Copenhagen Business School
Center for Statistics, August 2010.

Tue Tjur:

Inequalities related to the Stringer bound

Manuscript for retirement lecture August 2010

1. Introduction to the problem.

A more informative title for the present paper could be "Exact lower confidence bounds for the common expectation of a set of independent nonnegative random variables"; for independent, random variables $X_1, \ldots, X_n$, nonnegative and with common, unknown mean value $\mathrm{E}X_1 = \cdots = \mathrm{E}X_n = \mu$, we are going to study the problem of constructing lower confidence bounds for μ.

Notice that we make no further distributional assumptions, not even that the variables are identically distributed. Thus, even though the whole idea is to make inference about the "parameter" μ, the problem is what statisticians usually refer to as "non–parametric".

Notice also that we are not proposing to construct confidence *intervals* for μ. It makes no sense, under these conditions, to talk about *upper* confidence bounds for the mean value, because even if we knew that the variables were identically distributed and observed (most extremely) $X_1 = X_2 = \cdots = X_n = 0$, this would only allow us to infer that the common distribution must have so and so much of its mass concentrated at zero. The distribution of the (even so small) remainder of the mass, which could result in arbitrarily large values of μ, would still be entirely out of control. As you will see, if you try to change the following definition of a lower confidence bound to a definition of an upper confidence bound, there does not exist such a thing.

DEFINITION. For $0 < \alpha < 1$, a $1-\alpha$ *lower confidence bound for* μ is a function $\underline{\mu}(x_1, \ldots, x_n)$ of n non–negative variables, which is

(1) *increasing in any of its arguments* (when the remaining arguments are kept fixed),

(2) *symmetric* (i.e. $\underline{\mu}(x_1, \ldots, x_n) = \underline{\mu}(x_{p(1)}, \ldots, x_{p(n)})$ for any permutation p) ,

(3) *homogeneous* (i.e. $\underline{\mu}(\lambda x_1, \ldots, \lambda x_n) = \lambda \underline{\mu}(x_1, \ldots, x_n)$ for $\lambda \geq 0$,

and fulfils the essential condition that

$$P\left(\underline{\mu}(X_1, \ldots, X_n) \leq \mu\right) \geq 1 - \alpha\,. \tag{4}$$

for any independent nonnegative random variables $X_1, \ldots, X_n$ with common expectation $\mathrm{E}X_1 = \cdots = \mathrm{E}X_n = \mu$.

It can be proved that a function $\underline{\mu}$ satisfying these conditions has some nice properties. It is not always continuous (though it will be so in all of our examples), but it is continuous at all inner points of its domain. Similarly, its restriction to subsets of the boundary where some of the coordinates are set to zero, while the remaining coordinates are kept positive, are continuous. Moreover, $\underline{\mu}$ turns out to be *lower semicontinuous,* which implies that sets of the form $A(\mu_0)$ below are closed.

Let $\underline{\mu}$ be such a lower $1-\alpha$ confidence bound. For $\mu_0 > 0$, define a subset $A(\mu_0)$ of $[0,+\infty)^n$ by

$$A(\mu_0) = \{(x_1,\ldots,x_n) \,|\, \underline{\mu}(x_1,\ldots,x_n) \leq \mu_0\}\,.$$

This is the set of observation vectors for which we would accept the hypothesis $\mu \leq \mu_0$ on level α by the one–sided statistical test constructed from the confidence bound in the standard manner. We shall refer briefly to $A = A(1)$ as the *(unit) acceptance region* associated with $\underline{\mu}$. Notice that we have $A(\mu_0) = \mu_0 A(1)$, due to the homogeneity of $\underline{\mu}$

The unit acceptance region A for a lower bound $\underline{\mu}$ is seen to be

star–shaped around the origin ($(x_1,\ldots,x_n) \in A$ implies $(\lambda x_1,\ldots,\lambda x_n) \in A$ for $0 \leq \lambda < 1$, follows from the homogeneity of $\underline{\mu}$),

symmetric (invariant under coordinate permutations, follows from the symmetry of $\underline{\mu}$),

downwards saturated (meaning that if a point lies in A, and we decrease some of its coordinates, then we still have a point in A), and

closed.

Conversely, let A be a closed subset of $[0,+\infty)^n$ which is star–shaped around the origin, symmetric and downwards saturated, and furthermore fulfils the essential condition that for any n independent, nonnegative random variables $X_1,\ldots,X_n$ with expectation 1, $P((X_1,\ldots,X_n) \in A) \geq 1-\alpha$. Define a function $\underline{\mu}$ of n non–negative arguments by

$$\underline{\mu}(x_1,\ldots,x_n) = \inf\{m | (x_1,\ldots,x_n) \in mA\}\,.$$

Then, $\underline{\mu}(X_1,\ldots,X_n)$ is a lower $1-\alpha$ confidence bound for the common expected value of $X_1,\ldots,X_n$, and A is the unit acceptance region associated with $\underline{\mu}$.

In this sense, we have a one–to–one correspondence between lower confidence bounds and their acceptance regions. In many cases, the concept of an acceptance region is much simpler to visualise than that of a confidence bound. In the constructions that we make in the following, we will usually construct the confidence bounds via their acceptance regions. Notice that *sharp* bounds — i.e. a bounds with the desirable property that they somehow take as large values as possible — correspond to

small acceptance regions. Whereas very conservative bounds (which, in this context, means bounds that are unnecessarily small) corespond to large acceptance regions. In particular, the property that a bound is uniformely dominated by another corresponds to the property that the acceptance region of the first contains the acceptance region of the second,

$$\underline{\mu}_1 \leq \underline{\mu}_2 \Leftrightarrow A_1 \supseteq A_2 \,.$$

2. The case $n = 1$.

Let X be a non–negative random variable with expectation μ. According to Markov's inequality, we have $P(X \leq \mu/\alpha) \geq 1 - \alpha$, or $P(\alpha X \leq \mu) \geq 1 - \alpha$, which means that the function

$$\underline{\mu}(x) = \alpha x$$

is a $1 - \alpha$ lower confidence bound for μ. The corresponding unit acceptance region is the interval $A = [0, 1/\alpha]$.

This bound is uniformely better than any other bound. Indeed, take another valid lower confidence bound $\underline{\mu}'$ with acceptance region A'. Since A' is closed and star–shaped around the origin, we must have $A' = [0, c]$ for some c. Suppose that A' does not contain all points in $A = [0, 1/\alpha]$. This means that $c < 1/\alpha$. For $c < a < 1/\alpha$, suppose that the distribution of X is concentrated on the set $\{0, a\}$, with $P(X = a) = 1/a$ and $P(X = 0) = 1 - 1/a$. Then we have $P(X \in A') = 1 - 1/a < 1 - \alpha$, in contradiction with the assumption that $\underline{\mu}'$ was a valid confidence bound.

Thus, the case $n = 1$ is trivial. There is essentially only one way of specifying a lower confidence bound for the expectation of a nonnegative random variable with unknown distribution, namely the one given by Markov's inequality. For example, having observed $X = x$, we can say with 95% confidence that $\mathrm{E}X \geq 0.05x$, or with 99% confidence that $\mathrm{E}X \geq 0.01x$.

3. The case $n = 2$.

A general way of establishing lower bounds in the case $n = 2$ (which is easily generalized to the case $n > 2$) is to write down some symmetric, non–negative function $f(x_1, x_2)$ of two nonnegative variables, which has the property

(E) $\quad \mathrm{E}f(X_1, X_2) = 1$ for X_1 and X_2 independent, $\mathrm{E}X_1 = \mathrm{E}X_2 = 1$.

It then follows from Markov's inequality that

$$P\left(f(X_1, X_2) \leq \frac{1}{\alpha}\right) \geq 1 - \alpha$$

which means that the set

$$A = \{(x_1, x_2) | f(x_1, x_2) \leq \frac{1}{\alpha}\}$$

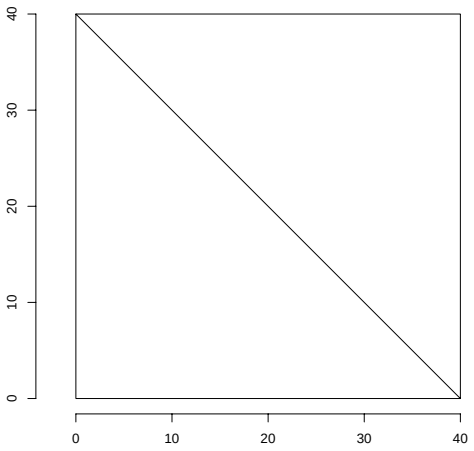
fulfils the essential condition (4) for being an acceptance region for some lower bound $\underline{\mu}$.

In the following examples, we follow the convention that f, μ, A etc. are equipped with the number of the example as subscript.

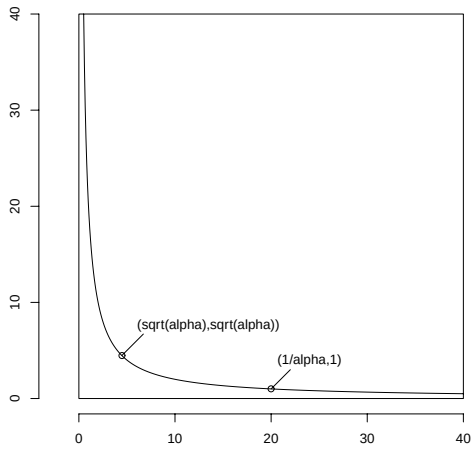
EXAMPLE 1. Let $f_1(x_1, x_2) = (x_1 + x_2)/2$. A_1 becomes the triangle spanned by the points $(0, 0)$, $(0, \frac{2}{\alpha})$ and $(\frac{2}{\alpha}, 0)$. This set is easily seen to fulfil the remaining conditions for being an acceptance region. The corresponding lower bound is

$$\underline{\mu}_1(x_1, x_2) = \frac{\alpha}{2}(x_1 + x_2)\,.$$

This bound is even valid without the assumption of independence. More interesting is, perhaps, the following bound which is a consequence of the multiplication rule for expectations of independent variables.



A_1 for $\alpha = 0.05$



A_2 for $\alpha = 0.05$

EXAMPLE 2. For $f_2(x_1, x_2) = x_1 x_2$, A_2 becomes the set bounded by the hyperbola $x_1 x_2 = 1/\alpha$. The lower bound is

$$\underline{\mu}_2(x_1, x_2) = \sqrt{\alpha x_1 x_2}$$

For example, if we observe $X_1 = 50$ and $X_2 = 10$, we can say with 95% confidence that the common expected value is at least

$$\underline{\mu}_2(50, 10) = \sqrt{0.05 \times 50 \times 10} = 5$$

whereas the bound of example 1 would give

$$\underline{\mu}_1(50, 10) = \frac{0.05}{2}(50 + 10) = 1.5\,.$$

Thus, $\underline{\mu}_2$ is sharper than $\underline{\mu}_1$ in this case; but as we can see (since none of the two acceptance regions is a subset of the other) it is not true that $\underline{\mu}_2$ is *uniformely* sharper than $\underline{\mu}_1$.

A relevant question in this context is whether there is a general way of describing the non–negative functions $f(x_1, x_2)$ with the property (E). A partial answer follows here:

Proposition 1. *Any function with the property* (E) *has the form*

$$f(x_1, x_2) = x_1x_2 + (1 - x_1)h(x_2) + (1 - x_2)h(x_1)$$

for some function $h(x)$ of a single non–negative variable.

For example, the function f_1 comes out for $h(x) = x/2$, and f_2 comes out for $h(x) = 0$. The trivial case $f(x_1, x_2) = 1$ (which is not useful for anything) is obtained for $h(x) = (1 + x)/2$.

The proof of the proposition will not be given here. It is relatively tedious, but elementary. The idea is to consider the situation where both distributions are concentrated on two points, and then make use of the relations coming out of this. The key lemma is that if two symmetric functions f and f' satisfy the condition (E) and $f(0, x) = f'(0, x)$ for all x, then the two functions must be identical.

It is obvious that the converse proposition is also true, in the sense that any non–negative function $f(x_1, x_2)$ of this form has the property (E), provided that the expectations involved are defined. The difficulty with the characterization is that it is not always so easy to tell when a given function h results in a function f which is non–negative and determines a set A that fulfils the remaining conditions for acceptance regions.

In all the cases that we are considering seriously here, we have $f(0, 0) = 0$, from which it follows that $h(0) = 0$ and $f(0, x) = h(x)$. Thus, h is simply the restriction of f to one of the coordinate axes.

It turns out to be a good idea to keep $h(x) = 0$ for $x \leq \frac{1}{\alpha}$. This implies that the intersection of A with the square $[0, \frac{1}{\alpha}] \times [0, \frac{1}{\alpha}]$ is bounded by the hyperbola $x_1x_2 = \frac{1}{\alpha}$, as in example 2. This gives the sharpest bound obtainable for distributions that are concentrated on $[0, \frac{1}{\alpha}]$, as we shall see later.

Example 3. Setting

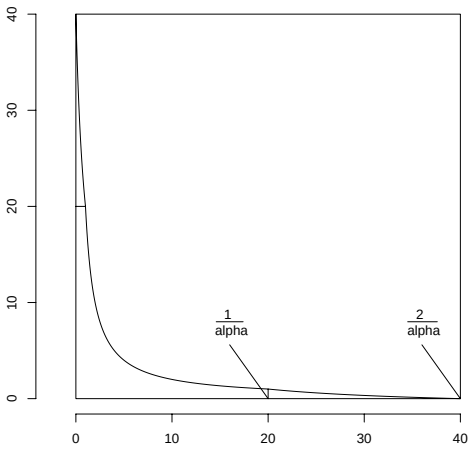
$$h(x) = \begin{cases} 0 \text{ for } x \leq \frac{1}{\alpha} \\ \frac{x}{2} \text{ for } x > \frac{1}{\alpha} \end{cases}$$

gives a bound which is quite close to the best we can do. The function f_3 becomes

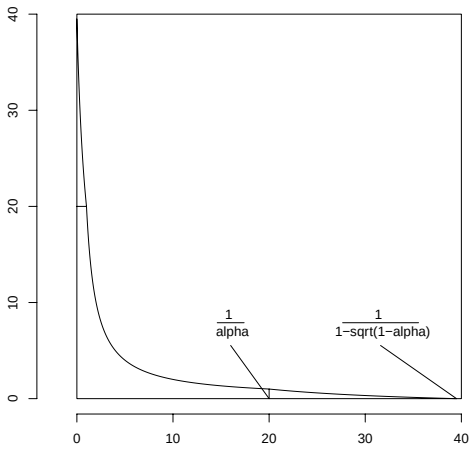
$$f_3(x_1,x_2)=\begin{cases} x_1x_2 \text{ for } x_1,x_2\le\frac{1}{\alpha} \\ \frac{x_1(1+x_2)}{2} \text{ for } x_2\le\frac{1}{\alpha}<x_1 \\ \frac{x_2(1+x_1)}{2} \text{ for } x_1\le\frac{1}{\alpha}<x_2 \\ \frac{x_1+x_1}{2} \text{ for } \frac{1}{\alpha}<x_1,x_2 \end{cases}$$

A_3 is the domain bounded by the three hyperbolae $x_1x_2=\frac{1}{\alpha}$, $x_1(1+x_2)=\frac{2}{\alpha}$ and $x_2(1+x_1)=\frac{2}{\alpha}$. The corresponding lower bound becomes

$$\underline{\mu}_3(x_1,x_2)=\begin{cases} \frac{\alpha x_2}{4}\left(1+\sqrt{1+\frac{8x_1}{\alpha x_2}}\right) \text{ for } \frac{x_1}{x_2}<\alpha \\ \sqrt{\alpha x_1x_2} \text{ for } \alpha\le\frac{x_1}{x_2}\le\frac{1}{\alpha} \\ \frac{\alpha x_1}{4}\left(1+\sqrt{1+\frac{8x_2}{\alpha x_1}}\right) \text{ for } \frac{x_1}{x_2}>\frac{1}{\alpha}\,. \end{cases}$$



A_3 for $\alpha=0.05$



A_4 for $\alpha=0.05$

EXAMPLE 4. An even sharper bound, which seems to the best that can be obtained in this way, comes out if we choose $h(x)$ as large as possible under the constraint that it should equal $\frac{1}{\alpha}$ for diagonal points (x,x) when $x>\frac{1}{\alpha}$. This gives us the identity

$$x^2+2(1-x)h(x)=\frac{1}{\alpha}$$

from which h is determined as

$$h(x)=\begin{cases} 0 \text{ for } x\le\frac{1}{\alpha} \\ \frac{x^2-\frac{1}{\alpha}}{2(x-1)} \text{ for } x>\frac{1}{\alpha}\,. \end{cases}$$

It turns out that this results in a non–negative function $f_4(x_1, x_2)$. The resulting acceptance region is extremely close to the region A_{opt} introduced later in this section as the best choice. So close, in fact, that if you draw the two regions in the same graph with $\alpha = 0.05$ or $\alpha = 0.1$, it is impossible to tell the difference. We have no explicit formula for $\underline{\mu}_4$, the computations involve the solution of a third degree equation.

A small technical remark is needed here. When A is constructed as the set of points for which $f(x_1, x_2) \leq \frac{1}{\alpha}$ in this case, it will actually include the infinite piece of the diagonal where $f(x, x) = \frac{1}{\alpha}$. But a simple limiting argument shows that the property $P((X_1, X_2) \in A) \geq 1 - \alpha$ also holds when A is constructed as the closure of the set given by the sharp inequality $f(x_1, x_2) < \frac{1}{\alpha}$, and this is the way it should be done here.

Notice (as can be seen from the figures, and also proved explicitely without too much work) that A_4 is contained in all other acceptance regions considered until now. Notice also, that A_4 is contained in the square $[0, \frac{1}{1-\sqrt{1-\alpha}}]^2$, since $\frac{1}{1-\sqrt{1-\alpha}}$ is the solution to the equation $h(x) = 0$. The importance of this number is further stressed by the following simpler example.

EXAMPLE 5. It is easy to show directly that $[0, \frac{1}{1-\sqrt{1-\alpha}}]^2$ is the smallest square that has the properties of a unit acceptance region. Indeed, for independent non–negative random variables with expectation 1 and a given constant $c > 0$ we have by Markov's inequality (with equality in the case where the two distributions are concentrated on $\{0, c\}$)

$$P\left(X_1 \leq c \text{ and } X_2 \leq c\right) = P\left(X_1 \leq c\right) \times P\left(X_2 \leq c\right) \geq \left(1 - \frac{1}{c}\right)^2$$

where the last quantity equals $1 - \alpha$ if and only if $c = \frac{1}{1-\sqrt{1-\alpha}}$. Hence,

$$A_5 = \left[0, \frac{1}{1 - \sqrt{1 - \alpha}}\right]^2$$

is a unit acceptance region. The corresponding lower bound becomes

$$\underline{\mu}_5(x_1, x_2) = (1 - \sqrt{1 - \alpha}) \max\{x_1, x_2\} .$$

EXAMPLE 6. This result can also be derived in a very simple way by the "$h(x)$–method", for example by defining

$$h(x) = \begin{cases} 0 \text{ for } x \leq \frac{1}{1-\sqrt{1-\alpha}} \\ \frac{1}{\alpha} \text{ for } x > \frac{1}{1-\sqrt{1-\alpha}} \end{cases}$$

resulting in an acceptance region A_6 which becomes the intersection of the "sub–hyperbola" of example 2 with the square A_5.

4. An optimal solution for $n = 2$.

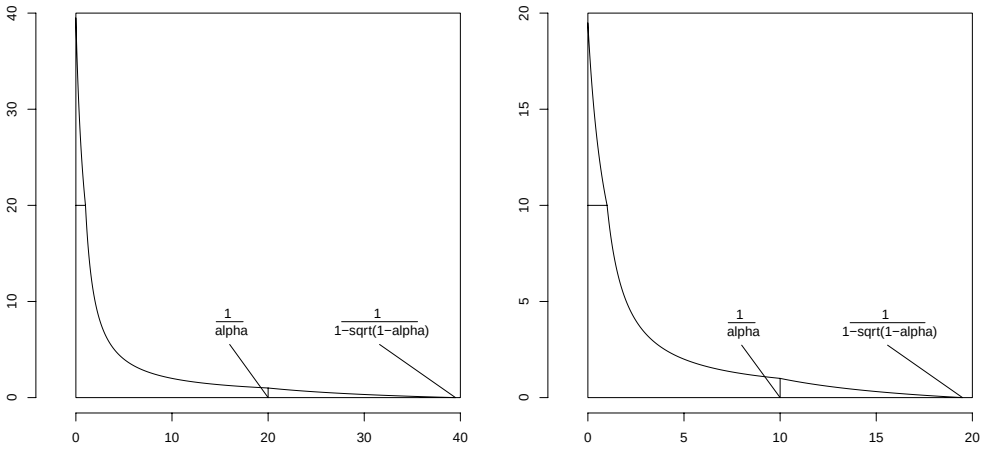
The last examples justify, to some extent, that we focus on acceptance regions that are subsets of the square $[0, \frac{1}{1-\sqrt{1-\alpha}}]^2$. And so does the following result, which roughly states that under this restriction there is a unique, uniformely sharpest lower bound:

Proposition 2. *The set*

$$A_{\text{opt}} = \left\{ (x_1, x_2) \;\middle|\; x_1, x_2 \geq 0 \text{ and } x_1 x_2 \leq \frac{1}{\alpha} \right.$$

$$\left. \text{and} \quad \frac{1 + x_1\sqrt{1-\alpha}}{1+\sqrt{1-\alpha}} x_2 \leq \frac{1}{\alpha} \quad \text{and} \quad \frac{1 + x_2\sqrt{1-\alpha}}{1+\sqrt{1-\alpha}} x_1 \leq \frac{1}{\alpha} \right\}$$

is the unit acceptance region for a lower bound. This set is minimal, in the sense that no proper subset of it is a unit acceptance region. Moreover, for any other unit acceptance region A which is a subset of $[0, \frac{1}{1-\sqrt{1-\alpha}}]^2$, we have $A \supseteq A_{\text{opt}}$.



A_{opt} for $\alpha = 0.05$ $\qquad$ A_{opt} for $\alpha = 0.10$

Remarks. The figure shows the set A_{opt} for $\alpha = 0.05$ and (to make the breakpoints of the bounding curve a little more visible) for $\alpha = 0.10$.

The formula for the corresponding lower bound is

$$\underline{\mu}_{\text{opt}}(x_1, x_2) = \begin{cases} \frac{(1-\sqrt{1-\alpha})x_2}{2} \left(1 + \sqrt{1 + \frac{4x_1\sqrt{1-\alpha}}{(1-\sqrt{1-\alpha})x_2}}\right) & \text{for } \frac{x_1}{x_2} < \alpha \\ \sqrt{\alpha x_1 x_2} \text{ for } \alpha \leq \frac{x_1}{x_2} \leq \frac{1}{\alpha} \\ \frac{(1-\sqrt{1-\alpha})x_1}{2} \left(1 + \sqrt{1 + \frac{4x_2\sqrt{1-\alpha}}{(1-\sqrt{1-\alpha})x_1}}\right) & \text{for } \frac{x_1}{x_2} > \frac{1}{\alpha}\,. \end{cases}$$

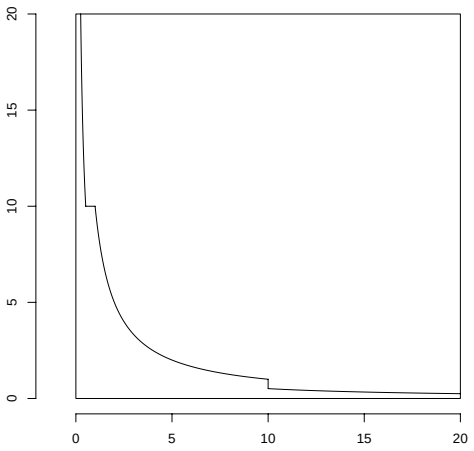
A surprisingly complicated answer to a relatively simple question ...

It would have been a lot nicer to come up with a result stating that there is a unique, uniformely best lower bound. However, this seems not

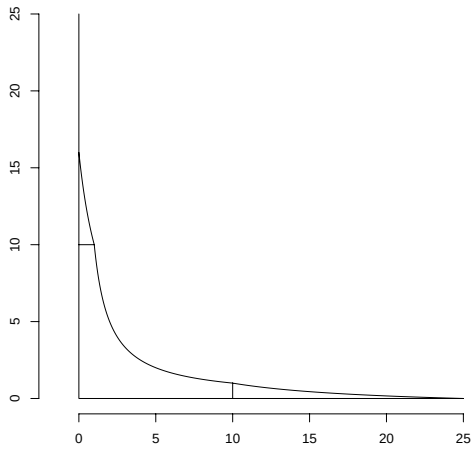
to be the case in any obvious sense. As will be obvious from the proof below, the part of the region that is contained in the square given by $x_1, x_2 \leq \frac{1}{\alpha}$ must always be chosen as the set bounded by the hyperbola $x_1 x_2 = \frac{1}{\alpha}$, but the shape of the tails can vary. For example, the set with tails bounded by the hyperbolae

$$x_1 = \frac{1}{\alpha} \times \frac{1 - \frac{1}{1+\sqrt{1+\alpha}}}{x_2 - \frac{1}{1+\sqrt{1+\alpha}}} \quad \text{and} \quad x_2 = \frac{1}{\alpha} \times \frac{1 - \frac{1}{1+\sqrt{1+\alpha}}}{x_1 - \frac{1}{1+\sqrt{1+\alpha}}}$$

is also a minimal acceptance region in the sense of the proposition. A very ugly one (see the figure below), unbounded and with some obscure breakpoints, and even with the very questionable property that the corresponding lower bound becomes zero if one of the two observations is zero; but, nevertheless, a minimal acceptance region. If we relax our conditions in such a way that acceptance regions are allowed to be assymmetric, many other solutions appear. The second figure below shows such an example, which is algebraically quite identical to the set A_{opt} of the propositon, except that the endpoints $\frac{1}{1-\sqrt{1-\alpha}}$ of the line segments on the axes are replaced with two *different* values B_1 and B_2 with the property that $\left(1 - \frac{1}{B_1}\right)\left(1 - \frac{1}{B_2}\right) = 1 - \alpha$. If we let B_1 tend to $+\infty$, B_2 will tend to $\frac{1}{\alpha}$, and in the limit we get an acceptance region which is A_2 (the "subhyperbola" of example 2) with one of its tails truncated at $\frac{1}{\alpha}$; which is also a minimal acceptance region, in this extended sense.



Another acc. reg. for $\alpha = 0.1$



A non–symm. acc. reg. for $\alpha = 0.1$

The justification of the choice of A_{opt} as the "optimal" acceptance region for $n = 2$ lies in the property that it is contained in the square $[0, \frac{1}{1-\sqrt{1-\alpha}}]^2$. As we shall see in the next section, this property is much more fundamental than explained here. Roughly, it means that the corresponding bound is the optimal one in the special case where the

variables are identically distributed rescaled Bernoulli variables (i.e. variables with two possible values, 0 and some $b > 1$).

The fact that A_{opt} is an acceptance region can also be stated as follow. Suppose that X_1 and X_2 are non–negative independent random variables with $\mathrm{E}X_1 = \mathrm{E}X_2 = 1$. Then the three inequalities

$$X_1 X_2 \leq \frac{1}{\alpha}, \quad \frac{1 + X_1\sqrt{1-\alpha}}{1+\sqrt{1-\alpha}} X_2 \leq \frac{1}{\alpha} \quad \text{and} \quad \frac{1 + X_2\sqrt{1-\alpha}}{1+\sqrt{1-\alpha}} X_1 \leq \frac{1}{\alpha}$$

hold *simultanously* with probability $\geq 1-\alpha$. This is remarkable, because it follows immediately from Markov's inequality that the inequalities hold *one by one* with probability $\geq 1 - \alpha$. How often does it happen, that the usual confusion around these two concepts of confidence doesn't matter?

However, Markov's inequality seems to be of no use at all for the proof of proposition 2. The set A_{opt} can not be constructed by the "$h(x)$–method". Our proof relies on a quite different technique, as indicated by the following lemma. We state and prove the lemma for $n = 2$ because this is where we need it, but it is actually valid for arbitrary n, with only trivial modifications of the proof.

Definition. By a two point distribution we mean a distribution which has all its mass on a set consisting of two points.

Notice that for real numbers a and b such that $0 \leq a < 1 < b$ there exists a unique two point distribution concentrated on $\{a, b\}$ with mean value 1. The point probabilities are given by

$$P(X = a) = \frac{b-1}{b-a} \quad \text{and} \quad P(X = b) = \frac{1-a}{b-a}\,.$$

Lemma (essentially equivalent to lemma 1 in Samuels 1969). *Let A be a subset of $[0, +\infty)^2$ which is star–shaped around the origin, symmetric, downwards saturated and closed. For A to be an acceptance region, it suffices that the essential condition*

$$P((X_1, X_2) \in A) \geq 1 - \alpha \text{ for } X_1, X_2 \text{ independent, } \mathrm{E}X_1 = \mathrm{E}X_2 = 1$$

is satisfied in the special case where both variables X_1 and X_2 have two point distributions.

Proof. Consider the class $\mathcal{A}$ of pairs (π_1, π_2) of probability distributions on $[0, +\infty)$ for which $(\pi_1 \otimes \pi_2)(A) \geq 1 - \alpha$. Notice that if (π_1', π_2) and (π_1'', π_2) belong to $\mathcal{A}$, then for any $\lambda \in [0, 1]$ we have also $(\lambda\pi_1' + (1 - \lambda)\pi_1'', \pi_2) \in \mathcal{A}$. In words, $\mathcal{A}$ is closed under the formation of convex combinations of the first factor π_1. In the same way, we see that it is closed under the formation of convex combinations of its second factor

π_2. By repeated use of this, we see that $\mathcal{A}$ must contain all pairs of probability measures that are convex combinations of finitely many two point distributions with mean value 1. But the set of such distributions is simply the set of all discrete distributions with mean value 1. The proof of this (which is quite elementary, use induction by the number of atoms) is left to the reader. Thus, $\mathcal{A}$ contains all pairs of discrete distributions with mean values 1. But since the set of discrete distributions with mean value 1 is dense in the set of all distributions with mean value 1 in the weak topology, and since the set of pairs (π_1, π_2) for which $(\pi_1 \otimes \pi_2)(A) \geq 1 - \alpha$ is weakly closed (or, to be more precise, closed in the product of the weak topologies), the lemma follows.

Proof of proposition 2. We construct $A = A_{\text{opt}}$ step by step, adding "bricks" to it in such a way that the essential condition $P((X_1, X_2) \in A) \geq 1 - \alpha$ is met for more and more possible pairs of two point distributions. In each step, the construction is made in such a way that all points added to A are necessary, in the sense that without these points A would not satisfy the conditions for a unit acceptance region.

Our notation is given by

$$P(X_i = a_i) = \frac{b_i - 1}{b_i - a_i} \quad \text{and} \quad P(X_i = b_i) = \frac{1 - a_i}{b_i - a_i}\,, i = 1, 2\,.$$

Now, notice first of all that if there exists a pair of two point distributions for which one of the four points (a_1, a_2), (a_1, b_2), (b_1, a_2) and (b_1, b_2) has mass $> \alpha$, then this point must be an element of *any* $1 - \alpha$ acceptance region A, because if it was not there would obviously be probability $< 1 - \alpha$ left for A. From this we can conclude the following.

(1) For any $a_1, a_2 \in [0, 1)$, we can always, by choosing b_1 and b_2 sufficiently large, obtain

$$P((X_1, X_2) = (a_1, a_2)) = \frac{b_1 - 1}{b_1 - a_1} \frac{b_2 - 1}{b_2 - a_2} > \alpha$$

From this it follows that the "half open" unit square $[0, 1)^2$ must be contained in A. The missing points on the boundary must also be included, because A should be closed. Thus, we must have $[0, 1]^2 \subseteq A$.

(2) For any $(a_1, b_2) \in [0, 1) \times (1, \frac{1}{\alpha})$ we can, by choosing $a_2 = 0$ and b_1 sufficiently large, obtain

$$P((X_1, X_2) = (a_1, b_2)) = \frac{b_1 - 1}{b_1 - a_1} \frac{1 - a_2}{b_2 - a_2} > \alpha$$

Thus, $[0, 1) \times (1, \frac{1}{\alpha}) \subseteq A$, and since A should be closed and symmetric, we conclude that

$$\left([0, 1] \times \left[1, \frac{1}{\alpha}\right]\right) \cup \left(\left[1, \frac{1}{\alpha}\right] \times [0, 1]\right) \subseteq A\,.$$

We have now showed that all points in the square $[0, \frac{1}{\alpha}]^2$ with at least one coordinate ≤ 1 must belong to A.

(3) Now consider a point $(b_1, b_2) \in (1, \frac{1}{\alpha}]^2$. The maximal probability mass concentrated at this point is $\frac{1}{b_1}\frac{1}{b_2}$, which is obtained for $a_1 = a_2 = 0$. From this we conclude that we *must* include in A the points (b_1, b_2) for which $b_1 b_2 \leq \frac{1}{\alpha}$. Conversely, for points over/to the right of this hyperbola (still restricting to the case where both b_1 and b_2 are $\leq \frac{1}{\alpha}$, we see that the three other points (a_1, a_2), (a_1, b_2) and (b_1, a_2) belong to the "bricks" that were already included under (1) and (2), and since the point mass at (b_1, b_2) is $< \alpha$ here, we can safely decide already that these points should *not* be included in A.

In summary, we have now taken care of all points in the square $[0, \frac{1}{\alpha}]^2$. We have shown that A must necessarily include the closed set bounded by the hyperbola $x_1 x_2 = \frac{1}{\alpha}$, and also that all points outside this set can remain outside.

We now turn to the case where either b_1 or b_2 or both exceed $\frac{1}{\alpha}$.

(4) First, consider the case where $a_1 = a_2 = 0$ and $b_1 = b_2 = b$. Here, we see that $P((X_1, X_2) = (0, 0)) \geq 1 - \alpha$ if and only if $b \geq \frac{1}{1-\sqrt{1-\alpha}}$. Thus, for $b < \frac{1}{1-\sqrt{1-\alpha}}$ we are forced to include the points $(0, b)$ and $(b, 0)$ (we can not just include one of them, due to the symmetry condition, and it would not help to include (b, b) because that would force us to include the two other points anyway, due to the requirement that A should be downwards saturated). Since the closedness condition implies that the limiting value $b = \frac{1}{1-\sqrt{1-\alpha}}$ must also be included, we have shown that any acceptance region must necessarily contain the set

$$\left(\left[\frac{1}{\alpha}, \frac{1}{1-\sqrt{1-\alpha}}\right] \times \{0\}\right) \cup \left(\{0\} \times \left[\frac{1}{\alpha}, \frac{1}{1-\sqrt{1-\alpha}}\right]\right)$$

consisting of two line segments on the coordinate axes.

(5) We now consider the case where one of the coordinates b_1 and b_2 — let us say b_1 — exceeds $\frac{1}{1-\sqrt{1-\alpha}}$. But to simplify things, we begin with the case $a_2 = 0$. Thus, (b_1, a_2) is a point on the x_1–axes to the right of $(\frac{1}{1-\sqrt{1-\alpha}}, 0)$ which means that neither this point nor the point (b_1, b_2) are potential elements of A. Since (a_1, a_2) is a born element of A (as belonging to "brick" (1)), the question is whether the fourth point (a_1, b_2) should be included in A. Which it should, in the first place, if the mass already in A is less than $1 - \alpha$. That is if

$$\frac{b_1 - 1}{b_1 - a_1}\frac{b_2 - 1}{b_2 - 0} < 1 - \alpha$$

Or, to be more precise: The point (a_1, b_2) should be included in A if this relation holds for *some* value of b_1. Since the left hand side is an

increasing function of b_1, this means that we should include (a_1, b_2) in A if the left hand side becomes smaller than $1-\alpha$ when b_1 comes close enough to $\frac{1}{1-\sqrt{1-\alpha}}$. And since A should be closed, we may as well include the limiting case from the beginning, which means that (a_1, b_2) should be included if

$$\frac{\frac{1}{1-\sqrt{1-\alpha}} - 1}{\frac{1}{1-\sqrt{1-\alpha}} - a_1} \frac{b_2 - 1}{b_2 - 0} \leq 1 - \alpha$$

Now, the left hand side here is increasing, both as a function of a_1 and as a function of b_2. This means, that the area we are trying to identify, the subset of $[0,1] \times \left[\frac{1}{\alpha}, \frac{1}{1-\sqrt{1-\alpha}}\right]$ consisting of the points (a_1, b_2) that must be included in A, is bounded (upwards/to the right) by the curve given by the equation

$$\frac{\frac{1}{1-\sqrt{1-\alpha}} - 1}{\frac{1}{1-\sqrt{1-\alpha}} - a_1} \frac{b_2 - 1}{b_2} = 1 - \alpha$$

which is "easily" rewritten to

$$\frac{1 + a_1\sqrt{1-\alpha}}{1 + \sqrt{1-\alpha}} b_2 = \frac{1}{\alpha}\,.$$

I have put "easily" in quotes because it is not easy at all, but in principle it is elementary, and I am not going to use a page for it here.

This (and its symmetric twin, of course) was the last brick. We have now included all points af the set A_{opt} described by the proposition, and we have seen that it is necessary to include these points. It remains to show that the cases not yet considered can be taken care of without further extension of A. In particular,

(6) Now consider the case where both a_1 and a_2 vary freely in the unit interval, and where (again) b_1 and b_2 are $> \frac{1}{\alpha}$. Suppose we can prove that A is star–shaped around (1,1). In that case, we can proceed as follows. Suppose (without loss of generality) that $a_2 \leq a_1$, and put $\lambda = \frac{1}{1-a_2}$. From the pair of random variables (X_1, X_2) with the prescribed distribution, construct a new pair by

$$(X_1', X_2') = (1 + \lambda(X_1 - 1), 1 + \lambda(X_2 - 1)\,.$$

Obviously, this results in a new pair of independent, two point distributed non–negative random variables with mean value 1, but now one of them $(1 + \lambda(X_2 - 1))$ has an atom at 0, which means that we are in the case covered by (5) above. The starhapedness implies that an atom can be "pushed out of A" by this transformation, but not the other way around. Since the point probabilities are unchanged, we conclude

that the sum of the point masses for points originally in A is at least $1-\alpha$.

It remains to prove that A is star–shaped around (1,1). If we look at the figure and use our geometric intution (or, if we really go into details, the fact that the boundary is pieced together of hyperbolae pieces that are concave, when seen from outside A), we can see that the only critical question has to do with the behaviour of the line from $(1,1)$ to $(\frac{1}{1-\sqrt{1-\alpha}},0)$. The question is whether this line cuts the bounding curve or stays under it. Or, equivalently, whether this line runs over or under the tangent to the hyperbola $\frac{1+x_2\sqrt{1-\alpha}}{1+\sqrt{1-\alpha}}x_1 = \frac{1}{\alpha}$ at the point where it meets the x_1–axis. And the answer is that the line actually coincides with that tangent, could it be nicer? Anyway, it is an elementary exercise to differentiate $\frac{1+x_2\sqrt{1-\alpha}}{1+\sqrt{1-\alpha}}x_1$ with respect to x_1 and x_2, and check that the resulting gradient vector at the point $(x_1,x_2) = (\frac{1}{1-\sqrt{1-\alpha}},0)$ is orthogonal to the connecting line.

(7) Only trivialities remain to be handled now. We have not yet considered the case where, say, $b_1 \leq \frac{1}{\alpha} < b_2$. But in this case, we obviously have $P(X_2 = a_2) = \frac{b_2-1}{b_2-a_2} \geq 1-\alpha$, and since both points of this event belong to A, there is no problem here. Also, we have entirely ignored (also in the formulation and proof of the lemma) the degenerate case where one or both variables are deterministic and $= 1$. However, this is no problem at all (this can be checked directly, or considered a consequence of the closedness of A and the fact that these degenerate distributions are limits of two–point distributions concentrated on two points that come closer and closer to 1).

5. The case $n > 2$ and The Stringer Bound.

It is not obvious at all what the generalisation of proposition 2 to three or more dimensions should look like. It is not even obvious that there exists such a generalisation.

It is easy, however, to generalize the method referred to as the "$h(x)$–method" for $n = 2$ and use it for construction of acceptance regions or lower bounds that appear to be quite efficient. One of the more promising ideas goes as follows. We explain it here for $n = 3$, but the construction can easily be transferred to the case of arbitrary n.

If X_1, X_2 and X_3 are non–negative, independent with expectation 1, then the sequence

$$(M_1, M_2, M_3) = (X_1,\ X_1X_2,\ X_1X_2X_3)$$

is a martingale. For given constants c_1 and c_2, define the stopping time τ as the first t where $M_t \geq \frac{c_t}{\alpha}$ (setting $\tau = 3$ if this never happens). Then

we have (by the most trivial version of the optional sampling theorem) $\mathrm{E}M_\tau = 1$. In this way, we have constructed a random variable

$$M_\tau = \begin{cases} X_1 \text{ if } X_1 > \frac{c_1}{\alpha} \\ X_1X_2 \text{ if } X_1 \leq \frac{c_1}{\alpha} \text{ and } X_1X_2 > \frac{c_2}{\alpha} \\ X_1X_2X_3 \text{ if } X_1 \leq \frac{c_1}{\alpha} \text{ and } X_1X_2 \leq \frac{c_2}{\alpha} \end{cases}$$

This is not a symmetric function of X_1, X_2 and X_3, but if we symmetrize it by formation of the average $\bar{M}$ of the six variables M_τ that can be constructed in this way by ordering of the three variables in the six possible ways, we obtain a new variable $\bar{M}$ with expectation 1, for which (by Markov's inequality) the condition $\bar{M} \leq \frac{1}{\alpha}$ defines an acceptance region corresponding to a lower bound for the mean. The same construction for $n = 2$ (with $c_1 = 1$) results in the bound of example 3, which (as can be seen from the figures) is quite close to the best we can do there. Thus, we have reasons to expect that this construction will result in bounds that – in practice — are close to the best one can do.

However, from a mathematical point of view this is not satisfactory. We would like to have a unique bound which is optimal, in some sense. For $n = 2$, we have seen that this is possible when a further condition is imposed, namely the condition that the acceptance region should be contained in the square $x_1, x_2 \leq \frac{1}{1-\sqrt{1-\alpha}}$. Actually, this property can also be stated as follows.

Suppose we know that the two variables are identically distributed rescaled Bernoulli variables, i.e. $P(X_i = b) = 1 - P(X_i = 0) = p$, where the mean value then becomes $\mu = bp$. In this case, the problem of giving a lower confidence bound for the mean reduces to the standard problem of finding a lower confidence bound for the probability parameter in a binomial distribution. Indeed, if $\underline{\mu}(X_1, X_2)$ is a lower bound for the mean in this case, we have $P(\underline{\mu}(X_1, X_2) \leq bp) \geq 1 - \alpha$, which is equivalent to $P(\frac{1}{b}\underline{\mu}(X_1, X_2) \leq p) \geq 1 - \alpha$, which means that $\frac{1}{b}\underline{\mu}(X_1, X_2)$ is a lower confidence bound for the probability parameter p, which in turn means that $\frac{1}{b}\underline{\mu}(X_1, X_2)$ must be smaller than the usual exact lower confidence bound in this case.

Suppose that we observe $(X_1, X_2) = (0, b)$. Then, the above argument can be repeated as follows. The observed number of "successes" $X_i = b$ is 1. Hence, we have the observation 1 of a binomal variable with probability parameter $p = \frac{\mu}{b}$, binomial total $n = 2$. A lower $1 - \alpha$ confidence bound $\underline{p}$ for p in this case is given by $2\underline{p}(1 - \underline{p}) + \underline{p}^2 = 1 - \alpha$, or $(1 - \underline{p})^2 = 1 - \alpha$, i.e. $\underline{p} = 1 - \sqrt{1 - \alpha}$. Hence, we can say with $1 - \alpha$ confidence that $p \leq \underline{p} = 1 - \sqrt{1 - \alpha}$, or that $\mu = bp \leq b\underline{p} = b(1 - \sqrt{1 - \alpha})$. Which is exactly what we get when the acceptance region's intersection with the axes are the interval $[0, \frac{1}{1-\sqrt{1-\alpha}}]$. The similar property when

(b, b) is observed is that the point $(\frac{1}{\sqrt{\alpha}}, \frac{1}{\sqrt{\alpha}})$ lies on the boundary of A. The acceptance region A_{opt} has both properties (and so does, in fact, the region A_4).

It is a remarkable fact, that a general, non–parametric lower bound for the common expectation of two independent variables can be constructed in such a way that it coincides with the exact bound for the (somehow rather fundamental) parametric situation, where the variables are known to be i.i.d. rescaled Bernoulli variables. Intuitively, we would perhaps expect that knowledge about the distributional form would enable us to make sharper confidence statements. But the fact that there *are* bounds with this property strongly suggests that we should restrict our attention to such bounds. What proposition 2 says is that for $n = 2$ there is a unique uniformely sharpest bound among those that have this canonical property.

Thus, in our search for a sort of optimal solution in the case $n > 2$, it seems to be a good idea to take the corresponding property for n i.i.d. rescaled Bernoulli variables into account.

The Stringer Bound. A famous proposed solution to a very similar problem was given by Stringer (1963). The applied context here had to do with random sampling in auditing. Bickel (1992) translated the original explanation into a mathematical formula and a proper mathematical conjecture about i.i.d. variables on the unit interval. This conjecture has never been proved, but it is generally agreed (and confirmed by lots of numerical simulations) that this socalled Stringer's conjecture holds, and even that the resulting bound seems to be rather conservative, suggesting that sharper bounds can be found.

Stringer's bound was originally stated as a non–parametric *upper* confidence limit for the common mean of n variables on the unit interval, which were assumed to be independent and *identically* distributed. Now, it obviously doesn't matter whether we talk about upper or lower bounds here, it is just a matter of replacing X_i with $(1 - X_i)$, μ with $1 - \mu$ etc. Moreover, if we stick to lower bounds, it appears to be a rather useless and irrelevant assumption that the variables cannot exceed 1. It will not help us to improve the bounds, and conversely, any bound which is valid in general will obviously hold also in the special case where the variables are known to be ≤ 1. If we furthermore relax the condition of identical *distributions* to that of identical *expectations*, Stringer's idea fits nicely into our framework, where it can be explained as follows. Let $X_1, X_2, \ldots, X_n$ be independent, non–negative with common expectation μ and let $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$ denote the ordered observations as usual. For given non–negative coefficients $a_1, \ldots, a_n$, consider the function

$$\underline{\mu}(X_1, \ldots X_n) = a_1 X_{(1)} + a_2 X_{(2)} + \cdots + a_n X_{(n)} \, .$$

This function is obviously increasing in any of its arguments, symmetric and homogeneous. Thus, it is a candidate for a $1-\alpha$ lower confidence bound of a particularly simple form. The question is what the coefficients should be. The general problem of determining the set of $(a_1, \ldots, a_n)$ for which $\underline{\mu}$ is a $1-\alpha$ lower confidence bound has never been solved. But a set of *necessary* conditions, in terms of upper bounds for the coefficients, is easy to derive from the case where the X's are i.i.d. rescaled Bernoulli variables, and the maximal coefficients satisfying these conditions are generally believed to result in a lower bound, which is valid for *all* distributions with equal mean values. This is Stringer's conjecture, or The Stringer Bound, in a slightly generalized version (since the distributions are usually assumed to be identical).

The maximal coefficients can be computed as follows. Suppose that the X_i's are identically distributed with $P(X_i = b) = 1 - P(X_i = 0) = p$, $\mu = \mathrm{E}X_i = pb$. Let K denote the number of variables that take the value b. K is binomially distributed with parameter p, binomial total n. Since the ordered variables are 0 for the first K and b for the remaining last $n-K$, we have

$$\underline{\mu}(X_1, \ldots, X_n) = (a_{n-K+1} + \cdots + a_n)b\,.$$

Thus, the essential condition $P(\underline{\mu}(X_1, \ldots, X_n) \le \mu) \ge 1-\alpha$ can be written

$$P((a_{n-K+1} + \cdots + a_n)b \le pb) \ge 1-\alpha\,.$$

or

$$P(a_{n-K+1} + \cdots + a_n \le p) \ge 1-\alpha\,.$$

In other words, the condition is that $A(K) = a_{n-K+1} + \cdots + a_n$, the sum of the last K coefficients, is a lower $1-\alpha$ confidence bound for the probability parameter p. Or, equivalently, the tail sum $A(k)$ must be less than or equal to the usual exact lower $1-\alpha$ confidence bound for the probability parameter in a binomial distribution where k successes out of n trials is observed. Moreover, by setting the tail sums *equal* to these lower confidence bounds, we obviously get the sharpest (pointwise maximal) bound of this form. Thus, we should define $A(k)$, $k = 1, \ldots, n$ as the solutions to

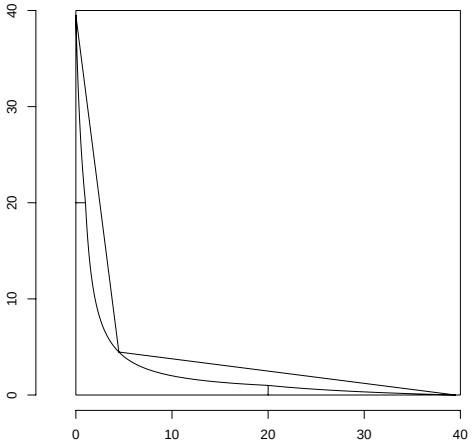
$$\sum_{j=k}^{n} \binom{n}{j} A(k)^j (1 - A(k))^{n-j} = \alpha$$

and then define $a_n = A(1)$ and $a_j = A(n-j+1) - A(n-j)$ for $j = 1, \ldots, n-1$.

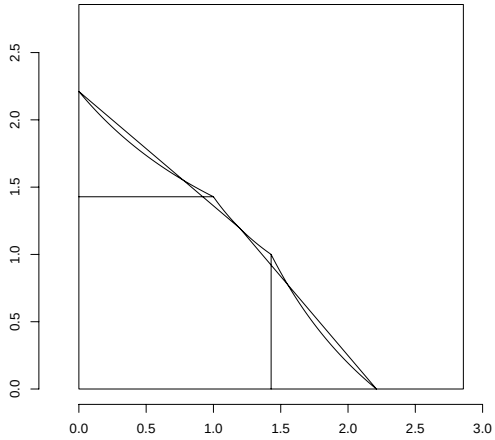
For $n = 2$ this results in the lower bound

$$\mu_{\mathrm{Str}}(X_1, X_2) = \sqrt{\alpha} X_{(1)} + (1 - \sqrt{1-\alpha})(X_{(2)} - X_{(1)})\,.$$

The corresponding acceptance region in spe A_{Str} is bounded by the two lines that connect the diagonal point $(\frac{1}{\sqrt{\alpha}}, \frac{1}{\sqrt{\alpha}})$ with the two points $(\frac{1}{1-\sqrt{1-\alpha}}, 0)$ and $(0, \frac{1}{1-\sqrt{1-\alpha}})$ on the axes. The first figure below shows this region together with the region determined by proposition 2 (or the bound from example 4, you wouldn't be able to tell the difference from the figure). As we can see, $A_{\text{opt}} \subseteq A_{\text{Str}}$, which means that **we have actually proved the Stringer bound for** $n = 2$ **!** However, the figure to the right shows the same regions for $\alpha = 0.7$, and here we don't have the inclusion. A more careful analysis shows that $A_{\text{opt}} \subseteq A_{\text{Str}}$ if and only if $\alpha \leq 0.5$.



Stringer B., $\alpha = 0.05$



Stringer B., $\alpha = 0.70$

Hence, we have only proved Stringer's conjecture for $n = 2$ and $\alpha \leq 0.5$. The last restriction is no problem in practice, where values of α greater than 0.1 are hardly of any relevance. But — talking about practice — we must admit that the case $n = 2$ is completely irrelevant. If we have proved anything interesting at all, it is that Stringer's conjecture is really a tough problem, and that the generalisation of this proof to higher values of n is not likely to be easy. We don't even know what A_{opt} should be when $n > 2$.

Notice also that we have not even proved that the original version of Stringer's conjecture is *false* for $\alpha > 0.5$. It follows from the way A_{opt} was constructed that there exists a pair of two–point distributions distributions with mean 1 for which $P((X_1, X_2) \in A_{\text{Str}}) < 1 - \alpha$ when, say, $\alpha = 0.7$. But it is not obvious how to construct an example where X_1 and X_2 are *identically* distributed, as they were in the original statements of Stringer's conjecture. However, Pap and Zuijlen (1995) have found examples for larger values of n where the Stringer bound does not hold for certain values of $\alpha > 0.5$, even in the case where the observations are uniformely distributed on the unit interval.

6. A conjecture about Bernoulli variables and geometric variables.

A reasonable condition to impose on the general solution seems to be that the lower confidence bound should — like the Stringer bound — take the maximal values possible in the case where the variables are i.i.d. rescaled Bernoulli variables. Thus, if the observed Xes take only the values 0 and b, we would expect that

$$\underline{\mu}(X_1, \ldots, X_n) = \underline{\mu}(0, 0, \ldots, 0, b, b, \ldots, b) = bA(k)$$

where k is the number of observations in the sample with the value b, and $A(k)$ is the exact lower $1 - \alpha$ confidence bound for the probability parameter in a binomial distribution where k successes out of n trials is observed. Clearly, this is the solution we would choose if we *knew* that the observations were i.i.d. rescaled Bernoulli, in which case the whole problem reduces to ordinary inference about the parameter p in a binomial distribution.

An immediate question, and perhaps a necessary first step if we want to solve the general problem, is "what should the solution be if the variables are known to be rescaled Bernoulli variables, but not necessarily identically distributed?"

I think I know the answer to this question. To support intuition, rephrase the problem as follows. Rather than thinking of lower confidence bounds for the common expectation μ, we think of it as a test problem, where the hypothesis is that $\mu = 1$, and the alternative is $\mu > 1$. In the case where the variables are known to be rescaled Bernoulli, our problem is equivalent to the following peculiar statistical problem, stated in terms of the Bernoulli variables before the rescaling.

Suppose that we observe Bernoulli variables $I_1, \ldots, I_n$ with probability parameters $p_j = P(I_j = 1)$. These parameters are unknown to us, but whenever a "success" $I_j = 1$ is observed, we are told what the corresponding p_j is (namely $1/X_j$). However, when $I_j = 0$ we are not informed about this. The question is: When can we say, on the basis of these informations, that "this can not be true", in the sense that we have far too many 1's with far too small p_j's? Surely, if $n = 100$ and we observe 50 successes, all with $p < 0.0001$, it must be pretty safe to conclude that the model is misspecified. Whatever the remaining 50 unknown p's are, I can't believe that this would happen in my lifetime. But how can this argument be formalized?

My answer to this is

Conjecture 1. *For independent geometric variables $G_1, \ldots, G_n$ with parameters $p_1, \ldots, p_n$ (i.e. $P(G_j = g) = p_j(1 - p_j)^g$ for $g = 0, 1, 2, \ldots$)*

define, for any $(i_1, \ldots, i_n) \in \{0,1\}^n$

$$g(i_1, \ldots, i_n) = P\left(\sum_{j=1}^{n} i_j(1+G_j) \leq n\right).$$

Then, for independent Bernoulli variables $I_1, \ldots, I_n$ *with parameters* $P(I_j = 1) = p_j$*, the random variable* $g(I_1, \ldots, I_n)$ *is superrectangular.*

REMARKS.

The random variables $G_1, \ldots, G_n$ are only used in the definition of the function g. They should not be regarded as random variables that are related to $I_1, \ldots, I_n$.

By a *superrectangular* random variable Z on the unit interval we mean a variable for which $P(Z \leq z) \leq z$. Or, equivalently, a variable that dominates in stochastic order a variable which is uniformely distributed on the unit interval. This is the property that you usually expect (sometimes only approximately) from a P–value when you test a hypothesis that is true. In the present context it means that if we decide to "reject" the model specification when $g(I_1, \ldots, I_n) \leq \alpha$, then the probabilty of rejection when the model specification is actually correct does not exceed α.

For some of the examples considered in the following, the (discrete) superrectangular random variable Z is such that $P(Z \leq z) = z$ in all points Z with $P(Z = z) > 0$, i.e. in all points where the c.d.f. makes a jump. Meaning that Z is somehow as close as possible to being uniformely distributed, given that it is stochastically larger than a uniform variable and has its mass concentrated in atoms of given sizes. As a simple example, take the distribution which assigns probability $\frac{1}{n}$ to each of the points $\frac{1}{n}, \frac{2}{n}, \ldots, \frac{n-1}{n}, 1$. In this case we say, briefly, that the property of superrectangularity holds "with equality".

For $p_1 = \cdots = p_n = p$ the conjecture holds with equality in this sense, and is just the wellknown connection between the tail probabilities of the binomial and the negative binomial distributions. This means that the proposed test procedure (or the corresponding unit acceptance region, or the corresponding lower bound for the common expectation) has the property emphasized earlier that it produces the obvious result if the observations are known to be i.i.d. rescaled Bernoulli. Also, if some of the probabilites are 1 and the rest equals a common p, the same argument holds.

Unfortunately I can not prove this result (except for $n = 2$, where it follows easily from some of the bounds in section 3, and for $n = 3$ where I once managed to work out a very tedious proof). However, I am pretty convinced that the conjecture holds, because I have verified

it by thousands and thousands of computer simulations, where for randomly chosen p's both sides of the inequality are computed for suitable values of α (only, however, with $n \leq 15$). I have also made lots of independent checks by another computer programs that draws the c.d.f. of $g(I_1, \ldots, I_n)$ for a randomly selected example and checks that it actually stays under the diagonal. Moreover, I have the following intuitive argument.

Suppose that the n observations are ordered by increasing values of the parameters, $p_1 \leq p_2 \leq \cdots \leq p_n$. This assumption is OK, provided that our criterion for "rejection" does not make use of the ordering. We observe the n outcomes $I_1, \ldots, I_n$ of the Bernoulli variables one by one. When $I_j = 1$ is observed, we are told in addition what the value of the "generating probability" p_j was, but for $I_j = 0$ we are not. Now, the first time we observe $I_j = 1$, say in the j_1'th trial, we can interpret j_1 as the value of a waiting time which is stochastically greater than $1 + G(p_{j_1})$, where $G(p)$ in general denotes a geometric variable with parameter p. Indeed, if we had $p_1 = p_2 = \cdots = p_{j_1}$, it would be true, in some sense, that we have observed $1 + G(p_{j_1}) = j_1$; if the preceding p's are smaller, we have observed a waiting time which is even "more extremely short" in its distribution. Next time we get a success, say $I_{j_2} = 1$, we can similarly argue that we have observed the value $j_2 - j_1$ of a waiting time which is stochastically greater than $1 + G(p_{j_2})$. And so on and so on, until no further successes are observed. However, when summarizing all this we cannot refer to the actual values $j_1, j_2, \ldots$ because they depend strongly on the ordering of the p's. What is true, however, is that the *sum* of all the waiting times we have observed is $\leq n$, and this sum is, in some sense, stochastically greater than

$$(1 + G(p_{j_1})) + \cdots + (1 + G(p_{j_k}))$$

where $p_{j_1}, \ldots, p_{j_k}$ are the probabilities of the successes that we actually observed. Thus if we find, after the experiment has been carried out, that (say)

$$P\left((1 + G(p_{j_1})) + \cdots + (1 + G(p_{j_k})) \leq n\right) \leq 0.0001$$

then we must conclude that we have observed far too many successes with far to small p_j. What happened here should usually only happen at most 1 out of 10000 times, and therefore we "reject" the model with a P–value ≤ 0.0001, in favour of the vaguely defined alternative that some or all of the p_j must be greater than we were told.

It is tempting to look for a proof based on some sort of coupling argument related to this idea. However, I have not been able to do this.

Conjecture 1 is (if it is true) a somewhat controversial result, with some interesting consequences that have little to do with the main topic of this paper. Some of these follow here.

An alternative way of expressing the same thing in terms of a Markov chain goes as follows. Consider a Markov chain with state space the set of c.d.f.'s on $\mathbf{Z}$, controlled by a sequence $I_1, I_2, \ldots$ of independent Bernoulli variables with parameters $P(I_t = 1) = p_t$ as follows. The initial state at time $t = 0$ is the c.d.f. of the one–point–mass at 0,

$$F_0(s) = \begin{cases} 0 & \text{for } s < 0 \\ 1 & \text{for } s \geq 0\,. \end{cases}$$

The state at time $t+1$ is given by I_{t+1} and the previous state F_t as

$$F_{t+1}(s) = \begin{cases} F_t(s+1) & \text{if } I_{t+1} = 0 \\ \sum_{i=0}^{+\infty} p_{t+1}(1-p_{t+1})^i F_t(s-i) & \text{if } I_{t+1} = 1\,. \end{cases}$$

In words, and in terms of distributions on the integers rather than their c.d.f.'s F_t, the event $I_t = 0$ translates the distribution one unit to the left, wheras $I_t = 1$ convolves it with a geometric distribution with the same probability parameter as I_t.

Then,

Conjecture 2. *For any t, the random variable $F_t(0)$ is superrectangular.*

There is also a continuous version of conjecture 1, which comes out as a limiting case when the probabilities p_j become smaller and smaller while $n \to \infty$. I have not worked out the details, but I am pretty sure that the two versions are equivalent (go from discrete to continuous by a limiting argument, the other way by interpretation of geometric variables as rounded exponential variables). The continuous version goes as follows.

Conjecture 3. *Let $\lambda > 0$ and $\mu > 0$ be given. For nonnegative real numbers $b_1, \ldots, b_k$ define*

$$g(k; b_1, \ldots, b_k) = P\left(b_1 V_1 + \cdots + b_k V_k \leq \lambda\mu\right).$$

where $V_1, V_2, \ldots$ are independent normalized exponential variables. Then, if K is Poisson distributed with parameter λ and $B_1, B_2, \ldots$ are i.i.d. nonnegative with $EB_j = \mu$ and independent of K, the random variable $g(K; B_1, \ldots, B_K)$ is superrectangular.

Remarks. Notice that if the B's are degenerate (all equal to μ with probability 1), the conjecture holds with equality, by the usual construction of a Poisson variable from a sequence of exponential waiting times.

Also, if the common distribution of the B's is a two–point distribution with one of its atoms in 0, the conjecture holds with equality, as an easy consequence of the fact that the number of non–zero B becomes Poisson with parameter $\lambda P(B_i > 0)$.

The difficult step, which I have been working on without any success for quite a long time, is obviously to prove the result when the common distribution of the B's is discrete. Here, the conjecture can also be stated as follows (with the notation a_i for the possible values of the B's, $\frac{\lambda_i}{\lambda_1+\cdots+\lambda_m}$ for the corresponding point probabilites).

Conjecture 3a. *Let $\lambda_1, \ldots, \lambda_m > 0$ be given. For fixed nonnegative real numbers $a_1, \ldots, a_m$ define a function of m nonnegative integer arguments by*

$$f(k_1, \ldots, k_m) =$$

$$P\big(a_1(V_1^{(1)} + \cdots + V_{k_1}^{(1)}) + \cdots + a_m(V_1^{(m)} + \cdots + V_{k_m}^{(m)})$$

$$\leq a_1\lambda_1 + \cdots + a_m\lambda_m\big)$$

where all the V's involved are i.i.d. normalized exponential. Then, for $K_1, \ldots, K_m$ independent Poisson with parameters $\lambda_1, \ldots, \lambda_m$, the random variable $f(K_1, \ldots, K_m)$ is superrectangular.

Even for $m = 2$, both atoms $a_i > 0$, I have no idea how to prove this.

References.

Bickel, P. J. (1992)
Inference and Auditing: The Stringer Bound.
International Statistical Review **60** 2, pp. 197–209.

Pap, G. and van Zuijlen, M. C. A. (1995)
The Stringer Bound in Case of Uniform Taintings.
Computers Math. Applic. **29**, 10, pp. 51–59.

Samuels, S. M. (1969)
The Markov Inequality for Sums of Independent Random Variables
The Annals of Mathematical Statistics **40**, 6, pp. 1980–1984.

Stringer, K. W. (1963)
Practical aspects of statistical sampling in auditing
Proceedings of Business and Economic Statistics Section, American Statistical Association.