i}_{\hat{\alpha}_1 = 0} \rightarrow & i = 2 \equiv prof \; \hat{y}_{2j} = \overline{y}_2 = \hat{\mu} + \hat{\alpha}_2 = 35.527 + 32.321 = 67.848 \\ & i = 3 \equiv wc \quad \hat{y}_{3j} = \overline{y}_3 = \hat{\mu} + \hat{\alpha}_3 = 35.527 + 6.716 = 42.244 \end{array}}$$

these right most values are the means for each parameter !!

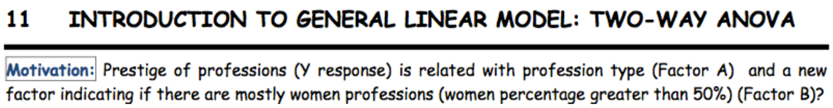
```
> m0 <- lm(prestige~ 1,data=df[!is.na(df$type),])
> m1 <- lm(prestige~ type, data=df[!is.na(df$type), ] )
```

//NOW TO compare null model m0 (with no type) to m1 which contains the type to determine if they are equally
we use the Fisher test anova(m0, m1) where m0 is the big model and m1 is the smaller model contained in the bigger one

```
> anova(m0,m1)
```

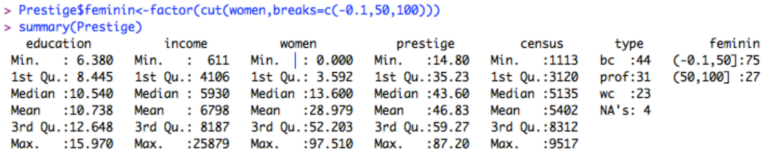
```
Analysis of Variance Table
Model 1: prestige ~ 1
Model 2: prestige ~ type
  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1     97 28346.9
2     95  8571.3  2     19776 109.59 < 2.2e-16 ***
        // null hypothesis is they are equivalent, because p-value is so low, we reject the hypothesis, and say type is in fact important to explaining prestige!!
```

//what would happen if we had a second factor?   considering nesting models, the inference is always the same.



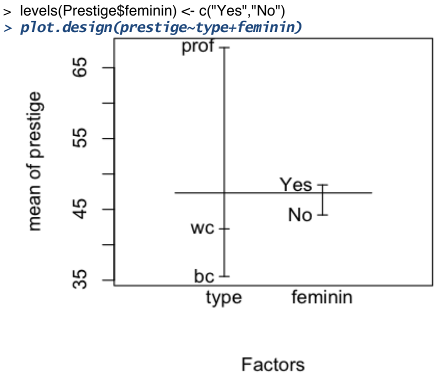
11 INTRODUCTION TO GENERAL LINEAR MODEL: TWO-WAY ANOVA

Motivation: Prestige of professions (Y response) is related with profession type (Factor A) and a new factor indicating if there are mostly women professions (women percentage greater than 50%) (Factor B)?

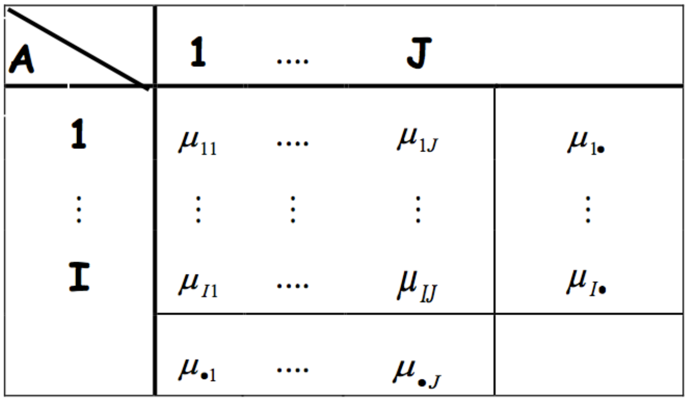


```
> Prestige$feminin<-factor(cut(women,breaks=c(-0.1,50,100)))
> summary(Prestige)
   education          income          women           prestige          census        type        feminin
 Min.   : 6.380   Min.   :  611   Min.   : 0.000   Min.   :14.80   Min.   :1113   bc  :44   (-0.1,50]:75
 1st Qu.: 8.445   1st Qu.: 4106   1st Qu.: 3.592   1st Qu.:35.23   1st Qu.:3120   prof:31   (50,100] :27
 Median :10.540   Median : 5930   Median :13.600   Median :43.60   Median :5135   wc  :23
 Mean   :10.738   Mean   : 6798   Mean   :28.979   Mean   :46.83   Mean   :5402   NA's: 4
 3rd Qu.:12.648   3rd Qu.: 8187   3rd Qu.:52.203   3rd Qu.:59.27   3rd Qu.:8312
 Max.   :15.970   Max.   :25879   Max.   :97.510   Max.   :87.20   Max.   :9517
```

```
> levels(Prestige$feminin) <- c("Yes","No")
> plot.design(prestige~type+feminin)
```



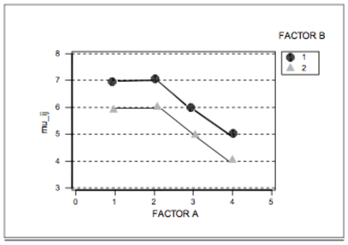
The **analysis of variance of 2 factors** examines
the relationship between **a quantitative response variable**
and **two qualitative explanatory variables** .
The inclusion of the second factor allows the modelling and standardisation of dependence relations and introduces interactions.
Assuming in Two-way ANOVA that population means for each cell in the combinations of the levels of the factors patterns of usual relationship appear clearly.

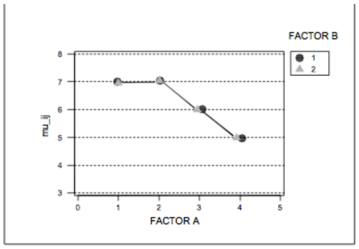


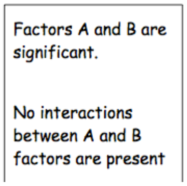
A	1	....	J	
1	μ_{11}	....	μ_{1J}	$\mu_{1\bullet}$
⋮	⋮	⋮	⋮	⋮
I	μ_{I1}	....	μ_{IJ}	$\mu_{I\bullet}$
	$\mu_{\bullet 1}$	....	$\mu_{\bullet J}$	

If Factors A and B do not interact,
then the partial relationship between each factor and the variable of response does not depend on the level of the other factor,
that is, **the difference between levels is constant**.

It is supposed I = 4 and J = 2 in the following diagrams. (slide 59 and 60) <-- i don't complete understand this, maybe ask Lidia

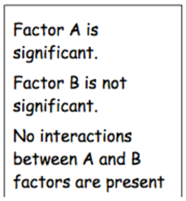






Factors A and B are significant.

No interactions between A and B factors are present

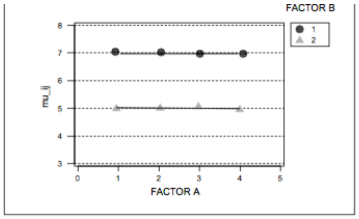


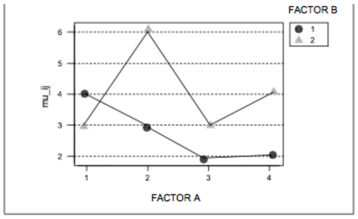
Factor A is significant.

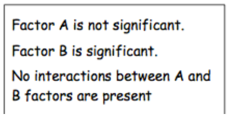
Factor B is not significant.

No interactions between A and B factors are present

IN ALL THESE EXAMPLES THE Y axis is MU_ij



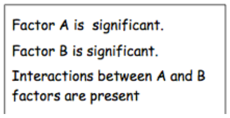




Factor A is not significant.

Factor B is significant.

No interactions between A and B factors are present



Factor A is significant.

Factor B is significant.

Interactions between A and B factors are present

ANCOVA models (multiple general linear regression)

```
>m4  <- lm(prestige~ income+education+women+type, data=df[!is.na(df$type) ,])
>m4i <- lm(prestige~ (income+education+women)*type, data=df[!is.na(df$type) ,])
```

61 - 76 is a little obtuse frankly

-> READ UP ON CONTRASTS!! (what is the difference between slide 55 and 56 in Intro to GLM pdf)

- Lab Session 6

```
//how to get rid of NA's
df<-Prestige[!is.na(Prestige$type),]

#### ANOVA 1- WAY
m1<-lm(prestige~ type,data=df)
summary(m1)
tapply(prestige,type,mean)

m0<-lm(prestige~ 1,data=df)
summary(m0)

anova(m0,m1)
attach(df)
df$femenin<-factor(cut(women,breaks=c(-0.1,30,100)))
summary(df)

# Models Two-Way anova
m11<-lm(prestige~ femenin,data=df[!is.na(df$type),])
m2<-lm(prestige~ type+femenin,data=df[!is.na(df$type),])
m2i<-lm(prestige~ type*femenin,data=df[!is.na(df$type),])

anova(m2,m2i)
anova(m11,m2)
anova(m1,m2)

step(m2i)
```

- SMDE exercises GLM using R

  look at this for DOE HOMEWORK and cause it seems good in general

* Input_Data_Analysis.pdf

  and then a little review this

** then DOE, and a some queuing theory

SECOND PART: Queuing Theory

general structure of queuing models, birth/death processes, generalized q models with non exponential distributions, exponential models in series

example (miri 3b)

The **service of medical emergencies** at a hospital has a doctor on duty permanently.
In spite of this, inappropriate waiting times have been detected and the Management wants to evaluate the benefits of assigning a second doctor to the service.

The arrival rate of patients is one each 30 minutes and
the average time required by the doctor to attend a patient is 20 minutes per.
Evaluate $\rho, P0, Lq, L, Wq$ and W with s=1 y s=2 doctors on duty.
Calculate the probability distribution of the waiting time of a patient until he/she is attended by a doctor in both situations.

for M/M/1
E[t] = 30   E[x] = 20   , p = E[x] / E[t] = 20 / 30 = 2/3   <--- implies λ = 2 and μ = 3   for math (1 / 5) / (1 / 4) = 4 / 5   so (1 / μ) / (1 / λ) = λ / μ
PO = 1 - p = 1 - 2/3 = 1/3
Lq = p^2 / (1 - p) = (2/3)^2 / 1 - 2/3 = 4/3 patients
L = p / 1 - p = 2 patients
W = L / λ = 2 / 2 = 1 hr
Wq = p * W = 2/3 * 1 = 2/3 hr

for M/M/2
p = λ / s * μ = 2 / 2 * 3 = 1 / 3
P0 =

ANOTHER EXAMPLE!

A l**awyer a**ttends their costumers in his office with a capacity of 8 seats for waiting;
a new costumer does not enter if he does not find a seat available.
Inter-arrival time of clients can be considered exponentially distributed with a parameter $\lambda = 20$ **clients** per hour.
The time required by a client is also distributed exponentially with an average of **12 minutes**,
1. How many clients will be attended per hour on average?
2. Which is the average sojourn time of a client on average?

M/M/1/8
λ = 20 clients / hr
μ = 5 clients / hr
p = 20 / 5 = 4

average clients attended per hr

$$\bar{\lambda} = \sum_{n=0}^{\infty} \lambda \cdot P_n$$

average sojourn time
W = L / λhat

A small **airline company** located in the Antilles has a flet size of 5 aircraft.
Each of these aircraft must revisioned each 30days on average.
A staff of two repairmen is available for this task.
Each of them requires 3 days on average in order to carry out a revision.
All these times are random following an exponential law of probabilities.
1. Calculate the average number of aircraft on service.
2. Calculate the average time that an aircraft is out of service due to a revision.
3. Calculate the fraction of time that a repairmen is idle.

M/M/2//5
E[x] = 1 / 30   E[t] = 1 / 3   so p = 3 / 30 = 1 / 10   λ=3 and μ = 30

1. The average number of aircraft on service is the total number N, minus L, the expected number being revisioned and/or waiting for revision.
2. Average time that an aircraft is out of service is its time in W where W = L / λhat
3. fraction of the time that repairman is idle is P0 + .5 * P1

SECOND PART: Prior Exams / Prior Examples
X / Y / s / K / N
- Arrival distribution / Service Distribution / # of servers / Capacity / Source
M exponential, En Erlang, G any

* ZZZZ.pdf

Some repairmen at a workshop are specialist in a special type of engines.
For these engines a sample has been taken of repair times in hours needed for his team of mechanics to repair 4000 engines.
Average values (Mean) and standard deviation (StDev) of the times are listed in the following table:

```
Variable N         Mean Median   TrMean      StDev     SE Mean
t_rep       4000  19,899   16,675   18,736      13,938    0,220
```

The type of **repair requires two separate stages** that need to be carried out consecutively;
both stages are equally **distributed following an exponential distribution**.
A new engine cannot start a new repair until the second stage of the previous repair has not been completed.

The number of engines **arriving to the workshop is distributed following a Poisson distribution with average 1 every 30 hours.**
Also the figure below shows the histogram of repair times in the sample are.
The right part of the figure shows a table with the probabilities of finding N motors in the workshop for N = 0,1,2, ... 9.

N	Prob
0	.333
1	.259
2	.165
3	.099
4	.058
5	.035
6	.021
7	.012
8	.007
9	.004

a. establish a queuing model for the number of engines in the workshop and
calculate the parameters of the probability law for the service time:

M / E2 / 1
The service distribution T (E2) has E[T] = 19.899 hours per car (so μ = 1 / 19.8999)
arrival distribution M with E[M] = 30 hours per car (so λ = 1 / 30)
loading factor: **p = λ * E[T]**
p = λ / μ = 1 car / 30 hours / 1 car / 19.899 hours = .03333 / .05025 = .66328

so l**oading factor** is **how long it takes to service one thing / divided by how long it takes for it arrive** (in above case its 19.899 / 30)

b. At any given time the average number of engines in the workshop is two.
Calculate the probability that the repair time for these two engines exceed 50 hours.
Trep = T1 + T2
Trep follows a k-erlang with k=4;   <--- why cause its two engines which each take two stages
thus E[Trep] = 2 * 19.899 = 39.798

P(T ≥ t) = e ^ (-k * μ * t) * ∑(i=0 to k-1) ((k * μ * t)^i) / i!

μ * k * t = (1 / 39.798) * 4 * 50 = 5.025
P(Trep ≥ 50) = e ^ (-5.025) * ∑(i=0 to 3) (5.025)^i / i!
= .00657 * (5.025 + (5.025^2)/2 + (5.025)^3 / 6)   = .00657 * (5.025 + 12.625 + 21.147) = **0.25497**

c. Calculate the average number of engines in the workshop.
from miri 3c:

Pollaczek-Khintchine's formula provides an approximation for the average occupancy in a queue Lq

Lq = ((σ^2 * λ^2) + ρ^2) / 2(1 - ρ)

The case M/Ek/1:

service times distribute accordingly to an Erlang distribution with parameters k and μ = 1/E[x],

its **variance** is **1/(kμ^2)**

and, when P-K formula is applied: Lq = ((1 + k) / 2k) * (ρ^2 / (1 - ρ))

so σ^2 = 1 / (2 * (1/19.899)^2) = 197.9851
p = 19.899 / 30 = .6633
and then Lq = (197.9851 / 30^2) + (19.899/30)^2) / (2(1 - 19.899/30)
= 0.2199 + 0.4399 / 0.6734 = .9798

and **L = Lq + p** = .9798 + .6633 = 1.6431 engines

d. Calculate the average residence time of an engine in the workshop.

W = L / λ

so W = 1.6431 / 1/30 = 49.28 hours

e. At the time instant at which an engine is sent to the workshop it is known that, at most there are three engines in the workshop.

Calculate the probability that the sojourn time in the workshop by that engine exceeds 30 hours
P(N ≤ 3) = P(0) + P(1) + P(2) + P(3)
= 0.3333 + 0.2592 + 0.1646 + 0.0992 = 0.8563
P(N=0 | N ≤ 3) = .3333 / .8563 = .3892
P(N=1 | N ≤ 3) = .2592 / .8563 = .3027
P(N=2 | N ≤ 3) = .1646 / .8563 = .1992
P(N=3 | N ≤ 3) = .0992 / .8563 = .1158

CURRENTLY HERE

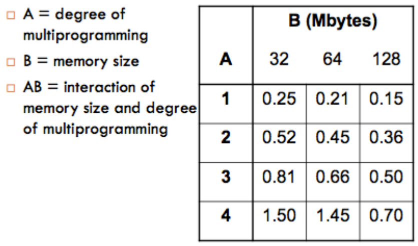
done ex1.pdf

done: ex2.pdf

THIRD PART: **Design of Experiments**

randomized blocks, latin squares and related designs (?), incomplete block design, factorial design

design_of_experiments.pdf



- □ A = degree of multiprogramming
- □ B = memory size
- □ AB = interaction of memory size and degree of multiprogramming

	B (Mbytes)		
A	32	64	128
1	0.25	0.21	0.15
2	0.52	0.45	0.36
3	0.81	0.66	0.50
4	1.50	1.45	0.70

Factor A – *a* input levels
Factor B – *b* input levels
n measurements for each input combination
abn total measurements

One factor ANOVA:

Each individual measurement is composition of: Overall mean, Effect of alternatives, Measurement errors

$y_{ij} = y_{..} + \alpha_i + e_{ij}$

WHERE $y_{..}$ =overall mean, α_i =effect due to A , and e_{ij} =measurement error

Two factor ANOVA

Each individual measurement is composition of: Overall mean, Effects, Measurement errors, and Interactions

$y_{ijk} = y_{...} + \alpha_i + \beta_j + \gamma_{ij} + e_{ijk}$

WHERE $y_{...}$ =overall mean , α_i =effect due to A, β_j =effect due to B

, γ_{ij} =effect due to interaction of A and B , e_{ijk} = measurement error

SUM OF SQUARES

SST = SSA + SSB + SSAB + SSE <-- total = factora + factorb + factorab + error (within each row)

Degrees of freedom:

df(SSA) = *a* – 1 , *df*(SSB) = *b* – 1
df(SSAB) = (*a* – 1)(*b* – 1)
df(SSE) = *ab*(*n* – 1)
df(SST) = *abn* - 1

$$\underbrace{\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{r}\left(Y_{ijk}-\bar{Y}_{...}\right)^2}_{SS_{Total}} = \underbrace{r\cdot b\cdot\sum_{i=1}^{a}\left(\bar{Y}_{i..}-\bar{Y}_{...}\right)^2}_{SS_A} + \underbrace{r\cdot a\cdot\sum_{j=1}^{3}\left(\bar{Y}_{.j.}-\bar{Y}_{...}\right)^2}_{SS_B} + \underbrace{r\times\sum_{i=1}^{a}\sum_{j=1}^{b}\left(\bar{Y}_{ij.}-\bar{Y}_{i..}-\bar{Y}_{.j.}+\bar{Y}_{...}\right)^2}_{SS_{A\times B}} + \underbrace{\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{r}\left(Y_{ijk}-\bar{Y}_{ij.}\right)^2}_{SS_{within}}$$

a = rows (experiments, i indexed)
b = columns (factors, j indexed)
r = replications (k indexed)
$\bar{Y}_{...}$ = overall average

$$MS_{within} = SS_{within} \Big/ df_{within}$$

ANOVA TABLE

Source	Degrees of Freedom	SS	MS	F
A	a-1	SS_A	MS_A	MS_A / MS_{within}
B	b-1	SS_B	MS_B	MS_B / MS_{within}
$A \times B$	(a-1)(b-1)	$SS_{A\times B}$	$MS_{A\times B}$	$MS_{A\times B} / MS_{within}$
Within	ab(r-1)	SS_{within}	MS_{within}	
Total	abr-1	SS_{Total}		

Factorial designs

Take in consideration the interactions.

A effect: $$\frac{A_1B_0 - A_1B_1}{2} - \frac{A_0B_0 - A_0B_1}{2}$$

B effect: $$\frac{A_1B_1 - A_0B_1}{2} - \frac{A_0B_0 - A_1B_0}{2}$$

Controlling “k” factors. (columns)
“l” levels for each factor (“li” levels for the l factor) (values each factor can take)
so $l_1 \cdot l_2 \cdot \ldots \cdot l_k$ experiments
The easiest factorial design is the 2_k with li = 2 ∀i = 1,..,k.

Problem with: *Full factorial design with replication gets huge quick*
Measure system response with all possible input combinations
Replicate each measurement *n* times to determine effect of measurement error
k factors, *v* levels, *n* replications → *n* v^*k* experiments
for example, if k = 5 input factors, *v* = 4 levels, *n* = 3, then → 3(4^5) = 3,072 experiments!

Fractional Factorial Designs: *n*2^k Experiments

Special case of generalized *m*-factor experiments
Restrict each factor to two possible values: High, low / On, off
1) Find factors that have largest impact
2) Full factorial design with only those factors

For n * 2^k Experiments (m factors (columns), 2 levels (so 2^k rows) , n replications)

	A	B	AB	Error
Sum of squares	*SSA*	*SSB*	*SSAB*	*SSE*
Deg freedom	1	1	1	$2^m(n-1)$
Mean square	$s_a^2 = SSA/1$	$s_b^2 = SSB/1$	$s_{ab}^2 = SSAB/1$	$s_e^2 = SSE/[2^m(n-1)]$
Computed *F*	$F_a = s_a^2/s_e^2$	$F_b = s_b^2/s_e^2$	$F_{ab} = s_{ab}^2/s_e^2$	
Tabulated *F*	$F_{[1-\alpha;1,2^m(n-1)]}$	$F_{[1-\alpha;1,2^m(n-1)]}$	$F_{[1-\alpha;1,2^m(n-1)]}$	

SLIDE 28-9, what are CONTRASTS!

n2^m , with m = 2 factors (as 2^2 = 4 experiment rows)

Measurements	**Contrast**		
	w_a	w_b	w_{ab}
y_{AB}	+	+	+
y_{Ab}	+	-	-
y_{aB}	-	+	-
y_{ab}	-	-	+

$$w_A = y_{AB} + y_{Ab} - y_{aB} - y_{ab}$$
$$w_B = y_{AB} - y_{Ab} + y_{aB} - y_{ab}$$
$$w_{AB} = y_{AB} - y_{Ab} - y_{aB} + y_{ab}$$

n2^m, with m = 3 factors (would have 2^3 = 8 experiment rows)

Meas	**Contrast**						
	w_a	w_b	w_c	w_{ab}	w_{ac}	w_{bc}	w_{abc}
y_{abc}	-	-	-	+	+	+	-
y_{Abc}	+	-	-	-	-	+	+
y_{aBc}	-	+	-	-	+	-	+

$$w_{AC} = y_{abc} - y_{Abc} + y_{aBc} - y_{abC} - y_{ABc} + y_{AbC} - y_{aBC} + y_{ABC}$$

for n * 2^m, with m = 3 factors:

$$SSAC = \frac{w_{AC}^2}{2^3 n}$$

df(each effect) = 1, since only two levels measured
SST = SSA + SSB + SSC + SSAB + SSAC + SSBC + SSABC
df(SSE) = (n-1)2^3
then perform ANOVA as before
easily generalizes to m > 3 factors

Important Points

Experimental design is used to:
- Isolate the effects of each input variable.
- Determine the effects of interactions.
- Determine the magnitude of the error
- Obtain maximum information for given effort

Expand 1-factor ANOVA to *k* factors

Use $n2_k$ design to reduce the number of experiments needed

But loses some information, Useful to underline the tendency with economy of experiments.

Yates Algorithm

simplifying the interaction calculus on a 2k factorial design

2k factorial designs:

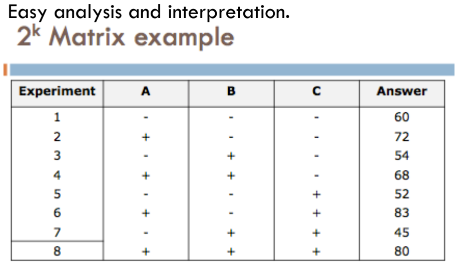
Advantages

Determination of the tendency with experiments economy (smoothness).

Possibility to evolve to composite designs(local exploration).

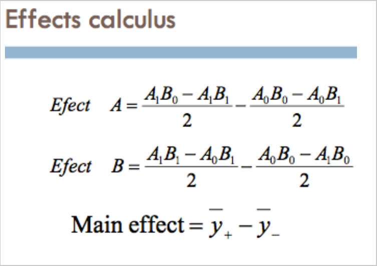
Basis for factorial fractional designs(rapidvision of multiple factors).

Easy analysis and interpretation.



2^k Matrix example

Experiment	A	B	C	Answer
1	-	-	-	60
2	+	-	-	72
3	-	+	-	54
4	+	+	-	68
5	-	-	+	52
6	+	-	+	83
7	-	+	+	45
8	+	+	+	80

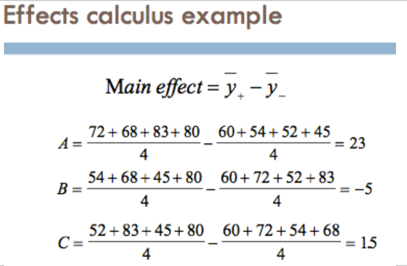


Effects calculus

$$Efect \quad A = \frac{A_1B_0 - A_1B_1}{2} - \frac{A_0B_0 - A_0B_1}{2}$$

$$Efect \quad B = \frac{A_1B_1 - A_0B_1}{2} - \frac{A_0B_0 - A_1B_0}{2}$$

$$\text{Main effect} = \overline{y}_+ - \overline{y}_-$$



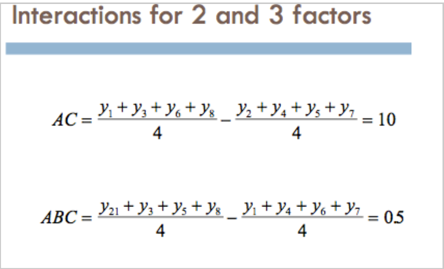
Effects calculus example

$$Main\ effect = \overline{y}_+ - \overline{y}_-$$

$$A = \frac{72+68+83+80}{4} - \frac{60+54+52+45}{4} = 23$$

$$B = \frac{54+68+45+80}{4} - \frac{60+72+52+83}{4} = -5$$

$$C = \frac{52+83+45+80}{4} - \frac{60+72+54+68}{4} = 15$$



Interactions for 2 and 3 factors

$$AC = \frac{y_1+y_3+y_6+y_8}{4} - \frac{y_2+y_4+y_5+y_7}{4} = 10$$

$$ABC = \frac{y_{21}+y_3+y_5+y_8}{4} - \frac{y_1+y_4+y_6+y_7}{4} = 0.5$$

YATES ALGORITHM

·to make systematic the interactions calculus using a table.

1)add the **answer** in the column "i" in the standard form of the matrix of the experimental design.

2)add **auxiliary columns** as factors exists.

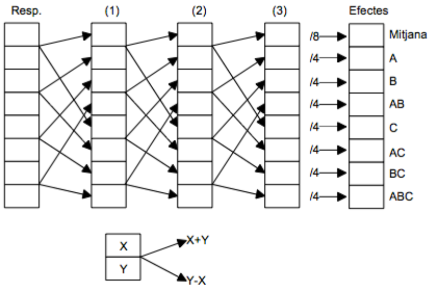
3)add a new column dividing the first value of the last auxiliary column by the number of experimental conditions "E", and the others by the half of "E".

4) in the last column the first value is the mean of the answers, the last values are the effects.

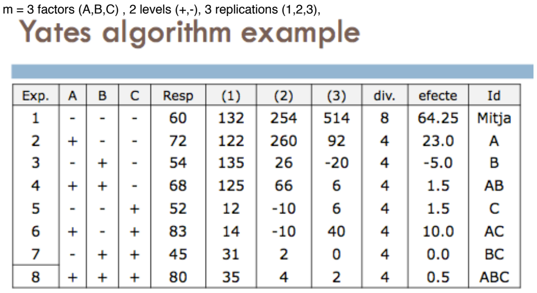
5) the correspondence between the values and effects is done through locating the + values in the corresponding rows of the matrix.

A value with a single + in the B column is representing the principal effect of B.

A row with two + on A and C corresponds to the interaction of AC, etc.



m = 3 factors (A,B,C) , 2 levels (+,-), 3 replications (1,2,3),

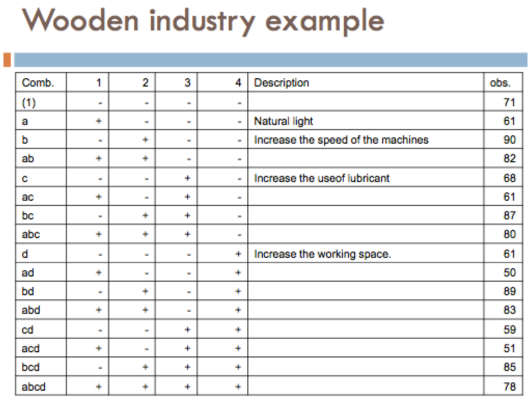


Yates algorithm example

Exp.	A	B	C	Resp	(1)	(2)	(3)	div.	efecte	Id
1	-	-	-	60	132	254	514	8	64.25	Mitja
2	+	-	-	72	122	260	92	4	23.0	A
3	-	+	-	54	135	26	-20	4	-5.0	B
4	+	+	-	68	125	66	6	4	1.5	AB
5	-	-	+	52	12	-10	6	4	1.5	C
6	+	-	+	83	14	-10	40	4	10.0	AC
7	-	+	+	45	31	2	0	4	0.0	BC
8	+	+	+	80	35	4	2	4	0.5	ABC

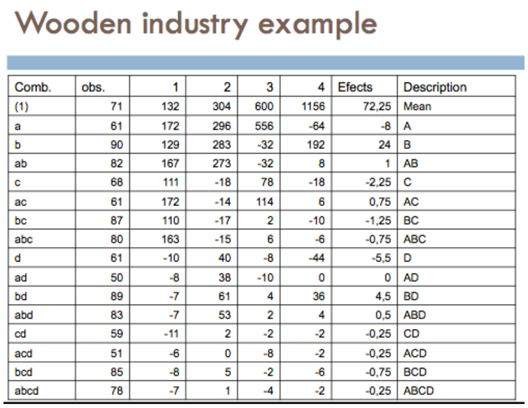
for (1) use Resp = x1 + x2, x3 + x4, x5 + x6, x7 + x8 , x2 - x1, x4 - x3, x6 - x5, x8 - x7
for (2) use (1) = x1 + x2, x3 + x4, x5 + x6, x7 + x8 , x2 - x1, x4 - x3, x6 - x5, x8 - x7
for (3) use (2) = x1 + x2, x3 + x4, x5 + x6, x7 + x8 , x2 - x1, x4 - x3, x6 - x5, x8 - x7
col = n rows for first spot, n/2 for rest
effects = (3) / col
ID = Overall Mean, names of each experiment

ANOTHER EXAMPLE (this time with 4 factors)



Wooden industry example

Comb.	1	2	3	4	Description	obs.
(1)	-	-	-	-		71
a	+	-	-	-	Natural light	61
b	-	+	-	-	Increase the speed of the machines	90
ab	+	+	-	-		82
c	-	-	+	-	Increase the useof lubricant	68
ac	+	-	+	-		61
bc	-	+	+	-		87
abc	+	+	+	-		80
d	-	-	-	+	Increase the working space.	61
ad	+	-	-	+		50
bd	-	+	-	+		89
abd	+	+	-	+		83
cd	-	-	+	+		59
acd	+	-	+	+		51
bcd	-	+	+	+		85
abcd	+	+	+	+		78



Wooden industry example

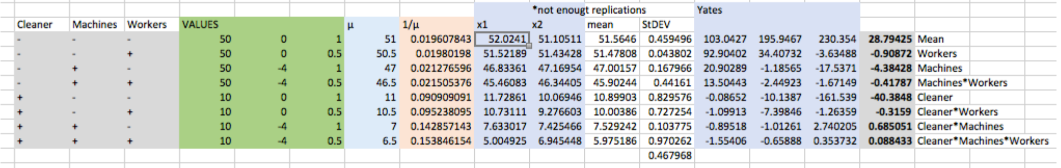
Comb.	obs.	1	2	3	4	Efects	Description
(1)	71	132	304	600	1156	72,25	Mean
a	61	172	296	556	-64	-8	A
b	90	129	283	-32	192	24	B
ab	82	167	273	-32	8	1	AB
c	68	111	-18	78	-18	-2,25	C
ac	61	172	-14	114	6	0,75	AC
bc	87	110	-17	2	-10	-1,25	BC
abc	80	163	-15	6	-6	-0,75	ABC
d	61	-10	40	-8	-44	-5,5	D
ad	50	-8	38	-10	0	0	AD
bd	89	-7	61	4	36	4,5	BD
abd	83	-7	53	2	4	0,5	ABD
cd	59	-11	2	-2	-2	-0,25	CD
acd	51	-6	0	-8	-2	-0,25	ACD
bcd	85	-8	5	-2	-6	-0,75	BCD
abcd	78	-7	1	-4	-2	-0,25	ABCD

CLEAN INDUSTRY EXAMPLE (see excel file!)

We have a system that processes some kind of pieces.
The time needed to process these pieces can be represented by
an **exponential distribution** with a parameter μ that depends on the technology used on the process.
This parameter μ can be calculated depending on several factors that affect it. Each factor adds time to the process:

1) the time needed to clean the pieces by a cleaner machine (range from 10 to 50 seconds).
2) the amount of machines that can be used to glue the different pieces (ranging from 1 to 5)
   -> each machine over 2 reduces the time needed by 1 second.
3) the amount of workers that take the finished pieces (1 or 2),
   -> with one worker the time is 1 second, with two workers its 0,5 seconds.

YATES tab:
CLEANER (- / +) -> (50, 10)
MACHINES(- / +) -> (0 , -4)
WORKERS(- / +) -> (1, .5)



									*not enougt replications			Yates				
Cleaner	Machines	Workers	VALUES			μ	1/μ	x1	x2	mean	StDEV					
-	-	-	50	0	1	51	0.019607843	52.0241	51.10511	51.5646	0.459496	103.0427	195.9467	230.354	**28.79425**	Mean
-	-	+	50	0	0.5	50.5	0.01980198	51.52189	51.43428	51.47808	0.043802	92.90402	34.40732	-3.63488	**-0.90872**	Workers
-	+	-	50	-4	1	47	0.021276596	46.83361	47.16954	47.00157	0.167966	20.90289	-1.18565	-17.5371	**-4.38428**	Machines
-	+	+	50	-4	0.5	46.5	0.021505376	45.46083	46.34405	45.90244	0.44161	13.50443	-2.44923	-1.67149	**-0.41787**	Machines*Workers
+	-	-	10	0	1	11	0.090909091	11.72861	10.06946	10.89903	0.829576	-0.08652	-10.1387	-161.539	**-40.3848**	Cleaner
+	-	+	10	0	0.5	10.5	0.095238095	10.73111	9.276603	10.00386	0.727254	-1.09913	-7.39846	-1.26359	**-0.3159**	Cleaner*Workers
+	+	-	10	-4	1	7	0.142857143	7.633017	7.425466	7.529242	0.103775	-0.89518	-1.01261	2.740205	**0.685051**	Cleaner*Machines
+	+	+	10	-4	0.5	6.5	0.153846154	5.004925	6.945448	5.975186	0.970262	-1.55406	-0.65888	0.353732	**0.088433**	Cleaner*Machines*Workers
											0.467968					

x1 = NORM.S.INV() + μ, x2= NORM.S.INV() + μ, mean = (x1 + x2) / 2; last col = prior / 8 (for top row) and prior / 4 for rest

NORMSINV(p) returns the value z such that, with probability p, a standard normal random variable takes on a value that is less than or equal to z.
A standard normal random variable has mean 0 and standard deviation 1 (and also variance 1 because variance = (standard deviation) squared).

REPLICATIONS TAB

Perform a DOE for the proposed system

- Set the objectives.
- Select the process variables.
- Define an experimental design.
- Execute the design.
- Check that the data are consistent with the experimental assumptions.
- Analyze and interpret the results, detect effects of main factors and interactions.

Replications

Number of replications calculus. Methods to perform the replications.

Experimentation

let x be an interest variable x11,...x1i,...,x1m
x21,...x2i,...,x2m
....
xn1,...xni,...,xnm

n is the number of replications.
xi is the value of each one of the replications.

Sample mean μ

$$\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Sample variance σ2

$$S^2 = \frac{\sum_{i=1}^{n} x_i - \overline{X}}{n-1}$$

Confidence interval

Need to know how far are μ and X
use Student's t-distribution of n-1 df.

$$\overline{X} \pm \mathrm{t}_{1-\alpha/2,n-1}\sqrt{\frac{S^2}{n}}$$

ABOVE THREE ARE IMPORTANT FOR REPLICATION CI calculation!

Example, given:

$\overline{X} = 32.4818$

$S = 3.5149$

n = 10 (chosen at random)

confidence interval: (use S because S = sqrt(s^2))

$$\mathrm{h} = \mathrm{t}_{1-\alpha/2,n-1}\frac{S}{\sqrt{n}}$$

$t_{9,0.975}$ = 2,26
h = 2,512

CI = Xmean ± h
32.4818 - 2.512 = 29.9698,
32.4818 + 2.512 = 34.9938

The interpretation is that with a probability of 0.95,
the random interval (29.9698, 34.9938)
includes the real value of the mean.

More replications needed.

If we specify that we want an interval within 5% of the sample mean with a confidence level of a 95%, we need more replications.
because 0.05·(32.4818) = 1.62 but we have 2.512

Number of needed replications

the next expression is used to determine if the number of replications is enough.

$$n^* = n(\frac{h}{h^*})^2$$

where:
n = initial number of replications.
n^* = total replications needed.
h = half-range of the confidence interval for the initial number of replications.
h^* = half-range of the confidence interval for all the replications (ie, the desired half-range).

so we have n = 10 replications now,
and we want half range, h* = 1.62 around mean, and we have h = 2.512

$$n^* = 10(\frac{2.512}{1.62})^2 = 24.04$$

We now have that we need 25 replications (you have to round up).
After doing that many replications,
imagine we have mean = 32.1094 and variance S = 3.1903

now if calculate h we get
h = t(25-1)(1 - 05 / 2) * S / sqrt(N)
= t(24)(.975) * 3.1903/ sqrt(25) = 2.064 * .638 = 1.317

now this h of 1.317 is less than the 1.62 that we wanted, but that is okay
because it just means we have shrunk the distance around the mean to be x * (32.1094) = 1.317 , so x = .041 % instead of .05

Methods to execute the replications.

Kind of simulations

Finite simulations: Simulations where a condition defines the end of the execution. Usually time.
Non finite simulations: Simulations without this condition.

Independent repetitions

From the same initial state of the model, ie with the same parameterizations and behavior,
only random numbers to be used un the GAV are changed.
These different random number generators (RNG) allow us to test again and again the new system
with the different possible values of the variables that are not controlled (random variables).

Batch means

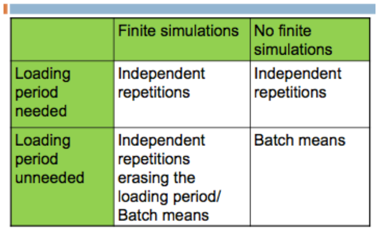
1) Execute a long simulation and then divide it into different blocks, or execution bags.
We work with the mean values of these observations.
Each one of these observations are considered as independent.
2) determine the required length of each one of these execution blocks, to assure the correctness of the experiment.

Regenerative methods

If the variables observed in the execution of the simulation model represent in some way a cyclical restart,
that implies the possible existence of cycles (in the life of the variable). We can consider each one of theses cycles as a replication

This method is not always applicable as it depends on the existence of cycles in the variables.
Also the longitude of this replications must be small; if the longitude of this cycles is big we obtain a small sum of replications.

Applicability



	Finite simulations	No finite simulations
Loading period needed	Independent repetitions	Independent repetitions
Loading period unneeded	Independent repetitions erasing the loading period/ Batch means	Batch means

Variance reduction techniques:

to reduce the number of replications needed

Interest: to reduce the variability introduced in the answer variable due to the use of RNG.
The value that estimates a specific answer variable, represented by its confidence interval, must be adjusted (as possible).

$$(\bar{x} - k\ {}^{s}/_{\sqrt{n}}, \bar{x} + k\ {}^{s}/_{\sqrt{n}})$$

where k = h from before and h = T - test with df N-1, and 1 - alpha/2 where alpha is usually 5 percent .

Obviously, increasing n, the number of observations, decreases the standard error.
Variance reduction techniques try to reduce this variability however without the need for increasing the number of observations.

Antithetic variables

- Use of antithetic values o the random numbers stream used.
- In the first execution the random numbers used can be (a, b, c, ..) ∈ [0,1).
  In the second execution we use it's antithetic values, that means (1-a, 1-b, 1-c, ..) ∈ [0,1).

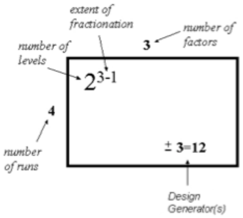
- ▫ Is needed to establish a synchronization method between both streams

Fractional factorial design

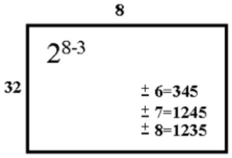
-because there may be still too many experiments with *n*2^*m*

def: a factorial experiment in which only an **adequately chosen fraction** of the treatment combinations required for the complete factorial experiment **is selected to be run.**

Confounding patterns:
1 = 23 (means column one is the multiplation of 2 and 3



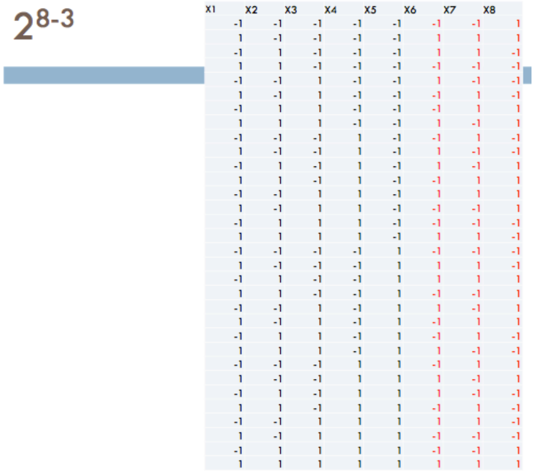
example: How to construct this experiment:



Construct a Fractional Factorial Design From the Specification Above

1) write down a full factorial design in standard order for *k-p* factors (8-3 = 5 factors for the example above).
In the specification above we start with a 2^5 full factorial design. Such a design has 2^5 = 32 rows.
2) add a sixth column to the design table for factor 6, using 6 = 345 (or 6 = -345) to manufacture it
-->(i.e., create the new column by multiplying the indicated old columns together).
3) do likewise for factor 7 and for factor 8, using the appropriate design generators.

The resultant design matrix gives the **32 trial runs** for an 8-factor fractional factorial design!



2^{8-3}

X1	X2	X3	X4	X5	X6	X7	X8
-1	-1	-1	-1	-1	-1	-1	1
1	-1	-1	-1	-1	1	1	1
-1	1	-1	-1	-1	1	1	-1
1	1	-1	-1	-1	-1	-1	-1
-1	-1	1	-1	-1	1	-1	-1
1	-1	1	-1	-1	-1	1	-1
-1	1	1	-1	-1	-1	1	1
1	1	1	-1	-1	1	-1	1
-1	-1	-1	1	-1	-1	1	-1
1	-1	-1	1	-1	1	-1	-1
-1	1	-1	1	-1	1	-1	1
1	1	-1	1	-1	-1	1	1
-1	-1	1	1	-1	1	1	1
1	-1	1	1	-1	-1	-1	1
-1	1	1	1	-1	-1	-1	-1
1	1	1	1	-1	1	1	-1
-1	-1	-1	-1	1	-1	-1	-1
1	-1	-1	-1	1	1	1	-1
-1	1	-1	-1	1	1	1	1
1	1	-1	-1	1	-1	-1	1
-1	-1	1	-1	1	1	-1	1
1	-1	1	-1	1	-1	1	1
-1	1	1	-1	1	-1	1	-1
1	1	1	-1	1	1	-1	-1
-1	-1	-1	1	1	-1	1	1
1	-1	-1	1	1	1	-1	1
-1	1	-1	1	1	1	-1	-1
1	1	-1	1	1	-1	1	-1
-1	-1	1	1	1	1	1	-1
1	-1	1	1	1	-1	-1	-1
-1	1	1	1	1	-1	-1	1
1	1	1	1	1	1	1	1

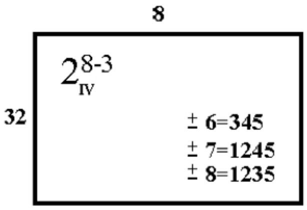
words (slide 110) ?

There are seven "*words*", or strings of numbers, in the defining relation for the 2_{8-3} design,
starting with the original three generators and
adding all the new "words" that can be formed by multiplying together any two or three of these original three words.
These seven turn out to be I = 3456 = 12457 = 12358 = 12367 = 12468 = 3478 = 5678.

in general, there will be (2_p -1) words in the defining relation for a 2_{k-p} fractional factorial.

Resolution

- ▫ The length of the shortest word in the defining relation is called the resolution of the design.
- ▫ Resolution describes the degree to which estimated main effects are confounded (or aliased) with estimated 2-level interactions, 3-level interactions, etc.
- ▫ Resolution is added as a Roman numeral to the experiment definition.



Plackett-Burman designs

- very efficient screening designs when only main effects are of interest.

-Effects of main factors only

-Logically minimal number of experiments to estimate effects of m input parameters (factors)

-Ignores interactions

Requires O(m) experiments, Instead of O(2m) or O(vm)

PB designs exist only in sizes that are multiples of 4

Requires X experiments for m parameters

X = next multiple of 4 ≥ m

PB design matrix

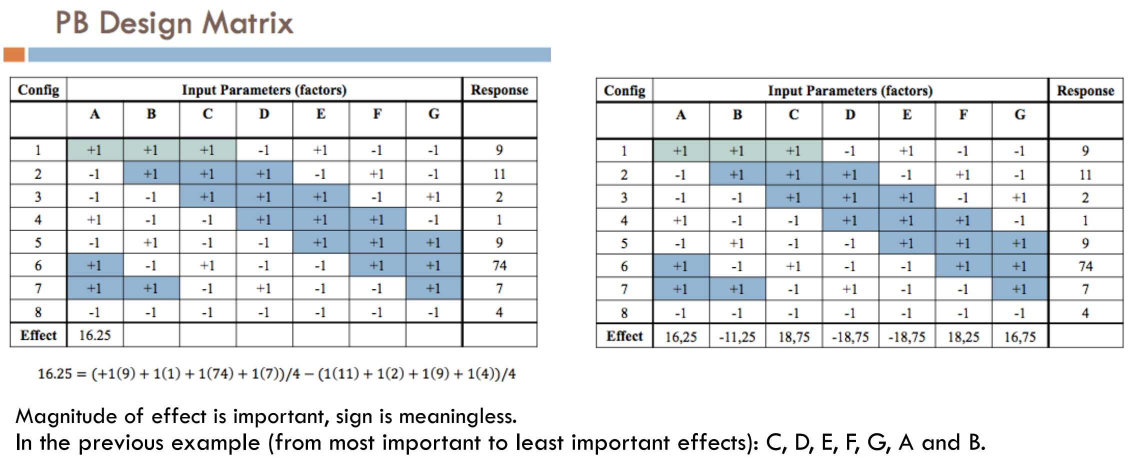
Rows = configurations

Columns = factor's values in each configuration -> High/low = +1/ -1

First row = from P&B paper

Subsequent rows = circular right shift of preceding row

Last row = all (-1)



PB Design Matrix

Config	Input Parameters (factors)							Response
	A	B	C	D	E	F	G	
1	+1	+1	+1	-1	+1	-1	-1	9
2	-1	+1	+1	+1	-1	+1	-1	11
3	-1	-1	+1	+1	+1	-1	+1	2
4	+1	-1	-1	+1	+1	+1	-1	1
5	-1	+1	-1	-1	+1	+1	+1	9
6	+1	-1	+1	-1	-1	+1	+1	74
7	+1	+1	-1	+1	-1	-1	+1	7
8	-1	-1	-1	-1	-1	-1	-1	4
Effect	16.25							

Config	Input Parameters (factors)							Response
	A	B	C	D	E	F	G	
1	+1	+1	+1	-1	+1	-1	-1	9
2	-1	+1	+1	+1	-1	+1	-1	11
3	-1	-1	+1	+1	+1	-1	+1	2
4	+1	-1	-1	+1	+1	+1	-1	1
5	-1	+1	-1	-1	+1	+1	+1	9
6	+1	-1	+1	-1	-1	+1	+1	74
7	+1	+1	-1	+1	-1	-1	+1	7
8	-1	-1	-1	-1	-1	-1	-1	4
Effect	16,25	-11,25	18,75	-18,75	-18,75	18,25	16,75	

$$16.25 = (+1(9) + 1(1) + 1(74) + 1(7))/4 - (1(11) + 1(2) + 1(9) + 1(4))/4$$

Magnitude of effect is important, sign is meaningless.

In the previous example (from most important to least important effects): C, D, E, F, G, A and B.

Assume you don't need to worry about the Randomness.pdf, the Principal Components one or the Bonferroni pdfs for the test.!!

THIRD PART: Prior Exams

MIRI. SMDE. Academic year 2012-13. Q1

1. given MB/time of sample calculate what it would be for 50 MB? (linear regression - see intro GLM slide 7)

n = X*B

1) get mean of MB(x) and mean of time(y)

2) get standard deviation of each row Smb and Stime which is Smb - Mmb and Stime - Mtime for each row

3) then calculate Sxy (where x = Smb , and y = Stime) and calculate (Sx)^2 for each row.

4) then get sums/means for those two columns

5) then divide each sum by N - 1 to get sum/n-1

6) then calculate b1 = sum(Sxy) / sum(Sx2)

in example we get Mmb = 10.7 Mtime = 4.69
sumSxy = 58.67 sumSx2 = 106.1
Sxy/n-1 = 6.51889 and Sx2 / n-1 = 11.7889

we get b1 = .552969

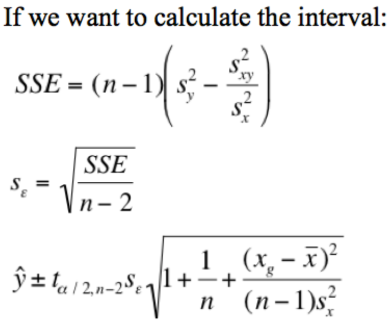
FOR LINEAR REGRESSION: y = b1X + b0

y = .552969x + b0

for sample mean (10.7, 4.69) we get 4.69 = .552969(10.7) + b0
so b0 = -1.226768

so our linear regression equation is y = .552969x - 1.226768

HENCE: for 50mb we get, y = .552969(50)- 1.226768
y = 26.42168 secs



If we want to calculate the interval:

$$SSE = (n-1)\left(s_y^2 - \frac{s_{xy}^2}{s_x^2}\right)$$

$$s_\varepsilon = \sqrt{\frac{SSE}{n-2}}$$

$$\hat{y} \pm t_{\alpha/2,n-2} s_\varepsilon \sqrt{1 + \frac{1}{n} + \frac{(x_g - \bar{x})^2}{(n-1)s_x^2}}$$

2 Queuing theory

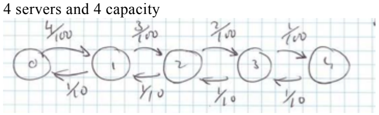
A simple model for a computer system.
An analyst wants to evaluate the performance of a computer system.
In order to make these evaluations, the analyst assumes that the system is composed by a single processor unit and an I/O unit.
The system is intended to run simultaneously a number N of applications, each of them requiring a large number of I/O requests.
The analyst assumes that during a large period of time the N applications are running without finishing.
The time required **to satisfy an I/O request is exponentially distributed with an average of 100 ms.**
The time between **two consecutive I/O requests in any of the applications is exponentially distributed with an average of 10 ms.**
All these times have been measured running a single application on the computer (i.e., with N=1)

The operating system can be configured in two different modes of operation accordingly to a predefined CPU burst.
First mode of operation (infinite or very large CPU burst) once an application has finished its I/O request, it enters into a processor queue waiting for its turn to be processed. When the application reassumes its execution, the processor attends it until a new I/O request happens; then the task goes to the I/O unit and the next application in the queue is then attended by the processor.
Second mode of operation (very small CPU burst). The processor executes a CPU burst (quantum) with a very short time if compared with the time between I/O requests of the applications and the time taken for the change of context can be neglected. All applications in the processor pool are attended following a multitasking operation. The application is given a small CPU quantum (burst), then left, the next application in the processor receives a CPU quantum and so on. If during a CPU quantum an I/O request is found in the application code, then the application goes out of the processor pool and enters in the I/O unit waiting for its I/O request to be completed.
In both modes of operation the applications that are in I/O state can be considered attended in parallel by the I/O unit.
Perform an analysis for N=4 applications.

For the 1st mode of operation:

a) State a queuing model for the number of applications in the processor,
depict the transition's diagram,
and calculate the probability of finding no applications waiting for I/O.
Will the queuing system be in steady state?
b) Evaluate the average number of applications that are waiting in the I/O system for its I/O to be completed.
c) Evaluate the average number of I/O requests per unit of time that the I/O system must satisfy.
d) Evaluate the average time that a task is either in I/O state or waiting to be attended by the processor (sleep).

answer a) so M/M/4/4
Service time ~ Exp with E[x] = 100 ms per io request so μ = 1/100ms
Arrival time ~ Exp with E[τ] = 10 ms per request so λ = 1/10ms
4 servers and 4 capacity



P(n < 4);
= Cn * P0 where P0 = 1 / sum(Cn)
so C(0) = 1, C(1) = 4/100 / 1/10 = .4 , C(2) = 3/100 / 1/10 = .3 , C(3) = 2/100 / 1/10 = .2 , C(4) = 1/100 / 1/10 = .1
P0 = 1 / (1 + .4 + .3 + .2 + .1) = 1/2
and P(4) = C4 * P0 = 1/10 * 1/2 = 1/20

finite diagram = steady state

b) L = sum n* Pn
= 0 * P0 + 1 * P1 + 2 * P2 + 3 * P3 + 4 * P4
= P1 + 2P2 + 3P3 + 4P4
= (C1 * P0) + 2(C2 * P0) + 3(C3 * P0) + 4(C4 * P0)
= (.4 * .5) + 2(.3 * .5) + 3(.2 * .5) + 4(.1 * .5)
= .2 + .3 + .3 + .2 = 1
N - L = 4 - 1 = 3 average number waiting to complete I/O request

c) average number of I/O requests per unit of time
λhat = μhat = P4 * 1/10 + P3 * 1/10 + P2 * 1/10 + P1 * 1/10 = 1/20 requests / ms
d)

For the 2nd mode of operation:

e) Using the general equilibrium equation, depict the transitions diagram of the queuing model for the number of applications in the processor and calculate the probability of finding all the applications in I/O state.
Which of the two modes of operation is more efficient (i.e. makes the applications to be processed in less time)?
f) Repeat e) if the time required for the change of context is not negligible and it is 1/10 of the CPU burst.
(time for processing the application context not included in the CPU burst granted to the application)

3 DOE

We want to determine the effect of machining factors on ceramic strength, our response variables is the ceramic strength.
We have 3 factors and for each factors different values.
Factor 1, the table speed is going from .025 m/s to .125 m/s, a real value.
Factor 2, the down feed rate is going from .05 mm to .125 mm, a real value.
Factor 3, the direction have two levels, longitudinal and transverse.
1) Define a DOE to determine what is the best scenario regarding our response variable.
2) How do you deal with the randomness of the experiment?

3) What are you going to apply to determine the best scenario? Justify your answers.

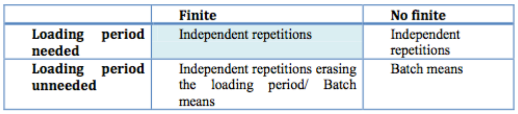
ANSWER:
In that case we need to define a design that constrains the amount of experiments we ca perform, since Factor 1 and Factor 2 are real values.
We propose to define a 2^k factorial design with the next levels for the 3 factors we have.

Factor	Positive	Negative
1	.025 m/s	.125 m/s
2	.05 mm	.125 mm
3	longitudinal	transverse

With this the table we have is composed by 2^3 = 8 experiments as is shown in the next table.

Experiment	Factor 1	Factor 2	Factor 3	Answer
1	-	-	-	?
2	-	-	+	?
3	-	+	-	?
4	-	+	+	?
5	+	-	-	?
6	+	-	+	?
7	+	+	-	?
8	+	+	+	?

Since the answer depends on an experiment that deals with randomness, we need to replicate the scenario.
In this case we are **in a FINITE scenario,** and we want to analyze the loading process,
hence **INDEPENDENT REPETITIONS** will be the best technique to deal with randomness.



	Finite	No finite
Loading period needed	Independent repetitions	Independent repetitions
Loading period unneeded	Independent repetitions erasing the loading period/ Batch means	Batch means

To <u>determine the number of replications needed in each experiment</u> (row) we need to:
1) calculate the half range for each experiment, and
2) the desired half range.

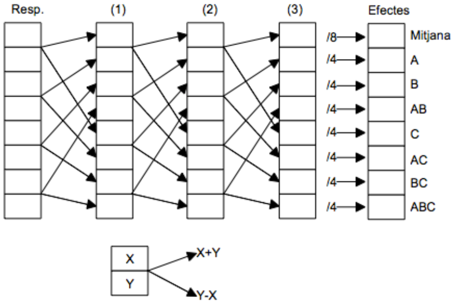
We can apply the next expression to determine if the number of replications is enough.

$$n* = n(\frac{h}{h*})^2$$

where: n = initial number of replications.
$n*$ = total replications needed.
h = half-range of the confidence interval for <u>the initial number</u> of replications.
$h*$ = half-range of the confidence interval for <u>all the replications</u> (ie, the desired half- range).

Once we have the data correctly taken for each scenario we can go further and <u>apply Yates algorithm</u> **to determine the interaction and the effects**.
to do so, we:
1. Add the answer in the column "i" in the standard form of the matrix of the experimental design.
2. Add an auxiliary column as factors exists.
3. Add a new column dividing the first value of the last auxiliary column by the number of experimental conditions "E", and the others by the half of "E".

In the last column the first value is the mean of the answers, the last values are the effects.
The correspondence between the values and effects is done through localize the + values in the corresponding rows of the matrix.
A value with a single + in the B column represents the principal effect of B.
A row with two + on A and C corresponds to the interaction of AC, etc.



Once we have some best candidates,
we can **apply an ANOVA test** to assure that our best alternative is different from others, and makes sense to be applied in the industry.

MIRI. SMDE. Academic year 2012-13. Q2

1. SSB = ∑ Na(Xa –XG)^2 <-- Xa is row average XG is overall average
2. SSW = ∑ (Xi –Xa)^2 <-- Xi is a column observation for person i , Xa is row average
3. MSE = SSW / (N - K) <-- where K is total groups, N is total observations

4. F = SSB / (K - 1) / MSE <-- calculated test statistic
5. use F a,k-1,n-k <-- looked up critical value (a is significance level, k-1 column, n-k row,
Decision rule: If test statistic > critical value, reject H0. HO is that u1 = u2 = u3 = … = uK, for K Groups

SSB = sum N * (each row averages - overall all average) ^2
SSW = sum(each observation in a row - row average) ^2
MSE = SSW/(N-K)
F = SSB / K - 1 / MSE or F = Sa^2 / Sb^2 hmmm.

	Storage Time	Observations	Sum	Average
0	months	58.75 57.94 58.91 56.85 55.21 57.30	344.96	57.49333
1	months	58.87 56.43 56.51 57.67 59.75 58.48	347.71	57.95167
2	months	59.13 60.38 58.01 59.95 59.51 60.34	357.32	59.55333
3	months	62.32 58.76 60.03 59.36 59.61 61.95	362.03	60.33833
			1412.02	58.83417

HO: u0 = u1 = u2 = u3
H1: means unequal

SSB = sum(N * (rowaverage - overall avg)^2)
SSB = (6 * (57.49333 - 58.83417)^2) + (6 * (57.95167 - 58.83417)^2) + (6 * (59.55333 - 58.83417)^2) + (6 * (60.33833 - 58.83417)^2)
= 32.138

SSW = forall rows sum(each observation - row avg)^2
= (58.75 - 57.49333)^2 + (57.94 - 57.49333)^2 + (58.91 - 57.49333)^2 + (56.85 - 57.49333)^2 + (55.21 - 57.49333)^2 + (57.30 - 57.49333)^2
+(58.87 - 57.95167)^2 + (56.43 - 57.95167)^2 + (56.51 - 57.95167)^2 + (57.67 - 57.95167)^2 + (59.75 - 57.95167)^2 + (58.48 - 57.95167)^2
+(59.13 - 59.55333)^2 + (60.38 - 59.55333)^2 + (58.01 - 59.55333)^2 + (59.95 - 59.55333)^2 + (59.51 - 59.55333)^2 + (60.34 - 59.55333)^2
+(62.32 - 60.33833)^2 + (58.76 - 60.33833)^2 + (60.03 - 60.33833)^2 + (59.36 - 60.33833)^2 + (59.61 - 60.33833)^2 + (61.95 - 60.33833)^2
= 9.450533 + 8.829683 + 4.022533 + 10.59828
= 32.90103

MSE = SSW / N - K = 32.90103 / (24 - 4) = 1.645051
F = SSB/ K - 1 / MSE = 32.138 / 3 / 1.645051 = 6.512057 <-- this is our test statistic

now look up, F(a,k-1,n-k) so F(.05,3,20) = so column with 3, and row with 20 at .05 significance = 3.0984 <-- this is our critical value

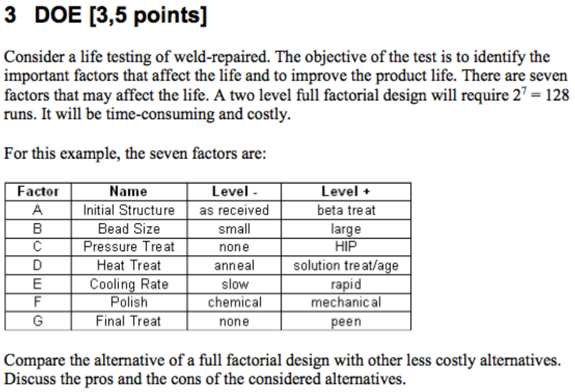
i used this table, http://www.statisticshowto.com/tables/f-table/#f05

since 6.512057 > 3.0984, we reject H0

This looks fine according to your answer, but you get a different pvalue of 0.003. Am I looking it up wrong?

2.

3.



3 DOE [3,5 points]

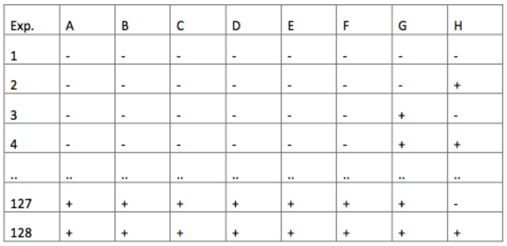
Consider a life testing of weld-repaired. The objective of the test is to identify the important factors that affect the life and to improve the product life. There are seven factors that may affect the life. A two level full factorial design will require $2^7 = 128$ runs. It will be time-consuming and costly.

For this example, the seven factors are:

Factor	Name	Level -	Level +
A	Initial Structure	as received	beta treat
B	Bead Size	small	large
C	Pressure Treat	none	HIP
D	Heat Treat	anneal	solution treat/age
E	Cooling Rate	slow	rapid
F	Polish	chemical	mechanical
G	Final Treat	none	peen

Compare the alternative of a full factorial design with other less costly alternatives. Discuss the pros and the cons of the considered alternatives.

Defining the table for this experimental full factorial 2^7 design we obtain:

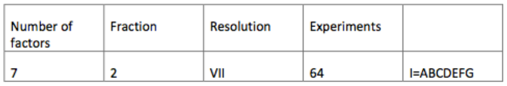


Exp.	A	B	C	D	E	F	G	H
1	-	-	-	-	-	-	-	-
2	-	-	-	-	-	-	-	+
3	-	-	-	-	-	-	+	-
4	-	-	-	-	-	-	+	+
..	..	..	..	..	..	..	..	..
127	+	+	+	+	+	+	+	-
128	+	+	+	+	+	+	+	+

For each one of the different experiments it is needed to calculate the number of replications, and depending on the resources needed for each replication this can be unfeasible.

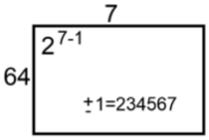
In order to reduce the amount of experiments to be considered two alternatives can be done, a fractional design or a Plackett and Burman (PB) design.

For a **fractional design** it is needed to
1) define the "fraction" of the experiments that are going to be executed, and
2) the confounding factors.
Depending on the desired "resolution" of the experiment we are losing information related to the interaction between the different factors.
The maximum resolution for this example needs 64 experiments and could be defined as follows:



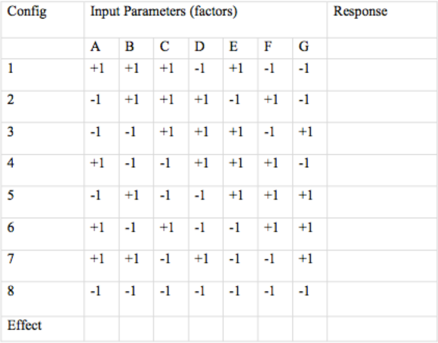
Number of factors	Fraction	Resolution	Experiments	
7	2	VII	64	I=ABCDEFG

Hence:



PRO: we can reduce the number of experiments depending on the desired resolution.
CONS: we lose some interactions information.

For the **Plackett and Burman (PB) design** the table that we obtain is: (how do we get this first row?)



Config	Input Parameters (factors)							Response
	A	B	C	D	E	F	G	
1	+1	+1	+1	-1	+1	-1	-1	
2	-1	+1	+1	+1	-1	+1	-1	
3	-1	-1	+1	+1	+1	-1	+1	
4	+1	-1	-1	+1	+1	+1	-1	
5	-1	+1	-1	-1	+1	+1	+1	
6	+1	-1	+1	-1	-1	+1	+1	
7	+1	+1	-1	+1	-1	-1	+1	
8	-1	-1	-1	-1	-1	-1	-1	
Effect								

PRO: less experiments to be analyzed that in the previous alternative.
CONS: only the main effects are analyzed.