CAP 5768 – Intro to Data Science

Lecture 4: Hypothesis Testing

Oge Marques, PhD

Professor

College of Engineering and Computer Science

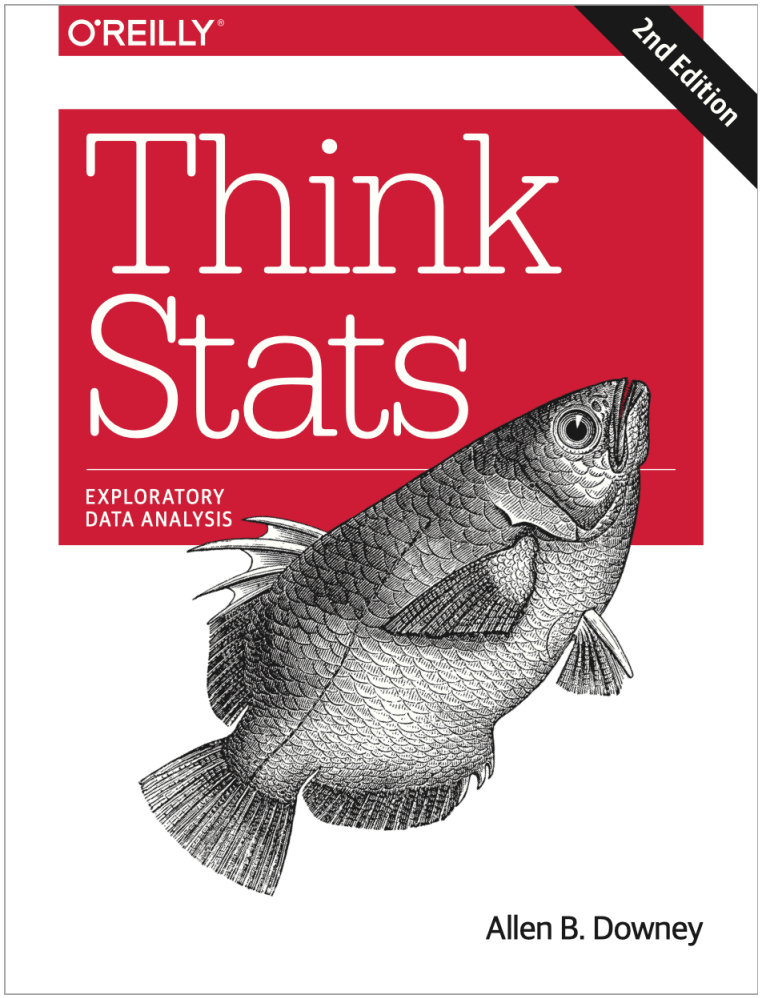
College of Business

FAU FLORIDA ATLANTIC UNIVERSITY

Highlights from

"Think Stats 2/e"

by Allen B. Downey



Chapter 9: Hypothesis testing

Classical hypothesis testing

Testing a difference in means

Other test statistics

Testing a correlation

Testing proportions

Chi-squared tests

Classical hypothesis testing

- Scope: going from "apparent effects" to "rigorous hypothesis testing"
- The fundamental question we want to address is *whether the effects we see in a sample are likely to appear in the larger population*.

- Example:
  - In the NSFG sample we see a difference in mean pregnancy length for first babies and others.
  - We would like to know if that effect reflects a real difference for women in the U.S., or if it might appear in the sample by chance.

Classical hypothesis testing

- The goal of classical hypothesis testing is to answer the question:

  "Given a sample and an apparent effect, what is the probability of seeing such an effect by chance?"

- The logic of the process is similar to a proof by contradiction.
  - To prove a mathematical statement, A, you assume temporarily that A is false.
  - If that assumption leads to a contradiction, you conclude that A must actually be true.

Classical hypothesis testing: steps

1. Quantify the size of the apparent effect by choosing a **test statistic**.
2. Define a **null hypothesis**, which is a model of the system based on the assumption that the apparent effect is not real.
3. Compute a **p-value**, which is the probability of seeing the apparent effect if the null hypothesis is true.
4. **Interpret the result**.

   If the p-value is low, the effect is said to be **statistically significant**, which means that it is unlikely to have occurred by chance.

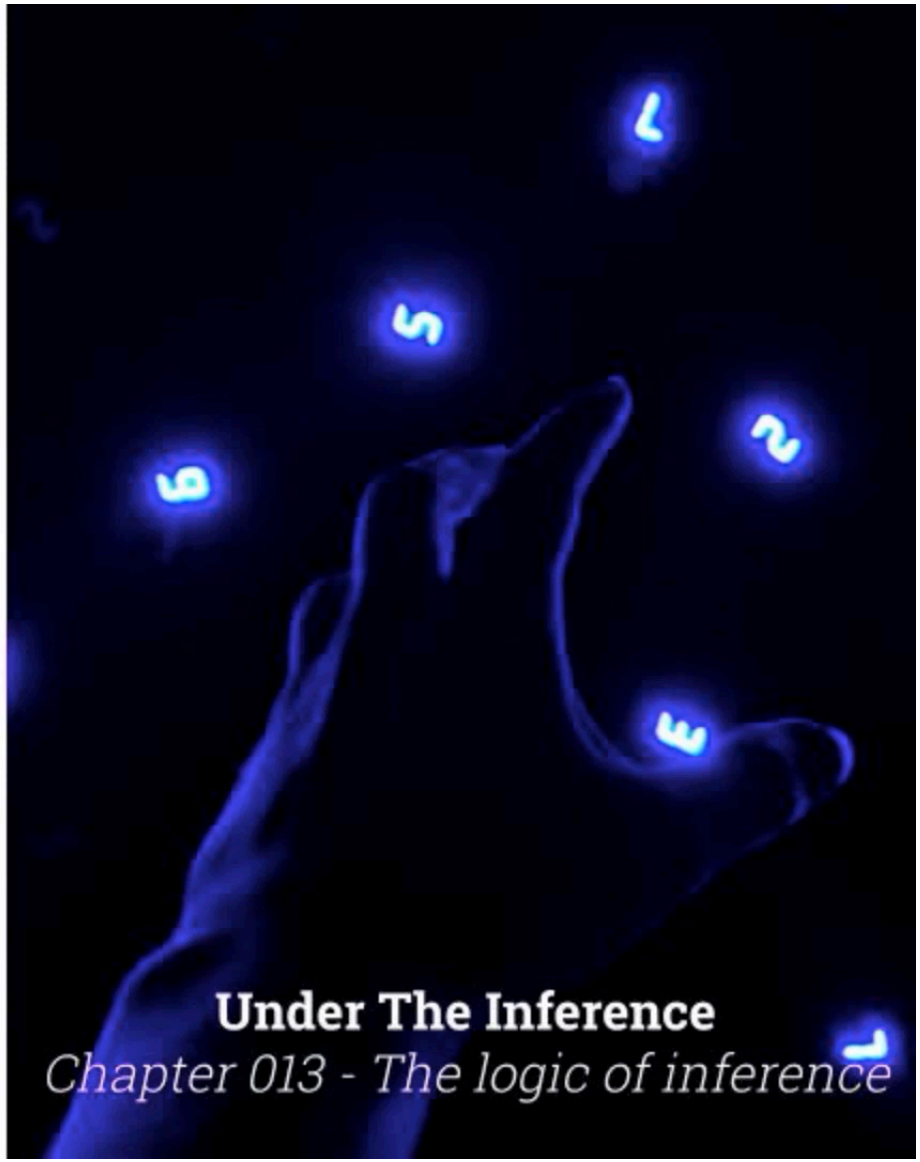
   In that case we infer that the effect is more likely to appear in the larger population.

Classical hypothesis testing: interpreting the result

- By convention, 5% is the threshold of statistical significance.
  - If the p-value is less than 5%, the effect is considered significant; otherwise it is not.
- But **the choice of 5% is arbitrary**, and the p-value depends on the choice of the test statistics and the model of the null hypothesis.
  - So **p-values should not be considered precise measurements**.
- Downey recommends interpreting p-values according to their order of magnitude:
  - if the p-value is **less than 1%**, the effect is **unlikely to be due to chance**;
  - if it is **greater than 10%**, the effect can **plausibly be explained by chance**;
  - p-values **between 1% and 10%** should be considered **borderline**.

Statistical Thinking

Cassie Kozyrkov
Head of Decision Intelligence, Google
@quaesita



Testing a difference in means

- Hypothesis: the pregnancy length is longer for first babies.
- Test statistic: **difference in mean** pregnancy length between 2 groups.
- Null hypothesis: the distributions for the two groups are the same.
- One way to model the null hypothesis is by **permutation**; that is, we can take values for first babies and others and shuffle them, treating the two groups as one big group.
- If you run the code, you'll see that the computed p-value is ~0.17, which means that we expect to see a difference as big as the observed effect about 17% of the time.
- So this effect is **not statistically significant**.

Testing a difference in means

The CDF intersects the observed difference at 0.83, which is the complement of the p-value, 0.17.

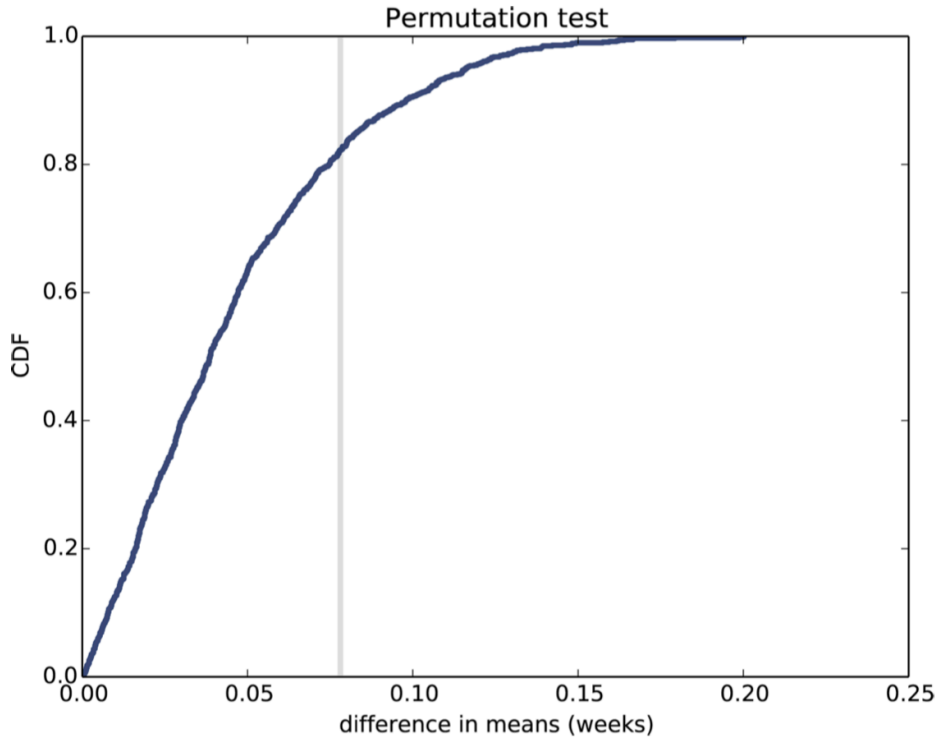


Figure 9-1. CDF of difference in mean pregnancy length under the null hypothesis

Testing a difference in means

- If the hypothesis under test is that **first babies come late**, the appropriate test statistic is **the raw difference** between first babies and others, rather than the absolute value of the difference.
- This is **a one-sided test** because it only counts one side of the distribution of differences.
- The previous test, using both sides, is **two-sided.**

- In that case, **the p-value is smaller (~0.09)**, because we are testing a more specific hypothesis.
- But in this example, the result is **still not statistically significant**.

Testing a difference in means

- If we run the same analysis with **birth weight**, the computed p-value is 0.

- After **1000 attempts**, the simulation never yields an effect as big as the observed difference, 0.12 lbs.

- So **we would report p < 0.001**, and conclude that **the difference in birth weight is statistically significant**.

Difference in standard deviation

- Earlier in the course (in an informal/EDA fashion) we saw some evidence that first babies are more likely to be early or late, and less likely to be on time.
- So we might **hypothesize that the standard deviation is higher**.

- This is **a one-sided test** because the hypothesis is that the standard deviation for first babies is higher, not just different.
- The p-value is 0.09, which is **not statistically significant**.

Testing correlation

- Example: In the NSFG data set, the correlation between birth weight and mother's age is about 0.07. It seems like older mothers have heavier babies. But could this effect be due to chance?
- Hypothesis: older mothers have heavier babies.
- Test statistic: **absolute value** of the **Pearson's correlation** between birth weight and mother's age.
- Null hypothesis: there is no correlation between mother's age and birth weight.
- By **shuffling** the observed values, we can simulate a world where the distributions of age and birth weight are the same, but where the variables are unrelated.

Testing correlation

- **Results**
  - The actual correlation is 0.07.
  - The computed p-value is 0; after 1000 iterations the largest simulated correlation is 0.04.
  - So **although the observed correlation is small, it is statistically significant**.

- This example is a reminder that **"statistically significant" does not always mean that an effect is important, or significant in practice**.
- It only means that it is **unlikely to have occurred by chance**.

Testing proportion

- **Example**: Is this die crooked?

Value	1	2	3	4	5	6
Frequency	8	9	19	5	8	11

- Hypothesis: it is a crooked die.
  - To test this hypothesis, we can compute the expected frequency for each value, the difference between the expected and observed frequencies, and the total absolute difference.
  - In this example, we expect each side to come up 10 times out of 60; the deviations from this expectation are -2, -1, 9, -5, -2, and 1; so **the total absolute difference is 20.**
  - How often would we see such a difference by chance?

Testing proportion

- **Example**: Is this die crooked?

Value	1	2	3	4	5	6
Frequency	8	9	19	5	8	11

- The data are represented as a list of frequencies: the observed values are [8, 9, 19, 5, 8, 11]; the expected frequencies are all 10.
- Test statistic: **the sum of the absolute differences**.
- Null hypothesis: the die is fair.
- **Results**: The p-value for this data is 0.13, which means that if the die is fair we expect to see the observed total deviation, or more, about 13% of the time.
  - So **the apparent effect is not statistically significant**.

Chi-squared tests

- Popular test statistic for proportions.

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$$

- O_i are the observed frequencies and E_i are the expected frequencies.

Chi-squared test for the suspicious die

```
class DiceChiTest(DiceTest):

    def TestStatistic(self, data):
        observed = data
        n = sum(observed)
        expected = np.ones(6) * n / 6
        test_stat = sum((observed - expected)**2 / expected)
        return test_stat
```

- Squaring the deviations (rather than taking absolute values) gives more weight to large deviations.
- Dividing through by expected standardizes the deviations, although in this case it has no effect because the expected frequencies are all equal.
- Using this test, we get a substantially smaller p-value (~0.04).
- What should we conclude from it?

The suspicious die: conclusions

- This example demonstrates an important point: **the p-value depends on the choice of test statistic and the model of the null hypothesis**, and sometimes these choices determine whether an effect is statistically significant or not.
- The p-value using the chi-squared statistic is 0.04, substantially smaller than what we got using total deviation, 0.13.
  - If we take the 5% threshold seriously, we would consider this effect statistically significant.
  - But considering the two tests together, we could say that the results are borderline.
  - We should not rule out the possibility that the die is crooked, but… would you convict the accused cheater?

Chi-squared test of pregnancy length

- A few slides ago, we looked at pregnancy lengths for first babies and others, and concluded that the apparent differences in mean and standard deviation are not statistically significant.
- But, earlier in the course, we saw several apparent differences in the distribution of pregnancy length, especially in the range from 35 to 43 weeks.
- To see whether those differences are statistically significant, we can use a test based on a chi-squared statistic.

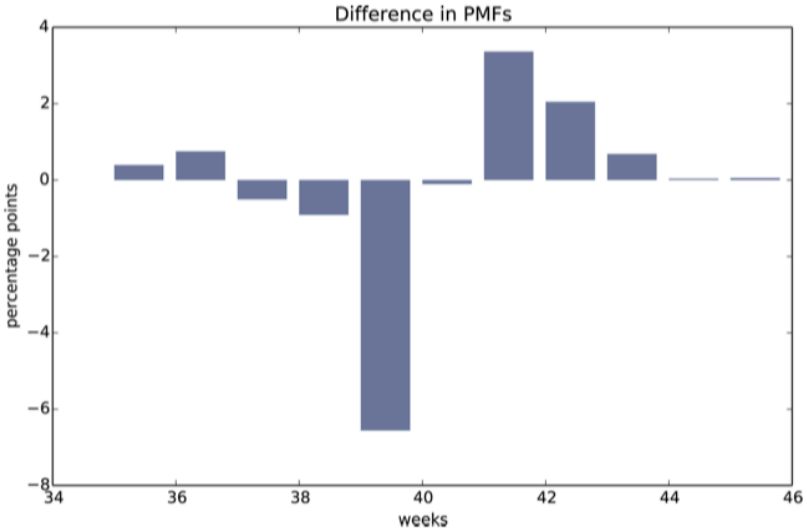


Figure 3-2. Difference, in percentage points, by week

Chi-squared test of pregnancy length

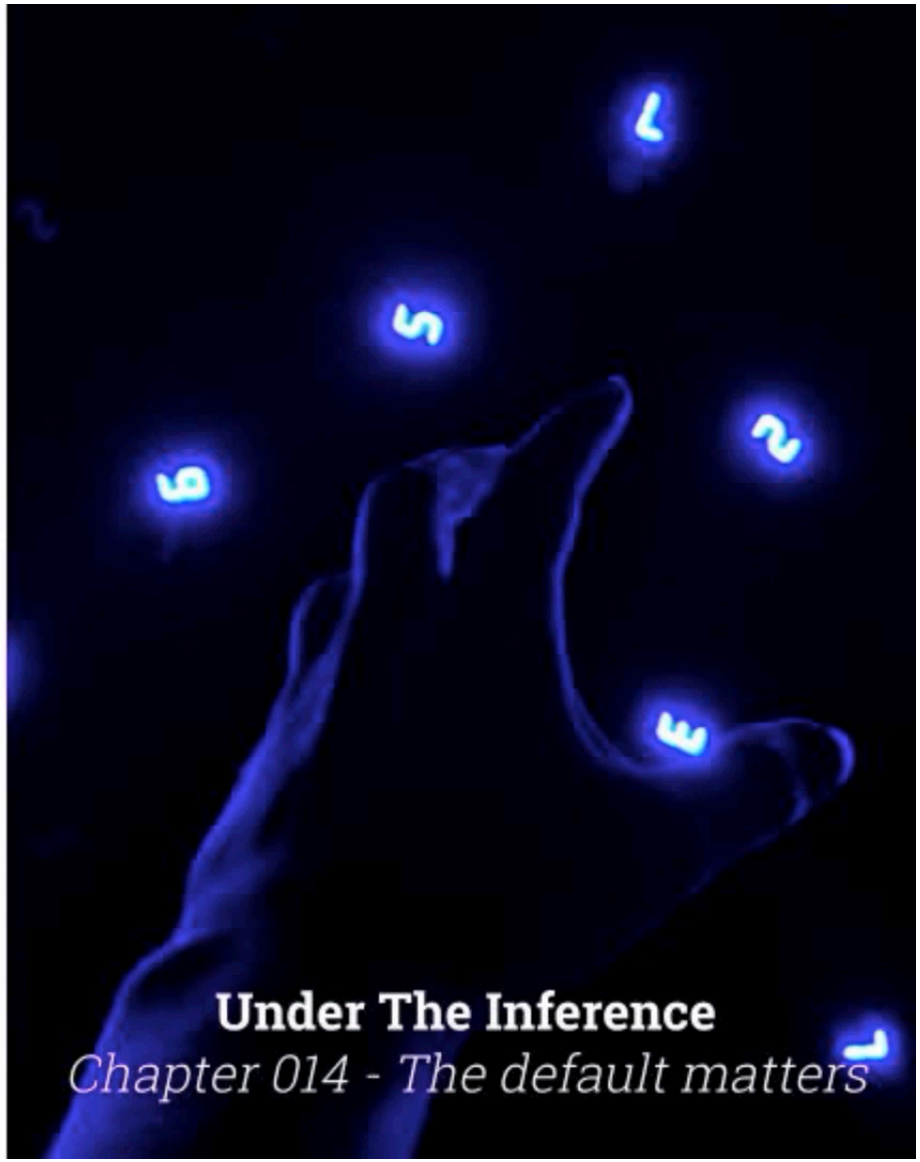
- The data are represented as two lists of pregnancy lengths.
- The null hypothesis is that **both samples are drawn from the same distribution**.
- We generate simulated data by shuffling the pooled sample and splitting it into two parts.
- For the NSFG data the total chi-squared statistic is 102, which doesn't mean much by itself.
- But after 1,000 iterations, the largest test statistic generated under the null hypothesis is 32.
- We conclude that the observed chi-squared statistic is unlikely under the null hypothesis, so **the apparent effect is statistically significant**.

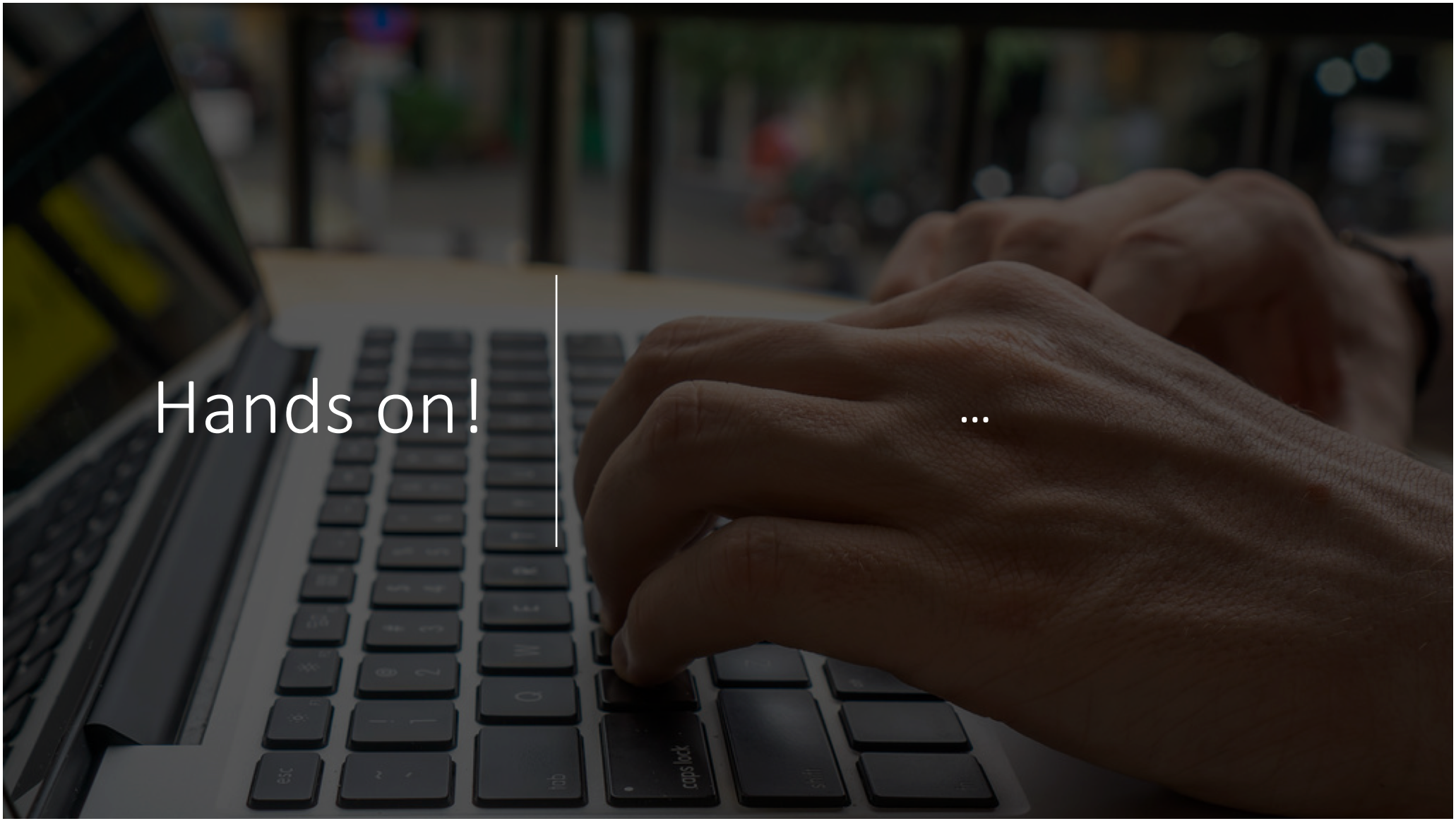
Chi-squared test of pregnancy length

- This example demonstrates a **limitation of chi-squared tests**:
  - They indicate that there is a difference between the two groups, but they don't say anything specific about what the difference is.
- From the companion code:
  - "If we specifically test the deviations of first babies and others from the expected number of births in each week of pregnancy, the results are statistically significant with a very small p-value.
    But **at this point we have run so many tests, we should not be surprised to find at least one that seems significant**."

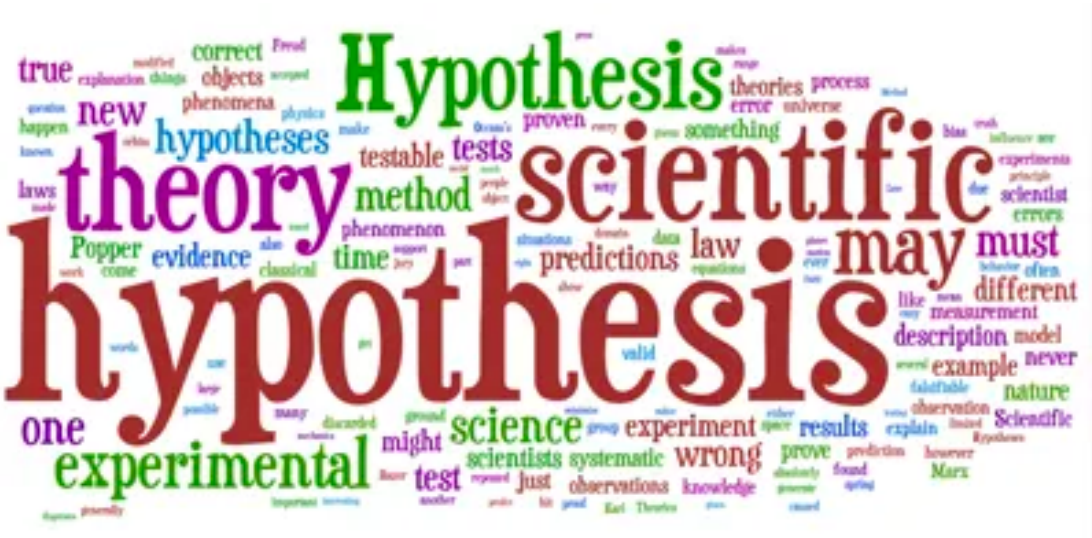
Statistical Thinking

Cassie Kozyrkov
Head of Decision Intelligence, Google
@quaesita




Hands on!
…

Assignment 5: Hypothesis testing



Concluding remarks