COMPARING TWO MEANS USING SIMULATION

Applied Data Science 2

Gedi Lukšys

November 6, 2023

OUTLINE

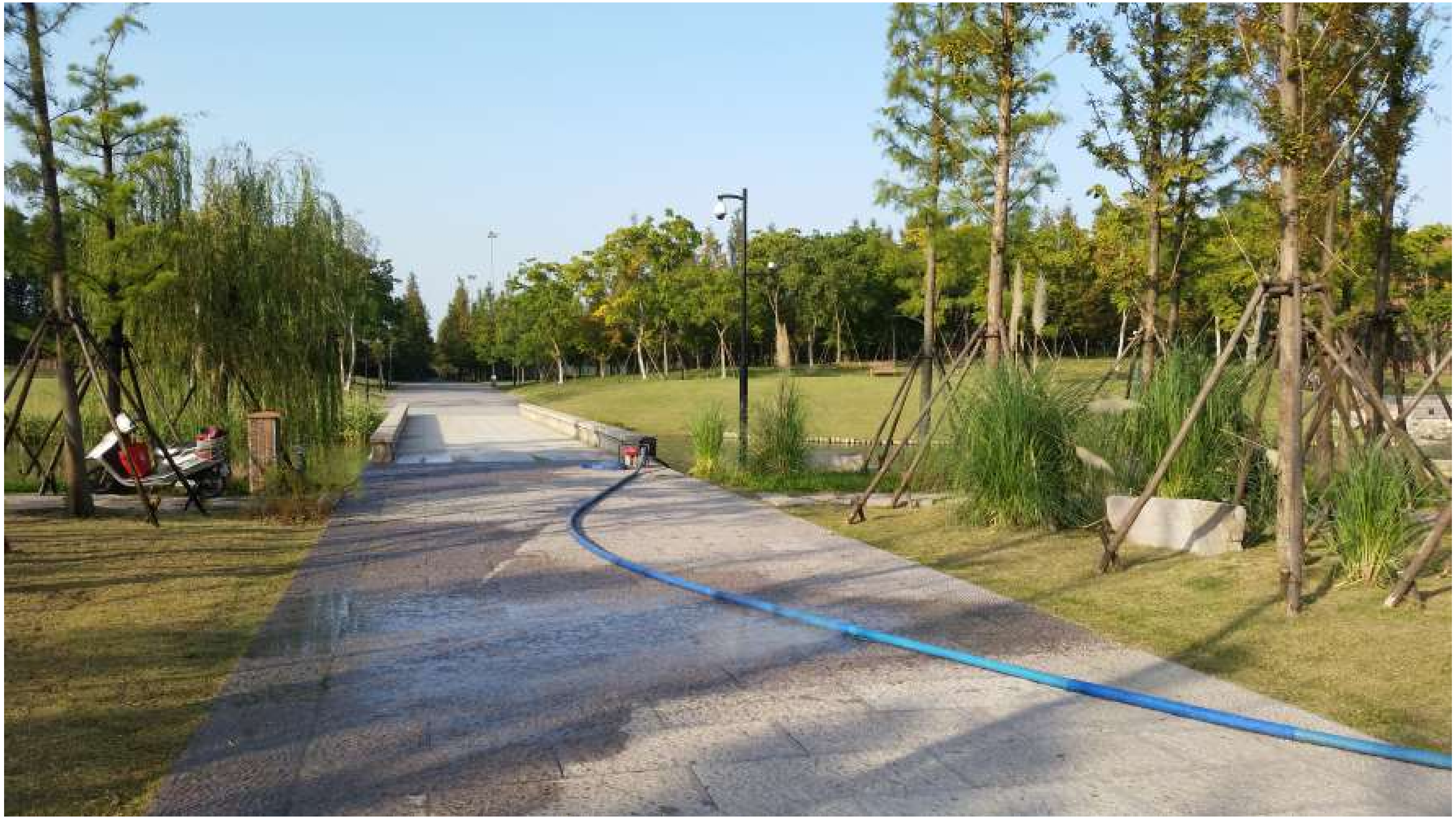
Comparing two means vs. two distributions

- Statistical inference
- Effect size vs. statistically significant difference
- Central limit theorem

When and why do we use simulations?

How do we use simulations?

WHAT DO YOU THINK ABOUT THIS?



IMAGINE YOU WANT TO COMPARE 2 QUANTITIES

Happiness ratings of UoE students based in Edinburgh and in Haining

- 2 independent samples

Satisfaction ratings of catering on ZJU international campus in 2018/19 and 2019/20 (assume the same people provide ratings)

- 2 paired samples => take the difference, treat as 1 sample

Average activity of PFC neurons with a baseline (4 Hz)

- 1 sample, compare with a number

WHAT DO WE ACTUALLY COMPARE?

Two means

A mean with a value (<=> mean–value with 0)

What exactly does it mean?

One mean is greater than the other (or a value)

Well that’s just yes or no…

But what are we interested in statistically speaking?

WHAT ARE WE INTERESTED IN STATISTICALLY SPEAKING?

(most of the time) **Future prediction!**

If we get one student in Edinburgh and another in Haining, what's the **probability** that the former will be **more** happy **than** the latter?

If we take a random member of ZJU international campus, what's the **probability** that s/he was **more** happy with catering in 2018/19 **than** in 2019/20?

If we take a random PFC neuron in our experiment, what's the **probability** that its activity is **above** 4 Hz?

Effect size! How to estimate this?

DISTRIBUTIONS!

Sampling

Comparing area under curve to the right/left of a value, e.g. can I pass through the right side without getting significantly wet

e.g. getting less than 1% of water

HOW CAN WE KNOW THIS?

Ideally, know the underlying distribution, e.g. normal

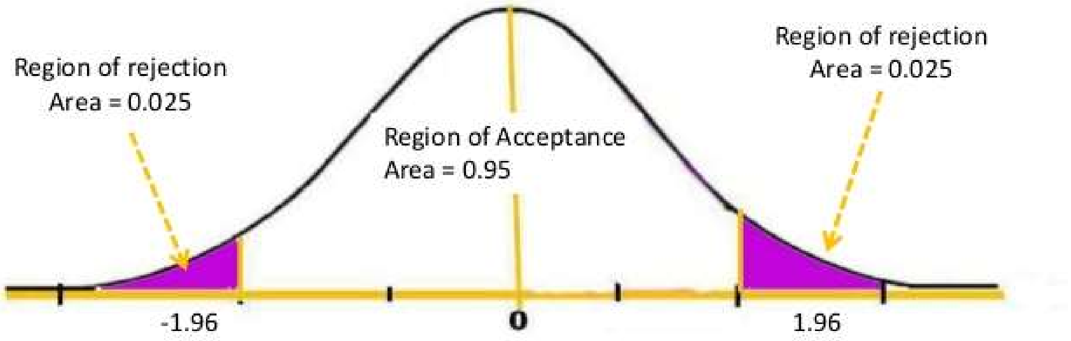
compute its Cumulative Distribution Function value based on known parameter(s)

e.g. $\text{cdf}_{\text{normal}}(1.96) = 0.975$

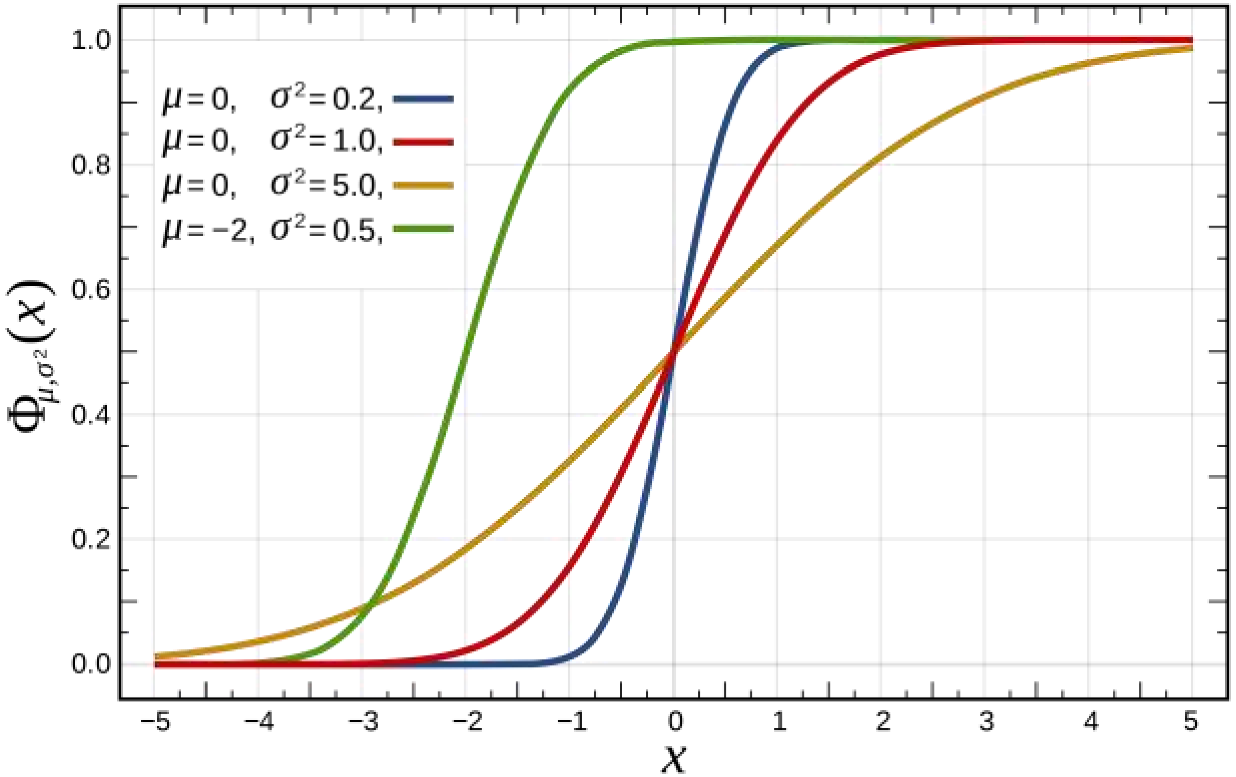
in R, *pnorm(1.96)*

or its inverse, if cdf value is known

in R, *qnorm(0.975)*

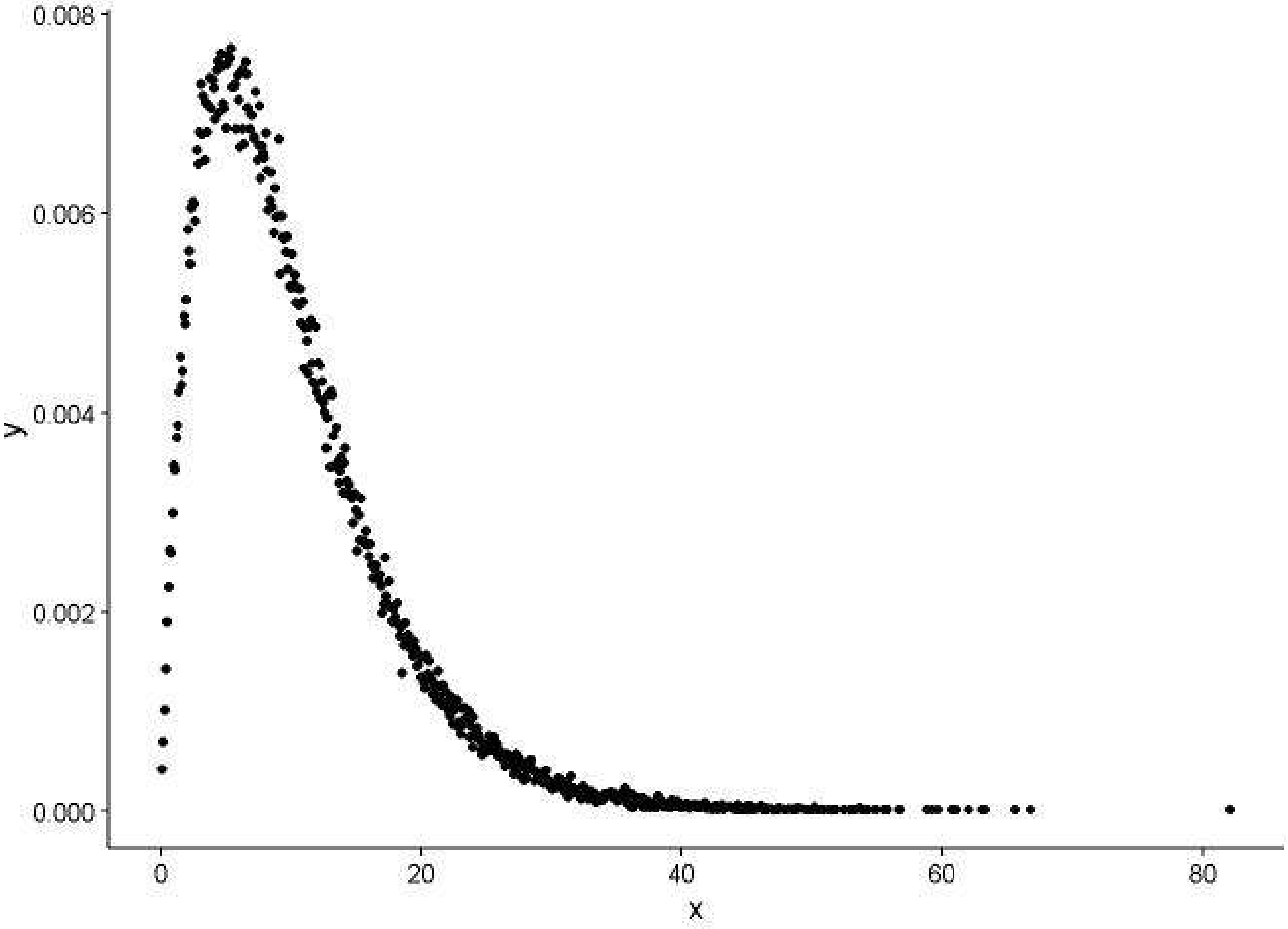


https://image.slidesharecdn.com/statisticshypothesistesting-130115004420-phpapp02/95/statistics-hypothesis-testing-14-638.jpg?cb=1358210733



GAMMA DISTRIBUTION

for the water pipe example…



https://stats.stackexchange.com/questions/190465/how-to-draw-fitted-graph-and-actual-graph-of-gamma-distribution-in-one-plot

SOMETHING HAPPENED HERE... ☺

Sampling

Comparing a known distribution with a sample

How likely is it that **their means** are **the same**?

COMPARING THE MEANS

What will it depend on:

- Effect size (distribution-based)

e.g. r^2, Cohen's d $= (\mu_1 - \mu_2)/\sigma$

- Sample size n

s.e.m. $= \sigma$ / sqrt(n)

Hence, we can have a fairly large effect size but no significant difference between the means or a very small effect size but a significant difference between the means

REMEMBER CENTRAL LIMIT THEOREM!

AFTER ALL, WE WANT TO COMPARE MEANS…

https://www.youtube.com/watch?v=YAIJCEDH2uY

Even if you’re not normal...

WHY SIMULATIONS?

Try to understand how distributions work ☺

What if we just want to compare 2 means?

Two sample Student’s T test (next week)

What if our data doesn’t satisfy the assumptions

E.g. our distributions are not normal but known, e.g. uniform or exponential

$$f(x;\lambda) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0, \\ 0 & x < 0. \end{cases} \qquad F(x;\lambda) = \begin{cases} 1 - e^{-\lambda x} & x \geq 0, \\ 0 & x < 0. \end{cases}$$

$$f(x) = \begin{cases} \frac{1}{b-a} & \mathrm{for}\ a \leq x \leq b, \\ \\ 0 & \mathrm{for}\ x < a\ \mathrm{or}\ x > b \end{cases} \qquad F(x) = \begin{cases} 0 & \mathrm{for}\ x < a \\ \frac{x-a}{b-a} & \mathrm{for}\ a \leq x \leq b \\ 1 & \mathrm{for}\ x > b \end{cases}$$

our sample size is small

Convert data to ranks, use a rank-based test, e.g. Wilcoxon rank sum

In doing so, we lose their underlying distributions…

WHY SIMULATIONS?

If distributions can be mathematically expressed, there could be a mathematical solution for the comparison between two means (or a mean and a value)

But, do we want to try to find it?

Sometimes, our distributions may not be mathematically expressed but could be easy to sample

Sometimes the two samples we compare may have very different properties

Sometimes our random variables may not be independent and identically distributed, but may follow some peculiar rule, as in the example later

In all these cases, performing standard statistical tests may not be a good idea

Instead, we can use simulations to estimate our predictions (p values)!

SO WE FINALLY WANT TO USE A SIMULATION. WHAT NEXT?

Need to have one or two samples

Know what kind of distribution they follow

Estimate its parameters for each sample (e.g. μ, σ for normal, λ for exponential)

Be able to generate equivalent random samples of the same size!

e.g. *rnorm(N, mean(our_sample), sd(our_sample))*

Sometime we may instead have an empirical distribution and use it for sampling

e.g. *sample(empirical_distribution, N)*

THE SIMULATION PROCEDURE

1. Need to have a random variable to sample from (either a parametric distribution or an empirical one)

2. Sample it many times (the more the merrier)

3. Count how many times something we're interested in happens

e.g. random variable is above/below certain value, if we're interested in effect size

e.g. mean of our generated sample is above/below the mean of our other generated sample or a certain value, if we're interested in significant difference between two means or between a mean and a value

HOW MANY TIMES?

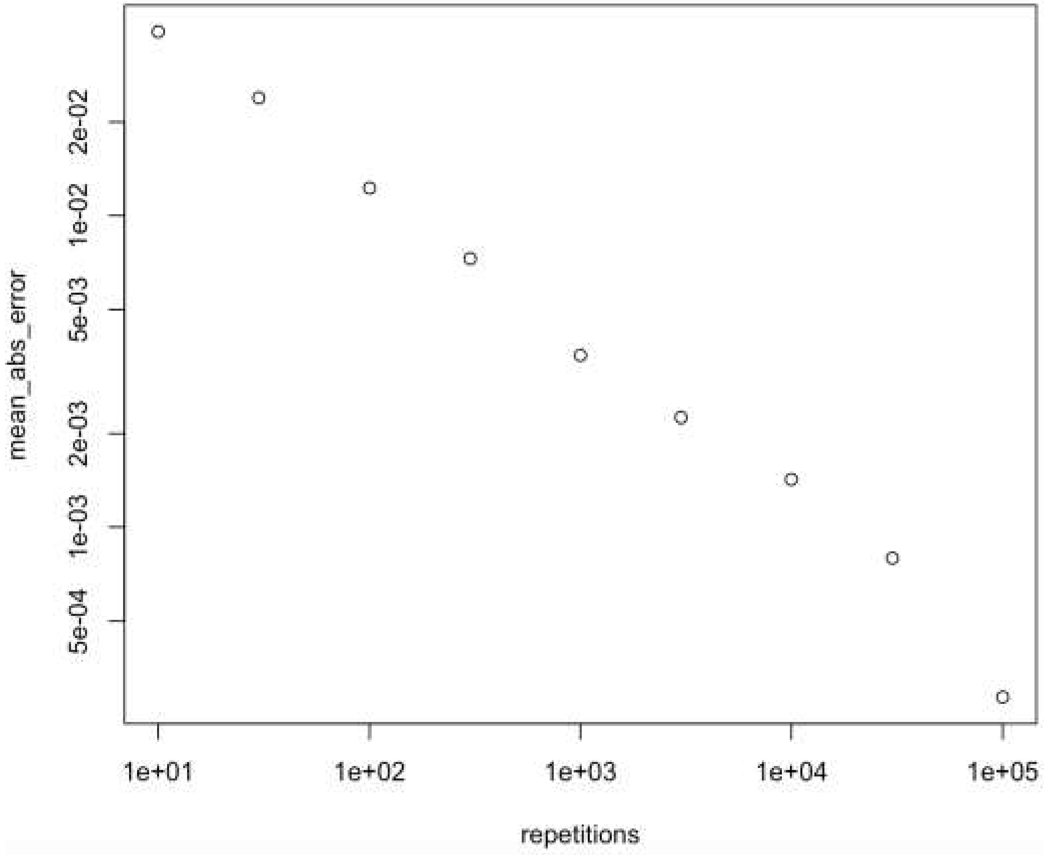
It can take up memory and/or computing time (often you can exchange them!)

In the worst case, crash your computer or take ages to complete…

Errors can be estimated!

e.g. let's look at mean values of *abs(mean(rnorm(repetitions,0,1)<1.96) – pnorm(1.96))*

For more complicated simulations not exactly the same but similar trend



HOW MANY TIMES? INHERENT LIMITATIONS

Usually 10000-100000 is enough, but can depend on task demands

When is it definitely not good enough?

Very small P-values!

What can be done?

No simple solution, but some shortcuts possible, e.g.

- Sample only a part of the distribution where the event of interest can occur
- Need to know which exactly part, also be sure!

AN EXAMPLE WHERE SIMULATION CAN HELP… ☺

How would you solve this?

AVOIDING LES ENTARTEURS

Billy Goats, Chairman and CEO of industry giant Mycowsoft, is visiting his European branch offices. But alas, the notorious Le Gloupier is at it still, and has placed his agents (the entarteurs) in the streets of Europe, waiting to throw cream pies in celebrities' faces. Luckily, Mycowsoft Corporate Intelligence has determined how many entarteurs lie in wait along each of the routes between offices, and has asked you to plan Billy's visits so as to minimize the number of pies Billy has to wipe off his face.

Billy needs to visit all N of Mycowsoft's corporate offices, each one exactly once, starting from any of them and ending in any other. You should create such a route and print the office numbers visited in order. Your solution will be scored based on how many entarteurs were encountered; the better (smaller) that number is, the more points you get for that test case.

http://www.angelfire.com/ca2/lorddave/usa984.html

Starring…

More in the practical & problem set!