A critique of the statistical protocol in ISO 20072 for aerosol drug delivery device design verification.

CUED/F-INFENG/TR.703

October 8, 2016

M.S. Christescu, M.A. Tirlea, D.J.C. MacKay

ISSN 0951-9211

University of Cambridge
Department of Engineering
Trumpington Street
Cambridge CB2 1PZ
United Kingdom

Email: msc64 / mat58 @cam.ac.uk

A Critique of the Statistical Protocol in ISO 20072 for Aerosol Drug Delivery Device Design Verification

Matei S. Christescu, Marius A. Tirlea
Trinity College, Cambridge

David J. C. MacKay*
Department of Engineering, University of Cambridge

October 8, 2016

Abstract

ISO 20072 is an international standard detailing the methods and testing protocols required to ensure the quality of aerosol drug delivery devices (ADDDs), for use in humans [1]. Our paper examined the pitfalls and errors in this standard, and the difficulties in its application to ADDD-producing factories. We find that the likelihood of many imperfect factories which satisfy the requirements listed in the standard actually passing the prescribed testing is minuscule. We suggest other testing protocols, which achieve the stated goals of the standard. In the specific example we consider, they also reduce the workload by around 25-fold compared with the methods specified in ISO 20072, as well as giving crucial additional advantages.

*Sadly David MacKay FRS, Regius Professor of Engineering, died in April 2016 during the preparation of this report.

Contents

1 Introduction

In 2009, the International Standards Organization published a standard, ISO 20072 (see [1]), which describes a protocol for specifying and verifying the properties of a factory producing inhalers.

A paper prepared by the International Pharmaceutical Aerosol Consortium on Regulation and Science (IPACRS), the European Pharmaceutical Aerosol Group (EPAG) and other experts (see [2]) criticizes the 2007 draft of ISO 20072 in the following statement: "the values for [the minimum probability content] p proposed in ISO DIS 20072:2007 are set so high that meeting the statistical requirements will be challenging or even impossible for many ADDD-based products.". We investigated whether or not this statement is true.

The intentions of ISO 20072 are noble. The aim of the standard is to require aerosol drug delivery device (ADDD) manufacturers to comply with stringent testing protocols, in order to ensure that the ADDDs are of a high quality. However, there are two problems with the standard: a minor error, which is easily correctable but has had the effect of satisfactory factories being rejected by the standard, and a major flaw, which renders the entire standard unfit for purpose. The second issue leads to manufacturers being tempted to manipulate their test data in order for their factories to be deemed compliant with this standard.

1.1 Overview of the paper

In sections 2, 3, and 4, we explain what a "Device Functionality Profile" is and define some terms we will use in this report, then using an example Device Functionality Profile, discuss the testing protocol mandated by ISO 20072.

We then explain the minor error which occurs in table D1, Annex D of ISO 20072. This table contains the values of k to be used in a particular two-sided test of a continuous variable, under the assumption that this variable follows a Gaussian distribution. This table contains values which are often incorrect to the first decimal place, and even before for some values of n. For instance, one of the correct k values is 48% lower than the incorrect one listed in table D1 of ISO 20072.

We discuss this issue further in section 4.2. However, this error can be simply remedied by replacing the incorrect table of k values with a correct one, which is why we consider it to be a minor problem.

In section 5 we examine the sample sizes needed, which leads us in sections 5.2, 5.3, and 5.4 to describe the major problem, which is that the statistical tests in the ISO 20072 protocol are deeply flawed - so flawed that it is essentially impossible for many factories that are in fact satisfactory to pass all non-trivial tests[1], unless the tests are rigged (for example, by censoring unfavourable results).

The public is then seen to be in the following highly unsatisfactory situation: all devices that are stated to have complied with ISO 20072 and have non-trivial DFP requirements are extremely likely to have done so by a testing process that in fact did not comply with the ISO standard:

[1] For examples of non-trivial requirements of the DFP, see section 2.1

some of the test results will have been censored or falsified. This not only implies that patients are at risk, but also that ISO 20072 is not fit for purpose. Note that we do not imply that the censoring or falsification of the results is done with malicious intent; rather, we note in section 5.3.1 that the workers in question may be bending over backwards to try to reconcile the commercial pressure to pass the test with the difficulty of doing so, but have insufficient knowledge of statistics to understand that what they are doing constitutes falsification of the results.

The root of the problem with the standard is that it is based on classical "frequentist" statistical tests, which calculate a parameter called the "(frequentist) confidence" of the test, and which do not pay attention to the properties that we actually care about, such as the probability that a "good" factory might fail the test, or the probability that a factory that has passed the test is actually "bad".

In section 6 we therefore propose an alternative approach, aiming for the same high standard of quality control, but using Bayesian statistics. In section 7 we see that this results in both enormously reduced workload and in a much higher probability of a compliant factory passing the test. In section 8 we look briefly at what may be motivating regulators, based on the comments of others, before concluding in section 9.

2 The Device Functionality Profile (DFP)

2.1 Example of a DFP

In order to understand how the major flaw in ISO 20072 arises, we must first consider how the this standard refers to a device functionality profile (DFP), developed by the manufacturer in conjunction with the requirements of ISO 14971. This DFP is a guide to the intended properties and acceptable limits of the parameters of the ADDD. An example DFP is shown in table 1: this DFP is for an inhaler which, in addition to its drug delivery functions, also comes to and reports judgments about whether the device has been correctly used, by monitoring the vibrations occurring in the device during use.

We make various comments on the specific DFP items in table 2.

2.2 Rationale behind the various basic items in the DFP

From a manufacturer's perspective, compliance with the items in the DFP shows that the device is working and safe to use. Therefore, a manufacturer is faced with a trade-off: to include many items in the DFP, thus enhancing the safety of the patients and at the same time increasing the amount of testing needed, or to sacrifice safety of the patients by reducing the number of items in the DFP, which in turn reduces the amount of testing needed.

We note several properties of the example DFP:

1. There are 16 basic items in the DFP to be tested within ISO 20072, since item 9 adds three separate basic items to the DFP.
2. There are four types of variable in the DFP:
   (a) Directly observable, real-valued, continuous variables, constrained to be between finite lower and upper bounds, l and u, respectively (e.g. item 1).
   (b) Directly observable, real-valued, continuous variables, constrained either to be below a finite upper bound u or to be above a finite lower bound l (e.g. item 3).
   (c) A probability, which cannot be directly observed (e.g. item 2).
   (d) A Poisson rate, which cannot be directly observed (e.g. item 8).
3. Each variable specified within the DFP has some distribution across the population of devices.

There are several additional specifications related to the example DFP:

1. Temperature range for correct operation must be at least $-15°C$ to $+45°C$, with relative humidity between 5% and 95%.
2. Storage temperature range must be at least $-30°C$ to $+60°C$.
3. Operating and storage atmospheric pressure range must be at least 0.5 to 1.2 atmospheres.

Item	Description	Type of variable
1	The force needed to remove the cap must be not less than 8 N, and not more than 12 N.	Real, two-sided
2	The probability, given a particular device, which has not warned of low battery, must be at least 0.99 that power comes on within 5 seconds of removal of the cap.	Probability parameter
3	The force needed to replace the cap must be not more than 15 N.	Real, one-sided
4	The probability, given a particular device, must be at least 0.99 that power turns off within 5 seconds of replacing the cap.	Probability parameter
5	The pressure drop to achieve 60 l/min flow must be between 0.9 and 1.2 kPa.	Real, two-sided
6	The probability, given a particular device, must be at least 0.93 that, after a correct inhalation, it gives the response appropriate to correct use.	Probability parameter
7	The probability, given a particular device, must be at least 0.93 that, after an incorrect inhalation, it gives the response appropriate to incorrect use.	Probability parameter
8	The Poisson rate of false detection of inhalations, given a particular device, must be less than 1 per hour of power-on time.	Poisson rate
9	The device must still operate correctly (i.e. satisfy items 6, 7, and 8 in the DFP) when 1 m in front of the standard acoustic background noise generator	(as 6, 7, 8)
10	The battery must last for at least 2 hours of cap-off time before needing replacement.	Real, one-sided
11	The device must give a low-battery warning between 10 and 20 minutes of power-on time before the battery will expire if not turned off.	Real, two-sided
12	Peak current drain from the battery must not exceed 500 μA.	Real, one-sided
13	The device must have mass not exceeding 12 g, including the battery.	Real, one-sided
14	The mean particle radius in the delivered airflow during a correct inhalation must not exceed that in the dose of drug powder put into the device by a factor of more than 1.2	Real, one-sided

Table 1: Example DFP. The third column indicates which type of variable the respective item in the DFP is referring to. Note that drug dose delivered is outside the scope of ISO 20072, although of course it does in practice also need to be tested.

DFP item	Comments
1	Lower bound required so that the cap does not fall off accidentally, and upper bound required so that a physically weaker user will never be unable to remove the cap
2	Required since, otherwise, a user may not be able to use the ADDD.
3	See comment on item 1.
14	This might happen if the device unintentionally charged the particles with static electricity.

Table 2: Comments on some of the items in the example DFP.

3 Technicalities

We now turn to various definitions and assumptions that are necessary to make sense of the rest of this report.

3.1 Definitions

For the avoidance of ambiguity we here define some specific terms for use in the rest of the report.

By a **factory** we mean whatever combination of product design, manufacturing method, and manufacturing facility is used to produce a population of inhalers whose properties are to be tested.

A **measurement** is a single experiment on a single device, done exactly once. Thus, for example, dropping a device once and seeing whether it breaks is a measurement. Applying a single inhalation to a single device and determining whether it gives the correct or incorrect result is a measurement.

A **basic item** is one numbered item in the DFP defined therein without reference to ISO 20072. Thus for example "The pressure drop to achieve 60 l/min flow must be between 0.9 and 1.2 kPa" is a basic item.

An **item** is one of the sets of measurements that results from combining a basic item defined in the DFP with one of the environmental and/or insulting test conditions defined in ISO 20072. Thus, for example, "The pressure drop to achieve 60 l/min flow must be between 0.9 and 1.2 kPa when testing at standard atmosphere after exposure to cold storage atmosphere" is an item.

A **test** is a procedure that involves testing all the different items invoked by the combination of the DFP and ISO 20072, the outcome of which is that the factory either passes or fails the test.

In this case a factory passes a Bayesian test, using a suitable prior, if we confirm for each item that given the data the probability that the item's constraints hold for the relevant fraction (95% or 97.5% depending on environmental conditions) of produced devices is at least 0.95 . We refer

to this posterior probability as the **confidence** that an item is complied with (see the definition of fconfidence below for some justification for this unusual nomenclature).

A factory passes a frequentist test if, for each item, we achieve 95% **fconfidence** that the item's constraints hold for the relevant fraction of produced devices. Note that, to avoid confusion between the Bayesian and frequentist concepts of confidence, we refer to the frequentist concept as **fconfidence**. As justification for this unusual terminology, we note that the Bayesian concept of "confidence" corresponds much more closely to the intuitive concept of "confidence" than does fconfidence; we will expand on this point in section 5.1 and appendix A below. We also mention in passing that the fconfidence level of frequentist hypothesis tests is often misinterpreted. The fconfidence level is widely thought of as "how likely it is that the tested devices comply with the requirements", or "the degree of belief that one should have that the tested devices are compliant", but this is a incorrect interpretation. If one wants to know *that* probability then one needs to calculate the Bayesian posterior probability, which is what we define here as "confidence" (without the f).

We say that a device is **good** for an item if the true parameter in question is within the specification limits for the item. For example, if the force required to remove the cap when testing at standard atmosphere after after exposure to cold storage atmosphere is between 8 and 12 N then the device is good for that item.

We say that a factory is **good** for an item if the actual distribution of the parameter over devices satisfies the probability content requirement. For example, if the fraction of the devices produced by the factory whose cap removal force under normal use conditions lies between 8 and 12 N is 97.5% or more, then the factory is good for that item.

We say that a factory is **good** for for the test (or for all items), or that it is **compliant**, if for each item it is good for that item.

The meaning of **bad** in each case is that the factory or device is not good.

We define a **superlative** factory to be a factory such that:

1. For each item in the DFP constraining a probability to be at least some value p, all devices produced by the factory have that probability exactly equal to $1 - \frac{1-p}{10}$.
2. For each item in the DFP constraining a Poisson rate to be at most λ, all devices produced by the factory have that Poisson rate exactly equal to $\frac{\lambda}{10}$.
3. For each item in the DFP constraining 2-sidedly a continuously distributed parameter that is intrinsically positive, that parameter of devices produced by the factory is log-Gaussianly distributed with median equal to the geometric mean of the two constraints and standard deviation such that the fraction of the devices with that parameter outside the range of constraint is one tenth of the maximum fraction permitted.
4. For each item in the DFP constraining 1-sidedly a continuously distributed parameter that is intrinsically positive, that parameter of devices produced by the factory is log-Gaussianly distributed with median strictly inside the constraint and standard deviation such that the fraction of the devices with that parameter outside the range of constraint is one tenth of the maximum fraction permitted.

5. For each item in the DFP constraining a parameter not so far discussed, i.e. a continuous parameter that is not intrinsically positive, the distribution of that parameter of devices produced by the factory is as required above for the exponential of that parameter (which is intrinsically positive).

By a **perfect** factory we mean one such that the probability of passing the test is in practice indistinguishable from 1.

We will use the **symbol** $\propto$ in the following precise sense: We will write $P(A, B \mid C, D) \propto \text{expression}_1 \propto \text{epression}_2$ to mean that at both $\propto$ symbols there is a 'constant' of proportionality that does not vary with any of the variables to the left of the | sign (here A and B), but may vary with any of the symbols to the right of the | sign (here C and D).

We will describe an item as **insulted** if it is performed under such (adverse) environmental conditions that ISO 20072 prescribes 95% minimum probability content, and as **uninsulted** if 97.5% minimum probability content is prescribed.

3.2 Assumptions

In order to simplify the analysis we make the following assumptions, which are of varying levels of realism, but in practice necessary.

1. The devices do not wear out, indeed each device has the same probability distribution of behaviours when it is new as at any other stage of its life other than "end of life" (though these distributions may be different from device to device). Note that this does not exempt us from end-of-life testing according to ISO 20072, and at that point we do not assume that the distribution is as it was at the start.
2. In each case the parameter specified is constant for the device, but may vary from device to device; in some cases the parameter specified is a probability, in other cases it is a force, pressure, duration, etc.
3. The distributions of each of the specified parameters are mutually independent.
4. Where a continuous parameter that ranges over $(0, \infty)$ is considered, we will consider instead the logarithm of that parameter when assumptions of Gaussianity etc. are considered.

4 The ISO 20072 testing protocol

Once a manufacturer has established the DFP, ISO 20072 lists multiple differing environmental conditions under which each basic item must be tested. There are 14 such sets of conditions[2] listed, in each of which a hypothesis test must be passed for each of the 16 basic items of the DFP, giving $224 = 16 \times 14$ items to be tested in total.

In this section, we discuss the attempts made by ISO 20072 to prescribe a valid testing protocol.

4.1 The approach prescribed

Under ISO 20072, the compliance of each of the four types of parameter is tested with a frequentist hypothesis testing protocol. This section describes the testing protocols laid out in ISO 20072 for each type of parameter. ISO 20072 requires that all frequentist hypothesis tests are performed at a fconfidence level of 95%, i.e. that the probability of any specific non-compliant factory passing any given item is at most 5%.

All of the hypothesis testing protocols in ISO 20072 involve a "probability content", p, which is a proportion of the underlying population, either 95% or 97.5%. Each hypothesis test aims to determine whether a proportion p of the underlying distribution of the given parameter is within the given acceptable bounds.

4.1.1 Two-sided, real variable

For the case that a real continuous variable follows an underlying normal distribution, and the DFP specifies both upper and lower bounds $l < u$ for that variable, ISO 20072 gives a specific test, performed as follows.

Exactly one measurement is taken from each of N devices, and the sample mean, $\bar{x} = \frac{\sum_{n=1}^{N} x_n}{N}$, and sample standard deviation, $s = \sqrt{\frac{\sum_{n=1}^{N} (x_n - \bar{x})^2}{N-1}}$, are calculated. These two values are then used to calculate $\bar{x} - ks$ and $\bar{x} + ks$, where the value of k is obtained from table D1, in Annex D of ISO 20072, depending on the value of N and the probability content required. The two aforementioned values are subsequently compared with l and u, respectively, to determine whether the factory "passes" the test. The critical region is the event

$$(\bar{x} - ks > l) \cap (\bar{x} + ks < u)$$

[2]The "free fall" condition in ISO 20072 states that the test must be performed with the cap on and the cap off. We consider the EMC environmental condition to involve one test.

4.1.2 One-sided, real variable

The testing protocol in this case is very similar to the two-sided real variable. The only two differences are as follows: either $l = -\infty$ or $u = \infty$, since there is either only a lower or only an upper bound on our parameter; additionally, the values of k to be used differ from those to be used in the two-sided real parameter test.

These values of k are not included in ISO 20072. The reader is instead referred to a corresponding table in ISO 16269-6:2005.

4.1.3 Probability parameter

For this type of parameter, a testing protocol is not explicitly stipulated in ISO 20072. However, the standard requires every test to have a fconfidence level of 95%, i.e. states that the probability of any specific non-compliant factory passing must be ≤ 0.05. We suggest a frequentist test protocol in section 5.1.3 below.

4.1.4 Poisson rate

Similarly to the probability parameter, a testing protocol for the Poisson rates is not explicitly mentioned in ISO 20072. However, we suggest a frequentist testing protocol in section 5.1.4 below.

4.2 Erroneous values in ISO 20072 table of 2-sided values of k

ISO 20072 refers to ISO 16269-6:2005 for how to calculate the k values used in its two-sided testing of real continuous variables.

ISO 20072 invites the user to assume that the value of each continuous variable (e.g. the force needed, in Newtons, to remove the cap from a device) (or alternatively its logarithm) has a Gaussian distribution. Given this assumption, the testing protocol consists of obtaining sample values of the variable from N different devices, and evaluating the sample mean $\bar{x}$ and standard deviation, s. This is then followed by performing a frequentist hypothesis test at a fconfidence level of 95%. The null and alternative hypotheses are

$$\begin{aligned} H_0 : \int_l^u f(x)dx &\leq p \\ H_1 : \int_l^u f(x)dx &> p, \end{aligned}$$

where $f(x)$ is the population probability density function of the continuous variable, and l and u are

the pre-determined lower and upper bounds, respectively, within which we wish a pre-determined proportion p (e.g. 97.5%) of our underlying population to lie.

The factory passes the hypothesis test, i.e. H_0 is rejected, and we say with 95% fconfidence that the factory is suitable, if

$$\begin{aligned} \bar{x} - ks &> l \\ \bar{x} + ks &< u \end{aligned}$$

The values of k used to perform this test are "...determined based upon the confidence level (95%), probability content, p, and the number of measurements, n..." (see [1]). These k values, contained within Annex D of ISO 20072, are copied from the table which appears in ISO 16269-6.

However, the values within ISO 16269-6 have been misused. ISO 16269-6:2005 states that the values of k are to be used in order to determine an interval $(\bar{x} - ks, \bar{x} + ks)$ such that "...the probability that [the] interval constructed in the prescribed manner will contain at least a proportion p of the population [is 95%]".

In the hypothesis test, we wish to use a value of k such that the probability of falsely rejecting H_0 is less than 5%. These two definitions of k produce values which differ to the first decimal place for $n < 100$, and, even for $n = 1500$, differ at the second decimal place. The disparity between these values of k result in compliant factories being rejected, since the correct values of k are lower than the values provided in ISO 20072.

In ISO 16269-6:2005, it is stated that "The tolerance limits discussed... can be used to compare the natural capability of a process with one or two given specification limits, either an upper one U or a lower one L or both...". This statement is a misleading indicator that the k values contained within table E.4 of this standard are appropriate for use in a hypothesis test, which is wasteful of good factories.

The second edition of this standard, ISO 16269-6:2014, does not correct this statement sufficiently. ISO 16269-6:2014 suggests that "...the appropriateness of the given specification limits U and L can be compared with the actual properties of the process", which is again a wasteful protocol for performing a frequentist hypothesis test.

However, the table of k values pertaining to the case in which we know neither the mean, μ, nor the standard deviation, σ, of our underlying population is not included in the 2014 edition of ISO 16269-6. It is possible that, if ISO 20072 had been revised upon the publication of ISO 16269-6:2014, the requirement to calculate the values of k could have resulted in the values of k being corrected.

The corrected table of k values is provided in Appendix C.

4.3 A comment about two DFP basic items

Now that we have seen the testing protocol proposed by ISO 20072, we can comment about our choice of two of the basic items in the example DFP, which we consider in the uninsulted case where the desired probability content is 0.975 .

Firstly, let us compare DFP 2 with the following statement:

"The device shall switch on within 5 seconds of the cap being removed" (*)

If we suppose 95% of devices have the probability parameter equal to 1, and the remaining 5% of devices have probability parameter equal to 0.5, then the population of devices does not satisfy DFP item 2, but does satisfy the ISO 20072 interpretation of (*) because the mean is 0.975 (and no device is tested more than once). Indeed, any population satisfying our DFP 2 will also satisfy (*), since $0.975 \times 0.99 + 0.025 \times 0 > 0.95$. Therefore our choice of basic item 2 in the DFP is stronger than (*) in the uninsulted case.

Secondly, let us compare DFP 6 with the following statement, again in the uninsulted case:

"The device shall signal a correct inhalation as correct." (**)

If we consider a population with all devices having the probability parameter equal to 0.94, then the population satisfies our DFP item 6; however it does not satisfy the ISO 20072 interpretation of (**) because the mean is $0.94 < 0.975$. Therefore our choice of basic item 6 is not stronger than (**). However, if we consider a population with 90% of devices each with the probability parameter equal to 1, and another 10% of devices with probability parameter equal to 0.9, the mean will be $0.99 > 0.975$, hence it will satisfy the ISO 20072 interpretation of (**) despite not satisfying our DFP item 6. Therefore, our DFP item 6 is also not weaker than (**). Therefore, our DFP item 6 is neither weaker nor stronger than (**).

So we should ask "Why would we use DFP 2 rather than (*) ?", and "Why would we use DFP 6 rather than (**) ?". In the DFP 2 case, the answer is that it is important that a very high proportion of devices turn on with very high probability, and it is not acceptable to permit 5% of devices to be only turn on half the time (which would be permitted by (*)). In the DFP 6 case, (**) requires a mean probability of correct answer that may be unachievable within the available technological constraints, so needs to be set weaker. However, ISO 20072's interpretation makes it impossible to specify a DFP item that merely requires the mean probability to exceed 0.93; and if one did that, it would then be permissible for up to 14% of devices to give random correct or incorrect answers with probability 0.5, which would be unacceptable.

5 Frequentist testing

In this section, we propose a frequentist approach to doing testing that is in accordance with ISO 20072. We start by explaining the rationale that leads to this approach, taking into account the manufacturer's perspective.

5.1 Details of an approach to designing a frequentist test compliant with ISO 20072

In the previous sections we have described in outline the extent to which ISO 20072 prescribes how statistical testing is to be done. Before designing the remaining details of such a procedure, it is also necessary to examine the perspective of an ADDD manufacturer who will have to comply with those requirements.

ISO 20072 does not stipulate how many devices should be examined in each test of each required item; it simply requires that every item should have an fconfidence level of 95%, i.e. the probability that any specific non-compliant factory passes the hypothesis test must be $\leq 5\%$.

However, any factory owner who has made a good factory cares about another probability – the probability that their good factory might fail one of the 224 items. Failure of any one of the items would lead to a manufacturer being unable to produce any ADDDs for human use, resulting in a loss of the millions of dollars used to design and build an ADDD-producing facility.

Retaking a test is not an option for a manufacturer, on account of the confidence level required by ISO 20072. Consider, for example, a specific non-compliant factory which has a probability of 5% of passing one item given a single run of the relevant testing procedure, but which is allowed to take the test a second time if it fails the first. The probability that this item either passes the first time, or fails the first time and subsequently passes a second one, is 9.75%, which implies that the confidence level of the extended "double" test is now less than 95% (it is, in fact, equal to 90.25%).

We strongly believe that a manufacturer would prefer that a **good** factory should have at least 0.95 probability to pass the ISO 20072 protocol. In order to satisfy this, the frequentist approach must be designed such that a good factory passes each of the 224 items with probability substantially higher than 0.95: indeed, if the design is such that each item has the same probability of failing, then the probability of passing a single item needs to be at least 0.9998 (to 4 decimal places), so that the aforementioned good factory will pass the whole protocol with probability 0.9998^{224}, which is about 0.95.

While this is not required by ISO 20072, it is essential for a manufacturer of ADDDs. From their point of view, if their factory is entirely compliant with their DFP in the manner required by ISO 20072, they would wish their probability of passing any required testing protocols to be as close to 1 as possible.

Now, it is possible to design a testing plan giving a probability of 0.95 of passing the entire test in many different ways — provided that one knows the true parameters of the factory. For simplicity, in what follows, we will assume that the design is such that each item is assigned the same pass

probability of $0.95^{1/224}$, or approximately 0.9998, but it is likely that significant savings can be made on the overall number of tests by allocating the pass probabilities non-uniformly; for our present purposes, however, the workload of doing so optimally would be excessive. Furthermore, whether one designs for equal pass probability for each item or not, even the idea of doing so assumes that one knows the true parameters of the factory, which in practice one doesn't.

We will return to this topic later, but for now we note that for the following sample size calculations **we will assume that the factory is superlative** according to the definitions in section 3.1 above; intuitively then, we will be assuming that the factory is ten times better than it needs to be (in a carefully defined sense).

5.1.1 Continuous parameter with 2-sided constraints

We have seen in section 4.1.1 how to test each item involving a continuous parameter that is (log-)Gaussianly distributed with a two-sided constraint.

We will examine only DFP basic item 1, as Gaussian and log-Gaussian distributions can be linearly transformed to match the other basic items in the DFP that contain a two-sided constraint.

Therefore, our data x will come from a log-Gaussian distribution $\log N(\mu, \sigma)$, i.e. $\log x$ will be Gaussian with median μ and standard deviation σ. The lower and upper constraints on $\log x$ are respectively $\log(8)$ and $\log(12)$. The parameters, according to section 3.1 and knowing that the median is e^{μ}, are as follows:

1. For probability content $p = 0.95$, $\mu = 2.2822$ and $\sigma = 0.0722$.
2. For probability content $p = 0.975$, $\mu = 2.2822$ and $\sigma = 0.0671$.

Taking into account what we discussed in section 5.1, the minimum number of measurements **per item** needed when testing a superlative factory in order to have 0.9998 probability of passing each item is:

1. For probability content $p = 0.95$, at least 108 measurements are needed.
2. For probability content $p = 0.975$, at least 154 measurements are needed.

Therefore, as we have three DFP basic items with two-sided constraints and ISO 20072 mentions 8 environmental conditions for 0.95 content, and a further 6 conditions for 0.975 content, the total number of measurements needed for passing the 2-sided basic items will be $3 \times (108 \times 8 + 154 \times 6) = 5364$.

5.1.2 Continuous parameter with one-sided constraint

As we have seen in section 4.1.2, this is very similar to the previous subsection.

Again, we will examine only DFP basic item 3, as Gaussian and log-Gaussian distributions can be linearly transformed to match the other basic items in the DFP that contain a one-sided constraint and a superlative factory.

Therefore, our data x will come from a log-Gaussian distribution $\log N(\mu, \sigma)$ and the lower and upper constraints on $\log x$ are $-\infty$ and $\log(15)$ respectively. The parameters, according to section 3.1 and choosing an arbitrary value below the upper constraing for the median e^{μ}, are as follows:

1. For probability content $p = 0.95$, $\mu = 1.3540$ and $\sigma = 0.5256$.
2. For probability content $p = 0.975$, $\mu = 1.3540$ and $\sigma = 0.4823$.

Taking into account what we discussed in section 5.1, the minimum number of measurements **per item** needed when testing a superlative factory in order to have 0.9998 probability of passing each item is:

1. For probability content $p = 0.95$, at least 93 measurements are needed.
2. For probability content $p = 0.975$, at least 135 measurements are needed.

Therefore, as we have 5 DFP basic items with one-sided constraints and ISO 20072 mentions 8 environmental conditions for 0.95 content, and other 6 conditions for 0.975 content, the total number of measurements needed for passing the one-sided basic items will be $5 \times (93 \times 8 + 135 \times 6) = 7770$.

5.1.3 Probability parameter

On account of what we discussed in sections 5.1 and 4.1.3, we suggest the following frequentist testing protocol.

We iterate the following procedure over a suitably wide range of values of N and N_n, picking the version giving the minimum total number of measurements.

We consider testing N devices, collecting N_n 0/1 measurements from device n; we consider only the possibility that N_n is the same for all n, although clearly there are other possibilities. We need to determine a critical region, i.e. a subset of $\{0, 1, \ldots, N_n\}^N$ which gives a frequentist fconfidence level of 95%.

As usual there are an enormous number of possible critical regions; we consider only those of the form "At most $N_{\max}$ devices give more than $n_{\max}$ measurements that are 0", and consider what values of $N_{\max}$ and $n_{\max}$ result in this defining a critical region for 95% fconfidence.

The null and alternative hypotheses are as follows, where q is the lower constraint on the probability parameter, p is the required probability content and $F(x_0) = P(x \leq x_0)$ is the cumulative distribution function of the probability parameter x:

$$\begin{aligned} H_0 : \ F(q) &\geq 1-p \\ H_1 : \ F(q) &< 1-p, \end{aligned}$$

The probability that the data lies in a critical region of the form being considered is maximised over $h \in H_0$ by the "critical distribution" given by

$$F(x) = \begin{cases} 0 & \text{if } x < q, \\ 1-p & \text{if } q \leq x < 1, \\ 1 & \text{if } x = 1. \end{cases}$$

and we therefore can determine, for each set of values of $N, N_n, N_{\max}$, and $n_{\max}$, whether or not the probability that data generated from the critical distribution lies in the critical region is less than 0.05 as required. Having picked the set of values of $N, N_n, N_{\max}$, and $n_{\max}$ minimising $N \times N_n$ over those satisfying this requirement on the critical region, we can then check whether if data is instead generated from the a superlative factory we get a sufficiently *high* probability (at least 0.9998) that *that* data lies in the critical region. If not, we discard that set of parameters and try the next best. Eventually we come to the set of parameters minimising the total number of measurements among those that are permissible.

We then measure each of the N devices N_n times. The hypothesis test for this item is then passed, i.e. the null hypothesis H_0 is rejected and we say, with 95% fconfidence, that the factory is good for this item, if the data produced lies within the critical region.

When testing a superlative factory (see section 3.1), we have found the following optimal minimal numbers of measurements and hypotheses needed in order to achieve what we discussed in section 5.

For items that with a lower constraint on the probability parameter of 0.99, we will say that a device "passes" if we observe at most 8 fails per device and the critical region is "every device passes", i.e. $N_{\max} = 0, n_{\max} = 8$. The following numbers are **per item**:

1. For content $p = 0.95$, it is optimal to test $N = 71$ devices 1171 times each, giving 83141 measurements.
2. For content $p = 0.975$, it is optimal to test $N = 141$ devices 1192 times each, giving 168072 measurements.

Therefore, as we have 2 DFP basic items with target 0.99 and ISO 20072 mentions 8 environmental conditions for 0.95 content, and another 6 conditions for 0.975 content, the total number of measurements needed for passing the target 0.99 basic items will be $2 \times (83141 \times 8 + 168072 \times 6) = 3347120$. With British understatement, we note that this number is rather large, particularly in a context where manufacturers are hoping to use no more than a couple of hundred devices to test and to test each only a handful of times. We note also that we have paid no attention to whether a device would in practice survive being tested over one thousand times.

For items with a lower constraint on the probability parameter of 0.93, we will say that a device "passes" if we observe at most 9 fails per device and the critical region is "every device passes", i.e. $N_{\max} = 0, n_{\max} = 9$. The following numbers are **per item**:

1. For content $p = 0.95$, it is optimal to test $N = 68$ devices 190 times each, giving 12920 measurements.
2. For content $p = 0.975$, it is optimal to test $N = 142$ devices 184 times each, giving 26128 measurements.

Therefore, as we have 4 DFP basic items with target 0.93 and ISO 20072 mentions 8 environmental conditions for 0.95 content, and other 6 conditions for 0.975 content, the total number of measurements needed for passing the target 0.93 basic items will be $4 \times (12920 \times 8 + 26128 \times 6) = 1040512$. Note that, while lower that the previous one, this number is still rather large.

5.1.4 Poisson rate

As in the previous case, we search for parameters N the number of devices to be tested, T_n the length of time for which each the response of device n is to be recorded, $N_{\max}$ and $n_{\max}$ specifying the critical region of the form "Not more than $N_{\max}$ devices gave more than $n_{\max}$ events while being observed". Again, we consider only the case where all T_n are the same, and we search for the minimum number of total hours of observation $N \times T_n$ compatible with a critical region giving a fconfidence level of 95% and a sufficiently high pass probability for a superlative factory.

The null and alternative hypotheses are as follows, where λ is the upper constraint on the Poisson rate, p is the probability content and $F(x_0) = P(x \leq x_0)$ is the cumulative distribution function of the Poisson rate x:

$$\begin{aligned} H_0 : \ & F(\lambda-) \ \leq \ p \\ H_1 : \ & F(\lambda-) \ > \ p. \end{aligned}$$

The probability that the data lies in a critical region of the form being considered is maximised over $h \in H_0$ by the "critical distribution" given by

$$F(x) = \begin{cases} p & \text{if } x < \lambda, \\ 1 & \text{if } \lambda \leq . \end{cases}$$

and we therefore can determine, for each set of values of $N, T_n, N_{\max}, \text{and } n_{\max}$, whether or not the probability that data generated from the critical distribution lies in the critical region is less than 0.05 as required. Having picked the set of values of $N, T_n, N_{\max}, \text{and } n_{\max}$ minimising $N \times T_n$ over those satisfying this requirement on the critical region, we can then check whether if data is instead

generated from the a superlative factory we get a sufficiently *high* probability (at least 0.9998) that *that* data lies in the critical region. If not, we discard that set of parameters and try the next best. Eventually we come to the set of parameters minimising the total number of measurements among those that are permissible.

We then observe each of the N devices for a time T_n. The hypothesis test for this item is then passed, i.e. the null hypothesis H_0 is rejected and we say, with 95% fconfidence, that the factory is compliant in the environmental condition in which the item was tested, if the data produced lies within the critical region.

When testing a superlative factory (see section 3.1), we have found the following optimal minimal device-hours of observation time needed in order to achieve the goals of section 5.

1. For content $p = 0.95$, each device "passes" if we register no more than 9 false inhalations during a testing time of 11.1 hours. The optimal testing time we found is 11.1 hours per device, and a total of 76 devices. This gives a total of 843 device-hours of observation.
2. For content $p = 0.975$. each device "passes" if we register no more than 8 false inhalations during a testing time of 9.9 hours. The optimal testing time we found is 9.9 hours per device, and a total of 182 devices. This gives a total of 1801 device-hours of observation.

Therefore, as we have 2 DFP basic items with upper constraint 1/hour and ISO 20072 mentions 8 environmental conditions for 0.95 content, and another 6 conditions for 0.975 content, the total number of device-hours of observation needed for passing the Poisson basic items will be $2 \times (843 \times 8 + 1801 \times 6) = 35100$ device-hours.

5.2 Summary of effort required for frequentist test

We now have all the necessary information to assess how much testing needs to be done, in order to have a probability at least 95% that a **superlative** factory passes the whole test, that is passes all items. Note that a superlative factory is an almost perfect one, and in reality such a factory may be almost impossible to design. Therefore, in reality, all the numbers are likely to be even higher.

Summing the minimal number of measurements needed for each category of items, we get that at least $3347120 + 1040512 + 7770 + 5364 = 4400766$ measurements are required plus 35100 device-hours of observation. If we reckon that a robot could perform one measurement per minute, this gives a total of almost 12.5 years of robot time. This makes it an incredibly difficult task for the reality we live in. But this is just the tip of the iceberg - in reality, the only frequentist way that we could know that our factory was superlative is by doing an amount of testing that is enormously larger still. This **cannot** be the right way of doing things !

Sadly, if we are determined to stay with frequentist rather than Bayesian methods, this naturally points the discussion to the topic of cheating and manipulating data, to which frequentist testing provides exquisitely acute temptation.

5.3 Cheating

In our opinion, the enormous amount of testing needed in order to **honestly** pass such a test naturally leads to the phenomenon of cheating and data manipulation.

The first and most serious consequence of this is, of course, that the safety of patients is in danger.

People with experience in the industry (who understandably wish to remain anonymous) tell us there there are several ways one can cheat while doing frequentist tests, some of which are pretty obvious, but nonetheless hard to check and control:

1. There is no assurance that the critical region was picked before rather than after collecting the data. In some instances, it is done **after** doing the measurements, which completely invalidates the results, although this is becoming more difficult now that the test plan is required to be recorded in advance.
2. Censoring of data: when the observed data does not fall into the critical region, hence the frequentist approach cannot infer anything, the temptation is to **discard** some or all of the unfavourable measurements and collect new measurements to take their place. In an extreme case, of course, one can discard every bad measurement and get 100% "correct" results. There are a battery of excuses available for discarding such data:

   (a) "I haven't actually started yet"
   (b) "I didn't do that one properly"
   (c) "I wasn't wearing goggles, so I need to do it again"
   (d) "That one was faulty"
   (e) "Operator error"

   Our suggestion is that if an independent observer, given other data on *all* the measurements, but deprived of their results, is able to determine precisely the set of measurements to be discarded, and is in agreement with the operators, then that is reasonable evidence that those measurements may be discarded. In reality, often, the decision to discard is considered only for those measurements which have given an undesirable result.
3. The "Concession Note": here it is admitted that some of the items did in fact fail, but only in small print and buried in a large report, while the executive summary states that "All the tests passed as described in detail in the body of this document". Such documents may even succeed in obtaining regulatory approval; we can never tell whether this is because the regulator hasn't noticed the concession note, or because they realise that it is not realistic to hope that the ISO 20072 test will truly be passed.

Because of the unrealistic number of measurements needed to pass a frequentist test of a realistic factory **honestly**, taking only a few measurements results in obvious cheating that is in principle easy to detect as follows:

1. Calculate a level of performance that although in H_1 gives a very low probability of passing the test.

2. Collect sufficient new data to show that the performance is indeed that bad.

In other words, if a factory passed such a frequentist test and the manufacturer claims that the reason is a good design, then there are a number of possibilities:

1. Their factory is not just superlative, but actually close to **perfect**. (We may believe that such a perfect factory is yet to be designed, and if so we note that proving that a factory is not perfect is far easier than proving it is perfect.)
2. Their DFP contains very few basic items. Noting that even one basic item gives rise to 14 items, getting a small number of items while still providing adequate control of the parameters of the ADDD is difficult.
3. Their DFP constraints are very slack. This is likely to mean that their ADDDs may not be safe.
4. They have cheated in one of the many possible ways.

However, there is also the danger of cheating unintentionally. In the following subsection we will discuss an anecdote brought to our attention by someone in the industry, where the intention of being honest resulted in a false claim to have passed a test.

5.3.1 An example of an ignorant attempt to be honest

Say that we want 95% fconfidence that 97.5% of the devices do not break when dropped. To do this, drop 119 devices. If none break, pass the item. If any break, fail the item.

Then the hypothesis space is $H = [0, 1]$, the null hypothesis is $H_0 = [0, 0.975]$ and $H_1 = (0.0975, 1]$. Then the data space of how many devices broke is $X = \{0, 1, \cdots 119\}$ and the critical region is $C = \{0\}$.

Consider $h_0 \in H_0$.

Then $P(x \in C \mid h_0) = h_0^{119}$ which achieves its upper bound of 0.975^{119}, which is about 0.05, hence the fconfindence is about 95% – if the data falls in C.

However, without planning to do this in advance, what was actually done was that when a device did break (and some did), it was replaced by 3 others and the data on the broken device discarded. While the intention was to be honest (the operator knew that just discarding the broken device and replacing its data with one new device would be dishonest), this procedure **drastically** reduces the fconfidence as shown in the following calculation.

Denote the statement "on 1 drop either 0 break or 1 breaks then 3 more do not" by (E). Then $P(E \mid h_0) = h_0 + (1 - h_0)h_0^3$ which achieves its upper bound of $0.975 + 0.025 \times 0.975^3$, which is about 0.9982. This happens 119 times with probability 0.9982^{119}, which is about 0.8043. Therefore the fconfidence is now at most 19.6%, when it was supposed to be 95%.

Since this "replace by three" procedure was not planned in advance, the test is invalid anyhow. To see that the situation is actually even worse than it appears from the above calculation, one must ask oneself "If one of the three replacements had also broken, would it have been replaced by a further set of three, and so on iteratively as needed ?". If the answer to this is Yes, then a further calculation shows that although the item will now pass with probability 1 (so long as $h_0 > 0$), the fconfidence then achieved would be zero. Thus we see that:

1. The fconfidence resulting depends on "what would have been done if...", which is why any test whose procedure and critical region have not been fixed in advance cannot be regarded as valid.
2. Those who are intending to be honest, but whose mathematics is not up to the job, are likely to falsify the results due to the extreme temptation posed by the conflict between the commercial pressure for the test to pass and the impossibility of getting it to pass honestly.

5.4 The frequentist results are in fact irrelevant

Apart from the huge numbers of devices required and the temptation to cheat, the frequentist approach has some other major flaws that makes its results in our view irrelevant.

5.4.1 Fconfidence misbehaves when applied to combinations of items

To see this, consider a simplified DFP consisting of just one basic item of the form "When dropped the device must not break", to be tested under only two environmental conditions (impossible under ISO 20072, of course). To make the diagrams drawable, suppose also that the probability content required is only 0.9, and that we require only 65% fconfidence of that. We then make the following observations.

1. If a factory passes the test, we have fconfidence 65% that the factory is good on each item taken one at a time. Any reasonable intuitive understanding of "confidence" would require that our confidence that the factory is simultaneously good on both items be not 65% but $0.65^2 = 0.4225$ or 42.25%. This is illustrated in Figure 1.
2. However, if we consider the combined simultaneous frequentist test of both items, with the alternate hypothesis being "the factory is simultaneously good on both items" and the null hypothesis therefore being "the factory is not good on at least one item", then the null hypothesis is the shaded area, i.e. everything apart from the top right small square of Figure 2. But now $P(\mathbf{x} \in C \mid h_0)$ is maximized over $h_0 \in H_0$ not by taking $h_0 = (0.9, 0.9)$, but by taking h_0 to be either of the points (0.9, 1) or (1, 0.9) (indicated by the red stars), which makes that probability $0.9^{10} \times 1^{10} = 0.35$ approximately, so again we have 65% fconfidence, not 42.25% fconfidence as we would expect from 1 above. Thus to think of fconfidence as similar to the intuitive concept of "confidence" is misleading.
3. Similarly, doing a test according to ISO 20072 with 224 items and the null hypothesis "the factory is bad on at least one item", given that the factory passes all items, the fconfidence that the factory is simultaneously good for all items is still 95%, where any reasonable person

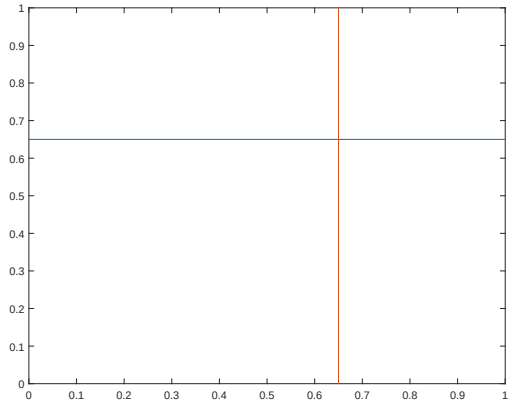

Figure 1: How the intuitive idea of "confidence" behaves when two independent items are combined. If we are 65% confident that a point chosen yesterday from the unit square uniform randomly is to the left of the red line, and 65% confident that it is below the blue line, then the confidence that it is both to the left of the red line and below the blue line should be not 65% but 42.25%.

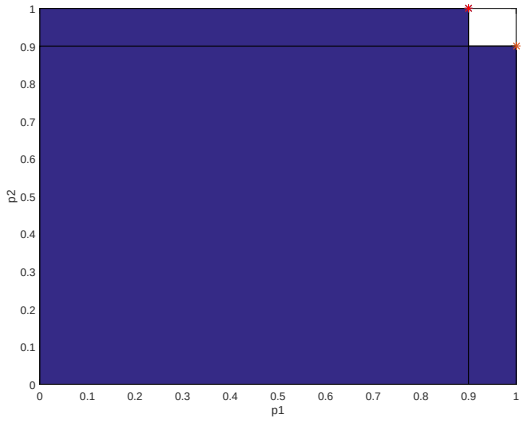

Figure 2: Illustration of the hypothesis space for the simplified 2-item test discussed in subsection 5.4. The null hypothesis H_0 is the shaded area, i.e. everything except for the small square at the top right. The two values of h_0 maximizing $P(\mathbf{x} \in C|h_0)$ are indicated by the red stars.

would expect it to be 0.95^{224}, which is approximately zero. Thus frequentist hypothesis tests seem unable to distinguish being, for each of N independent items, fconfident that the factory is good for that item, and being fconfident that the factory is simultaneously good for all N items.

4. Given a factory design that intends to make the factory conform to the limits in the DFP as interpreted by the probability content specifications in ISO 20072, the frequentist approach is overwhelmingly likely to reject such a factory. It will do so not because the factory is bad, but because of "bad luck", i.e. the data not falling in the critical region. Indeed, rather than calling this "bad luck", we would consider it to be "bad statistics".

5.4.2 Frequentist testing considers the wrong denominator

When we use a phrase such as "We are 90% confident that the factory satisfies property Z" we are implicitly considering some population L of factories and saying that 90% of that population satisfies Z. We need to consider what population is relevant for us to be thinking of.

Once we have collected our data, we would argue that the population L of interest is the subpopulation of all possible factories that would have yielded that data - as that is the subpopulation to which we have managed to determine our factory belongs.

A frequentist test, however, instead considers the subpopulation of factories that has one of the most difficult to detect non-compliances and is perfect on all other items. Even if a manufacturer wanted to produce such a factory, it would be impossibly difficult. No reasonable person would believe it remotely likely that the factory in question was such a factory. However, any reasonable person *would* believe that the factory we are dealing with is one that has produced the data we have observed - indeed this is a tautology.

This, above all other reasons, is why we believe that a frequentist test is inappropriate for this purpose. However, for yet another way in which fconfidence does not correspond to the intuitive understanding of confidence, see theorem A.2 in the appendix.

6 Bayesian testing

When designing a better testing protocol, the issues to be dealt with are:

1. The current protocol depends on an assumption of Gaussianity, which is very likely to be inappropriate. The protocol's statistical assertions are incorrect for non-Gaussian distributions. This assumption needs to be replaced.
2. The current protocol has a high chance of rejecting factories that are actually compliant. A satisfactory protocol must have at most a small probability of rejecting factories that are compliant. Of course, it must also be the case that, after testing, for each item, the probability that a passed factory is good for that item is high. One might also hope that the probability that a passed factory is good for all items is high, but this is not realistic; achieving e.g. 0.95 probability that a factory is good for all items would require achieving, for each item, a 0.9998 probability that the factory was good for that item, which would take a truly immense amount of testing.
3. The current protocol employs sample sizes that are fixed in advance, as we have seen in section 5, and for the protocol to have a good chance of accepting a superlative factory (let alone an arbitrary compliant factory), those sample sizes have to be very large. To reduce the expected cost of the testing process, it would be desirable to use sample sizes that are not fixed in advance, that to be able to stop the test when the evidence for one hypothesis or another is sufficiently strong.
4. The current protocol is designed around knowing that the factory is superlative (and any protocol on the same lines is designed around some assumption on the quality of the factory that is much more restrictive than that the factory is good for all items). It is desirable that the protocol is completely automatically adaptive, and that the design does not depend on any stronger assumption than that any factory that is good for all items should have a high probability of passing.

Now, we note that there exist methods of frequentist testing that design critical regions spanning a range of different sample sizes, and that *if* the precise properties of the factory under test are known, then it is possible to somewhat reduce the required sample size, at the expense of reducing still further the probability that a factory will pass the test that has different but still compliant true properties. However, we have chosen not to explore this route, as the fundamental issue is that true properties of the factory are *not* known before it is tested.

6.1 Sequential Bayesian testing

We propose a Bayesian iterative method, since Bayesian inference makes full use of all the data, given the assumptions; sequential hypothesis testing is particularly straightforward when a Bayesian approach is taken. We will first describe the general method, then explain in section 6.1.2 why it is valid — as this situation is so unlike that in a frequentist analysis, where repeated analysis and decision to stop when that analysis is favourable in general biases the results.

For both description and validity we will take a simpler example, that of determining whether the probability p that a biased coin lands on heads is at least p_0. The same principles apply to the ISO 20072 case.

6.1.1 Description of the method

If a fixed number N tosses are made, the number r of heads is expected, given p, to be binomially distributed with mean Np. The likelihood function is:

$$P(r \mid N, p) = \binom{N}{r} p^r (1-p)^{N-r} \tag{1}$$

In a sequential approach, the number of tosses N is not fixed, but, whatever the stopping rule, the likelihood function's dependence on p still has the same form as the binomial distribution (1) given N.

We will determine in advance a prior probability distribution on p. For simplicity, we will assume that this distribution has a density function $P(p)$.

Let $\mathbf{s}$ be the actual sequence of head/tail outcomes $s_1, s_2, \ldots s_N$, with $s_n = 1$ denoting heads and $s_n = 0$ denoting tails. Let S be the number of tosses we make before we decide to stop collecting data, and let S_N denote the event that S takes the value N.

Whatever the stopping rule, the likelihood function given N is:

$$\begin{aligned} P(\mathbf{s} \,|\, p, N) &= \left(p^{s_1}(1-p)^{1-s_1}\right)\left(p^{s_2}(1-p)^{1-s_2}\right)\ldots\left(p^{s_S}(1-p)^{1-s_S}\right) \\ &= p^r(1-p)^{(N-r)} \end{aligned}$$

Suppose we decide to analyse the data after N tosses, where N is chosen independently of the data and of the unknown p, but that we have not necessarily decided to stop collecting data at that point. For example, we might decide to analyse the data after every new toss. Then the posterior probability density of the probability p is:

$$P(p \,|\, \mathbf{s}, N) = \frac{P(\mathbf{s} \,|\, p, N) P(p \mid N)}{P(\mathbf{s} \mid N)} = \frac{P(\mathbf{s} \,|\, p, N) P(p)}{P(\mathbf{s} \mid N)}$$

where $P(\mathbf{s} \mid N)$ is the normalising constant

$$P(\mathbf{s} \mid N) = \int_0^1 P(\mathbf{s} \mid p, N) P(p)\, dp.$$

However, if our stopping rule has also decided to stop at this point, we instead need

$$P(p \,|\, \mathbf{s}, S_N, N) = \frac{P(\mathbf{s}, S_N \,|\, p, N) P(p)}{P(\mathbf{s}, S_N \mid N)},$$

where $P(\mathbf{s}, S_N|N)$ is the normalising constant

$$P(\mathbf{s}, S_N \mid N) = \int_0^1 P(\mathbf{s}, S_N \mid p, N) P(p)\, dp.$$

Note, however, that we have seen an equation for $P(\mathbf{s} \mid p, N)$ but that we need to use $P(\mathbf{s}, S_N \mid p, N)$. If $S = N$ is fixed in advance, these two quantities are the same, but otherwise we will need to consider how these two quantities are related.

Let us next define the proposed sequential protocol, that is, the rule for deciding whether to stop or continuing testing.

At arbitrary intervals during data collection (for example after each new toss or measurement), we compute two posterior probabilities, based only on the sequence of tosses we have seen so far:

$$a_N = \int_{p_0}^1 P(p \,|\, \mathbf{s})\, dp$$

which is the probability, given the data gathered thus far, that the probability p exceeds the required probability content p_0; and

$$b_N = \int_0^{p_0} P(p \,|\, \mathbf{s})\, dp$$

which is the probability that p is smaller than p_0.

We stop testing and declare the item to have passed with (Bayesian) confidence a_N if and when $a_N > \alpha$, and it stops testing and declares the item to have failed with confidence b_N, if $b_N > \beta$, where α and β are protocol parameters. Thus S_N occurs if and only if N is the smallest integer such that $a_N > \alpha$ or $b_N > \beta$.

If neither $a_N > \alpha$ nor $b_N > \beta$ then the protocol continues testing.

The protocol parameter α controls the required confidence level to pass, and in this application will always be 0.95 as prescribed by ISO 20072. If an item passes, then the probability that the factory is bad for that item is $1 - a$, which is less than $1 - \alpha$.

On the other hand, β specifies how sure the manufacturer has to be that the factory is bad before he gives up on it. It is usually catastrophic financially to give up on a factory and reject it, and the manufacturer is likely to set β very close to 1. In the following we will assume that it is set at exactly 1; in other words, that the manufacturer will continue testing until such time as the factory passes, or otherwise for ever.

6.1.2 Validity of the Bayesian sequential method

We now consider why the Bayesian sequential method is valid. This essentially involves a consideration of the relationship between $P(\mathbf{s} \mid p, N)$ and $P(\mathbf{s}, S_N \mid p, N)$, and hence that between $P(p \mid \mathbf{s}, S_N, N)$ and $P(p \mid \mathbf{s}, N)$.

But these are actually a very simple relationships. Consider

$$P(\mathbf{s}, S_N \mid p, N) = P(\mathbf{s} \mid p, N) P(S_N | \mathbf{s}, p, N)$$

by the chain rule of probability. Now, $P(S_N|\mathbf{s}, p, N)$ is either 0 or 1 depending on whether $(a_N > \alpha)$ or $b_N > \beta)$ or not — and the formula for a_N, b_N do not involve p (except as a dummy variable of integration). They do, of course, involve $\mathbf{s}$. Thus $P(S_N|\mathbf{s}, p, N) = P(S_N|\mathbf{s}, N)$, and we have

$$P(\mathbf{s}, S_N \mid p, N) = P(\mathbf{s} \mid p, N) P(S_N | \mathbf{s}, N),$$

and hence

$$\begin{aligned} P(p \,|\, \mathbf{s}, S_N, N) &= \frac{P(\mathbf{s}, S_N \,|\, p, N) P(p)}{\int P(\mathbf{s}, S_N \,|\, p, N) P(p)\, dp} \\ &= \frac{P(S_N \mid \mathbf{s}, N) P(\mathbf{s} \,|\, p, N) P(p)}{\int P(S_N \mid \mathbf{s}, N) P(\mathbf{s} \,|\, p, N) P(p)\, dp} \\ &= \frac{P(\mathbf{s} \,|\, p, N) P(p)}{\int P(\mathbf{s} \,|\, p, N) P(p)\, dp} \\ &= P(p \mid \mathbf{s}, N), \end{aligned}$$

so that the decision to stop, made in this way, does not affect the inference we draw about whether or not $p > p_0$.

Thus as we have seen, the key point is that the stopping rule is conditionally independent of the unknown parameters given the data.

In exactly the same way, this sequential testing and analysis protocol does not influence in any way the inference about whether an item is good or not — in total contrast to the effects of either repeated analysis or repeated data collection under the frequentist paradigm.

It is, however, not impossible to make a stopping rule which is *not* conditionally independent of the unknown parameters given the data, and with such a stopping rule a Bayesian analysis would be invalid without modification. Two examples of such a stopping rule are:

1. Suppose we take a peek at each new data item before including it in our data, and decide to stop if the next data item doesn't look good, but *without* including it in the data considered.

Correct calculation of the effect of such a stopping rule would unsurprisingly show that the inference resulting would be the same as if the data item had been included. Thus data censoring (which this would be a form of) is illegal under the Bayesian paradigm just as much as under the frequentist paradigm.

2. Suppose we hear that our rival up the road has stopped their data collection on exactly the same subject, and decide to stop ours as a result and rush to publication. Such an event is clearly dependent on our rival's data, and hence on the unknown parameters, but not via our own data, and stopping on this basis would therefore invalidate our analysis. (In this situation it should be obvious that the right thing to do is likely to be to cooperate with our rival and merge data sets.)

6.1.3 Further useful properties of the sequential Bayesian approach

In addition to giving us the ability to repeatedly analyse the data and automatically assess whether we have yet achieved the desired confidence level, such Bayesian testing also yields an additional advantage: as shown in appendix B, the probability of passing a good factory for a finite number of items now approaches 1 as the number of measurements approaches infinity (at least, so long as $\beta = 1$). Of course, that is no guarantee that the necessary number of measurements will be feasible - but it turns out that it is vastly smaller than the corresponding number of measurements needed under the frequentist paradigm.

Thus the sequential Bayesian approach solves issues 2, 3, and 4 of the list at the start of section 6. As we will see shortly, it is also able to address issue 1.

6.2 Mixture models

Before actually reporting the number of measurements needed using this method, we need to address some issues connected with issue 1 of the list at the start of section 6.

As motivation, we note that most ADDDs are made of plastic, and that in a factory it is typical to use not just one mould for any particular plastic part, but several, in order to increase the rate of production. In consequence, the population of ADDDs produced is often a mixture of several closely related but non-identical populations. This situation can be modelled using mixtures of distributions, as follows.

We will denote the number of data points by N, the data points themselves by x_n for $n = 1, 2 \ldots N$, and the (unknown) number of mixture components by K (we will also call them "clusters"). Each such component has its own set of parameters θ_k; depending on which item is being considered, θ_k may be a vector of several scalar parameters.

We denote by $\mathbf{q}$ the weights of the components, where q_k is the weight of component k. Hence $q_k \geq 0$ for $k = 1, 2 \ldots K$ and $\sum_{k=1}^{K} q_k = 1$.

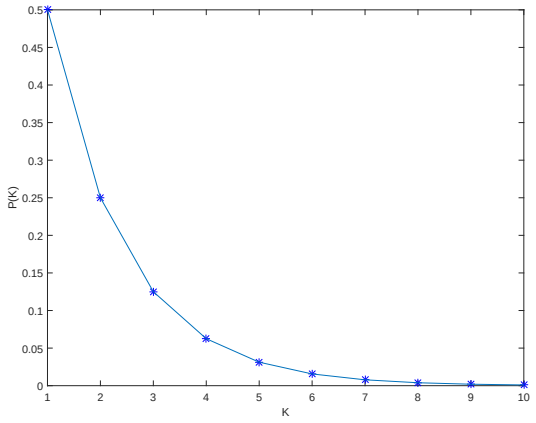

Figure 3: Probability mass function for K

Therefore $P(x_n) = \sum_{k=1}^{K} q_k P(x_n \mid \theta_k)$.

We also denote by k_n the mixture component from which x_n is taken, for $n = 1, 2 \dots N$, and by $\mathbf{k}$ the vector of these indices.

Thus $P(x_n \mid k_n) = P(x_n \mid \theta_{k_n})$.

We will denote by $\mathbf{N} = (n_1, \dots, n_K)$ the number of data points coming from each cluster, i.e. $N_k = |\{n : k_n = k\}|$.

6.3 Priors on K and q

We choose as prior for the number of clusters K the geometric distribution, i.e. $P(K) = (1-\lambda)\lambda^{K-1}$ with $\lambda = 0.5$. A plot of the probability mass function can be found in figure 3.

Because the distribution $P(\mathbf{N} \mid N, \mathbf{q})$ is the Multinomial, we will use its conjugate prior, that is the Dirichlet Distribution, for $\mathbf{q}$, with support on the K-simplex.

Therefore, $P(\mathbf{q} \mid \boldsymbol{\alpha}, K) = \frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\sqrt{K} \prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} q_k^{\alpha_k - 1}$. For simplicity, we choose to make $\alpha_k = \frac{1}{K}$ for all $k = 1, 2 \dots K$.

6.4 Markov Chain and Gibbs Sampling

Using the data and the priors, we will perform Markov Chain Monte Carlo to assess whether the population of devices tested is compliant with the restriction of a particular item. In this section, we will discuss the Gibbs sampling rules of the mixture model variables and how to sample the

next element of the Markov Chain.

For general information on Markov Chain Monte Carlo and Gibbs Sampling, we recommend the review by Neal in [3] (the Gibbs Sampling method is discussed in section 4.4).

We do Gibbs sampling as follows. Starting from:

$$P(\mathbf{x}, \mathbf{k}, \mathbf{q}, K, \boldsymbol{\theta}) = P(K)P(\mathbf{q} \mid K, \boldsymbol{\alpha})P(\mathbf{k} \mid \mathbf{q}, K, N)P(\mathbf{x} \mid \boldsymbol{\theta}, \mathbf{k})P(\boldsymbol{\theta})$$

1. First sample K. Before this, we need to rearrange the **empty** clusters to be at the end of the list of clusters. Then the probability mass function of K will be $P(K \mid \mathbf{x}, \boldsymbol{\theta}, \mathbf{k}) \propto \frac{P(K)\Gamma(\sum_{k=1}^{K} \alpha_k)}{\sqrt{K} \prod_{k=1}^{K} \alpha_k} \frac{\sqrt{K} \prod_{k=1}^{K} (\alpha_k + N_k)}{\Gamma(\sum_{k=1}^{K} (\alpha_k + N_k))}$ if K is greater than or equal to the old number of nonempty clusters, and $P(K) = 0$ otherwise.

   Then, we have three possible situations:

   (a) The new sampled K is the same as the old one. In this case nothing special needs to be done.

   (b) The new sampled K is less than the old K. In this case, we randomly select from the empty clusters which ones to discard.

   (c) The new sampled K is greater than the old K. In this case, we form extra clusters by taking their corresponding θ parameters from the priors on θ, which depend on the item.

2. Sample $\mathbf{q}$ from $P(\mathbf{q} \mid \mathbf{x}, \mathbf{k}, K, \boldsymbol{\theta}) \propto P(\mathbf{q} \mid K, \boldsymbol{\alpha})P(\mathbf{k} \mid \mathbf{q}, K, N)$. This is again a Dirichlet distribution with new parameters $a_k = \alpha_k + N_k$ for $k = 1, 2 \ldots K$.

3. Sample $\mathbf{k}$ from $P(\mathbf{k} \mid \mathbf{x}, \mathbf{q}, K, \boldsymbol{\theta}) \propto P(\mathbf{k} \mid \mathbf{q}, K, N)P(\mathbf{x} \mid \mathbf{k}, \boldsymbol{\theta}) = \prod_{n=1}^{N} q_{k_n} P(x_n \mid \theta_{k_n})$. Evaluate this for each possible value of k_n, then draw from the discrete distribution, for $n = 1, 2 \ldots N$.

4. Sample $\boldsymbol{\theta}$ from $P(\boldsymbol{\theta} \mid \mathbf{x}, \mathbf{k}, \mathbf{q}, K) \propto P(\boldsymbol{\theta})P(\mathbf{x} \mid \mathbf{k}, \boldsymbol{\theta}) = \prod_{k=1}^{K} P(\theta_k) \prod_{n=1}^{N} P(x_n \mid \theta_{k_n})$.

5. Then to ensure that detailed balance holds for the combination of moves, we make the sequence palindromic, by repeating all but the last step in reverse order.

6.5 Sampling and code

In the previous subsection, we have seen that our code samples from several types of distributions:

1. From a discrete (possibly unnormalised) distribution: our code normalises the probability mass function then uses an uniform random variable u in the classical way, i.e. if $u < p_1$, the sample is 1, if $p_1 + \ldots + p_{k-1} \leq u < p_1 + \ldots + p_k$, the sample is k etc.

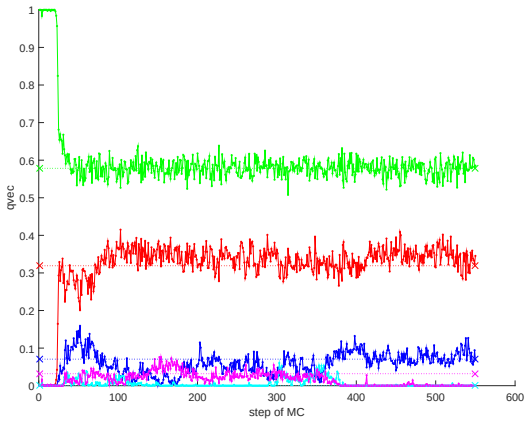

Figure 4: Behaviour of code when data is sampled from a true distribution with 5 mixture components (the dotted lines). Note that the magenta and cyan components are almost negligible, and that there is insufficient data to determine for definite whether some of the data points belong to the magenta, cyan, blue, or red clusters.

2. From a Dirichlet distribution, say with parameters $\alpha_1, \alpha_2, \ldots, \alpha_K$ we want to sample $\mathbf{q}$. Then sample r_k from the Gamma distribution Gamma(α_k, 1) using the standard Matlab function gamrnd() and set $q_k = \frac{r_k}{\sum_{j=1}^{K} r_j}$ for $k = 1, 2 \ldots K$.

3. From a continuous distribution on $(0, +\infty)$: here our code transforms the variables from the interval $(0, +\infty)$ to $(0, 1)$ using the map $x \to \frac{x}{x+1}$, then performs "Slice Sampling" according to Neal in [4], before transforming the sample back using $x \to \frac{x}{1-x}$. Similarly for other ranges of support using different transformations.

When monitoring the performance of the MCMC algorithm on synthetic data, there is an identification problem on the clusters; there is no hope of the algorithm determining which cluster it has found corresponds to which cluster number in the synthesis (and of course it may currently be considering more or fewer clusters than the true number). In order to ensure that the true and discovered clusters were colored with (as far as possible) the correct colors, the Hungarian algorithm [5] was used to associate true cluster numbers with clusters in a fashion maximizing the number of correct assignments of data points to true cluster numbers. A description of the Hungarian (or Munkres) Algorithm can be found in Harold Kuhn's article in [6].

A plot of the sample trajectory of $\mathbf{q}$ from the Markov Chain is shown in figure 4. Convergence was established by constructing synthetic data with a variety of true parameter values and ensuring that the duration of sampling run was sufficient that all these had converged to distributions covering the correct values.

6.6 Sample sizes needed for each type of DFP item

In this section, we will detail the method of determining the number of measurements likely to be needed using the Bayesian approach. Note that we report the **median** of 20 example runs rather than the mean, as when the distribution of something is unknown and we have only a small number

of samples to estimate it from it is much easier to estimate the median than the mean. Clearly when summing these various numbers we must then caution the reader that the sum of medians is not usually equal to the median of the sum of the variables.

6.6.1 Continuous variable with two-sided constraint

As we discussed before, our approach will be able to simulate reality better by not relying on the assumption of Gaussianity. Our Bayesian software will take into account the possibility of the data coming from a Student distribution, which is a generalisation of a Gaussian. Clearly this generalises to a log-Student instead of a log-Gaussian by taking the log of the data, and of course we will be considering mixtures of these distributions. In combination this makes the Bayesian method vastly more powerful and realistic than one assuming Gaussianity which ISO 20072 tempts the user to do.

The Student distribution with shape m (or with "$2m$ degrees of freedom") is the marginal distribution obtained by integrating out s from the following two: $P(x \mid \mu, s, \sigma) = \sqrt{\frac{s}{2\pi\sigma^2}} e^{-\frac{s}{2\sigma^2}(x-\mu)^2}$ where s is Gamma distributed $P(s \mid m, r = m) = \frac{m^m}{\Gamma(m)} s^{m-1} e^{-ms}$. Then the Student distribution has the form

$$P(x \mid \mu, \sigma, m) = \frac{\Gamma(m + \frac{1}{2})}{\Gamma(m)} \frac{1}{\sqrt{2\pi m \sigma^2}} (1 + \frac{1}{2m} \frac{(x-\mu)^2}{\sigma^2})^{-(m+\frac{1}{2})}$$

We will also denote $\frac{1}{\sigma^2}$ by τ.

We consider only items that contain DFP basic item 1. As the distributions can be scaled, we will get the same numbers for other items that contain a two-sided constraint.

In section 5.1.1 we already saw the parameters for a superlative factory: $\mu = 2.2822$, $\sigma = 0.0722$ when the content is 0.95, and $\sigma = 0.0671$ when content is 0.975. Also, we set the true $m = 1000$ because the Student distribution tends to the Normal as $m \to +\infty$.

The acceptable region of μ and σ is plotted in figure 5.

We now set reasonable dependent priors on μ, τ and m and plot them in figures 6, 7 and 8.

We choose the prior for τ to be a Gamma distribution with shape and scale parameters $k_0 = 0.9$ and $t_0 = 140$. This can be found in figure 6 with some comments.

We choose the prior for μ given τ to be Normal$(\mu_0, \frac{1}{\sqrt{\alpha_0 \tau}})$ with parameters $\mu_0 = 2.2822$ and $\alpha_0 = 0.02$. This is actually a conjugate prior to the Gaussian. See the contours of the prior on μ, σ in figure 7.

The conjugate prior on m takes the form of a so-called "Pro-Gamma" with parameters $a_0, b_0 > 0$: $P(m \mid a_0, b_0) \propto \frac{m^{b_0 m}}{\Gamma(m)^{b_0}} e^{-(a_0+b_0)m}$ with $a_0 = 0.2$ and $b_0 = 0.8$. See the plot in figure 8.

Therefore in this DFP basic item, the parameters of clusters are $\theta_k = (\mu_k, \tau_k, m_k)$ for $k = 1, 2 \ldots K$.

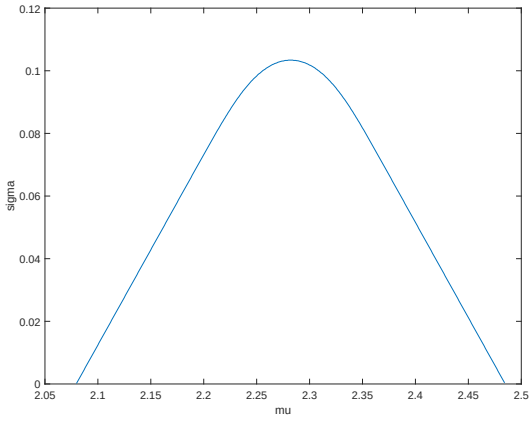

Figure 5: The acceptable region for μ and σ for DFP 1, in the special case that $m = \infty$ (i.e. distribution is (log-)Gaussian), wanting probability content of 0.975 lies below the curve, bearing in mind that the upper and lower limits are log(12) and log(8).

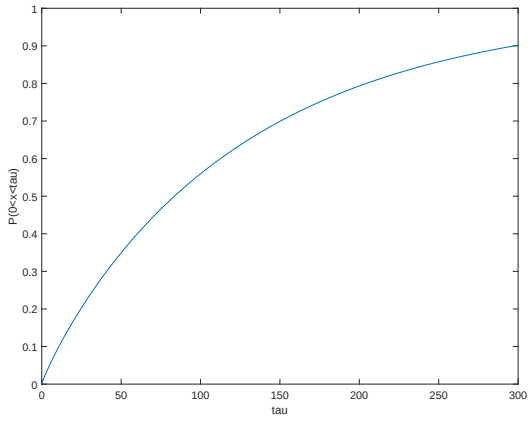

Figure 6: Plot of the **cumulative** distribution function on τ. It was taken such that the probability of τ being the in acceptable region in figure 5 to be about 0.2, so as not to bias the Bayesian approach towards what we know to be the true distribution.

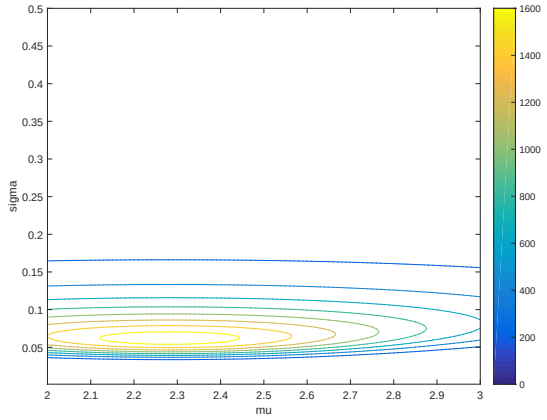

Figure 7: Plot of the contours of the prior on μ, σ. If we look at the acceptable region in figure 5, we can see that there is enough probability both inside and outside the acceptable region, so again the prior is not prejudiced in favour of H_1.

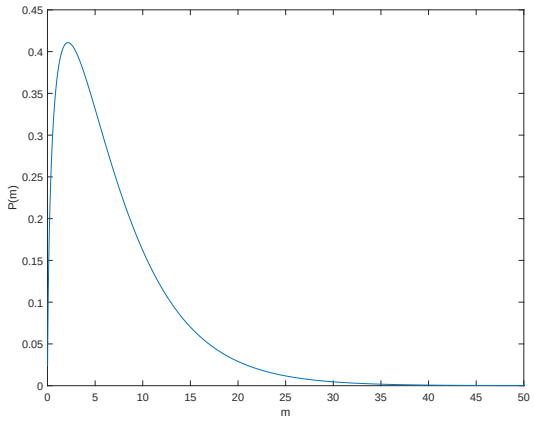

Figure 8: Plot of the selected conjugate prior for m.

Next we describe the Gibbs sampling updates which take the place of step 4 of section 6.4, for which we reintroduce $\mathbf{s} = (s_n)_{n=1,\ldots,N}$ in order to be able to sample more easily, so that (taking the case that there is only a single cluster for simplicity) $P(\mathbf{x}, m, \mu, \tau, \mathbf{s}) = P(\mathbf{x} \mid m, \mu, \tau, \mathbf{s})P(m)P(\mu \mid \tau)P(\tau)P(\mathbf{s} \mid m)$:

1. Sample $\mathbf{s}$ from Gamma distributions with shape and scale parameters for s_n being Gamma($m + \frac{1}{2}, m + \frac{\tau}{2}(x_n - \mu)^2$).
2. Sample m from the new Pro-Gamma with parameters $b_1 = b_0 + N$ and $a_1 = a_0 + \sum_{n=1}^{N}(s_n - \log(s_n)) - N$. Remark that this is valid since $a_1, b_1 > 0$ as a result of the inequality $u \geq 1 + \log(u)$ for $u > 0$.
3. Sample τ from the new Gamma distribution with the new shape and scale parameters $k_1 = k_0 + \frac{N+1}{2}$ and $t_1 = t_0 + \frac{\alpha_0(\mu-\mu_0)^2}{2} + \frac{\sum_{n=1}^{N} s_n(x_n-\mu)^2}{2}$.
4. Sample μ from the Normal$(\mu_1, \frac{1}{\sqrt{\tau\alpha_1}})$ with the new parameters $\alpha_1 = \alpha_0 + \sum_{n=1}^{N} s_n$ and $\mu_1 = \frac{\sum_{n=1}^{N} s_n x_n + \alpha_0\mu_0}{\sum_{i=n}^{N} s_n + \alpha_0}$.
5. Repeat all but the last step above in reverse order, to ensure palindromicity and thus detailed balance.

An example of the sample trajectories of μ and σ for a 5-cluster example can be found in figures 9 and 10.

Finally, we report the number of measurements needed as the **median** over 20 runs.

1. For content $p = 0.95$, we found that 37 devices are needed.
2. For content $p = 0.975$, we found that 41 devices are needed.

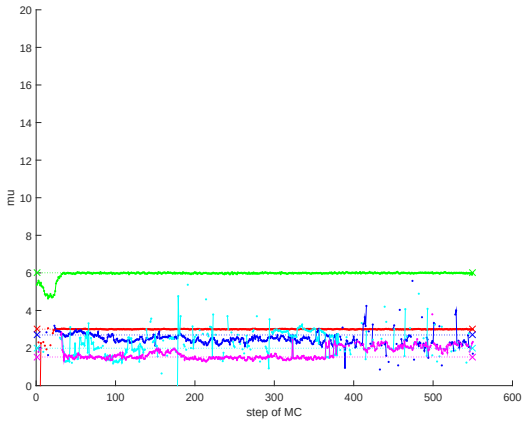

Figure 9: Markov Chain of μ when the data is taken from a true distribution with 5 mixture components (the dotted lines). Note that the blue, magenta and cyan components are negligible by figure 4, so the Markov Chain is focused on the red and green.

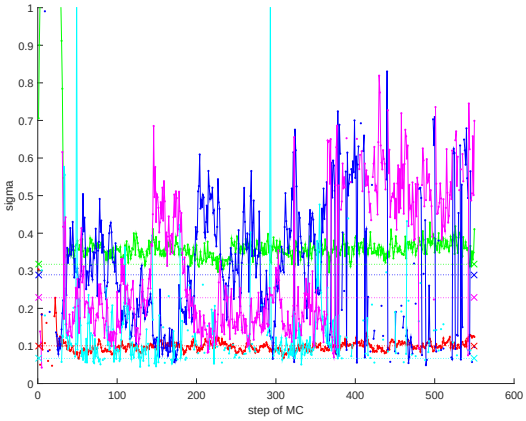

Figure 10: Markov Chain of σ when the data is taken from a true distribution with 5 mixture components (the dotted lines). Note that the blue, magenta and cyan components are negligible by figure 4, so the Markov Chain is focused on the red and green.

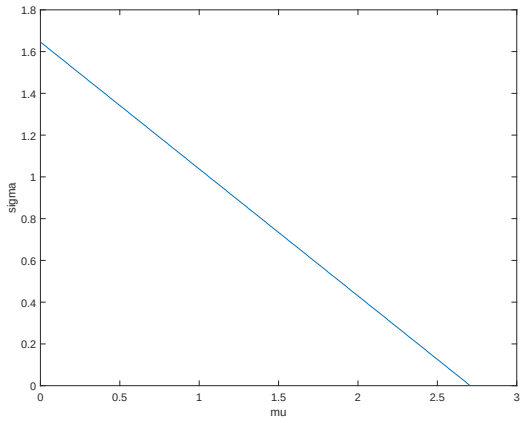

Figure 11: The acceptable region of μ, σ in DFP basic item 3 lies below the line.

Therefore, as there are 3 basic DFP items with a two-sided constraint, and as there are 8 environmental conditions for content 0.95 and another 6 for content 0.975 in ISO 20072, the total number of measurements needed to pass items that contain a two-sided constraint is $3 \times (37 \times 8 + 41 \times 6) = 1626$. We note that this is smaller than the corresponding frequentist number, even though we have allowed for both Student non-Gaussianity and for mixtures of distributions.

6.6.2 Continuous parameter with one-sided constraint

The only difference from the two-sided problem is that we will draw superlative data from a Normal distribution with $\mu = 1.3540$ and $\sigma = 0.5257$ for 0.95 content, respectively $\sigma = 0.4824$ for 0.975 content.

The corresponding acceptable region is in figure 11. The priors on τ, μ will scale accordingly.

Again, we report the **median** number of measurements needed over 20 runs:

1. For content $p = 0.95$, we found that 20 devices are needed.
2. For content $p = 0.975$, we found that 25 devices are needed.

Therefore, as there are 5 basic DFP items with a one-sided constraint, and as there are 8 environmental conditions for content 0.95 and another 6 for content 0.975 in ISO 20072, the total number of measurements needed to pass items that contain a one-sided constraint is $5 \times (20 \times 8 + 25 \times 6) = 1550$.

6.6.3 Probability parameter

We will consider the DFP basic items with the target probability 0.99, as the ones with target 0.93 are treated similarly. (Superlative test data will then be generated using a true probability parameter of 0.999 as in section 3.1.)

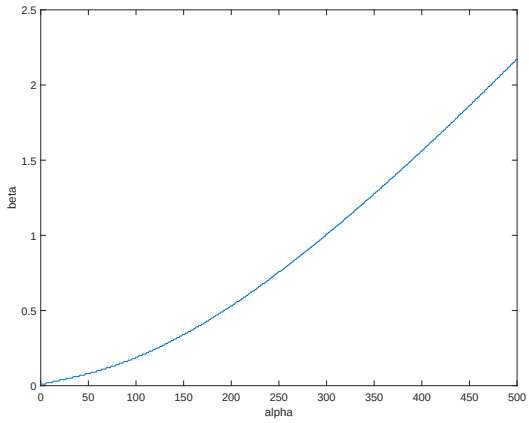

Figure 12: The acceptable region of α, β in DFP basic item 2, target 0.99, lies below the curve and is unbounded on the right.

We then assume a model in which each cluster in a mixture has parameter distributed Beta(α, β), then find the posterior distribution on α and β given some data and a suitable prior. Therefore the parameters of the clusters will be $\theta_k = (\alpha_k, \beta_k)$.

The acceptable region of α and β is plotted in figure 12.

We choose a prior on α and β from the "ProBeta" family, the conjugate distribution to the Beta distribution with respect to the joint parameters (α, β), whose probability density function is given by

$$P(\alpha, \beta \,|\, p_1, p_2, N_0) \propto \left(\frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \right)^{N_0} \cdot p_1^{N_0(\alpha-1)} \cdot p_2^{N_0(\beta-1)},$$

where p_1, p_2, and N_0 are independent parameters, constrained by $p_1, p_2, N_0 > 0$ and $p_1 + p_2 < 1$ so the expression has finite integral.

The values we take for p_1, p_2, N_0 affect the shapes of the resulting Beta distributions. To avoid prejudicing the outcome, we would like to get all intuitive types of Beta distributions when drawing from the selected prior. We found the prior with $p_1 = p_2 = 0.4525$ and $N_0 = 0.01$ satisfactory. See figure 13 of the contour plot and figure 14 of Beta distributions drawn from the prior.

We note that the closure of the space of mixture distributions so modelled includes the frequentist critical distribution for this item.

In this protocol, we measure the responses of N devices, testing the nth device N_n times and obtaining n_n successes. We obtain two N-dimensional vectors, $\mathbf{N}$ and $\mathbf{n}$. Since

$$\begin{aligned} P(\mathbf{N}, \mathbf{n} \mid \alpha, \beta) &= \int P(\mathbf{N}, \mathbf{n}, \mathbf{p} \mid \alpha, \beta)\, d\mathbf{p} \\ &= \int P(\mathbf{N}, \mathbf{n} \mid \mathbf{p}) P(\mathbf{p} \mid \alpha, \beta)\, d\mathbf{p} \end{aligned}$$

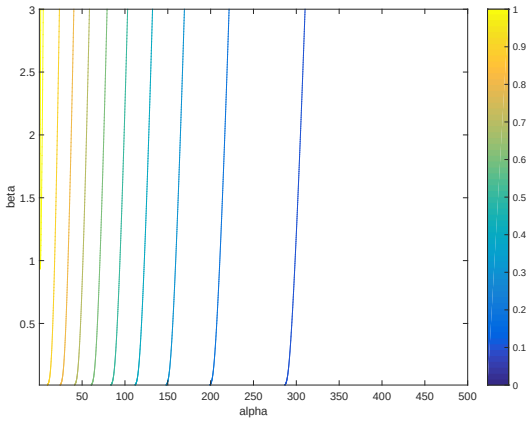

Figure 13: Contour plot of the unnormalised prior on α, β. Looking at the acceptable region in figure 12, we can see that there is a reasonable amount of probability both inside and outside of the acceptable region.

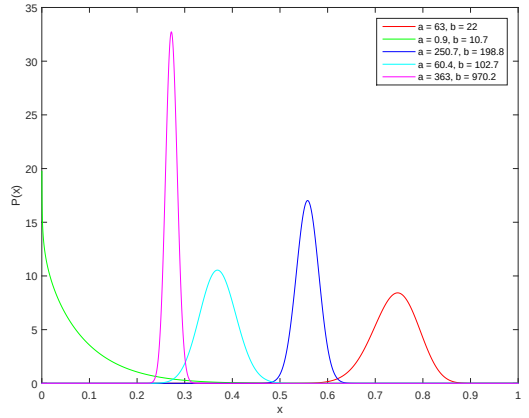

Figure 14: Sample Beta distributions drawn from the prior. They have various shapes and sizes, covering both narrow and broad and both those internal to the interval and those jammed up at one end. The values of α, β are specified in the top-right key.

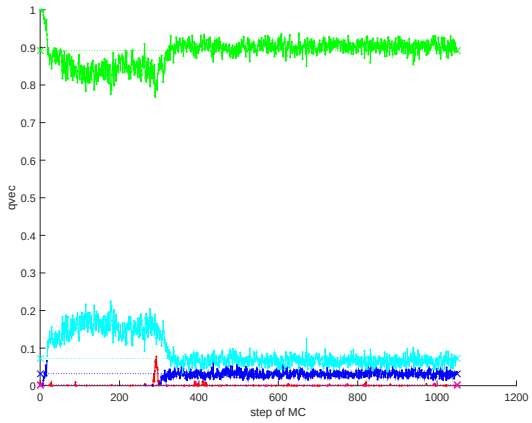

Figure 15: Example of the sample trajectories of the Markov chain for the cluster weights $\mathbf{q}$, when data is sampled from a true distribution with 5 mixture components (the dotted lines). Note that the magenta and red components are almost negligible, and that there is insufficient data to determine whether some of the data points belong to the magenta or red clusters.

our posterior probability distribution subsequently takes the following form:

$$P(\alpha, \beta \,|\, \mathbf{N}, \mathbf{n}) = \frac{P(\mathbf{N}, \mathbf{n} \,|\, \alpha, \beta) \cdot P(\alpha, \beta)}{P(\mathbf{N}, \mathbf{n})}$$
$$\propto \left(\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\right)^{N_0+N} p_1^{N_0(\alpha-1)} p_2^{N_0(\beta-1)} \prod_{n=1}^{N} \frac{\Gamma(\alpha+n_n)\cdot\Gamma(\beta+N_n-n_n)}{\Gamma(\alpha+\beta+N_n)}$$

This also tells us how to do the Gibbs Sampling step of the new element in the Markov Chain.

Plots of example Markov chain sample trajectories along with some comments can be found in figures 15, 16 and 17.

We considered only designs where the value of N_n was the same for all n, although exactly the same analysis would apply if one were to choose a variety of values for the various N_ns beforehand, or, by a similar analysis to that in section 6.1.2 above, even if one were to decide which device to test next based on the results obtained from the various devices so far. We considered various possible common values for the N_ns, and chose the considered value which gave smallest overall number of measurements.

Reporting as usual the median of the sample sizes required in 20 runs, we found the following.

For DFP basic item 2 with lower constraint 0.99 we needed, **per item**:

1. For content $p = 0.95$, testing each device 100 times, we needed to test a median of 45 devices, for a total of 4500 measurements needed.
2. For content $p = 0.975$, testing each device 200 times, we needed to test a median of 67 devices, for a total of 13400 measurements needed.

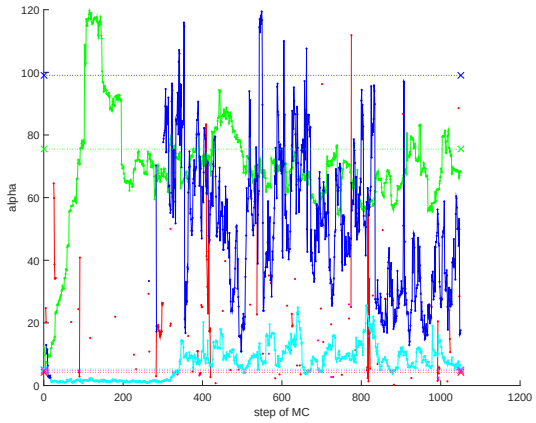

Figure 16: Example of the sample trajectories of the Markov Chain for the parameter $\boldsymbol{\alpha}$. Note that, according to figure 15, the most accurate are again green, blue and cyan, because there is not enough data to represent the almost negligible red and magenta. Note also that there is considerable uncertainty about the precise values of the green and blue cluster parameters, all of which represent quite narrow clusters; a 2-d plot with $\boldsymbol{\beta}$ would show trajectories of (α_k, β_k) moving in and out along radial lines through the origin, representing Beta distributions all of approximately the same mean, but of slightly varying widths. A corresponding plot of $\boldsymbol{\beta}$ is in Figure 17.

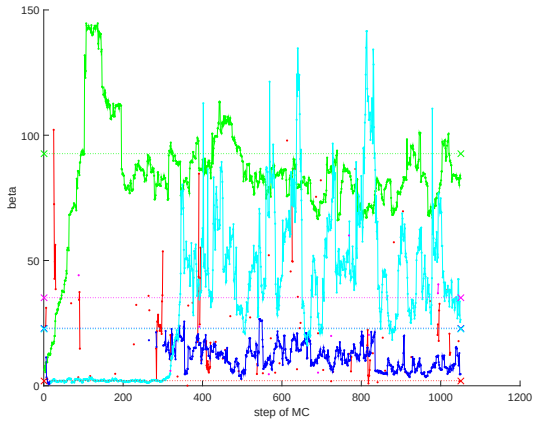

Figure 17: Markov Chain of $\boldsymbol{\beta}$. Note that the most accurate are again green, blue and cyan, because as seen in figure 15, there is not enough data to represent the almost negligible red and magenta.

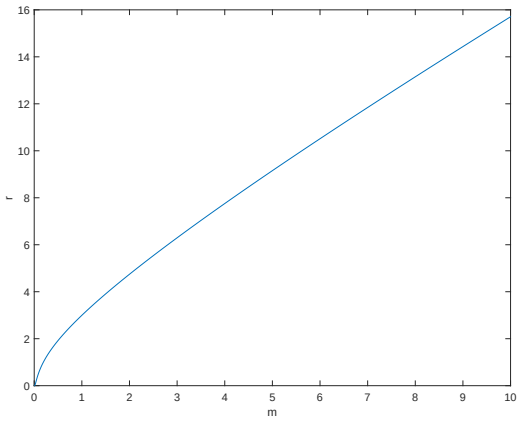

Figure 18: The acceptable region of (m, r) is above the line in DFP basic item 8 with upper constraint on the Poisson rate of 1 per hour.

As in ISO 20072 there are 8 environmental conditions that require content 0.95, and another 6 that require content 0.975, and there are 2 basic items in the DFP with target 0.99, a total number of $2 \times (4500 \times 8 + 13400 \times 6) = 232800$ measurements needed.

For DFP basic item 6 with lower constraint 0.93 we needed, **per item**:

1. For content $p = 0.95$, testing each device 7 times, we needed to test a median of 20 devices, for a total of 140 measurements.
2. For content $p = 0.975$, testing each device 10 times, we needed to test a median of 21 devices, for a total of 210 measurements needed.

As in ISO 20072 there are 8 environmental conditions that require content 0.95, and another 6 that require content 0.975, and there are 4 basic items in the DFP with target 0.99, a total number of $4 \times (140 \times 8 + 210 \times 6) = 9520$ measurements needed.

6.6.4 Poisson rate

We next consider the two DFP basic items with the Poisson rate constrained to be below 1 per hour. Superlative data will be generated from simulated devices with rate 0.1 per hour, as in section 3.1 above.

We then assume a model in which each cluster in a mixture has rate parameter distributed Gamma(m, r), then find the posterior distribution on $\mathbf{m}$, $\mathbf{r}$ given some data and a suitable prior. Therefore the parameters of the clusters with be $\theta_k = (m_k, r_k)$.

The acceptable region of (m, r) is plotted in figure 18.

We choose a prior on (m, r) from the "ProGamma" distribution family, the conjugate distribution

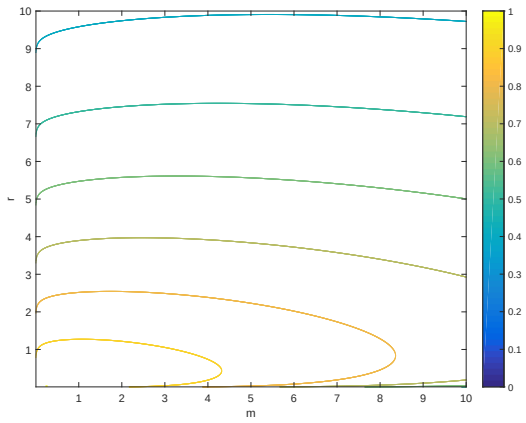

Figure 19: Contour plot of the prior on (m, r). Looking at the acceptable region in figure 18, we can see that there is a reasonable amount of probability both inside and outside of the acceptable region.

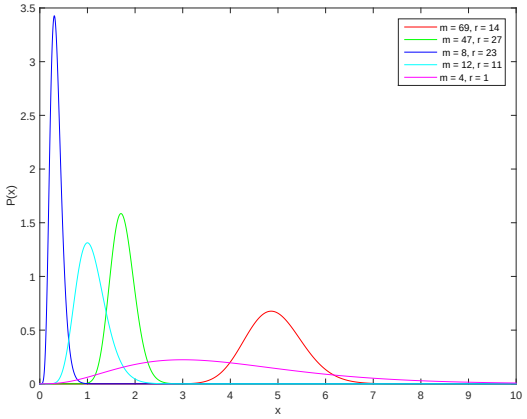

Figure 20: Sample Gamma distributions drawn from the prior. They have various shapes and sizes. The values of (m, r) are specified in the top-right key.

to the Gamma with respect to the joint parameters (m, r), whose probability density function is given by

$$P(m, r \,|\, \alpha, \beta, N_0) \propto \left(\frac{r^m}{\Gamma(m)}\right)^{N_0} \beta^{mN_0} e^{-\beta r} \frac{e^{-\alpha m}}{N_0^{mN_0}},$$

where α, β, and N_0 are independent parameters, constrained by $\alpha, \beta, N_0 > 0$ so that the expression has finite integral.

The values we take for α, β, N_0 affect the shapes of the resulting Gamma distributions. To avoid prejudicing the outcome, we would like to get all intuitive types of Gamma distributions from the selected prior. We found the prior with $\alpha = 0.03$, $\beta = 0.1$ and $N_0 = 0.01$ satisfactory. See figure 19 for the contour plot and figure 20 for Gamma distributions drawn from the prior.

In this protocol, we measure the responses of N devices, testing the nth device for T_n hours and

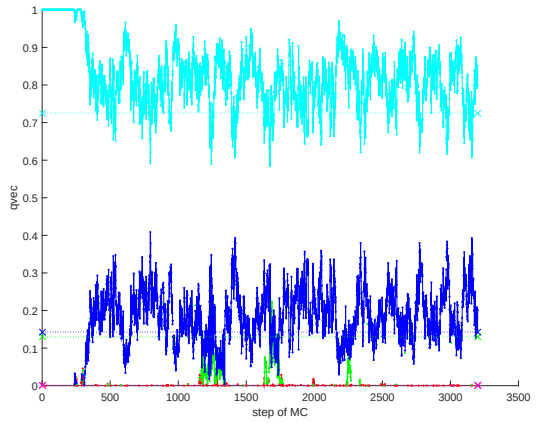

Figure 21: Example Markov chain sample trajectories of $\mathbf{q}$ when data is sampled from a true distribution with 5 mixture components (the dotted lines). Note that the blue and green clusters are virtually identical in all parameters (see also Figures 22 and 23), and it is therefore entirely reasonable for them to be interpreted as a single cluster in most samples from the chain. The red and magenta components are almost negligible, and there is insufficient data to determine whether some of the data points might belong to the red or magenta clusters.

observing n_n failures. We obtain two N-dimensional vectors, $\mathbf{T}$ and $\mathbf{n}$. Since

$$\begin{aligned} P(\mathbf{N}, \mathbf{n} \mid \alpha, \beta) &= \int P(\mathbf{N}, \mathbf{n}, \boldsymbol{\lambda} \mid \alpha, \beta)\, d\boldsymbol{\lambda} \\ &= \int P(\mathbf{N}, \mathbf{n} \mid \boldsymbol{\lambda}) P(\boldsymbol{\lambda} \mid \alpha, \beta)\, d\boldsymbol{\lambda} \end{aligned}$$

our posterior probability distribution subsequently takes the following form:

$$P(m, r \mid \mathbf{T}, \mathbf{n}) \propto \left(\frac{r^m}{\Gamma(m)}\right)^{N+N_0} \left(\prod_{n=1}^{N} \frac{1}{(r+T_n)^{m+n_n}}\right) \left(\prod_{n=1}^{N} \Gamma(m+n_n)\right) \beta^{mN_0} e^{-\beta r} \frac{e^{-\alpha m}}{N_0^{mN_0}}$$

This also tells us how to do the Gibbs Sampling step of the new element in the Markov Chain.

Plots of example Markov chain sample trajectories along with some comments can be found in figures 21, 22 and 23.

We considered only designs where the value of T_n was the same for all n, although exactly the same analysis would apply if one were to choose a variety of values for the various T_ns beforehand, or, by a similar analysis to that in section 6.1.2 above, even if one were to decide which device to observe next based on the results obtained from the various devices so far. We considered various possible common values for the T_ns, and chose the considered value which gave the smallest overall number of device-hours of observation needed.

Reporting as usual the median of the sample sizes required in 20 runs, we found the following.

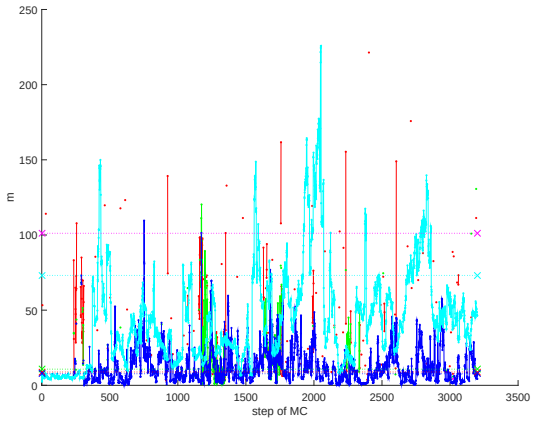

Figure 22: Example Markov chain sample trajectories of **m**. Note that the blue and green clusters are virtually identical in all parameters, and it is therefore entirely reasonable for them to be interpreted as a single cluster in most samples from the chain. As we see in Figure 21, the red and magenta clusters are almost negligible in weight, and there is in consequence not enough data to be sure of their existence.

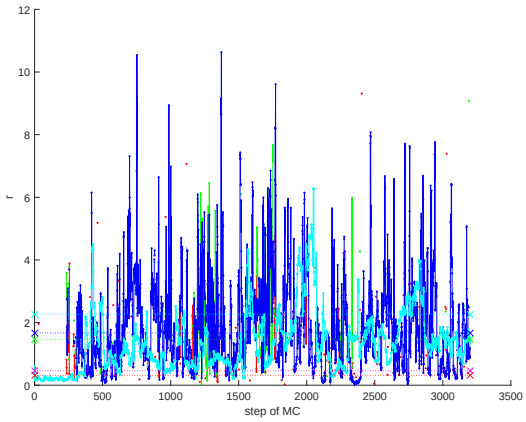

Figure 23: Example Markov chain sample trajectories of **r**. Note that the blue and green clusters are virtually identical in all parameters, and it is therefore entirely reasonable for them to be interpreted as a single cluster in most samples from the chain. As we see in Figure 21, the red and magenta clusters are almost negligible in weight, and there is in consequence not enough data to be sure of their existence.

For DFP basic item 8 with Poisson rate constrained to be below 1 per hour we obtained, **per item**:

1. For content $p = 0.95$, observing each device for 0.1 hours, we needed 60 devices, for a total of 6 device-hours of observation.
2. For content $p = 0.975$, observing each device for 0.1 hours, we needed 110 devices, for a total of 11 device-hours of observation.

As in ISO 20072 there are 8 environmental conditions that require content 0.95, and another 6 that require content 0.975, and there are 2 basic items in the DFP constraining a Poisson rate to be below 1 per hour, a total number of $2 \times (6 \times 8 + 11 \times 6) = 228$ device-hours needed.

6.7 Summary of effort needed in the Bayesian approach

We now have all the necessary information to assess how much testing needs to be done, in order to have a probability at least 95% that a **superlative** factory passes the whole test, that is passes all items. Note that a superlative factory is an almost perfect one, and in reality such a factory may be almost impossible to design. Therefore, in reality, these numbers may be higher — but equally, it is likely that for some items the factory will turn out to be better than superlative, while for others it is worse, and the Bayesian test will automatically increase the amount of testing only for those items that are worse than superlative, while it equally will reduce the amount of testing for those that turn out to be better than superlative.

Summing the number of measurements needed for each category of items, we get that $232800 + 9520 + 1550 + 1626 = 245496$ measurements are required plus 228 device-hours of observation. If we reckon that a robot could perform one measurement per minute, this gives a total of almost 180 days of robot time. This looks entirely feasible, in contrast with the frequentist result.

To be even-handed, we make some comments about safety and cheating.

The most important feature of the Bayesian approach is, in our opinion, the possibility of improving the quality control of the devices without massively increasing the required amount of testing. That is, a manufacturer may reasonably add several additional basic items to the DFP, thus making its device better and providing a safer product for its customers.

It is of course also *possible* to cheat in a Bayesian approach by censoring data just as in a frequentist approach. The question is whether the method provides an incentive to cheat or not. In section 7.2 below we argue that the incentives to cheat provided by the frequentist approach are not present in the Bayesian approach.

7 Comparison of the frequentist and Bayesian approaches

7.1 Workload required

We now compare the numbers of measurements needed when testing with the frequentist method according to ISO 20072 with the number needed with the Bayesian approach we proposed. We have also calculated the approximate time needed to test a factory. See the table in Figure 24 for a summary of these results.

type of item	cont. 0.95	cont. 0.975	no of items in DFP	total measurements	total time
Probability parameter p = 0.99 B	4500	13400	2	232800	160 days
Probability parameter p = 0.99 F	83141	168072	2	3347120	6.3 years
Probability parameter p = 0.93 B	140	210	4	9520	7 days
Probability parameter p = 0.93 F	12920	26128	4	1040512	2 years
One-sided B	20	25	5	1550	1 day
One-sided F	93	135	5	7770	5 days
Two-sided B	37	41	3	1626	1 day
Two-sided F	108	154	3	5364	4 days
Total measurements B				245496	169 days
Total measurements F				4400766	8.3 years
Poisson rate lambda = 1 B	6	11	2	228 hours	10 days
Poisson rate lambda = 1 F	843	1801	2	35100 hours	4 years
Total time B					180 days
Total time F					12.3 years

Figure 24: Table comparing the numbers in the frequentist vs Bayesian methods. The total number of measurements was calculated considering that there are 8 items in ISO 20072 that require 0.95 content, and another 6 that require 0.975 content. The total time required was computed assuming a robot that does one measurement per minute is available.

It is immediately obvious that the Bayesian method requires vastly fewer measurements, and overall approximately 25 times less robot-time, than the frequentist method. One of the facts which contributes to the enormous difference in workload required for the two methods is that

- When testing with the Bayesian approach, the number of measurements needed grows as the sum of the numbers of measurements needed for each item in the DFP, each of which is

independent of the existence of other items.

- On the other hand, when testing with the frequentist approach, increasing the number of items in the DFP **massively** increases the number of measurements needed because of the need to increase the pass probability of individual items.

7.2 Comparison of incentives to cheat

There are some ways in which the phenomenon of cheating is independent of the testing method, as the urge to cheat varies with personality, and it is possible to cheat in a Bayesian approach by censoring data just as in a frequentist approach. However, we think that the Bayesian method we described lacks the incentive to cheat observed with the frequentist approach for several reasons:

1. The amount of testing needed in the Bayesian approach is **much lower** than what was needed in the frequentist method. Thus, it is feasible to do the testing **honestly**, without censoring or manipulating the data.
2. The Bayesian method accurately reports the probability, given the data observed so far, that a factory is compliant (or isn't). Hence if a manufacturer tests a compliant factory, the design will pass with probability 1, while if a factory has not yet passed a test in progress, the manufacturer is continually aware of the probability that the factory is nonetheless compliant, and can make an informed decision on whether to continue testing.
3. Moreover, if the Bayesian method passes a factory, then for each item the probability that that factory is good is high (0.95 under the ISO 20072 conventions). Thus patients are protected under a Bayesian analysis just as under an honestly executed frequentist one — but in addition, they get the benefits that
   (a) Worthwhile treatments do actually become available to them, rather than be rejected by the minuscule pass probabilities of the frequentist approach, and
   (b) The probability that the manufacturer has cheated when "passing" their product is no longer close to 1.

7.3 Comparison of the incentives to produce a high quality factory

We now address an argument frequently raised by some regulatory authorities as justification of the use of frequentist methods. That argument is that a low pass probability for e.g. a superlative factory should incentivize the manufacturer to only design perfect factories. However, there are in our view a number of difficulties with this reasoning:

1. It is usually impossible to design a perfect factory.
2. The manufacturer has no way of knowing whether or not his factory is perfect before embarking on testing it. If it fails, he has then no recourse; he cannot retest it from scratch, and in the frequentist paradigm he cannot extend the test to handle the imperfections that he now knows must be present.

3. The Bayesian approach we have described also provides incentives to design better factories, as the number of measurements that will be needed for any item will automatically rise as the factory gets closer to the specified constraint. Given that it is thus possible to design tests that both provide an incentive to design better factories, and which will eventually pass good factories, such an approach is surely preferable.

7.4 Pseudo-Bayesian approaches

A "compromise" approach has also been described in [7], wherein such an approach is described as "Bayesian". However, as seen in section 4.8 of [7], this approach actually wants to handicap a Bayesian approach by considering also the Bayesianly-irrelevant probability that a factory for which all items but one are perfect but one is exactly borderline will pass the Bayesian test.

We have explained in section 5.4.2 above why we believe such probabilities to be irrelevant to the discussion. A Bayesian test as described in section 6 above will never satisfy restrictions on such a probability other than in rare special cases. Modification of a Bayesian test to satisfy such restrictions in fact turns it into a frequentist test, albeit one somewhat more complicated that the ones we have considered above, and in doing so sacrifices the most important properties of the Bayesian approach discussed above, and in particular sacrifices the probability of one of passing a good factory.

7.5 Discussion of choice of prior

The favourite argument of frequentists and institutions that advocate frequentist methods against the Bayesian approach is that the results of the Bayesian method depend on "the choice of prior" (see section 3.9 of [8]). To illustrate this, we compare the median of the number of measurements needed when testing the Poisson rate in section 6.6.4 for a range of priors. We choose this DFP item because the data consists mainly of periods where no events are observed, and consequently the results are more influenced by the prior than in other cases. Here we will vary only the parameter β in the prior, and will consider the case of the desired probability content being 0.975 and the number of hours of observation per device still being 0.1 .

1. As we have seen in section 6.6.4, with $\beta = 0.1$ the prior puts a reasonable amount of probability both inside and outside the acceptable region; in fact the prior probability that a single cluster factory is good for that item is 0.004, and the MCMC runs showed a median of 110 devices are needed.
2. When we take $\beta = 0.05$, the prior again puts a reasonable amount of probability both inside and outside the acceptable region, with the prior probability that the factory is good now being about 0.012, as seen in figure 25. Here we obtained a median of 104 devices needed.
3. When we take $\beta = 0.01$, the prior probability that the factory is good is now 0.12, as seen in figure 26. Here we obtained a median of 70 devices needed.
4. When we take $\beta = 0.0001$, the prior now puts more than 98% of the probability inside the acceptable region, as seen in figure 27. The prior is now totally prejudiced in favour of the factory being good for this item, and we need collect no data at all (since $0.98 > 0.95$).

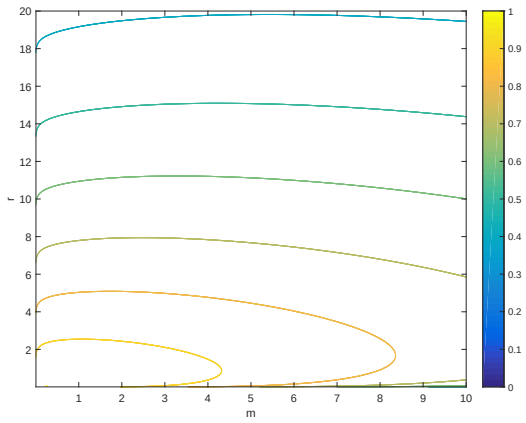

Figure 25: Prior on m, r when $\beta = 0.05$. The prior probability that the factory is good for this item is about 0.012 .

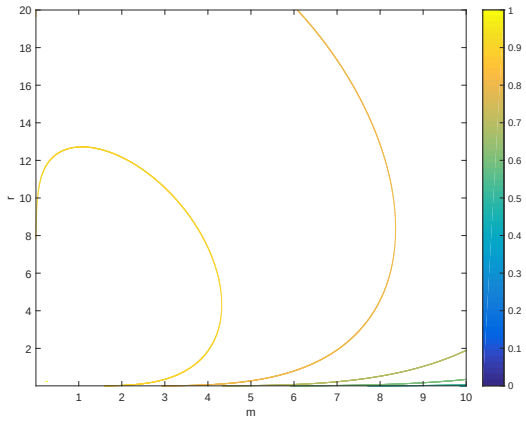

Figure 26: Prior on m, r when $\beta = 0.01$. The prior probabilty that the factory is good for this item is about 0.12 .

5. On the other hand, when we take $\beta = 1$, the prior probability that the factory is good is about 0.0004, as seen in figure 28. Here we obtained a median of 3000 devices.

Therefore, from the above analysis, we can conclude that indeed the Bayesian method depends on the selected prior. However, the important point is that when choosing reasonable priors that put between 0.1% and 5% of the probability inside the acceptable region and are generally well spread out, the required number of measurements is almost constant. The big differences arise when the priors are prejudiced either towards accepting or rejecting the design, but in practice this should **never** happen, if the intention is to test **honestly**. Moreover, it is entirely feasible for a regulator and manufacturer to come to agreement on an appropriate prior for each problem before embarking on data collection or finalising the design of the test plan.

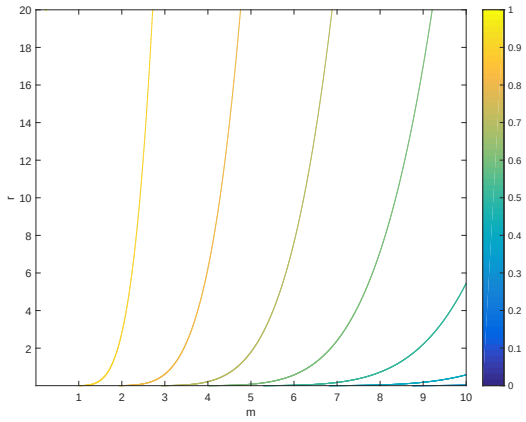

Figure 27: Prior on m, r when $\beta = 0.0001$. Looking at figure 18, we see that this prior puts more than 98% of the probability inside the acceptable region, and is so prejudiced that the factory is good for this item that we need collect no data at all.

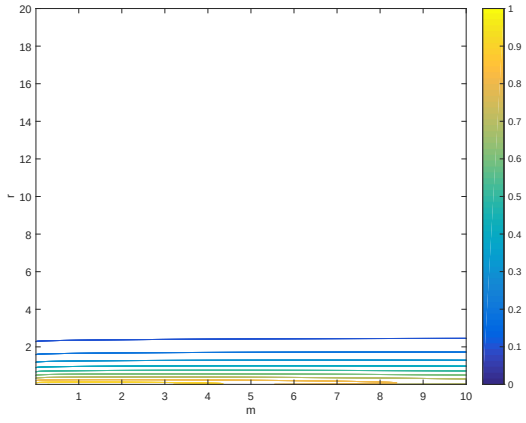

Figure 28: Prior on m, r when $\beta = 1$. The prior probability that the factory is good for this item is now 0.0004, so the prior is heavily prejudiced against the factory being good, and a lot of data is now needed before one can conclude otherwise.

8 Comments made in review by others

We note that the Independent Institute, a not for profit, non-partisan, scholarly research organisation, has assessed the actions of one regulator, the U.S. Food and Drug Administration (FDA), as follows [9]:

Defining type I and type II errors by

1. "Type I errors occur when the FDA approves a drug that ends up being a net bad for society."
2. "Type II errors occur when the FDA rejects a drug that would be a net good for society."

they find that the FDA cares more about type I errors mainly because they cause "major media attention, public concern, and congressional action"; their proposed testing procedures aim **not** to minimize the harm done by both types of errors, but only the harm done by type I error.

This supports our concern that better testing protocol is needed, preferably one that deals with both error types. The current protocol, as stated by the FDA, does not take into account the harm done to people that suffered because although a **good** treatment was produced, it failed a faulty testing procedure.

A further comparison is made in [10] by a so-called "whistle-blower" Henry Miller: a type I error attracts much attention and harms both patients and designers. However, a type II error only harms patients and is less vulnerable to attention, because "doctors, journalists and patients are usually unaware that a drug has been delayed or suppressed and that it would have saved individuals" [9].

Another testimonial from Vincent Kleinfeld, a Washington, D.C., attorney who worked on drug approval affairs in [10], wrote "the courts in this area of the law tend to equate the Food and Drug Administration with God, motherhood, and country". This status of "untouchable" achieved by institutions are another reason for a better protocol to be implemented, as there is no compensation for the defects of the current protocol.

9 Conclusion

In conclusion, our opinion is that in its current form, ISO 20072 is dangerous to patients because of both the frequentist protocol proposed, the way it encourages cheating when carrying out testing, and the risk of not passing good factories which would benefit patients. Instead, a sequential Bayesian protocol should be used.

We note that there are many analogies to this situation in the area of testing new drugs and medical devices. In [11] we see for example that in order to get a new drug accredited, drug companies typically run clinical trials with the intention of demonstrating a "statistically significant" benefit, but because frequentist statistical tests use data inefficiently, it is unnecessarily difficult to demonstrate statistical significance, and consequently:

1. It is in the drug company's interest to compare their new drug to a suboptimal treatment such as a placebo, rather than to the best available treatment - this is not in the interests of the patients in the trial.
2. Unnecessarily large numbers of patients are subjected to trials.
3. It is likely that many good drugs are failing to pass their tests (because frequentist "fconfidence" values and p-values do not actually control the probabilities we care about). Examples are also given in [11] of drugs that achieve "statistical significance" in test comparisons with placebos but which are not actually good for patients.

Unfortunately, many aspects of medicine are still dominated by classical frequentist statistical methods, so ISO 20072 is not an isolated example [11] of bad statistics driving organisations to behave in a way that is not in the public interest.

References

[1] International Standards Organization, Geneva, Switzerland, "ISO 20072:2009: Aerosol drug delivery device design verification — requirements and test methods." http://www.iso.org/iso/catalogue_detail.htm?csnumber=41989. Retrieved on 28.08.2016.

[2] International Pharmaceutical Aerosol Consortium on Regulation and Science (IPAC) and European Pharmaceutical Aerosol Group (EPAG), "Justification of the request for a negative vote on ISO DIS 20072, 'Aerosol drug delivery device design verification requirements and method'." http://ipacrs.org/assets/uploads/comments/On_ISO_20072_Nov_07.pdf, 2008. Retrieved on 28.08.2016.

[3] Radford M. Neal, "Probabilistic inference using Markov Chain Monte Carlo Methods," *1993*, Technical Report CRG-TR-93-1, Department of Computer Science, University of Toronto.

[4] Radford M. Neal, "Slice Sampling," *Annals of Statistics*, vol. 31, no. 3, pp. 705–776, 2003. http://projecteuclid.org/euclid.aos/1056562461, Retrieved on 28.08.2016.

[5] Yi Cao, "Munkres Assignment Algorithm." https://www.mathworks.com/matlabcentral/fileexchange/20328-munkres-assignment-algorithm, 2008. Retrieved on 28.08.2016.

[6] Harold W. Kuhn, "The Hungarian Method for the assignment problem," *Naval Research Logistics Quarterly*, vol. 2, pp. 83–97, 1955. http://onlinelibrary.wiley.com/doi/10.1002/nav.3800020109/epdf, Retrieved on 28.08.2016.

[7] U.S. Department of Health and Human Services, Food and Drug Administration, Center for Devices and Radiological Health, Office of Division of Biostatistics, Surveillance and Biometrics, Center for Biologics Evaluation and Research, "Guidance for the Use of Bayesian Statistics in Medical Device Clinical Trials." http://www.fda.gov/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/ucm071072.htm, 2010. Retrieved on 28.08.2016.

[8] Gerry Gray, PhD, Department of Health and Human Services, Food and Drug Administration, Center for Devices and Radiological Health, Division of Biostatistics, Surveillance and Biometrics, "Potential advantages and disadvantages to the Bayesian approach." http://www.fda.gov/ohrms/dockets/dockets/06d0191/06d-0191-ts00002-gray-vol1.pdf, 2006. Retrieved on 28.08.2016.

[9] The Independent Institute, "Glossary." http://www.fdareview.org/11_glossary.php#error_types, 2016. Retrieved on 28.08.2016.

[10] The Independent Institute, "Why the FDA has an incentive to delay the introduction of new drugs." http://www.fdareview.org/06_incentives.php, 2016. Retrieved on 28.08.2016.

[11] Ben Goldacre, *Bad Pharma: How drug companies mislead doctors and harm patients.* Faber and Faber (USA), 2012.

[12] G. E. Tauchen, "Maximum likelihood specification tests," *Journal of Econometrics*, vol. 30, pp. 415–443, 1985. Lemma 1.

A Ways in which the behaviour of frequentist confidence fails to match the intuitive behaviour expected

We give two theorems.

The first deals with the simultaneous occurrence of a number of hypotheses, each of which is the alternative hypothesis to some null hypothesis. It shows that whereas with the intuitive understanding of confidence the confidence of the intersection of multiple uncertain events is (usually much) smaller than the confidence of any of the individual events, the fconfidence of the intersection is always equal to the infimum of the fconfidence of the individual events - in other words it is far too large.

The second deals with the complementary situation that we are considering the union of a number of hypotheses, each of which is the alternative hypothesis to some null hypothesis. The situation here, though not relevant to the body of the work in this paper, is even worse. Whereas the intuitive confidence that at least one of a number of events has occurred must be at least as large as that of any one of the events, the fconfidence that at least one of multiple events has occurred can be strictly smaller than the fconfidence that a particular one of the events has occurred. In other words we can be less fconfident that a larger event occurred than that a small subset of that event occurred - totally contrary to any reasonable intuition.

In the following $F(A)$ denotes the fconfidence that an event A has occurred, it being understood that some particular critical region evident from the context is in use.

Theorem A.1 (Intersection of events)**.** *Suppose that $(H_{0,\iota})_{\iota\in I}$ is a family of events to be treated as null hypotheses, that $H_0 = \bigcup_{\iota\in I} H_{0,\iota}$, and that $H_{1,\iota}$ and H_1 are respectively the complements of $H_{0,\iota}$ and H_0, so that $H_1 = \bigcap_{\iota\in I} H_{1,\iota}$. Let C be a critical region in some data space for all the $H_{0,\iota}$ (for example, C might be $\prod_{\iota\in I} C_\iota$ if the data space is the product of one independent data space for each ι, or C might be $\bigcap_{\iota\in I} C_\iota$). Then $F(H_1) = F(\bigcap_{\iota\in I} H_{1,\iota}) = \inf_{\iota\in I} F(H_{1,\iota})$.*

Note that there is no requirement for I to be finite or even countable.

Proof. By definition, $F(H_{1,\iota}) = 1 - \sup_{h\in H_{0,\iota}} P(x \in C \mid h)$, while $F(H_1) = 1 - \sup_{h\in H_0} P(x \in C \mid h)$. But since $H_0 = \bigcup_{\iota\in I} H_{0,\iota}$, it follows that $1 - F(H_1) = \sup_{h\in H_0} P(x \in C \mid h) = \sup_{\iota\in I} \sup_{h\in H_{0,\iota}} P(x \in C \mid h) = \sup_{\iota\in I} (1 - F(H_{1,\iota}))$, whence the result.

□

Theorem A.2 (Union of events)**.** *There exist events $H_{1,1}$ and $H_{1,2}$ such that the fconfidence $F(H_1)$ that $H_1 = H_{1,1} \cup H_{1,2}$ has occurred is strictly smaller than $\min_{\iota\in\{1,2\}} F(H_{1,\iota})$, where $H_{1,\iota}$ is the alternative hypothesis to (and complement of) $H_{0,\iota}$, and critical regions C_ι are in effect for testing $H_{0,\iota}$ for $\iota \in \{1, 2\}$, so that $C = C_1 \cup C_2$ is in effect for testing $H_0 = H_{0,1} \cap H_{0,2}$.*

Comment: Any reasonable intuitive understanding of fconfidence as confidence would require that $F(H_1) = F(\bigcup_{\iota\in\{1,2\}} H_{1,\iota}) \geq \max_{\iota\in\{1,2\}} F(H_{1,\iota})$.

Proof. It suffices, of course, to simply provide an example.

We take our hypothesis space to be a four point space, $H = \{h_{0,1}, h_{1,1}\} \times \{h_{0,2}, h_{1,2}\}$, with the null hypotheses being $H_{0,1} = \{h_{0,1}\} \times \{h_{0,2}, h_{1,2}\}$, $H_{0,2} = \{h_{0,1}, h_{1,1}\} \times \{h_{0,2}\}$, so that $H_0 = \{(h_{0,1}, h_{0,2})\}$ and $H_1 = \{(h_{0,1}, h_{1,2}), (h_{1,1}, h_{0,2}), (h_{1,1}, h_{1,2})\}$.

Similarly we take our dataspace to be a four point space, $X = \{x_{0,1}, x_{1,1}\} \times \{x_{0,2}, x_{1,2}\}$, with the two factors independent, $P(x_{0,1} \mid h) = 0.95$ for all $h \in H_{0,1}$, and $P(x_{0,2} \mid h) = 0.95$ for all $h \in H_{0,2}$. (Note that the frequentist paradigm bizarrely puts no requirement on us to specify $P(x \mid h)$ for $h = (h_{1,1}, h_{1,2})$.) Thus for example $P((x_{1,1}, x_{1,2}) \mid (h_{0,1}, h_{0,2})) = 0.05^2 = 0.0025$.

Our critical regions will be $C_1 = \{x_{1,1}\} \times \{x_{0,2}, x_{1,2}\}$ and $C_2 = \{x_{0,1}, x_{1,1}\} \times \{x_{1,2}\}$, so that $C = \{(x_{0,1}, x_{1,2}), (x_{1,1}, x_{0,2}), (x_{1,1}, x_{1,2})\}$.

Then $P(x \in C_1 \mid h) = 0.05$ for all $h \in H_{0,1}$, and similarly $P(x \in C_2 \mid h) = 0.05$ for all $h \in H_{0,2}$, so that C_1 and C_2 are respectively 95% critical regions for $H_{0,1}$ and $H_{0,2}$.

However, somewhat unexpectedly, for $h = (h_{0,1}, h_{0,2})$, we have $P(x \in C \mid h) = 1 - 0.95^2 = 0.0975$, so that C is only a 90.25% critical region for H_1, and $0.9025 = F(H_1) < \min_{\iota\in\{1,2\}} F(H_{1,\iota}) = 0.95$ even though $H_1 \supset H_{1,\iota}$ for all ι.

□

B Proof that Bayesian pass probability of a good factory approaches 1 as the amount of data increases

We consider a single item only, noting that since the number of items is finite showing the desired result for a single item suffices to prove it for all 224 items.

Our proof will use the uniform strong law of large numbers [12], the proof of which carries over to the following more general statement with negligible modification and without invoking the separability condition subsumed in the assumption of regularity in [12]: If

- X is a measure space;
- K is a compact metric space;
- ϕ is a function from $X \times K$ to $\mathbb{R}$;
- for all $k \in K$, $x \mapsto \phi(x, k)$ is measurable on X;
- there exists an integrable function $b : X \to \mathbb{R}$ such that for all $k \in K$ and $x \in X$, $|\phi(x, k)| \leq b(x)$;
- for all $k \in K$, for almost all $x \in X$, ϕ is continuous in k at k;
- $x_1, x_2, \ldots$ is a sequence of independent identically distributed random variables taking values in X;

then

- for all $k \in K$, the expectation of $\phi(x_1, k)$ exists and is finite, and
- with probability 1, $(\frac{1}{N} \sum_{n=1}^{N} \phi(x_n, k) \to \mathrm{E}\phi(x_1, k)$ uniformly in k as $N \to \infty)$.

We assume throughout that $H = H_0 \cup H_1$, $H_0 \cap H_1 = \emptyset$.

Theorem B.1. *Suppose that:*

1. *X is a measure space;*
2. *H is a locally compact metric space carrying a measure denoted μ on its Borel sets, compact subsets having finite measure, and non-empty open sets of H having positive measure;*
3. *H_0 is compact;*
4. *h, x are random variables on the probability space Ω taking values in H, X respectively possessing respectively a density function and a conditional density given h;*
5. *for almost all $x \in X$, $P(x|h)$ is a continuous function of h;*
6. *$P(x|h)$ is bounded on X locally uniformly in h, while $P(h)$ is locally bounded;*

7. *for all* $h' \in H$, $(P(x|h_1) \stackrel{a.s.}{=} P(x|h') \implies h' = h_1)$;

8. *there exists an open subset* $U_1 \ni h_1$ *of* H_1 *such that P(h) is bounded away from zero on* U_1*; and*

9. *there exists a measurable and almost surely positive function* $a : X \to \mathbb{R}$ *and a neighbourhood* U_2 *of* h_1 *such that both* $\mathrm{E}_{x|h_1} \log a(x)$ *exists and is finite and also for all* $x \in X$ *and* $h \in H_0 \cup U_2$, $P(x|h) \geq a(x)$.

Under these conditions if $x_1, x_2, \ldots, x_N$ *are independently distributed according to* $P(x|h_1)$*, then* $P(h \in H_1|x_1, \ldots, x_N) \to 1$ $\mathbf{x}$*-almost surely as* $N \to \infty$.

Before proving this result, we make a few remarks on the conditions.

The basic conditions 1 and 2 on H and X are standard, and are fulfilled in particular by any finite-dimensional real vector spaces. They are also fulfilled by discrete spaces with H_0 finite, in which case the continuity and boundedness conditions 5 and 6 become vacuous.

The local boundedness conditions 6 can usually be achieved by continuous transformations of H or X if they do not already apply.

The compactness condition 3 on H_0 can usually be achieved, if does not already apply, either by adjoining a point at infinity or by deleting part of H_0 that is extremely unlikely. For example, if h is the standard deviation of a Gaussian likelihood, and H_0 is $[1, \infty)$, then replacing H by $(0, 10^{100}]$ and H_0 by $[1, 10^{100}]$ achieves compactness of H_0 without materially altering the problem.

However, note that the result specifically does not cover the case that h_1 is on the boundary of H_1, for which the theorem does not hold.

Without the continuity condition 5 it is very hard to control the union of the uncountable number of zero probability subsets of the underlying probability space Ω that arise as a result of the strong law of large numbers.

Condition 7 is required because otherwise the likelihood provides insufficient information to distinguish h_1 from other possibilities, while condition 8 is needed to ensure that the prior has not prejudicially excluded the right answer.

The remaining condition 9 can usually be achieved if not already present by replacing $P(x|h)$ by $(1 - \gamma)P(x|h) + \gamma b(x)$, where b is a bounded probability density on X such that $\int_X (b(x) + P(x|h_1)) \log b(x) dx$ exists and is finite and $\gamma > 0$ is a small positive constant. This is intuitively bizarre, as doing so makes the resulting random variable x contain *less* information about h. We therefore feel that some more able mathematician might be able to manage without this condition.

Proof. In the following we assume that $\mathbf{x}_N = (x_1, x_2, \ldots, x_N)$, that $\mathbf{x}$ means $\mathbf{x}_N$, and that x and all x_n are distributed according to $P(x|h_1)$. Thus for a fixed value $h \in H$, $P(x|h)$ is a random variable calculated by letting x be distributed according to $P(x|h_1)$ then calculating $P(x|h)$. We exclude the null subset of Ω on which $P(x|h_1) = 0$.

For any $h \in H$, let us define $q(h) = \mathrm{E}_{x|h_1} \log P(x|h)$.

From Jensen's inequality we note that for any $h \in H \setminus \{h_1\}$ the given conditions then imply $q(h_1) > q(h)$, with strict inequality.

Conditions 5, 6, and 9, and the dominated convergence theorem give us that for some neighbourhood U of h_1 in H_1, q is continuous on $H_0 \cup U$. Now the compactness of H_0 implies that q attains a maximum M on H_0, and we then have that $q(h_1) > M$. Indeed, for some compact neighbourhood $V \subseteq U \subseteq H_1$ of h_1 on which $P(x|h)$ is uniformly bounded and for some $\delta > 0$ and $\epsilon > 0$, we have $P(h) > \epsilon$ and $q(h) > M + \delta$ and $P(x|h) > 0$ on V a.s. (by conditions 5 and 9). We note that then on V we also have $P(\mathbf{x}|h) > 0$ a.s. for all N.

Applying the uniform strong law of large numbers with K equal to H_0 or V and $\phi(x, k)$ equal to $\log P(x|k)$, we find that almost surely, uniformly in $h \in V$ and $h_0 \in H_0$, $P(\mathbf{x}|h_0)^{1/N} \to e^{q(h_0)}$ and $P(\mathbf{x}|h)^{1/N} \to e^{q(h)}$, so $(\frac{P(\mathbf{x}|h_0)}{P(\mathbf{x}|h)})^{1/N} \to e^{q(h_0)-q(h)} < e^{-\delta} < 1$, and hence almost surely $(\frac{P(\mathbf{x}|h_0)}{P(\mathbf{x}|h)} \to 0$ uniformly in h_0 and h as $N \to \infty$).

We now fix any $\omega \in \Omega$ for which this uniform convergence occurs.

It suffices now to show that for that ω, $\frac{\int_{H_0} P(h_0)P(\mathbf{x}|h_0)dh_0}{\int_V P(h)P(\mathbf{x}|h)dh} \to 0$. Now $P(h_0)$ and $\sup_{x \in X} P(x|h_0)$ are locally uniformly bounded on the compact set H_0 and therefore uniformly bounded on H_0. By rescaling the measures on H and X we may assume that that bound is 1, in which case $P(\mathbf{x}|h_0)$ is likewise bounded by 1 for all N. Since also $P(h)$ is bounded away from zero on V, it even suffices to show that $\frac{\int_{H_0} P(\mathbf{x}|h_0)dh_0}{\int_V P(\mathbf{x}|h)dh} \to 0$. We already know that the integrand in the denominator is never zero.

Now fix $h \in V$. Since the integrand in the numerator is uniformly bounded and $\mu(H_0)$ is finite, by the dominated convergence theorem we have $\frac{\int_{H_0} P(\mathbf{x}|h_0)dh_0}{P(\mathbf{x}|h)} \to 0$. Let $F_N = \int_{H_0} P(\mathbf{x}|h_0)dh_0$. Then $F_N/P(\mathbf{x}|h) \to 0$, so $P(\mathbf{x}|h)/F_N \to +\infty$. But $F_N^{-1} \int_V P(\mathbf{x}|h)dh \geq \int_V \inf_{N' \geq N} (F_{N'}^{-1} P(\mathbf{x}_{N'}|h))dh \to +\infty$ by the monotone convergence theorem.

Therefore $\frac{\int_V P(\mathbf{x}|h)dh}{\int_{H_0} P(\mathbf{x}|h_0)dh_0} \to +\infty$, and hence $\frac{\int_{H_0} P(\mathbf{x}|h_0)dh_0}{\int_V P(\mathbf{x}|h)dh} \to 0$ as $N \to \infty$.

Thus almost surely $\frac{\int_{H_0} P(h_0)P(\mathbf{x}|h_0)dh_0}{\int_V P(h)P(\mathbf{x}|h)dh} \to 0$, and hence $P(h \in H_0|\mathbf{x}) \to 0$, so $P(h \in H_1|\mathbf{x}) \to 1$.

□

C Correct values of k to replace those in ISO 20072 table D1

We have computed the correct values for the two-sided tolerance factors k in the tables in figures 29, 30, 31 and 32.

N	p = 0.75	p = 0.9	p = 0.95	p = 0.975	p = 0.99	p = 0.995	p = 0.999
2	11.759	20.571	26.249	31.247	37.083	41.066	49.295
3	3.805	6.159	7.66	8.996	10.559	11.632	13.869
4	2.619	4.164	5.146	6.02	7.045	7.752	9.223
5	2.151	3.409	4.205	4.913	5.744	6.316	7.507
6	1.904	3.006	3.707	4.33	5.062	5.565	6.614
7	1.775	2.757	3.402	3.973	4.646	5.107	6.068
8	1.695	2.584	3.187	3.724	4.353	4.787	5.688
9	1.637	2.464	3.032	3.543	4.144	4.557	5.417
10	1.592	2.377	2.914	3.404	3.982	4.379	5.205
11	1.557	2.316	2.822	3.292	3.852	4.237	5.037
12	1.528	2.268	2.752	3.204	3.748	4.123	4.903
13	1.504	2.228	2.696	3.131	3.66	4.025	4.787
14	1.483	2.194	2.653	3.073	3.587	3.946	4.693
15	1.466	2.166	2.616	3.023	3.524	3.875	4.61
16	1.45	2.14	2.584	2.982	3.471	3.814	4.536
17	1.437	2.117	2.555	2.948	3.425	3.762	4.473
18	1.425	2.098	2.531	2.919	3.385	3.716	4.416
19	1.415	2.08	2.509	2.893	3.352	3.677	4.367
20	1.405	2.065	2.489	2.869	3.323	3.641	4.321
21	1.396	2.05	2.47	2.847	3.296	3.609	4.279
22	1.388	2.037	2.454	2.827	3.273	3.582	4.243
23	1.381	2.025	2.438	2.809	3.252	3.558	4.21
24	1.374	2.014	2.424	2.793	3.232	3.536	4.18
25	1.368	2.004	2.412	2.778	3.214	3.516	4.153
26	1.362	1.994	2.399	2.763	3.196	3.496	4.128
27	1.356	1.984	2.388	2.749	3.18	3.478	4.104
28	1.352	1.976	2.378	2.737	3.165	3.462	4.084
29	1.347	1.969	2.368	2.726	3.152	3.447	4.065
30	1.343	1.962	2.359	2.715	3.139	3.433	4.048
31	1.339	1.955	2.35	2.705	3.127	3.419	4.031
32	1.335	1.948	2.342	2.695	3.115	3.406	4.016
33	1.331	1.942	2.335	2.686	3.105	3.394	4.001
34	1.328	1.937	2.327	2.677	3.094	3.383	3.988
35	1.324	1.931	2.32	2.669	3.085	3.372	3.974
36	1.321	1.926	2.314	2.661	3.075	3.362	3.962
37	1.318	1.921	2.308	2.654	3.067	3.352	3.951
38	1.315	1.916	2.302	2.647	3.058	3.343	3.939
39	1.313	1.912	2.296	2.64	3.05	3.334	3.929
40	1.31	1.908	2.291	2.634	3.043	3.326	3.918
41	1.307	1.903	2.285	2.628	3.035	3.317	3.908
42	1.305	1.899	2.281	2.622	3.029	3.31	3.899
43	1.303	1.896	2.276	2.617	3.022	3.303	3.891
44	1.301	1.892	2.272	2.611	3.016	3.296	3.882
45	1.299	1.889	2.267	2.606	3.009	3.288	3.873
46	1.296	1.885	2.263	2.601	3.003	3.282	3.865
47	1.294	1.882	2.259	2.596	2.998	3.276	3.858
48	1.293	1.879	2.255	2.592	2.992	3.27	3.85
49	1.291	1.876	2.251	2.587	2.987	3.264	3.843
50	1.289	1.873	2.248	2.583	2.982	3.258	3.837

Figure 29: Part 1 of corrected k table.

N	p = 0.75	p = 0.9	p = 0.95	p = 0.975	p = 0.99	p = 0.995	p = 0.999
51	1.288	1.871	2.244	2.579	2.977	3.253	3.83
52	1.286	1.868	2.241	2.574	2.972	3.247	3.823
53	1.284	1.865	2.237	2.571	2.967	3.242	3.817
54	1.283	1.863	2.234	2.567	2.963	3.237	3.811
55	1.281	1.86	2.231	2.563	2.958	3.232	3.805
56	1.28	1.858	2.228	2.56	2.955	3.228	3.8
57	1.278	1.856	2.226	2.557	2.951	3.223	3.794
58	1.277	1.854	2.223	2.553	2.947	3.219	3.789
59	1.276	1.851	2.22	2.55	2.943	3.215	3.784
60	1.275	1.849	2.217	2.547	2.939	3.21	3.779
61	1.273	1.847	2.215	2.544	2.936	3.207	3.774
62	1.272	1.845	2.212	2.541	2.932	3.203	3.769
63	1.271	1.843	2.21	2.538	2.929	3.199	3.765
64	1.27	1.841	2.207	2.535	2.925	3.195	3.76
65	1.269	1.84	2.205	2.532	2.922	3.191	3.755
66	1.268	1.838	2.203	2.53	2.919	3.188	3.752
67	1.267	1.836	2.201	2.527	2.916	3.184	3.747
68	1.266	1.835	2.199	2.525	2.913	3.181	3.744
69	1.265	1.833	2.197	2.522	2.91	3.178	3.739
70	1.264	1.831	2.195	2.52	2.907	3.175	3.736
71	1.263	1.83	2.193	2.517	2.904	3.172	3.732
72	1.262	1.828	2.191	2.515	2.901	3.169	3.728
73	1.261	1.827	2.189	2.513	2.899	3.166	3.725
74	1.26	1.825	2.187	2.511	2.896	3.163	3.721
75	1.259	1.824	2.185	2.508	2.893	3.16	3.718
76	1.258	1.822	2.184	2.507	2.891	3.158	3.715
77	1.257	1.821	2.182	2.505	2.889	3.155	3.711
78	1.257	1.82	2.18	2.503	2.886	3.152	3.708
79	1.256	1.818	2.179	2.501	2.884	3.15	3.705
80	1.255	1.817	2.177	2.499	2.882	3.147	3.702
81	1.254	1.816	2.175	2.497	2.88	3.145	3.699
82	1.253	1.815	2.174	2.495	2.877	3.142	3.696
83	1.253	1.813	2.172	2.493	2.875	3.139	3.693
84	1.252	1.812	2.171	2.492	2.873	3.138	3.69
85	1.251	1.811	2.169	2.49	2.871	3.135	3.687
86	1.251	1.81	2.168	2.488	2.869	3.133	3.685
87	1.25	1.809	2.167	2.487	2.867	3.131	3.682
88	1.249	1.808	2.165	2.485	2.865	3.128	3.679
89	1.249	1.807	2.164	2.483	2.864	3.127	3.677
90	1.248	1.805	2.162	2.481	2.861	3.124	3.674
91	1.247	1.805	2.161	2.48	2.86	3.122	3.672
92	1.247	1.804	2.16	2.478	2.858	3.12	3.669
93	1.246	1.802	2.158	2.477	2.856	3.118	3.667
94	1.246	1.802	2.157	2.476	2.854	3.116	3.665
95	1.245	1.801	2.156	2.474	2.853	3.114	3.663
96	1.244	1.8	2.155	2.473	2.851	3.112	3.66
97	1.244	1.799	2.154	2.471	2.849	3.111	3.658
98	1.243	1.798	2.153	2.47	2.848	3.109	3.656
99	1.243	1.797	2.151	2.469	2.846	3.107	3.654

Figure 30: Part 2 of corrected k table.

N	p = 0.75	p = 0.9	p = 0.95	p = 0.975	p = 0.99	p = 0.995	p = 0.999
100	1.242	1.796	2.151	2.468	2.845	3.106	3.652
102	1.241	1.794	2.148	2.465	2.841	3.102	3.648
104	1.24	1.793	2.146	2.463	2.839	3.099	3.644
106	1.239	1.791	2.144	2.46	2.836	3.096	3.64
108	1.238	1.79	2.142	2.458	2.833	3.093	3.636
110	1.237	1.788	2.14	2.456	2.83	3.09	3.633
112	1.237	1.787	2.138	2.453	2.828	3.087	3.629
114	1.236	1.785	2.137	2.451	2.825	3.084	3.626
116	1.235	1.784	2.135	2.449	2.823	3.081	3.622
118	1.234	1.782	2.133	2.447	2.82	3.079	3.619
120	1.233	1.781	2.131	2.445	2.818	3.076	3.616
122	1.232	1.78	2.13	2.443	2.816	3.074	3.613
124	1.232	1.779	2.128	2.441	2.813	3.071	3.61
126	1.231	1.777	2.127	2.439	2.811	3.068	3.607
128	1.23	1.776	2.125	2.438	2.809	3.066	3.604
130	1.23	1.775	2.124	2.436	2.807	3.064	3.602
132	1.229	1.774	2.123	2.434	2.805	3.062	3.599
134	1.228	1.773	2.121	2.433	2.803	3.06	3.596
136	1.227	1.772	2.12	2.431	2.801	3.057	3.593
138	1.227	1.771	2.118	2.429	2.799	3.055	3.591
140	1.226	1.77	2.117	2.428	2.798	3.054	3.589
142	1.226	1.769	2.116	2.426	2.796	3.051	3.586
144	1.225	1.768	2.115	2.425	2.794	3.05	3.584
146	1.224	1.767	2.113	2.423	2.792	3.047	3.581
148	1.224	1.766	2.112	2.422	2.791	3.046	3.58
150	1.223	1.765	2.111	2.421	2.789	3.044	3.578
152	1.223	1.764	2.11	2.42	2.788	3.042	3.575
154	1.222	1.763	2.109	2.418	2.786	3.041	3.573
156	1.222	1.763	2.108	2.417	2.785	3.039	3.571
158	1.221	1.762	2.107	2.416	2.783	3.037	3.569
160	1.221	1.761	2.106	2.415	2.782	3.036	3.567
162	1.22	1.76	2.105	2.413	2.78	3.034	3.565
164	1.22	1.759	2.104	2.412	2.779	3.033	3.564
166	1.219	1.758	2.103	2.411	2.778	3.031	3.562
168	1.219	1.758	2.102	2.41	2.776	3.03	3.56
170	1.218	1.757	2.101	2.409	2.775	3.028	3.558
172	1.218	1.756	2.1	2.408	2.774	3.027	3.556
174	1.218	1.756	2.099	2.407	2.772	3.025	3.555
176	1.217	1.755	2.098	2.405	2.771	3.024	3.553
178	1.217	1.754	2.097	2.405	2.77	3.022	3.551
180	1.216	1.753	2.097	2.404	2.769	3.021	3.55
185	1.215	1.752	2.094	2.401	2.766	3.018	3.546
190	1.214	1.75	2.093	2.399	2.763	3.015	3.542
195	1.214	1.749	2.091	2.397	2.76	3.012	3.539
200	1.213	1.747	2.089	2.394	2.758	3.009	3.535
205	1.212	1.746	2.087	2.393	2.756	3.007	3.532
210	1.211	1.745	2.086	2.39	2.753	3.004	3.529
215	1.21	1.743	2.084	2.388	2.751	3.001	3.525
220	1.209	1.742	2.082	2.387	2.748	2.999	3.522

Figure 31: Part 3 of corrected k table.

N	p = 0.75	p = 0.9	p = 0.95	p = 0.975	p = 0.99	p = 0.995	p = 0.999
225	1.209	1.741	2.081	2.385	2.746	2.996	3.52
230	1.208	1.74	2.08	2.383	2.744	2.994	3.517
235	1.207	1.739	2.078	2.381	2.742	2.992	3.514
240	1.207	1.738	2.077	2.38	2.74	2.99	3.512
245	1.206	1.737	2.075	2.378	2.739	2.988	3.509
250	1.205	1.735	2.074	2.377	2.737	2.986	3.507
255	1.205	1.735	2.073	2.375	2.735	2.984	3.505
260	1.204	1.734	2.072	2.374	2.733	2.982	3.502
265	1.204	1.733	2.071	2.373	2.732	2.98	3.5
270	1.203	1.732	2.069	2.371	2.73	2.978	3.498
275	1.203	1.731	2.068	2.37	2.729	2.977	3.496
280	1.202	1.73	2.067	2.369	2.727	2.975	3.494
285	1.202	1.729	2.066	2.368	2.726	2.974	3.492
290	1.201	1.729	2.065	2.366	2.724	2.972	3.49
295	1.201	1.728	2.064	2.365	2.723	2.971	3.488
300	1.2	1.727	2.063	2.364	2.722	2.969	3.487
310	1.199	1.725	2.061	2.362	2.719	2.966	3.483
320	1.199	1.724	2.06	2.36	2.717	2.963	3.48
330	1.198	1.723	2.058	2.358	2.714	2.961	3.477
340	1.197	1.722	2.057	2.356	2.712	2.958	3.474
350	1.196	1.72	2.055	2.354	2.71	2.956	3.471
360	1.196	1.719	2.054	2.353	2.708	2.954	3.468
370	1.195	1.718	2.052	2.351	2.706	2.952	3.465
380	1.194	1.717	2.051	2.35	2.704	2.95	3.463
390	1.194	1.716	2.05	2.348	2.703	2.948	3.461
400	1.193	1.715	2.049	2.347	2.701	2.946	3.458
425	1.192	1.713	2.046	2.343	2.697	2.941	3.453
450	1.191	1.711	2.043	2.34	2.693	2.937	3.448
475	1.189	1.709	2.041	2.337	2.69	2.934	3.444
500	1.188	1.707	2.039	2.335	2.687	2.93	3.44
525	1.187	1.706	2.037	2.332	2.684	2.927	3.436
550	1.187	1.704	2.035	2.33	2.681	2.924	3.432
575	1.186	1.703	2.033	2.328	2.679	2.922	3.429
600	1.185	1.702	2.031	2.326	2.677	2.919	3.426
625	1.184	1.7	2.03	2.324	2.674	2.916	3.423
650	1.183	1.699	2.028	2.323	2.672	2.914	3.42
700	1.182	1.697	2.026	2.32	2.669	2.91	3.415
750	1.181	1.695	2.024	2.317	2.665	2.907	3.411
800	1.18	1.694	2.021	2.314	2.662	2.903	3.407
850	1.179	1.692	2.02	2.312	2.66	2.9	3.403
900	1.178	1.691	2.018	2.31	2.657	2.898	3.4
950	1.177	1.689	2.016	2.308	2.655	2.895	3.397
1000	1.177	1.688	2.015	2.306	2.653	2.893	3.394
1500	1.172	1.68	2.004	2.294	2.638	2.876	3.374

Figure 32: Part 4 of corrected k table.