CENTRE DE RECHERCHE PUBLIC

Microarray Center

APPLIED STATISTICS

Lecture 11

Advanced Topics

Petr Nazarov
petr.nazarov@crp-sante.lu

26-11-2009

PRINCIPLE COMPONENT ANALYSIS

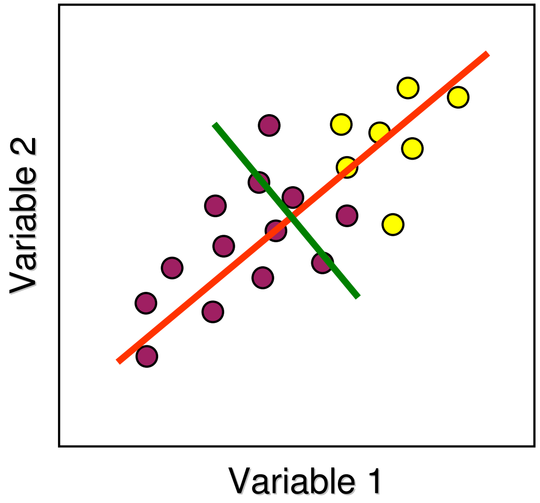
PCA Basics

Principal component analysis (PCA) is a vector space transform used to reduce multidimensional data sets to lower dimensions for analysis. It selects the coordinates along which the variation of the data is bigger.

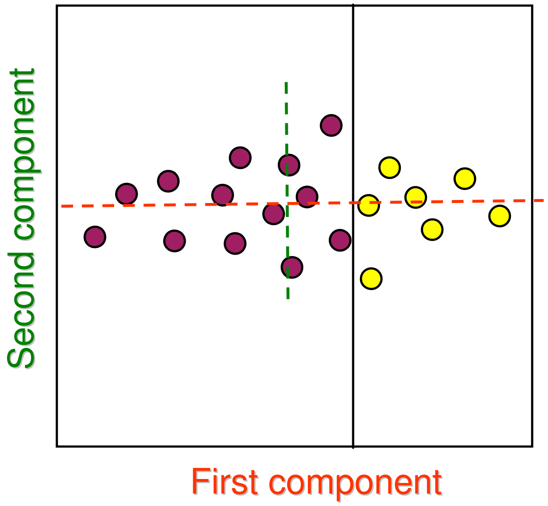
20000 genes → 2 dimensions

For the simplicity let us consider 2 parametric situation both in terms of data and resulting PCA.

Scatter plot in "natural" coordinates



Scatter plot in PC



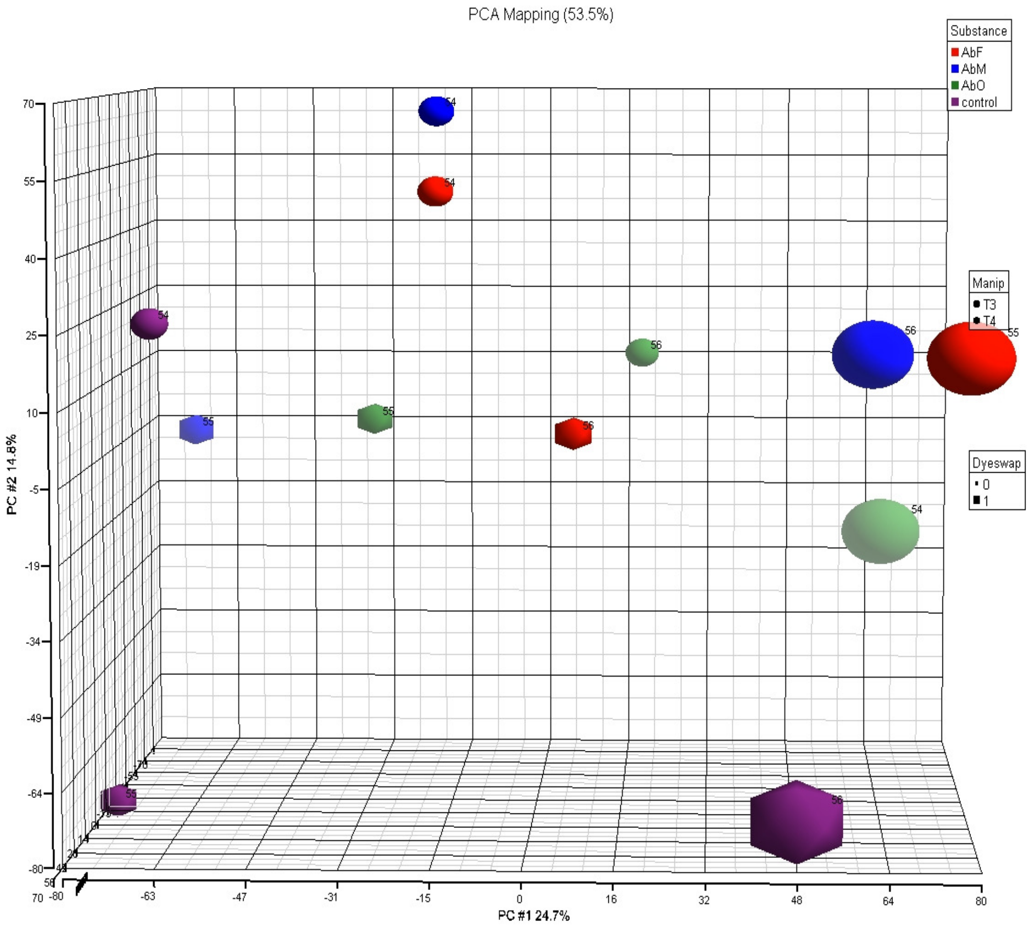
Instead of using 2 "natural" parameters for the classification, we can use the first component!

PRINCIPLE COMPONENT ANALYSIS

PCA Example in Partek

- Transcriptomic profile of a sample contains thousands of genes, i.e. thousands of coordinates/parameters.

- PCA is extremely useful for initial data analysis in transcriptomics, as it allows to depict thousands of parameters just in 2 or 3 dimension space.



3 factors can influence the distribution of the variability:

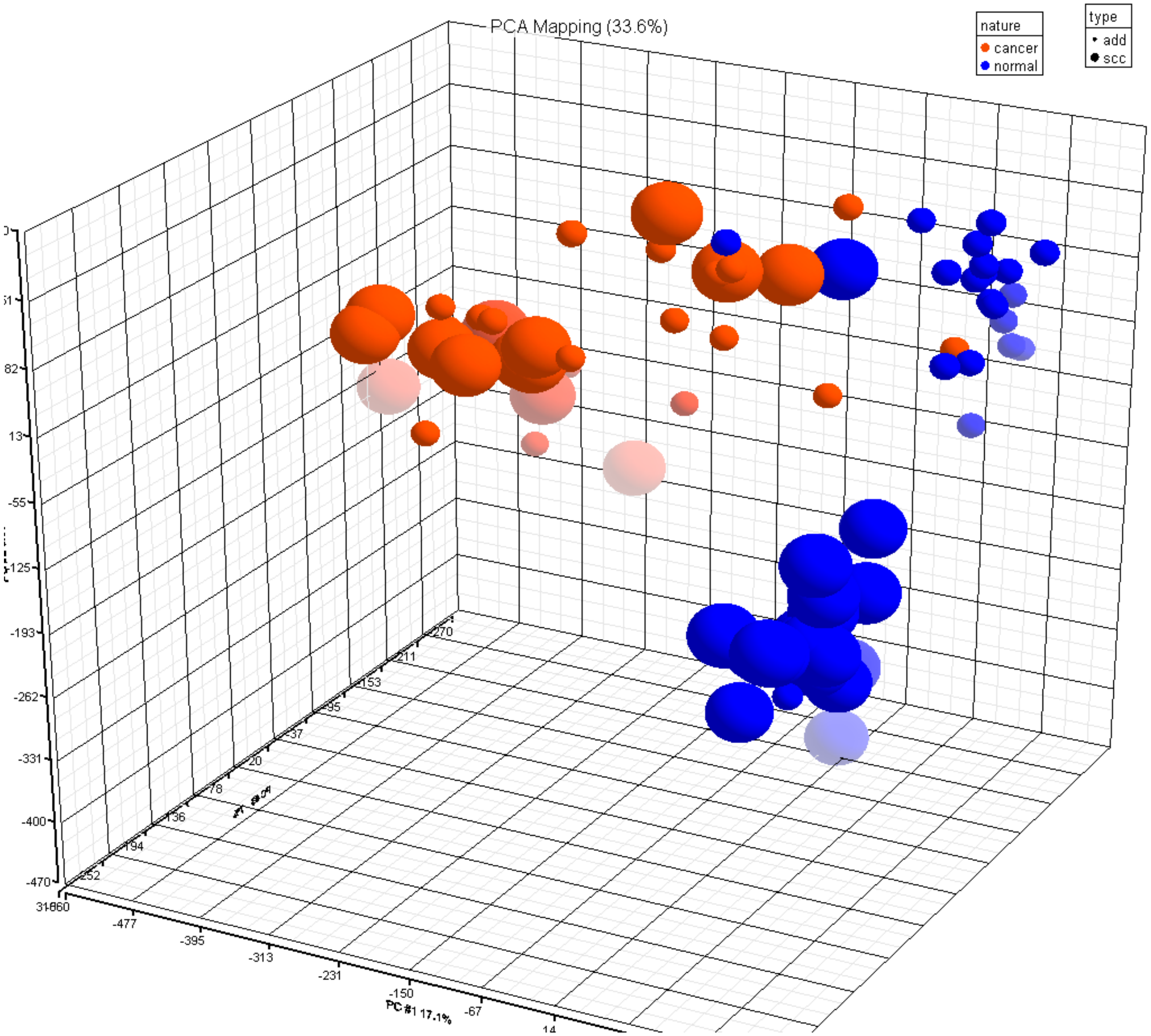
- Substance

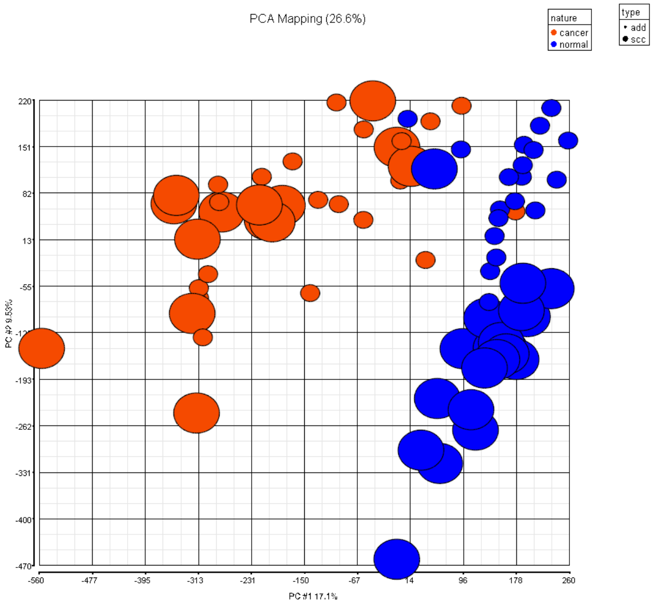
- Manip (bio replicate)

- Dye swap

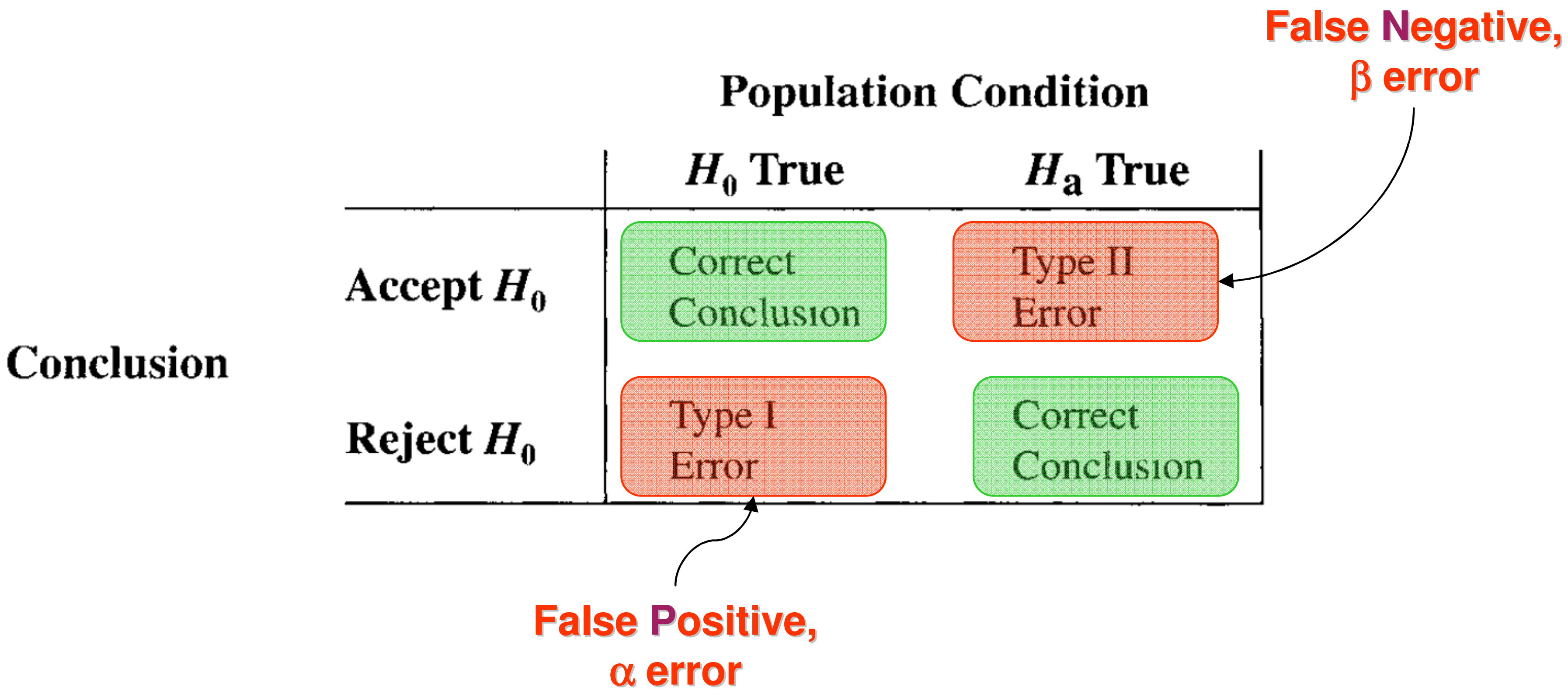
PRINCIPLE COMPONENT ANALYSIS

PCA Example in Partek 2





Correct Results and Errors



		Population Condition	
		H_0 True	H_a True
Conclusion	Accept H_0	Correct Conclusion	Type II Error
	Reject H_0	Type I Error	Correct Conclusion

False Negative, β error

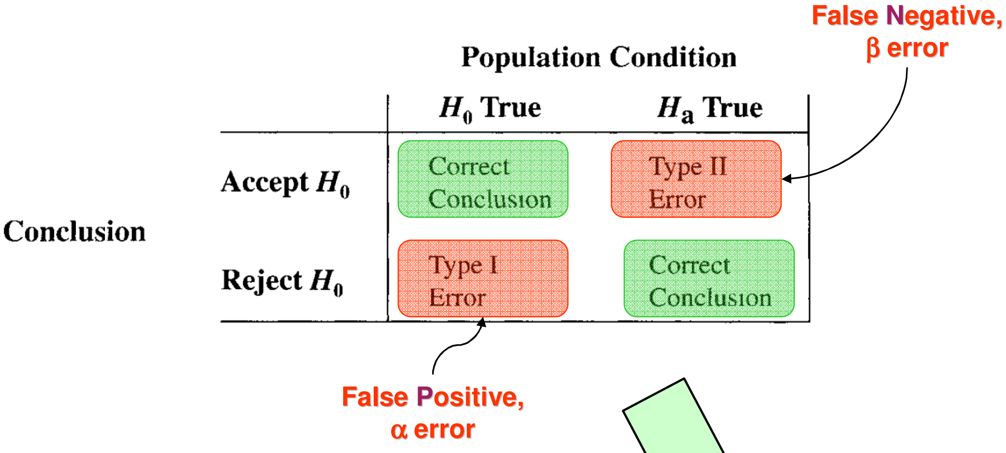
False Positive, α error

Probability of an error in a multiple test:

$$1-(0.95)^{\text{number of comparisons}}$$

MULTIPLE EXPERIMENTS

PCA Example in Partek 2



		Population Condition	
		H_0 True	H_a True
Conclusion	Accept H_0	Correct Conclusion	Type II Error
	Reject H_0	Type I Error	Correct Conclusion

False Negative, β error

False Positive, α error

Familywise error rate (FWER) is the probability of making one or more false discoveries, or type I errors among all the hypotheses when performing multiple pairwise tests.

		Population Condition		
		H_0 is TRUE	H_0 is FALSE	Total
Conclusion	Accept H_0 (non-significant)	$\boldsymbol{U}$	$\boldsymbol{T}$	$m-R$
	Reject H_0 (significant)	$\boldsymbol{V}$	$\boldsymbol{S}$	R
	Total	m_0	$m-m_0$	$\boldsymbol{m}$

$$FWER = 1 - \mathbf{P}(V = 0)$$

CENTRE DE RECHERCHE PUBLIC

MULTIPLE EXPERIMENTS

Bonferroni Correction

Bonferroni correction
if an experimenter is testing ***k*** dependent or independent hypotheses on a set of data, then one way of maintaining the FWER is to test each individual hypothesis at a statistical significance level of **1/*k*** times what it would be if only one hypothesis were tested.

If you would like to be sure **FWER** $< \alpha$, use pairwise testing with significance of α/k.

Assume we need to perform $k = 10$ comparisons, and selected $\alpha = 0.05$. Then

FWER(no correction) = $1–(1-\alpha)^k = 1–(0.95)^{10} = 0.401$

FWER(Bonferroni) = $1–(1-\alpha/n)^k = 1–(0.995)^{10} = 0.0489$

Too conservative
☹

Holm-Bonferroni method
more soft and precise method of significance adjustment.

Holm-Bonferroni Method

Holm-Bonferroni method
more soft and precise method of significance adjustment.

Assume we need to perform k = 6 comparisons, and selected FWER = α = 0.05

Tratments: A, B, C, D

Compare	p-value
A vs B	0.045
A vs C	0.02
A vs D	0.03
B vs C	0.009
B vs D	0.0001
C vs D	0.01

1. Order p-values of the pairwise t-test

p-value
0.0001
0.009
0.01
0.02
0.03
0.045

2. Compare first **p-value** with $\alpha/\boldsymbol{k}$.

IF: **p-value < α/k**, reject H_0 for this comparison and set ***k* = *k* – 1**
else: stop checking.

3. Repeat this comparison for all p-values while **p-value < α/k**

	alfa	
p-value	**0.05**	**k**
0.0001	0.008333	6
0.009	0.01	5
0.01	0.0125	4
0.02	0.016667	3
0.03		**Stop**
0.045		

MULTIPLE EXPERIMENTS

False Discovery Rate

False discovery rate (FDR)
FDR control is a statistical method used in multiple hypothesis testing to correct for multiple comparisons. In a list of rejected hypotheses, FDR controls the expected proportion of incorrectly rejected null hypotheses (type I errors).

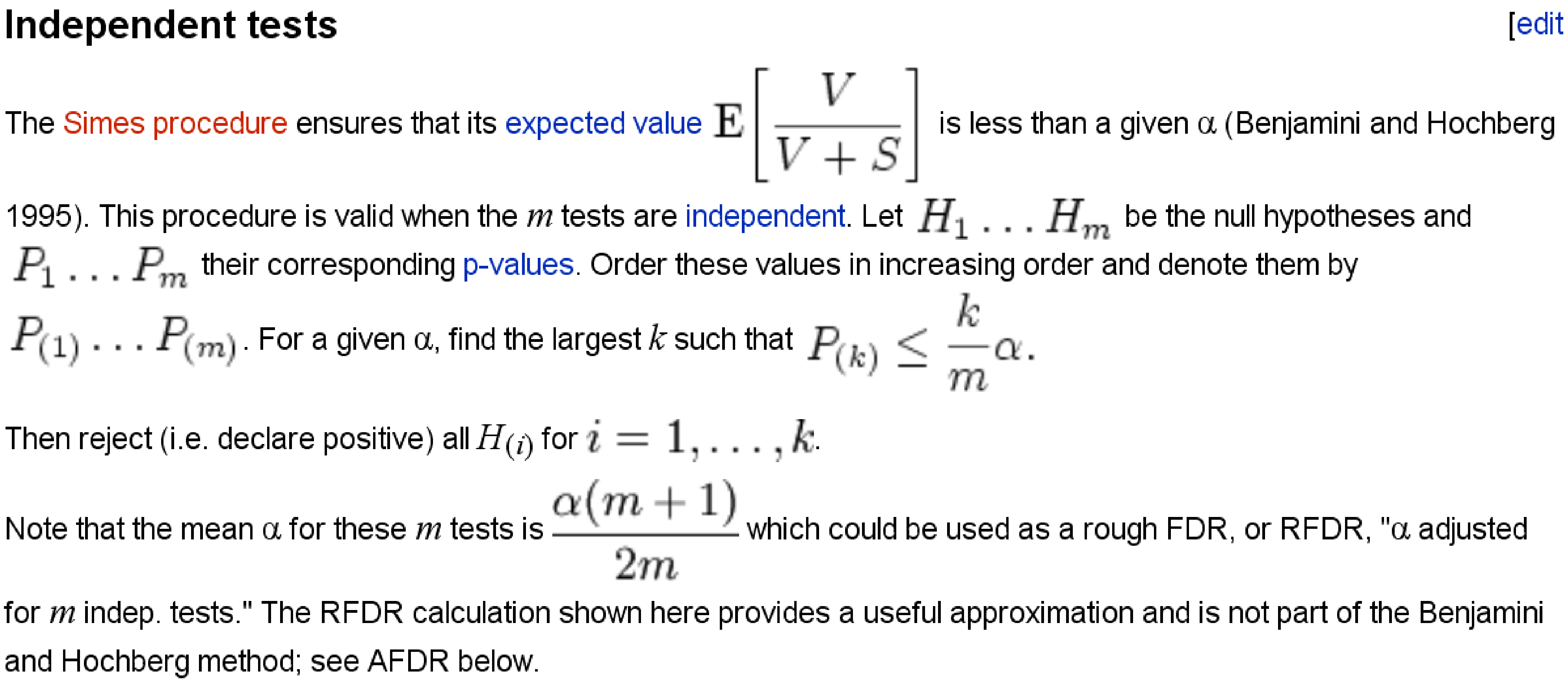
		Population Condition		
		H_0 is TRUE	H_0 is FALSE	Total
Conclusion	Accept H_0 (non-significant)	$\boldsymbol{U}$	$\boldsymbol{T}$	$m-R$
	Reject H_0 (significant)	$\boldsymbol{V}$	$\boldsymbol{S}$	R
	Total	m_0	$m-m_0$	$\boldsymbol{m}$

$$FDR = E\left(\frac{V}{V+S}\right)$$

MULTIPLE EXPERIMENTS

False Discovery Rate

Assume we need to perform *k* = 100 comparisons, and select maximum FDR = α = 0.05



Independent tests

[edit

The Simes procedure ensures that its expected value $\mathrm{E}\left[\frac{V}{V+S}\right]$ is less than a given α (Benjamini and Hochberg 1995). This procedure is valid when the m tests are independent. Let $H_1 \ldots H_m$ be the null hypotheses and $P_1 \ldots P_m$ their corresponding p-values. Order these values in increasing order and denote them by $P_{(1)} \ldots P_{(m)}$. For a given α, find the largest k such that $P_{(k)} \leq \frac{k}{m}\alpha.$

Then reject (i.e. declare positive) all $H_{(i)}$ for $i = 1, \ldots, k$.

Note that the mean α for these m tests is $\frac{\alpha(m+1)}{2m}$ which could be used as a rough FDR, or RFDR, "α adjusted for m indep. tests." The RFDR calculation shown here provides a useful approximation and is not part of the Benjamini and Hochberg method; see AFDR below.

Holm-Bonferroni Method: Example

ANOVAResults.xls

INTERVAL ESTIMATIONS FOR RANDOM FUNCTIONS

Sum and Square of Normal Variables

Distribution of sum or difference of 2 normal random variables
The sum/difference of 2 (or more) normal random variables is a normal random variable with mean equal to sum/difference of the means and variance equal to SUM of the variances of the compounds.

$$x \pm y \to \textit{Normal distribution}$$

$$E[x \pm y] = E[x] \pm E[y]$$

$$\sigma^2_{x \pm y} = \sigma^2_x + \sigma^2_y$$

Distribution of sum of squares on *k* standard normal random variables
The sum of squares of *k* standard normal random variables is a χ^2 with *k* degree of freedom.

$$\textit{if} \quad x_1, \ldots, x_k \to \textit{Normal distribution}$$

$$\sum_{i=1}^{k} x_i^2 \to \chi^2 \quad \textit{with } d.f. = k$$

What to do in more complex situations?

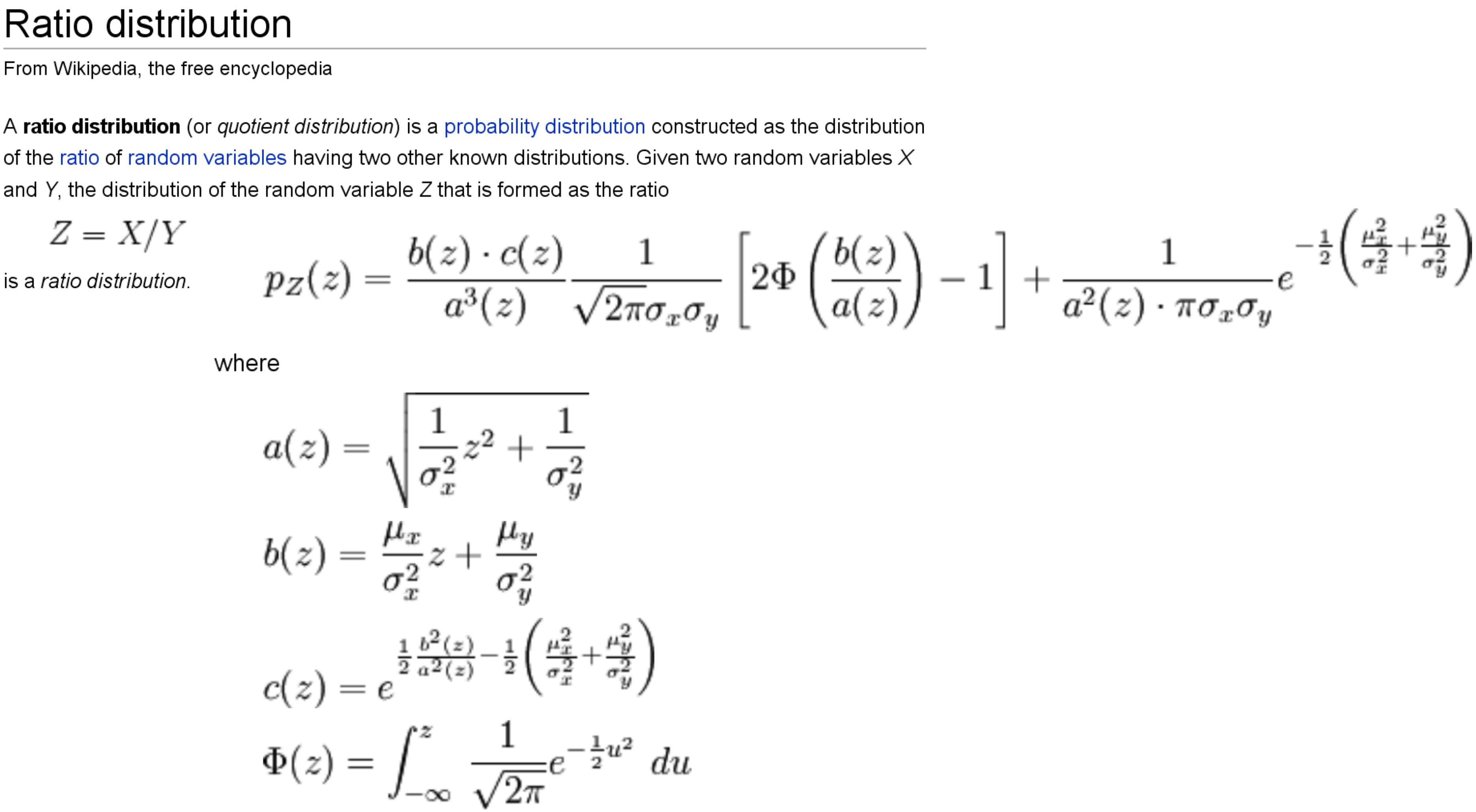
$$\frac{x}{y} \to ?$$

$$\sqrt{x} \to ?$$

$$\log(|x|) \to ?$$

Try to solve analytically?

Simplest case. E[x] = E[y] = 0



Ratio distribution

From Wikipedia, the free encyclopedia

A **ratio distribution** (or *quotient distribution*) is a probability distribution constructed as the distribution of the ratio of random variables having two other known distributions. Given two random variables *X* and *Y*, the distribution of the random variable *Z* that is formed as the ratio

$$Z = X/Y$$

is a *ratio distribution*.

$$p_Z(z) = \frac{b(z)\cdot c(z)}{a^3(z)} \frac{1}{\sqrt{2\pi}\sigma_x\sigma_y}\left[2\Phi\left(\frac{b(z)}{a(z)}\right) - 1\right] + \frac{1}{a^2(z)\cdot\pi\sigma_x\sigma_y} e^{-\frac{1}{2}\left(\frac{\mu_x^2}{\sigma_x^2} + \frac{\mu_y^2}{\sigma_y^2}\right)}$$

where

$$a(z) = \sqrt{\frac{1}{\sigma_x^2} z^2 + \frac{1}{\sigma_y^2}}$$

$$b(z) = \frac{\mu_x}{\sigma_x^2} z + \frac{\mu_y}{\sigma_y^2}$$

$$c(z) = e^{\frac{1}{2}\frac{b^2(z)}{a^2(z)} - \frac{1}{2}\left(\frac{\mu_x^2}{\sigma_x^2} + \frac{\mu_y^2}{\sigma_y^2}\right)}$$

$$\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2}\, du$$

INTERVAL ESTIMATIONS FOR RANDOM FUNCTIONS

Practical Approach

Two rates where measured for a PCR experiment: experimental value (X) and control (Y). 5 replicates where performed for each.

From previous experience we know that the error between replicates is normally distributed.

Q1: provide an interval estimation for the fold change X/Y (α=0.05)

Q2: provide an interval estimation for the log fold change $\log_2(X/Y)$

#	Experiment	Control
1	215	83
2	253	75
3	198	62
4	225	91
5	240	70

Mean	226.2	76.2
StDev	21.39	11.26

Let us use a ***numerical simulation…***

INTERVAL ESTIMATIONS FOR RANDOM FUNCTIONS

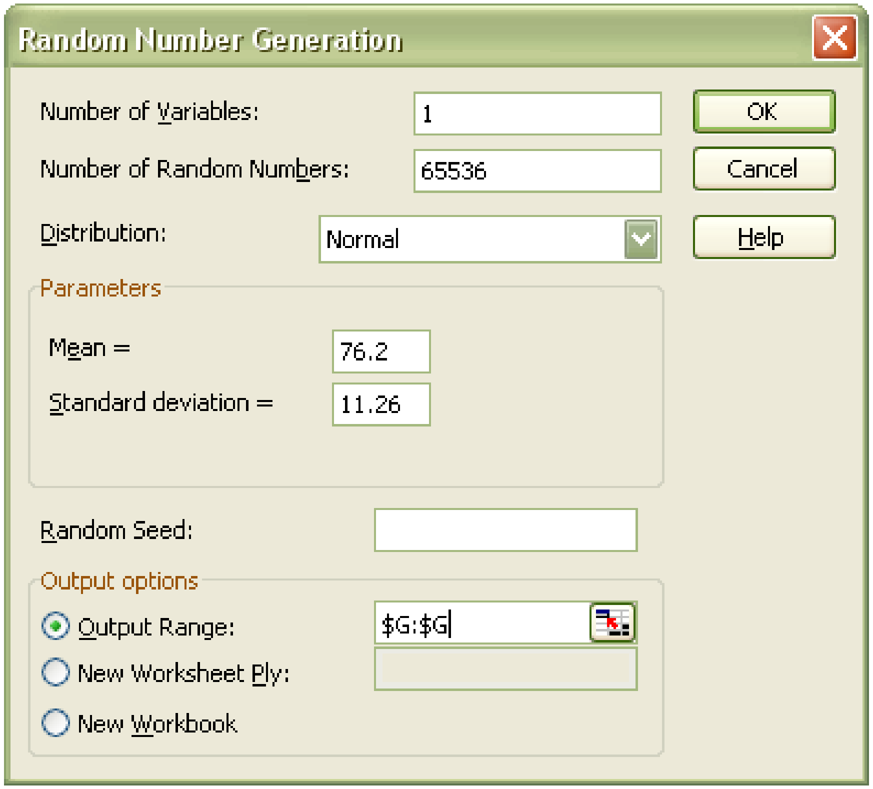
Practical Approach

1. Generate 2 sets of 65536 normal random variable with means and standard deviations corresponding to ones of experimental and control set.

Mean	226.2	76.2
StDev	21.39	11.26

In Excel go: Tools → Data Analysis:

- `Random Number Generation`



Random Number Generation

Number of Variables: 1

Number of Random Numbers: 65536

Distribution: Normal

Parameters

Mean = 76.2

Standard deviation = 11.26

Random Seed:

Output options

● Output Range: $G:$G

○ New Worksheet Ply:

○ New Workbook

If you do not have Data Analysis tool – approximate normal distribution by sum of uniform:

$$N(x,m_x,\sigma_x)=m_x+\sigma_x\left(\sum_{i=1}^{12}U(x_i)-6\right)$$

- `= RAND()` ← *U*(*x*)

INTERVAL ESTIMATIONS FOR RANDOM FUNCTIONS

Practical Approach

1. Generate 2 sets of 65536 normal random variable with means and standard deviations corresponding to ones of experimental and control set.

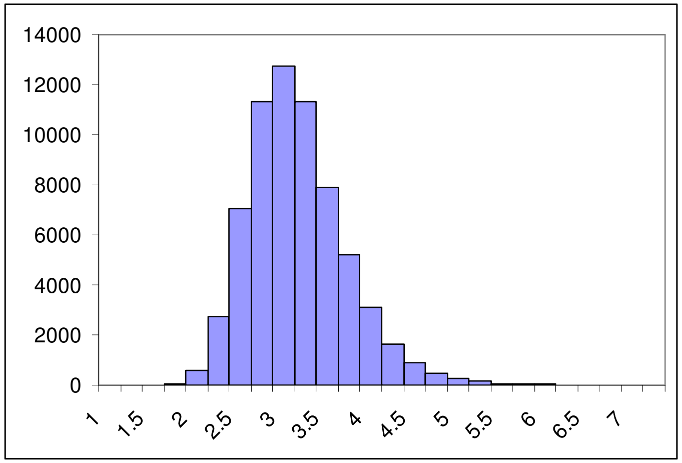
Mean	226.2	76.2
StDev	21.39	11.26

sim.m	226.088799	76.2823
sim.s	21.379652	11.2885

2. Build the target function. For Q1 build X/Y

X/Y.m	3.03289298
X/Y.s	0.566865
min	-8.14098141
max	7.72162205

3. Study the target function. Calculate summary, build histogram.



4. If you would like to have 95% interval, calculate 2.5% and 97.5% percentiles.

In Excel use function

- **=PERCENTILE(data,0.025)**

$X/Y \in [\ 2.13,\ \ 4.33\ ]$

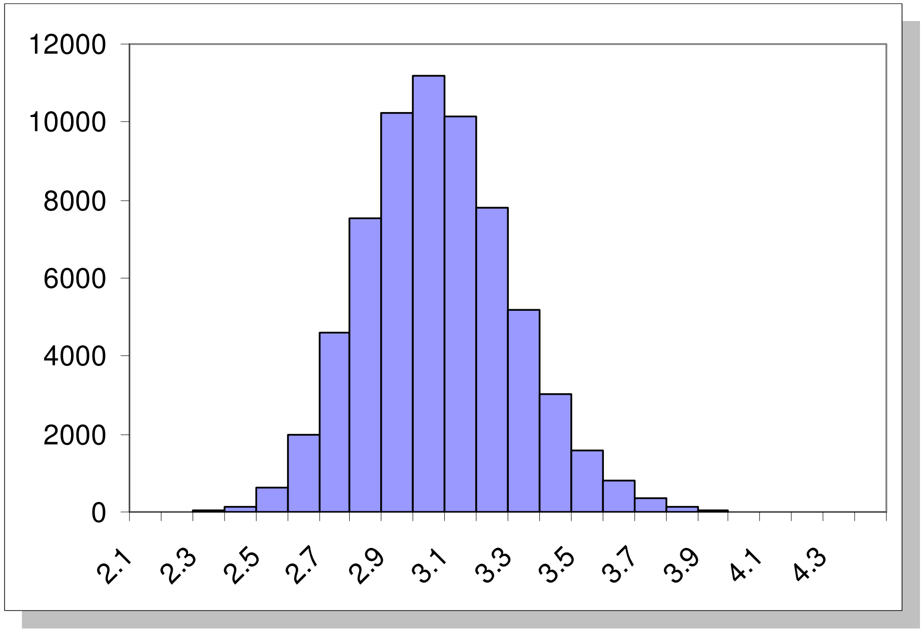
What was a "mistake" in the previous case?

There we spoke about **prediction interval** of X/Y. Now let's produce the **interval estimation for mean X/Y**

Mean	226.2	76.2
StDev	**9.57**	**5.03**

X/Y.m	2.98047943
X/Y.s	0.23616818
min	2.01556098
max	4.31131109

$E[X/Y] \in [\, 2.55,\ \ 3.48 \,]$



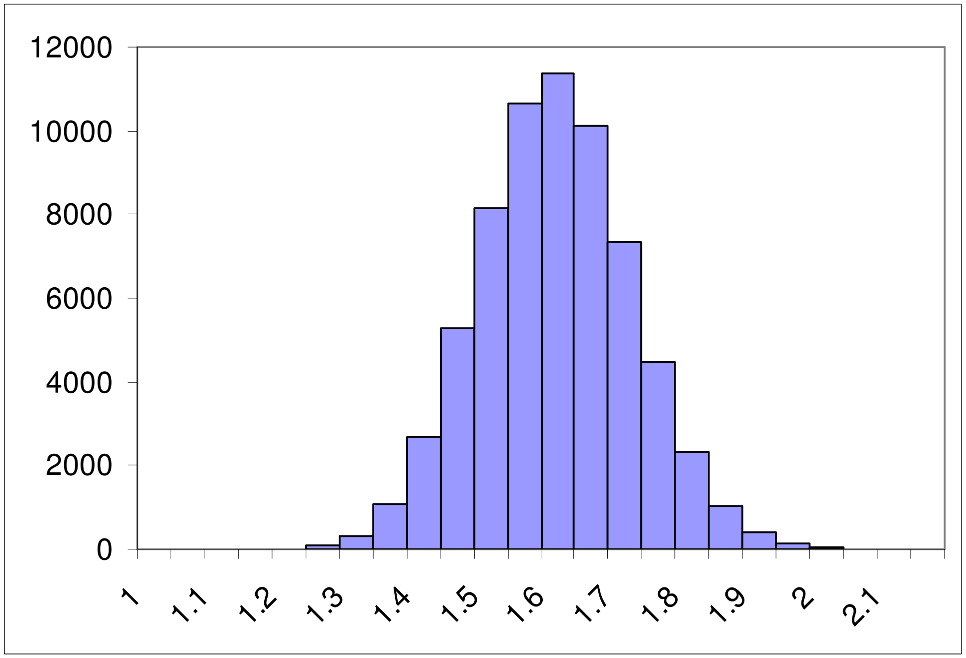
INTERVAL ESTIMATIONS FOR RANDOM FUNCTIONS

Practical Approach

Q2: provide an interval estimation for the log fold change log2(X/Y)

Mean	1.571052
Standard Devi	0.113705

E[log(X/Y)] ∈ [1.35, 1.80]



	Simulation	**Normal**
2.50%	1.3546	1.3482
97.50%	1.7998	1.7939

QUESTIONS ?

Thank you for your attention