Hurricane Clustering

Jack Devanney
Center for Tankship Excellence, USA, djw1@c4tx.org

Abstract

This paper takes a Bayesian view of hurricane occurrence. After some preliminaries, the interval between storms is assumed to be distributed according to a Gamma density with both parameters unknown. We begin with a non-informative, conjugate prior and compute the posterior on these parameters and the storm interval based on the storm record for the Florida Keys since 1900. This data is strongly over-dispersed relative to the Poission. In particular, this means the commonly held view that, since we have not had a storm in a long time, we are now due is incorrect. According to the model, immediately after the last Category 3 or better storm (Andrew) hit the Keys, the mean time to the next Cat 3+ storm for the Keys was 8.8 years. Now after 17 years without a Cat 3 storm, the mean time to the next storm is 14.8 years. The flip side is that this clustering has important negative implications for storm relief planning and insurance.

1 Poisson Sampling

1.1 Are we due for a hurricane?

In the Florida Keys where I live, the conversation often turns to hurricanes. You'll hear statements like "We haven't had a big hurricane in a long time. We're due." clearly implying that the fact that we have NOT observed a hurricane in a long time increases the probability of a hurricane in the near future. Heads nod solemnly all around.

Does this make sense? And by the way what is our probability we will have a big storm this year?

Let's begin with the simplest possible model. We will assume that Nature produces Keys bound hurricanes at a constant rate, λ. This rate in storms per year doesn't change just because she just produced a storm. She doesn't take a rest. Nor does she increase the rate, just because it has been a long time since she's hit the Keys. This is a big, erroneous assumption to which we will return. For now let's go with it, and see where it takes us.

If Nature is generating Keys Bound storms at a constant rate λ storms per year, then we have the familiar Poission process and the conditional probability of a r storms in T years *given* λ is

$$Pr(r|T,\lambda) = \frac{e^{\lambda T}(\lambda T)^r}{r!}$$

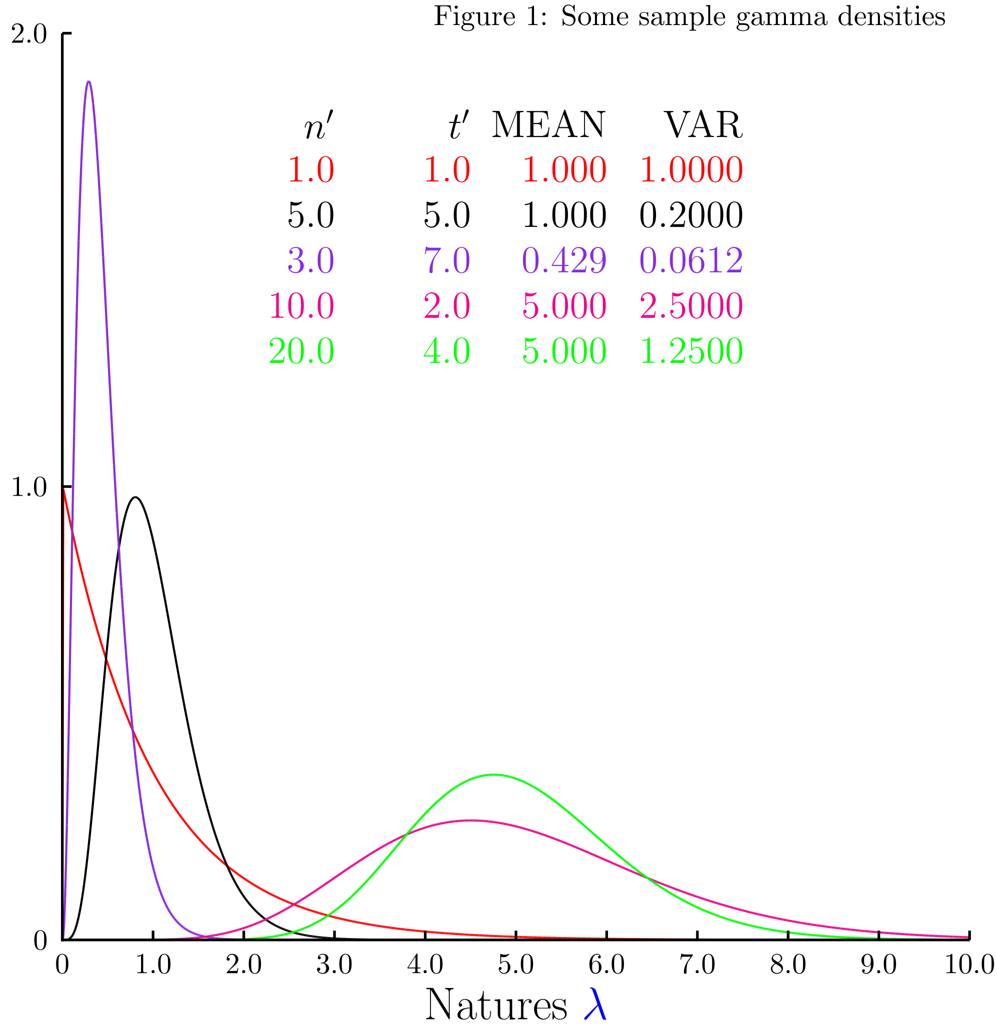
Of course, we don't know what Nature's λ is for the Keys. Therefore, we adopt a Bayesian view, specify a prior density on the unknown λ, and update this prior based on the data we have using Bayes Rule.

1.2 The Gamma Family of Priors

Before seeing any storm data, about all we can say about λ is that it is non-negative. The Gamma family is a rich set of densities on the non-negative half-line. The Gamma is a two parameter family and by varying these two parameters just about any smooth, unimodal density on λ greater than zero can be approximated. A few of these densities are shown in in Figure 1.

We will call the prior parameters, n' and t'. which, as we shall see, we can loosely think of as encapsulating our prior feelings in equivalent number of imagined storms n' in an imagined t' years.

Figure 1: Some sample gamma densities



You may think of the red density as our prior if our past experience is equivalent to seeing 1 storm in 1 year. The black density is a prior that is equivalent to seeing 5 storms in 5 years. The green density is equivalent to 20 storms in 4 years. And so on. The average or *mean* of a Gamma is the ratio of the two parameters, n'/t'. The larger n' is for a given mean, the tighter the probability is grouped around the mean.[1]

Despite the way Figure 1 is drawn, the densities don't stop at a λ of 10.0. They go on forever; but, for these five densities, the probability of λ being above 10.0 is too small to show up on the graph.

It turns out that, if we choose a Gamma prior with parameters n' and t', and then actually experience n real storms in t real years, the posterior density according to Bayes Rule is also a Gamma and it is the Gamma whose parameters are $n'' = n' + n$ and $t'' = t' + t$. ***Updating our feelings about Nature's λ after seeing some data is reduced to two simple additions.*** Our past experience and new data are combined in the simplest and most natural way imaginable.

[1] The variance of a Gamma is n'/t'^2. The Gamma mean and variance equations can easily be flipped to yield $n' = MEAN^2/VAR$ and $t' = MEAN/VAR$. The mode is $(n' - 1)/t'$ for $n' \geq 1$.

Priors that have this remarkable facilty are called *conjugate* priors. The Gamma is the conjugate prior for the Poisson process.

1.3 Non-informative Conjugate Priors

If n' and t' both get very large, then the Gamma converges toward a spike at n'/t'. *If we knew that Nature's was creating storms at a constant rate* and we had sample of 5,000 storms in 100,000 years, then we would be nearly certain that λ was very close to 0.05. Experiencing another storm in say another 2 years would change our ideas about λ hardly at all. This is reflected in the two simple sums. There is almost no difference between the Gamma with parameters (5,000, 100,000) and the Gamma with parameters (5,001, 100,002).

Conversely, as n' and t' both become very small, the Gamma becomes a very spread-out density. If n' and t' are both less than 1 (say 0.005 and 0.01), then it would take very little data to dominate the posterior. In this case if we experienced just one storm in 2 years, our posterior Gamma would have parameters (1.005, and 2.01) which parameters are made up almost entirely of the data.

The obvious extension of this thinking is a Gamma whose parameters are 0.0 and 0.0.[2] This is called the *non-informative* prior. It gives no weight to our prior feelings before seeing any data. In problems such as hurricane occurrence this is probably what we want to do.

1.4 Hurricane posteriors

Back to hurricanes. Table 1 is a list of all Category 3 and higher storms that have hit the Florida Keys since 1900, according to the National Hurricane Center.[3]

Table 1: Category 3 or Higher Storms that have hit Florida Keys

YEAR	CAT	NAME	DATE	INTERVAL	MIN PRES
1906	3		Oct 18		953
1909	3		Oct 11	2.98	957
1919	3		Sep 9	9.91	
1926	3		Oct 20	7.11	
1933	3		Oct 5	6.96	
1935	5	Labor Day	Sep 3	1.91	892
1935	3		Sep 28	0.07	
1948	3		Sep 21	12.98	963
1948	4		Oct 5	0.04	
1960	4	Donna	Sep 9	11.93	932
1965	3	Betsy	Sep 8	5.00	952
1992	4	Andrew	Aug 24	26.96	937

According to this table, the Keys have seen 12 Cat 3 or more storms in 109 years, 4 of which were Cat 4 or worse, and one was a Cat 5, the famous Labor Day Storm that killed about 500, mostly due to stupidity.

If we start out with a Gamma prior with $n' = 0$ and $t' = 0$ and observe 12 big storms in 109 years, our posterior is a Gamma with parameters $0 + 12$ and $0 + 109$. This density is sketched in Figure 2 together with the similar densities for Cat 4 or better and Cat 5. The first thing that jumps out at you is that the density on the Cat 4 or worse λ looks much tighter than the density

[2]The Gamma with parameters 0.0 and 0.0 does not actually exist; but, as long as we have some data, this is not a practical problem, for the posterior will be a perfectly good Gamma.

[3]http://maps.csc.noaa.gov/hurricanes/viewer.html. Some judgement has been used in deciding how close a storm had to some to the Keys to be called a hit.

on Cat 3 or worse λ. And the density on the Cat 5 λ looks tighter still.[4] However, ***relative to the mean***, the Cat 5 λ density is actually more spread out than the Cat 4 which is more spread out than the Cat 3. The best way to see this is to compare the ratio of the mean to the square root of the variance (known as the standard deviation) as we have done in Figure 2.

1.5 Yeah, but what about this fall

Of course, our real interest is not in Nature's λ but whether or not we are going to have a big storm this fall. More generally, given our current state of knowledge n'' and t'', we'd like to know what is the probability of N storms in the next T years.

To determine this, we need to to multiply Nature's probability of a storm given her λ times our probability of that λ and sum over all possible λ. In this case, after you grind through all the algebra, it turns out that our density on N storms in the next T years is something called the *negative binomial*.[5]

$$P(N|T, n'', t'') = f_{nb}(N|n'', t'', T)$$

Figure 3 shows the results for Cat3+, Cat4+, and Cat 5 given our data and a single year ($T = 1$). Most of the probability is at zero.

If you are thinking about buying a house in the Keys, then you might want to take a longer term view. Figure 4 shows the results for a 25 year time horizon, that is, $T = 25$. These densities are considerably less reassuring.[6]

If you are really paranoid or perhaps thinking about selling hurricane insurance in the Keys, then you might want to take a 50 year view. Figure 5 shows the results for a 50 year time horizon. There is of course a big difference between a storm hitting the Keys and damage to a particular location in the Keys. The Keys are over 100 miles long, and the swath of really major damage from even a Cat 5 storm is generally much narrower than this. These storm densities would be just the starting point for a complete study of insurance pricing.

However, before we get too carried away with ourselves, let's go back to the one year density. Almost all the probability in the one year graph, Figure 3, is at $N = 0$, so it is a little more useful to put the results in tabular form, as we've done in Table 2.

Table 2: Probability of N Storms in one year

Number of Storms	Category 3 or better	Category 4 or better	Category 5
0	0.89620	0.96413	0.99091
1	0.09777	0.03506	0.00900
2	0.00578	0.00080	0.00008
3	0.00025	0.00002	
4	0.00001		

According to Table 2, the one year probability of one or more Cat 3 or better storms is just a little over 0.1, the probability of a Cat 4 or better storm is about 0.04, and the probability of a Cat 5 storm is about 0.01. But we have a problem.

According to this table, having two or more Cat 3+ storms in one year is quite unlikely, less than 1 in a hundred. About this time, alarm bells should be going off in the fire house. Looking

[4] The very spikey Cat 5 density does not show up well. The peak is 109 at an λ of zero. (For $n' = 1$, the Gamma simplifies to an exponential function.)

[5] The mean of the negative binomial is $n''T/t''$ and the variance is $(n''T/t'')((T + t'')/t'')$.

[6] Despite the way these figures are drawn, these densities don't stop at 5 or 10 storms. They go on forever, but eventually the probabilities becomes too small to show up on the graph.

back at Table 1, we see that this supposedly rare event has happened twice in our sample which only includes 12 storms.

Time to rethink our model of Nature.

Figure 2: Posterior on Nature's λ for data in Table 1

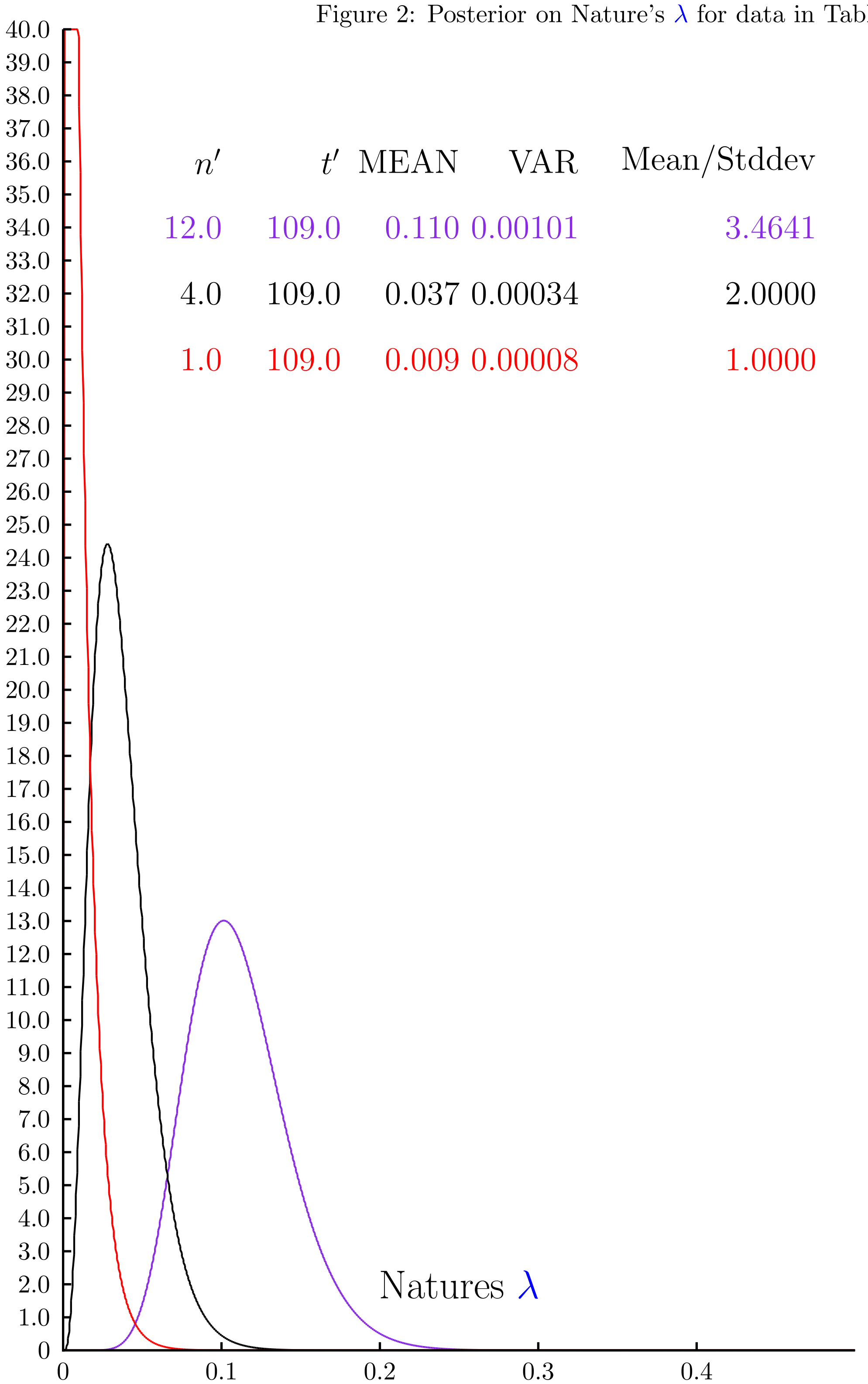


Figure 3: Our storm probabilities for this fall

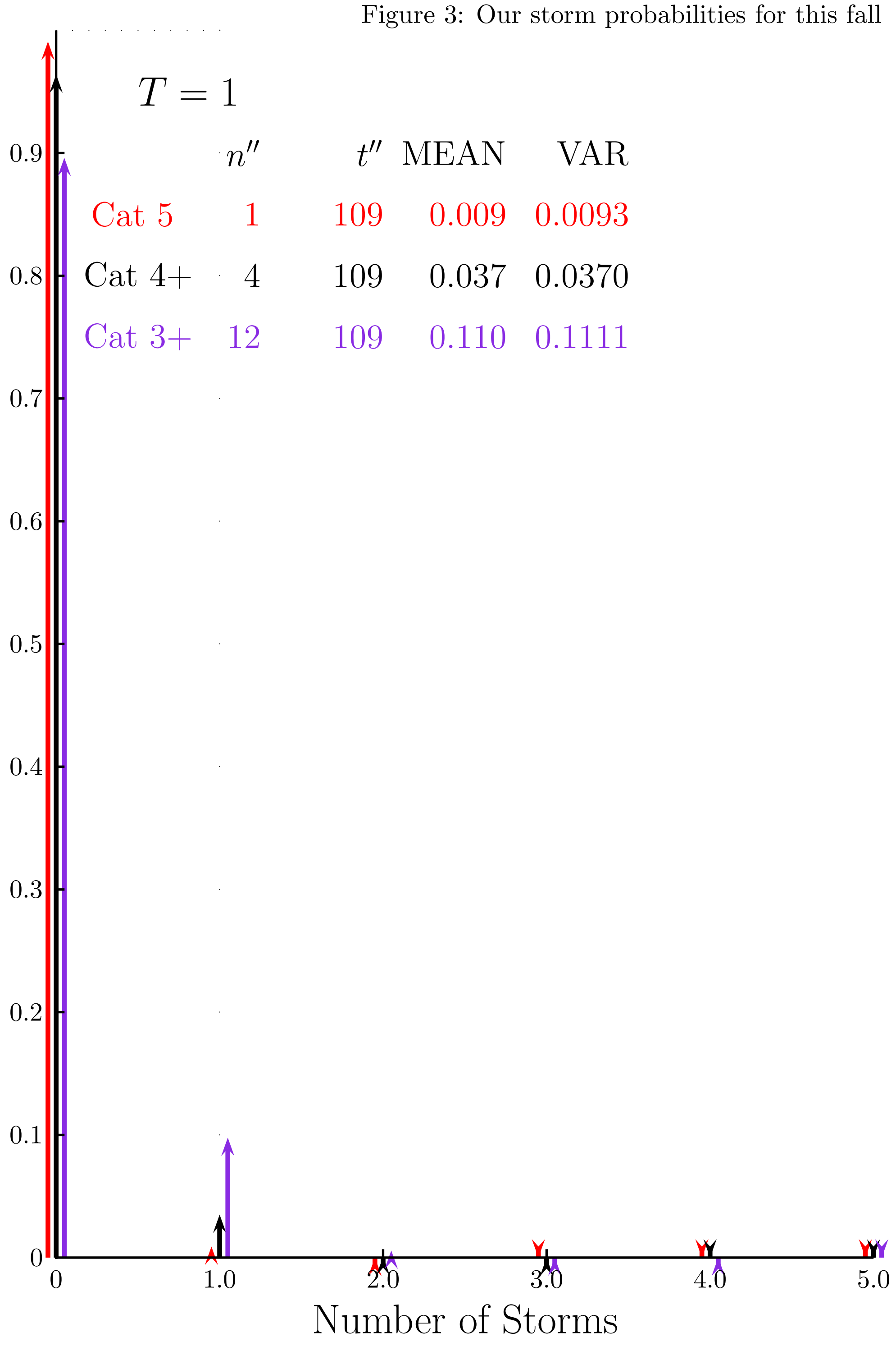


Figure 4: Our storm probabilities for the next 25 falls

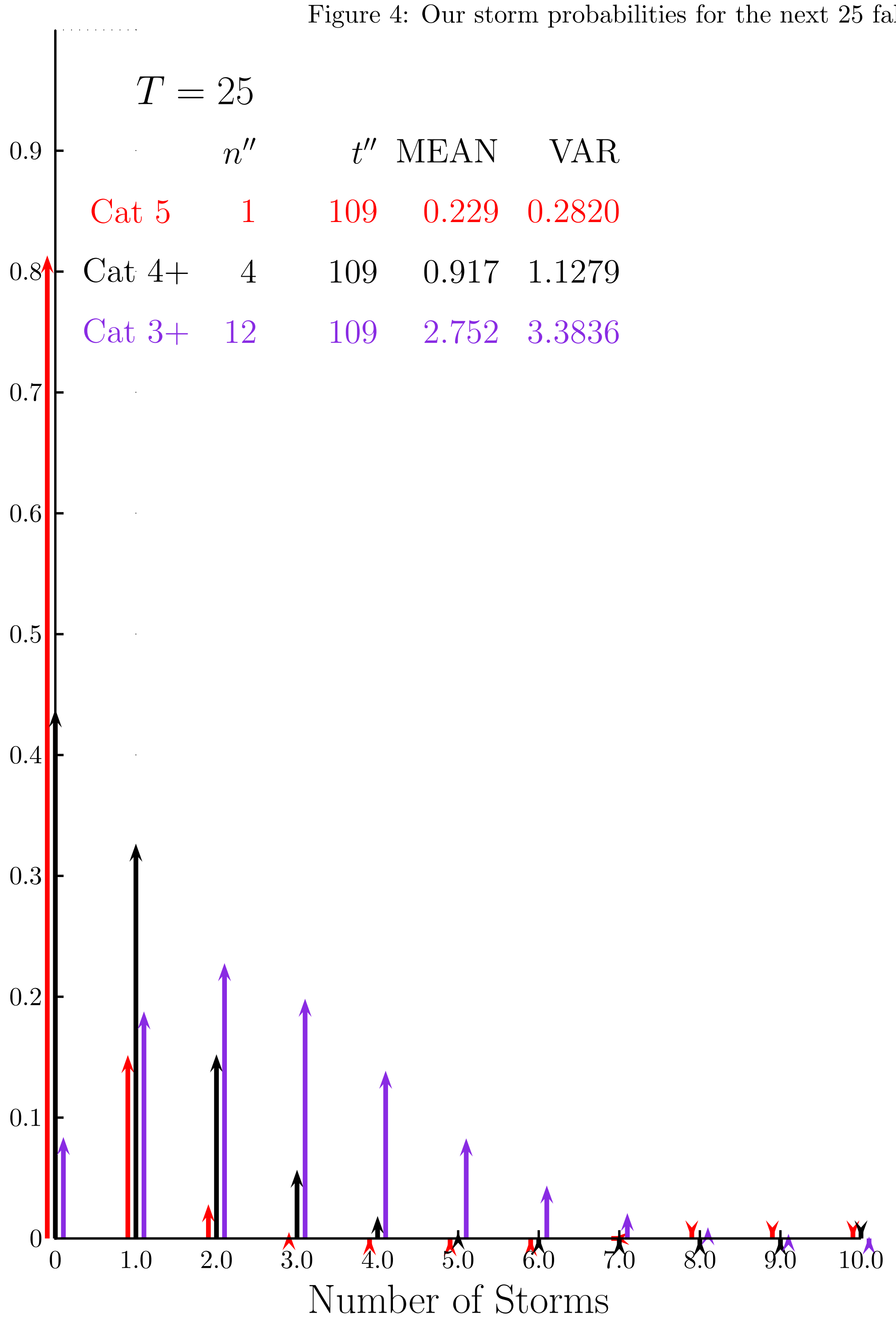
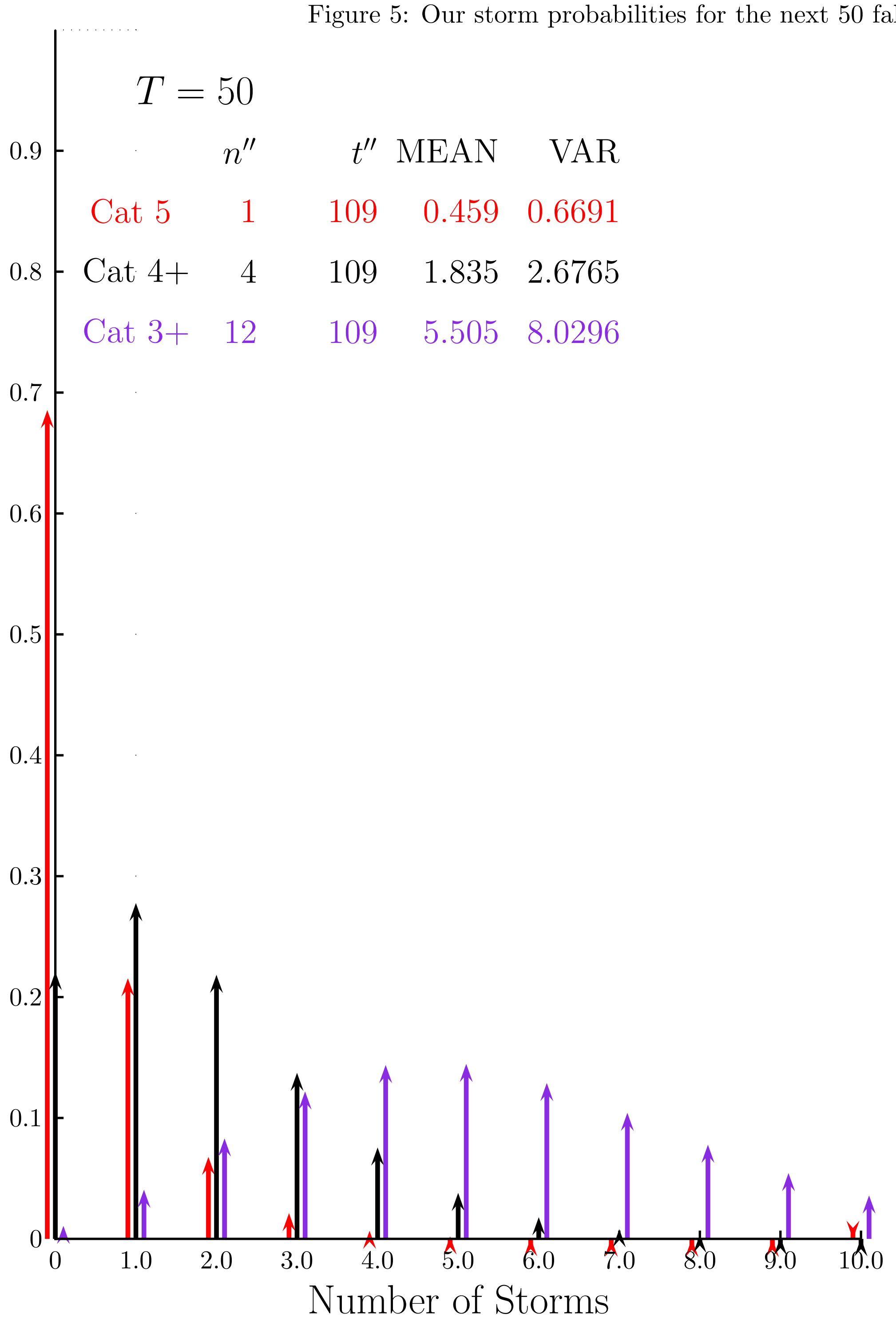


Figure 5: Our storm probabilities for the next 50 falls



2 Gamma Sampling

2.1 Giving Nature more freedom

It is not that our single rate model of storm generation is useless. I'd argue that we have obtained some insight about our storm risk from this simple model. But we have good reason for suspecting we can do better, so let's give it a try.

When we assumed in Section 1 that Nature produced storms at a constant rate regardless of whether or not she had recently produced a storm, we forced a very restrictive behavior on Mother Nature. In particular, this assumption means that Nature must pick each interval between storms y according to an Exponential density.

$$f_e(y|\lambda) = \lambda e^{-\lambda y}$$

where λ is Nature's unknown (to us) rate of storms per year.

The Exponential density has a remarkable and unique property. Suppose it has been z years since the last storm. We are interested in the remaining interval to the next storm. If and only if storm intervals are distributed according to an Exponential density, then according to Bayes Rule,

$$f(y|z,\lambda) = \frac{\lambda e^{-\lambda(y+z)}}{\int_z^\infty \lambda e^{-\lambda x} dx} = \frac{\lambda e^{-\lambda y} e^{-\lambda z}}{e^{-\lambda z}} = \lambda e^{-\lambda y}$$

Her probabilities haven't changed at all! With an Exponential, it does not matter how long it has been since the last storm.[7] Such a density is said to be *memoryless*. That is more or less what we meant when we assumed Nature produced storms at a constant rate regardless of whether or not she had recently produced a storm. We have seen from the Keys' storm data, there is good reason to suspect that Nature does not follow this rule. We must give Nature more freedom.

The Exponential density is a special case of our old friend, the Gamma density. The Gamma looks like:

$$f_\gamma(y|\nu,\tau) = \frac{e^{-\tau y}(\tau y)^{\nu-1}\tau}{(\nu-1)!}$$

As we have seen, the Gamma density is a two parameter family. But this time we are ***not*** going to use the Gamma as our prior, but rather ***as Nature's density by which she picks a storm interval*** y. Of course, we don't know Nature's parameters, so we give them Greek names and blue coloring. The interesting parameter is ν, which in this context we will call the *dispersion* parameter. If Nature's ν is 1.0, then the Gamma simplifies to an Exponential, and τ becomes Nature's storm generation rate. We will call τ, the *intensity* parameter.

As we have also seen, the mean of a Gamma with parameters ν and τ is ν/τ and the variance is ν/τ^2. If ν doubles, then in order to maintain the same mean, τ must also double, and the variance is halved. By increasing the dispersion parameter ν for a given mean, we can generate a density that is more and more tightly concentrated around the mean. In such situations, the fact that you have just had a storm decreases the probability that you will have a storm in the near-future.

Figure 2.1 shows a series of Gammas, all of which have a mean of 7.8 years which is the average interval in our sample of Cat 3 plus storms hitting the Keys. By making ν large enough, we can create a density that is just a spike at the mean. If storms were generated by such a density, they would come through with total regularity exactly every 7.8 years, like a perfectly scheduled train.

Of course, that's not what storms do. In fact, based on our data, the intervals appear to be over-dispersed relative to the Exponential, which is the black density in Figure 2.1. It looks like we are getting both more very short intervals and more very long intervals that we would expect if Nature was using a ν of 1.0. This over-dispersed situation is handled by ν's of less than 1.0. As Figure 2.1 shows, such densities put more probability at both the very short intervals and at

[7] The numerator in the first step says to have a storm y from now we need a total interval of $x = y + z$. The denominator as usual is the probability of the conditioning event, $x >= z$

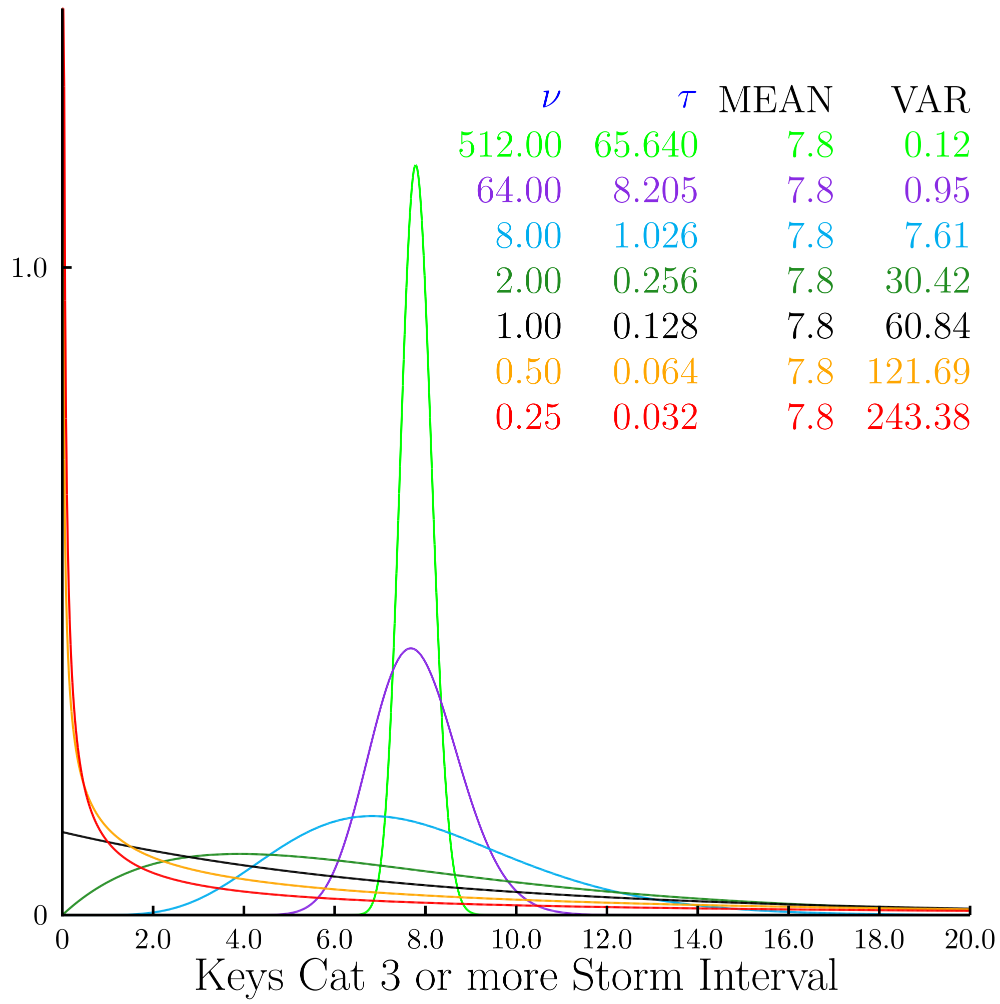


Figure 6: Some Possible Nature's Gammas with a mean of 7.8 years

the very long (although the latter is difficult to see in the figure) than the Exponential. In such situations, if you've just experienced a storm, the chances are higher than usual that you will shortly see another one. Conversely, if you haven't seen a storm in an unusually long time, your chances are better than normal.

By allowing Nature to use any Gamma to pick storm intervals, we create a far more robust model. A priori Nature could use a Gamma that produced storms with perfect regularity or she could use a Gamma that produces storms that are markedly over-dispersed relative to an Exponential, or anything in between, including if she likes an Exponential. By allowing her to choose any Gamma rather than just any Exponential, we are imposing far less constraint on her behavior. We have a model that can capture a much wider range of phenomena, a model that is less likely to lead us astray.

Of course, if Nature now has two parameters to play with, that means we Bayesians must somehow concoct a prior on both such parameters, and then combine this two-dimensional prior with the data according to Bayes Rule to produce both our posterior on Nature's ν and τ given the data, and our corresponding posterior probability for the next storm interval.[8]

The conjugate prior for this situation is a three parameter family called the *Gamma-hyperpoisson*. It turns out that, if before seeing any data, we start out with a non-informative Gamma-hyperpoisson on Nature's ν and τ, then we need to keep track only of the number of intervals in our sample of storms, n, the sum of the storm intervals, s, and the product of these intervals, p. And each time we experience another storm, we simply add one to n, add the interval since the last storm to s, and multiply p by this interval, and bingo! we have our updated parameters.[9] Bayes Rules has come through again.

Of course, as usual we are much more interested in the density on the next storm interval than we are on ν and τ. This is obtained by summing Nature's probability given ν and τ over all ν and τ weighted by our probability of each ν and τ. This calculation gets a little messy but with a modern computer it is not a big deal. The result is called a Stewart density. See Devanney and Stewart(1974) for the gory details.

When you input the Keys data, Table 1, for Cat 3 or better storms, to the Stewart density calculations, the result is shown in Figure 7. The red line is the *Stewart* density, $f_S(y|n, s, p)$ where n is the number of intervals, s is the sum of the intervals, and p is the product of the intervals. This density has been compared with an Exponential (the black line) with the same mean interval. It is a little difficult to see in this Figure, but, for this data, the Stewart puts considerably more probability at the very short intervals and more probability at the very long intervals than the Exponential. In this case, the cross-overs are at 0.48 years and at 12.1 years. The Exponential has more probability between these numbers. The Exponential has a variance of about 60; the Stewart has a variance of over 300.

In this case, we can gain more insight by looking at the cumulative distributions, Figure 8.

The Stewart cumulative rises much more quickly than the Exponential at short intervals, but then grows less quickly at very long intervals. As the little table in the figure shows, the result is that the Stewart probability of a very short interval is much higher than the Exponential. For the Stewart, the probability of an interval less than one-quarter year is three times that of the Exponential. But at the same time, the Stewart probability of a very long interval is also higher than the Exponential.

Figure 9 is a contour plot of our posterior on Nature's τ and ν after seeing the Key's storm data. It is a ridge whose axis is roughly along the $\nu = 7.8\tau$ line, especially for larger ν. The peak is at $\nu = 0.55$ and $\tau = 0.06$. The important point is that 89.5% of our probability lies below $\nu = 1$ reflecting the strongly over-dispersed nature of the data.[10]

[8]In this case, the this means summing over all possible combinations of Nature's ν and τ weighted by our current probability on these two unobservables.

[9]Since the product parameter can get very big, we normally work with the log of this product, in which case we merely add the log of the latest interval to the sum of the logs of the earlier intervals.

[10]It is interesting that the storm interval sample variance is only 54.8. The variance of an exponential with a mean of 7.8 would be 60.8. This might lead the unwary to conclude that the data is under-dispersed. But the variance is a very poor measure of dispersion for highly skewed densities such as this, and Bayes Rule is not fooled.

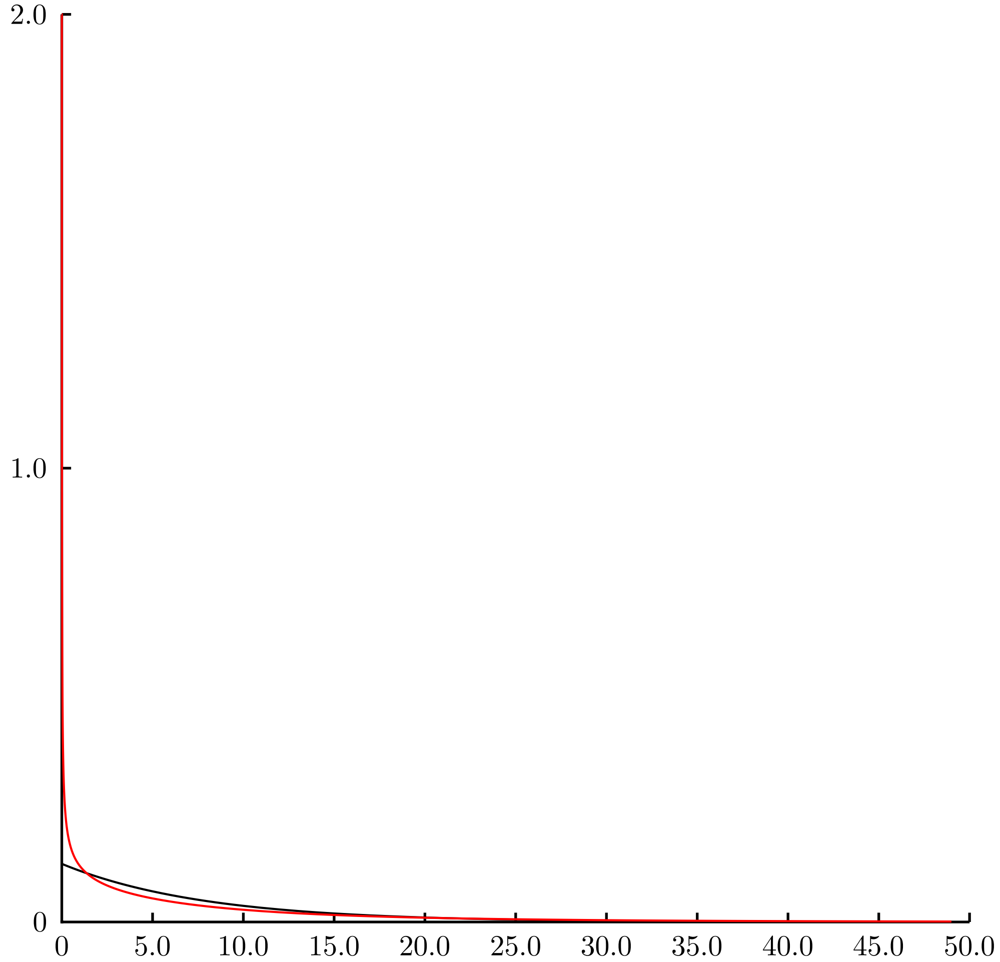


Figure 7: Stewart Density for Next Storm Interval

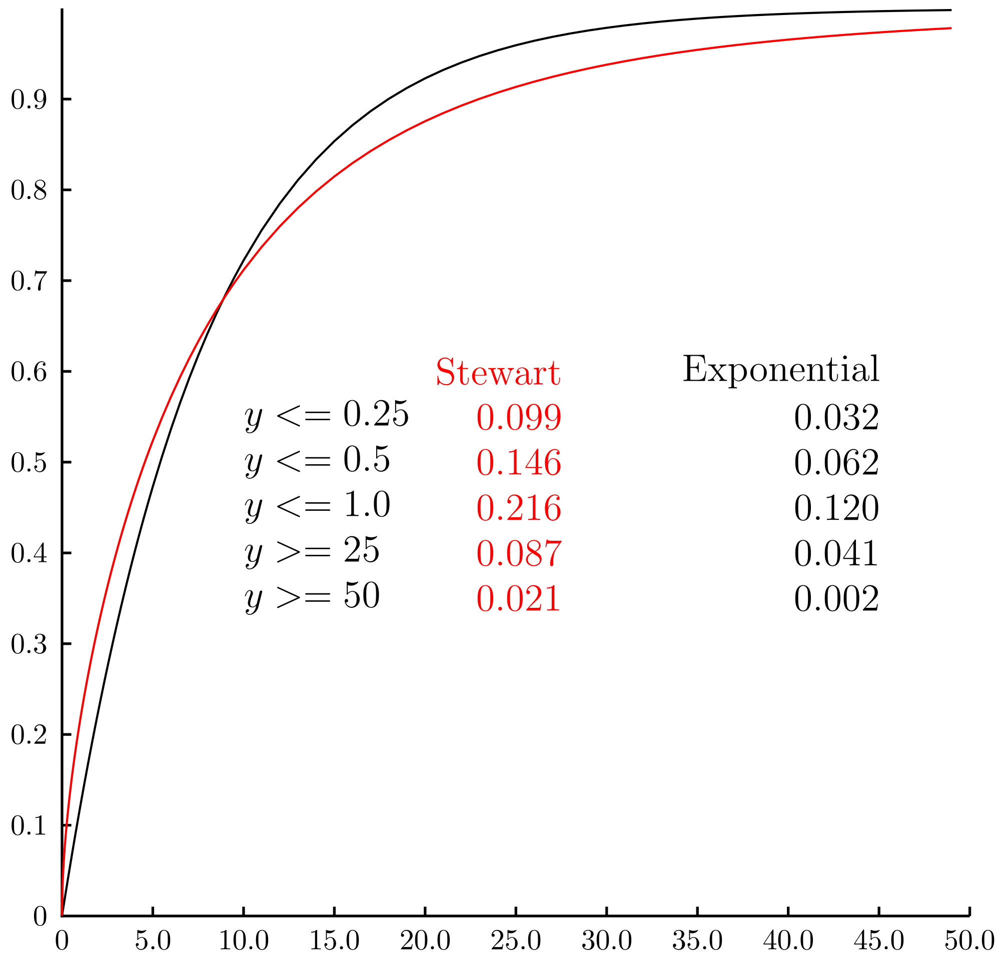


Figure 8: Cumulative Distributions for Next Storm Interval

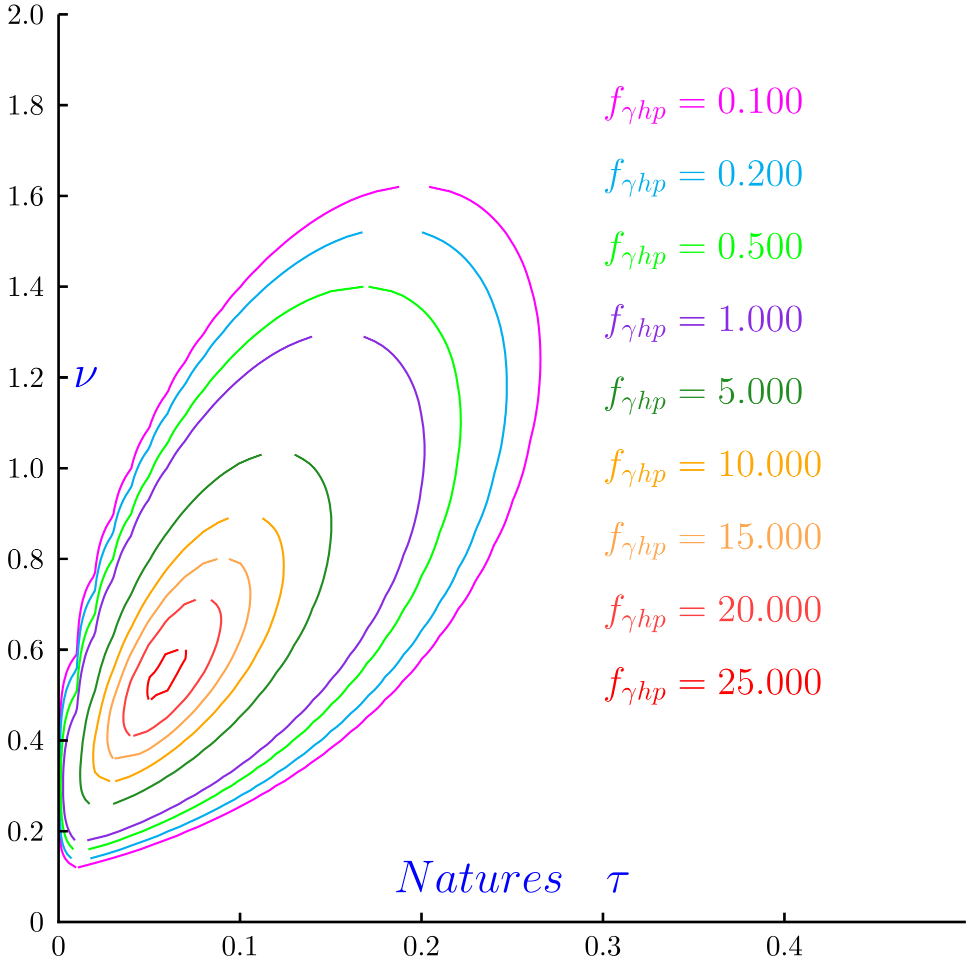


Figure 9: Our Posterior Density on Nature's Storm Parameters

This clustering of storms has important implications for storm relief and insurance, among others. Models that do not capture this clustering, such as our earlier single parameter model, will under-estimate the probability of agencies being forced to respond to multiple storms in a short period. Models that do not capture this clustering may confuse an unusually long interval with a change in the underlying process parameters, leading to unwarranted insurance premium reductions. Also clustering boosts the requirement for insurance reserves.

The Stewart density is anything but memoryless. In the case of the Keys, the last Cat 3 plus storm was Andrew in 1992, about 17 years ago as I write this. The density in Figure 7 is really our density on the next interval ***immediately after Andrew***. More precisely, it is the density on the ***total*** interval between Andrew and the next storm. However, we have gotten past the high probability short interval stage; so as I write this the relevent density is

$$f(y|z = 17, n, s, p) = \frac{f_S(y + z|n, s_p)}{\int_z^\infty f_S(y|n, s, p)dy}$$

The right side of this equation is based on the same argument that we used in showing the Exponential was memoryless. The density on the left side is called the *modified Stewart* $f_{mS}(y|z, n, s, p)$. Figure 10 plots the modified Stewart. for a number of z's given the Table 1 storm data.

The Reverend's Bayes Remarkable Rule has come up with a remarkable result: ***the longer that it has been since the last storm, the longer it will be on average to the next storm.*** Immediately after Andrew, the mean time to the next Cat 3+ storm was 8.8 years. 17 years later the mean *remaining* time to the next Cat3+ storm is 14.8 years. Thanks to clustering, the conventional wisdom that an unusually long storm-free period increases storm risk is not only wrong, but badly wrong. The answer to "It's been a long time since the last big storm, aren't we due?" is a resounding "***No***". The bad news is the opposite is true. Storms really do come in bunches. Of course, meterologists have known that hurricanes were clustered at least since 1780.[11] Gamma sampling is just one way of quantifying this phenomenon.

[11] In October of 1780, three of the most deadly Atlantic hurricanes in history occurred within a three week period. See Emmanuel, K. Divine Wind, pages 63-66. All three storms followed south to north tracks out of the Caribbean. On October 3rd, the first storm clobbered Jamaica from the south, headed north over eastern Cuba and NNE over eastern Bahamas. Emmanuel calls this the Savanna-la-Mar Hurricane, and puts the death toll at 3000, including the crews of four British naval ships. The Great Hurricane of 1780 hits Barbados on the 10th, turns NNW just to the west of the islands, clobbering each in turn; crosses the east end of Hispaniola on the 15th; and then north to Bermuda where it sank several ships. Emmnauel says 22,000 killed. Solano's hurricane hits the Caymans on the 16th, then the western tip of Cuba on the 18th, reaching the Florida Panhandle on the 22nd. This storm scattered Admiral Solano's fleet which had set out from Havana on the 16th to capture Pensacola. killing 2000 sailors/soldiers.

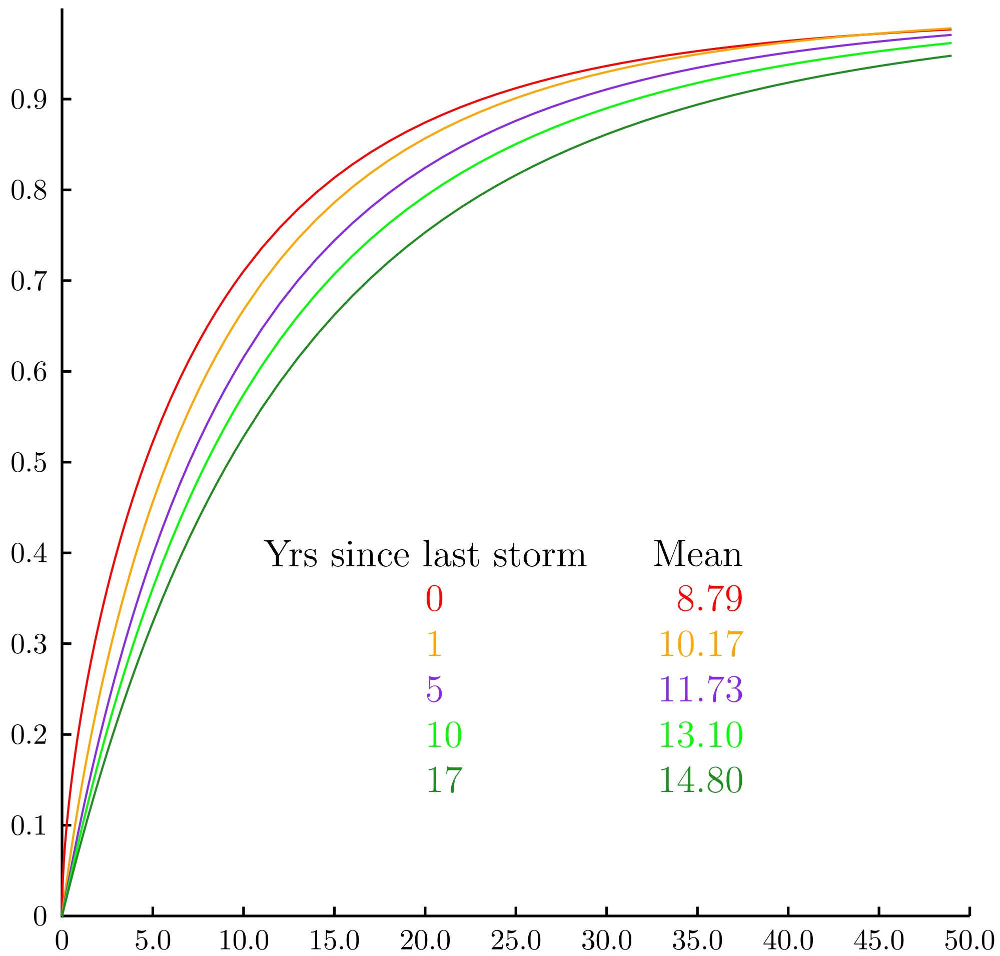


Figure 10: Cumulative Distributions for Remaining Interval to Next Storm

3 Conclusions and Future Work

Our initial single parameter storm model assumed that storms were generated according to a Poisson process. It is obvious that the two parameter, Gamma model, is much more robust, and far less likely to lead to misleading or incomplete results. Yet the Gamma sampling model has been used very rarely, while the Poisson process is ubiquitous in the literature.

The Poisson process is a reasonable starting point in many situations. But in almost all such situations, it is no more than a hypothesis, often introduced simply because the modeler knows how to do the numbers. In real world situations, where lives or large amounts of resources are at stake, such a hypothesis must be tested. One way of doing such testing is to expand to the Gamma process, and see if the posterior ν is grouped around 1.0. If not, the process is not Poisson. With packages available which will do the Gamma computations for you, there is no longer any excuse for not doing this test.

Having said this, the Gamma is certainly not a panacea. In the case of hurricanes, the obvious next question is: why are storm intervals so over-dispersed? The equally obvious answer is that the rate at which big storms are generated varies markedly from year to year. If for whatever reasons, conditions are unusually favorable for storm creation and growth — most importantly, high sea water temperature — then Nature will use a much higher storm generation rate than in seasons when conditions are unfavorable.

This suggests a two stage storm model:

1. In the first stage, Nature selects a λ_i just for the coming season. We can leverage the work we have already done by assuming that Nature's storm creation rates are distributed according to a Gamma, or, if the season is close enough, so that we have information on the conditions for this Fall, from a density that depends on those conditions.
2. In the second stage, Nature generates storms for this season by sampling from the density that resulted from the first stage.

Such multi-stage models are called hierarchical. Bayes rule is fully capable of handling such models. It is just a little more computational work. The CTX has a project underway to create such a hierarchical model.

References

Devanney, J (1974) and Stewart, R, "Bayesian Analysis of Oil Spill Statistics", Marine Technology, October 1974, p 365-382.

Emmanuel, K (2005), "Divine Wind, The History and Science of Hurricanes", Oxford University Press, 2005.