# Lesson1: How Big is the Web?

# Unit4: Probabilistic Simulated Generative Modeling

Rene Pickhardt

Introduction to Web Science Part 2

Emerging Web Properties

WeST

People and Knowledge Networks

**Institute for Web Science and Technologies · University of Koblenz-Landau, Germany**

# Completing this unit you should

- Understand the notion of a model parameter
- Realize that a probabilistic model needs to run more than once
- See how generative models **can** yield an explanation for a described phenomenon
- Know that the explanation is plausible but needs not to be true even if statistics match perfectly
- Be aware of a statistical test to see how many significant digits of two numbers are the same

# Goal: Explain why there are 1.2 Mio words on the simple English Wikipedia data set

- The descriptive model gave a clue for the size of the Simple English Wikipedia
- Find a way of explaining the size
- Build a simple, probabilistic generative Model
  - Simple: Few model parameter
  - Probabilistic: Create a random process
  - Generative: reproduces the statistics from the descriptive model
- Compare Descriptive and Generative Model

# Hypothesis

- The number of words in the Simple English Wikipedia can be explained because since the birth of the website every minute a new word is added with a probability of p.

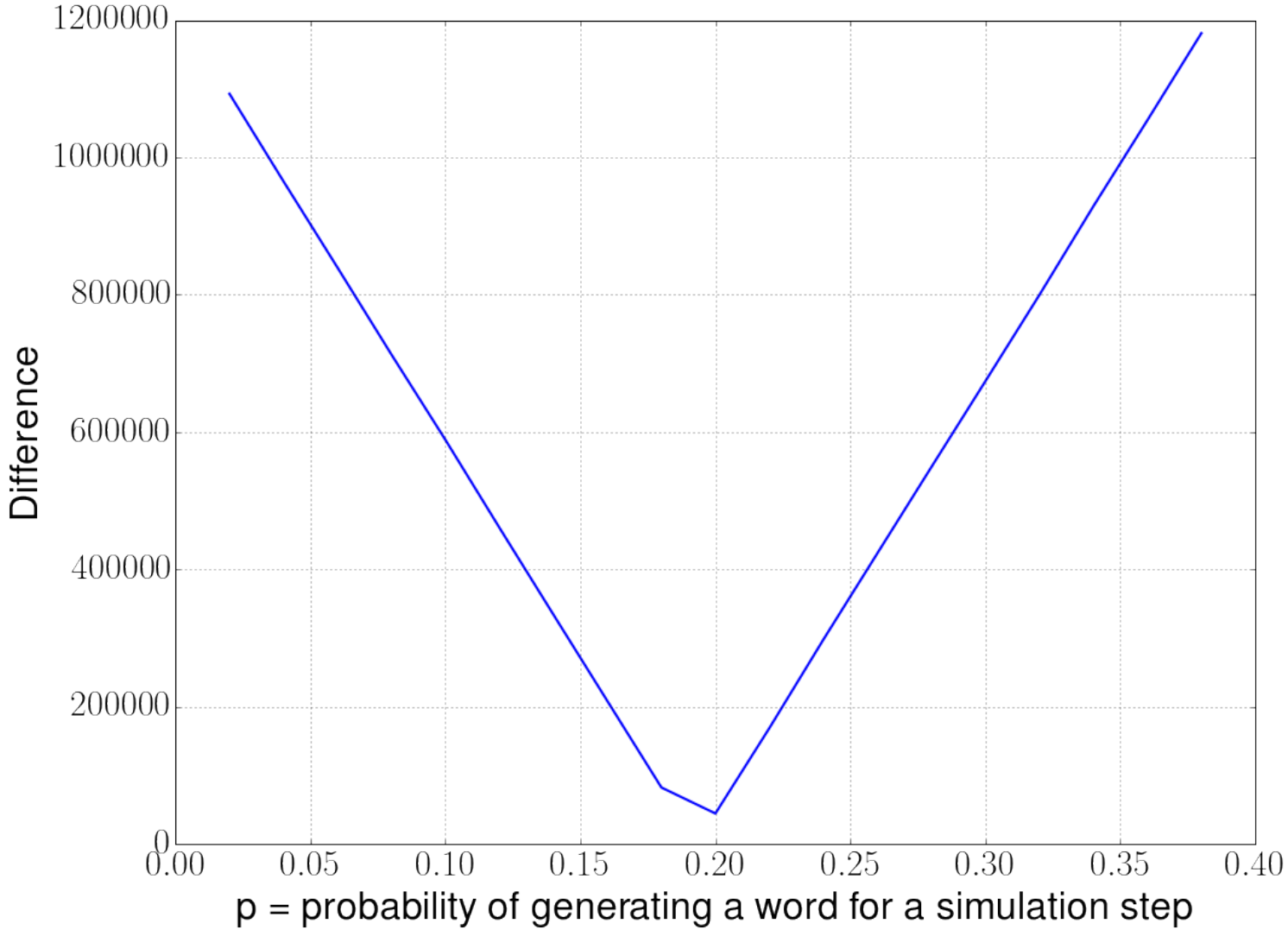
- Remarks
  - This hypothesis is oversimplified and will not reflect the reality
  - We will show with statistic methods how it can be verified.
  - It serves also as an example to see how conclusions from models should be taken carefully

# Define the Model(parameters)

- Words are added randomly.
- Edits can happen every minute
- Model parameter p stats the probability weather to generate a word or not in this minute
- Model simulates 12 * 365.25 * 24 * 60 Minutes
  - (12 years uptime of Simple English Wikipedia)
- Obviously these assumptions are pretty strong!

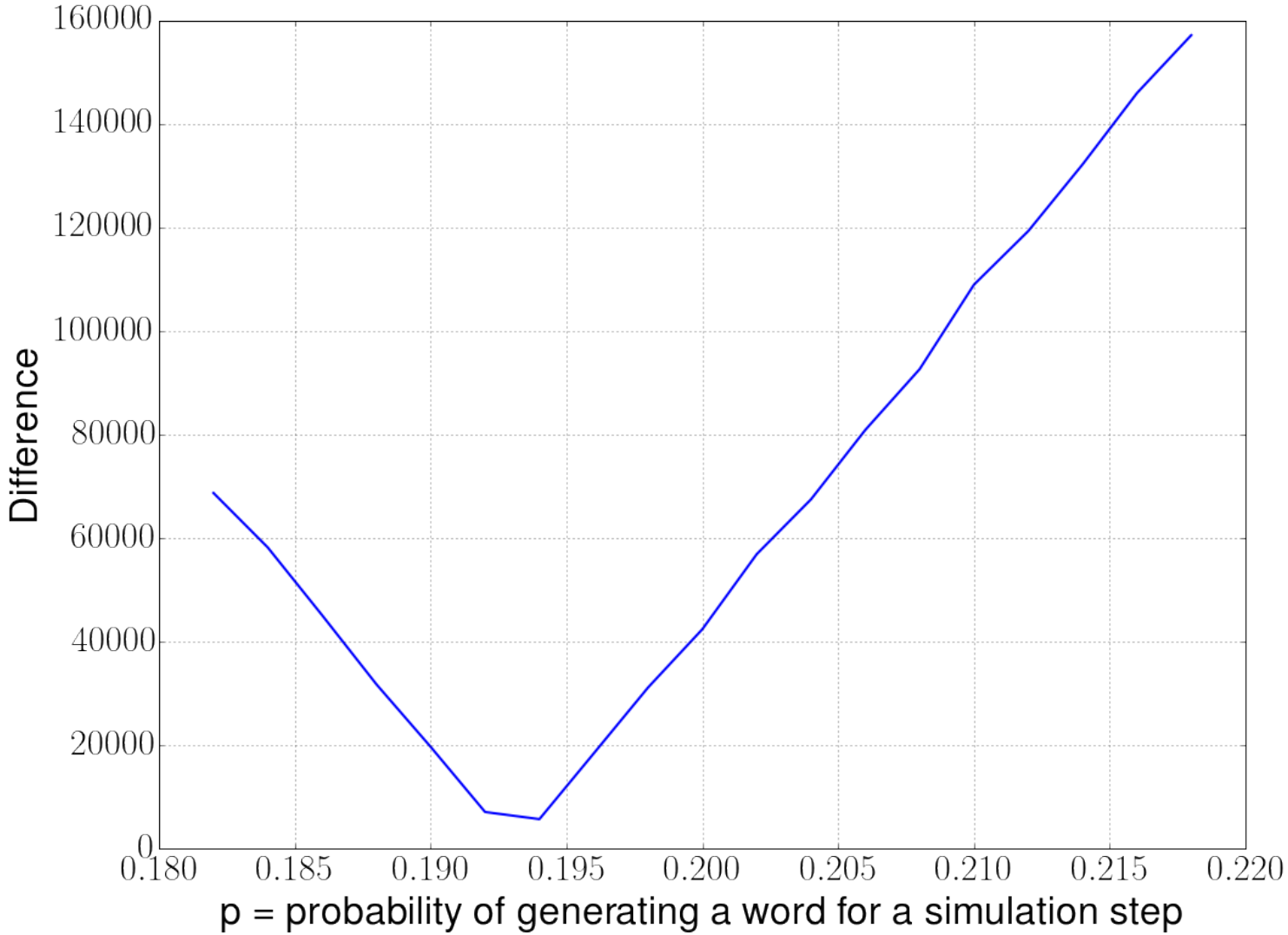
# Test for various model parameters

Difference of generated words and measured words on Simple English Wikipedia



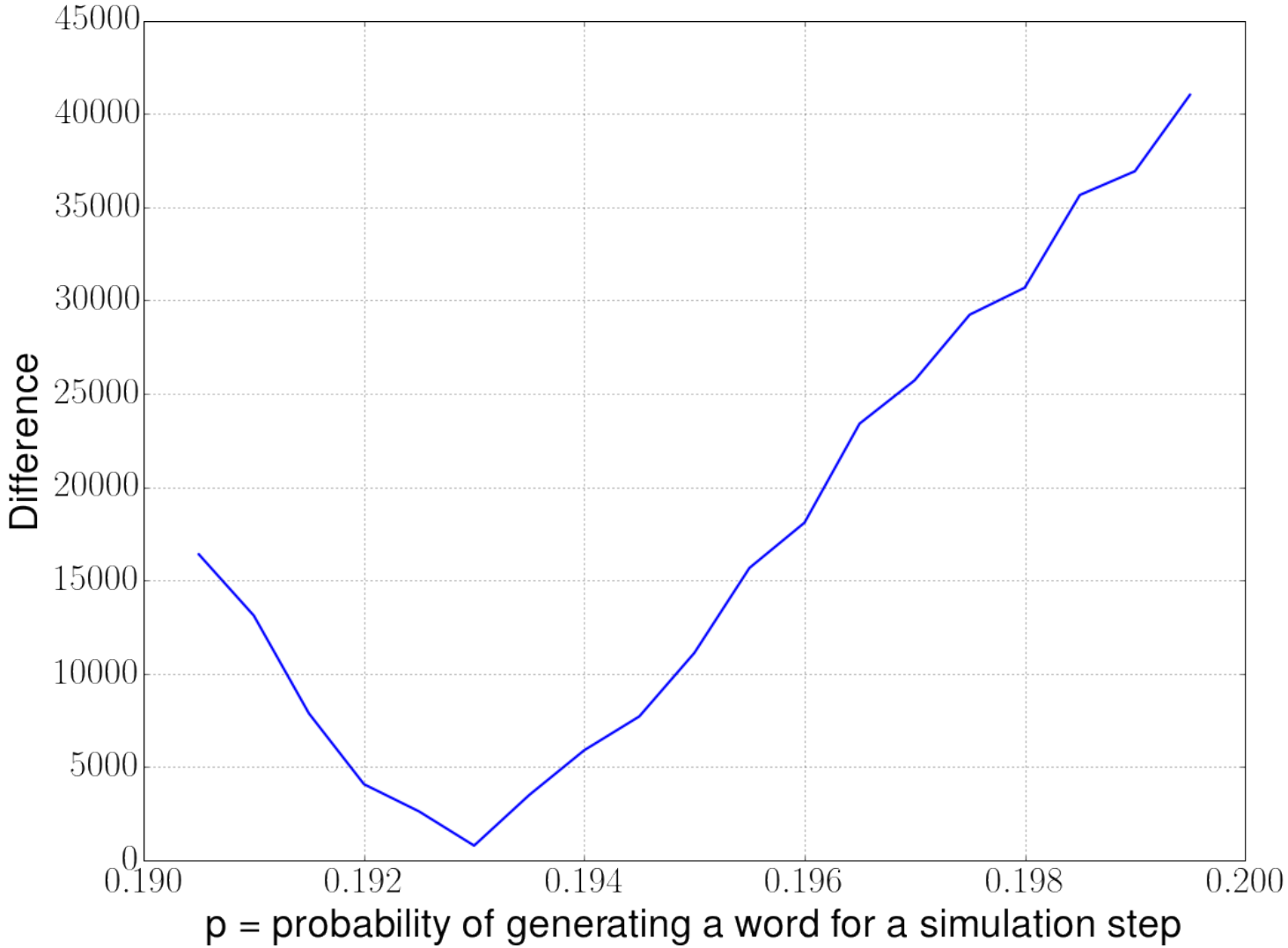
# Zooming in

Difference of generated words and measured words on Simple English Wikipedia



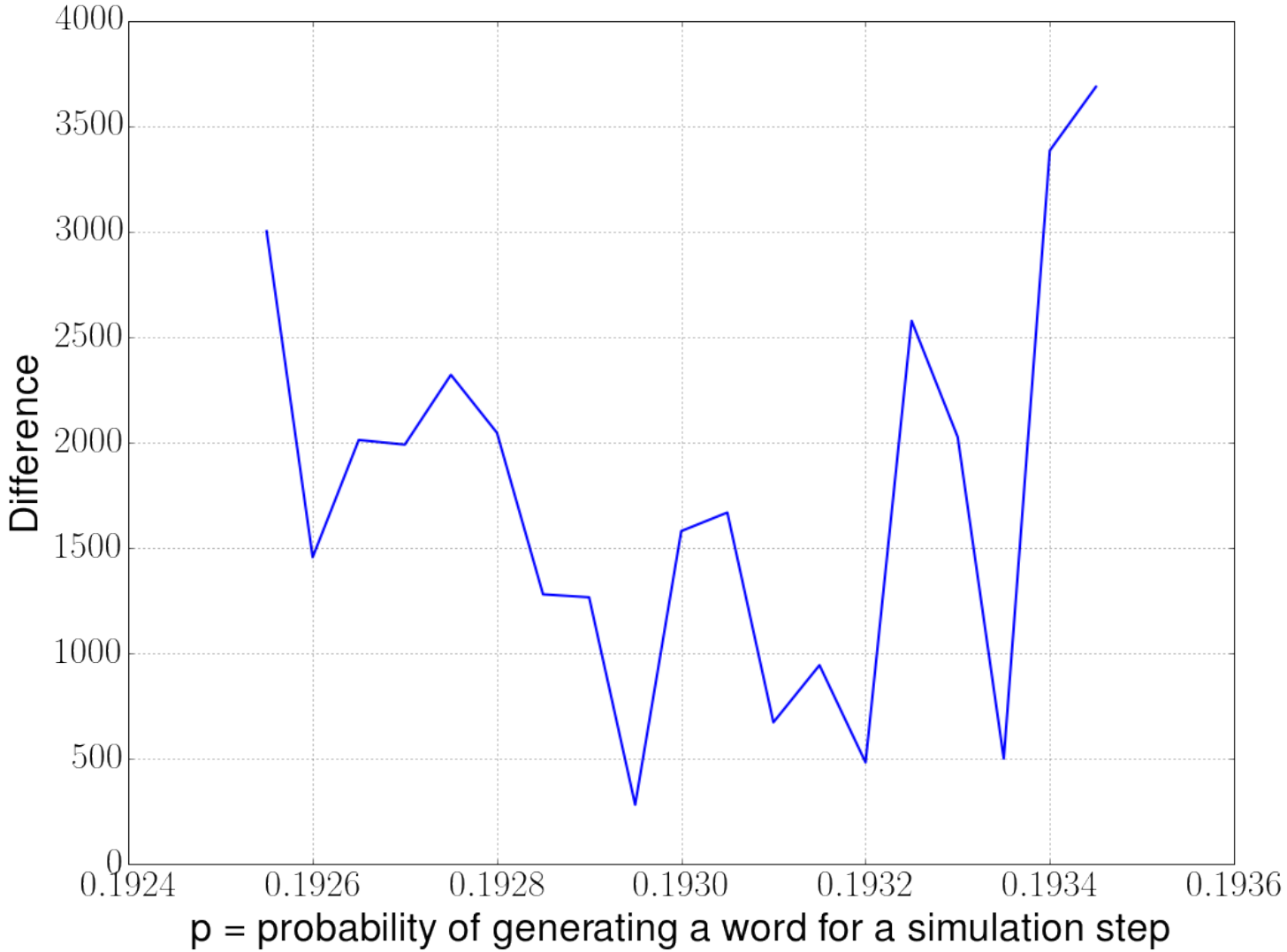
# 0.193 words seems to be the correct value

Difference of generated words and measured words on Simple English Wikipedia



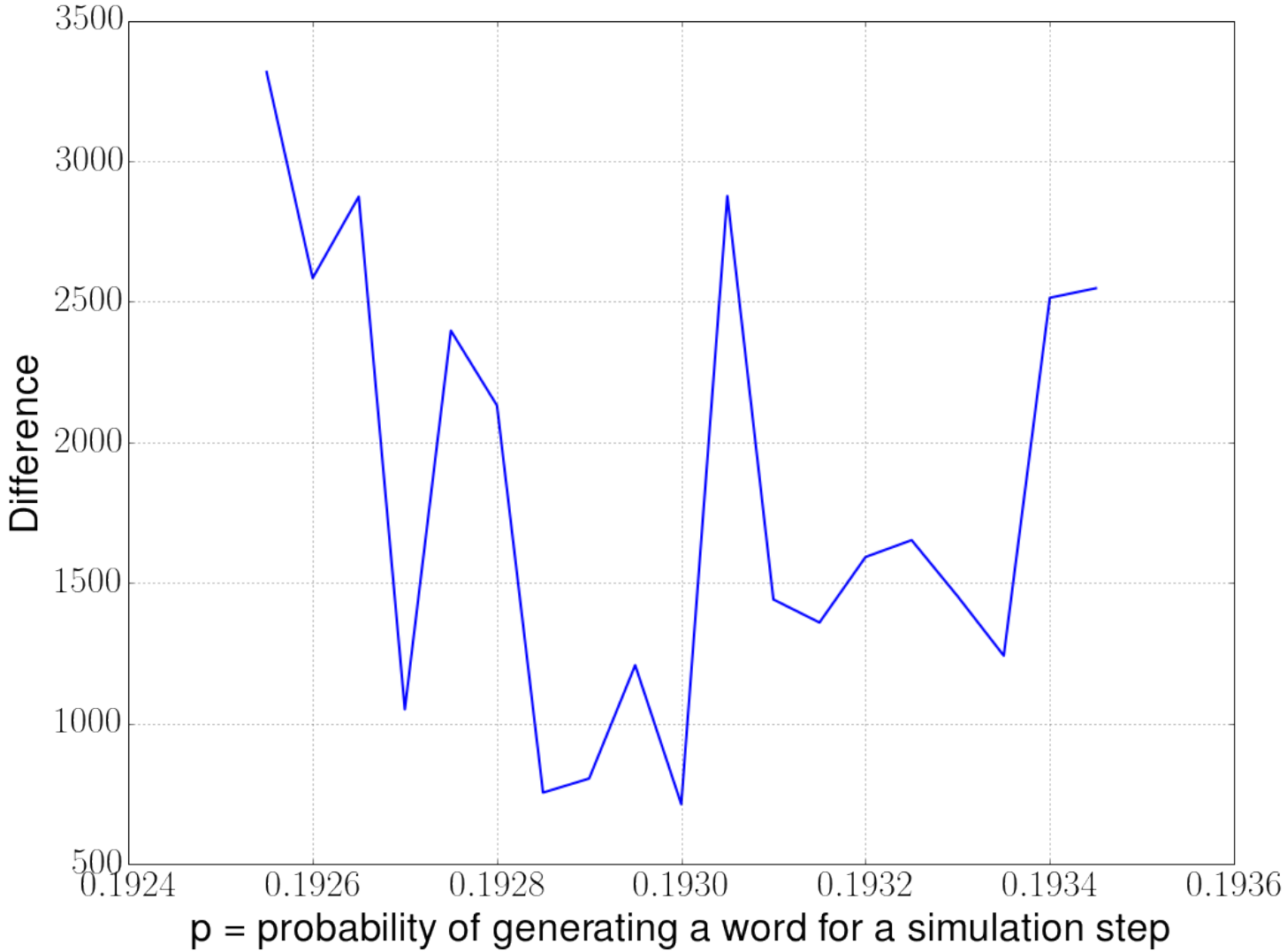
# Even closer view yields a surprising picture

Difference of generated words and measured words on Simple English Wikipedia



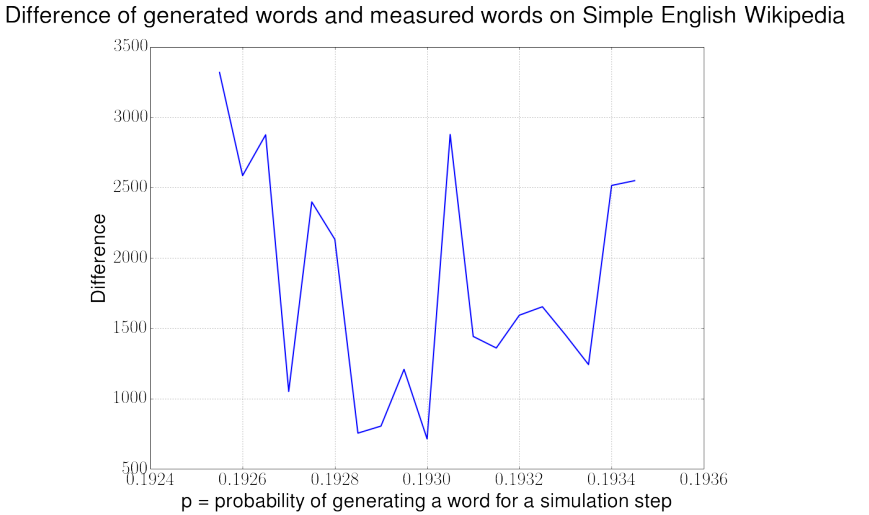
# A new simulation changes the results

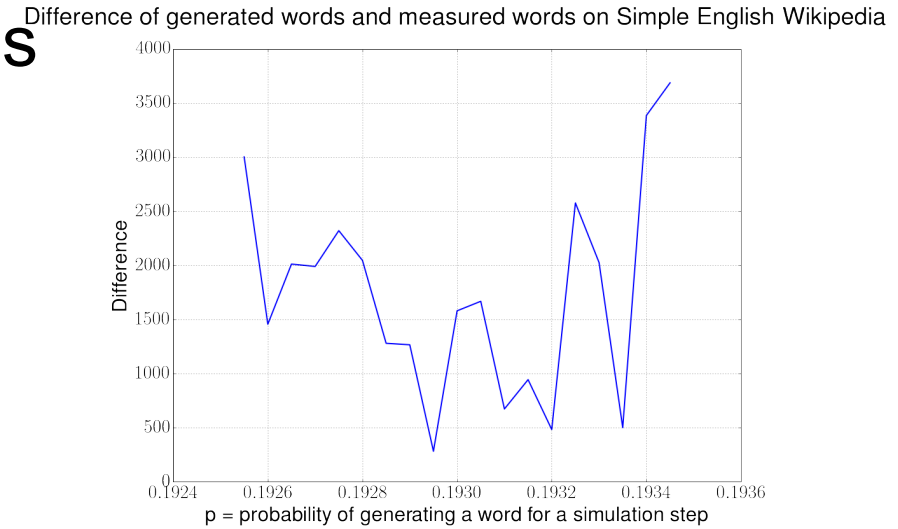
Difference of generated words and measured words on Simple English Wikipedia



# A new simulation changes the results

- Probabilistic models will not have the exact same results every time they run
- Even though every side of a dice has the same probability. Rolling a dice 6 times does not mean you will see all eyes
- How can we know the best model parameter for a simulated probabilistic model?



# Finding the best model parameter in a generative probabilistic model

- Apply statistics for comparing the results from the generative model with the ones from the descriptive model
  - Significance test
  - Averages
  - Histograms
  - Medians

- Run the experiment more than once to avoid outliers!

# How to decide weather two numbers are close to each other?

- 1218526 words found via our Descriptive Model

| Parameter (p) | number of generated words |
|---|---|
| 0.1928 | 1216395 |
| 0.19285 | 1217772 |
| 0.1929 | 1217722 |
| 0.19295 | 1217319 |
| 0.193 | 1219239 |
| 0.19305 | 1221402 |
| 0.1931 | 1217085 |
| 0.19315 | 1219885 |
| 0.1932 | 1220118 |

# How to decide weather two numbers are close to each other?

- Count number of significant digits two numbers have in common
- Find n such that

$$abs(x-y) < 0.1^n * max(x,y)$$

- Or $\log_{0.1} \frac{abs(x-y)}{max(x,y)} < n$
- Or $n = \left\lfloor \log_{0.1} \frac{abs(x-y)}{max(x,y)} \right\rfloor$

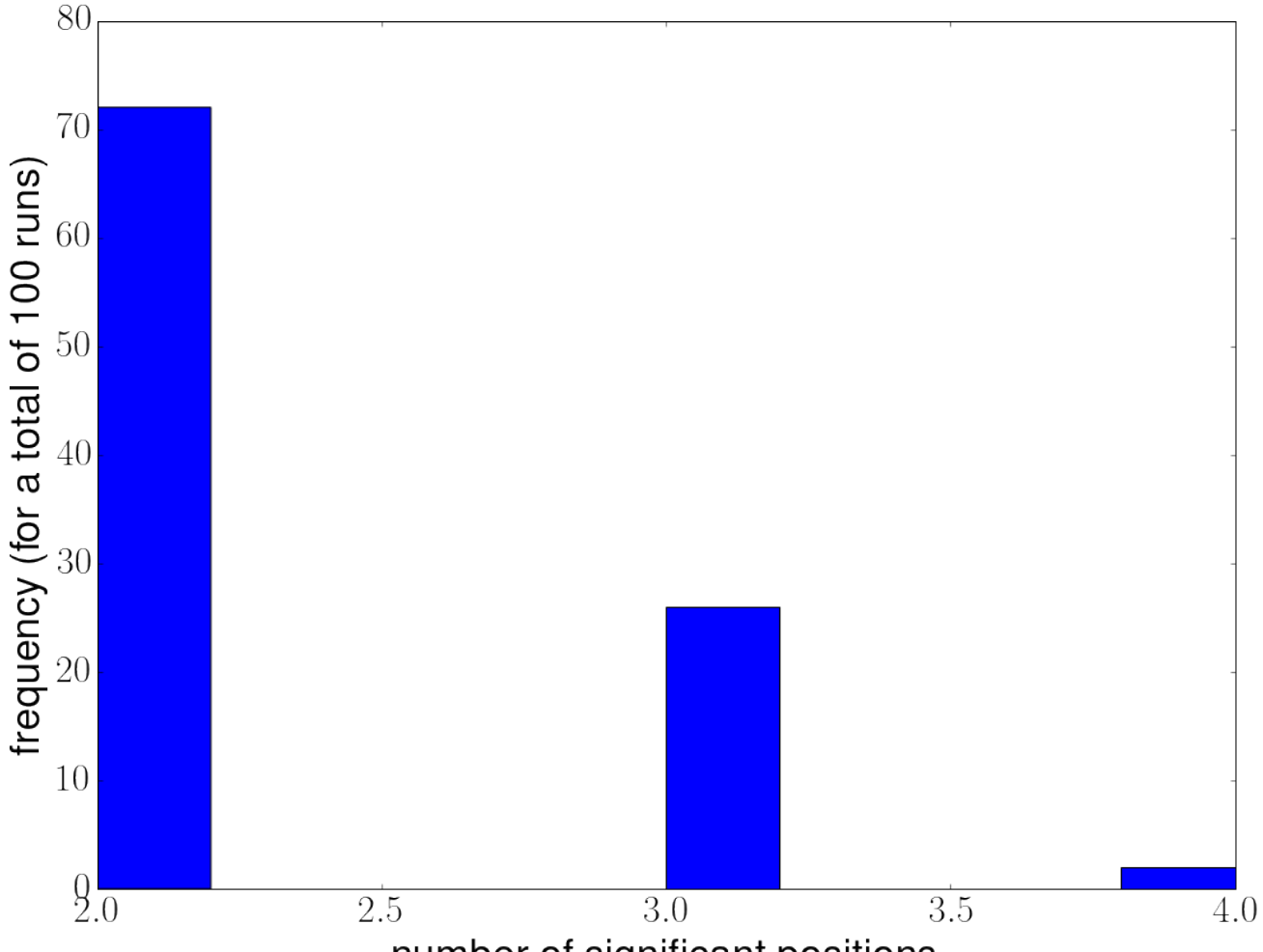
# How to decide weather two numbers are close to each other?

- **121**8526 words found via our descriptive model

| Parameter (p) | number of generated words | n |
|---|---|---|
| 0.1928 | 1216395 | 2 |
| **0.19285** | **1217772** | **3** |
| **0.1929** | **1217722** | **3** |
| **0.19295** | **1217319** | **3** |
| **0.193** | **1219239** | **3** |
| 0.19305 | 1221402 | 2 |
| 0.1931 | 1217085 | 2 |
| 0.19315 | 1219885 | 2 |
| 0.1932 | 1220118 | 2 |

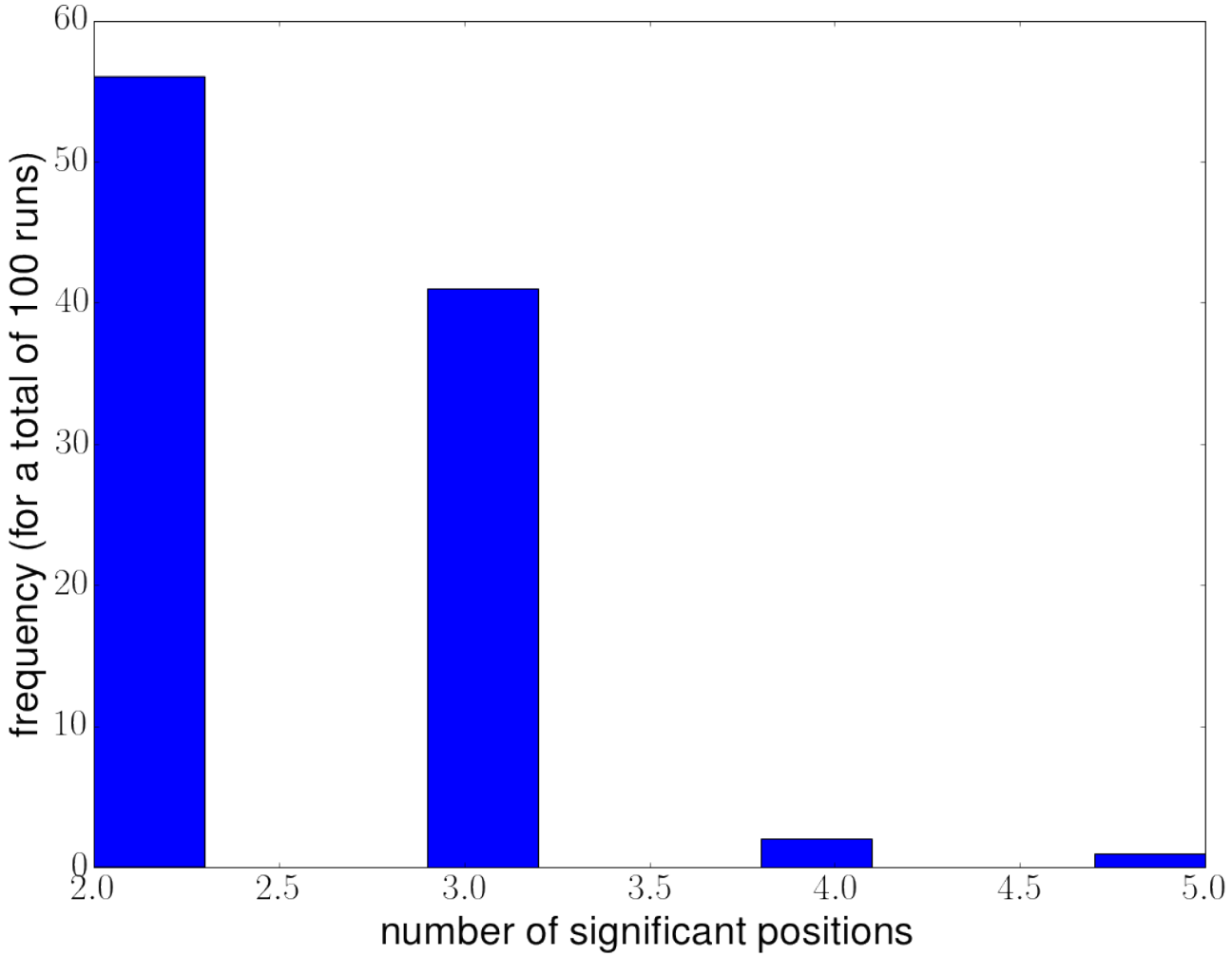
# On average 2.3 identical significant positions for 100 runs.

Comparing the descriptive model with the number of generated words for p=0.1928



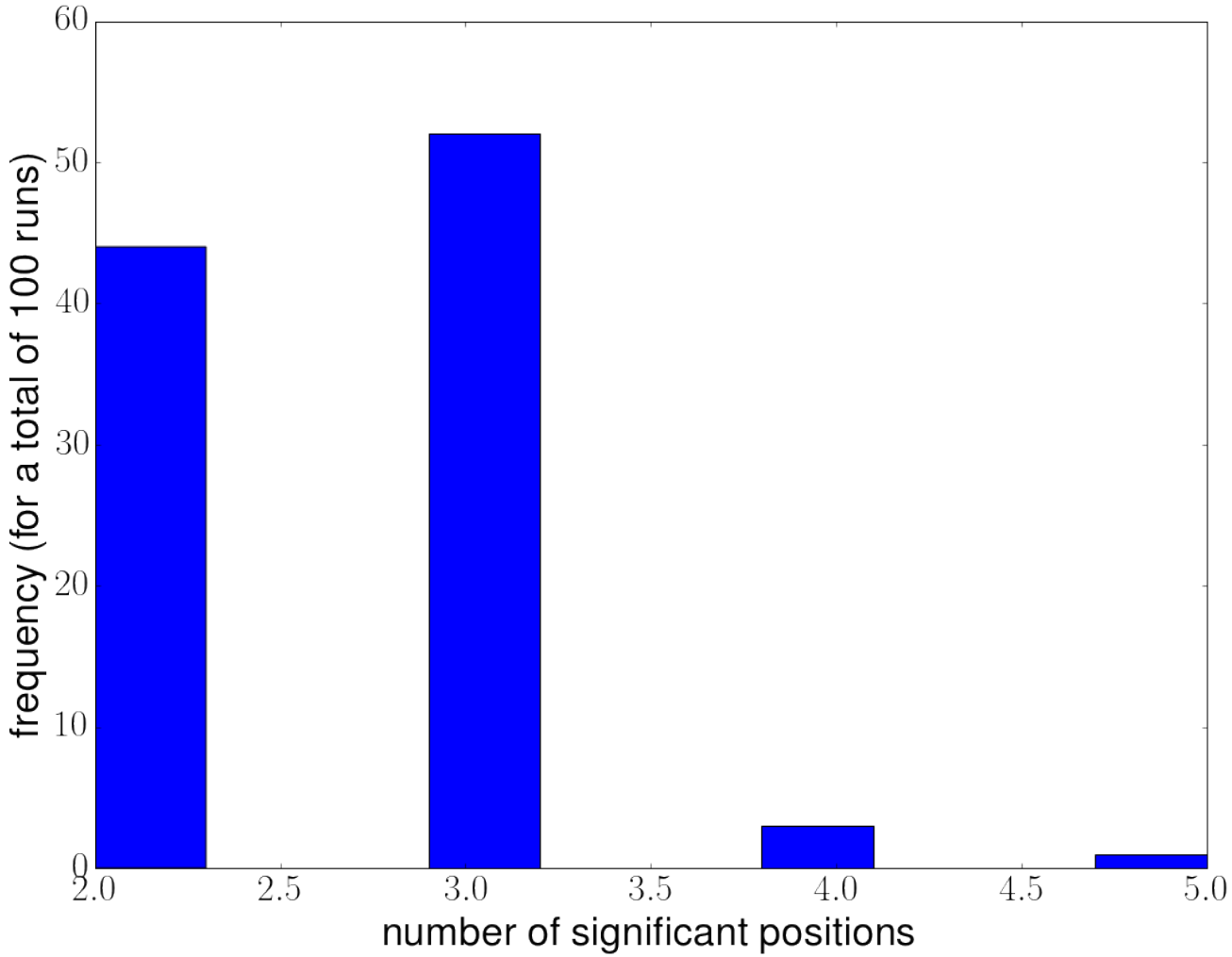
# On average 2.48 identical significant positions for 100 runs.

Comparing the descriptive model with the number of generated words for p=0.19285



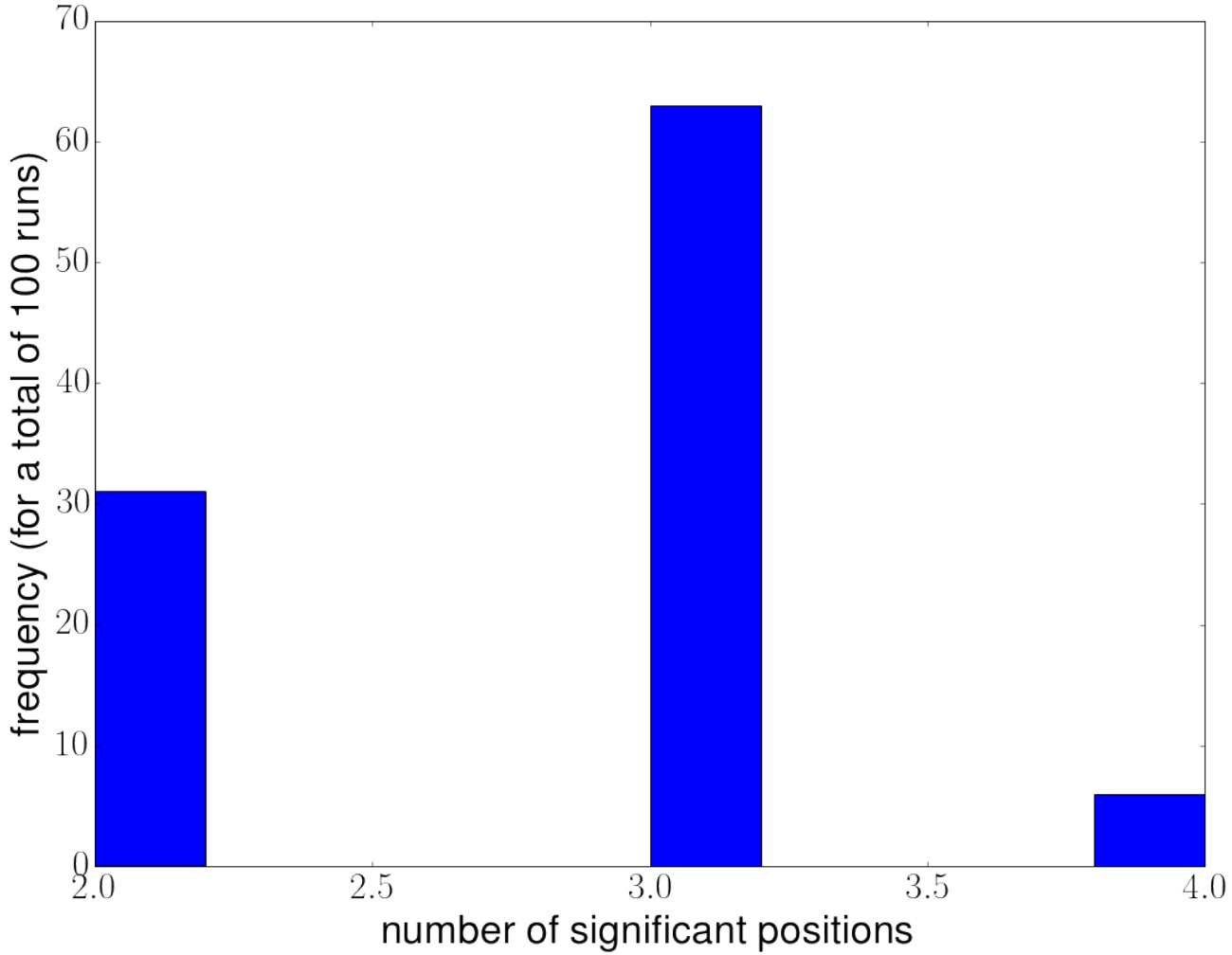
# On average 2.61 identical significant positions for 100 runs.

Comparing the descriptive model with the number of generated words for p=0.1929



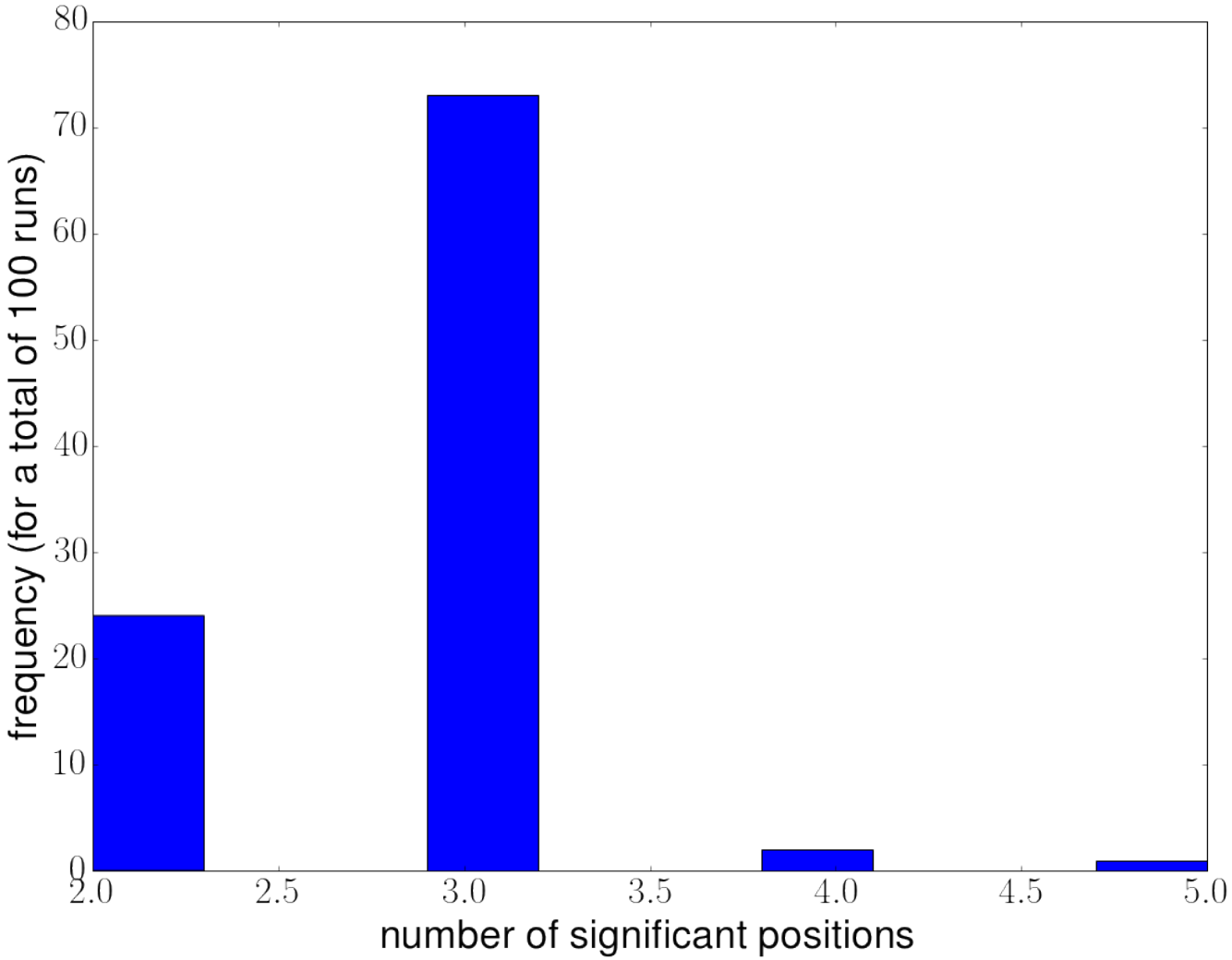
# On average 2.75 identical significant positions for 100 runs.

Comparing the descriptive model with the number of generated words for p=0.19295



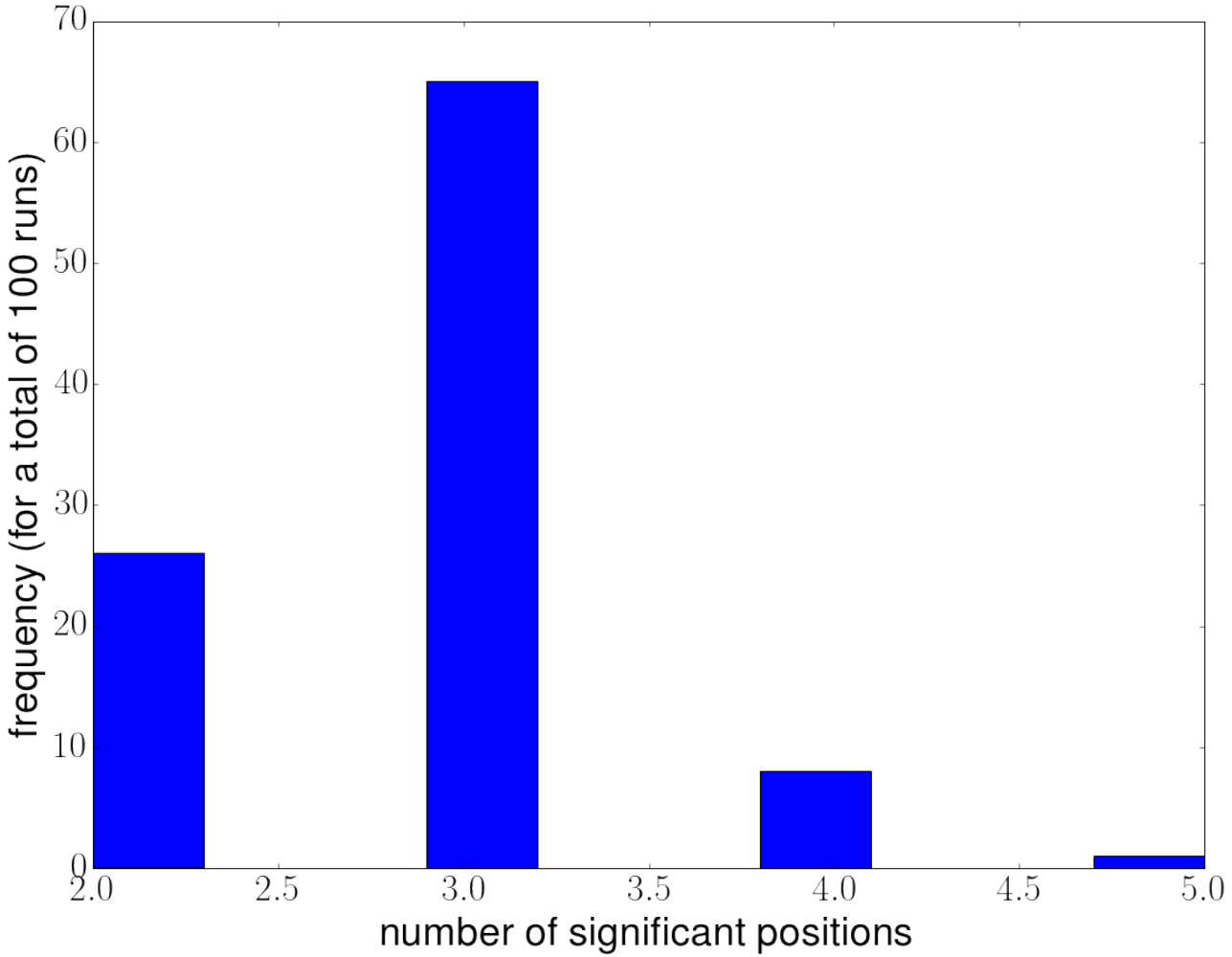
# On average 2.8 identical significant positions for 100 runs.

Comparing the descriptive model with the number of generated words for p=0.193



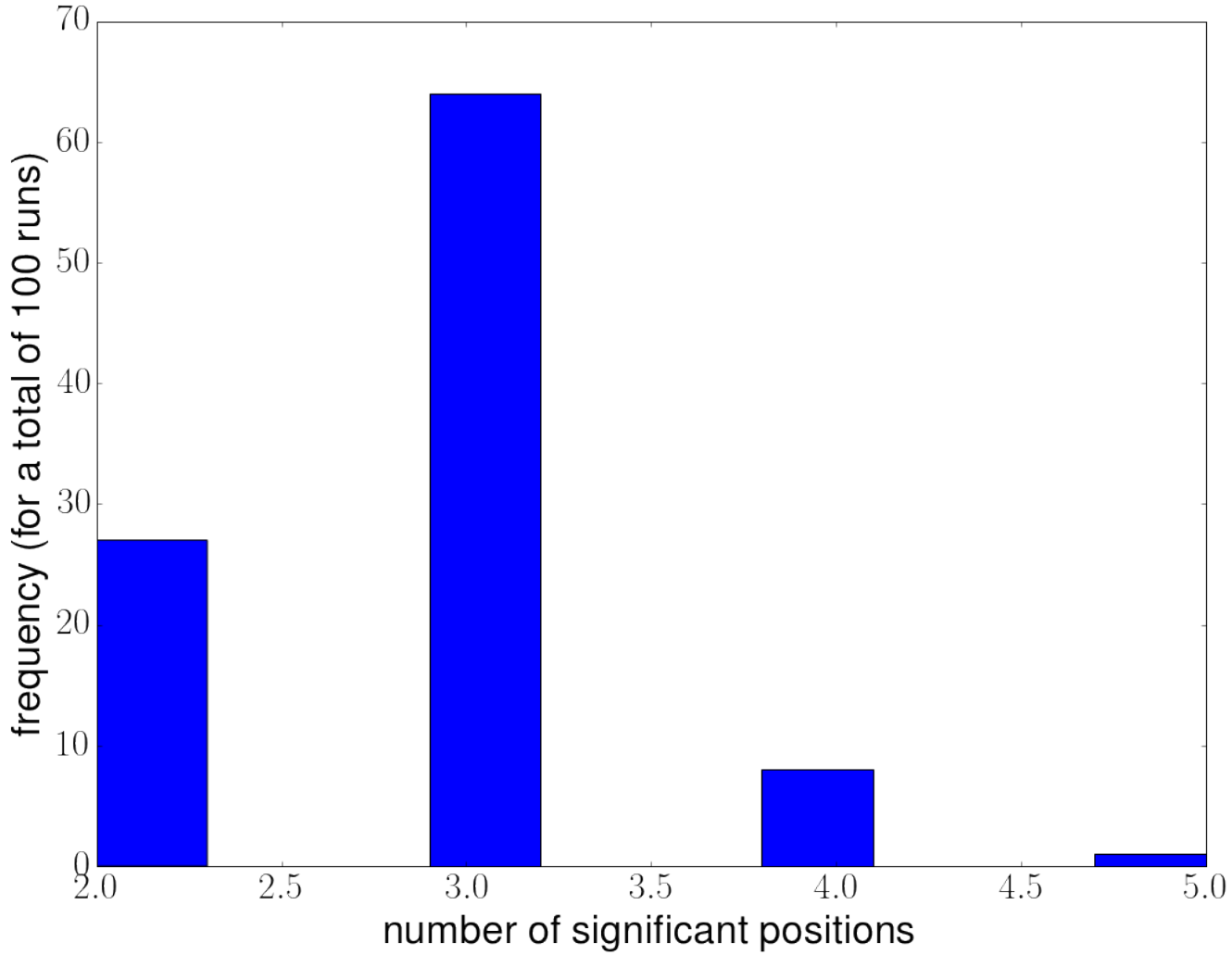
# On average 2.84 identical significant positions for 100 runs.

Comparing the descriptive model with the number of generated words for p=0.19305



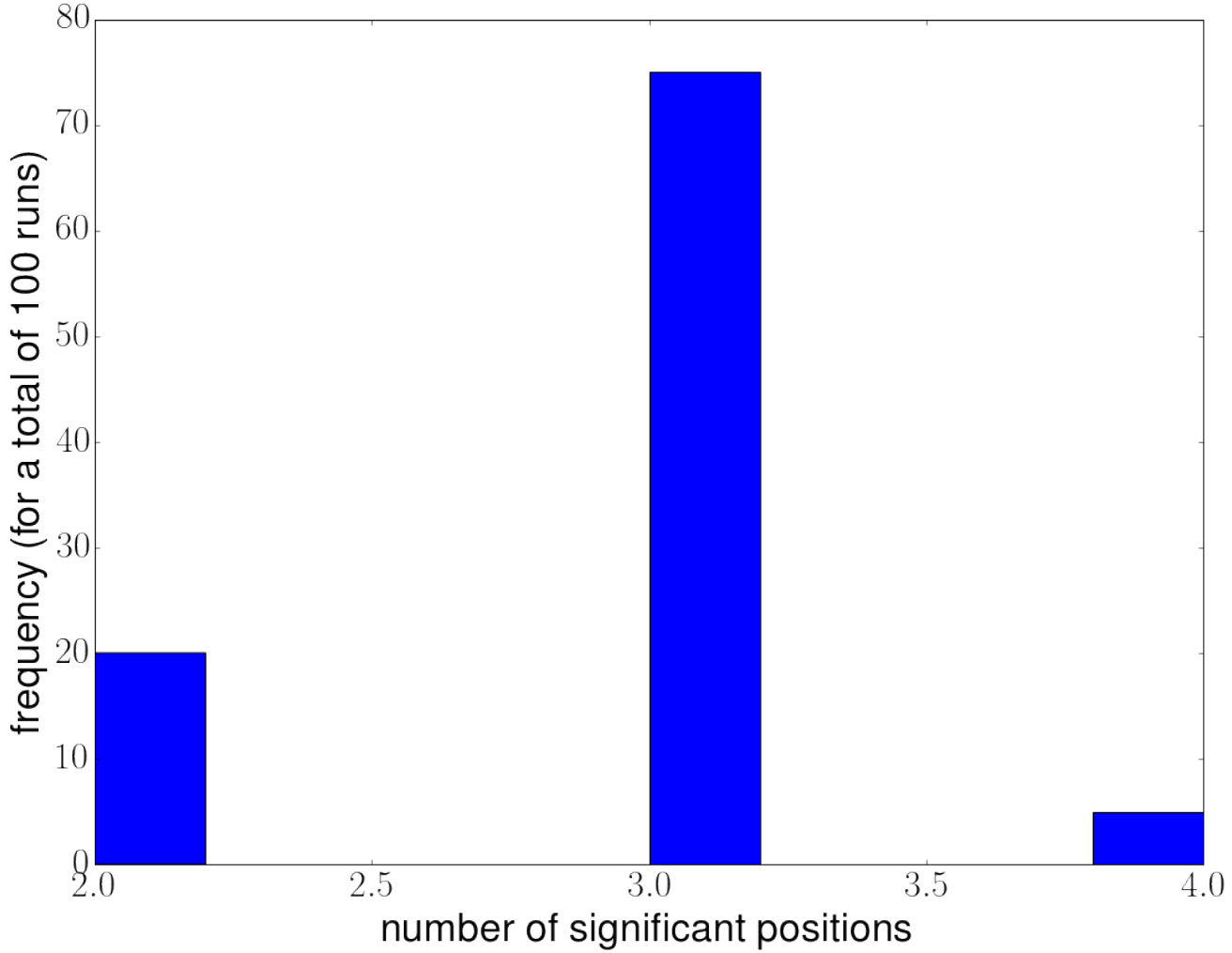
# On average 2.83 identical significant positions for 100 runs.

Comparing the descriptive model with the number of generated words for p=0.1931



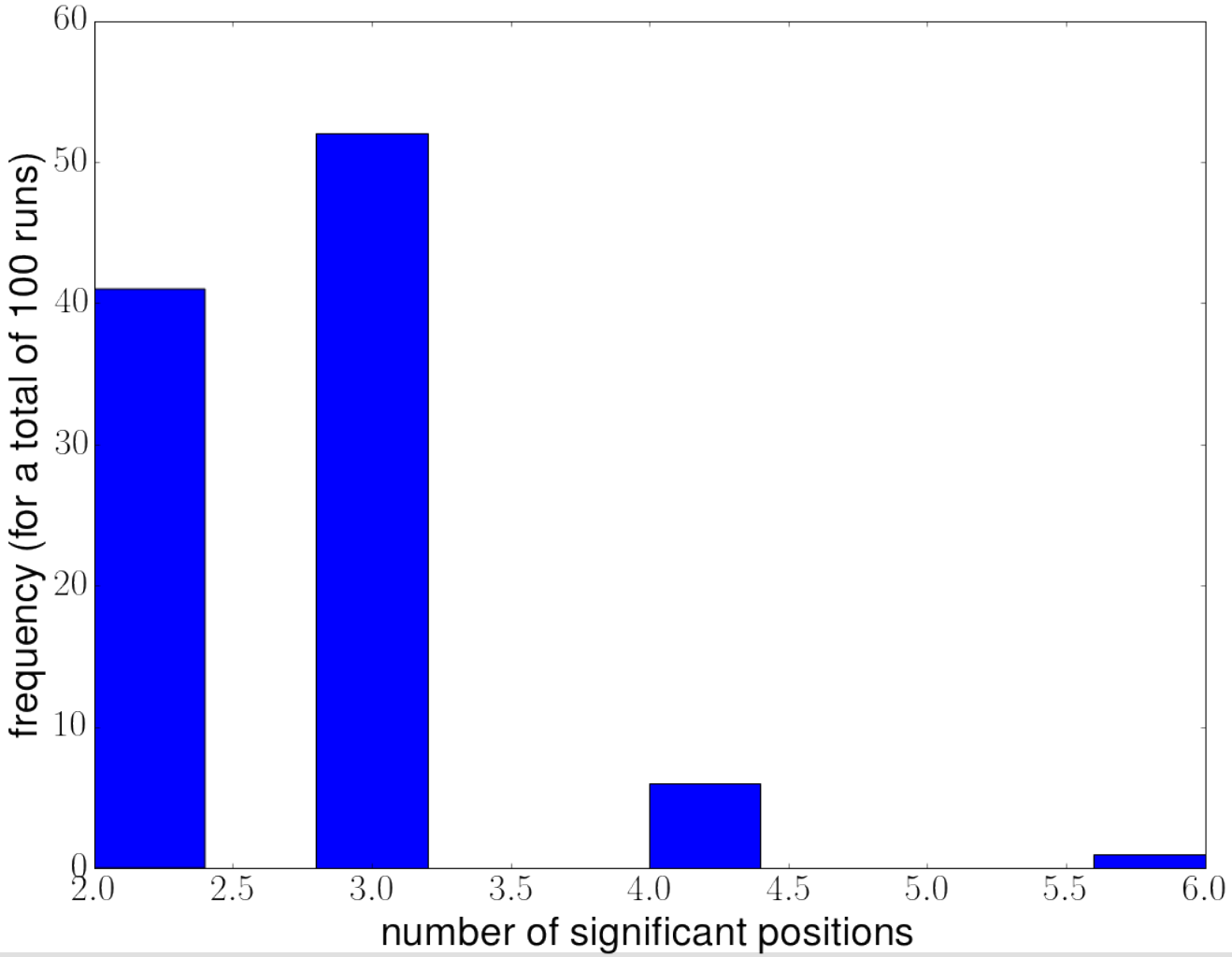
# On average 2.85 identical significant positions for 100 runs.

Comparing the descriptive model with the number of generated words for p=0.19315

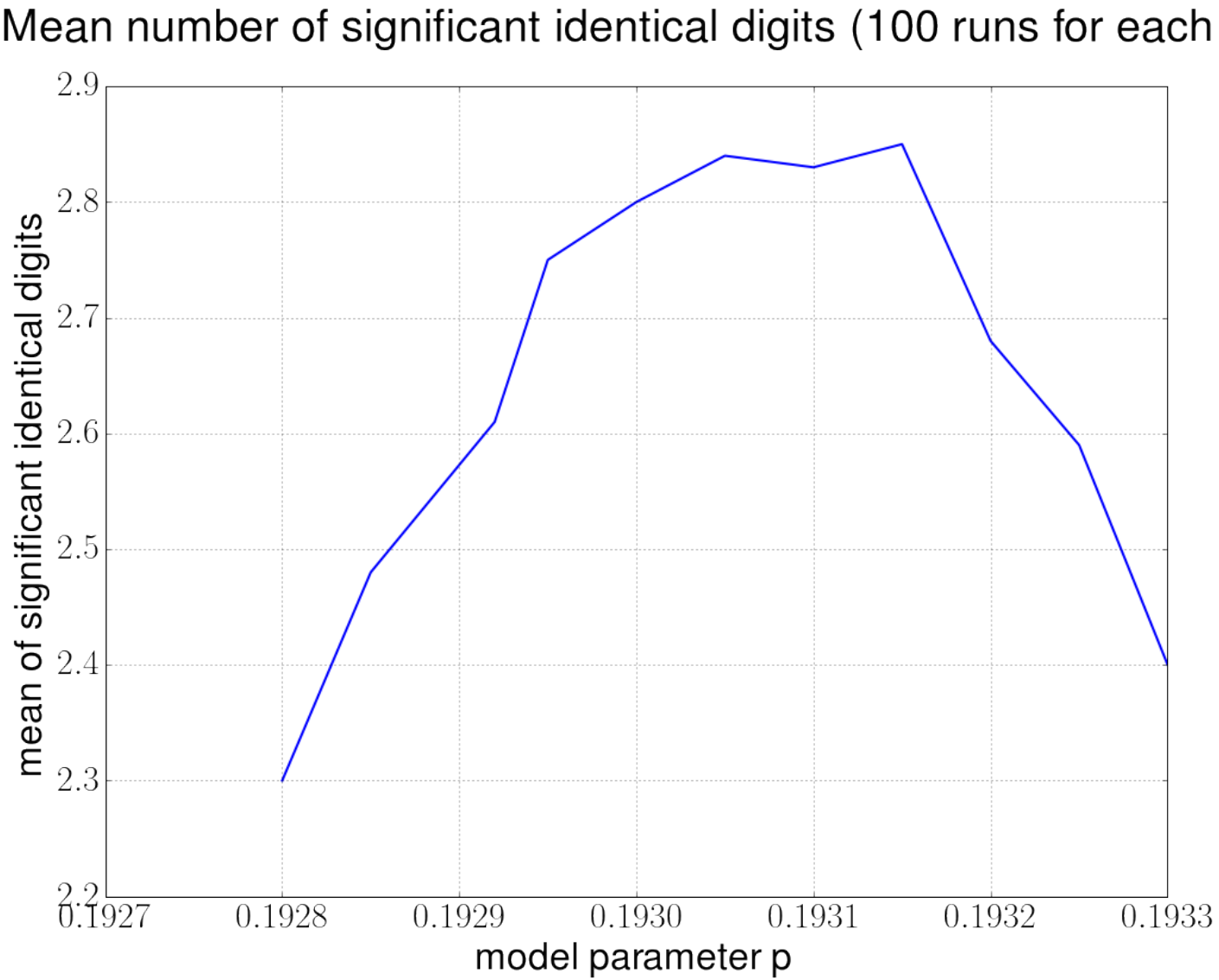


# On average 2.68 identical significant positions for 100 runs.

Comparing the descriptive model with the number of generated words for p=0.1932



# Run the model 100 times for each parameter



# Optimal model parameter somewhere between 0.193 and 0.19315

- Solutions for still existing problems
  - Run more than 100 times (maybe 1000 times)
  - Use other statistics than the mean

- Conclusions
  - About 0.1931 words are created every minute on the Simple English Wikipedia
  - This can explain the amount of words that we found
  - Number could have derived much easier by:
  - Number of words on Wikipedia / uptime in minutes

# 7 Steps for probabilistic, simulated, generative modelling

- Have a descriptive model for comparison
- Formulate a hypothesis
- Select as few model parameters as possible
- Vary the model parameter(s)
- Run the model several times for each parameter set
- For each set of model parameters compare the statistics of the simulated model with the statistics of the descriptive model
- The set of parameters that generate a model that is closest to the descriptive model might yield an explanation

# Thank you for your attention!


Contact:

Rene Pickhardt
Institute for Web Science and Technologies
Universität Koblenz-Landau
rpickhardt@uni-koblenz.de

WeST
People and Knowledge Networks

# Copyright: