An example of using Bayesian statistics: revised OSD heterogeneity testing under certain conditions

Kirk Remund and Jean-Louis Laffont

This work was carried out as part of the activities of the Vegetable Seed Industry Working group (VSI WG).

The problem:

For certain species with a well-established manufacturing process, historical Other Seeds data show a low level of OS and homogeneity of OS within lots.

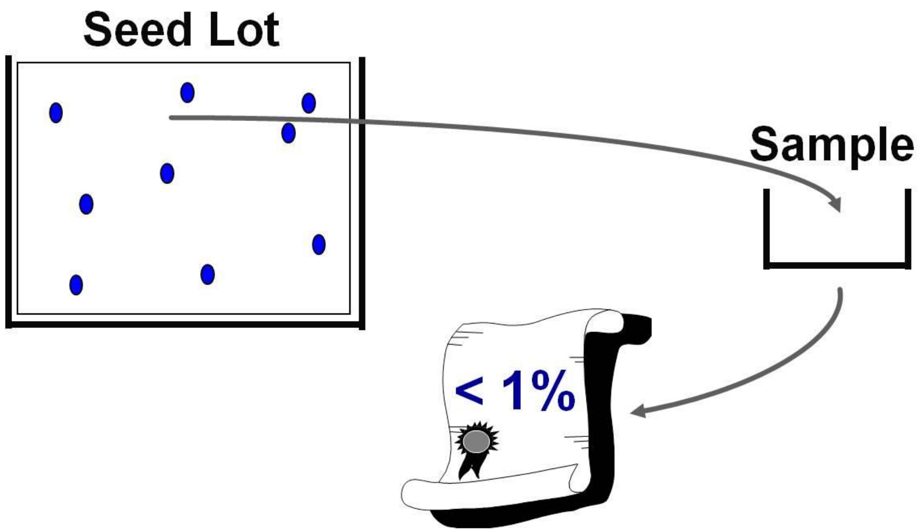
Given this information, can we reduce the level of testing (using fewer seeds) to test the homogeneity of a lot with respect to OS?

Using **Bayesian Statistics** to answer this question

Refresher on statistical inference

Drawing conclusions about populations from data collected samples

Example: find the probability that the adventitious presence (AP) of GM seeds in a conventional seed lot is below 1% given no GM seeds were found in a sample of 1,000 seeds



Why Bayesian statistics?

1. We can use external information (called prior information) to improve our statistical inference.

Example: a seed company tested a conventional seed lot for AP and stated:
we are 95% confident that the % of GM seeds in the lot is below 0.1%

A new sample of 1,000 seeds is tested by a 3rd party lab exhibiting no GM seeds.

Find the probability that AP is below 0.5% given this result and taking into account the information provided by the seed company.

Why Bayesian statistics?

2. We can compute probabilities associated to subsequent samples (called posterior predictive probabilities) given what we found in previous samples

Example: no GM seeds were found in a sample of 3,000 seeds.

Find the probability to find 1 GM seed in a new sample of 1,000 seeds.

Why Bayesian statistics?

3. The probabilities computed in Bayesian statistics are about the parameter given the data observed whereas Frequentist statistics usually compute probabilities of an event given an hypothesis

Example: purity of a seed lot

Bayesian statistics:
probability that the seed lot purity is above 99.5% given what we observed in the sample

This is really what we want

Frequentist statistics:
probability to get what we observe in the sample given the hypothetical purity is above 99.5%

What are Bayesian statistics ?

We have some data Y, and we want to know about a parameter θ

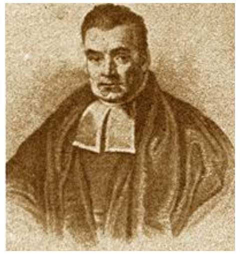
Example:

- Y : number of OS found in a sample of 2,500 seeds
- θ : OS proportion in the lot

What is the probability that θ is below 0.5% given Y ?

Bayes formula:

$$P(\theta \mid Y) = \frac{P(Y \mid \theta) \times P(\theta)}{P(Y)}$$



Reverend Thomas Bayes (1702-1761)

$P(\theta \mid Y)$: also called the **posterior**

$P(Y \mid \theta)$: what is usually computed in Frequentist statistics. Called the **likelihood**

$P(\theta)$: what we know about θ independently of the data. Called the **prior**

$P(Y)$: the probability of the data for all values of θ :

$\sum_{\theta \in \Theta} P(Y|\theta)P(\theta)$ or $\int_{\Theta} P(Y|\theta)P(\theta)d\theta$ (**constant**)

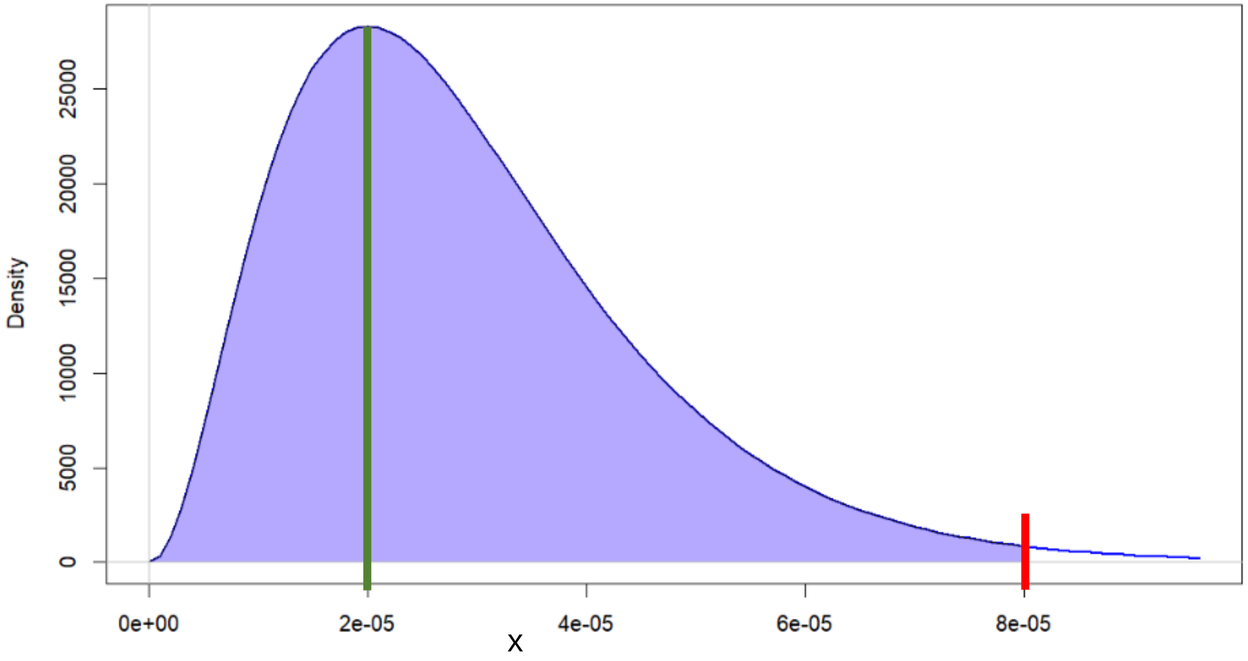
What are Bayesian statistics ?

- A convenient distribution to describe our **prior beliefs** about θ is the **beta distribution** with 2 parameters α and β

- From historical knowledge, we can state our beliefs about θ as follows:

  The most likely value is M and there is a $100P\,\%$ chance that the value is below D.

- This leads to the following beta distribution with parameters $\alpha = 3.18$ and $\beta = 108745$ for $M = 0.5/25000$, $D = 2/25000$, and $P = 99\%$:



What are Bayesian statistics ?

Example:

- Y : one OS found in a sample of 2,500 seeds → $Y \mid \theta \sim \mathrm{Bin}(n, \theta)$
- θ : OS proportion in the lot → $\theta \sim \mathrm{Beta}(\alpha, \beta)$

What is the probability that θ is below 0.5% given Y ?

posterior **likelihood** **prior**

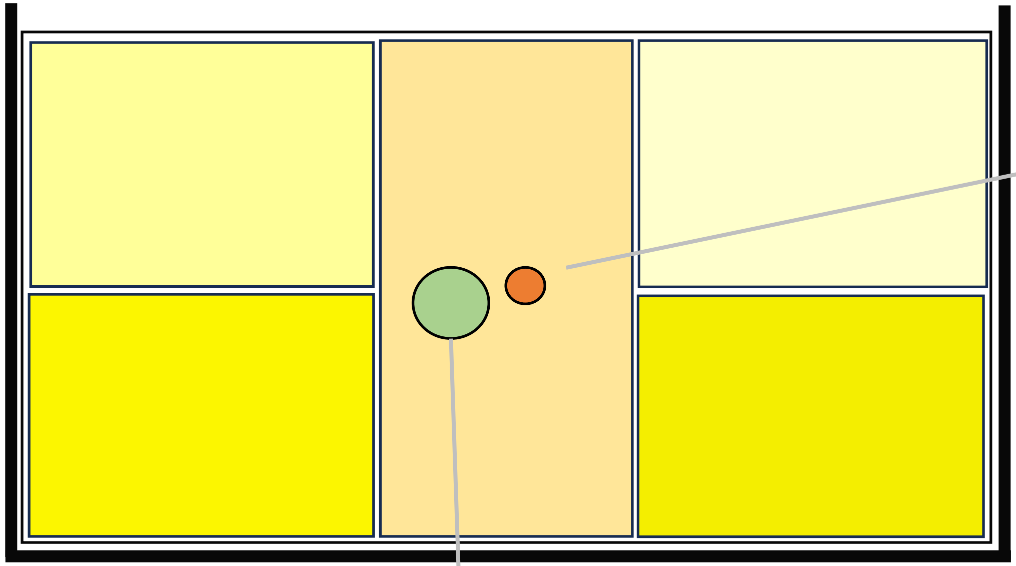
$$P(\theta \mid Y) = \frac{P(Y \mid \theta) \times P(\theta)}{P(Y)} = \frac{1}{\mathrm{B}(\alpha + y, \beta + n - y)} \theta^{\alpha + y - 1} (1 - \theta)^{\beta + n - y}$$

constant

$\mathrm{B}(x,y)$: Beta function

$$\rightarrow \quad \theta \mid y \sim \mathrm{Beta}(\alpha + y, \beta + n - y)$$

Now, consider the following problem:

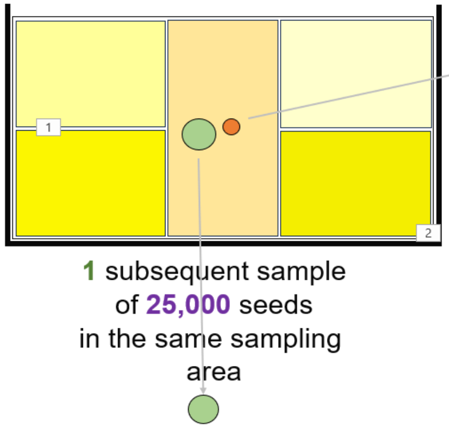


1 sample
of 2,500 seeds

y **OS** out of
2,500 seeds

1 subsequent sample
of **25,000** seeds
in the same sampling
area

What is the probability of having
0, 1, 2, 3, … **OS** among the **25,000** seeds in this new sample, given that y **OS** were found in the initial sample of **2,500** seeds?



1 sample of **2,500** seeds

y **OS** out of **2,500** seeds

1 subsequent sample of **25,000** seeds in the same sampling area

What is the probability of having 0, 1, 2, 3, … **OS** among the **25,000** seeds in this new sample, given that ***y*** **OS** were found in the initial sample of **2,500** seeds?

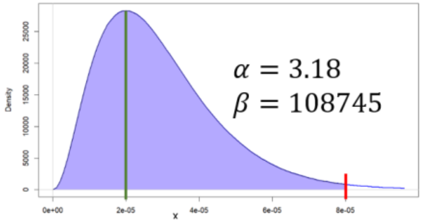
Posterior predictive density:

$$f(\tilde{y}|y) = \binom{m}{\tilde{y}} \frac{\mathrm{B}(\alpha + y + \tilde{y}, \beta + n - y + m - \tilde{y})}{\mathrm{B}(\alpha + y, \beta + n - y)}$$

Derived from the beta posterior distribution:

$$\begin{aligned} f(\tilde{y}|y) &= \int_\theta f(\tilde{y}|\theta) f(\theta|y) d\theta \\ &= \int_0^1 \binom{m}{\tilde{y}} \theta^{\tilde{y}} (1-\theta)^{m-\tilde{y}} \frac{1}{\mathrm{B}(\alpha+y, \beta+n-y)} \theta^{\alpha+y-1} (1-\theta)^{\beta+n-y-1} d\theta \\ &= \binom{m}{\tilde{y}} \frac{1}{\mathrm{B}(\alpha+y, \beta+n-y)} \int_0^1 \theta^{\alpha+y+\tilde{y}-1} (1-\theta)^{\beta+n-y+m-\tilde{y}-1} d\theta \\ &= \binom{m}{\tilde{y}} \frac{\mathrm{B}(\alpha+y+\tilde{y}, \beta+n-y+m-\tilde{y})}{\mathrm{B}(\alpha+y, \beta+n-y)} \end{aligned}$$

Example: **Prior**:



No OS found in the sample of **2,500** seeds ($y = 0$)

$$f(\tilde{y} = 0 | y = 0) = 0.5253904$$

$$f(\tilde{y} = 1 | y = 0) = 0.3060821$$

$$f(\tilde{y} = 2 | y = 0) = 0.1172369$$

$$f(\tilde{y} = 3 | y = 0) = 0.03710573$$

$$f(\tilde{y} = 4 | y = 0) = 0.01050974$$

99.6% chance to have less than **5 OS** in a subsequent sample of **25,000** seeds

Heterogeneity test * considered in the following:

- Test statistic:

# of samples used in the test (e.g. 5)

$$H = \frac{(K-1)s^2}{\sigma_0^2}$$

Observed variance among the results x_i (i = 1, …, K)

Reference variance for a Poisson distribution inflated by a factor of 1.4: $\sigma_0^2 = 1.4\bar{x}$

Table 2E. Factors for additional variation in seed lots to be used for calculating *W* and finally the H value

Attributes	Non-chaffy seeds	Chaffy seeds
Purity	1.1	1.2
Other seed count	1.4	2.2
Germination	1.1	1.2

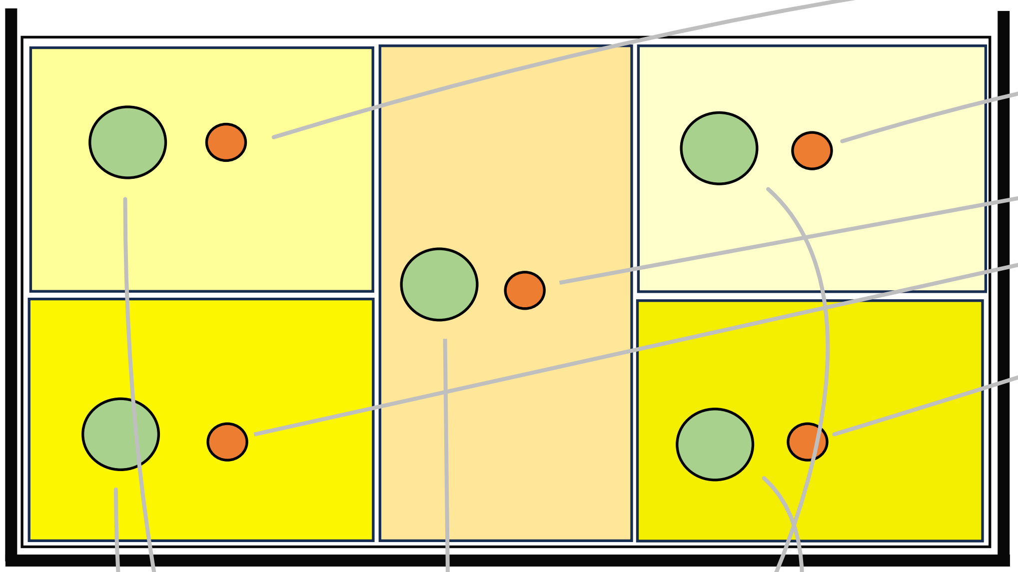
- We reject the hypothesis of homogeneity if:

$H > \chi^2_{1-\delta,K-1}$ (quantile with probability $1-\delta$ of chi-square distribution with $K-1$ degrees of freedom)

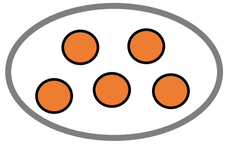
Example: $\chi^2_{0.99,5-1}$ = 13.2767

* This test is referenced in the ISTA Rules, Chapter 2, under the name 'H value test'

Now, a new problem:

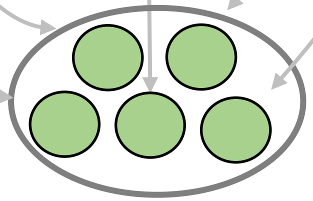


5 samples
of **2,500** seeds



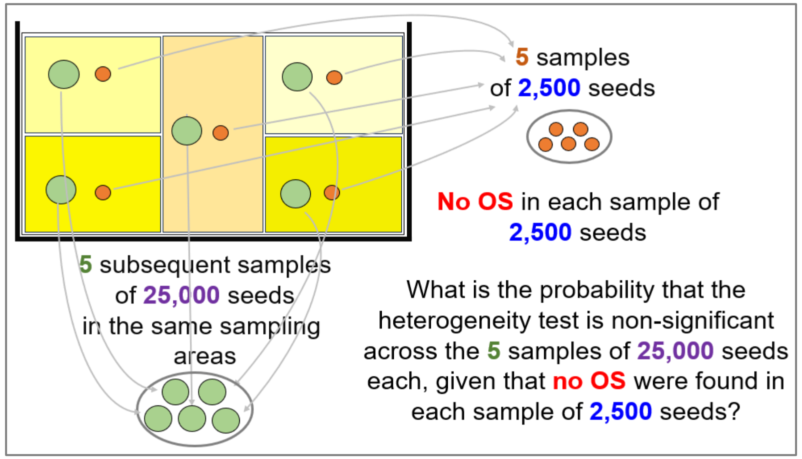
No OS in each sample of
2,500 seeds

5 subsequent samples
of **25,000** seeds
in the same sampling
areas

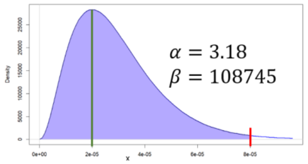


What is the probability that the heterogeneity test is non-significant across the **5** samples of **25,000** seeds each, given that **no OS** were found in each sample of **2,500** seeds?

Monte Carlo simulations



Select prior:



Generate * 5 sample values from posterior predictive distribution:

$$f(\tilde{y}|y=0) = \binom{25000}{\tilde{y}} \frac{B(3.18+0+\tilde{y}, 108745+2500-0+25000-\tilde{y})}{B(3.18+0\ ,\ 108745+2500-0)}$$

Determine if the heterogeneity test is non-significant for these 5 generated values

Repeated

N times

Compute the probability that the heterogeneity test is non-significant as $\frac{T}{N}$ where T is the number of times the test is non-significant

* R function:

```
rPostPred <- function (N = 10, y = 0, n = 2500, m = 25000, a = 1, b = 1)
{
    # Generate N random numbers from posterior predictive distribution
    # y: number of OS found in the sample of n seeds
    # m: number of seeds in the subsequent sample
    # a, b: parameters of the prior Beta distribution
    postPred.pdf<-function(x,y,n,m,a,b)
    {
        return(exp(lchoose(m,x)+lbeta(a+y+x,b+n-y+m-x)-lbeta(a+y,b+n-y)))
    }
    postPred.cdf <- function(x,y,n,m,a,b) {
        sum(sapply(0:x, function(k) postPred.pdf(k,y,n,m,a,b))) }
    inverse.postPred.cdf <- function(p,y,n,m,a,b) {
        k <- 0
        cdf.value <- postPred.cdf(k,y,n,m,a,b)
        while (cdf.value < p) {
            k <- k + 1
            cdf.value <- cdf.value + postPred.pdf(k,y,n,m,a,b)
        }
        return(k)
    }
    rn<-replicate(N, inverse.postPred.cdf(runif(1), y, n, m, a, b))
    return(rn)
}
```

Example: `rPostPred(5,0,2500,25000,3.18,108745)`

```
[1] 1 3 0 3 0
```

$$H = \frac{(5\ -1)2.3}{1.4 \times 1.4} = 4.693878$$

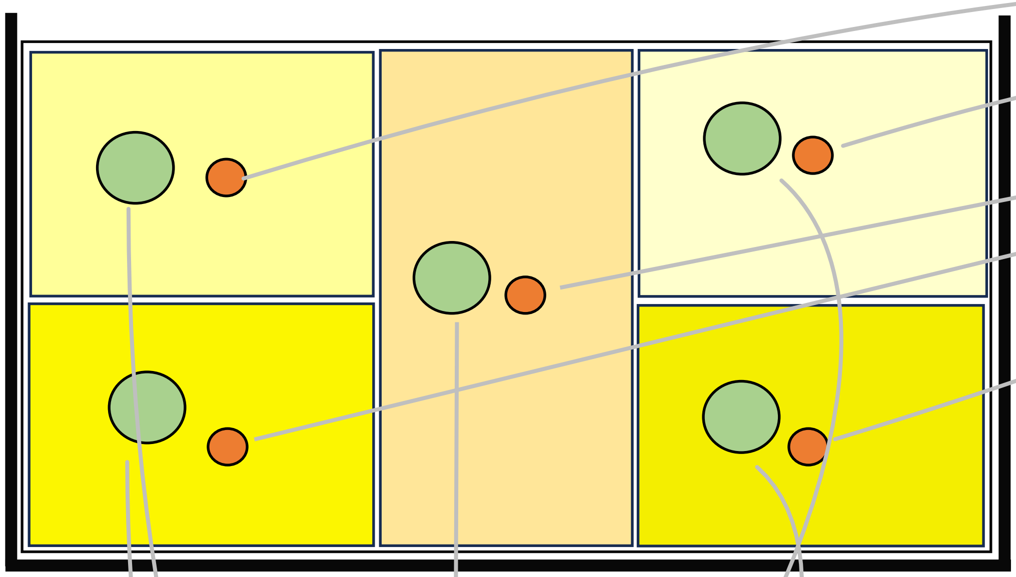
$$\chi^2_{0.99,5-1} = 13.2767$$

Non significant

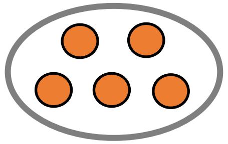
Example:

P(*Normal Het test NS* | y_i*'s all* = 0) = 0.99725

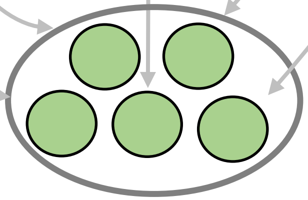
And finally, the complete problem:



5 samples
of 2,500 seeds

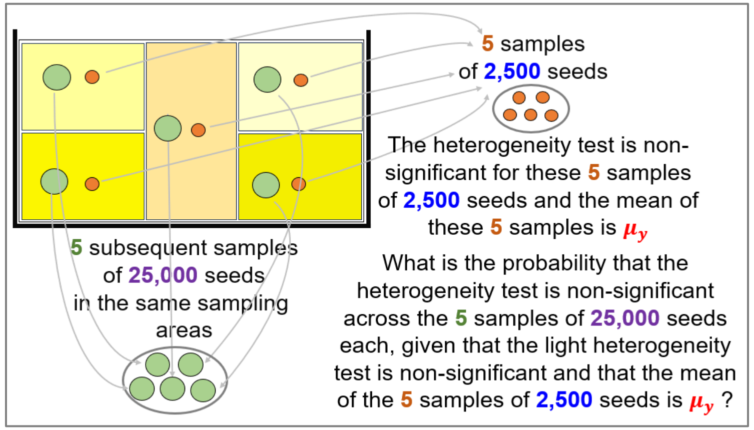


5 subsequent samples
of 25,000 seeds
in the same sampling
areas



The heterogeneity test is non-significant
for these 5 samples of 2,500 seeds
and the mean of these 5 samples is μ_y

What is the probability that the heterogeneity test is non-significant across the 5 samples of 25,000 seeds each, given that the light heterogeneity test is non-significant and that the mean of the 5 samples of 2,500 seeds is μ_y ?



Monte Carlo simulations

For a given μ_y:

Generate * K homogeneous values of the light test

Define the prior and generate K values of the normal test using the posterior predictive distribution

Determine if the heterogeneity test is non-significant for these K generated values

Repeated N_1 times

Compute the probability p_1 that the heterogeneity test is non-significant as $\frac{T}{N_1}$ where T is the number of times the test is non-significant

Repeated N_2 times

Compute the probability p that the heterogeneity test is non-significant as the mean of the p_1's

* Over-dispersed binomial data:

$$variance = 1.1 \times Binomial\_variance$$

Table 2E. Factors for additional variation in seed lots to be used for calculating W and finally the H value

Attributes	Non-chaffy seeds	Chaffy seeds
Purity	1.1	1.2
Other seed count	1.4	2.2
Germination	1.1	1.2

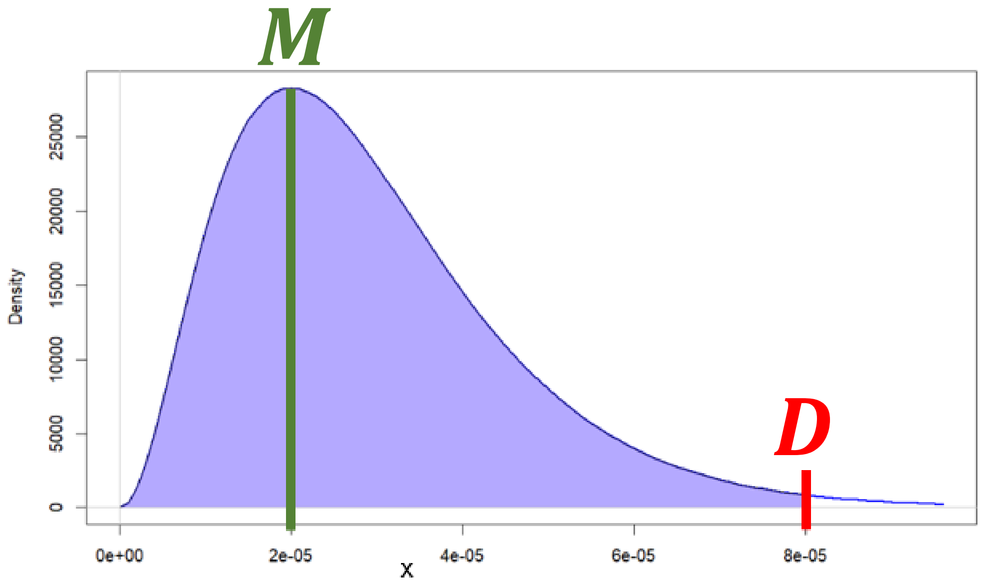
Beta-binomial distribution with parameters:

$$A = \frac{\mu_y}{2500}\left(\frac{2500-1}{1.1^2-1} - 1\right)$$

$$B = A\left(\frac{2500}{\mu_y} - 1\right)$$

Generating the K values ensuring they meet the constraint that the heterogeneity test is non-significant

Choice of prior based on historical data/expert feedback:



Most likely value M	Value D with 99% chance to be below
0%	1/25000 = 0.004%
0%	2/25000 = 0.008%
0%	5/25000 = 0.02%
0.5/25000 = 0.002%	1/25000 = 0.004%
0.5/25000 = 0.002%	2/25000 = 0.008%
0.5/25000 = 0.002%	5/25000 = 0.02%

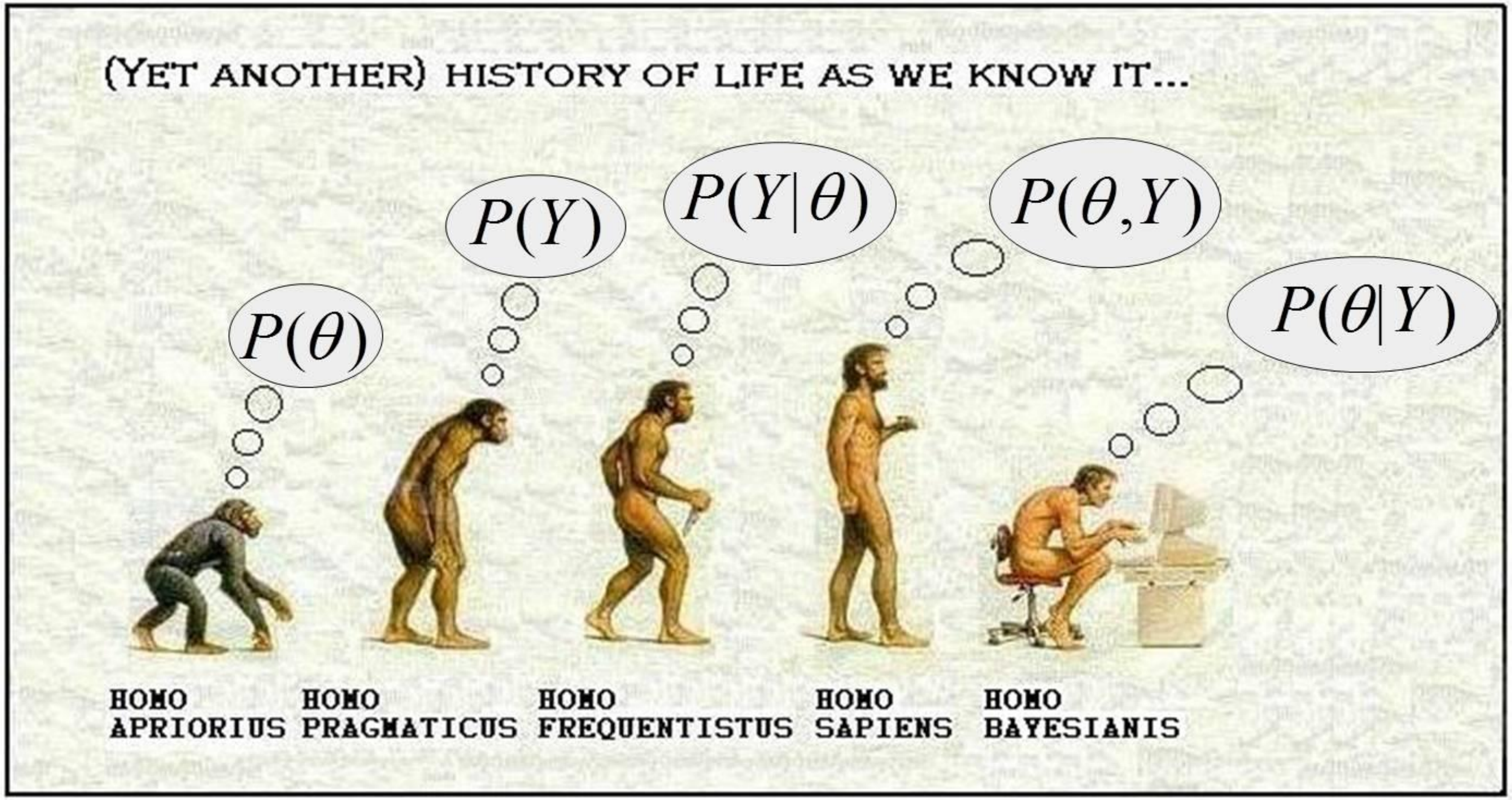
Results: probability that the heterogeneity test performed on 5 samples of 25,000 seeds is non-significant given that :

. An heterogeneity test performed on 5 samples of 2,500 seeds is non-significant with, on average, 0 (0%), 1 (0.04%), 2 (0.08%) OS found in this test

. Different hypotheses regarding the prior distribution

	$\mu_y = 0$ (0%)	$\mu_y = 1$ (0.04%)	$\mu_y = 2$ (0.08%)
$M = 0\%, D = 0.004\%$ ($P = 0.99$)	0.999664	0.997664	0.995584
$M = 0\%, D = 0.008\%$ ($P = 0.99$)	0.996048	0.986336	0.979264
$M = 0\%, D = 0.02\%$ ($P = 0.99$)	0.967056	0.918672	0.892688
$M = 0.002\%, D = 0.004\%$ ($P = 0.99$)	0.999616	0.999552	0.999568
$M = 0.002\%, D = 0.008\%$ ($P = 0.99$)	0.997488	0.996272	0.993984
$M = 0.002\%, D = 0.02\%$ ($P = 0.99$)	0.974752	0.949456	0.933264

Reproducibility probabilities of the 'normal' (25,000 seeds) OSD heterogeneity test derived from the 'light' (2,500 seeds) OSD are very high (above 89%)



Proba of an hypothesis no matter what data tells

Only interested in data

Proba of the data given the hypothesis

Proba of the data and the hypothesis

Proba of the hypothesis given the data

Image credit: from a talk from Mike West

ISTA ANNUAL MEETING 01-04 JULY 2024 CAMBRIDGE, UNITED KINGDOM

Thank you