Modelling the effects of domestication in Wheat through novel computer-vision techniques

Nathan Hughes

April 18, 2018

Outline

What is the project?

Description

The project is aiming to use computational methods to answer biologically significant questions on wheat grain morphology and domestication using μCT images.

How?

To do this, I will be using:

- Computer vision on 3D image sets
- Statistical analysis and data science
- Scientific theory to create reproducible results

About Wheat Domestication

Why Domestication?

- Answers to questions about diversity in the wheat genus is hidden in the ancestors Cockram et al. [2007]
- Crop breeding depends on making informed decisions, exploring domestication presents an opportunity to augment these decisions

Why ?

- In this project, μ-CT has enabled the study of individual seeds of wheat
- Particularly examining traits which are lost during other methods:
  - Depth, 3D shape, spike location, spikelet formation etc.

Population Diversity

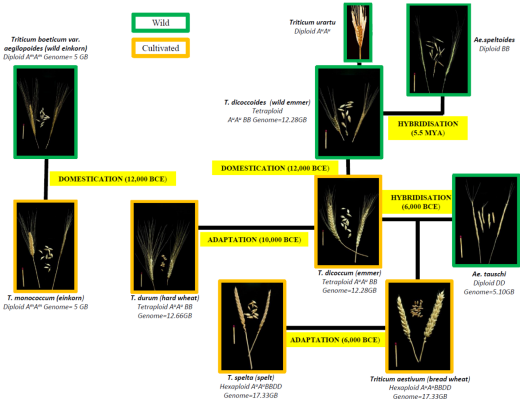

Figure 1: Phylogeny of wheat genotypes (Provided by Dr. Hugo Oliveira)

Research Question:

Is it possible to use μCT imaging to answer questions about Wheat domestication?

I hope so!

Main Groups that we are comparing

1. Einkorn Wild and Einkorn Domesticated
2. Emmer Wild and Emmer Domesticated
3. Spelt and Bread wheat
4. Emmer Domesticated and Pasta Wheat
5. Einkorn Wild and Emmer Wild

Aims

Primary Aims

I am wanting to produce:

- A software library (in Python) which can be used to help analysis of μCT scanned seeds
- A GUI application for researchers to use to auto analyse seeds
- Descriptions of the differences/similarities of the aforementioned groups

Extracted Features

Features List

The features I am collecting are:

- Length
- Width
- Depth
- Volume
- Surface Area
- X,Y,Z coordinates of grains

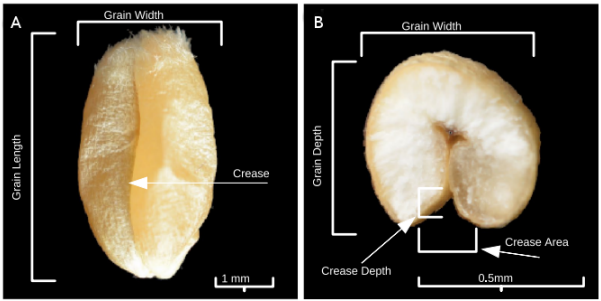

Figure 2: Major features extracted from analysis

Outline

Materials (Plant)

Wheat information

We have a wide range of Wheat genotypes, these are:

- Ranged between diploid, tetraploid and hexaploid
- 12 total genotypes
- Divided between domestication status

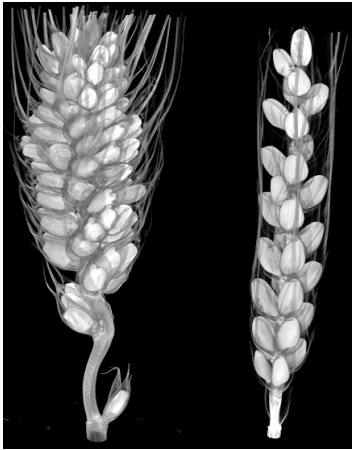

Figure 3: Two wheat spikes, showing diversity in Population, Club Wheat (6N) left, Pasta Wheat right (4N)

Methods

CT Scanning software

The features where extracted using an improved and optimised version of our software which was used in our previous study.

Modifications were implemented to handle the wide range of diversity in the population of this experiment

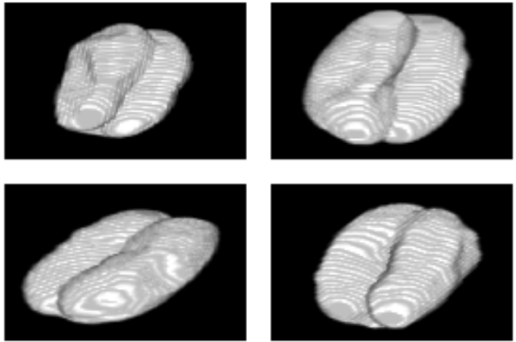

Figure 4: Grains extracted from our imaging software and displayed in 3D

Outline

Why I don't trust the T-Test

Proof of deception

Figure 5: Exhibit A - Why the T-Test is evil

Bayesian Hypothesis Testing

Why?

- "Despite their wide use in scientific journals . . . , statistical hypothesis tests add very little value to the products of research" - [Johnson, 1999]

Bayesian Hypothesis Testing

Why?

- "Despite their wide use in scientific journals . . . , statistical hypothesis tests add very little value to the products of research" - [Johnson, 1999]
- It provides interpretable answers, such as "the true parameter θ has a probability of 0.95 of falling in a 95% credible interval."

Bayesian Hypothesis Testing

Why?

- "Despite their wide use in scientific journals . . . , statistical hypothesis tests add very little value to the products of research" - [Johnson, 1999]
- It provides interpretable answers, such as "the true parameter θ has a probability of 0.95 of falling in a 95% credible interval."
- Allows for missing data points i.e. where a complete range of data is not possible i.e. ALL OF BIOLOGY

Bayesian Model Used

Bayes states that

- $P(A|B) \propto P(B|A) \times P(A)$
  - The posterior is proportional to the likelihood times the prior
- $P(mean.1|sample.1) \propto P(sample.1|mean.1) \times P(mean.1)$

Likelihood is described as

- $y_i^{(g)} \sim T(\nu, \mu, \sigma)$
  - ν (Degrees of freedom) is assumed similar for groups g
  - μ (mean) of groups is assumed the same
  - σ (S.D.) is assumed the same

Prior Mean μ

Mean

- Using the method described in [Kruschke, 2012]
- $\mu_k \sim N(\bar{x}, 2s)$
  - The data are real-values and normal priors are applied (to ensure the posterior follows suit)
  - $2s$ - twice the S.D. ensures no values are favoured in the model

Distribution

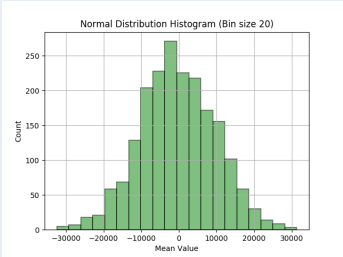

Figure 6: Normal($\bar{x}, 2s$)

Prior Standard Deviations σ

Standard Deviations

- Using the method described in [Kruschke, 2012]
- Uniform$(1, 10000)$ is used
- Whilst no values in the model will have this range, it makes no difference due to random sampling
- Figure:7 shows the distribution expected by random sampling

Distribution

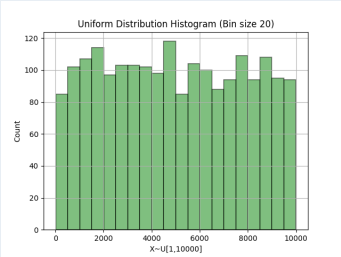

Figure 7: Uniform(1,10000)

Prior Degrees of freedom ν

Degrees of freedom

- Using the method described in [Kruschke, 2012]
- ν of 30 is used with an exponential distribution
- Shown in Figure:8

Distribution

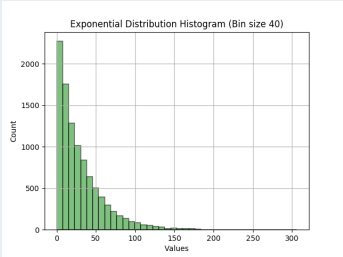

Figure 8: Exponential Distribution

Sampling and Testing

Markov chain Monte Carlo

- 1000 random samples are drawn using Markov chain Monte Carlo
  - This is done twice, independently to ensure convergence of randomness
- These provide a posterior of possibilities where the same mean could exist for the given data

Outline

Example Input Data

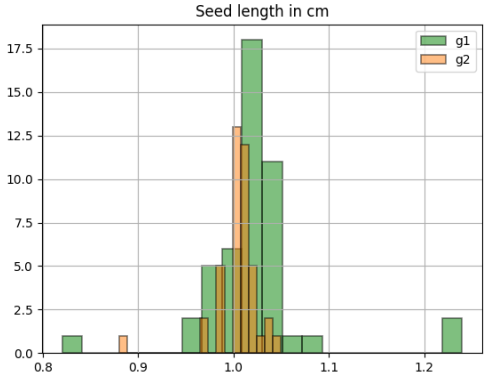

Figure 9: Histogram of input data

Example Posterior

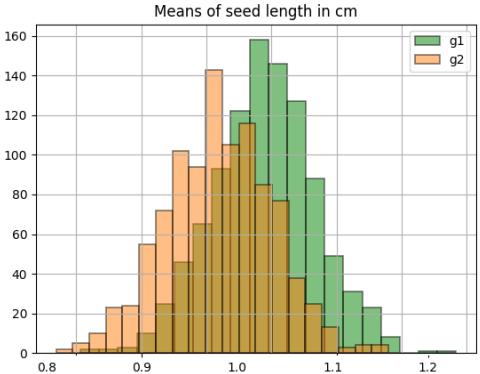

Figure 10: Histogram of posterior data

Example Difference of Means

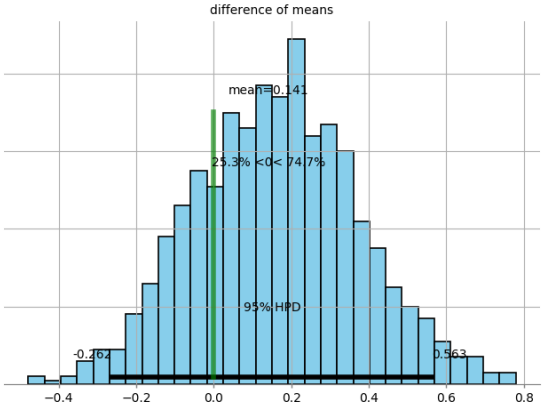

Figure 11: Histogram of posterior data subtracted

Example Forest Plot

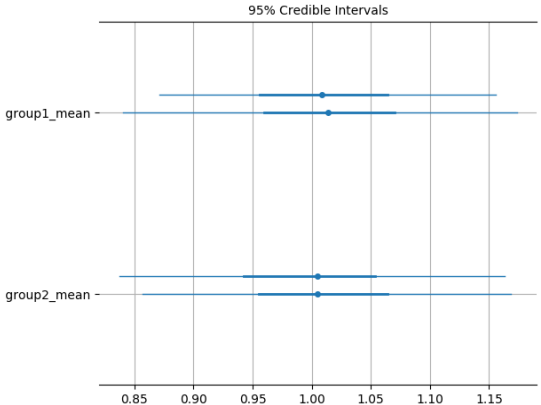

Figure 12: Forest Plot of both chains (bold is 95% of data)

Outline

Einkorn Wild and Einkorn Domesticate ($P < 0.01$)

Boxplots

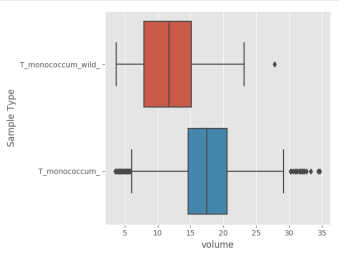

Figure 13: Boxplot for volume

Difference of means

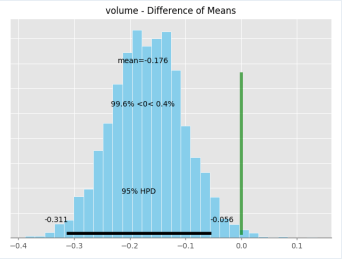

Figure 14: Difference of means

Emmer Wild and Emmer Domesticated ($P = 0.032$)

Boxplots

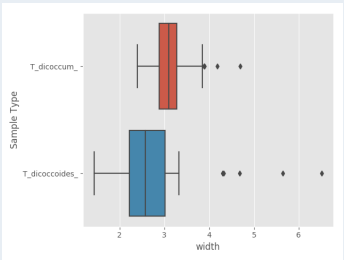

Figure 15: Boxplot for width

Difference of means

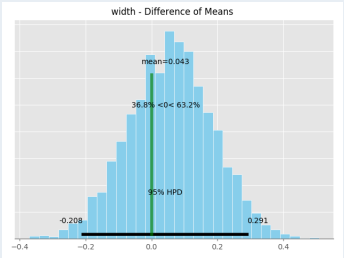

Figure 16: Difference of means

Spelt and Bread wheat ($P = 0.11$)

Boxplots

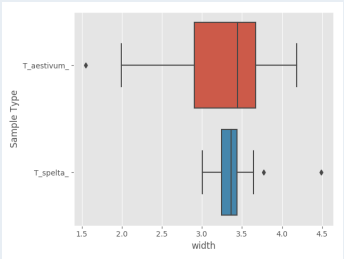

Figure 17: Boxplot for width

Difference of means

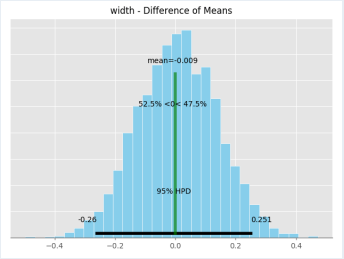

Figure 18: Difference of means

Outline

Loading CT Data

Figure 19: Showing the Data loading window

Investigating CT Data Distributions

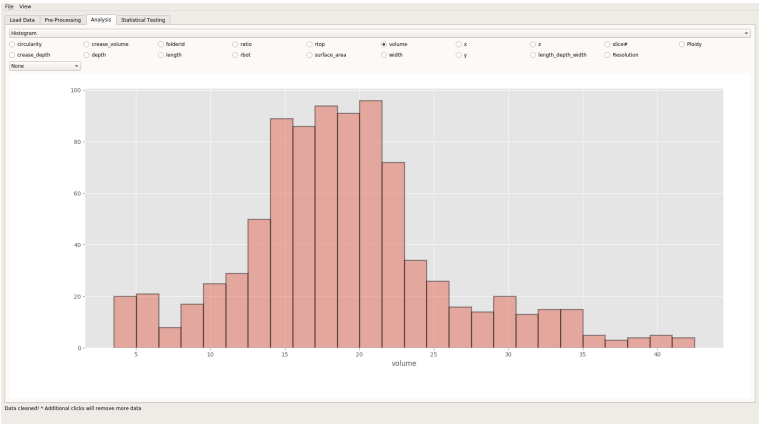

Figure 20: Histogram of some data attributes

Comparing CT Data Distributions

Figure 21: Grouping by data columns

Running T-Tests on CT Data

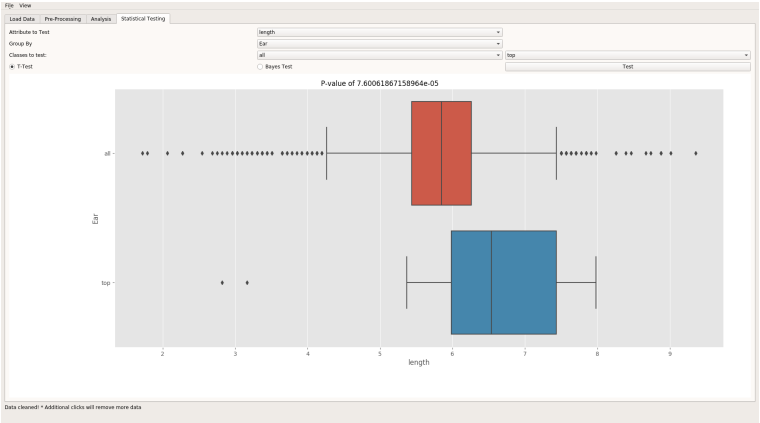

Figure 22: Running T-Tests

Running Bayesian Tests on CT Data

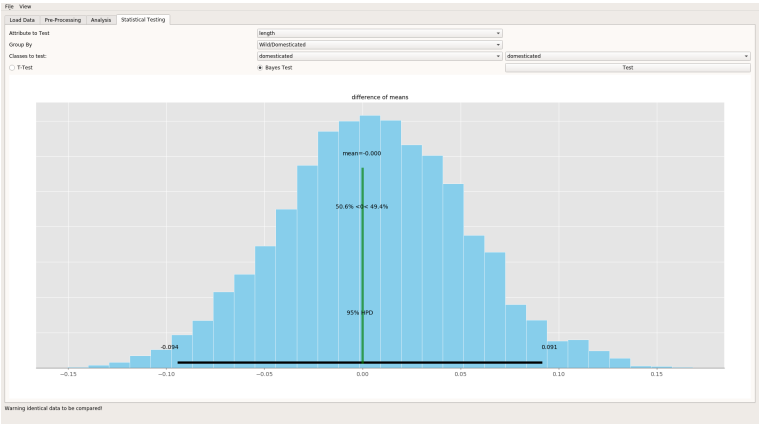

Figure 23: Running Bayesian Tests

Outline

Thanks to

All these people:

Dr. Wayne Aubrey	Prof. John Doonan
Dr. Candida Nibau	Dr. Kevin Williams
Mr. Jason Brook	Everyone at the NPPC

References

James Cockram, Huw Jones, Fiona J. Leigh, Donal O'Sullivan, Wayne Powell, David A. Laurie, and Andrew J. Greenland. Control of flowering time in temperate cereals: Genes, domestication, and sustainable productivity. *Journal of Experimental Botany*, 58(6):1231–1244, 2007. ISSN 00220957. doi: 10.1093/jxb/erm042.

Douglas H. Johnson. The Insignificance of Statistical Significance Testing. *The Journal of Wildlife Management*, 63(3):763, jul 1999. ISSN 0022541X. doi: 10.2307/3802789. URL `http://www.jstor.org/stable/3802789?origin=crossref`.

John K Kruschke. Bayesian estimation supersedes the t test. *Journal of Experimental Psychology: General Version of May*, 31, 2012. URL `http://www.indiana.edu/{~}kruschke/BEST/BEST.pdf`.