Stat215A

Discussion 2014-09-02

Topics Today

- About this lab
- Why Stat215?
- Readable code

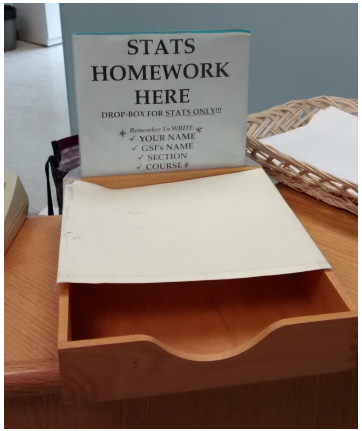
About this lab

About this lab

It will be a lot of work.

You will learn by doing, not by listening.

Turn-in



On the due date (Tuesday), before the lab:

← Homework and write-up here (Evans 367)

bCourses

← Writeup and all code here

Lab Grading

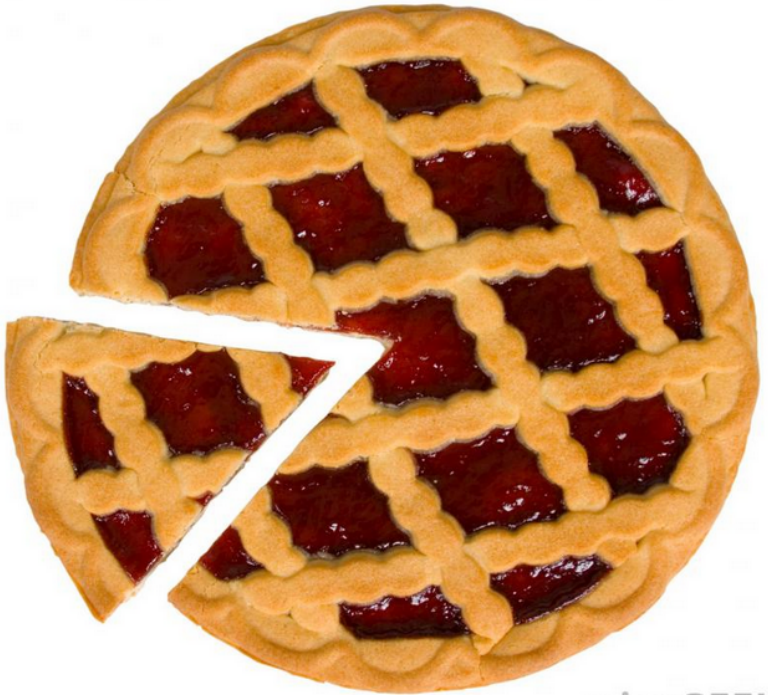
- A: Addressed all the lab questions with exceptional:
  - Clarity
  - Creativity and / or
  - Care
- B: Satisfactorily addressed all the lab questions
- C / D: Did not complete all the lab tasks

The labs are somewhat open-ended. Grading will, necessarily, be somewhat subjective.

Lab Grading

- Clarity
  - Clear writing, graphs, equations, code
- Creativity
  - Answer unasked questions, use new methods, discover new patterns, innovative graphics
- Care
  - Well-tested code, lots of sanity checks, different ways of answering the same question

Evaluation and Support



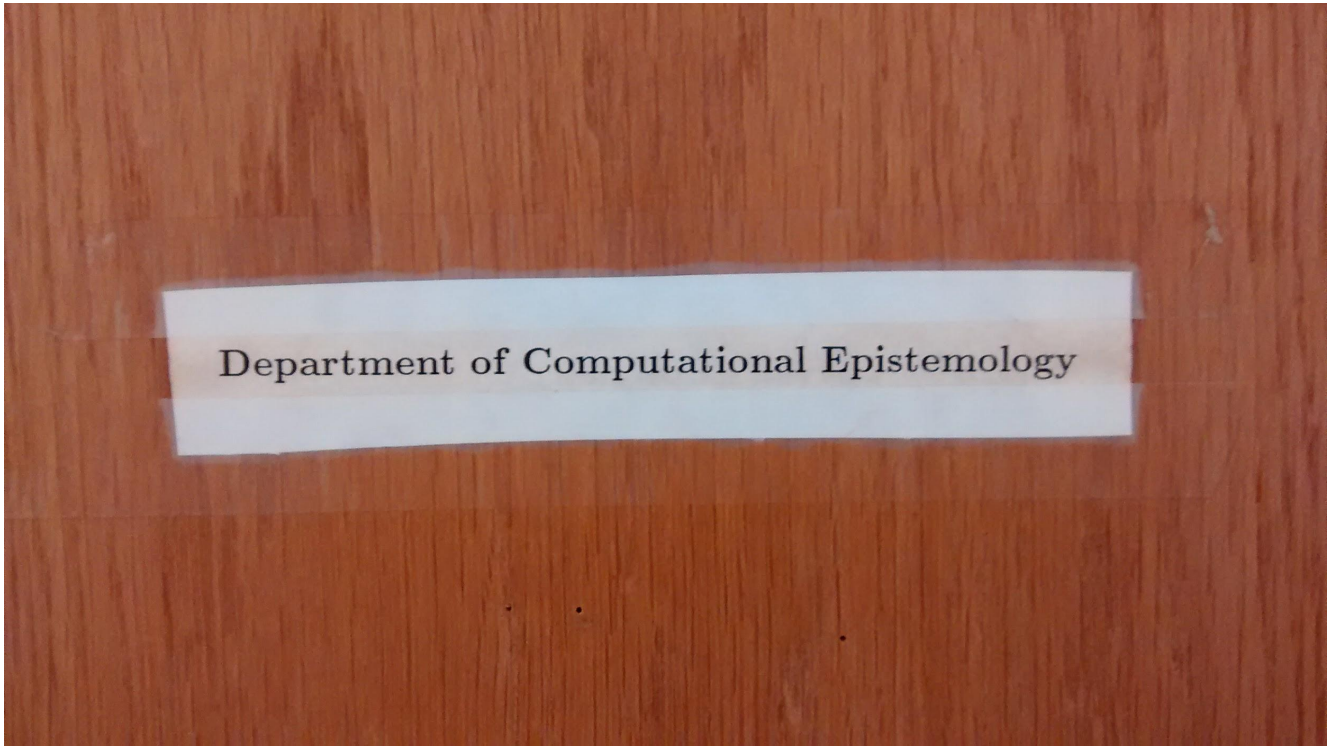
GSI office hours:
Evans 446
Mon 1-3
Fri 2-4

If you want more feedback on your labs, come see me in person.

(I'll be happy to extend office hours, but won't change grades.)

Why Stat215?

Why Stat215?



Statistics is philosophy

“Probability … [is] a quadruple success: metaphysical, epistemological, logical, and ethical. … Probability cannot dictate values, but it now lies at the basis of all reasonable choice made by officials. No public decision, no risk analysis, no environmental impact, no military strategy can be conducted without decision theory couched in terms of probabilities. By covering opinion with a veneer of objectivity, we replace judgement by computation. … Probability is, then, *the* philosophical success story of the first half of the twentieth century.”

- Ian Hacking, The Taming of Chance

Science Needs You

(Prof. Yu will focus on this a lot)

Silicon Valley Needs You

Disclaimer

The expressed interpretation of the follow exercises is the opinion of the GSI, not necessarily of Prof. Yu, Berkeley Statistics, or the statistical establishment.

You should consider making up your own mind.

Group Question 1

Describe *precisely* (as you would to an intelligent caveman*) the meaning of these statements:

- The probability of the coin coming up heads is 0.5
- The probability of rain tomorrow is 0.2
- The probability that my partner and I will get married is 0.95

* You may assume that the caveman understands the institution of marriage.

Moral of Question 1

We use randomness metaphorically to describe uncertainty. This analogy is more appropriate in some situations than in others.

When you describe uncertainty using randomness, think about whether that analogy is justified.

In fancy terms: know the difference between “aleatoric” and “epistemic” uncertainty.

Group Question 2

Which of these sequences are correlated?

- 3, 4, 3, 1, 4, 2
- 1, 2, 3, 4, 5, 6
- 6, 2, 1, 4, 5, 3
- 6, 6, 6, 6, 6, 6

Moral of Question 2

Processes have probabilistic properties. Numbers do not.

Numbers that come from a process can tell you things about that process.

When you describe statistical properties a dataset, describe the random process that you think generated it.

Group Question 3

You are a food inspector. You randomly inspect 100 salads and find one dead mouse.

Let *p* denote the rate of dead mice per salad.

What is your best estimate of *p*?

Can you reject the hypothesis that *p=0*?

Moral of Question 3

If the math tells you something ridiculous, the math is probably wrong.

Regularly check your mathematical and statistical reasoning against reality.

Avoid “pernicious reification”!

History

Consider reading some of the history of statistics!

“Probability has two aspects. It is connected with the degree of belief warranted by evidence, and it is connected with the tendency, displayed by some chance devices, to produce stable relative frequencies. Neither of these aspects were self-consciously and deliberately apprehended by any substantial body of thinkers before the time of Pascal.”

Ian Hacking, The Emergence of Probability

Readable Code

R, for better or worse

R has upsides and downsides. But we will be using it in this lab.

We will emphasize writing R for:

- Readability (style)
- Reliability (unit testing)
- Efficiency (Rcpp)

Code as Document

Write your code as if it is to be read, not only run.
Please follow the Google style guide.

- Organize your code
- Use meaningful variable names
- Comment liberally

What does this code do?

```
myfun <- function(x, y) sum(x*log(y)-y); myfun2 <- function(x) myfun(dd, exp(x)); dd <- rpois(100, 4)

exp(optim(0.693, myfun2, method="Brent", lower=-6.90775, upper=4.60517, control=list(fnscale=-1))$par)
```

What does this code do?

```
PoissonLogLik <- function(data, lambda) {
  # The unnormalized Poisson log likelihood.
  # Arguments:
  #   data: A numeric vector containing iid data
  #   lambda: The poisson parameter.
  # Returns:
  #   The unnormalized log likelihood of data if the parameter were lambda.
  return(sum(data * log(lambda) - lambda))
}

PoissonLogLikForOptim <- function(log.lambda) {
  # A wrapper function for optim, which takes only log(lambda) as an argument.
  return(PoissonLogLik(data=data, lambda=exp(log.lambda)))
}
```

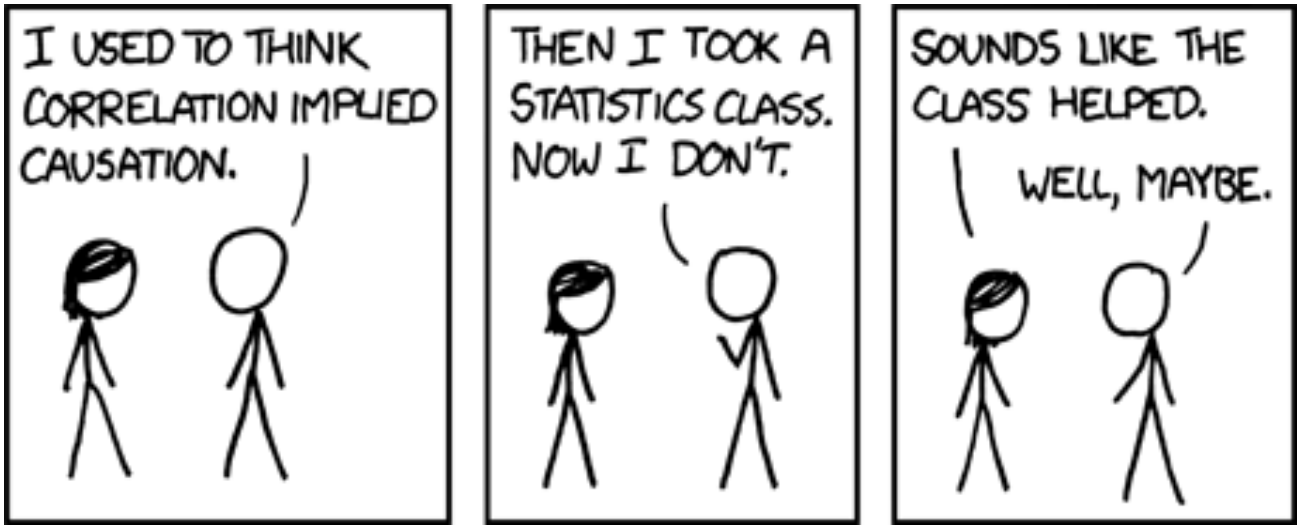
What does this code do?

```
true.lambda <- 4
starting.lambda <- 2
data.size <- 100
data <- rpois(n=data.size, lambda=true.lambda)

# optim's default behavior is to minimize, so fnscal=-1 maximizes the likelihood.
optim.result <- optim(par=log(starting.lambda), fn=PoissonLogLikForOptim,
                      control=list(fnscale=-1), method="Brent",
                      lower=log(0.001), upper=log(100))
print(exp(optim.result$par))
```

Why code readably?

- Other people (including your future self) can easily know what it does
- You don’t have to write and update extra documentation
- It helps you find mistakes
- It makes code more modular
- It quickly becomes habit
- Your GSI will know immediately what an excellent job you have done



In closing, we hope this class is useful.