06- Quantitative studies, statistics

Lorrie Cranor and Blase Ur

January 30, 2014

05-436 / 05-836 / 08-534 / 08-734
Usable Privacy and Security

Carnegie Mellon University
CyLab

institute for SOFTWARE RESEARCH

Engineering & Public Policy

Today!

- Rearranged lectures
- Homework question on designing studies
- Statistics!
  - The main idea
  - Hypothesis testing
  - Correlations
  - Non-independent data
- Homework question on statistics

Designing studies

- What did you pick?
  - A diary study
  - A survey
  - Interviews
  - A usability test
  - Collecting data in the field

Statistics

- In general: analyzing and interpreting data
- Statistical hypothesis testing: is it unlikely the data would like this unless there is actually a difference in real life?

Hypotheses

- **Null hypothesis**: There is no difference
- **Alternative hypothesis**: There is a difference
- You generally either “reject the null hypothesis” (find evidence in support of the alternative hypothesis) or “fail to reject the null hypothesis” (do not find evidence in support of the alternative hypothesis) except with very large samples

P values

- What is the probability that the data would look like this if there's no actual difference?
- Most often, $\alpha = 0.05$, but some people choose 0.01
  - If $p < 0.05$, reject null hypothesis; there is a "significant" difference between Foo and Bar
  - You don't say that something is "more significant" because the p value is lower

P values

- Type I error (false positive)
  - You would expect this to happen 5% of the time if $\alpha = 0.05$
- What happens if you conduct a lot of statistical tests in one experiment?
- Many methods for “correcting” p values
  - Bonferroni correction (multiply p values by the number of tests) is the easiest to calculate but most conservative

P values

- Type II error (false negative)
  - There is actually a difference, but you didn't see evidence of a difference
- Statistical power is the probability of rejecting the null hypothesis if you should
  - You could do a **power analysis**, but this requires that you estimate the effect size

What kind of data do you have?

- Quantitative
  - Discrete (The number of ponies we have)
  - Continuous (A pony's age)
- Categorical
  - Binary (Is it a pony or is it not a pony?)
  - Nominal- no order (Color of the pony)
  - Ordinal- ordered (Is the pony super cool, cool, a little cool, or uncool)

(Pearson's) Chi-squared (χ^2) Test

- (Not covered today) Goodness of fit: Does the distribution we observed differ from a theoretical distribution?
- Test of independence: Are two variables independent of each other?
  - Correlation example: Is gender (male, female) correlated with a pony's favorite color?
  - Causation example: If we feed a pony hay, is it more likely to think privacy is important than if we feed it pop-tarts?

Contingency tables

- Rows are one variable, columns the other

CreateAnnoying Counts:

	0	1
0	161	32
1	165	33
2	168	34
3	170	30
4	164	32
5	161	35
6	167	32
7	129	60
8	128	61
9	154	40
10	153	40
11	154	38
12	142	42
13	121	67
14	124	76

Percentages:

	0	1
0	"83.42%"	"16.58%"
1	"83.33%"	"16.67%"
2	"83.17%"	"16.83%"
3	"85%"	"15%"
4	"83.67%"	"16.33%"
5	"82.14%"	"17.86%"
6	"83.92%"	"16.08%"
7	"68.25%"	"31.75%"
8	"67.72%"	"32.28%"
9	"79.38%"	"20.62%"
10	"79.27%"	"20.73%"
11	"80.21%"	"19.79%"
12	"77.17%"	"22.83%"
13	"64.36%"	"35.64%"
14	"62%"	"38%"

- $\chi^2 = 97.013$, df = 14, p = 1.767e-14

Contrasts

- If we determine that the variables are dependent, we may compare conditions
- Planned vs. unplanned **contrasts**
- You can safely make the same number of planned contrasts as the degrees of freedom as long as you choose what you will compare **before** you look at the results
- If you perform unplanned/post-hoc comparisons, be sure to correct p values!

Contrasts in the meters paper

- “We ran pairwise contrasts comparing each condition to our two control conditions, no meter and baseline meter. In addition, to investigate hypotheses about the ways in which conditions varied, we ran planned contrasts comparing tiny to huge, nudge-16 to nudge-comp8, half-score to one-third-score, text-only to text-only half-score, half-score to text-only half-score, and text-only half-score to bold text-only half-score.”

Choosing a statistical test

- Use χ^2 if you are testing if one categorical variable (usually the assigned condition or a demographic factor) impacts another categorical variable
  - If you have fewer than 5 data points in a single cell, use Fisher's Exact Test
- Do not use χ^2 if you are testing quantitative outcomes!

Choosing a statistical test

- Do your data follow a normal (Gaussian) distribution? (You can calculate this!)

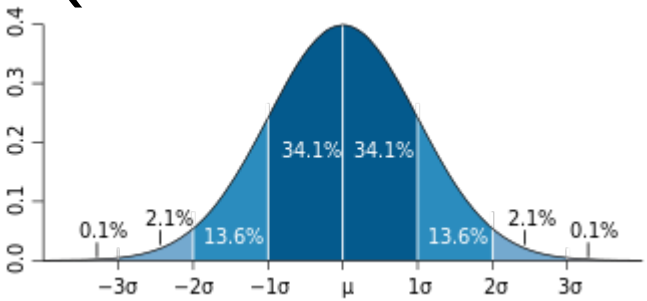


Image from http://www.wikipedia.org

  - If so, use parametric tests. If not, use non-parametric tests

- Are your data independent?
  - If not, repeated-measures, mixed models, etc.

Continuous/ordinal data

- If you want to compare “which is bigger?”
- Normal, continuous data (compare mean):
  - 2 conditions: t-test
  - 3+ conditions: ANOVA
- Non-normal data / ordinal data (does one group tend to have larger values?)
  - 2 conditions: Mann-Whitney U (AKA Wilcoxon rank-sum test)
  - 3+ conditions: Kruskal-Wallis

Continuous/ordinal data

```
Initial Kruskal-Wallis Test (uncorrected)
-----------------------------------------
######
Column name:  Length
######
Medians:
   0    1    2    3    4    5    6    7    8    9   10   11   12   13   14
 9.0 11.0 10.0 10.0 11.0 11.0 11.0 12.5 12.0 12.0 11.0 10.0 10.0 10.5 11.0
Means:
       0        1        2        3        4        5        6        7        8
10.37436 11.97475 11.54187 11.29500 11.44162 11.61929 11.42786 14.95789 14.34211
StdDevs:
       0        1        2        3        4        5        6        7        8
2.851724 3.693095 3.780090 3.592003 3.245545 3.333845 3.467855 7.263616 8.141852

      Kruskal-Wallis rank sum test

data:  num.data[[curcol]] and num.data$condition
Kruskal-Wallis chi-squared = 146.0215, df = 14, p-value < 2.2e-16
```

Continuous/ordinal data

```
Pairwise MWU tests
------------------
######
Column name:  Length
######
Uncorrected p-values:
           Pvalue
0-1   3.109043e-07
0-2   4.105790e-04
0-3   7.983785e-04
0-4   7.958007e-05
0-5   1.350688e-05
0-6   1.016164e-04
0-7   1.033295e-13
0-8   1.176549e-13
0-9   8.837726e-15
0-10  1.858190e-05
0-11  6.258968e-04
0-12  8.110757e-02
0-13  2.052447e-04
0-14  4.541805e-08
1-2   9.951006e-02
1-3   3.242047e-02
1-4   2.070811e-01
1-5   4.269510e-01
1-6   1.266861e-01
1-7   3.817210e-04
1-8   5.517246e-03
1-9   1.084607e-03
1-10  3.500755e-01
1-11  3.582721e-02
1-12  8.649624e-04
1-13  2.661746e-01
1-14  5.199768e-01
4-5   6.351934e-01
7-8   3.703923e-01
9-10  4.844768e-05
7-13  4.001252e-05
12-13 4.240731e-02
13-14 9.985554e-02

Holm corrected p-values:
           Pvalue Star
0-1   9.016223e-06    *
0-2   8.211580e-03    *
0-3   1.437081e-02    *
0-4   1.909922e-03    *
0-5   3.781925e-04    *
0-6   2.337178e-03    *
0-7   3.306545e-12    *
0-8   3.647303e-12    *
0-9   2.916449e-13    *
0-10  5.017114e-04    *
0-11  1.189204e-02    *
0-12  8.921833e-01
0-13  4.515384e-03    *
0-14  1.362542e-06    *
1-2   9.951006e-01
1-3   4.538866e-01
1-4   1.000000e+00
1-5   1.000000e+00
1-6   1.000000e+00
1-7   8.016142e-03    *
1-8   8.275868e-02
1-9   1.735370e-02    *
1-10  1.000000e+00
1-11  4.657537e-01
1-12  1.470436e-02    *
1-13  1.000000e+00
1-14  1.000000e+00
4-5   1.000000e+00
7-8   1.000000e+00
9-10  1.211192e-03    *
7-13  1.040326e-03    *
12-13 5.088877e-01
13-14 9.951006e-01
```

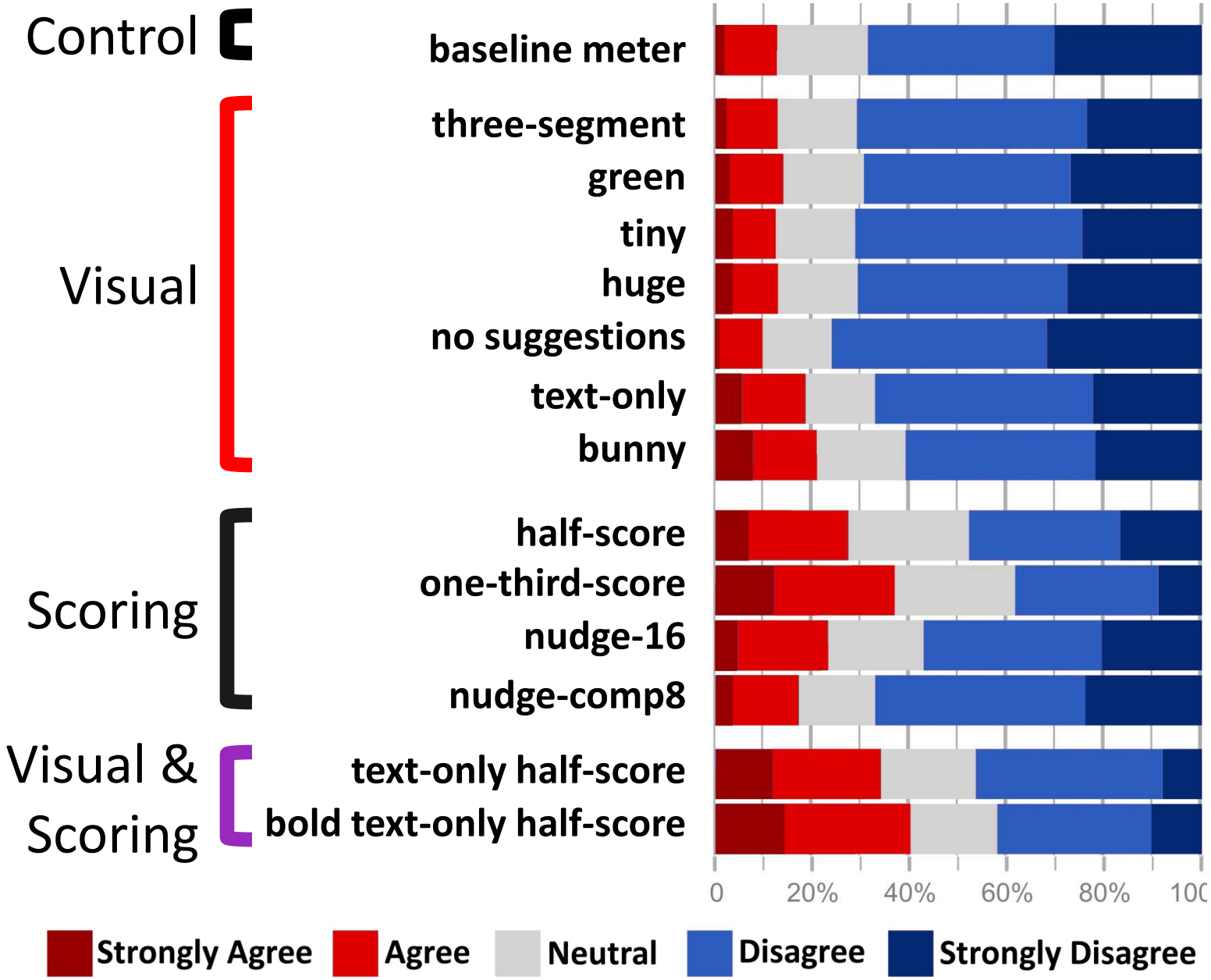
What are Likert-scale data?

- Respond to the following statement: Ponies are magical.
  - 7: Strongly agree
  - 6: Agree
  - 5: Mildly agree
  - 4: Neutral
  - 3: Mildly disagree
  - 2: Disagree
  - 1: Strongly disagree

What are Likert-scale data?

- Some people treat it as continuous (meh!)
- Other people treat it as ordinal (ok!)
  - You can use Mann-Whitney U / Kruskal-Wallis
- A simple way to compare the data is to “bin” (group) the data into binary “agree” and “not agree” categories (ok!)
  - You can use χ^2

Password meter annoying



Regressions

- What is the relationship among variables?
  - Generally one outcome (dependent variable)
  - Often multiple factors (independent variables)
- The type of regression you perform depends on the outcome
  - Binary outcome: logistic regression
  - Ordinal outcome: ordinal / ordered regression
  - Continuous outcome: linear regression

Example regression

- Outcome: completed pony race (or not)
- Independent variables:
  - Age
  - Number of prior races
  - Diet: hay or pop-tarts
  - (Indicator variables for color categories)
  - Etc.

Interactions in a regression

- Normally, outcome = $ax_1 + bx_2 + c + \ldots$
- Interactions account for situations when two variables are not simply additive. Instead, their interaction impacts the outcome
  - e.g., Maybe brown horses, and only brown horses, get a much larger benefit from eating pop-tarts before a race
- Outcome = $ax_1 + bx_2 + c + d(x_1x_2) + \ldots$

Example regression

```
model- binary
************************

Cumulative Link Mixed Model fitted with the adaptive Gauss-Hermite
quadrature approximation with 20 quadrature points

formula: correct ~ gender + chosen + programming + age + alreadydid +      experiment + chosen * programming + alreadydid *
experiment + chosen * experiment + (1 | uid)
data:    data

 link  threshold nobs logLik  AIC     niter     max.grad cond.H
 logit flexible  1832 -745.62 1565.24 53(13128) 1.23e-05 6.2e+05

Random effects:
       Var Std.Dev
uid 0.7885   0.888
Number of groups:  uid 223

Coefficients:
                              Estimate Std. Error z value Pr(>|z|)
genderI prefer not to answer  0.475650   1.308540   0.363 0.716234
genderMale                   -0.017708   0.205080  -0.086 0.931192
chosenb                      -1.739132   0.472334  -3.682 0.000231 ***
chosenc                       0.644282   0.630716   1.022 0.307014
chosend                       0.571554   0.600672   0.952 0.341339
chosene                       1.541800   0.778734   1.980 0.047717 *
chosenf                      -0.481121   0.510956  -0.942 0.346393
choseng                      -3.726763   0.503302  -7.405 1.32e-13 ***
chosenh                      -1.706179   0.479596  -3.558 0.000374 ***
choseni                      -0.280454   0.530171  -0.529 0.596813
chosenj                      -0.348918   0.521329  -0.669 0.503313
programming1                 -0.208038   0.580828  -0.358 0.720213
age                          -0.017786   0.008671  -2.051 0.040242 *
alreadydid                    0.173464   0.041030   4.228 2.36e-05 ***
experiments                   0.139865   0.534377   0.262 0.793527
chosenb:programming1          0.485281   0.656680   0.739 0.459913
chosenc:programming1          0.278906   0.893211   0.312 0.754849
chosend:programming1          1.243753   0.958374   1.298 0.194365
```

What if you have lots of questions?

- If we ask 40 privacy questions on a Likert scale, how do we analyze this survey?
- One technique is to compute a “privacy score” by adding their responses
  - Make sure the scales are the same (e.g., don’t add agreement with “privacy is dumb” and “privacy is smart”… reverse the scale)
  - You should verify that responses to the questions are correlated!

Correlation

- Usually less good: Pearson correlation
  - Requires that both variables be normally distributed
  - Only looks for a linear relationship
- Often preferred: Spearman's rank correlation coefficient (Spearman's ρ)
  - Evaluates a relationship's monotonicity (always going in the same direction or staying the same)

What if you have lots of questions?

- Another option: factor analysis, which evaluates the latent (underlying) factors
  - You specify *N*, a number of factors
  - Puts the questions into N groups based on their relationships
  - You should examine factor loadings (how well each latent factor correlates with a question)
  - Generally, you want questions to load primarily onto a single factor to be confident

In groups:

- What statistical analysis would you do?
  - You randomly assign ponies to have private stalls or public stalls. Does this assignment impact whether they finish their next race?
  - …and does this impact their finishing time?
  - You are analyzing interviews of 10 pony trainers and are reporting what these trainers think ponies say (“neigh,” “ring-ding-ding,” etc.)
  - Do gender, state of residence, and education level impact ponies’ level of privacy concern?

Independence

- Why might your data in UPS experiments not be independent?
  - Non-independent sample (bad!)
  - The inherent design of the experiment (ok!)
- If you have two data points of ponies' race completion times (before and after some treatment), can you actually do a single test that assumes independence to compare conditions?

Non-independence

- Repeated measures (multiple measurements of the same thing)
  - e.g., before and after measurements of a pony's time to finish a race
- Paired t-test (two samples per participant, two groups)
- Repeated measures ANOVA (more general)

Non-independence

- For regressions, use a mixed model
  - "Random effects" based on hierarchy/group
- Case 1: Many measurements of each pony
- Case 2: The ponies have some other relationship. e.g., there are 100 ponies each trained by one of 5 trainers. The identity of the trainer might impact a whole class of ponies' performance.