5. Empirical Methods in Security

ENEE 657

Prof. Tudor Dumitraș
Assistant Professor, ECE
University of Maryland, College Park

http://ter.ps/enee657

Today's Lecture

- Where we've been
  - Memory corruption exploits
  - Cryptography
  - OS protection mechanisms
- Where we're going today
  - Empirical methods in security
- Where we're going next
  - Measurements module
  - **Pilot projects: proposals due on Wednesday**

2

Pilot Project Proposals

- No class on Wednesday
  - Focus on developing your proposals for the pilot project
- **Post concise (2-3 paragraphs) proposal** on Piazza
  - Problem statement
  - Approach considered for tackling the problem
    - Must describe **concrete tasks**, not vague directions
    - Must **demonstrate that you've thought about the first steps**, and you are not simply paraphrasing the project idea
  - **Deadline: Wednesday**

3

Goals of Security Mechanisms

- Eliminate an **entire class** of attacks
  - Example: harvesting credit card numbers by sniffing network packets used to be common in the '90s. HTTPS stopped that.
  - Challenges:
    - **Arms race**: adversaries find new attacks (e.g., harvesting credit card numbers by hacking point-of-sale systems)
    - Mechanism may not address the **capabilities of real-world adversaries** (we've seen: attacking crypto without breaking the math)
- Make it **less likely** for an attack to succeed
  - Increases the attacker's **work factor**
  - Challenges:
    - Requires understanding attack techniques (we've seen: mitigations for memory-corruption exploits)
- **Distinguish** between benign and malicious behavior
  - Increasingly using **statistical techniques**

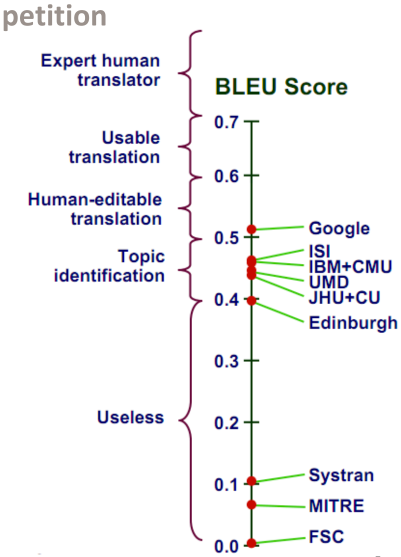
4

The "Unreasonable Effectiveness" of Data

2005 NIST Machine Translation Competition

English-Arabic competition

- Google's first entry
  - None of the engineers spoke Arabic
- Simple statistical approach
- Trained using United Nations documents
  - 200 million translated words
  - 1 trillion monolingual words



5

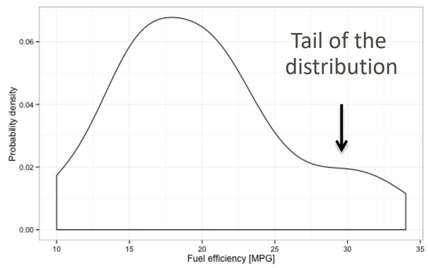
> The world's most valuable resource is no longer oil, but data
>
> *The Economist, 2017*

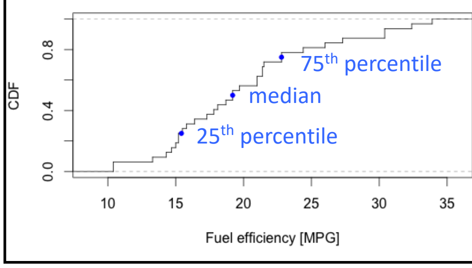
6

Statistical Distributions

What does the data look like? (empirical distribution)

- Probability density function (PDF) of the values you measure
  - PDF(x) is the probability that the metric takes the value x
  - $\int_a^b PDF(x)\,dx = \Pr[a \le metric \le b]$
  - Estimation from empirical data (Matlab: `ksdensity` R: `density`)





- Cumulative density function (CDF)
  - CDF(*x*) is the probability that the metric takes a value less than *x*

$$CDF(x) = \int_{-\infty}^{x} PDF(u)\,du = \Pr[metric \le x]$$

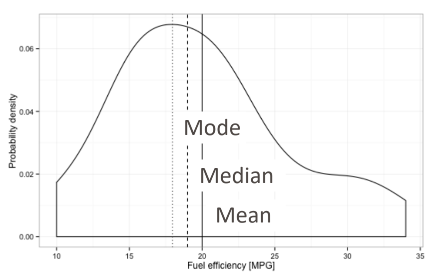
  - Estimation (R: `ecdf`)

9

Summary Statistics

What does the data look like? (in summary)

- Measures of centrality
  - Mean = sum / length (`mean`)
  - Median = half the measured values are below this point (`median`)
  - Mode = measurement that appears most often in the dataset



> "80% of analytics is sums and averages."
>
> ***Aaron Kimball, wibidata***

- Measures of spread
  - Range = maximum – minimum (`range`)
  - Standard deviation (σ) (Matlab: `std` R: `sd`) $\sigma = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}$
  - Coefficient of variation = σ / mean
    - Independent of the measurement units

10

Percentiles and Outliers

What does the data look like? (in summary)

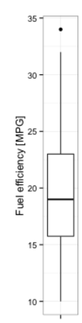
- Percentiles
  - N^{th} percentile: X such that N% of the measured samples are less than X
    - The median is the 50^{th} percentile
    - The 25^{th} and 75^{th} percentiles are also called the 1^{st} and 3^{rd} quartiles (Q_1 and Q_3), respectively
  - Matlab: `prctile` R: `quantile`
  - ! **The "five number" summary of a data set: <min, Q_1, median, Q_3, max>**

- Outliers
  - "Unusual" values, significantly higher/lower than the other measurements
  - ! **Must reason about them: Measurement error? Heavy-tailed distribution? An interesting (unexplained) phenomenon?**
  - Simple detection tests:
    - 3σ test $X_{outlier} > \overline{X} + 3\sigma$
    - 1.5 * IQR $X_{outlier} > Q_3 + 1.5(Q_3 - Q_1)$
    - R package outliers
  - ! **The median is more robust to outliers than the mean**

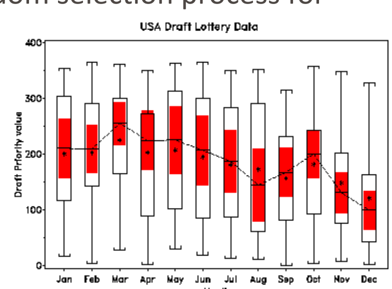
11

Boxplots

What does the data look like? (comparisons)



- Box-and-whisker plots are useful for comparing probability distributions
  - The box represents the size of the inter-quartile range (IQR)
  - The whiskers indicate the maximum and minimum values
  - The median is also shown
  - Matlab: `boxplot` R: `ggplot(..)+geom_boxplot()`
- In 1970, US Congress instituted a random selection process for the military draft
  - All 366 possible birth dates were placed in a rotating drum and selected one by one
  - The order in which the dates were drawn defined the priority for drafting
  - ! **Boxplots show that men born later in the year were more likely to be drafted**



From http://lib.stat.cmu.edu/DASL/Stories/DraftLottery.html

Statistical Inference

- You must understand how to interpret data correctly
- Statistical inference: Methods for drawing conclusions about a population from sample data
- Two key methods
  - Confidence intervals
  - Hypothesis tests (significance tests)

13

Confidence Intervals

What is the range of likely values?

- 95% confidence interval for the sample mean
  - If we repeated the experiment 100 times, we expect that this interval would include the mean 95/100 times
  - $CI = \mu \pm 1.96 \frac{\sigma}{\sqrt{n}}$

μ: mean
σ: standard deviation
n: number of elements

- Why 95%?
  - No good reason, but widely used
- You can compute confidence intervals for many statistical measures
  - Variance, slope of regression line, effect size, etc.

14

Hypothesis Tests

Is a result statistically significant?

- Compare an **experimental group** and a **control group**
  - **H_0**: Null Hypothesis = No difference between the groups
  - **H_1**: Alternative Hypothesis = Significant difference between the groups

- Hypothesis tests
  - **t-test**: are the means significantly different? (R: `t.test`)
    - One-tailed ($\mu_1 > \mu_2$), two-tailed ($\mu_1 \neq \mu_2$)
    - Paired (difference between pairs of measurements)
  - **χ^2 goodness-of-fit test**: does the empirical data match a probability distribution (or some other hypothesis about the data)? (R: `chisq.test`)
  - **Analysis of Variance (ANOVA)**: is there a difference among a number of treatments? Which factors contribute most to the observed variability? (R: `anova`)

15

Hypothesis Tests – How Different is Different?

Is a result statistically significant?

- How do we know the difference in two treatments is not just due to chance?
  - We don't. But we can calculate the odds that it is.

- The ***p*-value** = likelihood that H_0 is true
  - In repeated experiments at this sample size, how often would you see a result at least this extreme assuming the null hypothesis?
  - $p < 0.05$: the difference observed is **statistically significant**
  - $p > 0.05$: the result is **inconclusive**
  - Why 5%? Again, no good reason but widely used.

! **A non-significant difference is not the same as no difference**

! **A significant difference is not always an interesting difference**

16

The Truth Wears Off

Jonah Lehrer, The New Yorker, 2010

- **John Davis, University of Illinois**
  - "Davis has a forthcoming analysis demonstrating that the efficacy of antidepressants has gone down as much as threefold in recent decades."
- **Jonathan Schooler, 1990**
  - "subjects shown a face and asked to describe it were much less likely to recognize the face when shown it later than those who had simply looked at it."
  - The effect became increasingly difficult to measure.
- **Joseph Rhine, 1930s, coiner of the term extrasensory perception**
  - Tested individuals with card-guessing experiments. A few students achieved multiple low-probability streaks.
  - But there was a "decline effect" – their performance became worse over time.

http://www.newyorker.com/reporting/2010/12/13/101213fa_fact_lehrer

Sampling

What can you tell about a population by observing a sub-sample?

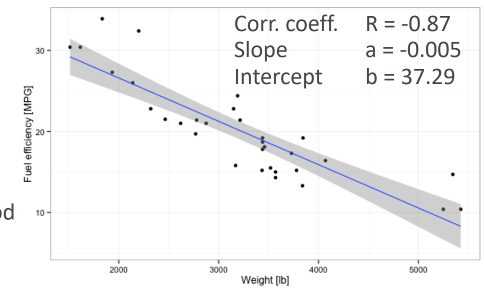
- Sometimes you may choose your sample size (or sampling rate)
  - Rule of thumb: 10% is usually OK for large data
  - Strategies:
    - Uniform sampling: randomly keep 1 out of 10 data points (R: `sample`)
    - Stratified sampling: for each city, keep equal number of rows
  - Useful trick: sample based on output of crypto hash (e.g. MD5)
    - Output bits of hash are uniformly distributed regardless of the input
- Bootstrapping: how to extrapolate property **Q**
  - Want **Q**(sample) → **Q**(whole population)
  - Key idea: observe the distribution of **Q** on several sub-samples
    - How well can you extrapolate **Q**(sub-sample) → **Q**(sample)?
  - Useful when the sample size is insufficient for inference

18

Correlation and Regression

Are two factors related?

- Correlation coefficient R (R: `cor`)
  - ~ 1: positive correlation (when X grows, Y grows too)
  - ~ -1: negative correlation (when X grows, Y goes down)
  - ~ 0: no correlation
  - – *p*-value: Pr[R ≠ 0], dependent on sample size (R: `cor.test`)
  - ! **Compute the correlation coefficient only you think that the relationship between X and Y is linear**
  - ! **Correlation is not causation**



- Regression (R: `lm`)
  - – Fit linear model $y = ax + b$
    - Typically using least squares method
    - Some methods are robust to outliers (R package: `minpack.lm`)

Effect Size

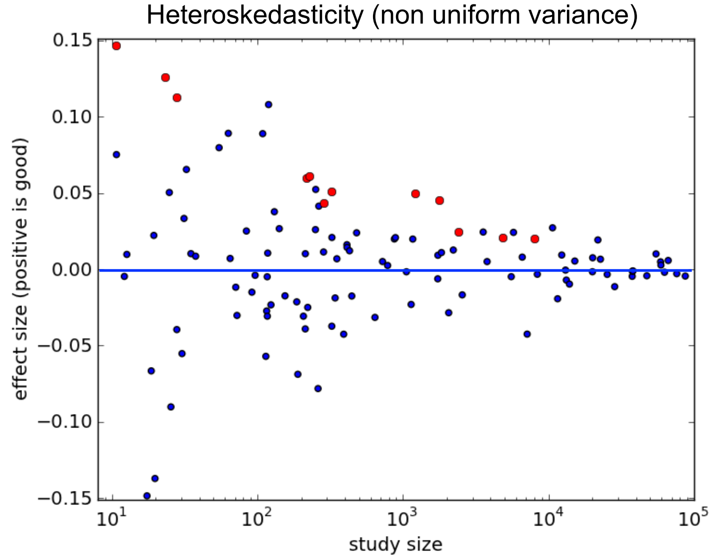
"Significant" is not good enough – how significant?

$$\text{Effect size} = \frac{[\text{Mean of experimental group}] - [\text{Mean of control group}]}{\text{standard deviation}}$$

- Used prolifically in meta-analysis to combine results from multiple studies
  - – The aggregate result may have an increased confidence level
  - – Example: weighted average, using inverse variance weights
  - ! **Averaging results from different experiments can produce nonsense if you violate the assumptions of those experiments**
  - – Other definitions of effect size exist: odds ratio, correlation coefficient

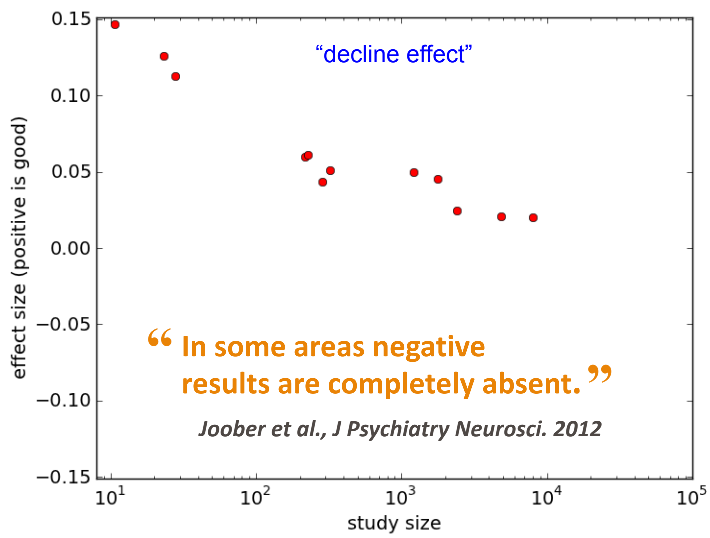
20

So Why Does the Truth Wear Off?



21

Publication Bias



22

A Note on Paper Critiques and Discussions

- **Think critically!**

- Extract the essence – what you want to remember from the paper
  - What did the authors try to achieve?
  - What are the contributions of the research?
  - What are the weaknesses?

- Some papers are tutorial in nature
  - Summarize them, instead of writing strengths / weaknesses

- Write the critiques down, but don't submit them yet
  - Next week, I will ask you to write these points on the blackboard

23

Review of Lecture

- What did we learn?
  - Data exploration
  - Statistical inference
  - Correlation and regression
  - Evaluating statistical predictions

- Sources
  - Some slides from Bill Howe and Vitaly Shmatikov

- Good reference: NIST Engineering Statistics Handbook `http://www.itl.nist.gov/div898/handbook/index.htm`

- What's next?
  - Pilot project proposals due on Wednesday
  - Measurement module starts next week
    - Focus on paper discussions
  - Paper discussion: 'Mining Your Ps and Qs: Detection of Widespread Weak Keys in Network Devices'

24