COMP 150-4: Human Factors in Security and Privacy

Quantitative Methods

Lecture 8

Prof. Daniel Votipka
Spring 2021

(some slides courtesy of Prof. Adam Aviv and Prof. Michelle L. Mazurek)

Administrivia

- Project proposals due next Tuesday (3/9)
- Need groups by the end of the day
- For current research, check out the SOUPS program
  - https://www.usenix.org/conference/soups2020/technical-sessions
- Diary entries for HW2 are in the box

What we did last time!

- Overview of quantitative data
- Surveys
    - Crowdsourcing

What are we doing today?

- Field studies
- Statistical analysis

FIELD STUDIES

Why a field study?

- Better ecological validity
  - Validate a lab study result
- Because you can’t get the data any other way

Why not a field study?

- Logistically difficult
- Limited piloting / not easy to adjust
  - One shot at your participant pool
- Expensive (money and time)

Plan extremely carefully!

PhishGuru in the real world

- Evaluate phishing training in a real company
- Why use a field study here? Is it necessary?

Logistical problems

- Didn’t include a legitimate email before training to compare click rates
- Control and experimental not 100% parallel
- Participants talked to each other, sharing the training materials
- No one turned in the post-study questionnaire!

- How could these have been avoided?

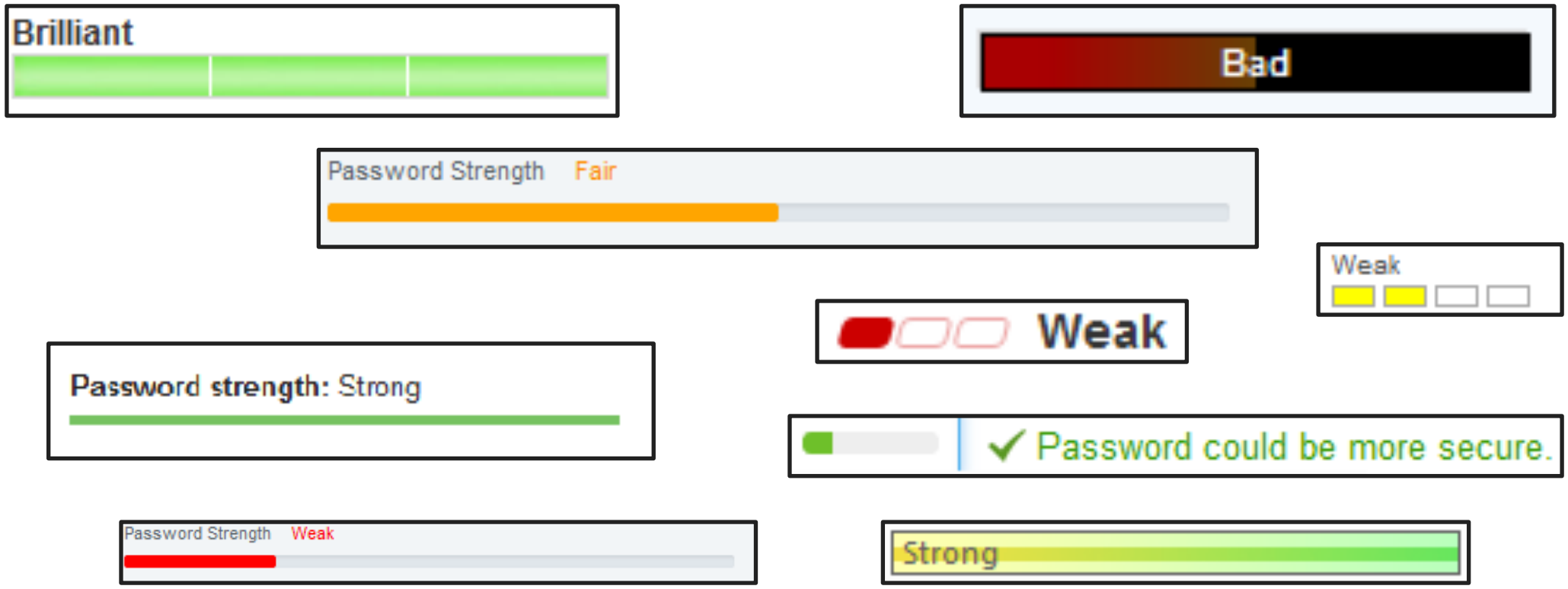
STATISTICS / HYPOTHESIS TESTING

“Classical” hypothesis testing

- Frame a mathematical model describing relationship between input and output variable(s)
- Specify null and alternative hypotheses within this model
- Choose a statistic that (hopefully) can discriminate between the null and alternative hypotheses (probabilisitically)

Running example: Password meters

- Do different password meters help users to create better passwords?



[Ur et. al, USENIX Sec 2012]

Framing a model and hypotheses

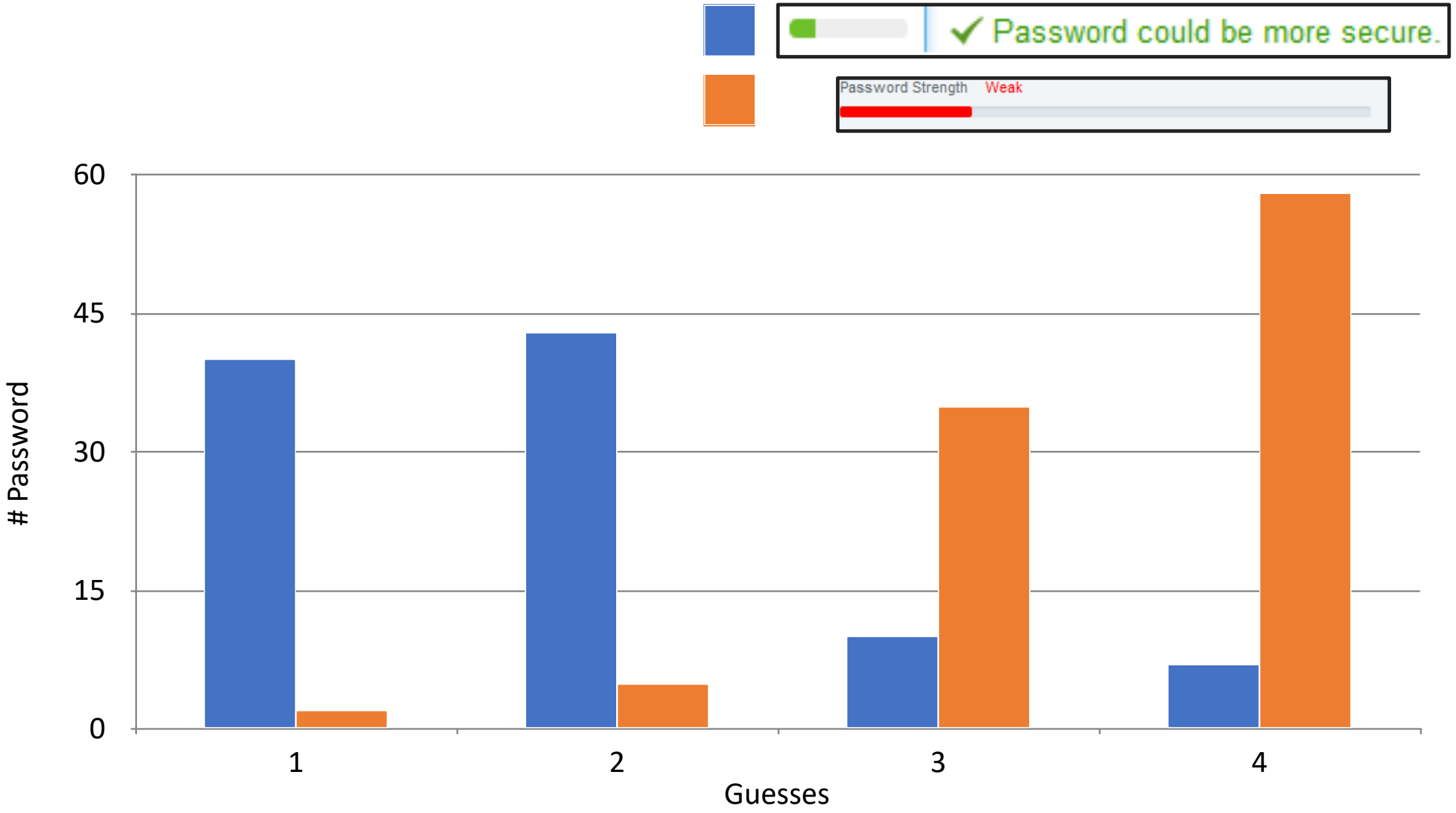
- In general: DV varies with IV(s)
  - Do you expect a direction?
  - Categorical vs. numeric inputs and outputs

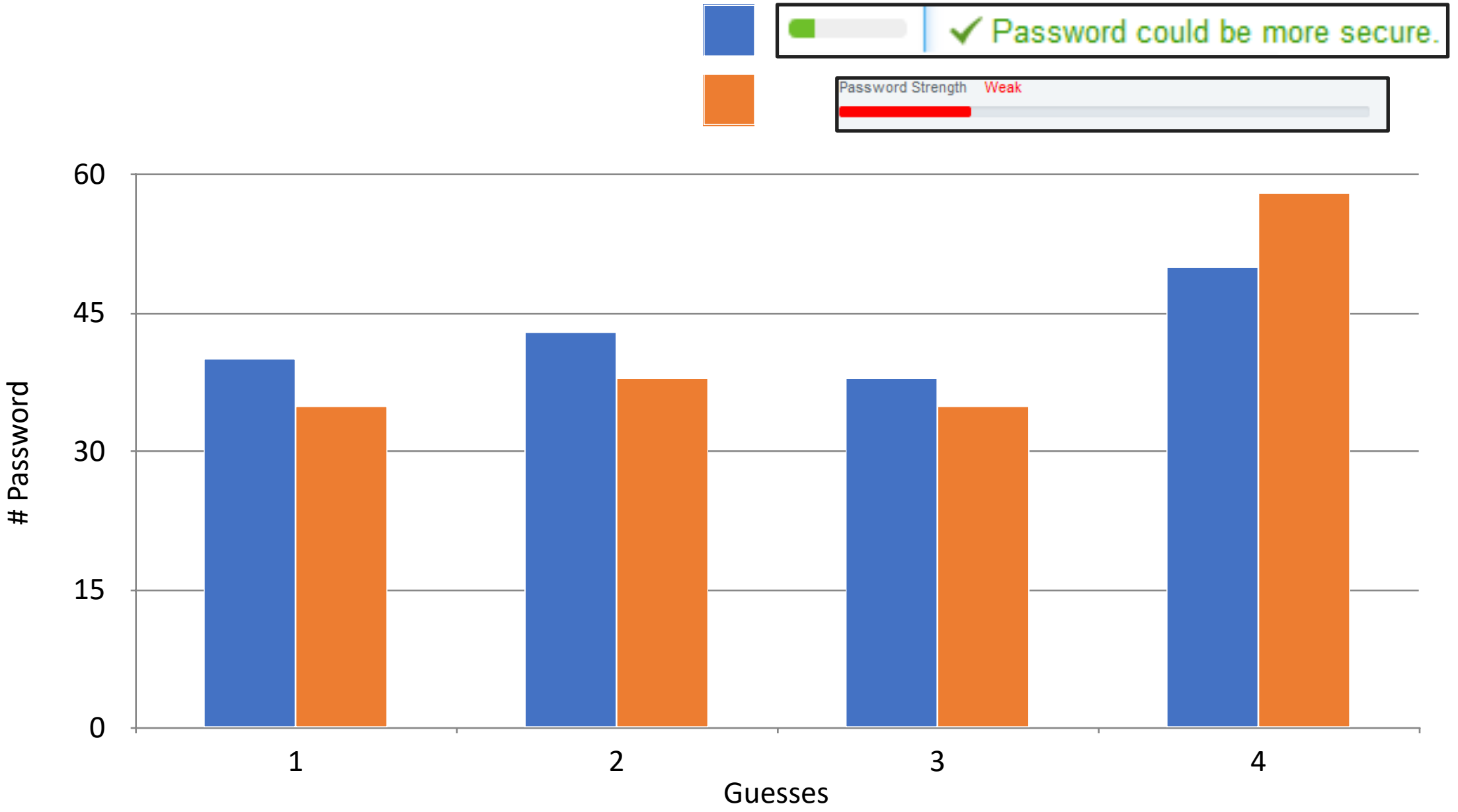
- Null hypothesis: IV *does not* influence DV
- Alt. hypothesis: IV *does* influence DV
  - As framed in model

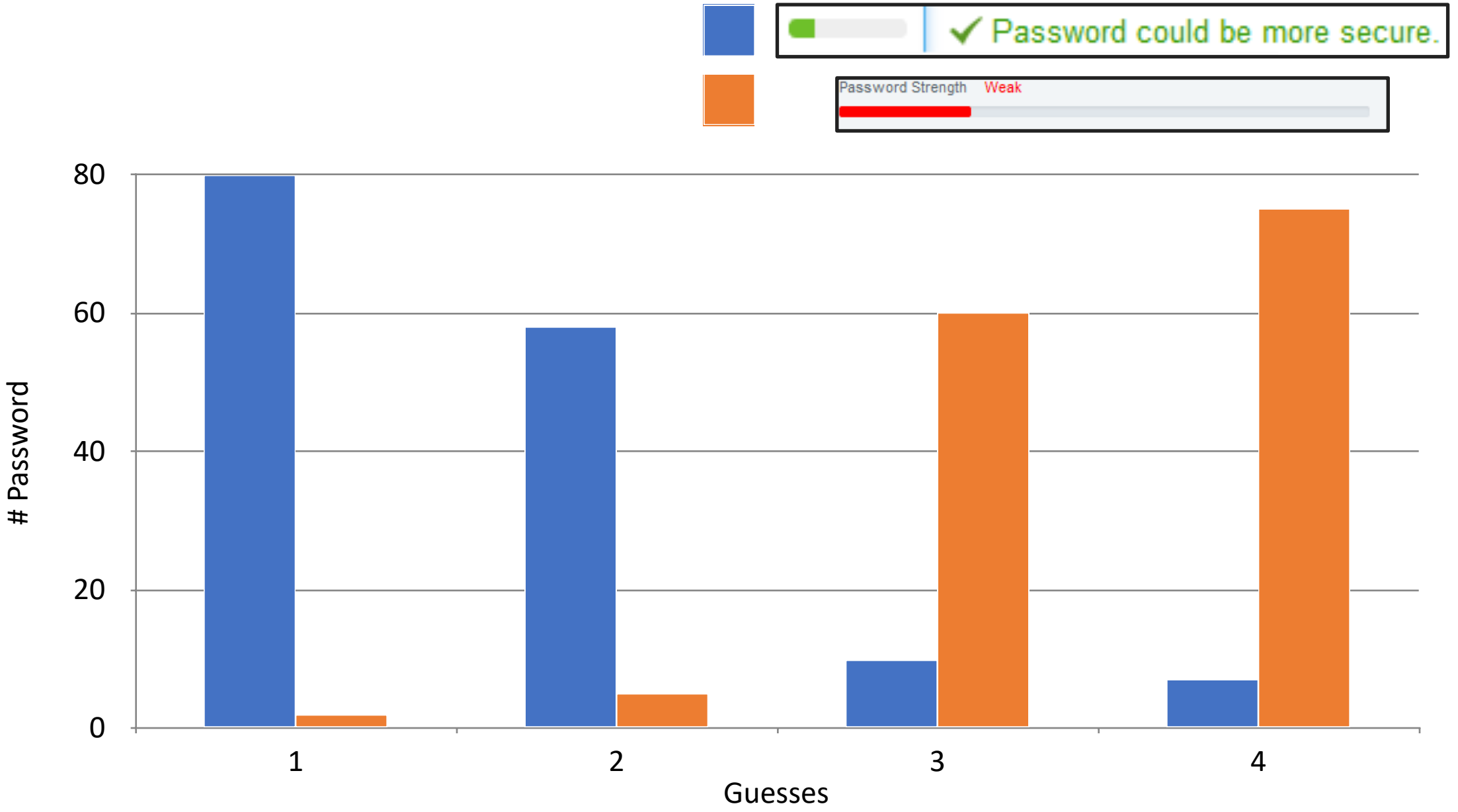
(Running example)

- Password meter that changes colors will produce stronger passwords than all-green meter
  - Has a direction: color >> green
  - IV: categorical; DV: numeric (guess score)
- H0: No difference in strength between meters
- H1: Multicolor >> Green

- What does it mean for one set to be “stronger”?

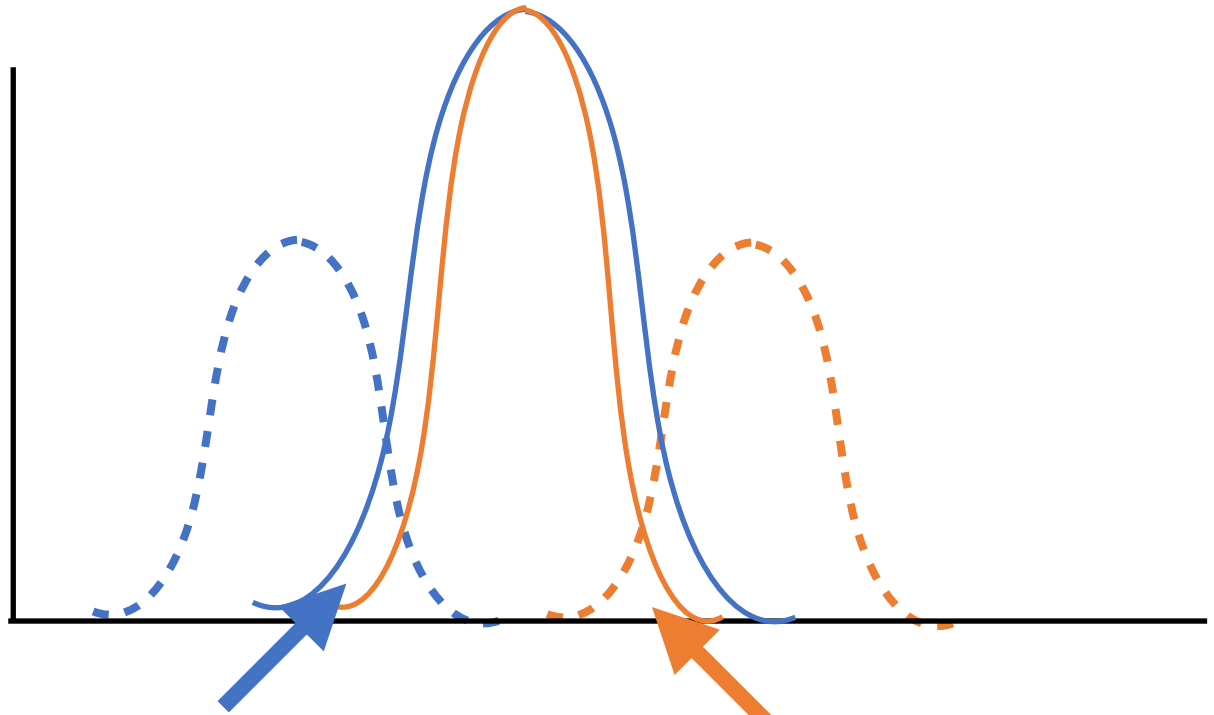

Password could be more secure.
Password Strength Weak
60
45
30
15
0
Password
1
2
3
4
Guesses


Password could be more secure.
Password Strength Weak
60
45
30
15
0
Password
1
2
3
4
Guesses


Password could be more secure.
Password Strength Weak
80
60
40
20
0
Password
1
2
3
4
Guesses

What does a statistical test do?

- Compares tendencies in the data set
  - Typically central tendency: mean, median, etc.
  - Start w/ mean b/c simplest to talk about



What does a statistical test do?

- Compares tendencies in the data set
  - Typically central tendency: mean, median, etc.
  - Start w/ mean b/c simplest to talk about
- We have samples; they have errors
  - Measurement errors
  - Sampling errors
  - Random noise, variation in people, etc.

Running example

- H_0: $m_M = m_G$
- H_1: $m_M > m_G$

What does a statistical test do?

- Find evidence to *reject* the null
  - Or not!
- Does not find evidence to *support* the null
  - In practice: Evidence things are different, but not finding evidence of difference != evidence that things are the same

What kind of evidence?

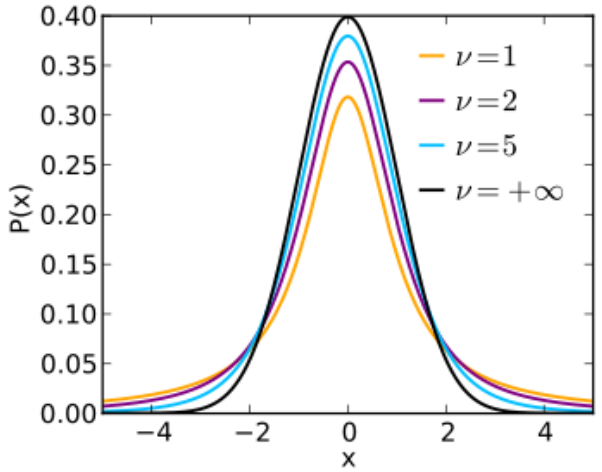
- Pick a statistic that should be different under H_0/H_1
- Calculate theoretical sample distribution for this statistic, under null hypothesis
- “Formally, a p-value is the probability that any given experiment will produce a value of the chosen statistic equal to the observed value in our actual experiment, or something more extreme (in the sense of less compatible with the null hypotheses), when the null hypothesis is true and the model assumptions are correct.”
- Decision rule: p < alpha (typically alpha = 0.05)

Example: Independent samples t-test

- Assumes DV is normally distributed for each condition
- Assumes common variance between conditions
- Assumes means: mu, mu+delta
  - H0: delta = 0, or $mu_A = mu_B$

T-test continued

- Calculate sample distribution under H0
- T-statistic (for null hyp):
  - $(m_A – m_B) / \text{sqrt}(var_A/n_A + var_B/n_B)$
  - Denominator based on variance, sample size
  - Follows T-distribution based on assumptions, degrees of freedom (N-1)
- P-value = area under curve more extreme than observed T-stat
  - Tailedness (in our running example)



T distributions, from Wikipedia. Depends on total N

Calculating p-value (p = 0.382 one sided or 0.764 two sided), from Seltman book

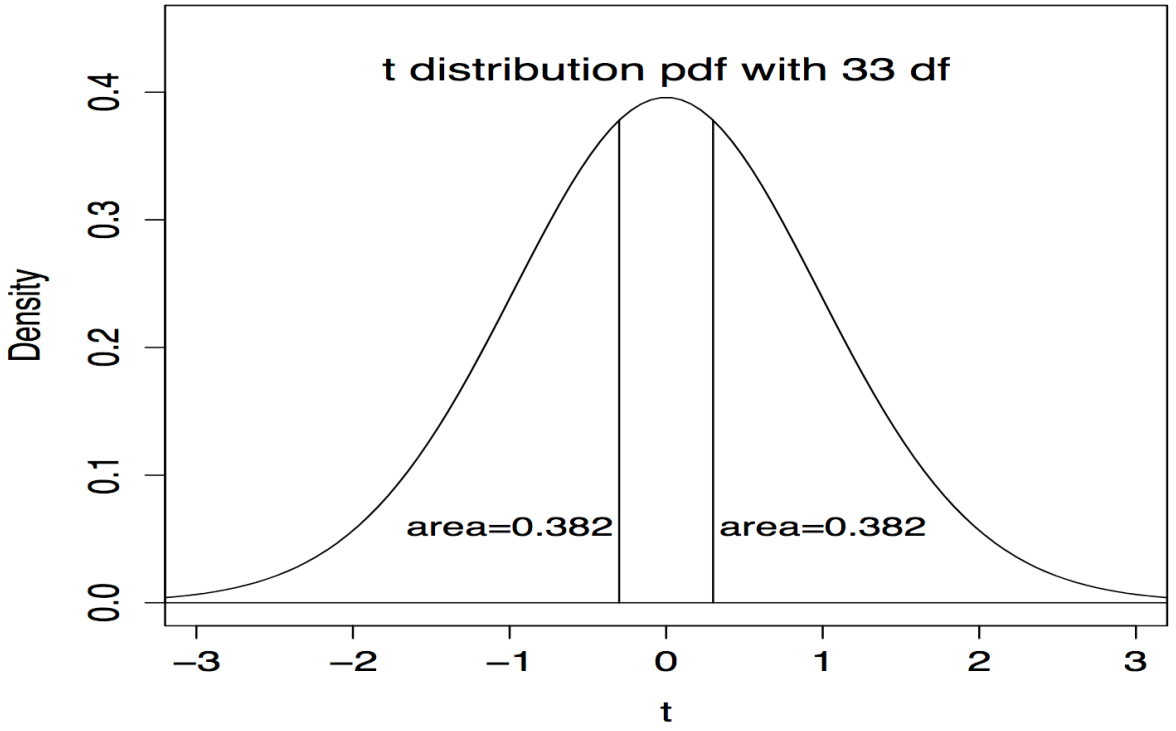


Figure 6.3: Calculation of the p-value for the HCI example

Interpreting p-values

- Model assumptions matter!
- NO ASSUMPTIONS about chance that H0 is true – comes from dist. assuming H0 is true!
- Type I error: We found a difference that isn’t real
- Type II error: We failed to find a real difference
- P-values ONLY BOUND TYPE I

Interpreting p-values

- Small p-value is evidence that H1 is likely
  - Otherwise, bad luck or wonky sample
- Large p-value: H0 is true, or type 2 error
  - Can use power analysis to interrogate this a bit
  - ”Statistical power”: prob. of rejecting null if you should (more later)

- P-values don’t PROVE anything

Interpreting p-values

- Generally, reject null w/ $p < 0.05$
- A p-value is not magic, just probability, and the threshold is arbitrary
- But, reported TRUE or FALSE: *You don't say something is "more significant" because the p-value is lower*

Defining significance

- Statistically significant: It would be unlikely to observe this data if the underlying distributions were the same (e.g. if they had the same mean)

- This doesn’t mean the difference is meaningful!
  - Effect size == strength of the effect
  - Sufficiently large samples can find real but small effects

Running example

- We find that $m_M >> m_G$ (H0 is rejected)
  - Attacker has to guess 2 billion more passwords?
  - Attacker has to guess 2 more passwords?
  - Etc.

P values and multiple testing

- P-values bound Type I error (false positive)
  - You expect this to happen 5% of the time if $p = 0.05$
- What happens if you conduct a lot of statistical tests in one experiment?
- Your cumulative probability of a Type I error can increase dramatically!
- You can and *correct* for this
  - Correctio t p-value thresh
  - Bonferro reshold is 0.05/

>>

p-values and confidence intervals

- 95% CI: we are 95% confident the “true” parameter is between the CI bounds
  - 95% of experiments, true param is within calc. bounds
  - NOT parameter is 95% likely within these bounds!
- Calculated from your sample, with assumptions
- Related to p-value: if CIs do not overlap, $p < 0.05$
- Human judgment about whether this is narrow/wide and how to interpret result

CIs help to interpret results

- Running example: delta guess number
- [2, 10]: significant, not meaningful
- [-10, 5]: not significant, but even if a real difference exists, not likely to be meaningful
- [-1e8, 50]: not significant, don’t have enough info, because meaningful diff is possible

CHOOSING THE RIGHT TEST

Planning

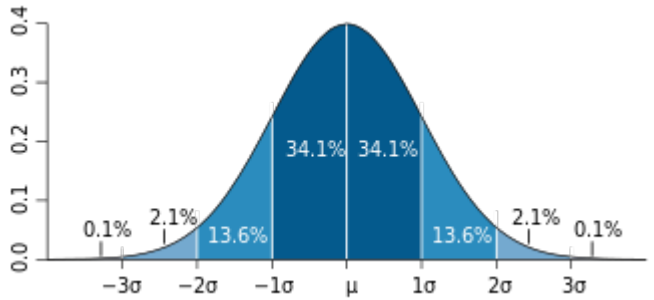
- Choose the test(s) before you collect the data
  - Especially before you look at it!
- Interplay between: what question am I asking? How could I demonstrate a result? What data must be collected for that to work? Etc.
- But, do exploratory analysis / visualization to sanity check your plan and results

What kind of data do you have?

- For input and outcome variables
- Quantitative
  - Discrete (Number of times pw meter was used)
  - Continuous (Average guess number of pw)
- Categorical
  - Binary (pw “passes” or “fails”)
  - Nominal: No order (Color of the pw meter)
  - Ordinal: Ordered (Is the meter lenient, medium, stringent)
- How many of each?!

What kind of data do you have?

- Does your dependent data follow a normal distribution? (You can calculate this!)



http://www.wikipedia.org

- Choice of test depends on normality
  - If so, use parametric tests.
  - If not, use non-parametric tests.

What kind of data do you have?

- Are your data independent?
  - Within vs. between subjects; group effects; time series
  - If not, repeated-measures, mixed models, etc.
  - Can you make them independent? E.g., after – before
  - Independence is usually the least robust assumption!

- Other assumptions (and robustness) about distribution, errors, variance, etc.

http://www.stat.cmu.edu/~hseltman/309/Book/

Reporting a statistical test

- Make it clear what the IV and DVs were
  - Fails surprisingly often
- Be clear which test was used and why
- Supply:
  - p-value
  - Effect size
  - Test statistic (sometimes redundant but people like it)
  - df / sample size

What we did today!

- Field studies
- Statistical analysis

What’s next?

- Ethics!
  - Belmont principles
  - Examples of unethical research
  - IRB @ Tufts

Logistics + Questions?