Research Methods

Empirical Sampling Distributions 2

Dr. Sven Magg, Prof. Dr. Stefan Wermter

Knowledge Technology

http://www.informatik.uni-hamburg.de/WTM/

Plan for today!

1. Quick Recap
   a) Statistical tests and sampling distributions
   b) Monte-Carlo Tests and Bootstrapping
2. Randomization
3. Data Collection Summary
4. The Final Experiment

Hypothesis Testing

Following Neyman-Pearson:

1. State null Hypothesis H_0 and alternate hypothesis H_1
2. Determine acceptable α and β errors
3. Gather a sample statistic x (*run experiment*)
4. Find sampling distribution N_h, assuming H_0 is true
5. Determine cut-off points c^+ and c^- such that $\mathrm{P}(N_h \geq c^+) + P(N_h \leq c^-) \leq \alpha$
6. **Decide:** If $(x \geq c^+)\ or\ (x \leq c^-)$, reject H_0
   - Reject H_0 if x falls into rejection regions defined by α

- Problem is always to find a suitable sampling distribution

Sampling distributions

- Different ways to get sampling distributions
  - Exact distributions
    - Derived analytically/mathematically
  - Estimated distributions
    - Central Limit Theorem (CLT)
    - Z-distribution (standard normal distribution)
    - t-Distributions
    - Fisher’s z-distribution
  - Determining sampling distributions empirically
    - Monte-Carlo Tests
    - Bootstrapping
    - Randomization

Monte-Carlo Simulation

- If we know the parameters of the population we draw from, we can
  - treat sampling as a stochastic simulation
  - create a probability distribution by drawing pseudo-samples

- **Monte-Carlo Simulation**:
  1. Determine population parameters and test statistic θ
     a) For i = 1 to K
     b) Draw pseudo-sample of size N from the population
     c) Calculate and record test statistic $\theta_i{}^*$ for pseudo-sample
  2. Use the distribution of θ^* to determine probability of original sample under H_0

Monte-Carlo Example

- We have two populations A and B and two samples S_A and S_B of sizes N_A and N_B
- As statistic θ we use the difference of the median
  $\theta = median(S_A) - median(S_B)$

- Generate probability distribution:
  For i = 1 to K
  a. Draw pseudo-samples ${S_A}^*$ of size N_A from population A and ${S_B}^*$ of size N_B from B
  b. Calculate and record ${\theta_i}^* = f({S_A}^*, {S_B}^*)$
- Find probability of θ using the distribution of θ^*

Monte-Carlo Sampling

- Advantages
  - Straightforward and usually simple to calculate
  - Cheap for most computer science problems
  - Can be used for any statistic

- Disadvantages
  - We have to know the population parameters to know where to draw samples from
  - Often the population parameters are not known

Bootstrapping

- Let's assume
  - We have sample(s) S of a reasonable size N
  - We don't know the population parameters
- We can perform Monte-Carlo Sampling on the sample
  - Treat the sample as the population
  - Run Monte-Carlo Simulation with replacement

1. For i = 1 to K
   a) Select a sample ${S_i}^*$ of size N from S with replacement
   b) Calculate and record statistic ${\theta_i}^*$ for ${S_i}^*$
2. Determine and use probability distribution of θ^*

Bootstrapping

- Reminder:
  - S and therefore $S_i{}^*$ are taken from the population that belongs to H_1
  - θ^* is the probability distribution under H_1!

- Since we want the sampling distribution under H_0, we have to transform it:
  - If we can assume that the shapes of the population distributions under H_0 and H_1 are similar: **Shift-Method**
  - If we can assume that $\bar{x} - \mu$ is normally distributed: **Normal Approximation Method**

Bootstrapping

- Advantages
  - Straightforward and usually simple to calculate
  - One important assumption: The original sample is representative of the population
  - Works well in many situations
  - Can be used to bootstrap confidence intervals for distributions that are not normal (see Cohen 5.6)

- Disadvantages
  - Bootstrapping is dependent on the quality of the sample
  - It is hard to decide whether we have a good sample

Randomization Tests

- Sometimes we don't need to draw conclusions about populations
- Question: Do two samples S_A and S_B significantly differ?
  - Parametric and bootstrap tests: Indirect answer through inference about population parameters

- Can we answer the question directly?
  - We only want to use information from the samples
  - We do not want to make assumptions about the populations those samples come from

Randomisation Example

- We have two samples and a statistic

	1	2	3	4	5	6	7	8	9	10	median
s_A	32	32	45	45	23	67	53	67	41	53	45
s_B	43	24	42	23	23	43	23	60	32	41	36.5

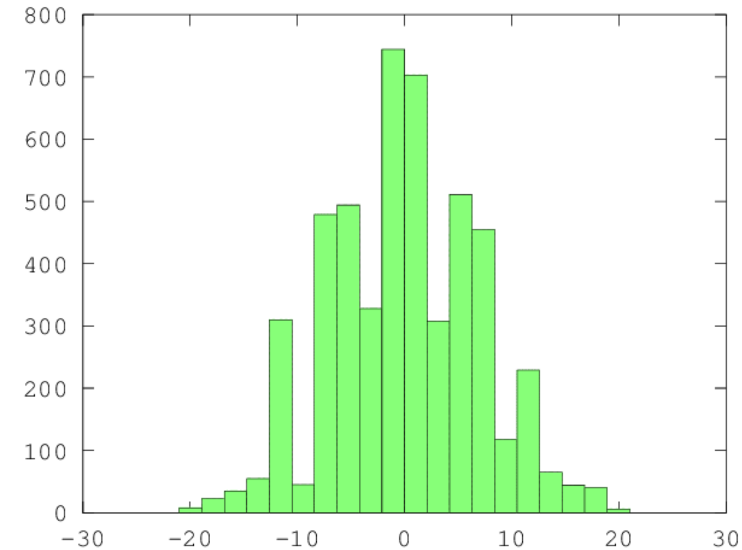
- We want to know whether the difference in median of 8.5 means a significant difference

- **Approximate randomisation**
  - We want to know whether both samples are drawn from the same population
  - We have two samples of size $N_A = N_B = 10$ and our test statistic is $\theta = median(S_A) - median(S_B)$
  - $S_{A+B} = S_A + S_B$ is the concatenation of the two samples

Approximate randomisation

1. For i = 1 to K
   a) Shuffle the elements of S_{A+B} to create $S_{A+B}{}^*$
   b) Assign first N_A values to the randomised pseudo-sample $S_A{}^*$ and remaining to $S_B{}^*$
   c) Calculate and record test statistic $\theta_i{}^*$
2. Use distribution of θ^* to determine probability of the sample result θ under H_0

- In our example:
  - K=5000
  - 502 elements are ≥ 8.5
  - $p = 502/5000 = 0.1004$



Why approximate?

- If we would use all possible outcomes to create the probability distribution, we would perform exact randomisation
- In our example we can draw $\frac{20!}{10! * 10!} = 184756$ possible samples ${S_A}^*$
- We only used 5000 (=2.7%), therefore "*approximate*"

- Exact randomisation may not be feasible due to the large number of possible randomised samples
- We have to use a smaller distribution and arrive only at an approximate probability

Randomisation Tests

- Advantages
  - Can always be used if we have 2 samples
  - Does not need assumptions about population parameters
  - "suited to test hypotheses about arrangements of data and statistics that characterize the arrangements"[Cohen]
- Disadvantages
  - Does not create a real "sampling distribution"
  - We can't infer general results about the underlying populations

Randomisation Test of Independence

x	1	2	3	4	5	6	7	8	9	10
y	54	66	61	44	60	55	51	45	63	52

- Correlation coefficient: $r = -0.255$
- Question: Are x and y independent?
- **Randomisation test of independence:**
  1. Repeat 5000 times
     a) Shuffle y to create y^*
     b) Calculate and record $r^* = corr(x, y^*)$
  2. Use distribution of r^* to calculate probability of $r = -0.255$
- After 5000 repetitions, 1200 values are below -0.255
- For which number of values would we have rejected H_0?

Bootstrap vs. Randomisation

- **Both**
  - generate distributions from the original sample
  - can be used if parametric assumptions can't be met
  - can be used for statistical tests where no estimated sampling distribution is available (unconventional statistics)
- **Bootstrapping**
  - Resampling with replacement
  - Simulates the process of drawing from an infinite population
  - Assumes that the sample is representative of the population, i.e. that the frequency distribution is the same
  - Can be used to construct confidence intervals
  - Becomes very robust with larger sample sizes

Bootstrap vs. Randomisation

- **Randomisation**
  - Resampling without replacement
  - Needs at least two samples to generate a combined sample
  - Tests whether a particular arrangements is unusual relative to the distribution under the null hypothesis
  - Does not produce “real” sampling distributions
  - We can therefore not infer population parameters (i.e. also no confidence intervals, etc.)

Bootstrap vs. Randomisation

- **Why use randomization then at all?**
  - Perform as well as parametric tests when parametric assumptions hold
  - Outperform them when assumptions don’t hold!
  - Bootstrapping is as accurate as t-test for larger samples and generally equal when parametric assumptions are violated
- **Computer-Intensive vs. parametric**
  - No need to check assumptions
  - Usually not inferior to parametric tests in ideal conditions
  - Ideal for computer scientists where large sample sizes are the standard and require no knowledge about specific sampling distributions

What have we learned?

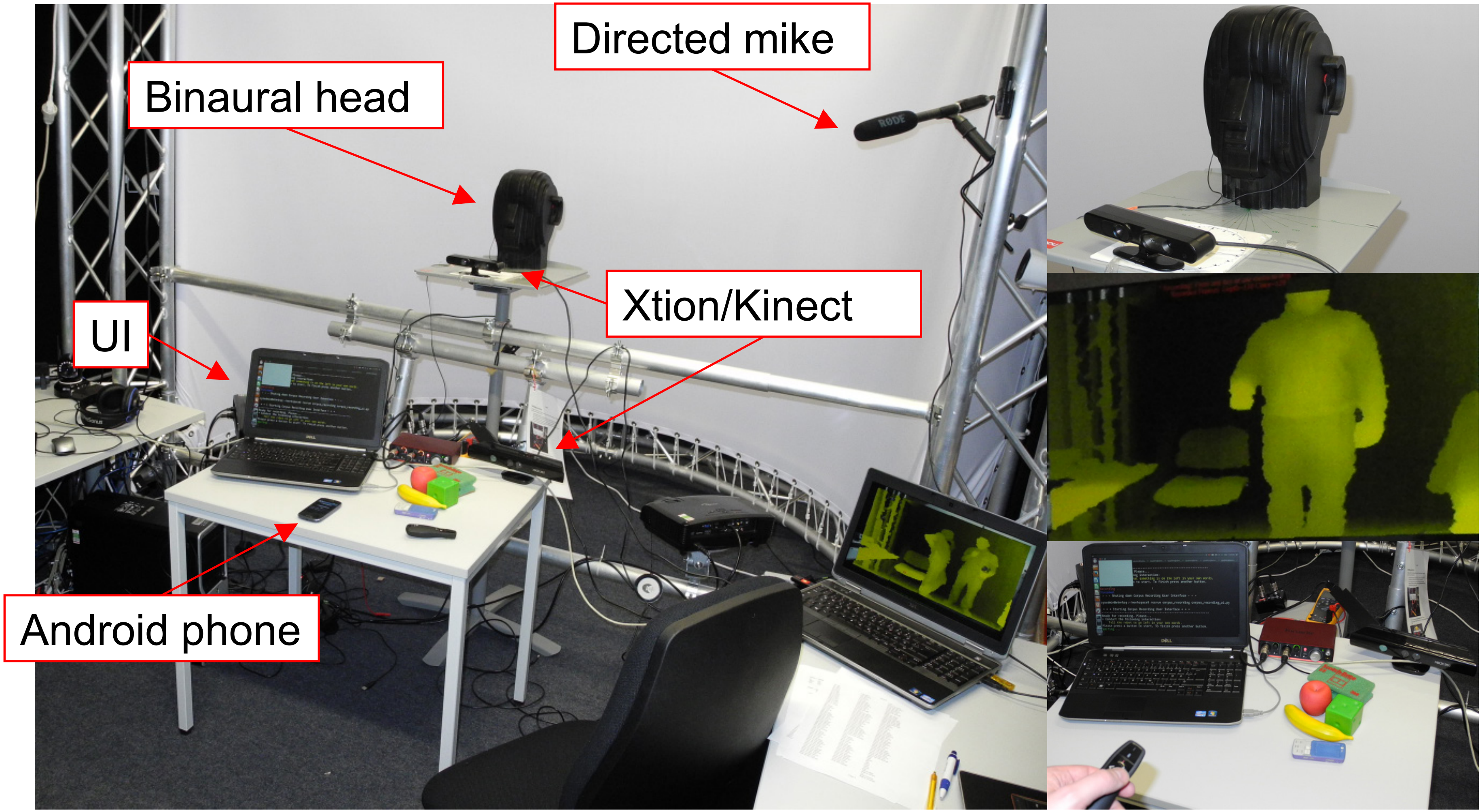
1. Randomisation allows us to compare two samples without knowledge about the population
2. Variants of randomisation can replace parametric tests (2-sample t-test, paired t-test, test of independence)
3. Computer-intensive tests are not inferior to parametric tests
4. For you it is often a trade-off between practicability and the perfect test
5. **Always:** Once you have chosen a test, be careful about assumptions, limits, transformations, etc.
6. **If in doubt:** Run a second test (e.g. 2-sample t-test & randomisation test if not sure parametric assumptions hold)

Data-Collection Wrap-Up

- Consent form and questionnaire
  - 15 participants, 13 male, 2 female
  - Only right handed people!
  - 13 different native languages, none English
  - 13/15 consented to possible public usage of the data
  - Why did we offer a “Don’t use” option?

- Trying to prepare for confounding variables:
  - Conditions that affect movement/speech (injuries, sports)
  - Pre-existing knowledge on robots/HRI (field of study, previous HRI studies, specific order of people)

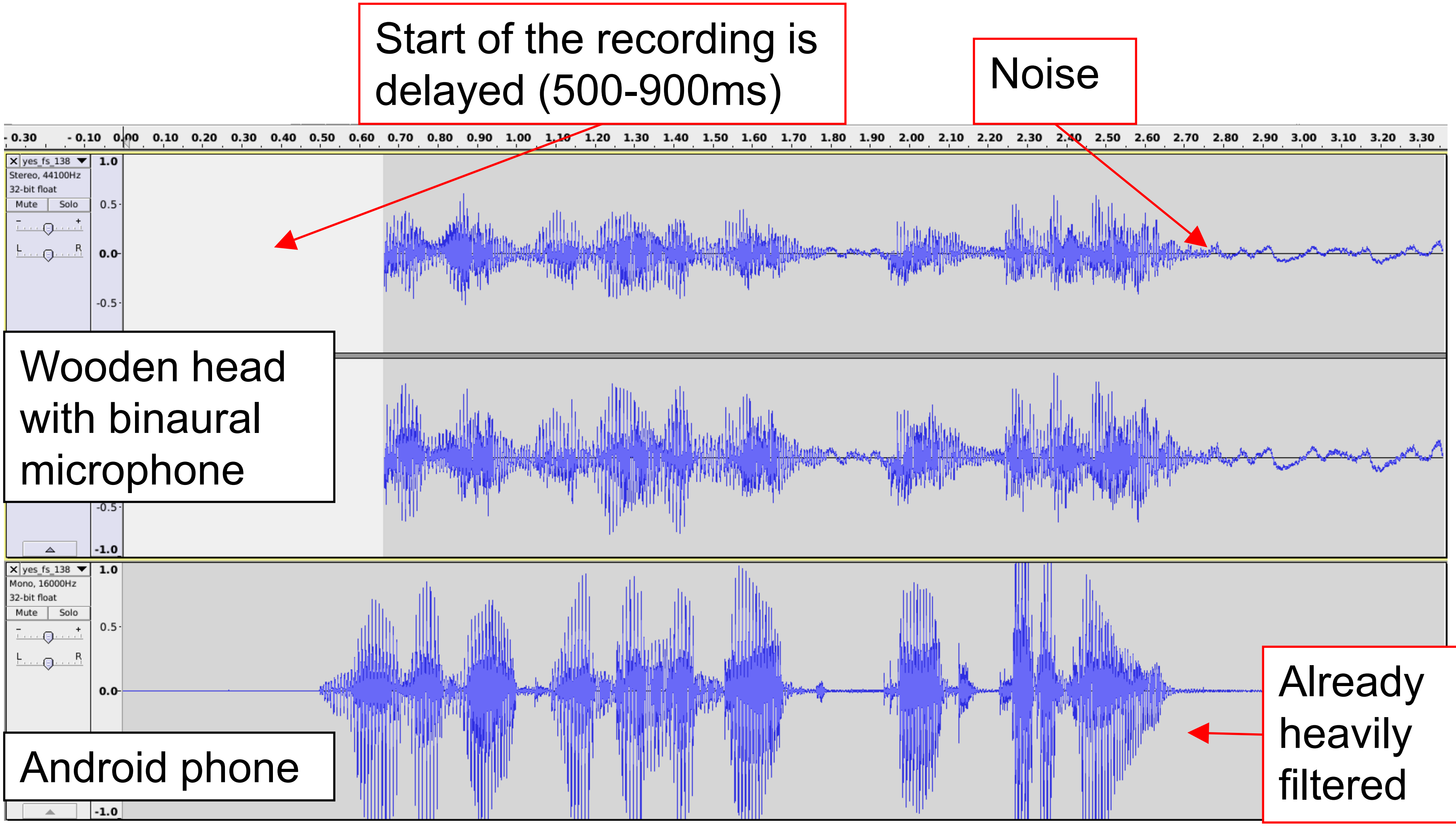
Speech and Gesture Recording



Speech and Gesture Recording: Data Set

- Overall:
  - 489 data points – including test runs
  - 4,14 GB video data, 382 MB sound recordings
  - Recordings session (in sum): 3h, 53min
  - Each participant labored 16,64min on average
- Conducted 4 different recordings:
  - Free speech & gesture: 206 recs, with Binaural head + Kinect
  - Free speech: 171 recs, with Binaural head + Android Phone
  - Free gesture: 26 recs, with Kinect
  - Speech & gesture from grammar: 86 recs, with Binaural head + Kinect
- Overall recordings with useful quality: estimated 80%

Speech Recording: Result Quality



Free Speech Recording: Interesting Results

- Example: Robot should stop its action.
  - Expected Utterance: `Robot stop.`
  - Recorded Utterance: `Robot please don’t do that.`
- Example: Robot should put down the object.
  - Expected Utterance: `Robot put down object.`
  - Recorded Utterance: `Hi robot, please put the object that you are carrying down.`
- Example: Robot should move to the left.
  - Expected Utterance: `Robot go left.`
  - Recorded Utterance: `Robot I want you to go left, right now.`
- Example: Robot’s action was correct.
  - Expected Utterance: `This is correct.`
  - Recorded Utterance: `Robot you are doing good, keep going.`

Unexpected vocabulary

Complex and long utterances

Ambiguities

Slang

Depth and Video recording

Things we have encountered

- Always provide cookies to keep people happy!
- **Run pilot study!**
  - To avoid technical problems (directed microphone, Kinect interference, delay in recording, ….)
  - To adjust timing and synchronisation between setups

- Organisation
  - Provide quiet place to fill consent form and questionnaire
  - Prepare schedule that can deal with asynchronous setups
  - Perform complete test run of setup before start
  - Better preparation (e.g. forms for comments of investigator)

Things we have encountered

- Full body movements
  - Kinects interfere with each other, creating noise
  - People try to be especially expressive when moving
    - Slower motion?
    - Looking towards the sensor / investigator
  - Some assumptions were wrong, e.g. people don’t want to fall realistically
  - We always have the same sequence, maybe improve setup towards random order of actions
  - Potentially 100% of the data can be used, although standard libraries, e.g. skeleton model fitting, fail for some poses
  - Streams were not automatically synchronised

Things we have encountered

- Speech & Gesture setup
  - Quality of sound recording inverse to expectation (head vs. phone)
  - Giving user control of begin & end of gesture worked and made post-processing easier
    - Cutting points unbiased by investigator
    - Automatic segmentation of video/audio streams
    - Automatic labelling

The Final Experiment

- Aim
  - You are designing, setting up and running an HRI pilot study
  - Possible questions to investigate: What is a human baseline for
    a) detecting the gesture/command that was given
    b) accuracy in understanding an utterance
- Organisation
  - Two groups run two different (but similar) studies
  - Each group is accompanied by an advisor and supported by WTM in terms of rooms/equipment/expertise
  - Deadline: Open for discussion

The Final Experiment

- You work:
  - Break the whole task down into subtasks and prioritise
  - Define the hypotheses
  - Define and discuss the experiment protocol and the procedures that you want to use
  - Specify how to collect the data and how to analyse it
  - Organise and run the experiments.